PEDECIBA Informática

Instituto de Computación - Facultad de Ingeniería

Universidad de la República

Montevideo - Uruguay

Evaluación de la Calidad de Datos en un Sistema de Data Warehousing:

Un enfoque basado en Contextos

Ing. Flavia Serra
fserra@fing.edu.uy
Tesis de Maestría

Supervisora:
Dra. Adriana Marotta
amarotta@fing.edu.uy

Tribunal:
Dr. Héctor Cancela
Dra. Verónika Peralta
Dr. Alejandro Vaisman

Diciembre 2015

A mi hijita, Mili, gracias por llegar a mi vida.
A Álvaro, gracias por darme fuerza y apoyo;
amor, sin tu ayuda esto hubiese sido imposible.

Agradecimientos

En primer lugar quiero agradecer a la Dra. Adriana Marotta, mi tutora, por todo su apoyo, tanto en el plano profesional como en el plano personal. Aprendí mucho bajo su supervisión. Gracias por su guía y por recordarme que siempre es necesario parar para "afilar el hacha". También quiero agradecer a los Porfesores Dr. Ismael Caballero y Dr. Alejandro Vaisman por todos los valiosos aportes que realizaron en distintas etapas de este trabajo. A todos mis compañeros del curso Fundamentos de Bases de Datos, por su ayuda y apoyo, que fue fundamental para poder finalizar esta tesis. Al Profesor Dr. Raúl Ruggia por su eterna paciencia, que tantas veces me permitió postergar mi trabajo para poder llevar a cabo esta tarea. Finalmente, no puedo dejar de agradecer a mi familia, que gracias a ellos ha sido posible llegar hasta aquí. A mis cuatro hermanos, por su inmenso apoyo al comenzar mi carrera. A mis padres, gracias por su ejemplo constante, por lo que soy. Porque cada día me ayudan para poder seguir adelante con mi trabajo. Esta tesis es para y por mis padres, a ellos mi eterno agradecimiento.

Resumen

Los Sistemas de *Data Warehousing* son de gran relevancia para el apoyo en la toma de decisiones y el análisis de los datos. Esto ha quedado demostrado a lo largo del tiempo, a través de la generalización de su desarrollo y uso a nivel industrial en todo tipo de organizaciones y mediante la gran cantidad de trabajos científicos que se han centrado en el estudio de este tipo de sistemas. Muchos investigadores han presentado la necesidad de incorporar y mantener la calidad de los datos en los Sistemas de *Data Warehousing*. Sin embargo, en las investigaciones no se encuentra un consenso acerca de cómo hacerlo, ni acerca de si es posible definir un único conjunto de dimensiones de calidad en el entorno de un *Data Warehouse*, dado que dicho conjunto puede depender del propósito con el cual se utilizan los datos. Por otro lado, una vez que los datos están en el *Data Warehouse* surge otro desafío, cómo serán utilizados los mismos. Los requerimientos de calidad pueden variar entre los diferentes dominios y entre los diferentes usuarios, no sólo por el propósito de la tarea que necesiten realizar, sino también porque la calidad percibida por un usuario puede diferir respecto a la calidad percibida por otro usuario. Dado que, los datos vienen de diversas fuentes con niveles de calidad distintos, los dominios de análisis pueden ser variados y los usuarios pueden percibir la calidad de distintas formas, dependiendo esto de múltiples factores (su perfil, la tarea que va a realizar, etc.). Para la evaluación de la Calidad de Datos en los Sistemas de *Data Warehousing*, se considera un enfoque basado en el Contexto de los datos.

En este trabajo se ejecuta una metodología de búsqueda bibliográfica para obtener una visión general de la investigación existente acerca del uso de contextos en los Sistemas de *Data Warehousing* y/o en la evaluación de Calidad de Datos. A partir de los resultados obtenidos con la aplicación de dicha metodología, se obtiene una visión general del estado del arte, lo que permite realizar el primer planteo de una propuesta para evaluar la Calidad de Datos en los Sistemas de *Data Warehousing*, con un enfoque basado en Contextos. Este primer planteo, es el punto de partida de una investigación más amplia y profunda que permita la gestión de la calidad en este tipo de Sistemas.

Palabras Clave: Calidad de datos, Sistemas de *Data Warehousing*, *Data Warehouse*, Contextos.

Abstract

Data Warehousing Systems are of great relevance for supporting decision making and data analysis. This has been proven over time, through the generalization of its development and use at industrial level in all kind of organizations. Moreover, the large number of scientific studies that have focused on the study of such systems have also proven the importance of them. Many researchers have presented the need to incorporate and maintain data quality in Data Warehousing Systems. However, there is no consensus in the research community on how or whether it is possible to define a single set of quality dimensions for Data Warehouse systems, due to the fact that this set of dimensions may depend on the purpose for which the data are used. On the other hand, once the data are in the Data Warehouse another challenge arises, how they will be used. Quality requirements may vary among different domains and among different users, not only due to the task they need to perform, but also because the quality perceived by a user may differ from the quality perceived by another user. Since data come from different sources with different levels of quality, analysis domains can vary and users can perceive the quality in different ways, depending on many factors (their profile, the task to be performed, etc.), for the evaluation of Data Quality in Data Warehousing Systems it is considered a data-context based approach.

In this thesis a systematic literature review is executed to obtain an overview of existing research on the use of contexts in Data Warehousing Systems and/or on the evaluation of Data Quality in this kind of systems. From the results obtained with the application of this methodology, an overview of the state-of-the-art is performed, which allows to do the first proposal to assess data quality in Data Warehousing Systems with an approach based on Contexts. This first proposal is the starting point of a broader and deeper investigation that will allow quality management in Data Warehousing Systems.

Keywords: Data Quality, Data Warehousing System, Data Warehouse, Context.

Índice general

1. Introducción **17**
- 1.1. Contexto . . . 17
- 1.2. Motivación . . . 17
- 1.3. Enfoque . . . 18
- 1.4. Objetivos y aportes . . . 19
- 1.5. Organización del documento . . . 20

2. Primera revisión bibliográfica **21**
- 2.1. Calidad de datos y Sistemas de DW . . . 21
  - 2.1.1. Conceptos básicos . . . 22
  - 2.1.2. Trabajos sobre Calidad de Datos en SDW . . . 32
- 2.2. Contextos . . . 35
  - 2.2.1. Trabajos sobre Contextos . . . 35
- 2.3. Conclusiones . . . 40

3. Aplicación de una metodología de búsqueda **43**
- 3.1. Definición de las *Research questions* . . . 47
- 3.2. Creación de las cadenas de búsqueda . . . 48
- 3.3. Selección de las bibliotecas digitales . . . 49
- 3.4. Desarrollo de la estrategia de búsqueda . . . 50
- 3.5. Ejecución del *Mapping Study* . . . 50
- 3.6. Análisis por *Research question* . . . 59
  - 3.6.1. RQ1: ¿Cómo son usados y definidos los contextos en los sistemas de DW? . . . 59
  - 3.6.2. RQ2: ¿Cómo se maneja la calidad de los datos en los sistemas de DW? . . . 68
  - 3.6.3. RQ3: ¿Cómo se consideran los contextos para la evaluación de calidad de datos? . . . 72
  - 3.6.4. RQ: ¿Cómo pueden ser usados los contextos para evaluar la calidad de datos en *Data Warehouse*? . . . 77
- 3.7. Otros resultados relevantes . . . 83
- 3.8. Conclusiones . . . 99

4. Propuesta:
Evaluación de la calidad en SDWs basada en contextos **103**
4.1. Contexto en componentes del SDW 105
4.2. Calidad de datos de acuerdo a su contexto 107
4.3. Caso de estudio 109
4.3.1. Calidad en el *Data Warehouse* 114
4.3.2. Calidad en el *Data Mart* 118
4.3.3. Calidad en uso 121
4.3.4. Resumen del caso de estudio 124
4.4. Prueba de concepto:
Implementación en Datalog 125
4.5. Conclusiones 134

5. Conclusiones y trabajo a futuro **137**
5.1. Aportes 141
5.2. Limitaciones 141
5.3. Trabajo a futuro 142

A. Definición de las cadenas de búsqueda **159**

Índice de figuras

2.1. **Jerarquía de conceptos de DQ** 24
2.2. **Cubo tridimensional de los datos de ventas. Con las dimensiones "Tienda", "Tiempo", "Producto" y la medida "Cantidad"** 26
2.3. **Notación del modelo MultiDim** 27
2.4. **CMDM: Nivel "Sucursal"** 27
2.5. **CMDM: Dimensiones "Sucursal" y "Promoción"** 28
2.6. **CMDM: Relación dimensional "Ventas"** 28
2.7. **Tabla de hecho y dimensiones en un modelo dimensional** . 30
2.8. **Arquitectura típica de un Sistema de *Data Warehousing*** 31
2.9. **Arquitectura en dos capas del Sistema de *Data Warehousing*** 32

3.1. **Definición de las cadenas de búsqueda parciales** 52
3.2. **Resultados por cadena de búsqueda parcial** 55
3.3. **Resultados por biblioteca digital** 56
3.4. **Proceso de selección de artículos** 57
3.5. **Cantidad de artículos por objeto contextualizado** 88
3.6. **Cantidad de artículos por contexto considerado** 90
3.7. **Tareas de calidad de datos y aspectos del contexto** 95

4.1. **Sistema de *Data Warehousing*** 103
4.2. **Contextos en un Sistema de *Data Warehousing*** 105
4.3. **Enfoques de calidad en un SDW** 108
4.4. **Calidad en un Sistema de *Data Warehousing*** 108
4.5. **Jerarquías y sus niveles, para cada una de las dimensiones** 110
4.6. **Relación dimensional "Ventas"** 110
4.7. **Relación dimensional "Rebajas"** 110
4.8. **Estructura del documento que contiene las sucursales de cada ciudad.** 115
4.9. **Dimensiones y niveles que dan contexto a la tabla de hechos "Rebajas"** 116
4.10. **Resultados obtenidos para la métrica dwq_Example1** . . . 127
4.11. **Resultados obtenidos para la métrica dwq_Example2** . . . 129

4.12. **Resultados obtenidos para la métrica dmq_Example3** . . . 131
4.13. **Resultados obtenidos para la métrica dmq_Example4** . . . 132
4.14. **Resultados obtenidos para la métrica qiu_Example5** 133

Índice de tablas

3.1. **Diferencias entre MS y SLR** 46
3.2. ***Research questions* parciales** 47
3.3. **Términos alternativos a las palabras claves** 48
3.4. **Cadena de términos por palabra clave** 49
3.5. **SS. Cadena de búsqueda principal** 49
3.6. **Bibliotecas digitales** 50
3.7. **Criterios de inclusión y exclusión** 51
3.8. **SS1. Cadena de búsqueda parcial** 52
3.9. **SS2. Cadena de búsqueda parcial** 52
3.10. **SS3. Cadena de búsqueda parcial** 53
3.11. **Resultados por bibliotecas digitales** (con duplicados) 54
3.12. **Artículos seleccionados por palabras claves** 58
3.13. **Contextos** 87
3.14. **Objetos contextualizados** 88
3.15. **Contextos considerados** 90
3.16. **Modelos** 91
3.17. **Reglas** 93
3.18. **Agrupación de las reglas** 93
3.19. **Tareas de Calidad de Datos** 94
3.20. **Dimensiones de calidad en artículos de DQ y CTX** 97
3.21. **Dimensiones de calidad en artículos de DQ y DW** 98

4.1. **Tabla de dimensión "Producto"** 111
4.2. **Tabla de dimensión "Tiempo"** 111
4.3. **Tabla de dimensión "Sucursal"** 112
4.4. **Tabla de dimensión "Promoción"** 113
4.5. **Tabla de hechos "Ventas"** 114
4.6. **Tabla de hechos "Rebajas"** 114
4.7. **Calidad en el DW. Contexto: Documento de la organización** 116
4.8. **Calidad en el DW. Contexto: Dimensiones del DW** 117
4.9. **Calidad en el DM. Dominio: Ventas** 119
4.10. **Calidad en el DM. Dominio: Publicidad** 120
4.11. **Calidad en uso, de acuero al perfil de usuario** 122
4.12. **Calidad en uso. Usuario: Gerente de publicidad** 123

4.13. **Resumen del caso de estudio** 124

A.1. **Adaptación de SS1 para la librería digital ACM** 159
A.2. **Adaptación de SS2 para la librería digital ACM** 160
A.3. **Adaptación de SS3 para la librería digital ACM** 160
A.4. **Adaptación de SS1 para la librería digital IEEE** 161
A.5. **Adaptación de SS2 para la librería digital IEEE** 161
A.6. **Adaptación de SS3 para la librería digital IEEE** 161
A.7. **Adaptación de SS1 para la librería digital ScienceDirect** . 162
A.8. **Adaptación de SS2 para la librería digital ScienceDirect** . 162
A.9. **Adaptación de SS3 para la librería digital ScienceDirect** . 162

Capítulo 1

Introducción

1.1. Contexto

Los Sistemas de *Data Warehousing* (SDW) son de gran relevancia para el apoyo en la toma de decisiones y el análisis de los datos. Esto ha quedado demostrado a lo largo del tiempo, a través de la generalización de su desarrollo y uso a nivel industrial en todo tipo de organizaciones y mediante la gran cantidad de trabajos científicos que se han centrado en el estudio de este tipo de sistemas. Más allá del objetivo de investigación, todos los autores coinciden en el valor agregado que los SDW aportan a las organizaciones. En particular, muchos de estos trabajos resaltan ampliamente la importancia de la Calidad de Datos para dichos sistemas y el peso que ésta tiene en la toma de decisiones. Por esta razón, muchos investigadores han presentado la necesidad de incorporar y mantener la calidad en los SDW. Sin embargo, en las investigaciones no se encuentra un consenso acerca de cómo hacerlo. La mayoría de los trabajos sólo abordan la limpieza de los datos en la etapa de extracción, transformación y carga de los datos, ignorando la tarea de evaluación de la calidad de los datos a lo largo de todo el ciclo de vida del *Data Warehouse* (DW). Además, también surge de la bibliografía la afirmación de que aún no se ha identificado cuál es el conjunto de dimensiones de calidad pertinentes para estos Sistemas de Información. A partir de esto último surge el cuestionamiento de si es imposible definir un único conjunto de dimensiones de calidad en el entorno de un DW, dado que dicho conjunto puede depender del propósito con el cual se utilizan los datos.

1.2. Motivación

Como se menciona en la sección anterior, los Sistemas de *Data Warehousing* son de gran relevancia para el apoyo en la toma de decisiones y el análisis de los datos, en particular en organizaciones de mediano y gran porte. Por ejemplo, consideremos una cadena de supermercados cuyas sucursales se distribuyen a lo

largo de todo el mundo y donde los datos son integrados desde diferentes fuentes. Las fuentes de datos son heterogéneas y presentan diferentes niveles de calidad en los datos que van a ser integrados. Además, dado que las fuentes aportan datos de distintas partes del mundo contienen, entre los diferentes problemas de calidad posibles, diferencias en los formatos (como puede ser el caso de la fecha), en la semántica (cuándo utilizan valores nulos), etc. Por todo esto, la integración de los datos presenta un desafío en sí mismo.

Por otro lado, una vez que los datos están en el DW surge otro desafío, cómo serán utilizados los mismos. En esta etapa, aparecen las necesidades de cada dominio de análisis. Dado que una organización consta de varias áreas o secciones (Ventas, Publicidad, Recursos Humanos, etc.), cada una define un dominio de análisis diferente con necesidades distintas. Se puede observar claramente que la sección de Ventas necesita realizar análisis de datos y estadísticos diferentes a los realizados por la sección Publicidad.

Los usuarios finales, de cada dominio de análisis, presentan necesidades y/o requerimientos diferentes, dependiendo de su perfil, de la tarea que van a realizar, etc. A su vez, los requerimientos de calidad pueden variar entre los usuarios, no sólo por el propósito de la tarea que necesiten realizar, sino también porque la calidad percibida por un usuario puede diferir respecto a la calidad percibida por otro usuario. Esto es así por la naturaleza subjetiva de la calidad.

Por todo esto, surge el interés de evaluar la calidad de los datos, en los SDW, teniendo en cuenta cómo van a ser utilizados dichos datos, por quién van a ser usados, cuándo y/o con qué propósito. Dado que, como se observa en la realidad antes presentada a modo de ejemplo, los datos vienen de diversas fuentes con niveles de calidad distintos, los dominios de análisis pueden ser variados y los usuarios pueden percibir la calidad de distintas formas, dependiendo esto de múltiples factores (su perfil, la tarea que va a realizar, etc.).

1.3. Enfoque

Para la evaluación de la Calidad de Datos en los SDW, se considera un enfoque basado en el Contexto de los datos. Más allá de los requisitos que presentan los SDW, de contemplar las distintas necesidades que presentan los diferentes usuarios a la hora de utilizar los datos almacenados en este tipo de sistemas, existe evidencia científica, que se desprende de la bibliografía, respecto a la subjetividad de la calidad de los datos. Aunque se observa claramente dicha subjetividad desde el punto de vista de los usuarios finales, la misma se mantiene a lo largo de todo el ciclo de vida del DW. Esto es así, porque los datos no son utilizados con un único propósito y en un único dominio de análisis. Por lo tanto, la evaluación de la Calidad de los Datos en los Sistemas de *Data Warehousing*, se realiza en base

al Contexto de dichos datos. Este enfoque se apoya en la visión de que los datos, durante su ciclo de vida (desde que son extraídos de las fuentes y hasta que son recibidos por un usuario), pueden recorrer diversos contextos.

1.4. Objetivos y aportes

El objetivo de esta tesis es encontrar un enfoque adecuado para la evaluación de la Calidad de Datos en SDW. Para alcanzar este objetivo, se definen dos objetivos específicos y se presentan a continuación:

- Analizar el Estado del Arte en el tema Calidad de Datos en SDW, teniendo en cuenta el tema Contextos como un posible enfoque.
- Proponer una primera aproximación a la resolución del problema de la evaluación de la Calidad de Datos en SDW.

Por otro lado, los aportes de la investigación realizada en esta tesis de Maestría son los que se presentan a continuación:

- Se presenta un Estado del Arte que relaciona las áreas Calidad de Datos, Sistemas de *Data Warehousing* y Contextos, utilizando una metodología de búsqueda bibliográfica rigurosa y reproducible llamada *Mapping Study*. Esta revisión bibliográfica permite entender cuál es el estado actual de la investigación referida a las áreas de interés e identificar cuáles son los desafíos existentes en dichas áreas.
- Las metodologías de búsqueda tienen como característica la posibilidad de reproducir las búsquedas. Se ejecuta la metodología *Mapping Study* en dos etapas, lo que permite demostrar su capacidad de reproducción. De esta manera, se obtiene un estado del arte exhaustivo de una forma incremental.
- Se realiza un primer planteo de una propuesta para la evaluación de la Calidad de Datos en Sistemas de *Data Warehousing*, con un enfoque basado en el Contexto de los datos. Se define un marco donde se situan los problemas de calidad de datos en un Sistema de *Data Warehousing*. Para esto, se tienen en cuenta tres de los componentes que forman parte de estos sistemas: el Data Warehouse, los Data Marts y las aplicaciones del cliente, y los diferentes contextos que tienen influencia sobre ellos. Luego se define la calidad para cada uno de estos componentes, basada en el contexto correspondiente.
- Se ejecuta una prueba de concepto, con la cual es posible demostrar la aplicabilidad de la propuesta planteada.

1.5. Organización del documento

En el Capítulo 2 se presentan conceptos básicos y una primera revisión bibliográfica, no sólo acerca de los temas que conciernen al objetivo de este trabajo, sino también de los temas Calidad de Datos, Sistemas de *Data Warehousing* y Contextos, que son temas de gran interés, por sí sólos, para esta tésis.

En el Capítulo 3 se presenta la aplicación de una metodología de búsqueda, con la cual se realiza un análisis más profundo del Estado del Arte. Este capítulo es un complemento del Capítulo 2, ya que este último permitió obtener una primera aproximación del estado actual de la investigación relacionada con el objetivo de esta tesis.

En el Capítulo 4 se realiza un primer planteo de una propuesta para evaluar la Calidad de Datos en los SDW, con un enfoque basado en Contextos. Además, se define un Caso de Estudio para el cual se ejecuta una prueba de concepto implementada en *Datalog*.

En el Capítulo 5 se presentan las conclusiones obtenidas al final de la realización de esta tesis de Maestría. Además, se incluyen los aportes, limitaciones y futuras líneas de trabajo que surgen de la misma.

Este documento, además contiene al Apéndice A, el cual muestra cómo son definidas las cadenas de búsqueda, para cada biblioteca digital considerada, en la metodología de búsqueda aplicada en el Capítulo 3.

Capítulo 2

Primera revisión bibliográfica

El objetivo de la primera revisión bibliográfica es conocer el estado actual de la literatura existente acerca de la Calidad de Datos (del inglés *Data Quality*, DQ) en los Sistemas de *Data Warehousing*.

Inicialmente, se obtiene una visión general y los diferentes aspectos de DQ, independientemente del Sistema de Información (del inglés *Information System*, IS) en el cual se aplican los principios de DQ. Por otro lado, se investigan los trabajos que abordan conceptos y problemas de los SDW y una vez presentados los conceptos básicos de cada uno de los temas de interés, se analiza la literatura que se enfoca en el estudio de la Calidad de Datos en los Sistemas de *Data Warehousing*.

A partir del análisis de los diferentes trabajos de investigación, relacionados con la Calidad de datos en un SDW, surge la necesidad de centrar los estudios en la importancia del usuario y en todo lo que este implica (sus preferencias, la tarea que éste realiza, su perfil, etc.). Por esto, se plantea el estudio de un nuevo concepto: Contexto (CTX). En primer lugar, se aborda el contexto desde la perspectiva de los usuarios, pero una vez que se abordan los conceptos que lo definen se observa que dicho concepto va más allá del usuario, ya que los contextos son objetos de numerosos trabajos, donde son considerados con diferentes enfoques. Por esta razón, se incluye una revisión bibliográfica para entender qué son los contextos, en qué consisten y cómo se representan. Además, se analizan las circunstancias en las cuales éstos son aplicados y los resultados obtenidos a partir de su aplicación.

2.1. Calidad de datos y Sistemas de DW

En esta sección se presentan al lector los conceptos básicos de Calidad de datos y de los SDW, para posteriormente analizar la bibliografía que aborda la investigación de la calidad de datos en los Sistemas de DW.

2.1.1. Conceptos básicos

Los conceptos básicos que definen a cada una de las áreas: Calidad de Datos y Sistemas de DW, introducen al lector en las nociones elementales que permiten abordar esta tesis.

Calidad de datos

Calidad de Datos es un área de investigación muy amplia, que implica muchos aspectos, problemas y desafíos diferentes. Además, tiene una enorme relevancia para la industria, debido a su gran impacto en los sistemas de información de utilidad en todos los dominios de aplicación. El término "Calidad de Datos" se usa con referencia a un conjunto de características que deben poseer los datos, tales como su correctitud, su grado de actualización, etc [1]. Las consecuencias de la mala calidad de los datos a menudo se experimentan en la vida cotidiana, una dirección incorrecta o duplicada es ejemplo de un problema de DQ. Una de las razones para abordar los problemas de calidad de datos, es la creciente necesidad de integrar la información a partir de fuentes de datos heterogéneas, ya que la mala calidad de los datos obstaculiza la integración de los mismos [2].

En [3] definen DQ como la capacidad de cumplir con los requerimientos necesarios para el uso de los datos. Por tanto, los datos carecen de calidad en la medida en que no satisfacen los requerimientos. En otras palabras, según los autores, la calidad de datos depende tanto del uso que se le vaya a dar a dichos datos, como de los datos en sí. En [4] proponen una definición similar, ya que consideran que la calidad de datos es la adecuación de los datos para su uso (del inglés *fitness for use*). Este concepto es ampliamente adoptado en la bibliografía referente a calidad de datos, algunos ejemplos son [1] [5] [6] [7] [8]. En particular, los autores de [1] se centran específicamente en la definición de DQ en el área de *Computer Science* y creen necesaria una definición de DQ que incluya características consideradas por la mayoría de las propuestas realizadas. En dicha investigación, presentan un análisis de las distintas definiciones que han sido propuestas desde los años 90.

Por otro lado, también en [1], resaltan que la calidad de los datos puede ser relacionada a un conjunto de dimensiones. Estas dimensiones de calidad son generalmente definidas como propiedades o características de calidad. También en [9] identifican y discuten dimensiones de DQ, finalmente agrupan las mismas en cuatro categorías: exactitud, actualidad, completitud y consistencia. Además, afirman que éstas son las dimensiones más a menudo asociados con DQ, lo que coincide con [1], ya que en el análisis realizado a partir de varias propuestas, definen la Calidad de Datos como un conjunto de dimensiones, que incluye:

- Exactitud (del inglés *accuracy*): Mide la distancia entre un valor v1 y un valor v2 que se considera correcto [2].

- Completitud (del inglés *completeness*): El grado en que los datos son suficientemente amplios y el alcance que éstos tienen para la tarea en cuestión [2].
- Consistencia (del inglés *consistency*): Captura la violación de reglas semánticas (restricciones de integridad), definidas sobre un conjunto de datos [2].
- Actualidad (del inglés *timeliness*): Qué tan actuales son los datos para una tarea específica [2].
- Interpretabilidad (del inglés *interpretability*): Mide el grado de claridad de la información (lenguaje, símbolos, definiciones, etc.) [10].
- Accesibilidad (del inglés *accessibility*): Mide la disponibilidad de la información, cuán fácil y rápida es su recuperación [10].

Los autores de [1] subrayan que en la literatura no hay un acuerdo sobre el conjunto de dimensiones que caracterizan a DQ, incluso agregan que si algunas dimensiones son consideradas como importantes, no hay un acuerdo en sus significados. Además, mencionan que si bien muchas propuestas se han hecho, nadie ha podido establecer un estándar. Sin embargo, mencionan que aunque hay varias dimensiones diferentes en las distintas propuestas analizadas, es posible destacar algunas de ellas que, básicamente, definen el concepto de Calidad de Datos.

En el momento de realizar una evaluación de la calidad de los datos, esta tesis se basa en el enfoque de DQ [11] [12] que caracteriza la calidad de acuerdo a varias dimensiones que ayudan a calificar los datos y, como se observa en la Figura 2.1, define una jerarquía de conceptos de calidad. Este enfoque se apoya fuertemente en los conceptos antes presentados y coincide en que cada dimensión captura una faceta de la calidad. A su vez, cada **dimensión** puede verse como un conjunto de factores de calidad que tienen un mismo propósito y un **factor** representa un aspecto particular de una dimensión. Por otro lado, un mismo factor de calidad puede ser medido con distintas métricas, mientras que una **métrica** es un instrumento que define la forma de medir a cada factor de calidad. Una misma métrica puede ser medida por diferentes métodos. Cada **método** es un proceso que implementa a cada métrica. Finalmente, cuando se realiza una **medición**, mediante la aplicación de un método, sobre un dato o un conjunto de datos, se obtiene una **medida de calidad**.

La definición de las dimensiones está relacionada con la tarea de evaluación de la calidad de los datos. Sin embargo, existe otra tarea de DQ ampliamente utilizada, la misma es la limpieza de los datos (del inglés *Data cleaning, Data cleansing o Scrubbing*). En [13] clasifican problemas de calidad de datos y proporcionan distintos enfoques de solución a partir de la limpieza de datos. Los autores mencionan que la tarea de limpieza se ocupa de la detección y eliminación de errores e inconsistencias de los datos, con el fin de mejorar la calidad de

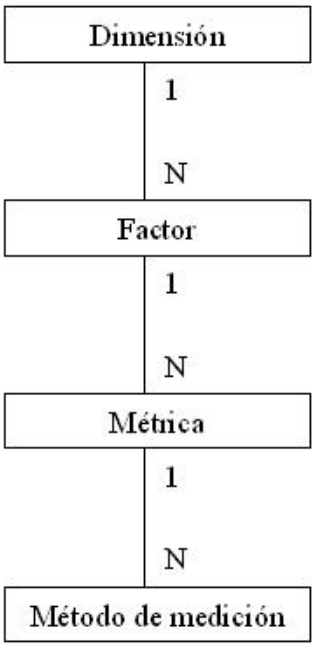

Figura 2.1: **Jerarquía de conceptos de DQ**

los mismos. La tarea de limpieza de datos es especialmente requerida cuando se realiza la integración de fuentes de datos heterogéneas.

No es una novedad el impacto que la mala calidad de los datos tiene sobre una organización. Por ejemplo, en el trabajo [14] del año 1998, los autores ya observaban diferentes impactos y en su investigación afirman que dentro de los mismos se incluye la insatisfacción del cliente, el aumento de los costos operativos, la toma de decisiones ineficaz y se reduce la capacidad de hacer y ejecutar estrategias. Desde una visión más general, mencionan que la mala calidad de los datos genera desconfianza sobre la organización.

Sistema de *Data Warehousing*

En esta sección se presentan los componentes y conceptos que definen a un Sistema de *Data Warehousing*. Entre todos los Sistemas de Soporte de Decisión (del inglés Decision Support Systems, DSS), según los autores de [15], los Sistemas de *Data Warehousing* son probablemente los sistemas a los cuales las comunidades académicas e industriales han prestado mayor atención. En este mismo trabajo definen *Data Warehousing* como una colección de métodos, técnicas y herramientas usadas para apoyar el proceso de toma de decisiones.

Los SDW se han convertido en una tecnología clave para la integración de fuentes de información distribuidas. Conceptualmente, un DW es una colección de datos orientados a temas, integrados, no volátiles y variables en el tiempo,

organizados de tal forma que facilitan el proceso de la toma de decisiones [16]. Por otro lado, los datos que ofrece un DW fueron extraídos de diversas fuentes, integrados, limpiados y transformados para finalmente ser usados en el momento de la toma de decisiones [15] [16] [17]. Uno de los desafíos de este tipo de sistemas es, justamente, la dinámica de integración de las diferentes fuentes de datos. En [18] identifican los problemas que se presentan cuando hay cambios en las fuentes de información de las cuales son tomados los datos procesados y analizados en un DW. Además, mencionan que algunos de los problemas característicos son la actualización de los datos, los cambios de esquema y las modificaciones de restricciones, entre otros.

Los DWs están basados en un **Modelo Multidimensional** [19]. Este modelo permite una mejor comprensión de los datos y proporciona un mejor rendimiento en consultas complejas. Los datos, en los modelos multidimensionales, se presentan en un espacio n-dimensional, generalmente llamado **cubo de datos** o hipercubo. Un ejemplo de cubo de datos se muestra en la Figura 2.2. Un cubo de datos está definido por **dimensiones** y **hechos** (del inglés *facts*). Las dimensiones representan a las distintas perspectivas que se utilizan para analizar los datos. Por ejemplo, el cubo de datos de la Figura 2.2, tomado de [19], se utiliza para analizar ventas y tiene tres dimensiones: "Tienda", "Tiempo", "Producto". Las instancias de una dimensión se denominan **miembros**. Por ejemplo, "París", "Niza", "Roma" y "Milán" son miembros de la dimensión "Tienda". Las dimensiones tiene **atributos** asociados que describen a la dimensión. Por ejemplo, la dimensión "Producto" podría contener atributos tales como "número de producto", "nombre del producto", "descripción" y "tamaño", estos atributos no se muestran en la figura. Por otro lado, las celdas del cubo de datos o hechos, están asociadas con valores numéricos, llamados **medidas**. Estas medidas permiten la evaluación cuantitativa de los diversos aspectos del problema de análisis. Por ejemplo, los números mostrados en el cubo de datos de la Figura 2.2 representan una cantidad medida e indica el importe total de las ventas se especifica en, por ejemplo, miles de dólares. Un cubo de datos normalmente contiene varias medidas, por ejemplo, otra medida que no se muestra en el cubo en la Figura 2.2 podría ser la cantidad de unidades vendidas.

En el **diseño conceptual** de un DW se construye un esquema conceptual, éste es la descripción de los requerimientos de los usuarios. Para la representación del esquema conceptual, se tienen en cuenta dos modelos:

- **MultiDim** [19]: Este modelo permite representar conceptualmente todos los elementos requeridos (dimensiones y hechos con sus medidas asociadas), en el almacenamiento de los datos en un DW y las aplicaciones OLAP (del inglés *Online Analytical Processing*). En este modelo, un **esquema** está compuesto por un conjunto de dimensiones y un conjunto de hechos. Una **dimensión** es un concepto abstracto que agrupa datos que comparten

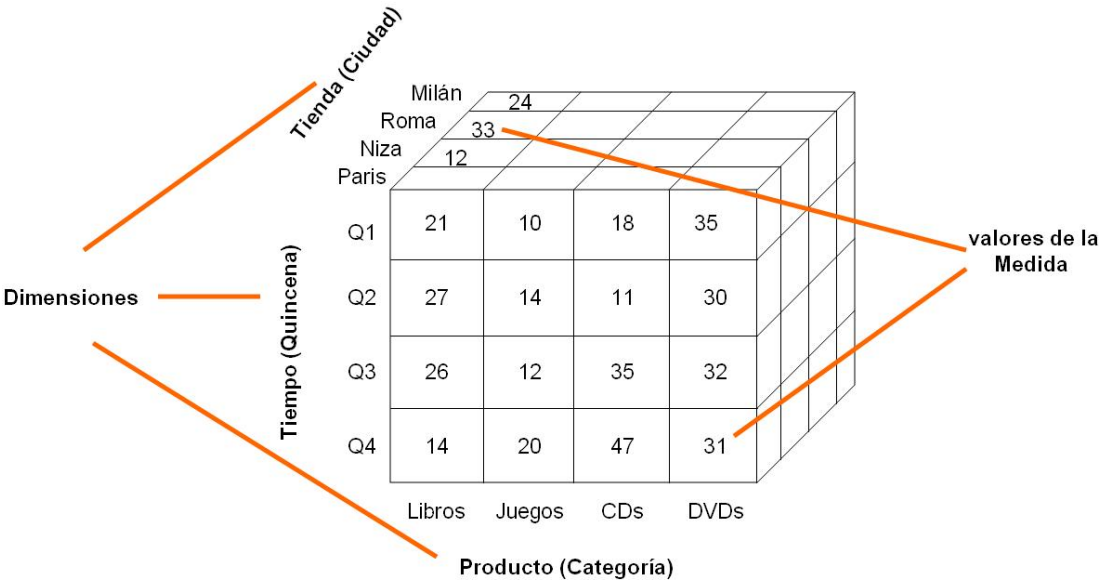

Figura 2.2: **Cubo tridimensional de los datos de ventas. Con las dimensiones "Tienda", "Tiempo", "Producto" y la medida "Cantidad"**

un significado semántico en el dominio que se modela. Una dimensión se compone de un conjunto de jerarquías y una **jerarquía** está compuesta por un conjunto de niveles. Un **nivel** describe una serie de conceptos del mundo real que tienen características similares. Cuando dos niveles se relacionan en una jerarqua, el nivel inferior se denomina nivel hijo y el nivel más alto se llama nivel padre. Por esto, las relaciones que componen las jerarquías se denominan relaciones padres-hijos. Por otro lado, una **relación de hecho** expresa un enfoque de análisis y representa una relación n-aria entre los niveles. Ejemplos de notación MultiDim se muestran en la Figura 2.3.

- **CMDM** [20]: El objetivo de este modelo es especificar bases multidimensionales, lo que significa especificar datos para OLAP y *Data Warehousing*. Fundamentalmente especifica una determinada realidad en términos multidimensionales y para lograr esto, este modelo presenta tres estructuras básicas: Niveles, Dimensiones y Relaciones Dimensionales. Un **nivel** representa un conjunto de objetos que son de un mismo tipo. Cada nivel debe tener un nombre y un tipo, un ejemplo de nivel se presenta en la Figura 2.4. Una **dimensión** está determinada por una jerarquía de niveles. En la Figura 2.5 se observan dos dimensiones con sus respectivas jerarquías, el nivel "Sucursal" de la Figura 2.4 pertenece a la jerarquía de la dimensión "Sucursal" de la Figura 2.5. Una **relación dimensional** representa el conjunto de todos los cubos que se pueden construir a partir de los niveles de un conjunto dado de dimensiones. Un ejemplo de relación dimensional se presenta en la Figura 2.6, el esquema está dado por un grafo en forma de estrella. El nodo central es de forma oval y tiene el nombre de la relación

dimensional y los nodos rectangulares "satélites" son las dimensiones que participan de la dicha relación. La medida está indicada con la punta de la flecha, en este caso, la medida se llama "cantidad de ventas".

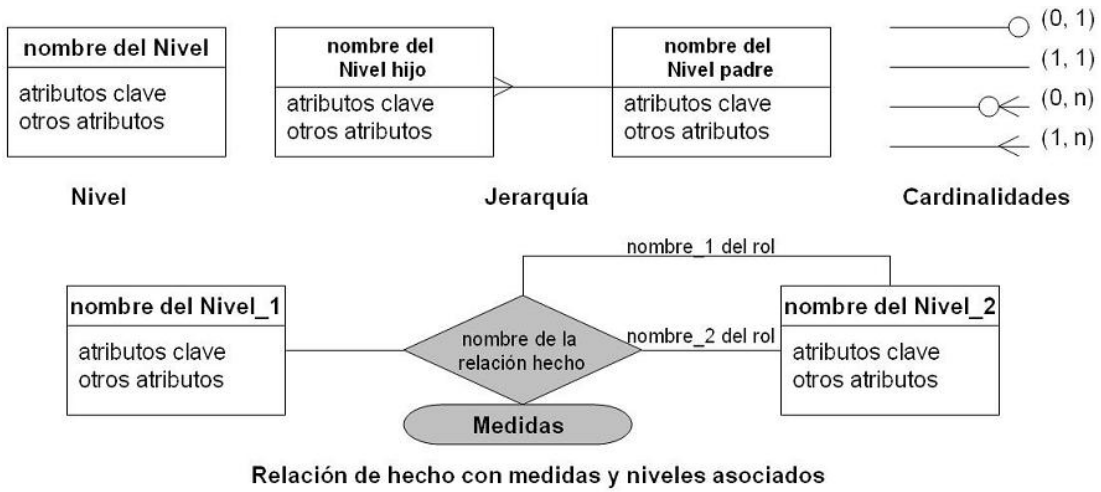

Figura 2.3: **Notación del modelo MultiDim**

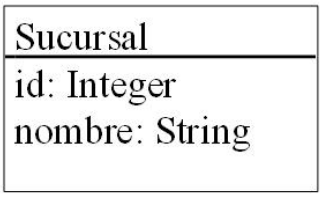

Figura 2.4: **CMDM: Nivel "Sucursal"**

Existen varios enfoques diferentes para la implementación de un modelo multidimensional. Dicha implementación se conoce como el **diseño lógico** del DW y este depende de la forma en la que un cubo de datos es almacenado [19]:

- OLAP Relacional (ROLAP): Los servidores OLAP almacenan los datos en bases de datos relacionales.
- OLAP Multidimensionales (MOLAP): Los servidores OLAP multidimensionales almacenan los datos multidimensionales directamente en estructuras de datos especiales (por ejemplo, *arrays*) y ponen en práctica las operaciones OLAP sobre dichas estructuras de datos.
- OLAP Híbrido (HOLAP): Los servidores OLAP híbridos combinan las tecnologías anteriores, beneficiándose de la capacidad de almacenamiento de ROLAP y las capacidades de procesamiento de MOLAP.

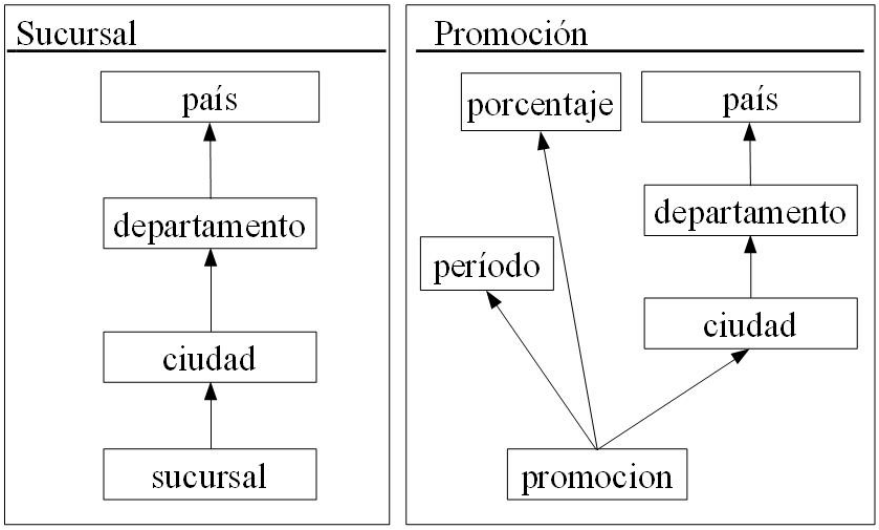

Figura 2.5: **CMDM: Dimensiones "Sucursal" y "Promoción"**

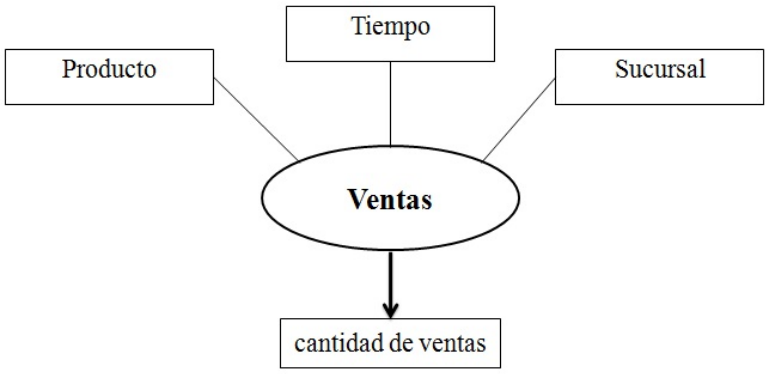

Figura 2.6: **CMDM: Relación dimensional "Ventas"**

En los sistemas ROLAP, los datos multidimensionales se implementan como tablas relacionales, organizados en estructuras llamadas: esquema estrella y esquema *snowflake*. En un esquema en estrella, sólo hay una tabla de hechos central y un conjunto de tablas dimensión, uno para cada dimensión del esquema conceptual. En el esquema *snowflake* se evita la redundancia de los esquema estrella por la normalización de las tablas de dimensiones. Por lo tanto, una dimensión está representada por varias tablas relacionadas por restricciones de integridad referenciales.

Los esquemas antes definidos, para los sistemas ROLAP, hacen referencia a las tablas de hechos y a las tablas dimensión, estos conceptos son tomados de [17] y se describen a continuación:

- Tabla de hechos: es la tabla principal de un modelo dimensional, donde se almacenan las mediciones de rendimiento numéricos de la empresa. El término hecho (o *fact*), es utilizado para representar una medida de negocios.Una fila en la tabla de hechos se corresponde con una medición, es decir, una medida es una fila de la tabla de hechos. Todas las mediciones en una tabla de hechos deben tener la misma granularidad.

- Tabla dimensión: Las tablas dimensión contiene descriptores del negocio y cada tabla tiene muchas columnas o atributos. Cada dimensión se define por su clave primaria (del inglés *primary key*, PK), que sirve como base para obtener la integridad referencial con cualquier tabla de hechos.

El ejemplo de la Figura 2.7 resume los conceptos antes presentados, muestra un esquema estrella en el cual se observan las tablas de dimensiones ("Tiempo", "Sucursal" y "Producto") y la tabla de hechos ("Ventas"). El ejemplo representa a los productos que se venden en una sucursal, de una cadena de supermercados, y se anota la cantidad de ventas cada día para cada producto en cada sucursal. La intersección de todas las dimensiones (el día, el producto y la sucursal) representa a una medida.

Por otro lado, una vez definidos los conceptos más importantes de los modelos multidimensionales, en los cuales se basan los DWs, interesa describir los componentes que determinan la arquitectura de un SDW. Los autores de [19] presentan la **arquitectura típica de un SDW**, la misma se muestra en la Figura 2.8 y consta de los siguientes componentes:

- Las **fuentes de datos** contienen la materia prima de un SDW y pueden ser internas (sistemas operacionales, documentos electrónicos, etc) o externas a la organización (indicadores demográficos, etc.).

- Las **herramientas de extracción, transformación y carga** (del inglés *Extraction, Trasformation and Load*, ETL) se utilizan para cargar el DW

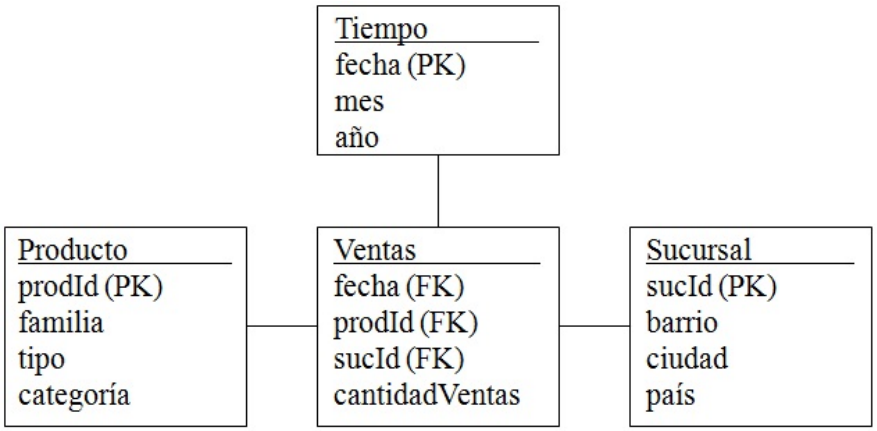

Figura 2.7: **Tabla de hecho y dimensiones en un modelo dimensional**

con los datos obtenidos a partir de las fuentes. Los procesos de ETL pueden unir distintos esquemas heterogéneos, resolviendo todos los problemas que eso implica. Dado que el proceso de extracción puede tener un alto impacto en la performance de los sistemas fuente y no siempre se encuentran todos disponibles en forma simultanea, es que se utiliza lo que se conoce como ***Data Staging***. El *Data Staging* es un almacenamiento temporal que guarda los datos extraídos y sobre el cual se ejecutan todos los procesos de integración y transformación de los datos, antes de que estos sean cargados en el DW.

- Un ***Data Warehouse*** abarca todas las secciones de una organización. El DW puede ser directamente accesido y usado como fuente para crear ***Data Marts*** (DM). Un *Data Mart* (DM) replica un subconjunto de datos o una agregación del DW y es diseñado para una sección específica de la organización, es decir, para un dominio de análisis específico [15].

- El servidor OLAP brinda el soporte necesario para organizar los datos de forma multidimensional y para ejecutar operaciones de manipulación de los datos contenidos en el DW [17]. Una característica fundamental del modelo multidimensional es que permite ver los datos desde múltiples perspectivas y en diferentes niveles de detalle. Hay una conjunto de operaciones OLAP que permiten estas perspectivas y niveles de detalle que se materializan mediante la explotación de las dimensiones [19]. Ejemplos de algunas de las operaciones de manipulación de datos son: *slice* (permite definir un subconjunto del cubo especificando las dimensiones sobre las qué se desea trabajar), *roll-up* (esta operación consolida medidas detalladas en medidas resumidas aplicando las funciones de agregación definidas para cada medida. Esto se lleva a cabo cuando se asciende en una jerarquía o se elimina una dimensión), *drill-down* (esta operación navega el cubo sobre sus di-

mensiones), *pivot*(esta operación rota los ejes de un cubo para brindar un presentación alternativa de los datos).

- Las herramientas **clientes** permiten que los usuarios realicen análisis interactivos de la información. Para esto utilizan clientes OLAP, aplicando técnicas de Minería de datos (del inglés *Data Mining*), etc. Los datos integrados son utilizados para generar informes, analizar la información de forma dinámica y simular hipotéticos escenarios de negocio [15].
- Los **Metadatos** (MD) son información acerca de los datos del DW. Los repositorios de Metadatos (MD) almacenan información sobre las fuentes, los procedimientos de acceso, de los usuarios, de los esquemas de los *Data Marts*, etc. [15]. Los metadatos son fundamentales para el control de calidad de los datos y para la explotación eficaz del DW [17].

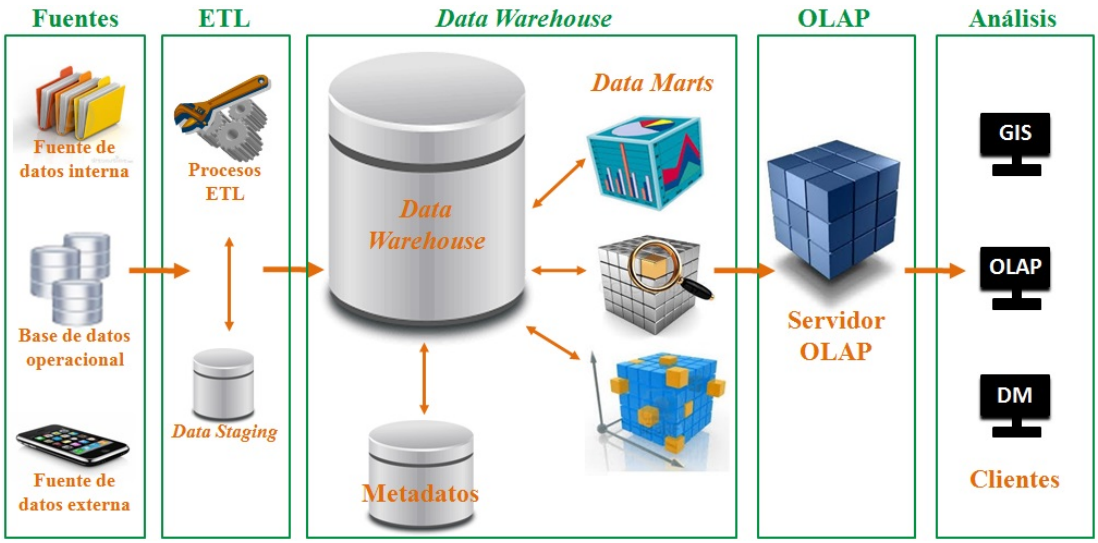

Figura 2.8: **Arquitectura típica de un Sistema de *Data Warehousing***

Por otro lado, en [15] presentan diferentes arquitecturas de un SDW y mencionan que la más referenciada por la bibliografía y en la cual se apoyan los conceptos de esta tesis, es la denominada arquitectura de dos capas y se muestra en la Figura 2.9. Esta arquitectura es llamada así para resaltar la separación entre las fuentes y el *Data Warehouse* (DW), pero en realidad, según los autores, consiste en cuatro etapas de flujo de datos:

- Capa de las fuentes
- Capa de *Data staging*
- Capa del *Data Warehouse*
- Capa de análisis

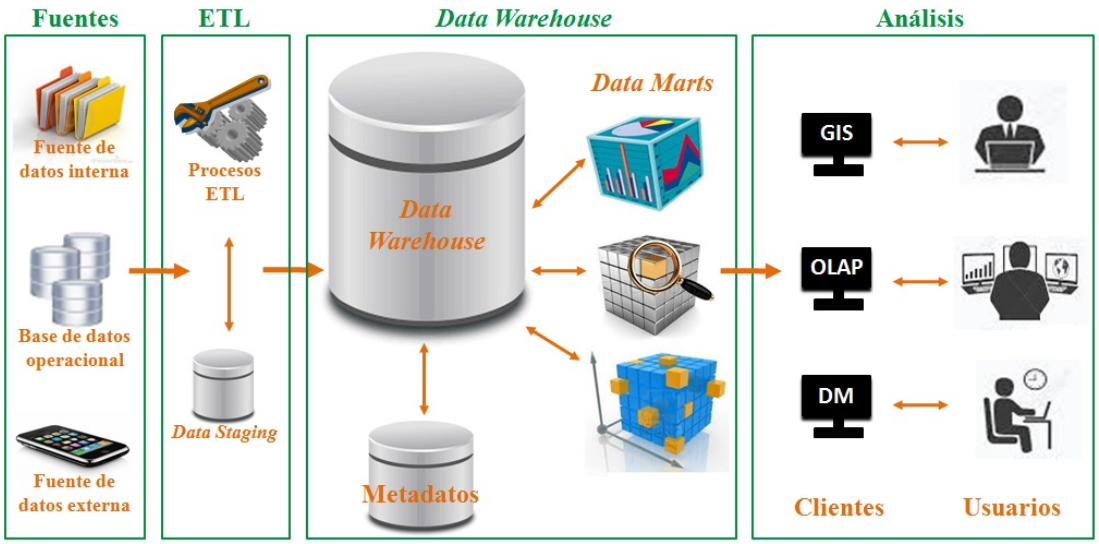

Figura 2.9: **Arquitectura en dos capas del Sistema de *Data Warehousing***

2.1.2. Trabajos sobre Calidad de Datos en SDW

Una vez que los aspectos generales de DQ y de los SDW han sido presentados, se analiza la bibliografía referente a los trabajos que abarcan ambas áreas de investigación: Calidad de Datos en los Sistemas de *Data Warehousing*.

En la búsqueda de la información, los autores de [21] sugieren que los metadatos son importantes. En particular, resaltan la importancia de los metadatos de calidad, que manifiestan pueden guiar a los usuarios finales de los DW en sus tomas de decisiones y ayudarlos a determinar si ciertos datos son adecuados para un propósito específico. También en [22] se centran en el manejo de los metadatos que se encuentran en los DW, de forma de utilizarlos en el análisis sistemático de la calidad y el diseño orientado a la calidad. Por otro lado, los autores de [23] se refieren a otra forma de uso de los metadatos. Señalan que todos los componentes de un DW, procesos y datos, deberían tener asociado un repositorio de metadatos, de esta forma, el repositorio sería una forma de rastreo de todas las opciones de diseño y del historial de cambios realizados en la arquitectura y sus componentes, lo que consideran puede ayudar en la gestión de calidad de datos .

Los autores de [24] coinciden en que los datos deben ser explotados utilizando al DW como una herramienta poderosa en el momento de apoyar la toma de decisiones. Sin embargo, consideran que enfoques como los presentados anteriormente, imponen tareas adicionales (como lo es la tarea de consultar el repositorio de metadatos), a los usuarios finales, para decidir si la calidad de la información de interés es pertinente, olvidando que los usuarios finales podrían no tener conocimientos acerca del dominio del DW. Teniendo esto en cuenta, proponen una extensión a la arquitectura típica del DW con el objetivo de capturar los

cambios de calidad y notificando dichos cambios a los usuarios finales. Para esto, introducen dos tipos de servicios en el diseño de arquitectura: *Quality factory* y *Quality notification service*, para proveer medidas de calidad y para notificar a los usuarios finales sobre los cambios de calidad, respectivamente.

En el trabajo presentado en [25], también se aborda la dificultad de diseñar DWs de buena calidad y plantean una propuesta que provee ayuda a los diseñadores mediante la vinculación de los principales componentes de la arquitectura de un DW con el modelo formal de calidad de datos presentado en [26]. En [27], extienden el trabajo presentado en [25], examinando y describiendo herramientas que se centran en los aspectos de la gestión de metadatos y en la mejora de la interacción cliente-DW. En este caso, también se hace foco en el uso de los metadatos para apoyar el manejo de la calidad de datos. También en [28] [29] abordan la necesidad de contar con buenos diseños y se enfoca en la calidad de los modelos de datos, en particular, en la calidad de los modelos conceptuales.

Por otro lado, el artículo [30] se centra en el uso de los datos y ofrece una metodología que mejora la calidad de un conjunto de datos de uso frecuente. Dado que los datos almacenados son accedidos por diversos usuarios con diferentes necesidades, consideran que las actividades de calidad de datos deben ser distribuídas a lo largo de todo el proceso, para apoyar mejor las decisiones de una organización. Por esto, subrayan que la calidad de datos del DW debe ser asegurada en todas las etapas del mismo, no solo en la etapa del diseño como mencionan los trabajos anteriores, sino también en las etapas de implementación y mantenimiento. Además, señalan que una de las características más importantes de un DW es el predominio de los datos históricos y de los datos blandos, siendo estos últimos datos cuya calidad es inherentemente incierta. Un ejemplo de datos blandos es la asignación de tareas futuras, a empleados de una organización, que involucra una valoración subjetiva. Si bien esta asignación puede ser imperfecta, porque podría cambiar en el futuro, la misma debe ser tenida en cuenta en la organización. En muchos sistemas se excluyen los datos blandos, sin embargo, los autores consideran que en muchas ocasiones estos datos pueden llegar a ser cruciales para la toma de decisiones.

El trabajo presentado en [31] coincide con el enfoque anterior. Los autores plantean una propuesta para definir, crear y mantener la gestión de calidad de datos a lo largo de todo el ciclo de vida de un SDW. Relacionan cada una de las etapas de un SDW con los factores que definen en el modelo de calidad propuesto: exactitud, completitud, actualidad, integridad, consistencia, conformidad y registro de duplicación. Además, presenta un caso de uso para demostrar cómo la calidad de datos puede ser lograda aplicando dicho modelo. En [32] aplican métricas de calidad a la información de origen para mejorar la calidad de la información de salida. Desarrollan un modelo definido como un cubo de datos y un álgebra que apoya las operaciones de evaluación de la calidad sobre este cubo.

Los autores aseguran que la metodología presentada permitiría a los usuarios determinar la confiabilidad de la información recibida, logrando mejores resultados en la toma decisiones. En [33] identifican un conjunto de datos que produzca los valores óptimos en el proceso de toma de decisiones de una organización. Para esto, ofrecen un modelo que presenta técnicas para determinar qué tablas de datos y qué vistas deben ser incluidos en el DW, basándose en la calidad, costo y valor. Además, las técnicas propuestas, contemplan la posibilidad de cambios en el DW, por lo que dichas técnicas pueden ser aplicadas a medida que el DW evoluciona.

A diferencia de los trabajos anteriores, en [34] se enfocan en una de las etapas del SDW, ya que el objetivo de este trabajo es integrar el proceso de calidad dentro de las tareas ETL. Para esto, presenta una propuesta para el control de calidad y limpieza, mediante la creación de módulos de calidad que corren automáticamente después de la carga en la capa de *Data Staging* y antes de la carga del DW. Cada módulo DQ marca registros sospechosos (registros de advertencia) y corrige valores inválidos (registros erróneos), buscando tener la menor cantidad posible de registros erróneos, en comparación con el número total de registros. La propuesta también incluye el estudio de reglas de calidad y la mayoría de estas reglas son aplicadas en la capa *Data Staging*, para evitar que registros erróneos lleguen a la capa del DW. Algunas de las reglas corrigen valores inválidos y otras simplemente marcan registros incorrectos. Los registros de advertencia se realizan en el momento de transformación y carga, en la capa del DW, mientras que los registros erróneos se encuentran en la capa de *Data Staging*. Finalmente, la corrección de los registros requiere la intervención de los usuarios del negocios. En la misma línea, los autores de [35] indican que la calidad de datos sólo puede ser producida a través de la limpieza de los mismos, antes de que éstos sean cargados en el DW. Consideran que una propuesta de limpieza de datos es satisfactoria cuando detecta y elimina la mayor cantidad de datos erróneos e inconsistencias importantes, tanto en fuentes de datos individuales como en la integración de múltiples fuentes. En este caso, también proponen reglas para la transformación de los datos, que son las que realizan la limpieza de los mismos.

La etapa de análisis de los datos es de gran importancia en los SDW, los autores de [36] [37] se enfocan en esta etapa y proponen una herramienta de recomendación para sesiones OLAP. Dicha herramienta basa sus recomendaciones en sesiones previas (del mismo y/o de otros usuarios) sobre el mismo cubo de datos. La propuesta se basa en un modelo probabilístico del comportamiento del usuario, que permite estimar su comportamiento futuro, y en la definición de una métrica de similaridad de consultas. Por otro lado, en el trabajo [38] presentan una arquitectura para soluciones de calidad en aplicaciones BI, haciendo foco en la relación entre calidad de datos y reportes de calidad. Para esto, introducen la noción de reportes de calidad (*quality-aware reports*) y los mismos son definidos como consultas sobre el DW que exponen la calidad de los resultados obtenidos. Presentan el desafío que representa la cantidad de diferentes problemas potencia-

les de calidad y la importancia de que el usuario tenga conciencia de por qué un reporte es considerado de baja calidad. Subrayan que la calidad es subjetiva en dos sentidos, por un lado los problemas de calidad pueden ser o no significativos para ciertas decisiones y, por otro lado, los analistas pueden tener (y en general es así), conocimiento personal u opiniones sobre la calidad de datos. En el trabajo investigan el problema de la baja calidad de datos en BI, centrándose en el usuario final. Los datos de baja calidad utilizados como entrada en las aplicaciones de BI afectan la calidad de datos de salida, como pueden ser los reportes. De acuerdo a esto, se propone el uso de reportes de calidad, permitiendo al usuario final interactuar con los metadatos de calidad de dichos reportes.

En esta primera revisión bibliográfica se observa la necesidad de considerar metadatos de calidad a lo largo de todo el ciclo de vida de un SDW. En particular, resaltan aquellas propuestas [21] [24] [30] [38] en las cuales el usuario final y sus necesidades presentan un rol de importancia a la hora de considerar la calidad de los datos en un SDW.

2.2. Contextos

En la sección anterior se observa que varios trabajos resaltan la importancia del usuario y/o las características que lo determinan (sus preferencias, la tarea que éste realiza, su perfil, etc.). Por esta razón, en esta tesis surge la necesidad de incorporar un nuevo concepto: Contexto. En principio, se aborda la investigación teniendo en mente el contexto del usuario, pero una vez avanzado el estudio de los trabajos científicos que se refieren al área de los Contextos, se observa que dicho concepto va más allá del usuario, esto se observa a partir del análisis bibliográfico que se presenta a continuación.

2.2.1. Trabajos sobre Contextos

En esta sección se presentan los conceptos básicos que definen a un Contexto. Por lo tanto, se incluye bibliografía que permita entender qué son los contextos, en qué consisten y cómo se representan. Además, se analizan las circunstancias en las cuales éstos son aplicados y los resultados obtenidos a partir de su aplicación.

En la bibliografía se destaca repetidas veces el problema del crecimiento de la información, que se encuentra en forma digital, y el ruido que dicho crecimiento provoca en los datos [39] [40] [41]. A ésto se suma la evolución de las redes inalámbricas, la aparición de nuevos dispositivos móviles y sistemas integrados, que hacen que la informática se mueva de los hogares y oficinas a nuevos dominios de la vida diaria. Esto último se define en [42] [43] [44] [45] como Computación Ubicua (del inglés *Pervasive Computing*), y demanda aplicaciones capaces de

operar en ambientes altamente dinámicos.

Las situaciones antes mencionadas motivan la búsqueda de mecanismos para poder seleccionar únicamente los datos relevantes para el usuario, siendo éste una persona, una aplicación, un dispositivo, etc. Esto significa poder seleccionar la información requerida o "preferida" por el usuario, por lo que inmediatamente surge el cuestionamiento de cuáles son y cómo se determinan dichas preferencias. En principio, podría decirse que los datos preferidos pueden ser determinados en base a las tareas que ya han sido realizadas por el usuario en cuestión o por tareas que han realizado usuario con características similares. En todo caso, el ambiente en el cual se mueve el usuario siempre es necesario para determinar los datos que son de interés para el mismo. Inclusive, en [40] subrayan que la misma información puede ser considerada diferente, para un mismo usuario, en situaciones o lugares diferentes. Por lo tanto, los datos dependen del ambiente de trabajo y, según los autores de [40], el ambiente de trabajo está determinado por el contexto del usuario.

El contexto [39] [40] [42] [46] y las preferencias del usuario [39] [47], surgen como la solución para manejar los problemas mencionados al principio, buscando presentar información relevante para el usuario. Según [39], el contexto es la posibilidad de seleccionar datos según el ambiente en el cual se encuentra el usuario. En [40] consideran el contexto como un conjunto de variables que son de interés e influencian las acciones de un agente. Por otro lado, lo definen como las circunstancias o la situación en la cual una tarea informática [42] o un cálculo [45] es llevaba a cabo. En sistemas de Computación Ubicua, los sensores son comunmente empleados en la captura de datos ambientales. Por esta razón, el trabajo presentado en [48] introduce el concepto de contexto sensado y lo define como las propiedades que caracterizan a un fenómeno y que son potencialmente relevantes para cierta tarea. Además, agregan que el contexto por sí solo, es cualquier información que se pueda utilizar para caracterizar la situación de una entidad y que es considerada relevante para la interacción entre un usuario y una aplicación.

En [49] analizan distintas definiciones del concepto "Contexto", tomadas de la Web, y las distintas problemáticas acerca de la comprensión del contexto. Los autores consideran difícil encontrar una definición relevante que satisfaga a cualquier disciplina. Por esto surgen los siguientes planteos:

- ¿El contexto es un marco para un objeto determinado?
- ¿El contexto es el conjunto de elementos que tienen influencia sobre algún objeto?
- ¿Es posible definir el contexto a priori o sólo observar sus efectos a posteriori?
- ¿El contexto es estático o dinámico?

- ¿Cuál es el contexto relevante? ¿El contexto de la persona, de la tarea, de una situación dada?

Una vez analizados estas y otras preguntas, los investigadores concluyen que el contexto actúa como un conjunto de restricciones que influyen en el comportamiento de un sistema (una persona, una aplicación, un dispositivo, etc.) embebido en una determinada tarea.

Tal cual lo destacan los autores de [49], las definiciones antes presentadas no coinciden exactamente, lo que resalta las distintas formas de percibir el concepto "Contexto". Sin embargo, coinciden en la importancia de detectar los datos relevantes. Más allá de la gran variedad de conceptos, como mencionan en [42], aún se ofrecen pocas pistas acerca de las propiedades que son interesantes destacar a la hora de modelar el contexto. No es trivial la tarea de modelar, ya que es vital determinar cómo se obtendrán y cuáles serán los datos más relevantes para un usuario o para un grupo de usuarios. Por esta razón, se destacan los aspectos y características más importantes en el modelado de contextos, de acuerdo a la bibliografía consultada. En primer lugar, el trabajo presentado en [46] realiza un reporte en el que no sólo distinguen aspectos a tener en cuenta en el modelado, sino que también realiza la comparación de algunos *frameworks* propuestos para el modelado de contextos. Dichos *frameworks* pueden ser consultados en [41] [43] [44]. Por ejemplo, el tiempo y el espacio son aspectos generalmente tenidos en cuenta, los datos históricos también son parte importante del contexto, ya que muchas veces el estado del contexto actual depende fuertemente de contextos anteriores. Además de los datos necesarios para definir un contexto, es clave identificar quién o qué será el sujeto del mismo. Algunos modelos se definen de acuerdo a la percepción del usuario, mientras que otros consideran que el usuario forma parte del contexto. Finalmente, aparecen nuevamente las preferencias y características personales del usuario, también denominadas perfil del usuario y, en este caso, el modelo debe especificar si el contexto está basado en un único usuario, en un grupo de usuarios o en un rol.

Por otro lado, los autores de [43] agregan una clasificación para los contextos. Distinguen el caso en el cual el contexto es adquirido directamente del proveedor, del caso en el que el contexto se deduce a partir del contexto directo. Un ejemplo de esta clasificación considera un conjunto de parámetros, en el cual el estado actual de una persona (contexto directo: duchándose) puede ser inferido a partir de su ubicación (contexto indirecto: baño). Otro factor a destacar es la naturaleza de la información, como especifican en [42], ésta depende fuertemente de los sistemas en los cuales será utilizada. En particular, en Computación Ubicua, la naturaleza de la información es determinante para los requerimientos de diseño en el modelo de contexto. La característica temporal de la información siempre está presente, ya que los datos del usuario pueden ser estáticos (fecha de nacimiento) o dinámicos (ubicación). Estos conceptos también son considerados por

los autores de [43]. Otros aspectos importantes de la información son su imperfección y su característica de interrelacionamiento. Imperfecta porque parte del contexto puede ser desconocido e incorrecto, si no refleja la realidad representada, o inconsistente si tiene contradicciones. Fuertemente interrelacionada porque existen muchas relaciones entre las personas, sus dispositivos y sus canales de comunicación.

En base a la variedad de aspectos que pueden ser tenidos en cuenta en el momento de definir el modelo de un contexto y pensando en cuál será el sujeto del contexto, se presenta un listado con diferentes alternativas:

- **Contexto orientado a la presentación**: En este caso el contexto se percibe como la capacidad que tiene un sistema para adaptarse a la presentación de contenidos, según el usuario o dispositivo [40] [46].
- **Contexto orientado a la ubicación**: En este grupo de modelos las coordenadas de tiempo y espacio son los párametros más importantes [40]. En [46] también son llamados contextos de localización y entorno.
- **Contexto centrado en el usuario**: En este caso el contexto está basado en lo que está haciendo el usuario. En muchos casos se busca deducir la actividad del usuario teniendo en cuenta la historia del contexto o a través de registros de sensores [40].
- **Contexto basado en comunidades**: En esta propuesta el contexto es considerado como un conjunto de variables que son compartidas por un grupo de usuarios [40]. El enfoque presentado en [46], para este caso, está centrado en el problema de lograr un acuerdo sobre un contexto compartido entre pares.
- **Contexto basada en datos a medida**: Esta perspectiva se basa en el concepto de *data tailoring* y es introducida en [40]. Los autores apuntan a la reducción del tamaño de los datos utilizando preferencias contextuales, considerando únicamente datos relevantes para obtener un subconjunto personalizado de la información disponible. Por otro lado, en [46] extienden esta definición y consideran, no sólo datos relevantes, sino también funcionalidades y servicios.

Una vez especificado el modelo de contexto, interesan las características de representación del mismo [46]:

- **Formalismo**: Indica qué tipo de formalización es aplicada, una ontología, grafos, etc.
- **Nivel de formalidad**: Si considera una definición formal, para la representación del modelo de contexto.

- **Flexibilidad**: Esta característica indica si el modelo es adaptable a diferentes contextos.
- **Granularidad**: Indica la capacidad del modelo de representar un contexto en distintos niveles de detalle.
- **Restricciones de validación**: Determinan la posibilidad de reducir el número de contextos admisibles, mediante la imposición de restricciones semánticas que deben satifacer los contextos para una aplicación dada.

Las características que van a ser priorizadas en un modelo en particular, dependen de los problemas que interesen resolver. En [39], dado un conjunto de preferencias, los autores investigan cómo se propagan éstas en una jerarquía de contextos y el impacto que dicha propagación tiene sobre las consultas a una base de datos. El proceso de propagación consiste en que los contextos más específicos, dentro de cierta jerarquía, prevalecen sobre los menos específicos, siempre y cuando los contextos sean comparables. Es decir, cuanto más abajo en la jerarquía, más específico es el contexto. En este caso, cada contexto está representado por un conjunto finito de dimensiones contextuales y cada una de ellas está descrita por un conjunto de niveles que determina la granularidad del contexto. En [43] [44] ponen énfasis en la formalidad de la representación, basando el modelo de contexto en una ontología para representar, manipular y acceder a la información del contexto. La necesidad de considerar el contexto de la información en las bases de datos continúa presente en [40]. En este caso, la principal característica del modelo utilizado también es la granularidad, ya que plantean un árbol de dimensiones del contexto, donde cada dimensión captura una característica diferente del mismo. En la etapa de diseño, todos los posibles contextos deben ser identificados. Además, se debe determinar el conjunto de datos que es relevante para cada actor en cada contexto antes definido.

Por otro lado, muchas de las investigaciones basadas en el uso de contextos consideran inevitable tener en cuenta la calidad de los datos que determinan a dichos contextos. Al pensar en calidad de datos y contextos surge el siguiente cuestionamiento: ¿cómo determinar la calidad de los datos si inclusive ésta podría depender del contexto? [50]. [42] es un ejemplo de este tipo de trabajos, donde los autores afirman que los errores en la información de contexto aparecen porque generalmente los datos son derivados de sensores y, muchas veces, los datos sensados son erróneos, por lo que las aplicaciones deben tener la capacidad de evaluar si los datos son confiables o no. Por esta razón, observan que es necesario incorparar medidas de calidad en los modelos de contexto. En [45] plantean que los datos de un contexto generalmente involucran entidades del mundo real, por lo que tiene sentido medir la calidad de dichos datos o el grado en que éstos se corresponden con la realidad.

Otro ejemplo es el trabajo presentado en [51], en el cual proponen un modelo para la evaluación de la calidad de los datos desde la perspectiva del usuario. Destacan que dado que la calidad es *"fitness for use"* o adecuada para su uso, en el momento de la evaluación de las dimensiones de la calidad de los datos, se debe considerar el grado de satifacción del usuario. Presentan un ejemplo en el cual consideran la dimensión de calidad completitud y una base de datos en la cual podrían estar faltando datos. La base de datos podría ser considerada completa por algunos usuarios e incompleta por otros, de acuerdo a los requerimietnos de los mismos. Por ésto, definen el concepto de "perfil de usuario", como una representación de la información que describe al usuario y sus preferencias. A partir de su perfil, el usuario es asignado a una clase, donde cada clase contiene usuarios con características similares.

2.3. Conclusiones

Con esta primer revisión bibliográfica, si bien es posible tener una noción general de la investigación actual acerca de la Calidad de Datos en los SDW, no es suficiente para identificar los desafíos más importantes que se presentan en el área. Por otro lado, se observa que varios trabajos, más allá del foco de su investigación, resaltan la importancia del usuario y/o las características que lo determinan (sus preferencias, la tarea que éste realiza, su perfil, etc.). Por esta razón, en esta tesis surge la necesidad de incorporar un nuevo concepto: Contexto.

En el análisis de la bibliografía se identificaron contextos orientados a la presentación, la ubicación, el usuario, las comunidades, etc., y se observó que los mismos son objetos de numerosos trabajos científicos, donde son considerados con diferentes enfoques. Por lo tanto, se plantea el propósito de entender cómo son definidos y usados los contextos en los Sistema de *Data Warehousing*, en particular, para la evaluación de la calidad de los datos presentes en este tipo de sistemas.

Estos cuestionamientos generan la necesidad de realizar una revisión bibliográfica más exhaustiva, que permita conocer los desafíos actuales del área Calidad de Datos en los Sistema de *Data Warehousing* y cómo son aplicados los contextos (si son aplicados) en dicha área. Por esta razón, se plantea la aplicación de una metodología que guíe en la búsqueda de la bibliografía, que no sólo permita identificar los desafíos presentes en las áreas de interés, sino que sea rigurosa y ofrezca la posibilidad de reproducir las búsquedas realizadas.

Se plantea la siguiente interrogante:

¿Cómo pueden ser usados los contextos para evaluar la calidad de datos en *Data Warehouse*?

A partir de esta pregunta se define y ejecuta la metodología de búsqueda bibliográfica. La aplicación de dicha metodología es presentada en el siguiente Capítulo.

Capítulo 3

Aplicación de una metodología de búsqueda

A partir de la primer revisión bibliográfica, se obtienen varios trabajos de las tres áreas de interés, *Data Quality*, *Data Warehouse* y *Context*. En la mayoría de los casos, son trabajos que relacionan calidad de datos y Data Warehouse, pero sin tener en cuenta el contexto. En otros casos, profundizan en el uso de los contextos, en un entorno particular o en casos generales. Por todo esto, no se localizaron trabajos que investiguen o relacionen a estos tres grandes temas. Las características y propiedades que cada tema ofrece, analizados en el Capítulo 2, apoyan el enfoque de esta tesis, en el cual se considera posible relacionarlos entre sí. Por esto, surge la inquietud de analizar de forma ordenada y exhaustiva, la existencia de trabajos que se enfoquen en estas tres áreas de investigación. En particular, como se observa en el Capítulo 2, muchos trabajos se enfocan en el análisis y/o evaluación de la calidad de los datos en los SDW, sin embargo, no se encontraron trabajos que tuvieran en cuenta a los contextos en este tipo de trabajos. Con la motivación de saber si existen investigaciones que consideran contextos para la evaluación de la calidad de datos en un SDW y cuáles son éstas, se plantea la necesidad de utilizar algún método que permita una visión general de la investigación actual en estos temas.

Por lo tanto, se toma la decisión de aplicar una metodología de búsqueda en las áreas temáticas de interés: *Data Quality*, *Data Warehouse* y *Context*. En primer lugar, se considera la realización de una revisión sistemática (de su nombre en inglés, *Systematic Literature Review* o *Systematic Review*, SLR), tradicionalmente utilizada en el área de la Medicina y, según los autores de [52], muy poco aplicada en el área de la Informática (o *Computer Science*). La mayoría de las revisiones sistemáticas, en el área de la Informática, se presentan en trabajos relacionados con la Ingeniería de Software, [53] [54] [55]. Sin embargo, también en [56] muestran los resultados de una revisión sistemática sobre las distintas propuestas de evaluación de la calidad de los datos de *Linked Open Data*. Por

otro lado, [57, A60][1] es un trabajo muy reciente que presenta una revisión de la literatura de métricas de calidad para modelos de datos de DW. Si bien los autores presentan una SLR como metodología de búsqueda, no incluyen un análisis profundo para respoder las preguntas científicas que guían las búsquedas, ni evalúan la calidad de los trabajos seleccionados en la revisión, tal como lo exige la SLR (Tabla 3.1). Sin embargo, interesa destacar el uso de este tipo de metodologías en áreas cercanas a esta tesis. En particular, se destaca que los autores concluyen, a partir de la revisión de la bibliografía, que existe una falta de validación de métricas y factores de calidad para los modelos de datos conceptuales de DW. Otro ejemplo, es el trabajo presentado en [58], donde realizan una revisión sistemática sobre la gestión de la calidad de datos guiada por procesos. Según los autores, con los procesos se busca identificar las causas de los errores, eliminarlos y sostener las mejorías logradas en el largo plazo.

En [53] definen esta metodología como un estudio secundario, que se basa en el análisis de investigación previa y que es utilizada para encontrar, evaluar y unir todos los artículos de investigación (o estudios primarios), que son relevantes para un tema o pregunta de investigación específica. El objetivo de este tipo de estudios es garantizar que la revisión de la literatura es imparcial, rigurosa y reproducible. Posteriormente, durante el estudio de la bibliografía referida a la SLR, surge el concepto *Mapping Study* (también llamado *Systematic Mapping Study* o *Scoping Study*, MS) como otro tipo de revisión que complementa la SLR. También en [53], señalan que el MS es una revisión que utiliza la misma metodología que la SLR, cuyo objetivo es identificar y clasificar toda la investigación relacionada con un tema. Además, agregan que el MS proporciona una visión general de un área temática e identifica si hay temas con suficientes estudios primarios como para realizar una SLR. A su vez, identifica temas en los que se necesitan más estudios primarios y proporciona una visión global de la literatura. De esta forma, permite a los nuevos investigadores conocer qué hay que leer en el área de interés y cuáles son los autores más importantes. Según los autores de [59], hay un número importante de MS, pero poca discusión acerca de su valor como herramienta de investigación. Ejemplos de trabajos que presentan un MS son [60] [61] [62] [63] [64].

Dados los objetivos de las dos metodologías de revisión bibliográfica, SLR y MS, se analizan las diferencias más importantes que éstas presentan. De tal forma, es posible analizar cuál de las revisiones, con sus características, objetivo y alcance, es la que mejor se ajusta a las necesidades de esta tesis. En [53] [59], los autores contrastan sus características, las principales diferencias se destacan en la Tabla 3.1. Si bien cada metodología tiene objetivos diferentes, en casos excepcionales se solapan. Por ejemplo, en [59] presentan un MS que además, incluye una

[1]Los artículos seleccionados siguiendo la metodología de búsqueda presentan, a la izquierda el número de referencia y a la derecha una etiqueta con el número asignado en el momento de la selección.

evaluación de los resultados de los estudios primarios. Por otro lado, destacan que algunas SLR incluyen un sistema de clasificación para ordenar la literatura obtenida.

La desventaja que presentan estas metodologías es que requieren de un esfuerzo mayor que las revisiones bibliográficas tradicionales. Sin embargo, la metodología bien definida hace que los resultados de la literatura estén menos sesgados, aunque no se pueda asegurar el sesgo de los estudios primarios. Por otro lado, la documentación que permite generar la estrategia de búsqueda hace que los lectores puedan evaluar su rigurosidad y qué tan exhaustiva es la misma. A su vez, es muy importante destacar que con dicha documentación es posible reproducirla.

Por lo tanto, en base a los conceptos antes presentados y de acuerdo al alcance de esta tesis, se decide aplicar un *Mapping Study*. Este tipo de revisión tiene un amplio alcance y su objetivo se ajusta a la necesidad que se presenta a partir de la primera revisión bibliográfica: clasificar y analizar la literatura en áreas temáticas determinadas.

Un *Mapping Study* implica varias etapas y las mismas se presentan a continuación:

- Definición de la *Research question*: Las *Research questions* son las preguntas que se plantean inicialmente para guiar la búsqueda de artículos o estudios primarios, como son llamados en la metodología.
- Creación de las cadenas de búsqueda: Para poder crear las cadenas de búsqueda, es esencial determinar las palabras claves, que surgen a partir de las *Research questions* antes planteadas. Además, es necesario seleccionar un conjunto de términos alternativos a las palabras claves que determinan, mediante la aplicación de conectivas booleanas, las cadenas de búsqueda.
- Selección de las bibliotecas digitales: Se seleccionan las bibliotecas digitales sobre las cuales se realizan las búsquedas. En el caso del MS, es un número reducido de bibliotecas.
- Desarrollo de la estrategia de búsqueda: En esta etapa se deciden los criterios de inclusión y exclusión que determinan la selección de los artículos encontrados a partir de las cadenas de búsqueda.
- Ejecución de la metodología: Se realizan las búsquedas, aplicando las cadenas de búsqueda definidas, sobre las bibliotecas digitales seleccionadas. Una vez que los artículos son devueltos, se aplican los criterios de exclusión e inclusión para la selección de los mismos.

Una vez que fueron ejecutadas todas las etapas del MS, se realiza un análisis por *Research question*. Para esto, se clasifican los artículos obtenidos a partir de la

	MS	**SLR**
Objetivo	Clasificación y análisis temático de la literatura en un área.	Identificación de las mejores prácticas en relación a los procedimientos específicos, tecnologías, métodos o herramientas.
Research questions	Más amplias y genéricas, relacionadas con tendencias en la investigación. Son del tipo "¿Cuáles investigadores?" "¿Qué tipo de estudios?". En general, múltiples preguntas.	Específicas. Relacionadas a estudios empíricos. Son del tipo "¿La tecnología/método A es mejor o no que el B?"
Proceso de búsqueda	Definido por área temática.	Definido por las research questions.
Alcance	Amplio. Todos los trabajos referidos a un área temática son incluidos.	Enfocado. Se incluyen sólo trabajos empíricos relacionados con una pregunta de investigación específica.
Requerimientos de la estrategia de búsqueda	Menos estrictos, sobretodo cuando lo que más interesa son las tendencias en la investigación de cierta área. Por ejemplo, sólo artículos de journals, se consideran una o dos librerías digitales, etc.	Muy estricto. Deben encontrarse todos los estudios pertinentes. Por lo general, también es necesario mirar las referencias en los estudios primarios identificados y/o contactar a investigadores para averiguar si nuevas investigaciones se están llevando a cabo en el área.
Etapa de análisis	Resume los datos para responder las *Research questions*.	Incluye un análisis profundo.
Evaluación de la calidad	No es esencial. Por la naturaleza de inclusión de la búsqueda, que considera tanto estudios teóricos como estudios empíricos, hace complicada la evaluación de la calidad de los estudios primarios.	Es importante asegurar que los resultados se basan en evidencia de calidad.
Resultados	Un conjunto de trabajos relacionados con un área temática que se categorizan en una variedad de dimensiones. Además, se presenta el número de artículos en cada categoría.	Los resultados de los estudios primarios se unen para responder la(s) reasearch question(s).

Tabla 3.1: **Diferencias entre MS y SLR**

ejecución del MS. La clasificación de los trabajos se realiza en base a la *Research question* que intentan responder. Finalmente, se extraen y analizan una serie de resultados, que si bien no se desprenden directamente de las *Research questions* planteadas, son de gran interés para esta tesis.

A continuación se desarrolla cada una de las etapas de la metodología de búsqueda: *Mapping Study*

3.1. Definición de las *Research questions*

El objetivo de este Mapping Study es encontrar trabajos científicos en los que se haya definido el contexto para la evaluación de la calidad de los datos en un SDW. Para esto, se plantea la pregunta general o principal, que motiva a la investigación. La metodología de búsqueda la denomina *Research Question* (RQ). Para este trabajo, la RQ es la que se muestra a continuación:

RQ: ¿Cómo pueden ser usados los contextos para evaluar la calidad de datos en *Data Warehouse*?

Dado que la pregunta RQ es muy específica se considera, sin perder el foco de la investigación, la posibilidad de plantear cuestionamientos más generales que den como resultado un conjunto de respuestas más amplio. Por lo tanto, una vez identificados los temas que determinan la RQ, se definen las *Research Questions* parciales. Los temas de interés para esta investigación, que surgen de la RQ, son *Context*, *Data Quality* y *Data Warehouse*. Las *Research Questions* parciales son RQ1, RQ2 y RQ3, las mismas se presentan en la Tabla 3.2 y son las que conducen al Mapping Study.

***Research questions* parciales**	**Resultados esperados**
RQ1: ¿Cómo son usados y definidos los contextos en los SDW?	Encontrar definiciones de "contextos", sus componentes, sugerencias de modelos y sus diferentes formas de representación. Investigar si el concepto "contexto" ha sido definido y cómo, en un entorno de DW.
RQ2: ¿Cómo se maneja la calidad de los datos en los sistemas de DW?	Encontrar trabajos que tengan en cuenta la calidad de los datos en un ambiente de DWs.
RQ3: ¿Cómo se consideran los contextos para la evaluación de calidad de datos?	Encontrar trabajos que utilicen contextos en la evaluación de la calidad de los datos.

Tabla 3.2: ***Research questions* parciales**

3.2. Creación de las cadenas de búsqueda

Como se mencionó anteriormente, los temas que motivan esta investigación son *Context*, *Data Quality* y *Data Warehouse*. Por lo tanto, estas son las **palabras claves** a partir de las cuales se realizan las búsquedas de los estudios primarios. Es importante destacar, que las palabras claves están en inglés porque las búsquedas son realizadas en inglés. Por esta misma razón, tanto la *Research question* principal RQ, como las *Research questions* parciales RQ1, RQ2 y RQ3, también son definidas en inglés:

- **RQ**: *How contexts can be used for assessing DQ in DWs?*
- **RQ1**: *How contexts are defined and managed in a DW environment?*
- **RQ2**: *How DQ is managed in DW systems?*
- **RQ3**: *How contexts are considered for the assessment of DQ?*

Es necesario realizar una búsqueda exhaustiva y por esta razón, las palabras claves por sí solas no son suficientes para lograr un número de trabajos significativo. Por lo tanto, se seleccionan otras palabras, las cuales son sinónimos y palabras alternativas a las palabras claves. La metodología de búsqueda denomina a este nuevo grupo de palabras **términos alternativos**. En la Tabla 3.3 se muestra el conjunto de términos alternativos seleccionado para cada una de la palabras claves.

Palabras claves	Términos alternativos
Context	*data tailoring, pervasive computing, ubiquitous computing, preference*
Data Quality	*information quality, quality factor, quality metric, quality dimension, quality measure, quality attributes*
Data Warehouse	*warehousing, business intelligence, multidimensional database, dimension hierarchies, fact table, cube*

Tabla 3.3: **Términos alternativos a las palabras claves**

Una vez identificados los términos alternativos a las palabras claves, se crean las **cadenas de términos**. Estas cadenas enlazan todos los términos alternativos que se relacionan entre sí y esto se realiza mediante el uso de la conectiva booleana OR. Esto se muestra en la Tabla 3.4.

Finalmente, se define la **cadena de búsqueda principal** y se construye uniendo las cadenas de términos relacionados, definidos en la Tabla 3.4, mediante la conectiva booleana *AND*. La cadena de búsqueda principal (SS) se muestra en la Tabla 3.5.

Palabras claves	Cadenas de términos
Context	*data tailoring OR pervasive computing OR ubiquitous computing OR preference*
Data Quality	*information quality OR quality factor OR quality metric OR quality dimension OR quality measure OR quality attributes*
Data Warehouse	*warehousing OR business intelligence OR multidimensional database OR dimension hierarchies OR fact table OR cube*

Tabla 3.4: **Cadena de términos por palabra clave**

(Context** OR data tailoring OR pervasive computing OR ubiquitous computing OR preference**) AND (Data Quality** OR information quality OR quality factor OR quality metric OR quality dimension OR quality measure OR quality attributes**) AND (Data Warehouse** OR warehousing OR business intelligence OR multidimensional database OR dimension hierarchies OR fact table OR cube**)

Tabla 3.5: **SS. Cadena de búsqueda principal**

3.3. Selección de las bibliotecas digitales

Luego de la construcción de la cadena de búsqueda principal, se seleccionan las **bibliotecas digitales**, sobre las cuales se realiza la búsqueda de trabajos científicos. Es necesario limitar el alcance, por lo que se selecciona un número reducido de bibliotecas, siendo esto además, una característica de los MS. La elección de las bibliotecas digitales es realizada en base a la relación de las mismas con el área de investigación y de acuerdo con el correcto funcionamiento de las funcionalidades que sus motores de búsqueda proveen. Además, es importante que dichos motores permitan reproducir exactamente las cadenas de búsqueda o, en el peor de los casos, que permitan expresar las cadenas de la forma más parecida posible a las cadenas originales. Por otro lado, interesa la posibilidad de utilizar filtros en las búsquedas, limitando así el número de trabajos devueltos.

Las bibliotecas digitales seleccionadas, para la realización del *Mapping Study*, son las que se presentan en la Tabla 3.6. Se considera de interés realizar a futuro un trabajo más abarcativo respecto al número de bibliotecas digitales. Para esto, las búsquedas deberían ser realizadas por un grupo de investigadores, organizando varios grupos de trabajo, donde cada uno de ellos podría estar enfocado en una búsqueda determinada, definida por área, *Research question*, etc. De esta forma, podría realizarse una SLR como complemento de este MS.

ACM Digital library	http://dl.acm.org/
IEEE*Xplore*	http://www.ieeexplore.ieee.org
ScienceDirect	http://www.sciencedirect.com/

Tabla 3.6: **Bibliotecas digitales**

Las bibliotecas digitales que no cumplen alguna de las condiciones necesarias antes mencionadas, más allá de la importancia de las mismas, no son consideradas. Un ejemplo de esto es la base de datos interactiva *Springer*, que si bien es de gran importancia para el área, presenta una serie de dificultades a la hora de realizar las búsquedas. Dichas dificultades son las que se presentan a continuación:

- la construcción de las cadenas de búsqueda, mediante las conectivas booleanas *AND* y *OR*, no funciona correctamente.
- el uso de las comillas (“ ”), para la construcción de las cadenas de búsqueda, no funciona correctamente.
- la búsqueda debe realizarse en el contenido de los artículos solamente, sin poder considerar el título ni el resumen (*abstract*).
- el número de trabajos devueltos es muy grande en todos los casos, sin poder aplicar todos los filtros necesarios para acotar dicho número.

3.4. Desarrollo de la estrategia de búsqueda

Cuando la cadena de búsqueda principal ha sido definida y las bibliotecas digitales han sido seleccionadas, es necesario limitar la búsqueda. Para ésto se definen los **criterios de inclusión y exclusión** que son los que permiten decidir qué artículos serán finalmente seleccionados. Además, artículos que son seleccionados en las búsquedas como de interés para la investigación pueden, una vez analizados, ser considerados irrelevantes para el trabajo en desarrollo. Es por todo esto que se plantea un conjunto de criterios de inclusión y de exclusión que se presentan en la Tabla 3.7.

3.5. Ejecución del *Mapping Study*

Intentando responder las *Research questions* parciales planteadas inicialmente y dada la longitud de la cadena principal SS, se decidió definir a partir de ésta, nuevas cadenas de búsqueda parciales, denominadas SS1, SS2 y SS3.

Para definir las cadenas SS1, SS2 y SS3 se adapta y utiliza el método de intersección de conjuntos utilizado en [52]. El mismo considera el conjunto de todos

Criterios	**Condiciones**
Inclusión	Publicaciones realizadas en el rango 2008 - 2015 Idioma del artículo: Inglés Tipo de publicación: Journal, Conferencia, Workshop o Libro Tipo de trabajo: artículo o capítulo de Libro Artículo completo disponible Artículo gratuito
Exclusión	El trabajo sólo está enfocado en DW y no hace referencia ni a DQ ni a CTX El trabajo sólo está enfocado en DQ y no hace referencia ni a DW ni a CTX El trabajo sólo está enfocado en CTX y no hace referencia ni a DQ ni a DW Los términos "contexto" y/o "calidad" son usados con otro sentido Las palabras claves aparecen en el *abstract* del artículo, pero no aparecen en el resto del trabajo Trabajos en formato ppt o poster El artículo aparece duplicado en la selección El artículo no es gratuito El artículo aparece como "*in press*"

Tabla 3.7: **Criterios de inclusión y exclusión**

los trabajos relacionados con la palabra clave *Data Warehouse*, el conjunto de todos los trabajos relacionados con la palabra clave *Data Quality* y el conjunto de todos los trabajos relacionados con la palabra clave *Context*. Estos tres grupos se diseñan para recuperar diferentes conjuntos de la literatura relevante. El objetivo principal es encontrar la literatura que se encuentra en la intersección de los tres conjuntos. En principio, se hace foco en la intersección dos a dos de estos conjuntos, donde cada intersección determina el conjunto de trabajos relacionados con 2 de las palabras claves. Por lo tanto, se define a SS1 como la cadena de búsqueda que devuelve todos los trabajos correspondientes con *Data Warehouse* y *Context*, que es el conjunto de trabajos que pertenecen a la intersección de los conjuntos de la literatura deterninada por las palabras claves *Data Warehouse* y *Context*, junto a todos sus términos alternativos. De igual forma se definen SS2 y SS3, y se explica gráficamente en la Figura 3.1. De esta forma, los trabajos hallados con cada una de las cadenas de búsqueda parciales, intentan responder las *Research questions* parciales. Finalmente, se observa que los artículos que deberían ser devueltos por la cadena de búsqueda principal, SS, son aquellos artículos que pertenecen a la intersección de los tres conjuntos de trabajos que se relacionan con cada una de las palabras claves.

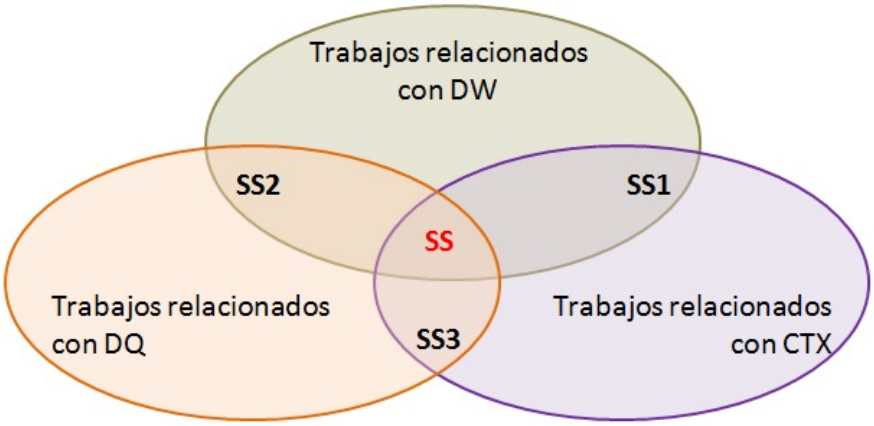

Figura 3.1: **Definición de las cadenas de búsqueda parciales**

A continuación, en las Tablas 3.8, 3.9 y 3.10 se presentan cada una de las cadenas de búsqueda parciales.

(Context** OR data tailoring OR pervasive computing OR ubiquitous computing OR preference**) AND (Data Warehouse** OR warehousing OR business intelligence OR multidimensional database OR dimension hierarchies OR fact table OR cube**)

Tabla 3.8: **SS1. Cadena de búsqueda parcial**

(Data Quality** OR information quality OR quality factor OR quality metric OR quality dimension OR quality measure OR quality attributes**) AND (Data Warehouse** OR warehousing OR business intelligence OR multidimensional database OR dimension hierarchies OR fact table OR cube**)

Tabla 3.9: **SS2. Cadena de búsqueda parcial**

La metodología es aplicada en dos etapas, la primer etapa considera trabajos publicados desde el año 2008 hasta el año 2014, mientras que la segunda etapa considera trabajos desde el año 2014 hasta el año 2015. La primer etapa se ejecuta en el período marzo-junio del año 2014 y la segunda en octubre de 2015. Esto es así porque dado que la revisión bibliográfica es finalizada en junio de 2014 y la tesis se extiende hasta octubre de 2015, se considera importante hacer una revisión del estado del arte para el período 2014-2015, ya que a más de un año de realizada la última búsqueda, podrían haber surgido trabajos de mucho interés para este trabajo.

***(Context** OR data tailoring OR pervasive computing OR ubiquitous computing OR preference)* ***AND (Data Quality** OR information quality OR quality factor OR quality metric OR quality dimension OR quality measure OR quality attributes)*

Tabla 3.10: **SS3. Cadena de búsqueda parcial**

Para la realización de la búsqueda de trabajos se adapta cada cadena parcial a las funcionalidades que ofrecen los motores de búsqueda de las bibliotecas digitales que fueron seleccionadas. Dado que cada biblioteca ofrece diferentes niveles de búsqueda fue necesario, en cada uno de los casos, armar cadenas de búsqueda acordes con las posibilidades ofrecidas, respetando la estructura de la cadena original. En el Apéndice A se pueden consultar las cadenas de búsqueda ejecutadas en cada una de las bibliotecas digitales.

Una vez finalizadas las búsquedas, se obtienen 619 trabajos. Posteriormente, se aplican los criterios de inclusión y exclusión y se seleccionan 86 trabajos. Al finalizar la etapa de análisis, se descartan aquellos artículos que no se corresponden con los temas de interés, obteniendo un total de 64 trabajos. Cuando los artículos verifican los criterios de inclusión, se busca evaluar cuán relacionados están con las palabras claves, de forma tal que permitan responder las *Research questions* parciales que guían la metodología. Por ejemplo, todos aquellos trabajos basados en calidad de esquemas son descartados. Esto es así porque, si bien esta investigación está centrada en "calidad", interesan específicamente aquellos trabajos que se enfocan en la calidad de los datos. Por otro lado, varios trabajos hacen referencia a los SDW y, aunque dichos sistemas son mencionados, los trabajos no están desarrollados basándose en los mismos. Situaciones similares se presentan para el caso de los contextos y de la calidad de los datos. Por lo tanto, este tipo de artículos no es considerado relevante para la investigación. Para llevar a cabo la selección de trabajos, en primer lugar, se consulta el resumen o *abstract* del mismo y cuando éste se considera relevante se prosigue con la lectura del contenido. Sin embargo, para los casos en que el resumen no contiene información suficiente, como para decidir la elección o no del trabajo, se consultan las conclusiones y se analiza del uso de las palabras claves dentro del contenido del documento. Si el analisis devuelve un resultado relevante se prosigue con la lectura completa del trabajo, de lo contrario, el trabajo es descartado.

Por otro lado, se seleccionan artículos que, si bien sus objetivos no son exactamente los temas que motivan a esta tesis, cumplen con los criterios de inclusión y, además, se apoyan fuertemente en dos o todos los temas de interés. Por ejemplo, algunos trabajos se basan en los conceptos multidimensionales de los sistemas de DW y de las aplicaciones OLAP. Otros no mencionan explícitamente el concepto de calidad de datos, pero destacan la necesidad de obtener resultados de análisis

	ARTÍCULOS							
	ACM Digital library		IEEEXplore		ScienceDirect		TOTAL	
	devueltos	seleccionados	devueltos	seleccionados	devueltos	seleccionados	devueltos	seleccionados
SS1	58	7	176	16	53	4	287	27
SS2	10	1	57	14	20	5	87	20
SS3	44	2	114	12	87	3	245	17
							619	64

Tabla 3.11: **Resultados por bibliotecas digitales** (con duplicados)

sólidos y con valor. También en el caso de los contextos se observan casos similares, por ejemplo autores que no mencionan al contexto, pero hacen referencia a la importancia de considerar las necesidades y las preferencias de los usuarios. Por lo tanto, en base a esto, dichos trabajos son seleccionados y analizados, ya que los resultados obtenidos a partir de ellos, ayudan a responder directa o indirectamente alguna de las *Research questions* parciales.

En la Tabla 3.11 se observa en detalle la cantidad de artículos devueltos para cada *Research question* parcial y la cantidad de trabajos seleccionados en cada caso, para cada una de las bibliotecas digitales. Estos resultados se representan gráficamente por *Research question* parcial, como se muestra en la Figura 3.2 y por biblioteca digital, como se presenta en la Figura 3.3. Una vez finalizada la tarea de remoción de artículos irrelevantes, se realiza la búsqueda de artículos duplicados, lo que permite eliminar 2 artículos más. Por lo tanto, el número final, de artículos seleccionados, es 62. Este número representa el 10 % de los artículos devueltos en las búsquedas iniciales. En la Figura 3.4 se presenta gráficamente el proceso de selección completo, descripto anteriormente.

En la Tabla 3.12 se observan los trabajos seleccionados para cada cadena de búsqueda parcial. El artículo [65, A20][2] es el único trabajo encontrado en las búsquedas realizadas con las tres cadenas SS1, SS2 y SS3. En los tres casos, [65, A20] fue devuelto por la biblioteca digital IEEE*Xplore*. En julio de 2014, se realiza una búsqueda con la cadena principal SS, en cada una de las bibliotecas digitales de la Tabla 3.6, para corroborar el resultado obtenido con las cadenas de búsqueda parciales. Efectivamente, el artículo [65, A20] es devuelto por todos los motores de búsqueda de todas las bibliotecas, con la cadena principal SS. Esto se refleja en la Tabla 3.12. En octubre de 2015, en la segunda etapa de la revisión, no se realiza la búsqueda con la cadena de búsqueda principal SS, en todos las bibliotecas digitales. Se toma esta decisión porque ninguno de los trabajos seleccionados en esta etapa fue devuelto por las tres cadenas de búsqueda

[2]Los artículos seleccionados a partir del MS presentan, a la izquierda el número de referencia y a la derecha una etiqueta con el número asignado en el momento de la selección.

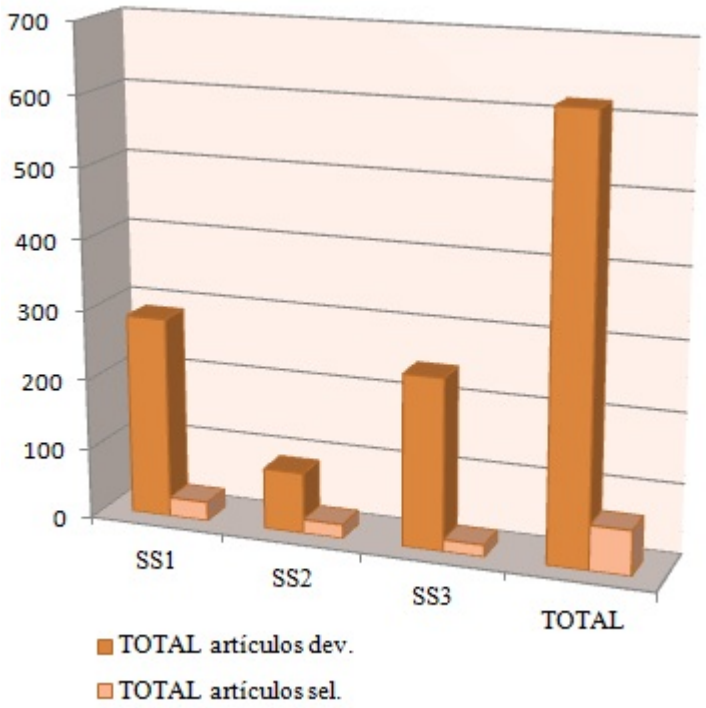

Figura 3.2: **Resultados por cadena de búsqueda parcial**

parciales SS1, SS2 y SS3, al mismo tiempo. Por tanto, considerando que cada una de las cadenas de búsqueda parciales es menos restrictiva que la cadena de búsqueda principal, como se observa en la Figura 3.1, se opta por no realizar ninguna búsqueda con la cadena SS.

Por otro lado, el artículo [66, A04] de ACM Digital library, los artículos [67, A10] [68, A14] [69, A28] [70, A29] [71, A34] [72, A36] [73, A48] [74, A59] de la biblioteca IEEE*Xplore* y [75, A40] de la biblioteca ScienceDirect, fueron devueltos utilizando únicamente las cadenas de búsqueda parciales SS1 ([66, A04] [67, A10] [68, A14]), SS2 ([69, A28] [70, A29] [71, A34] [72, A36] [75, A40] [74, A59]) y SS3 ([73, A48]). Sin embargo, una vez analizado el contenido de los mismos, se concluye que los trabajos se corresponden, directa o indirectamente, con todos los temas que motivan esta investigación, *Context*, *Data Quality* y *Data Warehouse*. Por lo tanto, se toma la decisión de cambiar la clasificación de estos artículos, por lo que son analizados y evaluados buscando responder la *Research question* principal, RQ.

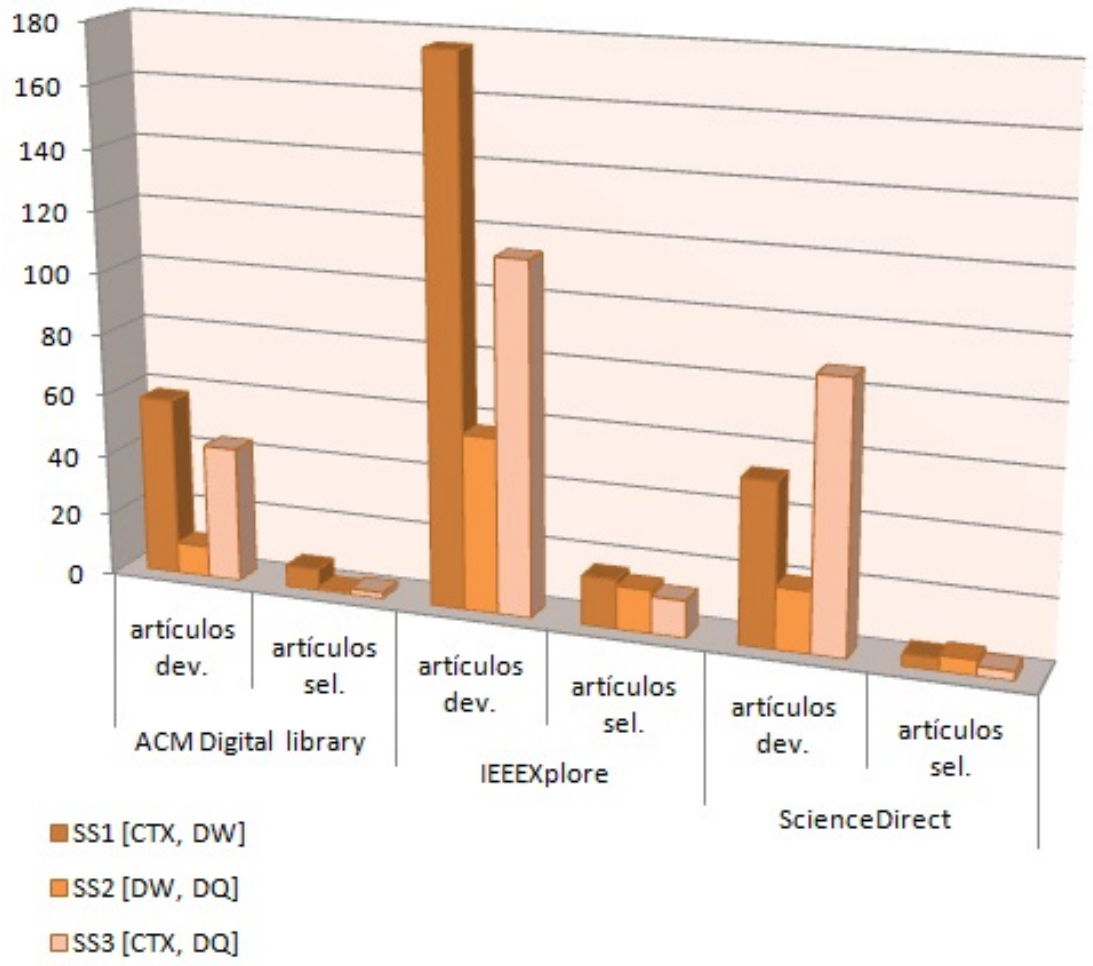

Figura 3.3: **Resultados por biblioteca digital**

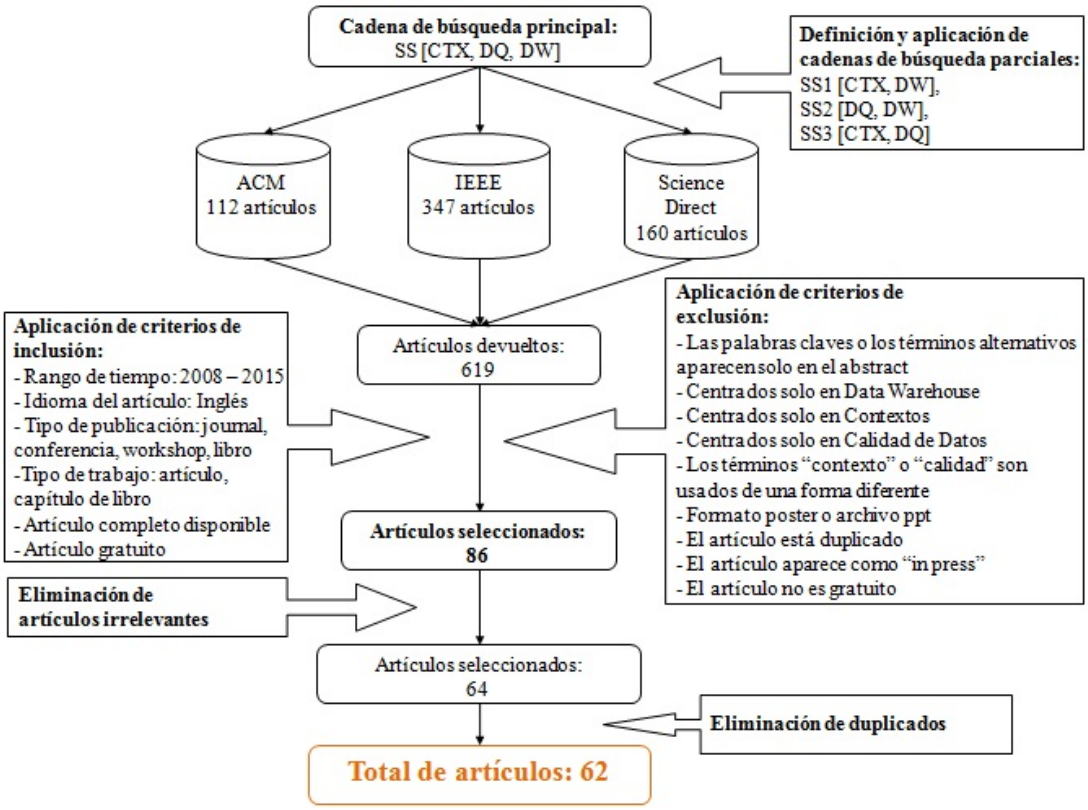

Figura 3.4: **Proceso de selección de artículos**

		ARTÍCULOS		
	Palabras claves	**ACM Digital library**	**IEEEXplore**	**ScienceDirect**
SS1	*Data Warehouse* *Context*	[79, A01] [80, A02] [81, A03] [88, A05] [89, A06] [90, A57]	[76, A07] [77, A08] [78, A09] [82, A11] [83, A12] [84, A13] [91, A15] [92, A16] [93, A17] [95, A18] [96, A19] [97, A21] [98, A22]	[85, A23] [86, A24] [87, A25] [94, A58]
SS2	*Data Warehouse* *Data Quality*	[34, A26]	[99, A27] [100, A30] [101, A31] [104, A32] [105, A33] [106, A35] [108, A37] [109, A38]	[102, A39] [103, A41] [57, A60] [107, A61]
SS3	*Data Quality* *Context*	[113, A42] [114, A62]	[110, A43] [111, A44] [112, A45] [115, A46] [116, A47] [117, A49] [121, A50] [122, A51] [123, A52] [124, A53]	[118, A54] [119, A55] [120, A56]
SS	*Data Warehouse* *Data Quality* *Context*	[66, A04]	[67, A10] [68, A14] [65, A20] [69, A28] [70, A29] [71, A34] [72, A36] [73, A48] [74, A59]	[75, A40]

Tabla 3.12: **Artículos seleccionados por palabras claves**

3.6. Análisis por *Research question*

En esta sección se presenta el análisis de cada uno de los trabajos seleccionados. Los artículos fueron agrupados de acuerdo con la categoría en la cual fueron clasificados, es decir, se agruparon en base a la cadena de búsqueda parcial con la cual fueron obtenidos (SS1, SS2 o SS3). Como se explicó anteriormente, dado que cada cadena de búsqueda está asociada a una *Research question* (RQ1, RQ2 o RQ3), cada artículo fue clasificado en base a la *Research question* que intenta responder. Finalmente, se presenta el análisis de los trabajos que se consideraron relevantes para responder la *Research question* principal, RQ. Si bien tanto la cadena de búsqueda principal (SS) como las cadenas de búsqueda parciales devolvieron un único trabajo, un conjunto de trabajos fueron seleccionados buscando encontrar la respuesta a esta pregunta. Estos trabajos relacionan entre sí, explícita o implícitamente, los tres temas de interés: *Data Warehouse*, *Data Quality* y *Context*.

3.6.1. RQ1: ¿Cómo son usados y definidos los contextos en los sistemas de DW?

Buscando responder RQ1 se seleccionaron trabajos que se apoyaran en el uso de contextos en el desarrollo de un SDW. Por tanto, los artículos a continuación fueron organizados según la etapa de desarrollo del DW (definición de requerimientos, diseño, modelización, implementación y/o análisis de los datos) en la cual es tenido en cuenta el contexto. Además, se presenta una descripción de los contextos que fueron definidos y/o utilizados (de datos, de usuarios, de las tareas realizadas, de documentos, etc.), en cada trabajo.

Por otro lado, también se analizaron artículos que abordan las características de los contextos. En estos trabajos los investigadores relacionan dichas características con las propiedades multidimensionales de los sistemas de DW. Si bien estos artículos no están centrados en responder la pregunta planteada, el vínculo que estos presentan entre los sistemas de DW y los contextos, resulta de interés para la investigación planteada.

Contexto en la etapa de diseño conceptual

Los sistemas de DW están definidos por las dimensiones y medidas que describen al negocio. En base a esto, los autores de los siguientes trabajos consideran el contexto como las dimensiones que permiten analizar las medidas. Por esta razón, desde las etapas iniciales del desarrollo de un DW, en particular en la etapa del diseño conceptual, es de gran importancia el análisis de los requerimientos teniendo en cuenta las necesidades de los usuarios, ya que estos son quienes aportan la

información requerida para obtener dichas medidas y dimensiones.

Un ejemplo de cómo dar más protagonismo a los usuarios en el momento del diseño del DW se presenta en los trabajos [88, A05] [76, A07]. En el primer artículo destacan que la etapa de análisis tiene como objetivo la obtención de los requerimientos de información de los usuarios que toman las decisiones. Según los autores, esta información está relacionada con las medidas de interés del proceso de negocio y con el contexto que permite analizar dichas medidas. Este contexto está determinado por las dimensiones del DW. Sin embargo, consideran que los usuarios que toman las decisiones a menudo ignoran cómo describir adecuadamente las necesidades de información. Por lo tanto, en la etapa de análisis de requerimientos, se debe comenzar obteniendo los objetivos de los usuarios para luego obtener los requerimientos de información. Subrayan que los diversos elementos multidimensionales, como lo son los hechos y las dimensiones, se obtienen a partir de estos requerimientos de información y denominan, a esta forma de especificar el modelo conceptual multidimensional de un DW, "Sensibilidad a los Requerimientos" (del inglés *Awareness Requirements*).

También en [76, A07] se enfocan en la definición de objetivos en las distintas actividades del proceso de ingeniería de requerimientos y agregan que los requerimientos se descomponen en contextos y medidas, dichos contextos son determinados por las dimensiones del DW. Además, consideran que estos contextos describen los datos adicionales necesarios para analizar un proceso de negocio determinado. Finalmente, subrayan que los requerimientos para un DW evolucionan a medida que la información de la organización va cambiando. Por esta razón, mencionan que el modelado de las necesidades de los usuarios es un aspecto muy importante para el buen diseño de un DW. Por otro lado, en el trabajo [90, A57] utilizan un DW para representar información acerca de un sistema para el análisis estructural y funcional de genomas. El esquema del DW se define en términos de objetos primarios (las *facts*), que son caracterizados en el contexto de otros objetos (las dimensiones del DW). En este trabajo, los genes son los objetos primarios que definen los hechos del DW. En este trabajo también consideran que las dimensiones del DW dan contexto a los hechos.

Contexto en la etapa de identificación y selección de las fuentes

En estos trabajos tienen en cuenta el contexto del usuario y lo definen a través de los datos internos y externos de las organizaciones. Los datos internos se obtienen una vez que las fuentes de los mismos han sido identificadas y que las transformaciones necesarias, que permiten definir el modelo lógico de datos compuesto por entidades y relaciones que determinan las necesidades del negocio, han sido definidas.

Algunas investigaciones como [79, A01] [80, A02] [82, A11] [85, A23] consideran que los datos internos podrán ser explotados de una forma más exitosa si los mismos son combinados con datos externos al DW. Los autores de [79, A01] [82, A11] consideran que el contexto está definido por el contenido en documentos. Según [79, A01], las fuentes de información más comunes para un usuario son la Intranet de su empresa, la Web y los correos electrónicos. Por esto, agregan que es muy frecuente que los datos no estructurados, presentes en estas fuentes (o documentos), puedan estar relacionados con las entidades y relaciones almacenadas en los DWs.

Contexto en la etapa de extracción, transformación y carga

En este caso también utilizan el contexto de los usuarios y agrupan aquellos usuarios que comparten intereses en el proceso de extracción, transformación y carga de los datos. En el trabajo [91, A15] se concentran en el proceso ETL y muestran cómo contextualizar dichos procesos en base a diferentes puntos de vista de los expertos. Los autores hacen hincapié en que los usuarios deben participar en el proceso ETL.

Por tanto, buscando agrupar usuarios que tienen en cuenta puntos de vista similares, presentan el Contexto Compartido (del inglés *Sharing Context*), que contiene elementos del conocimiento (que son más o menos relevantes para el foco de atención) que permiten construir el contexto del procedimiento (que es la parte del contexto que se invoca, organizado, estructurado y es utilizado en un paso concreto). Estos contextos compartidos son construidos, y enriquecidos gradualmente, a partir de elementos contextuales (por ejemplo parámetros para la conexión a una base de datos, estado de dicha conexión, etc.), que provienen de los contextos individuales (ya que cada experto tiene una representación mental del contexto compartido). En esta propuesta, cuando un elemento contextual es propuesto por un experto, entrará en el contexto compartido si es aceptado o validado por otros expertos.

Contexto en la etapa de análisis de los datos

En esta etapa de análisis de los datos, los trabajos seleccionados abordan diferentes definiciones de contexto. En su mayoría, los artículos consideran el contexto del usuario, teniendo en cuenta diferentes perspectivas del mismo. Por ejemplo, en algunos casos el contexto del usuario es determinado por el contenido en documentos externos que son, según los autores, los que permite comprender los datos almacenados en un DW. En otros casos, el contexto es determinado por las características propias del usuario (idioma, lugar de trabajo, etc.) y su ubicación geográfica. Sin embargo, otros autores lo definen a través del perfil del usuario, el comportamiento del mismo al realizar una consulta OLAP y los dispositivos

utilizados durante dichas consultas.

Algunos investigadores destacan que dado que los usuarios que toman las decisiones analizan datos multidimensionales, las agregaciones de los datos (SUM, AVG, MIN, MAX, etc.) son de gran importancia. Por tanto, definen el contexto de la agregación como el cubo de datos, la medida que el usuario desea agregar y el eje en el cual el usuario desea realizar la agregación. Otros investigadores tienen una percepción más amplia del contexto y lo definen como toda la información que cubre la tarea OLAP (información sobre los documentos analizados, contenido y semántica de los documentos, información del usuario, etc.). En otro caso, si bien la propuesta también está centrada en el usuario, utilizan el contexto de las anotaciones, éstas son realizadas por dichos usuarios en sus diferentes tareas. El análisis de los datos presenta muchos desafíos, ya que los usuarios intentan explotar al máximo los datos que se almacenan en el DW y, para dicha tarea, se apoyan en diferentes técnicas de análisis: OLAP, IR (del inglés *Information Retrieval*), *Data Mining*, etc. Todos estos trabajos son descriptos a continuación.

En el trabajo [79, A01] buscan enriquecer documentos con información que proviene de un DW. Con este propósito presentan un sistema que sugiere gráficos para ilustrar el contenido textual de documentos utilizando datos estructurados que provienen de un DW. Para obtener resultados relevantes, consideran el contexto del usuario, que en este caso está descrito por el contenido de dichos documentos. Los autores proveen un método en el cual medidas, dimensiones o instancias de las dimensiones son reconocidas en un texto para combinarlas en consultas que permiten explotar la semántica del DW. Las dimensiones y medidas reconocidas se combinan y a partir de esto sugieren diferentes consultas basadas en técnicas OLAP.

También en [82, A11] están centrados en textos que se encuentran fuera del DW. En este caso, los autores tienen en cuenta los foros Web donde los usuarios dejan su opinión acerca de los productos o servicios que ofrecen ciertas organizaciones. El objetivo de este artículo es explotar estos documentos de opinión a través de warehouses contextualizados. Para esto, presentan un framework para contextualizar un DW tradicional. Una vez más, en la bibliografía subrayan que el contexto que permite comprender los datos almacenados en un DW es usualmente descrito en documentos separados y su valor no es correctamente explotado. Estos documentos proveen información acerca de los hechos del DW, y según los autores, dichos documentos describen el contexto de los hechos. Por tanto, en [82, A11] cada hecho está relacionado con una lista ordenada de documentos (la posición del documento en la lista está relacionada con la relevancia del mismo respecto al hecho). En este caso, para la selección y clasificación de los documentos, los autores se apoyan en técnicas de recuperación de información (IR).

En la misma línea que los trabajos anteriores, en [85, A23] afirman que los sistemas de DW tradicionales permiten adquirir conocimientos útiles a partir de los datos de sus organizaciones, por medio de una variedad de tecnologías, como lo son OLAP o *Data Mining*. Sin embargo, si se desea proporcionar conocimientos más ricos en la dinámica de los negocios de hoy, también consideran importante combinar los datos internos de la organización con los datos del exterior, buscando así complementar la información. Por esto, los autores presentan el uso de tecnologías XML en entornos de DWs para asegurar el intercambio de información entre diferentes aplicaciones y personas. Además, resaltan que las dimensiones de un DW representan el contexto en el cual los hechos son analizados, por medio de los atributos de dichas dimensiones, que son organizados jerárquicamente.

Trabajos como [92, A16] también destacan el uso de las herramientas OLAP, pero a su vez resaltan la necesidad de adaptación de las mismas. Mencionan que los principales dispositivos de hoy día generalmente son incompatibles con la necesidad de consultar y navegar la información extraída de enormes cantidades de datos que se encuentran en la Web. Además, agregan que un desafío de las técnicas OLAP será poder utilizarlas en dispositivos móviles. Para superar esto, los autores proponen un sistema de recomendación contextual. De esta forma, aparte de la utilización de computadoras personales de escritorio, los usuarios pueden conectarse al servidor OLAP con sus notebooks, celulares, etc., para consultar, en muchas ocaciones, los mismos contenidos. La personalización y contextualización de los sistemas OLAP es una manera, según los autores, de responder a esta necesidad. Además, agregan que un sistema de recomendación sensible al contexto no sólo debe hacer uso de las preferencias del usuario, sino que también debe explotar la información de la situación contextual específica en la cual se consumirá el elemento recomendado. Como en [80, A02], consideran que en los sistemas de DW, este es un tema emergente y presenta muchos desafíos. Por esta razón, los investigadores de [92, A16] proponen un sistema de recomendación contextual con el fin de apoyar a los usuarios en sus actividades de navegación OLAP, mediante la explotación del contexto determinado por el usuario. Dicho contexto tiene en cuenta el comportamiento del usuario al consultar el servidor OLAP, los dispositivos utilizados durante dichas consultas y sus preferencias. En [94, A58] también se basan en el comportamiento del usuario, en sus preferencias y en sus características para detectar las consultas, que han realizado otros usuarios, y que pueden ser sugeridas al usuario actual.

Por otro lado, la definición de reglas para la representación de la información es ampliamente utilizada en trabajos de investigación. En particular, se destacan aquellos trabajos que definen reglas para la representación de información contextual, ejemplos de esto son los trabajos [86, A24] [87, A25]. En [87, A25] presentan la necesidad de desarrollar DWs espaciales, ya que los consideran útiles para mejorar el proceso de la toma de decisiones. La propuesta se basa en un lenguaje de reglas de personalización que utilizan para especificar las necesida-

des espaciales requeridas, adaptándolos a cada usuario (responsable de la toma de decisiones). Esto es así porque, según los autores, las necesidades espaciales de cada usuario podrían cambiar con el tiempo o dependiendo del contexto. Los investigadores consideran que la personalización espacial puede estar influenciada por los siguientes factores: las características específicas de los usuarios que son independientes del dominio (idioma, departamento en el que trabaja, etc.), el comportamiento espacial del usuario (con el fin de derivar las preferencias o intereses sobre diferentes elementos del sistema) y los cambios en el contexto espacial del usuario (ubicación geográfica). Los datos de interés son representados como medidas, mientras que los contextos de análisis para dichas medidas están determinados por las dimensiones. La información relativa a la ubicación geográfica, junto al contexto de análisis, caracteriza a los cambios del entorno que rodea a un usuario. Finalmente, la estructura de los datos necesarios para la personalización es especificada en un modelo de usuario sensible a los datos espaciales.

Con otro enfoque de contexto, los autores de [86, A24], proponen el uso de reglas para representar el conocimiento y resaltan que los usuarios toman las decisiones mediante el uso de herramientas OLAP en diferentes niveles de agregación. Por esta razón, subrayan que el conocimiento debe estar correctamente representado en los modelos multidimensionales y la importancia de las agregaciones de los datos (SUM, AVG, MIN, MAX, etc.). Agregan que una de las ventajas del uso de reglas es que son un medio apropiado para tomar en cuenta el contexto, ya que éste juega un papel crucial en la orientación al usuario durante el análisis OLAP. En este caso, interesa el contexto de una agregación, que consiste en el cubo de datos, la medida que el usuario desea agregar y la dimensión del DW en el cual el usuario desea realizar la agregación. Los investigadores de [86, A24] destacan que el concepto de agregación es central en el diseño de los DW y el modelado multidimensional. Sin embargo, dicho concepto generalmente es poco representado, dado que los modelos multidimensionales se enfocan en la representación estática del conocimiento, mientras que éste es contextual y tiene una estructura compleja y dinámica. Apoyándose en esas debilidades, el trabajo plantea que el conocimiento de las agregaciones no puede ser representado sólo por medio de restricciones en el esquema multidimensional y si bien los autores consideran que estas restricciones son útiles, no son suficientes para expresar toda la riqueza, dinámica y complejidad del conocimiento en las agregaciones.

En algunos trabajos se menciona que es notorio que las consultas de grandes volúmenes de datos mediante técnicas OLAP son eficaces para analizar datos numéricos. Sin embargo, en [80, A02] [82, A11] [92, A16] resaltan que la mayoría de los datos de una empresa son complejos, tales como texto, imágenes, vídeos, etc; y desafortunadamente, las herramientas estándares de los sistemas de toma de decisiones no son suficientes para analizar dichos datos. En [79, A01] [80, A02] [82, A11] observan que, en su mayoría, estos datos son textuales y se presen-

tan en informes, e-mails, etc; sin un conocimiento semántico claro de su contenido. Como en trabajos que se mencionaron anteriormente, se utilizan técnicas de recuperación de información (IR) usando datos de texto, para evaluar la pertinencia de los datos en diferentes consultas. Generalmente, como subrayan en [79, A01], esta relevancia se basa en la frecuencia de los términos en el documento que se está procesando. Por todo esto, en [80, A02] proponen el uso de técnicas de IR junto con procesamiento OLAP para analizar mejor los datos textuales y extraer así su semántica. Además, mencionan que la tarea OLAP determina el contexto de la toma de decisiones y el mismo está formado por toda la información que cubre a dicha tarea (información sobre los documentos analizados, contenido y semántica de los documentos, información del usuario, etc.). Los autores afirman que, a la hora de explotar un DW, la información contextual debe ser tenida en cuenta. Sin embargo, agregan que si bien ha habido una variedad de aplicaciones sensibles al contexto, se ha investigado muy poco la integración del contexto en los sistemas de DW.

A la hora de optimizar la toma de decisiones en los negocios es inevitable mencionar los sistemas de *Business Intelligence* (BI). En [89, A06] presenta una propuesta centrada en el usuario, buscando mejorar la visualización de los datos en base a un sistema basado en anotaciones que soportan diferentes tareas de análisis. Los requerimientos de esta propuesta fueron tomados a partir de entrevistas realizadas a usuarios expertos analistas en BI. En particular, se basan en el desafío que presenta el cambio en el contexto de los datos que se van a visualizar. Para los autores, el contexto de una anotación incluye el contenido (texto provisto por el usuario), la información capturada automáticamente (autor y la fecha de creación de la anotación), las propiedades definidas por el usuario (tiempo de vida de una anotación, reglas de validez, etc.), las entidades a las cuales la anotación se refiere (el gráfico o la tabla de datos en la que fue hecha la anotación) y los datos específicos que el usuario está anotando. Finalmente, tienen en cuenta otras anotaciones conectadas a la anotación de interés (por ejemplo, un hilo de discusión). En base a las necesidades contextuales, también en el trabajo [97, A21] mencionan que los sistemas BI aún son débiles para soportar múltiples paradigmas, problemas multi-dominio y el mantenimiento eficaz de la visualización bajo contextos dinámicos. Los autores consideran que el contexto incluye información sobre la situación de los problemas relevantes, el tiempo, el espacio, el contexto social y el contexto tecnológico (incluyendo hardware y software). Por otro lado, el contexto también incluye información relacionada con los perfiles de visualización de los usuarios, tales como su conocimiento previo del dominio del problema, estilos cognitivos, características personales y preferencias, edad, género e incluso su estado de ánimo al tomar decisiones. En base a estos conceptos, en [97, A21] plantean un framework conceptual para desarrollar visualizaciones y adaptarlas a las necesidades planteadas por los cambios de contexto. Inclusive en el trabajo [98, A22] se presenta una investigación sobre los desafíos que se presentan en los sistemas de BI y, si bien no se apoyan en ninguna definición de contexto,

afirman en base a la literatura que el uso de estos sistemas es dependiente del contexto.

La propuesta de [95, A18] está centrada en indicadores financieros y presentan un DW que proporciona una visión integrada de dichos indicadores de acuerdo con diferentes perspectivas de análisis. En este trabajo se enfocan en aplicaciones basadas en servicios y en el impacto que generan los cambios del negocio sobre dichos servicios. El enfoque busca mejorar la toma de decisiones empresariales, de acuerdo con el contexto determinado por los usuarios que hacen uso de los servicios, y utilizan un modelo de DW que permite analizar los servicios desde varias perspectivas. Por otro lado, en [96, A19] presentan una herramienta que guía a instituciones reguladoras en el análisis de normas. Este tipo de organizaciones generalmente recogen grandes cantidades de datos acerca del cumplimiento de las normas. Por lo anterior, los autores consideran que las herramientas de BI son claves para visualizar el cumplimiento de las normas y el rendimiento general de una institución. Por otro lado, en este trabajo consideran que las dimensiones del DW son el contexto clave de un dominio, contra el que los hechos (del inglés *facts*) pueden ser analizados. A su vez, mencionan que las herramientas de BI han proporcionado durante mucho tiempo la capacidad de razonar sobre las dimensiones del tiempo, de la ubicación, de estructuras organizativas, etc. Sin embargo, según los investigadores de [96, A19], la dimensión correspondiente a las normas no ha sido discutida hasta ahora. En base a esto y considerando que las dimensiones definen al contexto de los hechos, los autores creen necesario incorporar a las normas en la definición del contexto.

Modelado de contextos con modelos multidimensionales

En la bibliografía analizada se observa la necesidad clara que presentan los distintos sistemas de información, de definir y utilizar un contexto. Muchos de estos sistemas de información se apoyan en técnicas de *data warehousing* para modelar el contexto de interés, de esta forma es posible consultar naturalmente sus diferentes dimensiones. Esto es así porque se apoyan en las características multidimensionales de dichos contextos.

Un ejemplo de esto, es representado en los trabajos [78, A09] [83, A12]. En [78, A09] procesan expresiones dependientes del contexto. En particular, está enfocado en programas que se ejecutan en un contexto multidimensional en el que pueden manipular explícitamente las dimensiones de ese contexto. En [83, A12] plantean la recopilación de los datos del entorno, el cual representa el contexto de la toma de decisiones. Además, proponen un modelo multidimensional del contexto en el que cada dimensión del DW representa una dimensión del mismo. Por otro lado, los autores de [81, A03] [77, A08] proponen generar cubos de datos multidimensionales a partir de datos obtenidos de sensores y, dado que manejan grandes cantidades de datos, consideran útil utilizar un DW para el manejo de los cubos

de datos multidimensionales. En [84, A13] presentan propuestas que han utilizado SDW para explotar las propiedades de un esquema multidimensional estructurando la información de interés (presente en los archivos de registro o logs), en medidas y dimensiones. Estas últimas, según los autores, representan el contexto para analizar dichas medidas. A su vez, en [93, A17] dan contexto a tareas remotas en dispositivos móviles y dicho contexto esta determinado por los usuarios de los dispositivos. Mencionan que las limitaciones de recursos y las necesidades de los usuarios deben ser integradas en un modelo sensible al contexto, con el fin de presentar datos precisos. Subrayan que el uso de técnicas de DW es eficiente para afrontar dicho desafío. Inclusive, agregan que el contexto es un ingrediente clave para apoyar la interacción entre el usuario de un sistema y los dispositivos de informática.

Resumen

Como se observa en el análisis antes presentado, los autores consideran que si bien ha habido una variedad de aplicaciones sensibles al contexto, se ha investigado muy poco la integración del contexto en los sistemas de DW. Por lo tanto, no hay un acuerdo claro sobre cómo definir y manejar contextos en este tipo de sistemas. Los trabajos identifican, en general, el contexto en base a las necesidades del negocio. En muchos de ellos planten que las dimensiones del DW determinan el contexto de las medidas del mismo. Sin embargo, no se ha encontrado ninguna representación formal para esto. Por otro lado, algunos destacan la importancia de dar más protagonismo al usuario, teniendo en cuenta los datos del contexto del mismo, mientras que otros definen y usan el contexto de los datos, de las tareas, de los documentos, etc. Por otro lado, se observa que en mayor o menor medida, en todas las etapas del desarrollo de un sistema de DW se presenta la necesidad de tener en cuenta un contexto particular, esto es así en base a los requerimientos de cada etapa. Por ejemplo, en la etapa de extracción, transformación y carga generalmente interesa considerar el contexto de las fuentes, mientras que en la etapa de análisis de los datos se tiene en cuenta el contexto del usuario. En la etapa de explotación y de análisis de los datos de un DW es donde se observa más claramente la dependencia del sistema respecto a algún contexto.

Teniendo en cuenta el proceso de la toma de decisiones de un DW, algunos autores afirman que el concepto de agregación es poco representado, debido a que los modelos multidimensionales se enfocan en la representación estática del conocimiento, mientras que éste es contextual y tiene una estructura compleja y dinámica. Otros autores consideran que los sistemas BI aún son débiles para soportar el mantenimiento eficaz de la visualización bajo contextos dinámicos. Por otro lado, en otra línea de trabajo, se destacan las características multidimensionales de los contextos y cómo muchos sistemas de información, aparte de los sistemas de DW, se apoyan en técnicas de data warehousing para definir métodos

que permitan seleccionar las dimensiones pertinentes para el contexto de interés.

Por todo esto y aunque ha habido muy poca investigación acerca de la integración de contextos en los sistemas de DW, se puede afirmar que es importante definir un contexto (o varios contextos, si fuera necesario), en el ambiente de este tipo de sistemas. Dicho contexto puede estar enfocado en los datos, en los usuarios, en el negocio, etc; o en una combinación de estos.

3.6.2. RQ2: ¿Cómo se maneja la calidad de los datos en los sistemas de DW?

El problema de la calidad de los datos, en los sistemas de DW, es ampliamente investigado en la bibliografía. Sin embargo, aunque muchos trabajos estén centrados en resolver los problemas de calidad de datos en la etapa de ETL, la literatura indica que aún no está claro cuál es el mejor momento para abordar la evaluación de la calidad en este tipo de sistemas. Por todo esto, también intentando responder RQ2, resulta interesante organizar los trabajos seleccionados de acuerdo al momento en el cual se plantea la evaluación de la calidad de los datos. Por otro lado, también se presentan aquellos trabajos que presentan y/o clasifican los problemas de calidad más comunes que se encuentran en los sistemas de DW.

Calidad de datos en la etapa de identificación y selección de las fuentes

El trabajo [104, A32] está centrado en la medición de la credibilidad de los datos (del inglés *believability*) y esta dimensión es descompuesta en tres subdimensiones de calidad: *trustworthiness*, *reasonableness* y *temporality*. Los autores deducen, a partir del concepto de credibilidad, que la credibilidad de los datos depende de su procedencia, o sea de las fuentes. Por esto, desarrollan un modelo basado en la procedencia de los datos aplicable en dominios como DW y BI. También en el trabajo [109, A38] destacan la importancia de atender las fuentes de los datos y afirman que los errores e inconsistencias se deben detectar y eliminar tanto en las fuentes de datos individuales como en la integración de múltiples fuentes. A su vez, mencionan que inicialmente se debe realizar el análisis de los datos (buscar detectar qué tipos de errores e inconsistencias se buscarán), luego se debe definir el flujo del proceso de limpieza, se deben verificar los datos (la correctitud de las transformaciones debe ser evaluada), y finalmente los datos erróneos deben ser sustituidos por los datos corregidos en las fuentes originales de los datos.

Por otro lado, en [102, A39] presentan un caso de estudio en el que listan los desafíos planteados en el diseño e implementación de un DW, uno de estos desafíos es la calidad de los datos. En la descripción del DW, los autores presentan a los hechos como los datos numéricos usados para satisfacer todas las opciones

de cálculo que son de interés para el usuario final y lo que provee contexto a dichos hechos son las tablas dimensionales. Los autores subrayan la importancia de chequear la calidad y la consistencia de los datos que se mueven desde las fuentes de datos. Un aspecto interesante de [102, A39] es que separan la consistencia de los datos de la calidad de los mismos. Es decir, consideran importante tener en cuenta la consistencia de los datos, pero la misma no es presentada como un aspecto de la calidad en sí.

Calidad de datos en la etapa de diseño conceptual

En el trabajo [57, A60], realizado durante el año 2015, consideran que varios autores han propuesto varias métricas para medir los factores de calidad de los modelos de datos conceptuales para los DWs. Por esta razón, con el fin de conocer el estado actual del estado del arte y buscando explorar oportunidades para futuras investigaciones, consideran necesario realizar una revisión de la bibliografía acerca de las métricas de calidad para modelos de datos de DW. Para dicha revisión bibliográfica aplican como metodología de búsqueda una SLR y concluyen que existe una falta de validación de métricas y factores de calidad para los modelos de datos conceptuales de DW.

Calidad de datos en la etapa de extracción, transformación y carga

Muchos trabajos reconocen que el proceso ETL es un proceso crítico para lograr calidad de datos en un DW. Un ejemplo de esto es el trabajo [105, A33], en el cual los autores destacan que tanto en el proceso de ETL como en el procesamiento de transacciones en línea (del inglés *Online Transaction Processing*, OLTP), se detectan tablas con registros duplicados. Por esto, discuten las distintas estrategias de deduplicación (que se refiere a la eliminación de los datos redundantes), y algunos métodos utilizados para prevenirla. Los autores concluyen que la deduplicación es importante para la mejora de la calidad de los datos antes de que estos sean cargados en el DW. Por otro lado, en [108, A37] se enfocan en el análisis de la calidad de los datos (según su estructura, su integridad, su consistencia, atomicidad, etc.), antes y después que el proceso de ETL haya sido realizado.

En el trabajo [99, A27] son más terminantes y afirman que, si se desea obtener calidad de datos en un sistema de DW, la limpieza de los datos es vital en el proceso ETL. Proponen un framework de reglas de configuración, el cual tiene dos entradas: un análisis de los datos (*Data Profiling*) que permite seleccionar las fuentes de datos a las cuales se les aplicará la limpieza y opiniones y experiencias de usuarios sobre problemas de los datos que ya son conocidos. A su vez, dicho framework es dividido en dos partes: configuración de reglas de limpieza de datos y procesamiento de la limpieza.

Por otro lado, en [34, A26] mencionan que si bien muchos proyectos de calidad de datos son integrados en los proyectos de DW, los mismos no asignan suficiente tiempo a la parte de calidad de datos. Con el objetivo de abordar esta problemática, los autores presentan un generador de reglas de calidad. En base a esto, integran procesos de calidad de datos en las tareas de ETL. El procedimiento comienza con la creación de módulos de calidad para cada tabla en la staging area (tablas temporales), estos módulos se ejecutan automáticamente después de la carga de cada tabla y antes de la carga los datos en el DW. En la ejecución se aplican reglas de calidad sobre las tablas, las cuales son creadas con un programa generador de reglas. Los módulos de calidad marcan los registros que se consideran con errores y corrigen los valores no válidos cuando sea posible. Los registros correctos, los corregidos y los que tienen advertencias son cargados en el DW, mientras que los registros con errores se mantienen en la staging area para ser corregidos manualmente por usuarios expertos del dominio.

En la misma línea, los autores de [103, A41] presentan una propuesta para el modelado de las tareas de limpieza utilizando, como herramienta de especificación, álgebra relacional. Consideran que las operaciones de álgebra relacional pueden ser implementadas en cualquier plataforma y sistema, en el cual los datos puedan ser almacenados en un archivo estructurado (como por ejemplo XML). Además, agregan que la ventaja de este enfoque es la capacidad de utilizar infraestructura tecnológica común (por ejemplo computadoras de escritorio), para ejecutar el proceso de ETL, sin necesidad de utilizar una base de datos. En particular, mencionan que las tareas de limpieza del proceso ETL deben tratar de identificar los problemas e inconsistencias de datos y realizar las transformaciones necesarias con el fin de garantizar una calidad mínima de datos y prepararlos para la carga en el DW.

Calidad de datos en la etapa de toma de decisiones

Los autores de [107, A61] consideran que la calidad de los datos y la calidad de la información tiene efectos directos y/o indirectos sobre la gestión de los sistemas de BI. En el trabajo afirman que la gestión de los datos, para garantizar la confianza en los mismos, es un requerimiento importante para alcanzar altos niveles de calidad en la etapa de la toma de decisiones. En particular, mencionan que los resultados de la investigación revelan una trayectoria significativa desde la calidad de los datos y la calidad de la información, pasando a través de la calidad de la gestión en sistemas de BI, hasta la toma de decisiones.

Clasificación de los problemas de calidad

En [100, A30] [101, A31] resaltan que no es una novedad el impacto que la mala calidad de los datos tiene en la toma de las decisiones, en la confianza de las organizaciones y en la satisfacción de los clientes. A pesar de esto, consideran

que los problemas de calidad en los repositorios multidimensionales aún no han sido correctamente ordenados. Por un lado, en [100, A30] comparan diferentes modelos de calidad de datos y presentan ventajas y desventajas de los mismos. Si bien no todos los modelos discutidos en el trabajo están centrados en los sistemas de DW, los autores se apoyan en las necesidades de este tipo de sistemas y consideran que cada uno de estos modelos tiene un aporte para la evaluación de la calidad de un DW. Finalmente, sin profundizar, proponen un modelo que combina algunas de las características de los modelos estudiados. Por otro lado, en [101, A31] proponen una clasificación de los problemas de calidad existentes en los sistemas de DW y presentan la misma como una taxonomía. Con este fin, el trabajo comienza con una revisión general de las taxonomías presentes en la bibliografía y mencionan que ninguna de ellas ha demostrado ser totalmente adecuada para representar los problemas de calidad presentes en los sistemas de DW. Luego listan, a partir del análisis de los requerimientos de calidad de este tipo de sistemas, cinco dimensiones de calidad que han sido identificadas: *completeness*, *timeliness*, *uniqueness*, *consistency* y *accuracy*. Los autores subrayan que la taxonomía que proponen tiene como punto de partida las dimensiones de calidad y la relación de las mismas en el nivel de granularidad en el cual el problema se manifiesta de acuerdo al modelo multidimensional (valor, tupla, columna, tabla, etc.). En [106, A35] también clasifican y discuten las causas de los problemas de calidad de datos hallados en un DW. En este caso para un dominio particular como lo es la medicina. Además, agregan que como estrategia de gestión es importante que quien proporcione los datos sea quien se haga responsable de ellos. Esto, con el fin de obtener datos precisos desde las fuentes de origen.

Resumen

La importancia de la calidad de los datos en un sistema de DW ya ha sido ampliamente demostrada en la bibliografía, ya que muchos autores han presentado la necesidad de incorporar y mantener la calidad en dichos sistemas. De todas formas, en las investigaciones no se encuentra un consenso acerca de cómo hacerlo. La mayoría de los trabajos sólo han abordado la limpieza de los datos en la etapa de ETL, ignorando la tarea de evaluación de la calidad de los datos a lo largo de todo el ciclo de vida de un DW. Además, aunque muchos trabajos han asociado dimensiones de calidad a los sistemas de DW, según los investigadores, aún no se ha identificado cuál es el conjunto de dimensiones de calidad pertinentes para estos sistemas de información. Por otro lado, es importante cuestionar si esto ocurre porque es imposible definir un único conjunto de dimensiones de calidad en el entorno de un DW, dado que dicho conjunto de dimensiones puede depender del propósito con el cual se utilizan los datos.

3.6.3. RQ3: ¿Cómo se consideran los contextos para la evaluación de calidad de datos?

Para responder esta pregunta se seleccionaron trabajos que abordaran, para cualquier sistema de información los conceptos de calidad de datos y contextos de forma conjunta. Si bien varios trabajos coinciden en el uso del contexto en la limpieza y/o en la evaluación de la calidad de los datos, no todos comparten la misma noción de contexto y no todos realizan dicha limpieza y/o evaluación con el mismo propósito. Por ejemplo, algunos de estos trabajos tienen en cuenta el proceso de producción de los datos, los datos en sí y su procesamiento, las necesidades de los consumidores de la información y las circunstancias que rodean al uso de la misma. La noción de contexto mayormente utilizada es aquella que tiene en cuenta al usuario en una tarea específica o, como una variante de este caso, a un conjunto de usuarios que tienen un propósito específico. Por otra parte, muchos investigadores hacen hincapié sólo en el usuario o sólo en la tarea que se está realizando. Estas últimas formas de describir el contexto coinciden con el enfoque de "adecuación para el uso" (del inglés *fitness for use*), muchas veces referenciado en la bibliografía, que permite afirmar que la calidad de los datos es por naturaleza contextual.

Una vez identificado qué es el contexto, los diferentes trabajos plantean distintos objetivos en relación a la calidad de los datos. Los propósitos son variados, por ejemplo, en un caso plantean mejorar la producción de productos de información. Otros casos se centran en identificar las dimensiones de calidad que permitan evaluar la calidad respecto a una tarea. Algunos autores combinan un conjunto de dimensiones de calidad para la toma de decisiones o buscan dar respuestas que se aproximen a las necesidades de información. Por otro lado, también se analizaron artículos que están centrados en la evaluación de la calidad de los contextos. Aunque este no es el foco de la pregunta de interés, dichas investigaciones se tuvieron en cuenta con el fin de mostrar, desde otro punto de vista, la relación existente entre la calidad y los datos que determinan un contexto.

Calidad de datos y contextos

El concepto de "adecuación para el uso" es ampliamente usado en la literatura relacionada con la Calidad de Datos. Los trabajos [99, A27] [69, A28] [70, A29] [112, A45] [115, A46] [122, A51] [118, A54] se basan y/o destacan este enfoque y, según los autores, este concepto implica que la calidad depende del contexto en el cual son utilizados los datos. En [122, A51] agregan que esto es así porque se toman en cuenta las necesidades de los consumidores de la información y las circunstancias que rodean al uso de la misma. Por otro lado, en [115, A46] examinan las relaciones y dependencias existentes entre el contexto, determinado por el ambiente, y las dimensiones de calidad utilizadas en los frameworks de calidad actuales. En el trabajo señalan que un análisis de la literatura reveló que la ma-

yoría de los enfoques son dependientes del contexto, pero aún así la dimensión contextual no suele ser representada en los frameworks de calidad. Apoyándose en esta contradicción, proponen un framework que, según los autores, puede ser aplicado en diversos contextos. Si bien la propuesta es aplicada en un entorno experimental, concluyen que el contexto es importante para la evaluación de la calidad de la información. Finalmente, agregan que la aplicación del framework de calidad les permitió tener en cuenta a los usuarios y a la tarea que se está realizando. En [112, A45] también abordan las dimensiones de calidad y presentan un caso de estudio en el cual evalúan la aplicabilidad de un modelo particular. Dicho modelo, relaciona dimensiones de calidad con procesos de producción de productos de información. Para los autores, el contexto de interés está determinado por el proceso de producción y su propósito y afirman que la calidad de los datos, en particular las dimensiones de calidad, deben ser ajustadas al contexto. Por esta razón, las dimensiones de calidad aplicadas en el modelo considerado les permitieron identificar los requerimientos de calidad de datos de los consumidores.

En base a que todos los usuarios tienen diferentes requerimientos de datos, el nivel de calidad satisfactorio varía con la perspectiva de cada usuario y los parámetros que especifican el contexto de interés. De acuerdo con esto, en [124, A53] presentan un modelo para incorporar una combinación de las dimensiones de calidad *reliability*, *credibility* y *timeliness*, basado en el contexto y en la toma de decisiones secuenciales para el reconocimiento de patrones. En este caso son los usuarios que toman las decisiones, en particular sus requerimientos, lo que definen al contexto. Por otro lado, en [113, A42] los investigadores destacan que la gran cantidad de información que se encuentra en la Web es a menudo subutilizada. Esto es debido a las dificultades que se presentan en el acceso a las fuentes de datos heterogéneas y dinámicas. Buscando abordar esta problemática, muchas veces se definen consultas complejas que intentan satisfacer las necesidades de los usuarios, pero el procesamiento de dichas solicitudes generalmente tiene un alto costo en este tipo de fuentes y no garantizan la satisfacción del usuario. Por esta razón, presentan el concepto de consulta usable (del inglés *wearable query*) que captura las características específicas del usuario y de la solicitud. El trabajo presenta un enfoque para proporcionar respuestas que se aproximan a las necesidades de información. Los autores consideran que el elemento clave de la solución propuesta es un nuevo concepto de adaptabilidad basada en tres puntos claves: los datos, la calidad de procesamiento y el contexto del usuario (ubicación, intereses, necesidades, etc.).

En el trabajo [116, A47] hacen referencia al concepto de etiquetas de calidad de datos (del inglés *Data Quality tags*). Los autores mencionan que la semántica de una etiqueta, es decir su significado, se refiere a las características específicas de la calidad de los datos, por ejemplo la consistencia, cuyo valor está representado por una etiqueta de calidad de datos. A su vez, consideran un framework en el cual diferentes etiquetas de calidad pueden ser definidas en tres categorías

del mismo. Dichas categorías describen la conformidad de los datos, respecto a reglas de integridad, la correspondencia con el mundo real y la utilidad para un usuario en una tarea específica. Las dos primeras categorías son inherentes del conjunto de datos, mientras que la última depende del uso específico de los datos y del usuario, por tanto es considerada una categoría contextual y los contextos están determinados por los usuarios. Agregan que la información de la calidad de los datos, basada en medidas subjetivas de calidad, debe estar asociada a información contextual adicional para poder ser interpretada o utilizada. También en [110, A43] exploran la naturaleza contextual de la calidad de la información y los autores utilizan el término "contexto" para referirse a un conjunto de usuarios con un propósito específico, en una tarea específica. En el trabajo consideran que sería más apropiado evaluar la calidad de la información como un proceso más que como una medida estática. A su vez, consideran que esto permitiría captar mejor su esencia como "información", yendo desde los datos hasta el conocimiento, pasando a través de diversos contextos.

Por otro lado, en [122, A51] diferencian la calidad de los datos de la calidad de la información y mencionan que las métricas de la calidad de los datos pueden ser absolutas o relativas. A las métricas relativas también las denominan "*fitness for use*" y para el caso de la calidad de la información consideran que las medidas son principalmente "*fitness for use*", ya que la información sólo tiene sentido en un contexto. En el artículo indican que mientras una determinada pieza de información tiene un valor de calidad único asociado a cada uno de sus atributos absolutos, las valoraciones "*fitness*" de la misma "pieza" de información pueden diferir ampliamente. Los autores afirman que dichas valoraciones pueden variar en función de los usuarios que consumen los datos, de la tarea en cuestión y de las circunstancias dadas. Además, agregan que todas estas consideraciones de la calidad de los datos determinan al contexto de la toma de decisiones.

En el trabajo [118, A54], como muchos trabajos presentes en la literatura, se apoyan en la clasificación de Wang y Strong, 1996 [4]; para las dimensiones de calidad. Dicha clasificación agrupa las dimensiones de calidad en: intrínsecas (el grado en el que los valores de los datos se ajustan a los valores reales), contextuales (la medida en que los datos son aplicables a la tarea del usuario de los datos), representacionales (el grado en el que los datos se presentan de una manera inteligible y clara) y de accesibilidad (el grado en que los datos están disponibles). Por ejemplo, mencionan que las dimensiones *accuracy* y *objectivity* son objetivas, ya que son intrínsecas a los datos en sí e independientes del contexto en el cual son utilizados los datos. Sin embargo, consideran que no todas las dimensiones de calidad pueden ser medidas objetivamente, ya que dimensiones como *relevance* y *believability* tienden a variar de acuerdo al contexto en el cual son usadas. Además, agregan que a pesar de las amplias discusiones en la bibliografía acerca de la calidad de los datos, no existe un único conjunto de dimensiones de calidad de datos. Esto último se debe, según los autores, a que la calidad de los datos

es dependiente del contexto. Los investigadores estudian la calidad de los datos según el propósito de los usuarios en una tarea específica. En particular, el objetivo presentado en [118, A54] es identificar las dimensiones de calidad apropiadas para evaluar la calidad de datos de una institución financiera. Los resultados obtenidos indicaron que hay una diferencia entre los sectores financieros y otros sectores respecto a las dimensiones de calidad necesarias. Esto confirma, según los autores, el comportamiento contextual de la calidad de los datos.

Al recorrer la bibliografía es fácil hallar diferentes clasificaciones de las dimensiones de calidad, y sin duda las categorías más destacadas son la intrínseca y la contextual. Los autores de [120, A56], al igual que en los trabajos [112, A45] [122, A51], se apoyan en la definición de dichas categorías y afirman que la exactitud (del inglés *accuracy*) es una dimensión intrínseca de la calidad de los datos. En este artículo consideran que para esta dimensión ningún contexto es necesario para los datos, ya que subrayan que los datos son exactos o no. En contraposición con esto, los autores en [119, A55] destacan que en la mayoría de los trabajos todavía confían en el conocimiento del experto para evaluar qué tan exacta es una fuente de datos. Por esto, proponen un enfoque de cuantificación para la dimensión *accuracy* en entornos de múltiples fuentes. Para esto, los autores se concentran en el contexto de las fuentes de datos, el cual está definido por la estructura de las mismas, denominada por los investigadores sintáxis de las fuentes. Es decir, las fuentes de datos son comparadas a nivel de columna y/o registro, buscando similitudes estructurales entre ellas. Finalmente resaltan que la dimensión *accuracy* está fuertemente relacionada con el contexto y que diferentes contextos darán lugar a valores diferentes de la exactitud para la misma fuente de datos.

Por otro lado, en [114, A62] consideran requerimientos de calidad que son considerados restricciones sobre los datos. Dichas restricciones se aplican a un conjunto de datos y se identifica el subconjunto de esos datos, para los cuales se sospecha podrían violar los requerimientos de calidad. Los autores mencionan que los datos pueden ser erróneos o no dependiendo de las restricciones que se pongan sobre ellos. Por lo tanto, los requerimientos de calidad dan contexto a los datos para poder determinar la validéz de los mismos. De esta forma, datos que se consideran no tienen buena calidad, según ciertas restricciones, podrían tenerla de acuerdo a otros requerimientos de calidad.

Calidad de los datos del contexto

Según [117, A49] [121, A50], la tarea principal para las aplicaciones sensibles al contexto es la recopilación de la información contextual, y en la mayoría de las ocasiones, la misma se realiza a partir de diferentes sensores físicos. Para los autores de [123, A52] las aplicaciones *pervasive*, como se denomina a las aplicaciones sensibles al contexto, se basan en la adquisición y el consumo de los datos en diferentes ambientes, en tiempo real. En [111, A44] [117, A49] presentan pro-

puestas en las cuales se evalúa la calidad del contexto en sí mismo.

Los autores de [111, A44] destacan que la diversidad de las fuentes de información de contexto y las características de los entornos *pervasive*, entre otros, suponen todo un reto para la gestión eficiente de los datos del contexto. Esto se debe a la detección de una gran cantidad de datos redundantes y contradictorios y, pensando en este desafío, en el trabajo proponen un sistema que detecta y elimina duplicados y conflictos en la información de contexto, determinado por el ambiente de los datos. A su vez, hacen referencia a la definición de calidad de contexto, que mencionan es cualquier información que describe la calidad de los datos usados como información de contexto. Por otro lado, los autores de [117, A49] se centran en la gestión eficiente de los datos de contexto. Para esto último definen varios factores que le permiten evaluar la calidad de los mismos.

Resumen

En muchos trabajos señalan la naturaleza contextual de la calidad de los datos lo que, según algunos autores, justifica que no se encuentre un único conjunto de dimensiones de calidad. A pesar de esto, varios trabajos consideran que la dimensión contextual no suele ser representada en los frameworks de calidad y a su vez, coinciden en el uso del contexto en la tarea de limpieza y/o evaluación de la calidad de los datos. Sin embargo, no todos comparten la misma noción de contexto y no todos consideran los contextos de la misma forma.

Por otro lado, cabe destacar la diferencia existente entre el punto de vista de algunos investigadores. Esto se observa claramente en la discusión que se presenta acerca del aspecto contextual de la dimensión de calidad *accuracy*. Una vez más, esto resalta la dependencia de la calidad de los datos respecto al entorno en el cual se va a realizar la limpieza y/o evaluación de la calidad. Además, interesa resaltar el enfoque de la calidad de datos como un proceso más que como una medida estática. Los investigadores consideran que esto permitiría captar mejor la esencia de los datos, en particular de la información, yendo desde los datos hasta el conocimiento, y pasando a través de distintos contextos.

Finalmente, se puede afirmar en base a la evidencia hallada, que en algunas ocasiones los contextos son considerados para la limpieza de los datos y/o evaluación de la calidad de los mismos. Sin embargo, más allá de la clara naturaleza contextual de la calidad, no se ha establecido aún bajo qué circunstancias y de qué forma deben ser considerados y/o representados dichos contextos.

3.6.4. RQ: ¿Cómo pueden ser usados los contextos para evaluar la calidad de datos en *Data Warehouse*?

La búsqueda de trabajos que integrara simultáneamente los temas de investigación *Data Warehouse*, *Data Quality* y *Context* no arrojó buenos resultados, dado que sólo un trabajo fue devuelto en este caso. En la etapa de análisis y evaluación de los artículos que fueron devueltos en las búsquedas parciales (donde cada una de las búsquedas combinó pares de temas: *Data Warehouse* y *Data Quality*, *Data Warehouse* y *Context*, *Data Quality* y *Context*), se seleccionó un conjunto de trabajos. Algunos de dichos trabajos, aunque no hacen referencia explícita a los tres temas de interés, están relacionados con ellos en mayor o menor medida. Por esta razón, se consideraron de importancia para intentar responder RQ. Por ejemplo, en algunos de los casos los trabajos se centran en la evaluación de la calidad de datos de los sistemas de DW, sin embargo, para la investigación tienen en cuenta las necesidades de los usuarios o la tarea realizada, lo cual determina un contexto.

Por lo tanto, aunque en la mayoría de las ocasiones los autores no incorporan una definición o descripción explícita del contexto, resaltan la necesidad de adaptación de la calidad de los datos. Algunas de las formas propuestas para abordar esto es mediante la definición de modelos de usuario, teniendo en cuenta los datos de la tarea que se lleva a cabo, mediante la inclusión de los intereses de los usuarios, en base a las decisiones que van a ser tomadas en el proceso del negocio, etc; y todas estas formas permiten determinar un contexto. Los diferentes autores presentan distintos propósitos, algunos consideran que los datos deben ser ajustados de acuerdo al uso que le darán los consumidores de los mismos. Otros autores, en un ambiente de toma de decisiones, consideran la calidad de los datos para el análisis estadístico en un entorno sensible al contexto.

Otras formas de relacionar el contexto y la calidad de los datos son aquellas que presentan los trabajos que, en base a los modelos multidimensionales, abordan los desafíos que presenta el modelado del contexto teniendo en cuenta la calidad de los datos.

Contexto y calidad de datos en la etapa de análisis de requerimientos

En [70, A29] presentan una propuesta en la cual integran la calidad de los datos en toda la fase de desarrollo del DW, en particular en la etapa de análisis de requerimientos. Por otro lado, agregan que la calidad de los datos ya no es una característica opcional, sino un requerimiento para el efectivo desempeño del negocio. Mencionan que una dimensión de calidad puede contener información objetiva sobre las características de los datos y su proceso, como por ejemplo la última fecha de modificación para la medida "actualidad" (del inglés *timeliness*). Pero además, especifican que las dimensiones de calidad también pueden contener

un punto de vista subjetivo. Esto es así porque diferentes usuarios podrían definir los criterios de calidad en función de sus propios requerimientos de calidad, es decir, de acuerdo con sus necesidades. En este caso, también hacen referencia al concepto de "adecuación para el uso" (del inglés *fitness for use*), para describir la calidad de los datos. Según los autores, este concepto implica que la calidad es relativa, ya que datos que son adecuados para un uso pueden no serlo para otro.

Contexto y calidad de datos en la etapa de diseño conceptual

Otra forma de considerar el contexto de los datos es teniendo en cuenta que la semántica de los datos puede cambiar. En base a esto, los autores de [66, A04] consideran el concepto conocido como dimensiones que cambian lentamente (del inglés *Slowly Changing Dimensions*, SCDs), que estudia la evolución de los datos de las dimensiones. Además, destacan que si bien las medidas y funciones de las medidas son consideradas tradicionalmente como estables dentro del esquema de un DW, el cambio en las mismas también puede estar sujeto a cambios en el diseño del DW. Los autores afirman que los cambios en la semántica conducen a valores de medidas incomparables, lo que significa resultados de análisis poco sólidos y sin valor. Por esta razón, abordan la evolución de las medidas, por lo que proponen el concepto de medidas que cambian lentamente (del inglés *Slowly Changing Measures*, SCMs), como un concepto adicional en el diseño del DW. Si bien en este caso no mencionan el concepto de calidad de datos, subrayan la importancia de obtener resultados sólidos, lo que implica resultados de análisis que contengan buena calidad. Por otro lado, en el trabajo mencionan que las dimensiones del DW representan entidades u objetos que en su conjunto determinan el contexto semántico de las medidas. Por tanto, los autores se basan en el contexto de los datos, es decir, en el contexto de las medidas del DW.

Contexto y calidad de datos en la etapa de extracción, transformación y carga

En [69, A28] presentan una solución para manejar el proceso de limpieza de los datos en la etapa de ETL. Se refieren a la calidad de datos, en un DW, como una propuesta "libre de errores" y la definen como el grado en que los datos cumplen con las necesidades específicas de los clientes. Dado que los intereses de los clientes varían, los autores hacen referencia a la necesidad de adaptación de la calidad de los datos por lo que consideran, implícitamente, el contexto de los usuarios. También en base a las tareas de limpieza de datos, en [72, A36] presentan un resumen de los problemas de calidad de datos que se observan en la construcción y en la integración de los datos de un DW. Además, analizan el desarrollo de las tareas de limpieza y algunas herramientas de limpieza de datos y posteriormente proponen un framework de limpieza de datos basado en un modelo de usuario. Además, presentan la necesidad de construir un modelo de calidad universal que permita definir un modelo de calidad que se ajuste a las necesidades de cada

usuario. Para esto, definen un modelo de usuario cuya función es adaptar los datos en relación con los requerimientos. Los autores tienen en cuenta la diversidad existente entre los distintos sistemas de información, por lo que subrayan que la limpieza de los datos es específica de cada dominio. En [72, A36] no mencionan al contexto explícitamente, sin embargo, la consideración de las distintas necesidades de los usuarios y el interés de especificar cada dominio, constituyen claramente la consideración de algún contexto.

Los autores de [71, A34] introducen su trabajo señalando que han descubierto que la calidad de los datos es incierta, ya que la calidad percibida de los datos está influenciada por la tarea que se lleva a cabo. Además, consideran que esos mismos datos pueden ser vistos de diferente forma respecto a su calidad, dependiendo del usuario en una tarea específica. Por esta razón, construyeron una base de conocimientos para atender diferentes requerimientos de calidad de los datos en base a las necesidades del usuario. Con este enfoque, una vez que el proceso de ETL finaliza, proponen aplicar técnicas de data profiling (para el análisis de los datos) y de limpieza de datos mediante el uso de diferentes reglas. Finalmente, destacan que los datos deben ser ajustados según el uso que le darán los consumidores de los mismos y no de acuerdo con los investigadores de la calidad de los datos. A su vez, resaltan que no se puede poseer datos perfectos y en muchas ocasiones tampoco se necesita que lo sean.

Contexto y calidad de datos en la etapa de análisis de los datos

El trabajo [67, A10] considera el análisis estadístico de los datos en un DW, en el nivel más bajo de agregación, por lo que tienen en cuenta únicamente estadísticos tales como *sum*, *count*, *mean* y *average*. En esta propuesta buscan comprender mejor el importante rol que juega la confianza en los sistemas de DW, con este fin, aplican los principios de gestión de la confianza y definen un modelo de confianza para un DW estadístico. Presentan la relación de confianza como la relación asimétrica entre dos partes y mencionan que en un DW dichas partes pueden ser: las fuentes de datos (personas o sistemas que proporcionan los datos), el administrador (que es responsable de recolectar y gestionar los datos del DW) o el usuario de los datos (que proporciona datos de alta calidad sobre los que luego se realizan diversas consultas estadísticas). Dado que la relación de confianza, en un DW, determina la calidad de los datos, dicha relación puede ser vista como una dimensión de calidad. A su vez, los autores afirman que el contexto es muy importante para cualquier definición de confianza y se apoyan en una definición del mismo que lo considera "una situación dada".

Los autores de [73, A48], como otros, resaltan que la calidad de los datos es por naturaleza sensible al contexto y por tanto, debe ser evaluada en el contexto del negocio objetivo, en el cual los datos serán utilizados. También afirman que la investigación de la evaluación de la calidad de los datos hasta ahora solo ha

sido centrada en la identificación de un conjunto de factores de calidad de datos o en el análisis de ciertas dimensiones (como por ejemplo, *timeliness* o *correctness*). Por esta razón, consideran que dichos enfoques tienen huecos importantes. En primer lugar, mencionan que los factores de calidad de datos relevantes, para sistemas de soporte de toma de decisiones, no han sido definidos y el contexto no ha sido considerado en la medición de la calidad de los mismos. Luego, como en [67, A10], subrayan que el nivel de confianza de los usuarios, en la calidad de los datos y en las decisiones, tampoco ha sido tenido en cuenta. En base a todo esto, presentan un enfoque en el cual introducen una medida de calidad en sistemas para la toma de decisiones y se apoyan en la hipótesis de que existe una relación directa entre los resultados del negocio y los factores de calidad de los datos. Además, agregan que el grado de impacto varía en función de la decisión tomada, por ejemplo, el impacto del factor de calidad "actualidad" (del inglés *timeliness*) sobre las decisiones tácticas es diferente al impacto que tiene sobre las decisiones de crédito. Por esta razón, los autores introducen el concepto de "categoría de decisión", que es lo que permite obtener un análisis de calidad de los datos sensible al contexto. A modo de resumen, en [73, A48] interesa el contexto de la calidad de los datos que, según los autores, se determina en la toma de decisiones ya que, las diferentes decisiones y sus categorías, permiten definir los factores de calidad pertinentes para cada negocio.

Por otro lado, en [74, A59] combinan las tres áreas de interés, definiendo un modelo de calidad que considera el análisis multidimensional y dando contexto a los datos de un DW, donde dicho contexto está determinado por el dominio de análisis. El trabajo se enfocan en un *Web Warehouse* (WW), que es un DW que consolida datos de la Web. Los autores mencionan que el objetivo de este tipo de sistemas es actuar como intermediario entre la publicación de los datos y los usuarios, pre-procesando los datos para darles un valor agregado. El pre-procesamiento implica la integración, la agregación y la reestructuración de los datos, también considera la medición y mejora de la calidad de los datos. Además, un modelo de procesos de negocio (del inglés *Business Process Model*, BPM) ayuda a especificar los usuarios, las actividades, las relaciones de precedencia entre las actividades y las restricciones, que se llevan a cabo con el fin de obtener el resultado deseado. En particular, para los datos del DW, construyen un modelo de calidad teniendo en cuenta las funciones de los datos en el análisis multidimensional (medidas, dimensiones, jerarquías, etc.), el contexto del análisis, las reglas de dominio que se aplican a los datos del DW y al conocimiento del dominio de usuario final.

Contexto y calidad de datos a través de modelos multidimensionales

Por otro lado, los trabajos [68, A14] [65, A20] relacionan fuertemente a los datos con la calidad de los mismos, apoyándose en el concepto multidimensional tomado de los sistemas de DW y las aplicaciones OLAP para la visualización

de la información contextual. Por un lado, en [68, A14] proponen integrar información contextual (datos del perfil del cliente y preferencias, entre otros) en un modelo de recomendación multidimensional. El método de recomendación convencional considera sólo dos dimensiones: cliente y artículo (o producto) y, según los autores, los resultados de la información contextual facilitan el aumento de la exactitud del resultado de la recomendación. Aunque los investigadores no mencionan la calidad de los datos, consideran la importancia de la exactitud de los resultados, haciendo referencia implícita a la dimensión de calidad *accuracy*, que junto a las dimensiones *timeliness* y *completeness*, son las dimensiones de calidad más investigadas en la bibliografía consultada.

Por otro lado, en [65, A20], único artículo que se obtuvo mediante la cadena de búsqueda principal y mediante las cadenas de búsqueda parciales, representan contextos con dimensiones (incluyendo sus respectivas jerarquías y niveles) y con esto, según los autores, hacen posible una evaluación multidimensional de la calidad de los datos. El trabajo es introducido con la afirmación de que la calidad de los datos no puede ser evaluada sin un conocimiento contextual de la producción y/o el uso de los datos. Los investigadores se apoyan y extienden el modelo de datos multidimensional Hurtado-Mendelzon [125], cuya creación fue motivada principalmente por los sistemas de DW y las aplicaciones OLAP. Para la extensión del modelo proponen una representación ontológica. A su vez, plantean mecanismos para la evaluación de la calidad de los datos en base a las consultas realizadas a la ontología, a través de la navegación dimensional. La propuesta incluye relaciones (asociadas a categorías en diferentes niveles de las jerarquías dimensionales), restricciones y reglas dimensionales.

Para finalizar interesa destacar la investigación realizada en el trabajo [75, A40], cuya propuesta tiene como objetivo proporcionar un mejor entendimiento respecto al éxito de los sistemas de BI. Para esto examinan el impacto de sus capacidades, como lo es la calidad de los datos, entre otros, en presencia de diferentes ambientes de decisión. Agregan que el mantenimiento de datos pobres y errores en el proceso de migración de un sistema a otro son algunos de los procesos que pueden provocar poca confianza en los datos. Una vez más, como en los trabajos [67, A10] [73, A48], insisten en la importancia que tiene la confianza en los datos en este tipo de sistemas. Los autores afirman que si la información que se analiza no es exacta o consistente, las organizaciones no pueden satisfacer las expectativas de sus clientes. Por otro lado, también resaltan que las necesidades de acceso a los datos, de parte de los usuarios, varían en una misma organización. Concluyen diciendo que algunas funciones críticas, como lo es la calidad de los datos, parecen haber alcanzado un nivel aceptable y que nuevas mejoras parecen no traducirse en un mayor éxito para un sistema de BI. Sin embargo, como en [73, A48] agregan que es necesario considerar más factores en los ambientes de tomas de decisiones, ya que hasta el momento sólo consideran dos factores (tipos de decisiones y necesidades de procesamiento de la información). Mencionan

ejemplos de otros factores importantes que deben ser considerados; estos son el rol de los responsables de la toma de decisiones, las preferencias e intereses de los mismos en la toma de decisiones y las jerarquías de las decisiones, entre otras. Esto último, permite deducir el interés que demuestran los autores en tener en cuenta información más detallada de los usuarios y sus preferencias. En definitiva, consideran que se deben tener en cuenta los diferentes ambientes de decisión, es decir, interesa la información del contexto de los usuarios y de las decisiones.

Resumen

La evaluación de los trabajos seleccionados, que permitió analizar la respuesta a las preguntas planteadas en las secciones anteriores (RQ1, RQ2 y RQ3), permite observar la compatibilidad existente entre la calidad de los datos, los contextos y los sistemas de DW. De todas formas, la búsqueda de bibliografía enfocada en la evaluación de la calidad de los datos, teniendo en cuenta el contexto en los sistemas de DW, no devolvió un número interesante de trabajos que presentara antecedentes en el área. Por esta razón, fue necesario realizar una investigación más general, combinando los temas de interés de dos en dos (*Data Warehouse* y *Data Quality*, *Data Warehouse* y *Context*, *Data Quality* y *Context*). Como se mencionó al inicio, un único trabajo fue devuelto por la cadena de búsqueda principal (que combina todos los temas que motivan a esta investigación: *Data Warehouse*, *Data Quality* y *Context*) y el mismo también fue devuelto para cada una de las cadenas de búsqueda parciales. Este resultado, hacía suponer que un solo trabajo era apto para intentar hallar una respuesta a RQ. Sin embargo, otros trabajos encontrados a partir de alguna de las búsquedas parciales, fueron seleccionados con el propósito de responder la *Research question* principal. Aunque en muchas ocasiones estos artículos no hacen referencia explícita a alguno de los temas de esta investigación, por su contenido y/o objetivo, se consideran de importancia para el trabajo abordado.

Del análisis realizado anteriormente, se observa que uno de los trabajos seleccionados propone integrar la calidad de los datos en toda la fase de desarrollo del DW y tiene en cuenta que diferentes usuarios pueden tener distintos criterios de calidad en función de sus requerimientos. A su vez, se encontró la afirmación de que la calidad es relativa, por lo que datos que son adecuados para un uso pueden no serlo para otro. Por otro lado, se obtuvieron trabajos que consideran las tareas de limpieza en la etapa de ETL y mencionan que la calidad es el grado en que los datos cumplen con las necesidades específicas de los usuarios. Dado que los intereses varían, los autores hacen referencia a la necesidad de adaptación de la calidad de los datos y en esta misma línea, otros investigadores, subrayan que los datos pueden ser vistos de distintas maneras respecto a su calidad, dependiendo del usuario y de la tarea en la cual están siendo utilizados. También se observó que en algunos trabajos se apoyan en la gestión de la confianza en los sistemas de DW, para determinar la calidad de los datos y, en este caso, consideran que el

contexto (para los autores, una situación dada), es muy importante para definir dicha relación de confianza. Adicionalmente, trabajos que se basan en la toma de decisiones, se enfocan en el negocio objetivo. Es decir, el dominio que determina el contexto de la toma de decisiones y que además permite definir los factores de calidad necesarios para dicho negocio. Como se puede observar, todos estos trabajos, en mayor o en menor medida, relacionan la calidad de los datos, los contextos y los sistemas de DW. En particular, sobresale el trabajo [74, A59] que combina las tres áreas de interés, definiendo un modelo de calidad que considera el análisis multidimensional y dando contexto a los datos de un DW, donde dicho contexto está determinado por el dominio de análisis.

Por otro lado, no se destaca un gran número de trabajos enfocados en la evaluación de la Calidad de los Datos en los SDW, para el período definido (2008-2015). Sin embargo, resaltan aquellos que hacen uso de los conceptos multidimensionales para abordar los temas de Calidad de Datos y/o Contextos. En particular, se observa como los investigadores relacionan al contexto de los datos con la calidad de los mismos apoyándose, para la visualización de los datos, en el concepto multidimensional de los SDW y de las aplicaciones OLAP

En base a todos estos trabajos se puede afirmar que es importante y presenta ventajas el tener en cuenta un contexto para la evaluación de la calidad de los datos, independientemente del sistema de información que se esté considerando. Sin embargo, no está determinado cuál es el contexto más adecuado, el de los datos, el de la tarea, el de los usuarios, etc; ya que se observa que la definición del contexto puede ser muy variada. Además, para los casos en los que el contexto de interés está bien definido, no existe una única forma de utilizar dicho contexto. En particular, para los sistemas de DW, no se encuentra una definición formal de contexto y para los casos en los cuales está definido, no se plantea cómo el mismo puede ser usado. Por lo tanto, es posible plantear, a partir de esta búsqueda bibliográfica, que hay una falta de abordaje al problema de definir y aplicar contextos para la evaluación de la calidad en el entorno de los SDW.

3.7. Otros resultados relevantes

En esta sección se presentan resultados que, si bien no son respuestas directas a las *Research questions* planteadas, se consideran de sumo interés para el problema general de la evaluación de la calidad de los datos, teniendo en cuenta el contexto de los mismos, en un SDW. Una vez finalizado el análisis de los artículos seleccionados, se realizó la extracción de información que demuestra las distintas necesidades y/o tendencias de las investigaciones consultadas.

En la Tabla 3.13 se presenta una clasificación de los artículos de acuerdo con la definición de contexto que los distintos autores presentan en la bibliografía con-

sultada. Además, se muestra para qué o quién definen dicho contexto y cómo éste está determinado. En esta tabla, el título de la columna "Objeto" se corresponde con el concepto de para qué o para quién definen contexto, en la columna "Contexto" se muestra el concepto que determina qué es el contexto y en la columna "Componentes del contexto" se describe cómo está formado dicho contexto.

Objeto	**Contexto**	**Componentes del Contexto**	**Artículo**
Agregación	El cubo de datos, la medida y las dimensiones específicas del DW	Información del cubo de datos, las preferencias del usuario, la función de agregación y las especificaciones acerca de cómo la función será realizada	[86, A24]
Calidad de datos o información	Datos	Aspectos dimensionales de los datos, relaciones entre categorías jerárquicas, restricciones dimensionales y reglas.	[65, A20]
	Proceso de producción de productos de información	Observaciones, documentos y conversaciones informales de la empresa.	[112, A45]
	Ambiente	Información del ambiente, del usuario y de la tarea que se está realizando	[115, A46]
	Toma de decisiones	Categorías de decisiones (por ejemplo, decisiones tácticas o de crédito), el peso con el que un factor de calidad de datos influye sobre una categoría de decisión y el nivel de confianza del usuario para cada factor de calidad.	[73, A48]
	Usuarios	Requerimientos del usuario.	[124, A53]
		El propósito de los usuarios en una tarea específica.	[71, A34] [110, A43] [116, A47] [118, A54]

Continúa en la página siguiente

Tabla 3.13: **Contextos**

Continuación de la página anterior

Objeto	Contexto	Componentes del Contexto	Artículo
Consultas	Usuarios y especificaciones de las solicitudes	Información del usuario (personal, sus hábitos, intereses y necesidades). Las coordenadas espacio-temporales de la solicitud, su motivación y su entorno (por ejemplo, urgencia).	[113, A42]
	Usuarios	Características y preferencias del usuario.	[94, A58]
Datos	Ambiente	Identificación de la fuente (sensor), identificación de la entidad para la cual el sensor recoge datos, el tiempo en el que se recogen dichos datos y los datos en sí.	[111, A44]
		Un sujeto, un predicado y un objeto que dan nombre y contenido al contexto (ej.: Juan entra a una habitación), y el alcance determinado por la relevancia del objeto en el predicado.	[121, A50]
	Calidad de los datos, usuario, sensor	Características (CPU, memoria, energía, etc.) y medición del sensor, preferencias del usuario, información de la tarea que se está realizando.	[117, A49]
	Requerimientos de calidad	Restricciones definidas a través de reglas.	[114, A62]
Entorno de recomendación multidimensional	Usuario	Perfil de usuario y sus preferencias.	[68, A14]
Lenguaje de programación de flujo de datos	Datos	Etiquetas del sistema de flujo de datos.	[78, A09]

Continúa en la página siguiente

Tabla 3.13: **Contextos**

Continuación de la página anterior

Objeto	**Contexto**	**Componentes del Contexto**	**Artículo**
Fuentes de datos	Estructura de las fuentes	Todas las fuentes de datos con estructura similar, determinada por columnas, registros, etc.	[119, A55]
Medidas del DW	Documentos	Contenido de los documentos.	[82, A11]
	Dimensiones del DW	Datos de las dimensiones.	[81, A03] [66, A04] [88, A05] [76, A07] [84, A13] [96, A19] [85, A23] [87, A25] [102, A39] [90, A57]
Procesos ETL	Usuarios	Elementos compartidos entre expertos con iguales intereses en las distintas etapas del proceso ETL. Por ejemplo, los parámetros para conectarse a una fuente de datos.	[91, A15]
Tarea	Conjunto de variables ligadas en técnicas de *Data Mining*	Datos del banco, de sus clientes y sus proveedores (localidades, actividades, etc).	[77, A08]
	Usuarios	Detalles de los clientes y sus ubicaciones.	[93, A17]
Toma de decisiones	Tarea	Información que describe la tarea OLAP: contenido de documentos, metadatos de los documentos (autor, fecha de creación, temas, etc.), preferencias y perfil (historial de consultas OLAP) de usuario.	[80, A02]
	Ambiente	Agentes de software, sistemas de información, proveedores, clientes, etc.	[83, A12]
		Los tipos de decisiones y las necesidades de información del usuario.	[75, A40]

Continúa en la página siguiente

Tabla 3.13: **Contextos**

Continuación de la página anterior

Objeto	**Contexto**	**Componentes del Contexto**	**Artículo**
	Usuarios	Preferencias del usuario y sus especificaciones (hechos y medidas del DW que han sido seleccionadas).	[92, A16]
		Necesidades financieras y de uso.	[95, A18]
	Calidad de los datos	Aspectos de los datos/información disponible, como completitud, correctitud, relevancia, etc.	[122, A51]
Usuario	Documentos	Contenido de los documentos.	[79, A01]
Visualización en BI	Usuarios	Anotaciones: texto provisto por el usuario, autor, fecha de creación, propiedades definidas por el autor (reglas de validez y tiempo de vida), objetivo de los datos, entidades a las cuales la anotación se refiere (el gráfico o la tabla de datos en la que fue hecha la anotación) y otras anotaciones relacionadas.	[89, A06]
		Perfil, capacidad y propósitos del usuario.	[97, A21]

Tabla 3.13: **Contextos**

Como se observa en la Tabla 3.13, en muchos trabajos resulta intuitivo entender cuál es el contexto que está siendo considerado y cómo éste está descrito. Un ejemplo de esto es el trabajo [124, A53], el cual tiene en cuenta el contexto de la calidad de los datos. Dicho contexto está determinado por los usuarios que participan en la toma de decisiones y está formado por los requerimientos de estos usuarios.

Por otro lado, en el trabajo [86, A24], se concentran en el contexto de las agregaciones. Cabe destacar que una agregación es realizada para una medida específica de un DW, para una dimensión y un nivel de dimensión, y el resultado de la misma es un nuevo cubo de datos. Los autores consideran que las agregaciones son dependientes del contexto, ya que las funciones de agregación que se pueden aplicar en un momento dado muchas veces dependen de las funciones aplicadas

previamente. Además, tienen en cuenta que se debe especificar cómo elegir la función de agregación y cómo realizar la agregación una vez que la función de agregación ha sido elegida. Otro caso similar es el trabajo [113, A42], en el cual destacan que la gran cantidad de información generada por los usuarios y sus aplicaciones es a menudo subutilizada. Además, la recuperación de información de las fuentes de datos muchas veces es irrelevante. Por esta razón, los autores consideran que el procesamiento de solicitudes complejas sobre tales fuentes de información es costoso y no garantiza la satisfacción del usuario. Por todo esto, presentan una solución al problema antes planteado, teniendo en cuenta el contexto de las consultas planteadas a las diferentes fuentes de datos.

Se da contexto a:	TOTAL
Calidad de datos/información	9
Datos	4
Fuentes de datos	1
Medidas del DW	11
Procesos ETL	1
Tarea	2
Toma de decisiones	6
Usuario	1

Tabla 3.14: **Objetos contextualizados**

Figura 3.5: **Cantidad de artículos por objeto contextualizado**

Los trabajos [111, A44] [117, A49] [121, A50] están centrados en contextos "sensados", concepto manejado en los entornos de la computación ubicua (del inglés Ubiquitous computing). Es de gran importancia, para este tipo de entorno, tener en cuenta el contexto en el cual los datos son capturados, y en este caso, la fuente de los datos son sensores. También en [119, A55] se enfocan en un entorno de múltiples fuentes en el cual los autores consideran que una misma entidad es almacenada de forma redundante en las diferentes fuentes de datos que participan. Por tanto, consideran que entidades con igual estructura representan a una misma entidad. Dicha estructura (o sintáxis como le llaman los autores del trabajo), determina el contexto de las fuentes de datos. Con otros objetivos el trabajo [78, A09] está centrado en un lenguaje de programación de flujo de datos. En este caso los valores que toman las variables dan contexto a los programas de flujo de datos y el conjunto de las etiquetas que definen al sistema de flujo determinan los contextos posibles para cada programa.

Por otro lado, en el trabajo [68, A14] proponen integrar información contextual en un entorno de recomendación multidimensional, teniendo en cuenta su contexto y buscando aplicar dicha integración en sistemas OLAP. Siguiendo en el entorno de los sistemas de BI, en el trabajo [97, A21] consideran que el contexto de visualización es necesario para ayudar a las organizaciones en la toma de decisiones. En particular, buscando resolver problemas multidominio que, según los autores, aún no han sido resueltos.

Finalmente, se destacan los trabajos que resultan de mayor relevancia para el problema de investigación planteado en esta tesis, ya sea por el enfoque dado por los autores o por los temas abordados. De acuerdo con la Tabla 3.13, estos trabajos son aquellos que definen el contexto de las medidas de un DW. Según los autores de [81, A03] [66, A04] [88, A05] [76, A07] [84, A13] [96, A19] [85, A23] [87, A25] [102, A39] [90, A57], el contexto de las medidas de un DW está determinado por las dimensiones del mismo. En particular para los autores de [82, A11], el contexto está determinado por los documentos relevantes de la organización y sus contenidos son los que dan forma a dicho contexto. Las medidas del DW son datos y por tanto los trabajos que se enfocan en el contexto de las mismas podrían estar agrupados con los trabajos que específicamente consideran contextos de datos. Sin embargo, por el interés antes mencionado, resulta relevante considerar la separación de estos trabajos, poniendo énfasis en la existencia de los mismos.

De la Tabla 3.13 se extrajeron los objetos para los cuales, en los trabajos analizados, se ha definido un contexto y que a su vez tienen algún punto de contacto con esta tesis. En la Tabla 3.14 se listan dichos objetos y se adjunta la cantidad de artículos que se han enfocado en el estudio de los mismos. Por otro lado, también resulta interesante destacar cuales fueron los contextos utilizados, asimismo, se seleccionaron aquellos que presentan algún tipo de relación con esta tesis. Esto

último se muestra en la Tabla 3.15.

Contexto	TOTAL
Ambiente	4
Calidad de datos/información	2
Datos	3
Dimensiones del DW	11
Documentos	2
Requerimientos de calidad	1
Tarea	1
Usuario	15

Tabla 3.15: **Contextos considerados**

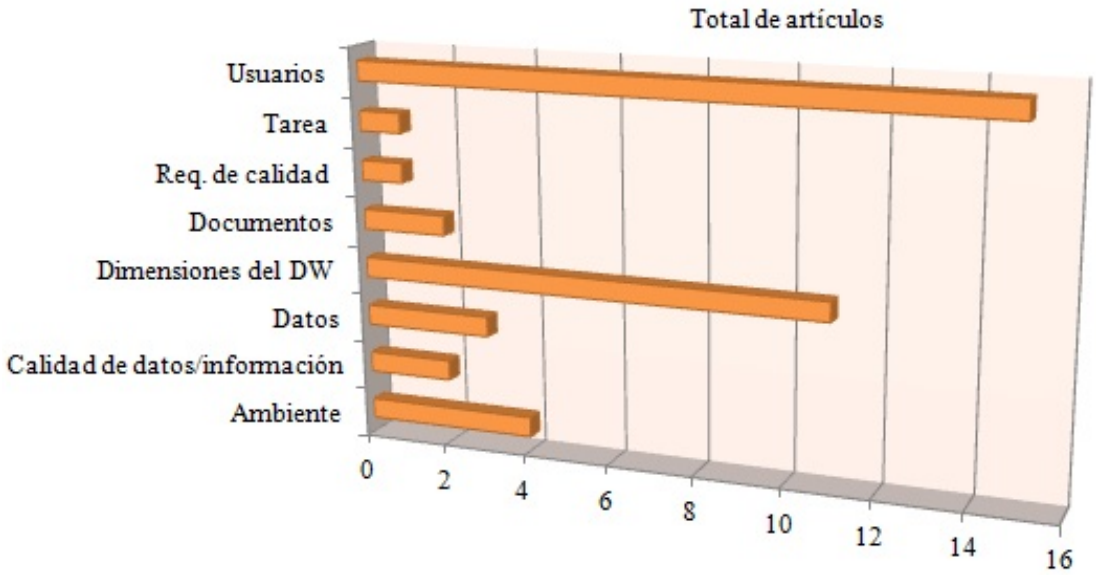

Figura 3.6: **Cantidad de artículos por contexto considerado**

Las Figuras 3.5 y 3.6 representan los datos mostrados en las Tablas 3.14 y 3.15 respectivamente. Resulta interesante destacar como objetos para los cuales, en algunos trabajos, se consideró importante definir un contexto, en otros trabajo han decidido usarlos para determinar el contexto de otros objetos. Un ejemplo de esto se observa para los casos: "Usuarios", "Calidad de los datos/información" y "Datos". En la Figura 3.5 es posible observar que la mayor cantidad de artículos consultados buscan dar contexto al objeto "Medidas del DW" mientras que, a partir de la Figura 3.6, se destaca que los "Usuarios" fueron los más empleados por los investigadores para dar contexto a diversos objetos.

Modelos presentados

Muchos trabajos no solo se enfocan en definir el contexto de determinado objeto, sino que además desarrollan una representación formal de dicho contexto. Por esta razón, en la Tabla 3.16 se destacan aquellos artículos que presentan un modelo del contexto analizado. Además, resultan de interés algunos trabajos que incorporan otros modelos, de los cuales se seleccionan aquellos que pueden apoyar la investigación de esta tesis. En particular, sólo se consideran trabajos que definen modelos para la representación de Contextos, Calidad de Datos y Usuarios. Trabajos que referencian y/o analizan modelos de otros trabajos, no son tenidos en cuenta. Como se observa en la Tabla 3.16, para la mayoría de los trabajos seleccionados, los investigadores encuentran necesario presentar un modelo de Calidad de Datos, en algunos casos como foco central del artículo y en otros como un aporte adicional.

Modelo	Artículos	TOTAL
Contexto	[77, A08] [83, A12] [91, A15] [92, A16] [93, A17] [65, A20] [97, A21] [111, A44] [117, A49] [121, A50]	10
Calidad de datos	[34, A26] [99, A27] [69, A28] [100, A30] [72, A36] [103, A41] [113, A42] [110, A43] [112, A45] [116, A47] [73, A48] [124, A53] [74, A59]	13
Usuarios	[87, A25] [72, A36]	2

Tabla 3.16: **Modelos**

Utilización y/o definición de reglas

Durante el análisis de los artículos se observó que muchos trabajos presentaban la inclusión de reglas para la solución de los problemas planteados. Por tanto, surgió el interés de clasificar aquellas reglas planteadas en los artículos consultados, independientemente del tema central de los mismos, ya que podrían ser de utilidad en la continuidad de la investigación planteada en este trabajo.

En la Tabla 3.17 se presenta una clasificación para las reglas seleccionadas, que de acuerdo con la finalidad para la cual han sido planteadas, son agrupadas en tres grandes ítems: “Calidad de datos”, “Negocio” y “Usuario”. Por otro lado, si bien las reglas de “Calidad de datos” pueden ser clasificadas como reglas de “Negocio”, porque describen características esenciales del dominio que se está

considerando, en este trabajo resulta de interés separar estos ítems, ya que Calidad de datos es uno de los tres grandes temas abordados en esta tesis.

Clasificación	**Utilizadas para:**	**Artículo**
Calidad de datos	Detectar si los datos fuente están/son contenidos/iguales en/a los datos de referencia.	[71, A34]
	Chequear los datos, marcar registros erróneos o hacer cambios de datos inválidos.	[34, A26]
	Filtrar datos de acuerdo con las expectativas de calidad de los datos.	[71, A34]
	Determinar qué datos son confiables y cuáles no cuando se perciben iguales datos de diferentes fuentes de datos.	[106, A35]
	Extraer datos, de un conjunto de datos mayor, que necesitan ser limpiados.	[99, A27]
	Mejorar la calidad de los contextos.	[117, A49]
	Refinar los procesos intentando encontrar nuevas restricciones.	[85, A23]
	Resolver inconsistencias como registros duplicados, valores faltantes, etc.	[99, A27]
	Determinar el período de tiempo de vida de una anotación (por ejemplo, válida mientras el valor de las ventas este entre $3000 y $5000).	[89, A06]
	Ser satisfechas por todos los datos de una fuente de datos.	[123, A52]
Negocio	Utilizar razonamiento que se aproxime al razonamiento humano.	[71, A34]
	Ilustrar un proceso sistemático cercano a los requerimientos en lenguaje natural.	[88, A05]
	Enriquecer los datos, generando nuevos datos mediante su ejecución.	[65, A20]
	Distinguir datos erróneos o perdidos en base a cierta definición del negocio.	[34, A26]
	Aplicar a los datos de acuerdo con la realidad del dominio.	[99, A27]
	Medir el grado de cumplimiento de los valores de los datos.	[69, A28]
	Verificar la consistencia de los datos respecto al dominio de interés.	[71, A34]

Continúa en la página siguiente

Tabla 3.17: **Reglas**

		Continuación de la página anterior
Clasificación	**Utilizadas para:**	**Artículo**
	Mantener el vínculo entre dos unidades de texto, definen reglas de propagación del contexto de análisis de datos.	[79, A01]
Usuarios	Personalizar el sistema y adaptarse a las preferencias de cada usuario que toma las decisiones. Se generan automáticamente.	[92, A16]
	Especificar necesidades espaciales de los usuarios.	[87, A25]
	Aclarar preferencias de los usuarios con exigencias especiales.	[68, A14]
	Seleccionar la función de agregación más correcta y preferida por el usuario.	[86, A24]
	Proporcionar al usuario sugerencias generadas por agentes de cooperación.	[77, A08]
	Aclarar los objetivos de clientes y proveedores.	[83, A12]

Tabla 3.17: **Reglas**

Clasificación	**TOTAL**
Calidad de datos	10
Negocio	8
Usuarios	6

Tabla 3.18: **Agrupación de las reglas**

En la Tabla 3.18 se observa que la mayor cantidad de reglas definidas en el conjunto de trabajos seleccionados son utilizadas para representar diferentes aspectos de la calidad de los datos.

Abordaje de la calidad de los datos

De acuerdo con los resultados obtenidos en las Tablas 3.19 y 3.20, independientemente de la consideración de un contexto, se observa que las investigaciones ponen énfasis en resolver las distintas problemáticas que se presentan a la hora de abordar la Calidad de los Datos. En base a esto, interesa detectar cuáles de estos trabajos consideran aspectos del contexto para llevar a cabo tareas de calidad de datos.

Tarea	**Consideran aspectos del Contexto**		**TOTAL**	
	SI	**NO**	**SI**	**NO**
Análisis	[65, A20] [122, A51]	[100, A30] [104, A32] [106, A35] [75, A40] [110, A43] [111, A44] [112, A45] [115, A46] [116, A47] [73, A48] [124, A53] [118, A54] [119, A55] [120, A56]	2	14
Limpieza de datos	[114, A62]	[65, A20] [34, A26] [99, A27] [69, A28] [100, A30] [105, A33] [71, A34] [106, A35] [72, A36] [108, A37] [109, A38] [103, A41]	1	12
Medición	[65, A20] [73, A48] [117, A49] [121, A50] [122, A51] [124, A53] [119, A55] [114, A62]	[100, A30] [104, A32] [106, A35] [75, A40] [110, A43] [115, A46] [116, A47] [123, A52] [118, A54] [120, A56]	8	10

Tabla 3.19: **Tareas de Calidad de Datos**

En la Figura 3.7 se observa la cantidad de trabajos que involucran aspectos del contexto en las diferentes tareas de Calidad de Datos. Además, se consideran aquellos artículos que consideran tareas de Calidad sin tener en cuenta aspectos del contexto. De esta forma, se puede observar la diferencia que se presenta entre estos trabajos que tienen en común el abordaje de alguna o todas las tareas de Calidad de datos. Resulta interesante observar como la medición de la calidad de los datos es la tarea en la que mayormente se tienen en cuenta distintos aspectos de un determinado contexto.

Como es notorio, no son muchos los trabajos que plantean la medición de la calidad de los datos teniendo en cuenta elementos y/o aspectos del contexto, como pueden ser las necesidades del usuario en una situación específica, distancia entre un sensor y un objetivo, confianza de los usuarios en la calidad de los datos de determinada empresa, etc. Sin embargo, aquellos trabajos que involucran al contexto en la definición de las métricas de calidad, resaltan la importancia de la consideración del mismo. El trabajo [73, A48] es un ejemplo de esto y si bien no está centrado en medir la calidad explícitamente, los autores miden la relación existente entre los resultados de un negocio y los factores de calidad de datos. Para esto, consideran elementos del contexto como lo es la categoría de la decisión del negocio que se está evaluando. Además, tienen en cuenta la importancia aplicada a cada factor de calidad y la confianza que el usuario le asigna a los datos. Los

autores destacan la importancia de incluir elementos del contexto en la medición de la calidad de los datos. Por otro lado, los investigadores de [122, A51] analizan la relación existente entre una medida de la calidad de la información y aspectos de la tarea que se está ejecutando. En este trabajo consideran que elementos del contexto, como lo es la tarea en cuestión, pueden alterar los resultados de la medición de la calidad de los datos.

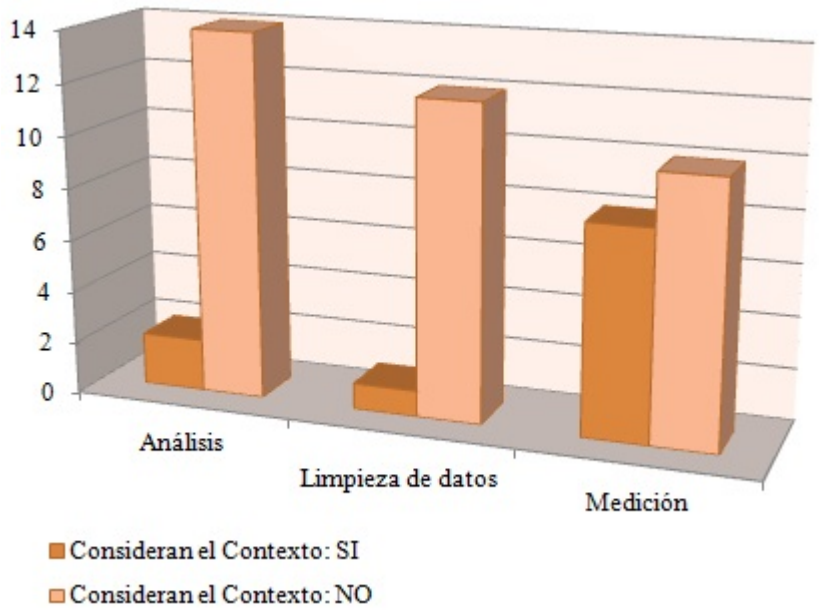

Figura 3.7: **Tareas de calidad de datos y aspectos del contexto**

Dimensiones de Calidad de Datos

Otro resultado adicional y de interés para el área Calidad de Datos es la identificación de las distintas dimensiones de calidad que son abordadas en los artículos consultados. Este estudio se realiza por separado, ya que interesa distinguir las dimensiones de calidad abordadas en los artículos centrados en los temas DQ y DW de las dimensiones de calidad presentes en artículos centrados en los temas DQ y CTX. La Tabla 3.20 presenta todas las dimensiones de calidad estudiadas en los trabajos que se enfocan en la Calidad de los Datos y Contextos, mientras que la Tabla 3.21 muestra las dimensiones tenidas en cuenta en los artículos centrados en los temas Calidad de Datos y los Sistemas de *Data Warehousing*.

Por otro lado, interesa resaltar que en las tablas 3.20 y 3.21 no se presenta una conceptualización propia, sino que se muestra cómo los distintos autores presentan los diferentes conceptos. Es decir, interesa señalar la variedad de conceptos utilizados en la bibliografía. Por ejemplo, en la Tabla 3.20 se observan las dimensiones *accuracy*, *correctness* y *free of error*, que refieren a una misma di-

mensión de calidad que generalmente es llamada *accuracy*. Por lo tanto, diferente terminología es usada para un mismo concepto, lo que confirma una vez más la no estandarización de los conceptos en el área de la calidad de los datos.

Dimensión	Artículos	TOTAL
Accessibility	[112, A45] [118, A54]	2
Accuracy	[113, A42] [112, A45] [117, A49] [118, A54] [119, A55] [120, A56]	6
Actionable	[118, A54]	1
Aggregation	[73, A48]	1
Alignment	[118, A54]	1
Appropriate-amount	[118, A54]	1
Authority	[117, A49]	1
Availability	[123, A52]	1
Certainty	[117, A49]	1
Complacency	[116, A47]	1
Completeness	[111, A44] [112, A45] [115, A46] [117, A49] [118, A54] [120, A56]	6
Concisely-Represented	[118, A54]	1
Confidence	[116, A47]	1
Consensus	[116, A47]	1
Consistency	[112, A45] [115, A46] [120, A56]	3
Correctness	[110, A43]	1
Cost	[117, A49]	1
Credibility	[124, A53]	1
Currency	[112, A45] [123, A52]	2
Delay	[117, A49]	1
Distortion	[117, A49]	1
Easily-understandable	[118, A54]	1
Efficiency	[116, A47] [117, A49]	2
Equivalence	[117, A49]	1
Free-of-error	[115, A46]	1
Freshness	[117, A49] [121, A50]	2
Granularity	[73, A48]	1
Interpretability	[118, A54]	1
Locality	[117, A49]	1
Objectivity	[112, A45] [118, A54]	2
Precision	[117, A49]	1

Continúa en la página siguiente

Tabla 3.20: **Dimensiones de calidad en artículos de DQ y CTX**

Continuación de la página anterior

Dimensión	Artículos	TOTAL
Priority	[121, A50]	1
Reasoning Accuracy	[117, A49]	1
Relativity	[117, A49]	1
Relevancy	[73, A48] [118, A54]	2
Reliability	[112, A45] [117, A49] [121, A50] [124, A53]	4
Representational-consistent	[118, A54]	1
Reputability	[118, A54]	1
Resolution	[117, A49]	1
Responsiveness	[122, A51]	1
Scope	[112, A45]	1
Security	[117, A49] [118, A54]	2
Shared understanding	[122, A51]	1
Significance	[111, A44] [117, A49]	2
Timeliness	[112, A45] [115, A46] [122, A51] [124, A53] [118, A54] [120, A56]	6
Traceability	[118, A54]	1
Trust-worthiness	[111, A44] [117, A49]	2
Up-to-dateness	[111, A44] [117, A49]	2
Usability	[117, A49]	1
Usage of Resource	[117, A49]	1
Usefulness	[110, A43]	1
Validity	[117, A49] [123, A52]	2
Value-added	[118, A54]	1
Volatility	[117, A49]	1

Tabla 3.20: **Dimensiones de calidad en artículos de DQ y CTX**

Como se observa, en los trabajos centrados en los temas Calidad de Datos y Contextos, se ha definido un mayor número de dimensiones de calidad, mientras que en los trabajos enfocados en las áreas Calidad de Datos y *Data Warehouse* el conjunto de dimensiones tenidas en cuenta es mucho más pequeño. Por otro lado, para todos los casos, las dimensiones de calidad *accuracy*, *completeness* y *timeliness* son las más utilizadas.

Dimensión	Artículos	TOTAL
Accuracy	[100, A30] [101, A31] [71, A34] [106, A35] [74, A59]	5
Completeness	[100, A30] [101, A31] [71, A34] [74, A59] [107, A61]	5
Complexity	[57, A60]	1
Consistency	[100, A30] [101, A31] [74, A59] [107, A61]	4
Correctness	[57, A60] [107, A61]	2
Effectiveness	[57, A60]	1
Efficiency	[57, A60]	1
Expressiveness	[57, A60]	1
Freshness	[74, A59]	1
Integrity	[106, A35]	1
Legibility	[57, A60]	1
Reasonableness	[104, A32]	1
Reliability	[75, A40] [74, A59]	2
Simplicity	[57, A60]	1
Temporality	[104, A32]	1
Timeliness	[101, A31] [104, A32] [71, A34]	3
Transparency	[107, A61]	1
Trust-worthiness	[104, A32] [107, A61]	2
Understandability	[57, A60]	1
Uniqueness	[101, A31] [74, A59]	2
Validity	[75, A40]	1

Tabla 3.21: **Dimensiones de calidad en artículos de DQ y DW**

3.8. Conclusiones

La metodología aplicada permite tener una visión general de la investigación actual en las áreas temáticas de interés: *Data Quality*, *Data Warehouse* y *Context*, aplicando unas serie de pasos bien definidos. En particular, los resultados obtenidos permiten destacar las necesidades y tendencias que presentan los trabajos de investigación publicados dentro del período 2008-2015. Si bien las bibliotecas digitales utilizadas para las búsquedas fueron solo tres, por una cuestión de alcance y de acuerdo con las pautas del MS, se considera que los resultados que se presentaron en la Sección 3.7 son lo suficientemente relevantes para esta tesis.

Es importante destacar que un número muy bajo de trabajos fue seleccionado, como se observa en la Tabla 3.11, sólo un 10 % del total de los artículos devueltos por las búsquedas, es relevante para esta tesis. Estos resultados se dan así por lo abarcativas que son las cadenas de búsqueda. Cada una de ellas consta de una palabra clave y una serie de términos alternativos a las mismas, que cada uno en sí mismo, abarca un espectro muy amplio de investigaciones. Por ejemplo, si solo se tuvieran en cuenta las palabras claves: *Data Quality*, *Data Warehouse* y *Context*, cada una de ellas determina un área de investigación, lo que implica un gran rango de trabajos. Por esta razón, muchos trabajos fueron descartados durante las búsquedas y a partir de la aplicación de los criterios de exclusión. Por otro lado, cabe resaltar la naturaleza del significado de la palabra "Contexto" y el uso que se da a la misma. Por esta razón, la palabra "Contexto" es ampliamente utilizada para hacer referencia a distintos dominios por lo que, en muchas ocasiones, se encontraron resultados que si bien consideraban las áreas de investigación de DW o DQ, nada tenían que ver con el área de investigación: Contextos. Esto último también contribuye a que el número de trabajos encontrados sea tan alto respecto al número de trabajos seleccionados.

Por otro lado, es importante destacar que la metodología se aplicó en dos etapas, la primer etapa considera trabajos publicados desde el año 2008 hasta el año 2014, mientras que la segunda etapa considera trabajos desde el año 2014 hasta el año 2015. La primer etapa se realiza en el período marzo-junio del año 2014, mientras que la segunda etapa se lleva a cabo al finalizar la tesis, en octubre de 2015. El MS se ejecutó en dos etapas porque dado que la revisión bibliográfica finalizó en junio de 2014 y a más de un año de realizada la última búsqueda, podrían haber surgido trabajos de mucho interés para este trabajo. Por lo tanto, esto demuestra la alta capacidad de reproducir las búsquedas que ofrece la metodología. Además, cabe destacar que la ejecución del MS en la segunda etapa, fue realizada más ágilmente, ya que una vez definidas todas las cadenas de búsquedas basta con establecer el período para el cual se va a ejecutar la metodología.

Respecto a los trabajos seleccionado, en el caso de la *Research question* principal, si bien se identifica un único trabajo que relaciona explícitamente las tres

áreas temáticas, es posible seleccionar algunos trabajos que hacen referencia implícita a algunas de estas áreas, enfocándose al fin en todos los temas. Cabe destacar que algunos artículos seleccionados mediante la metodología de búsqueda, también fueron consultados para la realización de la primera revisión bibliográfica. Aunque dichos trabajos, [34, A26] [99, A27] [101, A31] [106, A35], ya habían sido analizados, se toma la decisión de tenerlos en cuenta, una vez más, en esta nueva etapa de análisis. Se considera importante aplicar la metodología independientemente del análisis realizado al inicio de esta tesis.

La evaluación de todos los trabajos seleccionados permite el análisis de las respuestas a las preguntas planteadas en las secciones anteriores: RQ1, RQ2 y RQ3. Esto, permite observar la compatibilidad existente entre la calidad de los datos, los contextos y los sistemas de DW, aunque la búsqueda de bibliografía enfocada en la evaluación de la calidad de los datos, teniendo en cuenta el contexto en los sistemas de DW, no devolvió un número interesante de trabajos que presentara antecedentes en el área. Además, no se destaca un gran número de trabajos enfocados en la evaluación de la Calidad de los Datos en los SDW, para el período definido (2008-2015). Sin embargo, son de gran importanica aquellos que hacen uso de los conceptos multidimensionales para abordar los temas de Calidad de Datos y/o Contextos. En particular, se observa como los investigadores relacionan al contexto de los datos con la calidad de los mismos, apoyándose para la visualización de los datos en el concepto multidimensional de los SDW y de las aplicaciones OLAP

En base a todos estos trabajos se puede afirmar que es importante y presenta ventajas el tener en cuenta un contexto para la evaluación de la calidad de los datos, independientemente del sistema de información que se esté considerando. Sin embargo, no está determinado cuál es el contexto más adecuado, el de los datos, el de la tarea, el de los usuarios, etc; ya que se observa que la definición del contexto puede ser muy variada. Además, para los casos en los que el contexto de interés está bien definido, no existe una única forma de utilizarlo. En particular, para los sistemas de DW, no se encuentra una definición formal de contexto y para los casos en los cuales está definido, no se plantea cómo el mismo puede ser utilizado. Por lo tanto, es posible plantear, a partir de esta búsqueda bibliográfica, que hay una falta de abordaje al problema de definir y aplicar contextos para la evaluación de la calidad en el entorno de los Sistemas de *Data Warehousing*.

Por lo tanto, la cantidad de estudios primarios obtenidos, las propuestas que los diferentes autores desarrollan y los resultados adicionales que se desprenden del análisis global, permiten concluir que la definición de Contextos presenta muchas ventajas para los SDW, en particular, para la evaluación de la Calidad de los Datos a lo largo de todo el ciclo de vida del DW. Por esta razón, en el siguiente capítulo se plantea una propuesta y la prueba de concepto correspondiente, que surge a partir de los resultados obtenidos en la realización del estado del arte, mediante la aplicación de la metodología *Mapping Study*.

Capítulo 4

Propuesta: Evaluación de la calidad en SDWs basada en contextos

Una vez finalizada la etapa de análisis de los trabajos seleccionados, y en base a los resultados obtenidos, se plantea una propuesta cuyo propósito es definir un marco donde se sitúen los problemas de calidad de datos en un Sistema de *Data Warehousing*. Además, interesa que este marco sirva como base para la gestión de la calidad. Para esto, se tienen en cuenta los distintos componentes que forman parte de este tipo de sistemas y los diferentes contextos que tienen influencia sobre ellos.

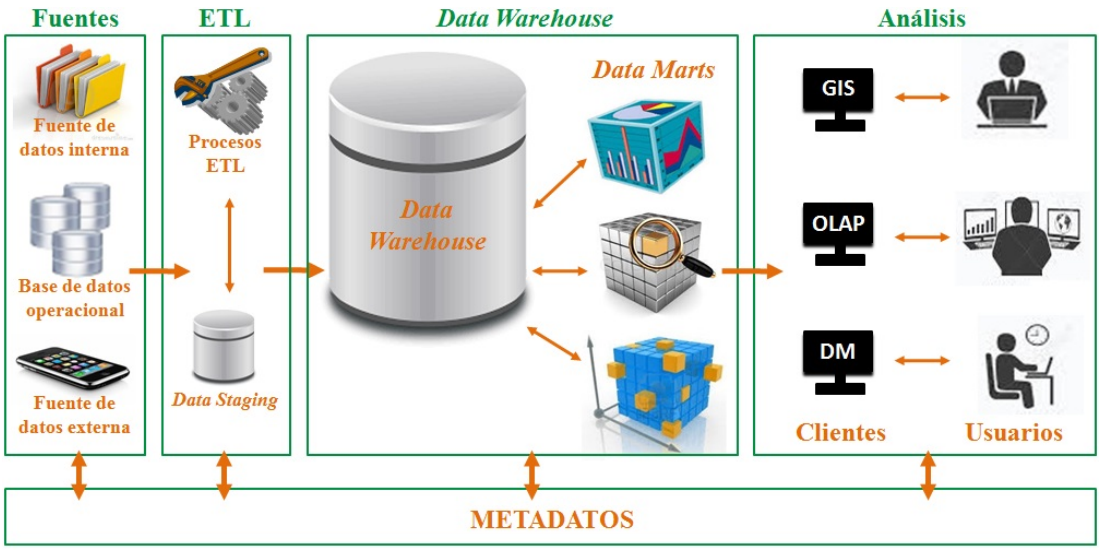

Figura 4.1: **Sistema de *Data Warehousing***

En la Figura 4.1 se observa la arquitectura de un SDW y los distintos componentes que lo conforman. Como se describe en la Sección 2, en primer lugar se observan las fuentes de datos de las cuales se obtienen los datos que, una vez realizado el proceso de extracción y transformación de los mismos, serán cargados en el DW. Después que los datos han sido procesados es posible generar todos los *Data Mart* necesarios, permitiendo así que los usuarios apliquen las herramientas de exploración y análisis (OLAP, *Data Mining*, reportes, estadísticas, etc.), que consideren pertinentes. Es importante destacar la presencia del repositorio de Metadatos en el SDW, ya que el mismo es de utilidad a la hora de almacenar los resultados obtenidos en la evaluación de la calidad de los datos, en los componentes del sistema.

Uno de los propósitos de esta tesis es definir contextos, presentes a lo largo de todo el ciclo de vida de un SDW, que son de interés para la evaluación de la calidad de datos, ya que pueden influir sobre la misma. Por lo tanto, esta propuesta está centrada en definir los distintos contextos que atraviesan los datos, desde que estos son cargados en el DW y hasta que los mismos son utilizados por los usuarios finales.

La visión de que los datos recorren diversos contextos mientras son utilizados, es compartida con los autores de [110, A43]. En dicho trabajo, los investigadores consideran más conveniente evaluar la calidad de la información como un proceso, más que como una medida estática. Además, agregan que esto permitiría captar mejor la esencia de la "información", yendo desde los datos hasta el conocimiento, pasando a través de distintos contextos. Además, interesa destacar los trabajos [91, A15] y [119, A55] que se enfocan en los contextos de los procesos de ETL y de las fuentes de datos, respectivamente.

Si bien estos dos componentes del SDW, fuentes y ETL, no serán considerados en esta propuesta para la definición de contextos, es de importancia resaltar la bibliografía que ya ha realizado este tipo de consideraciones. Por otro lado, tampoco es el objetivo de esta tesis el estudio de la calidad de datos en estos componentes, sin embargo, en secciones anteriores se presenta el análisis de la bibliografía referente a dicho tema de investigación. Además, muchos trabajos ya se han centrado en el estudio de la calidad de las fuentes de un SDW [35] [126] [25] [109, A38] [102, A39], así como también, se encuentra otro número importante de trabajos que se enfocan en la calidad, en especial en la limpieza de los datos, en los procesos de ETL [127] [34, A26] [99, A27] [69, A28] [105, A33] [71, A34] [108, A37] [103, A41].

4.1. Contexto en componentes del SDW

En esta sección se presentan y definen los contextos para los componentes que participan desde el momento en que los datos son cargados en el DW y hasta que estos son utilizados por los usuarios finales:

- Contexto en el *Data Warehouse*
- Contexto en el *Data Mart*
- Contexto en uso

En la Figura 4.2 se muestra cada componente con su respectivo contexto. A continuación se describen los elementos que conforman a cada uno de dichos contextos.

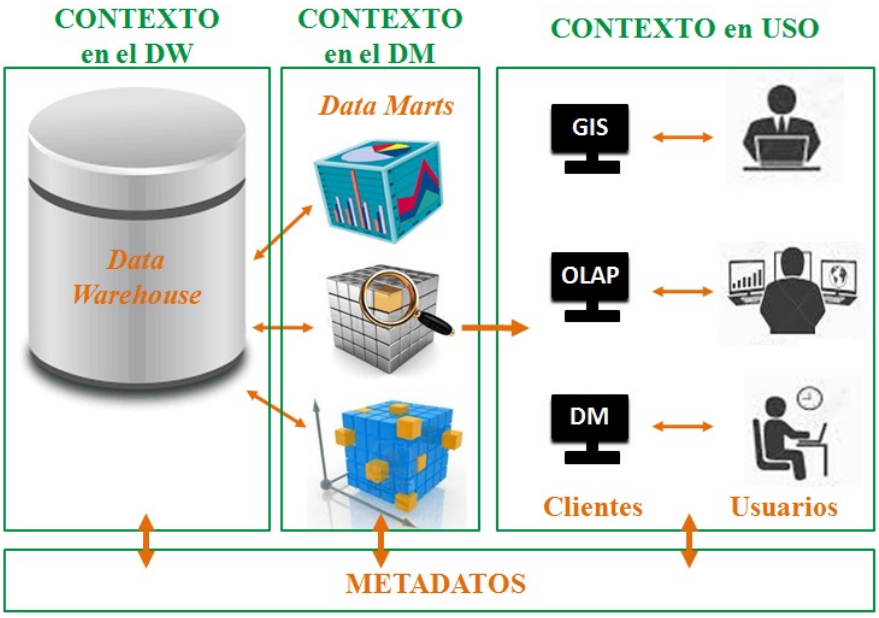

Figura 4.2: **Contextos en un Sistema de *Data Warehousing***

- Contexto en el *Data Warehouse*
  En esta propuesta, el contexto en el DW (DWC) está formado por los propios datos del DW, por documentos, correos electrónicos y otros datos. Estos últimos, pertenecientes a tablas de bases de datos personales de los distintos usuarios que integran la organización, que pueden estar vinculados con los datos del DW. Por otro lado, los documentos y correos electrónicos que se comparten dentro de la organización también contienen información fuertemente relacionada con los datos almacenados en el DW.

- Contexto en el *Data Mart*
  El DM contiene a un subconjunto de los datos almacenados en el DW, que pueden haber sufrido transformaciones, probablemente como resultado de la aplicación de agregaciones. A su vez, los datos que aquí se consideran están dirigidos a un dominio de análisis específico. Cada dominio de análisis está definido por un tema, una sección de la organización, un conjunto de usuarios, etc. En esta propuesta, el contexto en el DM (DMC) está determinado por un conjunto de reglas. Estas reglas describen propiedades, restricciones y requerimientos de calidad del dominio de análisis.

- Contexto en uso
  El contexto considerado en el cliente para la presentación de los datos está formado por los datos que describen al usuario final y se llama contexto en uso (CiU). En esta propuesta, el CiU está determinado por el usuario, por los datos que lo describen, como puede ser su ubicación geográfica, su idioma, su edad, su rol, una descripción de la tarea que está realizando, sus preferencias y sus requerimientos. Estos últimos pueden ser requerimientos de los datos o de calidad. Ejemplos de requerimientos de calidad pueden ser qué tan exactos deben ser ciertos datos o cuán completa debe estar la información requerida. En definitiva, el CiU es el contexto que está determinado por el uso que le da el usuario a los datos.

La información referente a los contextos podría ser almacenada en el repositorio de metadatos mostrado en la Figura 4.2. Un ejemplo de dicha información son las reglas de dominio, que determinan el contexto en el *Data Mart*.

Por otro lado, esta propuesta, que define contextos para componentes del SDW, se basa fuertemente en algunos de los trabajos que fueron analizados en la sección 3.7. Por ejemplo, en [79, A01] subrayan cómo los datos no estructurados, presentes en documentos y correos electrónicos, pueden vincularse con los datos y relaciones almacenados en un DW. También los autores de [82, A11] se apoyan en la información presente en este tipo de fuentes para dar contexto a las medidas del DW. En particular, interesa resaltar los trabajos [81, A03] [66, A04] [88, A05] [76, A07] [84, A13] [96, A19] [85, A23] [87, A25] [102, A39], para los cuales el contexto de las medidas de un DW está formado por los datos de las dimensiones que determinan a dicha medida. Por otro lado, la definición de reglas resulta de gran apoyo a la hora de definir un contexto. En [87, A25] utilizan un lenguaje de reglas de personalización, para especificar las necesidades de cada usuario. También en los trabajos [77, A08] y [83, A12] buscan colaborar con los usuarios mediante la definición de reglas. En el primer caso, proponen sugerencias, mientras que en el segundo buscan aclarar sus objetivos. Pero no solo se utilizan reglas centradas en los usuarios, trabajos como [99, A27] y [71, A34] presentan la utilización de las mismas para la validación de los datos en torno al dominio de interés. Finalmente, en el momento del análisis de los datos es cuando se hace más evidente el protagonismo del usuario y la necesidad de tener en cuenta datos que lo describan. Por

esta razón, trabajos como [71, A34] [110, A43] [116, A47] [118, A54] definen un contexto teniendo en cuenta aspectos que describen al usuario y a la tarea que está realizando.

4.2. Calidad de datos de acuerdo a su contexto

Para la evaluación de la calidad en los componentes del SDW, teniendo en cuenta los contextos antes presentados, esta tesis se apoya en dos enfoques de calidad que se presentan a continuación. La Figura 4.3 muestra gráficamente donde se aplican, en el SDW, dichos enfoques.

- *Meeting Requirements* de Crosby [128]: Este enfoque pone énfasis en el cumplimiento de los requerimientos del sistema y se centra en la prevención más que en la corrección, donde la única forma de lograr un buen rendimiento es teniendo cero defectos. La esencia de este enfoque está en que se debe conocer los requerimientos y traducir los mismos en características medibles del producto/servicio. Por tanto, la calidad de un producto/servicio equivale a la satisfacción de los criterios de especificación por parte de todas las características medibles de dicho producto/servicio [129]. El enfoque denominado *Meeting Requirements* es aplicado para evaluar la calidad de datos en el *Data Warehouse* y en los *Data Marts*.

- *Fitness for Use* de Juran [130]: Este enfoque está centrado en satisfacer las necesidades del usuario a través de la adecuación (del producto) al uso. El enfoque denominado *Fitness for Use* es aplicado para evaluar la calidad de datos en los clientes responsables de la presentación de los datos al usuario final.

Por lo tanto, a partir de los enfoques de calidad antes presentados, a continuación se define la calidad en los componentes del SDW, como se observa en la Figura 4.4.

- Calidad en el *Data Warehouse*
  La calidad en el DW (DWQ) depende del contexto del DW y los elementos que permiten definir métricas de calidad en el DW son aquellos que conforman a dicho contexto. En este caso, el contexto está formado por los datos del DW, por documentos, correos electrónicos y otros datos propios de la organización.

- Calidad en el *Data Mart*
  La calidad en el DM (DMQ) depende del contexto en el DM y los elementos que permiten definir métricas de calidad en el DM son aquellos que conforman a dicho contexto. En este caso, el contexto está determinado por un conjunto de reglas que caracterizan al dominio de análisis.

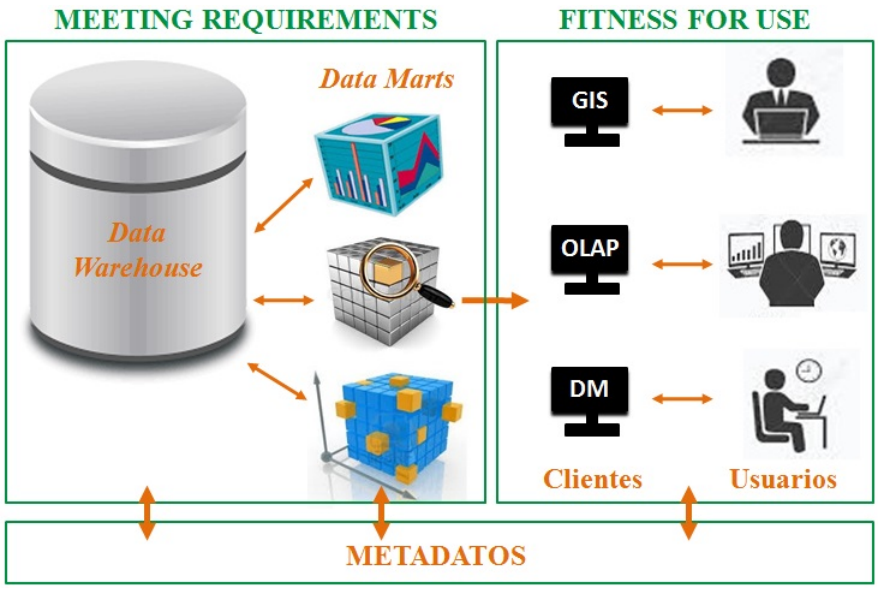

Figura 4.3: **Enfoques de calidad en un SDW**

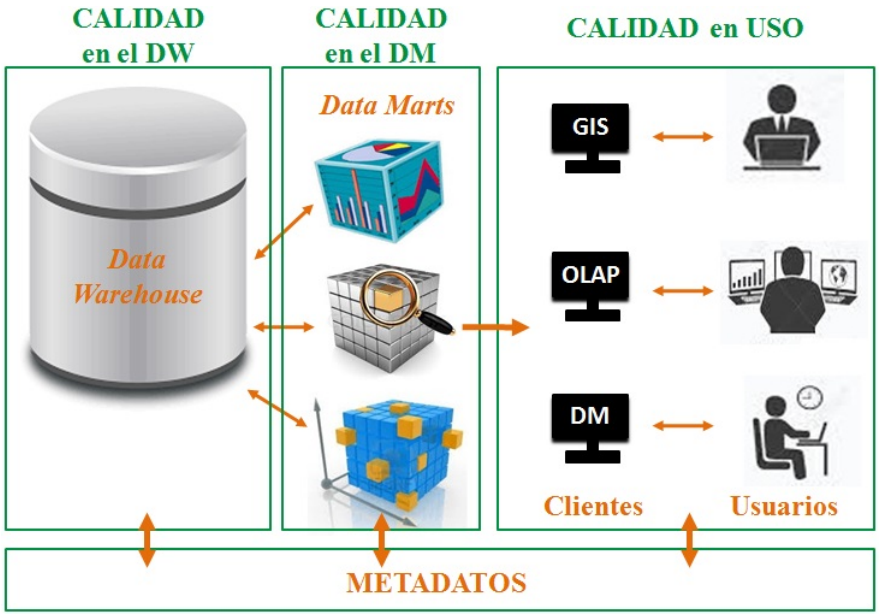

Figura 4.4: **Calidad en un Sistema de *Data Warehousing***

- Calidad en uso
  La calidad en uso (QiU) depende del contexto del usuario y los elementos que permiten definir métricas de calidad en el cliente responsable de la presentación de los datos son aquellos que conforman a dicho contexto. En este caso, el contexto está determinado por los datos que describen al usuario, por sus requerimientos y/o por la tarea que está realizando.

El concepto de calidad, aplicado en el cliente para la presentación de los datos del SDW, coincide con la definición de calidad en uso de la norma 25010 de ISO [131]. Dicha norma define calidad en uso como la calidad desde el punto de vista del usuario. Además, la norma define un conjunto de características de calidad, algunas de las cuales son divididas en subcaracterísticas. Un ejemplo de estas características es *usability* y el conjunto de subcaracterísticas para esta, según el estándar, son *effectiveness*, *efficiency* y *satisfaction*. Sin embargo, estos no son los conceptos que más interesan para esta tesis, sino que lo más relevante para este trabajo es que el modelo de calidad en uso evalúa la calidad en un contexto de uso particular que depende del punto de vista del usuario, tal como se define la "Calidad en uso" en esta sección. Si bien la norma 25010 de ISO se enfoca en un modelo de calidad para *Software*, el mismo puede ser adaptado y aplicado en otras áreas. Un ejemplo de esto es la aplicación de este estándar para la definición de un modelo de calidad en uso en portales Web [132].

4.3. Caso de estudio

En esta sección se presenta un caso de estudio que permite aplicar los conceptos antes introducidos. El ejemplo utilizado representa a una cadena de supermercados y dicha cadena mantiene información, acerca de sus ventas y promociones, en un DW. En las Figuras 4.5, 4.6 y 4.7 se presenta el esquema conceptual correspondiente al ejemplo. En primer lugar, en la Figura 4.5, se muestran las dimensiones participantes en el caso de estudio: "Producto", "Tiempo", "Sucursal" y "Promoción", con sus respectivas jerarquías. En las Figuras 4.6 y 4.7, se presentan las relaciones dimensionales "Ventas" y "Rebajas", respectivamente. Para la representación de las dimensiones y de las relaciones dimensionales se utiliza el modelo CMDM [20], el cual fue introducido en la Sección 2.1.1.

En las Tablas 4.1, 4.2, 4.3 y 4.4, se muestran ejemplos de instancias de las tablas de dimensión "Producto", "Tiempo", "Sucursal" y "Promoción", respectivamente. En cada tupla de la tabla de dimensión "Producto", Tabla 4.1, se presenta el identificador del producto, la familia a la cual pertenece, el tipo de producto y su categoría. En la tabla de dimensión "Tiempo", Tabla 4.2, cada tupla contiene la fecha, el mes y el año correspondiente. En la tabla de dimensión "Sucursal", Tabla 4.3, cada tupla contiene el identificador, nombre, ciudad, departamenteo o estado y país de cada sucursal. Además, en la tabla de dimensión

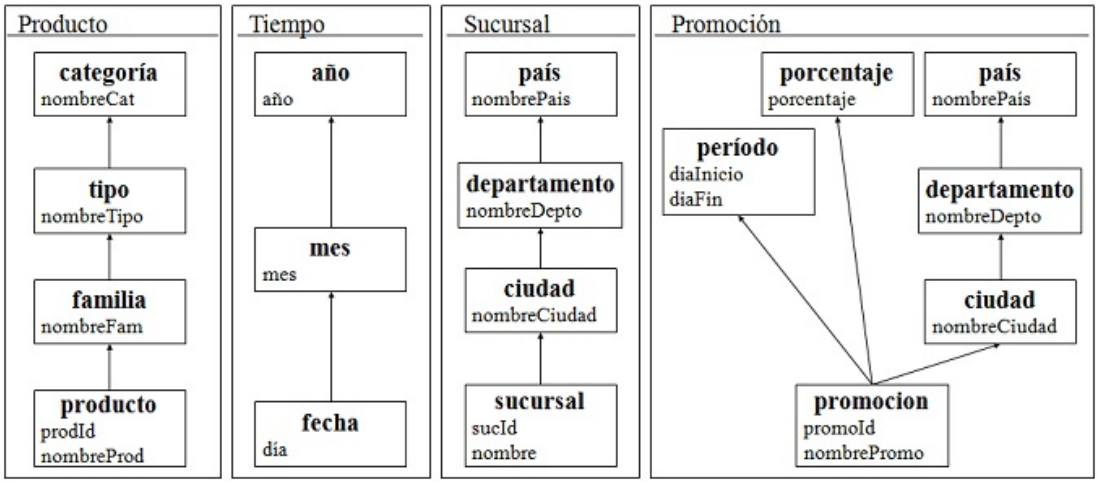

Figura 4.5: **Jerarquías y sus niveles, para cada una de las dimensiones**

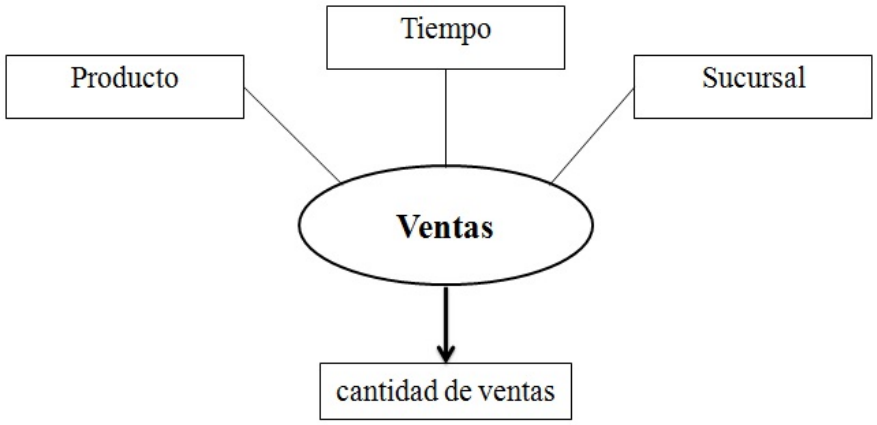

Figura 4.6: **Relación dimensional "Ventas"**

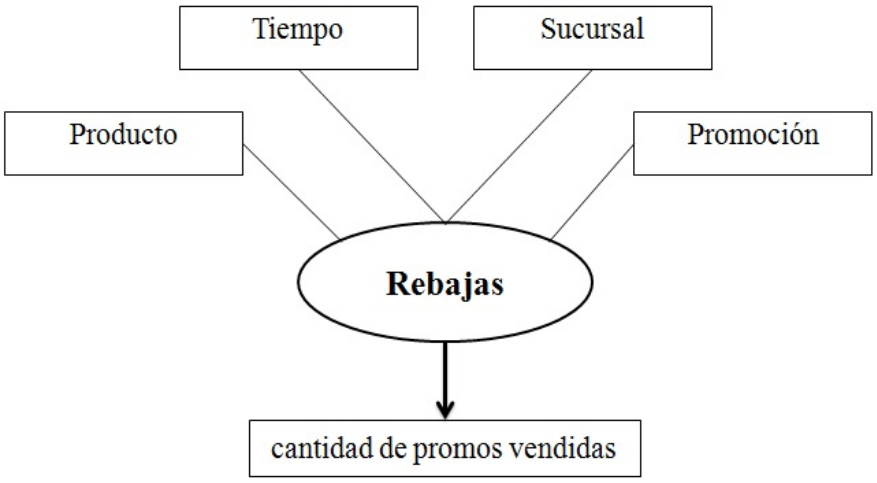

Figura 4.7: **Relación dimensional "Rebajas"**

prodId	**nombreProd**	**nombreFam**	**nombreTipo**	**nombreCat**
23	barraChocolate	chocolate	alimento	premium
15	surtidoChocolates	chocolate	alimento	premium
40	trufasChocolate	chocolate	alimento	premium
...	...	...	...	...
56	caramelosSurtidos	golosina	alimento	premium

Tabla 4.1: **Tabla de dimensión "Producto"**

día	**mes**	**año**
3/1/2014	1/2014	2014
...	...	...
5/4/2015	4/2015	2015
6/4/2015	4/2015	2015
7/4/2015	4/2015	2015
8/4/2015	4/2015	2015
...	...	...
10/11/2015	11/2015	2015

Tabla 4.2: **Tabla de dimensión "Tiempo"**

"Promoción", Tabla 4.4, cada tupla contiene el identificador de la promoción, su nombre y el período de duración de la misma (fecha de inicio y fecha de fin).

Por otro lado, se presentan las tablas de hechos "Ventas" y "Rebajas" en las Figuras 4.5 y 4.6, respectivamente. En la tabla de hechos "Ventas" se muestran las cantidades vendidas para cada producto en cada sucursal en una fecha dada. Por esto, en esta tabla cada tupla contiene un identificador del producto, la fecha de la venta, el identificador de la sucursal y la cantidad de unidades vendidas. En la tabla de hechos "Rebajas" se muestran las cantidades vendidas, en cada sucursal (de la ciudad en la cual se hace dicha promoción), en una fecha correspondiente al período de promoción. En esta tabla cada tupla contiene un identificador de la promoción, el identificador del producto que está en rebaja, la fecha de venta de dicho producto, el identificador de la sucursal en el cual se hizo la venta y la cantidad de unidades en promoción que han sido vendidas.

En la cadena de supermercados se identifican dos dominios de análisis de datos, Ventas y Publicidad, en los cuales se manifiestan diferentes requerimientos. A continuación se presentan ejemplos de calidad en el DW, de calidad en el DM y de calidad en uso para los dominios antes mencionados.

sucId	**nombre**	**nombreCiudad**	**nombreDepto**	**nombrePais**
25	SupermercadoTata	Maldonado	Maldonado	Uruguay
26	tata_Salto_26	Salto	Salto	Uruguay
28	tata_Florida_28	Florida	Florida	Uruguay
29	Florida_29	Florida	Florida	Uruguay
30	Florida_30	Florida	Florida	Uruguay
...	...	...	...	...
31	tata_Rocha_31	Rocha	Rocha	Uruguay

Tabla 4.3: **Tabla de dimensión "Sucursal"**

promoId	**nombrePromo**	**nombreCiudad**	**nombreDepto**	**nombrePais**	**diaInicio**	**diaFin**	**porcentaje**
p1	día del Niño	Florida	Florida	Urugay	1/1/2014	3/1/2014	20 %
p2	día del Padre	Colonia	Colonia	Urugay	10/4/2015	15/4/2015	50 %
p3	día de la Madre	Maldonado	Maldonado	Urugay	9/8/2015	11/8/2015	10 %
...	...	...	...	...	...	...	...
p56	outlet	Salto	Salto	Uruguay	9/10/2015	12/10/2015	18.03 %

Tabla 4.4: **Tabla de dimensión "Promoción"**

vId	prodId	día	sucId	cantVentas
v1	23	3/1/2014	26	50
v2	15	3/1/2014	26	20
v3	40	3/1/2014	28	35
...	...	...	...	...
v50	56	5/4/2015	28	35

Tabla 4.5: **Tabla de hechos "Ventas"**

pvId	promoId	prodId	día	sucId	promVend
pv1	p1	15	3/1/2014	28	35
pv2	p2	56	29/3/2015	31	5
pv3	p3	40	5/8/2015	29	20
...	...	...	...	...	...
pv40	p56	23	10/10/2015	26	10

Tabla 4.6: **Tabla de hechos "Rebajas"**

4.3.1. Calidad en el *Data Warehouse*

En esta sección se muestra cómo los diferentes elementos que dan contexto al componente DW, determinan la calidad en el mismo. A continuación se presentan dos ejemplos.

Ejemplo 1. Contexto: Datos del DW, documentos

En la Tabla 4.7 se muestra, para la dimensión de calidad **exactitud** y su factor **correctitud semántica**, la evaluación de calidad de la **tabla de dimensión "Sucursal"**, respecto al **atributo "ciudad"**. Dicho atributo representa la ciudad a la cual pertenece cada sucursal de la cadena de supermercados. Para este caso, se utiliza un documento de la organización que actua como referencial, el cual contiene para cada ciudad la lista de sucursales que pertenecen a la misma. Parte del documento, para el supermercado "Tata", se presenta en la Figura 4.8.

En la evaluación de la calidad de la tabla de dimensión "Sucursal", respecto al atributo "ciudad", se verifica que la ciudad que aparece en dicha tabla, es la ciudad a la cual efectivamente pertenece la sucursal que está siendo considerada. Por ejemplo, observando la Tabla 4.3 se distingue lo siguiente:

- sucursal **"tata_Rocha_31"**: esta sucursal pertenece a la ciudad "Rocha", de acuerdo con el documento referencial de la organización. Por otro lado, el campo "ciudad" para esta sucursal dice "Rocha", por lo tanto, esta tupla es correcta.

Figura 4.8: **Estructura del documento que contiene las sucursales de cada ciudad.**

- sucursal **"SupermercadoTata"**: esta sucursal pertenece a la ciudad "Salto", de acuerdo con el documento de la organización. Sin embargo, el campo "ciudad" dice "Maldonado", por lo tanto, esta tupla es incorrecta.

Una vez que todas las tuplas de la tabla de dimensión "Sucursal" han sido evaluadas, los valores obtenidos a partir de la métrica **dwq_Ejemplo1**, presentada en la Tabla 4.7, son almacenados como metadatos del DW que posteriormente pueden ser utilizados. Por ejemplo, una de sus utilidades podría ser la corrección de la tabla de dimensión "Sucursal", para contar así, con la ciudad correcta de cada sucursal.

Ejemplo 2. Contexto: Dimensiones del DW

En la Tabla 4.8 se muestra, para la dimensión de calidad **consistencia** y su factor **integridad intra-relación**, la evaluación de calidad de la **tabla de hechos "Rebajas"**, respecto a **la ciudad de la sucursal** en la cual se llevó a cabo la promoción y a **la ciudad de la promoción**, en la cual efectivamente se hizo dicha promoción. Por lo tanto, interesa verificar si dos atributos (ciudad de la sucursal y ciudad de la promoción) de una misma tabla ("Rebajas") satifacen una regla de integridad (la ciudad de la sucursal en la cual se realiza la promoción es efectivamente la ciudad para la cual se creó la promoción). En este caso, el contexto está determinado por los datos de algunas de las dimensiones del DW y las mismas son: "Sucursal" y "Promoción". En la Figura 4.9 se presentan los niveles, de cada dimensión que participa en la definición del contexto. De la tabla de dimensión "Sucursal" se considera el nivel "ciudad" (que contiene las ciudades en las cuales se encuentran las sucursales) y de la tabla de dimensión "Promoción" se considera el nivel "ciudad" (el cual contiene todas las ciudades en las

Componente	DW
Dimensión	Exactitud
Factor	Correctitud semántica: Refiere a la correctitud de los datos respecto al mundo real.
Métrica	Nombre: **dwq_Ejemplo1**
	Contexto considerado: Documento de la organización en el cual se listan las sucursales que pertenecen a cada ciudad.
	Objeto al que da contexto: Tabla de dimensión "Sucursal"
	Granularidad: A nivel de tupla
	Descripción: Revisa cada tupla de la tabla de dimensión "Sucursal", 4.3. Para cada sucursal s_i de la tabla, considera su ciudad en la tupla: c_x. Luego, se fija en el documento, para s_i la ciudad a la cual efectivamente pertenece dicha sucursal: c_y. Finalmente, controla que se cumpla $c_x = c_y$. Si se verifica la igualdad, la tupla es correcta. De lo contrario, es una tupla incorrecta.

Tabla 4.7: **Calidad en el DW. Contexto: Documento de la organización**

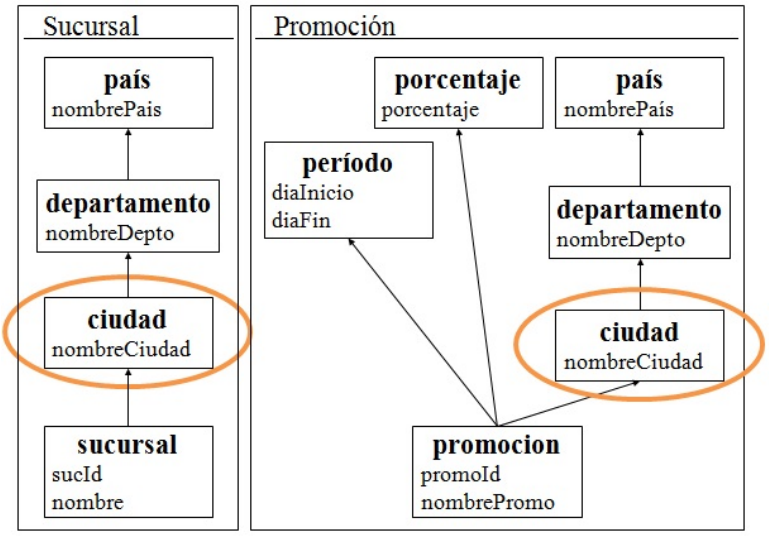

Figura 4.9: **Dimensiones y niveles que dan contexto a la tabla de hechos "Rebajas"**

cuales se han hecho promociones).

Componente	DW
Dimensión	Consistencia
Factor	Integridad intra-relación: Refiere a qué tan bien se satifacen las reglas de integridad.
Métrica	Nombre: **dwq_Ejemplo2**
	Contexto considerado: Dimensiones del DW (Promoción.ciudad, Sucursal.ciudad)
	Objeto al que da contexto: Tabla de hechos "Rebajas"
	Granularidad: A nivel de tupla
	Descripción: Revisa cada tupla de la tabla de hechos "Rebajas", 4.6. Para cada tupla considera el identificador de la promoción p_i de la tabla y luego, en la tabla dimensión "Promocion" obtiene la ciudad c_i para dicha promoción. Luego, para la misma tupla de p_i, considera el identificador de la sucursal s_j y, en la tabla dimensión "Sucursal", obtiene la ciudad c_j para dicha sucursal. Finalmente, controla que se cumpla $c_i = c_j$. Si se verifica la igualdad, la tupla es correcta. De lo contrario, es una tupla incorrecta.

Tabla 4.8: **Calidad en el DW. Contexto: Dimensiones del DW**

Dado que las promociones se realizan por ciudades, en la evaluación de la calidad de la tabla de hechos "Rebajas", se verifica que la ciudad de la sucursal, en la cual se hizo la promoción, sea la misma ciudad para la cual se creó dicha promoción. Es decir, se controla que las sucursales que hacen promociones, efectivamente pertenecen a la ciudad para la cual se creó dicha promoción. Por ejemplo, observando la Tabla 4.6 se distingue lo siguiente:

- tupla **pv1**: esta tupla se corresponde con la promoción p1 y la sucursal 28. La promoción p1 se creó para la ciudad "Florida" (como se observa en la Tabla 4.4) y la sucursal 28 pertenece a la ciudad "Florida" (como se observa en la Tabla 4.3). La promoción se realizó en una sucursal de la ciudad para la cual fue definida la promocion p1. Por lo tanto, esta tupla es consistente.

- tupla **pv2**: esta tupla se corresponde con la promoción p2 y la sucursal 31. La promoción p2 se creó para la ciudad "Colonia" (como se observa en la Tabla 4.4) y la sucursal 31 pertenece a la ciudad "Rocha" (como se observa en la Tabla 4.3). La promoción se realizó en una sucursal de una ciudad que no se corresponde con la ciudad para la cual fue definida la promoción p2. Por lo tanto, esta tupla es incorrecta.

Una vez más, luego que todas las tuplas de la tabla de hechos "Rebajas" han sido evaluadas, los valores obtenidos a partir de la métrica **dwq_Ejemplo2**, presentada en la Tabla 4.8, son almacenados en el repositorio de Metadatos, para posteriormente poder ser utilizados.

4.3.2. Calidad en el *Data Mart*

En esta sección se muestra cómo las reglas de dominio de análisis, que dan contexto a cada componente DM, determinan la calidad en el mismo. A continuación se presentan dos ejemplos.

Ejemplo 3. Dominio: Ventas

En la Tabla 4.9 se muestra, para la dimensión de calidad **exactitud** y su factor **correctitud sintáctica**, la evaluación de calidad de la **tabla de dimensión "Sucursal"**, respecto al **atributo "nombre"**. Dicho atributo, representa el nombre de cada sucursal. En este caso, se considera el dominio correspondiente al área de ventas, en el cual existe una regla de dominio denominada R_{Ventas} que expresa lo siguiente: "El nombre de cada sucursal debe contener el nombre del supermercado, el nombre de la ciudad a la cual pertenece y el número de su identificador, en ese orden". Dicha regla es necesaria en este dominio porque los datos para este *Data Mart*, deben ser integrados con otros datos que verifican esta regla.

Por ejemplo, observando la Tabla 4.3, correspondiente a la Tabla dimensión "Sucursal" se distingue lo siguiente:

- sucursal con **id 29**: esta sucursal tiene el nombre "Florida_29", la sucursal se encuentra en la ciudad "Florida" (como se observa en el campo "ciudad") y el identificador de la sucursal es el número 29, pero el nombre de esta sucursal no contiene el nombre del supermercado. Por lo tanto, esta tupla viola la regla R_{Ventas}, por lo que es una tupla incorrecta para el dominio "Ventas".

- sucursal con **id 31**: esta sucursal tiene el nombre "tata_Rocha_31", el supermercado se llama "Tata", la sucursal se encuentra en la ciudad "Rocha" (como se observa en el campo "ciudad") y el identificador de la sucursal es el número 31. Esta tupla contiene los datos exigidos por la regla R_{Ventas}. Por lo tanto, es una tupla correcta para el dominio "Ventas".

Después que todas las tuplas de la tabla de hechos "Rebajas" han sido evaluadas, los valores obtenidos a partir de la métrica **dmq_Ejemplo3**, presentada en la Tabla 4.9, son almacenados en el repositorio de Metadatos, para posteriormente poder ser utilizados.

Componente	DM
Dominio	Ventas
Regla de dominio	R_{Ventas}: "El nombre de cada sucursal debe contener el nombre del supermercado, el nombre de la ciudad a la cual pertenece y el número de su identificador, en ese orden"
Dimensión	Exactitud
Factor	Correctitud sintáctica: Refiere a la correctitud sintáctica de los datos.
Métrica	Nombre: **dmq_Ejemplo3**
	Contexto considerado: Regla de dominio R_{Ventas}
	Objeto al que da contexto: Tabla de dimensión "Sucursal"
	Granularidad: A nivel de tupla
	Descripción: Revisa cada tupla de la tabla de dimensión "Sucursal", 4.3. Para cada tupla, verifica que su nombre tenga la forma p_1-p_2-p_3, donde p_1 es la parte del nombre de la sucursal que contiene el nombre del supermercado, p_2 es la parte que contiene la ciudad a la cual pertenece la sucursal y p_3 es la parte que contiene su identificador. Si el nombre de la sucursal tiene esta estructura, entonces la tupla correspondiente es correcta para el dominio "Ventas". De lo contrario, es una tupla incorrecta.

Tabla 4.9: **Calidad en el DM. Dominio: Ventas**

Ejemplo 4. Dominio: Publicidad

En la Tabla 4.10 se muestra, para la dimensión de calidad **exactitud** y su factor **correctitud sintáctica**, la evaluación de calidad de la **tabla de dimensión "Sucursal"**, respecto al **atributo "nombre"**. Dicho atributo, representa el nombre de cada sucursal. En este caso, se considera el dominio correspondiente al área de publicidad, en el cual existe una regla de dominio denominada $R_{Publicidad}$ que expresa lo siguiente: "El nombre de cada sucursal debe contener el nombre de la ciudad a la cual pertenece". Pensando en el marketing de la cadena de supermercados, este dominio organiza promociones periódicamente, éstas se realizan por ciudad, inclusive promociones que se hacen en una o algunas ciudades, no se hacen en otras. Por tanto, por una cuestión de organización, en este dominio exigen que se cumpla la regla $R_{Publicidad}$.

Componente	DM
Dominio	Publicidad
Regla de dominio	$R_{Publicidad}$: "El nombre de cada sucursal debe contener el nombre de la ciudad a la cual pertenece"
Dimensión	Exactitud
Factor	Correctitud sintáctica: Refiere a la correctitud sintáctica de los datos.
Métrica	Nombre: **dmq_Ejemplo4**
	Contexto considerado: Regla de dominio $R_{Publicidad}$
	Objeto al que da contexto: Tabla de dimensión "Sucursal"
	Granularidad: A nivel de tupla
	Descripción: Revisa cada tupla de la tabla de dimensión "Sucursal", 4.3. Para cada tupla, verifica que su nombre tenga la forma $p_1p_2p_3$, donde p_2 es la parte que contiene la ciudad a la cual pertenece la sucursal, independientemente de los valores que contengan las partes p_1 y p_3. Si el nombre de la sucursal tiene esta estructura, entonces la tupla correspondiente es correcta para el dominio "Publicidad". De lo contrario, es una tupla incorrecta.

Tabla 4.10: **Calidad en el DM. Dominio: Publicidad**

Por ejemplo, observando la Tabla 4.3, correspondiente a la Tabla dimensión "Sucursal" se distingue lo siguiente:

- sucursal con **id 25**: esta sucursal tiene el nombre "SupermercadoTata", por lo que no contiene la ciudad a la cual pertenece dicha sucursal, "Maldonado" (como se observa en el campo "ciudad"). Esta tupla viola la regla $R_{Publicidad}$. Por lo tanto, esta es una tupla incorrecta para el dominio "Publicidad".

- sucursal con **id 29**: esta sucursal tiene el nombre "Florida_29" y la sucursal se encuentra en la ciudad, "Florida" (como se observa en el campo "ciudad"). Por lo tanto, el nombre de la sucursal contiene el dato exigido por la regla $R_{Publicidad}$, por lo que esta es una tupla correcta para el dominio "Publicidad".

En este caso, es importante destacar que tuplas de la tabla dimensión "Sucursal", que eran incorrectas para el dominio "Ventas" (por ejemplo, tupla con id 29), son correctas para el dominio "Publicidad". Esto es así porque diferentes reglas de dominio fueron aplicadas para evaluar la calidad en dos dominios distintos. Por lo tanto, se observa que diferentes contextos, en este caso determinados por las reglas de dominio, determinan diferentes valores de calidad.

Por otro lado, una vez que todas las tuplas de la tabla dimensión "Sucursal" han sido evaluadas, los valores obtenidos a partir de la métrica **dmq_Ejemplo4**, presentada en la Tabla 4.10, son almacenados en el repositorio de Metadatos, para posteriormente poder ser utilizados.

4.3.3. Calidad en uso

En esta sección se muestra cómo los diferentes elementos que dan contexto al cliente responsable de la presentación de los datos, determinan la calidad en uso en dicho cliente. A continuación se presentan dos ejemplos.

Ejemplo 5. Usuarios del área de ventas

En la Tabla 4.11 se presenta la métrica **qiu_Ejemplo5**, para la dimensión de calidad **frescura** y su factor **actualidad**, que refiere a cuán actualizado está el dato con respecto a su fuente. La evaluación de calidad se realiza sobre la **tabla de hechos "Ventas"**, respecto a la **medida "cantVentas"**. Dicha medida, representa la cantidad de ventas realizadas para cada producto, en cada sucursal en un determinado día. En este caso, se considera que la tabla de hechos "Ventas" cuenta con un campo "timestamp" que indica la última actualización del valor de la medida "cantVentas". La evaluación de la calidad se realiza en dos contextos diferentes: respecto a un usuario gerente general de ventas de la cadena del supermercado y respecto a un usuario director de ventas de sucursal. Por lo tanto, para la evaluación de la calidad, es necesario considerar el perfil de cada usuario de acuerdo a los requerimientos de los mismos:

- **usuario u_1**: dado que el usuario gerente general de ventas, de la cadena del supermercado, necesita realizar diferentes estadísticas, exige que la última actualización de los datos haya sido realizada, a lo sumo, el día anterior a utilizar los datos. O sea, que los datos deben tener 24 hs. de edad como máximo.

Componente	Cliente para presentación de datos
Usuario u_1	Rol: Gerente general de ventas
	Requerimiento de calidad: Los datos deben tener 24 hs. de edad como máximo.
Usuario u_2	Rol: Director de ventas de sucursal
	Requerimiento de calidad: Los datos deben tener 1 h. de edad como máximo.
Dimensión	Frescura
Factor	Actualidad: Refiere a cuán actualizado está el dato con respecto a su fuente.
Métrica	Nombre: **qiu_Ejemplo5**
	Contexto considerado: Perfil de usuario
	Objeto al que da contexto: Tabla de hechos "Ventas"
	Granularidad: A nivel de tabla
	Descripción: Obtiene la última actualización de la tabla de hechos "Ventas", 4.5. Sea t_s el valor correspondiente al campo "timestamp" de la última actualización y t_a el valor correspondiente al momento actual.
	Resultado para u_1: Si t_a - t_s <= 24 hs., entonces el resultado es 1. De lo contrario es 0.
	Resultado para u_2: Si t_a - t_s <= 1 hs., entonces el resultado es 1. De lo contrario es 0.

Tabla 4.11: **Calidad en uso, de acuero al perfil de usuario**

- **usuario u_2**: dado que el usuario director de ventas de sucursal, necesita realizar distintos análisis comparativos en tiempo real, exige que la última actualización de los datos haya sido realizada, a lo sumo, una hora antes de utilizar los datos. O sea, que los datos deben tener 1 h. de edad como máximo.

Como se observa, el requerimiento de calidad del usuario u_2 es más exigente que el del usuario u_1. Por lo tanto, datos que son válidos para el usuario u_2 serán válidos para el usuario u_1. Sin embargo, datos válidos para el usuario u_1, no serán válidos para el usuario u_2. En este caso, si bien es un requerimiento de calidad el que tiene influencia sobre el resultado de la métrica **qiu_Ejemplo5**, es el perfil del usuario el que determina cuál de los dos requerimientos será tenido en cuenta. Por ende, el perfil de usuario es el que determina el resultado obtenido a partir de la métrica **qiu_Ejemplo5**, por tanto, es el usuario el que determina el contexto en el cual se realizará la evaluación de la calidad.

Ejemplo 6. Usuario: Gerente de publicidad

Componente	Cliente para presentación de datos
Usuario	Rol: Gerente de publicidad
	Requerimiento de calidad: El 100 % de las sucursales debe tener el nombre escrito correctamente.
Dimensión	Exactitud
Factor	Correctitud sintáctica: Refiere a la correctitud sintáctica de los datos.
Métrica	Nombre: **qiu_Ejemplo6**
	Contexto considerado: Requerimiento de calidad
	Objeto al que da contexto: Tabla de dimensión "Sucursal"
	Granularidad: A nivel de tabla
	Descripción: Esta métrica consulta los resultados obtenidos en la métrica **dmq_Ejemplo4**.
Resultado	Devuelve 1 si todas las sucursales tienen el nombre correcto. De lo contrario, devuelve 0.

Tabla 4.12: **Calidad en uso. Usuario: Gerente de publicidad**

En la Tabla 4.12 se presenta, para la dimensión de calidad **exactitud** y su factor **correctitud sintáctica**, la evaluación de calidad de la **tabla de dimensión "Sucursal"**, en el **contexto de un usuario** gerente de publicidad. En este caso, para evaluar la calidad de los datos, se consultan los valores de calidad obtenidos anteriormente, en el componente DM, mediante la ejecución de la métrica **dmq_Ejemplo4** de la Tabla 4.10. El usuario tiene un requerimiento de calidad

con el cual exige que el 100 % de las sucursales tengan el nombre escrito correctamente. Dado que el usuario es un gerente de publicidad, el mismo pertenece al dominio "Publicidad". Por lo tanto, que las sucursales tengan el nombre correcto, en este dominio, implica que dicho nombre debe contener el nombre de la ciudad al cual pertenece la misma. Esto así, de acuerdo a la regla de dominio que se presenta en la Tabla 4.10. Por todo esto, la métrica **qiu_Ejemplo6** consulta la métrica **qiu_Ejemplo4** para verificar dicho requerimiento de calidad.

4.3.4. Resumen del caso de estudio

En esta sección se presenta, a través de la Tabla 4.13, un resumen de todos los ejemplos antes presentados. En la tabla se observa el tipo de calidad, el número de ejemplo, el objeto al cual se da contexto, el contexto considerado y la dimensión de calidad, con su respectivo factor, que es evaluada.

Calidad	**Ejemplo**	**Objeto**	**Contexto**	**Dimensión, Factor**
en el DW	1	Tabla dimensión "Sucursal"	Documento de la organización	Exactitud, Correctitud semántica
	2	Tabla de hechos "Rebajas"	Valores de dimensiones del DW	Consistencia, Integridad intrarelación
en el DM	3	Tabla dimensión "Sucursal"	Regla de dominio R_{Ventas} sobre los nombres de las sucursales	Exactitud, Correctitud sintáctica
	4	Tabla de hechos "Sucursal"	Regla de dominio $R_{Publicidad}$ sobre los nombres de las sucursales	Exactitud, Correctitud sintáctica
en USO	5	Tabla de hechos "Venta"	Perfil de usuario	Frescura, Actualidad
	6	Tabla dimensión "Sucursal"	Requerimiento de calidad	Exactitud, Correctitud sintáctica

Tabla 4.13: **Resumen del caso de estudio**

4.4. Prueba de concepto: Implementación en Datalog

En esta sección se muestran los resultados obtenidos en la prueba de concepto. Si bien el caso de estudio es basado en el modelo relacional, que es un modelo ampliamente conocido y esto ayuda a la comprensión del mismo, se toma la decisión de representar el *Data Warehouse* y evaluar la calidad de este utilizando contextos, con un enfoque basado en reglas, el cual es propuesto en [133]. Por lo tanto, para el desarrollo de la prueba de concepto se utiliza el modelo y se adapta el ejemplo presentados en el artículo [133]. A continuación se muestra con un ejemplo[3] cómo es utilizado dicho modelo.

El predicado *aggr* se utiliza para representar las instancias de las tablas de dimensión. Las siguientes reglas representan una instancia de la dimensión *store* (sucursal), con su jerarquía y sus respectivos niveles:

storeName(31, "tata_Rocha_31").
storeStructName1(31, "tata").
storeStructName2(31, "Rocha").
storeStructName3(31, "31").
aggr(X, Store, store_city, "Florida") $\leftarrow$ *aggr*(X, Store, store_id, 31).
aggr(X, Store, store_state, "Florida") $\leftarrow$ *aggr*(X, Store, store_city, "Florida").
aggr(X, Store, country, "Uruguay") $\leftarrow$ *aggr*(X, Store, store_state, "Florida").

El predicado "*storeName*" asocia a cada identificador de sucursal el nombre correspondiente. Mientras que los predicados "*StructName1*", "*StructName2*" y "*StructName3*" contienen las partes que conforman la estructura del nombre de cada sucursal. Es decir, asocian a cada identificador de sucursal la primera, la segunda y la tercera parte de la estructura del nombre de esa sucursal, respectivamente.

Por otro lado, la siguiente regla representa la tabla de hechos "Ventas", donde "*Sales*" es el nombre correspondiente a la tabla de hechos, "s1" es el identificador del hecho y "50" es el valor de la medida (cantidad vendida) del hecho.

$AFact_{\text{Qty}}$(Sales, s_1, 50).

Dado que la dimensión "*Store*" participa del hecho "*Sales*", se tiene el siguiente paso base:

aggr(s1, Store, store_id, 31).

[3]Todas las reglas fueron definidas en inglés, porque los datos obtenidos para realizar la prueba de concepto están todos en inglés

Además, se define el predicado "*infoFact*", el cual contiene información administrativa de las tablas de hechos: nombre de la tabla, la fecha de creación y un timestamp con la fecha de la última actualización de dicha tabla. Estos dos últimos valores se presentan expresados en segundos. Esto es así porque, en *Datalog*, todos los valores que refieren a tiempos son expresados en segundos. A continuación se presenta un ejemplo para la tabla de hechos "*Sales*".

infoFact(Sales, 1448551353.401108, 1448648722.769284).

A continuación se muestran las ejecuciones de las reglas definidas en *Datalog*.

Ejemplo 1. Contexto: Datos del DW, documentos

La métrica **dwq_Example1** utiliza la regla *documentCity*, la cual contiene información acerca de la ciudad a la cual pertenece efectivamente cada sucursal. Por ejemplo, la sucursal con identificador 30 pertenece a la ciudad "Florida".

documentCity(30, 'Florida').

La regla *documentCity* determina el contexto para la métrica **dwq_Example1**:

```
contextEx1(X, Y)    :- documentCity(X, Y).

dwq_Example1(X, S, N, C, Z)   :- contextEx1(S, Z),
                                 aggr(X, store, store_city, C),
                                 aggr(X, store, store_id, S),
                                 storeName(S, N),
                                 not(C = Z).
```

En este ejemplo se define el contexto **contextEx1** y la métrica **dwq_Example1** que usa dicho contexto, el nombre de las sucursales, su identificador y su ciudad, para la evaluación de la calidad. La métrica devuelve todas las sucursales cuya ciudad es incorrecta, mostrando el valor de ciudad que tiene y el que debería tener. Los parámetros devueltos son los siguientes:

- X: es el identificador del hecho
- S: identificador de la sucursal
- N: nombre de la sucursal
- C: ciudad de la sucursal en el nivel "ciudad", de la dimensión "Sucursal"
- Z: ciudad correcta de la sucursal de acuerdo al documento de la sucursal, o sea, el contexto

En la Figura 4.10 se presentan los resultados luego de la ejecución de la métrica **dwq_Example1**.

```
29 ?- dwq_Example1(X,S,N,C,Z).
X = s1,
S = 2,
N = 'Store_2',
C = 'Bellingham',
Z = 'Bremerton' ;
X = s903,
S = 19,
N = 'Store_19',
C = 'Vancouver',
Z = 'Victoria' ;
X = s1001,
S = 25,
N = 'SupermercadoTata',
C = 'Maldonado',
 Z = 'Salto' ;
false.
```

Figura 4.10: **Resultados obtenidos para la métrica dwq_Example1**

Ejemplo 2. Contexto: Dimensiones del DW

Para la definición de la métrica **dwq_Example2** es necesario el siguiente predicado: *PromotionName*. Dicho predicado se muestra a continuación:

```
...
promotionName(1476, 'Dia del Nino').
promotionName(1477, 'Fin de Temporada').
promotionName(1479, 'Dia de la Madre').
promotionName(1480, 'Dia del Padre').
...

contextEx2(X, S, Z, P, C)   :- aggr(X, store, store_id, S),
                               aggr(X, store, store_city, Z),
                               aggr(X, promotion, promotion_id, P),
                               aggr(X, promotion, promotion_city, C).

dwq_Example2(X, M, Z, N, C)  :- contextEx2(X, S, Z, P, C),
                                storeName(S, M),
                                promotionName(P, N),
                                not(Z = C).
```

La regla **contextEx2** es el contexto para la evaluación de la calidad a partir de la métrica **dwq_Example2**. Dicho contexto devuelve los siguientes parámetros:

- S: identificador de la sucursal
- Z: ciudad de la sucursal
- P: identificador de la promoción
- C: ciudad en la cual se realizó la promoción

Por otro lado, la métrica **dwq_Example2** considera el contexto **contextEx2**, el nombre de las sucursales, el nombre de las promociones y devuelve las tuplas en las cuales la ciudad de la promoción no coincide con la ciudad de la sucursal en la cual se realizó dicha promoción. La métrica devuelve los siguientes parámetros:

- X: es el identificador del hecho
- M: nombre de la sucursal
- Z: ciudad de la sucursal
- N: nombre de la promoción
- C: ciudad en la cual se realizó la promoción

En la Figura 4.11 se presentan los resultados luego de la ejecución de la métrica **dwq_Example2**.

Ejemplo 3. Dominio: Ventas

En este caso, se cuenta con información del dominio, que indica la estructura que debe tener el nombre de cada sucursal. De esta forma, dicha información permite verificar que se cumpla la regla de dominio R_{Ventas}. A su vez, esta información da contexto a la evaluación de calidad a partir de la métrica **dmq_Example3**. Dicho contexto es **contextEx3** y devuelve los siguientes parámetros:

- N: nombre de la sucursal
- A: parte 1 del nombre de la tienda (nombre del supermercado)
- B: parte 2 del nombre de la tienda (nombre de la ciudad)
- C: parte 3 del nombre de la tienda (id de la tienda)

Por otro lado, la métrica **dmq_Example3** devuelve el nombre de todas las sucursales que cumplen la regla de dominio R_{Ventas}: "El nombre de cada sucursal debe contener el nombre del supermercado, el nombre de la ciudad a la cual pertenece y el número de su identificador, en ese orden".

Para la definición de la métrica **dmq_Example3** son necesarias las siguientes reglas: *documentStore*, *documentCity* y *documentIdent*.

```
6 ?- dwq_Example2(X,M,Z,N,C).
X = s1001,
M = 'SupermercadoTata',
Z = 'Maldonado',
N = 'Price Savers',
C = 'Orizaba' ;
X = s1005,
M = tata_Salto_26,
Z = 'Salto',
N = 'Price Savers',
C = 'Orizaba' ;
X = s1006,
M = tata_Salto_27,
Z = 'Salto',
N = 'Price Savers',
C = 'Orizaba' ;
X = s1004,
M = tata_Florida_28,
Z = 'Florida',
N = 'Price Savers',
C = 'Orizaba' ;
X = s1002,
M = 'Florida_29',
Z = 'Florida',
N = 'Price Savers',
C = 'Orizaba' ;
X = s1003,
M = 'Florida_30',
Z = 'Florida',
N = 'Price Savers',
C = 'Orizaba' ;
X = s1007,
M = tata_Rocha_31,
Z = 'Rocha',
N = 'Price Savers',
C = 'Orizaba' ;
X = s1008,
M = tata_Rocha_32,
Z = 'Rocha',
N = 'Price Savers',
C = 'Orizaba' ;
false.
```

Figura 4.11: **Resultados obtenidos para la métrica dwq_Example2**

- *documentStore*: contiene el nombre del supermercado al cual pertenece efectivamente cada sucursal. Por ejemplo, la sucursal con identificador 30 pertenece al supermercado "tata".
- *documentCity*: contiene el nombre de la ciudad a la cual pertenece efectivamente cada sucursal. Por ejemplo, la sucursal con identificador 30 pertenece a la ciudad "Florida".
- *documentIdent*: contiene la cadena de caracteres que representa al identificador de cada sucursal. Por ejemplo, la sucursal con identificador 30 tiene la cadena "30".

```
documentStore(30, 'tata').
documentCity(30, 'Florida').
documentIdent(30, '30').
```

Por otro lado, se consideran los predicados que indican la estructura que tiene el nombre de cada sucursal. Por ejemplo, de acuerdo con estos predicados, la sucursal con identificador 30, se llama "Florida_30". Dichos predicados se presentan a continuación:

```
storeStructName1(30, '').
storeStructName2(30, 'Florida').
storeStructName3(30, '30').
```

Las reglas *documentStore*, *documentCity* y *documentIdent* determinan el contexto para la métrica **dmq_Example3**:

```
contextEx3(X, A, B, C)   :- documentStore(X, A),
                            documentCity(X, B),
                            documentIdent(X, C).

dmq_Example3(X, N)       :- contextEx3(S, A, B, C),
                            aggr(X, store, store_id, S),
                            storeName(S, N),
                            storeStructName1(S, Y),
                            storeStructName2(S, Z),
                            storeStructName3(S, W),
                            A = Y, B = Z, C = W.
```

En la Figura 4.12 se presentan los resultados luego de la ejecución de la métrica **dmq_Example3** y los parámetros de la misma son:

- X: es el identificador del hecho
- N: nombre de la sucursal

```
6 ?- dmq_Example3(X,N).
X = s1005,
N = tata_Salto_26 ;
X = s1006,
N = tata_Salto_27 ;
X = s1004,
N = tata_Florida_28 ;
X = s1007,
N = tata_Rocha_31 ;
X = s1008,
N = tata_Rocha_32.
```

Figura 4.12: **Resultados obtenidos para la métrica dmq_Example3**

Ejemplo 4. Dominio: Publicidad

Para este ejemplo también se cuenta con información del dominio, que indica la estructura que debe tener el nombre de cada sucursal. De esta forma, dicha información permite verificar que se cumpla la regla de dominio $R_{Publicidad}$: "El nombre de cada sucursal debe contener el nombre de la ciudad a la cual pertenece". A su vez, esta información da contexto a la evaluación de calidad a partir de la métrica **dmq_Example3**. Dicho contexto es **contextEx4**.

La métrica **dmq_Example4** utiliza la regla *documentCity*, la cual contiene el nombre de la ciudad a la cual pertenece efectivamente cada sucursal. Por ejemplo, la sucursal con identificador 30 pertenece a la ciudad "Florida".

documentCity(30, 'Florida').

Por otro lado, se considera el predicado que indica parte de la estructura que tiene el nombre de cada sucursal. Por ejemplo, de acuerdo con este predicado, la sucursal con identificador 30, tiene en su nombre la palabra "Florida":

storeStructName2(30, 'Florida').

La regla *documentCity* determina el contexto para la métrica **dmq_Example4**:

```
contextEx4(S, C)    :- documentCity(S, C).

dmq_Example4(X, N, B)   :- contextEx4(S, B),
                           aggr(X, store, store_id, S),
                           storeName(S, N),
                           storeStructName2(S, Z),
                           B = Z.
```

La métrica **dmq_Example4** devuelve los siguientes parámetros:

- X: es el identificador del hecho
- N: nombre de la sucursal
- B: parte del nombre de la tienda que indica la ciudad

En la Figura 4.13 se presentan los resultados luego de la ejecución de la métrica **dmq_Example4**.

```
9 ?- dmq_Example4(X,N,B).
X = s1005,
N = tata_Salto_26,
B = 'Salto' ;
X = s1006,
N = tata_Salto_27,
B = 'Salto' ;
X = s1004,
N = tata_Florida_28,
B = 'Florida' ;
X = s1002,
N = 'Florida_29',
B = 'Florida' ;
X = s1003,
N = 'Florida_30',
B = 'Florida' ;
X = s1007,
N = tata_Rocha_31,
B = 'Rocha' ;
X = s1008,
N = tata_Rocha_32,
B = 'Rocha'.
```

Figura 4.13: **Resultados obtenidos para la métrica dmq_Example4**

Ejemplo 5. Usuarios del área de ventas

En este ejemplo, es necesario definir la métrica **qiu_Example5** para cada uno de los usuarios u_1 y u_2, dado que dependiendo del perfil de cada uno, se aplica un requerimiento de calidad u otro. La métrica devuelve las medidas M que verifica el requerimiento de calidad correspondiente a dicho usuario.

La métrica **qiu_Example5** utiliza la regla *userCurrent*, la cual contiene el máximo valor exigido por cada usuario respecto a la edad de los datos y el nombre de la tabla de interés para el usuario. Observar, que una vez más, los tiempos están expresados en segundos.

userCurrent(u1, Sales, 86400).
userCurrent(u2, Sales, 3600).

Además, se propone un predicado que devuelve la diferencia de dos números:
difference(Z, W) :- (N - Z) = W.

Donde:

- N: es el valor actual del tiempo. En *Datalog* se utiliza el predicado *get_time(Now)* y el valor devuelto está expresado en segundos.
- Z: es el valor del timestamp grabado en el momento de la última actualización de los datos
- W: es el valor correspondiente a la edad de los datos

Las reglas *userCurrent* y *difference* determinan el contexto para la métrica **qiu_Example5**:

```
contextEx5(U, T, Z, W, F)  :- userCurrent(U, F, T),
                              difference(Z, W).
```

Como se observa, la métrica **qiu_Example5** utiliza el predicado *infoFact* que contiene el valor de la última actualización de cada tabla de hechos.

```
qiu_Example5(U, F, T)  :- contextEx5(U, T, Z, W, F),
                          infoFact(F, Y, Z),
                          (W<T).
```

En la métrica **qiu_Example5** se identifican los siguientes parámetros:

- U: representa al usuario
- F: es el nombre de la tabla de hecho
- T: es el valor máximo exigido para la edad de los datos

En la Figura 4.14 se presentan los resultados obtenidos luego de la ejecución de la métrica **qiu_Example5**.

```
2 ?- qiu_Example5(U,F,T).
U = u1,
F = sales,
T = 86400 ;
U = u2,
F = sales,
T = 3600.
```

Figura 4.14: **Resultados obtenidos para la métrica qiu_Example5**

Ejemplo 6. Usuario: Gerente de publicidad

Para este caso, se propone **dmq_Example4(X,N,B)_incorrect**, que devuelve el nombre de todas las sucursales que no verifican la regla de dominio $R_{Publicidad}$, ya que el 100 % de las sucursales deben tener su nombre con la estructura correcta. Si **dmq_Example4(X,N,B)_incorrect** devuelve algun valor, **qiu_Example6** no se satisface.

```
dmq_Example4(X, N, B)_incorrect   :- contextEx4(S, B),
                                     aggr(X, store, store_id, S),
                                     storeName(S, N),
                                     storeStructName2(S, Z),
                                     not(B = Z).

qiu_Example6(X, N, B)   :- qiu_Example4_incorrect(X, N, B).
```

4.5. Conclusiones

Para la evaluación de la calidad de datos en un SDW, se consideran los componentes del sistema y se define la calidad en cada uno de ellos. Para dicha tarea, esta tesis se apoya en dos enfoques de calidad: *Meeting Requirements* y *Fitness for Use*. El primer enfoque se aplica en el DW y en los *Data Marts* (DM), poniendo énfasis en el cumplimiento de los requerimientos del sistema. El segundo enfoque, se considera en el cliente para la presentación de los datos, buscando satisfacer las necesidades de cada usuario. Si bien la evaluación de la calidad de datos en este tipo de sistema no es una novedad, ya que mucha es la bibliografía que investiga y subraya la necesidad de la gestión de la calidad de datos en los SDW, no hay mucha evidencia acerca del enfoque que presenta esta tesis. Este primer planteo de propuesta define contextos en cada uno de los componentes del SDW para la evaluación de la calidad de datos en cada uno de ellos.

Como se observa en el Capítulo 3, existe evidencia científica respecto a la subjetividad de la calidad de los datos. Aunque dicha subjetividad es clara desde el punto de vista de los usuarios finales, la misma se mantiene a lo largo de todo el ciclo de vida del SDW. Esto es así, porque los datos no son utilizados con un único propósito y en un único dominio de análisis. Por lo tanto, la evaluación de la calidad de los datos en cada componente del SDW, que participa desde el momento en que los datos son cargados en el DW y hasta que los mismos son utilizados por los usuarios finales, se realiza en base al contexto de dichos datos.

Para mostrar cómo son usados los contextos en la evaluación de la calidad, se plantea un caso de estudio en el cual se considera cada componente del SDW: el DW, el DM y el cliente para la presentación de los datos. Una vez seleccionado el componente a estudiar, se elige la dimensión de calidad (y su respectivo factor de calidad), que se quiere evaluar. A partir de este momento es necesario definir el contexto para dicha medición de calidad. Dicho contexto, determina los resultados de calidad obtenidos a partir de la ejecución de la métrica, ya que diferentes contextos determinan distintos valores de calidad para una misma métrica.

Para validar los ejemplos definidos en el caso de estudio, se ejecuta una prueba de concepto implementada en *Datalog*, basada en el modelo propuesto en [133] Los contextos y métricas fueron ejecutadas sobre los datos de un DW implementado por la estudiante de grado Carmela Beiro en el contexto de la asignatura "Módulo de Taller". Si bien se presentó un conjunto pequeño de ejemplos, se considera que los resultados obtenidos son suficientes para demostrar la aplicabilidad del enfoque utilizado para evaluar la calidad de datos en los SDWs.

Finalmente, es importante destacar que algunas de las ideas base de esta propuesta fueron validadas, mediante la presentación de un resumen extendido, en las Jornadas Chilenas de Computación, en la Pontificia Universidad Católica de Valparaíso, Chile [134].

Capítulo 5

Conclusiones y trabajo a futuro

No es una novedad que los Sistemas de *Data Warehousing* (SDW) son de gran relevancia para el apoyo en la toma de decisiones y el análisis de los datos. Esto ha quedado demostrado a lo largo del tiempo, a través de la generalización de su desarrollo y uso a nivel industrial en todo tipo de organizaciones y mediante la gran cantidad de trabajos científicos que se han centrado en el estudio de este tipo de sistemas. Más allá del objetivo de investigación, todos los autores coinciden en el valor agregado que los SDW aportan a las organizaciones. En particular, muchos de estos trabajos resaltan ampliamente la importancia de la Calidad de Datos para dichos sistemas y el peso que ésta tiene en la toma de decisiones. Por esta razón, muchos autores han presentado la necesidad de incorporar y mantener la calidad en los SDW, sin embargo, en las investigaciones no se encuentra un consenso acerca de cómo hacerlo. La mayoría de los trabajos sólo abordan la limpieza de los datos en la etapa de ETL, ignorando la tarea de evaluación de la calidad de los datos a lo largo de todo el ciclo de vida de un DW. Además, según los investigadores, aún no se ha identificado cuál es el conjunto de dimensiones de calidad pertinentes para estos Sistemas de Información. A partir de esto último surge el cuestionamiento de si es imposible definir un único conjunto de dimensiones de calidad en el entorno de un DW, dado que dicho conjunto puede depender del propósito con el cual se utilizan los datos.

Por otro lado, existe un grupo de investigadores enfocados en el área de la Calidad de los Datos en sí misma, que subrayan la naturaleza contextual de la calidad de los datos lo que, según ellos, justifica que no se encuentre un único conjunto de dimensiones de calidad, independientemente del Sistema de Información. A pesar de esto, varios trabajos consideran que la dimensión contextual no suele ser representada en los frameworks de calidad y a su vez, coinciden en la importancia del uso del contexto en la tarea de limpieza y/o evaluación de la calidad de los datos. Sin embargo, no todos comparten la misma noción de contexto y no todos establecen aún bajo qué circunstancias y de qué forma deben ser considerados y/o representados dichos contextos. Otra discrepancia que se presenta entre los investigadores, es la discusión acerca del aspecto contextual de

la dimensión de calidad *accuracy*, esto resalta, una vez más, la dependencia de la calidad de los datos respecto al entorno en el cual se va a realizar la limpieza y/o evaluación de la calidad. Otro aspecto a resaltar es el enfoque de la calidad de datos como un proceso más que como una medida estática. Los investigadores consideran que esto permitiría captar mejor la esencia de los datos, en particular de la información, yendo desde los datos hasta el conocimiento y pasando a través de distintos contextos.

El tema central de esta tesis es la evaluación de la Calidad de Datos en los SDW. Sin embargo, una vez iniciada la primera instancia de la revisión bibliográfica, surgieron los distintos cuestionamientos antes mencionados, lo que permitió observar la necesidad de incorporar el análisis de los Contextos para dicha evaluación de calidad. En particular, surge el siguiente cuestionamiento: **¿Cómo pueden ser usados los contextos para evaluar la calidad de datos en *Data Warehouse*?** Dado que no estaba claro la definición ni el uso de los contextos en los SDW y/o en la evaluación de DQ, surge la necesidad de analizar el estado del arte actual respecto a las tres áreas de interés: *Data Quality* (DQ), *Data Warehouse* (DW) y *Context* (CTX). Para esto, se plantea el uso de una metodología de búsqueda, que permita aplicar una serie de pasos bien definidos y reproducibles. Por lo tanto, se realiza el *Mapping Study*, MS.

Se ejecutó un MS para obtener una visión general de la investigación existente acerca del uso de contextos en los SDW y/o en la evaluación de DQ. Se identificaron 62 trabajos publicados entre el año 2008 y el año 2015. Los trabajos seleccionados permiten destacar las necesidades, tendencias y desafíos que presentan en conjunto estas tres áreas de investigación. Si bien el número de bibliotecas digitales es pequeño (tres), por una cuestión de alcance y de acuerdo con las pautas de la metodología, se considera que los resultados presentados en el Capítulo 3 son lo suficientemente relevantes para esta tesis. El número de trabajos seleccionado tampoco es muy grande, sólo un 10 % del total de los artículos devueltos en las búsquedas, se destaca el amplio espectro de trabajos que abarcan las cadenas de búsqueda. Por ejemplo, si solo se tuvieran en cuenta las palabras claves: *Data Quality*, *Data Warehouse* y *Context*, cada una de ellas determina un área de investigación, lo que implica un gran rango de trabajos. Por esta razón, muchos trabajos fueron descartados durante las búsquedas y a partir de la aplicación de los criterios de exclusión. Por otro lado, cabe resaltar la naturaleza del significado de la palabra "Contexto" y el uso que se da a la misma. Por esta razón, la palabra "Contexto" es ampliamente utilizada para hacer referencia a distintos dominios por lo que, en muchas ocasiones, se encontraron resultados que si bien consideraban las áreas de investigación de DW o DQ, nada tenían que ver con el área de investigación: Contextos. Esto último también contribuye a que el número de trabajos encontrados sea tan alto respecto al número de trabajos seleccionados.

La metodología MS fue aplicada en dos etapas, la primera considera trabajos publicados desde el año 2008 hasta el año 2014, mientras que la segunda etapa considera trabajos desde el año 2014 hasta el año 2015. La primer etapa se realizó en el período marzo-junio del año 2014, mientras que la segunda etapa se llevó a cabo al finalizar la tesis, en octubre de 2015. Las razones que motivaron la decisión de seguir este procedimiento fueron dos. En primer lugar, era necesario conocer el estado del arte al momento de cerrar la tesis, principalmente porque la última revisión bibliográfica se había realizado hacía casi un año y medio, y en este período podían haber surgido trabajos interesantes y de importancia para esta tesis. Por otro lado, la nueva ejecución de las cadenas de búsqueda antes definidas, adaptándolas al período de interés (2014-2015), permitirá demostrar si la metodología aplicada es reproducible, como se menciona en la bibliografía correspondiente. En la segunda etapa se obtuvieron 50 trabajos más, de los cuales 6 fueron incluídos en el grupo de trabajos seleccionados, mediante la aplicación de los criterios de exclusión e inclusión definidos en el Capítulo 3. Efectivamente, con la experiencia se pudo demostrar la posibilidad, que presenta la metodología, de reproducir todas las búsquedas una vez que éstas son definidas.

Por otro lado, la desventaja que presenta este tipo de metodologías respecto a las revisiones bibliográficas tradicionales, es que requieren de un esfuerzo inicial mayor. Esto se debe al tiempo requerido para la elaboración de cada una de las etapas que debe cumplir la metodología. Sin embargo, este tiempo es compensado a la hora de realizar nuevas búsquedas, ya que la posibilidad de reproducirlas permite contar con una metodología de trabajo incremental, con la cual es posible mantener actualizado el estado del arte de una forma sistemática. A su vez, si bien la aplicación de las metodologías de búsqueda son originarias del área de Medicina, a lo largo del tiempo se observa un crecimiento notorio de su uso en el área de *Computer Science*, en particular en investigaciones de Calidad de Datos, lo que demuestra la utilidad de su aplicación. Actualmente, se está trabajando en la escritura de un artículo que incluye los resultados obtenidos a partir de la aplicación del MS, para su posterior publicación en una revista internacional especializada. Este artículo está siendo realizando en conjunto con mi supervisora Dra. Adriana Marotta y el Dr. Ismael Caballero profesor de la Universidad de Castilla-La Mancha, España.

Los resultados obtenidos a partir del MS, permitieron entender cuáles son los desafíos que se presentan en la investigación de los temas DQ, DW y CTX. Una vez que se logró una visión general del estado del arte, fue posible realizar el primer planteo de una propuesta para evaluar la Calidad de Datos en los SDW, con un enfoque basado en Contextos. Este primer planteo, es el punto de partida de una investigación más amplia y profunda que permita la gestión de la calidad en los SDW.

Para la evaluación de la calidad de datos en un SDW, se consideran los componentes del sistema y se define la calidad en cada uno de ellos. Para dicha tarea, esta tesis se apoya en dos enfoques de calidad: *Meeting Requirements* y *Fitness for Use*. El primer enfoque se aplica en el DW y en los *Data Marts* (DM), poniendo énfasis en el cumplimiento de los requerimientos del sistema. El segundo enfoque, se considera en el cliente para la presentación de los datos, buscando satisfacer las necesidades de cada usuario. Si bien la evaluación de la calidad de datos en este tipo de sistema no es una novedad, ya que mucha es la bibliografía que investiga y subraya la necesidad de la gestión de la calidad de datos en los SDW, no hay mucha evidencia acerca del enfoque que presenta esta tesis. Esta primera propuesta define contextos en cada uno de los componentes del SDW para la evaluación de la calidad de datos en cada uno de ellos. Este enfoque se apoya en los resultados obtenidos a partir del análisis del estado del arte, entre los que se destaca la naturaleza contextual de la calidad de los datos.

Para mostrar cómo son usados los contextos en la evaluación de la calidad, se planteó un caso de estudio en el cual se considera cada componente del SDW: el DW, el DM y el cliente para la presentación de los datos. Una vez seleccionado el componente a estudiar, se elige la dimensión de calidad (y su respectivo factor de calidad), que se quiere evaluar. A partir de este momento es necesario definir el contexto para dicha medición de calidad. Dicho contexto, determina los resultados de calidad obtenidos a partir de la ejecución de la métrica, ya que diferentes contextos determinan distintos valores de calidad para una misma métrica. Esto último se observó mediante el desarrollo de los ejemplos presentados en el Capítulo 4. Algunas de las ideas base de esta propuesta fueron validadas, mediante la presentación de un resumen extendido, en las Jornadas Chilenas de Computación, en la Pontificia Universidad Católica de Valparaíso, Chile [134].

Finalmente, se realizó una prueba de concepto, implementada en *Datalog* y basada en el modelo propuesto en [133], para ejecutar los ejemplos definidos en el caso de estudio. Si bien se presentó un conjunto pequeño de ejemplos, se considera que los resultados obtenidos son suficientes para demostrar la aplicabilidad del enfoque utilizado para evaluar la calidad de datos en los SDW.

5.1. Aportes

Los aportes de la investigación realizada en esta tesis de Maestría son los que se presentan a continuación:

- Se presentó un Estado del Arte que relaciona las áreas Calidad de Datos, Sistemas de *Data Warehousing* y Contextos, utililzando una metodología de búsqueda bibliográfica rigurosa y reproducible llamada *Mapping Study*. Esta revisión bibliográfica permitió entender cuál es el estado actual de la investigación referida a las áreas de interés e identificar cuáles son los desafíos existentes en dichas áreas.

- Las metodologías de búsqueda tienen como característica la posibilidad de reproducir las búsquedas. La metodología *Mapping Study* se ejecutó en dos etapas y la experiencia permitió demostrar su capacidad de reproducción. De esta manera, se obtuvo un estado del arte exhaustivo de una forma incremental.

- Se realizó un primer planteo de una propuesta para la evaluación de la Calidad de Datos en SDW, con un enfoque basado en el Contexto de los datos.

- Se ejecutó una prueba de concepto, con la cual fue posible demostrar la aplicabilidad de la propuesta planteada.

5.2. Limitaciones

Una vez finalizada la tesis, es de gran importancia poder hacer una autocrítica del trabajo realizado, buscando identificar los puntos débiles de la investigación. La limitaciones se presentan siguiendo el objetivo principal de este trabajo: aprender, no sólo reconociendo los resultados positivos, sino también localizando aquellos resultados que presentan ciertas debilidades y que aún pueden ser mejorados.

- Si bien la metodología utilizada, MS, puede ser ejecutada sobre un número pequeño de bibliotecas digitales, se podría obtener un resultado más relevante para la investigación si la misma se ejecutara sobre un número más importante de bibliotecas digitales. En particular, sería de gran interés incluir más bibliotecas de alto interés científico, como lo es por ejemplo *Springer*.

- La propuesta es una primera aproximación a la solución planteada para evaluar la Calidad de Datos en los SDW. Por esta razón, el caso de estudio presenta pocos ejemplos. Por lo tanto, sería importante definir un número más significativo de ejemplos, a través de los cuales se pudiera observar más claramente la influencia del contexto sobre la calidad de los datos.

- La realización de la prueba de concepto fue llevada a cabo en el marco de un taller realizado por una estudiante de grado. Si bien fue de gran valor el trabajo aportado para la validación del caso de estudio, sería mucho más representativo poder realizar una prueba de concepto sobre un caso real.

5.3. Trabajo a futuro

A partir de los objetivos logrados y de las limitaciones que presenta el trabajo final, surgen las distintas líneas de trabajo que podrían ser abordadas para continuar esta investigación. Los nuevos desafíos planteados a partir de esta tesis se presentan a continuación:

- Incluir, a corto plazo, más bibliotecas digitales para una nueva ejecución de la metodología MS, logrando así un alcance aún mayor del Estado del Arte.

- Finalizar la escritura del artículo que incluye los resultados obtenidos a partir de la aplicación de la metodología de búsqueda para su posterior publicación en un *Journal*. Este artículo está siendo realizando en conjunto con mi supervisora Dra. Adriana Marotta y el Dr. Ismael Caballero profesor de la Universidad de Castilla-La Mancha, España.

- Realizar la publicación de los resultados obtenidos a partir de la tesis en sí misma, en una conferencia o revista internacional.

- Aplicar, a partir de la metodología MS, la metodología *Systematic Literature Review* (SLR). Los trabajos que se enfocan en la investigación de este tipo de metodologías en sí mismas, recomiendan, una vez finalizado el MS, la ejecución de una SLR. De esta forma, sería posible formular *Research questions* más específicas logrando así un alcance más enfocado. Además, para la SLR es importante asegurar que los resultados se basan en evidencia de calidad. Por ejemplo, en la aplicación del MS se obtuvo un amplio espectro de trabajos científicos. Algunos de ellos, aunque analizaban las áreas de interés, eran demasiado superficiales. Sin embargo, satisfacían el objetivo de conocer el estado del arte, pero una vez logrado ésto se podrían refinar las búsquedas, mediante la aplicación de una SLR, seleccionando únicamente aquellos trabajos que presenten resultados de calidad para el objetivo de esta tesis.

- Definir un modelo formal del sistema, de los contextos y de las calidades definidas para cada uno de los componentes del SDW. Esto, permitiría la realización de una propuesta más detallada y profunda del enfoque abordado.

- Analizar la relación existente entre la calidad de los datos y la calidad en uso, abordadas a partir de los enfoques *Meeting Requirements* y *Fitness for*

Use, respectivamente. Si existiera una relación entre dichas calidades, sería interesante estudiar la representación gráfica de la misma. Graficando los valores de calidad de datos para cada factor de calidad, de la dimensión de calidad correspondiente, contra el valor de calidad en uso a lo largo del tiempo. Posteriormente, y dependiendo de la forma de la gráfica, sería posible mejorar la calidad de los datos para obtener una mejor calidad en uso, logrando así la satisfacción del usuario final.

- Plantear el desarrollo de un marco para aplicar las definiciones de Calidad Interna (CI) y Calidad Externa (CE), propuestas en la norma ISO25010, en un SDW. La CI se enfoca en el proceso de fabricación, mientras que la CE se centra en la percepción que se tiene de la calidad, en particular, en el cumplimiento de los requerimientos del sistema. En este marco se podría analizar el enfoque *Meeting Requirements* desde la perspectiva de la CI y la CE. De esta forma, se podría medir la CI y la CE, simultáneamente, en los componentes DW y DM. Este enfoque ha sido ampliamente utilizado en el área de Ingeniería de Software, por lo que surge el interés de estudiar su aplicabilidad en el dominio de los SDW.

Bibliografía

[1] Monica Scannapieco and Tiziana Catarci. Data quality under a computer science perspective. *Archivi & Computer*, 2:1–15, 2002.

[2] Monica Scannapieco, Paolo Missier, and Carlo Batini. Data quality at a glance. *Datenbank-Spektrum*, 14:6–14, 2005.

[3] Jack E Olson. *Data quality: the accuracy dimension*. Morgan Kaufmann, 2003.

[4] Richard Y. Wang and Diane M. Strong. Beyond accuracy: What data quality means to data consumers. *J. Manage. Inf. Syst.*, 12(4):5–33, March 1996.

[5] M Pamela Neely. The product approach to data quality and fitness for use: a framework for analysis. *Proc. of 10th International Conference on Information Quality*, 2005.

[6] Diane M. Strong, Yang W. Lee, and Richard Y. Wang. Data quality in context. *Commun. ACM*, 40(5):103–110, May 1997.

[7] Arthur D Chapman. *Principles of data quality*. GBIF, 2005.

[8] Giri Kumar Tayi and Donald P. Ballou. Examining data quality. *Commun. ACM*, 41(2):54–57, February 1998.

[9] Christopher Fox, Anany Levitin, and Thomas Redman. The notion of data and its quality dimensions. *Information Processing & Management*, 30(1):9 – 19, 1994.

[10] Fatimah Sidi, Abdullah Ramli, Marzanah Jabar, Lilly Suriani Affendey, Aida Mustapha, Hamidah Ibrahim, et al. Data quality comparative model for data warehouse. In *Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on*, pages 268–272. IEEE, 2012.

[11] Lorena Etcheverry, Verónika Peralta, and Mokrane Bouzeghoub. Qbox-foundation: a metadata platform for quality measurement. In *proceeding of the 4th Workshop on Data and Knowledge Quality (QDC2008)*, 2008.

[12] Jacky Akoka, Laure Berti-Equille, Omar Boucelma, Mokrane Bouzeghoub, Isabelle Comyn-Wattiau, Mireille Cosquer, Virginie Goasdoué-Thion, Zoubida Kedad, Sylvaine Nugier, Verónika Peralta, et al. A framework for quality evaluation in data integration systems. In *ICEIS (3)*, pages 170–175, 2007.

[13] Erhard Rahm and Hong Hai Do. Data cleaning: Problems and current approaches. *IEEE Data Engineering Bulletin*, 23:2000, 2000.

[14] Thomas C. Redman. The impact of poor data quality on the typical enterprise. *Commun. ACM*, 41:79–82, February 1998.

[15] Matteo Golfarelli and Stefano Rizzi. *Data Warehouse design: Modern principles and methodologies*. McGraw-Hill, Inc., 2009.

[16] William H Inmon. *Building the data warehouse*. John Wiley & sons, 2005.

[17] Ralph Kimball and Margy Ross. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. Wiley, 2nd edition, April 2002.

[18] Elke A. Rundensteiner, Andreas Koeller, and Xin Zhang. Maintaining data warehouses over changing information sources. *Commun. ACM*, 43:57–62, June 2000.

[19] Elzbieta Malinowski and Esteban Zimnyi. *Advanced Data Warehouse Design: From Conventional to Spatial and Temporal Applications (Data-Centric Systems and Applications)*. Springer Publishing Company, Incorporated, 1 edition, 2008.

[20] Fernando Carpani. CMDM, un modelo conceptual para la especificación de Bases Multidimensionales. Master's thesis, PEDECIBA. Universidad de la República. Montevideo, Uruguay, 2000.

[21] Neil Foshay, Avinandan Mukherjee, and Andrew Taylor. Does data warehouse end-user metadata add value? *Commun. ACM*, 50:70–77, November 2007.

[22] Matthias Jarke, Manfred Jeusfeld, Christoph Quix, and Panos Vassiliadis. Architecture and quality in data warehouses. In Barbara Pernici and Costantino Thanos, editors, *Advanced Information Systems Engineering*, volume 1413 of *Lecture Notes in Computer Science*, pages 93–113. Springer Berlin / Heidelberg, 1998. 10.1007/BFb0054221.

[23] A. Gosain and J. Singh. Achieving data warehouse quality using gdi approach. In *Applications of Digital Information and Web Technologies, 2008. ICADIWT 2008. First International Conference on the*, pages 494 –499, 2008.

[24] Yan Li and Kweku-Muata Osei-Bryson. Quality factory and quality notification service in data warehouse. In *Proceedings of the 3rd workshop on Ph.D. students in information and knowledge management*, PIKM '10, pages 25–32, New York, NY, USA, 2010. ACM.

[25] Matthias Jarke and Yannis Vassiliou. Data warehouse quality: A review of the DWQ project. In *IQ*, pages 299–313, 1997.

[26] Richard Y. Wang, Veda C. Storey, and Christopher P. Firth. A framework for analysis of data quality research. *IEEE Trans. on Knowl. and Data Eng.*, 7(4):623–640, August 1995.

[27] M. Gebhardt, M. Jarke, M.A. Jeusfeld, C. Quix, and S. Sklorz. Tools for data warehouse quality. In *Scientific and Statistical Database Management, 1998. Proceedings. Tenth International Conference on*, pages 229 –232, July 1998.

[28] Manuel Serrano, Coral Calero, Juan Trujillo, Sergio Lujn-Mora, and Mario Piattini. Empirical validation of metrics for conceptual models of data warehouses. In Anne Persson and Janis Stirna, editors, *Advanced Information Systems Engineering*, volume 3084 of *Lecture Notes in Computer Science*, pages 493–510. Springer Berlin / Heidelberg, 2004. 10.1007/978-3-540-25975-6_36.

[29] Gema Berenguer, Rafael Romero, Juan Trujillo, Manuel Serrano, and Mario Piattini. A set of quality indicators and their corresponding metrics for conceptual models of data warehouses. In A Min Tjoa and Juan Trujillo, editors, *Data Warehousing and Knowledge Discovery*, volume 3589 of *Lecture Notes in Computer Science*, pages 95–104. Springer Berlin / Heidelberg, 2005. 10.1007/11546849_10.

[30] Donald P. Ballou and Giri Kumar Tayi. Enhancing data quality in data warehouse environments. *Commun. ACM*, 42(1):73–78, January 1999.

[31] Rao R. Nemani and Ramesh Konda. A framework for data quality in data warehousing. In Will Aalst, John Mylopoulos, Norman M. Sadeh, Michael J. Shaw, Clemens Szyperski, Jianhua Yang, Athula Ginige, Heinrich C. Mayr, and Ralf-D. Kutsche, editors, *Information Systems: Modeling, Development, and Integration*, volume 20 of *Lecture Notes in Business Information Processing*, pages 292–297. Springer Berlin Heidelberg, 2009. 10.1007/978-3-642-01112-2_30.

[32] Ying Su and Zhanming Jin. A methodology for information quality assessment in data warehousing. In *Communications, 2008. ICC '08. IEEE International Conference on*, pages 5521–5525, May 2008.

[33] Lila Rao and Kweku-Muata Osei-Bryson. An approach for incorporating quality-based cost—benefit analysis in data warehouse design. *Information Systems Frontiers*, 10:361–373, July 2008.

[34] Jasna Rodic and Mirta Baranovic. Generating Data Quality Rules and Integration into ETL Process. In *Proceedings of the ACM Twelfth International Workshop on Data Warehousing and OLAP*, DOLAP '09, pages 65–72, New York, NY, USA, 2009. ACM.

[35] K. Ali and M.A. Warraich. A framework to implement data cleaning in enterprise data warehouse for robust data quality. In *Information and Emerging Technologies (ICIET), 2010 International Conference on*, pages 1 –6, 2010.

[36] Marie-Aude Aufaure, Nicolas Kuchmann-Beauger, Patrick Marcel, Stefano Rizzi, and Yves Vanrompay. Predicting your next olap query based on recent analytical sessions. In *Data Warehousing and Knowledge Discovery*, pages 134–145. Springer, 2013.

[37] Julien Aligon, Kamal Boulil, Patrick Marcel, and Veronika Peralta. A holistic approach to olap sessions composition: The falseto experience. In *Proceedings of the 17th International Workshop on Data Warehousing and OLAP*, pages 37–46. ACM, 2014.

[38] Florian Daniel, Fabio Casati, Themis Palpanas, and Oleksiy Chayka. Managing data quality in business intelligence applications. In *QDB/MUD*, pages 133–143, 2008.

[39] Paolo Ciaccia and Riccardo Torlone. Modeling the propagation of user preferences. In Manfred Jeusfeld, Lois Delcambre, and Tok-Wang Ling, editors, *Conceptual Modeling ER 2011*, volume 6998 of *Lecture Notes in Computer Science*, pages 304–317. Springer Berlin Heidelberg, 2011.

[40] C. Bolchini, C. A. Curino, G. Orsi, E. Quintarelli, R. Rossato, F. A. Schreiber, and L. Tanca. And what can context do for data? *Commun. ACM*, 52(11):136–140, November 2009.

[41] Cristiana Bolchini, Giorgio Orsi, Elisa Quintarelli, Fabio A. Schreiber, and Letizia Tanca. Context modeling and context awareness: steps forward in the context-addict project. *IEEE Data Eng. Bull.*, 34(2):47–54, 2011.

[42] Karen Henricksen, Jadwiga Indulska, and Andry Rakotonirainy. Modeling context information in pervasive computing systems. In *Proceedings of the First International Conference on Pervasive Computing*, Pervasive '02, pages 167–180, London, UK, UK, 2002. Springer-Verlag.

[43] Tao Gu, Hung Keng Pung, and Da Qing Zhang. A service-oriented middleware for building context-aware services. *J. Netw. Comput. Appl.*, 28(1):1–18, January 2005.

[44] Davy Preuveneers, Jan Van den Bergh, Dennis Wagelaar, Andy Georges, Peter Rigole, Tim Clerckx, Yolande Berbers, Karin Coninx, Viviane Jonckers, and Koen De Bosschere. Towards an extensible context ontology for ambient intelligence. In Panos Markopoulos, Berry Eggen, Emile Aarts, and JamesL. Crowley, editors, *Ambient Intelligence*, volume 3295 of *Lecture Notes in Computer Science*, pages 148–159. Springer Berlin Heidelberg, 2004.

[45] Maria R. Ebling, G. Hunt, and Hui Lei. Issues for Context Services for Pervasive Computing. In *Proceedings of the Advanced Workshop on Middleware for Mobile Computing*, Heidelberg, Germany, 2001. Springer.

[46] Cristiana Bolchini, Carlo A. Curino, Elisa Quintarelli, Fabio A. Schreiber, and Letizia Tanca. A data-oriented survey of context models. *SIGMOD Rec.*, 36(4):19–26, December 2007.

[47] Jan Chomicki. Preference formulas in relational queries. *ACM Trans. Database Syst.*, 28(4):427–466, December 2003.

[48] Philip D. Gray and Daniel Salber. Modelling and using sensed context information in the design of interactive applications. In *Proceedings of the 8th IFIP International Conference on Engineering for Human-Computer Interaction*, EHCI '01, pages 317–336, London, UK, UK, 2001. Springer-Verlag.

[49] Mary Bazire and Patrick Brézillon. Understanding context before using it. In *Modeling and using context*, pages 29–40. Springer, 2005.

[50] Leopoldo E. Bertossi, Flavio Rizzolo, and Lei Jiang. Data quality is context dependent. In *BIRTE*, pages 52–67, 2010.

[51] Cinzia Cappiello, Chiara Francalanci, and Barbara Pernici. Data quality assessment from the user's perspective. In *Proceedings of the 2004 international workshop on Information quality in information systems*, IQIS '04, pages 68–73, Paris, France, 2004. ACM.

[52] Anders Kofod-Petersen. How to do a structured literature review in computer science, 2012.

[53] Staffs Keele. Guidelines for performing systematic literature reviews in software engineering. In *Technical report, Ver. 2.3 EBSE Technical Report. EBSE*. 2007.

[54] Barbara Kitchenham, O Pearl Brereton, David Budgen, Mark Turner, John Bailey, and Stephen Linkman. Systematic literature reviews in software engineering–a systematic literature review. *Information and software technology*, 51(1):7–15, 2009.

[55] Barbara Kitchenham, Rialette Pretorius, David Budgen, O. Pearl Brereton, Mark Turner, Mahmood Niazi, and Stephen Linkman. Systematic literature reviews in software engineering a tertiary study. *Information and Software Technology*, 52(8):792 – 805, 2010.

[56] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, Sören Auer, and Pascal Hitzler. Quality assessment methodologies for linked open data. *Submitted to Semantic Web Journal*, 2013.

[57] Anjana Gosain and Heena. Literature review of data model quality metrics of data warehouse. *Procedia Computer Science*, 48:236 – 243, 2015. International Conference on Computer, Communication and Convergence (ICCC 2015).

[58] Paul Glowalla and Ali Sunyaev. Process-driven data quality management: A critical review on the application of process modeling languages. *Journal of Data and Information Quality (JDIQ)*, 5(1-2):7, 2014.

[59] Barbara A Kitchenham, David Budgen, and O Pearl Brereton. Using mapping studies as the basis for further research–a participant-observer case study. *Information and Software Technology*, 53(6):638–651, 2011.

[60] John Bailey, David Budgen, Mark Turner, Barbara Kitchenham, Pearl Brereton, and Stephen Linkman. Evidence relating to object-oriented software design: A survey. In *Proc. of the 1st Int. Symp. on Empirical Software Engineering and Measurement (ESEM 2007)*, pages 482–484. IEEE, 2007.

[61] Kai Petersen, Robert Feldt, Shahid Mujtaba, and Michael Mattsson. Systematic mapping studies in software engineering. In *12th International Conference on Evaluation and Assessment in Software Engineering*, volume 17. sn, 2008.

[62] Emelie Engström and Per Runeson. Software product line testing–a systematic mapping study. *Information and Software Technology*, 53(1):2–13, 2011.

[63] Javier Portillo-Rodríguez, Aurora Vizcaíno, Mario Piattini, and Sarah Beecham. Tools used in global software engineering: A systematic mapping review. *Information and Software Technology*, 54(7):663–685, 2012.

[64] Renato Lima Novais, André Torres, Thiago Souto Mendes, Manoel Mendonça, and Nico Zazworka. Software evolution visualization: A systematic

mapping study. *Information and Software Technology*, 55(11):1860–1883, 2013.

[65] M. Milani, L. Bertossi, and S. Ariyan. Extending contexts with ontologies for multidimensional data quality assessment. In *Data Engineering Workshops (ICDEW), 2014 IEEE 30th International Conference on*, pages 242–247, March 2014.

[66] Mathias Goller and Stefan Berger. Slowly Changing Measures. In *Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP*, DOLAP '13, pages 47–54, New York, NY, USA, 2013. ACM.

[67] H. Giggins and L. Brankovic. Statistical Disclosure Control: To Trust or Not to Trust. In *Computer Science and its Applications, 2008. CSA '08. International Symposium on*, pages 108–113, October 2008.

[68] R.G. Tiwari, M. Husain, B. Gupta, and A. Agrawal. Amalgamating Contextual Information into Recommender System. In *Emerging Trends in Engineering and Technology (ICETET), 2010 3rd International Conference on*, pages 15–20, November 2010.

[69] M.M. Hamad and A.A. Jihad. An Enhanced Technique to Clean Data in the Data Warehouse. In *Developments in E-systems Engineering (DeSE), 2011*, pages 306–311, December 2011.

[70] M. Munawar, N. Salim, and R. Ibrahim. Towards Data Quality into the Data Warehouse Development. In *Dependable, Autonomic and Secure Computing (DASC), 2011 IEEE Ninth International Conference on*, pages 1199–1206, December 2011.

[71] Zhimao Huang and Hong Peng. Improving Uncertain Data-Quality through Effective Use of Knowledge Base. In *Wireless Communications, Networking and Mobile Computing, 2008. WiCOM '08. 4th International Conference on*, pages 1–4, October 2008.

[72] Huang Yu, Zhang Xiao-yi, Yuan Zhen, and Jiang Guo-quan. A universal data cleaning framework based on user model. In *Computing, Communication, Control, and Management, 2009. CCCM 2009. ISECS International Colloquium on*, volume 2, pages 200–202, August 2009.

[73] A. Sundararaman. A framework for linking Data Quality to business objectives in decision support systems. In *Trendz in Information Sciences and Computing (TISC), 2011 3rd International Conference on*, pages 177–181, December 2011.

[74] A. Delgado, A. Marotta, and L. Gonzalez. Towards the construction of quality-aware web warehouses with bpmn 2.0 business processes. In *Research Challenges in Information Science (RCIS), 2014 IEEE Eighth International Conference on*, pages 1–6, May 2014.

[75] Oyku Isk, Mary C. Jones, and Anna Sidorova. Business intelligence success: The roles of BI capabilities and decision environments. *Information & Management*, 50(1):13 – 23, 2013.

[76] X. Franch, A. Mate, J.C. Trujillo, and C. Cares. On the joint use of i* with other modelling frameworks: A vision paper. In *Requirements Engineering Conference (RE), 2011 19th IEEE International*, pages 133–142, August 2011.

[77] Liu Xiang. An Agent-Based Architecture for Supply Chain Finance Cooperative Context-Aware Distributed Data Mining Systems. In *Internet and Web Applications and Services, 2008. ICIW '08. Third International Conference on*, pages 261–266, June 2008.

[78] J. Plaice, B. Mancilla, G. Ditu, and W.W. Wadge. Sequential Demand-Driven Evaluation of Eager TransLucid. In *Computer Software and Applications, 2008. COMPSAC '08. 32nd Annual IEEE International*, pages 1266–1271, July 2008.

[79] Raphal Thollot, Falk Brauer, Wojciech M. Barczynski, and Marie-Aude Aufaure. Text-to-query: Dynamically Building Structured Analytics to Illustrate Textual Content. In *Proceedings of the 2010 EDBT/ICDT Workshops*, EDBT '10, pages 14:1–14:8, New York, NY, USA, 2010. ACM.

[80] Lamia Oukid, Ounas Asfari, Fadila Bentayeb, Nadjia Benblidia, and Omar Boussaid. CXT-cube: Contextual Text Cube Model and Aggregation Operator for Text OLAP. In *Proceedings of the Sixteenth International Workshop on Data Warehousing and OLAP*, DOLAP '13, pages 27–32, New York, NY, USA, 2013. ACM.

[81] Muntazir Mehdi, Ratnesh Sahay, Wassim Derguech, and Edward Curry. On-the-fly Generation of Multidimensional Data Cubes for Web of Things. In *Proceedings of the 17th International Database Engineering & Applications Symposium*, IDEAS '13, pages 28–37, New York, NY, USA, 2013. ACM.

[82] J.M. Perez, R. Berlanga, M.J. Aramburu, and T.B. Pedersen. Towards a Data Warehouse Contextualized with Web Opinions. In *e-Business Engineering, 2008. ICEBE '08. IEEE International Conference on*, pages 697–702, October 2008.

[83] Liu Xiang. A Multiple Criteria Decision-Making Method for Enterprise Supply Chain Finance Cooperative Systems. In *Systems, 2009. ICONS '09. Fourth International Conference on*, pages 120–125, March 2009.

[84] P. Hernandez, I. Garrigos, and J. Mazon. Modeling Web Logs to Enhance the Analysis of Web Usage Data. In *Database and Expert Systems Applications (DEXA), 2010 Workshop on*, pages 297–301, August 2010.

[85] B. Vela, J. N. Mazn, C. Blanco, E. Fernndez-Medina, J. Trujillo, and E. Marcos. Development of secure xml data warehouses with qvt. *Information and Software Technology*, 55(9):1651 – 1677, 2013.

[86] Nicolas Prat, Isabelle Comyn-Wattiau, and Jacky Akoka. Combining objects with rules to represent aggregation knowledge in data warehouse and OLAP systems. *Data & Knowledge Engineering*, 70(8):732 – 752, 2011. Pushing Artificial Intelligence in Database and Data Warehouse Systems.

[87] Octavio Glorio, Jose-Norberto Mazn, Irene Garrigs, and Juan Trujillo. A personalization process for spatial data warehouse development. *Decision Support Systems*, 52(4):884 – 898, 2012. 1)Decision Support Systems for Logistics and Supply Chain Management 2)Business Intelligence and the Web.

[88] Víctor E. Silva Souza, Jose-Norberto Mazón, Irene Garrigós, Juan Trujillo, and John Mylopoulos. Monitoring Strategic Goals in Data Warehouses with Awareness Requirements. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, SAC '12, pages 1075–1082, New York, NY, USA, 2012. ACM.

[89] Micheline Elias and Anastasia Bezerianos. Annotating BI Visualization Dashboards: Needs & Challenges. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1641–1650, New York, NY, USA, 2012. ACM.

[90] I-Min A. Chen, Victor M. Markowitz, Ernest Szeto, Krishna Palaniappan, and Ken Chu. Maintaining a microbial genome & metagenome data analysis system in an academic setting. In *Proceedings of the 26th International Conference on Scientific and Statistical Database Management*, SSDBM '14, pages 3:1–3:11, New York, NY, USA, 2014. ACM.

[91] H. Tahir and P. Brezillon. A shared context approach for supporting experts in data ETL (Extraction, Transformation and Loading) processes. In *Intelligent Systems Design and Applications (ISDA), 2011 11th International Conference on*, pages 720–725, November 2011.

[92] R. Nachida and N. Fahima. A contextual personalized recommander system for mobile OLAP. In *Information Technology and e-Services (ICITeS), 2012 International Conference on*, pages 1–7, March 2012.

[93] P. Stack. Development of a Mobile Platform to Support Building Maintenance Engineering. In *Computer Software and Applications Conference Workshops (COMPSACW), 2012 IEEE 36th Annual*, pages 482–487, July 2012.

[94] Saida Aissi, Tarek Sboui, Mohamed Salah Gouider, Mohamed Ali Ben Hassine, and Lamjed Ben Said. A recommendation approach to enhance the interoperability between spatial datacubes. *Procedia Computer Science*, 56:558 – 565, 2015. The 10th International Conference on Future Networks and Communications (FNC 2015) / The 12th International Conference on Mobile Systems and Pervasive Computing (MobiSPC 2015) Affiliated Workshops.

[95] E. Silva, K. Becker, and R. Galante. Supporting Strategic Decision Making on Service Evolution Context Using Business Intelligence. In *Services Computing (SCC), 2013 IEEE International Conference on*, pages 240–247, June 2013.

[96] O. Badreddin, G. Mussbacher, D. Amyot, S.A. Behnam, R. Rashidi-Tabrizi, E. Braun, M. Alhaj, and G. Richards. Regulation-Based Dimensional Modeling for Regulatory Intelligence. In *Requirements Engineering and Law (RELAW), 2013 Sixth International Workshop on*, pages 1–10, July 2013.

[97] Xiaoyan Bai, D. White, and D. Sundaram. Context adaptive visualization for effective business intelligence. In *Communication Technology (ICCT), 2013 15th IEEE International Conference on*, pages 786–790, November 2013.

[98] P.R. Clavier, H.H. Lotriet, and J.J. van Loggerenberg. Business Intelligence Challenges in the Context of Goods- and Service-Dominant Logic. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 4138–4147, January 2012.

[99] K. Ali and M.A. Warraich. A framework to implement data cleaning in enterprise data warehouse for robust data quality. In *Information and Emerging Technologies (ICIET), 2010 International Conference on*, pages 1–6, June 2010.

[100] F. Sidi, A. Ramli, M.A. Jabar, L.S. Affendey, A. Mustapha, and H. Ibrahim. Data quality comparative model for data warehouse. In *Information Retrieval Knowledge Management (CAMP), 2012 International Conference on*, pages 268–272, March 2012.

[101] W. Gongora de Almeida, R.T. de Sousa, F.E. de Deus, G.D. Amvame Nze, and F.L. Lopes de Mendonca. Taxonomy of data quality problems in multidimensional Data Warehouse models. In *Information Systems and Technologies (CISTI), 2013 8th Iberian Conference on*, pages 1–7, June 2013.

[102] Anil Rai, Vipin Dubey, K. K. Chaturvedi, and P. K. Malhotra. Design and development of data mart for animal resources. *Computers and Electronics in Agriculture*, 64(2):111 – 119, 2008.

[103] Vasco Santos and Orlando Belo. Modeling ETL Data Quality Enforcement Tasks Using Relational Algebra Operators. *Procedia Technology*, 9(0):442 – 450, 2013. CENTERIS 2013 - Conference on ENTERprise Information Systems / ProjMAN 2013 - International Conference on Project MANagement/ HCIST 2013 - International Conference on Health and Social Care Information Systems and Technologies.

[104] N. Prat and S. Madnick. Measuring Data Believability: A Provenance Approach. In *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*, pages 393–393, January 2008.

[105] S. Maddodi, G.V. Attigeri, and A.K. Karunakar. Data Deduplication Techniques and Analysis. In *Emerging Trends in Engineering and Technology (ICETET), 2010 3rd International Conference on*, pages 664–668, November 2010.

[106] Bing Chen, Xuchu Weng, Beizhan Wang, and Xueqin Hu. Analysis and solution of data quality in data warehouse of Chinese materia medica. In *Computer Science Education, 2009. ICCSE '09. 4th International Conference on*, pages 823–827, July 2009.

[107] Bernhard Wieder and Maria-Luise Ossimitz. The impact of business intelligence on the quality of decision making a mediation model. *Procedia Computer Science*, 64:1163 – 1171, 2015. Conference on ENTERprise Information Systems/International Conference on Project MANagement/Conference on Health and Social Care Information Systems and Technologies, CENTERIS/ProjMAN / HCist 2015 October 7-9, 2015.

[108] Jie Zhang, Qiaoyan Wen, and Hua Zhang. The Research in Improving the Quality of DW Data: The Job-Scheduling and Checking Based Program in Upgrading DW Performance. In *Wireless Communications, Networking and Mobile Computing, 2009. WiCom '09. 5th International Conference on*, pages 1–4, September 2009.

[109] Gao Xiang and Wang Min. Applying data cleaning in Changqing Oilfield Company's data warehouse. In *Geoscience and Remote Sensing (IITA-GRS), 2010 Second IITA International Conference on*, volume 2, pages 605–607, August 2010.

[110] A.L. McNab and D.A. Ladd. Information Quality: The Importance of Context and Trade-Offs. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 3525–3532, January 2014.

[111] A. Manzoor, Hong-Linh Truong, and S. Dustdar. Quality Aware Context Information Aggregation System for Pervasive Environments. In *Advanced Information Networking and Applications Workshops, 2009. WAINA '09. International Conference on*, pages 266–271, May 2009.

[112] P. Glowalla, P. Balazy, D. Basten, and A. Sunyaev. Process-Driven Data Quality Management An Application of the Combined Conceptual Life Cycle Model. In *System Sciences (HICSS), 2014 47th Hawaii International Conference on*, pages 4700–4709, January 2014.

[113] Barbara Catania, Giovanna Guerrini, Alberto Belussi, Federica Mandreoli, Riccardo Martoglia, and Wilma Penzo. Wearable Queries: Adapting Common Retrieval Needs to Data and Users. In *Proceedings of the 7th International Workshop on Ranking in Databases*, DBRank '13, pages 7:1–7:3, New York, NY, USA, 2013. ACM.

[114] Tamraparni Dasu, Ji Meng Loh, and Divesh Srivastava. Empirical glitch explanations. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '14, pages 572–581, New York, NY, USA, 2014. ACM.

[115] M. Helfert and O. Foley. A Context Aware Information Quality Framework. In *Cooperation and Promotion of Information Resources in Science and Technology, 2009. COINFO '09. Fourth International Conference on*, pages 187–193, November 2009.

[116] R. Price and G. Shanks. DQ Tags and Decision-Making. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pages 1–10, January 2010.

[117] Di Zheng, Jun Wang, and B. Kerong. Evaluation of Quality Measure Factors for the Middleware Based Context-Aware Applications. In *Computer and Information Science (ICIS), 2012 IEEE/ACIS 11th International Conference on*, pages 403–408, May 2012.

[118] Helen-Tadesse Moges, Karel Dejaeger, Wilfried Lemahieu, and Bart Baesens. A multidimensional analysis of data quality for credit risk management: New insights and challenges. *Information & Management*, 50(1):43 – 58, 2013.

[119] Jingyu Han, Dawei Jiang, and Lingjuan Li. Automatic accuracy assessment via hashing in multiple-source environment. *Expert Systems with Applications*, 37(3):2609 – 2620, 2010.

[120] Benjamin T. Hazen, Christopher A. Boone, Jeremy D. Ezell, and L. Allison Jones-Farmer. Data quality for data science, predictive analytics, and big

data in supply chain management: An introduction to the problem and suggestions for research and applications. *International Journal of Production Economics*, 154(0):72 – 80, 2014.

[121] Xinkai Yang. An Adaptive Mechanism for Inconsistent Context Resolution in Ubiquitous Computing. In *Control Engineering and Communication Technology (ICCECT), 2012 International Conference on*, pages 703–706, December 2012.

[122] D.S. Alberts, M. Vassiliou, and J. Agre. C2 information quality: An enterprise systems perspective. In *MILITARY COMMUNICATIONS CONFERENCE, 2012 - MILCOM 2012*, pages 1–7, October 2012.

[123] Fei Li, S. Nastic, and S. Dustdar. Data Quality Observation in Pervasive Environments. In *Computational Science and Engineering (CSE), 2012 IEEE 15th International Conference on*, pages 602–609, December 2012.

[124] G. Rogova, M. Hadzagic, M. St-Hilaire, M.C. Florea, and P. Valin. Context-based information quality for sequential decision making. In *Cognitive Methods in Situation Awareness and Decision Support (CogSIMA), 2013 IEEE International Multi-Disciplinary Conference on*, pages 16–21, February 2013.

[125] Carlos A Hurtado, Claudio Gutierrez, and Alberto O Mendelzon. Capturing summarizability with integrity constraints in olap. *ACM Transactions on Database Systems (TODS)*, 30(3):854–886, 2005.

[126] Elke A. Rundensteiner, Andreas Koeller, and Xin Zhang. Maintaining data warehouses over changing information sources. *Commun. ACM*, 43:57–62, June 2000.

[127] D.J. Berndt, J.W. Fisher, A.R. Hevner, and J. Studnicki. Healthcare data warehousing and quality assurance. *Computer*, 34(12):56 –65, December 2001.

[128] Philip B Crosby and Quality Is Free. The art of making quality certain. *New York: New American Library*, 17, 1979.

[129] Robert W Hoyer and Brooke BY Hoyer. What is quality. *Quality Progress*, 34(7):53–62, 2001.

[130] Joseph Juran and A Blanton Godfrey. Quality handbook. *Republished McGraw-Hill*, 1999.

[131] ISO/IEC 25010:2011. Systems, software engineering. Systems, software Quality Requirements, Evaluation (SQuaRE). System, and software quality models. http://www.iso.org/iso/. Accedido el 08-10-2015.

[132] Mayte Herrera, MaÁngeles Moraga, Ismael Caballero, and Coral Calero. Quality in use model for web portals (qiuwep). In Florian Daniel and FedericoMichele Facca, editors, *Current Trends in Web Engineering*, volume 6385 of *Lecture Notes in Computer Science*, pages 91–101. Springer Berlin Heidelberg, 2010.

[133] Adriana Marotta and Alejandro Vaisman. Rule-based multidimensional data quality assessment using contexts. To be submitted, 2015.

[134] Flavia Serra and Adriana Marotta. Quality assessment in data warehouse: A context-based approach. Proceedings of XXXI International Conference of the Chilean Computer Science Society (SCCC). Valparaíso, Chile. 2012.

Apéndice A

Definición de las cadenas de búsqueda

El propósito de esta sección es permitir al lector reproducir la búsqueda en las librerías que han sido seleccionadas para llevar a cabo la Revisión Sistemática o Mapping Review. Para cada una de las librerías fue necesario adaptar las cadenas de búsqueda parciales, teniendo en cuenta únicamente las funcionalidades que éstas ofrecen. Las búsquedas que se realizaron con las cadenas que se presentan a continuación fueron ejecutadas durante el mes de junio de 2014.

ACM *Digital library*

Para el caso de la librería digital ACM fue necesario adaptar las cadenas de búsqueda parciales SS1 (que considera las áreas DW y CTX), SS2 (que considera las áreas DW y DQ) y SS3 (que considera las áreas DQ y CTX) como se muestra en las Tablas A.1, A.2 y A.3 respectivamente. La bandera *FtFlag* con valor *yes* indica que se solicitaron todos los trabajos que estaban completamente disponibles. En la primera etapa de la aplicación de la metodología de búsqueda se seleccionaron trabajos publicados en el período 2008-2014, mientras que en la segunda etapa se ejecutaron las mismas cadenas de búsqueda parciales, para un nuevo período de publicación, 2014-2015.

((Abstract:Context OR Abstract:"data tailoring" OR Abstract:"pervasive computing" OR Abstract:"ubiquitous computing" OR Abstract:"preference") and (Abstract:"Data warehouse" OR Abstract:warehousing OR Abstract:"business intelligence" OR Abstract:"multidimensional database" OR Abstract:"dimension hierarchies" OR Abstract:"fact table" OR Abstract:cube) and (Publisher:ACM) and (FtFlag:yes))

Tabla A.1: **Adaptación de SS1 para la librería digital ACM**

((Abstract:"Data Quality" OR Abstract:"quality factor" OR Abstract:"quality metric" OR Abstract:"quality dimension" OR Abstract:"quality measure" OR Abstract:"quality attributes") and (Abstract:"Data warehouse" OR Abstract:warehousing OR Abstract:"business intelligence" OR Abstract:"multidimensional database" OR Abstract:"dimension hierarchies" OR Abstract:"fact table" OR Abstract:cube) and (Publisher:ACM) and (PublishedAs:journal OR PublishedAs:proceeding OR PublishedAs:magazine) and (FtFlag:yes))

Tabla A.2: **Adaptación de SS2 para la librería digital ACM**

(((Abstract:Context OR Abstract:"data tailoring" OR Abstract:"pervasive computing" OR Abstract:"ubiquitous computing" OR Abstract:"preference") and (Abstract:"Data Quality" OR Abstract:"quality factor" OR Abstract:"quality metric" OR Abstract:"quality dimension" OR Abstract:"quality measure") and (Publisher:ACM) and (PublishedAs:journal OR PublishedAs:proceeding OR PublishedAs:magazine) and (FtFlag:yes)))

Tabla A.3: **Adaptación de SS3 para la librería digital ACM**

IEEE*Xplore*

Las funcionalidades del motor de búsqueda de la IEEE permitieron especificar el tipo de contenido, y si bien no se anexaba a la cadena de búsqueda, también fue posible seleccionar trabajos completos indicando el príodo de publicación. Las cadenas de búsqueda parciales SS1, SS2 y SS3 como se muestra en las Tablas A.4, A.5 y A.6 respectivamente. Para esta biblioteca digital se sigue el mismo procedimiento, en la primera etapa de la aplicación de la metodología de búsqueda se seleccionaron trabajos publicados en el período 2008-2014, luego, en la segunda etapa se ejecutaron las mismas cadenas de búsqueda parciales, para un nuevo período de publicación, 2014-2015.

ScienceDirect

Las Tablas A.7, A.8 y A.9 presentan las cadenas de búsqueda para SS1, SS2 y SS3 respectivamente, que fueron utilizadas en los motores de búsqueda de la Librería ScienceDirect. En este caso, se pudo especificar las áreas de investigación, *Computer Science* e *Engineering*. Una vez más, en la primera etapa de la aplicación de la metodología de búsqueda se seleccionaron trabajos publicados en el período 2008-2014 y luego, en la segunda etapa, se ejecutaron las mismas cadenas de búsqueda parciales, para un nuevo período de publicación, 2014-2015.

(((“Abstract”:Context OR “Abstract”:“data tailoring” OR “Abstract”:“pervasive computing” OR “Abstract”:“ubiquitous computing” OR “Abstract”:preference) AND (“Abstract”:“Data warehouse” OR “Abstract”:warehousing OR “Abstract”:“business intelligence” OR “Abstract”:“multidimensional database” OR “Abstract”:“dimension hierarchies” OR “Abstract”:“fact table” OR “Abstract”:cube))) You Refined by: Content Type: Conference Publications, Journals & Magazines, Books & eBooks Publisher: IEEE Publication Year: 2008 - 2014 Full text

Tabla A.4: **Adaptación de SS1 para la librería digital IEEE**

(((“Abstract”:“Data quality” OR “Abstract”:“quality factor” OR “Abstract”:“quality metric” OR “Abstract”:“quality dimension” OR “Abstract”:“quality measure”) AND (“Abstract”:“Data warehouse” OR “Abstract”:warehousing OR “Abstract”:“business intelligence” OR “Abstract”:“multidimensional database” OR “Abstract”:“dimension hierarchies” OR “Abstract”:“fact table” OR “Abstract”:cube))) You Refined by: Content Type: Conference Publications, Journals & Magazines, Books & eBooks Publication Year: 2008 - 2014 Full text

Tabla A.5: **Adaptación de SS2 para la librería digital IEEE**

((“Abstract”:Context OR “Abstract”:“data tailoring” OR “Abstract”:“pervasive computing” OR “Abstract”:“ubiquitous computing” OR “Abstract”:preference) AND (“Abstract”:“Data quality” OR “Abstract”:“information quality” OR “Abstract”:“quality factor” OR “Abstract”:“quality metric” OR “Abstract”:“quality dimension” OR “Abstract”:“quality measure”)) You Refined by: Content Type: Conference Publications , Journals & Magazines Publisher: IEEE Publication Year: 2008 - 2014 Full text

Tabla A.6: **Adaptación de SS3 para la librería digital IEEE**

pub-date > 2007 and ABSTRACT(Context OR "data tailoring" OR "pervasive computing" OR "ubiquitous computing" OR preference) and ABSTRACT("Data warehouse" OR warehousing OR "business intelligence" OR "multidimensional database" OR "dimension hierarchies" OR "fact table" OR cube)[All Sources(Computer Science,Engineering)]

Tabla A.7: **Adaptación de SS1 para la librería digital ScienceDirect**

pub-date > 2007 and ABSTRACT("Data quality" OR "information quality" OR "quality factor" OR "quality metric" OR "quality dimension" OR "quality measure") and ABSTRACT("Data warehouse" OR warehousing OR "business intelligence" OR "multidimensional database" OR "dimension hierarchies" OR "fact table" OR cube)[All Sources(Computer Science,Engineering)]

Tabla A.8: **Adaptación de SS2 para la librería digital ScienceDirect**

pub-date > 2007 and ABSTRACT(Context OR "data tailoring" OR "pervasive computing" OR "ubiquitous computing" OR preference) and ABSTRACT("Data quality" OR "information quality" OR "quality factor" OR "quality metric" OR "quality dimension" OR "quality measure")[All Sources(Computer Science,Engineering)]

Tabla A.9: **Adaptación de SS3 para la librería digital ScienceDirect**