DEFINICIÓN DE UNA ARQUITECTURA DE REFERENCIA PARA PLATAFORMAS DE SERVICIOS DE DATOS

EDISSON ESTELIO GUTIÉRREZ JIMÉNEZ

UNIVERSIDAD DE MEDELLÍN
FACULTAD DE INGENIERÍA
MAESTRÍA EN INGENIERÍA DE SOFTWARE
MEDELLÍN
2021

DEFINICIÓN DE UNA ARQUITECTURA DE REFERENCIA PARA PLATAFORMAS DE SERVICIOS DE DATOS

EDISSON ESTELIO GUTIÉRREZ JIMÉNEZ

Trabajo de grado para optar al título de
Magister en Ingeniería de Software

Director
Juan Bernardo Quintero
Doctor en Ingeniería Electrónica

Codirectora
Bell Manrique Losada
Doctora en Ingeniería

UNIVERSIDAD DE MEDELLÍN
FACULTAD DE INGENIERÍA
MAESTRÍA EN INGENIERÍA DE SOFTWARE
MEDELLÍN
2021

Dedico este trabajo principalmente a Dios, quien siempre ha estado presente en el camino de mi vida, bendiciéndome y dándome fuerzas para trabajar por mis sueños. A mi madre por su amor y esfuerzo por sacarme adelante y por su apoyo incondicional. A mis hermanos que siempre han estado presentes en todo momento de mi vida. A mi esposa que siempre me motivó y me ayudó a lograr este sueño. Una dedicatoria especial a la memoria de mi tío José, quien entrego toda su vida para ayudarnos, me enseñó a vivir sin apegos y a tener una vida de servicio.

AGRADECIMIENTOS

Al finalizar este trabajo deseo utilizar este espacio para expresar mis agradecimientos. En primer lugar, agradezco al Padre Celestial, fuente de todo bien, por ser el soporte y guía en cada proyecto de mi vida, por ser la fuente inagotable de sabiduría que alimenta cada decisión de mi existencia.

Así mismo, agradezco al director de esta tesis Dr. Juan Bernardo Quintero, por la dedicación y apoyo que ha brindado a este proyecto, por el respeto a mis ideas y por la dirección y el rigor que han permitido que este trabajo llegue a feliz término.

Mi agradecimiento a la Dra. Bell Manrique Losada por la orientación y atención a mis consultas sobre metodología, siempre me ha impulsado a obtener nuevos y mejores resultados.

Pero este trabajo de investigación también es fruto del reconocimiento y del apoyo de personas que me dan la fuerza y energía para crecer personal y profesionalmente.

Gracias a mi madre por su ejemplo de responsabilidad, honestidad y servicio, por promover mis sueños y sobre todo por los consejos, valores y principios que me ha inculcado.

Finalmente, un agradecimiento especial a mi esposa por su paciencia, comprensión y apoyo con este proyecto, por soñar conmigo y sobre todo por ser mi compañera de vida.

A todos los que tuvieron alguna incidencia en el resultado final de este proyecto, infinitas gracias por todo el apoyo y colaboración.

CONTENIDO

pág.

LISTA DE TABLAS

pág.

LISTA DE FIGURAS

pág.

ACRÓNIMOS

BI:	Business Intelligence
ETL:	Extract, Transform and Load
TIC:	Tecnologías de la Información y la Comunicación
PaaS:	Plataforma como Servicio
IaaS:	Infraestructura como Servicio
EPS:	Empresa Promotora de Salud
IPS:	Institución Prestadora de Servicios de Salud
UCE:	Unidad de Cuidados Especiales o Intermedios
UCI:	Unidad de Cuidados Intensivos

RESUMEN

Big Data se refiere a conjuntos de datos cuyo volumen, velocidad y variedad dificultan su captura, gestión y procesamiento mediante tecnologías y herramientas convencionales. Este concepto ha generado nuevas necesidades en las organizaciones para permitir la captura, almacenamiento y análisis de datos con estas características y así obtener información relevante para la toma de decisiones. Un reto para las organizaciones es la implementación de una arquitectura que permita cubrir estas necesidades, ya que deben considerar las diferentes tecnologías existentes y deben establecer las políticas para el gobierno de datos que están en manos de los usuarios. Una arquitectura de referencia de una plataforma de analítica de datos, que se desvincule de herramientas tecnológicas es una guía que le permite a las organizaciones trazar un camino para lograr la gestión de grandes volúmenes de datos y así tener herramientas efectivas para la toma de decisiones empresariales.

La arquitectura de referencia es lo suficientemente general como para implementarse con diferentes tecnologías, paradigmas informáticos y software analítico, dependiendo de los requisitos y propósitos de cada organización. En el proyecto desarrollado se realizó la implementación de la arquitectura con datos de la atención de urgencias en centros hospitalarios de la ciudad de Medellín.

Uno de los resultados del trabajo de investigación es que la arquitectura propuesta considera diferentes tipos de usuario y de fuentes de datos, no genera dependencia por el tipo de herramientas tecnológica que se utilizan y establece una capa para el gobierno de datos.

Palabras clave: Big data, arquitectura, analítica de datos, gobierno de datos.

ABSTRACT

Big Data refers to data set whose volume, velocity, and variety make it difficult to capture, manage and process using conventional technologies and tools. This concept is generating new needs in organizations to allow the capture, storage, and analysis of data with these characteristics and thus obtain relevant information for decision-making. A challenge for organizations is the implementation of an architecture that covers these needs, since they must consider the different existing technologies and must establish the policies for data governance that will be available to users. A reference architecture of a data analytics platform that is capable of decoupling from technological tools will be a guide that will allow organizations to define a path to achieve the management of these data and thus have effective tools for make decisions in the company.

The reference architecture is general enough to be implemented with different technologies, computing paradigms and analytical software, depending on the requirements and purposes of each organization. In the developed project, the architecture was implemented with data from emergency care in hospitals in the Medellín city.

One of the results of the research work is that the proposed architecture considers different types of user and data sources, does not generate dependency due to the type of technological tools used and establishes a layer for data governance.

Keywords: Big data; architecture; data analytics, data government

PARTE I
INTRODUCCIÓN

"Difunde el amor donde quiera que vayas. No dejes que nadie se aleje de ti sin ser un poco más feliz" – Madre Teresa de Calcuta

El descubrimiento de conocimiento y la toma de decisiones a partir de datos de gran volumen, variados, de diferente tipo y con un rápido crecimiento es un reto para las empresas en términos de almacenamiento, administración y procesamiento (Bibri, 2019). El concepto de Big Data es un conjunto de datos cuyo volumen, variedad y velocidad dificultan su captura, gestión y procesamiento mediante tecnologías y herramientas convencionales (Madden, 2012). Este concepto está generando en las organizaciones la necesidad de recolectar, almacenar, analizar y exhibir datos con estas características para obtener información que ayude a la toma de decisiones y a la proyección en el mercado (Oussous et al., 2018). Unido a esto, los avances tecnológicos de la Cuarta Revolución Industrial o Industria 4.0 y su aplicación en las organizaciones, los cuales buscan la automatización y el intercambio de datos, generarán más volumen de datos y es en este punto donde el Big Data cobra relevancia ayudando a mejorar el desarrollo de productos y a la innovación de los servicios (Wan et al., 2015).

Con el fin de gestionar datos con las características propias del Big Data se han desarrollado diferentes herramientas y arquitecturas tecnológicas, lo que ha generado un desafío para los profesionales de este campo (Mazumder, 2016), ya que ellos deben comprender y seleccionar las herramientas para abordar un problema organizacional especifico relacionado con Big Data y así obtener el conocimiento que se puede extraer de los datos.

Varias dificultades se presentan en las organizaciones al definir una arquitectura para Big Data, una es que las organizaciones no identifican la forma y las herramientas tecnológicas para gestionar y procesar datos con esas características (Jovanovic et al., 2016), otra es que las plataformas desarrolladas no facilitan el aumento de la capacidad de trabajo sin comprometer el funcionamiento (Oussous et al., 2018), y la última es el desconocimiento del dominio del problema que se desea atender y de los datos que se procesarán (Zhelev & Rozeva, 2017).

Los servicios de datos se refieren a plataformas para el almacenamiento y distribución de datos con características relacionadas al Big Data (Pollock & Dietrich, 2009), los cuales permiten la diagramación para facilitar el análisis de los datos. La dificultad actual se presenta con la implementación de los servicios de datos, debido a las diferentes tecnologías que existen alrededor de estos servicios sin tener pautas que indiquen cómo será su construcción. Esta misma situación la exponen Jovanovic et al. (Jovanovic et al., 2016) al mencionar que la variedad de motores de ejecución para la analítica de datos y la complejidad de las transformaciones son un desafío para las arquitecturas de Big Data y los sistemas de inteligencia de negocio de próxima generación.

Es necesario plantear la definición de una arquitectura de referencia de una plataforma de Big Data, que sea capaz de desvincularse de los detalles técnicos (Blazquez & Domenech, 2018), dicha arquitectura será la base para gestionar datos con características relacionadas al Big Data, no debe estar ligada a

herramientas técnicas particulares y debe incluir los posibles tipos de usuario que la pueden utilizar (Zaghloul et al., 2015), los componentes tecnológicos serán determinados por las organizaciones al momento de implementar la arquitectura. En este trabajo se presenta una aproximación a una arquitectura de referencia para plataformas de Big Data, la cual incluye dos componentes: los lineamientos que permiten guiar el proceso de implementación y una fase de gobierno de datos, que como lo exponen Khatri et al. (Khatri & Brown, 2010), permitirá tener control sobre la información que se expone, definiendo los usuarios que tienen derechos de decisión y se hacen responsables de los activos de datos de la organización.

CAPÍTULO 1

Introducción

El crecimiento en el volumen, diversidad y velocidad de los datos han ocasionado que las organizaciones busquen proporcionar mayor capacidad de almacenamiento y mejorar el procesamiento para el análisis y la toma de decisiones (Oussous et al., 2018), para realizar estas actividades las organizaciones deben identificar las fuentes y tipos de datos que se manejan y definir la forma en que se gestionarán, con el fin de generar valor en la empresa.

1.1. Campo de acción y enfoque de la investigación

Las organizaciones son más conscientes que el análisis de datos se está convirtiendo en un factor para ser competitivos y para descubrir nuevos conocimientos y así lograr la personalización de los servicios (Najafabadi et al., 2015), por esta razón, establecer una arquitectura de Big Data que les permita la ingestión, almacenamiento, análisis y visualización de grandes volúmenes de datos es un desafío que implica escalabilidad, disponibilidad, integridad, transformación y gobernanza de datos (Blazquez & Domenech, 2018).

Los servicios de datos son plataformas para el almacenamiento y distribución de datos con características relacionadas al Big Data (Pollock & Dietrich, 2009) y le permiten a las organizaciones realizar las actividades anteriormente mencionadas.

Contar con lineamientos que permitan construir una arquitectura de referencia que soporte plataformas de servicios de datos, ajustadas a las necesidades de la empresa (Blazquez & Domenech, 2018), ayudará a los departamentos de sistemas de las organizaciones a trazar un camino para lograr la gestión de grandes volúmenes de información y a tener un enfoque organizacional basado en DataOps (Ereth, 2018), para acoger rápidamente los continuos cambios tecnológicos y arquitectónicos generados por el crecimiento en el volumen, diversidad y velocidad de los datos.

En la Figura 1 se ilustra el importante papel que tiene la gestión de los datos en las organizaciones.

Figura 1. Ciclo de vida de los datos en las organizaciones

1.1.1. Big Data

Big data es un fenómeno caracterizado por un aumento continuo en el volumen, la variedad, la velocidad y la veracidad de los datos que requieren técnicas y tecnologías avanzadas para capturar, almacenar, distribuir, administrar y analizar estos datos (Ebner et al., 2014). Los cambios rápidos y continuos en el volumen y variedad de los datos requieren de una infraestructura técnica avanzada y de arquitecturas que abarquen estas nuevas características.

1.1.2. Arquitectura de software

La arquitectura de software se define como la estructura de los componentes de un programa o sistema, sus interrelaciones, y los principios y guías que controlan su diseño y evolución en el tiempo (Kazman et al., 1993). Dentro del desarrollo de software esta área juega un papel importante debido a la continua evolución de sistemas de información y a la creciente evolución de nuevas tecnologías.

Las propiedades particulares del Big Data y el desafío en las organizaciones para gestionar datos con sus características requiere arquitecturas específicas que posean componentes que almacenen, procesen y analicen el volumen y variedad de los datos (Blazquez & Domenech, 2018).

1.1.3. Arquitecturas para Big Data

Con el fin de atender los retos tecnológicos que demanda el Big Data se han definido arquitecturas para atender sus características particulares, algunas de esas arquitecturas son:

- Arquitecturas intensivas de datos, se componen generalmente de múltiples dispositivos computacionales que pueden resumirse en términos de trabajos, adquiriendo datos de una o más fuentes y produciendo datos en uno o más sumideros. Las fuentes y sumideros de datos pueden ser de

varios tipos. Los trabajos dentro de esta arquitectura pueden funcionar por lotes, secuencia o interactivo (Artac et al., 2018).

- Arquitecturas Lambda y Kappa. La arquitectura Lambda combina el procesamiento de datos por lotes y en tiempo real, se enfoca principalmente en la ingestión de datos y presenta dificultades en la implementación y el soporte, debido a que mantener estos procesamientos es complejo; la arquitectura Kappa simplifica la arquitectura Lambda combinando el procesamiento por lotes y en tiempo real en una sola capa, llamada capa de procesamiento, la segunda capa que posee es la de publicación, que se utiliza para consultar resultados (Zhelev & Rozeva, 2017). En la Figura 2 se ilustra la arquitectura Lamba y en la Figura 3 se ilustra la Kappa.

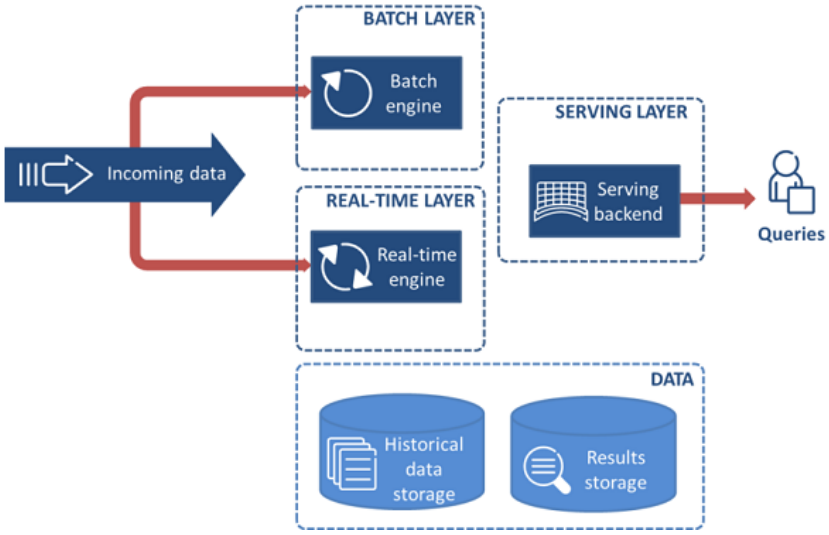


Figura 2. Arquitectura Lambda (Zhelev & Rozeva, 2017)

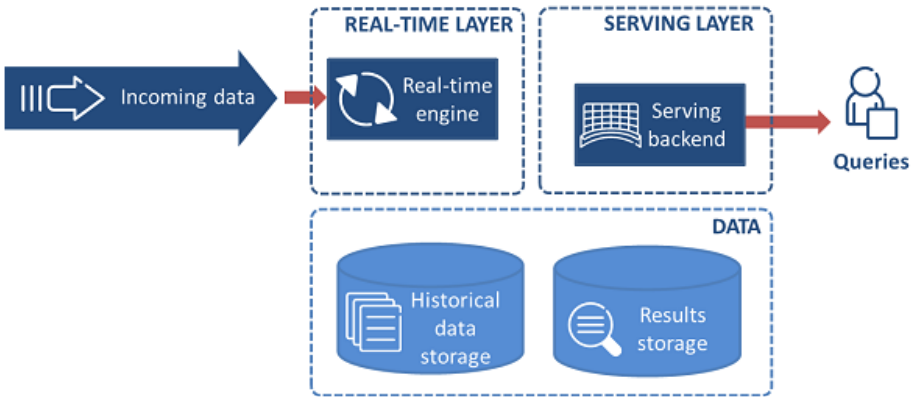


Figura 3. Arquitectura Kappa (Zhelev & Rozeva, 2017)

- Arquitectura basada en modelos, fue desarrollada para agilizar los procesos de analítica desde la preparación de los datos hasta la visualización. Esta última etapa logra ser más concreta por medio de la automatización de los objetivos que el usuario desea visualizar (Golfarelli & Rizzi, 2019). En la Figura 4 se ilustra esta arquitectura.

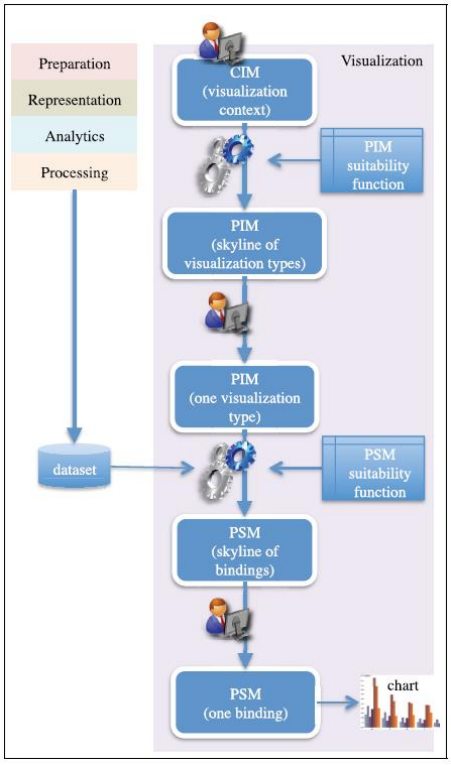


Figura 4. Arquitectura basada en modelos (Golfarelli & Rizzi, 2019)

- Arquitectura de procesamiento de Big Data, consiste en cuatro etapas: recolección o recopilación, carga, transformación y extracción de datos (Krishnan, 2013). En la Figura 5 se ilustra el diagrama de esta arquitectura.

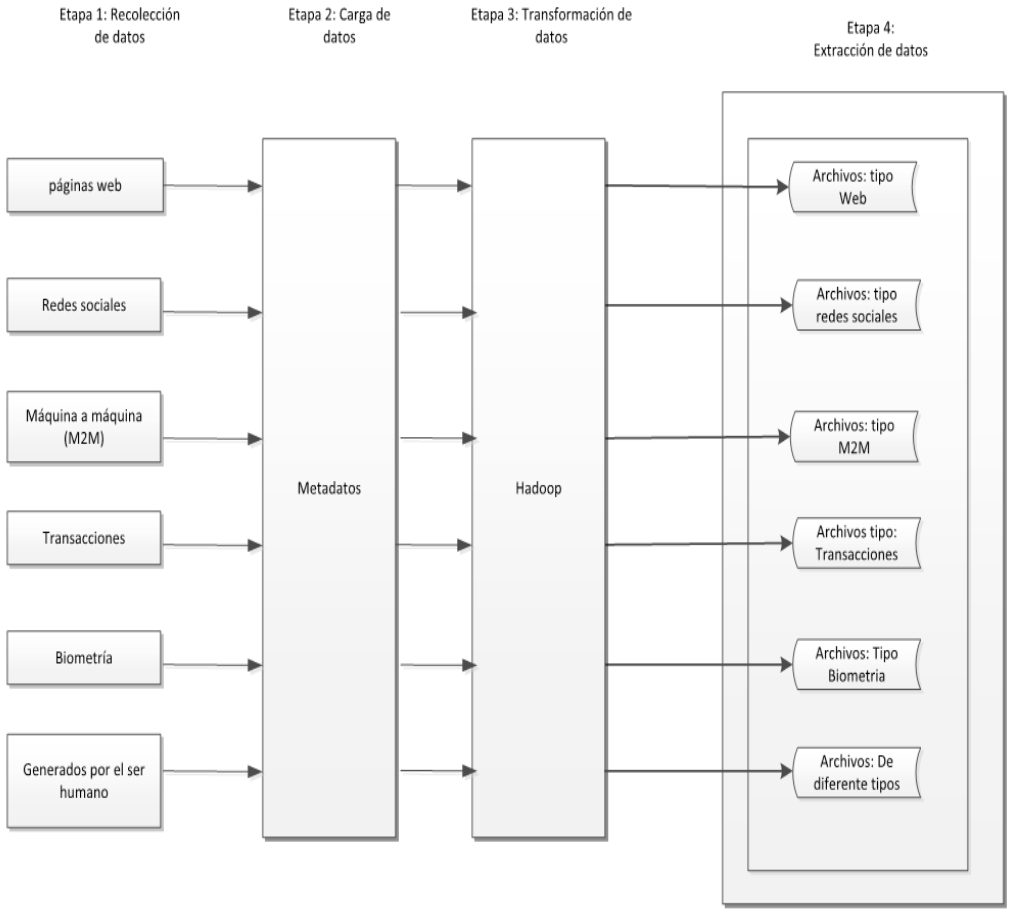


Figura 5. Arquitectura de procesamiento de Big Data (Krishnan, 2013)

- Arquitectura de Big Data de Marcus (Camargo Vega et al., 2015), es un modelo compuesto por 7 niveles: fuentes de datos externos, secuencia y procesamiento, fundación altamente escalable, bases de datos operacionales y de analítica, análisis e interfaces de bases de datos, aplicaciones e interfaces de usuario, servicios de apoyo. En la Figura 6 se ilustra el modelo por niveles de esta arquitectura.

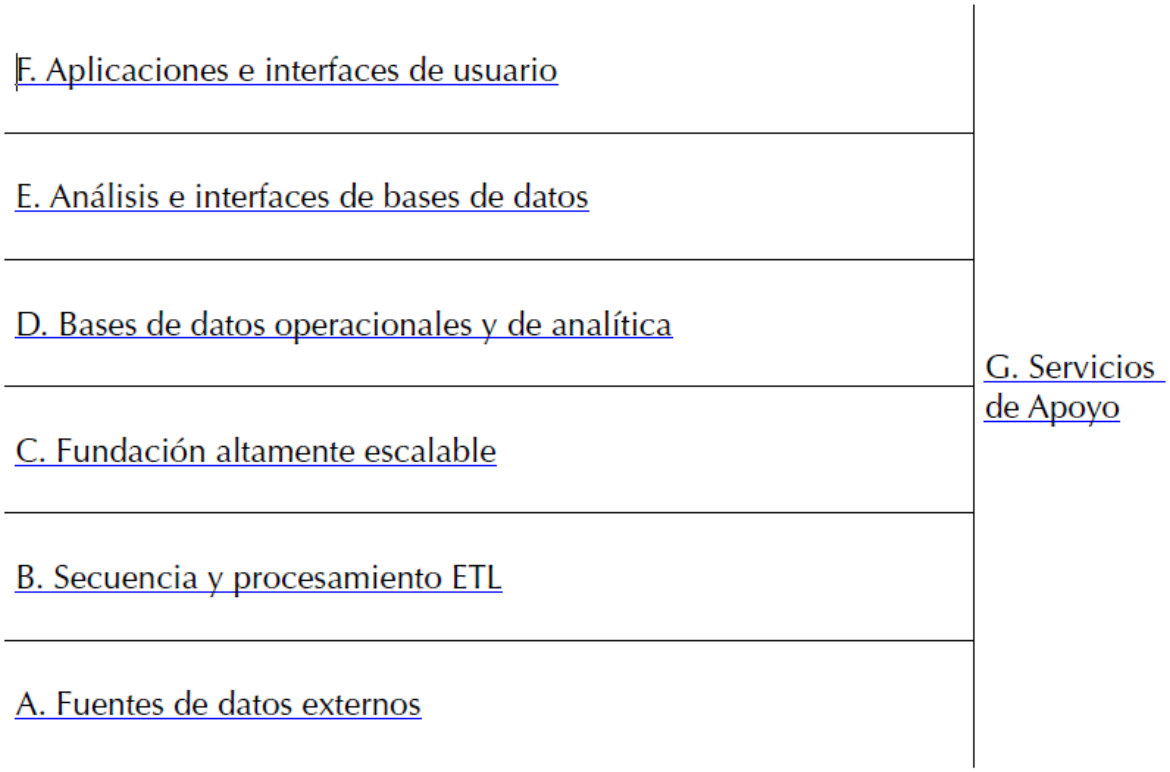


Figura 6. Arquitectura de Big Data de Marcus (Camargo Vega et al., 2015)

El potencial del Big Data reside en permitir que las organizaciones aprovechen toda su información para comprender, supervisar, sondear y planificar eficazmente sus procesos (Bibri, 2019), una de las propuestas para lograr este objetivo es alinear el Big Data y la computación en la nube, debido a las nuevas oportunidades que brinda la combinación de estos dos componentes para realizar inteligencia de negocios y análisis de datos, y a que sus características proporcionan capacidades computacionales para gestionar grandes volúmenes de datos (Dabbéchi et al., 2016).

1.1.4. Inteligencia de negocios (BI)

La inteligencia de negocios es un conjunto de aplicaciones, tecnologías y procesos para recopilar, almacenar, acceder y analizar datos, para ayudar a los usuarios empresariales a tomar decisiones (Schlesinger & Rahman, 2016). Para lograr esto

es necesario que los usuarios cuenten con herramientas que les faciliten el autoservicio, dentro de las plataformas de inteligencia de negocio, para que ellos puedan manipular los datos y obtener respuestas a sus preguntas, sin hacer requerimientos a las áreas de informática o de sistemas de las compañías.

El autoservicio se está convirtiendo en la norma en datos y análisis, cada vez más usuarios empresariales exigen acceso a los datos para obtener sus propios conocimientos e impulsar iniciativas (Clarke et al., 2016).

Dar un manejo correcto de los datos es un propósito que debe ser tenido en cuenta dentro del procesamiento de los datos, por esta razón para comprender y promover el valor de los activos de datos se define el gobierno de datos dividiéndolo en 5 pilares: principios de datos, calidad de datos, metadatos, acceso a datos y ciclo de vida de datos. La gobernanza de datos se refiere a quién tiene los derechos de decisión y se hace responsable de la toma de decisiones de una organización sobre sus activos de datos (Khatri & Brown, 2010).

1.1.5. Datos abiertos

Un aspecto por considerar dentro de la arquitectura de plataformas de servicios de datos es el de datos abiertos. Un dato o contenido está abierto si alguien es libre de usarlo, reutilizarlo y compartirlo, sujeto solo al requisito de hacer énfasis en la fuente original. Aplicado a los datos, requiere que un conjunto de datos sea accesible sin costo y sin restricciones técnicas que eviten su uso (Europe & Foundation, 2011).

1.1.6. DataOps

Con el propósito de obtener calidad y reducir los tiempos en los ciclos de análisis de datos en la inteligencia de negocios se define DataOps. Esta metodología es un conjunto de prácticas, procesos y tecnologías que se ajustan al flujo continuo

de trabajo de la inteligencia de negocios para reaccionar a requisitos imprevistos de una solución, y así permitir el despliegue e integración continua (Ereth, 2018).

1.2. Objetivos de la investigación

Para formular la arquitectura de referencia se proponen los siguientes objetivos:

Objetivo General:
Proponer una arquitectura de referencia para una plataforma de servicios de datos que permita a los usuarios finales integrar, clasificar y analizar los datos para generar información que les facilite la resolución de problemas y predecir tendencias futuras.

Objetivos específicos: para alcanzar el objetivo general, se plantean los siguientes objetivos específicos:

1. Identificar características comunes de las plataformas de referencia para arquitecturas de analítica de datos, que deban ser tenidas en cuenta dentro de una arquitectura de referencia.
2. Establecer el proceso a seguir dentro de una arquitectura para una plataforma de servicios de datos.
3. Proponer los lineamientos para el diseño de una arquitectura de referencia para una plataforma de servicios de datos.
4. Diseñar la arquitectura de referencia para plataformas de servicios de datos enfatizando aspectos de auto servicio y datos disponibles o abiertos a los usuarios.
5. Implementar la arquitectura de referencia propuesta para evaluar los lineamientos y procesos establecidos.

Para lograr el desarrollo de los objetivos se ha tomado como referencia la metodología *Design Science in Information Systems Research* (Hevner et al., 2008), la cual tiene como principio que el conocimiento y comprensión de un problema y su solución, se adquieren en la aplicación y construcción de un artefacto. Este principio será aplicado en el diseño de la arquitectura de referencia para una plataforma de servicio de datos y el diseño inicial será refinado en cada avance del proceso de investigación. En la Figura 7 se ilustra el proceso de investigación, mostrando las diferentes fases adelantadas y el principal resultado de cada una.

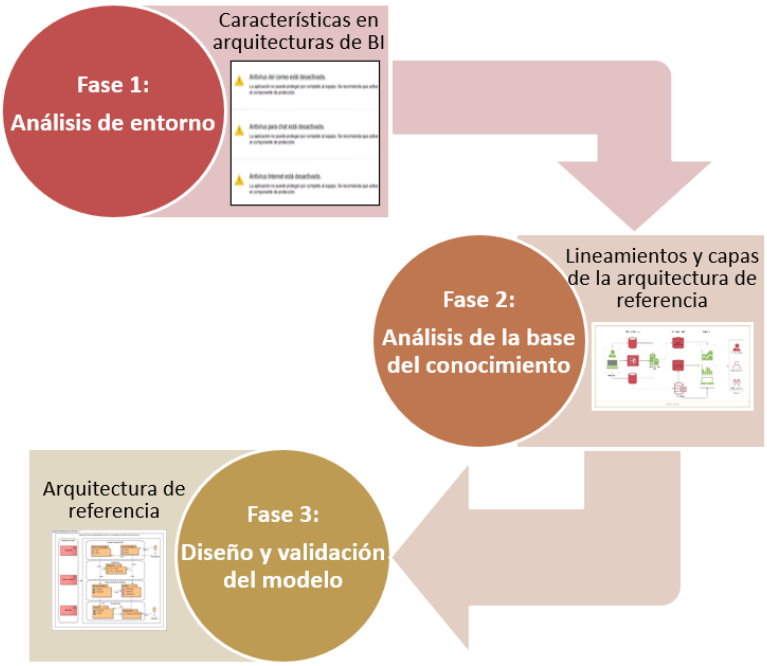


Figura 7. Proceso de investigación para alcanzar los objetivos

A continuación, se describen brevemente las fases a realizar con el principal resultado de cada una:

Fase 1 – Análisis del entorno: Durante esta fase se realiza la revisión de bibliografía para identificar las características comunes de las arquitecturas de referencia actuales para Big Data.

Fase 2 – Análisis de la base del conocimiento: En esta fase se definen las capas y los lineamientos de la arquitectura de referencia para plataformas de servicios de datos.

Fase 3 – Diseño y validación del modelo: En esta fase se diseña la arquitectura de referencia y se hace su implementación en una organización para realizar la evaluación.

Este documento consigna el trabajo adelantado para cubrir estas tres fases, con una presentación estructurada en partes para dar mayor claridad con respecto a su contenido.

1.3. Estructura de la tesis

La realización de este trabajo deja muchos testimonios y experiencias que necesitan ser documentadas de forma organizada para facilitar su estudio y comprensión. Por tal motivo este trabajo está organizado en 6 capítulos agrupados en 4 partes como se ilustra en la Figura 8.

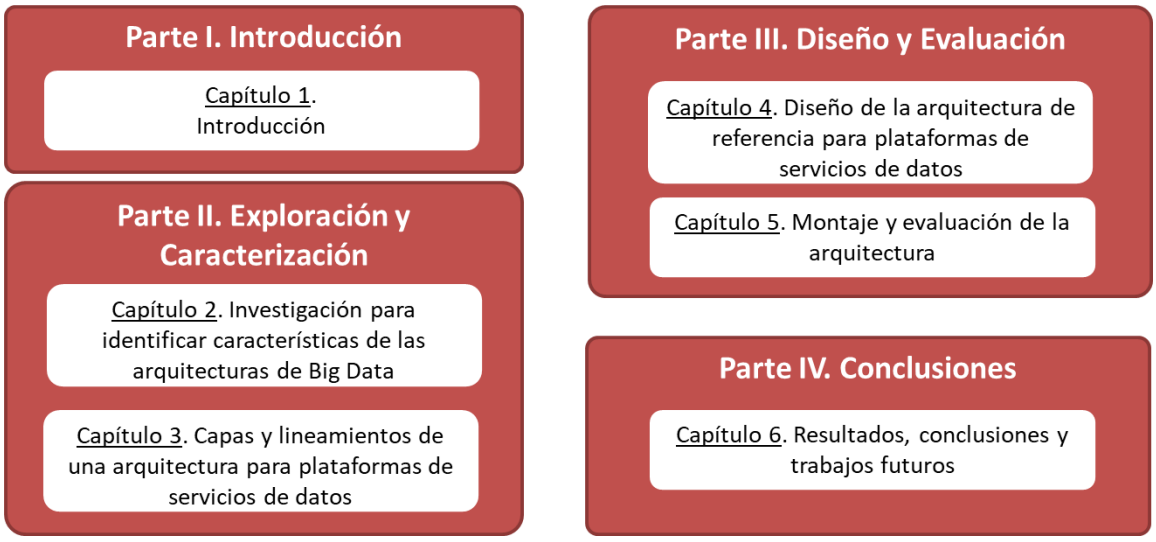


Figura 8. Estructura de la tesis

A continuación, se describen brevemente los capítulos de este trabajo con su respectivo contenido:

PARTE I - INTRODUCCIÓN

CAPÍTULO 1. Introducción: pretende dar los fundamentos conceptuales de este trabajo, presentando su campo de acción y explicando el proceso investigativo que sigue. Contiene una breve descripción de arquitecturas para Big Data, los objetivos y la estructura de la tesis.

PARTE II – EXPLORACIÓN Y CARACTERIZACIÓN

CAPÍTULO 2. Investigación para identificar características de las arquitecturas de Big Data: presenta la revisión de literatura sobre arquitecturas de Big Data. Las investigaciones se agrupan en tres categorías: aplicación de arquitecturas en un contexto especifico, propuestas de combinación de arquitecturas y propuestas de nuevas arquitecturas. Adicionalmente muestra una síntesis de estas investigaciones con las capas que cada una cubre.

CAPÍTULO 3. Capas y lineamientos de una arquitectura para plataformas de servicios de datos: se definen las capas y el proceso que se sigue para establecer la arquitectura de referencia para una plataforma de servicios de datos. Contiene la explicación del proceso colaborativo que se sigue en los grupos focales para el diseño de la arquitectura de referencia.

PARTE III – DISEÑO Y EVALUACIÓN

CAPÍTULO 4. Diseño de la arquitectura de referencia para plataformas de servicios de datos: con base en las investigaciones realizadas y en los resultados de los grupos focales, se diseña la arquitectura de referencia para plataformas de servicios de datos haciendo énfasis en el auto servicio y en los datos abiertos.

CAPÍTULO 5. Montaje y evaluación de la arquitectura: para realizar la validación de la arquitectura se selecciona una empresa y se realiza el montaje de acuerdo con las características de la organización. Contiene una evaluación que realiza el equipo humano de la organización.

PARTE IV. CONCLUSIONES

CAPÍTULO 6. Resultados, conclusiones y trabajos futuros: inicia con un análisis del cumplimiento de los objetivos específicos que se plantearon en la investigación. Muestra los aportes de este trabajo y sus beneficiarios. Concluye con las principales consideraciones de una arquitectura de referencia para una plataforma de servicios de datos; adicionalmente plantea los trabajos futuros.

PARTE II
EXPLORACIÓN Y CARACTERIZACIÓN

"Lo que cuenta en la vida no es el mero hecho de que hayamos vivido; es la diferencia que hemos hecho en la vida de los demás lo que determinará el significado de la vida que llevamos" – Nelson Mandela

CAPÍTULO 2

Investigación para identificar las características de las arquitecturas de Big Data

El proceso de revisión de literatura se planificó a partir de la definición de las cadenas de búsqueda “Data services AND architecture” y “Big Data AND architecture”. La búsqueda se realizó en las siguientes bases de datos:

- Scopus
- Science Direct
- Springer Journals
- IEEE Xplore

La revisión se realizó ejecutando las cadenas de búsqueda en cada una de las bases de datos mencionadas, luego se hicieron filtros para tener resultados desde el año 2014, con temas relacionados a *”Computer Science”* y *”Engineering*”, con documentos de tipo artículo, conferencia o libros y que su estado fuese publicado. Los artículos fueron seleccionados leyendo inicialmente el *abstract,* al encontrar que tenían relación con el propósito del trabajo se procedió a la lectura de la introducción y de las conclusiones, de esta manera se eligieron los documentos que son base de este trabajo.

A continuación, se presenta una síntesis de los estudios previos priorizados, los cuales están directamente relacionados con el objeto de estudio de este proyecto y sus fundamentos teóricos ayudan a dar sustento a este trabajo. Se organizan en

tres categorías: aplicación de arquitecturas en un contexto especifico, propuestas de combinación de arquitecturas y propuestas de nuevas arquitecturas.

2.1. Aplicación de arquitecturas en un contexto especifico

En el campo de la atención médica, Wang y otros (Wang et al., 2018) desarrolló una arquitectura de Big Data, la cual se construyó con base en las experiencias de implementación de sistemas de Big Data en la industria, y se compuso de cinco capas: primero, la capa de datos, que incluye las fuentes de datos que se utilizarán para apoyar las operaciones y la resolución de problemas; segundo, la capa de agregación de datos, que se encarga de adquirir, transformar y almacenar datos; tercero, la capa analítica, que se encarga de procesar y analizar datos; cuarto, la capa de exploración de información, que funciona generando resultados para el apoyo a la decisión clínica. Por último, la capa de gobierno de datos, que se encarga de administrar los datos a lo largo de todo su ciclo de vida mediante la aplicación de las normas y políticas adecuadas de seguridad y privacidad. Esta capa es particularmente necesaria en este caso dada la sensibilidad de los datos clínicos. Esta arquitectura solo considera los usuarios con experiencia en analítica, dejando de lado aquellos que pueden requerir un mayor procesamiento de datos a la hora de visualizar la información. Por otra parte, solo son tenidas en cuenta las fuentes de datos estructuradas, quedando por fuera aquellas fuentes no estructuradas que son comunes en la gestión de las organizaciones. En la Figura 9 se ilustran cada una de las capas de esta arquitectura.

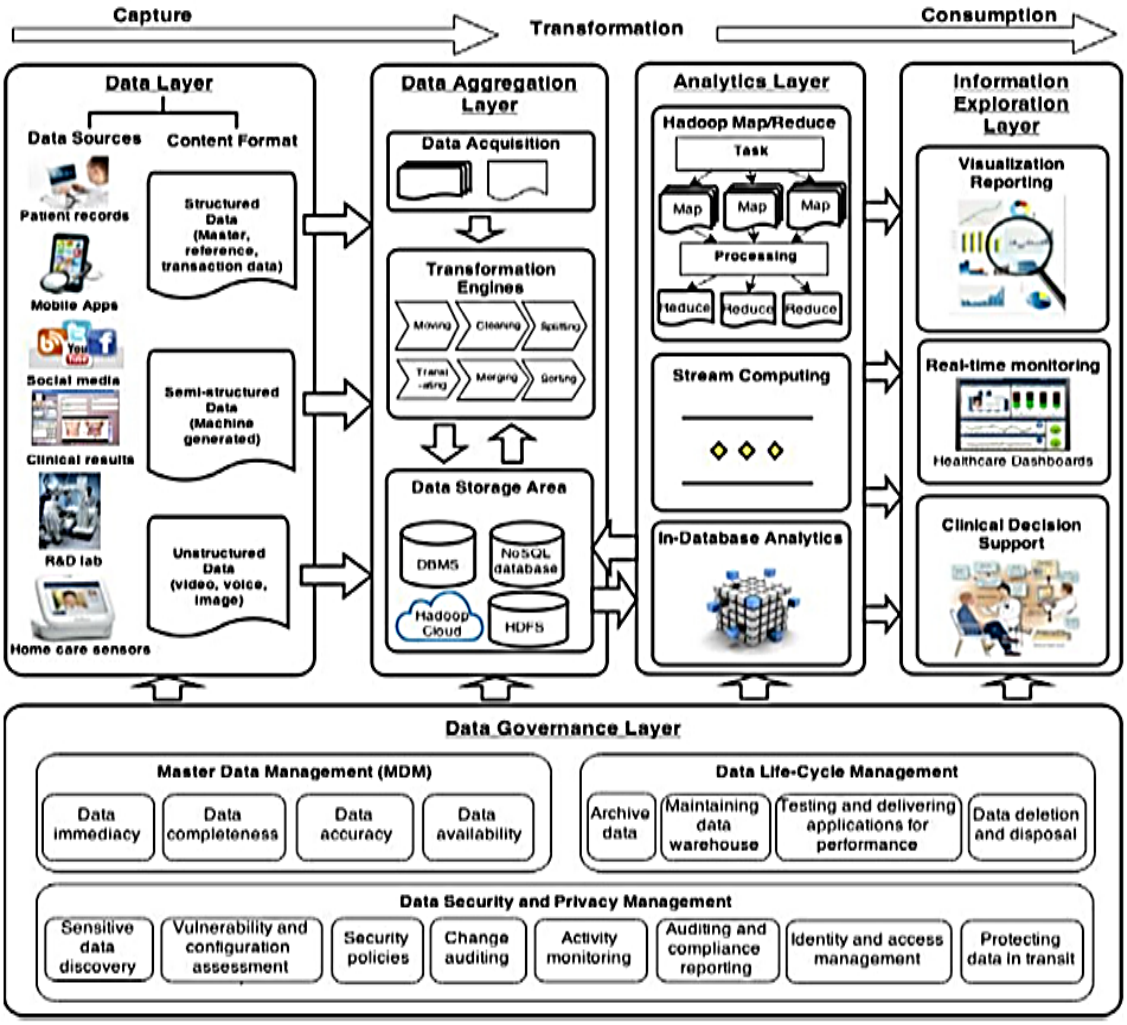


Figura 9. Arquitectura de Big Data para el campo de la salud (Wang et al., 2018)

La arquitectura de Big Data, en tiempo real, establecida e implementada por Gonçalves y otros (Gonçalves et al., 2017) en la unidad de cuidados intensivos del Centro Hospitalario de Porto, en Portugal, genera dependencia de las herramientas tecnológicas utilizadas, ya que la solución está basada solo en productos de código abierto, dejando de lado una gran variedad de herramientas que no tienen esta característica y que pueden ser útiles en otros escenarios. Por otra parte, la arquitectura solo puede ser implementada en unidades de cuidados intensivos que tengan especificaciones similares. Una arquitectura de referencia

debe ser útil para empresas de diferentes campos y no generar dependencia por el tipo de herramientas tecnológicas que se utilizan. En la Figura 10 se ilustra esta arquitectura.

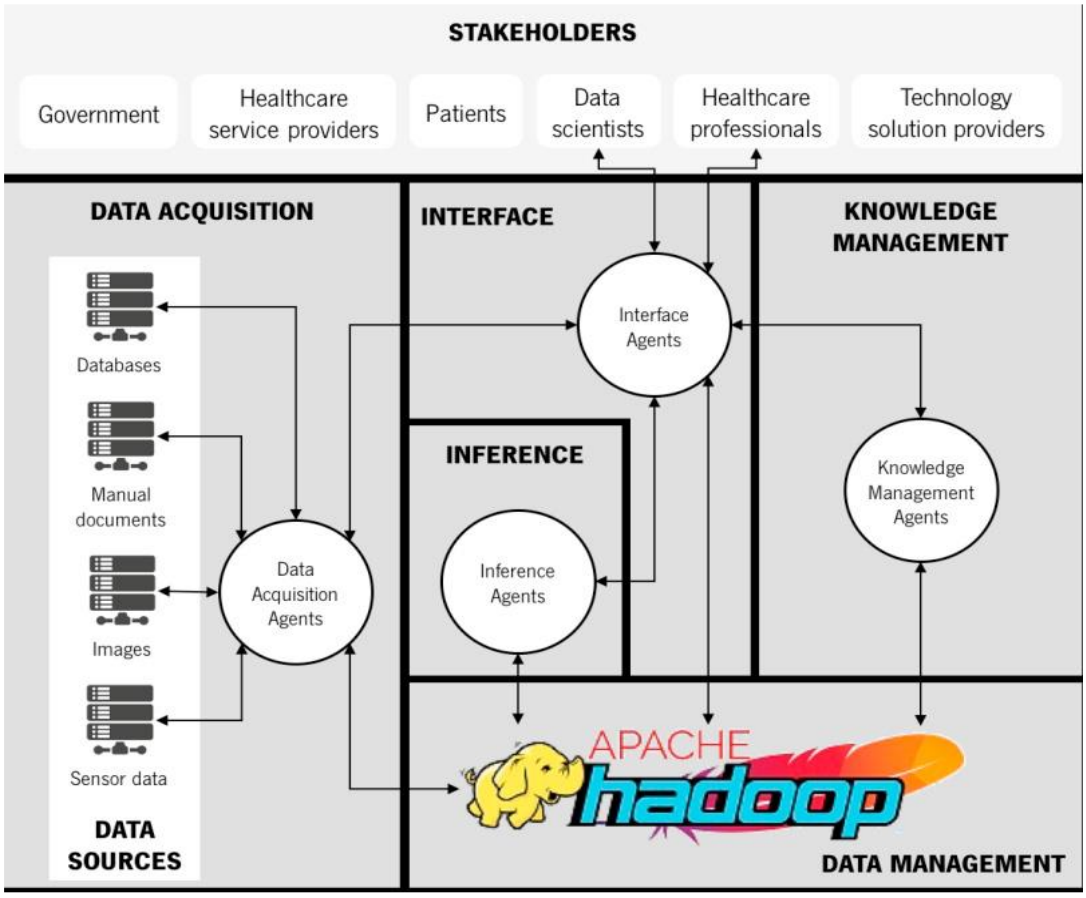


Figura 10. Arquitectura de Big data para la unidad de cuidados intensivos del Centro Hospitalario de Porto (Gonçalves et al., 2017)

La arquitectura de Big Data propuesta por Blazquez y Domenech (Blazquez & Domenech, 2018) explica las particularidades de los análisis de comportamiento económico y social en la era digital, basada en el enfoque del ciclo de vida de los datos. La primera particularidad está relacionada con la variedad de fuentes que podrían proporcionar información sobre temas económicos y sociales, en este punto proponen una taxonomía para clasificar las fuentes de acuerdo con el propósito del generador de datos. La segunda contribución está relacionada con

los métodos para procesar la información y permitir la gestión en una arquitectura robusta y flexible. En la Figura 11 se ilustra esta arquitectura.

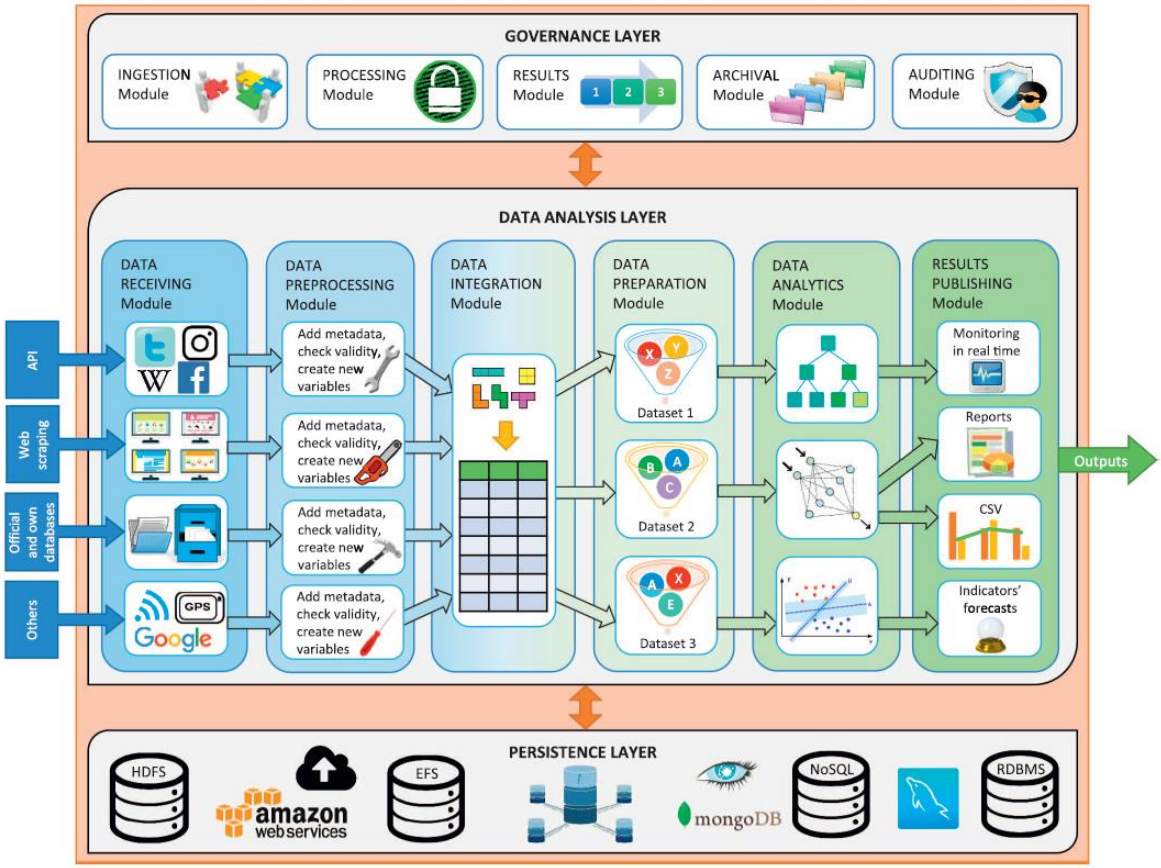


Figura 11. Arquitectura de Big Data para pronosticar cambios sociales y económicos (Blazquez & Domenech, 2018)

Las dificultades se presentan al integrar la arquitectura con los sistemas de información existentes en las organizaciones y con su implementación en un entorno de computación en la nube. Este último aspecto es una dificultad para implementar una plataforma de servicio de datos, que es el objetivo principal de esta investigación.

2.2. Propuestas de combinación de arquitecturas

Zhelev y Rozeva (Zhelev & Rozeva, 2017) exponen que es necesario contar con un amplio conocimiento para escoger la arquitectura de datos cuándo se trata de temas de Big Data y profundiza en algunas arquitecturas para la gestión y manejo de grandes volúmenes de datos, enfocándose en las arquitecturas Lambda y Kappa. Ambas arquitecturas presentan dificultades en la implementación y el soporte, debido a que mantener el procesamiento por lotes y en tiempo real se hace complejo. En la arquitectura de referencia, la organización será la responsable de definir el tipo de procesamiento que se debe realizar sobre los datos, según las necesidades. En la Figura 12 se ilustra esta arquitectura.

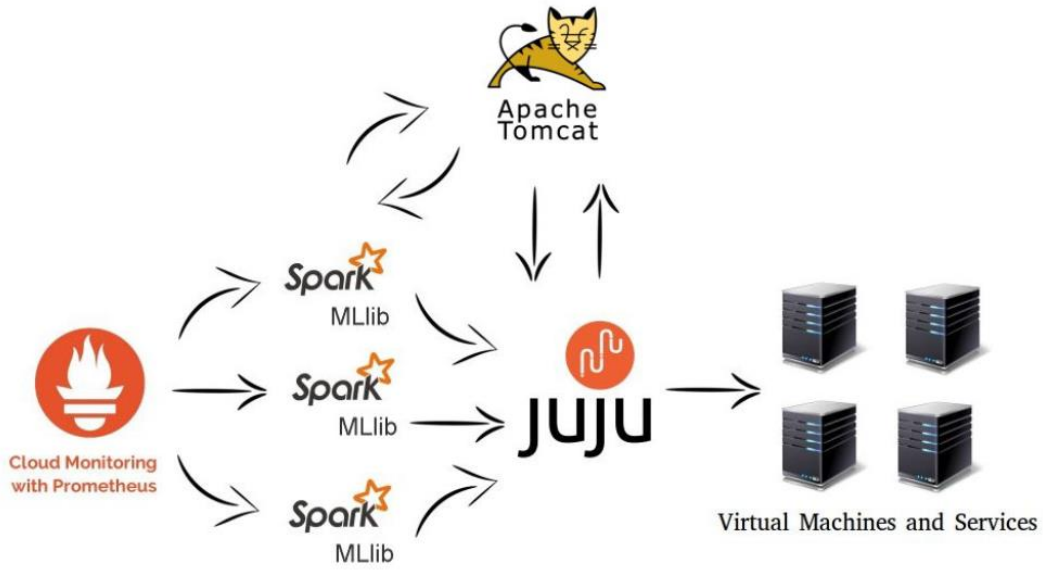


Figura 12. Combinación de arquitecturas Lambda y Kappa (Zhelev & Rozeva, 2017)

Camargo y otros (Camargo Vega et al., 2015) revisan tres propuestas de arquitectura de Big Data y proponen una que reúne características de cada modelo. La arquitectura propuesta contiene las siguientes etapas: recolección de datos, carga de datos, transformación de datos, extracción de datos y un aspecto de seguridad. En la etapa de carga de datos, la arquitectura que se propone

menciona el uso de Hadoop, lo que genera dependencia en la solución implementada, por otra parte, hace falta una etapa para el gobierno de datos, que no solo es seguridad, se deben mencionar aspectos como calidad del dato, datos maestros y metadatos, como lo sugieren Khatri y Brown (Khatri & Brown, 2010). Dentro de los lineamientos a establecer en el presente trabajo de grado se tendrá en cuenta el tema de gobierno de datos abarcando la seguridad, los datos maestros y los metadatos.

En un estudio realizado por Oussous y otros (Oussous et al., 2018), se analizan las tecnologías recientes desarrolladas para Big Data, el objetivo es ayudar a seleccionar y adoptar la combinación correcta de diferentes tecnologías, de acuerdo a las necesidades tecnológicas y a los requisitos de las aplicaciones, proporcionando una visión detallada de la arquitectura. Un resultado del estudio es que encuentran que existen muchas deficiencias en las herramientas tecnológicas, en la mayoría de los casos relacionadas con la arquitectura y técnicas adoptadas. Atender esta problemática es un objetivo del presente trabajo, a través de la definición de una arquitectura de referencia para plataformas de servicios de datos, estableciendo bases para que una organización pueda definir su arquitectura de Big Data.

2.3. Propuestas de nuevas arquitecturas

Passlick y otros (Passlick et al., 2017) describen un proceso llamado inteligencia del auto servicio para hacer inteligencia de negocio con grandes volúmenes de datos, el objetivo es que los interesados sean quienes hagan sus informes y los puedan analizar. Se resaltan dos componentes dentro del diseño planteado, las salas de colaboración y una base de datos de conocimiento para el autoaprendizaje, ya que hacen que el conocimiento implícito de los usuarios sea utilizable. La base de datos de conocimiento utiliza un algoritmo de autoaprendizaje, el cual se ve afectado si no se le suministra un gran volumen de datos, en este caso se aplicará una base de datos de conocimiento simple sin el

algoritmo; una alternativa a esto podría ser una base de datos de conocimiento basada en la nube, pero una dificultad que se presenta con esta arquitectura es el problema para trasladar los componentes a la nube. Además, dicha arquitectura no ha sido probada en un ambiente empresarial, fue establecida con base en los conceptos de expertos. En la Figura 13 se ilustra esta arquitectura.

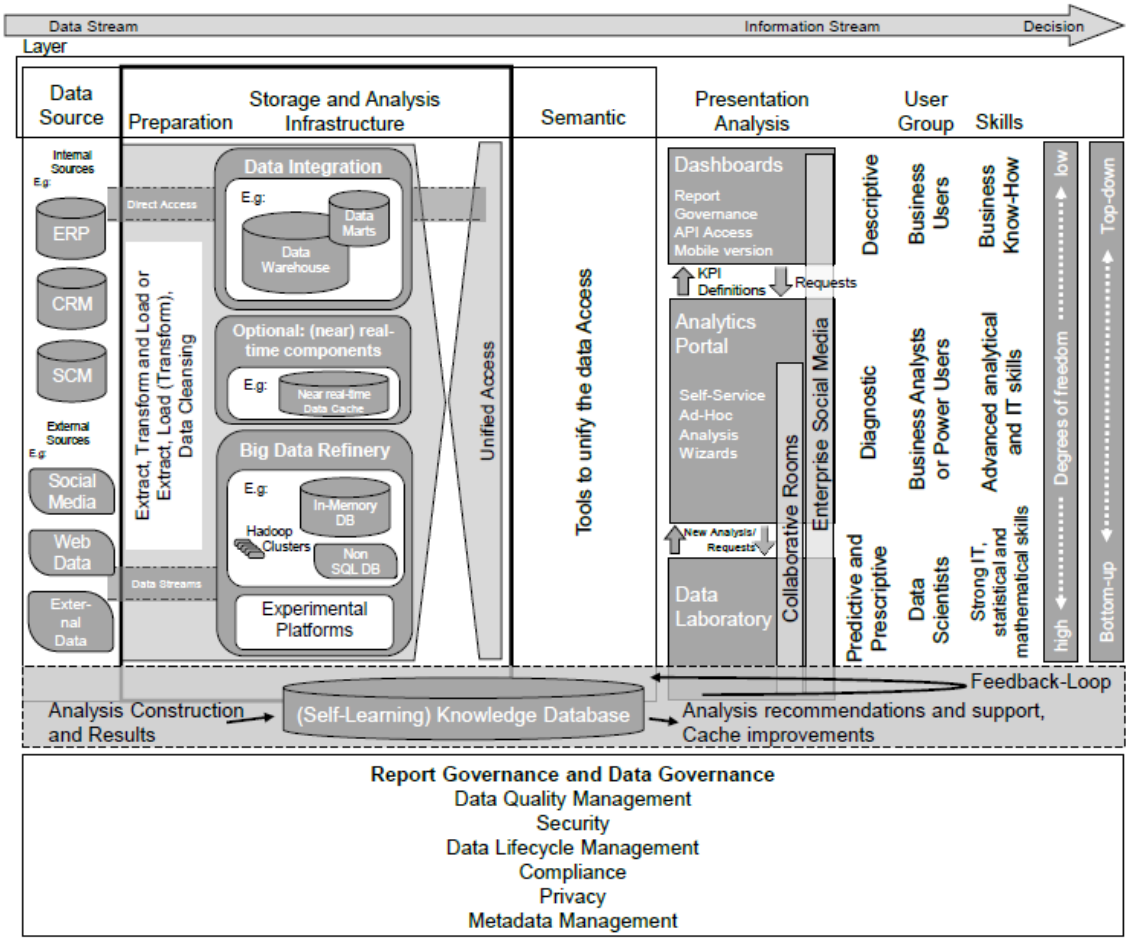


Figura 13. Arquitectura de Big Data propuesta por Passlick *et al.*(Passlick et al., 2017)

Pääkkönen y Pakkala (Pääkkönen & Pakkala, 2015) describen una arquitectura de referencia para sistemas de Big Data basada en el análisis de algunos casos de implementación en la industria. En este trabajo se describen las siguientes funcionalidades: fuentes de datos, extracción de datos, carga y preprocesamiento

de datos, análisis de datos, transformación de datos, interfaz y visualización. La arquitectura descrita en este trabajo no ha sido evaluada en un caso real con grandes volúmenes de datos, fue creada con base en algunos casos implementados en los cuales solo se utilizaron algunas herramientas tecnológicas, lo que puede restringir su implementación. Por otra parte, no se considera una capa de gobierno de datos para gestionar la seguridad y los metadatos. En la Figura 14 se ilustra esta arquitectura.

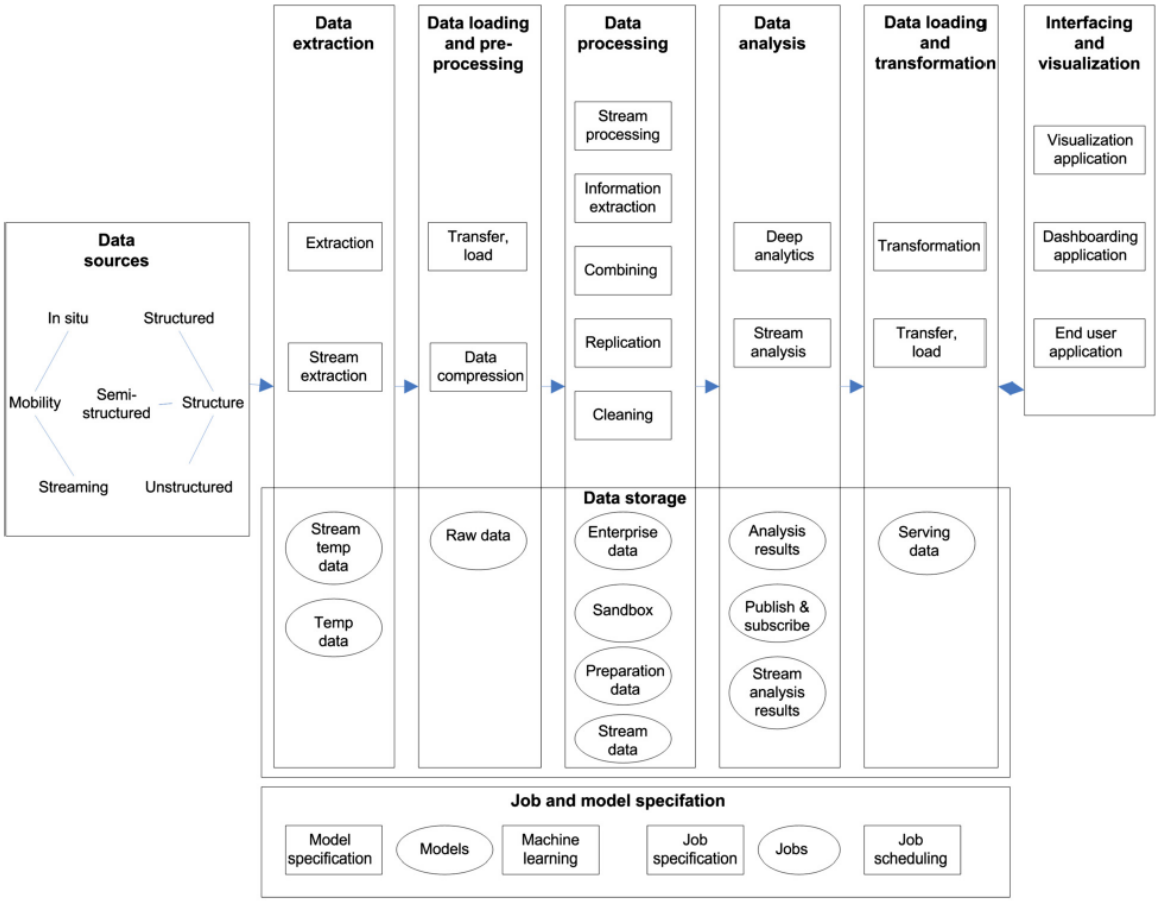


Figura 14. Arquitectura de Big Data propuesta por Pääkkönen *et al*. (Pääkkönen & Pakkala, 2015)

Assunção y otros (Assunção et al., 2015) identificaron los componentes que deberían estar presentes en cualquier arquitectura de Big Data al describir las

cuatro fases más comunes dentro de un flujo de trabajo de análisis de Big Data: fuentes de datos, gestión de datos (incluidas tareas de preprocesamiento), modelado y análisis de resultados y visualización. Esta arquitectura se propuso para ser implementada en un ambiente de computación en la nube; los beneficios que ofrece este componente para almacenar grandes cantidades de datos y realizar cálculos potentes lo están posicionando como una tecnología deseable para ser incluida en el diseño de una arquitectura de Big Data. Esta arquitectura resalta la importancia de la nube para el procesamiento de grandes volúmenes de datos, componente en el cual se ejecutarán los servicios de datos del presente trabajo y en el cual se podrán ejecutar las tareas de análisis y visualización. Además, dentro del trabajo futuro se deja descrito el establecimiento de estándares o lineamientos para establecer la arquitectura en una organización, los cuales serán definidos por la arquitectura de referencia. En la Figura 15 se ilustran las capas de esta arquitectura.

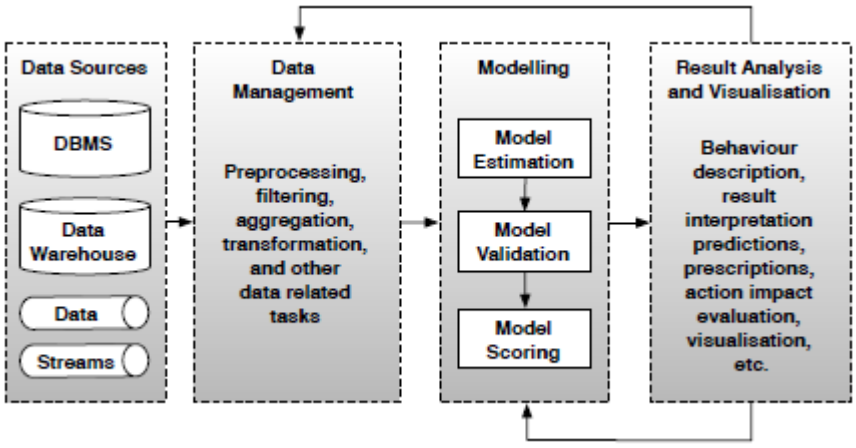


Figura 15. Arquitectura general de Big Data propuesta por Assunção *et al.*(Assunção et al., 2015)

2.4. Síntesis de estudios primarios

La revisión de los trabajos mencionados en los tres numerales anteriores permitió identificar e integrar las diferentes capas de un ciclo de vida de los datos. El

objetivo de las capas que se establecen es estandarizar el ciclo de vida de los datos y servir como marco de referencia cuando se debe diseñar una arquitectura de gestión de Big Data en una organización. Los resultados encontrados en cada uno de los estudios de referencia se pueden integrar en las siguientes capas:

- Fuentes de datos
- Carga de datos
- Transformación y almacenamiento
- Analítica
- Visualización
- Gobierno de datos

A continuación, se presenta la Tabla 1 con la síntesis de los estudios primarios y las capas de una arquitectura de Big Data que cada investigación ha contemplado.

Tabla 1. Síntesis de trabajos analizados

Investigación \ Capas	Fuentes de datos	Carga de datos	Transformación y almacenamiento	Analítica	Visualización	Gobierno de datos
Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations (Wang et al., 2018)	X		X	X	X	X
Towards of a Real-time Big Data Architecture to Intensive Care (Gonçalves et al., 2017)	X		X	X	X	
Big Data sources and methods for social and economic analyses (Blazquez & Domenech, 2018)	X	X	X	X	X	X
Big data processing in the cloud - Challenges and platforms (Zhelev & Rozeva, 2017)	X	X	X		X	

Arquitectura tecnológica para Big Data (Camargo Vega et al., 2015)	X	X	X	X	X	
Big Data technologies: A survey (Oussous et al., 2018)	X		X	X	X	
A Self-Service Supporting Business Intelligence and Big Data Analytics Architecture (Passlick et al., 2017)	X	X	X	X	X	X
Reference Architecture and Classification of Technologies, Products and Services for Big Data Systems (Pääkkönen & Pakkala, 2015)	X	X	X	X	X	
Big Data computing and clouds: Trends and future directions (Assunção et al., 2015)	X		X	X	X	

Una breve descripción de cada una de las capas permitirá un entendimiento global del ciclo de vida de los datos.

Fuentes de datos: Esta capa permite acceder a los orígenes o fuentes de los datos y recolectarlos. Las fuentes pueden ser: bases de datos internas o externas, documentos, imágenes u otros (Gonçalves et al., 2017).

Carga de datos: En este punto se cargan los datos y se debe aplicar el concepto de metadatos, es decir, se deben describir las características de los datos que se cargan, como el nombre, la importancia y la relación con otros datos u objetos de la empresa. Uno de los objetivos de esta etapa es obtener una estructura homogénea que ayude a que los datos sean trazables y fáciles de acceder (Camargo Vega et al., 2015).

Transformación y almacenamiento: Durante esta capa los datos se transforman mediante la aplicación de reglas de negocio y el procesamiento de los datos y luego se almacenan en el sitio dispuesto por la organización (Wang et al., 2018).

Analítica: Esta capa es responsable de procesar todo tipo de datos y realizar los análisis requeridos por la organización, el cual puede ser descriptivo, predictivo o prescriptivo (Assunção et al., 2015).

Visualización: En este punto se generan los resultados con informes de visualización y monitoreo de información. La generación de informes y tableros de control es una función crítica en las arquitecturas de Big Data, ya que esta capa debe permitir visualizar información útil que respalde las operaciones diarias de la compañía y debe ayudar a los directivos a tomar mejores decisiones y más rápido (Wang et al., 2018).

Gobierno de datos: Esta capa es transversal a las anteriores y es la encargada de la gestión de datos maestros, de la administración del ciclo de vida de los datos y de la seguridad de los datos (Wang et al., 2018).

Una vez identificadas las capas que cubren el ciclo de vida de los datos en una arquitectura de Big Data se procede a realizar una validación de dichas capas con personas con conocimientos y experiencia en este tipo de arquitecturas.

CAPÍTULO 3

Capas y lineamientos de una arquitectura para plataformas de servicios de datos

La creación de la arquitectura de referencia tendrá en cuenta diferentes usuarios del área de Big Data, con los cuales se tendrán sesiones de trabajo para conocer las arquitecturas que actualmente están manejando en sus organizaciones y luego plasmar una arquitectura ideal de Big Data que tenga en cuenta las capas presentadas en el capítulo anterior. La elección de una técnica para abordar las sesiones de trabajo, que permita obtener nuevos requisitos y validar las características de la arquitectura de referencia es el primer paso para lograr un acercamiento con los potenciales usuarios, que a futuro podrán hacer uso de la arquitectura para una plataforma de servicio de datos.

El estudio determinó que la combinación de grupos focales con la creación de prototipos facilita la interacción y participación de los usuarios, de forma que ellos puedan expresar abiertamente sus opiniones y proponer nuevas ideas que ayuden a la creación de la arquitectura de referencia. Un grupo focal es un medio para obtener ideas y opiniones sobre un producto, servicio u oportunidad específicos en un entorno grupal interactivo. Los participantes, guiados por un moderador, comparten sus impresiones, preferencias y necesidades. La creación de prototipos se utiliza para obtener y validar las necesidades de las partes interesadas a través de un proceso iterativo que crea un modelo o diseño de requisitos.

El grupo focal y el diseño de prototipos están alineados al propósito de la metodología elegida para el proyecto, *Design Science in Information Systems Research* (Hevner et al., 2008), ya que hace posible una discusión libre sobre los problemas y requisitos para una arquitectura de una plataforma de servicios de datos.

3.1. Bases y lineamientos de la arquitectura

Basados en los estudios e investigaciones analizados se abstraen las siguientes capas para una arquitectura de referencia para plataformas de servicios de datos:

- Fuentes de datos
- Carga de datos
- Transformación y almacenamiento
- Analítica
- Visualización
- Gobierno de datos

En la capa de visualización se utilizará el portal o plataforma de servicios de datos para facilitar el auto servicio y el análisis de los usuarios. La capa de gobierno de datos será transversal a todas las demás y permitirá definir quien se hace responsable de la toma de decisiones de la organización y quien debe responder por los datos, esta capa tendrá en cuenta la seguridad, los datos maestros y la metadata.

A continuación, se presentan un conjunto de características o lineamientos que se deben tener en cuenta para definir la arquitectura de Big Data en una organización:

- Una de las tareas en una arquitectura de Big Data que más tiempo y trabajo requiere es la preparación de los datos para el análisis, ya que en ocasiones la infraestructura existente es llevada hasta sus límites, inclusive

si se pretende hacer el despliegue en un entorno de computación en la nube, por esto se deben elegir métodos eficientes para almacenar, transformar y recuperar los datos (Assunção et al., 2015).

- Es necesario considerar la velocidad con la que los datos deben ser procesados, ya que en algunos escenarios el procesamiento puede hacerse por lotes, pero en otros se requiere continuo y en tiempo real, ya que se requiere una acción inmediata al procesar los flujos de datos entrantes (Zhelev & Rozeva, 2017).

- Dentro de las opciones para el almacenamiento se puede considerar la nube, ya que actualmente los sistemas de archivos que allí se ofrecen proporcionan la robustez, escalabilidad y redundancia que se requiere para esta tarea, aunque se debe analizar la concurrencia y rendimiento que se necesita en la ejecución de esta actividad (Assunção et al., 2015).

- En la visualización se debe contar con herramientas que faciliten la navegación para la presentación de los datos y en caso de requerir visualizaciones en tiempo real se deben definir técnicas eficientes de procesamiento de datos (Blazquez & Domenech, 2018).

Una mención aparte merece la computación en la nube debido a las características y bondades que ofrece para soportar cada una de las capas y lineamientos de una arquitectura de Big Data. A continuación, se describen las particularidades de la computación en la nube que favorecen una arquitectura de Big Data.

3.2. Computación en la nube para análisis de Big Data

El término “Computación en la nube” se refiere a un modelo de computación en el que las capacidades habilitadas para las Tecnologías de la Información y la Comunicación (TIC) se entregan a través de Internet, en forma de servicios, para

que sean estandarizadas, escalables y flexibles (Bibri, 2019). Por lo tanto, la computación en la nube consta de varios componentes que se pueden aprovisionar rápidamente con un esfuerzo mínimo de administración.

Lo anterior, ha facilitado que las organizaciones empleen la computación en la nube para una gran variedad de servicios como son almacenamiento, bases de datos y procesamiento de grandes volúmenes de datos; los buenos resultados de estos servicios ubican la computación en la nube como una de las opciones para tener la infraestructura disponible para realizar el análisis de Big Data (Bibri, 2019).

Esto implica tener la plataforma como servicio de Big Data (PaaS) y la infraestructura como servicio (IaaS), es decir, plataformas de desarrollo, servidores, almacenamiento y procesamiento en la nube (Assunção et al., 2015).

3.2.1. Infraestructura como servicio (IaaS) para Big Data

Los servicios de infraestructura de Big Data que se podrían considerar en la nube son: almacenamiento, acceso a datos y procesamiento (Bibri, 2019).

- **Servicios de almacenamiento:** proporcionan servicios para guardar datos en una infraestructura virtual, y permiten operaciones para crear, eliminar, modificar y actualizar datos, en un modelo que admite varios tipos de datos.

- **Servicios de acceso de datos:** permiten la gestión de recursos para todo tipo de datos, tales como selección, transformación de consultas, agregación y representación de resultados de consultas.

- **Servicios de procesamiento:** herramientas para acceder a los datos de interés, para transferir al sitio de procesamiento y técnicas para manejar la variedad de formatos de datos.

Los elementos que hacen parte de los servicios de infraestructura de Big Data se describen a continuación:

Nubes informáticas: permiten el aprovisionamiento bajo demanda de recursos informáticos, que pueden expandirse o reducirse según los requisitos.

Nubes de almacenamiento: ofrecen gran volumen de almacenamiento, incluido el sistema de archivos, permiten almacenamiento en bloque y almacenamiento basado en objetos, el cual maneja los datos como un objeto. Las nubes de almacenamiento ofrecen la posibilidad de crear el sistema de archivos preferido y son elásticamente escalables (Assunção et al., 2015).

Nubes de datos: son similares a las nubes de almacenamiento, pero a diferencia de la entrega de espacio de almacenamiento ofrecen datos como servicio. Las nubes de datos ofrecen herramientas y técnicas para publicar los datos, etiquetarlos, descubrirlos y procesar los datos de interés. Las nubes de datos operan en datos específicos del dominio que aprovechan las nubes de almacenamiento para servir datos como un servicio (Bibri, 2019).

3.2.2. Plataforma como servicio (PaaS) para Big Data

Las plataformas como servicios ofrecen mecanismos de consulta para la recuperación de datos y modelos de programación para abordar problemas analíticos de Big Data (Bibri, 2019).

3.3. Aproximación a la arquitectura de referencia

El diseño de la arquitectura de referencia para plataformas de servicios de datos fue guiado a partir de un proceso colaborativo con tres grupos focales. Se realizó la identificación previa de los participantes para garantizar su conocimiento en Big Data y analítica y se dividieron en grupos de trabajo con características similares, según el tipo de industria de la que hacen parte.

Los participantes de cada uno de los grupos focales fueron los siguientes:

- Cuatro empleados de empresas de tecnología de la ciudad de Medellín con conocimiento en Big Data y analítica.

- 11 personas con conocimiento y experiencia en entornos de Big Data y analítica de Instituciones de Educación Superior de la ciudad de Medellín.

- Dos arquitectos de Big Data de empresas de servicios de la ciudad de Medellín.

En el desarrollo de cada uno de los grupos focales se ejecutaron las siguientes actividades:

- Presentación de alternativas para implementar arquitecturas de Big Data y analítica basados en los estudios realizados.

- Participación de los asistentes informando la arquitectura y herramientas tecnológicas utilizadas en las plataformas de Big Data y analítica que tienen en las empresas.

- Diseño de modelos de arquitecturas para plataformas de servicios de datos, teniendo en cuenta 4 capas:
    - Repositorios
    - Preparación de datos
    - Modelado y evaluación.
    - Visualización.

- Análisis de los modelos de arquitectura construidos e identificación de características comunes.

- Calificación de las características comunes encontradas, para priorizar necesidades.

Las características comunes identificadas por los participantes para una arquitectura de Big Data fueron las siguientes:

- **Auto servicio de reportes:** Permite que los usuarios generen sus propios reportes según la necesidad.

- **Servicio de "Datos abiertos":** Característica que permite que los datos sean accesibles sin costo y sin restricciones técnicas.

- **Gobierno de datos:** Es la gestión de datos de una organización, los procesos y objetivos para garantizar la disponibilidad, el uso y la seguridad de la información.

- **Facilidad de software libre:** Permite el uso de software libre dentro de la implementación de la arquitectura.

- **Clúster de datos:** Característica de la arquitectura en la que varios equipos de computo trabajan como uno solo para almacenar los datos.

- **Auto servicios para preparación de datos:** Característica que facilita herramientas técnicas dentro de la arquitectura para la preparación de datos.

- **Datos no estructurados:** La arquitectura debe permitir la gestión de datos no estructurados.

- **Replicación:** La arquitectura debe permitir la replicación o copia constante de los datos para tener disponibilidad en caso de cualquier eventualidad.

Con el fin de calificar estas características, según la relevancia dentro de la arquitectura, se define una escala de medición de 1 a 5, siendo 1 la calificación más baja o de menos relevancia y 5 la más alta o la más relevante.

La calificación asignada por cada uno de los asistentes a los grupos focales arrojó el resultado que se plasma en la Tabla 2.

Tabla 2. Calificación de características para una plataforma de servicio de datos

Criterio / Calificación	Alto	Med-Alto	Medio	Med-Bajo	Bajo	Ponderación
	5	**4**	**3**	**2**	**1**	
Auto servicio de reportes	10	1	0	0	0	**54**
Servicios de "Datos abiertos"	10	1	0	0	0	**54**
Gobierno de datos	10	1	0	0	0	**54**
Factibilidad de software libre	5	3	1	0	0	**40**
Cluster de datos	0	5	6	0	0	**38**
Auto servicios para preparación de datos	2	3	3	3	0	**37**
Datos no estructurados	0	2	6	2	0	**30**
Replicación	0	1	2	1	7	**19**

Como se puede observar en la tabla anterior, el auto servicio, los datos abiertos y el gobierno de datos obtuvieron la mayor ponderación, evidenciando que son características con las cuales debe contar una arquitectura de Big Data.

Con el grupo de arquitectos de Big Data se hizo énfasis en las capas que debe tener una arquitectura de Big Data y analítica y sus características principales, como resultado se construyó la arquitectura que se ilustra en la Figura 16.

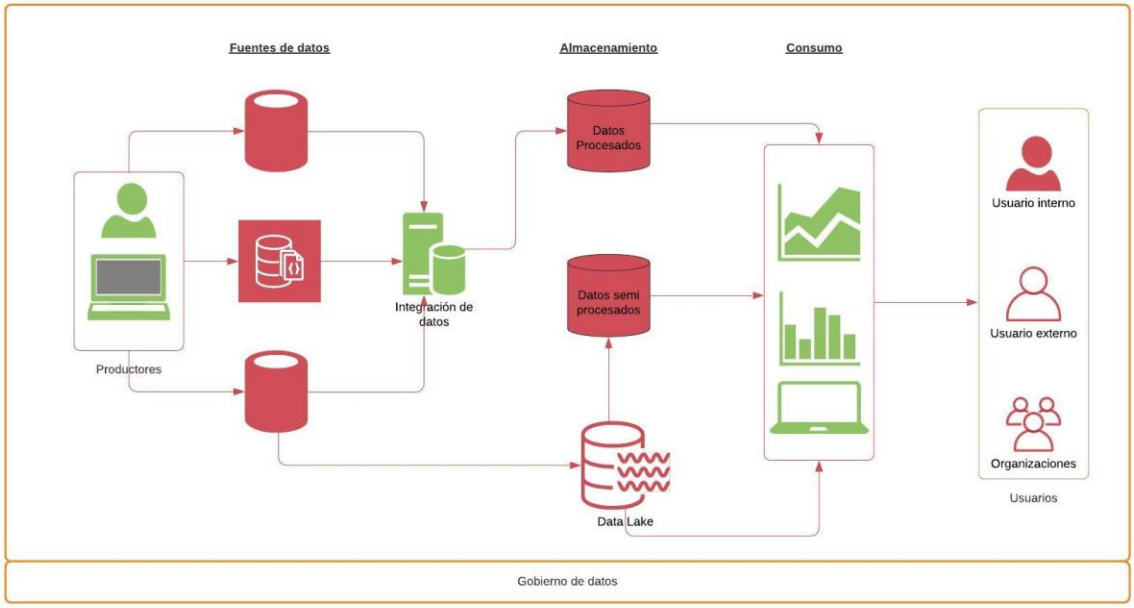


Figura 16. Arquitectura propuesta por arquitectos de Big Data.

Las principales conclusiones de las sesiones de trabajo que se deben tener en cuenta en una arquitectura de referencia de una plataforma de servicios de datos son:

- Es indispensable tener un gobierno de datos antes de que las organizaciones definan la arquitectura para la plataforma de servicio de datos, con el fin de tener control sobre la información que puede gestionar cada usuario.

- La arquitectura de referencia debe ser escalable y flexible para permitir la auto gestión de los usuarios finales.

- Se deben definir los metadatos, como parte del gobierno, para identificar y estandarizar los nombres de cada uno de los campos que pueden tratar y visualizar los usuarios.

PARTE III
DISEÑO Y EVALUACIÓN

"La verdad es como un león; no tienes que defenderlo. Deja que se pierda; se defenderá a sí mismo" – San Agustín

CAPÍTULO 4

Diseño de la arquitectura de referencia para plataformas de servicios de datos

4.1. Introducción

A continuación, se explica el diseño final desarrollado, el cual tiene sus bases en la revisión de literatura y en los hallazgos de los grupos focales. Inspirado en los modelos de arquitectura de análisis de Big Data consultados, el objetivo es describir todo el proceso desde las fuentes de datos hasta la presentación de la información.

En la primera parte de la arquitectura están las fuentes de datos. Las fuentes de datos se separan en fuentes no estructuradas y estructuradas. Los que alimentan estas fuentes de datos son los productores, los cuales pueden ser orígenes internos o externos, algunos ejemplos de estos orígenes son: los sistemas ERP o CRM de la organización, las redes sociales, diferentes tipos de archivos, entre otros.

El siguiente paso en el procesamiento de datos es la integración o preparación de los datos. Se presentan dos formas de realizar esta preparación, por medio de la extracción, transformación y carga de los datos, más conocida como ETL (*Extract, Transform and Load*), por sus siglas en inglés, la ETL en batch y la ETL en tiempo real, este último es utilizado para realizar un acceso directo a los datos y hacer análisis en tiempo real. Dentro del proceso ETL se presentan situaciones en las cuales la transformación no es necesaria, ya que existen usuarios que utilizan los datos sin procesar y para ellos esta tarea se puede omitir(Passlick et al., 2017),

esto es especialmente importante para los científicos de datos, que necesitan acceso a datos sin procesar.

Luego del procesamiento de los datos, en el modelo propuesto se presenta la capa de almacenamiento y analítica, en el componente de almacenamiento se puede utilizar una bodega de datos o un *data lake*, este componente es indispensable en una arquitectura de Big Data para lograr realizar los análisis requeridos; posteriormente se pasa a la analítica, la cual aplica métodos estadísticos y de aprendizaje automático para extraer conocimiento y hacer predicciones a partir de los datos preparados en las fases anteriores, para lograrlo se aplican diferentes técnicas descriptivas, predictivas y prescriptivas. Antes de utilizar alguno de los modelos, deben hacerse validaciones con un conjunto de datos diferente al que será utilizado en la estimación o el entrenamiento, esta validación proporciona una estimación de la robustez de los modelos y la calidad de las predicciones, de modo que se puede tener en cuenta el riesgo de una predicción inexacta (Blazquez & Domenech, 2018). Los modelos se pueden programar para ser entrenados continuamente con nuevos datos o se puede realizar el entrenamiento bajo demanda del usuario, la elección de una u otra opción depende de los recursos computacionales disponibles.

La presentación de los datos se divide dos componentes. Esta separación se realiza de acuerdo con la habilidad y la necesidad del usuario. En los tableros de control o cuadros de mando, los usuarios son consumidores de informes predefinidos y tienen un bajo grado de libertad, los informes están prediseñados, pero los usuarios pueden ajustarlos con restricciones. Los tableros deben ser fáciles de usar y permitir un simple salto a los detalles. Por lo tanto, debe tener una funcionalidad de desglose, la usabilidad es importante en este componente, al usuario le debe gustar usar el portal, porque es fácil de usar.

Por otra parte, está el grupo de científicos de datos, en el portal ellos deben tener un alto grado de libertad, así como derechos de acceso y herramientas para

construir completamente sus propios análisis e informes. Este grupo de usuarios necesita plataformas para experimentar con nuevos análisis porque se trata de conjuntos de datos grandes y no estructurados. El mismo portal debe brindar la opción para que los usuarios con un conocimiento medio en manejo y relacionamiento de los datos pueda obtener los análisis e informes deseados.

La capa de gobierno de datos es transversal a todas las mencionadas anteriormente y se compone de la gestión de datos maestros, de la gestión de la seguridad y privacidad de los datos y de administrar la información de las características de los datos que se procesan en todo el ciclo de vida de la arquitectura. Uno de los objetivos de esta capa es trazar las bases para aprovechar los datos en la organización. Es importante tener en cuenta los desafíos legales y regulatorios en el gobierno de los datos (Wang et al., 2018).

Finalmente, las definiciones para los grupos de usuarios se deben adaptar a las necesidades particulares de cada organización, teniendo en cuenta el grado de libertad que se les debe brindar en las visualizaciones. En general, los expertos estuvieron de acuerdo con esta representación.

Con la revisión de literatura, los hallazgos de los grupos focales y la anterior explicación del modelo se propone la arquitectura de referencia para plataformas de servicios de datos, la cual se ilustra en la Figura 17.

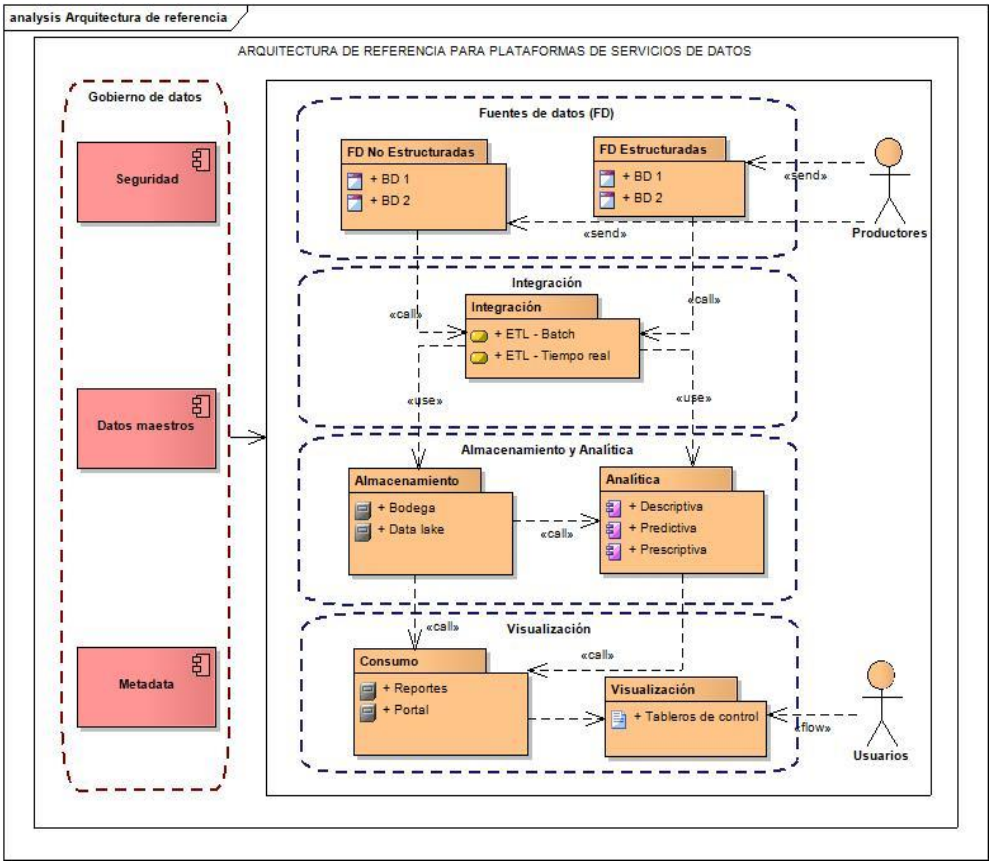


Figura 17. Arquitectura de referencia para plataformas de servicios de datos.

La arquitectura de referencia que se propone para plataformas de servicios de datos tiene en cuenta las siguientes capas:

- Fuentes de datos
- Integración o procesamiento
- Almacenamiento
- Analítica
- Visualización
- Gobierno de datos

4.2. Fuentes de datos

En esta capa se permite la recolección de los datos desde diferentes fuentes. El objetivo principal es leer los datos proporcionados por diversos canales, con

diversos tamaños y formatos, en esta capa se presenta el primer obstáculo en la implementación de una arquitectura de Big Data, ya que las características de los datos pueden variar considerablemente y la adquisición de componentes para almacenarlos puede exceder el presupuesto (Wang et al., 2018). Ante esta situación cada compañía debe decidir la cantidad y el tipo de datos a los que les dará tratamiento.

4.3. Integración o procesamiento

En esta capa se ejecuta la tarea de preparación de los datos para luego permitir su análisis, normalmente esta actividad es la más laboriosa y la que regularmente consume más tiempo (Assunção et al., 2015). En esta capa se fusionan los datos obtenidos de diferentes fuentes en una estructura coherente y homogénea, lo que ayuda a hacer que los datos sean rastreables, más fáciles de acceder y manipular (Blazquez & Domenech, 2018).

Las arquitecturas Lambda y Kappa pueden ser tenidas en cuenta en esta capa, dado que ambas permiten el procesamiento por lotes y en tiempo real, la Lambda combina los dos tipos de procesamiento en fases independientes y la Kappa los simplifica integrándolas en una fase (Zhelev & Rozeva, 2017). Elegir la arquitectura correcta depende de los requisitos de la solución a implementar, si los algoritmos que procesan datos en tiempo real e históricos son los mismos y la solución no necesita cálculos pesados que requieren mucho tiempo, entonces se podría optar por la arquitectura Kappa. En algunos casos, podría no ser posible simplificar el procesamiento de datos y sería necesario usar la arquitectura Lambda.

Esta es una de capas en las cuales el gobierno de datos debe hacer énfasis para incorporar restricciones de privacidad, ya que los datos integrados pueden facilitar el descubrimiento de algunos datos privados y personales que de otro modo serían anónimos. Una opción por considerar en esta capa es realizar la tarea de

integración en la nube, la cual demanda recursos especializados y por lo tanto habilidades especificas en las personas que la ejecutarán para utilizar los recursos de forma adecuada (Assunção et al., 2015).

4.4. Almacenamiento

En esta capa se almacenan los grandes volúmenes de datos para posteriormente permitir su recuperación y además se deben producir los metadatos asociados (Blazquez & Domenech, 2018). Actualmente existen varias opciones en la nube que permiten realizar esta tarea, pero deben ser analizadas con detalle, ya que en algunos casos no satisfacen las necesidades de concurrencia y rendimiento de algunas herramientas utilizadas en el análisis (Assunção et al., 2015).

Una de las actividades a analizar en esta capa es el costo computacional de transportar los datos del sitio de procesamiento al lugar de almacenamiento, dichos costos deben ayudar a determinar si el lugar de procesamiento debe ser más cercano al de almacenamiento.

El almacenamiento de los datos se puede realizar en un lago o en una bodega de datos, esto está determinado por los requisitos de la solución. El lago de datos generalmente se usa para el almacenamiento de datos de una dependencia específica de la organización y la bodega para almacenar los datos históricos de toda la organización (Assunção et al., 2015).

4.5. Analítica

Esta capa permite la ejecución de los análisis de datos y la obtención e interpretación de resultados. Los métodos estadísticos y de aprendizaje automático deben ser utilizados para extraer conocimiento y hacer predicciones a partir de los datos procesados y almacenados en las capas anteriores. Para lograr este objetivo, se aplican técnicas descriptivas y predictivas. El análisis descriptivo

ayuda a proporcionar algunas ideas sobre las características y la evolución de los diferentes tipos de variables de una organización, estos resultados se utilizarán en la capa de visualización y consumo para crear tablas y gráficos que representen la relación entre las variables (Blazquez & Domenech, 2018).

El análisis predictivo se basa en modelos que ayudan a explicar, clasificar, y pronosticar la actividad futura de la organización, el comportamiento y las tendencias. Para hacerlo, los modelos se entrenan usando métodos de aprendizaje automático para seleccionar las variables más significativas e ir mejorando las predicciones (Blazquez & Domenech, 2018).

El análisis prescriptivo ayuda a la toma de decisiones determinando acciones y evaluando su impacto con respecto a los objetivos, requisitos y limitaciones del negocio.

Existen diversas formas de realizar este análisis: procesamiento por lotes o en paralelo y procesamiento en tiempo casi real, se dice “casi real”, porque pueden existir diferencias en fracciones de segundo. La elección de cual técnica escoger estará basada en el tipo de análisis que se debe realizar y esto determinará la herramienta tecnológica que se debe usar, por ejemplo, MapReduce permite el procesamiento en paralelo de grandes volúmenes de datos y es el modelo más utilizado en análisis de Big Data (Wang et al., 2018).

En esta capa se deben definir los tipos de usuarios que realizarán el análisis de los datos para determinar qué tipos de herramientas tecnológicas se disponen para dicho fin y el grado de procesamiento que se debe realizar en los datos, de esta manera se podrán facilitar las tareas de autogestión de los diferentes grupos de usuarios (Passlick et al., 2017).

4.6. Visualización

En esta capa se generan los informes o reportes y se permite el monitoreo de los datos procesados. La presentación de informes es una característica crítica del análisis de Big Data que permite visualizar los datos de una manera útil para respaldar las operaciones diarias y ayudar a la toma de decisiones en las compañías (Wang et al., 2018).

La presentación de los datos se debe dividir según el tipo de usuario que los visualizará. La separación se realiza de acuerdo con la habilidad y la necesidad del usuario (Passlick et al., 2017). En los tableros de control, los usuarios son consumidores de informes predefinidos y allí visualizan y dan seguimiento a las métricas de desempeño de la organización. Los tableros de control son utilizados, generalmente, por usuarios ocasionales. Otro grupo de usuarios son los científicos de datos, ellos tienen un alto grado de libertad, así como los derechos de acceso y las herramientas para construir completamente sus propios análisis e informes. Entre las dos opciones mencionadas anteriormente se encuentra el portal de análisis, allí los informes están predefinidos, pero los usuarios pueden realizar ajustes con algunas restricciones. Las definiciones de grupos de usuarios pueden adaptarse a las necesidades individuales de cada organización.

En esta capa aparece la plataforma de servicios de datos para permitir la distribución y diagramación de grandes volúmenes de datos y así facilitar el análisis y la toma de decisiones.

La información o reportes que se presentan en esta capa deben permitir que los diferentes grupos de usuarios responsan las preguntas de interés del negocio y puedan hacer la creación y evaluación de escenarios hipotéticos para la toma de decisiones (Gonçalves et al., 2017).

Durante esta etapa el gobierno de datos debe establecer los derechos de autor de los datos e informes visualizados, citando fuentes y dando controles de acceso (Blazquez & Domenech, 2018).

4.7. Gobierno de datos

Esta capa es transversal a las demás y allí se deben aplicar las políticas y regulaciones de la organización a todo el ciclo de vida de los datos: desde la ingesta de datos hasta la eliminación (Blazquez & Domenech, 2018). Otros componentes de esta capa son la gestión de datos maestros, la gestión de la seguridad y la privacidad de los datos y los metadatos. Esta capa enfatiza en el "cómo" aprovechar los datos en la organización. La gestión de datos maestros está definida como los procesos, políticas y estándares para administrar los datos.

El componente de la seguridad de los datos y la gestión de la privacidad es la plataforma para proporcionar actividades de datos a nivel empresarial en términos de descubrimiento, evaluación de la configuración, monitoreo, auditoría y protección (Wang et al., 2018). Los metadatos permiten identificar y estandarizar los nombres de cada uno de los campos que pueden tratar y visualizar los usuarios.

La capa de gobierno debe tener en cuenta los siguientes módulos (Blazquez & Domenech, 2018):

- Módulo de ingestión: se ocupa de la gestión de las fuentes de datos, los permisos de usuario y la integridad de los metadatos.

- Módulo de procesamiento: realiza un seguimiento de las transformaciones, así como de los permisos para acceder a los datos y recursos informáticos.

- Módulo de resultados: se ocupa de la trazabilidad de los resultados (desde las fuentes hasta el informe final), los permisos para acceder a los informes y los resultados.

- Módulo de auditoría: inspecciona que la implementación de la arquitectura sea coherente con las políticas de seguridad y privacidad. También puede incluir la verificación del rendimiento general de la arquitectura para garantizar que el sistema tenga un tiempo de respuesta aceptable.

Con la arquitectura de referencia establecida y definidos los lineamientos para cada una de las capas se proceden a realizar el montaje con un caso de estudio.

CAPÍTULO 5

Montaje y evaluación de la arquitectura de referencia

5.1. Introducción

Con el fin de realizar la validación y evaluación de la arquitectura de referencia se debe ejecutar la implementación y montaje en algún sector de la industria, esto con el fin de mapear cada una de las capas de la arquitectura, teniendo en cuenta los lineamientos y así poder ilustrar la importancia de cada uno de los componentes a la hora de hacer la definición de la arquitectura.

Es necesario elegir un sector industrial que facilite grandes volúmenes de datos y la definición o explicación de cada uno de ellos, por esta razón se recurre a una fuente de información oficial del Gobierno Colombiano, en la que están disponibles diferentes fuentes de datos para el acceso de todos los ciudadanos. Una de estas fuentes es la atención en urgencias en las diferentes ciudades del país.

La atención de urgencias es uno de los servicios con mayor demanda en las instituciones de salud y al que cualquier persona tiene el derecho de acceder con el objetivo de resolver una situación de salud que ponga en riesgo su vida de manera rápida y transitoria.

La puerta de entrada a las urgencias es el triage, que corresponde a una clasificación del estado de salud medida en 5 categorías, de las cuales aplican para atención en urgencias a quienes por sus condiciones fisiológicas, físicas y clínicas apliquen para categoría 1, 2 y 3. Es de anotar, que las personas que ingresen en situación de riesgo de vida, es decir, categoría 1, pasan de inmediato a atención médica.

Luego del triage, el paciente pasa a consulta, en la que se definen posibles diagnósticos y su plan de manejo; en los casos más críticos de salud el paciente pasa a cirugía o a unidades de alta dependencia como lo son las Unidad de Cuidados Especiales o Intermedios (UCE) o Unidades de Cuidados Intensivos (UCI).

Luego de la estabilización del paciente, se define si este requiere ser ingresado a la Institución Prestadora de Servicios de Salud (IPS) en condición de hospitalización o si se da egreso.

Los costos de las atenciones de urgencias deben ser asumidos por las Empresas Promotora de Salud (EPS), pólizas de salud, otros tipos de aseguramiento o por los propios usuarios, en caso de no contar con una afiliación a las anteriormente mencionadas.

Teniendo en cuenta la alta demanda de la atención de urgencias, en algunas IPS se puede evidenciar una saturación de este servicio, poniendo en riesgo la calidad de la atención, la seguridad del paciente, las condiciones de bioseguridad para trabajadores de la salud y pacientes, generando insuficiencia de personal o de los espacios para la atención médica, entre otros factores que son objeto de análisis permanente para entes de control, IPS o EPS; quienes también miden las atenciones de urgencias prestadas en un periodo de tiempo, diagnósticos más comunes, especialidades médicas más demandadas, IPS con mayor volumen de atenciones de urgencias, costos de la atención, días de estancia, calidad en la atención y diagnósticos. Estos análisis tienen el objetivo de prestar servicios de urgencias seguros, oportunos, con calidad, eficacia, eficiencia y satisfacción.

5.2. Establecimiento de los niveles de interés

Los roles dentro de las IPS y EPS que necesitan realizar análisis de la información son:

- Director médico: el rol del director médico se relaciona con la parte administrativa de la institución y se basa en controlar y verificar el desempeño del personal médico sobre la atención brindada a los usuarios, además de relacionarse directamente con la finanzas y administración de los presupuestos para la prestación de los servicios.

- Director hospitalario: el rol del director hospitalario es brindar supervisión técnica y apoyo profesional al personal a cargo de su área, elaborar planes de trabajo para ejecutar las actividades necesarias para el correcto funcionamiento del equipo de trabajo, evaluarlo y hacer seguimiento al cumplimiento de los indicadores establecidos por la empresa.

- Director de urgencias: el rol de director de urgencias es garantizar todas las condiciones que permitan la atención ágil y oportuna a los pacientes que consultan en urgencias, cumpliendo con las políticas de calidad de la atención establecidas por la institución.

5.3. Análisis de indicadores y tableros de control

Los tableros de control que se diseñaron en la etapa de visualización de la arquitectura deben permitir al negocio responder las siguientes preguntas:

- ¿Cuáles son las IPS con mayor atención?
- ¿Cuáles son las EPS con mayor atención?
- ¿Cuáles son las causas externas más comunes en la atención en urgencias?
- ¿Cuáles son los diagnósticos más comunes en la atención en urgencias?
- Presentación geográfica de las IPS
- ¿Cuáles son los rangos de edad más propensos a la atención en urgencias?
- Atenciones por tipo de usuario
- Atenciones relacionando la IPS y la EPS.

- IPS clasificadas por rango de edad
- Diagnósticos más frecuentes por rango de edad
- Resumen por IPS
- Causas relacionando la IPS y EPS
- Tipo de Usuario por rango de edad
- Resumen Total

Los principales indicadores que pueden visualizar en los reportes o tableros de control de la arquitectura son:

Tabla 3. IPS con mayor número de atenciones

Nombre del Indicador: IPS con mayor atención	
Área de Análisis: Sector Salud – Urgencias	
Descripción	Se requieren identificar las IPS con mayor atención en el servicio de urgencias para establecer programas que agilicen y mejoren la atención.
Tipo de dato	Numérico
Estrategia de cálculo	Realizar carga de datos, sumar las atenciones por IPS para obtener las que tienen mayor número de servicios. *Suma de atenciones por IPS ordenadas de forma descendente.*

Tabla 4. EPS con mayor número de atenciones

Nombre del Indicador: EPS con mayor atención	
Área de Análisis: Sector Salud – Urgencias	
Descripción	Se requieren identificar las EPS que tienen el mayor número de atenciones en urgencias, con el fin de establecer programas de promoción y prevención que ayuden a

	descongestionar este servicio y para hacer vigilancia y control de los programas crónicos de las IPS adscritas a la EPS.
Tipo de dato	Numérico
Estrategia de cálculo	Realizar carga de datos, sumar las atenciones agrupando por EPS para obtener las que tienen mayor número de servicios. *Suma de atenciones por EPS ordenadas de forma descendente.*

Tabla 5. Causas externas más consultadas

Nombre del Indicador: Causas externas más consultadas	
Área de Análisis: Sector Salud – Urgencias	
Descripción	Se requieren identificar las causas más comunes por las cuales se utiliza el servicio de urgencias de las IPS para establecer programas de acción en conjunto con las EPS.
Tipo de dato	Numérico
Estrategia de cálculo	Realizar carga de datos, sumar las atenciones agrupando por la causa para obtener las que tienen mayor número de servicios. *Suma de atenciones por causa ordenadas de forma descendente.*

Tabla 6. Diagnósticos más comunes

Nombre del Indicador: Diagnósticos más comunes	
Área de Análisis: Sector Salud – Urgencias	
Descripción	Se requiere conocer los diagnósticos más comunes en la

	atención de urgencias para determinar los principales problemas de salud en la población, además de la oportunidad y calidad de la atención en los servicios ambulatorios.
Tipo de dato	Numérico
Estrategia de cálculo	Realizar carga de datos, sumar las atenciones por diagnóstico. *Suma de atenciones por diagnóstico ordenadas de forma descendente.*

Tabla 7. Rangos de edad más propensos a la atención en urgencias

Nombre del Indicador: Rangos de edad más propensos a la atención en urgencias	
Área de Análisis: Sector Salud – Urgencias	
Descripción	Se requiere conocer los rangos de edad más propensos a la atención en urgencias para evidenciar el comportamiento de las enfermedades según los diferentes grupos de edad
Tipo de dato	Numérico
Estrategia de cálculo	Realizar carga de datos, definir los límites de edad para cada rango, sumar las atenciones por rango de edad. *Suma de atenciones por rango de edad.*

Tabla 8. Atenciones por tipo de usuario

Nombre del Indicador: Atenciones por tipo de usuario	
Área de Análisis: Sector Salud – Urgencias	
Descripción	Se requiere conocer las atenciones por tipo de usuario para medir cobertura, comparación entre regímenes en cuanto a barreras de acceso, manejo de los programas de pacientes

	crónicos y programas de promoción de la salud y prevención de la enfermedad.
Tipo de dato	Numérico
Estrategia de cálculo	Realizar carga de datos, sumar las atenciones por tipo de usuario. *Suma de atenciones por tipo de usuario.*

5.4. Consideraciones de la implementación

La calidad de la información es fundamental para tener un indicador cercano a la realidad, por tanto, todas las IPS deben suministrar los servicios de urgencias que se han atendido.

Otras consideraciones para tener en cuenta para la implementación son:

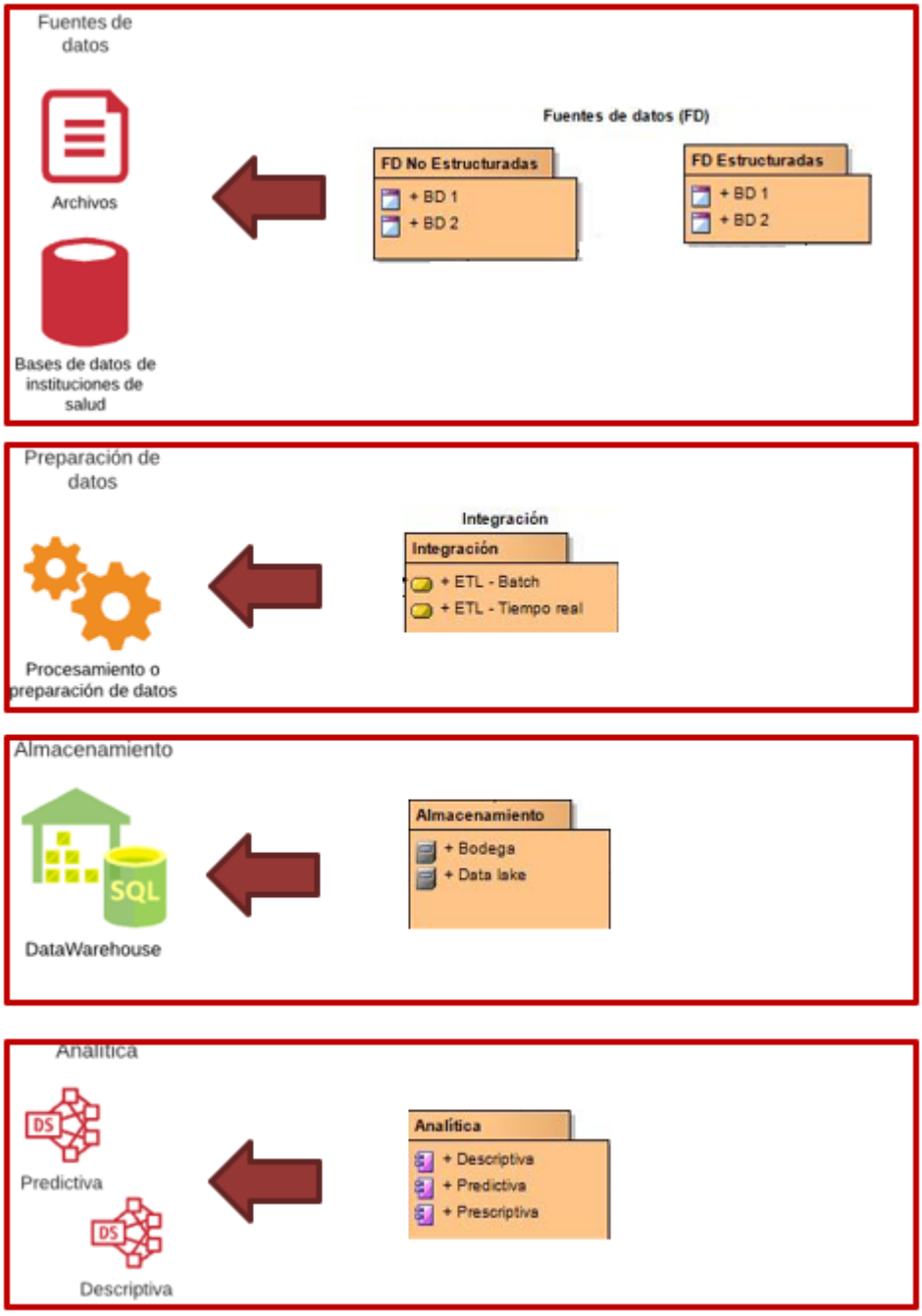
- El archivo contiene las atenciones de urgencias de 2016 por parte de las IPS de la ciudad de Medellín.
- La edad se presenta en años, meses o días, por lo cual para establecer los rangos de edad se debe establecer una fórmula para definir si es niño, joven, adulto o adulto mayor.
- Algunas personas atendidas no tenían EPS, pero el servicio de urgencias es prestado por la IPS.

Estrategia de construcción: Se realiza la descarga del archivo fuente que contiene las atenciones de urgencias de las IPS de la ciudad de Medellín en el año 2016, luego se procesa el archivo para definir los datos maestros, posteriormente se realiza un proceso de extracción, transformación y carga, después se

almacenan los datos y se desarrollan tareas programadas para realizar la carga automática, finalmente se desarrollan las visualizaciones.

5.5. Definición de la arquitectura

Basados en la información presentada anteriormente y en los objetivos de las IPS y EPS se hace un mapeo de la arquitectura de referencia a la arquitectura requerida, con el fin de analizar las atenciones en los servicios de urgencias, como se ilustra en la Figura 18


Fuentes de datos
Archivos
Bases de datos de instituciones de salud
Fuentes de datos (FD)
FD No Estructuradas
+ BD 1
+ BD 2
FD Estructuradas
+ BD 1
+ BD 2
Preparación de datos
Procesamiento o preparación de datos
Integración
Integración
+ ETL - Batch
+ ETL - Tiempo real
Almacenamiento
DataWarehouse
SQL
Almacenamiento
+ Bodega
+ Data lake
Analítica
Predictiva
DS
Descriptiva
DS
Analítica
+ Descriptiva
+ Predictiva
+ Prescriptiva

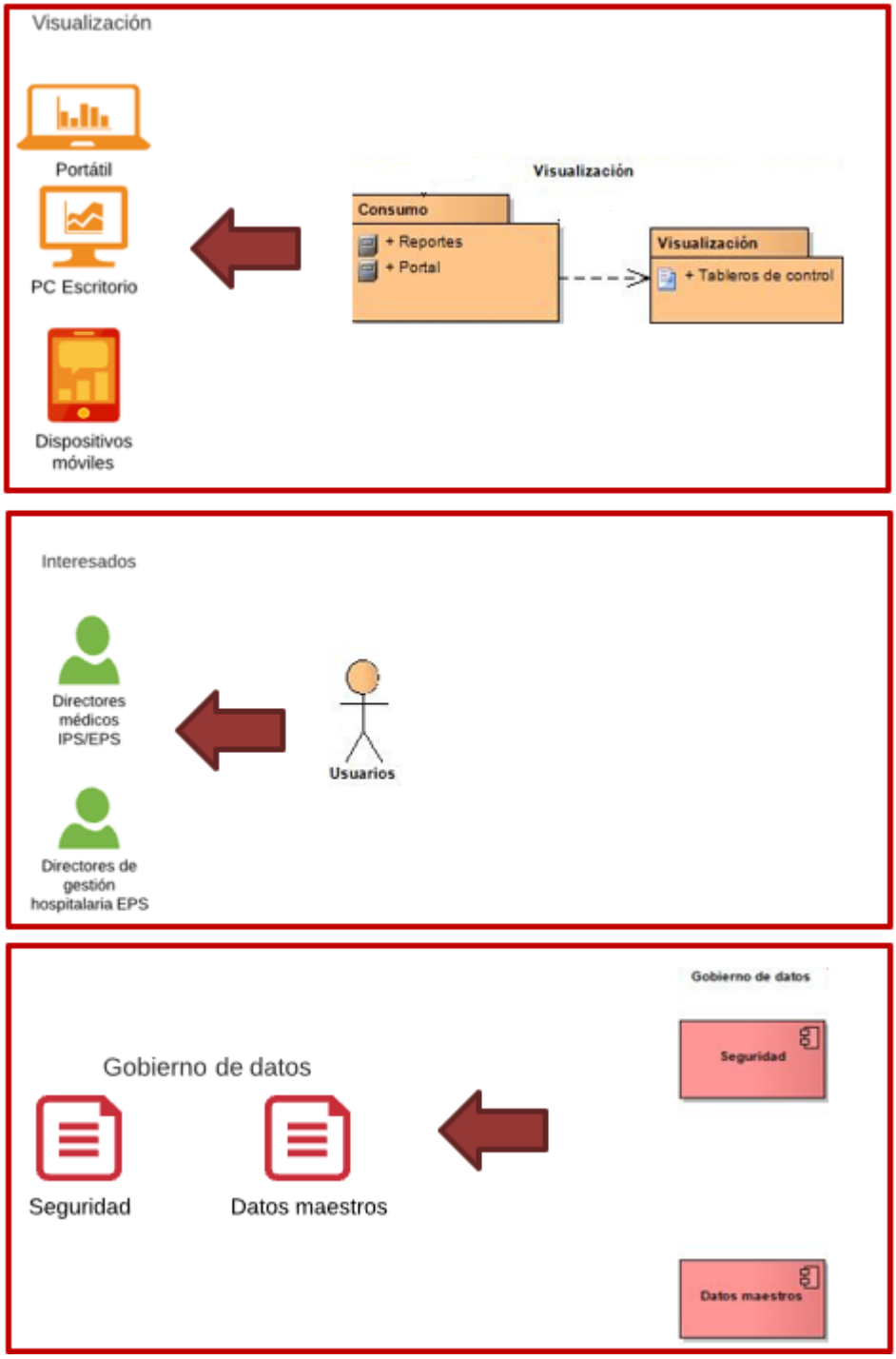


Figura 18. Mapeo de la arquitectura de referencia.

Con base en los objetivos estratégicos, en las necesidades de información de las IPS y las EPS, en la definición de los interesados y en el mapeo de cada una de las capas requeridas se establece la arquitectura de referencia. En la Figura 19 se ilustra esta arquitectura.

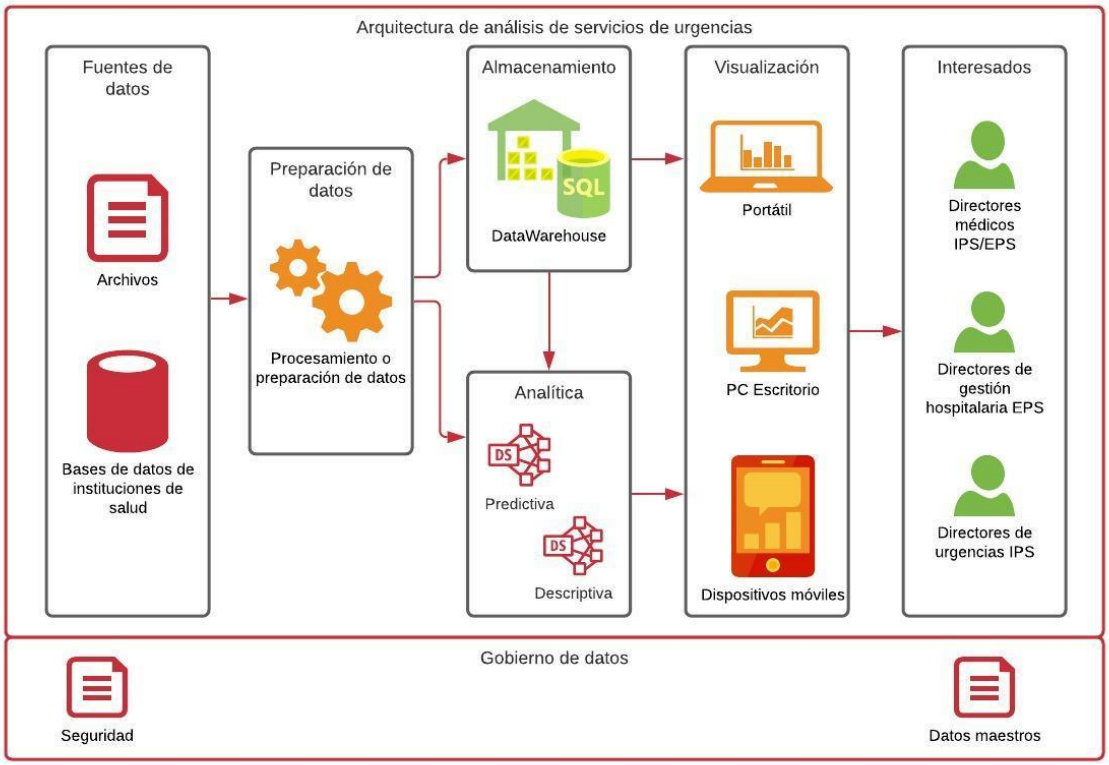


Figura 19. Arquitectura establecida para análisis de servicios de urgencias.

5.6. Evaluación de la arquitectura

Con el fin de realizar la evaluación de la arquitectura se toman datos del Ministerio de Salud y Protección Social de Colombia sobre las atenciones en urgencias en el municipio de Medellín en 2016 (Colombia, 2016), para el ejercicio se tomará como fuente de datos el archivo en formato .csv. La preparación de los datos se realiza por medio de Integration Services de Visual Studio, así como se ilustra en la Figura 20 y Figura 21

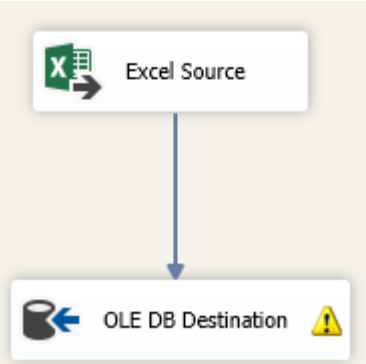


Figura 20. Procesamientos de los datos.

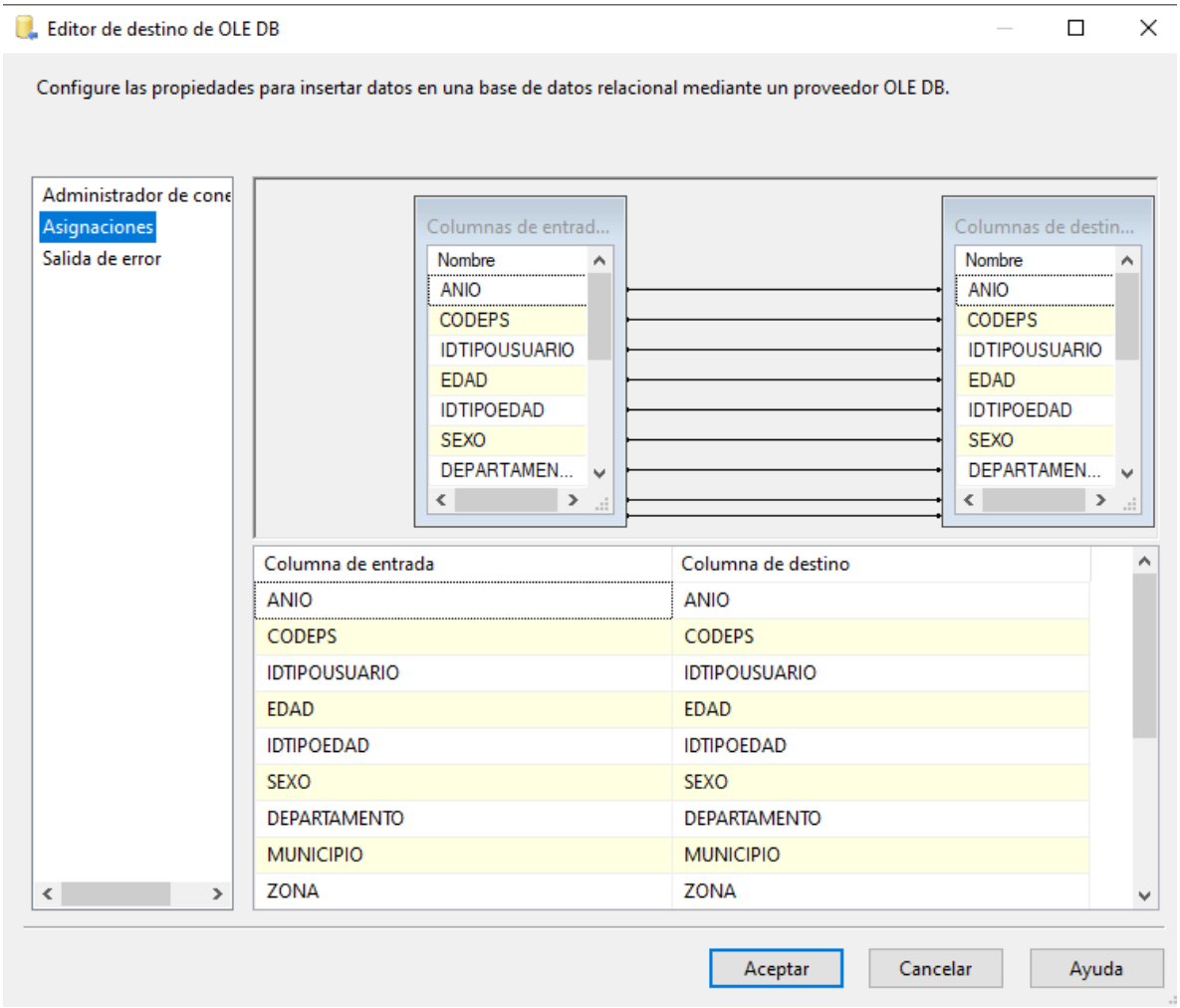


Figura 21. Mapeo de los datos.

Luego de esta transformación y carga, se almacenan los datos en una base de datos SQL Server, como se ilustra en la Figura 22

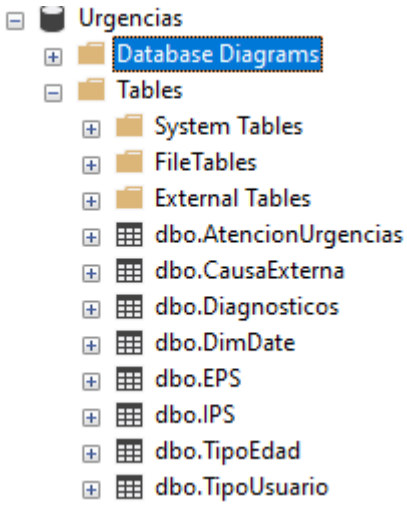


Figura 22. Almacenamiento de los datos.

Con el fin de automatizar la carga de los datos, se diseñan tareas programadas para que se ejecuten periódicamente y se guarden los datos en cada una de las tablas, como se ilustra en la Figura 23 y Figura 24

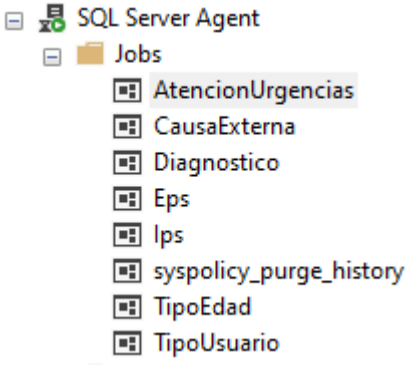


Figura 23. Tareas programadas

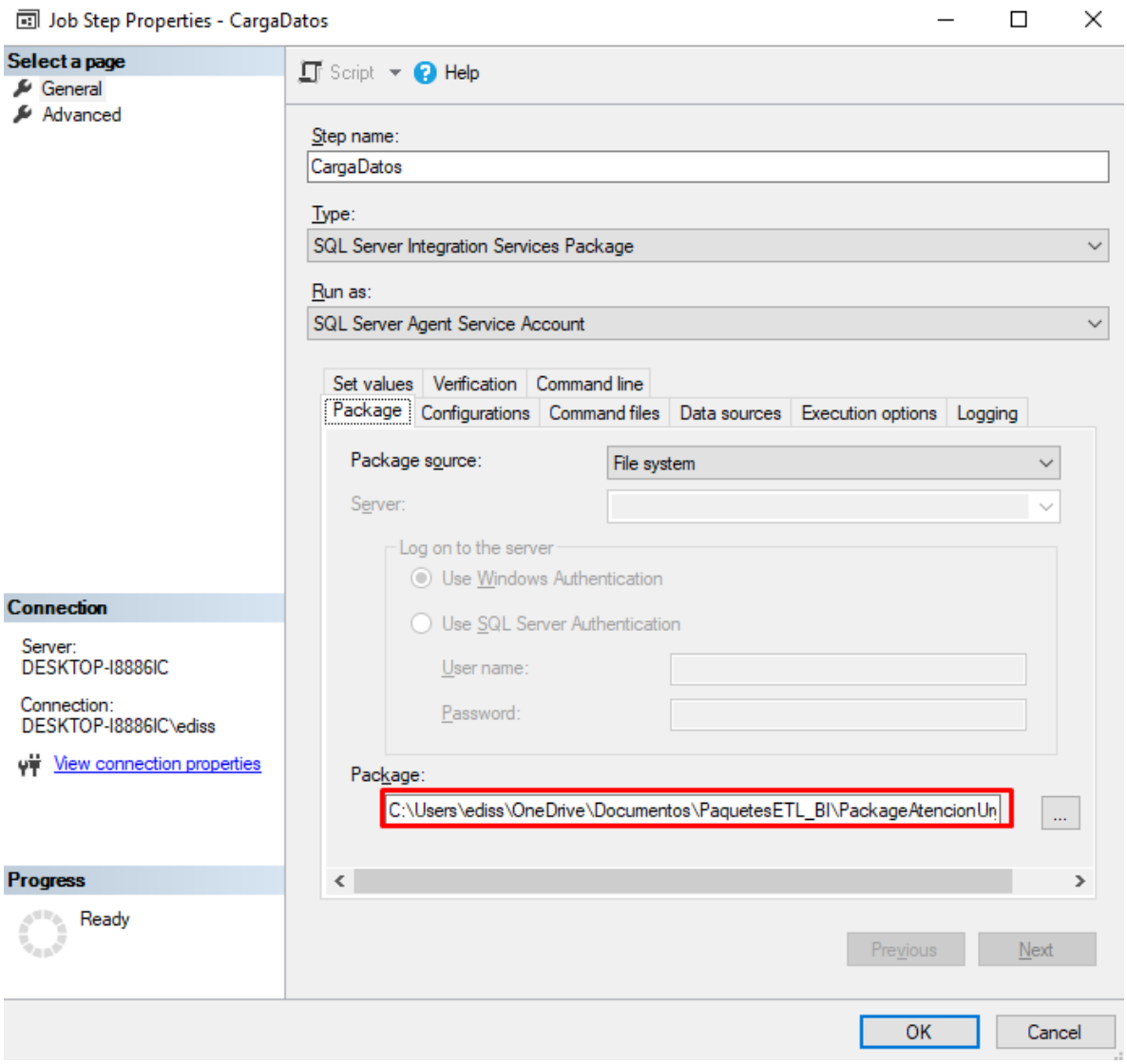


Figura 24. Configuración y programación de la tarea.

En la capa de analítica se desarrolla un cubo y sus dimensiones con un proyecto de “Analysis Services” de Visual Studio, como se ilustra en la Figura 25 y Figura 26

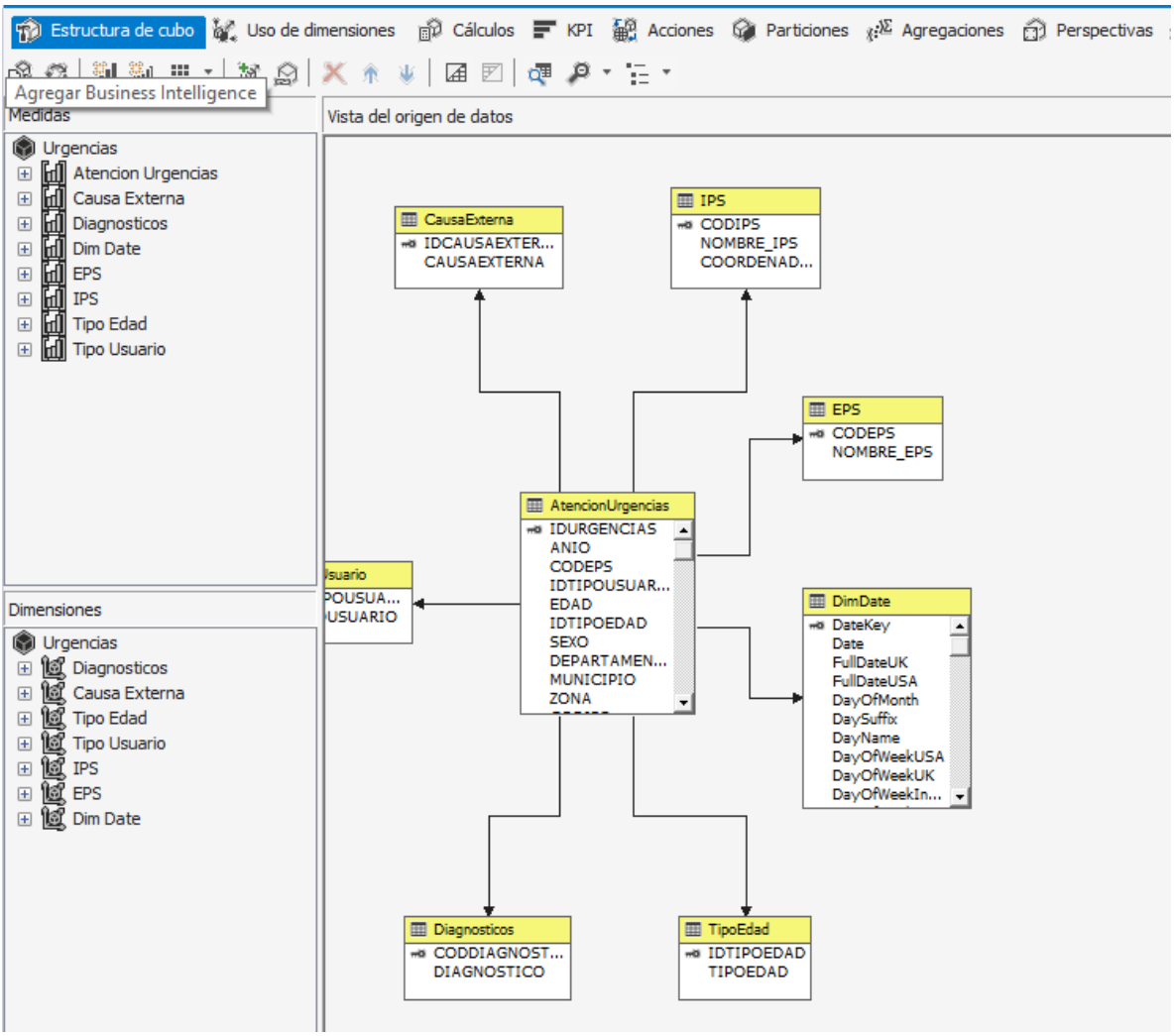


Figura 25. Estructura del cubo.

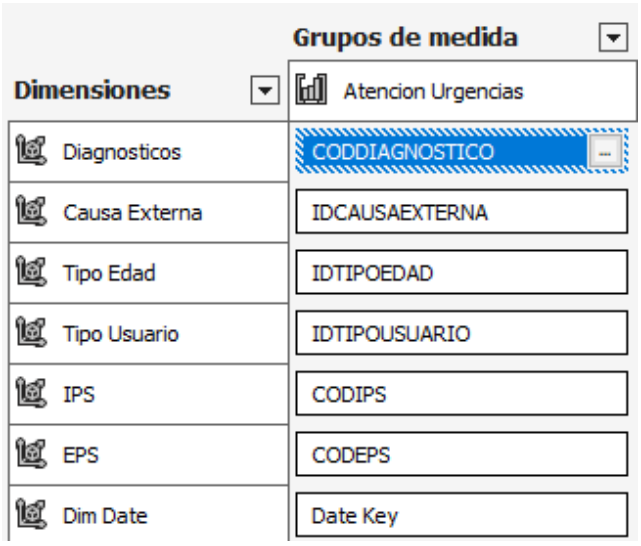


Figura 26. Dimensiones.

La seguridad y los datos maestros se establecen en el DataWarehouse y se definen políticas para el acceso a los tableros de control.

Las visualizaciones de cada uno de los tableros de control que dan respuesta a las preguntas de los usuarios se ilustran en la Figura 27, Figura 28, Figura 29, Figura 30 y Figura 31

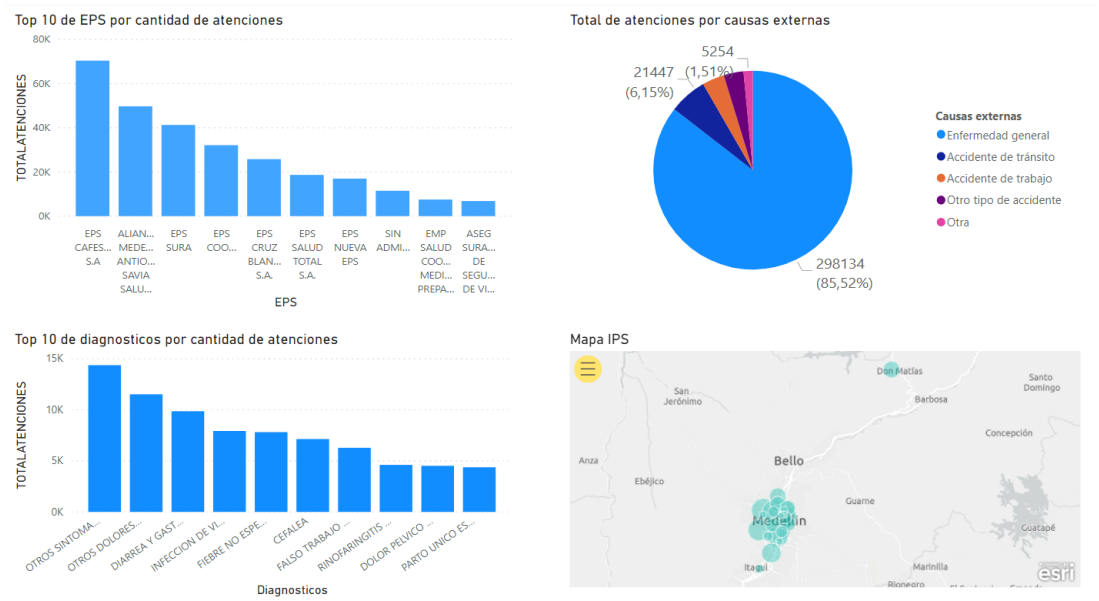


Figura 27. Tablero de control EPS – Causa – Diagnóstico.

En el tablero de control de la Figura 27 se ilustra la relación de la atención en urgencias entre EPS, diagnóstico y causa externa. Como se puede observar la EPS CafeSalud es la que más servicios de urgencias prestó. Es necesario hacer la observación a los médicos que sean más precisos en el diagnóstico, ya que el item de mayor atención es “Otros síntomas”.

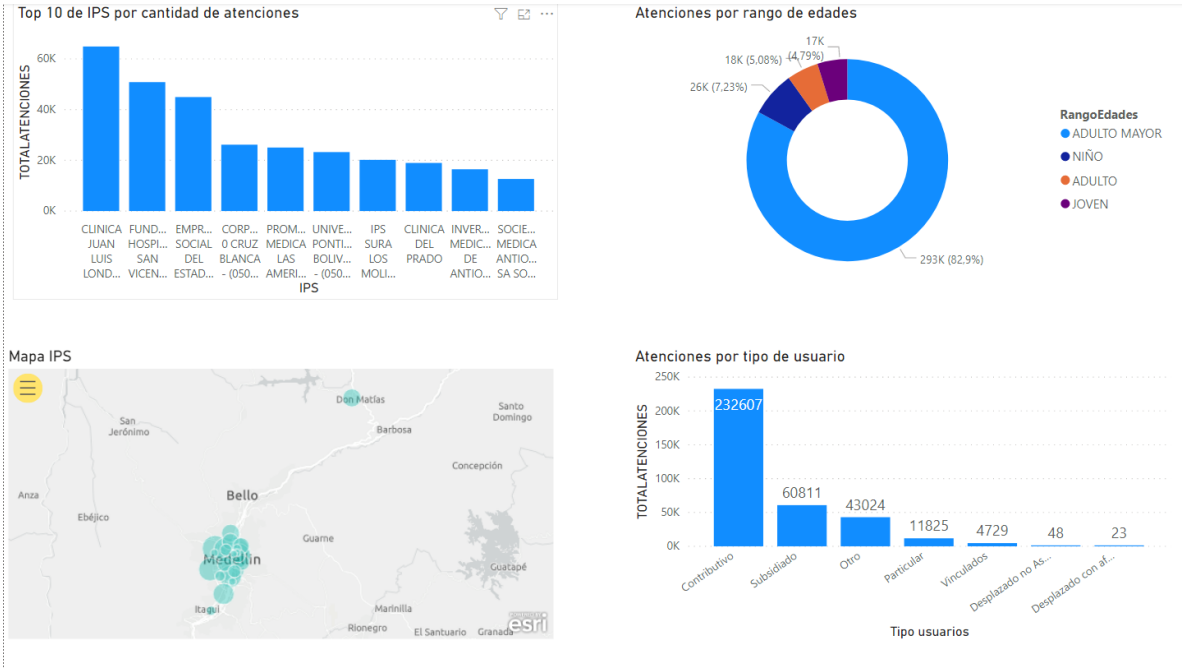


Figura 28. Tablero de control IPS – Rango edad – Tipo Usuario.

En el tablero de control de la Figura 28 se ilustra la relación de la atención en urgencias entre IPS, rango de edad y tipo de usuario. Se puede observar que el adulto mayor es quien más consultas realiza en urgencias, obteniendo un gran porcentaje. Esto resultado muestra que es necesario realizar programas de promoción y prevención sobre esta población, con el fin de evitar las enfermedades y así disminuir el número de atenciones en urgencias.

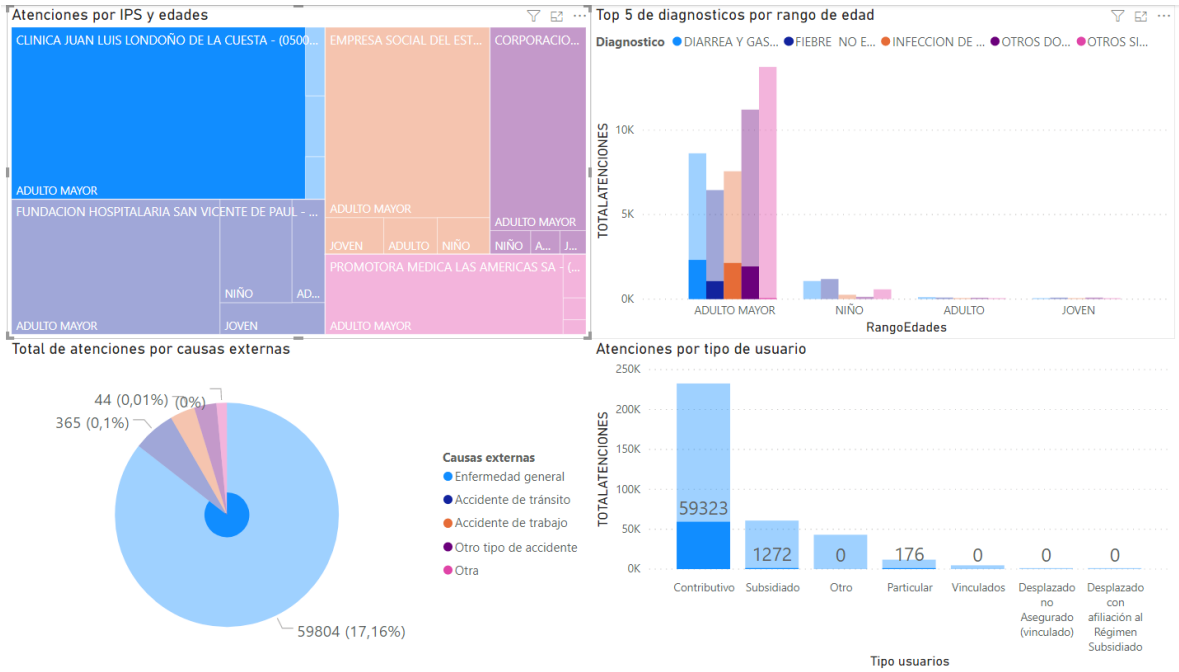


Figura 29. Tablero de control IPS – Rango edad – Causa - Diagnóstico.

En el tablero de control de la Figura 29 se ilustra la relación de la atención en urgencias entre IPS, rango de edad, causa y diagnóstico. Se evidencia que el adulto mayor es una población crítica en cada una de las IPS, demandando un alto número de servicios. Este resultado debe llevar a las directivas de las EPS a ser más estrictos con los planes preventivos que desarrollan con los pacientes.

TipoUsuario-RangoEdad

Tipo Usuario.TIPOUSUARIO	ADULTO	ADULTO MAYOR	JOVEN	NIÑO	Total
Vinculados	427	804	520	308	**2059**
Subsidiado	2359	6696	2806	3558	**15419**
Particular	947	1708	1136	913	**4704**
Otro	2750	4995	3060	1469	**12274**
Desplazado no Asegurado (vinculado)	10	8	15	7	**40**
Desplazado con afiliación al Régimen Subsidiado	8	7	4		**19**
Contributivo	6178	17755	6308	6342	**36583**
Total	**12679**	**31973**	**13849**	**12597**	**71098**

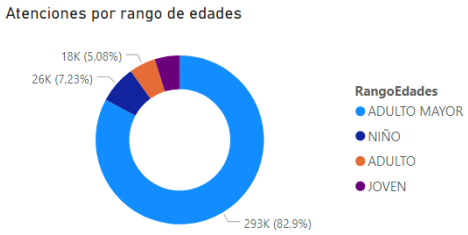


Figura 30. Tablero de control Tipo usuario – Rango edad.

En el tablero de control de la Figura 30 se ilustra la relación de la atención en urgencias entre el tipo de usuario y el rango de edad. Se evidencia nuevamente que la población que más atenciones demanda es el adulto mayor dentro de cualquiera de los regímenes, aunque el contributivo tiene un porcentaje muy alto, es necesario analizar el subsidiado para planear la forma de ofrecerles los programas preventivos y disminuir así la cantidad de atenciones.

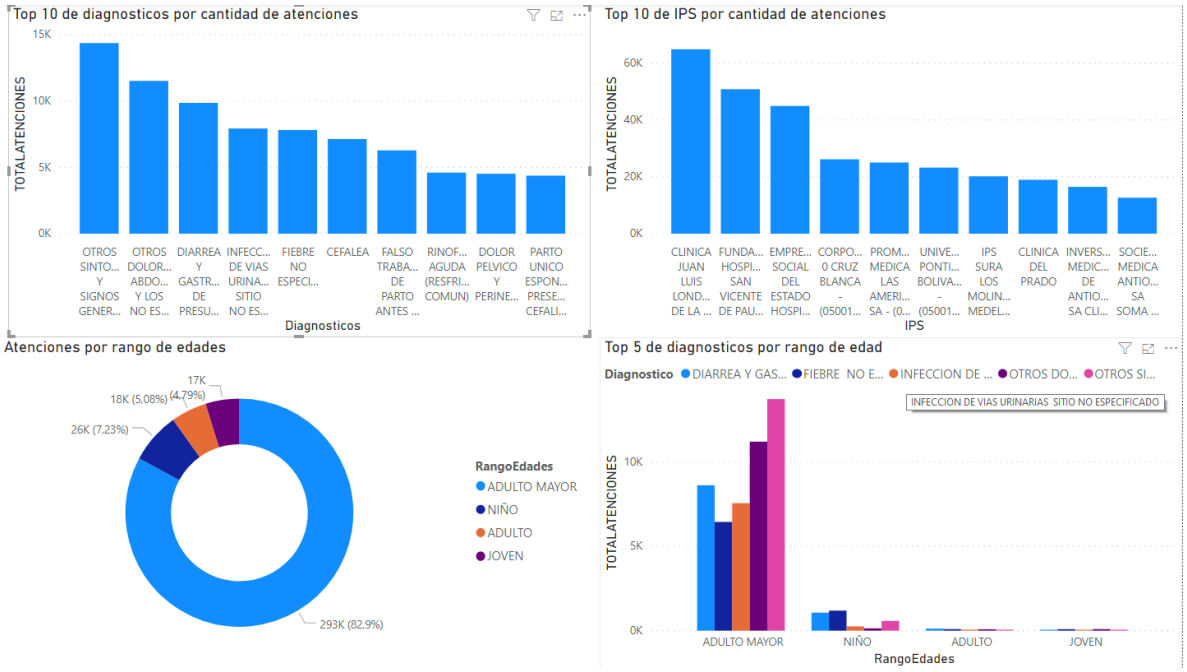


Figura 31. Tablero de control IPS – Rango edad - Diagnóstico.

Los científicos de datos o personas especializadas pueden ejecutar diferentes comandos con Python para obtener analíticas especializadas como la que se ilustra en la Figura 32 para relacionar diferentes variables y observar comportamiento futuro de los datos.

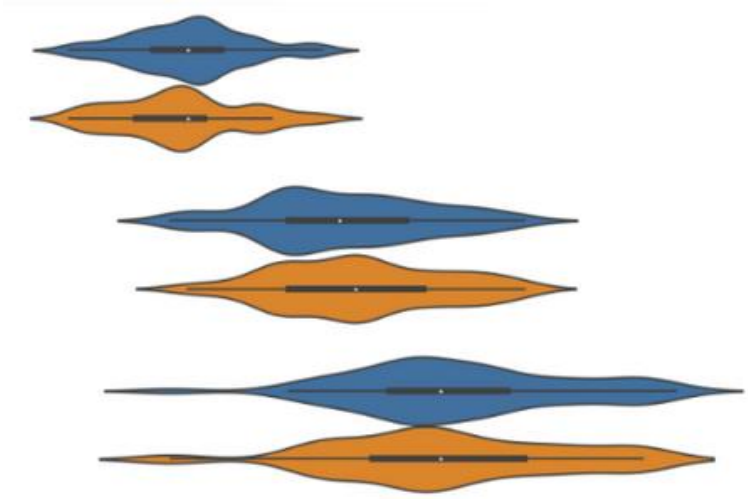


Figura 32. Distribución de datos.

En la Figura 33 se ilustra el código fuente que genera esta gráfica

```
# The following code to create a dataframe and remove duplicated rows is always executed and acts as a preamble for your script:

# dataset = pandas.DataFrame(Diagnosticos.DIAGNOSTICO, Causa Externa.CAUSAEXTERNA, RangoEdades, IPS.NOMBRE IPS)
# dataset = dataset.drop_duplicates()

# Paste or type your script code here:
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as pyplot

Visual = sns.violinplot(x='Recuento Atencion Urgencias', y='Diagnosticos.DIAGNOSTICO', data=dataset, hue='IPS.NOMBRE IPS', bw=.3)
plt.show()
```

Figura 33. Código para relacionar variables.

La implementación presentada muestra que la arquitectura de referencia se puede adaptar a las necesidades de una organización y a diferentes tipos de herramientas tecnológicas. Además, se puede hacer la presentación de los datos según el tipo de usuario que los utilizará, en la implementación presentada se puede observar que los científicos de datos o personas especializadas pueden hacer uso de otras herramientas y programas para hacer un análisis más profundo de los datos.

PARTE IV
CONCLUSIONES

"Nuestra mayor debilidad radica en renunciar. La forma más segura de tener éxito es siempre intentarlo una vez más" -- Thomas Alva Edison

CAPÍTULO 6

Resultados, conclusiones y trabajos futuros

6.1. Evaluación de la arquitectura de referencia

La evaluación de la arquitectura se realiza con base en aspectos planteados en la introducción. Primero, las organizaciones no identifican la forma y las herramientas tecnológicas para gestionar y procesar grandes volúmenes de datos (Jovanovic et al., 2016); frente a este punto, la arquitectura de referencia para plataformas de servicios de datos brinda una guía para que cada organización adopte el camino a seguir para establecer su arquitectura, según las fases que desee aplicar y las herramientas tecnológicas con la que cuente o desee adquirir.

Segundo, las plataformas desarrolladas no facilitan el aumento de la capacidad de trabajo sin comprometer el funcionamiento (Oussous et al., 2018) y se tiene desconocimiento del dominio del problema que se desea atender y de los datos que se procesarán (Zhelev & Rozeva, 2017); ante estas situaciones, identificar las fuentes de datos que serán tratadas por la arquitectura a implantar es el primer paso para definir todo el proceso de extracción, transformación y análisis de datos y así, no comprometer a futuro el funcionamiento de toda la arquitectura. La definición de las fuentes de datos debe incluir la cantidad y los tipos de datos a los que se les dará tratamiento, como primer paso para tener claridad del problema que se desea atender, el gobierno de datos ayudará en esta actividad estableciendo una tarea para la descripción de los datos.

Con relación al desconocimiento del dominio del problema, es necesario tener entendimiento de la organización, claridad en sus objetivos y comprensión de los datos que se procesarán. El propósito es convertir esos objetivos en objetivos técnicos dentro de la arquitectura y así establecer las capas y procesos que se

deben cubrir, así se podrán recolectar los datos correctos y después interpretar correctamente los resultados.

Por último, una arquitectura de referencia de una plataforma de Big Data debe ser capaz de desvincularse de los detalles técnicos (Blazquez & Domenech, 2018), dicha arquitectura no debe estar ligada a herramientas técnicas particulares y debe incluir los posibles tipos de usuario que la pueden utilizar (Zaghloul et al., 2015); frente a estos aspectos, la arquitectura de referencia pretende ser lo suficientemente general para implementarse con diferentes tecnologías, paradigmas informáticos y software analítico, dependiendo de los requisitos y propósitos de cada caso en particular.

6.2. Conclusiones

El gobierno de datos y la autogestión son dos características que deben ser tenidas en cuenta por las organizaciones para ser incluidas en una arquitectura de servicios de datos, con el fin de facilitar el uso de estas plataformas y establecer políticas para regular el ciclo de vida de los datos.

Es necesario identificar y agrupar los tipos de usuario que puede tener la plataforma de servicio de datos, ya que ellos serán una de las guías para la construcción de la arquitectura.

Frente a los hallazgos de la revisión del estado del arte, la propuesta de arquitectura de referencia tiene las siguientes diferencias: considera diferentes tipos de usuario y de fuentes de datos, no genera dependencia por el tipo de herramientas tecnológicas que se utilizan, las capas pueden ser implementadas en la nube y se establece una capa para el gobierno de datos.

Dentro de los componentes de la arquitectura establecida, uno de los principales retos pendientes es la gestión de los flujos intensivos de datos, y requiere una atención especial por parte del mundo académico y de la industria.

La arquitectura de referencia diseñada es la principal contribución de este trabajo, quedando como trabajo futuro la implementación y respectivas pruebas en algún tipo de industria y dentro de un contexto escalable.

Aunque la arquitectura propuesta es lo suficientemente general como para implementarse con cualquier tecnología, su adaptación no está exenta de obstáculos. Para mencionar alguno de ellos, la integración de la arquitectura con los sistemas de información de la organización se convierte en un proceso crítico, el cual debe asegurar la generación de pronósticos y predicciones sin problemas.

BIBLIOGRAFÍA

Artac, M., Borovsak, T., Di Nitto, E., Guerriero, M., Perez-Palacin, D., & Tamburri, D. A. (2018). Infrastructure-as-Code for Data-Intensive Architectures: A Model-Driven Development Approach. *Proceedings - 2018 IEEE 15th International Conference on Software Architecture, ICSA 2018*, 156–165. https://doi.org/10.1109/ICSA.2018.00025

Assunção, M. D., Calheiros, R. N., Bianchi, S., Netto, M. A. S., & Buyya, R. (2015). Big Data computing and clouds: Trends and future directions. *Journal of Parallel and Distributed Computing*, *79*, 3–15.

Bibri, S. E. (2019). The anatomy of the data-driven smart sustainable city: instrumentation, datafication, computerization and related applications. *Journal of Big Data*, *6*(1). https://doi.org/10.1186/s40537-019-0221-4

Blazquez, D., & Domenech, J. (2018). Big Data sources and methods for social and economic analyses. *Technological Forecasting and Social Change*, *130*(September 2017), 99–113. https://doi.org/10.1016/j.techfore.2017.07.027

Camargo Vega, J. J., Camargo Ortega, J. F., & Joyanes Aguilar, L. (2015). Arquitectura Tecnológica Para Big Data. *Revista Científica*, *1*(21), 7. https://doi.org/10.14483/udistrital.jour.rc.2015.21.a1

Clarke, P., Tyrrell, G., & Nagle, T. (2016). Governing self service analytics. *Journal of Decision Systems*, *25*, 145–159. https://doi.org/10.1080/12460125.2016.1187385

Colombia, M. de S. y P. S. (2016). *Atenciones en Urgencias Municipio de Medellín 2016*. Datos Abiertos Del Gobierno de Colombia. https://www.datos.gov.co/en/Salud-y-Protecci-n-Social/Atenciones-en-Urgencias-Municipio-de-Medell-n-2016/ew37-r3ft%0A

Dabbéchi, H., Nabli, A., & Bouzguenda, L. (2016). Towards cloud-based data warehouse as a service for big data analytics. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): Vol. 9876 LNCS* (pp. 180–189). https://doi.org/10.1007/978-3-319-45246-3_17

Ebner, K., Bühnen, T., & Urbach, N. (2014). Think big with big data: Identifying suitable big data strategies in corporate environments. *Proceedings of the*

Annual Hawaii International Conference on System Sciences, 3748–3757. https://doi.org/10.1109/HICSS.2014.466

Ereth, J. (2018). DataOps - Towards a Definition. *CEUR Workshop Proceedings*, *September*.

Europe, A. I., & Foundation, O. K. (2011). *Beyond Access : Open Government Data & the Right to (Re) use Public Information*. *January*, 89. http://www.access-info.org/documents/Access_Docs/Advancing/ Beyond_Access_7_January_2011_web.pdf

Golfarelli, M., & Rizzi, S. (2019). A model-driven approach to automate data visualization in big data analytics. *Information Visualization*. https://doi.org/10.1177/1473871619858933

Gonçalves, A., Portela, F., Santos, M. F., & Rua, F. (2017). Towards of a Real-time Big Data Architecture to Intensive Care. *Procedia Computer Science*, *113*, 585–590. https://doi.org/10.1016/j.procs.2017.08.294

Hevner, A. R., March, S. T., Park, J., & Ram, S. (2008). Design science in information systems research. *Management Information Systems Quarterly*, *28*(1), 6.

Jovanovic, P., Romero, O., & Abelló, A. (2016). A unified view of data-intensive flows in business intelligence systems: A survey. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* (Vol. 10120, pp. 66–107). https://doi.org/10.1007/978-3-662-54037-4_3

Kazman, R., Abowd, G., Bass, L., & Webb, M. (1993). Analyzing the Properties of User Interface Software Architectures. *Computer Science Technical Report CMU-CS-93-201, CMU*.

Khatri, V., & Brown, C. V. (2010). Designing data governance. *Communications of the ACM*, *53*(1), 148–152. https://doi.org/10.1145/1629175.1629210

Krishnan, K. (2013). Data Warehousing in the Age of Big Data. In *Data Warehousing in the Age of Big Data*. https://doi.org/10.1016/C2012-0-02737-8

Madden, S. (2012). From databases to big data. *IEEE Internet Computing*, *16*(3), 4–6.

Mazumder, S. (2016). Big data tools and platforms. In *Big Data Concepts, Theories, and Applications* (pp. 29–128). Springer.

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, *2*(1), 1.

Oussous, A., Benjelloun, F. Z., Ait Lahcen, A., & Belfkih, S. (2018). Big Data technologies: A survey. *Journal of King Saud University - Computer and Information Sciences*, *30*(4), 431–448. https://doi.org/10.1016/j.jksuci.2017.06.001

Pääkkönen, P., & Pakkala, D. (2015). Reference Architecture and Classification of Technologies , Products and Services for Big Data Systems. *Big Data Research*, *2*(4), 166–186. https://doi.org/10.1016/j.bdr.2015.01.001

Passlick, J., Lebek, B., & Breitner, M. H. (2017). A Self-Service Supporting Business Intelligence and Big Data Analytics Architecture. *Proceedings Der 13. Internationalen Tagung Wirtschaftsinformatik (WI 2017)*, 1126–1140. https://doi.org/10.1002/mds.22555

Pollock, R., & Dietrich, D. (2009). CKAN: apt-get for the Debian of Data. *26th Chaos Communication Congress*.

Schlesinger, P., & Rahman, N. (2016). *Self-Service Business Intelligence Resulting in Disruptive Technology SELF-SERVICE BUSINESS INTELLIGENCE RESULTING IN DISRUPTIVE TECHNOLOGY Intel Corporation*. *4417*(March). https://doi.org/10.1080/08874417.2015.11645796

Wan, J., Cai, H., & Zhou, K. (2015). Industrie 4.0: enabling technologies. *Proceedings of 2015 International Conference on Intelligent Computing and Internet of Things*, 135–140.

Wang, Y., Kung, L., & Byrd, T. A. (2018). Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change*, *126*, 3–13.

Zaghloul, M. M., Ali-Eldin, A., & Salem, M. (2015). A process-centric data analytics architecture. *2014 9th International Conference on Informatics and Systems, INFOS 2014*, DEKM34–DEKM39. https://doi.org/10.1109/INFOS.2014.7036705

Zhelev, S., & Rozeva, A. (2017). Big data processing in the cloud - Challenges and platforms. *AIP Conference Proceedings*, *1910*. https://doi.org/10.1063/1.5014007