TRABAJO DE GRADO DE MAESTRÍA

OBSERVATORIO DE PARTICIPACIÓN ELECTORAL Y SU REPRESENTACIÓN EN EL SENADO DE COLOMBIA A TRAVÉS DE UNA BODEGA DE DATOS Y LAS SOLUCIONES DE INTELIGENCIA DE NEGOCIOS

MONITORING OF ELECTORAL PARTICIPATION AND REPRESENTATION IN THE SENATE OF COLOMBIA THROUGH A DATA WAREHOUSE AND BUSINESS INTELLIGENCE SOLUTIONS

Aspirante

VICTOR MANUEL PARRA VARGAS Ing.

Supervisor

CESAR JESÚS PARDO CALVACHE PhD.

UNIVERSIDAD EAFIT

ESCUELA DE INGENIERÍA

DEPARTAMENTO DE INFORMÁTICA Y SISTEMAS

MAESTRÍA EN INGENIERÍA

NOVIEMBRE 2015.

TABLA DE CONTENIDO

TABLA DE ILUSTRACIONES

LISTA DE TABLAS

1. OBJETIVOS.

1.1. General

Desarrollar un observatorio a través de una plataforma web que permita conocer, analizar y gestionar la información relacionada a la participación, ejecución y transparencia de los senadores de la república de Colombia a través de Data Warehouse y business intelligence.

1.2. Específicos

1.2.1. Definir los lineamientos para el levantamiento de información con los datos abiertos de las entidades estatales como Registraduria, Planeación Nacional y Congreso.

1.2.2. Analizar y desarrollar los requerimientos funcionales y no funcionales planteados en el levantamiento de información para el sistema.

1.2.3. Definir una arquitectura para el proceso ETL para la correcta obtención de los datos.

1.2.4. Administrar y ejecutar los componentes que capturan los datos desde su origen hasta llevarlos hasta el repositorio del datamart.

1.2.5. Definir y diseñar una arquitectura robusta y eficiente para la explotación de los datos por parte del usuario final.

1.2.6. Configurar el motor de inteligencia de negocios y del proceso de ETL para la presentación de manera eficiente de los datos.

1.2.7. Elaborar un modelo de presentación amigable en un entorno web que permita la presentación de los análisis.

1.2.8. Preparar la documentación del sistema a implementar.

2. MARCO TEÓRICO Y ESTADO DEL ARTE.

En un artículo de 1958 (Luhn, 1958), el investigador de IBM Hans Peter Luhn utiliza la inteligencia de negocio a largo plazo. Él define la inteligencia como "la capacidad de aprehender las relaciones mutuas de los hechos presentados de una manera tal que orientar la acción hacia una meta deseada."

La inteligencia de negocios, como se entiende hoy en día, se dice que ha evolucionado a partir de los sistemas de apoyo a las decisiones que se iniciaron en la década de 1960 y se desarrollan a lo largo de la década de 1980. DSS se originó en los modelos asistidos por ordenador creados para ayudar en la toma de decisiones y la planificación. Desde DSS, almacenes de datos, sistemas de información ejecutiva, OLAP e inteligencia de negocios entró en foco a partir de finales de los 80.

En 1989, Howard Dresner propuso "inteligencia de negocios" (Dresner, 1989) como un término general para describir "los conceptos y métodos para mejorar la toma de decisiones de negocio mediante el uso de sistemas de apoyo basados en la realidad." No fue sino hasta finales de 1990 que este uso fue generalizado.

Según (Inmon, 1994) considerado por muchos el padre del Data Warehouse, “un Data Warehouse es un conjunto de datos orientados por temas, integrados, variantes en el tiempo y no volátiles, que tienen por objetivo dar soporte a la toma de decisiones.” (Llorente, 2012)

Ilustración 1 Enfoque Inmon - DW Corporativo

Según Ralph Kimball (Kimball, 2002) (considerado el principal promotor del enfoque dimensional para el diseño de almacenes de datos), "un Data Warehouse es una copia de los datos transaccionales específicamente estructurada para la consulta y el análisis".

Ilustración 2 Metodologia (KIMBALL, 2002)

Antes de comenzar nuestro proyecto de BI, vamos a determinar qué tipo de metodología vamos a utilizar. Existen diferentes métodos, todos relacionados con el ámbito del despliegue de sistemas de información, con alguna concreción referente a los sistemas de BI y DW.

Cuando diseñamos la arquitectura de un sistema de Data Warehouse nos hemos de plantear los diferentes entornos por los que han de pasar los datos en su camino hacia su Datamart o cubo de destino. Dada la cantidad de transformaciones que se han de realizar, y que normalmente el DWH, además de cumplir su función de soporte a los requerimientos analíticos, realiza una función de integración de datos que van a conformar el Almacén Corporativo y que van a tener que ser consultados también de la manera tradicional por los sistemas operacionales, es muy recomendable crear diferentes áreas de datos en el camino entre los sistemas origen y las herramientas OLAP.

Cada una de estas áreas se distinguirá por las funciones que realizan, de qué manera se organizan los datos en la misma y a qué tipo de necesidad pueden dar servicio. El área que se encuentra 'al final del camino' es importante, pero no va a ser la única que almacene los datos que van a explotar las herramientas de reporting.

Tampoco hay una convención estándar sobre lo que abarca exactamente cada área y la obligatoriedad de utilizar cada una de ellas. Cada proyecto es un mundo, e influyen muchos factores como la complejidad, el volumen de información del mismo, si realmente se quiere utilizar el Data Warehouse como almacén corporativo o Sistema Maestro de Datos, o si existen necesidades reales de soporte al reporting operacional.

Hoy en día, estamos en el auge del movimiento de la llamada inteligencia de negocio o Business Intelligence (BI). Casi todas las organizaciones se esfuerzan por crear y mejorar sus procesos y sistemas de toma de decisión. Una gran cantidad de nuevos proyectos de BI aparecen constantemente, pero la experiencia global en los últimos años no es tan buena. Algo por lo general va mal en la ejecución de proyectos de BI, ya que la mayoría de proyectos de BI (85%) no pudo lograr sus objetivos.

En este apartado vamos a revisar muy brevemente los principales enfoques metodológicos existentes para proyectos BI, así como sus principales limitaciones.

Con el Observatorio Electoral se pretende automatizar algunos de los procesos que se manejan en la Registraduria Nacional, con la finalidad de minimizar los tiempos, obtener automáticamente información actualizada y visualizarla en reportes dinámicos, otra se tomara del API del congreso visible.

Los procesos automáticos para obtener la información de elecciones para el senado serán diseñados con la herramienta de software libre Pentaho, para un fácil acceso a los usuarios y mejorar la calidad y credibilidad en la información publicada.

Con la aplicación de sistemas de Inteligencia de Negocios y procesos ETL se obtendrá:

Para la redundancia en datos:

OLAP: son bases de datos orientadas al procesamiento de transacciones.

Propiedad de la redundancia	**Efectos**	
	Base OLTP	**Data Warehouse**
Acelerar consultas con datos precalculados	X	X
Mayor tiempo en actualización con datos	-	-
Mayor probabilidad de generar inconsistencias	-	-
Mejorar controles	X	

Tabla 1 Cuadro comparativo de redundancia de datos

En obtención de datos como información

Datos Operacionales	Datos Informativos
* Orientados a una aplicación	* Orientados a un tema
* Integración limitada	* Integrados
* Constantemente actualizados	* No volátiles
* Solo valores actuales	* Valores a lo largo del tiempo
* Soportan operaciones diarias	* Soportan decisiones de administración

Tabla 2 Cuadro comparativo de los datos

Para el propósito de un Datamart, el modelo relacional (ER) presenta los siguientes problemas:

- Legibilidad limitada. Los usuarios finales no son capaces de entender el modelo ER. Por tanto, no pueden "navegar" por dicho modelo en busca de información.
- Dificultad para las herramientas de consulta en el acceso a un modelo ER general. Las herramientas de consulta a menudo poseen resultados pobres o inaceptables cuando se trabaja en entornos relacionales de grandes volúmenes de información.
- La utilización de la técnica de modelado ER frustra el principal atractivo del Data Mart, porque se impide la recuperación de información intuitiva y con alto rendimiento.
- Actualmente Colombia no cuenta con herramientas de software libre que permitan el acceso a todos los usuarios interesados y autorizados para la consulta de la información de votación.
- Los procesos empresariales pueden ser optimizados. El tiempo perdido esperando por información que finalmente es incorrecta o no encontrada, es eliminado y el acceso a esta información será eficiente y efectivo.

2.1. SISTEMAS DE INTELIGENCIA DE NEGOCIOS - BI

2.1.1. ¿QUÉ ES UN SISTEMA BI?

El Sistema de Inteligencia de Negocios es un software que tiene como objetivo convertir los datos de una empresa en información que sirva como conocimiento y crear un valor competitivo ante el mercado (Co., 2000), Básicamente, el objetivo de Business Intelligence es apoyar de forma razonable y continua a las organizaciones para mejorar su competitividad, facilitando la información necesaria en el momento adecuado para la toma de decisiones. Howard Dresner fue el primero que acuñó el término cuando era consultor de Gartner, popularizó Business Intelligence o BI como un término para describir un conjunto de conceptos y métodos que mejoraran la toma de decisiones, utilizando información sobre hechos.

2.1.2. ¿QUÉ PUEDE HACER CON UN SISTEMA BI?

Con un Sistema BI:

- Diseñar reportes para departamentos, áreas o globales en una empresa
- Generar una base de datos para clientes
- Dependiendo para la toma de decisiones crear escenarios
- Compartir entre áreas o departamentos de una empresa la información
- Estudios de diseños multidimensionales
- Extraer, transformar y procesar datos

- Dar un nuevo enfoque en la toma de decisiones
- Mejorar la calidad de servicio al cliente

Estos Sistemas de Inteligencia de Negocios permiten agrupar, analizar, escoger y transformar los datos que se ingresan en operaciones constantes que pueden ser: orden, municipios, votos, partidos y candidatos con información estructurada y convertirla en conocimiento que ayude en la toma de decisiones al momento de decidir el voto.

2.1.3. BENEFICIOS DE UN SISTEMA BI

Los beneficios que aporta un Sistema de Inteligencia de Negocios, es la unión de varias fuentes de información, crear perfiles de usuarios para el manejo de la información, disminuir la dependencia del Departamento de Sistemas, la reducción en los tiempos de obtención de la información, mejorar el análisis, la disponibilidad de la información en el momento en que se necesita y los criterios que se ajustan al estado actual de la empresa.

Cuando diseñamos la arquitectura de un sistema de Data Warehouse nos hemos de plantear los diferentes entornos por los que han de pasar los datos en su camino hacia su Data Mart o Cubo de Datos destino. Dada la cantidad de transformaciones que se han de realizar, y que normalmente el Data Warehouse, además de cumplir su función de soporte a los requerimientos analíticos, realiza una función de integración de datos que van a conformar el Almacén Corporativo y que van a tener que ser consultados también de la manera tradicional por los sistemas operacionales, es muy recomendable crear diferentes áreas de datos en el camino entre los sistemas origen y las herramientas OLAP.

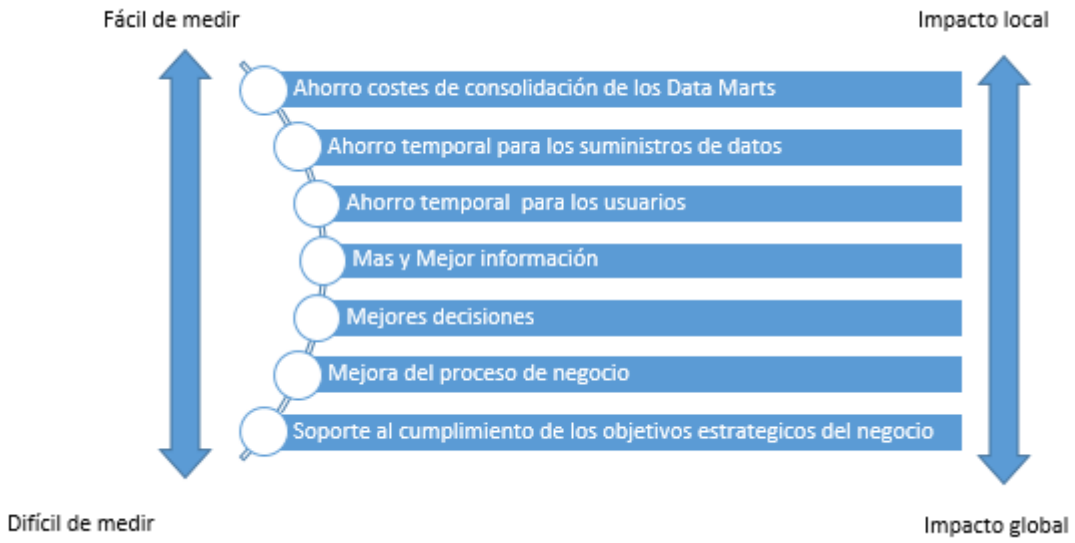


Ilustración 3 Los beneficios que aporta un Sistema de Inteligencia de Negocios

2.1.4. PLATAFORMA DE UN SISTEMA DE INTELIGENCIA DE NEGOCIOS

2.1.4.1. PENTAHO

Durante este trabajo se buscó una herramienta para la solución de inteligencia de negocios que cumpliera con las siguientes características:

- Adaptabilidad a todos los entornos.
- Sin costos de licencias
- Independencia del proveedor.
- Garantías de permanencia. Upgrades y mejoras frecuentes.
- Tendencia a la calidad y utilización de estándares (XML – MDX)

- Dedicación full a los servicios y a la solución de negocios.
- Gran aprovechamiento de Web Services.
- La comunidad del software libre.

AL buscar en el mercado de soluciones de inteligencia de negocios, se encontraron muchos productos que cumplían con algunas de estas características y tenían asistentes muy prácticos e integrados al manejador de bases de datos donde la tarea de extraer, trasformar y cargar era casi automática, en este ejercicio se debe contar con una herramienta libre por su uso.

Pentaho es un software de código, libre comercial para el "business intelligence" (BI) o "Inteligencia de Negocios".

Pentaho Open BI Suite proporciona reporting intuitivo, análisis OLAP, cuadros de mando, integración de datos, minería de datos y Plataforma BI. Esta suite se ha convertido en la líder mundial y la más ampliamente utilizada como herramientas BI de código libre.

El modelo de negocios de software libre y comercial de Pentaho proporciona soporte, servicios y mejoras del producto vía suscripciones anuales. Los módulos de la plataforma Pentaho son:

2.1.4.2. REPORTING

Un módulo de los informes: Ofrece una herramienta de diseño de reportes, informes ágiles y de gran capacidad acorde a las necesidades de los usuarios, solución basada en FreeReport.

Pentaho Reporting

Es una herramienta libre de acceso y en cualquier prestación de una manera fácil y rápida.

Es un editor basado en herramientas profesionales con la capacidad de desarrollo de informes a las necesidades que presente el negocio en estudio o destinado a desarrolladores.

Pentaho Reporting nos muestra de manera ordenada los resultados del análisis pudiendo imprimir o exportar en varios formatos (PDF, XLS, HTML y texto). Los reportes Pentaho permiten también programación de tareas y ejecución automática de informes con una determinada periodicidad.

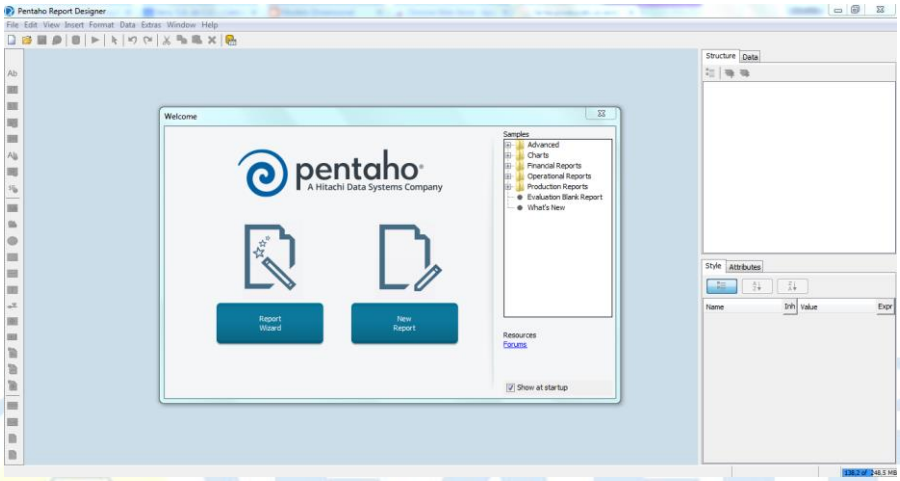


Ilustración 4 Captura Pantalla Pentaho Report Designer

2.1.4.3. ANÁLISIS

Un sistema avanzado de análisis de información para el uso de los usuarios. Con uso de las tablas dinámicas, generadas por herramienta Mondrian y JPivot, el usuario puede navegar y consultar los datos según lo requiera y ajustarlo a la necesidades y visión de los datos, los filtros de visualización, añadiendo o quitando los campos de agregación. Los datos pueden ser representados en una forma de SVG o Flash, los dashboards widgets, o también integrados con los sistemas de minería de datos y los portales web (portlets). Además, con el Microsoft Excel Analysis Services, se puede analizar los datos dinámicos en Microsoft Excel (usando la conexión a OLAP server Mondrian).

Pentaho Analysis

Este módulo ayuda a trabajar con máxima efectividad para ganar tiempo, sutileza y entender lo suficiente y poder tomar decisiones.

Las características generales son:

- Vista multidimensional de datos (por municipios, por candidatos, por partidos, etc.).
- Navegar y explorar (Análisis Ad Hoc, Drill-Down, etc.).
- Interactuar con alto rendimiento mediante tecnologías optimizadas para la rápida respuesta interactiva.

Pentaho Worbench

Para un análisis multidimensional, Pentaho tiene en su plataforma BI una módulo ROLAP a través de lo que llaman Pentaho Analysis Services. PAS está basado en Mondrian, que es el corazón de este, y en Jpivot, que es la herramienta de análisis de usuario, con el que realizamos la navegación multidimensional sobre los Cubos de Datos desde la plataforma BI y visualizamos los resultados de las consultas. Estas son ejecutadas por Mondrian, que traduce los resultados relacionales a resultados multidimensionales, que a su vez son mostrados al usuario en formato Html por Jpivot

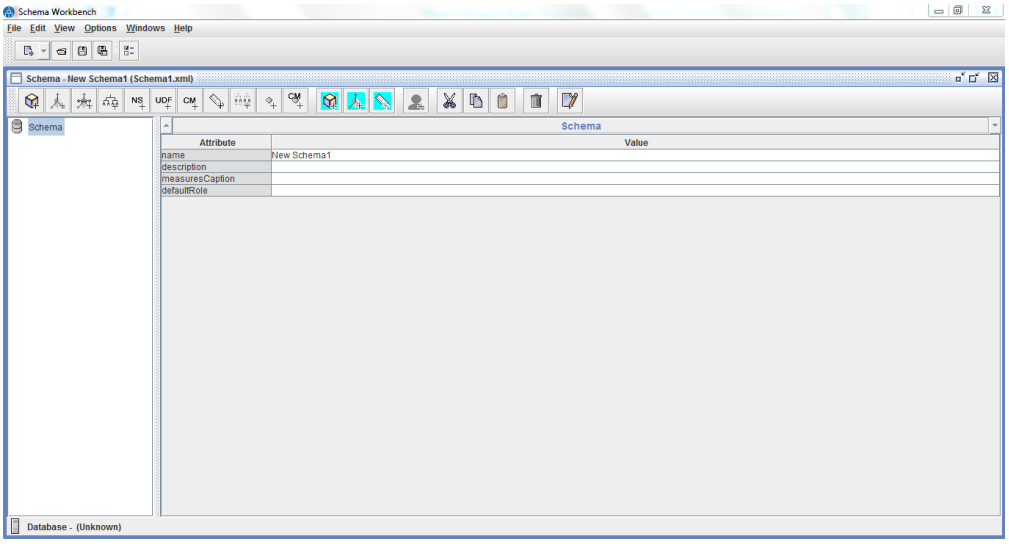


Ilustración 5 Schema Workbench

2.1.4.4. DASHBOARDS

Todos los componentes del módulo Pentaho Reporting y Pentaho Análisis pueden formar parte de un Dashboard. En Pentaho Dashboards es muy fácil de incluir una variedad de tipos de gráficos, tablas y velocímetros (dashboard widgets) e integrarlos con los Portlets JSP, en donde podrá visualizar informes, gráficos y análisis OLAP.

2.1.4.5. DATA MINING

Análisis en Pentaho: Se realiza con una herramienta WeKa.

Es un grupo de técnicas y tecnologías que admiten investigar bases de datos grandes de una manera automática o semiautomática, con el propósito de encontrar patrones en la información, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Básicamente, el Data Mining una herramienta que nos permite comprender toda la información que se encuentra en un repositorio de datos. Con este fin, hace uso de prácticas estadísticas y, en algunos casos, de algoritmos de búsqueda próximos a la Inteligencia Artificial y a las redes neuronales.

2.1.4.6. INTEGRACIÓN DE DATOS

Se realiza con una herramienta Kettle ETL (Pentaho Data Integration) que permite implementar los procesos ETL. Últimamente Pentaho lanzó una nueva versión que marcó un gran paso adelante en OSBI ETL y que hizo de Pentaho Data Integration una alternativa interesante para las herramientas comerciales.

Pentaho Data Integration

Kettle es un proyecto Open Source que incluye un conjunto de herramientas para realizar ETL ahora forma parte de la suite de Inteligencia de Negocios Pentaho y este a su vez consiste principalmente de las aplicaciones Spoon, Pan,kitchen.

Spoon

Herramienta gráfica de desarrollo para realizar las transformaciones y trabajos ETL una vez diseñados serán ejecutados ya sea manual o automáticamente, a su vez permite organizar las

trasformaciones en trabajos. Contiene un editor de consultas SQL para facilitar la creación de los datos que serán utilizados en los diferentes informes e indicadores.

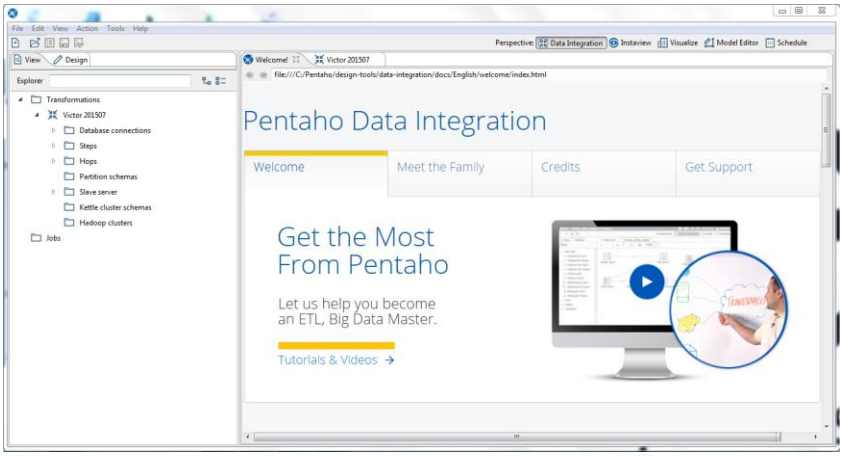


Ilustración 6 Spoon

2.1.5. DATA WAREHOUSE

Un Data Warehouse es una base de datos corporativa que se caracteriza por integrar y depurar información de una o varias fuentes distintas, con un análisis exhaustivo con varias perspectivas y las procesa a mejores velocidades de respuesta. La creación de un Data Warehouse representa en la mayoría de las ocasiones el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de Inteligencia de Negocios.

La ventaja principal de este tipo de bases de datos radica en las estructuras en las que se almacena la información (modelos de tablas en estrella, en copo de nieve, cubos relacionales, etc). Este tipo de persistencia de la información es homogénea y fiable, y permite la consulta y el tratamiento jerarquizado de la misma (siempre en un entorno diferente a los sistemas operacionales). Un Data Warehouse se caracteriza por ser:

Integrado

En un Data Warehouse los datos almacenados deben componer una estructura consistente, por las debilidades existentes entre los algunos sistemas operacionales los cuales debe ser eliminada. La información también se estructura en distintos niveles de referencia para ajustar a las necesidades de los usuarios.

Temático

Sólo datos importantes en el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por niveles para ayudar al acceso y fácil entendimiento por parte de los usuarios finales. Por ejemplo, todos los datos sobre clientes pueden ser consolidados en una única tabla del Data Warehouse. De esta forma, las peticiones de información sobre clientes serán más fáciles de responder dado que toda la información reside en el mismo lugar.

Histórico

El tiempo es parte implícita de la información contenida en un Data Warehouse. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada en el Data Warehouse sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el Data Warehouse se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

No volátil

El almacén de información de un Data Warehouse existe para ser leído, pero no modificado. La información es por tanto permanente, significando la actualización del Data Warehouse la incorporación de los últimos valores que tomaron las distintas variables contenidas en él sin ningún tipo de acción sobre lo que ya existía.

2.1.6. DATA MART

Un Data Mart es una base de datos que se aplica a un departamento específico de una empresa, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento. Un Data Mart puede ser alimentado desde los datos de un Data Warehouse, o integrar por sí mismo un compendio de distintas fuentes de información.

Los Data Mart que están dotados con estas estructuras óptimas de análisis presentan las siguientes ventajas:

- Menos cantidad de datos que un DW
- Mayor velocidad en las consulta
- Consultas sencillas de SQL

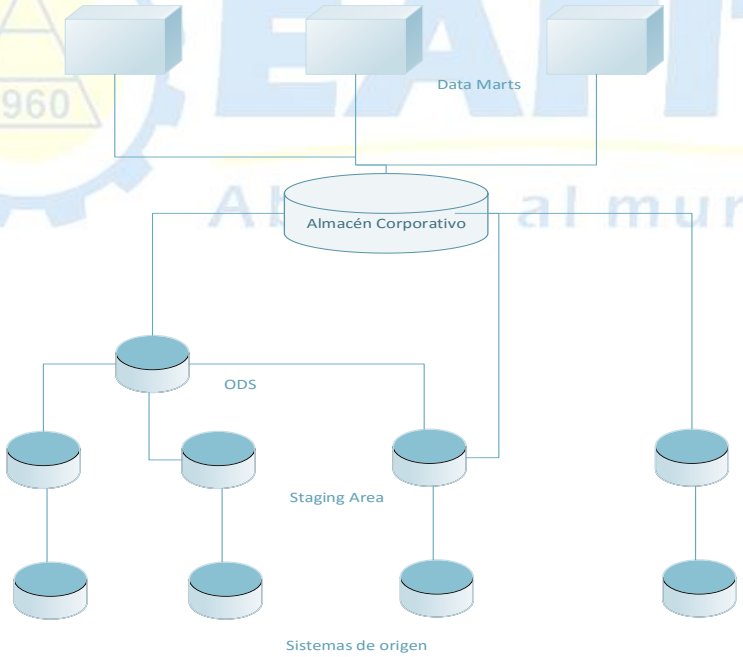


Ilustración 7 Arquitectura de un Data WareHouse

2.1.7. DATA WAREHOUSE VS DATA MART

Un Data Mart es una aplicación de Data Warehouse, construida rápidamente para soportar una línea de negocio simple. Los Data Mart, tienen las mismas características de integración, no volatilidad y orientación temática que el Data Warehouse. Representan una estrategia de "divide y vencerás" para ámbitos muy genéricos de un Data Warehouse.

2.1.7.1. TIPOS DE DATA MART

2.1.7.1.1. DATA MART DEPENDIENTES

Son los que se forman a partir de los Data Warehouse central, es decir reciben sus datos de un repositorio empresarial central.

2.1.7.1.2. DATA MART INDEPENDIENTES

Son aquellos que pueden recibir los datos directamente del ambiente operacional y no dependen de un Data Warehouse central, ya sea mediante procesos internos de las fuentes de datos o de almacenes de datos operacionales.

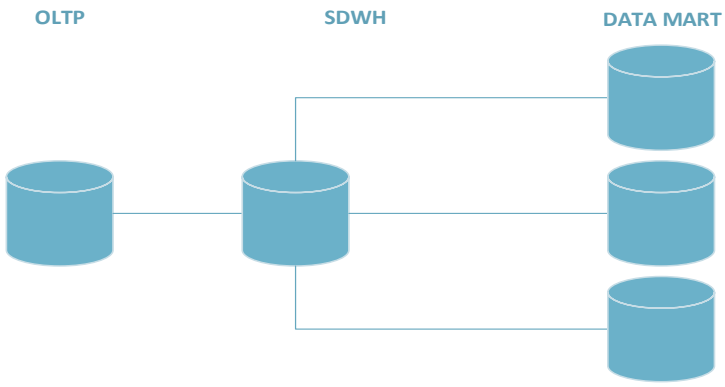


Ilustración 8 Data Warehouse central

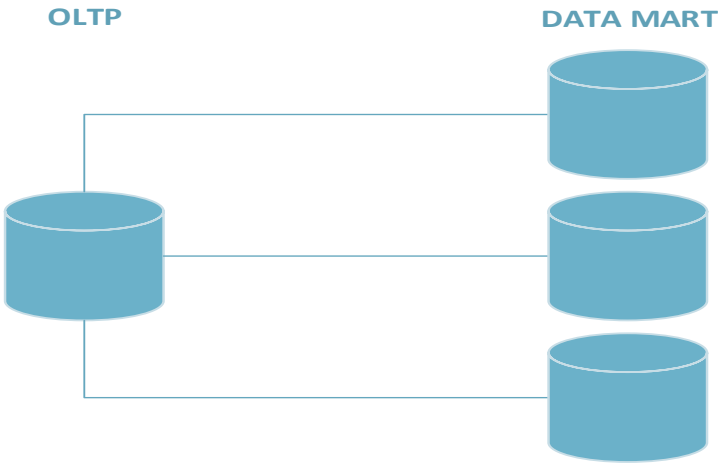


Ilustración 9 Data Mart –Independiente

2.1.8. ETL

ETL son las siglas en inglés de Extraer, Transformar y Cargar (Extract, Transform and Load). Es el proceso que admite a las empresas mover los datos desde algunas fuentes, transformarlos, eliminar datos que no aporten al estudio, y cargarlos en otra base de datos, Data Mart, o Data Warehouse o en otro sistema operacional para para analizarlos y apoyar un proceso de negocio.

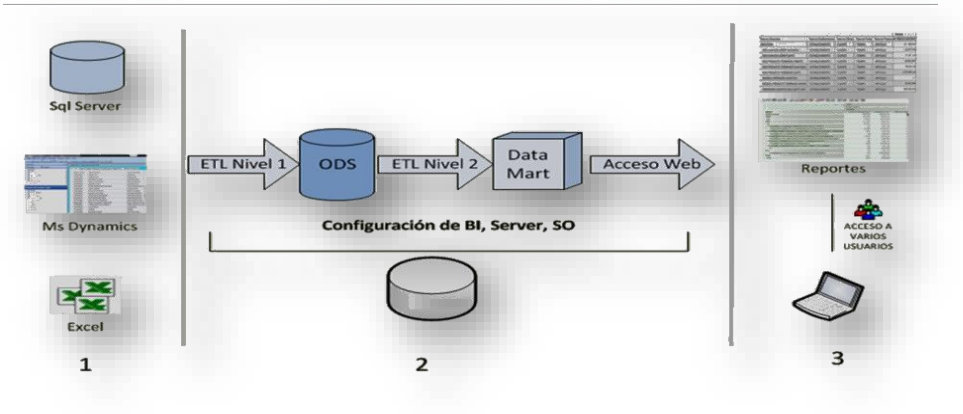


Ilustración 10 Diseño y Construcción -Proceso ETL

El proceso ETL

- Determina el éxito o fracaso de la implementación de un Data Mart.
- El proceso ETL agrega un valor significativo a los datos.
- Los subproceso que se manejan en el proceso ETL son:

2.1.8.1. EXTRACCIÓN

Este proceso trata de obtener los datos del sistema origen, realizando pasos completos o incrementales, actúa como pasarela entre los sistemas fuente origen y los sistemas destino, y cuyo objetivo principal radica en impedir la saturación de los servidores funcionales de la organización.

2.1.8.2. TRANSFORMACIÓN

Los datos originarios de bases distintas no suelen coincidir en formato. Por tanto, para conseguir integrarlos resulta necesario realizar operaciones de transformación. El objetivo no es otro que evitar duplicidades innecesarias e impedir la generación de fuentes de datos aislados.

2.1.8.3. CARGA

Se trata de cargar los datos, con un formato deseado y común dentro del sistema destino. Para la carga masiva de datos suele ser necesario desactivar temporalmente la integridad referencial de la base de datos destino.

Las más populares herramientas y aplicaciones ETL del mercado:

- IBM Websphere DataStage (anteriormente Ascential DataStage y Ardent DataStage)

- Pentaho Data Integration (Kettle ETL)
- SAS ETL Studio
- Oracle Warehouse Builder
- Informática PowerCenter
- Cognos Decisionstream
- Ab Initio
- BusinessObjects Data Integrator (BODI)
- Microsoft SQL Server Integration Services (SSIS)

2.1.9. REPORTES

Los reportes brindan beneficios como: la rapidez en la toma de decisiones inteligentes y acertadas, una mejor distribución de la información, etc. y se pagan por si solos sobre la marcha. Tanto los usuarios finales como el Departamento de Sistemas tienen fácil acceso a los datos, no importa dónde estos se encuentren. Los reportes traen múltiples análisis para obtener más respuestas a partir de los datos ofreciendo una distribución rápida y fácil de los reportes y sus resultados.

2.1.10. BASE DE DATOS OLTP Y OLAP

2.1.10.1. OLTP

Los sistemas OLTP son bases de datos encaminadas al procesamiento de transacciones. Una transacción genera un proceso elemental (que debe ser validado con un commit, o invalidado con

llback), y que puede involucrar operaciones de inserción, modificación y borrado de datos. El proceso transaccional es propio de las bases de datos operacionales.

- Se constituyen los datos según el nivel de aplicación (programa de gestión a medida, ERP o CRM implantado, sistema de información departamental.).
- El acceso a los datos está optimizado para tareas frecuentes de lectura y escritura. (Por ejemplo, la enorme cantidad de transacciones que tienen que soportar las BD de bancos o hipermercados diariamente).
- El historial de datos saben limitarse a los datos actuales o recientes.
- Los formatos de los datos no son necesariamente uniformes en los diferentes departamentos (es común la falta de compatibilidad y la existencia de fuentes de datos).

2.1.10.2. OLAP

Los sistemas OLAP son bases de datos encaminadas al procesamiento analítico. Este análisis implica, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil: tendencias de votos, patrones de comportamiento de los electores, elaboración de informes complejos, etc. Este sistema es típico de los Data Mart.

Se estructuran los datos según las áreas de negocio, y los formatos de los datos están integrados de manera uniforme en toda la organización.

El acceso a los datos suele ser de sólo lectura. La acción más común es la consulta, con muy pocas inserciones, actualizaciones o eliminaciones.

El historial de datos es a largo plazo, normalmente de dos a cinco años.

Las bases de datos OLAP se suelen cargar la información procedente de los sistemas operacionales existentes, mediante un proceso de extracción, transformación y carga (ETL).

2.1.11. MODELO ESTRELLA

Es una arquitectura de almacén de datos simple. En este diseño del almacén de datos la tabla de variables (hechos) está rodeada por dimensiones y juntos forman una estructura que permite implementar mecanismos básicos para poder utilizarla con una herramienta de consultas OLAP.

El objetivo de un modelo estrella, es optimizar el tiempo de respuesta de Base Datos y dar información a un usuario en menos tiempo posible para dejar de mantener las tablas en el modelo relacional y permitir el almacenamiento de información redundante. En este modelo, para obtener información solicitada no hay que construir una sentencia SQL muy compleja que enlace muchas tablas a la vez. Una herramienta de consultas sólo tiene que acceder una tabla.

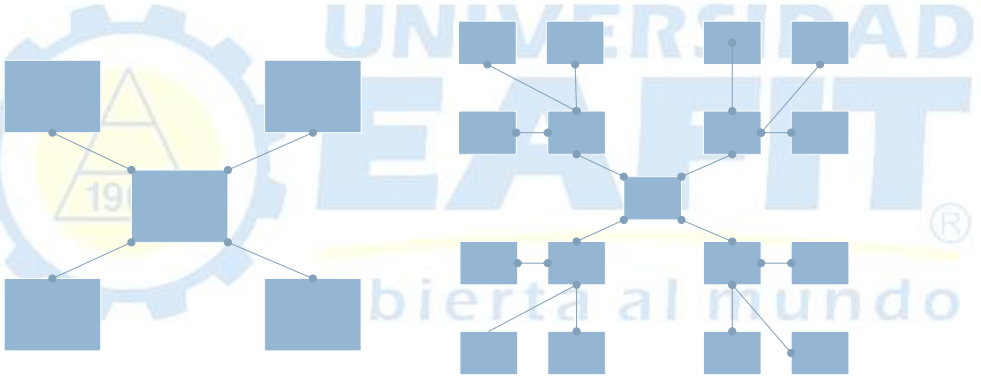


Ilustración 11 Modelos Estrella y Copo de Nieve

2.1.12. DIMENSIONES

Las dimensiones de un Cubo de Datos son atributos referentes a las variables, son las vistas de análisis de las variables (forman parte de la tabla de dimensiones). Son inventarios de información complementaria necesaria para la presentación de los datos a los usuarios, como por

ejemplo: descripciones, nombres, zonas, rangos de tiempo, etc. Es decir, la información general complementaria a cada uno de los registros de la tabla de hechos.

La clave primaria de una tabla de hechos está formada por todas las columnas que corresponden a las dimensiones.

2.1.13. VARIABLES

También llamadas "indicadores de gestión", son los datos que están siendo analizados. Forman parte de la tabla de hechos. Más formalmente, las variables representan algún aspecto cuantificable o medible los objetos o eventos analizar. Normalmente, las variables son representadas por valores detallados y numéricos para cada instancia del objeto o evento medido. En forma contraria, las dimensiones son atributos relativos a las variables, y son utilizadas para indexar, ordenar, agrupar o abreviar los valores de las mismas. Las dimensiones poseen una granularidad menor, tomando como valores un conjunto de elementos menor que el de las variables; ejemplos de dimensiones podrían ser: "productos", "localidades" (o zonas), "el tiempo" (medido en días, horas, semanas, etc.).

3. IDENTIFICACIÓN DE VARIABLES

3.1. VARIABLES INDEPENDIENTES

3.1.1. Reglas y Estándares

Normas de la organización que posiblemente pueda afectar al desempeño del proyecto y se encuentran fuera del alcance del mismo.

3.1.2. Organización

Por la falta de organización en la información se puede ver afectado el momento de tomar las correctas y acertadas de decisiones.

3.1.3. Tiempo

Magnitud física que se utiliza para realizar la medición de cuánto dura algo que es idóneo de cambios que se presenten; el tiempo permite ordenar los sucesos o actividades en secuencias.

3.2. VARIABLES DEPENDIENTES

3.2.1. Integración

El conjunto de varias fuentes de datos origen contenidas en los datos abiertos del gobierno entregados por la registradora, el resultado final debe ser un esquema único e imagen colectiva.

3.2.2. Eficiencia

Expresión que se emplea para medir la capacidad o cualidad de actuación de un producto o servicio, para lograr el cumplimiento de objetivos determinados, minimizando el uso de recursos.

3.2.3. Productividad

Termino que se utiliza para calcular el nivel de efectividad con el que se miden los logros y objetivos que se han alcanzado en el tiempo, costo y calidad de las tareas o actividades siendo complicado el manejo, localización y análisis de un gran número de fuentes de información.

3.3. VARIABLES INTERVINIENTES

3.3.1. Proceso ETL

Son probablemente los componentes más importantes y de mayor valor añadido en una infraestructura que implique la integración de varias fuentes de datos, especialmente se requiere mucha precisión y actualización en los datos.

4. MARCO METODOLÓGICO

La propuesta es tomar una metodología ágil como Scrum para afrontar los desarrollos "por partes". Es decir dividir el proceso verticalmente e ir desarrollando desde arriba hasta abajo sucesivamente dibujando una espiral, este es muy parecido a la metodología Data-Driven Metodology para BI.

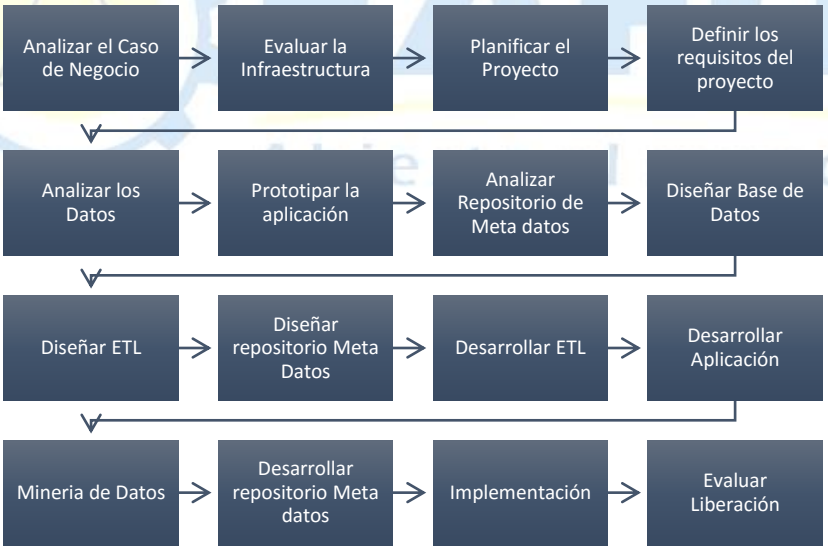


Ilustración 12 Ruta del Ciclo de vida para la implementación del Observatorio.

Todo esto debe tener la gestión de proyectos agiles como un eje trasversal, tomando tiempo, alcance y costo como líneas bases sin dejar a un lado los recursos humanos, stakeholders, riesgos y calidad.

4.1. Planificación

En este proceso se determina el propósito del proyecto de DW/BI, sus objetivos específicos y el alcance del mismo, los principales riesgos y una aproximación inicial a las necesidades de información. En la visión de programas y proyectos de Kimball, Proyecto, se refiere a una interacción simple del KLC (Kimball Life Cycle) (Kimball, 2002), desde el lanzamiento hasta el despliegue.

Esta tarea incluye las siguientes acciones típicas de un plan de proyecto:

- Definir el alcance (entender los requerimientos del negocio).
- Identificar las tareas.
- Programar las tareas.
- Planificar el uso de los recursos.
- Asignar la carga de trabajo a los recursos.
- Elaboración de un documento final que representa un plan del proyecto.

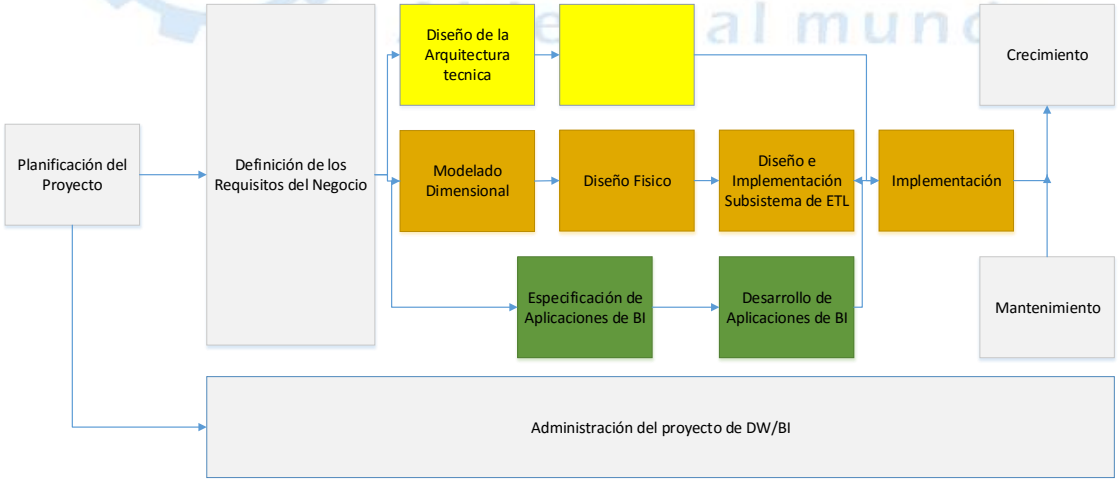


Ilustración 13 Kimball Lifecycle Methodology

Además en esta parte definimos cómo realizar la administración o gestión de esta sub fase que es todo un proyecto en sí mismo, con las siguientes actividades:

- Monitoreo del estado de los procesos y actividades.
- Rastreo de problemas.

4.2. Análisis de requerimientos:

La definición de los requerimientos es en gran medida un proceso de entrevistar al personal de negocio y técnico en este caso tomamos a estudiantes de ciencias políticas de la universidad externado de Colombia y personas tomadas al azar para ser consultadas sobre sus necesidades de información al momento de elegir a sus senadores, pero siempre conviene tener un poco de preparación previa. Se debe aprender tanto como se pueda sobre el negocio, los competidores, la industria y los clientes del mismo. Hay que leer todos los informes posibles de la organización; rastrear los documentos de estrategia interna; entrevistar a los empleados, analizar lo que se dice en la prensa acerca de la organización, la competencia y la industria. Se deben conocer los términos y la terminología del negocio.

Parte del proceso de preparación es averiguar a quién se debe realmente entrevistar. Esto normalmente implica examinar cuidadosamente el organigrama de la organización. Hay básicamente cuatro grupos de personas con las que hablar desde el principio: el directivo responsable de tomar las decisiones estratégicas; los administradores intermedios y de negocio responsables de explorar alternativas estratégicas y aplicar decisiones; personal de sistemas, si existen, la gente que realmente sabe qué tipos de problemas informáticos y de datos existen; y por último, la gente que se necesita entrevistar por razones políticas.

A partir de las entrevistas, podemos identificar temas analíticos y procesos de negocio. Los temas analíticos agrupan requerimientos comunes en un tema común. Por otra parte, a partir del análisis

se puede construir una herramienta de la metodología denominada matriz de procesos/dimensiones (Bus Matrix en inglés).

Una dimensión es una forma o vista o criterio por medio de cual se pueden sumariar, cruzar o cortar datos numéricos a analizar, datos que se denominan medidas (measures en inglés).

Esta matriz tiene en sus filas los procesos de negocios identificados, y en las columnas, las dimensiones identificadas.

4.3. Modelado Dimensional

La creación de un modelo dimensional es un proceso dinámico y altamente iterativo.

El proceso de diseño comienza con un modelo dimensional de alto nivel obtenido a partir de los procesos priorizados de la matriz descrita en el punto anterior.

El proceso iterativo consiste en cuatro pasos:

1. Elegir el proceso de negocio.
2. Establecer el nivel de granularidad.
3. Elegir las dimensiones.
4. Identificar medidas y las tablas de hechos.

4.3.1. Elegir el proceso de negocio

El primer paso es elegir el área a modelar. Esta es una decisión que sale del trabajo investigativo del autor, en las organizaciones las decisiones la toma el comité del negocio, y depende fundamentalmente del análisis de requerimientos y de los temas analíticos anotados en la etapa anterior.

Establecer el nivel de granularidad

La granularidad significa especificar el nivel de detalle. La elección de la granularidad depende de los requerimientos del negocio y lo que es posible a partir de los datos actuales. La sugerencia general es comenzar a diseñar el DW al mayor nivel de detalle posible, ya que se podría luego realizar agrupamientos al nivel deseado. En caso contrario no sería posible abrir (drill-down) las sumarizaciones en caso de que el nivel de detalle no lo permita.

Elegir las dimensiones

Las dimensiones surgen naturalmente de las discusiones del equipo, y facilitadas por la elección del nivel de granularidad y de la matriz de procesos/dimensiones. Las tablas de dimensiones tienen un conjunto de atributos (generalmente textuales) que brindan una perspectiva o forma de análisis sobre una medida en una tabla hechos. Una forma de identificar las tablas de dimensiones es que sus atributos son posibles candidatos para ser encabezado en los informes, tablas pivot, cubos, o cualquier forma de visualización, unidimensional o multidimensional.

Identificar las tablas de hechos y medidas

El último paso consiste en identificar las medidas que surgen de los procesos de negocios. Una medida es un atributo (campo) de una tabla que se desea analizar, sumarizando o agrupando sus datos, usando los criterios de corte conocidos como dimensiones. Las medidas habitualmente se vinculan con el nivel de granularidad, y se encuentran en tablas que denominamos tablas de hechos (fact en inglés). Cada tabla de hechos tiene como atributos una o más medidas de un proceso organizacional, de acuerdo a los requerimientos. Un registro contiene una medida expresada en números, como ser cantidad, tiempo, dinero, etc., sobre la cual se desea realizar una operación de agregación (promedio, conteo, suma, etc.) en función de una o más dimensiones. La granularidad es el nivel de detalle que posee cada registro de una tabla de hechos.

Modelo gráfico de alto nivel

Para concluir con el proceso dimensional inicial se realiza un gráfico denominado modelo dimensional de alto nivel (o gráfico de burbujas, Bubble chart, en el léxico de Kimball).

Identificación de atributos de dimensiones y tablas de hechos

La segunda parte de la sesión inicial de diseño consiste en completar cada tabla con una lista de atributos bien formada. Esta lista o grilla se forma colocando en las filas los atributos de la tabla, y en las columnas la siguiente información:

- Características relacionadas con la futura tabla dimensional del almacén de datos (target), por ejemplo tipo de datos, si es clave primaria, valores de ejemplo, etc. Por razones de espacio no describiremos todas las columnas, para mayor información puede consultarse la referencia (Rivadera, 2014).
- El origen de los datos (source, por lo general atributos de las tablas transaccionales).
- Reglas de conversión, transformación y carga (ETL rules), que nos dicen cómo transformar los datos de las tablas de origen a las del almacén de datos.

Implementar el modelo dimensional detallado

Este proceso consiste simplemente en completar la información incompleta de los pasos anteriores. El objetivo en general es identificar todos los atributos útiles y sus ubicaciones, definiciones y reglas de negocios asociadas que especifican cómo se cargan estos datos. Para este cometido se usa la misma planilla del punto anterior.

Prueba del modelo

Si el modelo ya esta estable, lo que se hace habitualmente es probarlo contra los requerimientos del negocio. Haciendo la pregunta práctica de ¿Cómo podemos obtener esta información en particular del modelo? Para las pruebas podemos usar diseños de reportes estructurados, de usuarios actuales, diseños de cubos prospectivos, etc.

Revisión y validación del modelo

Un vez que tenemos confianza plena en el modelo, ingresamos en esta etapa final, lo cual implica revisar el modelo con diferentes muestras de la audiencias, cada una con diferentes conocimientos técnicos y del evento electoral. En el área de sistemas deberían revisarlo los programadores y analistas de los sistemas, y el DBA si existe con una reunión de revisión tipo scrum. También debería revisarse con un senador o estudiantes de ciencias políticas que tengan mucho conocimiento de los procesos electorales y que quizás no hayan participado del diseño del modelo. Finalmente podemos hacer un documento que enuncie una serie de preguntas del proceso electoral (tomadas a partir de los requerimientos), y las conteste por medio del modelo.

Documentos finales

El producto final, son una serie de documentos (solo mencionamos los más importantes), a saber:

- Modelo de datos inicial de alto nivel
- Lista de atributos
- Diagrama de tablas de hechos
- Definición de campos de medida
- Diagrama de tablas de dimensiones
- Descripción de los atributos de las dimensiones
- Matriz DW (o DW Bus Matrix) completa
- Tabulación de encuesta de validación.
- Conclusiones de la reunión de revisión.

Diseño Físico

En esta parte, intentamos contestar las siguientes preguntas:

- ¿Cómo puede determinar cuán grande será el sistema de DW/BI?

- ¿Cuáles son los factores de uso que llevarán a una configuración más grande y más compleja?
- ¿Cómo se debe configurar el sistema?
- ¿Cuánta memoria y servidores se necesitan? ¿Qué tipo de almacenamiento y procesadores?
- ¿Cómo instalar el software en los servidores de desarrollo, prueba y producción?
- ¿Qué necesitan instalar los diferentes miembros del equipo de
- DW/BI en sus estaciones de trabajo?
- ¿Cómo convertir el modelo de datos lógico en un modelo de datos físicos en la base de datos relacional?
- ¿Cómo conseguir un plan de indexación inicial?
- ¿Debe usarse la partición en las tablas relacionales?

Diseño del sistema de Extracción, Transformación y Carga (ETL).

El sistema de Extracción, Transformación y Carga (ETL) es la base sobre la cual se alimenta el Datawarehouse. Si el sistema ETL se diseña adecuadamente, puede extraer los datos de los sistemas de origen de datos, aplicar diferentes reglas para aumentar la calidad y consistencia de los mismos, consolidar la información proveniente de distintos sistemas, y finalmente cargar (grabar) la información en el DW en un formato acorde para la utilización por parte de las herramientas de análisis.

Especificación y desarrollo de aplicaciones de BI

Una parte fundamental de todo proyecto de DW/BI está en proporcionarles a una gran comunidad de usuarios una forma más estructurada y por lo tanto, más fácil, de acceder al almacén de datos. Proporcionamos este acceso estructurado a través de lo que llamamos aplicaciones de inteligencia de negocios (Business Intelligence Aplications).

Las aplicaciones de BI son la cara visible de la inteligencia de negocios: los informes y aplicaciones de análisis proporcionan información útil a los usuarios. Las aplicaciones de BI incluyen un amplio espectro de tipos de informes y herramientas de análisis, que van desde informes simples de formato fijo a sofisticadas aplicaciones analíticas que usan complejos algoritmos e información del dominio. Kimball divide a estas aplicaciones en dos categorías basadas en el nivel de sofisticación, y les llama informes estándar y aplicaciones analíticas.

Informes estándar

Los informes estándar son la base del espectro de aplicaciones de BI. Por lo general son informes relativamente simples, de formato predefinido, y parámetros de consulta fijos. En el caso más simple, son informes estáticos pre almacenado. Los informes estándar proporcionan a los ciudadanos un conjunto básico de información acerca de lo que está sucediendo en la votación del senado según su región y como le beneficia. Este tipo de aplicaciones son el caballo de batalla de la BI. Son informes que los ciudadanos podrán usar día a día. La mayor parte de lo que piden las personas durante el proceso de definición de requisitos se clasificaría como informes estándar. Por eso es conveniente desarrollar un conjunto de informes estándar en el ciclo de vida del proyecto.

Aplicaciones analíticas

Las aplicaciones analíticas son más complejas que los informes estándar. Normalmente se centran en un proceso de negocio específico y resumen cierta experiencia acerca de cómo analizar e interpretar ese proceso de negocio. Estas aplicaciones pueden ser muy avanzadas e incluir algoritmos y modelos de minería de datos, que ayudan a identificar oportunidades o cuestiones subyacentes en los datos. Otra característica avanzada en algunas aplicaciones analíticas es que el usuario puede pedir cambios en los sistemas transaccionales basándose en los conocimientos obtenidos del uso de la aplicación de BI. En el otro extremo del espectro, algunas aplicaciones analíticas se venden como soluciones cerradas o enlatados, y son independientes de las aplicaciones particulares de la empresa. Algunas aplicaciones analíticas comunes incluyen:

5. DISEÑO DE LA SOLUCIÓN

5.1. DESCRIPCIÓN DE LA ARQUITECTURA

Para este proyecto se describe la arquitectura a utilizar en los Data Mart, detallando cada uno de los procesos o sub-sistemas que intervienen en el proyecto:

5.1.1. Fuente y Destino de datos

Los datos que alimentará el Data Mart de votaciones están alojados en la base de datos de la registraduria nacional del estado, secretaria del senado y datos del DNP sobre proyectos, puedo acceder a estos datos a través del programa de gobierno en línea en su eje de gobierno abierto, debemos tener en cuenta que el proceso electoral para el senado se hace cada 4 años, el último fue en el 2014.

Para crear las tablas del Cubo de votaciones y guardar los datos resultantes de la ejecución de los procesos ETLs se lo realizará en MySql.

5.1.2. Servidor Windows

Se tendrá dos máquinas virtuales con sistema operativo para separar tanto el servidor de base de datos como el servidor para alojar las herramientas de la solución Pentaho, la solución final estar alojado en un ec2 de Amazon Web Service AWS, para que los jurados y asesor de tesis pueda verla en línea le servidor de Business Analytics y el dashboard creado como ejemplo.

5.1.3. Proceso ETL

Permite implementar un grupo de procesos que realiza la extracción de las fuentes de datos, la transformación y finalmente cargar los datos actualizados en las tablas de los Data Mart.

5.1.4. Cubo de Datos

Una vez finalizado el proceso ETL se crea las tablas del modelo dimensional (modelo estrella) que se encuentran en el motor de base de datos MySql. En el Data Mart se almacenan los datos operacionales en estructuras multidimensionales (Cubo de Datos) los cuales son más flexibles para las consultas de los usuarios.

5.1.5. Presentación

Los Reportes e Indicadores es la interacción del usuario con la información procesada de lo ETLs, estos reportes se comunican directamente a los Cubos de Datos publicados en la plataforma de Pentaho y contienen la información a 2014 de las votaciones por municipio.

Los reportes serán creados desde la misma página de Pentaho y puedan ser visualizados por todos los usuarios que lo requieran.

5.1.6. Seguridad

Definir usuarios y roles para los administradores y ciudadanos que libremente pueden ver los reportes y análisis creados, dependiendo al perfil que tengan en la empresa y restricción de acceso a reportes e indicadores.

5.1.7. Administración

Perfiles de administradores para la manipulación de la aplicación.

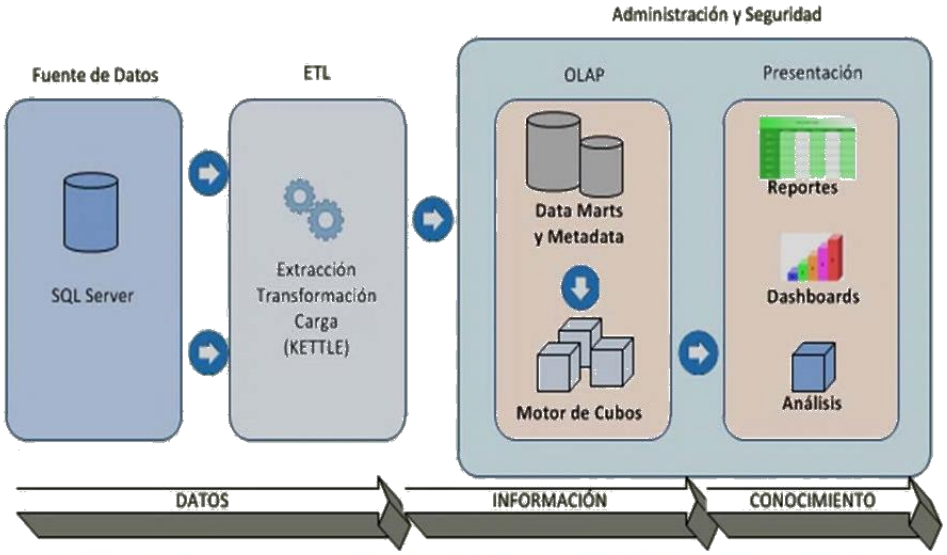


Ilustración 14 Arquitectura de la Solución

5.2. ANÁLISIS DE LAS FUENTES DE DATOS

Para definir qué información es importante para el ciudadano colombiano cuando va a elegir a un senador, se hizo una en cuenta a 100 ciudadanos vía online y presencial.

Se anexa la encuesta física y se muestra en la sigueinet ilustración el encabezado de la encuenta digital a través de un formulario de google.

Proyecto Observatorio Electoral Senado Colombia

Proyecto de inteligencia de negocio para un observatorio electoral de estudiante de EAFIT Maestría.
El objetivo de esta encuesta es solo académico. Recuerde que esta información es para uso academico

*Obligatorio

Cual es su nombre? *
No es necesario apellidos

Que edad tiene? *
Indique la edad cumplida en números enteros.

Voto usted por algún candidato al senado en el año 2010? *
- Si
- No

Por cual partido votó en esas elecciones ? *
- Partido de la U
- Partido Conservador Colombiano
- Partido Liberal Colombiano
- Partido de Integración Nacional

Ilustración 15 Captura de pantalla de la encuesta en linea.

Se tomaron varios aspectos de los temas trascendentales que a mi juicios pueden ser determinantes al momento de elegir un candidato a senador.

- Reforma Pensional
- Sistema de Justicia
- Sistema de Salud

- Fiscalía
- Agua
- La eutanasia
- El sexo antes del matrimonio
- El matrimonio homosexual
- El divorcio
- El consumo de marihuana
- Legalización de la Droga
- El aborto
- La paz
- Política de hidrocarburos
- Drogadicción en colegios
- Embarazo juvenil

También se observó el partido, la región y el género del candidato. Algunas de las respuestas en graficos analizados.

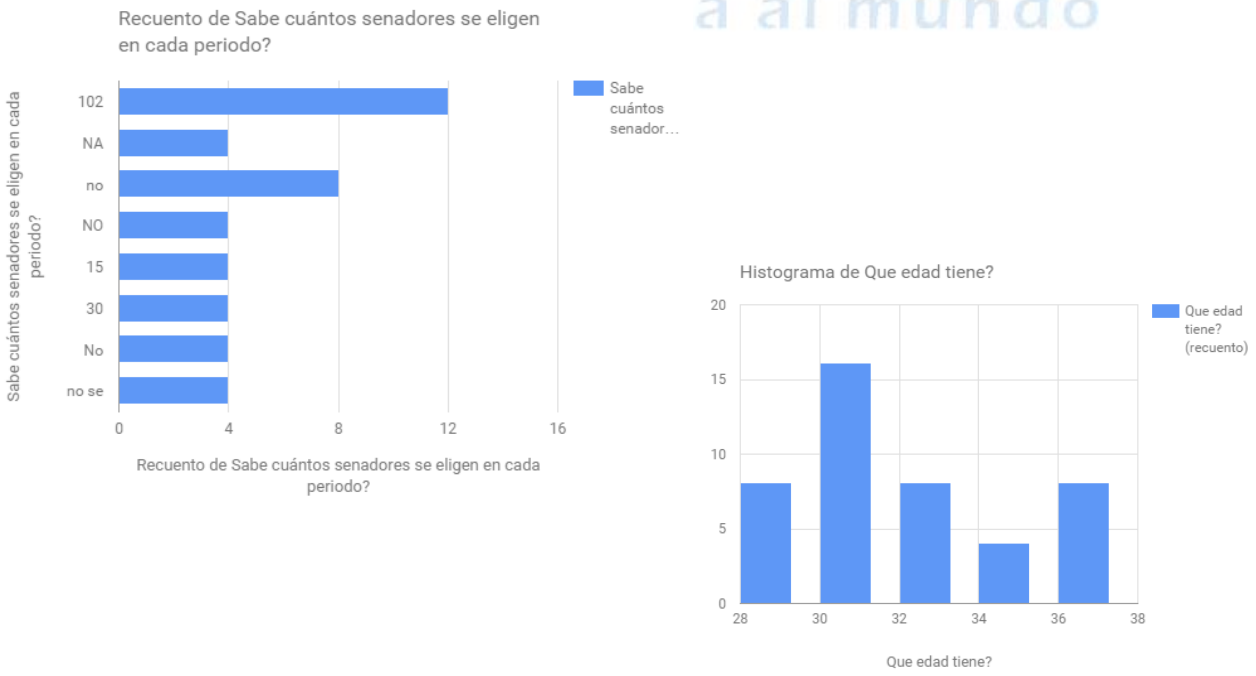
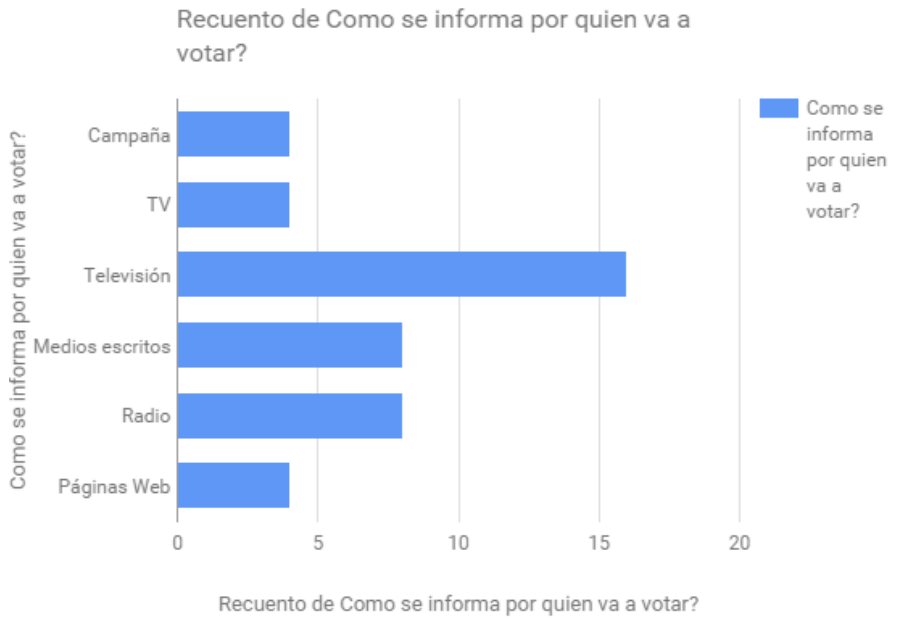
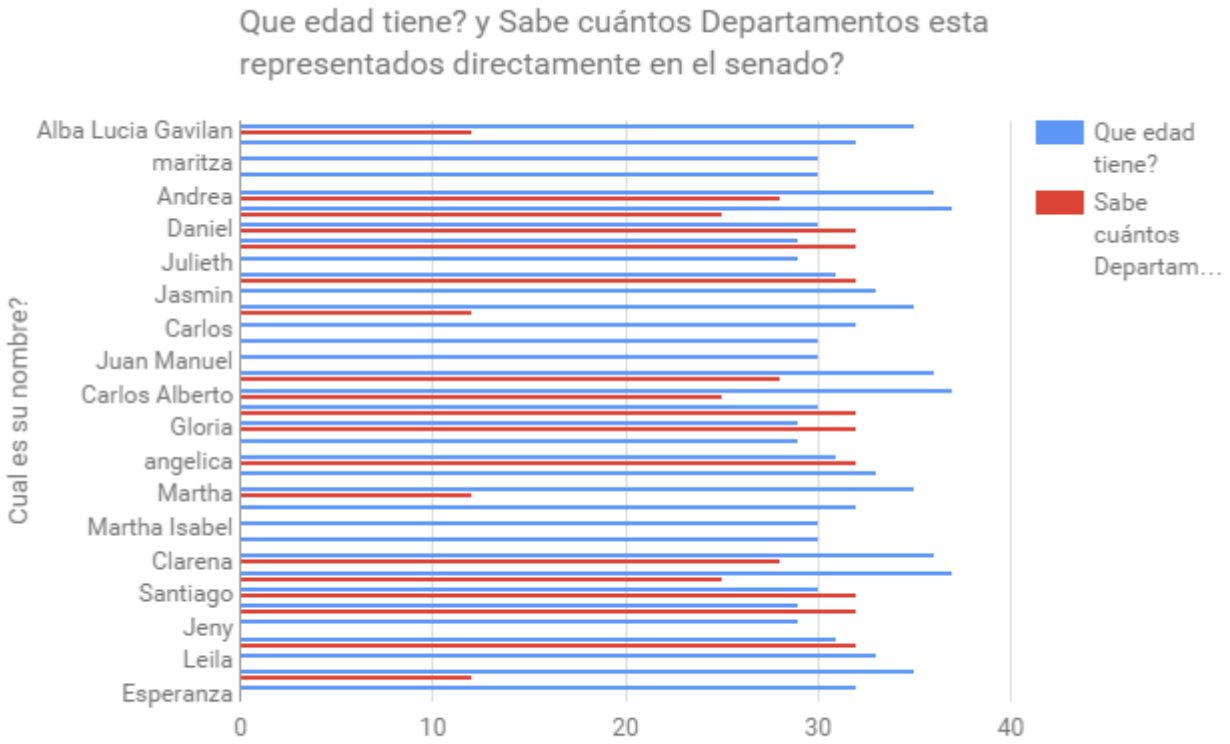


Ilustración 16 Ejemplos de Analytics


Recuento de Como se informa por quien va a votar?
Como se informa por quien va a votar?
Campaña
TV
Televisión
Medios escritos
Radio
Páginas Web
0
5
10
15
20
Como se informa por quien va a votar?
Recuento de Como se informa por quien va a votar?


Sabe cuántos Departamentos esta representados directamente en el senado? vs. Que edad tiene?
Sabe cuántos Departamentos esta representados directamente en el senado?
40
30
20
10
0
28
30.8
33.6
36.4
39.2
Que edad tiene?
Sabe cuántos Departam...



La fuente de datos que alimentará al Data Mart será la base de datos de las votaciones por municipio, este sistema tiene información relacionada a senadores, municipios, partidos políticos y votos.

Actualmente la información de votos se maneja en hojas de cálculo, o consultas directas a la base de datos, lo cual toma tiempo su ejecución y no siempre contienen la información correcta y requerida según la necesidad de cada usuario.

Dynamics es el principal sistema y fuente de información que se tomó para el proyecto, por lo cual se realizó un estudio minucioso de sus tablas y relaciones, así mismo las consultas y

hojas de cálculo ya existentes, para determinar las entidades, atributos más relevantes e importantes para formar parte de los reportes del proyecto.

En el grafico siguiente se puede observar algunos módulos que maneja el sistema Dynamics y de los cuales se obtuvo la información necesaria para armar los Data Mart.

	A	B	C	D	E
1	NOMBRE	PARTIDO	CURUL	EL ENCANTO-AMAZONAS	LA CHORRERA-AMAZONAS
2	SOLO LISTA CENTRO DEMOCRÃÂTICO	CDMFCG	0	3	17
3	ALVARO URIBE VELEZ	CDMFCG	1	0	0
4	MARIA DEL ROSARIO GUERRA DE LA ESPRIELLA	CDMFCG	1	0	0
5	PALOMA SUSANA VALENCIA LASERNA	CDMFCG	1	0	0
6	ANA MERCEDES GOMEZ MARTINEZ	CDMFCG	1	0	0
7	SUSANA CORREA BORRERO	CDMFCG	1	0	0
8	ALFREDO RANGEL SUAREZ	CDMFCG	1	0	0
9	IVAN DUQUE MARQUEZ	CDMFCG	1	0	0
10	FERNANDO NICOLAS ARAUJO RUMIE	CDMFCG	1	0	0
11	JOSE OBDULIO GAVIRIA VELEZ	CDMFCG	1	0	0
12	ORLANDO CASTAÃ‘EDA SERRANO	CDMFCG	1	0	0
13	DANIEL ALBERTO CABRALES CASTILLO	CDMFCG	1	0	0
14	EVERTH BUSTAMANTE GARCIA	CDMFCG	1	0	0
15	ALFREDO RAMOS MAYA	CDMFCG	1	0	0
16	JAIME ALEJANDRO AMIN HERNANDEZ	CDMFCG	1	0	0
17	ERNESTO MACIAS TOVAR	CDMFCG	1	0	0
18	RUBY THANIA VEGA DE PLAZAS	CDMFCG	1	0	0
19	CARLOS FELIPE MEJIA MEJIA	CDMFCG	1	0	0
20	PAOLA ANDREA HOLGUIN MORENO	CDMFCG	1	0	0
21	NOHORA STELLA TOVAR REY	CDMFCG	1	0	0
22	HONORIO MIGUEL HENRIQUEZ PINEDO	CDMFCG	0	0	0

Ilustración 17 Votos por Municipio de Colombia

En el gráfico siguiente se observa el nombre de la base de datos y sus tablas que se utiliza como fuente de información para obtener los datos y consultas para armar el Data Mart.

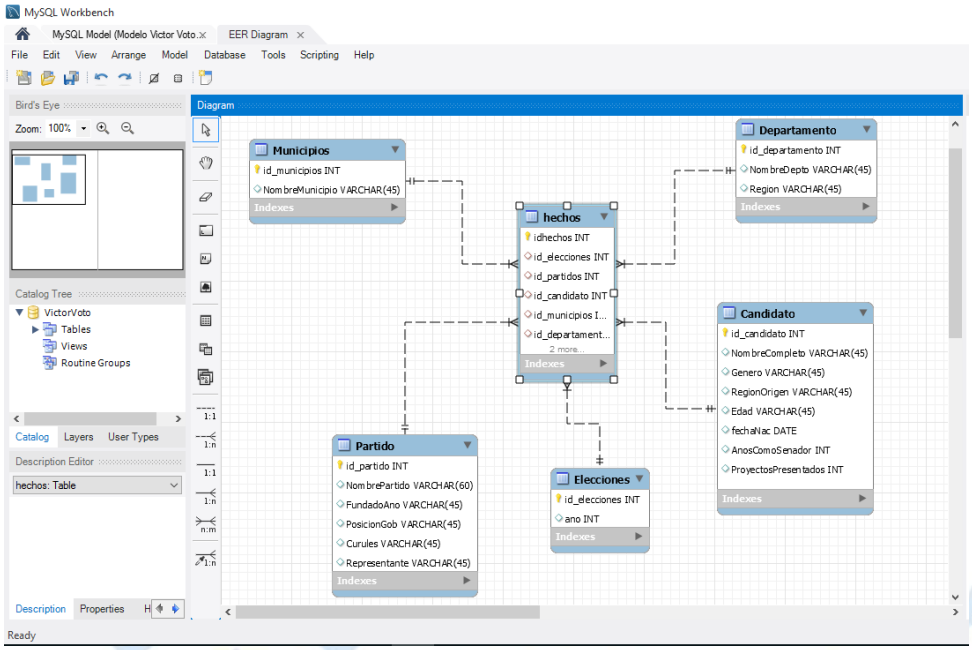


Ilustración 18 Mysql Workbench Matriz

5.3. CALIDAD DE DATOS

Para determinar una buena de calidad de los datos a mostrar, se fijan pautas en la recopilación y análisis de la información de la registraduria en el portal del gobierno www.datos.gov.co.

- La registraduria nacional de Colombia facilita todas las fuentes que contenga la información de las elecciones, para ser analizadas, comparadas entre si y seleccionar las mejores.
- Se analiza las fuentes seleccionadas para detectar los datos que aportan a una mejor y real visión de las votaciones del municipio separando los datos que no aportan en el estudio.
- Los reportes deben tener los suficientes y necesarios parámetros para una correcta interpretación de la información mostrada.

- La efectividad en la toma de decisiones depende de la herramienta a implementar, de la calidad y credibilidad de datos analizados.
- Los errores encontrados son:
- Al manejar las hojas de cálculo no permite traer información histórica por el limitante que existe en Excel lo cual no permite tomar decisiones acertadas sobre la información presentada.
- No existen tablas donde se guarda la codificación y explicación de los campos de las tablas de la base de datos origen, lo cual sería de gran ayuda para el usuario que lo utilice.
- Se utilizan campos de algunas tablas de la base de datos origen para guardar información que se considera importante pero no corresponde a ese campo, causando una confusión para la persona nuevas a manejar el sistema.

El proceso de control de calidad de los datos se lo realizó de la siguiente manera:

- Definir los filtros necesarios en las consultas para obtener solo valores requeridos y necesarios para el candidato en el Data Mart.
- Analizar los valores nulos, inválidos, faltantes y datos que no aportan para el estudio del proyecto.
- Con los análisis realizados a la información entregada, se notifica al usuario el límite y los problemas serían resueltos con la herramienta a implementar.

Con la calidad de datos se pretende tener un fuente de datos confiable, robusta como también mejorar el diseño de base de datos que va a soportar el Data Mart y a su vez la estructura de base de datos origen.

5.4. FRECUENCIA DE CARGA

Para la carga de los datos históricos en el nuevo modelo multidimensional, se ingresa los datos del año 2014, se valida los reportes que se encuentren correctos que mejoren los tiempos de respuesta al usuario, una vez finalizado esta validación, se procede a correr los mismos procesos para la carga de los datos desde el año 2010 hasta los de 2006, estos se guardarán en las tablas del modelo estrella en la base de datos Mysql que alimenta el Cubo de votaciones por municipios. Se diseña procesos automáticos para la carga de los datos diarios en el Cubo de Votaciones.

5.5. MODELO MULTIDIMENSIONAL

Para el diseño del modelo de datos se tomó en cuenta las consultas existentes como son: la consulta para obtener la información de votación y candidatos, estas estaban siendo utilizadas para entregar la información al usuario final, fueron analizadas y conjuntamente con lo solicitado por el usuario se obtuvo la siguiente información:

5.5.1. Dimensión de Elecciones:

Esta tabla filtra las consultas por el parámetro de tiempo, lo cual para la registraduria es muy importante como por ejemplo: ¿Cuántos votos se obtuvieron en un municipio por partido?

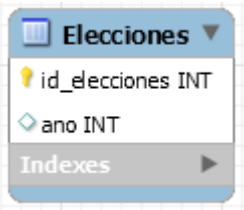


Ilustración 19 Dimensión Elecciones

5.5.2. Dimensión de Partidos

Para la dimensión DM_PARTIDOS las tablas del modelo de base de datos origen son:

- Partidos

Esta tabla proporciona la información del municipio en el cual se votó por un candidato y clasifica por partido.

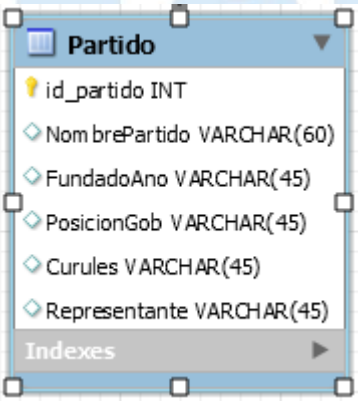


Ilustración 20 Dimensión Partidos

5.5.3. Dimensión de Candidato

Para la dimensión DM_CANDIDATO las tablas del modelo de base de datos origen son:

Candidatos

Esta tabla proporciona la información del municipio dónde se votó por el candidato

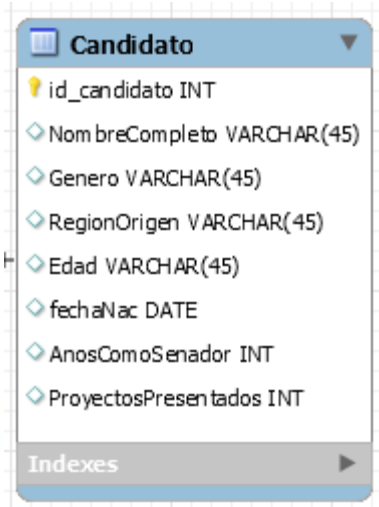


Ilustración 21 Dimensión Candidatos

5.5.4. Dimensión de municipios

Para la dimensión DM_MUNICIPIOS las tablas del modelo de base de datos origen son:

Municipios

Esta tabla describe el municipio, dónde está ubicado el punto de votación.

Ilustración 22 Dimensión Municipios

5.5.5. Dimensión de departamento

Para la dimensión DM_DEPARTAMENTO las tablas del modelo de base de datos origen son:

- DEPARTAMENTO

Esta tabla proporciona la información de la ubicación del muncipios dentro de Colombia, es decir, la región o provincia de cada departamento.

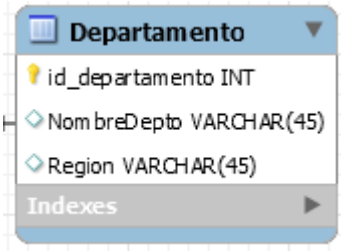


Ilustración 23 Dimensión Departamento

5.6. Esquema Multidimensional:

El esquema que se detalla a continuación fue el modelo estrella que se definió:

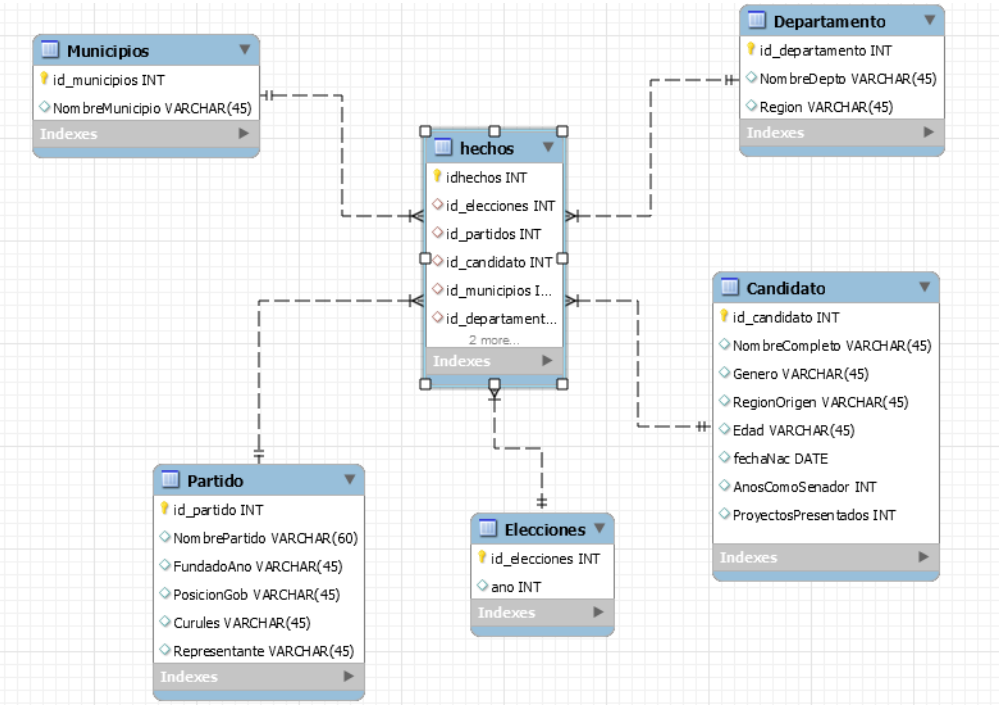


Ilustración 24 Modelo Multidimensional

5.7. DISEÑO RELACIONAL DE LA BASE DE DATOS QUE SOPORTA A LOS CUBOS

A continuación se detalla el modelo de datos que soportara el Cubo de Votaciones (Data Mart) para el diseño de los reportes.

El Cubo de Datos se deriva de las dimensiones y del esquema multidimensional. Se utiliza un esquema estrella ganando así simplicidad en el diseño y velocidad de acceso para obtener las distintas jerarquías.

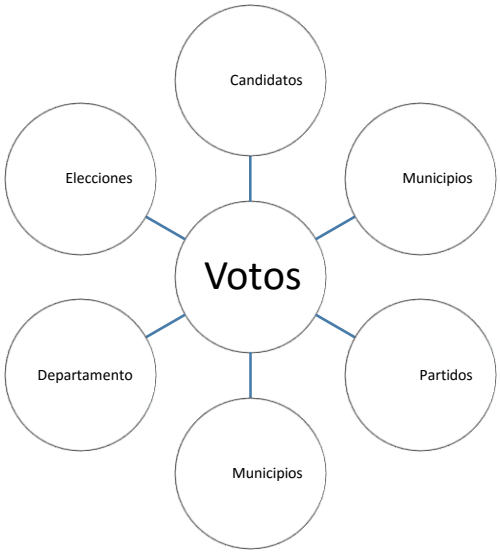


Ilustración 25 Modelo Multidimensional- Cubo de Votaciones

La unión de los cubos al análisis de negocios se da con la consola de usuario de pentaho.

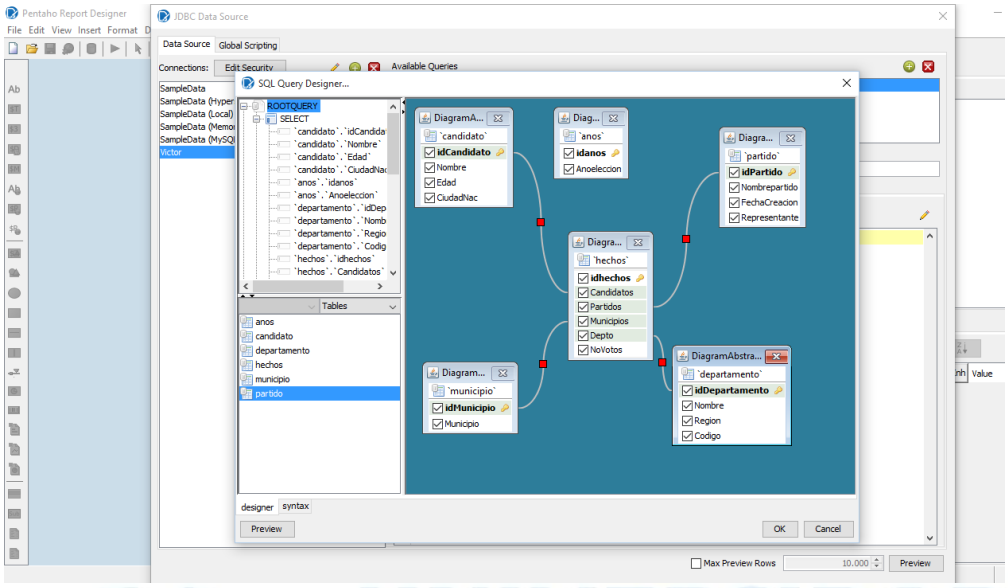


Ilustración 26 SQL Query Design

5.7.1. Reportes Esperados

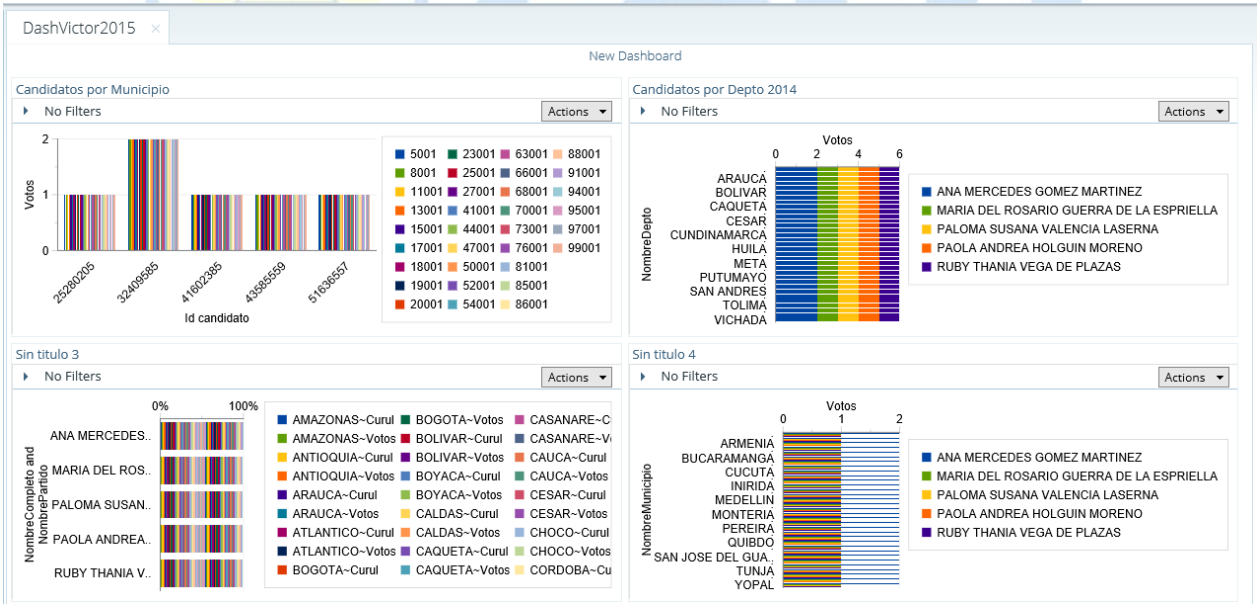


Ilustración 27 Dashboard en Pentaho

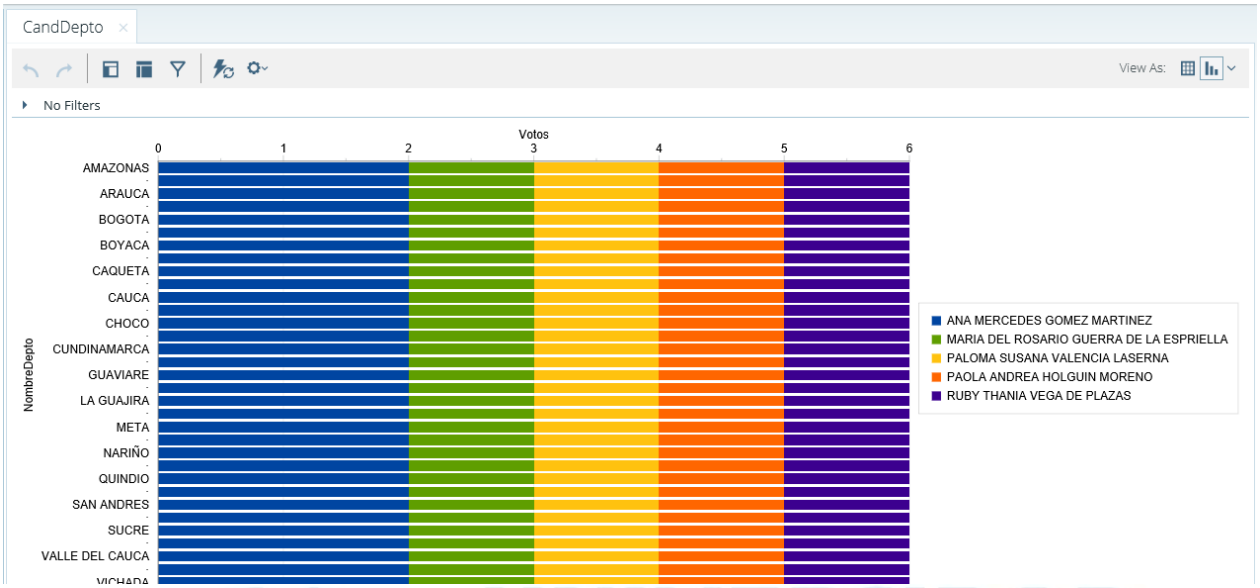


Ilustración 28 Ejemplo No 1 de Analytic en Pentaho

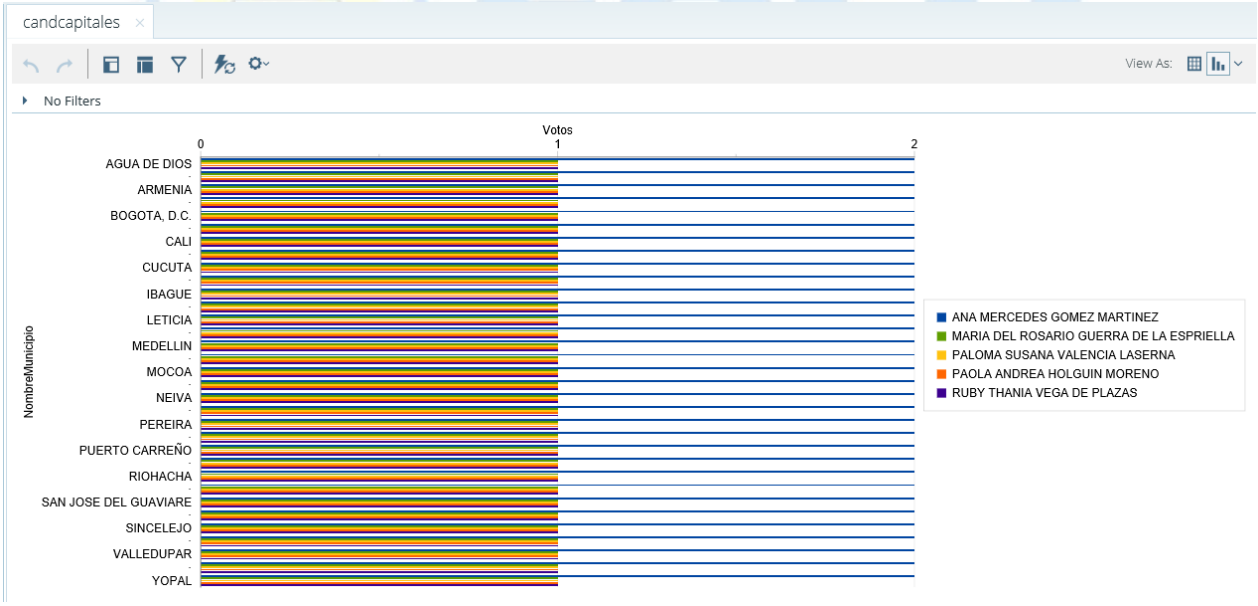


Ilustración 29 Ejemplo No 2 de Analytic en Pentaho

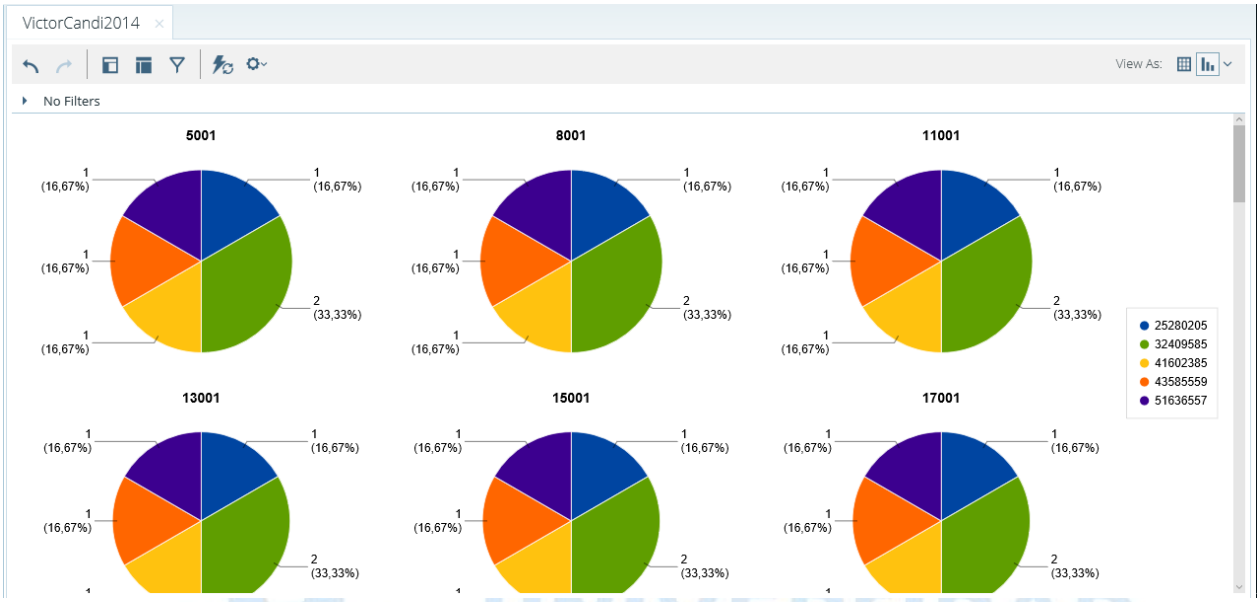


Ilustración 30 Ejemplo No 3 de Analytic en Pentaho

6. CONCLUSIONES

- ✓ La parte técnica de este trabajo me ha permito avanzar en una área de conocimiento, que sirve como solucione en muchos casos, pero a nivel estratégico y táctico puede ayudar a las organizaciones de nuestro país a estudiar su información anterior y presente para planear sus próximas juagas en el mercado.

- ✓ El datawarehouse y business intelligence aplicado a el Sistema electoral colombiano debería ser una función de la entidad encargada de realizar las elecciones donde se debe tener en una forma ordenada y sistemática la información de años de aplicación de este elecciones, generando los informes que en el presente trabajo se definieron como esenciales para la población muestra.

- ✓ La información de estos procesos es entregada a cuenta gotas por la registraduria y la secretaria del senado, pero se pude avanzar con la MOE(Misión de Observación Electoral), la cual libera un web services y entrega datos estadísticos analizados frecuentemente.

- ✓ Al iniciar este trabajo esperábamos colocar un granito de arena desde una solución tecnológica, para mejorar el proceso electoral en Colombia, sabemos que estamos lejos de cambiar la forma como los ciudadanos se informan sobre los candidatos y las practicas que se realizan alrededor de una elecciones, pero estamos seguros que con un trabajo fuerte por parte de organizaciones no gubernamentales (ONG) a las que se les entregue este prototipo y se desarrolle me voy a sentir satisfecho.

- ✓ La generación de un observatorio electoral a futuro es posible gracias al ejercicio establecido, vemos que los costó para un ONG no son altos tecnológicamente y se debe dedicar tiempo a poblarla y generar los analitics que los ciudadanos digitales cada año se suman a los que eligen por la información generada y evitar las dadivas temporales que no benefician a todos y terminan eligiendo a los menos preparados.

7. CONSIDERACIONES ADICIONALES.

7.1. Glosario

Almacén de Datos Operacionales	Base de datos diseñada para realizar consultas sobre datos transaccionales y/o lograr integración de información dispersa en varios sistemas operacionales. En general se convierten en fuentes o áreas de preparación para la Bodega de Datos.
Analítica Avanzada	Metodologías y técnicas que ofrecen mecanismos de simulación, predicción, optimización y otras capacidades avanzadas de análisis de información, encaminadas a enriquecer el proceso de toma de decisiones en la organización, en el tiempo y lugar óptimos.
BI	Business Intelligence (véase Inteligencia de Negocio)
BIMM	Business Intelligence Maturity Model (véase Modelo de Madurez de BI)

Bodega de Datos	Almacén de datos operacionales e históricos de una empresa y cuyo objetivo principal es el soporte a la toma de decisiones. Es una copia de los datos transaccionales de una Empresa, estructurada específicamente para consulta y análisis.
BPM	Business Process Management – Gestión de Procesos de Negocio
BSC	Balanced Scorecard – Cuadro de Mando Integral
Cuadro de Mando Integral	Sistema gerencial que vincula el logro de las metas estratégicas a largo plazo con las operaciones diarias de una organización. Propuesto por Kaplan y Norton en 1992, el sistema permite usar medidas en cuatro categorías – desempeño financiero, conocimiento de los clientes, procesos internos de negocio y aprendizaje/crecimiento – para alinear las iniciativas individuales, organizacionales y multi-departamentales, y para identificar nuevos procesos que cumplan los objetivos de los clientes y accionistas.
Data Warehouse	Véase Bodega de Datos
Datamart	Kimball: Componente de una Bodega de Datos, conformado por uno o varios esquemas de estrella (star schema), que ofrece servicios de información para un área o tema específico del negocio. La Bodega

	de Datos se convierte en la unión de datamarts, siendo dicha unión posible por la definición de dimensiones y medidas conformes.Inmon: Vista o subconjunto lógico de una Bodega de datos o extracto o subconjunto físico de la misma.
Datamining	Véase Minería de Datos
Dimensión	Es una característica de interés, utilizada para analizar un hecho del negocio, la cual puede tener asociados atributos y presentar jerarquías o niveles. Los ejemplos más comunes son cliente, producto, proveedor, vendedor, tiempo entre otras.
Dimensión conforme	Kimball: Es una dimensión homologada a nivel de toda la organización y por ende incluye todos los atributos y jerarquías necesarios para suplir los requerimientos de los diferentes datamarts que conforman la bodega de datos y que la incluyen en su estructura
Esquema de Estrella	Componente básico del modelo dimensional, que consiste de una tabla central denominada tabla de hechos (fact table) y un conjunto de tablas “satélite” que representan las dimensiones. La tabla de hechos corresponde a un subconjunto del producto cartesiano de las dimensiones, y cada elemento tiene asociado, en general, un conjunto de medidas.

Gestión de Procesos de Negocio	Conjunto de metodologías, técnicas y herramientas que soportan el diseño, modelamiento, automatización, administración, análisis y optimización de procesos de negocio.
Indicadores Clave de desempeño (KPI)	Medidas que reflejan el desempeño del negocio, y en general están asociadas con sus factores críticos de éxito
Inmon, William H.	Por muchos reconocido como el "padre de las bodegas de datos", recibió un grado en matemáticas de la Universidad de Yale y un Master of Science en Ciencias de la Computación de Nuevo México State University. Creó el sitio web "Corporate Information Factory" con el fin de educar profesionales en el tema de bodegas de datos. Prolífico autor de libros y conferencista.
Inteligencia de Negocio	Conjunto de tecnologías y aplicaciones que permiten recopilar, almacenar, analizar y tener acceso a datos, de tal manera que los usuarios de la organización pueden tomar mejores decisiones. Consiste en transformar los datos operacionales de una empresa en información "accionable", es decir, información que realmente habilite y optimice el proceso de toma de decisiones y la definición de estrategias y acciones encaminadas a mejorar el desempeño del negocio

IT	Information Technology (ver TI)
Kimball, Ralph	Ph.D de la Universidad de Stanford en Ingeniería Eléctrica (especializado en sistemas hombre-máquina), ha sido líder y visionario en la industria de las Bodegas de Datos desde 1982, y hoy en día un reconocido conferencista, consultor, profesor y autor de numerosos e ilustrativos libros sobre el tema.
KPI	Key Performance Indicator – Indicadores clave de desempeño
Medida	Atributo numérico que representa un evento del negocio
Medida Conforme	Kimball: Es una medida homologada a nivel de toda la organización y por ende suple los requerimientos de los diferentes datamarts que conforman la Bodega de Datos y que la incluyen en su estructura
Minería de Datos	Proceso de negocio cuyo objetivo es encontrar patrones en los datos, no evidentes y significativos, que generen conocimiento e ideas de cómo conducir el negocio de una manera más eficiente y eficaz
Modelaje Dimensional	Técnica de modelaje, alternativa al modelaje Entidad/Relación, propuesta por Ralph Kimball y

	específicamente utilizada para el diseño de Bodegas de Datos.
Modelo de Madurez de BI	Modelo que permite determinar en qué nivel se encuentra una organización o empresa con respecto a Inteligencia de Negocio y cómo debería evolucionar hacia un estado ideal
ODS	Operational Data Store (véase Almacén de datos operacionales)
OLAP	On line Analytical Processing o Procesamiento analítico en línea, es un término utilizado para referirse a herramientas de consulta de información que permiten el análisis dinámico de diferentes cifras del negocio denominadas "medidas", mediante la combinación de dos o más características de interés denominadas "dimensiones". Las herramientas OLAP permiten diferentes operaciones que facilitan la navegación, trasposición, combinación y en general la "maniobrabilidad" de la información, ofreciendo gran flexibilidad de consulta y análisis al usuario final.
OLTP	On line Transaction Processing o Procesamiento de Transacciones en línea, es un término utilizado para referirse a los sistemas de información de soporte del día a día o sistemas operacionales, enfocados al

	registro de transacciones realizadas en las diferentes áreas de una empresa.
Proceso de ETLC	Proceso de Extracción, Transformación, Cargue y Limpieza de información (por su denominación en inglés Extraction, Transformation, Load and Cleansing), que corresponde normalmente al primer paso para construir una bodega de datos y a los mecanismos posteriores para mantenerla actualizada.
Retorno sobre la Inversión	Tasa que compara el beneficio o la utilidad obtenida en relación con la inversión realizada
ROI	Return on Investments – Retorno sobre la inversión
Sistemas Operacionales	Sistemas transaccionales que dan soporte a las labores del día a día de una organización
SPREADMART	Término acuñado por TDWI para designar el hecho de que las organizaciones utilicen de manera no controlada, hojas electrónicas aisladas e independientes, como herramienta fundamental para el procesamiento de información para la toma de decisiones.
Star Schema	Véase Esquema de Estrella
Steakholder	Término en inglés que se refiere a quienes pueden afectar o son afectados por las actividades de una

	empresa (socios, empleados, accionistas, clientes, proveedores etc.)
Tablero de Control	Herramienta que permite visualizar, mediante diversos recursos gráficos, un conjunto de indicadores cuyo análisis, seguimiento y evaluación periódica refleja la situación del ente al que aplican (la empresa, un área particular, un sector etc.).
TDWI	The Data Warehousing Institute. Organización dedicada a la capacitación, entrenamiento, certificación, comunicación de noticias, e investigación orientada a ejecutivos y profesionales de TI, sobre los temas de inteligencia de negocio y bodegas de datos. Fundado en 1995. Sitio Web: www.tdwi.orghttp://www.tdwi.org/
TI	Tecnologías de Información

8. BIBLIOGRAFIA

al., R. C. (2002). A Comparison of Data Warehouse Development Methodologies Case Study of the Process Warehouse. *DEXA 2002, LNCS 2453*, 203–215.

B. Afolabi, O. T. (2010). *Using Users' Expectations to Adapt Business Intelligence Systems.*

Bäck., T. (2005). Adaptive business intelligence based on evolution strategies: some application examples of self-adaptive software. *Inf. Sci. 148(1-4)*, 113-121.

Co., C. T. (2000). *Business Intelligence The The Missing Link.* Minnesota.

Dresner, H. (1989). *Business Intelligence.*

Fowler, M. (2004). *UML distilled: a brief guide to the standard object modeling language.* Addison-Wesley Professional.

González, J. F. (2011). Factores críticos de éxito de un proyecto de Business Intelligence. *Novática: Revista de la Asociación de Técnicos de Informática*, 20-25.

Inmon, W. H. (1994). *Using the data warehouse. .* Wiley-QED Publishing.

J. Fernández, E. M. (2008). Agile Business Intelligence Governance: Su justificación y presentación. .

Kimball, R. &. (2002). *The data warehouse Toolkit.*

Kroll, P. &. (2004). *he rational unified process made easy: a practitioner's guide to the RUP.* Addison-Wesley Professional.

L.T. Moss, S. A. (2003.). *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision Support Applications.* Addison Wesley Longman.

Llorente, M. E. (2012). Sistema de soporte a la toma de decisiones basado en datawarehouse para pacientes. *XIV Workshop de Investigadores en Ciencias de la Computación*, (pp. 230-235).

Luhn, H. P. (1958). A Business Intelligence System. *IBM Journal of Research and Development (Volume:2 , Issue: 4)*, 314 - 319.

Moss., L. (2001). Ten Mistakes to avoid for Data Warehouse Projects Managers. *TDWI'S best of Business Intelligence Vol 3.*, 16-22.

Otero, L. W. (2014, Marzo 17). *razonpublica.com*. Retrieved from razonpublica.com: http://www.razonpublica.com/index.php/politica-y-gobierno-temas-27/7444-senado-2014-2018-%C2%BFrepresentaci%C3%B3n-nacional.html

PMI. (2013). *PMBOK Guide5th Spanish.* PMI.

Rivadera, G. R. (2014). *La metodología de Kimball para el diseño de almacenes de datos (Data warehouses).* Salta, Argentina: Cuadernos de la Facultad.

T.N. Huynh, J. S. (2001). Prototyping Data Warehouse Systems. . *DaWaK*, 195-207.

V. Stefanov, B. L. (2005). Bridging the Gap between Data Warehouses and Business Processes: A Business Intelligence Perspective for Event-Driven Process Chains. *EDOC*, 3-14.

http://www.kimballgroup.com: Este sitio contiene mucha información y artículos sobre la metodología, y además una serie de planillas de Excel usadas en cada paso de la metodología.

http://www.bi-bestpractices.com/view-articles/4768

http://todobi.blogspot.com.es/

http://www.businessintelligence.info/categoria/serie-dwh.html

http://www.dataprix.com/blogs/respinosamilla/base-datos-anal-tica-datawarehouse-o-almac-n-datos

http://www.pentaho.com/

http://www.dataprix.com/blogs