UNIVERSIDAD TECNOLÓGICA EQUINOCCIAL

FACULTAD DE CIENCIAS DE LA INGENIERÍA

CARRERA DE INGENIERÍA INFORMÁTICA Y CIENCIAS DE LA COMPUTACIÓN

TEMA:

"DISEÑO DE CUBO ANALÍTICO GENÉRICO PARA LAS CUENTAS DE BALANCES DE LAS COMPAÑÍAS DE SEGUROS DEL ECUADOR Y DE OTROS PAÍSES EN LATINO AMÉRICA UTILIZANDO LA HERRAMIENTA PENTAHO"

TRABAJO PREVIO A LA OBTENCIÓN DEL TÍTULO
DE INGENIERO EN INFORMÁTICA Y CIENCIAS DE LA COMPUTACIÓN

AUTOR:
PEDRO ARTURO GUALAVISI MADRID

DIRECTOR: ING. CIRO SAGUAY

Quito, Mayo 2015

DERECHOS DE AUTOR

DECLARACIÓN

Yo **PEDRO ARTURO GUALAVISI MADRID**, declaro que el trabajo aquí descrito es de mi autoría; que no ha sido previamente presentado para ningún grado o calificación profesional; y, que he consultado las referencias bibliográficas que se incluyen en este documento.

La Universidad Tecnológica Equinoccial puede hacer uso de los derechos correspondientes a este trabajo, según lo establecido por la Ley de Propiedad Intelectual, por su Reglamento y por la normativa institucional vigente.

_____

Pedro Arturo Gualavisí Madrid

C.I. 1720880739

CERTIFICACIÓN

Certifico que el presente trabajo que lleva por título "**Diseño de cubo analítico genérico para las cuentas de Balances de las compañías de Seguros del Ecuador y de otros países en Latino América utilizando la herramienta.**", que, para aspirar al título de **Ingeniera en informática y Ciencias de la computación** fue desarrollado por **Pedro Arturo Gualavisí Madrid**, bajo mi dirección y supervisión, en la Facultad de Ciencias de la Ingeniería; y cumple con las condiciones requeridas por el reglamento de Trabajos de Titulación artículos 18 y 25.

_____

Ing. Ciro Saguay
Director de Trabajo de Titulación
C.I. 0602692113

DEDICATORIA

Dedico mi tesis a mi madre, quien a lo largo de mi vida ha velado por mi bienestar y educación siendo mi apoyo en todo momento, depositando su entera confianza en cada reto que se me presenta sin dudar ni un solo momento en mi inteligencia y capacidad, a mi hermano que es un pilar fundamental, el cual me brinda todo su apoyo cuando lo requiero y a mi familia por apoyarme y ayudarme a cumplir cada reto. Es por ello que lograre llegar a cumplir mis metas propuestas.

AGRADECIMIENTO

Agradezco a Dios por la salud que me dio para poder seguir adelante y a mi madre por el apoyo que me brindo, por el sacrifico que día a día hace para yo poder culminar mi carrera y por creer en mí, por estar ahí dándome fuerzas para jamás darme por vencido por ser más que una madre una amiga en la cual deposite y deposito toda la confianza y jamás dudo de mí y por ser el segundo pilar más importante en mi vida.

Agradezco a la Universidad Tecnológica Equinoccial por abrirme sus puertas y darme la oportunidad de pertenecer y formar parte de ella, a los Ingenieros por brindarme todos sus conocimientos y sabiduría sobre los temas impartidos a lo largo de mi carrera.

Agradezco a mi hermano y familia por darme siempre ánimos y estar ahí brindándome todo su apoyo incondicional.

ÍNDICE DE CONTENIDOS

PÁGINA

ÍNDICE DE FIGURAS

PÁGINA

GLOSARIO

DW: Data WareHouse,……..12
ETL: Extract, Transform and Load, ……2
HOLAP: Hybrid Online Analytical Process, ……..……..21
MOLAP: Multidimensional Online Analytical Processing, ……21
OLAP: On-Line Analytical Processing, …… ……..……..……13
OLTP: OnLine Transaction Processing, …… …… ……..……13
ROLAP: Procesamiento Analítico OnLine Relacional, ……1
SCD: Dimensiones Lentamente Cambiantes, ……..……37
SGBD: sistema de gestión de bases de datos, …… 33
XLS: Archivos en formatos Excel, ……..……52

RESUMEN

En la presente tesis, se recopila información necesaria sobre Business Intelligence (BI), de esta manera se identifican metodologías para la implementación de esta infraestructura en el presente proyecto que serán importantes para el desarrollo del mismo. La metodología para en el presente proyecto es Kimball.

Se analiza el entorno del negocio y con la información obtenida se procede a crear las tablas de dimensiones y la tabla de hechos que permite almacenar la información de forma estándar para las cifras de Balances de las Compañías de Seguros.

Se analiza las fuentes de datos que nos servirán con insumo, estas fuentes de datos son obtenidas de la página web de la Superintendencia de Bancos y Seguros del Ecuador, la misma que publica de forma mensual las cifras de Balances para las compañías de seguros y los carga en formato xls y csv.

Se realizan los procesos de Extracción Transformación y Carga (ETL), estos procesos permiten la descarga de los archivos que son insumos para nuestro proyecto y también se realizan procesos de validación y carga a nuestra base de datos, en ellas contienen nuestras tablas de dimensiones y nuestra tabla de hechos que es poblada según el periodo que corresponda.

Se crean los Cubos analíticos con la herramienta Pentaho, la cual nos permite generar información que nos permitirá genera reportes de fácil acceso, así como gráficos que permiten analizar la información de manera dinámica para una correcta toma de decisiones.

ABSTRACT

In this thesis, necessary information on Business Intelligence (BI) is collected, so methodologies for the implementation of this infrastructure in our project will be important for its development are identified. The methodology to be implemented in this project is Kimball.

the business environment is analyzed and based on information obtained proceeds to create the dimension tables and the fact table to store the information in a standard way for figures Balance Sheet of Insurance Companies.

The data sources that will help us with input is analyzed, these data sources are obtained from the website of the Superintendency of Banks and Insurance of Ecuador, the same that publishes monthly figures Balances for insurance companies and load xls and csv format.

Extraction processes Transformation and Load (ETL) are performed, these processes allow downloading files that are inputs to our project and process validation and loading our database also perform in them contain our tables of dimensions and our fact table that is populated by the appropriate period.

Cubes with the analytic tool Pentaho are created, which allows us to generate information that will allow us easy access generates reports and graphs that analyze the dynamic information for proper decision making.

INTRODUCCIÓN

1 INTRODUCCIÓN

Los analistas financieros y gerentes de las diferentes compañías de seguros buscan obtener información vital para la toma de decisiones, esta información debe generase de manera rápida, de la misma forma que debe ser confiable, con una herramienta de fácil acceso y manipulación.

Desarrollar Cubos Analíticos de BI con la Herramienta Pentaho, necesarios para el análisis de la información que presenta la “Superintendencia de Bancos del Ecuador” para las Cuentas de Balances de las Compañías de Seguros del Ecuador.

La información que presenta la Superintendencia de Bancos y Seguros del Ecuador es una recopilación de información que deben presentar mensualmente las Compañías de Seguros, y tiene como finalidad proporcionar a los usuarios una visión de la situación financiera de cada una de ellas.

Actualmente muchas empresas ya sean grandes o pequeñas, se plantean implantar en su organización un sistema de información que les ayude en la toma de decisiones. Una de las tendencias actuales es implantar un sistema de BI, el cual procesa grandes cantidades de datos para obtener información valiosa y mediante la herramienta Pentaho generar reportes intuitivos que sirvan para la toma de decisiones.

Además de ser un proyecto académico, el proyecto busca mejorar la productividad y rendimiento en cuanto al reporteo de información para diferentes Compañías de Seguros que lo requieran.

Establecer el modelo de negocios para que los analistas financieros y gerentes de las compañías aseguradora obtenga de forma sencilla,

estandarizada y estructurada información estratégica para la toma de decisiones.

Desarrollar un diseño de procesos ETL con los archivos en que presente la Superintendencia de Bancos y Seguros obtenidos, para estandarizar dicho modelo y de esta manera lograr una sencilla migración de la información a nuestro nuevo modelo de datos.

Establecer e implementar la estructura para la construcción de las tablas de hechos y dimensiones de una forma estructurada utilizando como principal insumo los valores para las cuentas de balances obtenidas mediante los procesos ETL.

MARCO TEÓRICO

2 MARCO TEÓRICO

En esta sección, se realizará una introducción al tema que apunta este trabajo, es necesario antes que todo definir los conceptos claves y características sobre los que se basa. En otro punto se profundizará sobre lo que es un sistema BI, Data Mart (DM) y Cubos Analíticos OLPT (On-Line Analytical Processing).

2.1 HISTORIA DE BI

A continuación, se detallan los que se podrían denominar hitos más importantes de la historia de BI, retomando hechos anteriores a la creación de este concepto, pero que son antecedentes que han ayudado a la definición de la misma (Cano, 2007).

- 1969: Creación del concepto de Base de Datos por Codd.
- 1970-1979: Las primeras bases de datos se desarrollan al igual que las primeras aplicaciones (SAP, JD Edward, Siebel, PeopleSoft). Las mismas que ofrecieron la posibilidad de realizar data entry en los sistemas, aumentando la información disponible, pero no fueron capaces de ofrecer un acceso rápido y fácil a la misma.
- 1980-1988: Se crea el concepto de Almacén de datos (Ralph Kimball, Bill Inmon), y aparición de los primeros sistema de reporting. La información que estos ofrecían eran escasa y complicada para el usuario. Para ese momento ya existían sistemas de bases de datos relativamente potentes, pero no existían herramientas que faciliten la explotación de la misma.
- 1989: Introducción del término BI (Howard Dresner).

- 1990-199: BI 1.0. Proliferación de múltiples aplicaciones BI. Los proveedores resultaban costosos, pero el acceso a la información era fácil. En algunos casos agravaron el problema que pretendían resolver.
- 2000-2009: BI 2.0. Se consolidan aplicaciones de BI en algunas plataformas (Oracle, SAP, IBM, Microsoft, etc.). A parte de la información estructurada, se empieza a considerar otro tipo de información y documentos no estructurados.

2.2 DEFINICIÓN DE BI

Se denomina inteligencia empresarial, inteligencia de negocios o BI (del inglés business intelligence), al conjunto de estrategias y aspectos relevantes enfocados a la administración y creación de conocimiento sobre el medio, a través del análisis de los datos existentes en una organización o empresa.

“Las metodología BI utilizan la información para mejorar la gestión de empresas. Gracias al software de BI, los usuarios pueden acceder y analizar los datos con facilidad, y tomar mejores decisiones”(Gartner, 2012).

“BI es una alternativa tecnológica y de administración de negocios, que cubre todos los aspectos del manejo de información para la toma de decisiones, desde su extracción en los sistemas, depuración, transformación diseño de estructuras de datos o modelos especiales para el almacenamiento de datos, hasta la explotación de la información mediante herramientas comerciales de fácil uso para los usuarios. Este concepto es llamado también DataWarehouse” (Dongen V. , 2009).

“BI es un proceso interactivo para explorar y analizar información estructurada sobre un área (normalmente almacenada en un datawarehouse), para descubrir tendencias o patrones, a partir de los cuales derivar ideas y extraer conclusiones. El proceso de BI incluye la comunicación de los

descubrimientos y efectuar los cambios. Las áreas incluyen clientes, proveedores, productos, servicios y competidores (Dario, 2007).

A partir de las definiciones mencionadas anteriormente, podemos concluir que el BI nos brinda una herramienta que ayuda a la toma de decisiones, puesto que presenta la información de manera ordenada, rápida y oportuna, permitiendo a las empresas profundizar en su análisis, ya que se obtienen indicadores claves que ayudan a la gestión y manejo de la misma.

BI recolecta los datos y los presenta en información efectiva, de manera que el usuario pueda transformarlo en conocimiento necesario para tomar estrategias de mejora en la empresa que se implemente, permitiendo la toma de decisiones eficaz para una efectiva competencia en su entorno comercial.

Una frase popular en el mundo de BI dice: “Business Intelligence es el proceso de convertir datos en conocimiento, y el conocimiento en acción, para la toma de decisiones” (Davenport, 1999).

2.2.1 CARACTERÍSTICAS DE BI

Las principales características son:

- Ayuda en la toma de decisiones: posee herramientas de visualización avanzadas como: gráficas, tablas, velocímetro, que ayudan a obtener rápidos tiempos de respuestas, permitiendo una la fácil navegación, selección y manipulación de la información según el inter del usuario.
- Acceso a la información: Genera datos de calidad, completos, correctos y coherentes, también nos brinda el ingreso a los datos de manera independiente.
- Orientación al usuario final: se busca el manejo de interfaces amigables, que permitan al usuario la facilidad para intuir, sin la necesidad de previo conocimiento técnico para su uso.

2.2.2 PROCESOS DE BI

1. Dirigir y planear: Es la fase inicial en la cual se recolectan los requerimientos de la información específicos de los distintos usuarios, de esta manera se comprenderán sus distintas necesidades, que ayuden a generar las distintas preguntas que ayudarán a alcanzar los objetivos.

2. Recolección de Información: Se realiza el proceso de extracción de distintas fuentes de información de la empresa, puede ser de manera interna o externa, esto ayudará a encontrar las respuestas a las preguntas planteadas en el anterior paso.

3. Procesamiento de Datos: Se encarga de integrar los datos a partir de su forma más rústica en un formato que se utiliza para el análisis, se creará una base de datos completamente nueva para que se consolide la información.

4. Análisis y Producción: se procede a trabajar sobre datos extraídos e integrados, se utiliza las herramientas y técnicas que nos brinda el BI, para crear la inteligencia, el resultado a las preguntas planteadas en un inicio se generan mediante la creación de reportes, indicadores, entre otros.

5. Difusión: Es la fase final, la cual permite entregar al usuario final las herramientas adecuadas que le permitan interactuar con los datos de manera sencilla y rápida.

2.2.3 COMPONENTES DE BI

Definir que se requiere para la gestión y toma de decisiones es muy importante para un proyecto de BI, además de donde se obtendrán los datos y cuál es la disponibilidad que se requiere, establecer el formato y la navegabilidad que el usuario necesita.

2.2.4 DATAWAREHOUSE

“Un DataWareHouse es un conjunto de datos orientados a temas, integrados, no volátil, estable y que se usa para el proceso de tamo de decisiones” (Kimball, 2002).

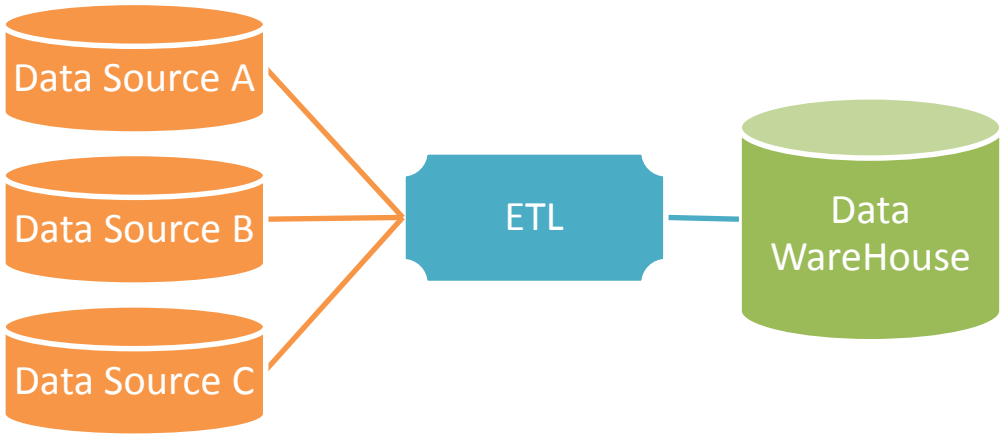

Figura 1. DataWareHouse

DataWareHouse almacena una gran cantidad de información histórica del negocio, aísla los sistemas operacionales de las necesidades de información para la gestión. Un cambio en los sistemas operacionales no debe afectar al DW/DM (Golden, 2004).

El DataWareHouse se alimenta a partir de los datos operacionales mediante las herramientas ETL(extract, transform, load).

Extraer: Hace referencia a la captura de la información desde los sistemas operacionales, necesidad de integración:

- Bases de datos
- Documentos impresos
- Ficheros planos

Transformar: Se refiere a la adaptación de los datos fuente con el formato destino definido en el DataWareHouse:

- Agregar datos numéricos
- Transponer información
- Componer

Cargar: Hace referencia al proceso en que los nuevos datos son finalmente almacenados en el Data WareHouse en su formato definitivo.

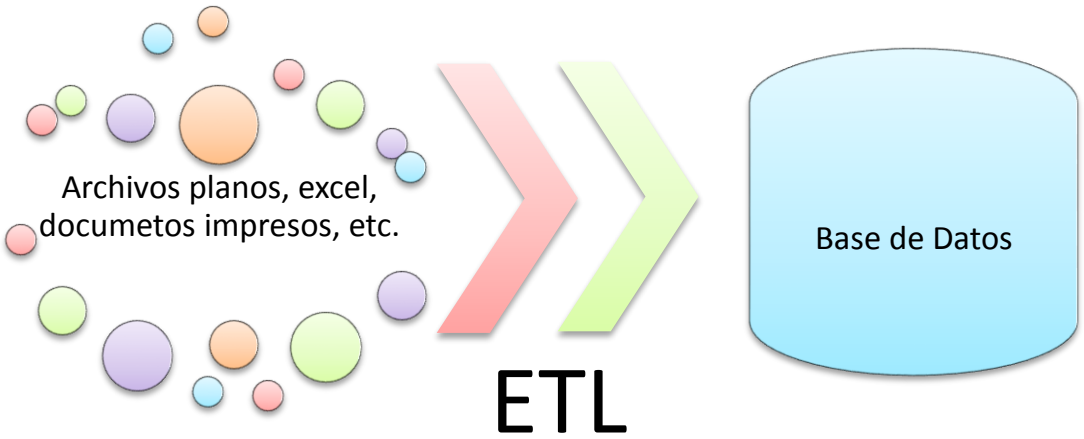

Figura 2. Extract, Transform and Load
(Dongen V. , 2009)

2.2.4.1 CARACTÉRISTICAS DE UN DATAWAREHOUSE

1. **Orientada al negocio:** La primera característica del DW, es que la información se clasifica en base a los aspectos que son de interés para la organización. Esta clasificación afecta al diseño y la implementación de los datos encontrados en el almacén de datos, debido a que la estructura del mismo difiere considerablemente a la de los clásicos procesos operacionales orientados a las aplicaciones.

2. **Integrada:** Existe la posibilidad que en DW, se maneje grades volúmenes de datos por esta razón el acceso puede demorar al realizar consulta, pero esta característica es muy distinta a la información que se encuentra en el ambiente operacional, la cual se obtiene en el momento mismo del acceso.

3. **Variante en el tiempo:** En un DW se maneja grades volúmenes de datos por esta razón el acceso puede demorar al realizar consulta, pero esta característica es muy distinta a la información que se encuentra en el ambiente operacional, la cual se obtiene en el momento mismo del acceso.

4. **No volátil:** El acceso y carga de los datos no cambia, por tal razón no se requiere un mecanismo de control de recuperación concurrente.

Figura 3. Características Data WareHouse

La actualización de los datos se hace de forma habitual en el ambiente operacional sobre una base de datos, registro por registro, en cambio en el depósito de datos la manipulación básica de los datos es mucho más simple, debido a que solo existen dos tipos de operaciones: la carga y el acceso a los mismos (Students, 2011).

Por tal motivo el DW no requiere un mecanismo de control de concurrencia y recuperación.

2.2.5 DATA MARTS

Los Data Marts, son almacenes de información específica que apuntan a un área de negocio en particular. El concepto en este caso se deriva de la certeza que cualquier usuario tiene la necesidad de información limitada, y aunque pueden existir requerimientos para análisis funcionales cruzados, el tamaño

de los mismos es reducido materialmente si limitamos el tamaño del DW en si mismo (Kavanagh, 2004).

Los principales beneficios de utilizar Data Marts son:

- Acelerar las consultas reduciendo el volumen de datos a recorrer.
- Estructurar los datos para su adecuado acceso por una herramienta.
- Segmentar los datos para asignar estrategias de control de acceso.
- Permite el acceso a los datos por medio de un gran número de herramientas del mercado, logrando independencia de estas.

Existen dos estrategias a partir del concepto de DataMarts, la de Data Marts dependiente y Data Marts independiente.

2.2.5.1 DATA MARTS DEPENDIENTES

Para esta arquitectura los datos son cargados desde los sistemas de producción hacia el DW empresarial y entonces subdivididos en Data Marts. Se llaman Data Marts dependientes porque se utilizan los datos y metadatos del DW en lugar de obtenerlos de los sistemas de producción (Carbone, 2001).

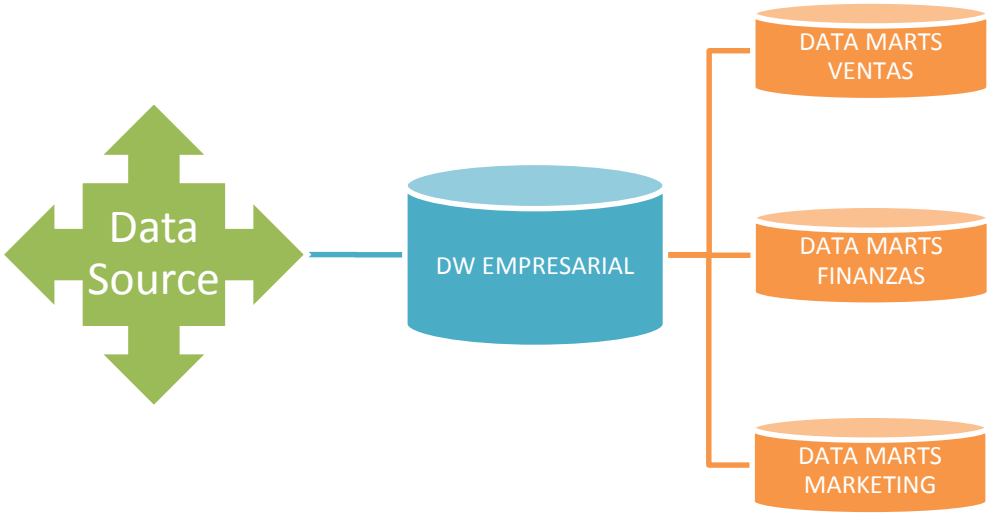

Figura 4. Data Marts Dependientes

2.2.5.2 DATA MARTS INDEPENDIENTES

Esta arquitectura es considerada por muchos como una alternativa del DW centralizado. Con ella es posible comenzar con un sistema pequeño, invirtiendo menos dinero y obteniendo resultados limitados entre tres a seis meses. Los que proponen esta arquitectura, argumentan que luego de comenzar con DM pequeños, otros DM pueden proliferar en otras líneas de negocios o departamentos que tengan las necesidades en su área. En este caso de DM multiples también se tienen procesos de carga múltiples donde los datos extraídos desde sistemas de producción quizás en forma redundante (Roberto, 2010).

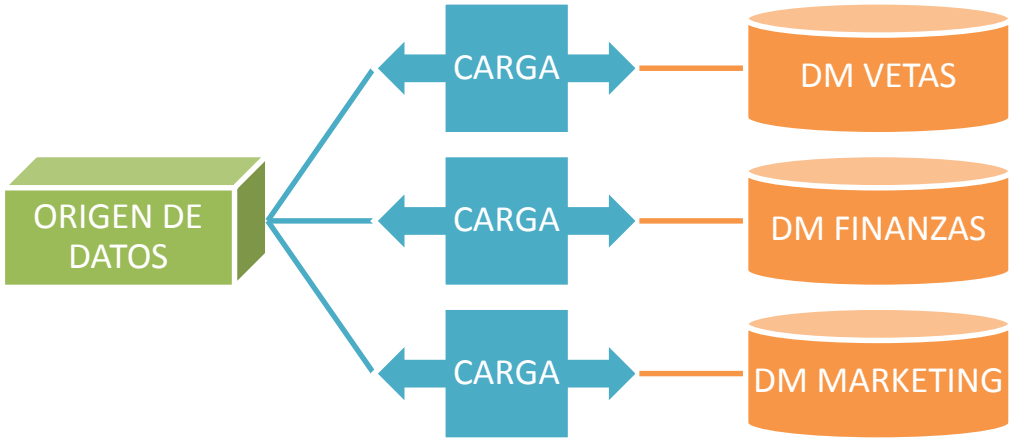

Figura 5. Data Marts Independiente

2.2.6 PROCESAMIENTO ANALÍTICO EN LÍNEA (OLAP)

La tecnología de OLAP permite un uso más eficaz de los DW para el análisis de datos en línea, lo que proporciona respuestas rápidas a consultas analíticas complejas e iterativas utilizando generalmente para sistemas de ayuda en la toma de decisiones.

OLAP muestra los datos a través de un modelo de datos intuitivo y natural. Con este estilo de navegación, los usuarios finales pueden ver y entender más efectivamente la información de sus bases de datos, permitiendo que la organización reconozca el valor de sus datos.

OLAP acelera la entrega de información a los usuarios finales que miran estas estructuras de datos como Cubos Analíticos o denominados también Cubos Multidimensionales debido a que la información es vista en varias dimensiones. Esta entrega es optimizada ya que se pre-calculan algunos valores, en vez de realizar el cálculo al momento de su solicitud. La

combinación de navegación fácil y rápida permite a los usuarios ver y analizar información más rápida y eficiente a diferencia de otras tecnologías de bases de datos relacionales (Kimball R. , 2002).

En conclusión, el usuario gasta menos tiempo analizando las bases de datos y dedica un porcentaje mayor de tiempo al análisis de la información presentada.

2.2.6.1 CARACTERÍSTICAS PRINCIPALES DEL OLAP

- El usuario puede ver la información de manera rápida y constante. La mayoría de peticiones que el usuario haga se deben responder en cinco segundos o menos.

- Se realiza análisis estadísticos y numéricos básicos de los datos, predefinidos por el desarrollador de la aplicación o definidos en “ad hoc” por el usuario.

- Se implementan requerimientos de seguridad que son necesarios cuando los usuarios desean compartir información confidencial a través de una gran población de los mismos.

- La multidimensionalidad es una característica esencial del OLAP, que es ver la información de determinadas vistas o dimensiones.

- El libre acceso a los datos y a la información necesaria y relevante para la aplicación, donde sea que este alojada y no este limitada por el volumen.

2.3 BASES DE DATOS MULTIDIMENSIONALES

Una base de datos multidimensional es un a base de datos donde su información se almacena en forma multidimensional, es decir, en varias dimensiones a través de tablas de hechos y tablas de dimensiones.

Proveen una estructura que permite, a través de la creación y consulta a una estructura de datos determinada generalmente cubos multidimensionales, tener acceso flexible a los datos, para explorar y analizar sus relaciones (Inmon W. H., 2002).

Las bases de datos multidimensionales implican tres variantes posibles de modelamiento, que permiten realizar consultas de soporte de decisión:

- Esquema en estrella (Star Scheme).
- Esquema copo de nieve (Snowflake Scheme).
- Esquema constelación o copo de estrellas.

Los mencionados esquemas pueden ser implementados de diversas maneras, independiente al tipo de arquitectura, requieren que toda la estructura de datos este desnormalizada o semi desnormalizada, para evitar desarrollar uniones complejas que impiden el acceso a la información, con el fin de agilizar la ejecución de consultas. Los diferentes tipos de implementación son los siguientes:

- Relación – ROLAP
- Multidimensionalidad – MOLAP
- Híbrido – HOLAP

2.3.1 TABLA DE DIMENSIONES

Se definen las tablas de dimensiones según su organización lógica de los datos y proveen el medio para analizar el contexto del negocio. Contienen

datos cualitativos. Representan los aspectos de interés, mediante los cuales los usuarios podrán filtrar y manipular la información almacenada en la tabla de hechos (Larissa T. Moss, 2003).

La siguiente figura muestra un ejemplo de tabla de dimensiones.

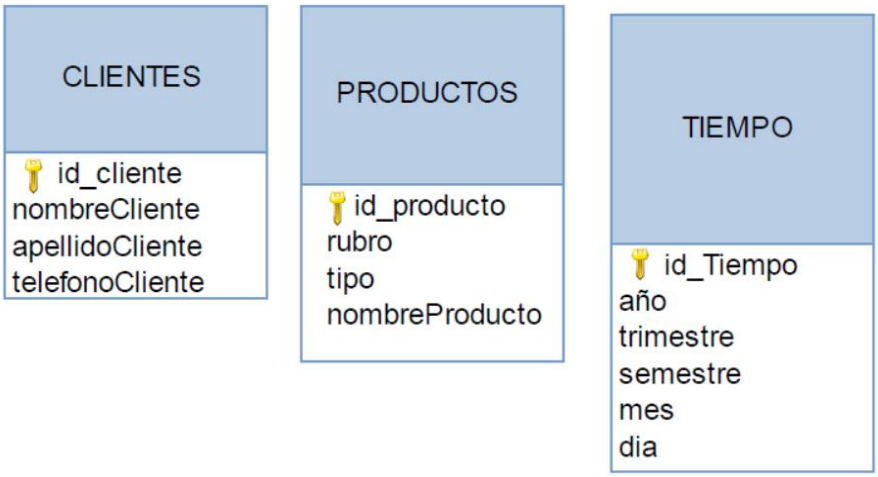

Figura 6. Ejemplo Tabla de dimensiones

Como se puede ver en el gráfico anterior, cada tabla posee un identificador único (primary key) y al menos un campo o atributo de referencia que describe los criterios de análisis relevantes para la organización, estos son por lo general de tipo texto.

Usualmente la cantidad de tablas de dimensiones, aplicadas a un tema de interés en particular, varían entre tres y quince.

Es recomendable manejar un sistema de claves en el DW (Claves Subrogadas) totalmente diferentes al de los OLTP, ya que si estos últimos son recodificados, el almacén quedaría inconsistente y debería ser poblado nuevamente en su totalidad.

2.3.2 TABLA DE DIMENSIÓN TIEMPO

En un DW, la creación y el mantenimiento de una tabla de dimensión Tiempo es obligatoria, y la definición de granularidad y estructuración de la misma depende de la dinámica del negocio que se esté analizando. Toda la información dentro del depósito, como ya se ha explicado, posee su propio sello de tiempo que determina la ocurrencia de hecho específico, representando de esta manera diferentes versiones de una misma situación.

Es importante tener en cuenta que la dimensión de Tiempo no es solo un secuencia cronológica representada de forma numérica, si no que mantiene niveles jerárquicos especiales que inciden notablemente en las actividades de la organización. Esto se debe a que los usuarios podrán por ejemplo analizar los socios ingresados en determinado periodo de año, trimestre, mes, semana, día.

2.3.3 TABLA DE HECHOS

Las tablas de hechos contienen hechos que serán utilizados por los analistas de negocios para apoyar el proceso de toma de decisiones, conformado generalmente por datos cuantitativos. Los hechos son datos instantáneos en el tiempo, que son filtrados, agrupados y explorados a través de condiciones definidas en la tabla de dimensiones.

La tabla de hechos posee una clave primaria que está compuesta por las claves primarias de las tablas de dimensiones relacionadas a este.

La siguiente figura muestra un ejemplo de tabla de hechos.

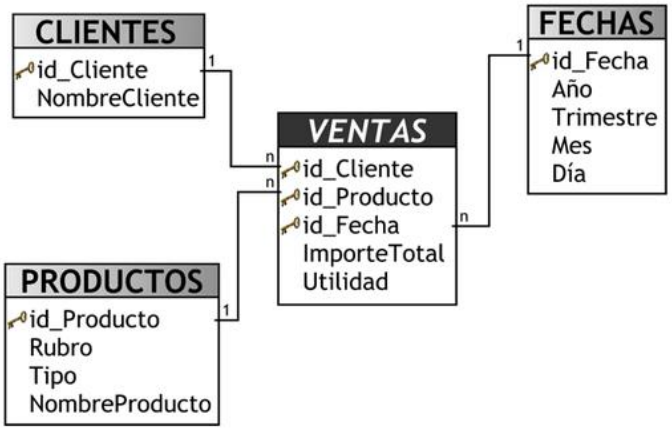

Figura 7. Ejemplo tabla de hechos

Como se muestra en la imagen la tabla de hechos VENTAS y a los costados se encuentran las tablas de dimensiones PRODUCTOS, CLIENTES Y TIEMPO que se encuentran conectadas a la tabla VENTAS por claves principales de las tablas de dimensiones. Los hechos en este ejemplo son utilidad y ventas totales.

Los hechos son aquellos datos que residen en una tabla de hechos y que son utilizadas para crear indicadores, a través de sumarizaciones preestablecidas al memento de crear un cubo dimensional.

2.4 TIPOS DE ESQUEMAS PARA UN DATAWAREHOUSE

A continuación se detallan los diferentes esquemas que se utilizan para la creación de un DW.

2.4.1 ESQUEMA EN ESTRELLA

Consiste en estructurar la información en procesos, vistas y métricas recordando a una estrella. Es decir, tendremos una visión multidimensional de un proceso que medimos a través de una métrica.

A nivel de diseño, consiste en una tabla de hechos (fact table) en el centro para el hecho objeto de análisis y una o varias tablas de dimensiones (dimension table) para cada dimensión de análisis que participa de las descripciones de ese hecho. En la tabla de hecho se encuentran los atributos destinados a medir (cuantificar) el hecho.

Mientras en las tablas de dimensiones, los atributos se destinan a elementos de nivel (que representan los distintos niveles de las jerarquías de dimensiones) y a atributos de dimensiones (encargados de la descripción de estos elementos de nivel).

El esquema en estrella, la tabla de hechos es la única tabla del esquema que tiene múltiples joins que la conectan con otras tablas (foreign key hacia otras tablas). El resto de las tablas del esquema llamadas tablas de dimensión únicamente hacen join con la tabla de hechos. Las tablas de dimensión se encuentran además totalmente denormalizadas, es decir, toda la información referente a un dimensión se almacena en la misma tabla.

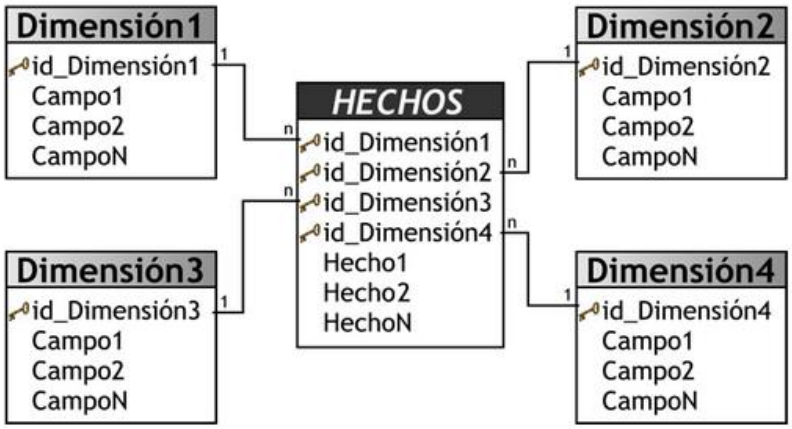

Figura 8. Esquema en Estrella

La mayoría de los Data WareHouse están diseñados en base al esquema en estrella para representar el sistema de datos multidimensional. El sistema en estrella se caracteriza por tener una o más tablas de hechos que contienen la información principal del Data WareHouse, y un número indeterminado de tablas de dimensión.

Cada una de las tablas de dimensión contiene la información sobre las entradas de un determinado atributo en la tabla de hechos. Cada tabla de dimensión está relacionada con la tabla de hechos mediante el sistema de clave primaria (clave ajena). Las dimensiones no se relacionan entre sí. Una tabla de hechos contiene claves y medidas.

2.4.2 ESQUEMA COPO DE NIEVE

Es una extensión del esquema en Estrella, esta posee una tabla de hechos central y las tablas de dimensión están relacionadas a este mediante claves, pero a su vez las tablas de dimensiones pueden relacionarse con otras tablas de dimensión.

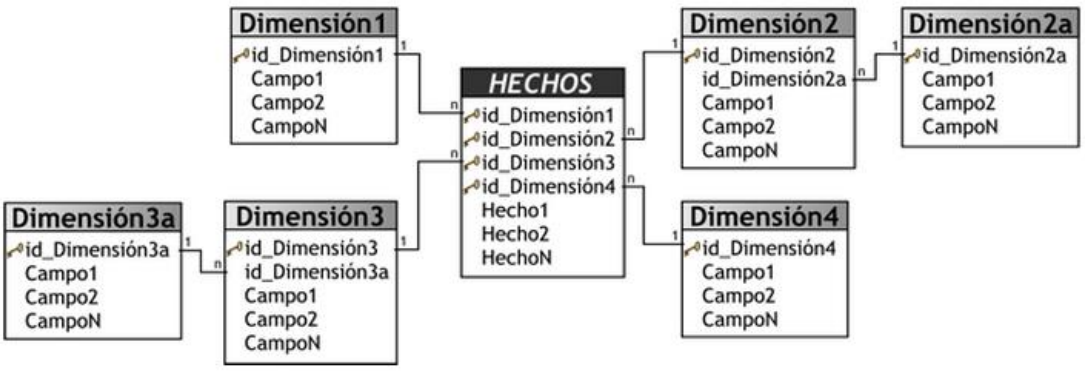

Figura 9. Esquema Copo de Nieve

De este esquema se puede decir que es el más parecido al modelo entidad relación ya que las tablas de dimensiones están normalizadas.

La finalidad de normalizar las tablas es reducir el espacio de almacenamiento al eliminar la redundancia de datos; pero tiene la contrapartida de generar peores rendimientos al tener que crear más tablas de dimensiones y más relaciones entre tablas (joins) lo que tiene un impacto directo sobre el rendimiento.

2.4.3 ESQUEMA CONSTELACIÓN

Este esquema es más complejo que las otras arquitecturas debido a que contiene múltiples tablas de hechos. Con esta solución las tablas de dimensiones pueden estar compartidas entre varias tablas de hechos.

El esquema de constelación de hechos tiene mucha flexibilidad y esta característica es su gran virtud.

Sin embargo, el problema es que cuando el número de las tablas vinculadas aumenta, la arquitectura puede llegar a ser muy compleja y difícil para mantener.

En la siguiente ilustración se indica la conformación de un esquema tipo constelación.

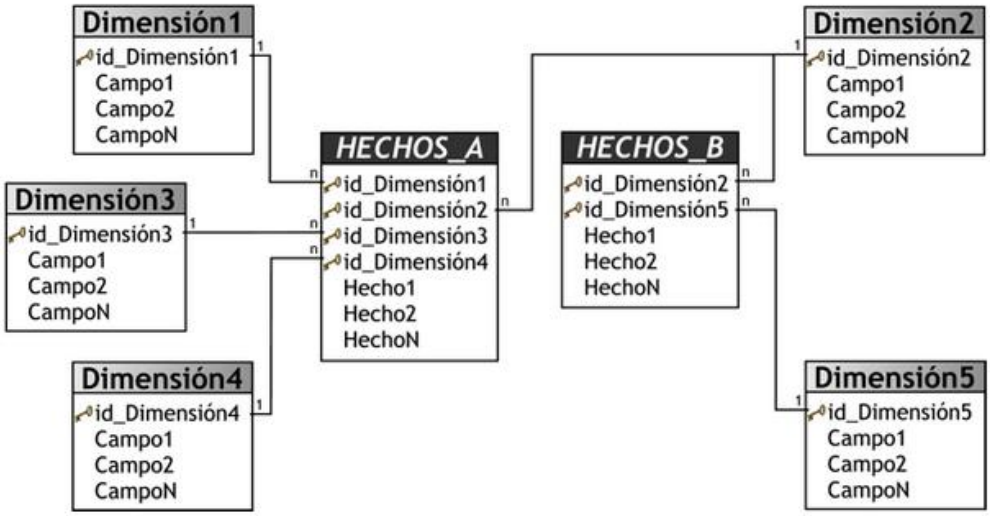

Figura 10. Esquema Constelación

2.5 TIPOS DE IMPLEMENTACIÓN DE UN DATA WAREHOUSE

La implementación de un Data Warehouse basa principalmente en la tecnología OLAP que permite trabajar sobre los datos al pensarlos como cubos multidimensionales o hipercubos con información de la empresa. Existen diferentes maneras en las que se puede implementar un Data Warehouse con la única diferencia que se transa menor tiempo de acceso a cambio de mayor espacio de utilización de disco y viceversa.

Existen tres modelos posibles:

- Procesamiento analítico relacional en línea (ROLAP)
- Procesamiento analítico multidimensional en línea (MOLAP)
- Procesamiento analítico en línea híbrido (HOLAP)

2.5.1 ROLAP

Significa Procesamiento Analítico OnLine Relacional, es decir, se trata de sistemas y herramientas OLAP (Procesamiento Analítico OnLine) construidos sobre una base de datos relacional. Es una alternativa a la tecnología MOLAP

(Multidimensional OLAP) que se construye sobre bases de datos multidimensionales. Ambos tipos de herramientas, tanto ROLAP como MOLAP, están diseñadas para realizar análisis de datos a través del uso de modelos de datos multidimensionales, aunque en el caso de ROLAP estos modelos no se implementan sobre un sistema multidimensional, sino sobre un sistema relacional clásico.

2.5.2 MOLAP

Significa procesamiento analítico multidimensional en línea. En esta implementación los datos se almacenan en forma multidimensional permitiendo operaciones rápidas de búsquedas y resúmenes gracias a los datos pre-calculados existentes. Como limitante, ocupa más espacio en disco para almacenamiento, pero como se sabe el costo de espacio en disco es cada vez más barato.
Consta de dos niveles, el nivel de la base de datos multidimensional (MDDB) que se encarga de manejar, acceder y obtener los datos y del motor analítico que ejecuta la consulta de los usuarios.

2.5.3 HOLAP

Significa procesamiento analítico en línea híbrido y es una combinación de ROLAP y MOLAP. HOLAP permite almacenar una parte de datos como ROLAP y otra como MOLAP. Utiliza las características de MOLAP para mejorar el acceso a las consultas y ROLAP para optimizar el tiempo en que se procesa el cubo.

2.6 CUBO ANALÍTICO

Un cubo multidimensional, representa o convierte los datos planos que se encuentran en filas y columnas, en una matriz de N dimensiones.

Los objetos más importantes que se pueden incluir en un cubo multidimensional, son los siguientes:

- Indicadores: sumarizaciones que se efectúan sobre algún hecho o expresiones basadas en sumarizaciones, pertenecientes a una tabla de hechos.
- Atributos: campos o criterios de análisis, pertenecientes a tablas de dimensiones.
- Jerarquías: representa una relación lógica entre dos o más atributos.

De esta manera en un cubo multidimensional, los atributos existen a lo largo de varios ejes o dimensiones, y la intersección de las mismas representa el valor que tomará el indicador que se está evaluando.

En la siguiente representación matricial se puede ver más claramente lo que se acaba de decir.

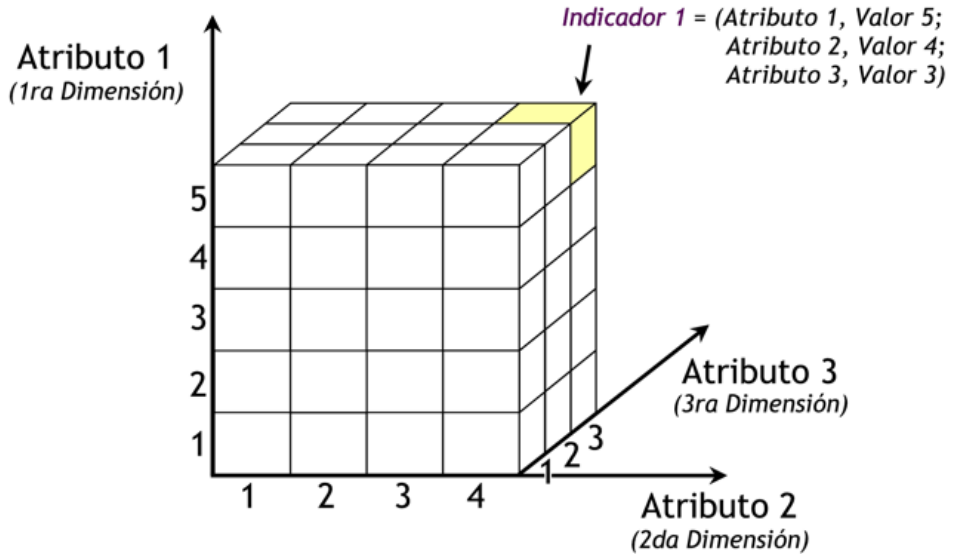

Figura 11. Cubo Multidimensional

Para la creación del cubo de la figura anterior, se definieron tres Atributos (“Atributo 1”, “Atributo 2” y “Atributo 3”) y se definió un Indicador (“Indicador 1”).

Entonces el cubo quedó compuesto por 3 dimensiones o ejes (una por cada Atributo), cada una con sus respectivos valores asociados. También, se ha seleccionado una intersección al azar para demostrar la correspondencia con los valores de las Atributos. En este caso, el indicador “Indicador 1”, representa el cruce del Valor “5” de “Atributo 1”, con el Valor “4” de “Atributo 2” y con el Valor “3” de “Atributo 3”.

Se puede observar, que el resultado del análisis está dado por los cruces matriciales de acuerdo a los valores de las dimensiones seleccionadas.
Más específicamente, para acceder a los datos del DW, se pueden ejecutar consultas sobre algún cubo multidimensional previamente definido. Dicho cubo debe incluir entre otros objetos: indicadores, atributos, jerarquías, etc., basados en los campos de las tablas de dimensiones y de hechos, que se deseen analizar. De esta manera, las consultas son respondidas con gran

performance, minimizando al máximo el tiempo que se hubiese incurrido en realizar dicha consulta sobre una base de datos transaccional.

2.6.1 INDICADORES

Los indicadores son sumarizaciones efectuadas sobre algún hecho o expresiones basadas en sumarizaciones, que serán incluidos en algún cubo multidimensional, con el fin de analizar los datos almacenados en el DW. El valor que estos adopten estará condicionado por los atributos/jerarquías que se utilicen para analizarlos.

Los indicadores, además de hechos, pueden estar compuestos por otros indicadores, pero no ambos simultáneamente. Pueden utilizarse para su creación funciones de sumarización (suma, conteo, promedio), funciones matemáticas, estadísticas, operadores matemáticos y lógicos.

2.6.2 ATRIBUTOS

Los atributos constituyen los criterios de análisis que se utilizarán para analizar los indicadores dentro de un cubo multidimensional. Los mismos se basan, en su gran mayoría, en los campos de las tablas de dimensiones y/o expresiones.

2.6.3 JERARQUÍAS

Una jerarquía representa una relación lógica entre dos o más atributos pertenecientes a un cubo multidimensional; siempre y cuando posean su correspondiente relación “padre-hijo”.

Las jerarquías poseen las siguientes características: Pueden existir varias en un mismo cubo.

Están compuestas por dos o más niveles.
Se tiene una relación “1-n” o “padre-hijo” entre atributos consecutivos de un nivel superior y uno inferior.

Por lo general, las jerarquías pueden identificarse fácilmente, debido a que existen relaciones “1-n” o “padre-hijo” entre los propios atributos de un cubo.

La principal ventaja de manejar jerarquías, reside en poder analizar los datos desde su nivel más general al más detallado y viceversa, al desplazarse por los diferentes niveles.

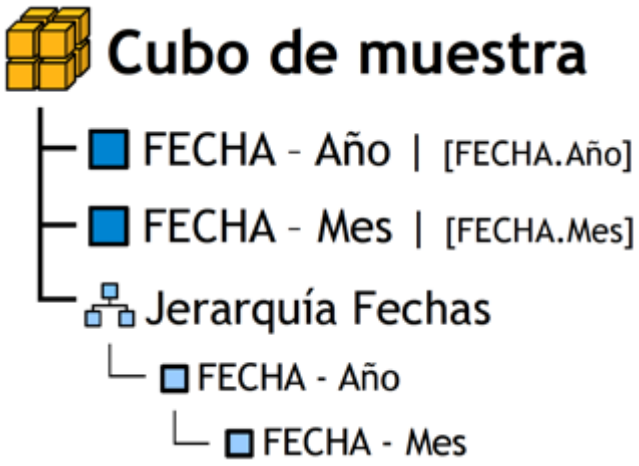

Figura 12. Cubo multidimensional, Jerarquía

2.7 METODOLOGÍAS

A continuación se detallan las diferentes metodologías que se utilizan para la construcción de un cubo analítico, y se detallan cada una de las etapas que deben cumplir para su correcta implementación.

2.7.1 METODOLOGÍA INMON

Innmon ve la necesidad de transferir la información de los diferentes OLTP (Sistemas Transaccionales) de las organizaciones a un lugar centralizado donde los datos puedan ser utilizados para el análisis a la Fábrica de Información Corporativa (CIF o Corporate Information Factory).

Insiste además en que ha de tener las siguientes características:

- Orientado a temas: Los datos en la base de datos están organizados de manera que todos los elementos de datos relativos al mismo evento u objeto del mundo real queden unidos entre sí.
- Integrado: La base de datos contiene los datos de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes.
- No volátil: La información no se modifica ni se elimina, una vez almacenado un dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas.
- Variante en el tiempo: Los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones.

La información debe estar a los máximos niveles de detalle. Los Data Warehouse departamentales o data marts son tratados como subconjuntos de este Data Warehouse corporativo, son construidos para cubrir las necesidades individuales de análisis de cada departamento, y siempre a partir del Data Warehouse central (Inmon B. , 1992).

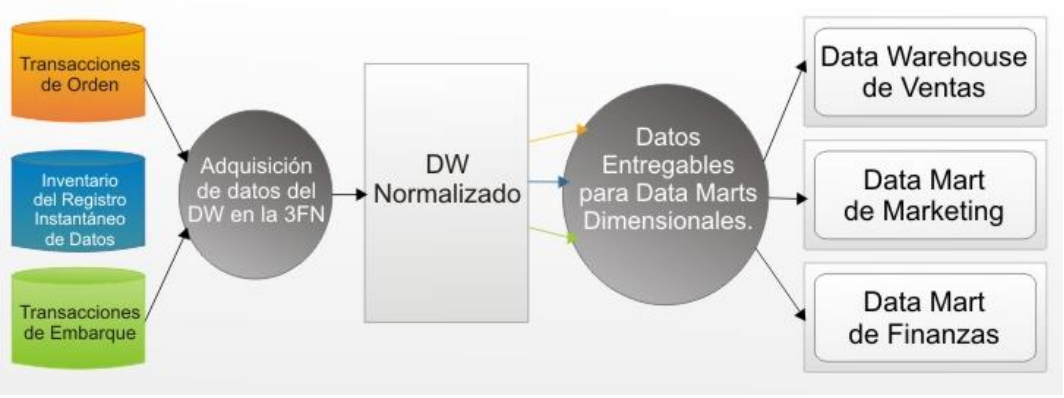

Figura 13. Metodología Inmon, DW Corporativo

La metodología Innmon también se referencia normalmente como Top-down. Los datos son extraídos de los sistemas operacionales por los procesos ETL y cargados en las “area stage”, donde son validados y consolidados en el DW corporativo, donde además existen los llamados metadatos que documentan de una forma clara y precisa el contenido del DW. Una vez realizado este proceso, los procesos de refresco de los Data mart departamentales obtienen la información de este, y con las consiguientes transformaciones, organizan los datos en las estructuras particulares requeridas por cada uno de ellos, refrescando su contenido.

La metodología para la construcción de un sistema de este tipo es la habitual para construir un sistema de información, utilizando las herramientas habituales (esquema Entidad Relación, DIS (Data Item Sets). Para el tratamiento de los cambios en los datos, usa la Gestión de las dimensiones continuas y discretas (inserta fechas en los datos para determinar su validez para la dimensión continua o bien mediante el concepto de snapshot o foto para la dimensión discreta).

Al tener este enfoque global, es más difícil de desarrollar en un proyecto sencillo (pues se intentará abordar el todo, a partir del cual luego se irá al detalle).

2.7.2 METODOLOGÍA HEFESTO

La metodología mencionada anteriormente da un enfoque diferente para implementar un Data Warehouse. Una metodología adicional que combina los objetivos de las metodologías anteriores se denomina Hefesto creada por Darío Bernabeu.

La metodología Hefesto parte de la recolección de requerimientos de información del usuario, seguido de los procesos de extracción, transformación y carga de datos (ETL) hasta definir un esquema lógico para la organización ya sean estos Data Marts o Data Warehouse.

La construcción e implementación de un DW puede adaptarse muy bien a cualquier ciclo de vida de desarrollo de software, con la salvedad de que para algunas fases en particular, las acciones que se han de realizar serán muy diferentes. Lo que se debe tener muy en cuenta, es no entrar en la utilización de metodologías que requieran fases extensas de reunión de requerimientos y análisis, fases de desarrollo monolítico que conlleve demasiado tiempo y fases de despliegue muy largas. Lo que se busca, es entregar una primera implementación que satisfaga una parte de las necesidades, para demostrar las ventajas del DW y motivar a los usuarios.

La metodología HEFESTO, puede ser embebida en cualquier ciclo de vida que cumpla con la condición antes declarada.

2.7.2.1 ETAPAS DE LA METODOLOGÍA HEFESTO

Las etapas que a continuación se detallan cumplen con las condiciones para la construcción de un DW y son una recopilación de las metodologías que se detallaron con anterioridad (Kimball e Inmon).

2.7.2.1.1 Análisis de requerimientos

Lo primero que se hará será identificar los requerimientos de los usuarios a través de preguntas que expliciten los objetivos de su organización. Luego, se analizarán estas preguntas a fin de identificar cuáles serán los indicadores y perspectivas que serán tomadas en cuenta para la construcción del DW. Finalmente se confeccionará un modelo conceptual en donde se podrá visualizar el resultado obtenido en este primer paso.

Es muy importante tener en cuenta que HEFESTO se puede utilizar para construir un Data Warehouse o un Data Mart a la vez, es decir, si se requiere construir por ejemplo dos Data Marts, se deberá aplicar la metodología dos veces, una por cada Data Mart. Del mismo modo, si se analizan dos áreas de interés de negocio, como el área de “Ventas” y “Compras”, se deberá aplicar la metodología dos veces.

a) Identificar preguntas

El primer paso comienza con el acopio de las necesidades de información, el cual puede llevarse a cabo a través de muy variadas y diferentes técnicas, cada una de las cuales poseen características inherentes y específicas, como por ejemplo entrevistas, cuestionarios, observaciones.

El análisis de los requerimientos de los diferentes usuarios, es el punto de partida de esta metodología, ya que ellos son los que deben, en cierto modo, guiar la investigación hacia un desarrollo que refleje claramente lo que se espera del depósito de datos, en relación a sus funciones y cualidades.

El objetivo principal de esta fase, es la de obtener e identificar las necesidades de información clave de alto nivel, que es esencial para llevar a cabo las metas y estrategias de la empresa, y que facilitará una eficaz y eficiente toma de decisiones.

b) Identificar indicadores y perspectivas

Una vez que se han establecido las preguntas de negocio, se debe proceder a su descomposición para descubrir los indicadores que se utilizarán y las perspectivas de análisis que intervendrán.

Para ello, se debe tener en cuenta que los indicadores, para que sean realmente efectivos son, en general, valores numéricos y representan lo que se desea analizar concretamente, por ejemplo: saldos, promedios, cantidades, sumatorias, fórmulas.

En cambio, las perspectivas se refieren a los objetos mediante los cuales se quiere examinar los indicadores, con el fin de responder a las preguntas planteadas, por ejemplo: clientes, proveedores, sucursales, países, productos, rubros, etc. Cabe destacar, que el Tiempo es muy comúnmente una perspectiva.

c) Modelo Conceptual

En esta etapa, se construirá un modelo conceptual a partir de los indicadores y perspectivas obtenidas en el paso anterior. Modelo Conceptual: descripción de alto nivel de la estructura de la base de datos, en la cual la información es representada a través de objetos, relaciones y atributos.

A través de este modelo, se podrá observar con claridad cuáles son los alcances del proyecto, para luego poder trabajar sobre ellos, además al poseer un alto nivel de definición de los datos, permite que pueda ser presentado ante los usuarios y explicado con facilidad.

2.7.2.1.2 Análisis de los OLTP

Seguidamente, se analizarán las fuentes OLTP para determinar cómo serán calculados los indicadores y para establecer las respectivas correspondencias entre el modelo conceptual creado en el paso anterior y las fuentes de datos.

Luego, se definirán qué campos se incluirán en cada perspectiva. Finalmente, se ampliará el modelo conceptual con la información obtenida en este paso.

a) Conformar Indicadores

En este paso se deberán explicitar como se calcularán los indicadores, definiendo los siguientes conceptos para cada uno de ellos:

- Hecho/s que lo componen, con su respectiva fórmula de cálculo. Por ejemplo: Hecho1 + Hecho2.
- Función de sumarización que se utilizará para su agregación. Por ejemplo: SUM, AVG, COUNT, etc.

b) Establecer correspondencias

El objetivo de este paso, es el de examinar los OLTP disponibles que contengan la información requerida, como así también sus características, para poder identificar las correspondencias entre el modelo conceptual y las fuentes de datos.

La idea es, que todos los elementos del modelo conceptual estén correspondidos en los OLTP.

c) Nivel de granularidad

Una vez que se han establecido las relaciones con los OLTP, se deben seleccionar los campos que contendrá cada perspectiva, ya que será a través de estos por los que se examinarán y filtrarán los indicadores.

Para ello, basándose en las correspondencias establecidas en el paso anterior, se debe presentar a los usuarios los datos de análisis disponibles para cada perspectiva. Es muy importante conocer en detalle que significa cada campo y/o valor de los datos encontrados en los OLTP, por lo cual, es conveniente investigar su sentido, ya sea a través de diccionarios de datos, reuniones con los encargados del sistema, análisis de los datos propiamente dichos.

Luego de exponer frente a los usuarios los datos existentes, explicando su significado, valores posibles y características, estos deben decidir cuáles son los que consideran relevantes para consultar los indicadores y cuáles no.

Con respecto a la perspectiva “Tiempo”, es muy importante definir el ámbito mediante el cual se agruparán o sumarizarán los datos. Sus campos posibles pueden ser: día de la semana, quincena, mes, trimestres, semestre, año.

Al momento de seleccionar los campos que integrarán cada perspectiva, debe prestarse mucha atención, ya que esta acción determinará la granularidad de la información encontrada en el DW.

d) Modelo Conceptual Ampliado

En este paso, y con el fin de graficar los resultados obtenidos en los pasos anteriores, se ampliará el modelo conceptual, colocando bajo cada perspectiva los campos seleccionados y bajo cada indicador su respectiva fórmula de cálculo.

2.7.2.1.3 Modelo Lógico del DW

A continuación, se confeccionará el modelo lógico de la estructura del DW, teniendo como base el modelo conceptual que ya ha sido creado. Para ello, primero se definirá el tipo de modelo que se utilizará y luego se llevarán a cabo las acciones propias al caso, para diseñar las tablas de dimensiones y de hechos. Finalmente, se realizarán las uniones pertinentes entre estas tablas. Modelo Lógico: representación de una estructura de datos, que puede procesarse y almacenarse en algún SGBD.

a) Tipo de Modelo Lógico del DW

Se debe seleccionar cuál será el tipo de esquema que se utilizará para contener la estructura del depósito de datos, que se adapte mejor a los requerimientos y necesidades de los usuarios. Es muy importante definir objetivamente si se empleará un esquema en estrella, constelación o copo de nieve, ya que esta decisión afectará considerablemente la elaboración del modelo lógico.

b) Tablas de dimensiones

En este paso se deben diseñar las tablas de dimensiones que formarán parte del DW.

Para los tres tipos de esquemas, cada perspectiva definida en el modelo conceptual constituirá una tabla de dimensión. Para ello deberá tomarse cada perspectiva con sus campos relacionados y realizarse el siguiente proceso:

- Se elegirá un nombre que identifique la tabla de dimensión.
- Se añadirá un campo que represente su clave principal.

Se redefinirán los nombres de los campos si es que no son lo suficientemente intuitivos.

c) Tablas de hechos

En este paso, se definirán las tablas de hechos, que son las que contendrán los hechos a través de los cuales se construirán los indicadores de estudio.

d) Uniones

Para los tres tipos de esquemas, se realizarán las uniones correspondientes entre sus tablas de dimensiones y sus tablas de hechos.

2.7.2.1.4 Integración de datos

Una vez construido el modelo lógico, se deberá proceder a poblarlo con datos, utilizando técnicas de limpieza y calidad de datos, procesos ETL; luego se definirán las reglas y políticas para su respectiva actualización, así como también los procesos que la llevarán a cabo.

a) Carga Inicial

Debemos en este paso realizar la Carga Inicial al DW, poblando el modelo de datos que hemos construido anteriormente. Para lo cual debemos llevar adelante una serie de tareas básicas, tales como limpieza de datos, calidad de datos, procesos ETL.

Se debe evitar que el DW sea cargado con valores faltantes o anómalos, así como también se deben establecer condiciones y restricciones para asegurar que solo se utilicen los datos de interés.

Cuando se trabaja con un esquema constelación, hay que tener presente que varias tablas de dimensiones serán compartidas con diferentes tablas de hechos, ya que puede darse el caso de que algunas restricciones aplicadas sobre una tabla de dimensión en particular para analizar una tabla de hechos, se puedan contraponer con otras restricciones o condiciones de análisis de otras tablas de hechos.

Primero se cargarán los datos de las dimensiones y luego los de las tablas de hechos, teniendo en cuenta siempre, la correcta correspondencia entre cada elemento. En el caso en que se esté utilizando un esquema copo de nieve, cada vez que existan jerarquías de dimensiones, se comenzarán cargando las tablas de dimensiones del nivel más general al más detallado.

Concretamente, en este paso se deberá registrar en detalle las acciones llevadas a cabo con diferente software. Por ejemplo, es muy común que

sistemas ETL trabajen con "pasos" y "relaciones", en donde cada "paso" realiza una tarea en particular del proceso ETL y cada "relación" indica hacia donde debe dirigirse el flujo de datos. En este caso lo que se debe hacer es explicar que hace el proceso en general y luego que hace cada "paso" y/o "relación". Es decir, se partirá de lo más general y se irá a lo más específico, para obtener de esta manera una visión general y detallada de todo el proceso.

Es importante tener presente, que al cargar los datos en las tablas de hechos pueden utilizarse pre-agregaciones, ya sea al nivel de granularidad de la misma o a otros niveles diferentes.

b) Actualización

Cuando se haya cargado en su totalidad el DW, se deben establecer sus políticas y estrategias de actualización o refresco de datos.

Una vez realizado esto, se tendrán que llevar a cabo las siguientes acciones:

- Especificar las tareas de limpieza de datos, calidad de datos, procesos ETL, que deberán realizarse para actualizar los datos del DW.

Especificar de forma general y detallada las acciones que deberá realizar cada software.

2.7.3 METODOLOGÍA RALPH KIMBALL

El Data Warehouse es un conglomerado de todos los Data marts dentro de una empresa, siendo una copia de los datos transaccionales estructurados de una forma especial para el análisis, de acuerdo al Modelo Dimensional (no normalizado), que incluye, como se explicó, las dimensiones de análisis y sus

atributos, su organización jerárquica, así como los diferentes hechos de negocio que se quieren analizar.

Por un lado se tiene las tablas para las representar las dimensiones y por otro lado tablas para los hechos. Los diferentes Data marts están conectados entre sí por la llamada estructura de bus, que contiene los elementos anteriormente citados a través de las dimensiones conformadas (que permiten que los usuarios puedan realizar consultas conjuntas sobre los diferentes data marts, pues este bus contiene los elementos en común que los comunican).

- Hechos: Es una colección de piezas de datos y datos de contexto. Cada hecho representa una parte del negocio, una transformación o evento.
- Dimensiones: Es una colección de miembros, unidades o individuos del mismo tipo.
- Medidas: Son atributos numéricos de un hecho que representan el comportamiento del negocio relativo o una dimensión.

Cada punto de entra a la tablas de hechos está conectada a una dimensión, lo que permite determinar el contexto de los hechos.

Una base de datos dimensional se puede concebir como un cubo de tres o cuatro dimensiones (OLAP), en el que los usuarios pueden acceder a un porción de la base de datos a lo largo de cualquier de sus dimensiones.

Esta metodología incluye las siguientes etapas:

- Planeación del Proyecto.
- Análisis de requerimientos.
- Selección e instalación de productos.
- Modelamiento Dimensional.
- Diseño Físico.
- Diseño y Desarrollo de la presentación de los datos.
- Especificación de aplicaciones para los usuarios finales.
- Mantenimiento y crecimiento.

2.8 PENTAHO

Pentaho Corp., empresa dueña de Pentaho, fue fundada en el año 2004 por pioneros en BI Open Source comercial provenientes de empresas como: Business Objects, Cognos, Hyperion, Jboss, Oracle, Red Hat y SAS.

Pentaho, provee una alternativa de soluciones de BI en distintas áreas como en la Arquitectura, Soporte, Funcionalidad e Implantación. Estas soluciones al igual que su ambiente de implantación están basados en JAVA, haciéndolo flexible en cubrir amplias necesidades empresariales. A través de la integración funcional de diversos proyectos de Open Source permite ofrecer soluciones en áreas como: Análisis de información, Reportes, Tableros de mando conocido como “DashBoards”, Flujos de Trabajo y Minería de Datos.

Pentaho ofrece dos versiones de su solución en base al modelo anteriormente descrito: una versión comunitaria gratuita orientada principalmente al mundo académico, y una versión Enterprise comercial orientada a la implementación profesional tanto en empresas privadas como en instituciones gubernamentales u otras sin fines de lucros que pretendan potenciar sus capacidades analíticas para mejorar su gestión. El siguiente esquema resume el modelo:

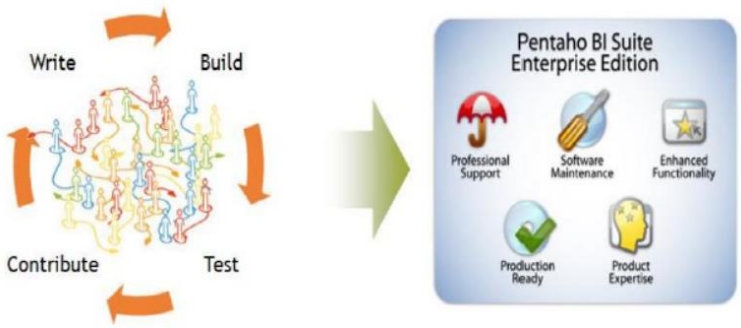

Figura 14. Modelo de las versiones de Pentaho
(Dongen B. &., 2009)

2.8.1 MÓDULO DE SISTEMAS

En esta parte de la tesis se describen los módulos con los que cuenta Pentaho y se realiza una descripción breve de su utilización dentro de la Suite.

Cabe destacar que todos los módulos aquí descritos sirven para crear una solución de Inteligencia de negocios completa e integral sin requerir de otros complementos.

2.8.1.1 REPORTING

Un módulo de los informes ofrece la solución adecuada a las necesidades de los usuarios. Pentaho Reporting es una solución basada en el proyecto JFreeReport y permite generar informes ágil y de gran capacidad. Pentaho Reporting permite la distribución de los resultados del análisis en múltiples formatos – todos los informes incluyen la opción de imprimir o exportar a formato PDF, XLS, HTML y texto. Los reportes Pentaho permiten también programación de tareas y ejecución automática de informes con una determinada periodicidad.

2.8.1.2 ANÁLISIS

Pentaho Análisis suministra a los usuarios un sistema avanzado de análisis de información. Con uso de las tablas dinámicas (pivot tables, crosstabs), generadas por Mondrian y JPivot, el usuario puede navegar por los datos, ajustando la 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 mineria 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).

2.8.1.3 DASHBOARDS

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

2.8.1.4 DATA MINING

Análisis en Pentaho se realiza con una herramienta WeKa.

2.8.1.5 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 – PDI 3.0 – que marcó un gran paso adelante en OSBI ETL y que hizo Pentaho Data Integration una alternativa interesante para las herramientas comerciales.

2.9 COMPAÑÍAS DE SEGUROS

Son instituciones financieras especializadas en asumir riesgos de terceros mediante la expedición de pólizas de seguros. Las partes que intervienen en un contrato de seguros son el tomador que es la persona que traslada los riesgos, el asegurado cuya vida o patrimonio se asegura y la empresa aseguradora que se encarga de asumir los riesgos (Efrén, 1998).

Las operaciones de seguros únicamente pueden ser realizadas por las empresas de seguros autorizadas por la Ley de Empresas de Seguros y Reaseguros.

Existen autorizaciones para operar en el ramo de seguros de vida o en uno o más ramos de seguros generales o en ambos. Los seguros de hospitalización, cirugía y maternidad y de accidentes personales se consideran seguros generales (Castelo, 1988).

2.9.1 ACTIVIDAD FINANCIERA

La actividad aseguradora es uno de los tres pilares de los mercados financieros, junto con el mercado de crédito o bancario y los mercados de valores o de instrumentos financieros. Su importancia estratégica, social y económica, lleva a que estén sometidas a estricta supervisión administrativa con reglas propias de funcionamiento, control e inspección.

Las empresas de seguros por su función mediadora en el sistema financiero son unos intermediarios financieros con unas características especiales que las diferencian de las empresas de otros sectores de la economía e incluso con las restantes empresas financieras.

2.9.2 CARACTERÍSTICAS DE LAS EMPRESAS DE SEGUROS

Las entidades aseguradoras, para poder afrontar los riesgos derivados de su actividad deben disponer de los recursos financieros suficientes y en consecuencia la legislación les impone determinadas restricciones (Carmen, 2015).

Dada la conveniencia de que exista permanencia y estabilidad en este sector, las normas legales suelen prohibir que esta actividad pueda ser desarrollada por personas naturales.

Para garantizar la solvencia de las empresas aseguradoras, la legislación rechaza que estas empresas puedan ejercer algún tipo de actividad distinta de la aseguradora.

El ejercicio de una actividad de intermediación financiera que tiene que inspirar la máxima confianza entre los asegurados e inversores conlleva que estas entidades estén sometidas a la tutela del Estado que las somete a control, tanto para el inicio de su actividad como del desarrollo.

2.9.2.1 JURÍDICO

La actividad aseguradora debe buscar las formas contractuales para regular esa actividad y la ley de Contratos de Seguros y la ley de Empresas de Seguros y Reaseguros.

2.9.2.2 ADMINISTRATIVO

El Estado regula esta actividad en beneficio y protección del asegurado que es el débil jurídico en la relación contractual.

2.10 SUPERINTENDECIA DE BANCOS Y SEGUROS DEL ECUADOR

La Superintendencia de Bancos y Seguros es un organismo técnico, con autonomía administrativa, económica y financiera, cuyo objetivo principal es vigilar y controlar con transparencia y eficacia a las instituciones del sistema financiero, de seguro privado y de seguridad social, a fin de que las

actividades económicas y los servicios que prestan se sujeten a la ley y atiendan al interés general. Asimismo, busca contribuir a la profundización del mercado a través del acceso de los usuarios a los servicios financieros, como aporte al desarrollo económico y social del país (Serrano, 2015).

2.10.1 MISIÓN

“Velar por la seguridad, estabilidad, transparencia y solidez de los sistemas financiero, de seguros privados y de seguridad social, mediante un eficiente y eficaz proceso de regulación y supervisión para proteger los intereses del público y contribuir al fortalecimiento del sistema económico social, solidario y sostenible” (Serrano, 2015).

2.10.2 VISIÓN

“Ser una Institución técnica de regulación y supervisión de alta productividad, prestigio y credibilidad para satisfacer con calidad los servicios que presta a los actores externos e internos, con recursos humanos competentes y tecnología de punta” (Serrano, 2015).

2.11 BALANCES

El balance es una imagen de la empresa en un momento determinado. Incluye los activos y pasivos, proporcionando información sobre el patrimonio neto de la empresa. En otras palabras un balance es un resumen de todo lo que tiene la empresa, de lo que debe, lo que le deben y de lo que realmente le pertenece a su propietario, a una fecha determinada.

Al elaborar el balance general el empresario obtiene la información valiosa sobre su negocio, como el estado de sus deudas, lo que debe cobrar o la

disponibilidad de dinero en el momento o en un futuro próximo. El balance general consta de dos partes, activo y pasivo. El activo muestra los elementos patrimoniales de la empresa, mientras que el pasivo detalla su origen financiero. La legislación exige que este documento sea imagen fiel del estado patrimonial de la empresa.

2.11.1 ACTIVOS

Los activos pueden definirse como el conjunto de bienes y derechos reales y personales sobre los que se tiene propiedad, así como cualquier costo o gasto incurrido con anterioridad a la fecha del balance, que debe ser aplicado a ingresos futuros. En otras palabras, los activos son todos los bienes que tiene la empresa y posee valor tales como:

- El dinero en caja y en bancos.
- Las cuentas por cobrar a los clientes
- Las materias primas en existencia o almacén
- Las máquinas y equipos
- Los vehículos
- Los muebles y enseres
- Las construcciones y terrenos

Los activos de una empresa se pueden clasificar en orden de liquidez en las siguientes categorías: Activos corrientes, Activos fijos y otros Activos.
Por otro lado y según el autor, existen otras formas de clasificar a los activos. Una de ellas lo clasifica en tres grupos principales: Circulantes, Fijos y Cargos diferidos. Otros reconoce dos grupos: Los Activos Circulantes y los No Circulantes. La base fundamental para hacer la distinción entre circulante y no circulante es primariamente el propósito con que se efectúa la inversión, es decir si es permanente o no (Fierro, Contabilidad de Activos, 2007).

2.11.2 PASIVOS

Es todo lo que la empresa debe. Los pasivos de una empresa se pueden clasificar en orden de exigibilidad en las siguientes categorías (Fierro, Contabilidad de Pasivos, 2009).

- Pasivos corrientes o a corto plazo.
- Pasivos a largo plazo
- Otros pasivos.

2.11.3 PATRIMONIO

En el lenguaje contable el patrimonio, puede definirse como el conjunto de bienes, derechos y obligaciones que posee una unidad económica en una fecha determinada, y que constituye precisamente el objeto material de estudio de la contabilidad. Es el valor de lo que le pertenece al empresario en la fecha de realización del balance. Está conformado por el Capital y Utilidades Acumuladas (Fierro, Contabilidad de Activos, 2007).

2.11.4 TIPOS DE BALANCES GENERALES

Los balances generales se formulan de acuerdo con un formato y un criterio estándar para que la información básica de la empresa pueda obtenerse uniformemente como por ejemplo: posición financiera, capacidad de lucro y fuentes de fondeo. Según estas características los balances se pueden clasificar en:

2.11.4.1.1 Balance General Comparativo

Estado financiero en el que se comparan los diferentes elementos que lo integran en relación con uno o más periodos, con el objeto de mostrar los

cambios ocurridos en la posición financiera de una empresa y facilitar su análisis (Guzman, 2006).

2.11.4.1.2 Balance General Consolidado

Es aquél que muestra la situación financiera y resultados de operación de una entidad compuesta por la compañía tenedora y sus subsidiarias, como si todas constituyeran una sola unidad económica.
Se formula sustituyendo la inversión de la tenedora en acciones de compañías subsidiarias, con los activos y pasivos de éstas, eliminando los saldos y operaciones efectuadas entre las distintas compañías, así como las utilidades no realizadas por la entidad (Guzman, 2006).

2.11.4.1.3 Balance General Estimativo

Es un estado financiero preparado con datos preliminares, que usualmente son sujetos de rectificación (Guzman, 2006).

2.11.4.1.4 Balance General Proforma

Estado contable que muestra cantidades tentativas, preparado con el fin de mostrar una propuesta o una situación financiera futura probable (Guzman, 2006).

2.11.4.1.5 Balance Operacional Financiero del Sector Público

Estado que muestra las operaciones financieras de ingresos, egresos y déficit de las dependencias y entidades del Sector Público Federal deducidas de las operaciones compensadas realizadas entre ellas. La diferencia entre gastos e ingresos totales genera el déficit o superávit económico (Calvache, 1995).

2.11.4.1.6 Balance Presupuestario

Saldo que resulta de comparar los ingresos y egresos del Gobierno Federal más los de las entidades paraestatales de control presupuestario directo (Calvache, 1995).

2.11.4.1.7 Balance Primario del Sector Público

El balance primario es igual a la diferencia entre los ingresos totales del Sector Público y sus gastos totales, excluyendo los intereses. Debido a que la mayor parte del pago de intereses de un ejercicio fiscal está determinado por la acumulación de deuda de ejercicios anteriores, el balance primario mide el esfuerzo realizado en el periodo corriente para ajustar las finanzas públicas (Guzman, 2006).

METODOLOGÍA

3 METODOLOGÍA

La metodología aplicada al proyecto es la de Ralph Kimball, debido a que se adapta a los requerimientos del proyecto y las etapas que a continuación se detallan, las mismas que son claras y se pueden identificar de manera sencilla.

3.1.1 ETAPAS DE LA METODOLOGÍA DE KIMBALL

Se detallan las diferentes etapas que se aplican con la metodología, estas etapas recopilan análisis, diseño y estructura para la construcción de un Cubo Analítico.

3.1.1.1 PLANEAMIENTO DEL PROYECTO

En esta etapa inicial se busca identificar el escenario del proyecto para determinar el alcance, definir el proyecto, incluyendo justificaciones del negocio, generando la información suficiente para poder dar seguimiento al progreso del proyecto.

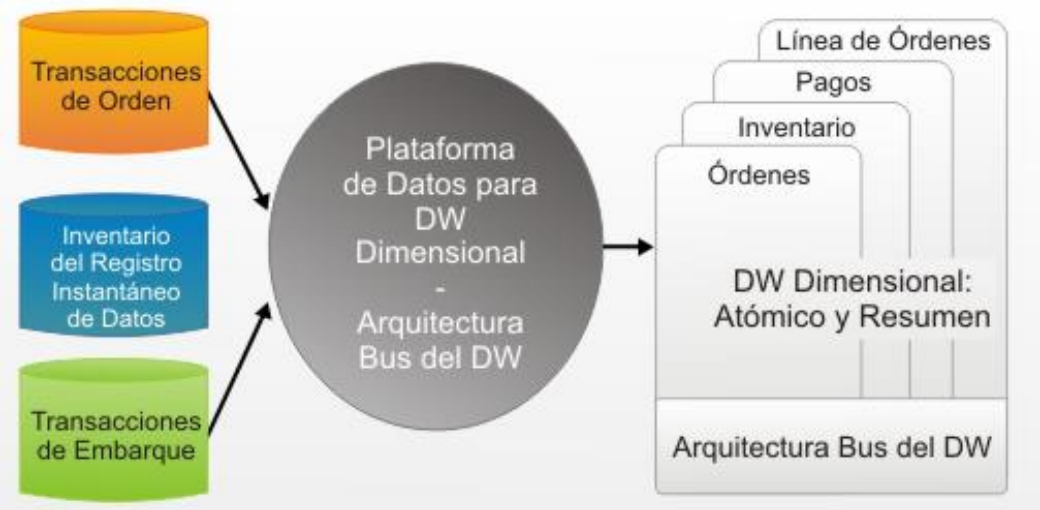

Figura 15. Metodología Kimball

3.1.1.2 ANÁLISIS DE REQUERIMIENTOS

Es un factor determinante para el éxito del proyecto, ya que se necesita de la interpretación correcta de los diferentes niveles de requerimientos expresados por los diferentes usuarios y establece la base de las tres etapas paralelas siguientes.

3.1.1.3 SELECCIÓN E INSTALACIÓN DE PRODUCTOS

En base al diseño de la arquitectura se evalúa y selecciona los componentes específicos de la arquitectura, como la plataforma, el motor de base de datos, herramientas ETL y herramientas de acceso.

3.1.1.4 MODELAMIENTO DIMENSIONAL

El diseño del modelo dimensional busca presentar los datos de una forma intuitiva y que proporcione acceso de alto desempeño.

3.1.1.5 DISEÑO FÍSICO

Esta etapa se focaliza sobre la selección de las estructuras que soporta el diseño lógico, que incluye los nombres de columnas, tipos de datos, declaraciones de claves.

3.1.1.6 DISEÑO Y DESARROLLO DE LA PRESENTACIÓN DE LOS DATOS

En esta etapa se ejecutan los procesos de extracción, transformación y carga (ETL), este proceso comprende varios aspectos que son determinantes en el

proyecto de inteligencia de negocios, por lo que para su desarrollo se debe seguir un plan para su correcto desarrollo.

3.1.1.7 ESPECIFICACIÓN DE APLICACIONES PARA LOS USUARIOS FINALES

En esta etapa se centra más en el front room, ya que se proporcionará la interfaz que se mostrará al usuario.

Una aplicación de usuario final, provee un diseño y estructura a los reportes, tomando como base los datos de la bodega de datos.

3.1.1.8 MANTENIMIENTO Y CRECIMIENTO

Cuando se desarrolla un proyecto Data Warehouse se debe pensar en el mantenimiento posterior, pues estas aplicaciones tienden a crecer a medida que crecen los datos de la organización.

ANÁLISIS Y DISCUSIÓN DE RESULTADOS

4 ANÁLISIS Y DISCUSIÓN DE RESULTADOS

Las etapas para el desarrollo y la construcción de los Data mart, cubos OLAP y reportes, se sigue la metodología de Ralph Kimball, dado que establece claros procesos para todo el ciclo del desarrollo del proyecto y garantiza la calidad y eficiencia de la solución de inteligencia de negocios.

En las siguientes secciones se describen los procesos realizados para cada fase del proyecto que garantizan su calidad y cumplimiento.

4.1 PLANEAMIENTO DEL PROYECTO

En esta sección se definen los objetivos, así como la justificación del proyecto y su alcance.

4.1.1 OBJETIVO DEL PROYECTO

Desarrollar Cubos Analíticos de BI con la Herramienta Pentaho, necesarios para el análisis de la información que presenta la “Superintendencia de Bancos del Ecuador” para las Cuentas de Balances de las Compañías de Seguros del Ecuador.

4.1.2 DEFINICIÓN DEL PROYECTO

Para el proyecto desarrollado se ha identificado un alto interés por parte de los Analistas financieros de las diferentes Compañías de Seguros, para el éxito de su implementación.

La demanda del proyecto se da debido a la necesidad de obtener mejor información del sistema financiero de las Compañías de Seguros del Ecuador

en Conjunto, para tomar mejores decisiones a nivel gerencial y así mejorar su competitividad y rendimiento. Esta demanda se satisface con un sistema de BI.

Dicha Solución BI está compuesta por un Data Marts, desarrollado por medio de tablas , en donde se encuentran ya no los simples datos operacionales de la base de datos sino la información útil lista para ser reporteada o como en nuestro caso información lista para el respectivo análisis multidimensional por medio de Cubos Analíticos OLAP.

4.1.3 ALCANCE DEL PROYECTO

El presente proyecto de tesis tiene todos los componentes de una solución BI con análisis multidimensional en la herramienta Pentaho.

Como fuente de datos principal se encuentran los archivos de Excel comprimidos (rar) por cada Compañía de seguros para cada mes del año, estos archivos son publicados por la Superintendencia de Bancos del Ecuador.

A partir de esto se define mediante la implementación de procesos ETL cada uno de los Data mart que se utilizan para el análisis multidimensional en Cubos Analíticos OLAP y su posterior explotación. El proceso ETL está realizado mediante la herramienta Pentaho Data Integration (Spoon) el cual carga la información útil en una base de datos MySql la cual contiene las dimensiones y tablas de hechos.

Posteriormente haciendo uso de la herramienta Pentaho Schema Workbench se realizará la creación de los Cubos Analíticos multidimensionales para su posterior explotación a través del servidor OLAP.

Todas estas herramientas soportan varios sistemas operativos, incluyendo Microsoft Windows y GNU/Linux lo que permite una gran portabilidad y la posibilidad de realizar el análisis OLAP vía web, gracias a que el análisis OLAP se realiza utilizando un navegador de Internet.

4.1.4 JUSTIFICACIÓN DEL PROYECTO EN EL NEGOCIO

El presente proyecto busca mejorar la productividad y rendimiento en cuanto al reporteo de información se refiere en las diferentes Compañías de Seguros que lo requieran.

a) Solución BI económicas

Ya que con la realización del denominado análisis multidimensional con la herramienta Pentaho, estamos dando una solución BI libre en cuanto a la capa BI se refiere.

b) Mejora de proceso de toma de decisiones

Es fundamental considerar este beneficio, puesto que los usuarios del sistema podrán obtener un mayor nivel en cuánto a tomar decisiones importantes, al tener reportes BI con información actualizada, clara y precisa cuyos reportes además ofrece un alto nivel estadístico y de interacción con el usuario final.

4.2 ANALISIS DE REQUERIMIENTOS

Para el levantamiento de requerimientos se realizaron reuniones con los Analistas Financieros de diferentes Compañías de Seguros debido a que ellos son los principales consumidores de información comparativa entre sus competidores directos e indirectos. Además y como se mencionó en la justificación del proyecto de tesis la creación del repositorio se basa principalmente en realizar los reportes que se generan mensualmente.

A continuación se definen los requerimientos de los usuarios:

4.2.1 REQUERIMIENTOS PARA GENERAR REPORTES

Uno de los requerimientos primordiales consiste en que el usuario pueda seleccionar una gran cantidad de datos para su análisis.

La principal necesidad de los analistas financieros es generar reportes en un menor tiempo posible con las siguientes características:

- Los reportes deben ser flexibles, esto quiere decir que el usuario puede seleccionar varias cuentas y varias compañías en una determinada línea de tiempo.
- Generar reportes por periodos, estos periodos pueden ser totalizados ya sean en semestres o trimestres.
- Filtros para cada selección debe ser dinámica, para que a partir de un reporte se pueda generar un filtro para cada dimensión o campo seleccionado.
- Se puede aplicar operaciones matemáticas a los valores resultantes.

4.2.2 REQUERIMIENTOS GRÁFICOS

Es parte fundamental para el usuario poder generar información gráfica con las siguientes especificaciones:

- Generar un gráfico a partir de un reporte generado.
- Los gráficos se deben generar de forma sencilla y rápida.
- Flexibilidad en cuanto a selecciones realizadas, esto quiere decir que no el usuario pueda graficar cualquier campo seleccionado.
- Facilidad en el uso de la herramienta para la generación de gráficos.

4.3 MODELAMIENTO DIMENSIONAL

Para iniciar el Modelamiento dimensional se debe tener en cuenta el principal objetivo de cualquier data mart es el análisis de la información. Este análisis es realizado por medio de reportes, por lo tanto al modelar el data mart se debe tener como objetivo la información deseada en los reportes.

Además se tomará en cuenta que el Data Marts ha sido realizado por medio de la creación de procesos ETL y creación de procedimientos almacenados. En las siguientes secciones se detalla cada componente del modelo dimensional.

4.3.1 DATA MARTS

El modelo diseñado e implementado es de un Data Marts en modelo Estrella lo cual hace que las dimensiones del Data Mart estén atadas directamente a la tabla de hechos.

El Data Marts implementado en este proyecto como fuente archivos XLS, por esta razón y por la facilidad se utilizará procesos ETL para la extracción, transformación y carga.

Es importante mencionar que en cada una de las dimensiones y la tabla de hechos de nuestros Data Marts se agregó información adicional la misma que sirve para permitir al usuario final el poder necesario para personalizar sus reportes, de manera que pueda habilitar y/o ocultar en tiempo de ejecución las dimensiones y medidas adicionales en cada uno de los reportes estadísticos BI que se ha desarrollado.

4.3.2 DEFINICIÓN DE GRANULARIDAD

Se definió la granularidad de la tabla de hechos como las más bajas o granulares posibles. Por ejemplo se puede consultar a detalle una cuenta en un período específico es así que la granularidad para este ejemplo es el día que se deriva directamente desde año->semestre->trimestre->mes.

De esta forma será posible llegar al grado de detalle que se desee y consultar registros de manera específica, aunque este no sea el objetivo de un Data Marts. Las medidas, que son los campos de valor de la tabla de hechos son los valores con la granularidad establecida.

4.3.3 DIMENSIONES

Se definen las dimensiones que soportan los requerimientos definidos, cumpliendo con la granularidad para la tabla de hechos. Las siguientes secciones relacionan la tabla diseñada para la base de datos con su dimensión correspondiente.

a) Dimensión Tiempo

La dimensión tiempo es una de las más importantes debido a que de ella depende mucho la granularidad, definimos como granularidad el mes debido a que es el nivel más bajo de tiempo del cual se presenta la información, también se definen atributos de año, semestre, trimestre debido a que siempre es necesario obtener información preestablecida de estos períodos de tiempo.

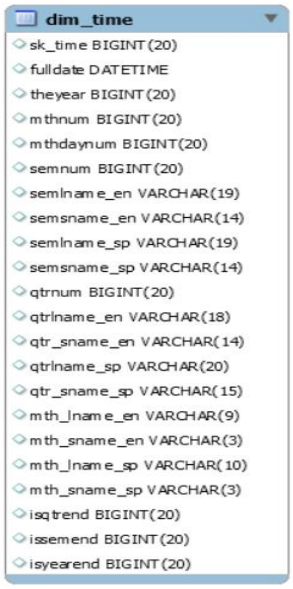

Figura 16. Dimensión Tiempo

b) Dimensión Entidad

Esta dimensión contiene la información básica de una Compañía de Seguros establecemos atributos que son relevantes para la dimensión como; nombres corto, nombre largo, tipo, estado, fecha de creación.

Figura 17. Dimensión Entidad

c) Dimensión Cuentas

Se definen los atributos que se utilizan para las cuentas, debido a que están pueden variar en el tiempo y se tienen diferentes niveles, también se asignan atributos propios como; nombre, fecha de creación, fecha de fin si lo tuviera en el caso de que la cuenta ya no aplica para fechas futuras.

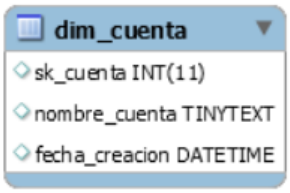

Figura 18. Dimensión Cuentas

4.3.4 TABLA DE HECHOS

A continuación se detalla la tabla de hechos al igual que se lo hizo para las dimensiones. Como solo tenemos una tabla de hechos llamada saldos_fact, esta contiene la información necesaria acerca de los valores que presenta cada compañía de seguros para una cuenta.

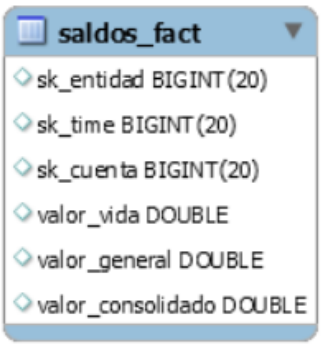

Figura 19. Tabla de Hechos Fact

4.3.5 DISEÑO DEL MODELO DIMENSIONAL

Luego de haber determinado las tablas que funcionarán como dimensiones y tabla de hechos, también de haber presentado una solución de Data Marts se presenta ahora el diseño multidimensional del Data Marts creado.

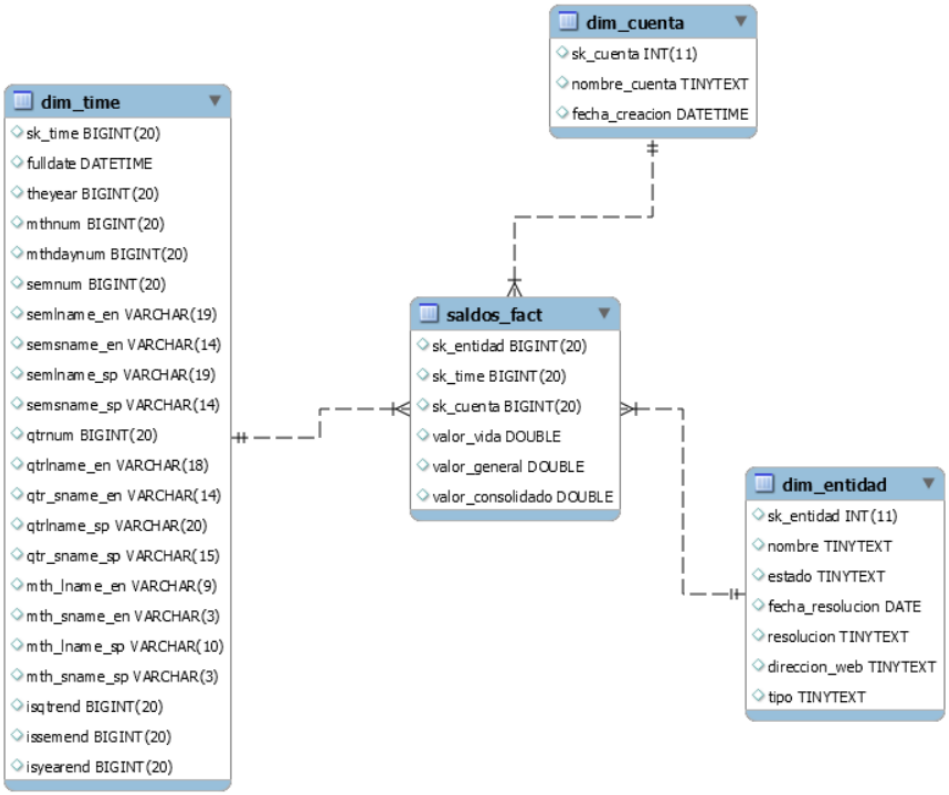

Figura 20. Modelo Dimensional

4.4 DISEÑO TÉCNICO DE LA ARQUITECTURA

La solución es un sistema de información que se conforma de varias tecnologías utilizadas para implementar la solución orientada al usuario final, con la capacidad de integrar los datos y transformarlos en información activa y productiva para la toma de decisiones.

Por lo expuesto, el sistema de información se enmarca en la categoría de un sistema del tipo BI apoyada en la Base de Datos MySql del cual se definen las estructuras del Data Marts para luego formar el Cubo Analítico con la herramienta Pentaho.

Figura 21. Diseño de la Arquitectura

4.4.1 OBTENCIÓN DATOS

La herramienta que se utilizará en todo el proceso de extracción de la data y su posterior carga a nuestro Data Marts es el Data Integration (PDI) en su versión actual.

En el siguiente gráfico se detallan las características de la herramienta PDI:

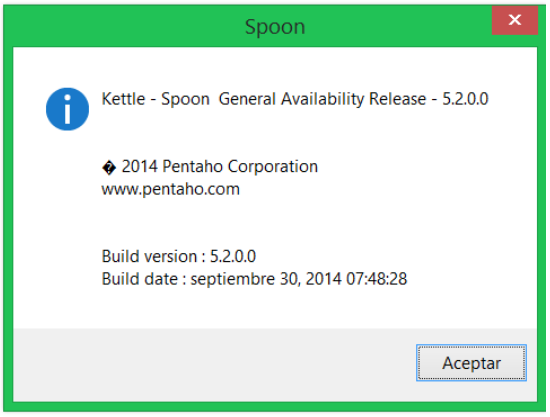

Figura 22. Características PDI

Como reseña general se definirán dos elementos que son importantes y necesarios en la construcción de Procesos ETL con la herramienta PDI, que a continuación se detallan:

- **Transformación**

  Es el elemento básico de diseño de los procesos ETL. Se compone de pasos o steps entrelazados entre si a través de los saltos o hops, de los cuales va fluyendo la información. Tenemos pasos para realizar múltiples actividades, como se mostrarán en detalle en el transcurso del proyecto.

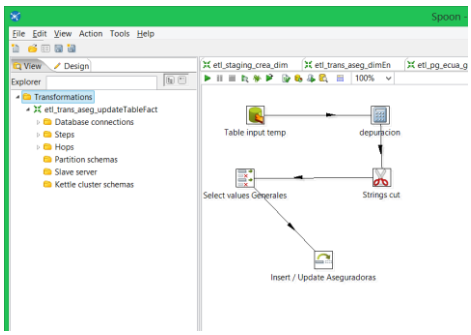

Figura 23. Ejemplo transformación

- Job

Es un conjunto complejo o sencillo de tareas para realizar una acción determinada. Igualmente se dispone de un conjunto de pasos (que son diferentes a los de las transformaciones) y los saltos (que en este caso determinan el orden de ejecución, y la gestión de resultados de la ejecución de cada paso). Dentro de los jobs se puede ejecutar una o varias transformaciones, los que nos permite ir dividiendo los procesos en partes y luego orquestar su ejecución mediante los jobs.

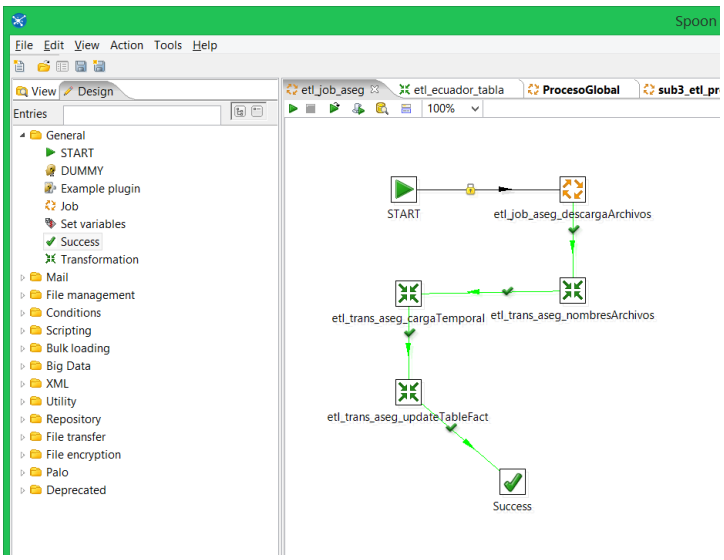

Figura 24. Ejemplo Job

Con la definición de estos elementos se continúa con los procesos, ya que era de vital importación profundizar en ellos.

Los datos se obtienen a partir de archivos XLS que publica la Superintendencia de Bancos del Ecuador. Debido a que los archivos se deben descargar por cada compañía resulta poco factible realizarlo de forma manual, por esta razón se crea un Proceso ETL que permite descargar los archivos por cada periodo de forma directa.

a) Se debe descargar el archivo de catastro

Las Compañías de Seguros se encuentran en la web de la Superintendencia de Bancos. Esto debido a que se necesita el catálogo general de las Compañías, para lo cual se debe descargar los archivos en el siguiente link:

http://www.sbs.gob.ec/practg/sbs_index?vp_art_id=&vp_tip=6&vp_buscr=/practg/pk_cons_bdd.p_bal_segr

Figura 25. Descarga archivo Catastro

b) Descarga de los Archivos de Balances

Los archivos de balances se seleccionan por cada Compañía debido a eso se realizó un proceso ETL que a continuación se detalla.

- El primer job en ejecutarse es el etl_job_aseg_descarga, este procesa una transformación que carga el período que se desea descargar y procesar durante todo el Proceso de Poblamiento Dimensional.

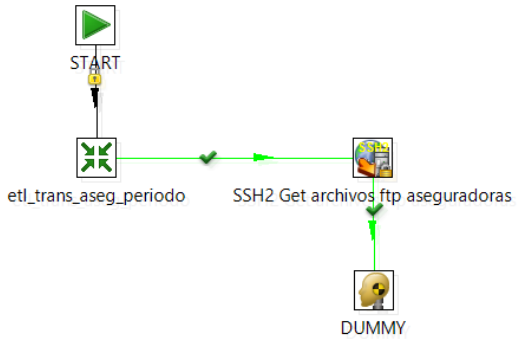

Figura 26.1 Descarga de Archivos

- La siguiente transformación es el etl_trans_aseg_periodo, obtiene el período que deseamos descargar la información, como se puede ver, asignamos el año en formato YY y el en YYYY, para el mes solo el formato es MM.

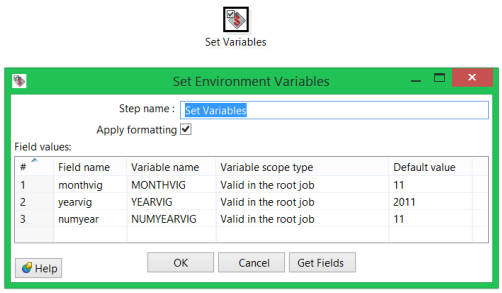

Figura 27.2 Definir fecha a ser descargada

- El paso SSH2 Get archivos ftp aseguradoras, especifica una dirección IP, usuario y contraseña que son necesarios para descarga de los archivos vía SHH, estos datos se obtienen directamente en la Superintendencia de Bancos y Seguro.

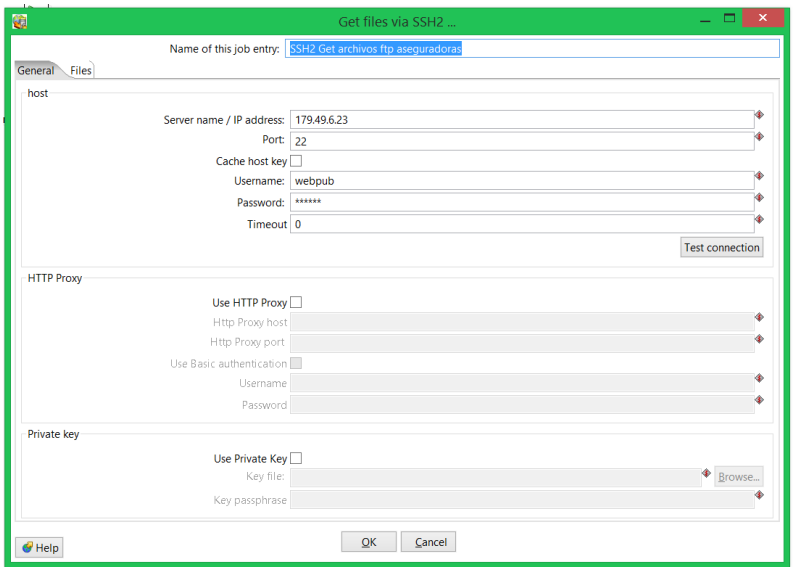

Figura 28. Parámetros para la descarga SHH

- Se especifica el directorio donde se guardar los archivos descargados, y el nombre que se les dará a las carpetas según el año y mes de descarga. El directorio es D:\PREYECTO\Archivos_Aseguradoras\${ANIOVIG}\${MESVIG}${ANIOVIG}, las variables corresponden año y la sub carpeta contendrá el mes y año en formato MMYY.

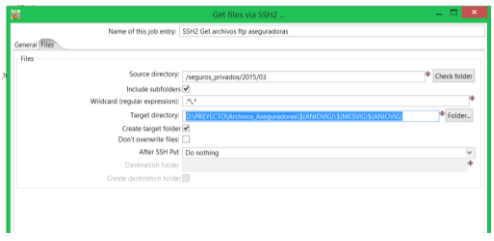

Figura 29. Directorio de descarga archivos

- La siguiente transformación detalla la iteración de los archivos descargados, también realizamos el truncamiento de la tabla saldos_aseg_temp, la cual se encuentra en la base de datos STAGIGN_AREA que sirve como una base de datos temporal para los archivos en crudo que se descargan.

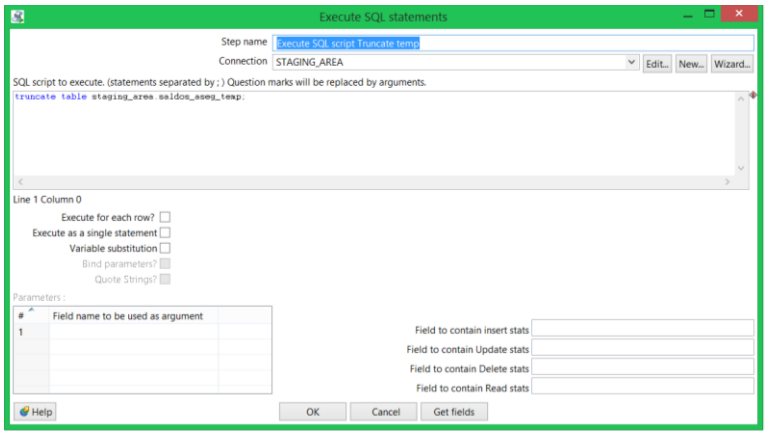

Figura 30. Truncamiento de la tabla saldos_aseg_temp

- El paso Get File Names Archivos, obtiene todos los archivos previamente descargados tomando del directorio y con las variables de año y mes que corresponden para saber la fecha que se están procesando los archivos.

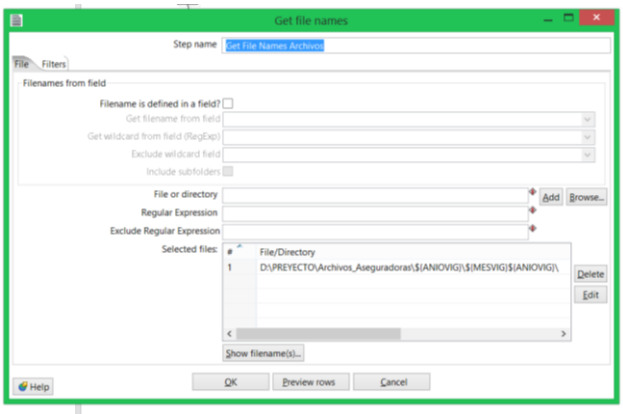

Figura 31. Traer Archivos descargados

- El paso Copy row to result, sirve para enviar en una serie de filas los nombre de los archivos obtenidos en el paso anterior, para que estos sean tomados por la transformación siguiente y se iteren de acuerdo al número de filas que se obtengan aquí.

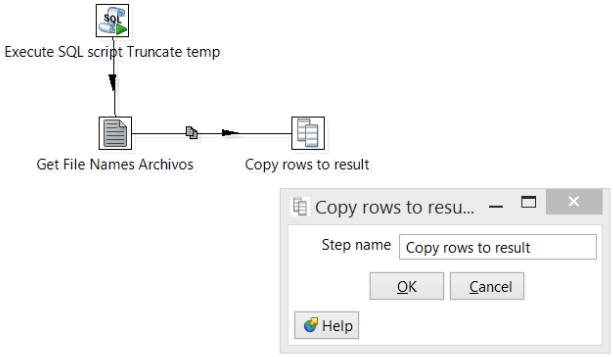

Figura 32. Nombres de archivos en filas.

- Llamada a la transformación etl_trans_aseg_cargaTemporal, aquí se especifica que esta transformación debe ejecutarse las veces que sean necesarias según el número de filas del paso anterior, que lo especificamos.

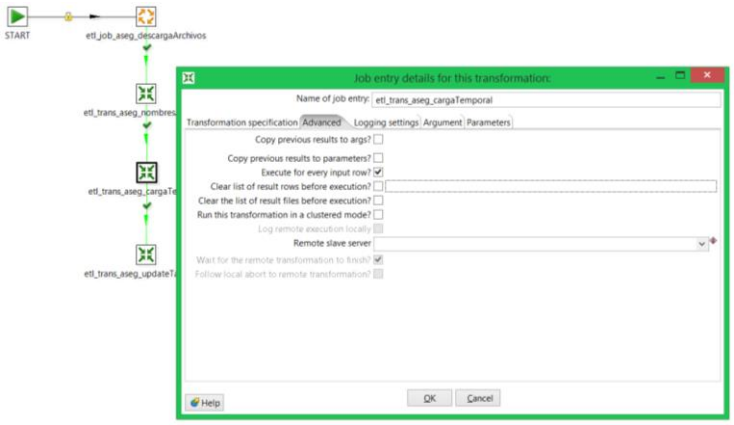

Figura 33. Opción para iterar filas de resultados

4.4.2 CARGA DE DATOS EN LA BASE DE DATOS TEMPORAL

Se generó una Base de Datos Temporal en la cual se almacenan los datos obtenidos a partir de los archivos descargados en los pasos anteriores.

Cabe recalcar que los datos que se almacenan en esta base de datos no son depurados ya que en esta Base de Datos solo debe contener y todos los campos que los archivos no retornen.

- En el siguiente paso, por cada nombre de archivo recibido en la transformación anterior, procedemos obtener las cabeceras y las cuentas. Luego se realizó un join para unir según corresponda las cuentas participantes y fecha con el valor de Generales, Vida y Consolidado. Para cargarlos en la base de datos temporal STAGING_AREA.

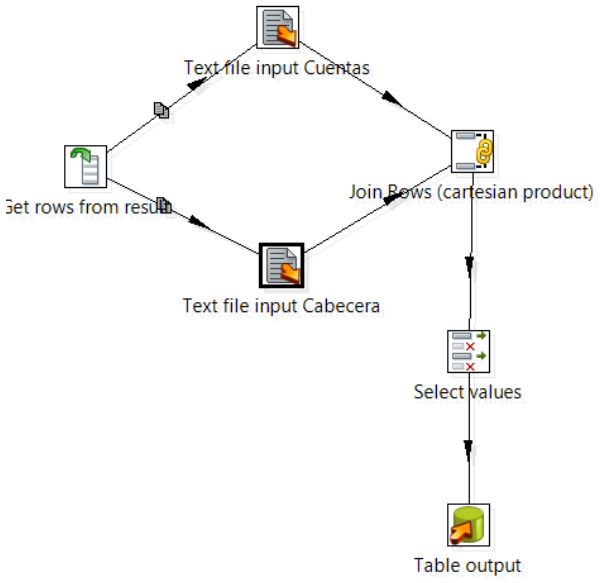

Figura 34. Transformación de carga temporal

- Para la obtención de cabeceras se procederá a la configuración de los campos necesarios.

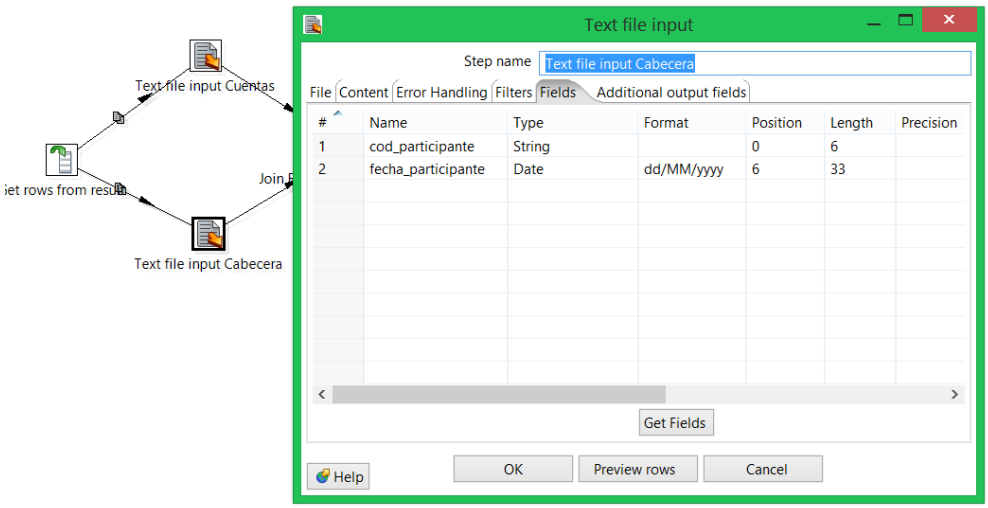

Figura 35. Campos para cabecera

- Para la obtención de cuentas se procede a configurar los campos necesarios.

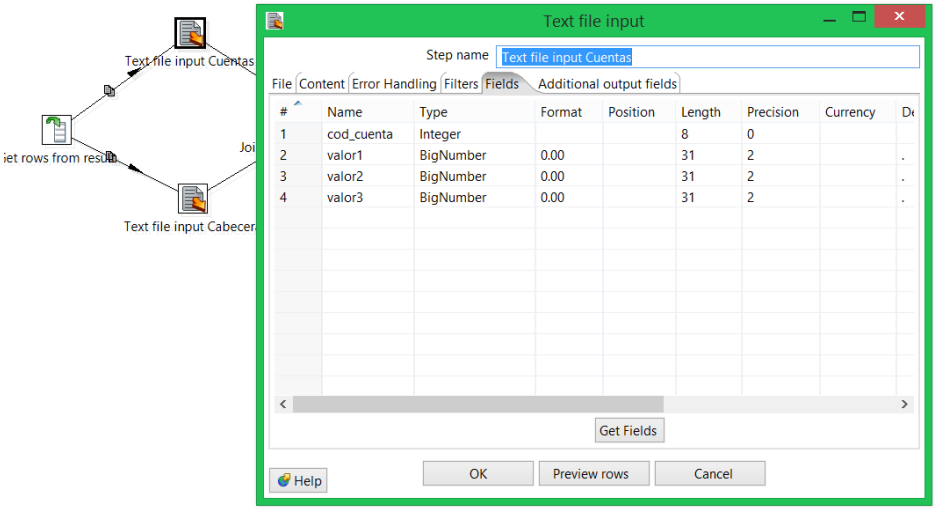

Figura 36. Campos para Valores de Cuentas

- Unión de los archivos cabeceras y cuentas según el nombre de archivos, debido a que estos se encuentran normalizados es fácil su unión sin tener problema alguno.

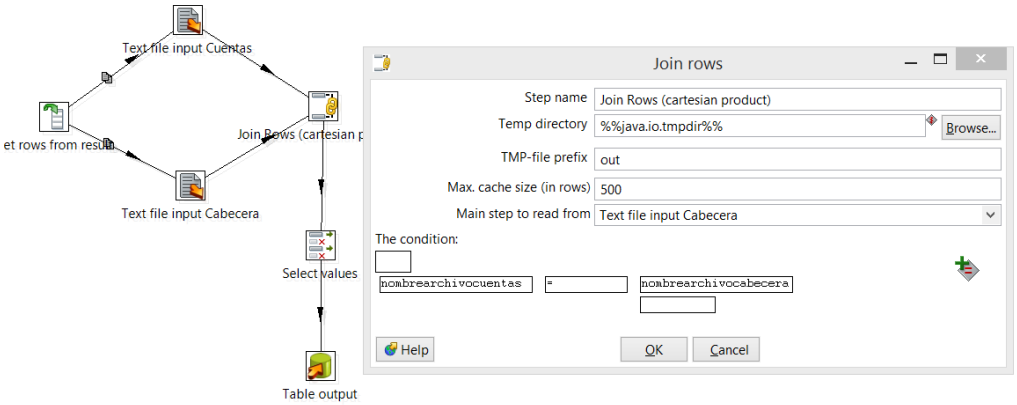

Figura 37. Unión cabecera y cuentas

- Se crea una conexión para la base de datos STAGING_ARE la cual nos servirá como un almacén temporal.

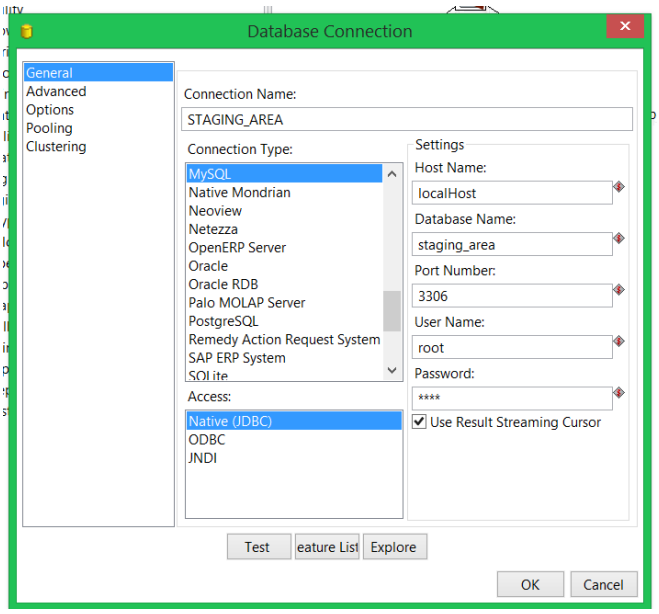

Figura 38. Conexión a base de datos staging_are

- Configuración de la tabla en la cual se insertarán los datos obtenidos del archivo que se itere.

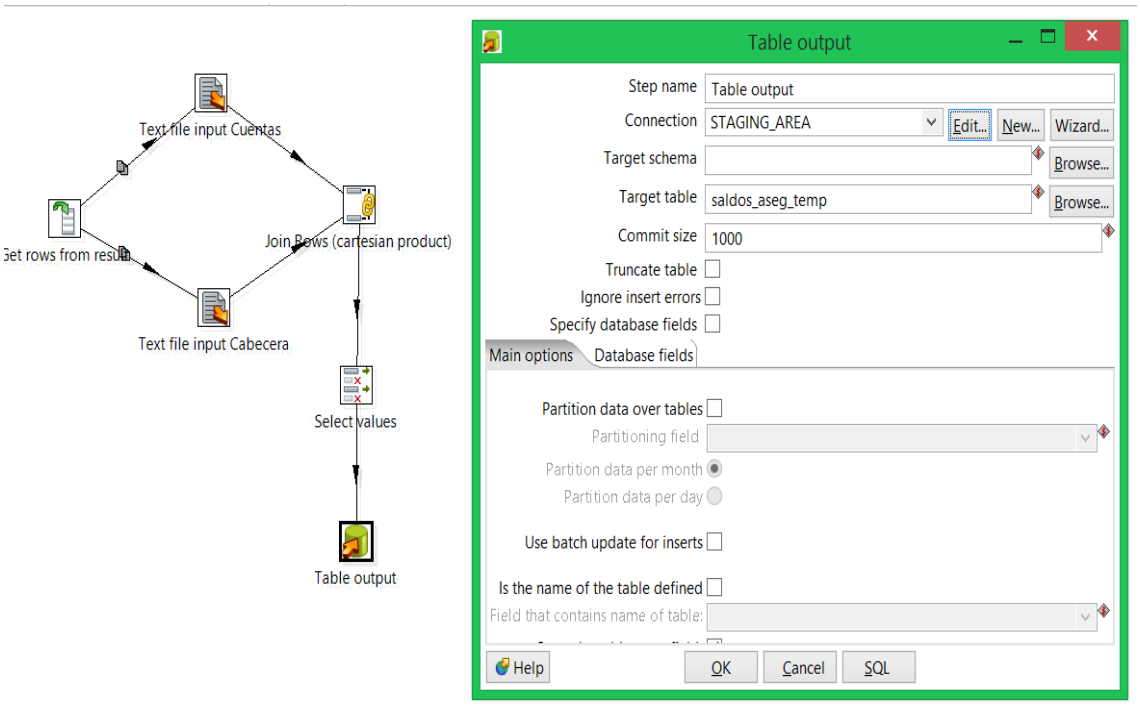

Figura 39. Conexión a la tabla saldos_aseg_temp

4.4.3 MAPEO DE LOS DATOS EN LOS MODELOS DIMENSIONALES

En esta sección se detallan los procesos que se realizaron para la población de las tablas de dimensiones, cabe recalcar que los procesos que a continuación se detallan realizan procesos de limpieza de datos que se especificarán en cada uno de ellos.

4.4.3.1 POBLACIÓN DE LA DIMENSIÓN TIEMPO

Los datos para la dimensión tiempo es genérica por tal razón se realiza un poblamiento de la misma a partir de un archivo plano llamado dim_time, esta dimensión es muy importante debido a que en esta se definió la granularidad.

Los campos que se obtienen del archivo pasan directamente a la tabla sin ningún tipo de depuración, debido a que en ella ya se encuentran estandarizados el formato de los mismos. Se detallan uno a uno los pasos y opciones que tiene cada uno.

A continuación se observa un gráfico de la transformación llamada etl_transf_aseg_dimTime.

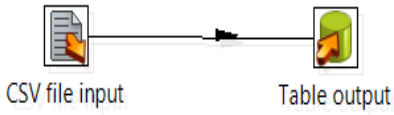

Figura 40. Población de dimensión Dim_Time

- Este paso especifica las rutas de acceso al archivo genérico para el poblamiento de la dimensión Tiempo, este paso es exclusivo para la entrada de archivos de tipo CSV, el cual define el formato del archivo.

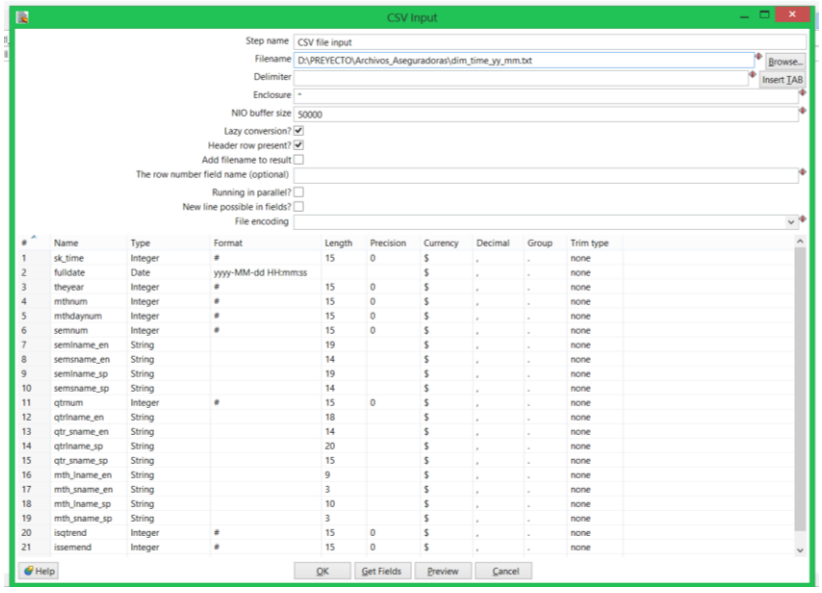

Figura 41. Archivo y campos necesarios para Dim_time

- Conexión a la tabla de Dim_time en la base de datos BALANCES_ASEG_ECUADOR se define en este paso, donde se especifica claramente la conexión seleccionada, la tabla que vamos a poblar y diferentes opciones que no es necesario especificar.

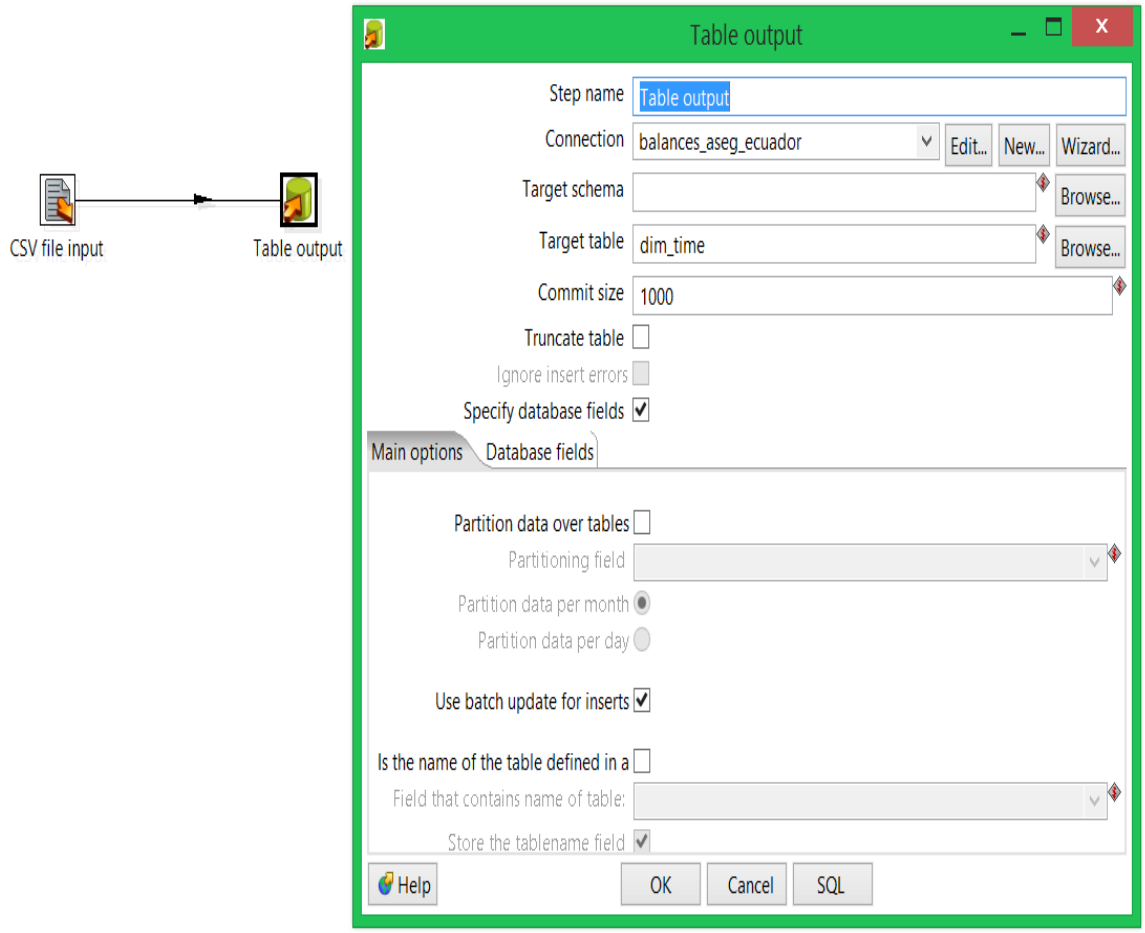

Figura 42. Conexión a la tabla dim_time.

4.4.3.2 TABLA TEMPORAL

Para consolidar los datos debemos poblar la tabla de saldos_aseg_temp esta tabla es una zona de aterrizaje primaria de los datos, para que a partir de ahí se puedan obtener la base para el poblamiento de las dimensiones y la tabla de hechos.

Se iteran por cada Aseguradora tomando el archivo que le corresponde y haciendo una serie de pasos para poblar la Base de Datos STAGING_SADOS.

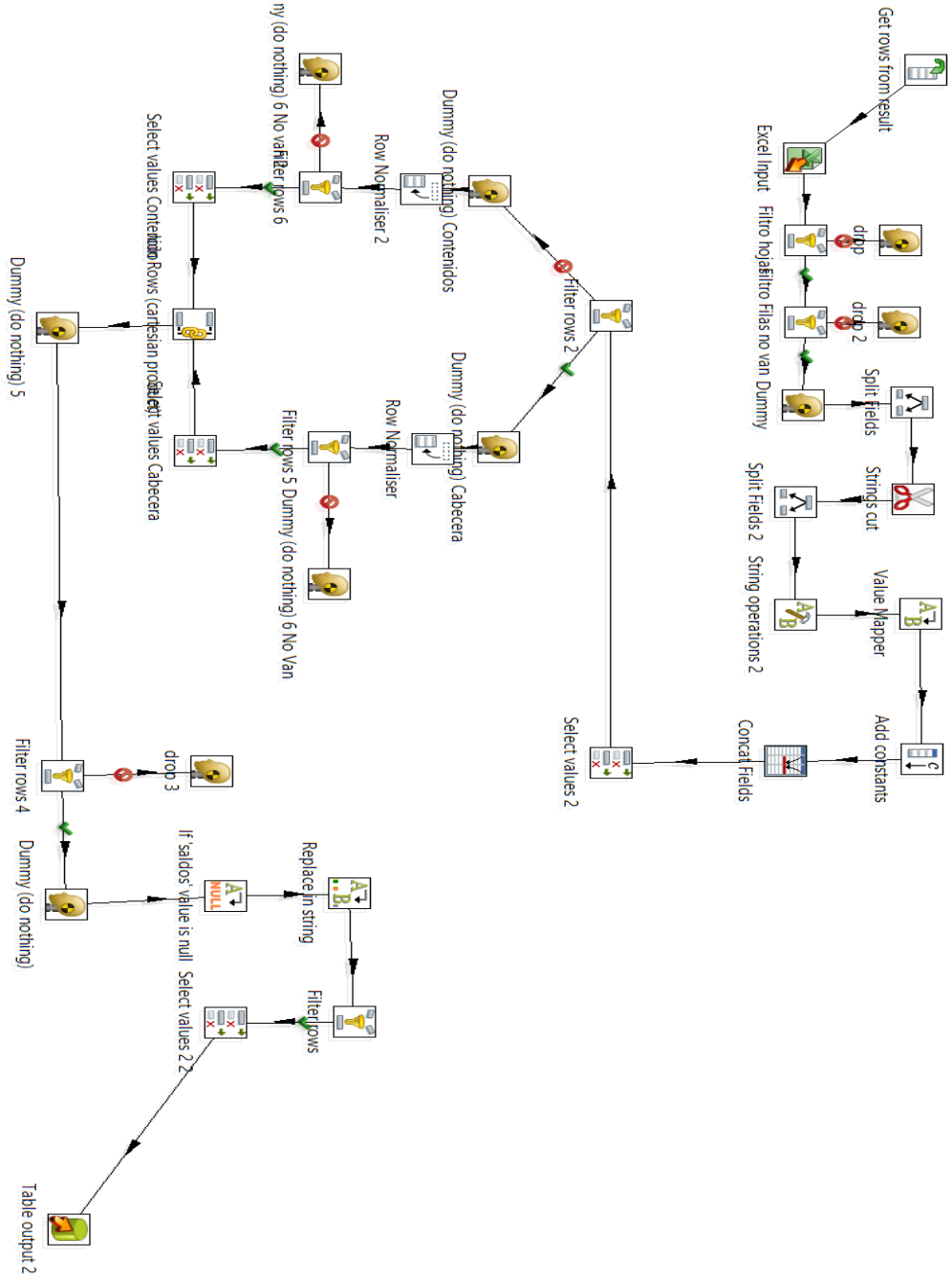

Figura 43. Población de datos temporales

4.4.3.2.1 Validación de Archivos

Para la validación de los archivos de Excel que se debe realizar, se hace una verificación del archivo de origen genérico, así como sus hojas y campos.

- Paso para tomar el archivo Excel en base al resultado de la transformación anterior que itera los diferentes archivos descargados.

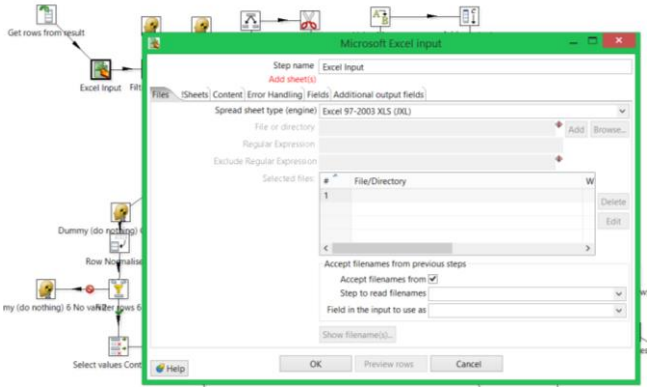

Figura 44. Traer archivo de paso anterior

- Filtro para la que solo permita el paso de las hojas con el nombre TOT o EPyG, que es la hoja que debe contener el archivo

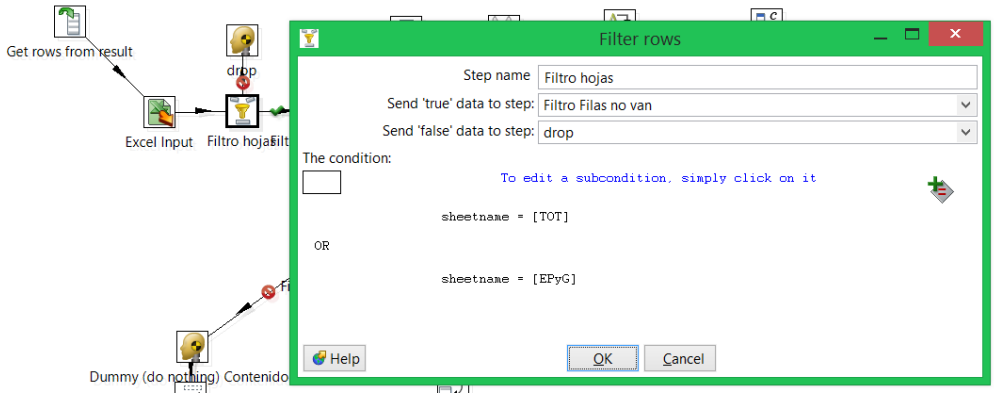

Figura 45. Filtros para hoja de archivo

- Filtramos filas que no contienen valores o que sus campos contengan null.

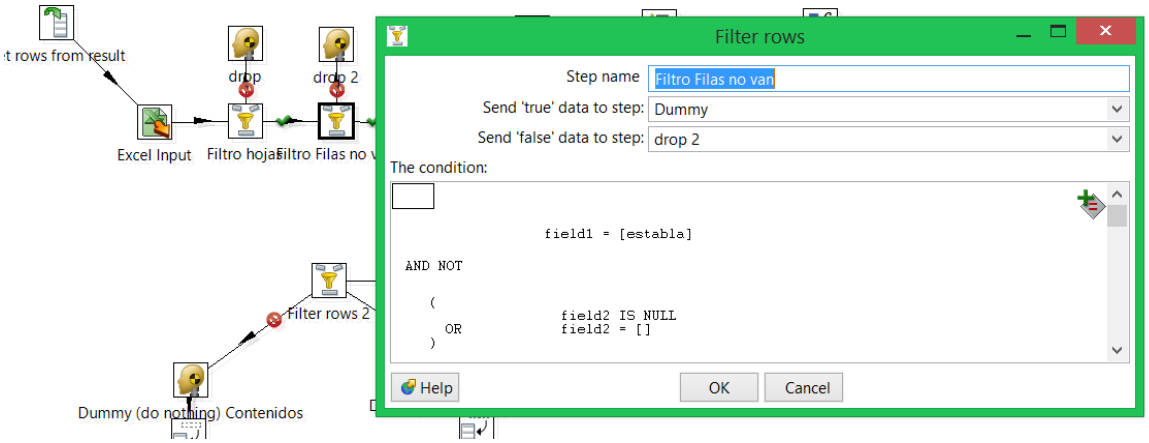

Figura 46. Filtro filas vacías

- Obtenemos solo los caracteres que se necesitan para el nombre del archivo que son del carácter 13 al 19.

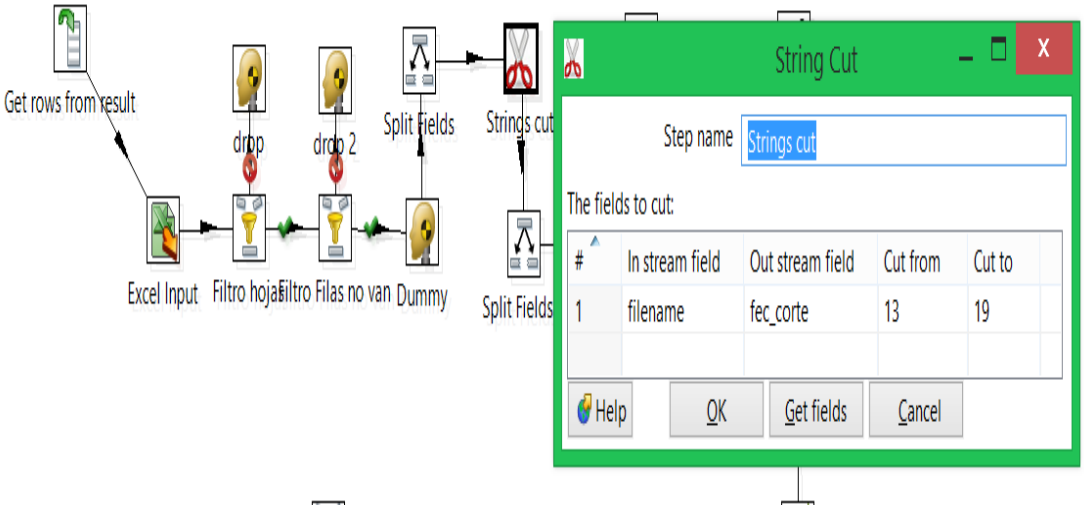

Figura 47. Caracteres del nombre del archivo

- Se mapea las fechas debido a que vienen en números y se desea la fecha en ISO 3.

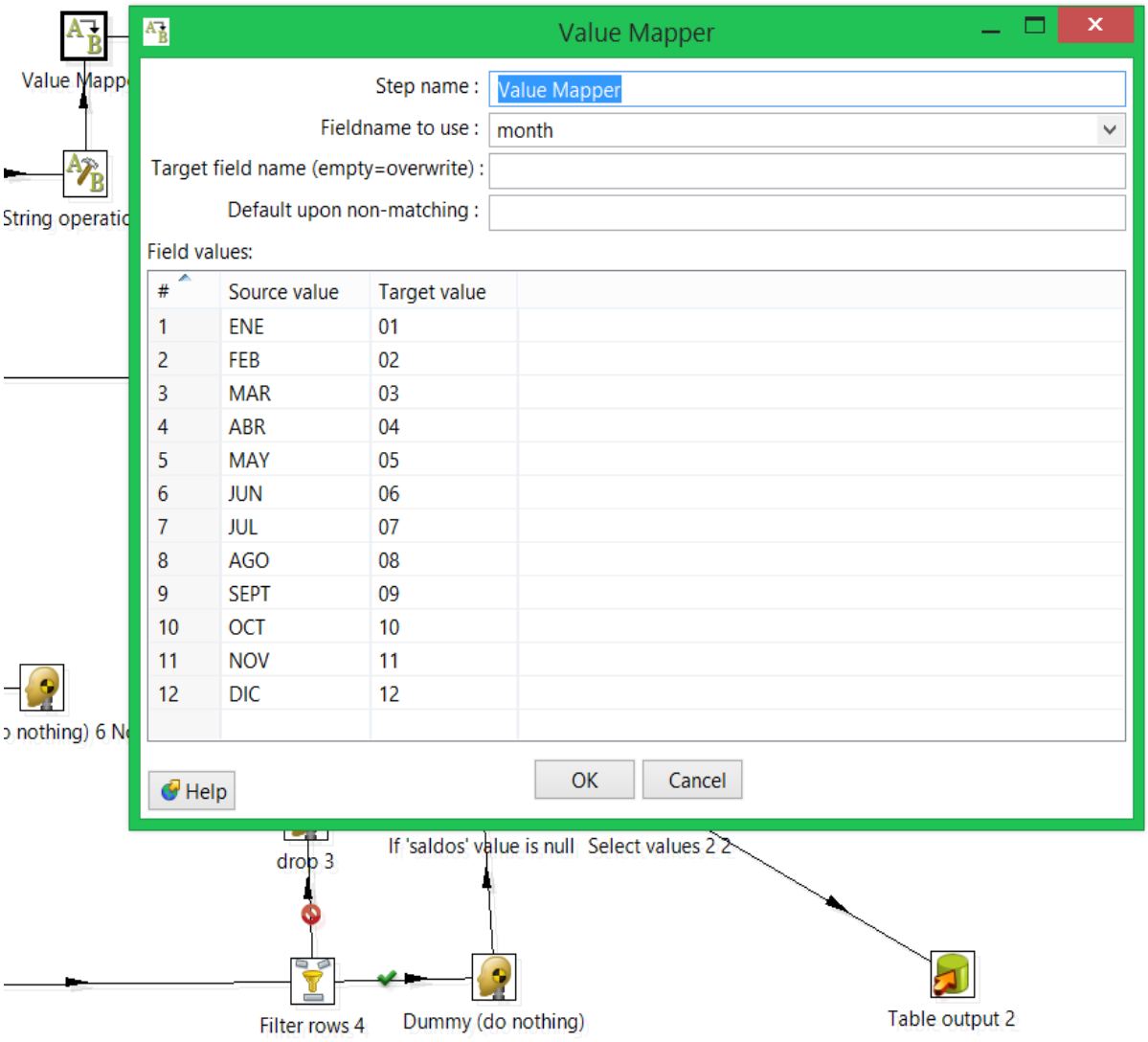

Figura 48. Mapeo de Fecha

4.4.3.2.2 Operadores especiales para campos

En estos pasos se va a realizar operaciones sobre los campos como limitar caracteres y validaciones de campos.

- Filtro para determinar que valores corresponden a la cabecera y cuales al contenido.

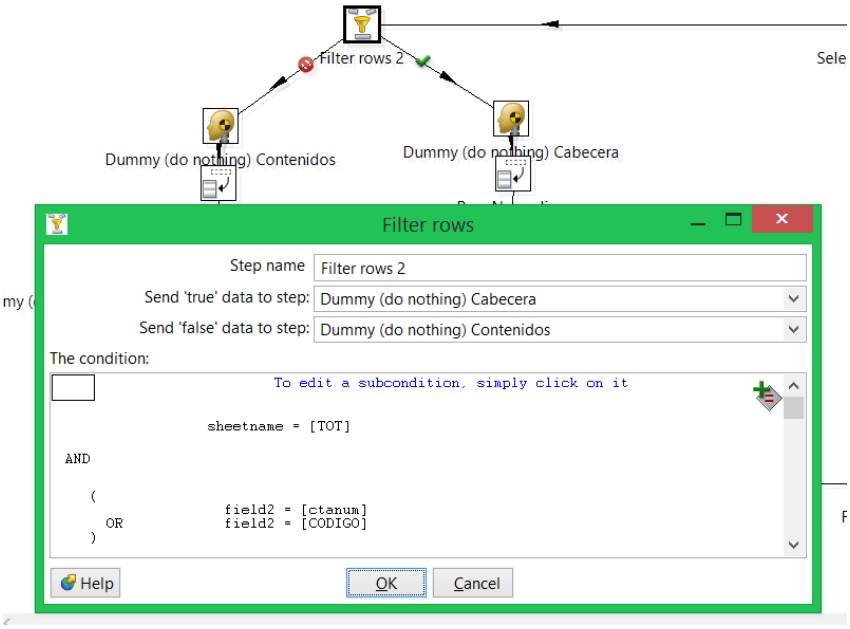

Figura 49. Filtro, Cabecera o Contenido

- Filtro para que en las cabeceras no haya valores null.

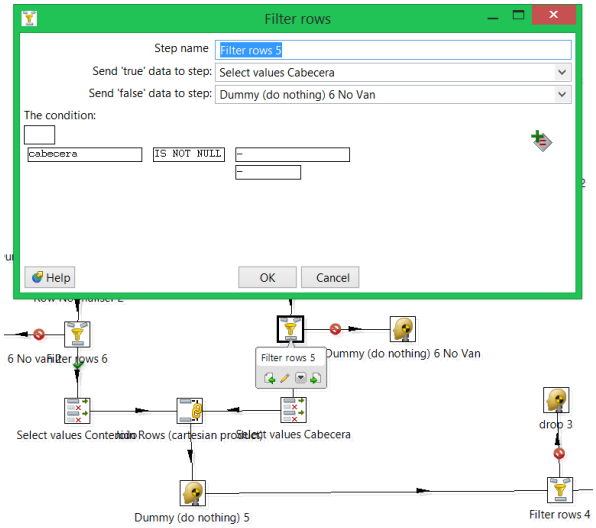

Figura 50. Filtro, Cabeceras sin null

- Join Rows que une las cabeceras con su correspondiente contenido, por los campos fieldCabecera y fieldCon

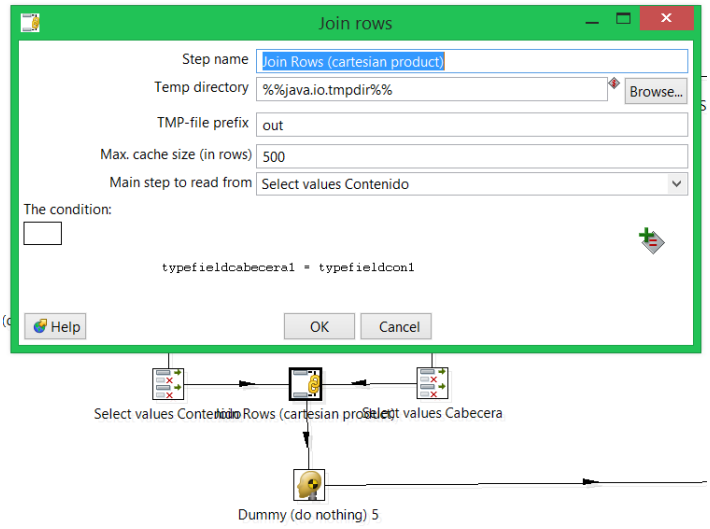

Figura 51. Join Row para unir cabecera con el contenido

- Reemplazo valores null por cero, si en el caso de que haya valores que no se hayan podido filtrar antes.

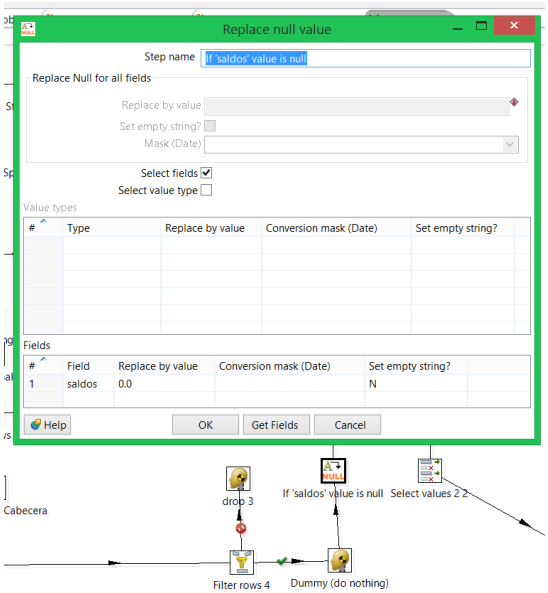

Figura 52. Reemplazo valores null por cero

4.4.3.3 POBLACIÓN DE LA DIMENSIÓN ENTIDAD

Cargamos la tabla de dimensiones Entidades, esto se realiza a partir del archivo Catastro descargado con anterioridad, también depuramos los nombres de las entidades para que no haya conflictos con los caracteres especiales y agregamos la fecha para mantener el histórico de la empresa en el caso de que haya cambiado de nombre.

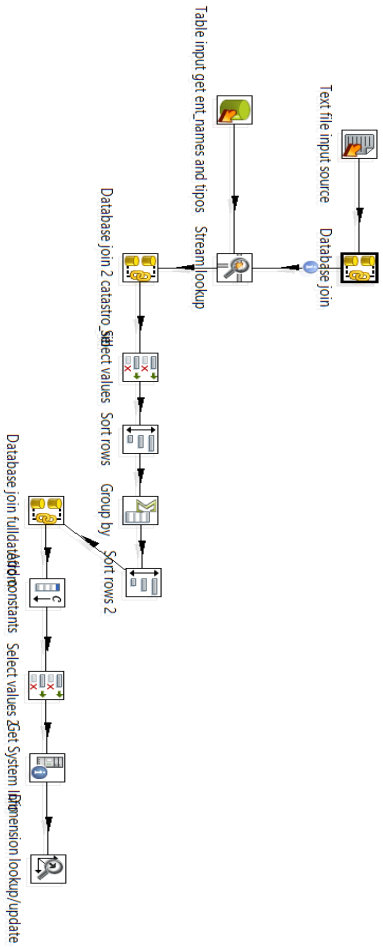

Figura 53. Población Dimensión Entidad

- Depuración de caracteres especiales para los nombres de las entidades que contengan algún tipo de carácter especial.

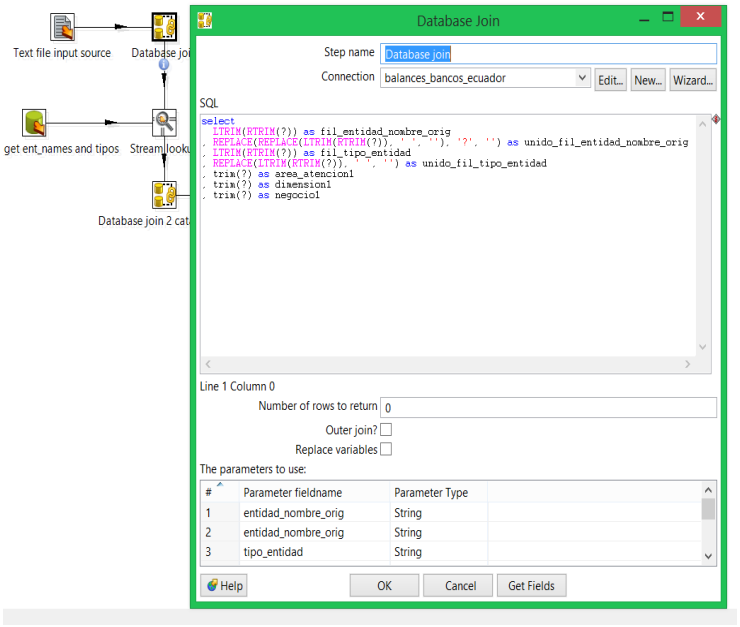

Figura 54. Depuración caracteres especiales

- Verificación de que las entidades que se están cargando se encuentren en la base de datos entidad esto sucede cuando la tabla de dimensiones ya se encuentra poblada.

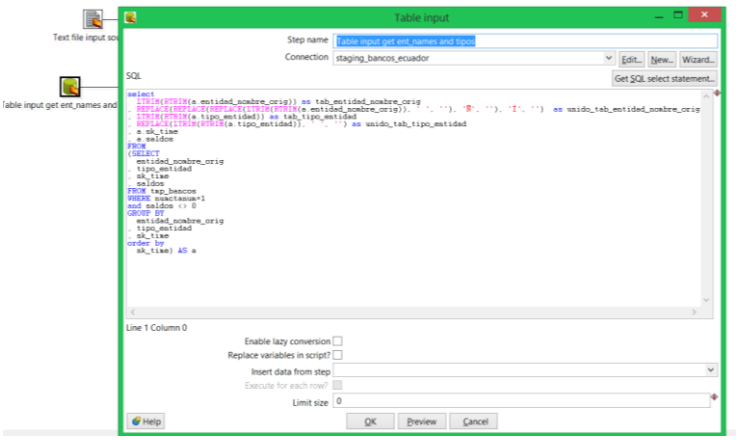

Figura 55. Validación entidades existentes

- Actualización dimensión en la tabla dim_entidad, con este paso se actualiza la tabla de una manera eficaz según su código único, si existiera una versión nueva la agregará.

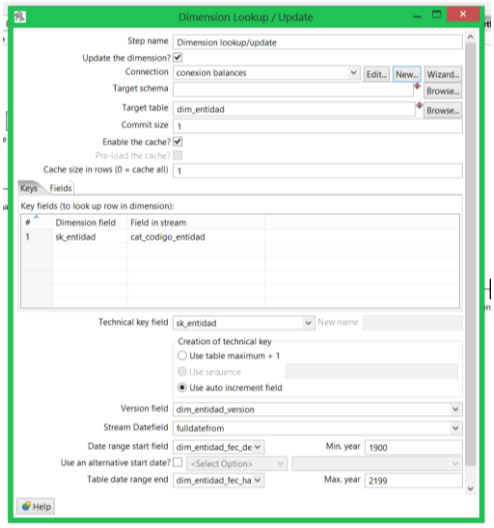

Figura 56. Actualización dimensión entidad

4.4.3.4 POBLACIÓN DE DIMESION CUENTAS

Se realiza la población de las cuentas que se obtienen a partir de la Base de Datos Pre-staging, de ahí se obtienen todos los datos y se realiza joins con las cuentas ya existentes.

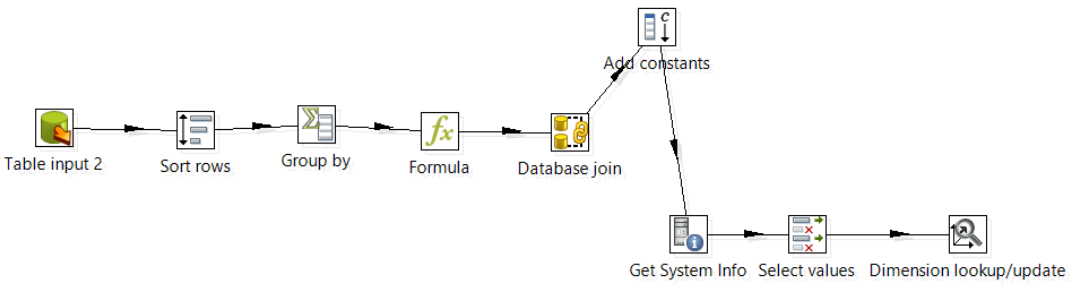

Figura 57. Población Dimensión Cuentas

- Fórmula para validar cuentas existentes que contengan el código que les corresponde.

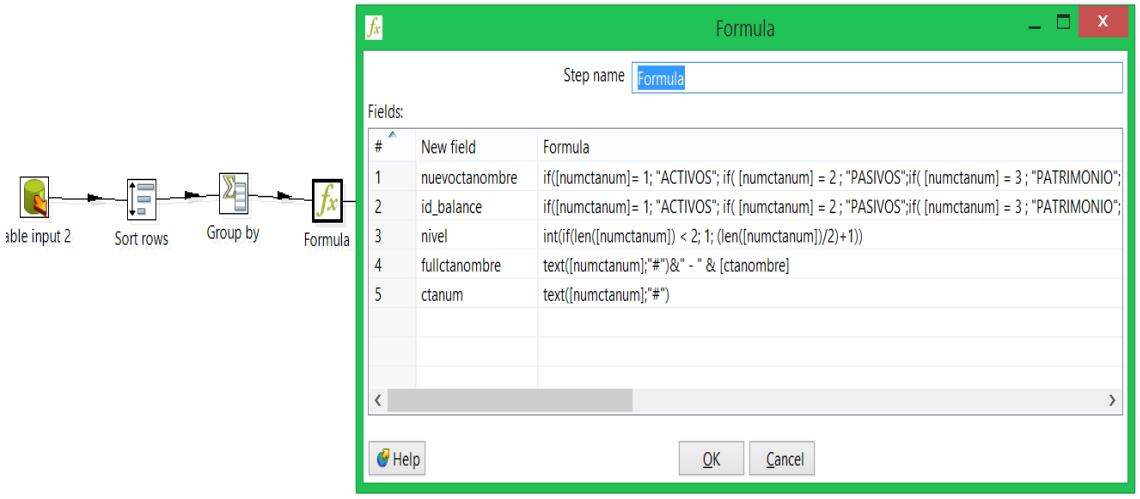

Figura 58. Validación código cuenta

- Validación de fecha sea DateTime, para que si alguna fecha de la cuenta este con un formato diferente.

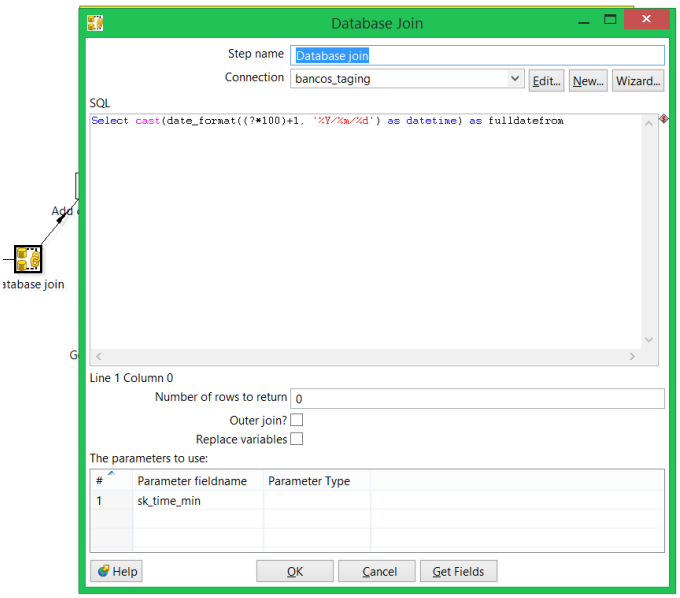

Figura 59. Validación fecha de cuenta

- Actualización dimensión en la tabla que dim_entidad, con este paso se actualiza la tabla de una manera eficaz según su código único, si existiera una versión nueva la agregara.

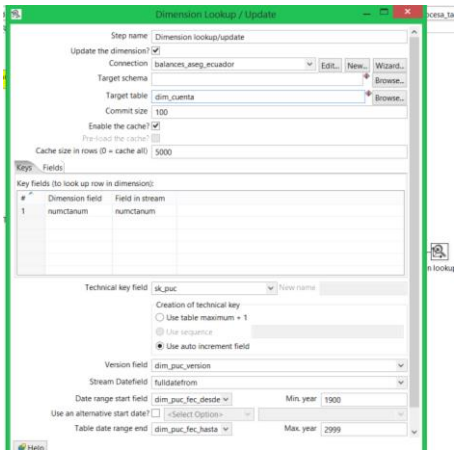

Figura 60. Actualización dimensión Cuenta

4.4.4 POBLACIÓN TABLA DE HECHOS

Como proceso final realizamos la población de la tabla de hechos la cual debe contener los valores de las cifras generales, vida y consolidado, las cuales corresponden al valor de las cuentas para cada Compañía de Seguros en un periodo.

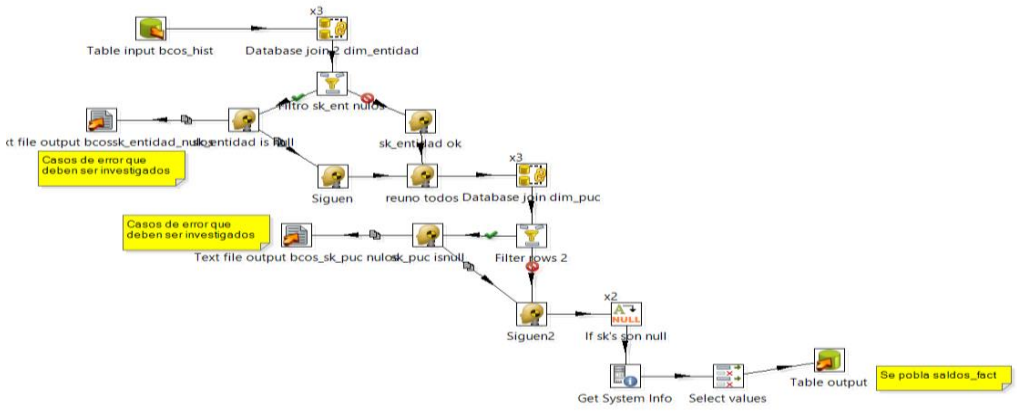

Figura 61. Población Tabla de Hechos

4.4.5 INFRAESTRUCTURA

Las bases de datos operacionales o también denominadas transaccionales o sistemas OLTP y el análisis multidimensional junto con los reportes web BI se ejecutarán en un equipo que hace de servidor de base de datos mysql y servidor de análisis respectivamente que se ejecutan en un solo equipo con las siguientes características:

- Windows 8.1
- Procesador Intel Core I7
- Memoria de 4GB en RAM

4.4.6 CREACIÓN DE CUBO ANALITICO

El análisis OLAP de la solución es realizado con la herramienta Schema WorkBench. Por lo tanto, los cubos que reflejan el diseño del modelo dimensional de los data marts serán construidos de acuerdo a los requerimientos de Mondrian para tal fin.

Se deben crear dos archivos para utilizar un cubo en Mondrian, en las siguientes secciones se describen las estructuras de estos archivos.
Luego de haber establecido las estructuras de estos archivos, un usuario puede navegar por las estructuras (dimensiones, atributos y medidas) de forma intuitiva e interactiva, utilizando únicamente el Mouse desde la interfaz gráfica ofrecida por Mondrian-Jpivot.

4.4.6.1 INSTALACIÓN HERRAMIENTA PENTAHO

- Primero se descarga el archivo .exe para la instalación de Pentaho business analytics 5.3

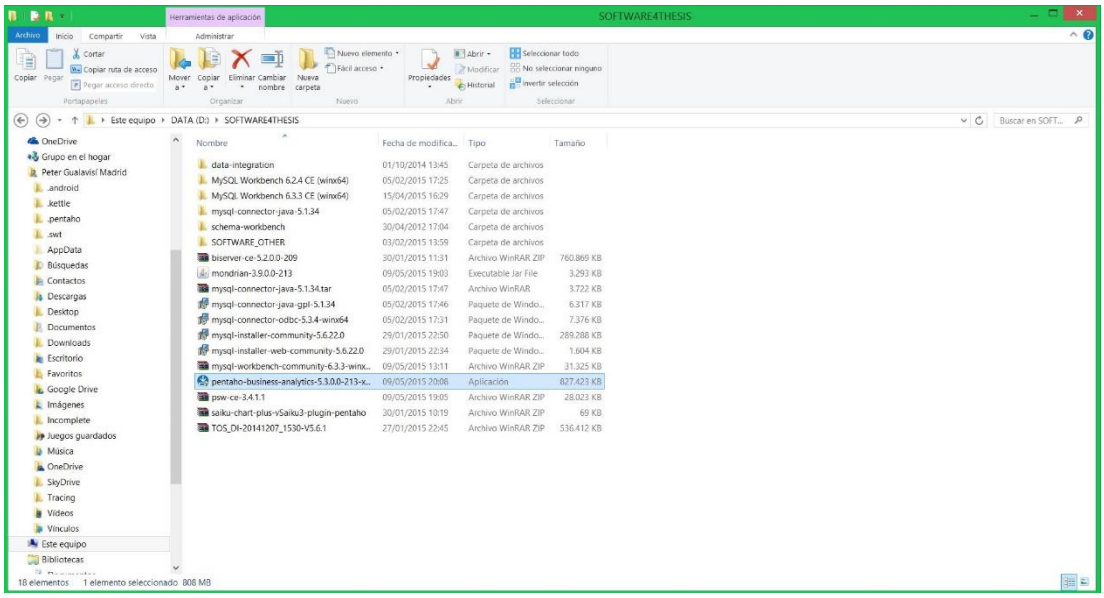

Figura 62. Instalador Pentaho 1.

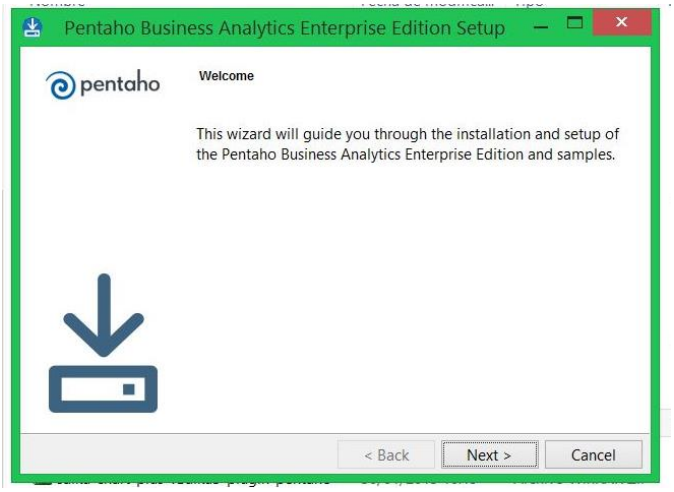

Figura 63. Instalación Pentaho 2

Directorio donde se alojará el servidor Pentaho

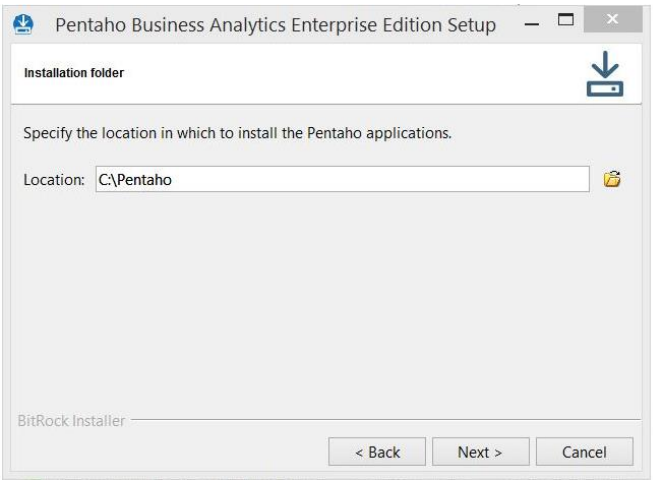

Figura 64. Directorio Servidor Pentaho

Definir usuario y contraseña para el servidor de Pentaho

Figura 65. Usuario y contraseña para servidor

4.4.6.2 Entrada a Pentaho

Una vez instalado el servidor de Pentaho, nos dirigimos al directorio en el cual se instaló para inciar el servidor, tomando en cuentan que esto se debe hacer cada vez que se inicie nuestro Sistema Operativo.

A continuación se descarga la librería de conexión mysql-connector-java-5.1.34-bin la cual de debe guardar en el directorio C:\Pentaho\server\biserver-ee\tomcat\lib. Esta librería es necesaria para la conexión al DataSource que se encuentra alojado en una Base de Datos MySql.

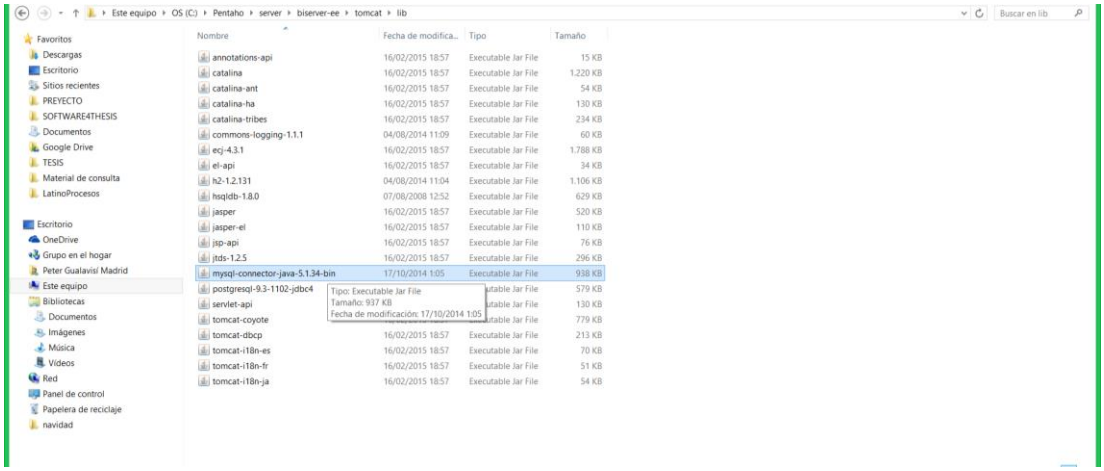

Figura 66. Librería para la conexión a MySql

Procedemos a ejecutar el archivo cmd que nos permite iniciar el servidor de Pentaho.

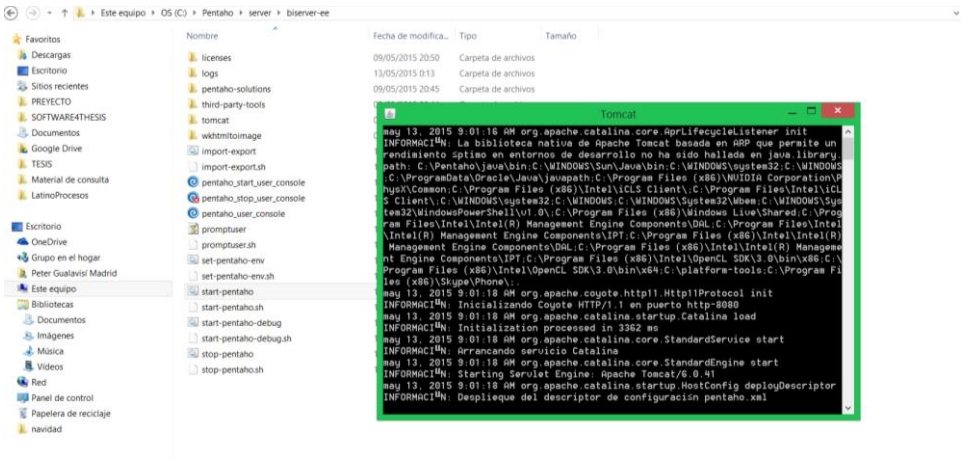

Figura 67.Inicio de Servidor

Nos dirigimos al directorio donde se instaló el sistema Pentaho, para iniciar el ejecutable que abrirá una página web donde se va a ingresar a los servicios de Pentaho.

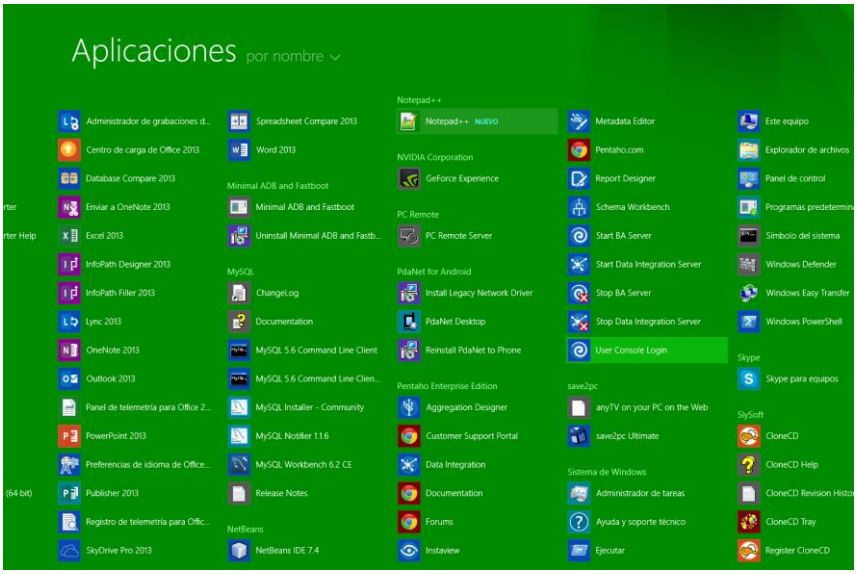

Figura 68. Ejecutable para abrir la Pagina de Pentaho

Ingreso al Portal Web de pentaho, aquí se ingresa con el usuario “admin” y la contraseña “password”.

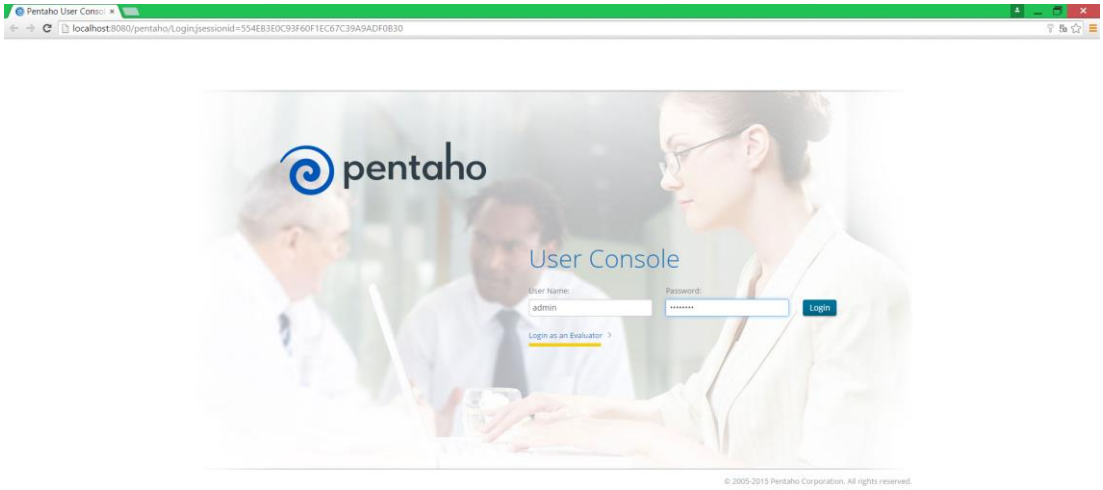

Figura 69. Login de al Portal Web

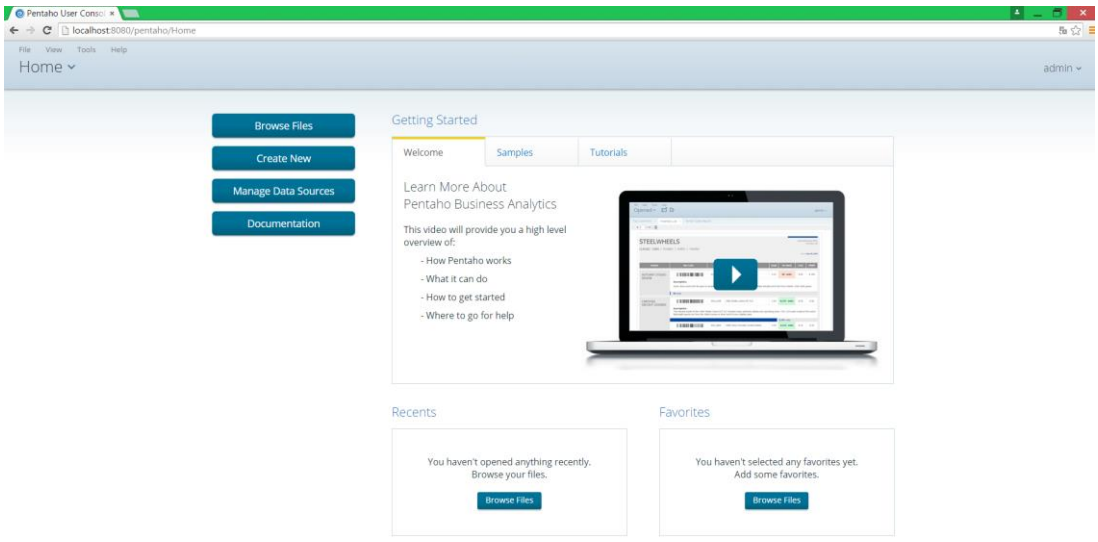

Figura 70. Home del Portal Web

4.4.6.2.1 Creación data source

Es de fundamental crear el data source, porque de ahí radica la conexión a nuestra base de datos que contienen las Dimensiones y Tablas de Hechos.

Primero se debe seleccionar la opción Crear Nuevo DataSource. Ahí se tiene que asignar el nombre del DataSource que se le dará al proyecto y seleccionar el tipo de origen para los datos, en este caso es una base de datos con tablas.

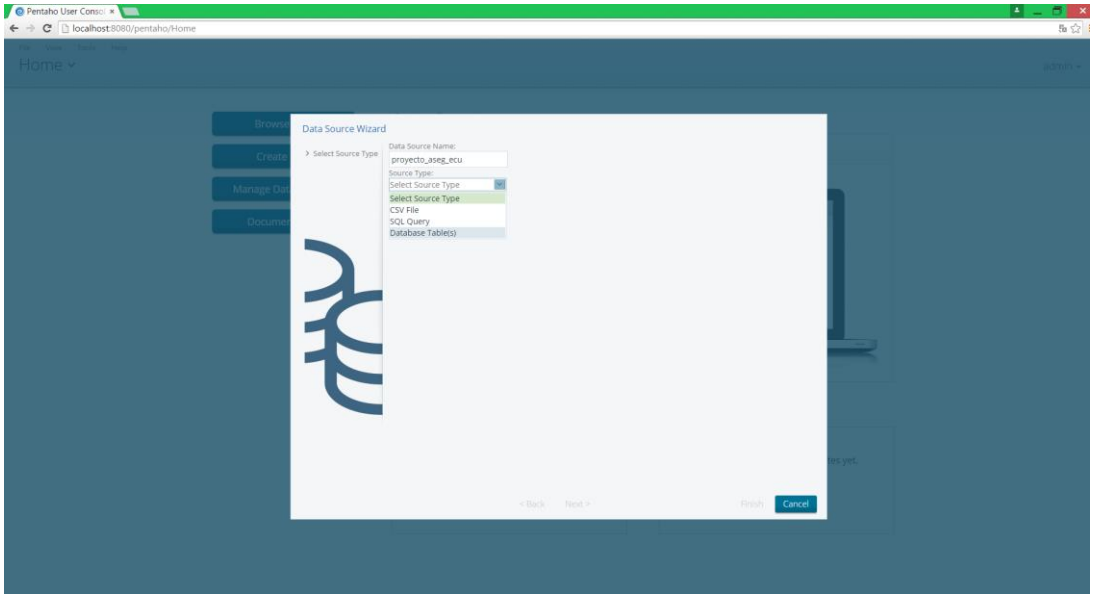

Figura 71. Nombre y tipo de Data Source

Se crea la conexión la cual nos permitirá acceder a nuestra base de datos que en este caso es MySql, se asigna como host “local host”, debido a que nuestro servidor es local.

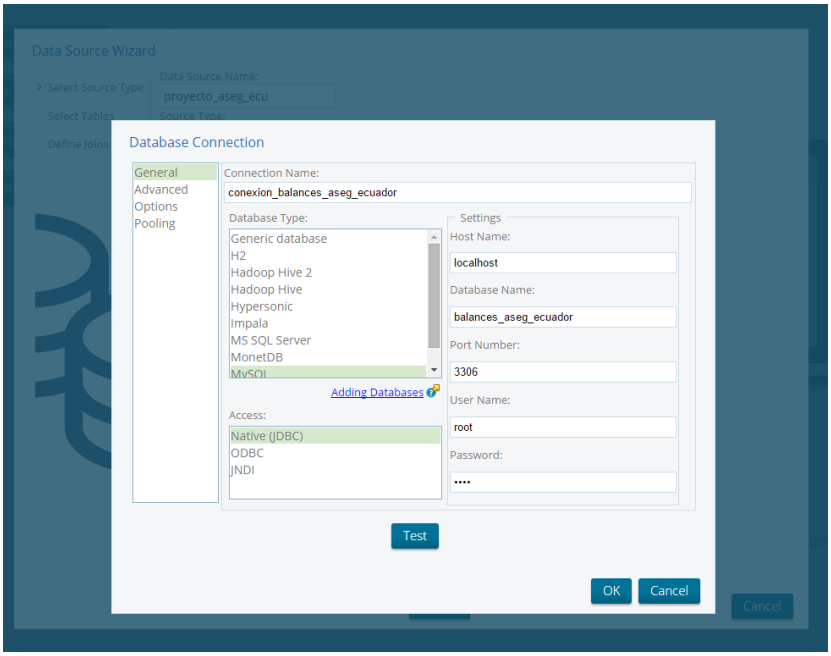

Figura 72. Data Source, Conexión a la Base de datos

Luego se asignan las tablas de dimensiones y nuestra tabla de hechos que hemos creada y debidamente poblado. Asignando según corresponda.

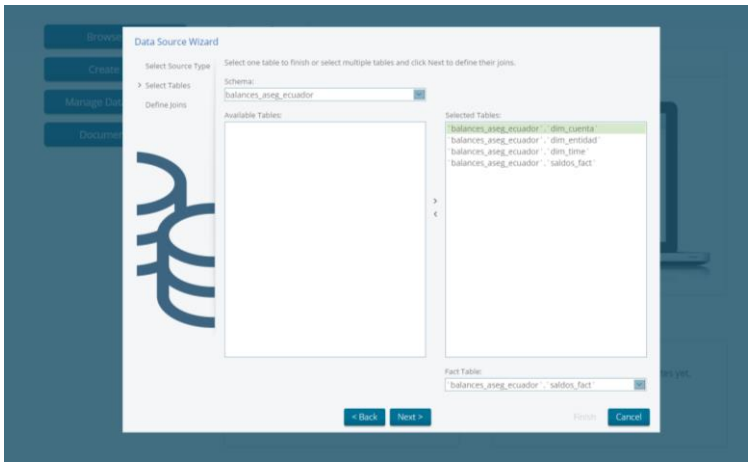

Figura 73. Data Source, Asignación de dimensiones y tabla de hechos

En estas descripciones se asocian los nombres de los campos que se utilizaron en los data marts de las bases de datos operacionales.

El archivo XML que define las dimensiones y las medidas usadas en el cubo que soporta el Data Marts.

```xml
<?xml version="1.0" encoding="ISO-8859-1"?>
<Schema name="CuboSEG">
    <Cube name="CuboSEG">
        <Table name="bcos_saldos_fact"/>
        <Dimension type="TimeDimension" name="Períodos" foreignKey="sk_time">
            <Hierarchy hasAll="true" allMemberName="Períodos Todos" primaryKey="sk_time">
                <Table name="dim_time"/>
                <Level name="Años" table="dim_time" column="theyear" uniqueMembers="true" levelType="TimeYears" type="Numeric"/>
                <Level name="Trimestres" table="dim_time" column="qtrnum" nameColumn="qtrabrev_sp" levelType="TimeQuarters" uniqueMembers="false"/>
                <Level name="Meses" table="dim_time" column="mthnum" nameColumn="mth_sname_sp" levelType="TimeMonths" uniqueMembers="false"/>
            </Hierarchy>
        </Dimension>
        <Dimension name="Entidades" foreignKey="sk_entidad">
            <Hierarchy name="Entidades" hasAll="true" allMemberName="Entidades Todos" primaryKey="sk_entidad">
                <Table name="dim_entidad"/>
                <Level name="TipoEntidad" table="dim_entidad" column="tipo_entidad" uniqueMembers="false"/>
                <Level name="NombreEntidad" table="dim_entidad" column="cat_codigo_entidad" nameColumn="cat_nombrecorto_entidad" uniqueMembers="false"/>
            </Hierarchy>
        </Dimension>
        <Dimension name="Categorias" foreignKey="sk_puc">
            <Hierarchy name="Categorias" hasAll="true" allMemberName="Categorias Todos" primaryKey="sk_puc">
                <Table name="dim_puc"/>
                <Level name="Categoria" table="dim_puc" column="id_balance" uniqueMembers="false" type="String"/>
            </Hierarchy>
        </Dimension>
        <Dimension name="Niveles" foreignKey="sk_puc">
            <Hierarchy name="Niveles" hasAll="true" allMemberName="Niveles Todos" primaryKey="sk_puc">
                <Table name="dim_puc"/>
                <Level name="Nivel" table="dim_puc" column="nivel" uniqueMembers="false" type="Numeric"/>
            </Hierarchy>
        </Dimension>
        <Dimension name="PlanDeCuentas" foreignKey="sk_puc">
            <Hierarchy name="PlanDeCuentas" hasAll="true" allMemberName="Cuentas Todos" primaryKey="sk_puc">
                <Table name="dim_puc"/>
                <Level name="Cuentas Contables" table="dim_puc" column="fullctanombre" uniqueMembers="false"/>
            </Hierarchy>
            <Hierarchy name="Catálogo" hasAll="false" allMemberName="" primaryKey="ID_CUENTA">
                <Table name="dim_puc" />
                <Level name="CtaNum" table="dim_puc" column="numctanum" uniqueMembers="false" />
            </Hierarchy>
        </Dimension>
        <Measure name="SaldosEnMiles" column="saldos" aggregator="sum" datatype="Numeric" formatString="#,##0.0"/>
    </Cube>
</Schema>
```

Figura 74. XML Creación Cubo Analítico

4.4.6.3 GENERACIÓN DE REPORTES

Según nuestros requerimientos podemos generar los reportes asignando a las columnas y filas los campos que se desean analizar tomando en cuenta que los measures son los hechos o valores que se deben de ir.

A continuación se muestra un reporte generado y que nos muestra las compañías en las filas y las cuentas correspondientes en las columnas así como el valor de las cifras para Generales.

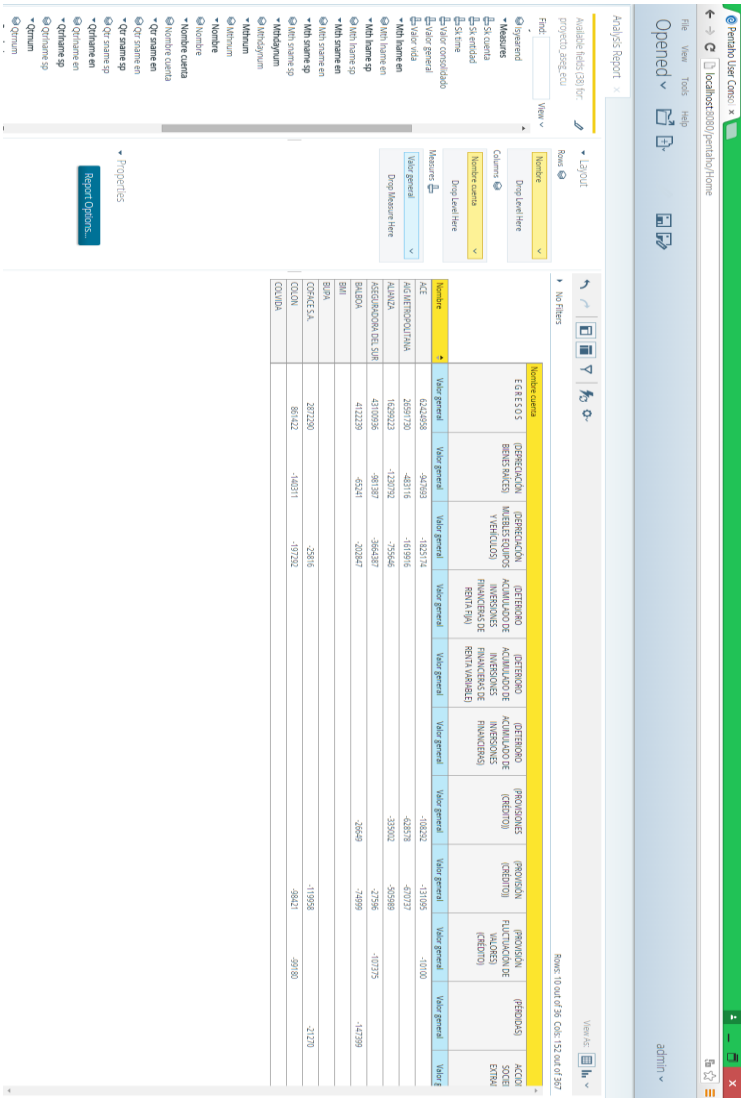

Figura 75. Reportes Cubo

De la misma manera podemos generar gráficos de vital importancia para la toma de decisiones o para analizar la información de forma dinámica.

El siguiente gráfico nos muestra los valores para cada Compañía de Seguros en las cuentas de Activos y Egresos.

Figura 76. Gráfico de Cubo Analitico

4.5 MANTENIMIENTO Y CRECIMIENTO

En cuanto al mantenimiento y crecimiento del repositorio de información, se puede decir que se hará actualizaciones cada mes que se publique un nuevo periodo de información, por el momento para analizar el funcionamiento y mantener la información mensual.

En cuanto al crecimiento, la tabla que tendrá un mayor crecimiento es la tabla de hechos de saldos (SALDOS_FACT), ya que esta es la tabla que contiene todos los valores de las cuentas y por tal motivo recibe una cantidad significativa de datos. No existe riesgo alguno de que los datos se pierdan o entren en conflicto con la aplicación del sistema, ya que los archivos de origen (fuente) están almacenados y se pueden repoblar las veces que sea necesaria y la información destino está en una Base de Datos que garantizan y soportan la cantidad de información que se pueda generar.

La carga actual de datos al repositorio se la realiza en un tiempo aproximado de 3 minuto, tal como lo muestra el histórico log de carga y análisis generado por el PDI.

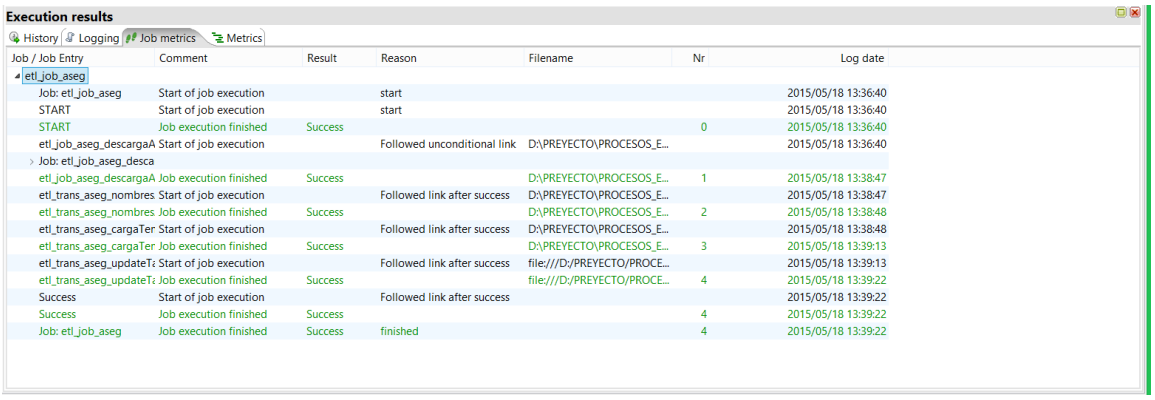

Figura 77. Duración de los procesos ETL

CONCLUSIONES Y RECOMENDACIONES

5 CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

Como resultado del desarrollo del presente documento de tesis en donde se realizó el estudio de metodologías de Data Warehouse para la implementación de repositorios de información para la toma de decisiones gerenciales, y luego la aplicación de la metodología seleccionada en la implementación de una aplicación utilizando la herramienta de Inteligencia de negocios Pentaho para crear reportes que permitan conseguir los objetivos planteados, se muestran las siguientes conclusiones:

- La implementación de herramientas de Inteligencia de negocios en las empresas colabora al mejoramiento de la administración y gestión de los datos, mostrando una mejor visión del estado actual e histórico de las empresas o negocios a través de la toma de decisiones oportunas.

- La utilización de una metodología de desarrollo tanto para implementaciones de software como de desarrollo de Data Warehouse permiten obtener productos de calidad y en tiempos relativamente cortos ya que se conoce los pasos a seguir y las posibles complicaciones que se puede tener en el transcurso.

- El uso de la metodología Ralph Kimball representa un proceso eficaz en tiempo y recursos debido a que se obtiene la solución al problema en corto plazo, acoplándose a la metodología tradicional de desarrollo de software.
- Existen excelentes herramientas de software libre para el desarrollo de sistemas de inteligencia de negocios. La versión community de Pentaho por ahora está siendo mantenido por la comunidad, pero actualmente existe alternativas a jpivot como visor OLAP.

- Cuando se trata de reportes que utilizan una gran cantidad de información el usuario debe tener un mayor conocimiento de las dimensiones que se utilizan.

5.2 RECOMENDACIONES

A partir del desarrollo del presente trabajo se dará algunas recomendaciones respecto del mismo, que los usuarios deberán tomar en cuenta para la correcta utilización del Cubo Analítico que ha sido objeto de este proyecto.

Se recomienda que si se utiliza bases de datos en cualquier DBMS difundido en el mercado, y se requiere realizar análisis multidimensional sobre esos datos, que además tienen estructuras complejas, lo haga con Mondrian el mismo que permite realizar conexiones JDBC.

Implementar Mondrian como herramienta OLAP en una empresa, requiere de un equipo que como mínimo tenga 8 GB de memoria RAM y un buen procesador para que ofrezca un mejor rendimiento en su reporteo dinámico, además de configurar la máquina virtual de java en el archivo de inicio del servidor sea este Tomcat el cual se utilizó para la implementación de esta solución o cualquiera para que el procesamiento y presentación de resultados sea eficiente.

Al trabajar con Mondrian y si no se conoce tan a fondo el desarrollo de un análisis OLAP se recomienda utilizar la herramienta Schema Workbench de Pentaho que es totalmente Open Source y es una gran ayuda para el desarrollo de cubos OLAP con Mondrian.

Implementar rutinas de carga de datos para otros países de la región debido a que las tablas de hechos actúan para cualquier país, y los procesos carga para las dimensiones actuaría de la misma manera.

Para el poblamiento de los datos a futuro se debe implementar un proceso de notificaciones para los periodos que se generen.

Notificaciones avanzadas que nos permitan identificar dificultades en los procesos de carga de datos.

Bibliografía

- Bajwa, F. R. (2002).
- Calvache, J. (1995). *Contabilidad General.* Bogotá: Unisur.
- Cano. (2007). *Dataprix Knowloedge.* Madrid: Fundación Cultural Banesto.
- Carbone, A. L. (2001). *Open Source Enterprise Solutions: Developing an E-Business Strategy.* Paperback.
- Carmen, M. (2015). El sector Asegurador. *El sector Asegurador*, 62.
- Castelo, M. (1988). Diccionario MAPFRE de Seguros. *Diccionario MAPFRE de Seguros*, 257.
- Dario. (2007). *DATA WAREHOUSING.* Córdoba.
- Davenport. (1999). *Working Knowledge.* USA.
- Desconocido. (2008). *SISTEMA DE INFORMACIÓN*. Obtenido de http://definicion.de/sistema-de-informacion/
- Dongen, B. &. (2009). *Pentaho.* Madrid.
- Dongen, V. (2009). *Pentaho Solutions Bussiness Intelligence and DataWarehousing with Pentaho and Mysql.* Indianapolis.
- Efrén, O. (1998). *Teoría General del Seguro.* Bogotá.
- Esteve, P. &. (1999).
- Fierro, A. (2007). *Contabilidad de Activos.* Bogotá: ECOE.
- Fierro, A. (2009). *Contabilidad de Pasivos.* Bogotá: ECOE.
- Golden, B. (2004). *Succeeding with Open Source.* USA: Paperback.
- Guzman, A. (2006). *Contabilidad financiera.* Bogotá: Universidad de Rosario.
- Inmon, B. (1992). *Building the Data Ware House.*
- Inmon, W. H. (2002). *Building the Data Warehouse (3rd Edition).*
- Inner, A. (2004). *Supply chain management*. Obtenido de http://www.monografias.com/trabajos89/supply-chain-management/supply-chain-management2.shtml

- Kavanagh, P. (2004). *Open Source Software: Implementation and Management.* Paperback .
- Kimball. (2002). *The Data Warehose Toolkit.* Willey.
- Kimball, R. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition).* Paperback.
- Larissa T. Moss, S. A. (2003). *Business Intelligence Roadmap: The Complete Project Lifecycle for Decision-Support Applications.*
- Orgill, D. S. (2000).
- Pastor, E. (1999).
- Pérez, D. (26 de Octubre de 2007). *¿Qué son las bases de datos?* Obtenido de http://www.maestrosdelweb.com/editorial/%C2%BFque-son-las-bases-de-datos/
- Roberto, E. (2010). *Aplicación de modelos dimesionales.* Lima.
- Serrano, M. (2015). *Superintendencia de Bancos y Seguros del Ecuador*. Obtenido de Superintendencia de Bancos y Seguros del Ecuador: http://portaldelusuario.sbs.gob.ec/contenido.php?id_contenido=66
- Students. (01 de 2011). *Students*. Obtenido de Students: http://businessintelligencemustudents.blogspot.com/2011/01/retos-y-desventajas-de-bussines.html
- Wallace. (2000).
- Wikipedia. (2012). *Wikipedia*. Obtenido de Wikipedia: http://es.wikipedia.org/wiki/Inteligencia_empresarial