ISMMM
INSTITUTO SUPERIOR MINERO
METALURGICO DE MOA
DR. ANTONIO NUÑEZ JIMENEZ

Ingeniería Informática
Facultad: Geología y Minas

Trabajo de Diploma

Para Optar por el Título de

Ingeniero Informático

Solución de Inteligencia de Negocio para los servicios gastronómicos en la Empresa de Servicios a la Unión del Níquel.

Autor: *Yanet Molina Hernández.*

Tutores: *Ing. Vladimir Torres Torres.*

MSc. Edgar Núñez Torres.

Moa, 2014
"Año 56 de la Revolución"

Pensamiento

"Ha llegado la hora en que la información disponible en su empresa se convierta en un auténtico activo capaz de generar negocio".

Emilio Arias

Dedicatoria

Este trabajo está dedicado especialmente a mi abuela Nilda Matos Rosabal y a mis padres Irais y Aida a los cuales les debo todo lo que soy.

Agradecimientos

A mis padres por su paciencia, comprensión y apoyo ese es el motivo por el cual llegué a este punto, son ustedes mi razón de ser y es por eso que les debo la vida.

A dos personas que quiero mucho las cuales han sabido educarme y ayudarme al igual que mis padres David y Marbelis gracias por todo.

A mis hermanos David, Leonardo, Lieti, Danyer y Yanilis a los que adoro con la vida.

A mi abuelita María Luisa y mi abuelo Irais gracias por todos los concejos, para ustedes es mi cariño y respeto.

A mi esposo Ernesto por soportarme y apoyarme en los momentos que más lo necesité, para el todo mi amor.

A Doris por ser como una madre para mí, por ayudarme, escucharme, su cariño e incluso por sus regaños.

A mi familia en general por todo el cariño y apoyo que he recibido de ella.

A mis tutores Vladimir y Edgar gracias por la ayuda que me brindaron.

A mis amistades Pikiri y Cala muchísimas gracias por todo lo que hicieron por mí.

A mis amistades Lily, Yeilin, Indira y Dailin gracias por soportarme y apoyarme.

A mis profesores y compañeros de aula por encontrarse en los momentos que más los necesité en especial a Indira, Edgar, Marcos, Eugenio y Jorge.

Gracias a los que de una manera u otra me ayudaron a realizar esta investigación y a los que no también.

Declaración de Autoría

Declaro que yo soy la única autora de este trabajo y autorizo al Instituto Superior Minero Metalúrgico de Moa “Dr. Antonio Núñez Jiménez” para que hagan el uso que estimen pertinente del mismo.

Para que así conste firmamos la presente a los días del mes de del.

_____

Yanet Molina Hernández

Firma del autor

_____ _____

MSc. Edgar Núñez Torres

Firma del tutor

Ing. Vladimir Torres Torres.

Firma del tutor

Resumen

La Inteligencia de Negocios es una estrategia que ha alcanzado un nivel elevado cuando de competitividad empresarial se trata. El aplicar una solución de Inteligencia de Negocios parte de los sistemas de origen de datos que posee una organización, apoyándose de un conjunto de herramientas encargadas de la extracción, depuración y consolidación de los datos. Esta información será almacenada en un Data Warehouse o en los Data Mart los cuales es una unidad más pequeña orientada a áreas específicas o un tema en particular.

En esta investigación se realiza el diseño e implementación de un Data Mart como solución de Inteligencia de Negocios para los servicios de alimentación prestados por la Empresa de Servicios a la Unión del Níquel la cual radica en el municipio de Moa. Se emplearon varias herramientas que permitieron la construcción del Data Mart y fue seleccionada la metodología HEFESTO para el desarrollo del mismo. Permitiendo que toda la información generada por los servicios gastronómicos se encuentre en un lugar específico, depurada y consolidada sirva como soporte a la toma de decisiones en los servicios gastronómicos de la ESUNI.

Abstract

The Intelligence of Business is a strategy that has reached a high level when of managerial competitiveness it is. Applying a solution of Business Intelligence it begin with the systems data origin that it possesses a company, leaning on a tools group in charge of the extraction, purification and consolidation of the data. This information will be stored in a Data Warehouse or in Data Mart which are smaller units guided in to specific areas or a particular topic.

In this investigation is carried out the design and implementation of a Data Mart like solution of Business Intelligence for gastronomic services for the Company of Services to the Union of Nickel which resides in the municipality of Moa. Several tools were used that allowed the construction of the Data Mart and HEFESTO was the methodology selected for the development of the same. Allowing that all the information generated by the gastronomic services is in a specific place purified and consolidated serves like support to the taking of decisions in the gastronomic services of the ESUNI.

Índice

INTRODUCCIÓN **1**

CAPÍTULO I: FUNDAMENTOS TEÓRICOS **5**

1.1 INTRODUCCIÓN 5
1.2 INTELIGENCIA DE NEGOCIO. 5
1.3 DATA WAREHOUSE 6
1.3.1 Características *7*
1.3.2 Estructura del Data Warehouse *9*
1.3.3 Arquitectura del Data Warehouse *10*
1.4 DATA MART 13
1.5 MODELO MULTIDIMENSIONAL. 14
1.5.1 Esquema en Estrella *17*
1.5.2 Esquema Copo de Nieve. *17*
1.5.3 Esquema Constelación *18*
1.5.4 Tipos de implementación para el diseño del Data Warehouse. *19*
1.5.5 Ventajas *21*
1.6 PROCESO DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA, PROCESO DE INTEGRACIÓN DE DATOS. 21
1.6.1 Función de Extracción. *22*
1.6.2 Función de Transformación. *22*
1.6.3 Función de Carga. *23*
1.7 ESTADO DEL ARTE. 23
1.8 AMBIENTE DE DESARROLLO. 24
1.9 JUSTIFICACIÓN METODOLÓGICA: METODOLOGÍA HEFESTO 24
1.9.1 Herramienta de diseño y Modelado de Almacenes de Datos: MySQL Workbench 5.1 ..... *26*
1.9.2 Herramienta realización de los procesos ETL: Pentaho Data Integration PDI 4.0 ..... *27*
1.9.3 Herramienta para Gestión de Base de Datos: MySQL 5.0 ..... *27*
1.9.4 Herramienta Servidor de BI: Pentaho BI Server 4.0 ..... *28*
1.10 CONCLUSIONES PARCIALES 30

CAPÍTULO II: ANÁLISIS Y DISEÑO DEL DATA MART. **31**

2.1 INTRODUCCIÓN. 31
2.2 ANÁLISIS DE LOS REQUERIMIENTOS. 31
2.2.1 Identificar preguntas. *31*
2.2.2 Identificar indicadores y perspectivas. *32*
2.2.3 Modelo conceptual. *33*
2.3 ANÁLISIS DE LOS OLTP. 34
2.3.1 Conformar indicadores. *34*
2.3.2 Establecer correspondencias. *34*
2.3.3 Nivel de granularidad. *36*
2.3.4 Modelo conceptual Ampliado. *37*
2.4 MODELO LÓGICO DEL ALMACÉN DE DATOS. 38
2.4.1 Tipo de modelo lógico *38*
2.4.2 Tablas de dimensiones *39*
2.4.3 Tablas de Hechos. *39*
2.4.4 Uniones *40*
2.5 MODELO DE ARQUITECTURA DEL MERCADO DE DATOS. 40
2.6 CONCLUSIONES PARCIALES 41

CAPÍTULO III: IMPLEMENTACIÓN Y PRUEBA. **42**

3.1 INTRODUCCIÓN. 42
3.2 IMPLEMENTACIÓN DEL DATA MART. 42
3.4 PROCESOS DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA DEL DATA MART. 42
3.5 SCHEMA OLAP. 48
3.6 ANÁLISIS DE LOS TIEMPOS DE CARGA DEL DATA MART. 51
3.7 PRUEBA DE INTEGRIDAD DE DATOS. 51
3.8 PRUEBA DE VALIDACIÓN. 52
3.9 CONCLUSIONES. 54

CONCLUSIONES GENERALES **55**

RECOMENDACIONES **56**

BIBLIOGRAFÍA **57**

ANEXOS **59**

Introducción

El desarrollo de la industria minera, se caracteriza por la trasformación de la minería, en un proceso de producción planificada para mejorar las condiciones de vida del pueblo. Al mismo tiempo, esta industria constituye una fuente estable de divisa, necesaria para el desarrollo.

Nuestro país, en su parte oriental, específicamente en la zona Nícaro-Moa-Baracoa, cuenta con grandes yacimientos minerales. Siendo el municipio Moa el más beneficiado puesto que en él se encuentra el gran grueso de ellos.

Para mantener esta industria estable se necesita servicios importantes donde la Empresa Servicios a la Unión del Níquel es la encargada de prestar los servicios de alimentación, hospedaje y transporte entre otros, a los trabajadores de tal sector.

Se cuenta con un grupo de control de los servicios gastronómicos que es encargado de lograr mayor dinamismo y eficiencia en el control de los servicios prestado por la Empresa de Servicios a la Unión del Níquel. Este control de los servicios gastronómicos desarrollado por los especialistas es una tarea agotadora que puede durar días cuando en ocasiones urgen resultados en periodos de tiempo extremadamente cortos y para ello cuenta con una forma que no es la más eficiente puesto que la información llega desde varios procesos en distintos formatos atentando contra la homogenización, consolidación y posterior procesamiento, todo ello conlleva a las siguientes dificultades:

- ✓ Descentralización de la información.
- ✓ Grandes cantidades de datos almacenados en hojas de cálculo Excel.
- ✓ Generación de informes muy extensos con hojas adheridas.
- ✓ Generación de informes no homogéneos.
- ✓ Duplicidad de Información entre el sistema informático y los archivos Excel.
- ✓ Desgaste físico de los especialistas.
- ✓ Dificultad en la toma de decisiones.
- ✓ Los procesos de análisis y elaboración de informes consumen tiempo excesivo.

Lo anteriormente expuesto da la medida de cuan necesaria se hace para la empresa una solución inteligente, por tanto en esta investigación se brindará solución a la situación antes mencionada que nos conduce al ***problema científico:*** ¿Cómo favorecer el análisis y procesamiento de la información para tomar decisiones más acertadas y objetivas en el control de los servicios de alimentación de la Empresa de Servicios a la Unión del Níquel "ESUNI"?

Para dar solución al problema antes mencionado se propone como ***objeto de estudio:*** La Inteligencia de Negocios.

El ***campo de acción*** viene enmarcado en el proceso de desarrollo de un Mercado de Datos para el análisis de los servicios de alimentación en la ESUNI.

Esta investigación se sustenta sobre la ***idea a defender:*** Con el desarrollo de un Mercado de Datos que sea capaz de centralizar, integrar, procesar y analizar la información favorecerá a que en el proceso de control de servicios de alimentación de la Empresa de Servicio a la Unión del Níquel se tomen decisiones más acertadas y objetivas.

De esta manera queda comprendido en un ***objetivo general:*** Desarrollar un Mercado de datos que facilite la centralización, análisis y procesamiento de la información para soporte a toma de decisiones en el proceso de control de servicios de alimentación en la Empresa de Servicios a la Unión del Níquel "ESUNI".

Para lograr los objetivos propuestos se ha trazado un ***sistema de tareas*** a realizar:

- ✓ Revisión de los principales conceptos asociados a las soluciones de inteligencia de negocios.
- ✓ Análisis para la definición de la metodología a seguir para la solución de inteligencia de negocios.
- ✓ Análisis de las herramientas de modelado, desarrollo y gestión de datos.
- ✓ Análisis de la arquitectura del mercado de datos.
- ✓ Entrevista con los especialistas encargados en el control de servicios alimentación de la ESUNI para la identificación de los requisitos de información.
- ✓ Análisis del modelo de datos de los sistemas OLTP.
- ✓ Identificación y descripción de las medidas dimensiones y hechos.

- ✓ Estructuración del modelo conceptual y modelo lógico de los datos.
- ✓ Definición de la arquitectura del mercado de datos.
- ✓ Diseño de los procesos de integración de los datos.
- ✓ Implementación del modelo de datos.
- ✓ Implementación de los cubos multidimensionales.
- ✓ Implementación de los reportes candidatos.
- ✓ Realización de las pruebas de integridad de datos al mercado de datos.
- ✓ Realización de pruebas de aceptación para validar la solución de inteligencia de negocios.

Para dar respuesta a estas tareas propuestas se emplearon los métodos científicos de la investigación: teóricos y empíricos.

Los métodos empíricos estudian las características y relaciones esenciales del objeto que son accesibles directamente desde la percepción sensorial (conocimiento sensorial).
Los métodos teóricos se aplican durante el proceso de explicación, predicción, interpretación y comprensión de la esencia del objeto. Posibilitan la interpretación conceptual de los datos empíricos, revelan las relaciones esenciales del objeto de investigación que no son observados a simple vista, participan en la construcción del modelo y la hipótesis de la investigación
Entre los métodos empíricos se encuentran:

- ✓ Entrevista: esencial para adquirir conocimientos acerca del problema en general. Para la elaboración de esta investigación fue necesario realizar varias entrevistas a los especialistas vinculados al tema, para lograr obtener los requerimientos fundamentales partiendo de las necesidades de la empresa.
- ✓ Revisión de documentos: fueron realizados varios estudios de diferentes bibliografías para alcanzar conocimiento de los conceptos básicos de esta investigación, así como el estudio de las herramientas y metodologías para de esta manera realizar la selección de la misma.
- ✓ La observación: es el principal objetivo para conocer como es gestionada la información en la Empresa de Servicios a la Unión del Níquel “ESUNI”.

Entre los métodos teóricos se encuentran:

- ✓ Análisis y síntesis: adoptado en la recolección y procesamiento de la información obtenidas en los métodos empíricos para de esta manera arribar a conclusiones.
- ✓ Método de Modelación: empleado para obtener la construcción de los modelos pertinentes entre los que se encuentran los modelos tanto lógico como físico de la Base de Datos.
- ✓ Inducción–deducción: empleado para la aplicación de la metodología de desarrollo y la interpretación de los resultados.

El documento consta de tres capítulos distribuidos de la siguiente manera:

Capítulo I FUNDAMENTOS TEÓRICOS: Se hace referencia al objeto de estudio de esta investigación, conceptos fundamentales relacionados al tema a tratar. Se ejemplificará la metodología aplicada para el desarrollo de un Data Mart así como las herramientas escogidas para el diseño e implementación del mismo.

Capítulo II ANÁLISIS Y DISEÑO DEL DATA MART: Se realizará el diseño del Mercado de datos partiendo de las perspectivas identificadas y los indicadores que apuntan a las necesidades del negocio. Para la construcción del mismo será aplicada la metodología y las herramientas seleccionadas en el capítulo anterior.

Capítulo III "IMPLEMENTACIÓN Y PRUEBA'': Se realizará la implementación del mercado de datos teniendo en cuenta las técnicas y procesos de extracción, transformación y carga de datos. Luego se procederá a conformar el diseño que caracteriza al mismo, así como el análisis de los tiempos de carga y para culminar las pruebas de integridad y aceptación de los de los datos.

CAPÍTULO I: FUNDAMENTOS TEÓRICOS

1.1 Introducción

En el presente capítulo se realizará un estudio del tema de la investigación, dando a conocer las definiciones principales que debemos tener en cuenta para llevar a cabo un buen proceso de Inteligencia de Negocios. Conjuntamente se analizarán las técnicas de integración, así como las herramientas y la metodología aplicada, de esta manera podremos cumplir con lo establecido por el proceso de Inteligencia de Negocio lo cual corresponderá a ser la vía más eficiente para la confección de un buen diseño e implementación del mercado de datos.

1.2 Inteligencia de Negocio.

La tecnología es una actividad que influye principalmente en el progreso social y económico de nuestra sociedad. Es en el aspecto económico que la competitividad en el negocio empresarial va alcanzando un nivel elevado, ya que como consecuencia de este desarrollo tecnológico las empresas se perfeccionan progresivamente. Las organizaciones precisan de mejoras con el objetivo de lograr mayor productividad y sobre todo disminuir los gastos originados, para lograr esto deben contar con un mecanismo eficiente capaz de mantener un control absoluto de su producción.

La Inteligencia de Negocios es una estrategia que permite que la información diaria de cualquier organización se encuentre de una manera centralizada y organizada para así brindar una mejor perspectiva de análisis. Es un sistema de herramienta capaz de recopilar toda la información de diferentes fuentes y diversos formatos para luego depurarlos, integrarlos y transformarlos con la finalidad que los datos almacenados sean certeros y seguros capaces de resolver las preguntas del negocio según las necesidades de la empresa.

Aplicada esta alternativa todos los datos que se generen serán convertidos en conocimientos y decisiones en tiempo real, favoreciendo la toma de decisiones a los usuarios y así ganar tiempo a la hora de realizar las consultas obteniendo respuestas con mayor rapidez. Para lograr que toda la información se encuentre en un lugar específico y de manera centralizada se crea el Data Warehouse capaz de almacenar toda la información en un mismo formato.

Podemos hacer mención de una definición sobre el Business Intelligence realizada en la tesis "Estudio de metodologías de Data Warehouse para la implementación de repositorios de información para la toma de decisiones gerenciales." cito:

"BI es un proceso interactivo para explorar y analizar información estructurada sobre un área (normalmente almacenada en un Data Warehouse), para descubrir tendencias o patrones, a partir de los cuales derivar ideas y extraer conclusiones. El proceso de Business Intelligence incluye la comunicación de los descubrimientos y efectuar los cambios." (Villarreal Rosero, 2013)

1.3 Data Warehouse

Las empresas luego de gestionar los datos obtenidos en diversos formatos, fuentes y tipos, para posteriormente depurarlos e integrarlos debían almacenarlos en un solo destino o base de datos. La creación del Data Warehouse en la década de los 90 resultó ser la alternativa para que las organizaciones almacenaran la información de una manera entendible y consistente. Este almacén de datos o colección de datos tiene una estructura multidimensional que está orientada al negocio, es variante en el tiempo y no volátil lo que favorece el soporte a la toma de decisiones. A continuación se especificaran algunas definiciones acerca del Data Warehouse (Vizuete Naranjo, y otros, 2006):

Ralph Kimball define el Data Warehouse:

- *"Una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis".*
- *"La unión de todos los Data Marts de una entidad".*

John Edwarts expresó:

- *"Un Data Warehouse toma información de múltiples sistemas y bases de datos y las almacena de una manera que está diseñada para dar a los usuarios accesos más rápidos, más fácil y más flexibles a los aspectos claves".*

La definición más famosa y más acertada fue la realizada por William Harvey Inmon en el año 1992 más conocido como el padre del Data Warehouse:

- *"Un Data Warehouse es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia".*

1.3.1 Características

La definición de W. H. Inmon resultó ser la más apropiada y enfocada a describir un almacén de datos, a partir de esta definición surgen las características esenciales del Data Warehouse las mismas serán detalladas a continuación (Vizuete Naranjo, y otros, 2006):

✓ Orientado a Tema:

La información almacenada en el depósito de datos se caracteriza por referirse a temas específicos del negocio según las necesidades de la empresa. Comúnmente las bases de datos constituyen una combinación de un entorno operacional diseñado alrededor de los datos orientados a las aplicaciones. El Data Warehouse está enmarcado en el modelamiento de los datos y en el diseño de la base de datos.

Este repositorio de datos orientado a temas se diferencia del ambiente orientado a las aplicaciones, ya que excluye los datos que no ofrecerá una fuente de información aceptable a los analistas para llevar a cabo el proceso de soporte a la toma de decisiones. Mientras que los datos operacionales en un entorno de aplicación, recolectará toda la información que satisfaga los requerimientos funcionales inmediatos y procesamientos que pueden ser de valor o no para los analistas.

✓ Integrado:

La información extraída de diversas fuentes y aplicaciones debe pasar por un proceso de depuración o limpieza, donde las inconsistencias de los datos analizados serán excluidas, de esta manera estará completa la integración de los datos para luego ser almacenados en el Data Warehouse. Con el objetivo de lograr este proceso de Integración de Datos, se cuenta con un sistema de tareas o técnicas encargadas de analizar los datos y consolidarlos, una de estas técnicas es conocida como ETL (Extracción, Transformación y Carga de datos). La Integración de Datos soluciona:

Las convenciones de nombres debido a las variedades de nombramiento existentes en las diversas aplicaciones.

Las unidades de medidas deberán ser convertidas en una manera estándar y entendible una vez almacenados los datos. No importa de la fuente que sea extraída la información, estas medidas deben coincidir al llegar al Data Warehouse.

La codificación de las estructuras es necesaria ya que los diseñadores tienen diferentes maneras de representar el género, esto no importa de cual manera llegue al Data Warehouse sino que debe almacenarse en un estado integrado uniforme.

Los atributos pueden ser tomados de múltiples fuentes lo que no significa que atentará contra la centralización e integración de los datos una vez almacenados. No importa de dónde se tome la información, la misma será analizada para brindar veracidad y confiabilidad a los analistas para el soporte a la toma de decisiones.

✓ Variante en el tiempo:

El Data Warehouse presenta datos antiguos conocidos como históricos, por lo que poseerá un elemento tiempo. Debido a esta situación la bodega de datos crecerá de manera considerable a medida que se vayan agregando informaciones referidas al negocio, es por esta razón que el tiempo de respuesta a las consultas realizadas por los usuarios tendrá un tiempo de espera considerable.

Los sistemas operacionales se diferencian del almacén de datos porque la información contenida en ellos puede ser cambiada o remplazada, lo cual representa la actividad presente en el negocio. La ventaja de obtener datos históricos es poder realizar comparaciones, predicciones, análisis de tendencias y pronósticos que ayudara de manera gradual la eficiencia y rendimiento de las organizaciones.

✓ No volátil:

Esta característica señala que al ser extraídas las informaciones de las diversas fuentes deberán ser transformadas y consolidadas, de manera tal que al ser almacenadas en el Data Warehouse se podrá agregar datos actualizados, realizar consultas pero nunca podrán ser modificados los datos.

En el ámbito operacional el manejo es totalmente diferente ya que la información contenida en los mismos se modificará o cambiará según las necesidades de la organización, lo que atentará contra la estabilidad de los datos. El almacén de datos soluciona lo anteriormente expuesto ya que la información será útil de manera que sea invariable.

Existen dos tipos de operaciones principales en proceso del Data Warehouse conocidas como carga inicial y acceso de los datos, por tanto no debe realizarse el proceso de actualización por sus funciones de insertar, eliminar y modificar los datos.

Según las características del Data Warehouse podemos resumir que son orientados a temas, sus datos se encuentran integrados, son variantes en el tiempo, no volátiles y que sobre todo se encuentra alejado de ser un ambiente operacional. Una vez tomada la información de diversas aplicaciones sujetas a un entorno operacional, se eliminara cualquier redundancia existente entre los datos de las aplicaciones y el depósito de datos.

1.3.2 Estructura del Data Warehouse

Debido a la estructura multidimensional que posee el Data Warehouse existen varios niveles diferenciados por el detalle que representa la información requerida por la empresa, los mismos serán abordados a continuación (Vizuete Naranjo, y otros, 2006):

- Detalle de datos actuales:

Se le llama datos actuales ya que se refiere a la información almacenada más reciente. Generalmente alcanzan un gran volumen de datos, debido al bajo nivel de granularidad que poseen los datos al ser guardados. Se recomienda almacenarlos en discos aunque resulte más costoso ya que el tiempo de respuesta emitida debe ser veloz.

- Detalle de datos históricos:

Son los datos antiguos que no son consultados con frecuencia, debido al volumen que presenta son guardados a nivel de detalle en un almacenamiento masivo. Mientras no sea prioritario el almacenamiento en un medio de almacenaje alterno, a causa del gran volumen de datos unido al acceso no frecuente de los mismos, es poco usual utilizar el disco como medio de almacenamiento.

- Datos ligeramente resumidos:

Es aquella que proviene desde un bajo nivel de detalle encontrado al nivel de detalle actual. Este nivel del Data Warehouse casi siempre se almacena en disco. Los puntos en los que se basa el diseñador para construirlo son:

  - ✓ Que la unidad de tiempo se encuentre sobre la esquematización hecha.
  - ✓ Qué contenidos (atributos) tendrá la data ligeramente resumida.

- Datos completamente resumidos

El siguiente nivel de datos encontrado en el Data Warehouse es el de los datos completamente resumidos. Estos datos son compactos y fácilmente accesibles por lo general son indicadores que son usados con más frecuencia para el análisis gerencial.

- Metadatos:

Lo que compone la estructura final del Data Warehouse son los metadatos, estos se encuentran en otra dimensión que difieren de otros datos del almacenes de datos ya que la información no es extraída directamente del ambiente operacional. Los metadatos juegan un papel mucho más importante en ambiente del Data Warehouse que en un ambiente operacional ya que son utilizados como (Vizuete Naranjo, y otros, 2006):

- ✓ Un directorio para ayudar al analista a ubicar los contenidos del Data Warehouse.
- ✓ Una guía para el mapping de datos de cómo se transforma, del ambiente operacional al de Data Warehouse.
- ✓ Una guía de los algoritmos usados para la esquematización entre el detalle de datos actual, con los datos ligeramente resumidos y éstos, con los datos completamente resumidos, etc.

Los metadatos contienen la estructura de los datos, el algoritmo utilizado para la esquematización, donde aquellos datos resumidos que se guardan permanentemente en el depósito de datos serán los frecuentemente consultados. También contienen el mapa de los datos desde el ambiente operacional hasta el Data Warehouse.

Flujo de datos

El Data Warehouse mantiene un flujo de datos normal y generalizado, los datos son cargados en el mismo desde el entorno operacional. Una vez extraídos son analizados y transformados para luego ingresarlos en el almacén de datos. La información almacenada se encuentra en un nivel de detalle actual, esta se mantiene y se utiliza en este nivel hasta que son purgadas, resumidas o archivadas. El proceso de los datos históricos se basa en mover los datos actuales detallados a los antiguos datos detallados teniendo en cuenta el tiempo de almacenamiento de la información.

1.3.3 Arquitectura del Data Warehouse

Los elementos que compone la arquitectura del Data Warehouse representan la estructura global de los datos almacenados, la comunicación y presentación de los datos a los usuarios

finales. La arquitectura se puede clasificar según el resumen de los datos las mismas pueden ser:

- Arquitectura de dos capas consta de un Data Warehouse que almacene los datos extraídos de las diversas fuentes y contenga las herramientas necesarias que permita a los usuarios analizar la información almacenada en el Data Warehouse. La ventaja de este tipo de arquitectura es que contiene toda la información referidas a las necesidades de la empresa, pero a su vez esto trae como consecuencia que cuando se desee consultar datos específicos demorara el tiempo de respuesta. (Cruz Guerrero, 2003)
- Arquitectura de tres capas contiene otro nivel de resumen o sumarización de los datos, ya que se compone de varios Data Mart que consiste en pequeños depósitos o departamentos de datos orientados a temas específicos conectados al Data Warehouse, posibilitando el fácil acceso a un tema en específico permitiendo que las consultas sean más rápidas. (Cruz Guerrero, 2003)

Se pueden especificar los elementos que constituyen la arquitectura del Data Warehouse (Pacheco, 2010):

- ✓ Base de datos operacional / Nivel de base de datos externo.
- ✓ Nivel de acceso a la información.
- ✓ Nivel de acceso a los datos.
- ✓ Nivel de directorio de datos (Metadato).
- ✓ Nivel de gestión de proceso.
- ✓ Nivel de mensaje de la aplicación.
- ✓ Nivel de data Warehouse.
- ✓ Nivel de organización de datos.

- Base de datos operacional: Se refiere a los sistemas transaccionales u operacionales y las fuentes de donde es extraída la información mediante el proceso de Data Warehousing.
- Nivel de acceso a la información: Este nivel es responsable de las interfaces entre las herramientas de acceso y las bases de datos. Podemos decir que es la capa de interacción del usuario cuya finalidad es la conversión de los datos almacenados en información fácil y transparente para las herramientas de los usuarios finales. La táctica del Data Warehousing consiste en facilitar a los usuarios finales tener acceso a datos universales sin tener en cuenta las herramientas de acceso ni la ubicación de los mismos, lo cual le permite acceder a todos los datos de la empresa que les resulte necesario.

- Nivel de directorio de datos: Con el fin de proveer el acceso a los datos universales es necesario contar con registros de datos o repositorios de información metadato. El metadato consiste en describir toda la información de los datos vinculados a la empresa contenida en el Data Warehouse, como son los cálculos realizados, de donde provienen los datos, etc. Es preciso contar con varios metadatos que sostenga la información del origen de los datos y las bases de datos de donde sean extraídos los mismos.
- Nivel de gestión de proceso: Para la construcción del y mantenimiento actualizado del Data Warehouse y los metadatos es preciso adoptar varias tareas. Este nivel depende del alto nivel de control para numerosos procedimientos que se deben tener en cuenta para mantener el Data Warehouse actualizado.
- Nivel de mensaje da aplicación: Este nivel se centra en el transporte de la información entorno red de la empresa. Este puede emplearse para aislar aplicaciones operacionales a partir de los datos correctos, o recolectar transacciones o mensajes y entregarlos en una sola ubicación.
- Nivel de Data Warehouse: Es la ubicación central de la información, generalmente los Data Warehouse son almacenados sobre mainframes aunque en ocasiones se encuentran en una plataforma cliente/servidor.
- Organización de los datos: este nivel es el más importante en el Data Warehouse, ya que se encarga de homogenizar la información que sea importante para el negocio capaz de solucionar las necesidades de la empresa. Una vez extraídos la información de las diferentes fuentes de datos operacionales se procede a la depuración, transformación y consolidación de los datos. En caso de que este proceso no se lleve a cobo con la eficiencia que requiere no será capaz de darle solución a los requerimientos del negocio.

Ventajas

En las empresas por lo general se precisan de vías y tecnologías capaces de mantener un alto control de los datos generados a diario. Son diversos los formatos y diversas fuentes en las que la información es almacenada, esto genera un gran número de informes y gran redundancia en los datos. Con el objetivo de solucionar este problema surge la aplicación del proceso de Inteligencia de Negocios y con esta el diseño e implementación del Data Warehouse. Este almacén de datos es capaz de mantener la información de manera homogénea y centralizada, de forma tal que resulte fácil de entender a los usuarios finales

tomar decisiones más acertadas y seguras. Podemos hacer mención de muchos de los beneficios ofrecidos por el Data Warehouse entre los que podemos enumerar los siguientes:

✓ Son capaces de almacenar grandes volúmenes de datos, lo que favorecerá a los analistas a tomar decisiones más acertadas y objetivas.

✓ Los datos que son almacenados tienen un largo plazo de vida, permitiendo así poder realizar predicciones y comparaciones de la información con el fin de lograr un mejor rendimiento en la producción.

✓ Permite obtener la información completa, correcta, consistente y accesible, de manera que los usuarios tengan un tiempo de respuesta rápida y segura de las consultas realizadas.

✓ Se establece una comunicación fiable entre los departamentos de la empresa, transformando los datos en información y la información en conocimiento.

✓ Surgen con el objetivo de tener mejor manejo y control de la información, de manera que los datos estén organizados y en un formato único brindando así un mejor entendimiento y análisis de los mismos.

1.4 Data Mart

Un Data Mart es una base de datos departamental especializada en el almacenamiento físico de los datos dirigidos a temas más específicos o a una parte del negocio, haciendo énfasis en un área en particular, departamento, tema o grupos de necesidades. Se caracterizan por poseer una estructura óptima de los datos para favorecer el análisis de la información al detalle desde todas las perspectivas que afecten a dicho departamento.

Puede ser alimentado de un Data Warehouse o integrar por sí mismo un compendio de distintas fuentes de información. Por tanto, para crear el Data Mart de un área funcional de una empresa es preciso encontrar la estructura óptima para el análisis de su información, estructura que puede estar montada sobre una base de datos OLTP (Procesamiento de Transacciones en Línea), o sobre una base de datos OLAP (Procesamiento analítico en Línea). La designación de una u otra dependerá de los datos, los requisitos y las características específicas de cada departamento.

El Mercado de Datos se puede definir como un Data Warehouse más pequeño, el cual permite el proceso de análisis de pequeñas cantidades de información útil sobre las cuales

se puedan realizar predicciones y clasificaciones, por lo que debe proporcionar las mismas funcionalidades de un Data Warehouse es por esta razón que no precisamente es más simple su implementación.

Teniendo en cuenta lo mencionado anteriormente se puede definir dos tipos de arquitectura para el diseño de un Data Mart (Bernabeu, 2010):

Arquitectura Buttom Up: Es la alternativa para que los datos extraídos de los sistemas operacionales no sean almacenados directamente en el Data Warehouse, por lo que se basa en definir primeramente los Data Marts y luego integrarlos en un Data Warehouse centralizado. Para la carga de los mismos los datos una vez extraídos deben pasar por una serie de procesos (Extracción, Transformación y Carga) que se encargarán de la limpieza y consolidación de los datos para luego distribuir la información a cada uno de ellos. Un mercado de datos puede ser implementado sin la necesidad que exista un Data Warehouse debido a las características similares que poseen.

Arquitectura Top Down: Esta arquitectura consiste en almacenar los datos en un Data Warehouse centralizado y a partir de este construir los Data Marts. El proceso es similar al anterior ya que los datos serán extraídos de las diversas fuentes, pasaran por las técnicas de integración o procesos ETL y luego serán almacenados en el almacén de datos. Partiendo de la información almacenada se distribuirá posteriormente a los Data Mart dependiendo del tema o departamento que traten.

Se pueden mencionar algunas de sus ventajas:

- Fácil acceso a los datos que se necesitan frecuentemente.
- Mejora el tiempo de respuesta del usuario final, son simples de implementar.
- Conllevan poco tiempo de construcción y puesta en marcha.
- Permiten manejar información confidencial.
- Reflejan rápidamente sus beneficios y cualidades.
- Reducen la demanda del depósito de datos.

1.5 Modelo Multidimensional.

El modelo multidimensional es el almacenamiento físico de los datos en un vector multidimensional, lo cual representa la relación entre las dimensiones. Las empresas

precisan de bases de datos para almacenar la información que genera periódicamente, las mismas pueden contar con dos modelos de datos diferentes. Estos modelos dependen del volumen de información que se requiera almacenar, se pueden clasificar en bases de datos relacionales orientadas a transacciones o bases de datos multidimensionales adecuadas para el análisis instantáneo de grandes cantidades de datos.

Estas bases de datos con fines analíticos están diseñadas por cubos OLAP (On-Line Analytical Processing o Procesamiento Analítico en Línea), los cuales tratan de organizar los datos por tablas o relaciones, pueden poseer tres ejes llamados dimensiones o pueden ser clasificados como hipercubos por contener un sin número de ejes o dimensiones. Estos cubos contienen los datos o parámetros de una determinada variable que se desee analizar, ofreciendo una vista lógica de los datos convirtiéndolos en información. Este tipo de modelamiento basado en cubos permite que las consultas sean más rápidas. Una vez creada la base de datos no se permite transformar su estructura por lo que será necesario rediseñar el cubo.

El modelo multidimensional cuenta con relaciones entre tablas de dimensiones y una tabla de hechos principal, las cuales representan indicadores como medidas o métricas, atributos y características del objeto de análisis, así como las jerarquías que se crean por las relaciones existentes entre las dimensiones.

Para lograr entender lo anteriormente expuesto es preciso tener en cuenta las siguientes definiciones acerca de los componentes del modelo multidimensional:

Tabla de hechos: Constituyen la tabla principal en el modelo dimensional, en las que se almacenan las medidas numéricas del negocio teniendo en cuenta el nivel de detalle en que es depositada la información, lo cual determina el nivel de granularidad. La clave de la tabla de hechos está compuesta por las llaves primarias de las tablas de dimensiones relacionada a ella, debido a esto son identificadas dos columnas una tabla fact y columnas key.

Tabla de dimensiones: Estas alimentan las tablas de hechos conectada a ellas, contienen un conjunto de atributos que ofrecen una perspectiva o forma de análisis sobre una medida en una tabla de hechos. Está compuesta por una llave primaria junto a un conjunto de atributos y dependiendo del diseño puede tener una llave foránea que representará la relación existente entre ella y otra dimensión. Dependiendo de la eficiencia de los atributos así será

la validez de la bases de datos. Existe también la dimensión tiempo determinada por la perspectiva histórica de la información.

Jerarquías: Establece un nivel de organización en una dimensión, representando una secuencia de atributos utilizados para simplificar los escenarios de aumento o reducción de detalle de las consultas. Surgen con las necesidades del usuario de efectuar análisis en función de ellas, ya que definen como los datos están sumarizados desde el primer nivel hasta el más detallado, estos niveles pueden designarse como nivel de agregación o simplemente nivel. Por esta razón se puede determinar que siempre existe al menos una jerarquía.

Granularidad: Representa el nivel de detalle en el que se desea almacenar la información sobre el negocio, esta puede ir desde lo más abarcador a lo más simple. La granularidad se mide según el grado de complejidad, cuando ésta soporta la lógica del negocio y las fuentes de datos será determinada como grano fino, y en caso de que contenga menor número de registro se le llama grano grueso esto significa que serán menos los atributos que tengan validez. Mientras más fino el grano que seleccionemos más segura será la información extraída.

Atributos: Son características de una propiedad considerados como metadatos, campos o criterios de análisis pertenecientes a tablas de dimensiones los que se utilizaran para la mejor comprensión de los indicadores en un cubo multidimensional.

Medidas: Son también conocidas como métricas que pueden definirse como variables que toman valores numéricos y sumarizaciones, las mismas no solamente serán representadas por sumas sino todos aquellos cálculos realizados (promedios, totales, fórmulas predefinidas u otros) que de una manera u otra permitirá a los usuarios un análisis más profundo de los hechos. En otras palabras podemos definir que son los conceptos cuantificables que nos posibilitará medir el proceso del negocio.

El Data Warehouse se caracteriza por poseer una estructura multidimensional orientada a operaciones analíticas, por esta particularidad existen varios tipos de modelamientos o esquemas los cuales posibilitan efectuar un buen diseño del mismo. Entre estos podemos mencionar el esquema en estrella, copo de nieve y constelación los cuales son detallados a continuación.

1.5.1 Esquema en Estrella

El diagrama en Estrella resulta el más sencillo para dar una visión multidimensional de los procesos medidos por métricas y sobre todo para realizar consultas frecuentes a las bases de datos multidimensionales. Debido a su implementación cuenta con una sola tabla de hechos o medidas relacionada a su vez con varias tablas de dimensiones. Una de las particularidades de este esquema es que no está normalizado lo que significa que uniones (joins) que pudieran existir son solamente desde la tabla de hechos hacia las tablas de dimensiones y debe existir una sola tabla de dimensiones para cada dimensión. Se puede decir que proporciona mejor rendimiento y velocidad al acceder a las dimensiones se encuentran individualizadas (Bernabeu, 2010).

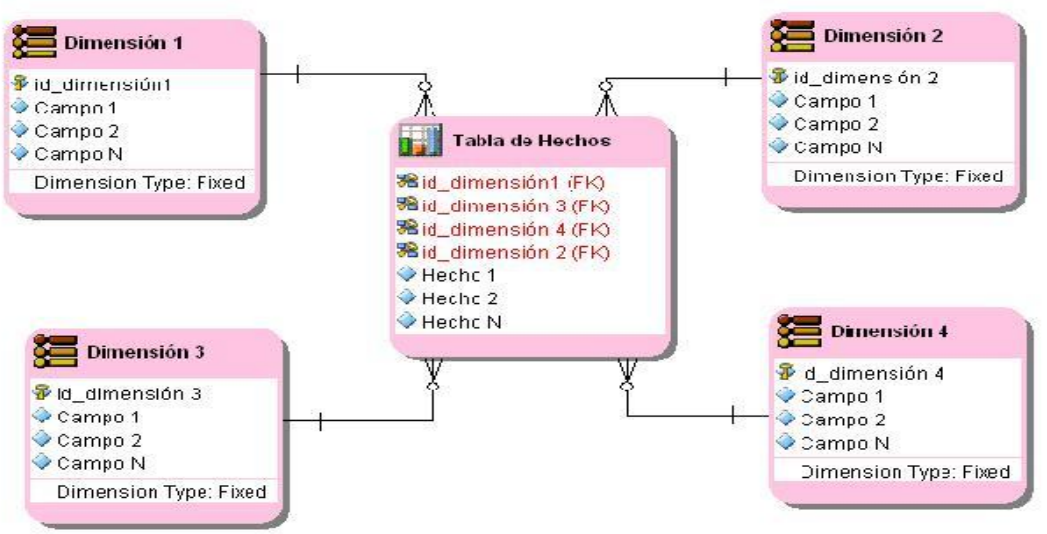


Figura 1.1: Esquema en Estrella.

1.5.2 Esquema Copo de Nieve.

Este esquema es un derivado del anterior solo que sería más complejo. Las uniones no son solamente entre la tabla de hechos principal y sus dimensiones, también se crearían joins entre las tablas de dimensiones. Esto indicaría que este modelo se encuentra normalizado lo que indica que se tratan de eliminar las redundancias, por consiguiente las consultas resultan más complejas. En este modelamiento la jerarquía está dada por la unión entre sus tablas de dimensiones.

En las aplicaciones OLAP implementadas sobre la base de datos relacionales ROLAP será fundamental tener en cuenta los cubos OLAP, los que almacenan gran cantidad de datos para luego ser analizados teniendo en cuenta varios datos. A la hora de crear la tabla donde se van a almacenar dichos datos es necesario aplicar la normalización, una vez normalizado se tendrá un modelo en copo de nieve (Bernabeu, 2010).

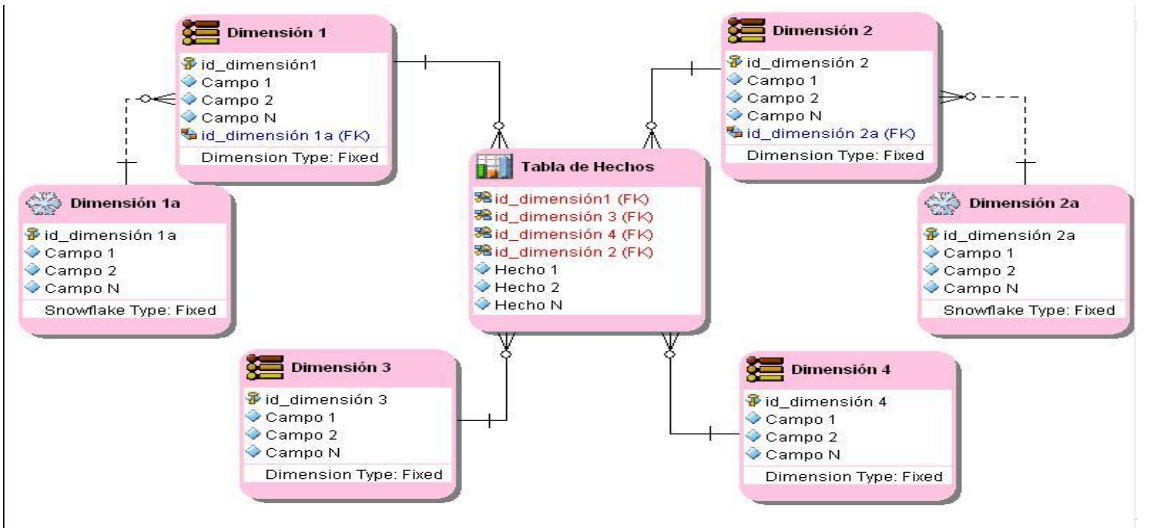


Figura 1.2: Esquema en Copo de Nieve.

1.5.3 Esquema Constelación

La diferencia de este esquema y los demás modelos esta dado en su estructura, la misma contiene más de una tabla de hechos lo que implica que sus dimensiones se dividen de manera que queden relacionadas con su respectiva tabla de hechos. Podemos referirnos que es la unión de los dos esquemas anteriores, el objetivo es aprovechar la normalización que brinda el copo de nieve y las desnormalizaciones de estrella.

Lo que trae consigo la aplicación del modelo constelación es que a medida que transcurre el tiempo su estructura va creciendo gradualmente, lo que implica que las consultas realizadas resulten ser más complejas (Bernabeu, 2010).

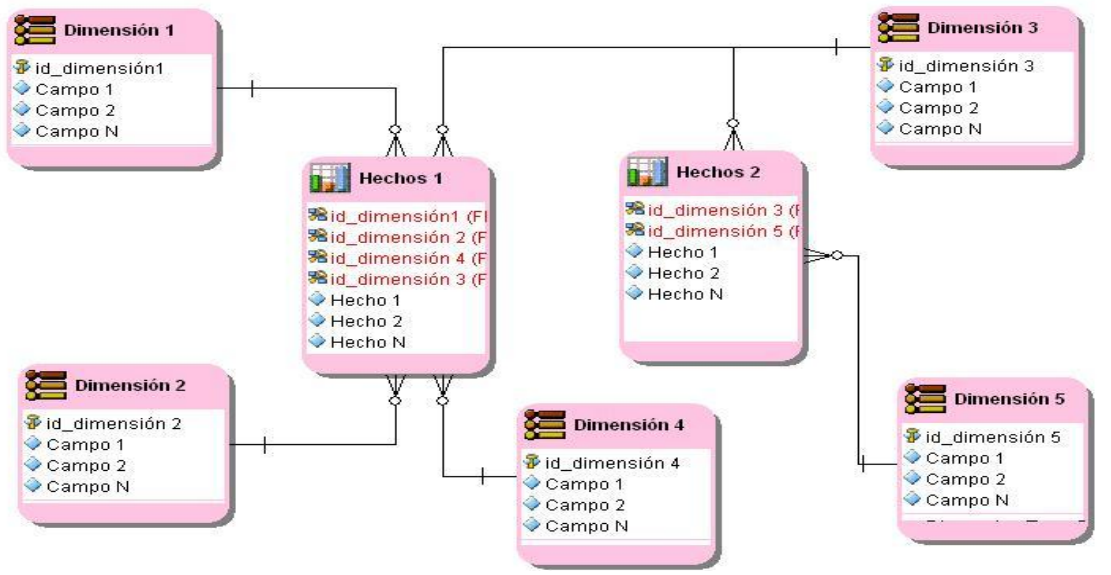


Figura 1.3: Esquema en Constelación.

Podemos concluir que un Data Warehouse bien diseñado debe ser capaz de responder a las necesidades de los usuarios ya sean actuales o futuras, el modelo en estrella no resolvería este problema ya que es un modelo muy simplificado. En el depósito de datos los cubos se multiplican, por lo que es necesario tener un entorno donde la información se encuentre detallada, integrada y actualizada, donde el modelo normalizado será el origen de los cubos y estrellas que convienen a la capa de presentación. Para realizar esta investigación se optó por el esquema en copo de nieve ya que permite obtener un modelo normalizado.

1.5.4 Tipos de implementación para el diseño del Data Warehouse.

Existen dos tipos de bases de datos que son utilizadas mundialmente en las empresas, una de estas es la base de datos relacional más adecuada para almacenar datos provenientes de transaccioneslo que se conoce como OLTP (Procesamiento de Transacciones en Línea). El otro tipo es el almacén de datos más conocido como Data Warehouse donde su diseño consiste en cubos OLAP (Procesamiento de Análisis en Línea) que contienen los datos a analizar creando así las dimensiones. Estos almacenes de datos no permiten cambios en su estructura por lo que resulta necesario rediseñar el cubo, para realizar este proceso se cuenta con diferentes métodos de creación de dichos cubos o hipercubos. Luego de varios análisis realizados muchas compañías descubrieron que se podían utilizar sistemas de administración de bases de datos relacionales para el OLAP del que se derivan los procesamientos de análisis multidimensional (MOLAP) y los relacionales (ROLAP), a

continuación detallaremos estos tipos de implementaciones (Villarreal Rosero, 2013) (Núñez Cárdenas, 2011).

- ✓ **OLAP**

El procesamiento de Análisis en Línea contiene los datos resumidos extraídos de bases de datos o sistemas transaccionales, por lo que utiliza una estructura multidimensional basada en cubos OLAP. Permitiendo tener acceso a datos agregados y organizados para su posterior análisis, se puede decir que obteniendo información generando consultas multidimensionales agiliza las consultas de grandes cantidades de datos. Los datos almacenados mediante este procesamiento tienen el fin de ser almacenados por un largo tiempo. Admite agrupar los atributos o parámetros orientado al negocio en dimensiones lo que posibilita a los usuarios efectuar consultas sin que exista la necesidad de realizar combinaciones entre todas las tablas que contenga el Data Warehouse.

- ✓ **MOLAP**

Es una base de datos propietaria multidimensional que proporciona el análisis, lo que conlleva a que se encuentre mejor implantada almacenando los datos multidimensionales. La estructura de almacenamiento de datos detalla técnicas de compactación, favoreciendo así el rendimiento del almacén de datos al reorganizar físicamente los datos. El Data Warehouse debe mantener un buen modelo relacional o esquemas, que favorezca la eficiencia de los cubos diseñados para almacenar los datos en una matriz de almacenamiento multidimensional optimizado. Algunas de estas implementaciones son propensas a la explosión de la base de datos lo que se debe al elevado número de dimensiones, pre calculados y escasos datos multidimensionales generados. Este procesamiento utiliza una arquitectura de dos niveles: bases de datos multidimensionales encargadas del manejo, acceso y obtención de los datos y el motor analítico. Se puede decir que el procesamiento analítico en línea multidimensional es una solución particular, adecuada para soluciones departamentales con menos volúmenes de información o número de dimensiones.

- ✓ **ROLAP**

El Procesamiento Analítico en Línea Relacional es otro derivado del OLAP, el cual esta implementado sobre la tecnología relacional accediendo a los datos almacenados en un Data Warehouse proporcionando análisis OLAP, por los que sus capacidades se soportan

mejor contra las bases de datos relacionales. Utiliza una arquitectura de tres niveles: la base de datos relacional que maneja los requerimientos de datos y el motor relacional que facilita la funcionalidad analítica. El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios, este al integrarse junto a niveles de presentación es el que favorece que la creación, análisis y consultas para los usuarios resulten sencillos. Este procesamiento es el más adecuado para las empresas que necesiten almacenar grandes volúmenes de información o número de dimensiones.

1.5.5 Ventajas

Podemos resumir que el modelo multidimensional brinda los siguientes beneficios:

✓ Los modelos multidimensionales proporcionan que el usuario quede aislado de los cálculos y procesos internos del almacén de datos, ya que pueden resultar bastante complejos.

✓ Contienen la tecnología de minería de datos para minar los datos y utilizar patrones descubiertos para la predicción.

✓ A medida que aumentan los datos van surgiendo nuevas interrogantes o se realizan otras acciones por parte de los usuarios. El modelo multidimensional posibilita que se escriban nuevos cambios de los datos o que existan acciones como ir a una dirección en específica (URL, Web, etc.) con el objetivo de obtener una información más ampliada.

✓ Se puede controlar el acceso al modelo multidimensional ya que proporciona una seguridad basada en funciones o lo que es lo mismo conceder permisos a las funciones donde se incluyan los usuarios como miembro de cada función.

✓ Proporciona consultas de alto rendimiento que admiten análisis iterativo, incluso con grandes cantidades de datos.

1.6 Proceso de Extracción, Transformación y Carga, proceso de integración de datos.

En la Inteligencia de Negocios es preciso adoptar técnicas y herramientas capaces de organizar los flujos de datos entre diferentes sistemas en una organización. Estas herramientas las brinda el proceso conocido como ETL (Extracción, Transformación y Carga de datos) las que están encargadas de obtener los datos de diversas fuentes, transformarlos, limpiarlos, consolidarlos y luego almacenarlos en un solo destino. Este

sistema de herramientas tiene varias etapas como ya su nombre lo indica las que especificaremos a continuación.

1.6.1 Función de Extracción.

Este primer paso consiste en seleccionar las fuentes de datos heterogéneas de donde será recuperada la información. Estas fuentes pueden ser desde archivos en diversos formatos (texto, hojas de cálculo, etc…) provenientes de sistemas transaccionales y los formatos de los distintos gestores de bases de datos orientadas a sistemas operacionales.

Las herramientas destinadas a la extracción de datos permiten varios formatos así como documentar el proceso, validar correctamente los datos y lograr una homogenización de los mismos. Deben tener la capacidad de mapear los diferentes orígenes de los datos, los distintos modelos ya sean relacionales o no relacionales, ficheros y formatos, formato de mensaje estándar, colas de mensajerías, etc.

Al realizar este primer proceso se crea un modo de almacenamiento intermedio donde se permita incorporar formatos no estructurados, con el objetivo de detectar los posibles errores y aplicar acciones sobre los mismos, además debe permitir múltiples frecuencias de actualización.

1.6.2 Función de Transformación.

Una vez realizado el proceso anterior, los datos deben ser procesados de manera tal que se les realice las transformaciones requeridas por la empresa. Los datos al encontrarse en dicho almacenamiento intermedio se realizan las operaciones de transformación como eliminación de inconsistencias, inserción de campos y tablas, así como la depuración, limpieza y consolidación de los datos, agregación y transformación de codificaciones, etc. Podemos mencionar algunas transformaciones realizadas con frecuencia:

- ✓ Estandarización de formatos: Mediante este proceso se puede obtener un formato o tipo de dato estándar, sin importar que se encuentren almacenados en diversos formatos y tipos de datos en distintas bases de datos transaccionales.
- ✓ Estandarización de nombramiento: Se hace transformaciones sobre los datos y no sobre el formato de los mismos, a partir de este es obtenida una forma estándar de nombrar un dato.

✓ Cuando algunos registros son repetidos con diferentes nombres se aplica una transformación para lograr fusionar estos registros y convertirlos en uno solo.

1.6.3 Función de Carga.

Una vez extraídos y transformados los datos se procede a realizar el almacenamiento físico del Data Warehouse, el cual dispone de dos tareas principales Carga Inicial y Actualización o Mantenimiento periódico. Donde la carga inicial se centra en cargar las tablas de dimensiones y las tablas de hechos, así como reconstruir índices. Para lograr dar un mantenimiento periódico al almacén de datos se realizan ingresos de nuevos datos lo cual provoca que el nivel de granularidad vaya creciendo de manera gradual, por lo tanto se debe tener en cuenta no afectar la estructura del Data Warehouse. Para lograr introducir nuevos datos en el Data Warehouse es preciso realizar verificaciones de los datos para de esta manera estar seguros que no han ocurrido nuevos cambios en los sistemas operacionales de donde es extraída la información. Se puede definir que los datos que serán ingresados en el almacén de datos son aquellos a los cuales se les ha realizado el proceso de transformación y se encuentran en el almacenamiento intermedio, así como los que se encuentran en los OLTP que mantienen correspondencia directa con el Data Warehouse [6].

Podemos resumir que para llevar a cabo las diferentes funciones del proceso ETL se deben tener en cuenta varios pasos reglamentarios, los cuales permitirán mantener un eficiente mantenimiento de la información que posteriormente será almacenada en el depósito de datos. Primeramente serán extraídos los datos principales de las diversas fuentes operacionales, los que se mantendrán en un almacenamiento intermedio. Una vez extraídos los datos serán consolidados, integrados y transformados en caso de requerirlo para de esta manera evitar las inconsistencias existentes. Por último será poblado el Data Warehouse con los datos encontrados en el almacenamiento intermedio.

1.7 Estado del arte.

A partir de los estudios realizados se encontraron muchos sistemas basados en el de desarrollo de un Data Warehouse, pero para el control de los servicios alimenticios solo fue identificado el sistema relacional siguiente:

En la Empresa de Servicios a la Unión del Níquel se cuenta con una herramienta nombrada SADESP v 1.0 donde se almacenan todas las actividades de los servicios gastronómicos, de hospedaje y transporte. Este sistema es utilizado en todas las unidades de la empresa, posee un módulo de análisis que es utilizado para gestionar el control de la producción

donde es consolidada y procesada la información. Este sistema transaccional se basa en las tareas de control de las actividades no se realizan análisis.

1.8 Ambiente de Desarrollo.

La Inteligencia de Negocios es una alternativa para aquellas empresas que requieran de gran almacenamiento y procesamiento de la información que respondan a las necesidades del negocio. Para llevar a cabo este proceso iterativo son agrupadas técnicas y herramientas capaces de extraer datos de diversas fuentes, entenderlos, limpiarlos e integrarlos de manera que se lograra un almacenamiento homogéneo y organizado de la información como el Data Warehouse.

Para cumplir con las condiciones de esta investigación, se realizaron estudios de varias herramientas capaces de alcanzar con eficiencia y rendimiento los requerimientos de la empresa. Para llevar a cabo el diseño metodológico se escogió la Metodología HEFESTO demostró ser la más eficiente y completa para efectuar un buen proceso del BI, para crear el Data Warehouse se aplicó la herramienta de diseño y modelado de almacenes Embarcadero ER/Studio 8.0, como gestor de bases de datos MySQL 5.0 por la simplicidad y considerable rendimiento que nos brinda. Así como Pentaho Data Integration (PDI) 4.0 para realizar la integración de los datos y Pentaho BI Server 4.0 como servidor de BI.

1.9 Justificación Metodológica: Metodología HEFESTO

Con el objetivo de realizar un buen diseño metodológico se estudiaron varias metodologías existentes para entender cuál de ellas sería la más apropiada para el desarrollo de esta investigación, podemos mencionar entre las metodologías analizadas la de Ralph Kimball la cual se centra en especificar que un Data Warehouse no es más que la unión de todos los Data Marts la cual se enfoca en la realización de proyectos pequeños. Además se encuentra la metodología de Innmon la misma es comúnmente usada en sistemas más complejos, según su ideología menciona que los Data Marts son subconjuntos del Data Warehouse que tienen la finalidad de cubrir las necesidades de análisis de cada departamento en particular, esta metodología al tener un enfoque global imposibilita realizar proyectos más sencillos.

En cambio la Metodología HEFESTO resuelve el problema de aplicar metodologías que requieran de fases muy extensas las cuales ocupan un tiempo excesivo en el negocio. Esta

metodología no es más que la comparación de otras metodologías, fue creada por Bernabeu Ricardo Darío y su nombre fue inspirado por el dios griego del fuego y la construcción, la misma se enfoca principalmente en el análisis de los requerimientos de la empresa.

Bernabeu planteó que la idea principal de esta metodología está fundamentada para comprender cada paso que se realizará, con el objetivo de emplear métodos paso a paso sin tener conocimiento de cómo, ni por qué son realizados. Para la aplicación de esta metodología se debe tener en cuenta varios pasos o fases que la identifican (Bernabeu, 2010):

Figura 1.4: Pasos de la Metodología HEFESTO.

La metodología HEFESTO cuenta con una serie de características mencionadas a continuación las que fueron extraídas del libro HEFESTO II:

- ✓ Los objetivos y resultados esperados en cada fase se distinguen fácilmente y son sencillos de comprender.
- ✓ Se basa en los requerimientos de los usuarios, por lo cual su estructura es capaz de adaptarse con facilidad y rapidez ante los cambios en el negocio.
- ✓ Reduce la resistencia al cambio, ya que involucra a los usuarios finales en cada etapa para que tome decisiones respecto al comportamiento y funciones del Data Warehouse.
- ✓ Utiliza modelos conceptuales y lógicos, los cuales son sencillos de interpretar y analizar.
- ✓ Es independiente del tipo de ciclo de vida que se emplee para contener la metodología.
- ✓ Es independiente de las herramientas que se utilicen para su implementación.
- ✓ Es independiente de las estructuras físicas que contengan el DW y de su respectiva distribución.
- ✓ Cuando se culmina con una fase, los resultados obtenidos se convierten en el punto de partida para llevar a cabo el paso siguiente.
- ✓ Se aplica tanto para Data Warehouse como para Data Mart.

1.9.1 Herramienta de diseño y Modelado de Almacenes de Datos: MySQL Workbench 5.1

MySQL Workbench es una aplicación para el diseño y documentación de bases de datos se puede decir que es la sucesora de la aplicación DBDesigner4, pensada para ser usada con el sistema de gestión de bases de datos MySQL. Existen dos versiones del producto, una es open source y la otra es una versión comercial. Evidentemente, la versión comercial proporciona algunas funcionalidades que pueden resultar de interés en algún ámbito, aunque la versión open source es más que suficiente para la realización de la práctica.

La herramienta podría usarse para realizar un diagrama EER, y esa es su principal función: primero diseñar el diagrama EER, implementándolo sobre la herramienta y a partir de él obtener el diagrama del esquema relacional y también las sentencias de creación de tablas, vistas e índices de manera automática. Se puede visualizar las tablas y sus campos, incluyendo las claves primarias, las claves foráneas y a quienes referencian, y todo ello mediante una interface gráfica que permitirá además generar las sentencias DDL de

creación de tablas e índices las cuales con algunas modificaciones, podrán adaptarse a lo solicitado en la práctica (MySql, 2009).

1.9.2 Herramienta realización de los procesos ETL: Pentaho Data Integration PDI 4.0

Esta herramienta también conocida como Kettle (KDE Extraction, Transportation, Transformation and Loading Environment) fue diseñada por Matt Casters en el año 2001, con el transcurso del tiempo obtuvo un mayor desarrollo y salieron nuevas versiones hasta que se incluyó en la ámbito de Pentaho y adquirió el nombre de Pentaho Data Integration ya que se puede categorizar como una herramienta ETL. Se pueden mencionar sus componentes principales los cuales no deben entenderse como programas sino un conjunto de metadatos en XML, estos metadatos le indican al motor PDI las formas de realizar acciones (Espinosa, 2010):

- ✓ Transformaciones: se basa en las colecciones de steps o pasos enlazados a través de saltos o hops. Estos saltos constituyen el elemento por donde circula la información entre los diferentes pasos, cada salida de estos pasos constituyen un registro. Los elementos a tratar y sus comportamiento son determinados en una ventana de configuración que tiene cada steps.
- ✓ Trabajos (Jobs): consiste en una colección de transformaciones o steps orientados a realizar una tarea determinada. En este componente los pasos son específicos, lo que muestra una tarea del proceso ETL y los saltos representan el orden de ejecución de cada uno de los componentes.

PDI está constituido por diferentes herramientas o aplicaciones cada una con funciones específicas entre las que podemos nombrar:

- ✓ Spoon: Es la principal herramienta de PDI mediante se puede realizar el diseño de las transformaciones y los trabajos. Permite la construcción y validación de los procesos ETL.
- ✓ Pan: Permite la ejecución de las transformaciones diseñadas en Spoon por línea de comandos, lo que nos proporciona preparar la ejecución mediante scripts.
- ✓ Kitchen: Posibilita ejecutar los trabajos por línea de comandos al igual que Pan.
- ✓ Carte: Es un servidor que posibilita la ejecución remota de transformaciones y trabajos.

1.9.3 Herramienta para Gestión de Base de Datos: MySQL 5.0

MySQL es una herramienta para la Gestión de Base de Datos relacional que ha alcanzado gran popularidad debido a su simplicidad y alto rendimiento, facilita operar todo el funcionamiento de una base de datos de manera visual. Es una herramienta que se encuentra en constante evolución, desarrollada por IBM como lenguaje de consulta, la misma ha llegado a ser estándar para el intercambio de datos. Cuenta con una serie de filtros de acceso a los datos, tales como EDA/SQL con la finalidad de acceder a casi todos los sistemas de gestión de base de datos y sistemas de operacionales siendo relacionales o no.

MySQL permite la introducción de sentencias SQL con el objetivo de trabajar en la base de datos, así como ejecución de comandos propios del SGBD para obtener información o ejecutar operaciones de administración. Permite el procesamiento por lotes de sentencias implícitas en un archivo de texto. Dispone de un amplio conjunto de tipos de datos los que nos permiten diseñar una tabla, estos datos pueden ser numéricos, cadenas de caracteres, fechas y horas.

Posibilita la creación de tablas temporales las que se pueden ver en la sesión abierta y guardar datos entre consultas, internamente también genera tablas temporales para resolver consultas. La administración y seguridad de esta herramienta se plantea sobre un esquema de privilegios y usuarios, con la finalidad de proteger el equipo servidor MySQL encerrando la asignación de contraseñas y la información de perfil de usuario (Casillas Santillán, y otros, 2010).

1.9.4 Herramienta Servidor de BI: Pentaho BI Server 4.0

Pentaho es una herramienta poderosa y la alternativa de software libre para aplicar el proceso de Inteligencia de Negocios, ya que nos permite realizar reportes, análisis OLAP, integración de los datos entre otros. Estas funcionalidades son muy necesarias actualmente ya que todas las empresas usan los informes de una manera u otra.

Pentaho Reporting

Los informes de Pentaho permiten acceder y formatear fácilmente a la información, como también distribuirlas a los empleados, clientes y socios de la organización. Estos formatos pueden encontrarse en formatos estándar como Excel, Html o Pdf. Entre las ventajas que proporciona podemos englobar las siguientes:

- Para los expertos en informes

✓ Permite acceder a fuentes de datos heterogéneas.

✓ Capacidad de integración en aplicaciones o portales.

✓ Definición modular de informes (presentación y consulta)

- Para diseño de informes flexibles

✓ Entorno de diseño gráfico y capacidad de uso templates.

✓ Acceso a datos relacionales, OLAP y XML

Pentaho Data Integration

Los datos generados en las empresas son almacenados en diversas fuentes los cual atenta contra la homogenización de los mismos. Para solucionar este problema Pentaho posee una poderosa herramienta de extracción, transformación y carga que nos brinda:

✓ Un entorno gráfico de desarrollo, así como el uso de las tecnologías estándar: Java, XML, JavaScript, etc.

✓ Mantiene una multiplataforma basada en dos tipos de objetos: Transformaciones y Trabajos.

Pentaho Analysis

Los análisis de Pentaho permiten consultar, explorar y analizar la información de la empresa y así adquirir conocimientos necesarios y específicos que favorecerá tomar decisiones acertadas y óptimas esto posibilita:

✓ Obtener una vista dimensional de los datos.

✓ Interactuar con alto rendimiento haciendo uso de las tecnologías perfeccionadas para dar una respuesta interactiva rápida.

Pentaho Dashboard

Pentaho provee la información necesaria a entender para mejorar el rendimiento organizacional, mostrando las métricas en una interfaz visual atractiva e intuitiva por lo tanto posibilita:

✓ Identificar las Métricas o Indicadores claves.

✓ Investiga detalles subyacentes y seguimientos de excepciones basadas en reglas del negocio.

Pentaho Data Mining

Pentaho es usado para alcanzar una mejor perspectiva del negocio y de esta manera mejorar el rendimiento futuro a través de analíticas predictivas. Lo anteriormente descrito posibilita:

✓ Descubrir patrones ocultos y correlaciónales en los datos, así como prevenir eventos futuros basados en patrones históricos.

✓ Contar con la tecnología de un poderoso motor de Data Mining, herramientas gráficas, seguridad, servicios web, repositorios y definiciones basadas en XML, rendimiento y escalabilidad.

1.10 Conclusiones parciales

En este capítulo consolidamos y adquirimos conocimientos acerca de los principales conceptos del proceso de Inteligencia de Negocios. Además se realizó un análisis contundente sobre todos los aspectos a tener en cuenta para el diseño de un Data Warehouse, así como sus características, estructura, entre otras cosas. Estudiamos como es procesada la información para lograr obtenerla en un solo lugar en una manera centralizada y organizada, para ello tuvimos en cuenta el estudio de las técnicas de integración, la metodología más segura y eficiente aplicada, así como las herramientas que proporcionaran el análisis, diseño e implementación del Data Warehouse. Lo antes mencionado favorecerá a obtener un almacenamiento de la información de manera segura y rápida, favoreciendo el análisis al soporte de la toma de decisiones en el negocio.

CAPÍTULO II: ANÁLISIS Y DISEÑO DEL DATA MART.

2.1 Introducción.

En el presente capítulo alcanzaremos un diseño lógico del mercado de datos obteniendo de esta manera una visión clara del almacenamiento de la información. Para lograr con eficiencia este proyecto se tendrá en cuenta cada paso o fase estudiada en la metodología escogida, de esta manera obtendremos los indicadores y las perspectivas identificadas a partir de las entrevistas y preguntas generadas por medio de los especialistas. Obtendremos el modelo conceptual el cual favorecerá el modelamiento de los datos y para el mismo tendremos en cuenta que esquema será el más apropiado.

2.2 Análisis de los requerimientos.

Primeramente se deben identificar los requerimientos del negocio a partir de las preguntas generadas por los usuarios, una vez identificados los requerimientos se definen los indicadores o perspectivas (dimensiones) que facilitaran el análisis para el soporte a la toma de decisiones y el diseño del Data Warehouse. Los usuarios podrán visualizar el resultado mediante al modelo conceptual que se diseñará. La investigación será guiada hacia un desarrollo a partir de la identificación de información de alto nivel para de esta manera cumplir los objetivos y estrategias trazadas por la empresa. Es necesario formular preguntas complejas sobre el negocio en las que influyan variables de análisis, lo cual posibilitará el análisis profundo de la información en diferentes perspectivas.

2.2.1 Identificar preguntas.

Con el objetivo de obtener una visión profunda sobre las necesidades de la empresa, fue preciso conversar con los especialistas encargados de mantener un control absoluto de los servicios gastronómicos. A partir de dichas entrevistas se pudo con conformar una serie de preguntas que ayudan a conocer los requerimientos del negocio.

- ✓ Se desea conocer la cantidad de servicios de alimentación solicitados por clientes en un tiempo determinado.
- ✓ Se desea conocer la cantidad de servicios de alimentación elaborados por unidades en un tiempo determinado.
- ✓ Se desea conocer la cantidad de servicios de alimentación prestados por unidades y por clientes tiempo determinado.

- ✓ Se desea conocer la cantidad de recursos utilizados por servicios y por clientes tiempo determinado.
- ✓ Se desea conocer la cantidad de recursos utilizados por servicios, unidades y clientes tiempo determinado.
- ✓ Se desea conocer la relación de clientes que reciben tiempo determinado los mismos productos en sus servicios.
- ✓ Se desea conocer la relación de servicios prestados por unidades y clientes.
- ✓ Se desea conocer en qué tiempo determinado se prestan más servicios agrupados por tipos.
- ✓ Se desea conocer el informe de un tiempo determinado por total de servicios prestados agrupados por tipos por unidades.
- ✓ Se desea conocer el informe de un tiempo determinado por total de servicios prestados agrupados por tipo por cliente.

2.2.2 Identificar indicadores y perspectivas.

Una vez comprendidas las necesidades del negocio a través de las preguntas obtenidas a partir de los encuentros con los especialistas, se procede a identificar los indicadores y perspectivas que formarán la base principal del diseño del mercado de datos.

Cantidad de servicios de alimentación solicitados por clientes por un tiempo determinado.

Cantidad de servicios de alimentación elaborados por unidades por un tiempo determinado.

Cantidad de servicios de alimentación prestados por unidades y por clientes por un tiempo determinado.

Cantidad de recursos utilizados por servicios y por clientes en un tiempo determinado.

Cantidad de recursos utilizados por servicios, unidades y clientes por un tiempo determinado.

Relación de clientes que reciben en un tiempo determinado los mismos productos en sus servicios.

Relación de servicios prestados por unidades y clientes en un tiempo determinado.

En qué tiempo determinado se prestan más servicios agrupados por tipos.

Informe de un tiempo determinado por total de servicios prestados agrupados por tipos por unidades.

Informe de un tiempo determinado por total de servicios prestados agrupados por tipo por cliente.

Indicadores	Perspectivas
Cantidad de servicios solicitados.	Clientes.
Cantidad de servicios elaborados.	Tiempo.
Cantidad de servicios prestados.	Unidades.
Cantidad de recursos.	Servicios.
Relación de clientes.	Recursos.
Relación de servicios.	Productos.
Total de servicios prestados por tipo.	Servicios por tipo.

2.2.3 Modelo conceptual.

Para lograr definir el alcance que tendrá el mercado de datos se realiza el modelo conceptual a partir de los indicadores y perspectivas obtenidas a causa de la descomposición realizada de las preguntas generadas por los especialistas, de esta manera se podrá entender con más claridad cómo estará diseñado el Data Mart.

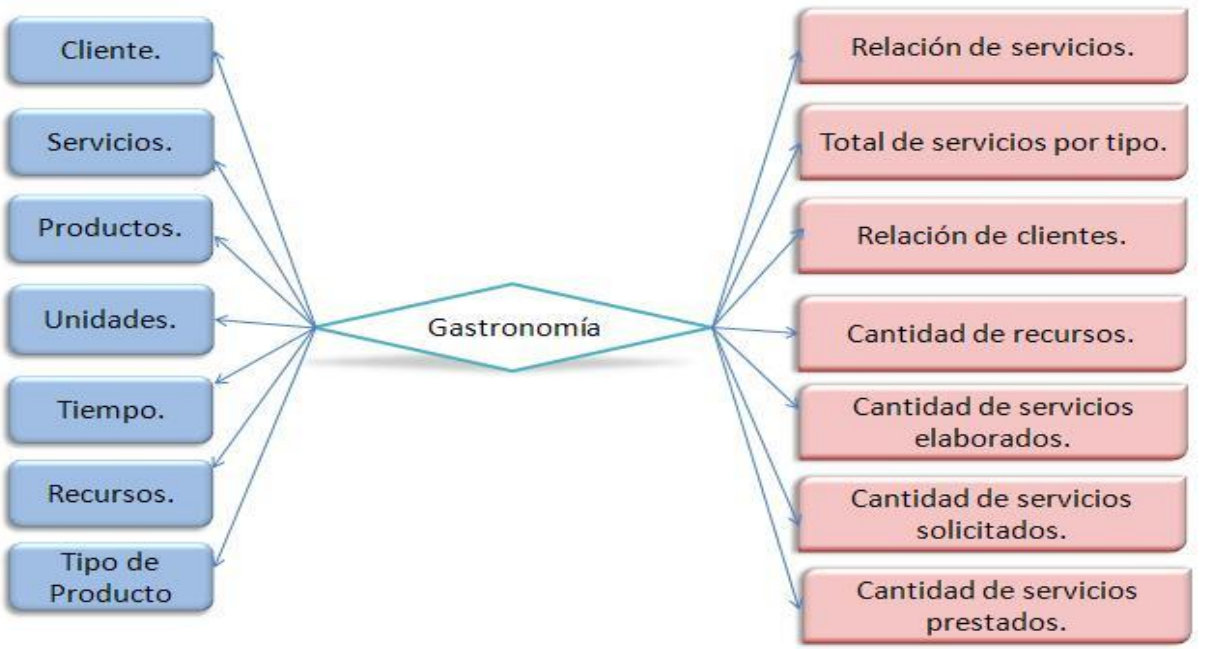


Figura 2.1: Modelo Conceptual.

2.3 Análisis de los OLTP.

Luego de culminar con esta fase se procede al análisis de los OLTP (Procesamiento Transaccional en Línea), lo cual permitirá tener una idea de cómo identificar los indicadores. Para lograr lo mencionado anteriormente es necesario adquirir conocimiento de los elementos o hechos que proporcionará realizar cálculos y los tipos de operaciones que se debe aplicar a los hechos para obtener el valor final. Se realizará un mapeo de los sistemas transaccionales y el modelo conceptual para establecer e identificar las correspondencias existentes entre ellos. Los usuarios serán los encargados de definir los campos que integrarán cada dimensión lo que representará el nivel de granularidad, una vez obtenida toda la información necesaria se realizará una ampliación al modelo conceptual.

2.3.1 Conformar indicadores.

Lo que nos permite realizar este paso es tener en cuenta los cálculos que deberán realizarse para obtener los indicadores identificados, a continuación se planteará cada indicador con su respectivo cálculo:

- Cantidad de servicios alimenticios solicitados es la sumatoria de los servicios que han sido solicitados por un cliente en particular.
- Cantidad de servicios alimenticios prestados es la sumatoria de los servicios que han sido prestado por una unidad específica.
- Cantidad de servicios alimenticios elaborados es la sumatoria de los servicios que ha elaborado una unidad específica.
- Cantidad de recursos es la sumatoria de los recursos utilizados por un servicio, una unidad y un cliente en particular.
- Relación de clientes representa la cantidad de clientes que recibe un mismo producto en específico.
- Total de tipos de servicios es la sumatoria de todos los servicios prestados por tipo por una unidad específica.

2.3.2 Establecer correspondencias.

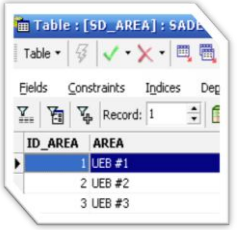
Una vez conocidos los cálculos de cada indicador se procede a establecer las correspondencias existentes en los sistemas operacionales de donde será extraída la información. Para lograr lo anteriormente expuesto es preciso examinar que los elementos

que contiene el modelo conceptual se correspondan a los sistemas operacionales. Las correspondencias existentes serian:

La tabla SD_P_PROD se encuentra relacionada con la perspectiva “Tiempo”.

La tabla SD_Cliente se encuentra relacionada con la perspectiva “Cliente”.

La tabla SD_AREA se encuentra relacionada con perspectiva “Unidades”.

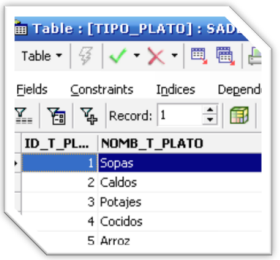


El campo SD_IND_PROD se encuentra relacionado con la perspectiva “Servicios”.

La tabla SD_IND_PROD se encuentra relacionada con la perspectiva “Producto”.

La tabla SD_INGREDIENTES se encuentra relacionada con la perspectiva “Producto”.

La tabla TIPO_PLATO se encuentra relacionada con la perspectiva “Tipo de Producto”.



2.3.3 Nivel de granularidad.

Analizadas ya las correspondencias se debe analizar cuáles serán los campos que integrarán cada perspectiva, para lograr esto es necesario consultar con los especialistas cuales serían los campos que ayudarán a consultar los indicadores de los mismos se obtuvieron los siguientes:

Perspectiva “Clientes”:

- ✓ id_cliente: representa la clave primaria de la perspectiva cliente.
- ✓ nombre_cliente: representa el nombre del cliente.
- ✓ siglas_cliente: representa las iniciales del nombre el cliente
- ✓ dirección_cliente: representa la dirección del cliente.

Perspectiva “Tiempo”:

- ✓ id_tiempo: representa la clave primaria de la perspectiva Tiempo.
- ✓ Año.
- ✓ Semestre.
- ✓ Trimestre.
- ✓ Mes.
- ✓ Quincena.
- ✓ Semana.

- ✓ Día.

Perspectiva “Unidades”:

- ✓ id unidades: representa la clave primaria de la perspectiva Unidades.
- ✓ nombre_unidad: representa el nombre de la unidad gastronómica.

Perspectiva “Servicios”:

- ✓ id_servicios: representa la clave primaria de la perspectiva Servicios.
- ✓ Id_servicios_tipo: representa la clave primaria de la perspectiva Tipo de Servicios que es obtenida por la unión con la perspectiva Servicios.
- ✓ descripción_servicio: representa la descripción del servicio.

Perspectiva “Producto”:

- ✓ id_producto: representa la clave primaria de la perspectiva Producto.
- ✓ nombre_producto: representa el nombre del producto.
- ✓ precio_cuc: representa el precio del producto en la moneda convertible.
- ✓ precio_cup: representa el precio del producto en la moneda nacional.
- ✓ precio_total: representa el precio total del producto.

Perspectiva “Recurso”:

- ✓ id_recursos: representa la clave primaria de la perspectiva Recurso.
- ✓ nombre_recurso: representa el nombre del recurso utilizado.

Perspectiva “Tipo Producto”:

- ✓ id_tipo_producto: representa la clave primaria de la perspectiva Tipo Producto.
- ✓ nombre_tipo_producto: representa el nombre del tipo de plato.

2.3.4 Modelo conceptual Ampliado.

Una vez realizado cada paso anterior con la mayor exactitud y eficiencia se debe realizar el modelo conceptual ampliado. Este modelo englobará cada perspectiva con los campos pertenecientes a la misma, y por otro lado se visualizaran los indicadores y los cálculos que les corresponde para lograr consultarlos, este modelo de manera ampliada podrá ser observado a continuación:

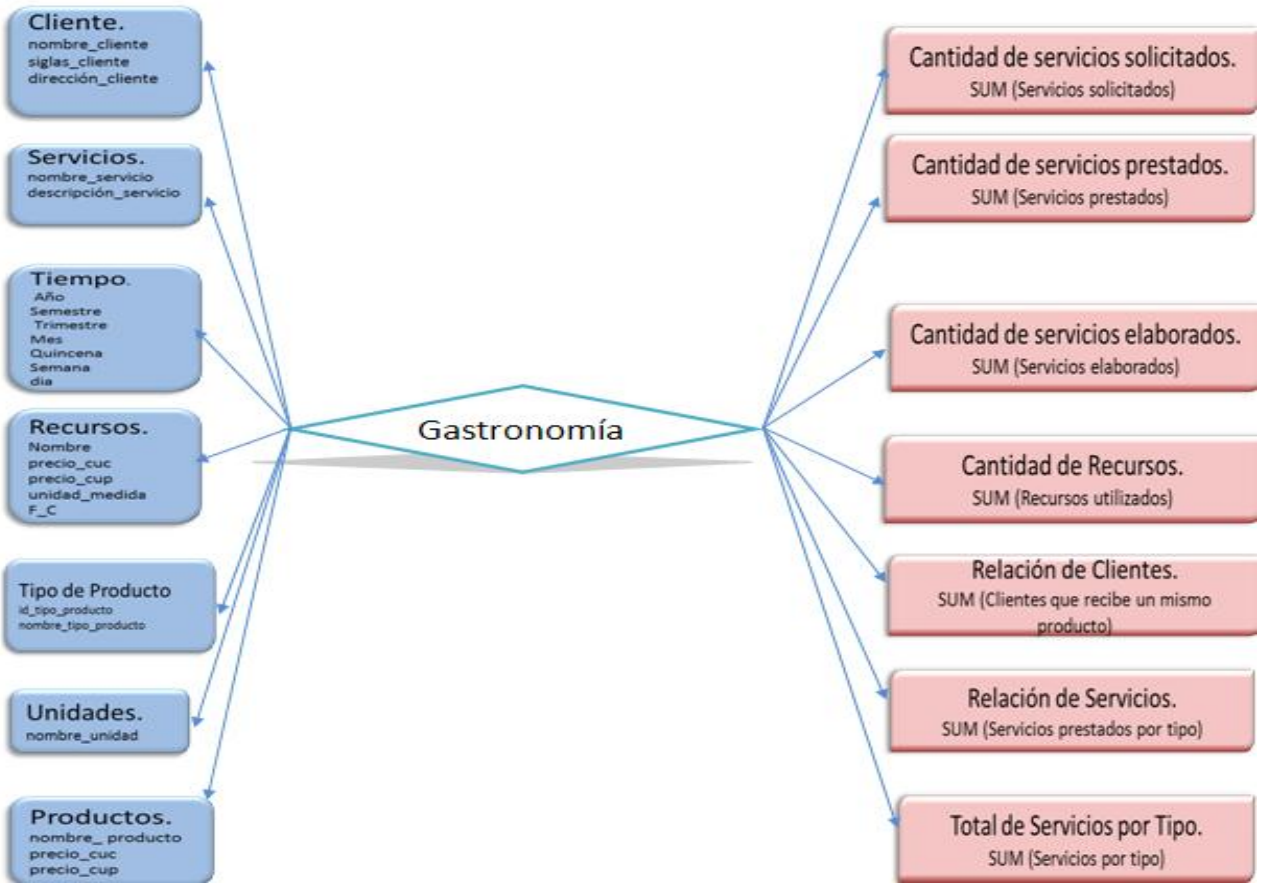


Figura 2.2: Modelo Conceptual Ampliado.

2.4 Modelo lógico del Almacén de Datos.

Una vez ampliado el modelo conceptual se inicia el modelo lógico del Data Warehouse, para lo que es necesario tener conocimiento del tipo de modelamiento aplicado en el diseño de su estructura. Estos modelos pueden diseñados mediante los esquemas en estrella, copo de nieve o constelación cada uno con su particularidad, ya definido el esquema serán diseñadas las tablas de dimensiones y las tablas de hechos identificándolas mediante un nombre, una clave primaria y definiendo los valores que suplirán los campos de las mismas, la clave primaria en la tabla de hechos resultará de las combinaciones de las claves primarias de las dimensiones relacionadas a esta.

2.4.1 Tipo de modelo lógico.

Partiendo de los problemas presentados por la empresa y las perspectivas e indicadores ya identificadas podemos realizar el diseño del Data Mart, para esto nos basamos en la estructura multidimensional que presenta y los esquemas que lo caracteriza.

Descartamos que el mercado de datos se encontrará diseñado sobre la base del esquema de modelamiento en copo de nieve, presentando así una tabla de hechos y varias tablas dimensiones relacionadas a ella. La característica principal de este esquema es la relación existente entre la tabla de hechos y las dimensiones pero a su vez estas dimensiones relacionadas con otras, esto implican las normalizaciones que precisan algunas de las dimensiones.

2.4.2 Tablas de dimensiones.

A partir de las perspectivas definidas se obtuvieron las dimensiones, significando que cada perspectiva será una dimensión específica. Se obtuvieron siete dimensiones con los campos o atributos que la componen las mismas serán especificadas a continuación.

Figura 2.3: Tablas de Dimensiones.

2.4.3 Tablas de Hechos.

La tabla de Hechos definida englobará la clave primaria de cada dimensión definida.

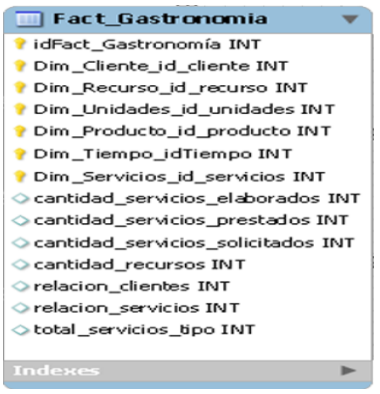


Figura 2.4: Tabla de Hechos.

2.4.4 Uniones

Luego de haber definido las dimensiones y la tabla de hecho principal es necesario realizar las uniones correspondientes entre ellas, las mismas pueden ser apreciadas en la siguiente tabla:

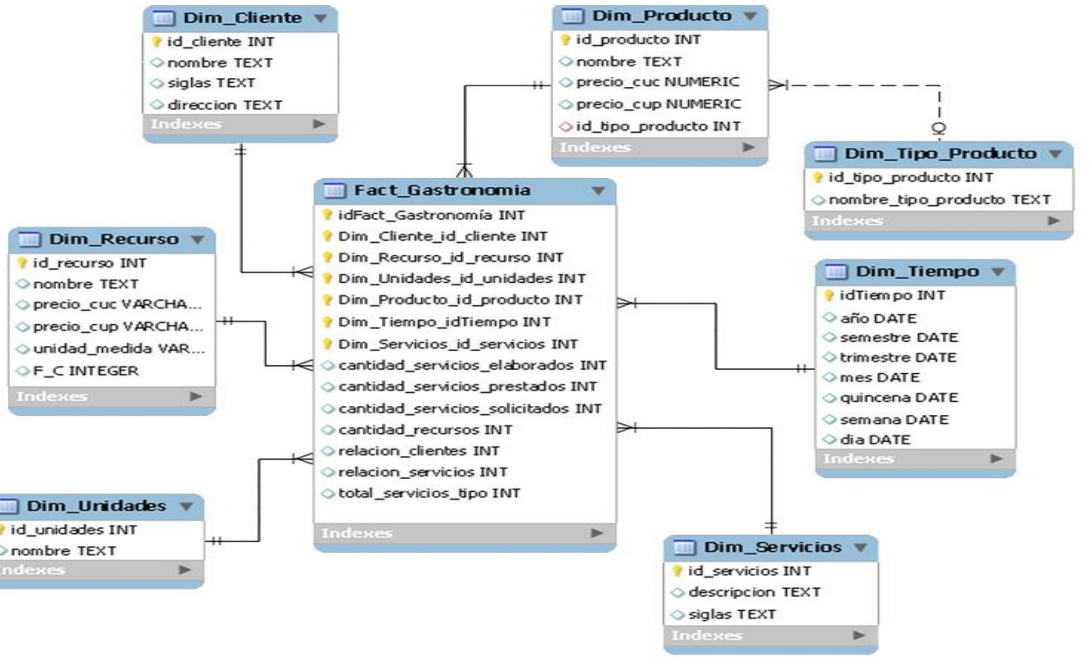


Figura 2.5: Diseño del Data Mart.

2.5 Modelo de Arquitectura del Mercado de Datos.

Para obtener la arquitectura del mercado de datos se extrajo la información de los diferentes sistemas operacionales en este caso la base de datos relacionales llamada Despacho y archivos Excel, estos datos se visualizaron a través de los sistemas gestor y administración de base de datos IBExpert y Firebird. Luego estos datos pasarán por una serie de técnicas o procesos encargados de depurarlos y consolidarlos y por último es poblado el Data Mart teniendo empleando el gestor de base de datos EMS MySQL(Figura 2.6).

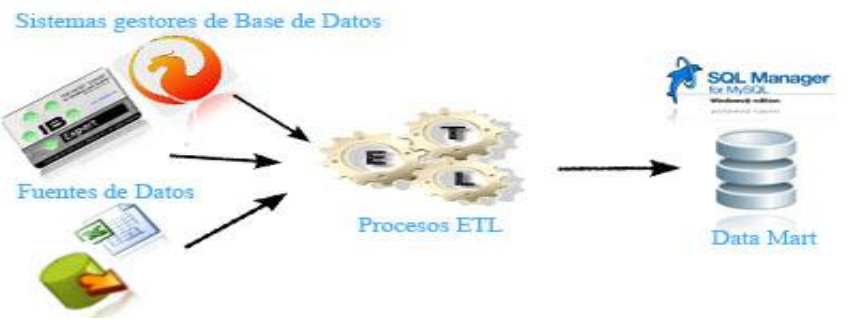


Figura 2.6: Modelo de Arquitectura del Mercado de Datos.

2.6 Conclusiones Parciales

Para culminar con la integración de los datos debe ser cargado inicialmente el Data Warehouse agregando datos al modelo lógico ya construido. Es necesario mantener un sistema de tareas o técnicas como la depuración y verificación de los datos, aplicar el proceso ETL, etc., así como establecer restricciones y condiciones para que los datos utilizados sean los fundamentales. Una vez poblado el Data Warehouse deben ser creadas estrategias de actualización o mantenimiento periódico de los datos lo que conlleva que se debe mantener el proceso de limpieza de datos, los sistemas ETL y consolidación de los datos.

CAPÍTULO III: Implementación y Prueba.

3.1 Introducción.

Este capítulo mostrará de manera clara los pasos llevados a cabo en la última fase de la metodología empleada como es la carga inicial y actualización del Data Mart. Una vez construido del modelo de datos es preciso poblarlo para lo cual es preciso aplicar una serie de procesos que permitan depurar y mejorar la eficiencia de los datos conocidos como ETL. Luego se procederá a conformar el cubo multidimensional, así como el análisis de los tiempos de carga y para culminar las pruebas de integridad de los de los datos y aceptación.

3.2 Implementación del Data Mart.

Para llevar a cabo la implementación primeramente debe ser poblado el mercado de datos. Para llevar a cabo este proceso se debe aplicar las técnicas de procesos de extracción, transformación para luego alcanzar cargar los datos siendo estos lo más consistentes posible. Para esto hacemos uso de la herramienta Pentaho Data Integration (KETTLE) que permite extraer los datos de los sistemas operacionales, sobre estos datos realizar una serie de transformaciones y luego cargarlos en el mercado de datos. Una vez poblado el Data Mart se debe realizar el diseño físico del mismo teniendo en cuenta la estructura multidimensional que lo caracteriza haciendo uso de los cubos de datos. El servidor OLAP requiere la especificación lógica de los esquemas y tablas que formarán estos cubos, así como las jerarquías y miembros de éstas que forman las dimensiones. Este diseño será elaborado mediante la herramienta de modelado Pentaho Schema Workbench, la cual permite confeccionar los cubos multidimensionales y serán efectuadas en ella las consultas que darán respuesta a las exigencias de los clientes, esta definición se hace a través de un archivo de esquema en XML. A partir de este esquema el servidor OLAP es capaz de traducir las consultas del usuario a SQL y ejecutarlas.

3.4 Procesos de Extracción, Transformación y Carga del Data Mart.

Primeramente para aplicar los procesos ETL pertinentes a los datos mediante la herramienta Pentaho Data Integration fue necesario realizar una conexión con las fuentes externas existentes. Para definir las conexiones estas fueron definidas mediante el Sistema de Gestor de Base de Datos Firebird SQL para las extracciones de los datos y MySql para la salida de los datos utilizadas en todas las transformaciones realizadas.

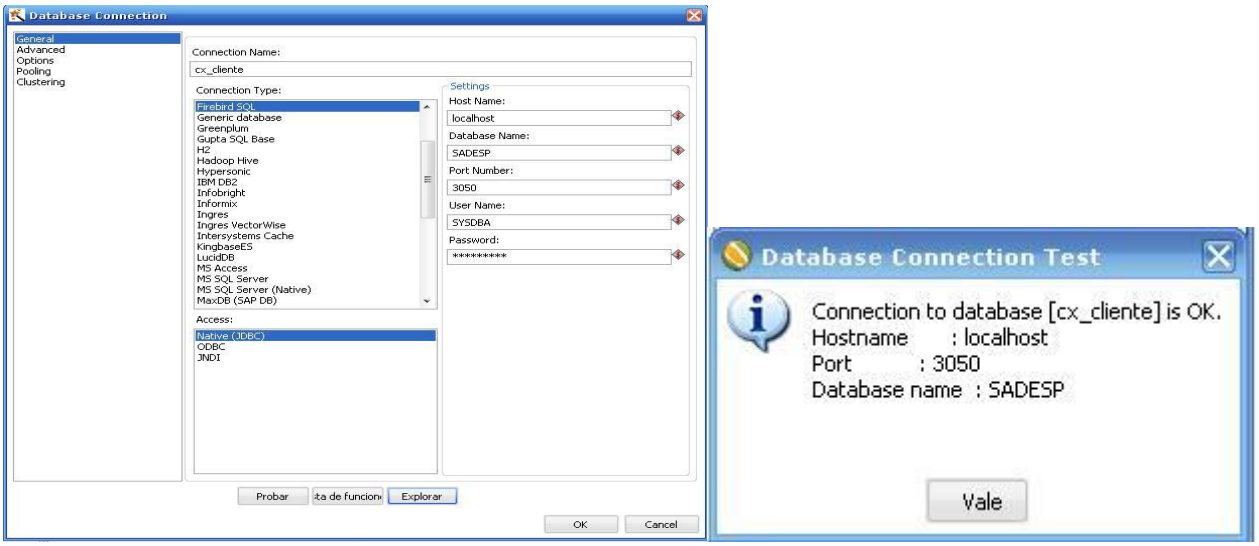


Figura 3.1: Conexión a la base de datos relacional.

Seguidamente de realizar la conexión se procede a aplicar los procesos de transformación, para llevar a cabo esta tarea se deben extraer los datos de las fuentes externas que cumplan con las necesidades, luego será necesario aplicar una serie de transformaciones y poblar el mercado de datos (figura).

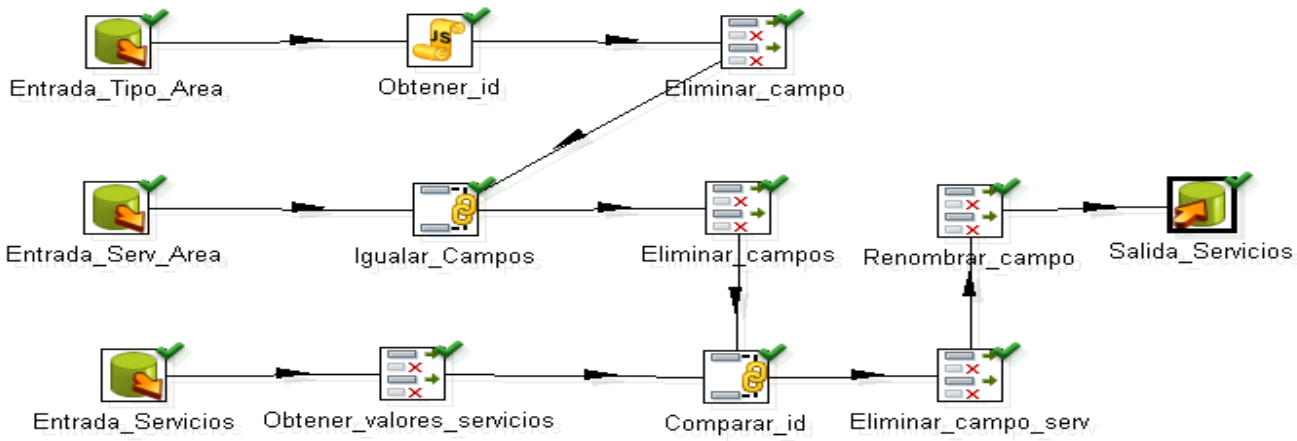


Figura 3.2: Transformación obtener servicios.

La idea principal de esta transformación se basa en extraer los servicios pertenecientes al área de Gastronomía en la empresa, donde los componentes Entrada_Tipo_Area, Entrada_Serv_Area y Entrada_Servicios serán los encargados de extraer los datos de base de datos relacional, definiendo así cuáles serán los datos a emplear (figura).

Figura 3.3: Búsqueda en la base de datos relacional.

Una vez extraídos los datos el procedimiento que continúa será realizar las transformaciones debidas, primeramente se define un componente de JavaScript el cual mediante un código permitirá almacenar en una variable el valor del identificador de la tabla Área.

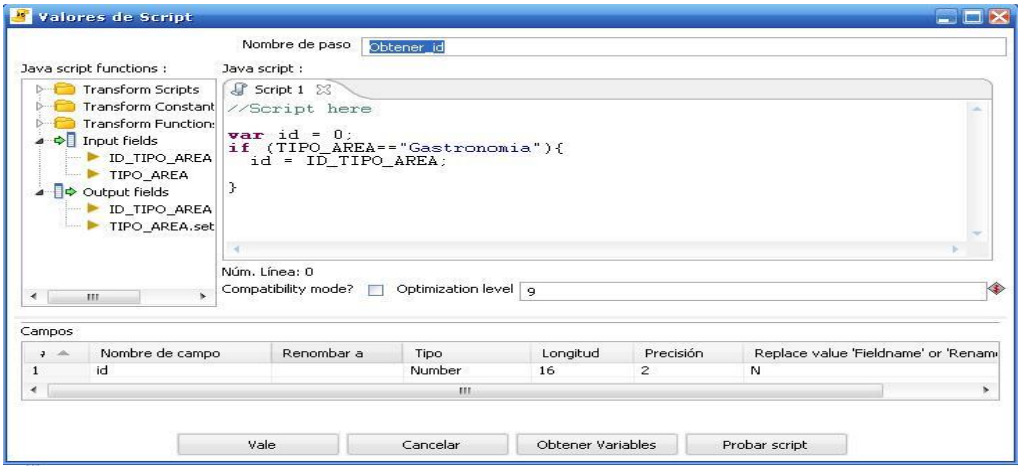


Figura 3.4: Código de JavaScript.

Luego de definir el área serán eliminados los campos innecesarios mediante el componente Eliminar_campos manteniendo así el valor de la variable, que se empleará para realizar una comparación de este con el identificador de los campos extraídos de la tabla tipo de área existente en la base de datos (figura 3.5) y luego el mismo procedimiento para obtener los servicios pertenecientes al área de Gastronomía, esta vez almacenado y comparando los identificadores de la tabla Productos (figura 3.6).

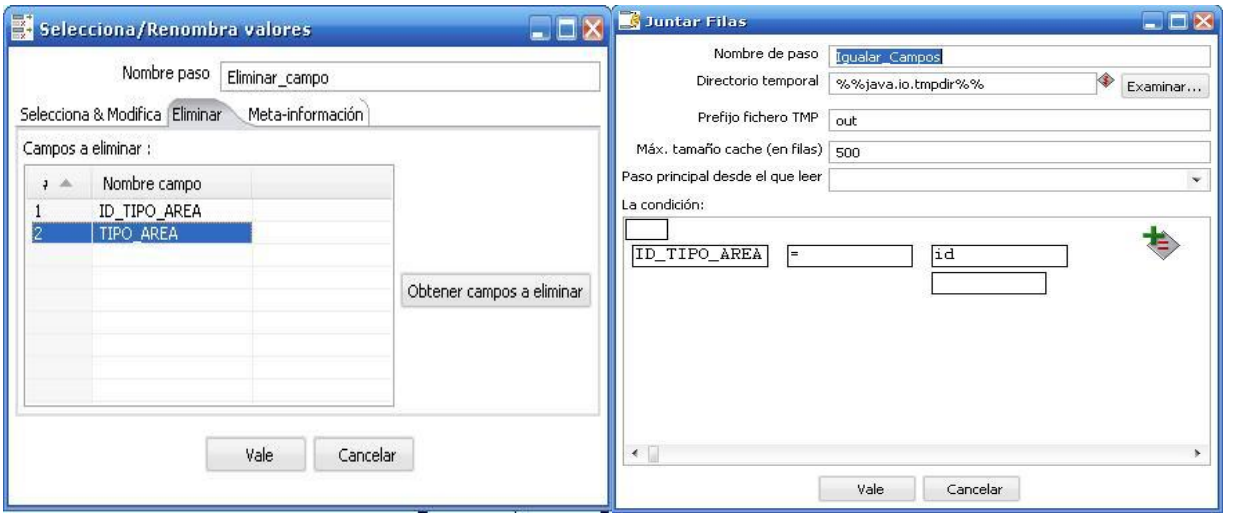


Figura 3.5: Eliminar e Igualar campos de la tabla Área.

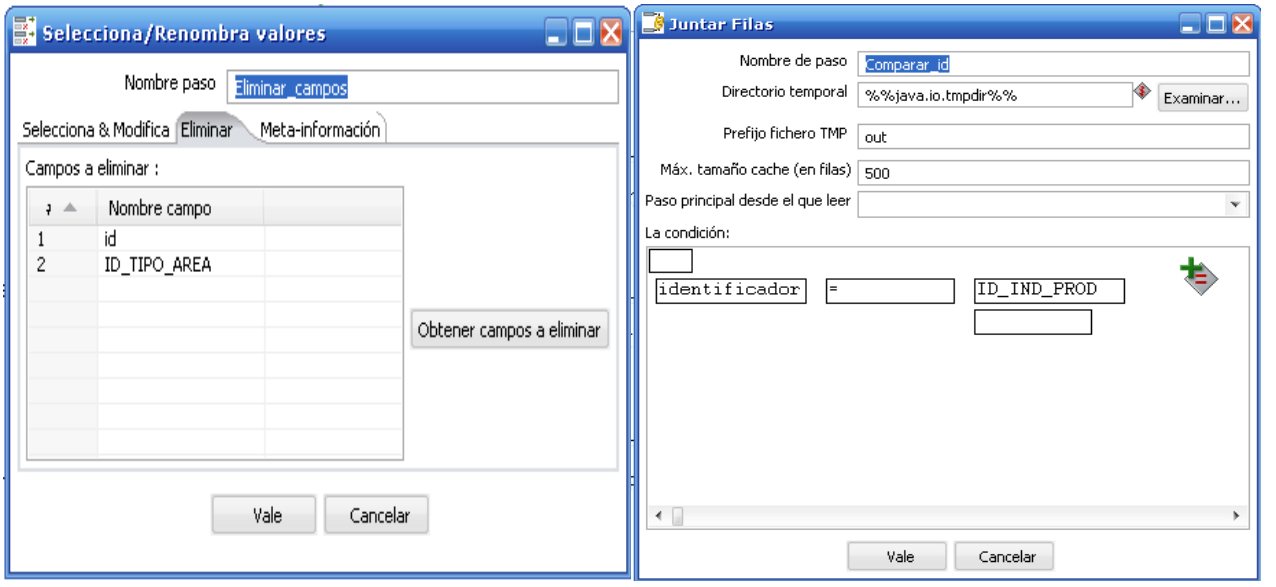


Figura 3.6: Eliminar e Igualar campos de la tabla Tipo_Área.

Finalmente con los componentes Eliminar_campo_serv, Renombrar_campo y Salida_Servicios siguiendo el orden, se eliminarán los campos no deseados, y los demás campos se modificarán asignándole el nombre de los campos definidos en la Dim_Servicios en el Mercado de Datos (figura 3.7):

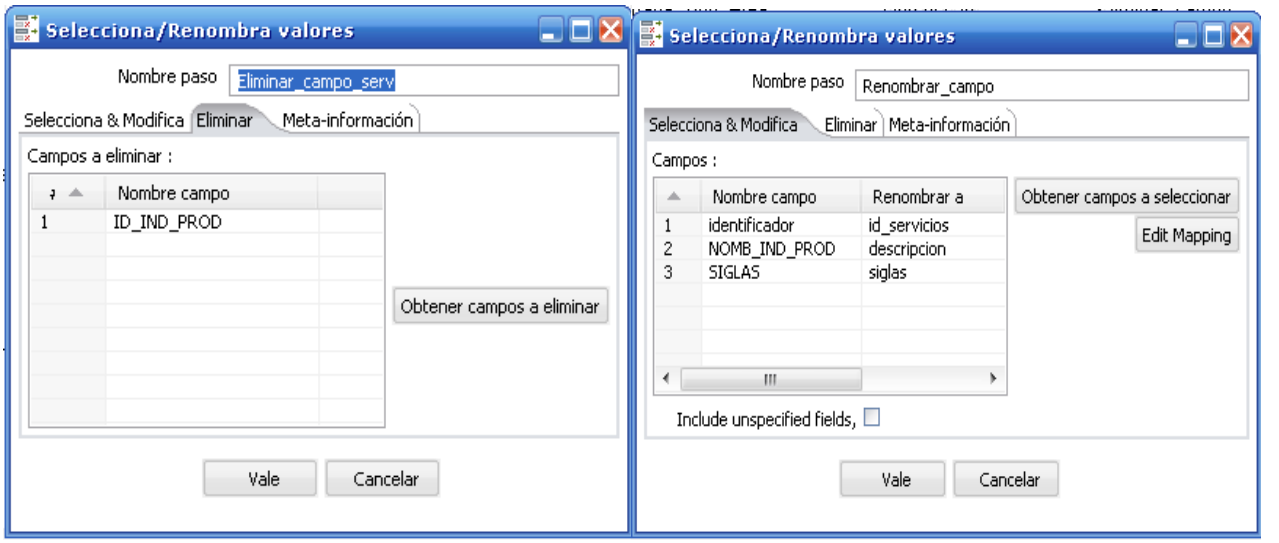


Figura 3.7: Eliminar y Renombrar campos servicios.

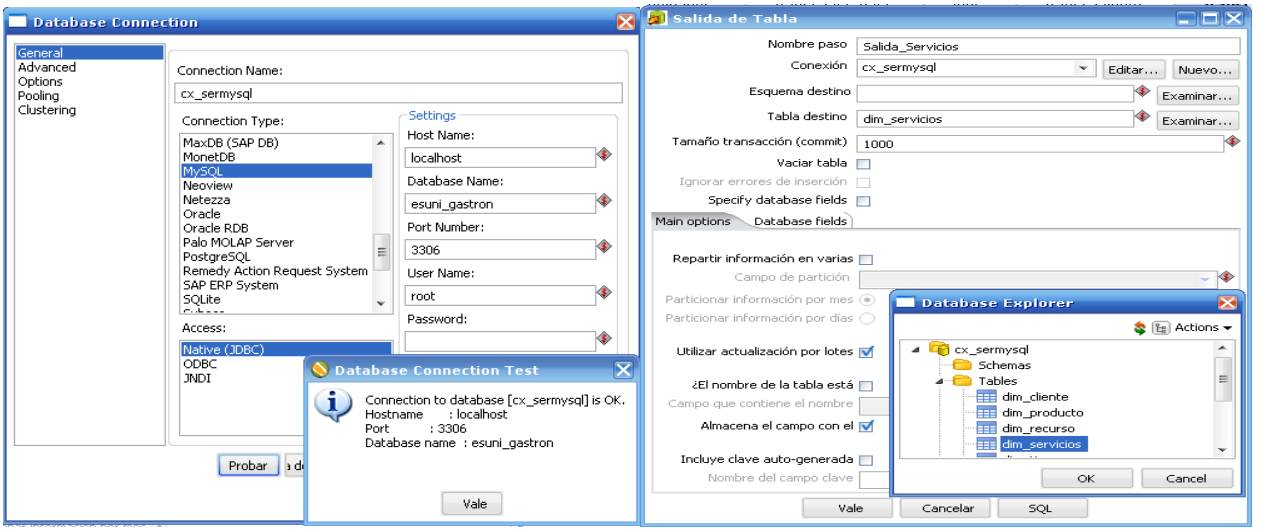


Figura 3.8: Conexión y Salida de datos al Data Mart.

En el caso de las transformaciones de la Dim_producto y Dim_Tipo_Producto se emplean los mismos componentes para la realizar las consultas a la base de datos y extraer los datos, pero en este caso se les asignará los nombres de Leer_Producto y Leer_Tipo_Producto. Es preciso mencionar que estos procesos son similares al ejemplo expuesto anteriormente la diferencia se encuentra dada en la carga de los datos ya que una depende de la otra (Figura 3.9, 3.10).

Figura 3.9: Transformación Tipo de Producto.

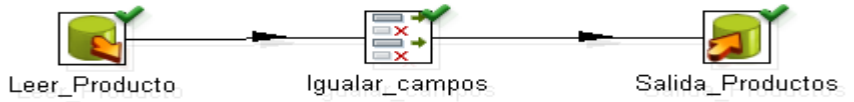


Figura 3.10: Transformación de Producto.

Los componentes de dichas transformaciones son similares luego de la extracción de los datos es aplicado el segundo componente Igualar_campos, esta transformación se basa específicamente en renombrar los campos obtenidos para de esta manera igualarlos a los campos de las salidas (Figura 3.11)

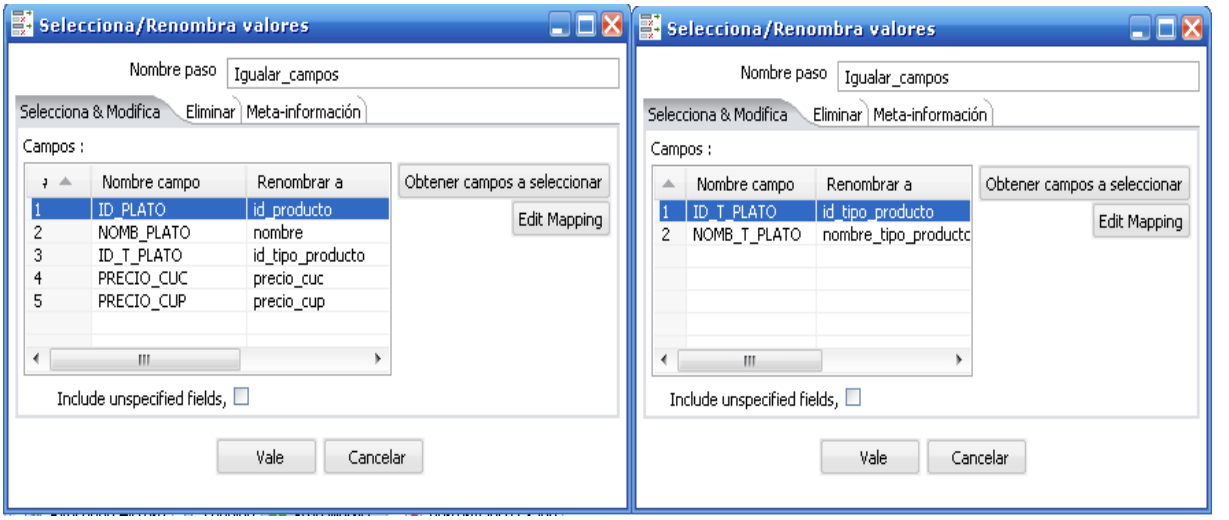


Figura 3.11: Igualar campos producto y tipo de producto.

Una vez culminadas las transformaciones y realizadas las cargas de las dimensiones, se procede a cargar aquellas dimensiones que dependen de otras es este el caso del ejemplo anterior. Lo antes expuesto significa que para lograr cargar la dimensión producto debe estar llena la dimensión tipo de producto, para esto es preciso originar los Jobs o más conocidos como Trabajos los cuales son ejecutados en función de un tiempo determinado (Figura).

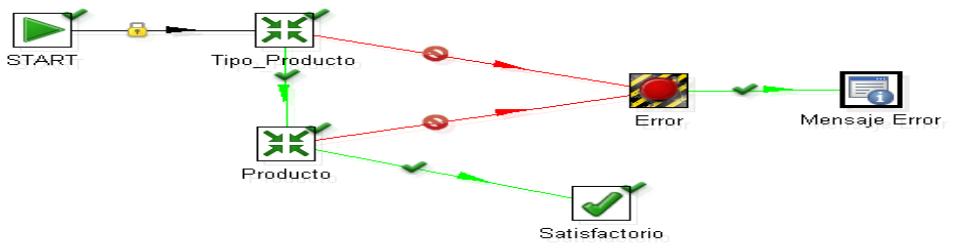


Figura 3.12: Trabajo o Jobs de las transformaciones tipo de producto y producto.

Inicialmente para la ejecutar el trabajo se emplea el componente START, luego se realizan las conexiones con las transformaciones según el orden de carga mediante el componente transformación (Figura 3.13):

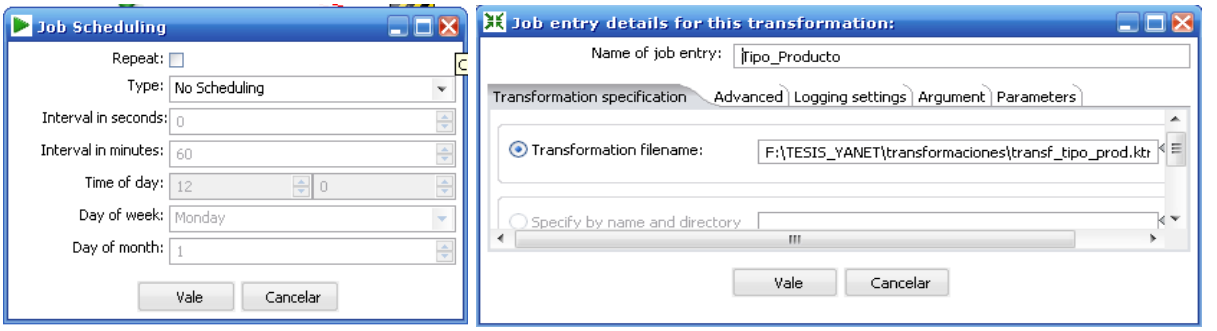


Figura 3.13: Ejecución de los componentes Jobs.

Después de especificar las transformaciones con el orden deseado se crea un componente Satisfactorio en caso de ser verdadero (figura3.14) y un componente Error en caso de ser falso el que contenga un mensaje describiendo el estado de error (figura 3.15).

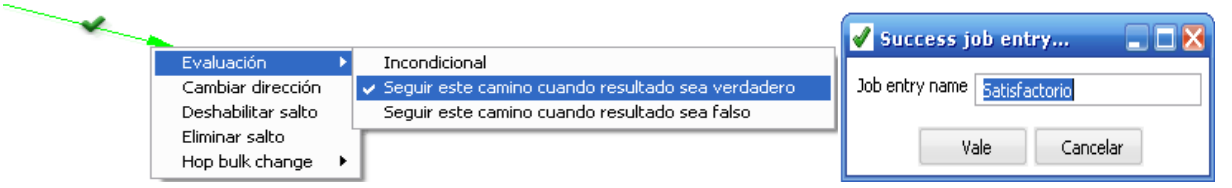


Figura 3.14: Salto verdadero.

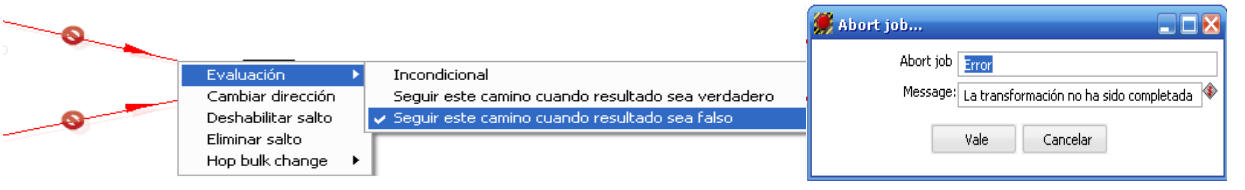


Figura 3.15: Salto falso.

3.5 Schema OLAP.

Para la confección del cubo OLAP del mercado de datos mediante la herramienta de modelado Schema Workbench fue necesario definir el cubo, las medidas y las dimensiones cada una con las jerarquías correspondientes a estas. Para lograr esta confección del cubo primeramente debemos conectarnos al Data Mart creado (Figura 3.16).

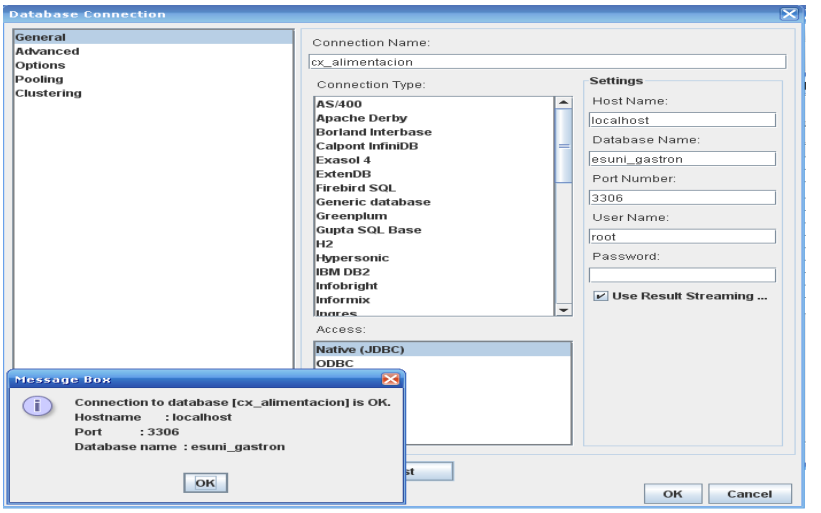


Figura 3.16: Conexión en la herramienta Schema Workbench.

Luego de la conexión se confecciona el Schema del mercado de datos el cual poseerá un cubo que estará confeccionado por la tabla de hechos que nombraremos Gastronomía. Un vez obtenido el cubo se le añadirá las dimensiones relacionadas, para esto se debe tener en cuenta que contengan el mismo nombre que en la base de datos, también se le atribuirá una llave foránea la cual será la llave primaria que representa esta dimensión en la tabla de hechos.

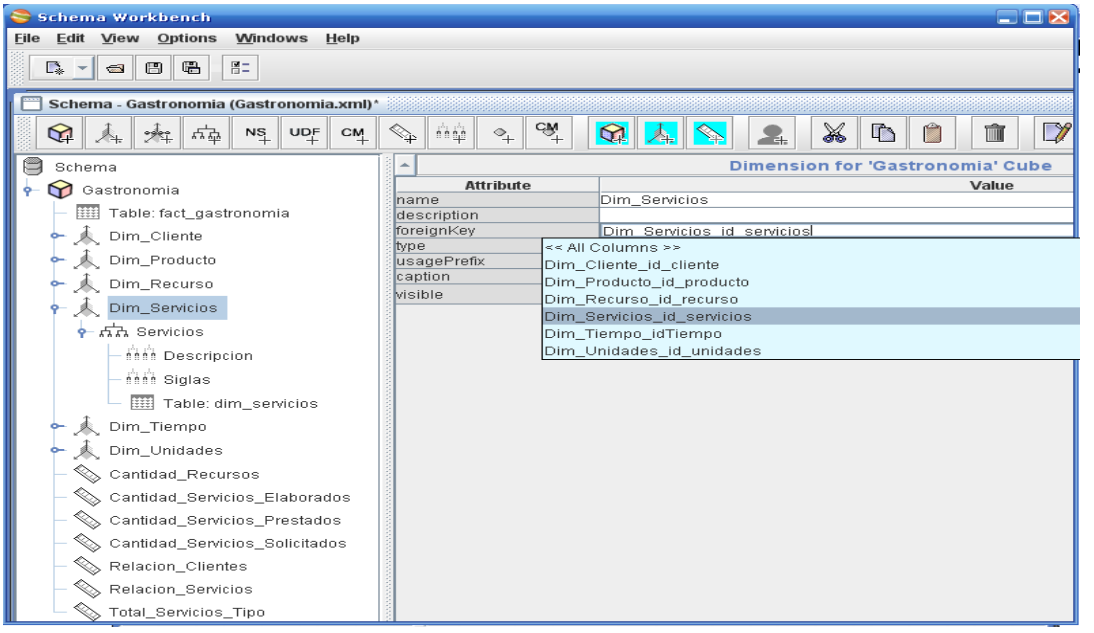


Figura 3.17: Cubos, dimensiones y jerarquías.

En el caso de las jerarquías poseerán varios niveles que representarán cada campo de las dimensiones, en este caso la jerarquía de la Dim_Servicios se llamará Servicios con la llave

primaria id_servicios y sus niveles serán Descripción y Siglas. Las medidas representan los hechos que fueron especificados en la tabla de hechos Cantidad_Recursos, Cantidad_Servicios_Elaborados, Cantidad_Servicios_Solicitados Total_Servicios_Tipo, Relacion_Clientes, Relacion_Servicios, Cantidad_Servicios_Prestados. El esquema XML generado en este esquema estará estructurado de la siguiente manera:

Schema

```
<Schema name="Gastronomia">
  <Cube name="Gastronomia" visible="true" cache="true" enabled="true">
    <Table name="fact_gastronomia">
    </Table>
    <Dimension type="StandardDimension" visible="true" foreignKey="Dim_Cliente_id_cliente" highCardinality="false" name="Dim_Cli
ente">
      <Hierarchy name="Cliente" visible="true" hasAll="true" primaryKey="id_cliente">
        <Table name="dim_cliente">
        </Table>
        <Level name="Nombre" visible="true" table="dim_cliente" column="nombre" type="String" uniqueMembers="true" levelType="Re
gular" hideMemberIf="Never">
        </Level>
        <Level name="Siglas" visible="true" table="dim_cliente" column="siglas" type="String" uniqueMembers="true" levelType="Regul
ar" hideMemberIf="Never">
        </Level>
        <Level name="Direccion" visible="true" table="dim_cliente" column="direccion" type="String" uniqueMembers="true" levelType="
Regular" hideMemberIf="Never">
        </Level>
      </Hierarchy>
    </Dimension>
    <Dimension type="StandardDimension" visible="true" foreignKey="Dim_Producto_id_producto" highCardinality="false" name="Dim
_Producto">
      <Hierarchy name="Producto" visible="true" hasAll="true" allMemberName="All Producto" primaryKey="id_producto">
        <Table name="dim_producto">
        </Table>
        <Level name="Nombre_Producto" visible="true" table="dim_producto" column="nombre" type="String" uniqueMembers="true" le
velType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Precio_CUC" visible="true" table="dim_producto" column="precio_cuc" type="Numeric" uniqueMembers="true" le
velType="Regular" hideMemberIf="Never">
        </Level>
        <Level name="Precio_CUP" visible="true" table="dim_producto" column="precio_cup" type="Numeric" uniqueMembers="true" le
velType="Regular" hideMemberIf="Never">
```

Figura 3.18: XML de la construcción del cubo.

Una vez culminado el diseño se elabora y especificadas las medidas se elaboran las consultas en MDX Query, es preciso que sean ejecutadas para conocer los resultados o si presenta algún error los que pueden darse por encontrarse sin datos el Data Mart o este mal confeccionada la consulta. A continuación se representa un ejemplo de estas consultas definida para la medida Relacion_Servicios (Figura 3.19).

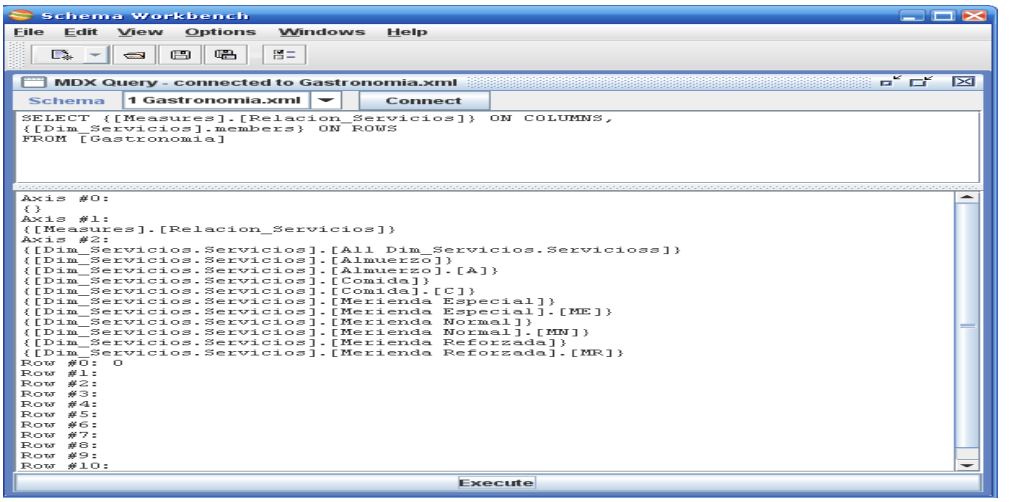


Figura 3.19: Consulta MDX Query.

3.6 Análisis de los tiempos de carga del Data Mart.

Una vez realizados los procesos de extracción, transformación y carga se procede a la carga inicial del Data Mart, verificando que todos los campos de cada dimensión sean poblados debidamente (ver figura 3.20).

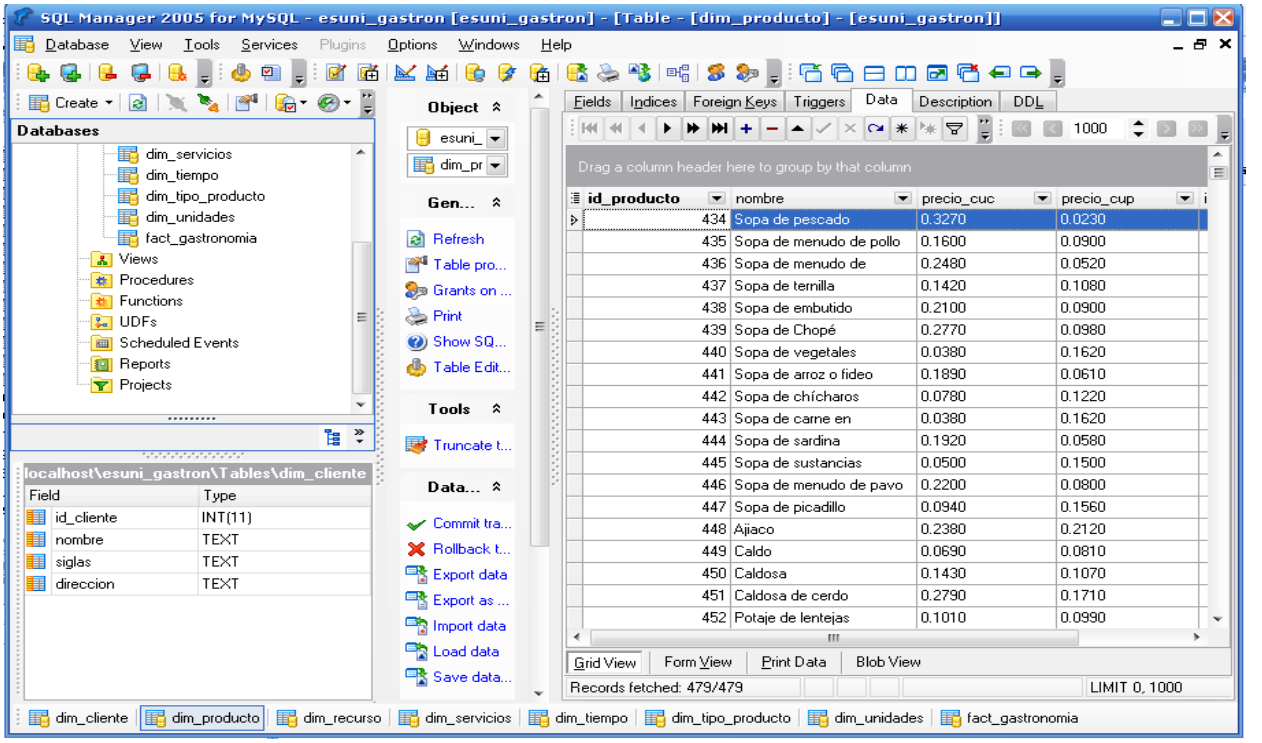


Figura 3.20: Datos cargados en Data Mart.

Como se puede observar cada dimensión del mercado de datos fue poblada satisfactoriamente, para esto se tuvo en cuenta algunas políticas, las mismas serán mencionadas a continuación.

- ✓ La información se refrescará todos los días a las seis de la mañana.
- ✓ Los datos de todas las tablas de dimensiones existentes en el Data Mart serán cargados desde 1 de Junio del 2014 hasta 30 de Junio del 2014.
- ✓ Antes de la carga de cada dimensión todos los datos serán borrados.
- ✓ Los datos de la tabla de dimensión “Tiempo” se cargarán de manera incremental teniendo en cuenta la fecha de la última actualización.
- ✓ Los datos de la tabla de hechos que corresponden al último mes (30 días) a partir de la fecha actual, serán reemplazados cada vez.

Estas acciones se realizarán durante un periodo de prueba, para analizar cuál es la manera más eficiente de generar las actualizaciones, basadas en el estudio de los cambios que se producen en los OLTP y que afectan al contenido del Data Mart.

3.7 Prueba de Integridad de Datos.

Para cualquier organización es fuente de valor contar con una estrategia bien definida y que todos los riesgos sean calculados. Cuando se trabaja con distintos sistemas, es primordial tomar su integración en consideración, dándole la importancia que merece y reconociendo su función armonizadora. La integración sentará las bases de la calidad de los datos, en especial en todo lo concerniente a integridad de la información y al Bussiness Intelligence. Mediante estas pruebas serán analizada que la solución Data Mart pueda integrarse a otras soluciones o aplicaciones a través del uso de un diseño documentado y de estándares de programación.

Se realizó un proceso de prueba probando cada consulta y procedimientos con el fin de obtener resultados correctos. Para conocer la eficiencia de las herramientas y técnicas empleadas se evaluará efectuando una consulta y lo esperado es que estas entreguen lo que se solicita. Las entradas de las consultas serán MDX devolviendo así los datos que se consulten. Una vez realizada la prueba se pudo observar que las respuestas a las consultas y procedimientos fueron los correctos.

3.8 Prueba de Validación.

Para realizar las pruebas de Validación se realizó una consulta a la una de las medidas definidas en la tabla de hechos Gastronomía con la finalidad de conocer si los resultados son correctos. La medida que fue consultada fue Cantidad de servicios para esto nos auxiliamos del navegador OLAP el cual presenta la información de los cubos y las medidas cuando se despliegan para seleccionarlas (Figura 3.21):

Figura 3.22: Cantidad de servicios.

Luego ejecutamos un código MDX para conocer si responde a esta consulta (Figura 3.22):

select{[Measures].[Relacion_Recurso]}ONCOLUMNS,
{[Dim_Recurso].[Precio_CUC].Members,[Dim_Recurso].[Precio_CUP].Members,[Dim_Recurso].[Nombre_Recurso].Members, [Dim_Recurso].[Unidad_Medida].Members} ON ROWS
from [Gastronomia]

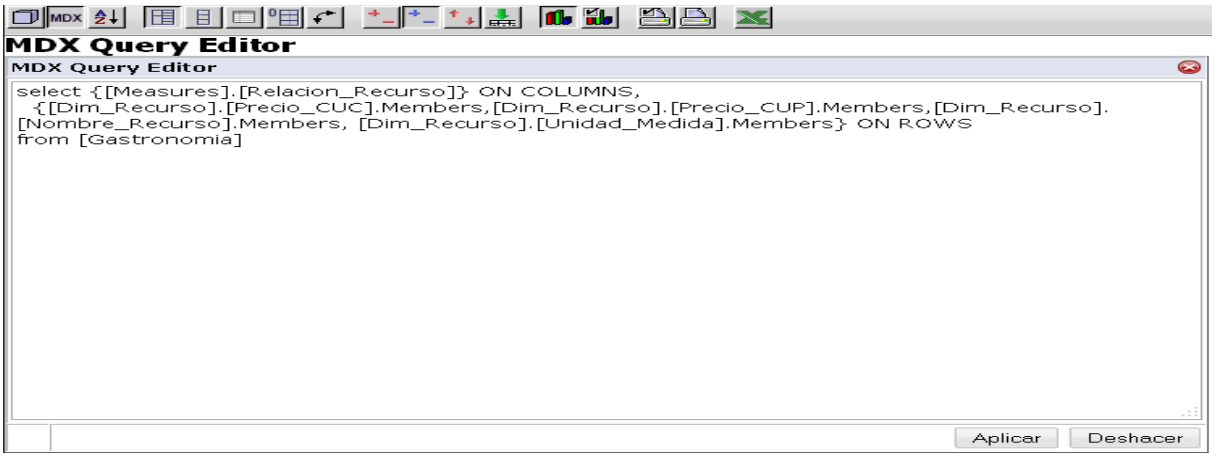


Figura 3.23: Consulta MDX.

Si la consulta esta correcta se podrá visualizar los siguientes resultados (Figura 3.23):

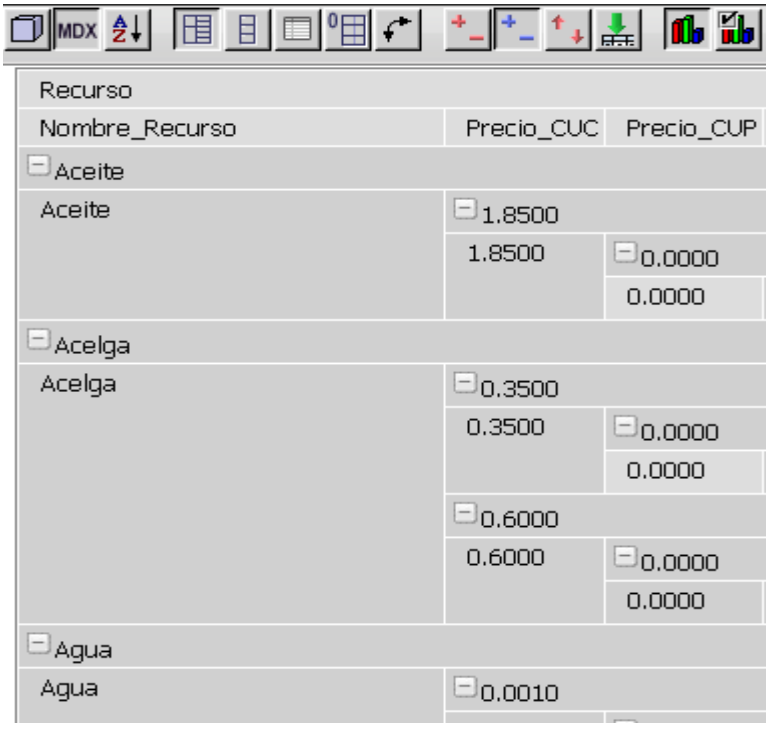


Figura 3.24: Registro de Recurso.

3.9 Conclusiones.

En este capítulo fue implementado el Data Mart haciendo uso de las herramientas adecuadas para el mismo. Luego de haber realizado los procesos de extracción, transformación y carga, además del diseño del mercado de datos se realizaron varias consultas donde se obtuvieron los resultados esperados. De manera que los usuarios finales que realizarán las consultas contarán con información integra y útil favoreciendo la toma de decisiones y que se puedan realizar predicciones en una línea de tiempo.

Conclusiones Generales

Con el desarrollo de un Mercado de Datos se dio cumplimiento al objetivo propuesto, se logró centralizar, integrar, procesar y analizar la información llevada a cabo en el proceso de control de servicios de alimentación de la Empresa de Servicio a la Unión del Níquel.

- ✓ Se realizó la investigación de los conceptos, metodologías y herramientas útiles para el desarrollo de la solución propuesta.
- ✓ Se desarrolló cada fase de la metodología HEFESTO basadas en las necesidades del negocio para así obtener un buen diseño e implementación del mercado de datos.
- ✓ Se realizó la implementación haciendo uso de las herramientas encargadas de los procesos de extracción, transformación y carga así como la de diseñar el cubo del mercado de datos.
- ✓ Fue validado por una serie de consultas realizadas para verificar que las respuestas obtenidas serían las correctas.

Recomendaciones

Con el objetivo de facilitar el rendimiento y uso del Data Mart para apoyar a la toma de decisiones es preciso recomendar:

- ✓ Realizar los reportes para mejorar la captura de la información sobre el almacén de datos, haciendo uso de la herramienta Pentaho Report Designer.
- ✓ Realizar el diseño e implementación de otros Data Marts para almacenar la información de los servicios de hospedaje y transporte en la ESUNI.

Bibliografía

Bernabeu, Ricardo Dario. 2010. DATA WAREHOUSING: Investigación y Sistematización de Conceptos. *HEFESTO.* Córdoba : s.n., 2010.

—. 2010. HEFESTO: Metodología para la Construcción de un Data Warehouse. *HEFESTO.* Córdoba : s.n., 2010.

Carpani, Fernando. 2000. *CMDM: Un Modelo Conceptual para la Especificación de Bases Multidimensionales.* s.l. : Universidad de la República, Instituto de Computación - Facultad de Ingeniería, 2000.

Casillas Santillán, Luis Alberto, Gibert Ginestà, Marc y Pérez Mora, Óscar. 2010. *Bases de datos en MySQL.* s.l. : UOC, 2010.

Cruz Guerrero, René. 2003. *Utilización de la tecnología Data Warehouse en instituciones educativas.* Hidalgo : Universidad Autónoma del Estado de Hidalgo, Instituto de Ciencias Básicas e Ingeniería, 2003.

Data Warehouse para la gestión por procesos en el sistema productivo. **Boza García, Andrés, y otros. 2004.** Cancún, México : s.n., 2004. Second World Conference on POM and 15tn Annual POM Conference.

Espinosa, Roberto . 2010. Construccion procesos ETL utilizando Kettle (Pentaho Data Integration). *El Rincon del BI.* [En línea] 10 de mayo de 2010. http://churriwifi.wordpress.com/2010/05/10/16-3-construccion-procesos-etl-utilizando-kettle-pentaho-data-integration/.

KEMP DE LA HOZ , ESTEBAN ALBERTO. 2005. *Desarrollo de una metodología que permita a empresas el desarrollo de un Data Warehouse y su integración con Sistemas Workflow utilizando herramientas de libre distribución y/o bajo costo.* VALDIVIA : Universidad Austral de Chile, Facultad de Ciencias de la Ingeniería, 2005.

Laitano , Rommell y Alvarado, Pedro . 2009. Modelo de base de datos con ER/studio. *Ilistrados.com.* [En línea] 2009. http://www.ilustrados.com/.

MySql. 2009. Minitutorial de MySQL Workbench. *MySql.* [En línea] 2009. http://dev.mysql.com/doc/workbench/es/index.html.

Núñez Cárdenas, Felipe de Jesús. 2011. s.l. : Universidad Autónoma del Estado de Hidalgo, Escuela Superior Huejutla, 2011. Sistemas ROLAP y MOLAP.

2011. OLAP, MOLAP, ROLAP…. *Aprendiendo Business Intelligence.* [En línea] 23 de febrero de 2011. http://www.bi.dev42.es/2011/02/23/olap-molap-rolap/.

Pacheco, Orlando. 2010. Arquitectura de un data Warehouse. *Blogspot.com.* [En línea] 2010. http://dwhucv.blogspot.com/p/estructuar-de-un-data-warehouse.html.

—. 2010. Estructura de un data warehouse. *Blogspot.com.* [En línea] 2010. http://dwhucv.blogspot.com/p/estructuar-de-un-data-warehouse.html.

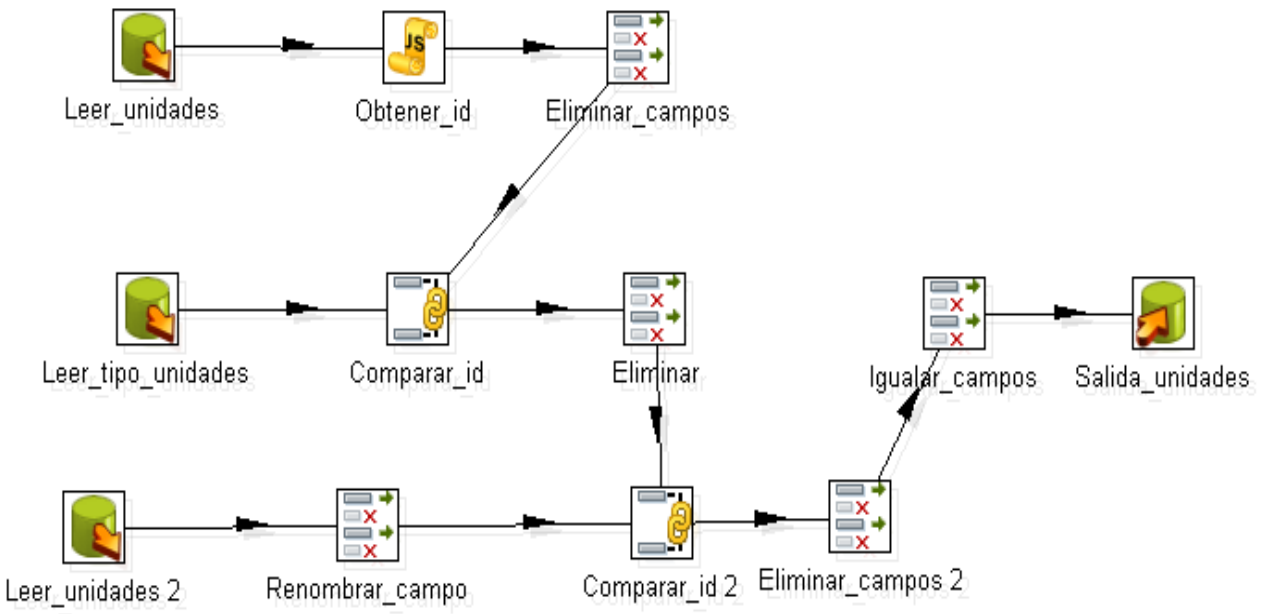
Rivadera, Gustavo R. 2010. *La metodología de Kimball para el diseño de almacenes de datos (Data warehouses).* 2010.

Tema 5 - PENTAHO: plataforma de Inteligencia de Negocio -. **ONUVA. 2008.** 2008. Inteligencia de Negocio: Conceptos y Aplicaciones con PENTAHO.

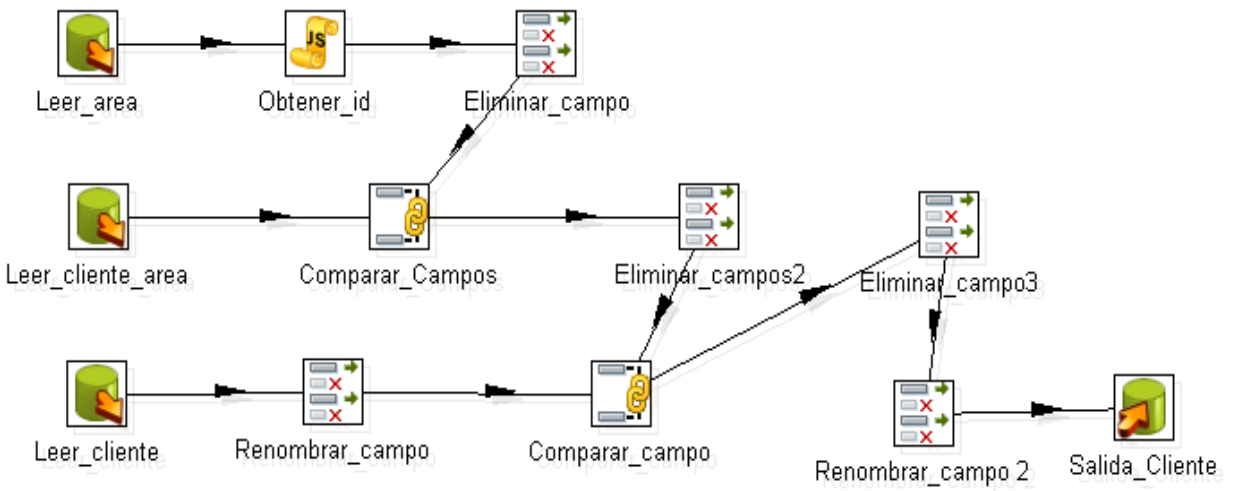
Villarreal Rosero, Roberto Xavier. 2013. *Estudio de metodologías de Data Warehouse para la implementación de repositorios de información para la toma de decisiones gerenciales.* Ibarra : UNIVERSIDAD TÉCNICA DEL NORTE, FACULTAD DE INGENIERÍA EN CIENCIAS APLICADAS, 2013.

Vizuete Naranjo, Michael Wladimir y Yela Shinín, Carlos Patricio. 2006. *Análisis, diseño e implementación de un Datamart para el Área de Sismología del Departamento de Geofísica de la Escuela Politécnica Nacional.* Quito : s.n., 2006.

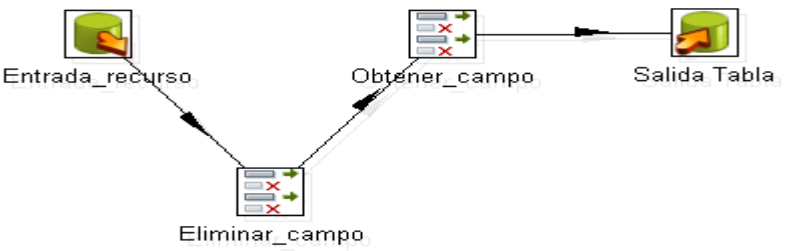
Anexos



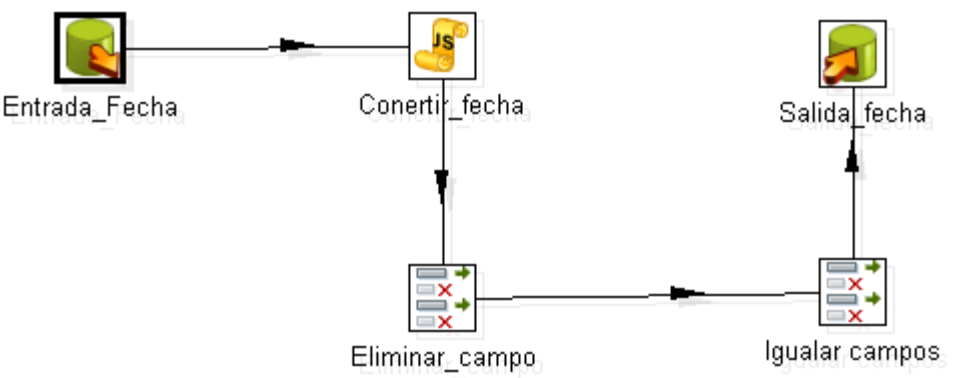
Anexo 1: Transformación unidades.



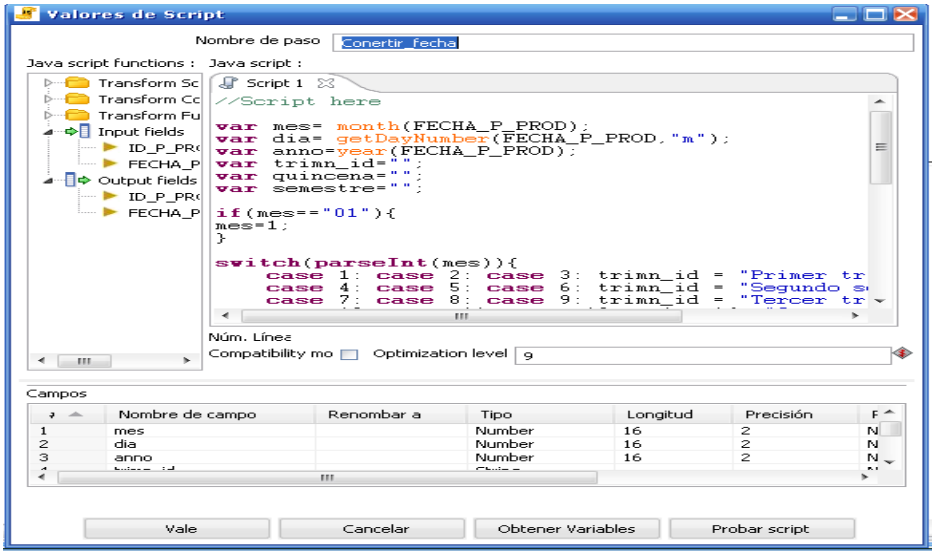
Anexo 2: Transformación Cliente.



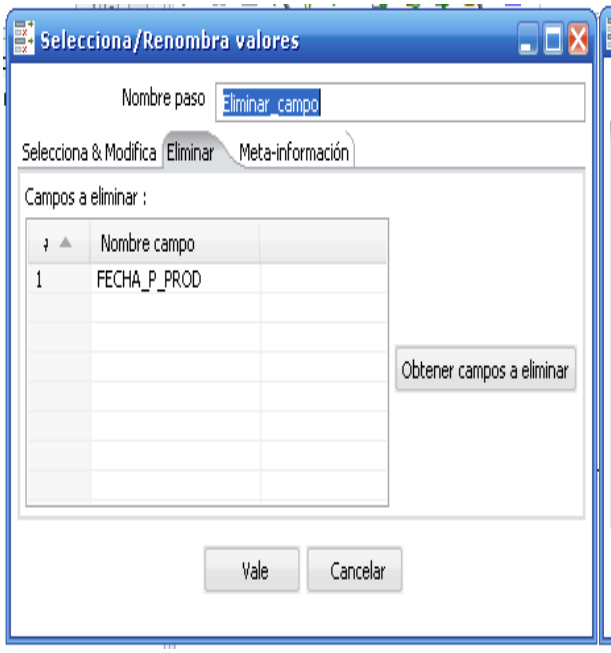
Anexo 3: Transformación Recuso.



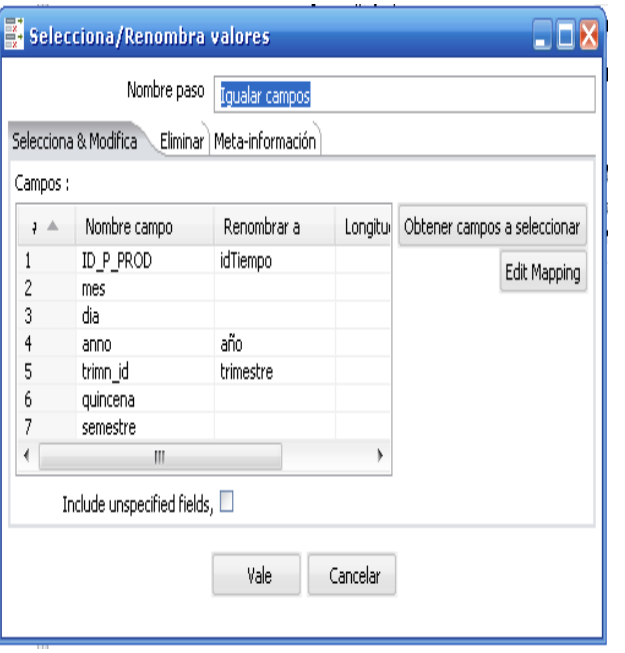
Anexo 4: Transformación Fecha.



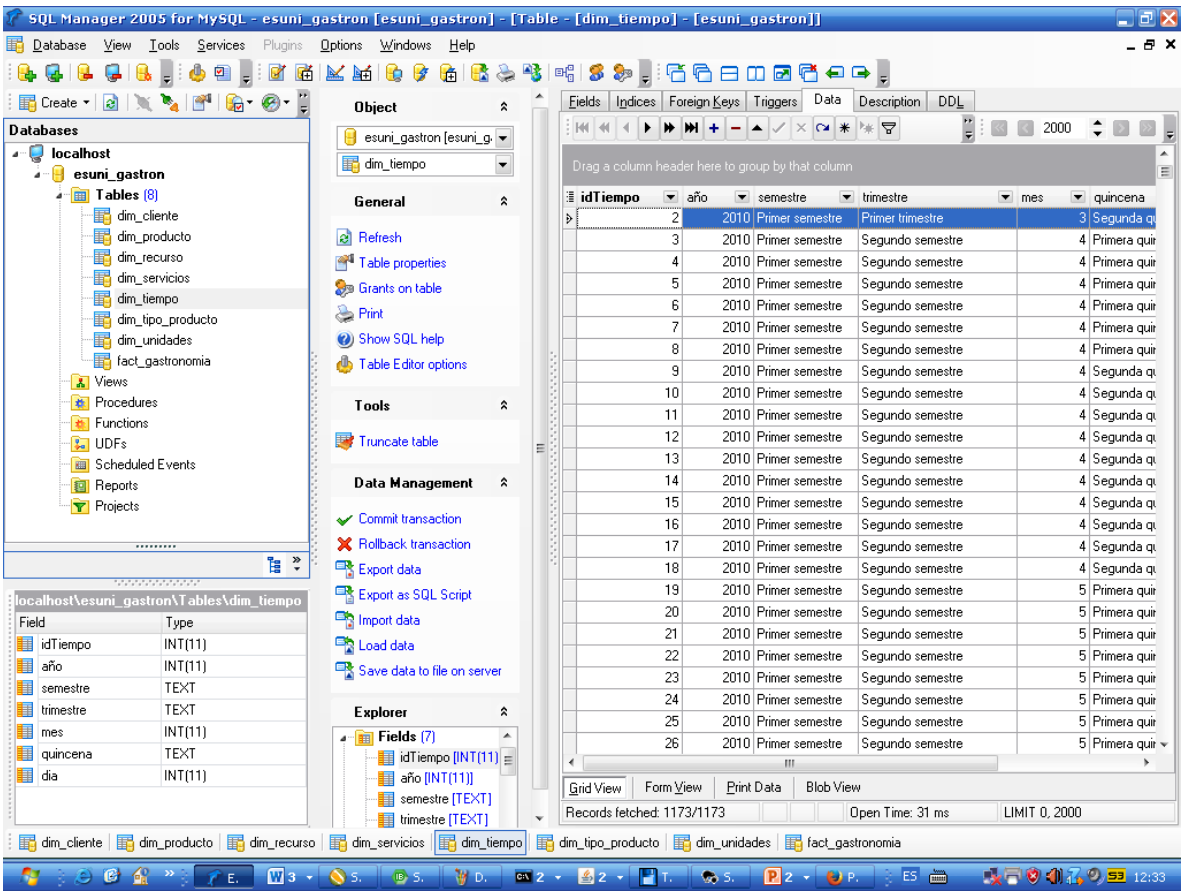
Anexo 5: Código de JavaScript para modificar fecha.



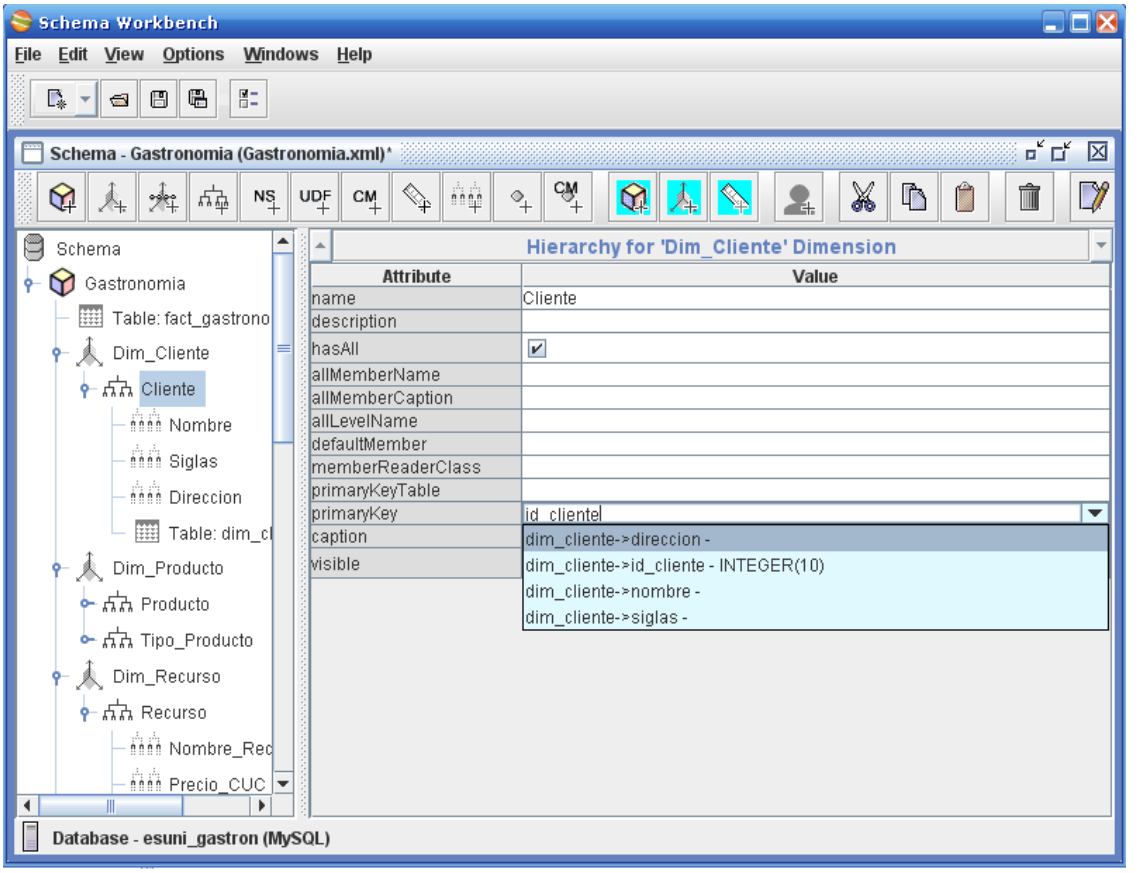
Anexo 6: Eliminar campo de fecha.



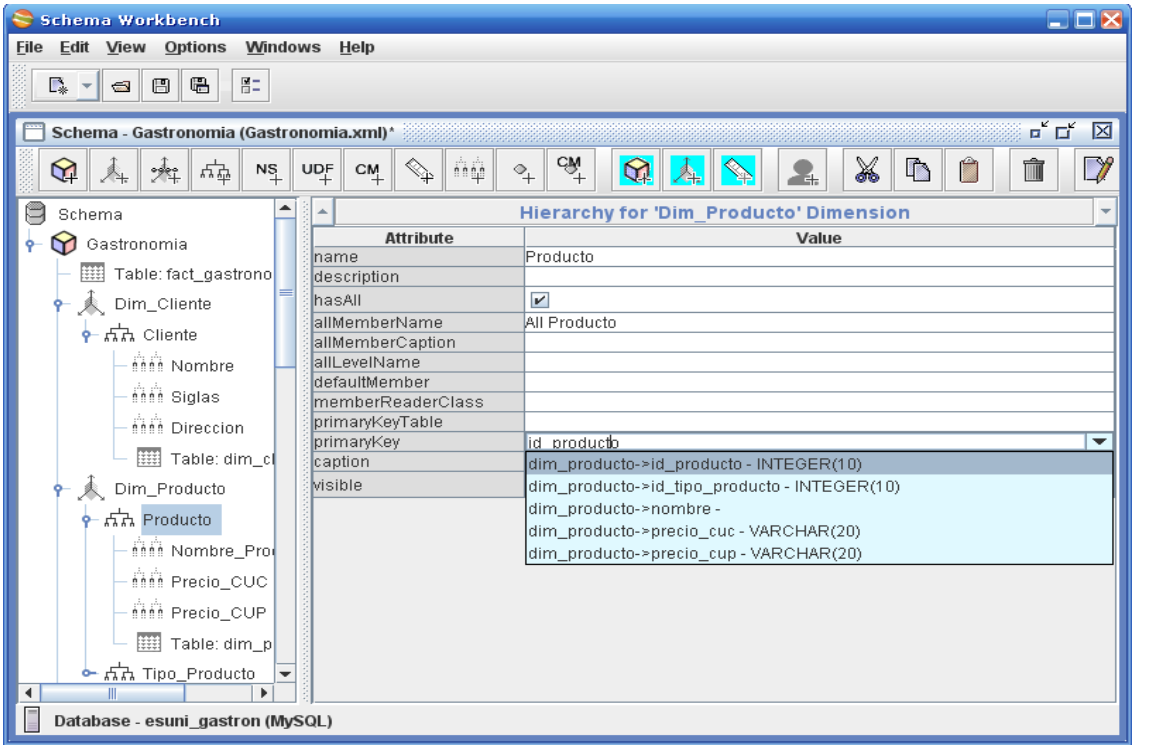
Anexo 7: Igualar campos fecha.



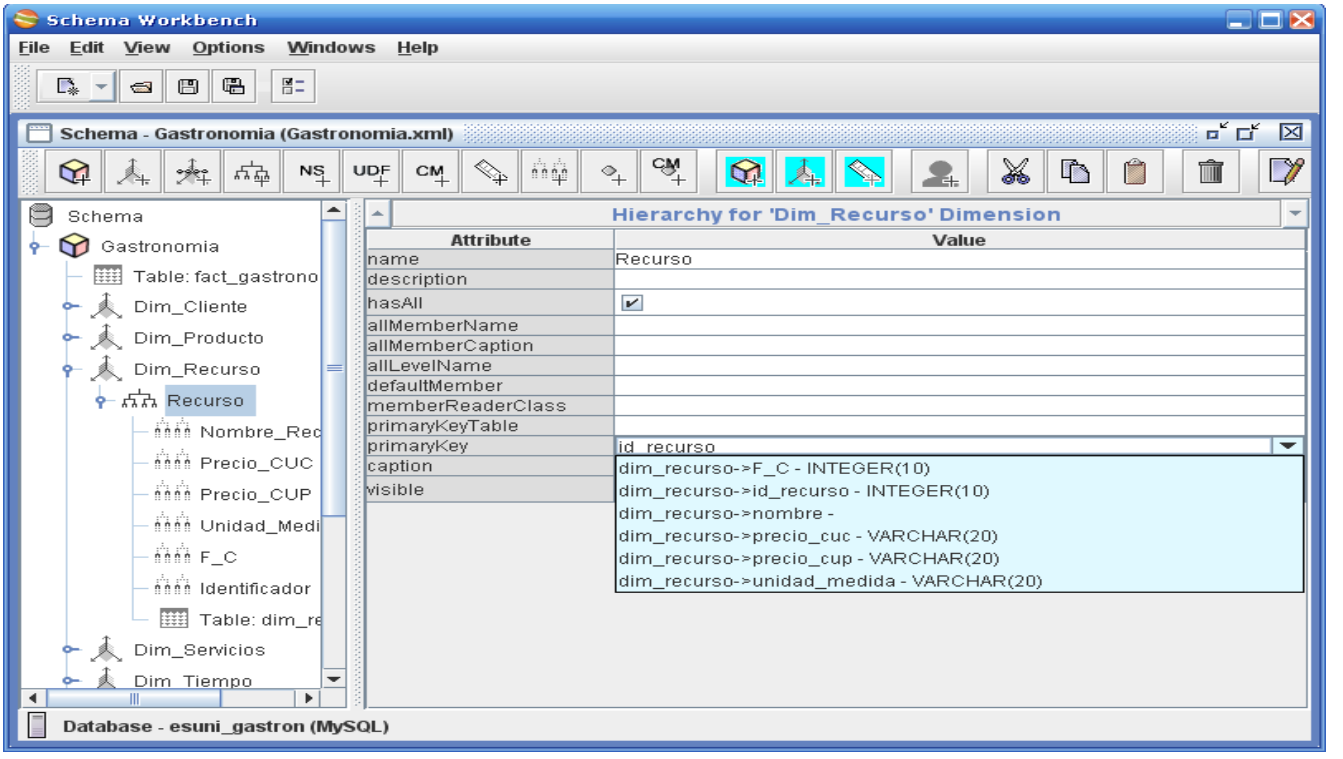
Anexo 8: Cargado el Mercado de datos.



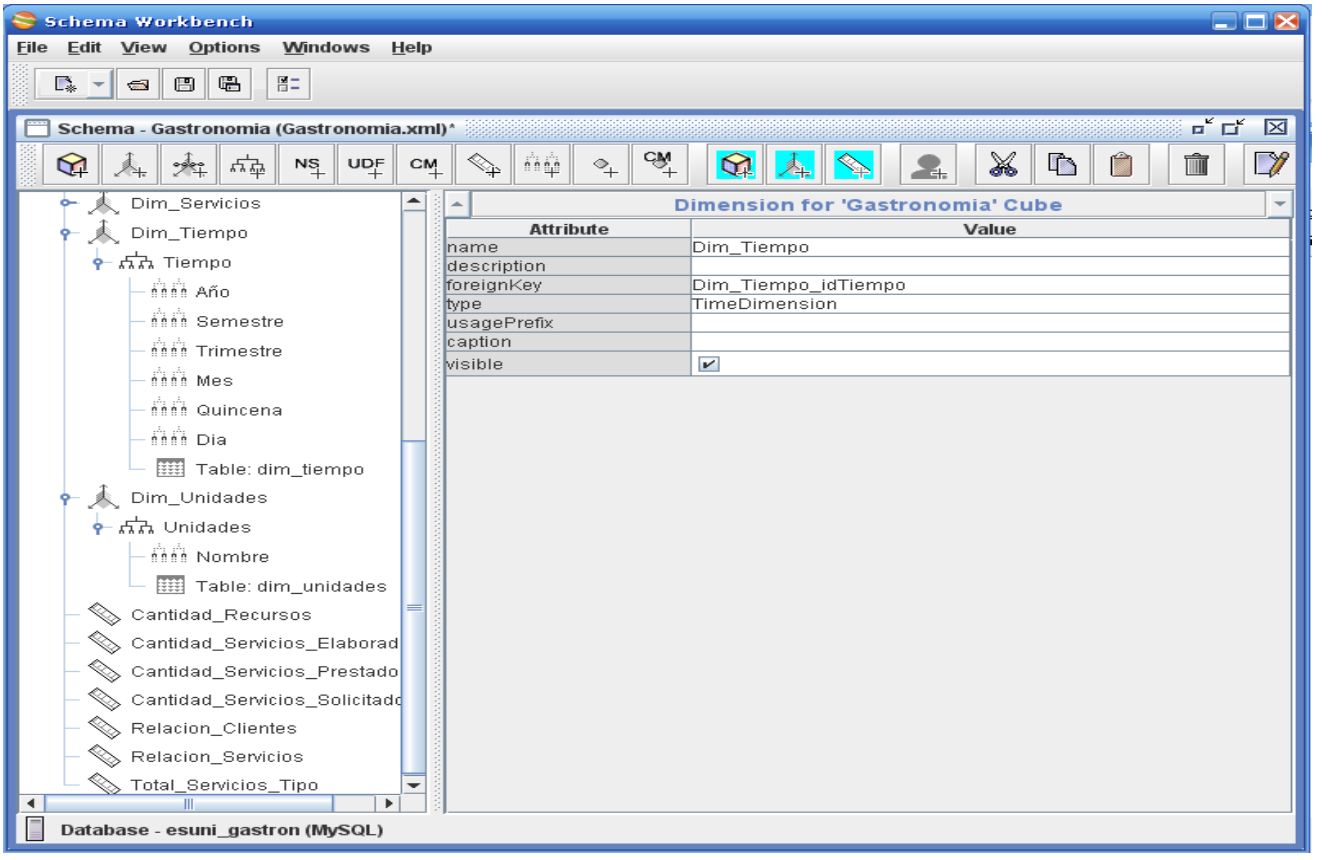
Anexo 9: Jerarquía Cliente en Workbench.



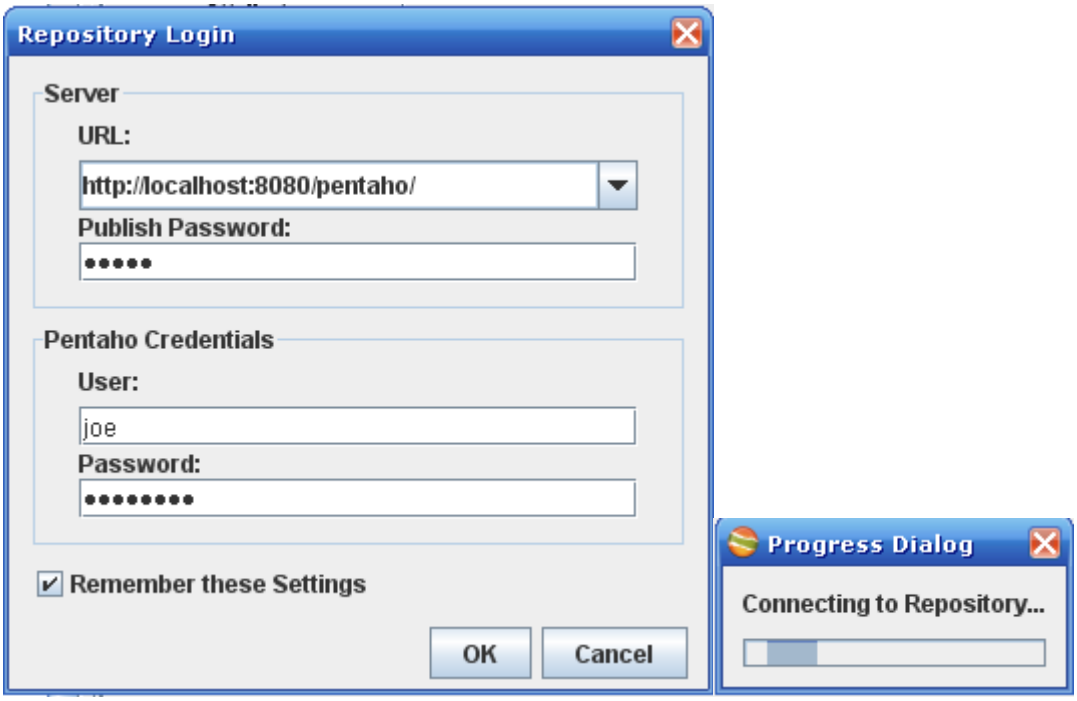
Anexo 10: Jerarquía Producto en Workbench.



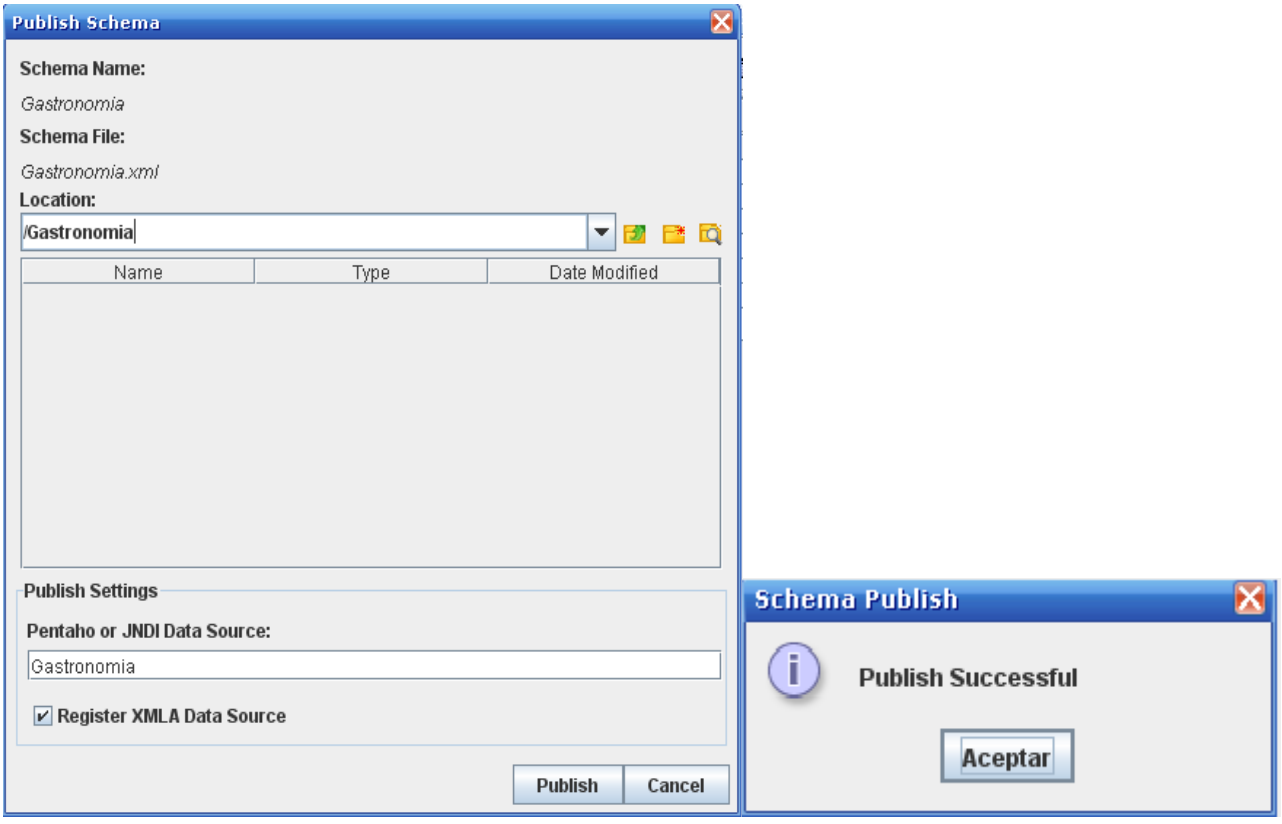
Anexo 11: Jerarquía Recurso en Workbench.



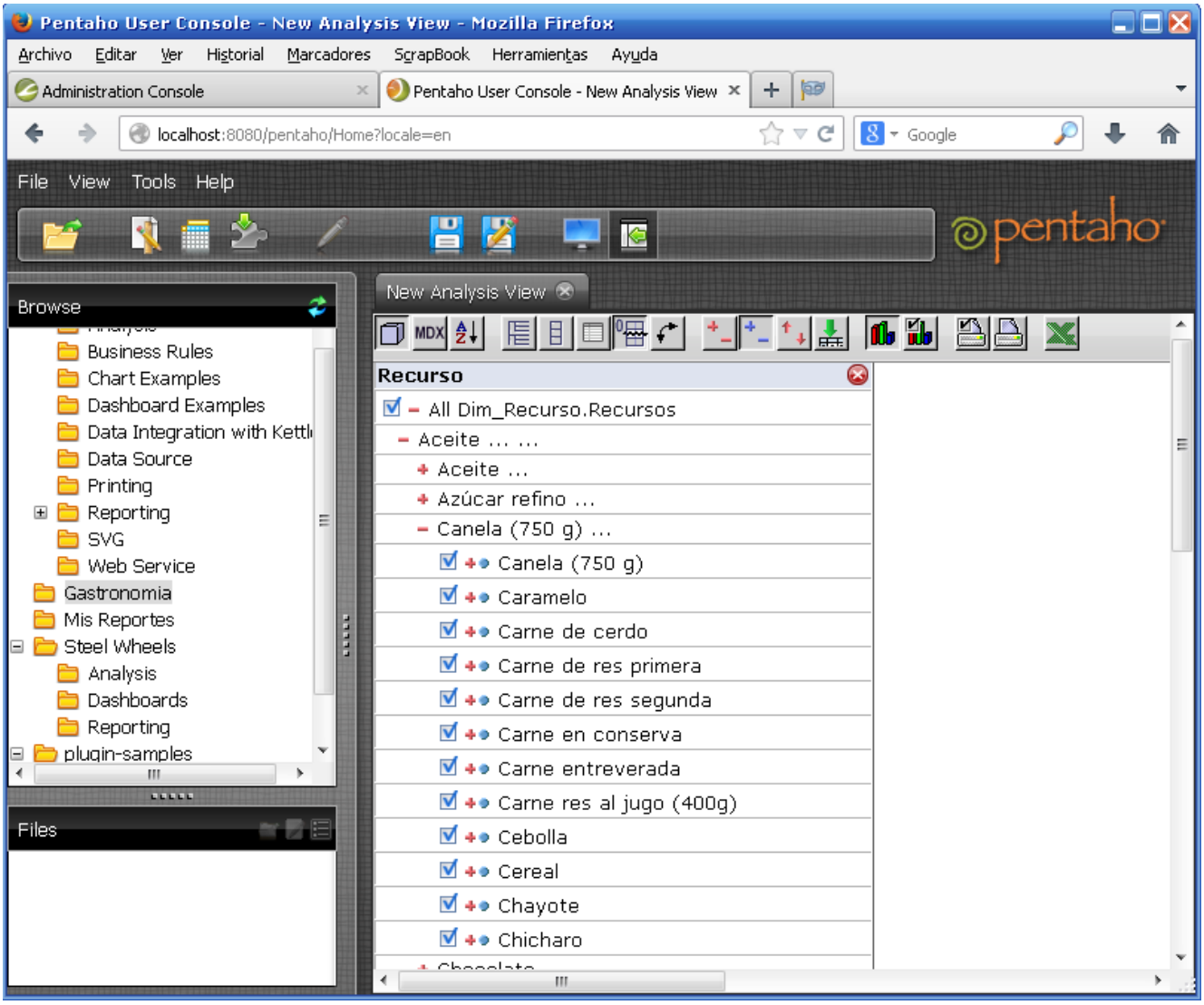
Anexo 12: Jerarquía Tiempo en Workbench.



Anexo 13: Ventana de publicación del esquema.



Anexo 14: Esquema publicado satisfactoriamente.



Anexo 14: Mercado de datos publicado en el servidor de Pentaho.