Universidad de las Ciencias Informáticas

Facultad 2

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas.

Título: Diseño e implementación de un mercado de datos para el análisis de eventos de navegación y mensajería.

Autor(es): *Osmari Hechevarría Delís*

Angel Rafael González Alvarez

Tutor: *MSc. Yasser Azan Basallo*

Co-Tutor: *Ing. Dayan Trujillo Marquez*

La Habana, 2013

"Año 55 de la Revolución"

DECLARACIÓN DE AUTORÍA

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmamos la presente a los ____ días del mes de _____ del año 2013.

Osmari Hechevarría Delis

_____

Firma del Autor

Angel Rafael González Alvarez

_____

Firma del Autor

MSc. Yasser Azan Basallo

_____

Firma del Tutor

Ing. Dayan Trujillo Marquez

_____

Firma del Tutor

AGRADECIMIENTOS

DEDICATORIA

RESUMEN

El presente trabajo tiene como objetivo principal desarrollar un mercado de datos para el departamento de Seguridad Informática de la empresa de telecomunicaciones de Cuba (ETECSA), debido a que la información que se genera a partir de los servicios que ofrece la misma no está establecida bajo un mismo formato, lo que retrasa en gran medida el proceso de toma de decisiones de la institución. El mercado de datos implementado posibilitará la realización de análisis estadísticos con la información que se encuentra en los ficheros de eventos de los servicios de navegación web y mensajería electrónica que brinda la empresa, así como determinar de forma rápida los accesos no permitidos y/o prohibidos, como por ejemplo, uso de las redes sociales, correo en servidores internacionales y el uso de proxys[1] libres que encubren la navegación. La construcción del mercado de datos está basada en el ciclo de vida dimensional del negocio que plantea Ralph Kimball en su metodología. La arquitectura conceptual de datos está conformada por dos capas, las cuales representan los diferentes entornos por los que pasan los datos en su camino hacia el mercado. Para iniciar la construcción del mercado, el equipo de desarrollo implementó una herramienta de apoyo para transformar la información inicial a un formato legible para la herramienta de creación del mercado. Para su elaboración se utilizó como gestor de base de datos PostgreSQL y como ambiente de desarrollo la suite[2] Pentaho.

Palabras claves: ETECSA, ficheros de eventos, Kimball, mercado de datos, toma de decisiones

[1] Proxys: Programa o dispositivo que realiza una acción en representación de otro

[2] Suite: Conjunto de programas libres para generar inteligencia empresarial (Business Intelligence). Incluye herramientas integradas para generar informes, minería de datos, ETL, etc.

ÍNDICE

INDICE DE FIGURAS

INTRODUCCIÓN

En la actualidad el acelerado desarrollo de las Tecnologías de la Información y las Comunicaciones (TIC) ha provocado que a nivel mundial, los volúmenes de información que se generan diariamente crezcan de manera exponencial, llegando a nombrarse la época actual como Era o Sociedad de la Información. Todo este volumen de datos es almacenado por diferentes empresas, quienes, desde un inicio utilizan las bases de datos relacionales para el manejo de los datos que se crean en sus operaciones diarias, por lo que en el transcurso de los años las bases de datos relacionales se convirtieron en una herramienta fundamental.

Sobre la información que manejan las empresas, la mayoría de las veces, se toman importantes decisiones. La toma de decisiones es en cualquier compañía un proceso complejo debido a la repercusión que puede tener, más aún si se trata de una institución influyente directamente en la economía de una nación.

El proceso de decisiones se considera como una serie de etapas que forman una decisión. La toma de decisiones es un proceso que lleva a cabo todo administrador y es considerado como una tarea central de la administración.[1]

Para el directivo o los administradores de una empresa, el proceso de toma de decisión es sin duda una de las mayores responsabilidades. Habitualmente, consideran este proceso como su principal trabajo; continuamente están decidiendo qué hacer, quién, cuándo, cómo y dónde se debe realizar.[2]

Justamente este enfoque permite destacar el papel que juegan en la actualidad las soluciones desarrolladas en el campo de la Inteligencia de Negocio (IN) o Business Intelligence como se le conoce en inglés. Para la autora Ivette Marrero Antunez, la IN no es más que el conjunto de sistemas, estrategias y herramientas informáticas cuyas funcionalidades están orientadas al apoyo de la toma de decisiones en una organización en aras del desarrollo exitoso de su negocio. Entre los sistemas que pueden estar comúnmente incorporados en las principales soluciones actuales de IN se encuentran los almacenes de datos (DWH)) y los mercados de datos[3].[3]

[3] Mercado de datos: Repositorios personalizados en función del análisis de datos de interés para un grupo específico de trabajadores o área de la entidad.

Los DWH constituyen uno de los soportes fundamentales para el proceso de toma de decisiones gerenciales. Las aplicaciones para soporte de decisiones basadas en DWH, pueden hacer más práctica y fácil la explotación de la información, pues permiten la integración y centralización de los datos que genera una empresa en todos los ámbitos de su actividad diaria. Organiza y almacena los datos que se necesitan para el procesamiento analítico e informático sobre una amplia perspectiva de tiempo, además de que es adaptable a los cambios que pueden ocurrir en el entorno del negocio.

Cuba no está ajena a esta revolución informática pues muchas de sus instituciones utilizan las soluciones de inteligencia de negocio para apoyar el proceso de toma de decisiones como lo son ETECSA, la Oficina Nacional de Estadísticas e Información (ONEI), el CIMEX[4], el Centro de Inmunología Molecular (CIM) entre otras. La Empresa de Telecomunicaciones de Cuba S.A (ETECSA), tiene la misión de: "Lograr en el período 2011-2015 una gestión efectiva que permita cada vez más brindar servicios de telecomunicaciones que satisfagan las necesidades de los usuarios y la población, así como respaldar los requerimientos de la defensa y del desarrollo socio-económico del país con resultados económicos que de la empresa demanda y espera el Estado cubano"[4]. Para lograr dicho objetivo es necesario mantener una seguridad informática adecuada por lo que el análisis de la información debe ser realizado de manera rigurosa.

La identificación y acción ante posibles vulnerabilidades mediante el monitoreo de eventos generados por los servicios telemáticos, constituye una misión del Departamento de Seguridad Informática (DSI) de la División de Tecnologías de la Información, en ETECSA. El cual tiene el propósito de garantizar la confidencialidad, integridad y disponibilidad de la información que se procesa e intercambia a través de las tecnologías.

El análisis de eventos de seguridad que se generan al realizar alguna actividad a partir de los servicios de navegación web y mensajería electrónica que ofrece la empresa, constituye uno de los elementos del proceso de vigilancia a la información que se genera en la empresa en el DSI, por lo que se necesita recolectar los datos pertenecientes a los mencionados ficheros, originados desde diferentes equipos para su integración y posterior análisis. Este paso permite un análisis de la información para la toma de

[4] CIMEX: Corporación Importadora y Exportadora

decisiones de la alta dirección y confeccionar informes objetivos que muestren el comportamiento de la seguridad en la entidad y adoptar oportunamente las medidas necesarias.

El monitoreo de los ficheros de los registros de eventos de seguridad se realiza de forma manual. La dificultad de este proceso está acentuada en que los registros de eventos de seguridad son generados por diferentes aplicaciones y dispositivos instalados relacionados con la seguridad y no están formalizados bajo un mismo formato. Esta situación retrasa en gran medida el análisis de la información obtenida de los ficheros para la adecuada toma de decisiones, ante los intentos de intrusión o cualquier otra amenaza o riesgos existentes.

A partir de todo lo anteriormente planteado surge como **problema a resolver:** ¿Cómo facilitar el proceso de toma de decisiones para la seguridad de la información presente en los servicios de navegación web y mensajería electrónica ofrecidas por ETECSA?

Para llevar a cabo la investigación se plantea como **objeto de estudio:** el análisis de la información en el proceso de toma de decisiones para la seguridad de la información en los servicios telemáticos y como **objetivo general:** desarrollar un mercado de datos que facilite el análisis de los eventos de navegación web y mensajería electrónica ofrecidas por ETECSA para el proceso de toma de decisiones. La investigación tiene como **campo de acción:** el análisis de la información en el proceso de toma de decisiones para la seguridad de la información en los servicios de navegación web y mensajería electrónica.

El **objetivo general** está dividido en los siguientes **objetivos específicos**:

1. Analizar los requisitos del mercado de datos.
2. Desarrollar una herramienta de apoyo para estructurar los datos obtenidos de las fuentes de origen.
3. Diseñar e implementar el mercado de datos.
4. Desarrollar los reportes y el análisis del mercado de datos a implementar.
5. Realizar las pruebas al mercado de datos.

Los objetivos específicos anteriormente planteados abarcan un grupo de **tareas** para darles cumplimiento, que a continuación se detallan:

- Elaboración del marco teórico de la investigación.
- Realización del estudio del arte de los sistemas dedicados al monitoreo de los servicios de navegación web y mensajería electrónica.
- Identificación de conceptos fundamentales vinculados a los mercados de datos.
- Descripción de la metodología de desarrollo del mercado de dato a implementar.
- Identificación de las herramientas a utilizar para el desarrollo del mercado de dato a implementar.
- Identificación y descripción de los procesos de negocio.
- Identificación de los requisitos de información del mercado de datos.
- Descripción de la herramienta de apoyo para estructurar los datos obtenidos de las fuentes de origen.
- Identificación de los niveles de granularidad, dimensiones, atributos de dimensiones y tablas de hechos del mercado de datos.
- Elaboración del diseño del modelo dimensional para implementar el proceso de extracción, transformación y carga (ETL) del mercado de datos.
- Investigación acerca de la implementación de los reportes y el análisis del mercado de datos.
- Comprobación de la lista de chequeo en el mercado de datos.

Posibles resultados:

- El diseño de un mercado de datos para correlacionar los datos según los intereses del DSI.
- Estandarización de los datos guardados en los ficheros de registro de eventos de seguridad de los servicios de navegación y mensajería.

- Reportes tomados del mercado de datos para facilitar la toma de decisiones con gráficos y tablas.

Durante la investigación se utilizaron los siguientes **métodos científicos**:

Métodos teóricos:

- Analítico Sintético: Análisis de la documentación existente sobre los procesos que se realizan en la toma de decisiones a partir de los datos recogidos en las auditorías de seguridad informática, además de la bibliografía sobre la construcción de los almacenes de datos y sintetizar los elementos necesarios a utilizar en el desarrollo del trabajo.
- Histórico Lógico: Se hace un estudio del estado del arte sobre las soluciones informáticas en el mundo que permitan el proceso de toma de decisiones a partir de los datos recogidos en las auditorías de seguridad informática, con el objetivo de ver cuál puede ser útil para la realización de este trabajo.

Métodos empíricos:

- Entrevistas: Se realizan un conjunto de entrevistas cerradas con el cliente para identificar los procesos de negocio que se llevan a cabo, los cuales son posteriormente agrupados por temas analíticos. Además se realiza la captura de los requisitos de información que son necesarios para darle respuesta a sus necesidades.

La presente investigación, está estructurada en tres capítulos, los cuales son:

Capítulo 1. "Marco teórico de la investigación" se realiza un estudio sobre soluciones informáticas similares existentes en sistemas de seguridad de Cuba y otros países. Además se analizan los principales elementos teóricos que constituyen la base de la investigación, entre los cuales se encuentran la arquitectura, la metodología de desarrollo para el mercado de datos y las herramientas a utilizar en la propuesta de solución.

Capítulo 2. "Análisis y diseño del mercado de datos" Se definen los temas de análisis, se realiza la especificación de los requisitos de información, se define la arquitectura y el modelado dimensional para los procesos del caso de estudio. Se describe la propuesta de solución y se especifican las dimensiones del mercado de datos y sus hechos asociados. También forma parte del capítulo la matriz dimensional, el

modelo físico de la base de datos y el diseño del proceso de Extracción, Transformación y Carga de los datos (ETL).

Capítulo 3. "Implementación y prueba del mercado de datos" Se desarrolla el subsistema de implementación y visualización, con el objetivo de darle solución a los requerimientos del sistema. Se valida el mercado de datos a través de las listas de chequeo.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA

Introducción

En el presente capítulo se realiza un estudio de las diferentes aplicaciones informáticas en sistemas dedicados al monitoreo de seguridad informática de diferentes países, incluyendo a Cuba. Además se enuncian conceptos, características, objetivos y ventajas de los almacenes de datos. Se hace un análisis de los diferentes enfoques que hay en la actualidad referente a la construcción de los almacenes de datos y además, se fundamentan las tecnologías que se utilizarán para el desarrollo de este trabajo.

1.1 La seguridad informática y la adecuada toma de decisiones

El éxito de una empresa está basado en parte en conservar una seguridad informática adecuada por lo que tomar las decisiones pertenecientes a esta área es un proceso fundamental para las repercusiones positivas de la empresa. Una de las etapas principales de dicho proceso es la investigación u obtención de información para luego ser analizada, en esta etapa se recopila toda la información necesaria para la adecuada toma de decisión; sin dicha información, el área de riesgo aumenta, porque la probabilidad de equivocarse es mayor, debido al desconocimiento de los elementos esenciales. [4]

1.2 Tendencias actuales de los sistemas dedicados al monitoreo de los servicios de mensajería y navegación

El desarrollo de las nuevas tecnologías, ha propiciado que diferentes organizaciones basen su desempeño corporativo en el uso de soluciones informáticas para garantizar la seguridad de la información. Varias de estas soluciones son capaces de facilitar el proceso de la toma de decisiones partiendo del análisis realizado a los datos obtenidos del uso de las tecnologías. Para la realización de este trabajo se analizaron un grupo de soluciones informáticas en el área de la seguridad. A continuación se describen estas soluciones:

1.2.1 Navegación Squid e ISA Server

Sarg (*Squid Analysis Report Generator* por sus siglas en inglés) es una de las herramientas más completas para el análisis. Está configurado para generar reportes web de los accesos a internet de forma periódica, además de, poder ejecutarlo manualmente para generar reportes de fechas, usuarios

o dominios en específico, usando como datos los logs [5]de Squid. Tiene soporte para generar los reportes en diferentes idiomas de toda la navegación realizada a través del proxy en un intervalo de tiempo. Es una de las herramientas más completas en ese sentido, mediante los reportes de uso web usted podrá obtener la siguiente información:

- Los diez primeros sitios más visitados
- Reportes diarios, semanales y mensuales
- Gráficas semanales y mensuales del consumo por usuario/host[6]
- Detalles de todos los sitios a los que accedió un usuario/host
- Descargas

Genera estadísticas en formato HTML[7] usando como datos los logs de Squid, de toda la navegación realizada a través del proxy en un intervalo de tiempo, es una de las herramientas más completas en ese sentido. En palabras de su programador: Sarg es un Squid Analysis Report

Squid Graph es un script de perl (Lenguaje de programación) que lee el archivo access.log de su servidor proxy Squid y genera una página web que muestra las estadísticas acerca de su proxy, accesos y transferencias, incluyendo el número de visitas de caché y el porcentaje de solicitudes que fueron atendidas por la memoria caché por sí solo. Squid Graph posee la ventaja de ser rápido y fácil de configurar. La desventaja es que requiere de la línea de comandos para interactuar en lugar de utilizar la web.[5]

ISA Stats es una herramienta para Microsoft ISA Server que permite controlar el uso de internet. Es un sistema analizador de tráficos y colector de estadísticas. Brinda la oportunidad de conocer a los usuarios más activos pues revela cuándo y qué sitios han visitado. Permite reunir la información acerca de los sitios más populares y utilizados de manera activa. El control sobre los usuarios se realiza tanto en tiempo real como en base a los datos recolectados. La herramienta permite realizar la recopilación

[5] Logs: Término anglosajón para referirse al registro oficial de eventos durante un rango de tiempo en particular

[6] Host: Computadoras conectadas a una red, que proveen y utilizan servicios de ella

[7] Html: ***HyperText Markup Language*** (lenguaje de marcado hipertextual)

de estadísticas, el seguimiento de los sitios web utilizados en tiempo real, el bloqueo de sitios indeseados y la administración remota.[6]

ISAWEB es una herramienta diseñada en Cuba por el centro ICID (Instituto Central de Investigación Digital). Se diseñó para un uso racional y controlado del internet en un ámbito empresarial. Esta aplicación permite que un usuario autenticado pueda filtrar información relevante sobre el acceso de internet. La información de tráfico de internet se extrae del servidor de internet "ISA Server", transformando los ficheros donde se almacenan las trazas de la navegación, a una base de datos en SQL Server. Esta es una aplicación web y su interfaz es sencilla, flexible, amigable e intuitiva. Se logra un fácil control individualizado del tráfico de internet con su uso.
Los reportes de la aplicación son:

- **General**: Muestra información general de la búsqueda que se está realizando con los siguientes campos: fecha, IP, host, usuario, cantidad de bytes descargados, cantidad de bytes enviados, URL.
- **Simplificado**: Muestra información general de la búsqueda, pero no haciendo visible todos los campos. Se muestran los siguientes campos: usuario, host, fecha.
- **Favorito**: Muestra un reporte adecuado a la información filtrada con un listado ordenado descendentemente por los sitios más visitados y la cantidad de visitas.
- **Navegación**: Muestra un listado de los usuarios que más sitios han visitado y la cantidad de sitios visitados.
- **Descargas**: Muestra un listado descendente de los usuarios que más volumen de información descarga de internet.[7]

1.2.2 Herramientas utilizadas actualmente en ETECSA para el análisis de los servidores Squid e ISA Server

Sawmill es una potente herramienta de análisis de logs (registros de sucesos/eventos). Está especialmente diseñada para analizar logs de acceso a servidores web, pero puede procesar casi cualquier log. Se ejecuta como un programa CGI[8] en un servidor web. Publica una interfaz gráfica de usuario que puede ser usada desde cualquier navegador para configurar y ejecutar Sawmill, o para ver

[8] CGI (Common Gateway Interface): Interfaz de entrada común que permite a un cliente solicitar datos de un programa ejecutado en un servidor web.

estadísticas de páginas. Las estadísticas son jerárquicas, atractivas y llenas de enlaces que facilitan la navegación. El programa incluye una completa documentación. Sawmill ofrece una gran cantidad de opciones, incluyendo una base de datos persistente, control sobre la apariencia de las páginas de estadísticas y opciones de filtrado sobre el log.[8]

Cyfin Reporter es un analizador de archivo de registro que genera informes sobre el acceso web del empleado. Admite múltiples formatos de archivo de registro y ubicaciones al tiempo que proporciona una visión clara de la actividad en línea de la fuerza laboral.[9]

Estas son algunas de las características claves del Cyfin.

Características funcionales:

- Controla el acceso web del empleado y organiza los datos en informes fáciles de leer.
- Los informes de acceso web son precisos, detallados y útiles.
- Clasifica el uso de la web en 81 categorías diferentes de contenido.
- Genera informes para la detección automática de abuso.
- El planificador de informes se ejecuta y distribuye informes automáticamente.
- Ideal para organizaciones de medio y gran tamaño: escalable a más de 100.000 usuarios.
- Solución sólida que puede manejar grandes cantidades de datos.
- Mejora la productividad, ahorra ancho de banda y evita amenazas de responsabilidad legal.

Características técnicas:

- Monitoriza e informa sobre qué sitios web están consumiendo más ancho de banda.
- Admite más de 100 formatos de registro diferentes y múltiples ubicaciones de archivo de registro.
- Tiene una base de datos interna que almacena y comprime los datos de archivo de registro.
- Actualizaciones de software, soporte técnico gratuito y actualizaciones diarias de la lista de URL.

1.2.3 Mensajería Exchange

Exchange Reporter Plus es una solución de generación de análisis y reportes basada en la web para MS Exchange, brindándole un conocimiento amplio de la infraestructura Exchange de su organización.

Este software no genera reportes planos solamente con propósitos de vigilancia, sino que ofrece un análisis & estadísticas del tráfico de correos, tasa de crecimiento del tamaño de la bandeja de correo, patrones de utilización del servidor y otros datos importantes, utilizando todos los recursos necesarios con los cuales un administrador pueda realizar decisiones informadas para optimizar la eficiencia de la configuración Exchange de su organización. En otras palabras, Exchange Reporter Plus es una encapsulación de todo el "análisis y supervisión" requerido por un administrador Exchange para administrar su infraestructura de comunicación de email eficientemente.[10]

Esta solución juega un papel vital en las siguientes actividades cruciales:

- En la optimización de los recursos para su mejor desempeño.
- En el balanceo de la carga del Servidor Exchange.
- Es un recurso para el análisis de los negocios.
- Es un recurso en la capacidad de planeación.

Exchange Reporter Plus provee más de 70 reportes diferentes de todos los aspectos del entorno de Microsoft Exchange. Incluidos algunos reportes de suma importancia. Algunas de las muchas tareas que pueden realizarse utilizando Exchange Reporter Plus se listan a continuación.

- Llevar un control del número de mensajes enviados y recibidos por cada servidor de Exchange, utilizando los reportes de tráfico.
- Llevar un seguimiento de las estadísticas vitales de las carpetas públicas de Exchange con los reportes correspondientes.
- Generar reportes de Listas de Distribución.

Exchange Reporter Plus presenta estos reportes en un formato fácilmente comprensible, que incluso usuarios con poca experiencia técnica no tendrían problemas en crear y comprender. Además de generar estos reportes, también es muy simple su exportación. Los mismos pueden ser descargados como archivos xls, csv, pdf o html para tareas adicionales.[11]

MailStat es una herramienta que le permite archivar el correo electrónico que se envía y recibe en su Empresa, brindando una variedad de funcionalidades adicionales. Le permite obtener una gran

variedad de estadísticas y reportes sobre el uso del email en su Empresa. Cada email procesado por MailStat recibe un número único que sirve para identificarlo. De este modo es fácil referirse a un email específico por su número, evitando mal entendidos.[12]

En cuestión de segundos MailStat contesta preguntas como:

- ¿Cuáles son las direcciones externas a las que más escribe Juan?
- ¿Cuáles dominios son los que más escriben a nuestra dirección de ventas?
- ¿Cuántos emails envió Pablo a hotmail.com en el día de ayer?

MailStat le permite entre otras cosas:

- Ver los emails que se enviaron y recibieron sobre un mismo tema, organizados como una conversación.
- Ver las respuestas a un determinado correo electrónico.

1.2.4 Herramientas utilizadas actualmente en ETECSA para el análisis del servidor de mensajería

SUECO es una herramienta diseñada en ETECSA, se desarrolló con el objetivo de poder monitorizar el servicio de mensajería que brinda la empresa. Contiene varios reportes de la información que transita a través de los correos.

1.2.5 Empresas dedicadas a brindar servicios de monitoreo

También existen compañías dedicadas a brindar servicios para la gestión de la navegación y la mensajería utilizando herramientas patrocinadas por ellas, a continuación se mencionan algunas de ellas.

SurfControl es una compañía norteamericana especializada en soluciones de filtrado de contenido web y Correo Electrónico. La Tecnología de SurfControl protege todos los puntos de entrada y todos los modos en que los empleados usan internet para su negocio: mensajería instantánea, navegación web, correo electrónico y comunicaciones punto a punto.

Herramientas utilizadas por SurfControl

- **SurfControl E-mail Filter** es una herramienta para la gestión y el control de mensajes de correo electrónico. Permite evitar la mayoría de abusos y amenazas que afectan a este tipo de comunicaciones y posibilita la implementación de políticas de uso del e-mail basadas en contenidos, direcciones de origen, listas negras, etc. Además también permite monitorizar y auditar, copiar, retener y generar alertas cuando se detectan mensajes que cumplen con ciertos criterios establecidos.[12]

  Características Principales

  - Incrementa la productividad de los usuarios
  - Impide emplear el tiempo en actividades no relacionadas con el trabajo
  - Optimiza el ancho de banda
  - Previene riesgos de seguridad
  - Evita situaciones embarazosas, incómodas o que podrían resultar en riesgos legales
  - Filtrado basado en reglas que incorporan análisis del contenido y direcciones de e-mail utilizadas
  - Análisis del contenido basado en herramientas léxicas, 15 categorías de contenido, 6 diccionarios con diferentes idiomas y un agente categorizador de contenido en tiempo real.
  - Monitorización en tiempo real de todos los mensajes
  - Políticas de acceso flexibles e inteligentes
  - Calendarización de acciones (Scheduling)
  - Informes predefinidos
  - Auditoria de mensajes, copia y suspensión en la entrega de los mismos

- **SurfControl web Filter** es una herramienta para la gestión y el control de la navegación web. SurfControl web Filter pone en sus manos la tecnología de filtrado y generación de informes más potente, garantiza un uso apropiado de internet y permite:[13]

  - Incrementar la productividad de los usuarios
  - Imposibilitar el empleo del tiempo en actividades no relacionadas con el trabajo
  - Optimizar el ancho de banda
  - Prevenir riesgos de seguridad

Características Principales:

- Soporte para un gran número de plataformas y entornos
- Filtrado preciso en base a múltiples parámetros (usuarios, grupos de usuarios, dominios de Windows, horarios, direcciones IP,... etc.)
- Control del tráfico y ancho de banda
- Monitorización en tiempo real
- Políticas de acceso flexibles e inteligentes
- Calendarización de acciones (Scheduling)
- Más de 55 informes predefinidos accesibles en formato pdf, Word y HTML
- Listas de contenidos clasificados en más de 40 categorías
- Determinación del contenido en tiempo real
- Arquitectura escalable y distribuida

Spamina es una compañía española que desarrolla y ofrece soluciones innovadoras de seguridad de email y tráfico web con el fin de garantizar a los administradores de sistemas y usuarios, la protección, gestión y el control total de sus comunicaciones.

Herramientas utilizadas por Spamina

- **Cloud web Security** es un servicio de filtrado de tráfico web (Web Filtering), incorpora numerosas funcionalidades como el control de tráfico sobre cualquier plataforma de acceso incluyendo smartphones, consola central de gestión en tiempo real con informes personalizados. Otras de las características clave de “Cloud web Security" son su rápido y sencillo despliegue, su facilidad de manejo y flexibilidad para que cada usuario tenga establecidas políticas de uso de internet personalizadas y dinámicas, el control del acceso a sitios y aplicaciones que consuman gran cantidad de ancho de banda y la facilidad de uso para los administradores. Además, esta solución proporciona completos y fiables informes sobre el uso de la web en las empresas y por parte de los trabajadores remotos lo que facilita la adecuación dinámica de los parámetros al cumplimiento de las políticas de seguridad.

- **Cloud Email Firewall** es una herramienta para gestionar la seguridad de seguridad del correo electrónico. Cloud Email Firewall destaca el acceso instantáneo para los administradores TI a

información de filtrado, con datos sintetizados de una forma muy visual y tablas resumen que se presentan a través del tablero de mando y generación de reportes de filtrado personalizados. Además, los administradores tienen ahora la posibilidad de realizar acciones masivas sobre los resultados de las búsquedas de logs de correos.[14]

1.2.6 Análisis de las soluciones estudiadas

Las herramientas estudiadas ofrecen muchas funcionalidades, pero no son factibles para dar solución al problema planteado. Su primera desventaja es que todas se especializan en un servicio telemático específico. La necesidad de ETECSA es utilizar una herramienta que permita unificar los dos servicios en un mismo sistema y de esta manera facilitar el proceso de toma de decisiones. Ninguna de estas soluciones cumple con exactitud los requisitos que plantea la empresa, como es el caso de Sarg, una herramienta muy potente generando reportes pero solo lo hace dependiendo de la información extraída de los logs del Squid, lo mismo sucede con Squid Graph. La herramienta ISA Stats posee un gran número de funcionalidades y tipos de reportes pero se enfoca solamente en la navegación ISA Server y es privativa por lo que implica un costo elevado su adquisición. ISAWEB es una herramienta dedicada solamente al monitoreo de la navegación ISA Server.

La empresa actualmente para darle seguimiento al servicio de navegación web utiliza las herramientas Sawmill y Cyfin Reporter. Estas aplicaciones tienen como características que pueden analizar información de cualquier servidor web, la dificultad es que además de no tener incluido el servicio de mensajería electrónica, no satisfacen las necesidades de la empresa pues los reportes, no muestran la información necesaria para un efectivo proceso de toma de decisiones, con esta herramienta los especialistas no conocen, por ejemplo, los usuarios a los que el proxy deniega por entrar en una dirección web no permitida. También se realizó un estudio de las herramientas dedicadas al monitoreo de la mensajería. De estas herramientas Exchange Reporter Plus y MailStat, son privativas y su utilización representaría un costo considerable y SUECO actualmente utilizada por ETECSA, presenta deficiencia de información en cuanto a los reportes por ejemplo no da cantidades de mensajes enviados que genera.

1.3 Almacén de datos

Cuando se habla de almacén de datos no se puede omitir a las personalidades que han dejado su historia, quienes son conocidos por sus grandes aportes en este tema, como lo son Ralph Kimball, considerado el principal promotor del enfoque dimensional y Bill Inmon. Inmon reconocido como el padre de los almacenes de datos, porque fue el primero en escribir un libro sobre este tema y es el que le da la definición:

"Los *almacenes de datos* o Data Warehouse (DW) por su nombre en inglés, es una colección de datos orientados a temas, integrados, no-volátiles y variante en el tiempo, organizados para soportar necesidades empresariales".

Un almacén de datos se crea al extraer datos desde una o más bases de datos de aplicaciones operacionales. La información extraída es transformada para eliminar inconsistencias y resumir si es necesario y luego ser cargadas en el almacén. El proceso de transformar, crear el detalle de tiempo variante, resumir y combinar los extractos de datos, ayudan a crear el ambiente para el acceso a la información institucional. Este nuevo enfoque ayuda a las personas individualmente, en todos los niveles de la empresa, a efectuar su toma de decisiones con más responsabilidad.[15]

Un almacén de datos se puede caracterizar partiendo de una comparación, de cómo los datos de una institución, al almacenarlos en un almacén de datos, difieren de los datos operacionales usados por las aplicaciones de producción, donde las instituciones hacen sus operaciones cotidianas de los procesos que se realizan en la misma.

Base de datos operacional	**Almacén de datos**
Datos operacionales	Datos del negocio para información
Orientado a la aplicación	Orientado al sujeto
Actual	Actual + histórico
Detallada	Detallada + más resumida
Cambia continuamente	Estable

Tabla 1: Diferencias en el almacenamiento de datos entre BD Operacionales y Almacén de datos

Sus **principales objetivos** son**:**

1. Proporcionar una vista total e integrada de la empresa.

2. Permitir que la información actual e histórica de la empresa esté disponible de manera práctica para el análisis.
3. Permitir ejecutar las operaciones de soporte a las decisiones sin afectar los sistemas operacionales.
4. Garantizar la consistencia de la información.
5. Constituir una fuente flexible e interactiva de información estratégica.

1.3.1 Las principales características de un almacén de datos son: [16]

- Orientado a temas
- No volátil
- Variante en el tiempo
- Integrado

Orientado a temas: La información es clasificada de acuerdo a los aspectos que son más importantes para la empresa. Existe una gran diferencia en la forma de diseñar la estructura para almacenar los datos en un DW con respecto a los clásicos procesos operacionales que están orientados a las aplicaciones. Los procesos que se llevan a cabo mediante las aplicaciones son los que responden al manejo de todos los datos de las operaciones realizadas en la empresa, mientras que un almacén de datos sólo tiene en cuenta aquella información de interés para una oportuna toma de decisiones. Por ejemplo, para una institución financiera el ambiente operacional se diseña a través de funciones tales como préstamos, ahorros, tarjetas bancarias y depósitos, porque son muy importantes para sus operaciones productivas, sin embargo, no se toma en cuenta en un almacén de datos ya que carecen de valor para la toma de decisiones. En este caso un ambiente de almacén de datos se organiza teniendo en cuenta la información que se relacione con el cliente, vendedor, producto y actividad. La principal ventaja que tiene el DW al estar orientado a temas es que los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.[16] ¡Error! No se encuentra el origen de la referencia. (Ver anexos)

Variante en el tiempo: Esta característica básica de un almacén de datos es muy diferente de la información que se encuentra en un ambiente operacional, en el cual los datos se obtienen en el

momento de acceder. En un almacén de datos como la información es solicitada en cualquier momento, los datos encontrados en el depósito se llaman de "tiempo variante". Esto es precisamente una de las principales ventajas de un DW, los datos que se almacenan tienen asociado una variable tiempo, esta cualidad no se encuentra en la base de datos operacionales, ya que simplemente los datos se insertan sin tener, necesariamente, en cuenta una variable tiempo. Además de que el almacén de datos al contar con el sello del tiempo permite acceder al dato histórico de la información que almacena, sin embargo en una base de datos operacionales los datos históricos son de poco uso. El almacenamiento de datos históricos, es lo que permite al DW desarrollar pronósticos y análisis de tendencias y patrones, a partir de una base estadística de información.[16]

El tiempo variante se puede ver de diversas maneras:

1° La información representa los datos sobre un horizonte largo de tiempo (de 5-10 años). El horizonte de tiempo representado para el ambiente operacional es mucho más corto (desde valores actuales hasta 60-90 días), debido al diseño de aplicaciones rígidas.

2° La segunda manera en la que se muestra el tiempo variante en el almacén de datos está en la estructura clave. Cada estructura clave en el almacén de datos contiene, implícita o explícitamente, un elemento de tiempo como día, semana y mes. En ocasiones, el elemento de tiempo existirá implícitamente, como el caso en que un archivo completo se duplica al final del mes, o al cuarto.

3° La tercera manera en que aparece el tiempo variante es cuando la información del almacén de datos, una vez registrada correctamente, no puede ser actualizada. La información del almacén de datos es, para todos los propósitos prácticos, una serie larga de "snapshots" (vistas instantáneas). En caso de que las vistas instantáneas de los datos se han tomado incorrectamente, entonces pueden ser cambiados, pero de no ser así, ellos no son alterados una vez hechos.[16] ¡Error! No se encuentra el origen de la referencia. (Ver anexos)

Integrado: Los datos almacenados en el almacén de datos deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La integración de datos se muestra de muchas maneras: en convenciones de nombres consistentes, en la medida uniforme de variables, en la codificación de estructuras consistentes, en atributos físicos de los datos consistentes, fuentes múltiples y otros. Cualquiera que sea la forma del diseño, el resultado es el mismo: la información necesita ser almacenada en el almacén de datos en un

modelo aceptable y singular, aun cuando los sistemas operacionales subyacentes almacenen los datos de manera diferente. [16]¡Error! No se encuentra el origen de la referencia. (Ver anexos)

No volátil: Es necesario para la toma de decisiones que la información que se utilice para ello sea estable. Los datos en los procesos operacionales cambian de un momento a otro mientras que en un DW una vez almacenados no se pueden cambiar. La actualización (insertar, borrar y modificar), se hace regularmente en el ambiente operacional sobre una base de registro por registro, en este aspecto también difiere el DW, ya que la manipulación sobre los datos que almacena es mucho más simple porque solamente puede hacer dos tipos de operaciones: la carga inicial de los datos y el acceso a los mismos. La mayoría de los datos que se extraen de la fuente se alteran física y radicalmente cuando se mueven al depósito, no es la misma información que reside en el ambiente operacional desde el punto de vista de integración.[16] ¡Error! No se encuentra el origen de la referencia. (Ver anexos)

1.3.2 Ventajas de los almacenes de datos

Entre las ventajas de los almacenes de datos se encuentran:

- Transforma datos orientados a las aplicaciones en información orientada a la toma de decisiones.
- Integra y consolida diferentes fuentes de datos (internas y/o externas) y departamentos empresariales, que anteriormente formaban islas, en una única plataforma sólida y centralizada.
- Provee la capacidad de analizar y explotar las diferentes áreas de trabajo y de realizar un análisis inmediato de las mismas.
- Permite reaccionar rápidamente a los cambios del mercado.
- Aumenta la competitividad en el mercado.
- Elimina la producción y el procesamiento de datos que no son utilizados ni necesarios, producto de aplicaciones mal diseñadas o ya no utilizadas.
- Mejora la entrega de información, es decir, información completa, correcta, consistente, oportuna y accesible. Información que los usuarios necesitan, en el momento adecuado y en el formato apropiado.[16]

1.3.3 Mercados de datos

Un mercado de datos o Data Mart (DM) por su nombre en inglés, es la implementación de un DW con alcance restringido a un área funcional, problema en particular, departamento, tema o grupo de necesidades en el negocio.[16] Un DM puede ser alimentado desde los datos de un almacén de datos, o integrar por sí mismo un compendio de distintas fuentes de información. Tienen como principales ventajas que son simples de implementar, además de que conllevan poco tiempo de construcción y despliegue.

Muchos almacenes de datos comienzan siendo un mercado de datos, para, entre otras cosas, minimizar riesgos, pero una vez que se han implementado y explotado con éxito su alcance se va ampliando a medida que pasa el tiempo, hasta conformar un DW.

Teniendo en cuenta las operaciones que se desean o requieran desarrollar, los DM pueden adoptar los siguientes enfoques: [16]

- **Top-Down:** primero se define el DW y luego se desarrollan, construyen y cargan los DM a partir del mismo. El DW es cargado a través de procesos ETL y luego este alimenta a los diferentes DM, cada uno de los cuales recibirá los datos que correspondan al tema o departamento que traten.

Esta forma de implementación cuenta con la ventaja de no tener que incurrir en complicadas sincronizaciones de hechos, pero requiere una gran inversión y una gran cantidad de tiempo de construcción. ¡Error! No se encuentra el origen de la referencia. (Ver anexos)

- **Bottom-Up:** en esta arquitectura, se definen previamente los DM y luego se integran en un DW centralizado. Los DM se cargan a través de procesos extracción transformación y carga (ETL), los cuales suministrarán la información adecuada a cada uno de ellos. En muchas ocasiones, los DM son implementados sin que exista el DW, ya que tienen sus mismas características pero con la particularidad de que están enfocados en un tema específico. Luego de que hayan sido creados y cargados todos los DM, se procederá a su integración con el depósito.

  La ventaja que trae aparejada este modelo es que cada DM se crea y pone en funcionamiento en un corto lapso de tiempo y se puede tener una pequeña solución a un costo no tan elevado. Luego que todos los DM estén puestos en marcha, se puede decidir si construir el DW o no. **¡Error! No se encuentra el origen de la referencia.** (Ver anexos)

1.4 Bases de datos OLTP y OLAP

Los sistemas de Procesamiento Transaccional y Analítico en Línea conocidos como OLTP y OLAP respectivamente, son conceptos a analizar para un mejor entendimiento y posterior desarrollo de los almacenes de datos, es de suma importancia reconocer cuándo se está en presencia de uno u otro sistema.

1.4.1 Sistemas OLTP- On-Line Transactional Processing

OLTP (Procesamiento de transacción en línea, base de datos orientados al procesamiento de transacciones).

Los sistemas de OLTP son los sistemas operacionales que capturan las transacciones de un negocio y las persisten en estructuras relacionales llamadas base de datos.

Las características principales de los sistemas OLTP son:

- Realizan transacciones en tiempo real del proceso de un negocio, con lo cual los datos almacenados cambian continuamente. Los sistemas OLTP en sus transacciones conducen procesos esenciales del negocio.
- Los sistemas OLTP son los responsables del mantenimiento de los datos, ya sea agregando datos, realizando actualizaciones o bien eliminándolos.
- Las estructuras de datos deben estar optimizadas para validar la entrada de los mismos y rechazarlos si no cumplen con determinadas reglas de negocio.
- Para la toma de decisiones, proporciona capacidades limitadas ya que no es su objetivo, por lo tanto no es prioridad en su diseño. Si se quisiera obtener determinada información histórica relativa al negocio consultando un sistema OLTP, se produciría un impacto negativo en el funcionamiento del sistema.[7]

1.4.2 Sistemas OLAP- On-Line Analytical Processing

OLAP (Procesamiento analítico en línea, Base de datos orientados al procesamiento analítico)

Los sistemas OLAP proporcionan una alternativa a los sistemas transaccionales, ofreciendo una visión de los datos orientada hacia el análisis y una rápida y flexible navegación por estos.

Las siguientes son características que la tecnología OLAP posee:

- Las bases de datos de OLAP tienen un esquema que está optimizado para que las preguntas realizadas por los usuarios sean respondidas rápidamente.
- Las preguntas que se le hacen a un OLAP, deben permitir un uso interactivo con los usuarios.
- Los cubos de OLAP almacenan varios niveles de datos conformados por estructuras altamente optimizadas que responden a las expectativas de negocio de la empresa.
- Un sistema OLAP está preparado para realizar informes complejos de una manera simple.
- OLAP proporciona una vista de datos multidimensional. Los cubos proporcionan una vista de los datos multidimensional que se extiende más allá del análisis de dos dimensiones que puede proporcionar una simple planilla de cálculo utilizada como tal.
- Los usuarios pueden cambiar fácilmente las filas, las columnas y las páginas en informes de OLAP, pudiendo leer la información de la manera que se crea más conveniente para el análisis.[7]

1.5 Modelado multidimensional

En el modelado de datos es donde se encuentra la principal diferencia entre los sistemas operacionales y el DW: cada uno de ellos es sostenido por un modelo de datos diferente. Los sistemas operacionales se sustentan en el Modelo Entidad Relación (MER) y el DW trabaja con el modelo multidimensional. El modelo dimensional se describe en el año 1996 por Ralph Kimball, partiendo de la visión multidimensional que los usuarios tienen de los datos empresariales cuando se enfrentan a ellos con propósito de análisis (de análisis multidimensional –OLAP– en concreto), el modelo dimensional es uno de los más utilizados para la construcción de los almacenes de datos. La estructura básica de un DW para el modelo multidimensional está definida por dos elementos: esquemas y tablas. Cada modelo dimensional se compone de una tabla de hechos central, que contienen los valores de las medidas de negocio y un conjunto de tablas pequeñas llamadas dimensiones, que contienen el detalle de los valores que se encuentran en la tabla de hechos. Los elementos de estas tablas se pueden definir como: [16]

Hechos: Los hechos son la representación en el almacén de datos de los procesos de negocio de la organización, son datos instantáneos en el tiempo, que son filtrados, agrupados y explorados a través

de condiciones definidas en las tablas de dimensiones. Las tablas de hechos contienen los hechos, medidas o indicadores que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones. Idealmente está compuesta por valores numéricos, continuamente evaluados y aditivos. Los hechos se podrán reconocer porque siempre tienen asociada una fecha y porque una vez registrados no se modifican ni se eliminan (para no perder la historia). [16]

Dimensiones: Son la representación en el almacén de datos de un punto de vista para los hechos de cierto proceso de negocio, definen cómo están los datos organizados lógicamente y proveen el medio para analizar el contexto del negocio. Contienen datos cualitativos. Representan los aspectos de interés, mediante los cuales los usuarios podrán filtrar y manipular la información almacenada en la tabla de hechos, es decir son las que alimentan a la tabla de hechos. La tabla de dimensión Tiempo es obligatoria, la definición de granularidad y estructuración de la misma depende de la dinámica del negocio que se esté analizando. Los atributos dimensionales son un rol determinante en un DW. Ellos son la fuente de todas las necesidades que debieran cubrirse. Esto significa que la base de datos será tan buena como lo sean los atributos dimensionales, mientras más descriptivos, manejables y de buena calidad, mejor será el almacén de datos.[16]

Cubos multidimensionales

"Los cubos son elementos claves en OLAP (On Line Analytical Processing), una tecnología que provee rápido acceso a datos en un almacén de datos. Los cubos proveen un mecanismo para buscar datos con rapidez y tiempo de respuesta uniforme independientemente de la cantidad de datos en el cubo o la complejidad del procedimiento de búsqueda."[17]

Los objetos más importantes que se pueden incluir en un cubo multidimensional, son los siguientes: [16]

Indicadores: sumarizaciones que se efectúan sobre algún hecho o expresiones, pertenecientes a una tabla de hechos, que serán incluidos en algún cubo multidimensional, con el fin de analizar los datos almacenados en el DW.[16]

Atributos: campos o criterios de análisis, pertenecientes a tablas de dimensiones, que se utilizarán para analizar los indicadores dentro de un cubo multidimensional.[16]

Jerarquías: representa una relación lógica entre dos o más atributos, pertenecientes a un cubo multidimensional; siempre y cuando posean su correspondiente relación “padre-hijo”. Las mismas están compuestas por dos o más niveles y pueden existir varias en un mismo cubo.[16]

1.5.1 Esquemas dimensionales

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

Esquema en estrella (Star Scheme): Un esquema en estrella se caracteriza por tener una tabla central de hechos rodeada por tablas de dimensiones que contienen información desnormalizada de los hechos.

Los siguientes elementos son característicos en un esquema en estrella:

• El centro del esquema es la tabla de hechos, contiene las métricas o medidas del negocio.

• Las puntas de la estrella son las tablas de dimensiones. Estas puntas son utilizadas para describir la información existente en un proceso específico del negocio y proveen el contexto a los datos numéricos. [18] ¡Error! No se encuentra el origen de la referencia. (Ver anexos)

Esquema copo de nieve (Snowflake Scheme): El modelo copo de nieve es una variación o derivación del modelo estrella. En este modelo la tabla de hechos deja de ser la única relacionada con otras tablas ya que existen otras tablas que se relacionan con las dimensiones y que no tienen relación directa con la tabla de hechos. El modelo fue concebido para facilitar el mantenimiento de las dimensiones, sin embargo esto hace que se vinculen más tablas a las secuencias SQL, haciendo la extracción de datos más difícil así como vuelve compleja la tarea de mantener el modelo.[18] ¡Error! No se encuentra el origen de la referencia. (y)

Esquema constelación o copo de estrellas (Starflake Scheme): Propone una combinación de los esquemas de federación y copos de nieve manteniendo las tablas de hechos normalizadas y las tablas de dimensiones desnormalizadas; es el modelo que brinda mayor flexibilidad al diseñador de un sistema informacional y es útil para dar respuesta a las necesidades de integración de aplicaciones, ya que su construcción se debe basar en el aprovechamiento adecuado de las características de cada tipo de modelo.[19] ¡Error! No se encuentra el origen de la referencia.

El esquema de modelado dimensional que se obtiene al modelar todos los procesos que engloba el Mercado de Datos, es la Constelación. Esto se debe a que cada proceso posee un cubo formando un modelo de estrella.

1.6 Arquitectura de un almacén de datos

"Una Arquitectura Data Warehouse (Data Warehouse Architecture - DWA) es una forma de representar la estructura total de datos, comunicación, procesamiento y presentación, que existe para los usuarios finales que disponen de una computadora dentro de la empresa."[20]

Arquitectura conceptual de los datos del mercado de datos

La arquitectura de un almacén de datos está formada por diversos elementos que interactúan entre sí, donde es conveniente tener en consideración los diferentes entornos por los que han de pasar los datos en su camino hacia el DW o DM de destino.[16]

1.7 Metodologías de desarrollo de un almacén de datos

Las metodologías de desarrollo de un almacén de datos, tienen como principal objetivo guiar todo el ciclo de vida para su construcción. En la actualidad existen varias metodologías a seguir, pero de manera general pueden ser englobadas en dos, las metodologías propuestas por Bill Inmon y Ralph Kimball.

1.7.1 Metodología de Ralph Kimball – Ciclo de vida dimensional

La metodología propuesta se basa en lo que Kimball denomina Ciclo de Vida Dimensional del Negocio, esta metodología se ha convertido en el estándar de facto en el área de apoyo a las decisiones empresariales.

La construcción de una solución de almacenes de datos con inteligencia de negocio es muy compleja, con esta metodología se simplifica esta complejidad y como respuesta se divide el ciclo de vida de su desarrollo en una serie de etapas, el flujo central de dichas etapas se describe a continuación:[21]

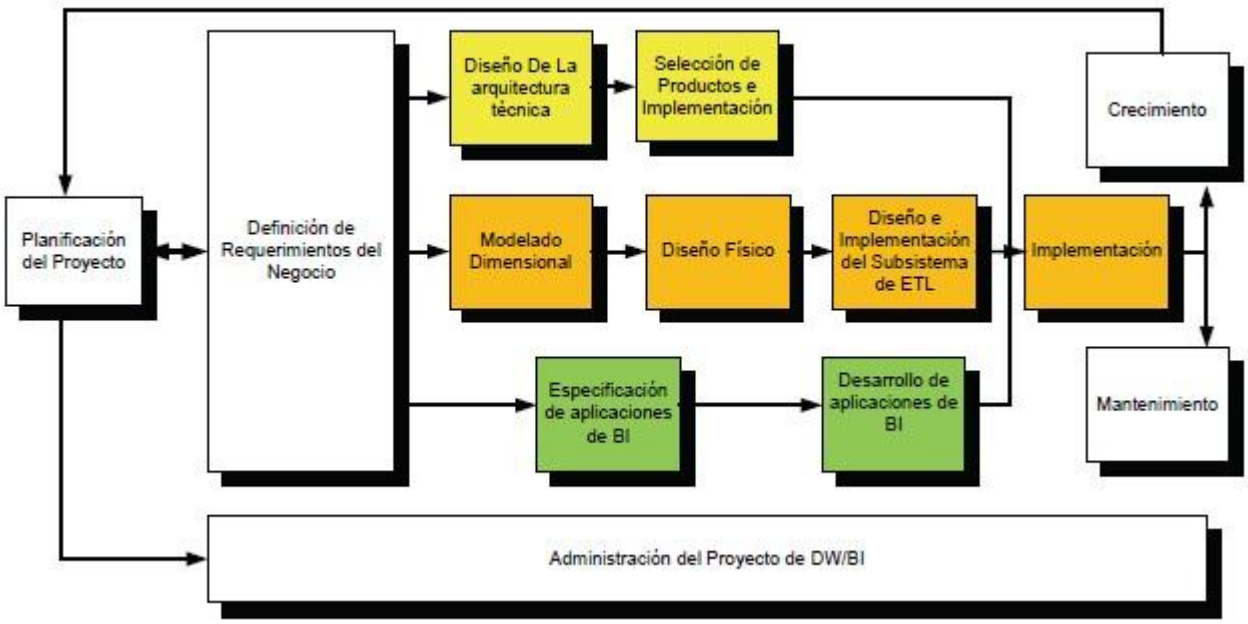


Figura 1: Etapas que componen la metodología de Kimball.

Planificación

Esta es la primera tarea del ciclo de vida de la metodología, donde se determina el propósito, objetivos específicos y el alcance del proyecto, los principales riesgos y una aproximación inicial a las necesidades de información. Además incluye un conjunto de acciones típicas de un plan de proyecto.

Análisis de requerimientos

En el análisis de los requerimientos se realizan entrevistas, se hacen encuestas y todo tipo de métodos para conocer a profundidad sobre el negocio en cuestión, los competidores, la industria y los clientes del mismo. Los requerimientos del negocio deben determinar el alcance del almacén de datos.

Modelado dimensional

En esta tarea es donde se diseña el modelo lógico del almacén de datos, se identifican las tablas de hechos, se eligen las dimensiones, se establece el nivel de granularidad y se realiza el modelo gráfico de alto nivel.

Diseño físico

El diseño físico se focaliza sobre la selección de las estructuras necesarias para soportar el diseño físico, responde a cuál es la estructura a seguir, cómo se va a configurar, etc.

Diseño e Implementación del Sistema de Extracción, Transformación y Carga (ETL)

El sistema de Extracción, Transformación y Carga (ETL) es la base sobre la cual se alimenta el almacén de datos. En este paso se extraen los datos de los sistemas de origen de datos, se aplican diferentes reglas para aumentar la calidad y consistencia de los mismos, finalmente se realiza la carga de los datos para poblar el almacén de datos.

Despliegue

El despliegue representa la convergencia de la tecnología, los datos y las aplicaciones del usuario final accesibles desde el escritorio de los usuarios de negocios. Una amplia planificación es necesaria para garantizar el buen funcionamiento de todos los elementos que intervienen. La educación de los usuarios de negocio integrando todos los aspectos de la convergencia, debe ser desarrollada y entregada. Además, el apoyo a los usuarios y la comunicación o las estrategias de retroalimentación debe quedar establecido antes de que los usuarios de negocios puedan tener acceso al almacén de datos.

Mantenimiento y crecimiento

Queda mucho trabajo a raíz de la implementación inicial del almacén de datos, por lo que se hace necesario seguir centrándose en los usuarios del negocio, proporcionándoles apoyo y educación. También es necesario asegurar que los procesos y procedimientos están en su lugar para el funcionamiento continuo y eficaz del almacén. La aceptación y desempeño del almacén de datos debe ser medido a través del tiempo, finalmente el plan de mantenimiento debe incluir una estrategia de comunicación de amplio alcance. **Figura 11**(Ver Anexo)

- **Justificación de la metodología propuesta**
  - ✓ La técnica de Kimball posee una gran cantidad de documentación y generalmente se puede encontrar una respuesta a casi todas las problemáticas que puedan aparecer.
  - ✓ Claridad de las actividades a realizar por cada rol propuesto.
  - ✓ Esta metodología de dividir el mundo de BI entre el hecho y las dimensiones es muy eficaz y conduce a una solución completa en un tiempo razonable.

- ✓ Es iterativo, donde se construye una pieza a la vez (mercado de datos) garantizando mayor velocidad de respuesta a los clientes.
- ✓ La forma de almacenar la información es de fácil entendimiento por parte del usuario lo que permite mayor comprensión para el análisis de los datos que se encuentran integrados.
- ✓ Es una metodología resistente y adaptable ante los cambios.

1.8 Herramientas utilizadas para la solución

1.8.1 Herramientas de creación de un almacén de datos

Para la construcción de un almacén de datos se necesita utilizar una serie de herramientas informáticas capaces de asistir y facilitar el trabajo. "La plataforma de Inteligencia de Negocio Pentaho cubre muy amplias necesidades de análisis de los datos y de los informes empresariales. Las soluciones de Pentaho están escritas en Java y tienen un ambiente de implementación también basado en Java. Eso hace que Pentaho sea una solución muy flexible para cubrir una amplia gama de necesidades empresariales, tanto las típicas como las sofisticadas y específicas al negocio".[22]

- **Plataforma BI**

Se optó por utilizar la suite Pentaho Open Source Business Intelligence como plataforma BI. Esto es debido a que es una alternativa de licencia libre. Pentaho funciona sobre cualquier navegador, permite gestionar usuarios, colgar documentación y finalmente, posee una versión de demostración que permite aprender rápido el manejo de la plataforma, reduciendo la curva de aprendizaje, contando con ejemplos de los cuales se puede partir. La documentación para consultar es basta. Tiene una interfaz amigable para los usuarios, la que es posible editar para ajustarse más a las necesidades de estos.

- **Herramientas para la integración de datos**

Pentaho Data Integration 4.1.0 (Kettle)

Pentaho Data Integration, es un proyecto belga de código abierto, ahora adoptado por Pentaho BI, que incluye un grupo de herramientas para realizar el proceso de ETL. Uno de sus objetivos es que dicho proceso sea más fácil de generar, mantener y desplegar. Kettle está compuesto por cuatro herramientas: SPOON, PAN, CHEF y KITCHEN.

- ✓ SPOON: permite diseñar de forma gráfica las transformaciones ETL.

- ✓ PAN: ejecuta un conjunto de transformaciones diseñadas con SPOON.
- ✓ CHEF: permite diseñar la carga de datos incluyendo un control de estado de los trabajos.
- ✓ KITCHEN: permite ejecutar los trabajos batch diseñados con CHEF.

El uso de esta herramienta permite evitar grandes cargas de trabajo manual frecuentemente difícil de mantener y de desplegar. Además, es una herramienta que permite definir transformaciones de forma gráfica, interconectando bloques que tienen diversas funciones. Es extremadamente versátil, ya que se tienen bloques que permiten leer y escribir de cualquier BD, fichero Excel, Access y otros que permiten operar con los campos renombrando, calculando campos en función de otros, mapeando valores, realizando búsquedas auxiliares en BD y normalizando los datos de distintas filas en una sola. Las transformaciones que se hacen con el Kettle se guardan en un fichero con extensión ktr que luego puede ser ejecutado mediante líneas de comandos o un fichero batch.[23]

- **Herramienta para la creación de los cubos**

Pentaho Schema Workbench 3.0

Es una herramienta para el desarrollo y prueba de cubos OLAP de forma visual. La definición del XML no es extremadamente compleja, pero en la práctica resulta engorroso recordar cada uno de los elementos junto a sus atributos y sub-elementos tal y como se encuentran en el almacén. Con esta aplicación, se puede configurar una conexión con el modelo físico, para luego elaborar el esquema lógico de manera simple y efectiva. Para ello la herramienta ofrece un editor de esquemas con la fuente de datos subyacente para su validación.[24]

- **Herramienta para reportes y estadísticas**

Pentaho BI Server

Con esta herramienta se suministra soporte e infraestructura para crear soluciones de inteligencia de negocio. Proporciona servicios básicos además de incluir autenticación, registro, auditoría y servicios web. Incorpora un motor de solución que integra reportes, análisis, tableros de comandos y componentes de minería de datos. Funciona como un sistema basado en administración web de informes, el servidor de integración de aplicaciones y un motor de flujo de trabajo ligero (secuencias de acción). Además, está diseñada para integrarse fácilmente en cualquier proceso de negocio. Permite que puedan ejecutarse los informes y aplicaciones, se puede usar como base para construir un sistema propio de inteligencia de negocios.[25]

1.8.2 Entorno de desarrollo

Visual Studio 2010 Professional

Es un entorno integrado que simplifica la creación, depuración e implementación de aplicaciones. Da rienda suelta a tu creatividad y lleva tu visión a la vida con superficies de diseño potentes y métodos innovadores de colaboración para los desarrolladores y diseñadores. Trabaja dentro de un entorno personalizado, dirigido a un número cada vez mayor de plataformas, incluidas las aplicaciones de Microsoft SharePoint y acelera el proceso de codificación mediante el uso de las habilidades existentes. Soporte integrado para la primera prueba de desarrollo y las nuevas herramientas de depuración te permiten encontrar y solucionar errores de forma rápida y sencilla para garantizar soluciones de alta calidad.[11]

1.8.3 Gestor de base de datos

PostgreSQL es un sistema gestor de base de datos objeto-relacional libre, liberado bajo la licencia BSD (del inglés Berkeley Software Distribution). Como muchos otros proyectos Open Source, el desarrollo de PostgreSQL no es manejado por una sola compañía, sino que es dirigido por una comunidad de desarrolladores y organizaciones comerciales las cuales trabajan en su desarrollo, dicha comunidad es denominada el PostgreSQL Grupo Global de Desarrollo (PGDG), sus siglas en inglés se definen como: PostgreSQL Global Development Group. PostgreSQL ha tenido una larga evolución, comenzando con el proyecto Ingres en la Universidad de Berkeley. Este proyecto, liderado por Michael Stonebraker, fue uno de los primeros intentos en implementar un motor de base de datos relacional.[26]

Ventajas encontradas en PostgreSQL:

- ✓ Soporta lenguajes: PHP, C, C++, Perl y Python.
- ✓ Drivers: ODBC, JDBC y.Net.
- ✓ Soporta: triggers, procedimientos almacenados, funciones, secuencias, relaciones, reglas, tipos de datos definidos por el usuario, vistas y vistas materializadas.
- ✓ Soporte de tipos de datos de SQL92, SQL99 y SQL2003.
- ✓ Soporte de protocolo de comunicación encriptado por SSL.
- ✓ Máximo de bases de datos: ilimitado.
- ✓ Máximo de tamaño de tabla: 32 TB.

- ✓ Máximo de tamaño de registro: 1.6 TB.
- ✓ Máximo de tamaño de campo: 1 GB.
- ✓ Máximo de registros por tabla: ilimitado.
- ✓ Máximo de campos por tabla: 250 a 1600 (depende de los tipos usados).
- ✓ Máximo de índices por tabla: ilimitado.
- ✓ Número de lenguajes en los que se puede programar funciones: aproximadamente 10 (pl/pgsql, pl/java, pl/perl, pl/python, tcl, pl/php, C, C++ y Ruby).

1.8.4 Metodología XP

La metodología XP perteneciente al grupo de metodologías ágiles, se basa en la retroalimentación continua entre el cliente y el equipo de desarrollo y la comunicación fluida entre todos los participantes, además utiliza las Historias de Usuario (HU) para especificar los requisitos del software.[27] Esta metodología que fue seleccionada para guiar el proceso de desarrollo de la herramienta de apoyo, ya que está pensada para equipos de desarrollo pequeños y un ciclo de desarrollo corto.

1.9 Conclusiones parciales

En el presente capítulo se ha realizado un estudio sobre los basamentos teóricos necesarios para la construcción de un mercado de datos. El mismo inició con el estudio del arte, donde se evidenció que ninguno de los sistemas analizados es factible para darle solución a la problemática propuesta. También se realizó un análisis sobre los diferentes elementos que se pueden utilizar para la creación del mercado de datos y se identificaron cuáles de estos se emplearán en su desarrollo: como metodología de desarrollo, Kimball, para la creación del mercado, la suite de Pentaho y como gestor de base de datos PostgreSQL.

CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS

Introducción

En este capítulo se presenta el análisis realizado al monitoreo actual de los eventos de navegación y mensajería que se realiza en el departamento de seguridad informática en la empresa ETECSA, para realizar el diseño del mercado de datos siguiendo los pasos que plantea la metodología Kimball. Se realiza un estudio de los diferentes procesos que se realizan y para ello se hace una breve descripción de los mismos, así como de los requisitos de información y de la herramienta de apoyo utilizada para estructurar los datos obtenidos de las fuentes de origen. Además de identificar las dimensiones y los hechos necesarios para diseñar el almacén de datos.

2.1 Procesos del negocio

- **Análisis de los eventos generados por los servicios de navegación web y mensajería**

El análisis de los eventos generados por los servicios de navegación web y mensajería es una tarea realizada por el departamento de seguridad informática de la empresa ETECSA. La institución como proxy de navegación utiliza las herramientas Squid e ISA Server y para la mensajería utiliza Microsoft Exchange. Cada una de estas herramientas genera un archivo donde guardan todos los eventos creados a partir de las peticiones realizadas por los usuarios de la empresa. Para lograr un efectivo seguimiento de dichos eventos, se hace necesario analizar los reportes que generan las herramientas de los servicios de navegación web y mensajería y así tomar las decisiones que garanticen el bienestar de la empresa. Para un mayor entendimiento de esta acción se divide el proceso en 3 subprocesos los cuales son descritos a continuación.

- **Análisis de los eventos generados por el servicio de navegación Squid**

Squid. Es un programa de software libre que implementa un servidor proxy y un dominio para caché de páginas web.[28] El servidor proxy Squid genera un log de navegación llamado Access.log. En este fichero se almacenan las peticiones hechas al servidor por un usuario, o sea toda la navegación que pueda tener un usuario en el servidor, de esta manera se puede obtener el tráfico de la navegación, así como las URL visitadas y el acceso de cada usuario a la navegación, toda esta información es analizada para posteriormente formar parte del proceso de toma de decisiones.

- **Análisis de los eventos generados por el servicio de navegación ISA Server**

(ISA) Server 2004 es la solución de caché web, servidor de seguridad avanzado en el nivel de aplicación y red privada virtual (VPN).[4] Al igual que el Squid esta aplicación genera un log del seguimiento de la navegación, este log es creado por peticiones de los usuarios que se conectan al servidor, obteniendo de igual manera el tráfico de la navegación, las URL[9] visitadas y el acceso a la navegación de cada usuario, toda esta información es analizada para posteriormente formar parte del proceso de toma de decisiones.

- **Análisis de los eventos generados por el servicio de mensajería ISA Server (Exchange)**

Es una potente solución que permite la gestión absoluta del servicio de mensajería, administrando buzones de manera inteligente y flexible. EXCHANGE al igual que Squid e ISA Server genera un log, el cual se actualiza cuando un usuario envía un mensaje y en él se guarda información referente al mensaje enviado, de este modo podemos obtener el tráfico de la mensajería de todos los usuarios de la empresa, toda esta información es analizada para posteriormente formar parte del proceso de toma de decisiones.

2.2 Propuesta del sistema

Con el propósito de diseñar un mercado de datos para mejorar el proceso de toma de decisiones para la seguridad de la información presente en los servicios de navegación web y mensajería electrónica ofrecidas por ETECSA se obtuvo la siguiente propuesta de sistema:

[9] URL: Uniform Resource Locator (Localizador Uniforme de Recursos)

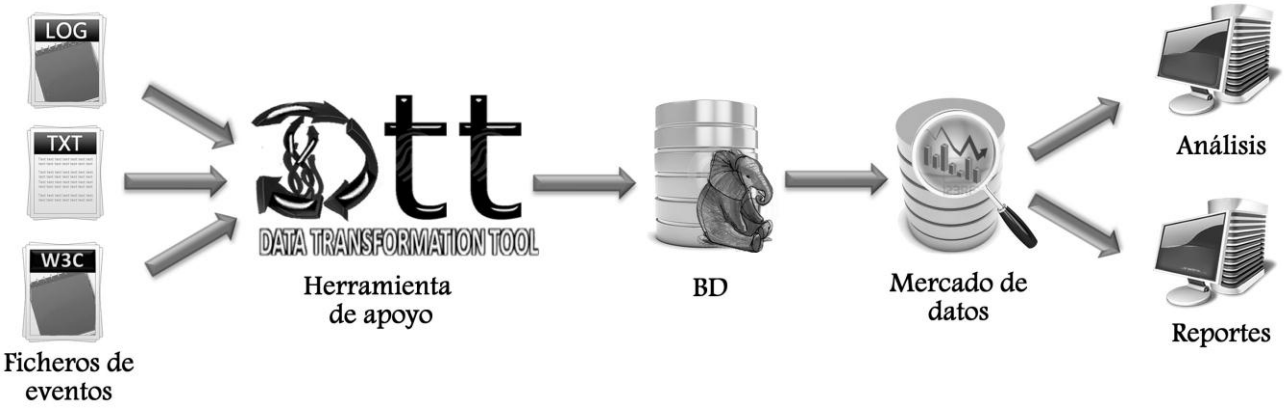


Figura 2 : Propuesta de sistema.

La solución propuesta está formada por las etapas que se muestran en la figura. Primeramente los ficheros de eventos (logs) son cargados para la herramienta de apoyo la cuál le realiza varias transformaciones a la información para estructurarla a un formato legible para la herramienta de creación del mercado. Posteriormente estos datos son llevados a una base de datos quién va a ser la fuente de alimentación directa del mercado, para poblarlo los datos vuelven a pasar por otra serie de transformaciones y por último con el mercado terminado el usuario puede realizar reportes o análisis de la información guardada en el mercado. A continuación se describe de forma más detallada este proceso por cada una de sus etapas.

2.2.1 Análisis y descripción de la herramienta de apoyo

Para el desarrollo de la solución se hizo necesario antes de la creación del mercado, la implementación de una aplicación que convierta los logs (fuente de información), en un formato legible para la herramienta de creación del mercado, en este caso la suite de Pentaho. Para su desarrollo se utilizará la metodología Programación Extrema (XP) pues es una aplicación sencilla que no posee muchas funcionalidades. Esta aplicación es la encargada de cargar los logs en una base de datos que se crea con la intención de hacer una copia de los datos pero ya en un formato capaz de ser leído por la herramienta Pentaho.También valida que el archivo a cargar no haya sido cargado anteriormente, a continuación se muestra una imagen para que se entienda con mayor facilidad el proceso que desarrolla la aplicación.

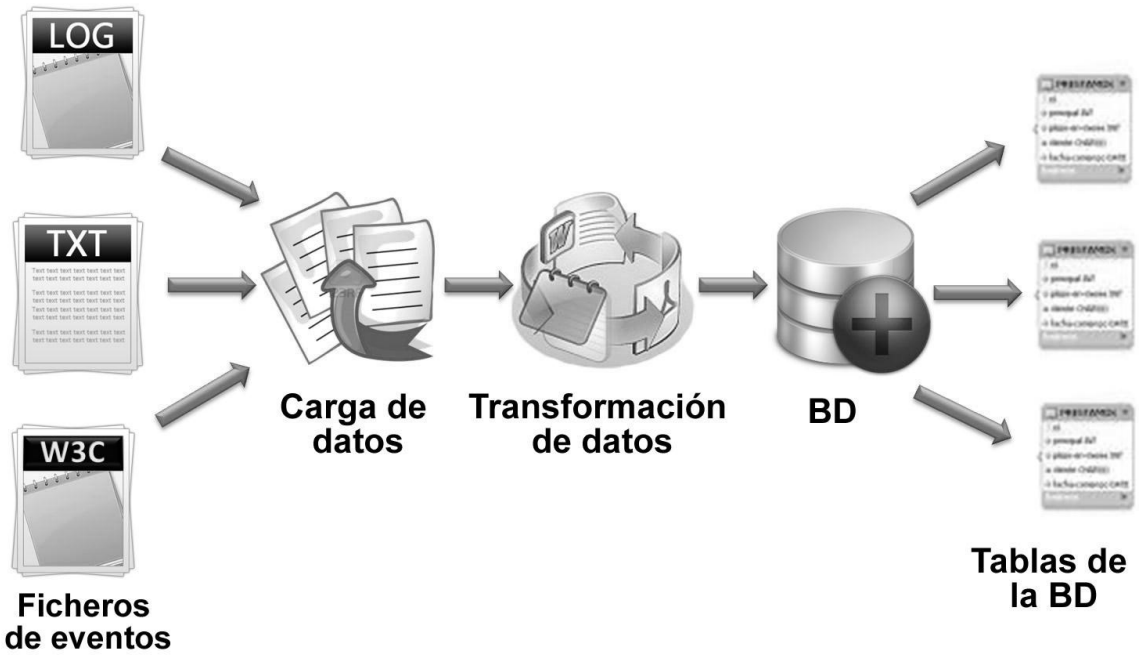


Figura 3: Proceso que desarrolla la herramienta de apoyo.

Ficheros de eventos: Fuente de datos originales, ficheros extraídos de los servidores utilizados por ETECSA**.**

Carga de datos: Se cargan los datos en la aplicación para proceder a realizar las transformaciones pertinentes.

Transformación de datos: Se le realizan las transformaciones a los ficheros hasta convertirlos en un formato legible para la herramienta de creación del mercado de datos.

BD (Base de datos): La información es almacenada en la base de datos de la cual se va a nutrir el mercado.

Tablas de BD: Se crean exactamente tres tablas en la base de datos, cada una corresponde a un fichero de eventos.

2.2.1.1 Requisitos de la aplicación

R1_ Autenticar usuario: El usuario se autentica en el sistema.

R2_Cargar los datos: El usuario selecciona el archivo a cargar y procede a cargarlo.

R3_Verificar última fecha de carga de los log: El sistema debe ser capaz de verificar que el archivo que se vaya a cargar no haya sido cargado anteriormente.

- ✓ Ultima fecha de carga de los log de Mensajería ISA Server.
- ✓ Ultima fecha de carga de los log de Navegación ISA Server.
- ✓ Ultima fecha de carga de los log de Navegación Squid.

2.2.1.2 Historias de usuarios

Las historias de usuarios (HU) son escritas por el equipo de desarrollo con apoyo del cliente, en su propio lenguaje, como descripciones cortas de lo que el sistema debe realizar. A continuación se presentan las HU determinadas por el cliente:

- ✓ Autenticar usuario.
- ✓ Verificar última fecha de carga de los log.
- ✓ Cargar los datos.

Historia de Usuario	
Número: 1	**Nombre historia de usuario:** Autenticar usuario
Modificación de Historia de Usuario:	
Usuario: Administrador	**Iteración asignada:** 1
Prioridad en negocio: Baja	
Riesgo en desarrollo: Baja	
Descripción: El usuario se autentifica en el sistema	
Historia de Usuario	
Número: 1	**Nombre historia de usuario:** Verificar última fecha de carga de

	los log
Modificación de historia de usuario:	
Usuario: Administrador	**Iteración asignada:** 1
Prioridad en negocio: Media	
Riesgo en desarrollo: Media	
Descripción: El sistema debe ser capaz de verificar que el archivo a cargar no exista en la base de datos.	
Historia de Usuario	
Número: 1	**Nombre historia de usuario:** Cargar los datos
Modificación de historia de usuario:	
Usuario: Administrador	**Iteración asignada:** 1
Prioridad en negocio: Media	
Riesgo en desarrollo: Media	
Descripción: El usuario carga el archivo que desee y lo introduce en la base de datos.	

- **Plan de Iteraciones**

A continuación se realiza el plan de iteraciones, que consiste en seleccionar las historias de usuario que se implementan en cada iteración, en este caso el sistema contará con una sola iteración por la sencillez del mismo.

Iteración 1 En esta iteración se implementan las 3 historias de usuarios propuestas. Al finalizar se contarán con todas las funcionalidades de las historias de usuarios (Ultima fecha de carga de los log, carga de los datos y autenticarse en la aplicación.).

2.2.2 Análisis y diseño del mercado de datos

2.2.2.1 Técnicas de identificación de requisitos de información

En la etapa del análisis de la construcción de los almacenes de datos, se realiza la tarea identificación de los requisitos, donde se aplican un conjunto de técnicas para la captura de los mismos. Siguiendo la metodología de Kimball, se realizan fundamentalmente las entrevistas cerradas con el cliente, aplicando una serie de preguntas enfocadas al rol que ejecuta el que se va a entrevistar. Para ver ejemplo ir a la ¡Error! No se encuentra el origen de la referencia. (Ver anexos)

Entrevistas: Las entrevistas se llevan a cabo individualmente o en grupos muy pequeños. Las mismas necesitan ser programadas en un tiempo limitado para no perjudicar el calendario de los entrevistados. Además promueven un alto grado de participación del cliente, lo cual posibilita generar una mayor cantidad de información, así como un mayor intercambio entre las partes involucradas.

Este tipo de técnica se utiliza con el propósito de conocer cuáles son las metas y objetivos del negocio, sus prioridades y los procesos que abarca. Otro motivo de la entrevista es saber cuál es el personal que trabaja en dicha área y las responsabilidades que le competen a cada uno, así como identificar sus necesidades de información.

2.2.2.2 Análisis de Requerimientos

El análisis de los requerimientos recoge todas las necesidades del usuario final, la misma está compuesta por los requisitos de información y la especificación de requisitos, los cuales deben ser descriptivos y orientados al usuario final. Además se describen los usuarios que van a interactuar con el sistema.

Requisitos de Información

1. Describen qué información debe almacenar el sistema para satisfacer las necesidades de clientes y usuarios. Identifican los conceptos relevantes sobre los que se debe almacenar información y los datos específicos que son de interés.[6]

Los requisitos de información representan toda la información necesaria para los especialistas del DSI de ETECSA, con el propósito de realizar los análisis pertinentes que serán tomados en cuenta en la proyección estratégica del centro. Para darle respuesta a las necesidades del cliente, se identificaron

un conjunto de tablas dimensiones y medidas, por los diferentes procesos que abarca el negocio. Los requisitos de información identificados se muestran a continuación desglosados por cada proceso, para un mejor entendimiento y organización, su enumeración es continua:

RI1_Exportar los reportes.

- **Procesos: Análisis de los eventos generados por el servicio de Navegación Squid**

RI2_Mostrar tráfico de navegación de direcciones IP (Internet Protocol): Muestra el IP y la cantidad de veces que accedieron desde ese IP en una fecha dada.

- ✓ Tráfico de navegación por direcciones IP diario.
- ✓ Tráfico de navegación por direcciones IP mensual.
- ✓ Tráfico de navegación por direcciones IP en un rango.

RI3_Mostrar tráfico de navegación de usuarios: Muestra el usuario, el IP al que accedió y la cantidad de veces que lo hizo en una fecha dada.

- ✓ Tráfico de navegación por usuario diario.
- ✓ Tráfico de navegación por usuario mensual.
- ✓ Tráfico de navegación por usuario en un rango.

RI4_Mostrar tráfico de navegación: Muestra el usuario, las direcciones a las que ha accedido y la cantidad de veces que lo ha hecho en una fecha dada.

- ✓ Tráfico de navegación diario.
- ✓ Tráfico de navegación mensual.
- ✓ Tráfico de navegación en un rango.

RI5_ Reporte de sitios de navegación denegados mensual: Muestra la cantidad de direcciones web denegadas en el mes, el usuario, la dirección web denegada, cantidad de veces que intentó acceder.

RI6_ Reporte de registro de direcciones web mensual. Muestra los usuarios y la cantidad de veces que se ha registrado en un mes.

- **Procesos: Análisis de los eventos generados por el servicio de navegación ISA Server**

RI7_Mostrar tráfico de navegación por direcciones ip: Muestra el IP y la cantidad de veces que accedieron desde ese IP en una fecha dada.

- ✓ Tráfico de navegación por direcciones IP diario.
- ✓ Tráfico de navegación por direcciones IP mensual.
- ✓ Tráfico de navegación por direcciones IP en un rango.

RI8_Mostrar tráfico de navegación por usuario: Muestra el usuario, el IP al que accedió y la cantidad de veces que lo hizo en una fecha dada.

- ✓ Tráfico de navegación por usuario diario.
- ✓ Tráfico de navegación por usuario mensual.
- ✓ Tráfico de navegación por usuario en un rango.

RI9_Mostrar tráfico de navegación: Muestra el usuario, las direcciones a las que ha accedido y la cantidad de veces que lo ha hecho en una fecha dada.

- ✓ Tráfico de navegación diario.
- ✓ Tráfico de navegación mensual.
- ✓ Tráfico de navegación en un rango.

- **Procesos: Análisis de los eventos generados por el servicio de mensajería ISA Server**

RI10_Mostrar tráfico de mensajería por usuario: Muestra el usuario y la cantidad de mensajes que envió en una fecha dada.

- ✓ Tráfico de mensajería por usuario diario.
- ✓ Tráfico de mensajería por usuario mensual.
- ✓ Tráfico de mensajería por usuario en un rango.

RI11_Mostrar tráfico de mensajería por direcciones ip: Muestra la dirección IP de donde fueron enviados los mensajes y la cantidad de mensajes enviados.

- ✓ Tráfico de mensajería por direcciones IP diario.

- ✓ Tráfico de mensajería por direcciones IP mensual.
- ✓ Tráfico de mensajería por direcciones IP en un rango.

RI12_Mostrar tráfico de mensajería por mensajes enviados: Muestra el usuario, el asunto del mensaje que envió, el destinatario y la cantidad de veces que envió el mensaje.

- ✓ Tráfico de mensajería por mensajes enviados diario.
- ✓ Tráfico de mensajería por mensajes enviados mensual.
- ✓ Tráfico de mensajería por mensajes enviados en un rango.

Técnicas de validación de requisitos

Posteriormente al proceso de captura de requisitos, se aplican técnicas para la validación de los mismos. Es necesario asegurar que el análisis realizado y los resultados obtenidos en la etapa de identificación de requisitos sean correctos. Para validar los requisitos se va a utilizar la técnica prototipo funcional, la cual se describe a continuación:

Técnica Prototipo: Los prototipos son simulaciones del posible producto, con el objetivo de mostrarle al cliente una versión reducida del producto final, permitiendo conseguir una importante retroalimentación en cuanto a saber si el producto diseñado con base a los requerimientos recolectados, le permite al usuario realizar su trabajo de manera eficiente. Para realizar el prototipo se utilizó como herramienta de apoyo los ejemplos que trae por defecto la Suite de Inteligencia de Negocios de Pentaho.

2.2.2.3 Usuarios del sistema

A partir de las entrevistas realizadas, dentro del proceso de captura de requisitos, se identificaron 3 usuarios principales que van a interactuar con el sistema, los cuales se describen a continuación:

1. Administrador del Sistema: Realiza el proceso de extracción, transformación y carga de los datos además de gestionar los usuarios y los roles.
2. Especialista del Departamento de Seguridad Informática: Encargado de gestionar los reportes.
3. Jefe del Departamento de Seguridad Informática: Consulta y analiza los reportes.

2.2.2.4 Arquitectura del mercado de datos

La arquitectura de los almacenes de datos está compuesta por una serie de procesos o subsistemas que definen, en su conjunto, el ambiente que estos poseen. La arquitectura del mercado de datos para el análisis de eventos de navegación web y mensajería electrónica en el DSI de ETECSA queda definida de la siguiente manera:

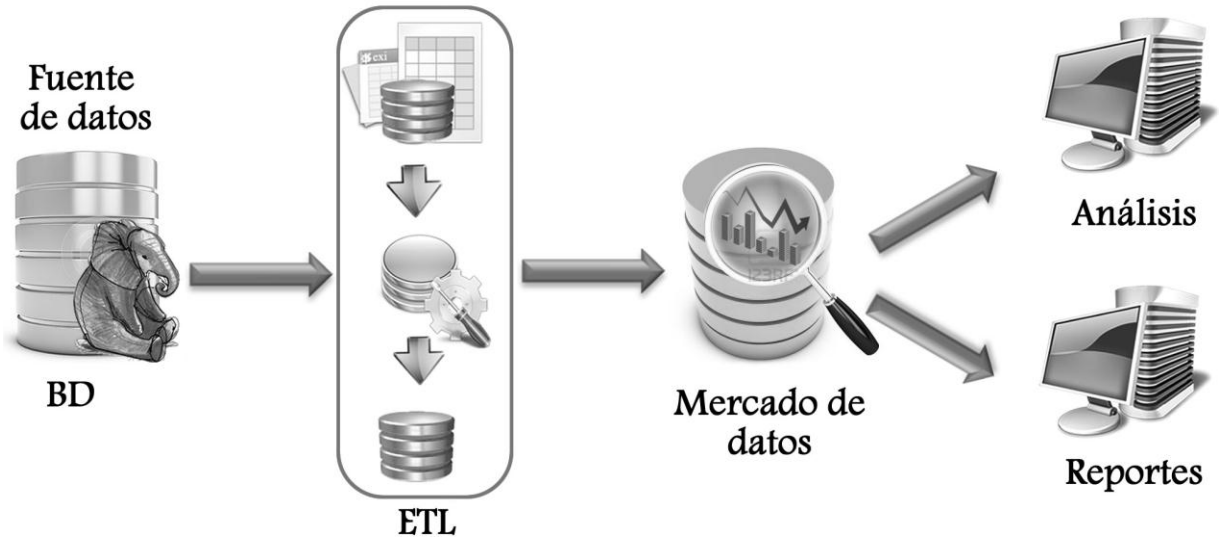


Figura 4: Arquitectura del mercado de datos.

Fuente de datos: La primera capa representa las fuentes de origen que van a alimentar al mercado. Los datos fuentes pueden estar en cualquier formato, desde archivos de texto, hojas excel, base de datos relacionales así como otros tipos de bases de datos, en este caso la fuente de origen no es más que la base de datos que crea la herramienta de apoyo con los datos estructurados. Estos datos son consultados en la heramienta de creación donde se le hacen las transformaciones necesarias para almacenarlos posteriormente en el mercado, con el objetivo de eliminar las incosistencias que puedan traer del sistema origen, además de integrarlos al formato establecido.

Mercado de datos: La última capa no es más que el mercado de datos obtenido. En esta capa residen los datos almacenados en una estructura dimensional, a partir de los cuales se realiza todo el análisis para la creación de los reportes con las herramientas de consultas.

2.2.2.5 Modelado Dimensional

Dentro del modelo de datos se encuentra el "*modelo dimensional*", que tiene como elementos principales las dimensiones y los hechos, los cuales recogen de los procesos que abarca el subsistema, los atributos necesarios para dar respuesta a los requisitos de información. Las dimensiones describen los diferentes puntos de vista del hecho y el hecho es el punto de intersección entre las diferentes dimensiones. Es importante conocer la composición de ambos y la relación que existe entre ellos para lograr un buen funcionamiento del mercado de datos.

- **Granularidad**

Se refiere al nivel de detalle de las unidades de datos en el mercado de datos. Mientras más detalle exista, más bajo será el nivel de granularidad. Esto se ve reflejado en las tablas de dimensiones por ejemplo en el caso de la dimensión tiempo se llega a un grano diario (año, mes, semana, día).[29]

- **Tablas dimensiones**

Las tablas de dimensiones definen como están los datos organizados lógicamente y proveen el medio para analizar el contexto del negocio. Contienen datos cualitativos. Representan los aspectos de interés, mediante los cuales los usuarios podrán filtrar y manipular la información almacenada en la tabla de hechos.[30]

Para el diseño del mercado de datos se identificaron 11 tablas dimensiones y 3 tablas de hechos de los 3 procesos que componen el monitoreo de los eventos de seguridad. A continuación se describen brevemente las dimensiones de los procesos del caso de estudio:

dim_ip_navegacion_isaserver_web: Contiene la información referente al IP del cliente de la navegación isaserver web.

dim_url_navegacion_isaserver_web: Contiene la información referente a la URL visitada por el cliente.

dim_servidor_navegacion_isaserver_web: Contiene la información referente al IP del servidor web.

dim_usuario_navegacion_isaserver_web: Contiene la información referente al usuario.

dim_ip_navegacion_squid: Contiene la información referente al IP del cliente de la navegación Squid.

dim_usuario_navegacion_squid: Contiene la información referente al usuario.

dim_url_navegacion_squid: Contiene la información referente a la URL visitada por el cliente.

dim_usuario_mensajeria_isaserver: Contiene la información referente al usuario.

dim_ip_mensajeria_isaserver: Contiene la información referente al IP cliente de la mensajería isaserver.

dim_mensaje_mensajeria_isaserver: Contiene la información referente a los mensajes del cliente. (asunto y destinatario).

dim_tiempo: Contiene todos los datos relacionados con el tiempo, hasta el nivel más bajo de la granularidad, que en este caso es hasta el día del mes.

- **Tablas de hechos**

Las tablas de hechos contienen, precisamente, los hechos que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones. Contienen datos cuantitativos y son datos instantáneos en el tiempo, que son filtrados, agrupados y explorados a través de condiciones definidas en las tablas de dimensiones.[30]

Para el modelo de los procesos se definieron 3 tablas de hechos, las cuales se describen a continuación:

hecho_navegacion_isaserver_web: Guarda el hecho correspondiente a la cantidad de navegación isa_server web realizada por un usuario en una fecha dada.

hecho_navegacion_squid: Guarda el hecho correspondiente a la cantidad de navegación Squid realizada por un usuario en una fecha dada.

hecho_mensajeria_isaserver: Guarda el hecho correspondiente a la cantidad de mensajes de mensajería isa_server realizada por un usuario en una fecha dada.

- **Matriz dimensional**

En la matriz dimensional o Bus Matrix por sus siglas en inglés, se recoge la relación existente entre las dimensiones y los procesos que componen el negocio. En los modelos propuestos existen dimensiones comunes y específicas de cada proceso. En la Tabla 2.1 que se muestra a continuación se observa esta matriz:

	Procesos		
Dimensiones	Navegación Squid	Navegación ISA Server .	Mensajería ISA Server
dim_ip_navegacion_isaserver_web		x	
dim_tiempo	x	x	x
dim_url_navegacion_isaserver_web		x	
dim_servidor_navegacion_isaserver_web		x	
dim_usuario_navegacion_isaserver_web		x	
dim_ip_navegacion_squid	x		
dim_usuario_navegacion_squid	x		
dim_url_navegacion_squid	x		
dim_usuario_mensajeria_isaserver			x
dim_ip_mensajeria_isaserver			x
dim_mensaje_mensajeria_isaserver			x

Tabla 2.1: Matriz Dimensional

- **Patrones de diseño**

El Modelo Dimensional de los procesos que se analizan, cuenta con un total de 11 tablas dimensiones y 3 tablas de hechos. Para el diseño del mismo se utilizó el siguiente patrón de almacenes de datos:

1. Claves subrogadas: Con la utilización de este patrón se posibilita a la hora de realizar el proceso de ETL, garantizar que no exista redundancia de información ni sobre escritura de ninguno de los datos almacenados. Al asignar llaves primarias independientes a las utilizadas en la base de datos no existe la posibilidad de generar errores de conflicto ni obtener

información que no tiene relación con las consultas realizadas, independizándose de las llaves primarias de las base de datos que almacenan la información. Es una manera adecuada de separar la lógica de los Sistemas OLTP del modelado dimensional.

En cada una de las transformaciones realizadas para crear las dimensiones fue necesario asignarles una llave subrogada para garantizar las prestaciones de este patrón. Con este propósito existe un paso dentro del componente transformación que permite realizar este procedimiento de asociación. Este paso es Búsqueda/Actualización en combinación, ideal para la creación de dimensiones llamadas basuras debido a su carácter no cambiante. De esta manera es posible escoger que campo resulta el indicado para establecer esta relación llave primaria del sistema OLTP y la Dimensión creada.

- **Modelo de datos**

En consecuencia de los requisitos de información definidos para los procesos, se elaboró su modelo de datos. Cada modelo está compuesto por la dimensión común, que es la dim_tiempo y las dimensiones propias de su proceso, en el caso de la Navegación Squid las dim_ip_navegacion_squid, dim_usuario_navegacion_squid, dim_url_navegacion_squid, para Navegación ISA Server las dim_ip_navegacion_isaserver_web,dim_url_navegacion_isaserver_web,dim_servidor_navegacion_isaserver_web, dim_usuario_navegacion_isaserver_web, para la Mensajería ISA Server las dim_usuario_mensajeria_isaserver, dim_ip_mensajeria_isaserver, dim_mensaje_mensajeria_isaserver. Además de las tablas de hechos que contienen las medidas, formando cada esquema una estrella.

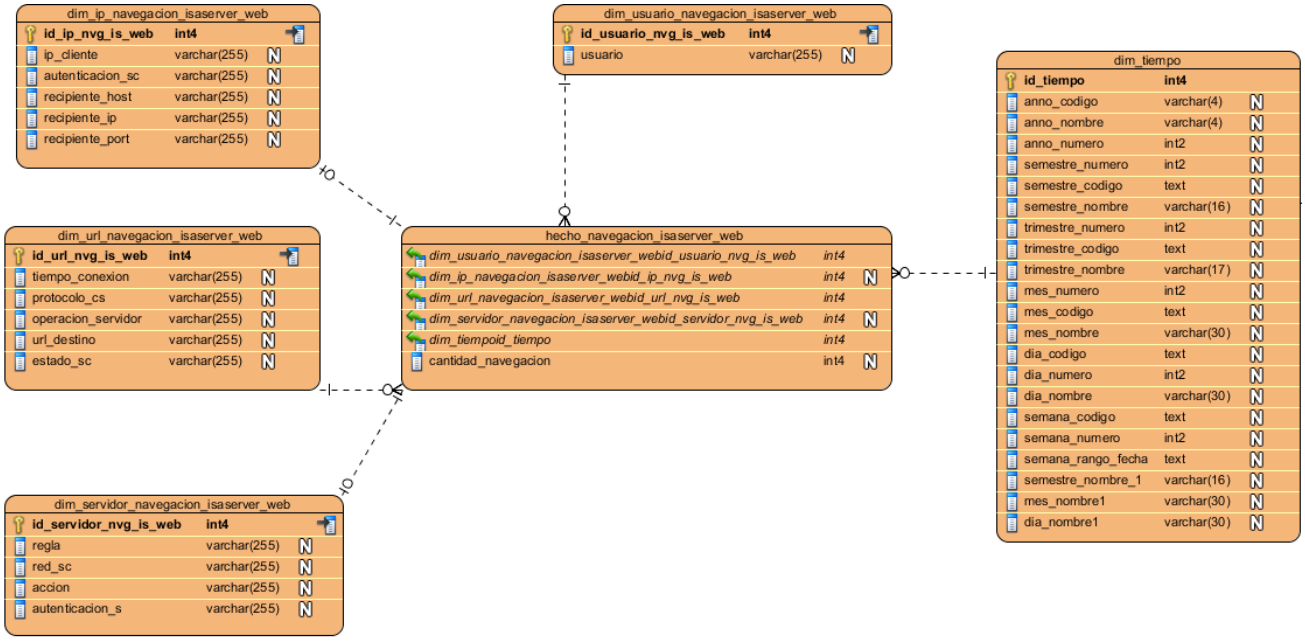


Figura 5: Modelo dimensional del proceso Navegación Squid.

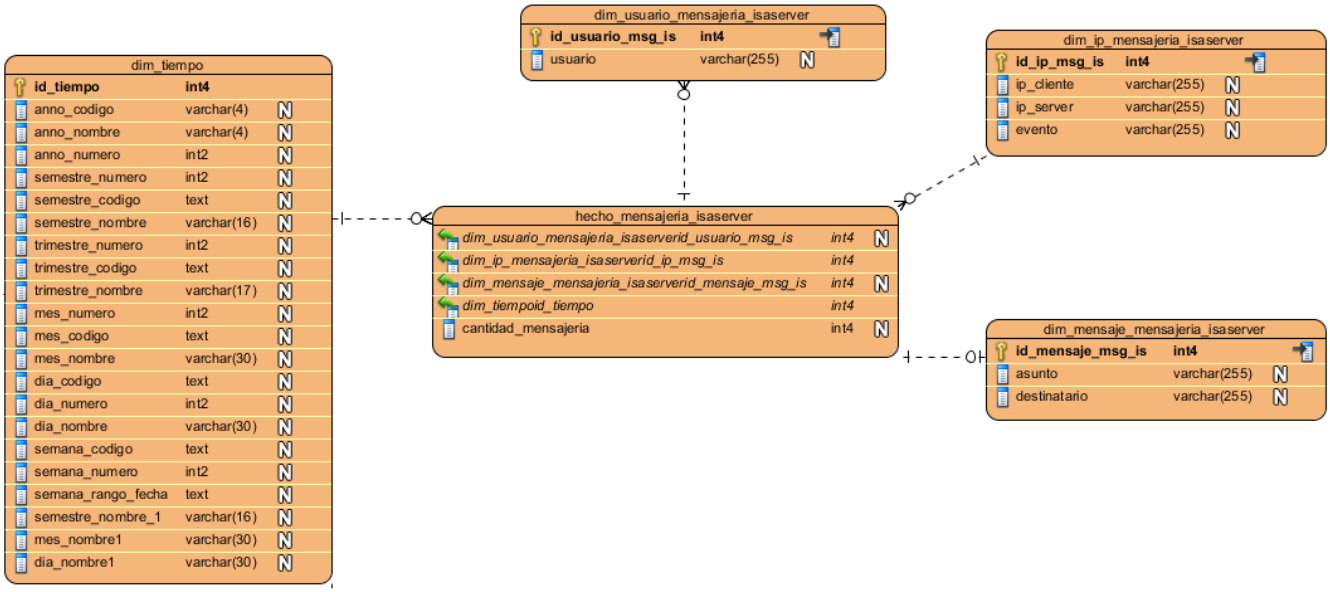


Figura 6: Modelo dimensional del proceso Mensajería ISA Server

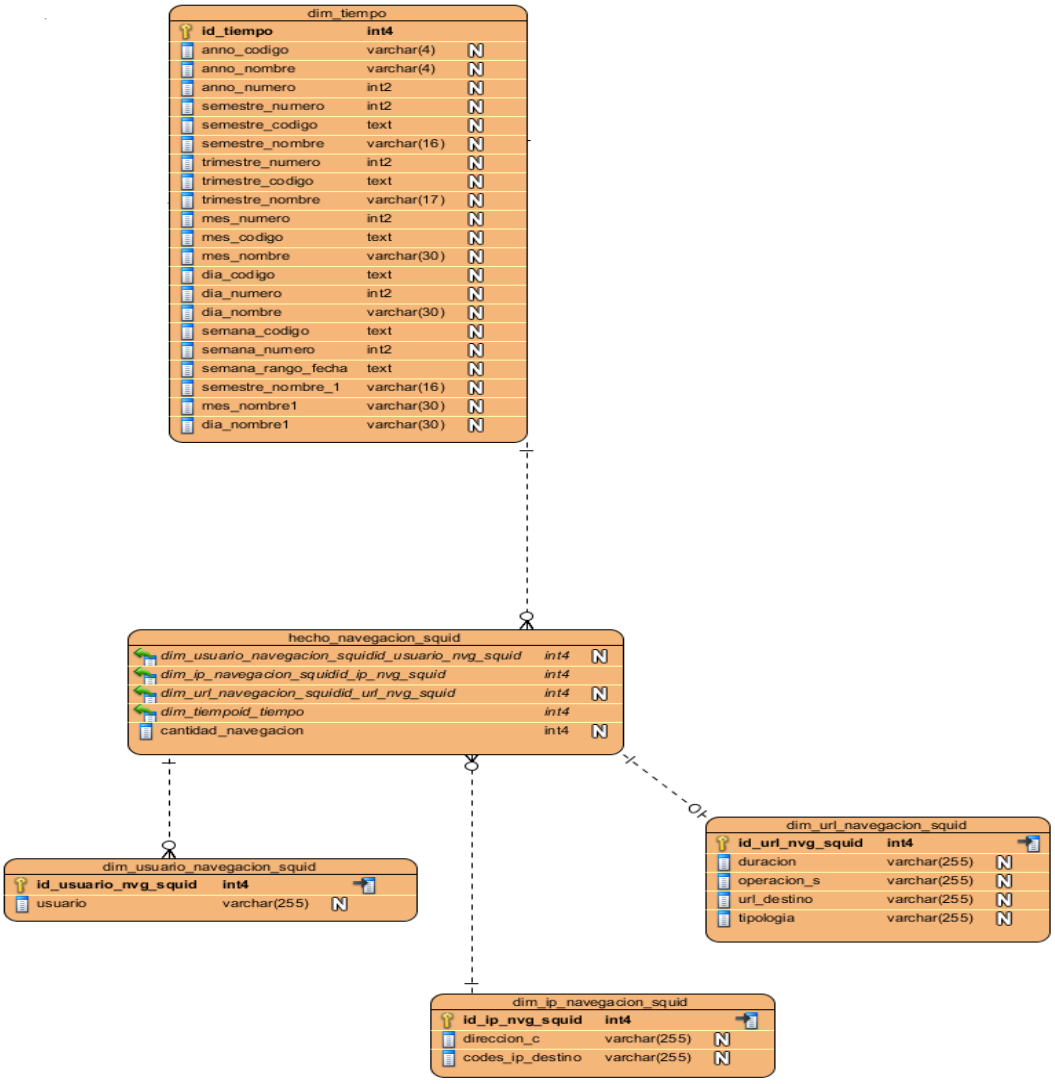


Figura 7: Modelo dimensionales del proceso Navegación ISA Server.

2.5 Conclusiones Parciales

En el presente capitulo se describieron los principales procesos que se llevan a cabo en el desarrollo del análisis y diseño del mercado de dato. Se describió la propuesta de solución brindada conformándose por una herramienta de apoyo y el mercado de datos. Se determinaron 12 requisitos de información y 3 requisitos con los que debe cumplir el sistema. Se especificaron 11 tablas de

dimensiones y 3 tablas hechos por los que está compuesto el modelo dimensional, así como su modelo de datos correspondiente.

CAPÍTULO 3: IMPLEMENTACIÓN Y PRUEBA

Introducción

En este capítulo se presenta la implementación del mercado de datos y las pruebas realizadas. Como parte de la implementación se muestra: la implementación del proceso ETL de los datos pertenecientes a los procesos analizados en el capítulo anterior, otro elemento es la configuración del repositorio de datos que será necesario para la construcción de los reportes. En consecuencia de lo que describe la metodología Kimball para probar el mercado de datos, se realizan las pruebas y se analiza el resultado de las mismas.

3.1 Implementación del subsistema de integración

El proceso de implementación engloba la realización de las ETL de los datos, la configuración del repositorio a través de la herramienta de administración y la creación de los reportes en la herramienta de consulta.

3.1.1 Perfilado de los datos

El perfilado de datos es el proceso de examinar los datos que existen en las fuentes de origen de una organización y recopilar estadísticas e información sobre los mismos. El propósito de dichas estadísticas es: [31]

- Determinar qué datos pueden ser usados para otros propósitos.
- Conseguir métricas de calidad de datos que incluyen si los datos cumplen los estándares de la organización.
- Reducir el riesgo de integrar información a nuevas aplicaciones dado que se conoce su estado.
- Permitir hacer un seguimiento de la calidad de datos.
- Entender problemas derivados de los datos en proyectos que hagan uso intensivo de los mismos.

La herramienta utilizada en esta investigación para el perfilado de los datos fue el Data Cleaner 1.5.3. DataCleaner es una aplicación de código abierto para el perfilado, validación y comparación de datos. Estas actividades ayudan a administrar y monitorear la calidad de la información, en el orden de

garantizar la utilidad de la información y de la aplicación para el negocio. Es una alternativa libre para la metodología de administración de datos, para proyectos de almacenes de datos, búsquedas estadísticas y para actividades de preparación de ETL.

A continuación se muestran las imágenes obtenidas del perfilado de datos en la tabla de mensajería.

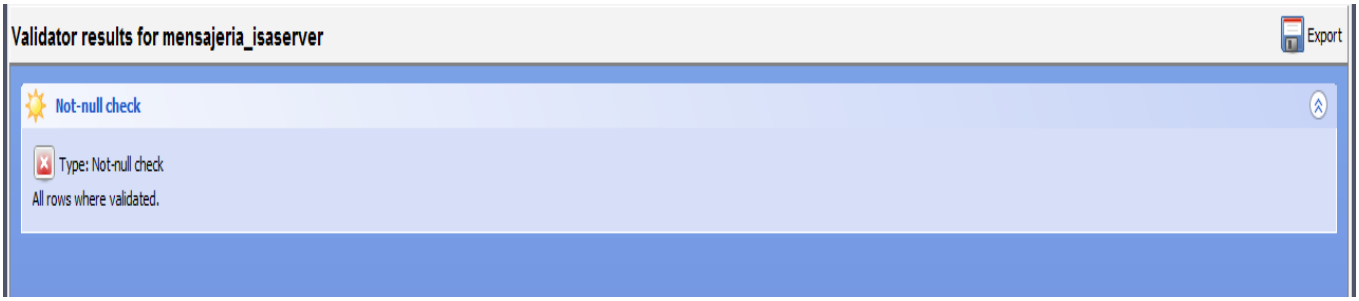


Figura 8: Campos nulos.

Campos nulos: Se realiza con el objetivo de garantizar que no existan campos nulos en la base de datos.

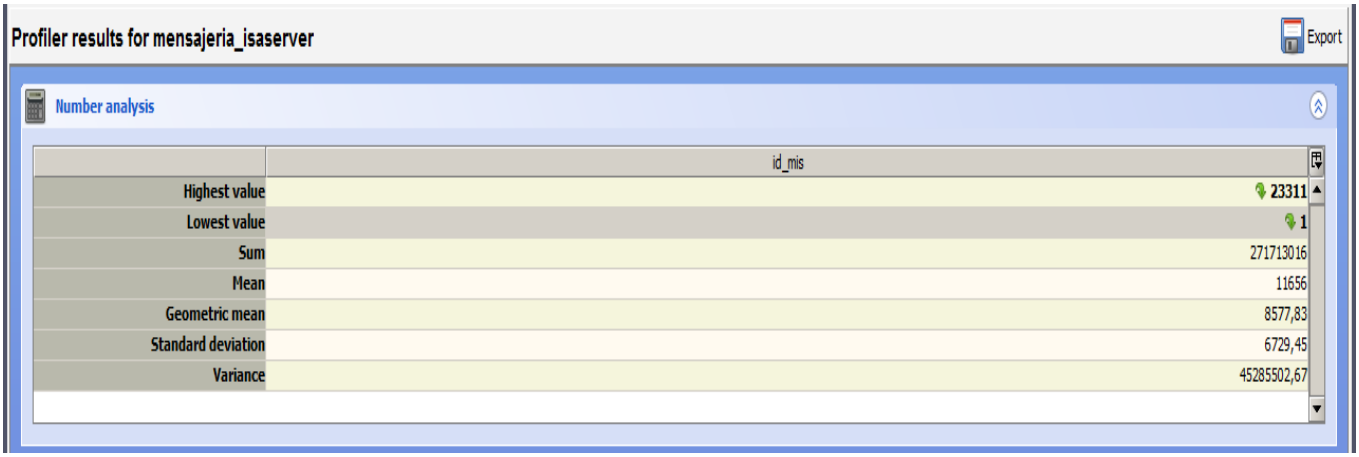
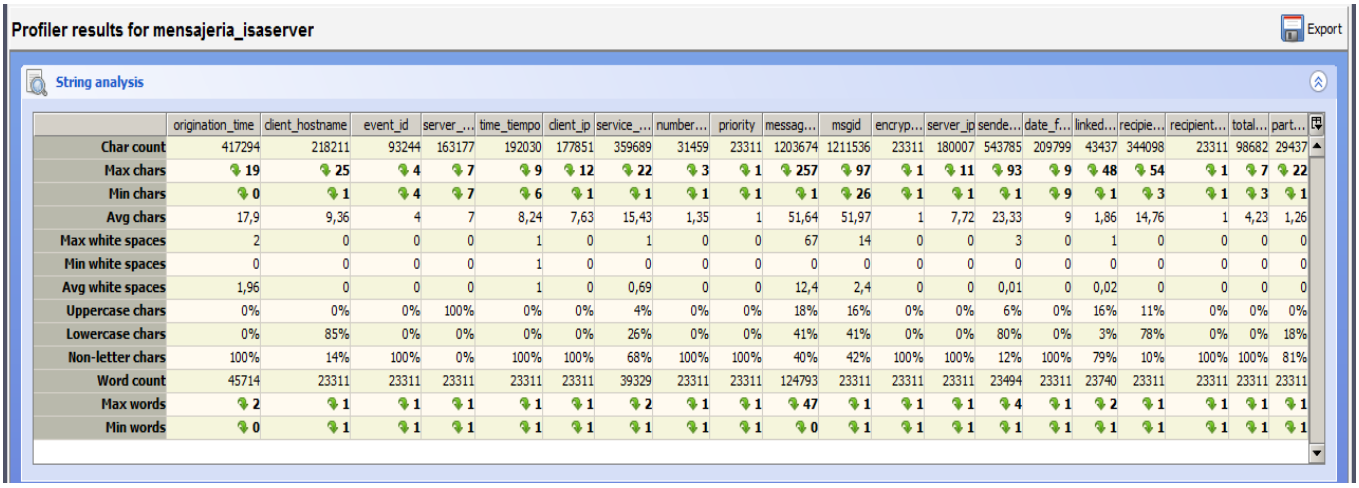


Figura 9: Análisis de números.

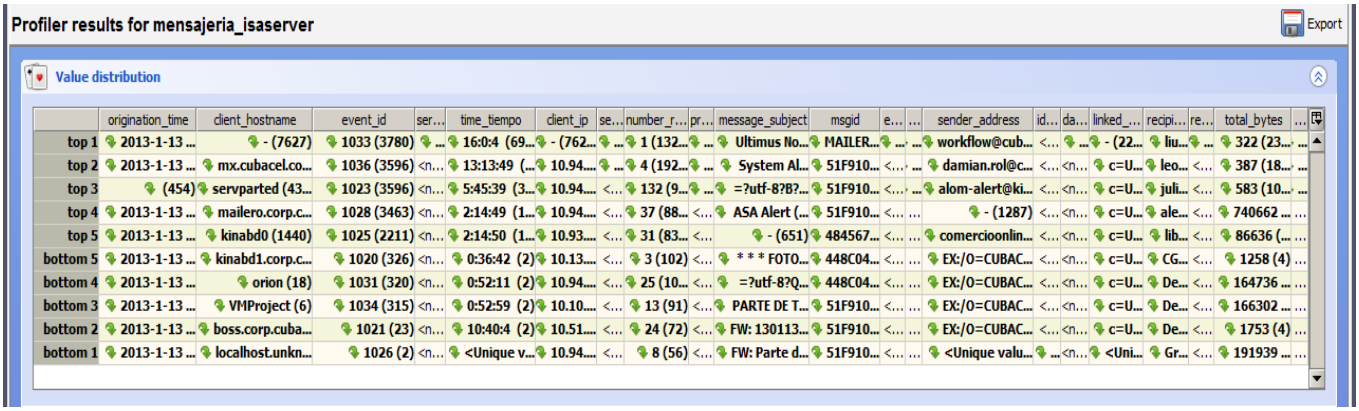
Análisis de números: Se realiza con el objetivo de garantizar que, en las celdas destinadas a contener números no exista otro tipo de datos.



	origination_time	client_hostname	event_id	server_...	time_tiempo	client_ip	service_...	number...	priority	messag...	msgid	encryp...	server_ip	sende...	date_f...	linked...	recipie...	recipient...	total...	part...
Char count	417294	218211	93244	163177	192030	177851	359689	31459	23311	1203674	1211536	23311	180007	543785	209799	43437	344098	23311	98682	29437
Max chars	19	25	4	7	9	12	22	3	1	257	97	1	11	93	9	48	54	1	7	22
Min chars	0	1	4	7	6	1	1	1	1	1	26	1	1	1	9	1	3	1	3	1
Avg chars	17,9	9,36	4	7	8,24	7,63	15,43	1,35	1	51,64	51,97	1	7,72	23,33	9	1,86	14,76	1	4,23	1,26
Max white spaces	2	0	0	0	1	0	1	0	0	67	14	0	0	3	0	1	0	0	0	0
Min white spaces	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Avg white spaces	1,96	0	0	0	1	0	0,69	0	0	12,4	2,4	0	0	0,01	0	0,02	0	0	0	0
Uppercase chars	0%	0%	0%	100%	0%	0%	4%	0%	0%	18%	16%	0%	0%	6%	0%	16%	11%	0%	0%	0%
Lowercase chars	0%	85%	0%	0%	0%	0%	26%	0%	0%	41%	41%	0%	0%	80%	0%	3%	78%	0%	0%	18%
Non-letter chars	100%	14%	100%	0%	100%	100%	68%	100%	100%	40%	42%	100%	100%	12%	100%	79%	10%	100%	100%	81%
Word count	45714	23311	23311	23311	23311	23311	39329	23311	23311	124793	23311	23311	23311	23494	23311	23740	23311	23311	23311	23311
Max words	2	1	1	1	1	1	2	1	1	47	1	1	1	4	1	2	1	1	1	1
Min words	0	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1

Figura 10: Análisis de cadenas.

Análisis de cadenas: Se realiza con el objetivo de garantizar que, en las celdas destinadas a contener cadenas no exista otro tipo de datos.



	origination_time	client_hostname	event_id	ser...	time_tiempo	client_ip	se...	number_r...	pr...	message_subject	msgid	e...	...	sender_address	id...	da...	linked_...	recipi...	re...	total_bytes	...
top 1	2013-1-13 ...	- (7627)	1033 (3780)	...	16:0:4 (69...	- (762...	...	1 (132...	...	Ultimus No...	MAILER...	...	...	workflow@cub...	<...	...	- (22...	liu...	...	322 (23...	...
top 2	2013-1-13 ...	mx.cubacel.co...	1036 (3596)	<n...	13:13:49 (...	10.94....	...	4 (192...	...	System Al...	51F910...	<...	...	damian.rol@c...	<...	<n...	c=U...	leo...	<...	387 (18...	...
top 3	(454)	servparted (43...	1023 (3596)	<n...	5:45:39 (3...	10.94....	<...	132 (9...	...	=?utf-8?B?...	51F910...	<...	...	alom-alert@ki...	<...	<n...	c=U...	juli...	<...	583 (10...	...
top 4	2013-1-13 ...	mailero.corp.c...	1028 (3463)	<n...	2:14:49 (1...	10.94....	<...	37 (88...	<...	ASA Alert (...	51F910...	<...	...	- (1287)	<...	<n...	c=U...	ale...	<...	740662 ...	...
top 5	2013-1-13 ...	kinabd0 (1440)	1025 (2211)	<n...	2:14:50 (1...	10.93....	<...	31 (83...	<...	- (651)	484567...	<...	...	comercioonlin...	<...	<n...	c=U...	lib...	<...	86636 (...	...
bottom 5	2013-1-13 ...	kinabd1.corp.c...	1020 (326)	<n...	0:36:42 (2)	10.13....	<...	3 (102)	<...	* * * FOTO...	448C04...	<...	...	EX:/O=CUBAC...	<...	<n...	c=U...	CG...	<...	1258 (4)	...
bottom 4	2013-1-13 ...	orion (18)	1031 (320)	<n...	0:52:11 (2)	10.94....	<...	25 (10...	<...	=?utf-8?Q...	448C04...	<...	...	EX:/O=CUBAC...	<...	<n...	c=U...	De...	<...	164736 ...	...
bottom 3	2013-1-13 ...	VMProject (6)	1034 (315)	<n...	0:52:59 (2)	10.10....	<...	13 (91)	<...	PARTE DE T...	51F910...	<...	...	EX:/O=CUBAC...	<...	<n...	c=U...	De...	<...	166302 ...	...
bottom 2	2013-1-13 ...	boss.corp.cuba...	1021 (23)	<n...	10:40:4 (2)	10.51....	<...	24 (72)	<...	FW: 130113...	51F910...	<...	...	EX:/O=CUBAC...	<...	<n...	c=U...	De...	<...	1753 (4)	...
bottom 1	2013-1-13 ...	localhost.unkn...	1026 (2)	<n...	<Unique v...	10.94....	<...	8 (56)	<...	FW: Parte d...	51F910...	<...	...	<Unique valu...	...	<n...	<Uni...	Gr...	<...	191939 ...	...

Figura 11: Valor de la distribución.

Valor de la distribución: Se realiza con el objetivo de garantizar que, los valores que sean distribuidos en las tablas lo hagan de la manera correcta

Posteriormente a cada una de estas pruebas se muestra una interfaz con su resultado, todas salieron de manera satisfactoria, a continuación se muestra una imagen de ejemplo.

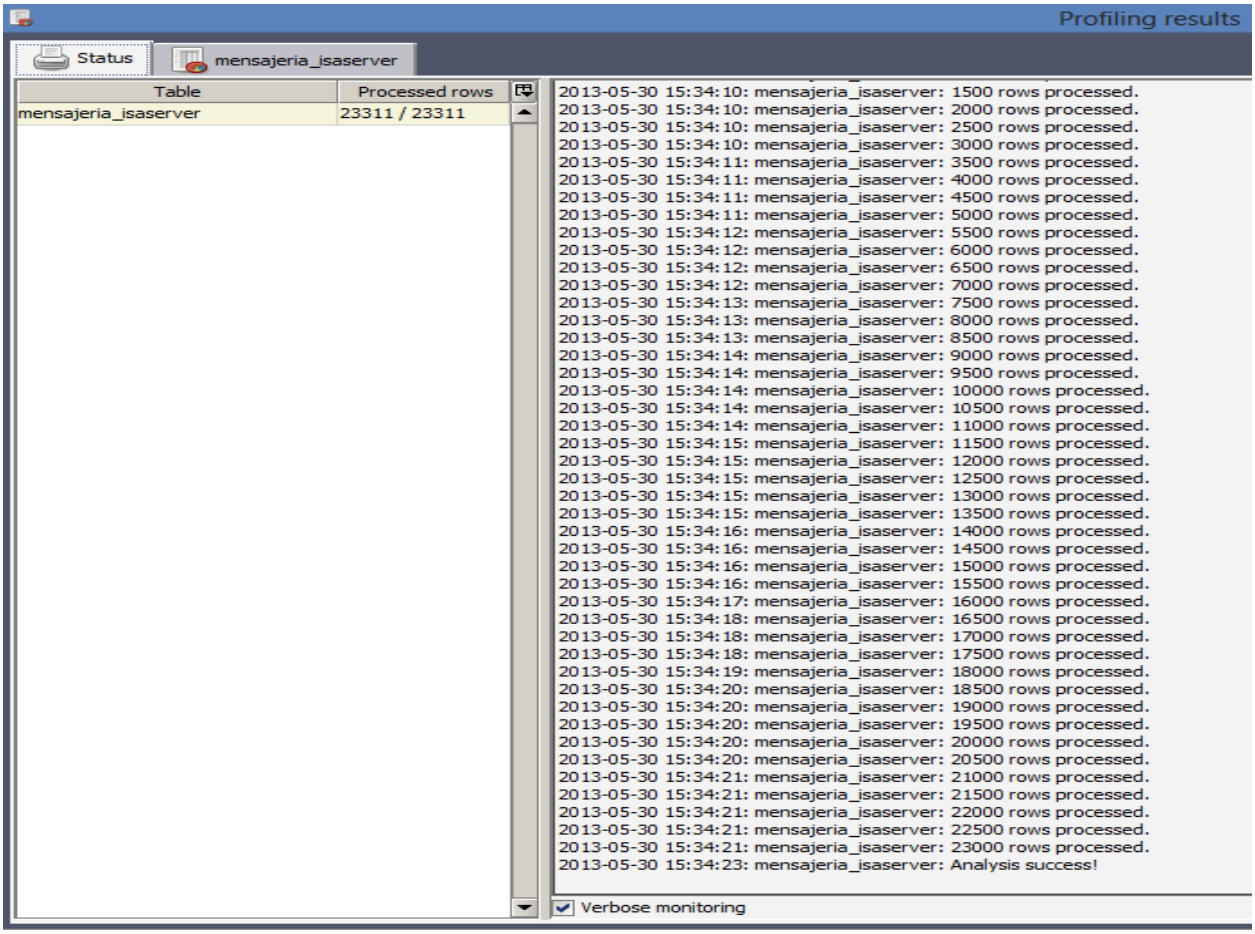


Figura 12: Éxito del análisis.

3.1.2 Implementación de la Extracción Transformación y Carga (ETL)

La implementación de las ETL fue realizado con la herramienta Spoon el diseñador gráfico de transformaciones y trabajos del sistema de ETTLs de Pentaho Data Integration (PDI), también conocido como Kettle (acrónimo recursivo: "Kettle Extraction, Transformation, Transportation, and Load Environment"). Está diseñado para ayudar en los procesos ETTLs, que incluyen la Extracción, Transformación, Transporte y Carga de datos. Spoon es una Interfaz Gráfica de Usuario (GUI), que permite diseñar transformaciones y trabajos que se pueden ejecutar con las herramientas de Kettle. Un trabajo es un conjunto sencillo o complejo de tareas con el objetivo de realizar una acción determinada, normalmente se planifican en modo batch (por lotes) para ejecutarlos automáticamente en intervalos

regulares. Las Transformaciones y Trabajos se pueden describir usando un archivo XML o se pueden colocar en un catálogo de base de datos de Kettle.[32]

La extracción es el proceso donde se extraen los datos de la base de datos para adaptarlos al modelo establecido. La transformación es de gran importancia, porque es la etapa donde se garantiza el resultado final de cómo se van a mostrar los datos. Se aplican las reglas de transformación definidas a partir de las reglas de transformación identificadas; se detectan otras posibles deficiencias de la fuente y se corrigen. El resultado es la carga al mercado de datos. La carga es el paso final del proceso, los datos son cargados al mercado para ser utilizados de forma satisfactoria.

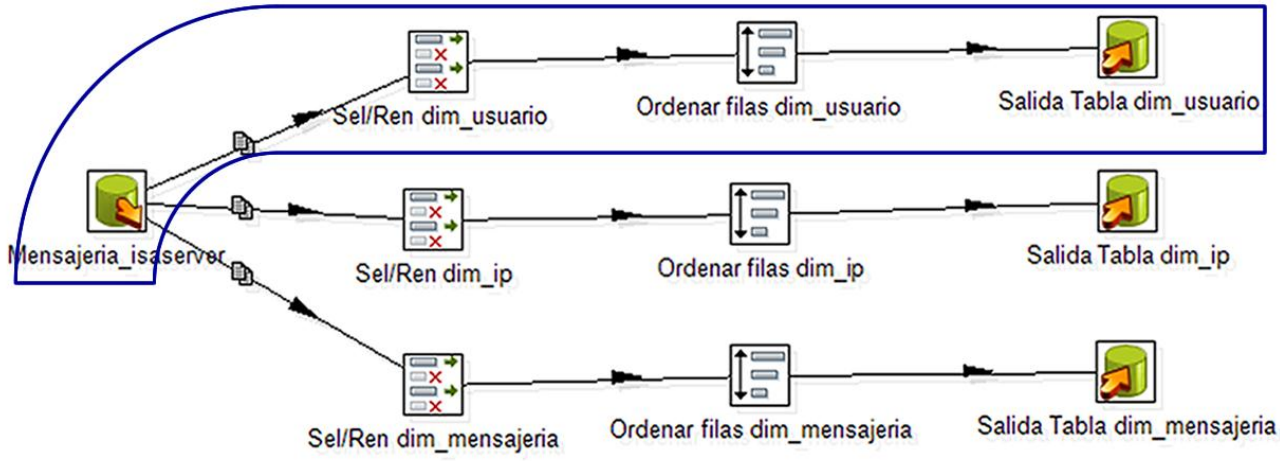


Figura 13: ETL del proceso análisis de los eventos generados por el servicio de Mensajería ISA Server.

La figura 14 muestra un ejemplo de una de las transformaciones implementadas correspondientes al proceso de Mensajería ISA Server. Primeramente se le realiza una consulta a la base de datos de la cual se está alimentando el almacén para extraer los datos necesarios y así realizar el proceso ETL. La consulta utilizada fue (SELECT * FROM mensajeria_isaserver), luego estos datos son reflejados en un componente de entrada que se crea en el Pentaho (estos componentes de entada son los encargados de buscar la información que se necesite de la base de datos que se utiliza como fuente), en este caso el componente creado fue Mensajeria_isaserver. Posteriormente en otro componente (Sel/Ren dim_usuario) se escogen de estos datos los que realmente se necesitan para colmar la dimensión y se le realiza una transformación de renombre (los datos escogidos pueden ser renombrados para un mejor entendimiento del usuario en un reporte final). Después estos datos pasan por otro componente

(Ordenar filas dim_usuario) en el cuál se ordena la información en la manera que se desee. Por último estos datos son llevados y cargados en una tabla que se crea automáticamente en el mercado la tabla dimensión (Salida Tabla dim_usuario). Este proceso no sucede hasta que se oprima un botón en la herramienta que lo mande a ejecutarse, así ocurre para cada una de las dimensiones del proceso hasta conformar el cubo.

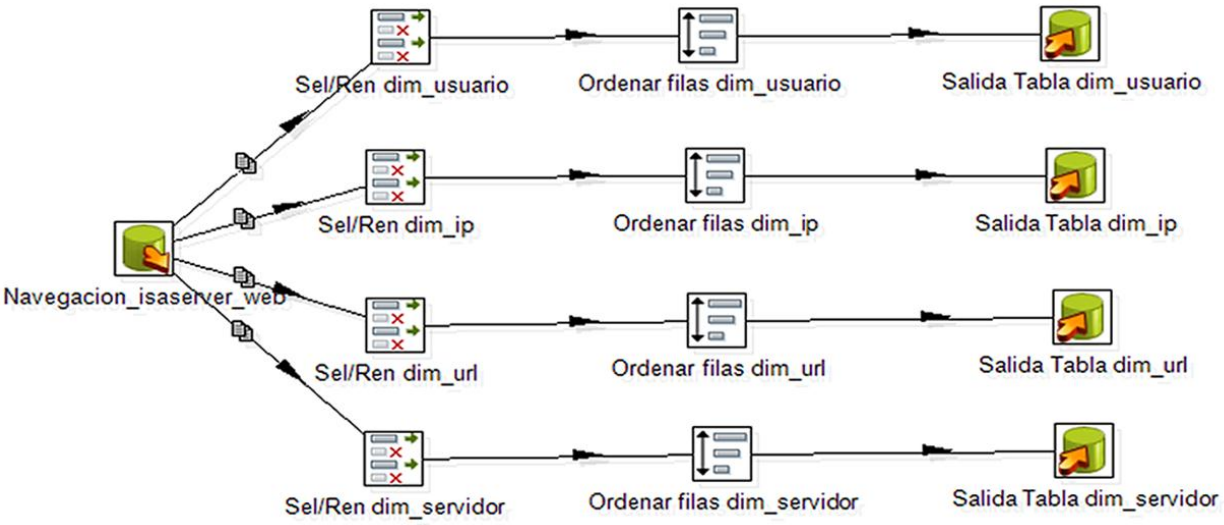


Figura 14: ETL del proceso análisis de los eventos generados por el servicio de Navegación ISA Server.

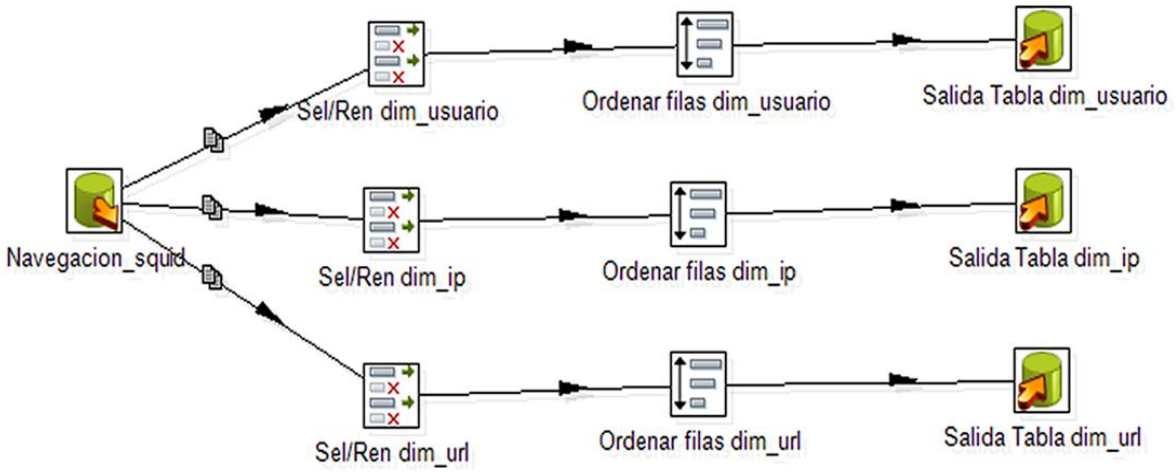
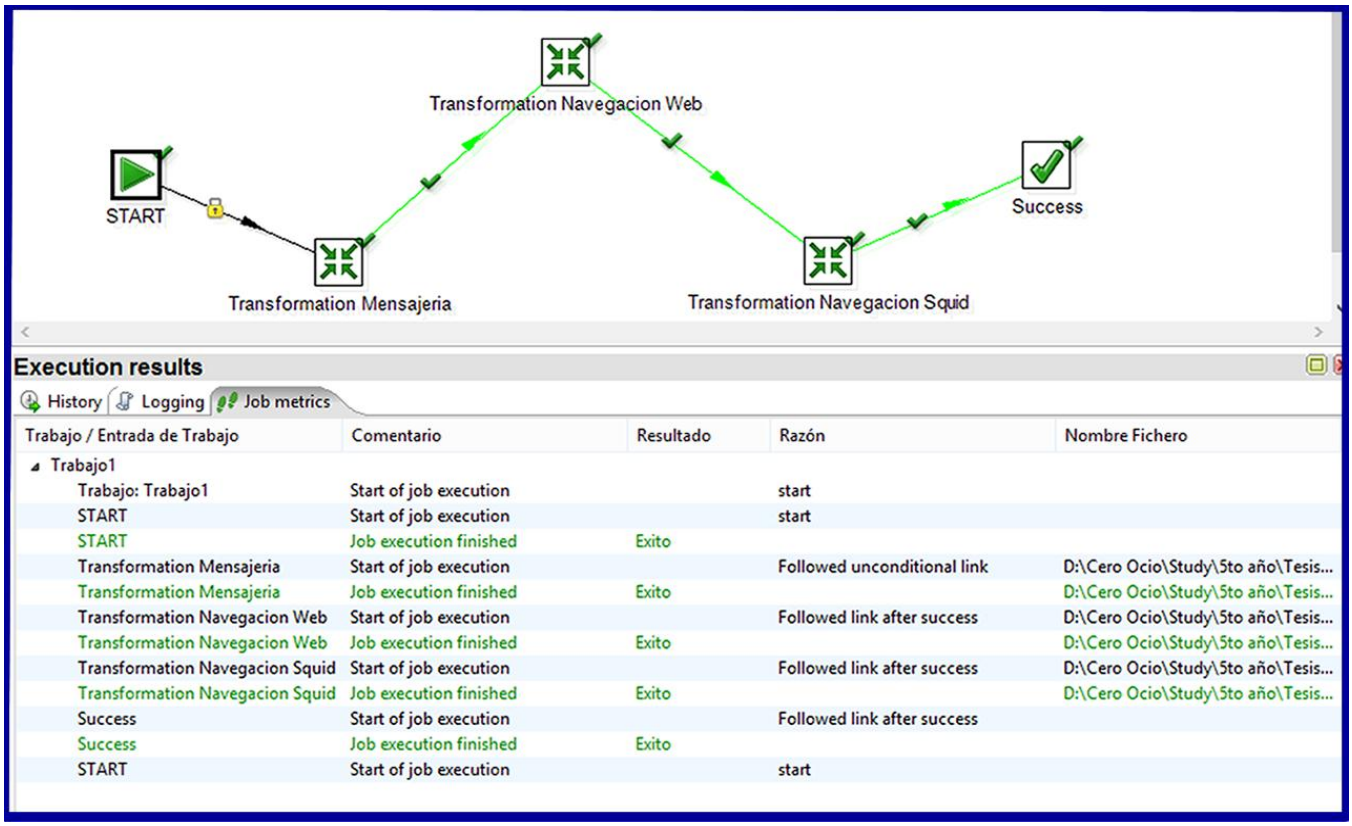


Figura 15: ETL del proceso análisis de los eventos generados por el servicio de Navegación Squid.

Después de realizar las transformaciones a todos los procesos del negocio se diseñó un trabajo con el objetivo de iniciar todos los procesos ETL de manera consecutiva y no ejecutarlos manual uno por uno. Para lograr la realización de esto se configura un botón (START) donde se especifica la fecha exacta de ejecución de las transformaciones, después se seleccionan las transformaciones que se quieren ejecutar y se oprime el botón ordenando a comenzar la ejecución. Por último el Spoon muestra si las transformaciones se ejecutaron satisfactoriamente o no. A continuación se muestra una imagen del proceso descrito.



Trabajo / Entrada de Trabajo	Comentario	Resultado	Razón	Nombre Fichero
Trabajo1				
Trabajo: Trabajo1	Start of job execution		start	
START	Start of job execution		start	
START	Job execution finished	Exito		
Transformation Mensajeria	Start of job execution		Followed unconditional link	D:\Cero Ocio\Study\5to año\Tesis...
Transformation Mensajeria	Job execution finished	Exito		D:\Cero Ocio\Study\5to año\Tesis...
Transformation Navegacion Web	Start of job execution		Followed link after success	D:\Cero Ocio\Study\5to año\Tesis...
Transformation Navegacion Web	Job execution finished	Exito		D:\Cero Ocio\Study\5to año\Tesis...
Transformation Navegacion Squid	Start of job execution		Followed link after success	D:\Cero Ocio\Study\5to año\Tesis...
Transformation Navegacion Squid	Job execution finished	Exito		D:\Cero Ocio\Study\5to año\Tesis...
Success	Start of job execution		Followed link after success	
Success	Job execution finished	Exito		
START	Start of job execution		start	

Figura 16: Trabajo para la ejecución de los procesos ETL.

3.1.3 Construcción de reportes en la herramienta de Respuesta

Para la construcción de los reportes se utilizó la herramienta Pentaho Report Designer. Pentaho Report Designer es una herramienta de reportes que permite crear informes, bien para ejecutarlos directamente o para publicarlos en la Plataforma BI y que desde allí puedan ser utilizados por los

usuarios. La herramienta es independiente de la plataforma y forma parte del conjunto de herramientas de la suite de Pentaho. El proceso de creación de reportes se describe a continuación.

El proceso de creación de reportes comienza con la conexión a la base de datos de donde se extraerán los datos para generar los reportes. Posteriormente se realizan las consultas para obtener los datos necesarios y por último utilizando estos datos, se diseña el reporte de manera sencilla. A continuación se muestra la realización de un reporte por proceso de negocio en este caso tres reportes y se especifica cada uno de los pasos a seguir.

- **Proceso análisis de los eventos generados por el servicio de Mensajería ISA Server**
  Consulta realizada para devolver los primeros 20 usuarios existentes en la base de datos y la cantidad de mensajes q han enviado.
    - Select distinct usuario, count(usuario) as cantidad from public.dim_usuario_mensajeria_isaserver inner join public.dim_tiempo on id_tiempo = id_usuario_msg_is where mes_nombre between ${mesnombre1} and ${mesnombre2} Group By usuario Order By cantidad Desc limit 20.

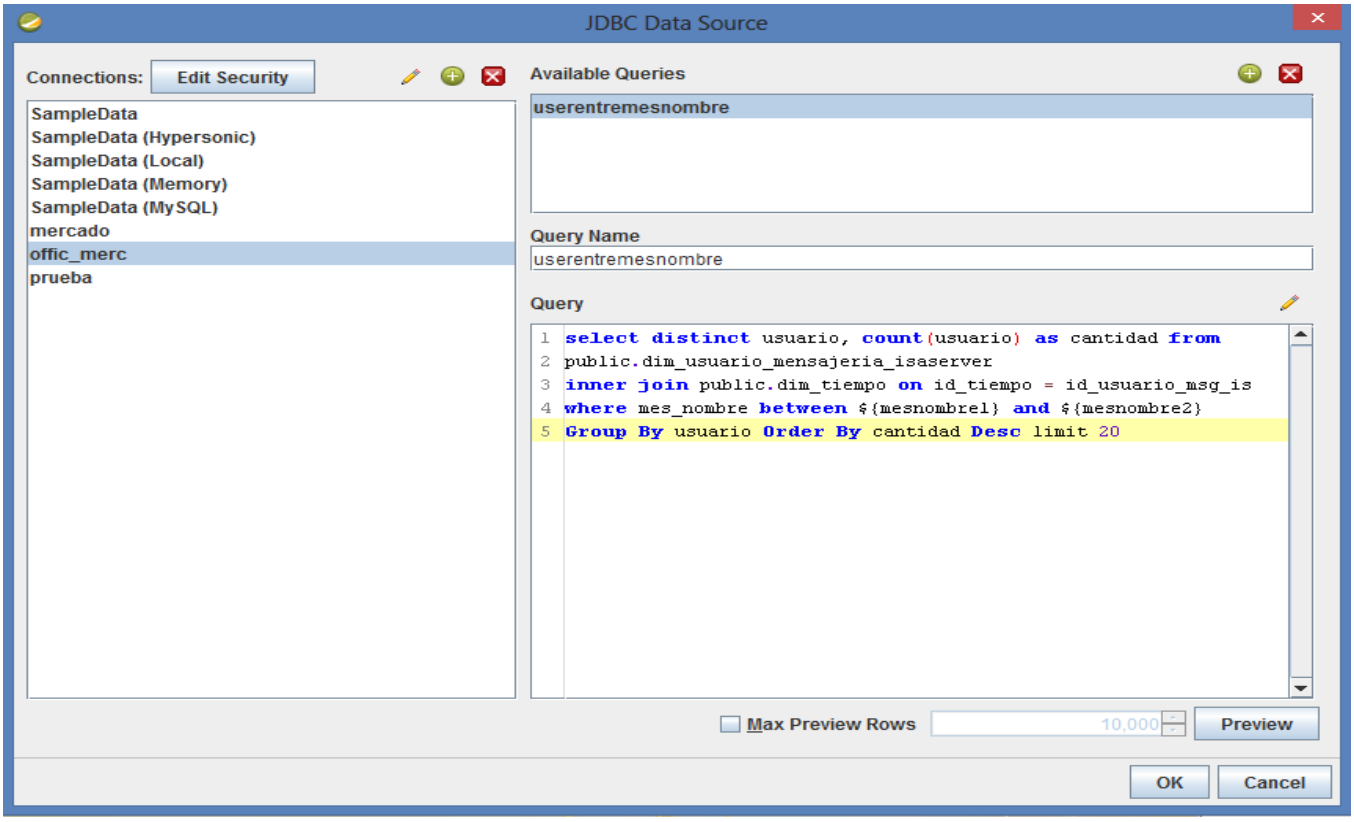


Figura 17: Consulta para extraer de la base de datos.

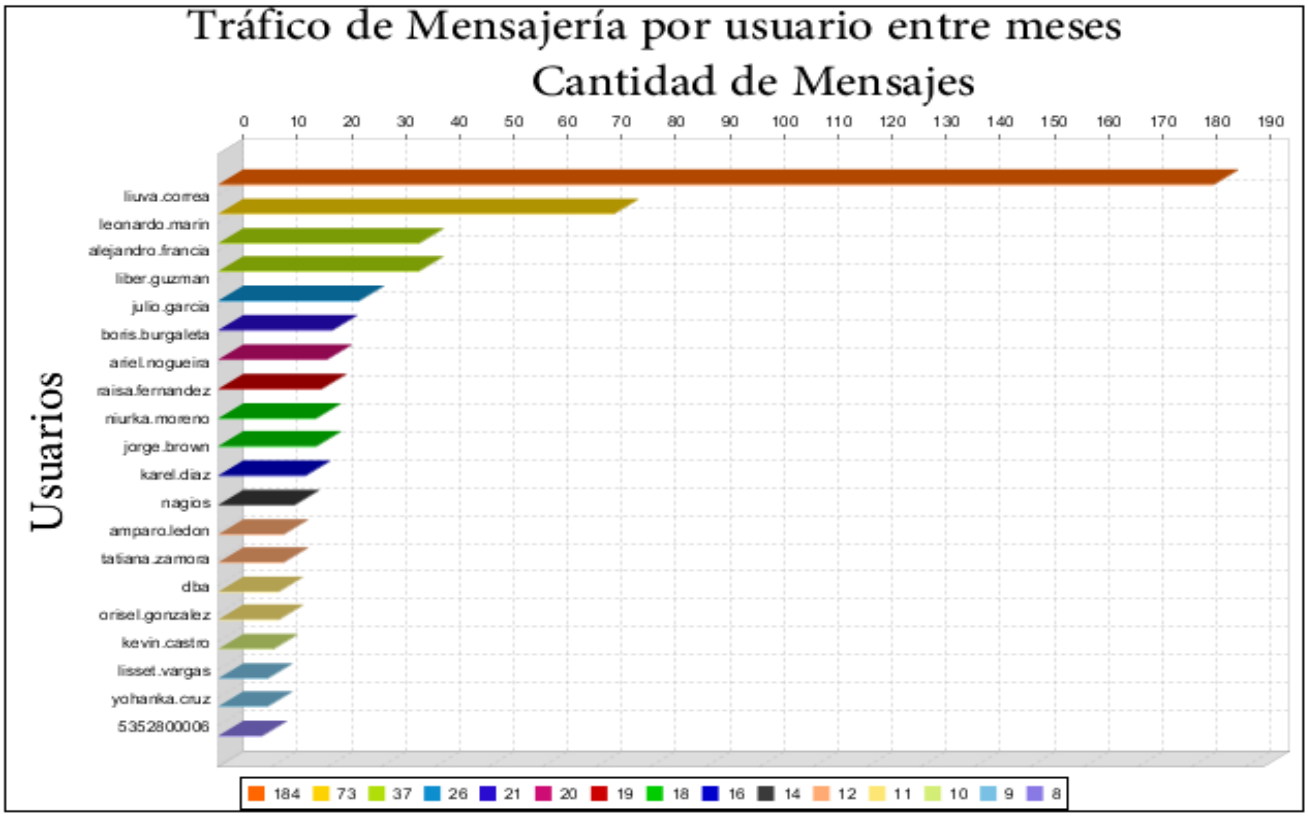


Figura 18: Reporte para el tráfico de mensajería por usuario entre meses.

3.1.4 Implementación de subsistema de visualización

3.1.4.1 Esquema OLAP

Para la implementación del subsistema de visualización, es necesaria la creación de los cubos multidimensionales, en los cuales se definen las dimensiones, los niveles de jerarquía de las dimensiones, las medidas físicas, las medidas calculables, las propiedades. Esto se realiza a través de la herramienta Pentaho Schema Workbench, donde la misma genera un fichero XML, el cual contendrá la estructura de todos los cubos OLAP.

En la realización del diseño de los cubos, se modeló el esquema MD_ETECSA integrado por tres cubos: hech_nav_isaserver, hech_nav_squid y hech_msg_is, donde hech_nav_isaserver está compuesto por cinco dimensiones mientras que los otros dos cubos están compuestos por cuatro. La dimensión *tiempo* es común para todos. Las siguientes imágenes muestran una vista de la aplicación.

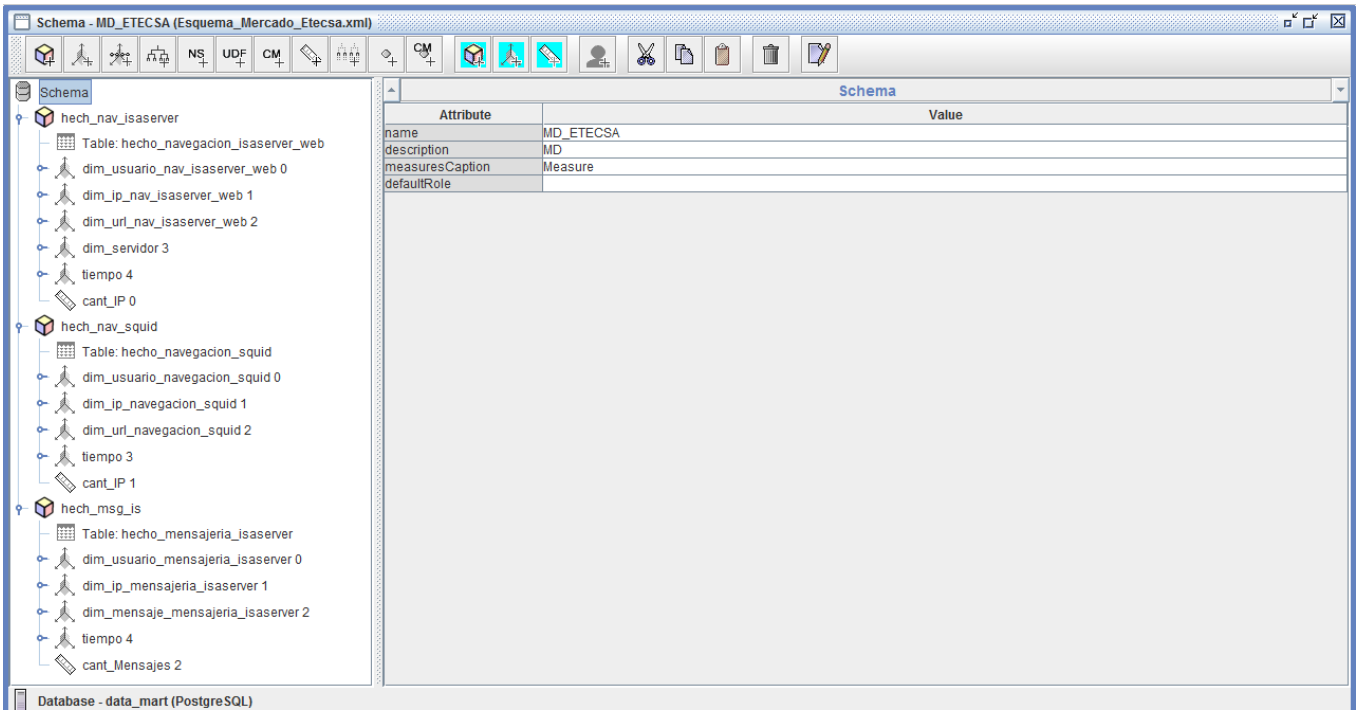


Figura 19: Diseño del esquema.

Las siguientes imágenes muestran la estructura de algunos cubos:

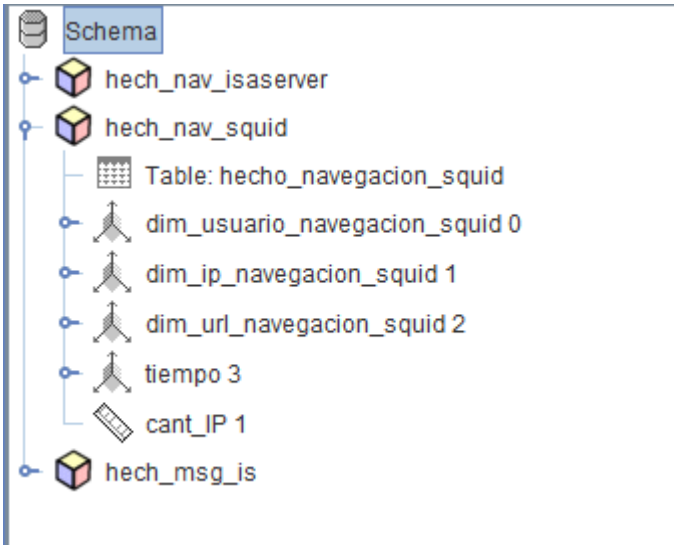


Figura 20: Cubo navegación ISA Server.

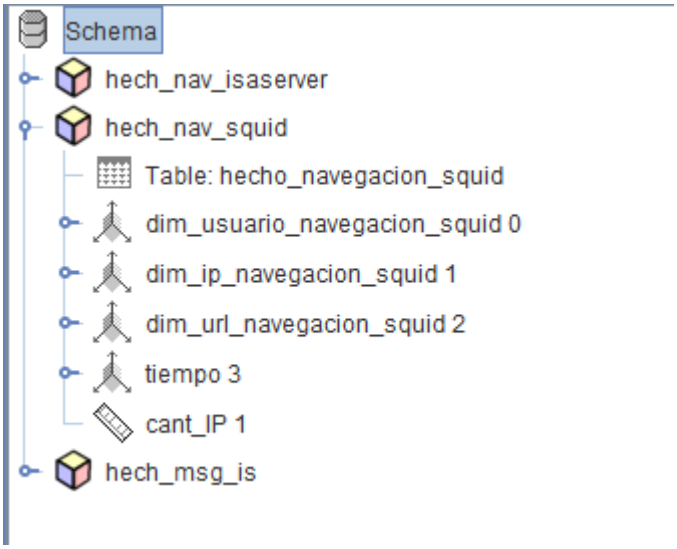


Figura 21: Cubo navegación Squid.

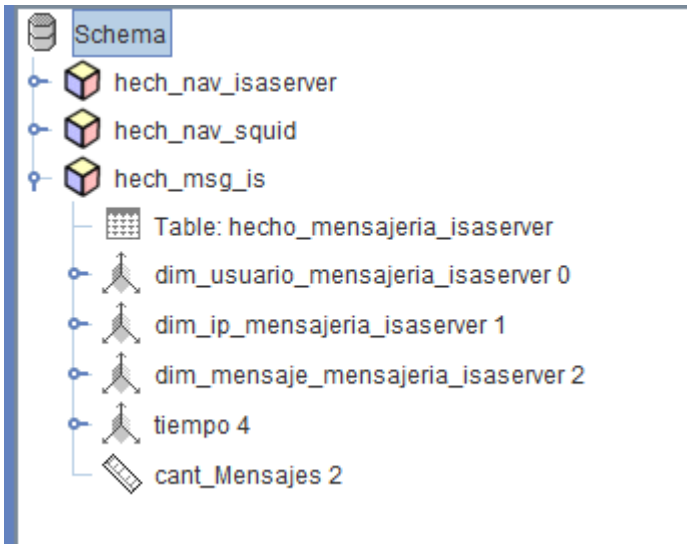


Figura 22: Cubo mensajería ISA Server.

3.1.4.2 Navegación de la capa de visualización

El mapa de navegación es la representación gráfica en que está organizada la información. El mercado de datos Comunicaciones está compuesto por un Área de Análisis General (A.A.G), tres Áreas de Análisis (A.A), nueve Libros de Trabajo (L.T) y 60 Archivos o Tablas de Salida (TS). A continuación se detalla la estructura en la cual es presentada la capa de visualización

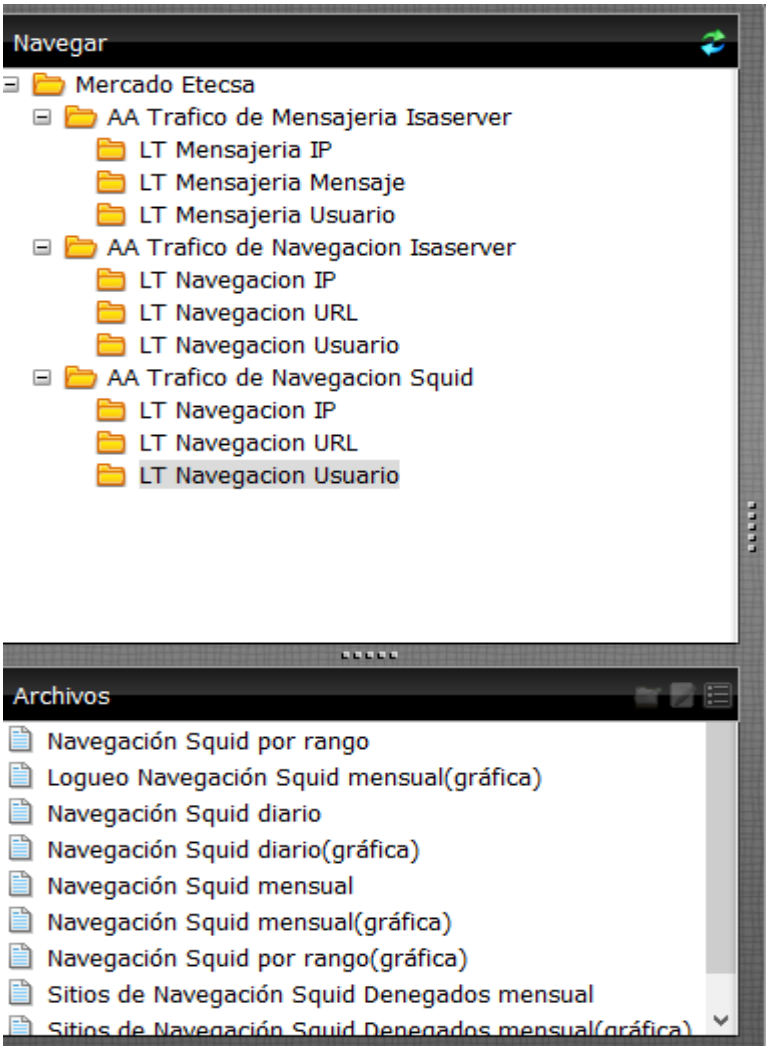


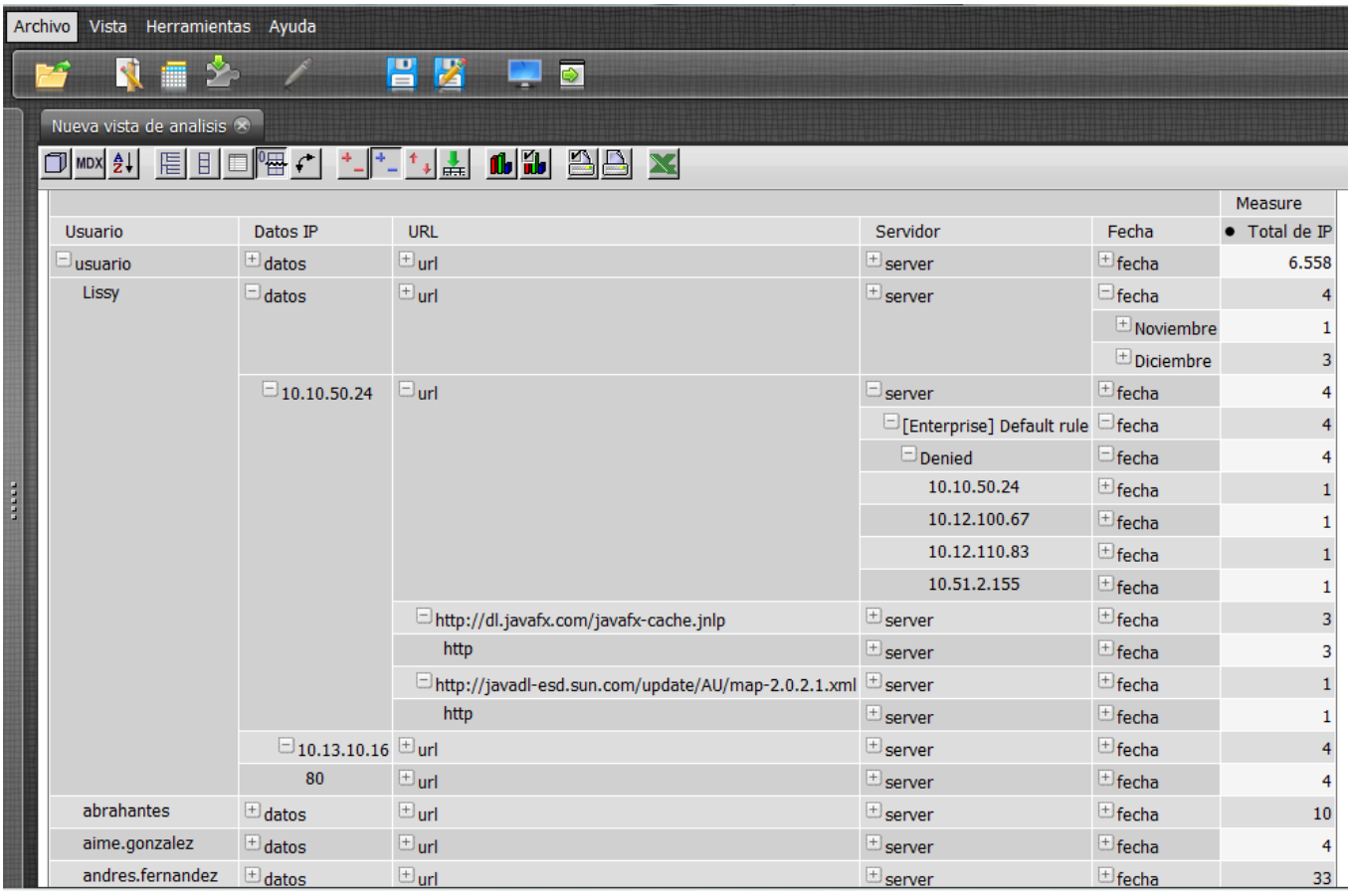
Figura 23: Mapa de navegación.

- **Descripción del A.A.G (Mercado ETECSA):** Agrupa toda la información de las tres áreas de análisis.
- **Descripción del A.A Tráfico de mensajería ISA Server:** Agrupa toda la información referente al Tráfico de mensajería ISA Server.
- **Descripción del A.A Tráfico de navegación ISA Server:** Agrupa toda la información referente al Tráfico de la navegación ISA Server.

- **Descripción del A.A Tráfico de navegación Squid:** Agrupa toda la información referente al Tráfico de la navegación Squid.

  - ✓ **LT Navegación IP**
  - ✓ **LT Navegación URL**
  - ✓ **LT Navegación Usuario:** Agrupa la información referente a las conexiones realizadas por los usuarios. Contiene reportes que responden a esta información.

Para la realización de estas vistas se utilizó el Mondrian, el Tomcat y el Pentaho BI Server. En la siguiente imagen se representa un ejemplo de cómo es la estructura de una vista de análisis.



Usuario	Datos IP	URL	Servidor	Fecha	Measure: Total de IP
usuario	datos	url	server	fecha	6.558
Lissy	datos	url	server	fecha	4
				Noviembre	1
				Diciembre	3
	10.10.50.24	url	server	fecha	4
			[Enterprise] Default rule	fecha	4
			Denied	fecha	4
			10.10.50.24	fecha	1
			10.12.100.67	fecha	1
			10.12.110.83	fecha	1
			10.51.2.155	fecha	1
		http://dl.javafx.com/javafx-cache.jnlp	server	fecha	3
		http	server	fecha	3
		http://javadl-esd.sun.com/update/AU/map-2.0.2.1.xml	server	fecha	1
		http	server	fecha	1
	10.13.10.16	url	server	fecha	4
	80	url	server	fecha	4
abrahantes	datos	url	server	fecha	10
aime.gonzalez	datos	url	server	fecha	4
andres.fernandez	datos	url	server	fecha	33

Figura 24: Vista de Análisis.

3.2 Pruebas

Las pruebas son un conjunto de actividades en las cuales un sistema es ejecutado bajo condiciones o requisitos específicos, donde los resultados son observados y registrados para dar una evaluación de algún aspecto del sistema determinar así la calidad del mismo.

3.2.1 Pruebas al mercado de datos

Las pruebas fueron realizadas con el objetivo de comprobar que el sistema cumple con los requisitos de información detectados en etapas anteriores. Se realizan para identificar posibles fallos en la implementación del sistema.

- **Descripción de los métodos de prueba a realizar**

Los métodos de pruebas definen que estrategia seguir en cuanto a verificación y validación del sistema, ya que están diseñados con el propósito de descubrir fallos y no para demostrar que el software funciona.

Existen dos métodos de prueba: Pruebas de caja blanca, estas pruebas comprueban los caminos lógicos dentro del sistema y son derivadas a partir de las especificaciones internas del diseño o del código. Otro de los métodos son las Pruebas de caja negra las cuales pretenden demostrar que las entradas al software se aceptan de forma adecuada y se produce el resultado correcto. Dada la necesidad de que las pruebas abarquen los requerimientos, las funciones y las respuestas de la aplicación, se decide utilizar las pruebas de caja negra. Este método presenta varias técnicas de pruebas a emplear. Para realizar las pruebas al Mercado de Datos se utilizó la técnica lista de chequeo, que es la definida por la metodología utilizada, a continuación se describe con más detalle el método y la técnica:

- **Lista de Chequeo**

La lista de chequeo realiza una serie de preguntas en forma de cuestionario para probar la solución. Su objetivo principal es dar una valoración de la calidad del producto.[33]

La lista de chequeo que se utilizó fue:

Peso	Criterio de evaluación	Eval	NP	Cantidad de elementos afectados	Comentarios

Crítico	¿Los subtítulos de las filas y las columnas corresponden a la descripción de sus atributos?				
Crítico	¿Los reportes son configurables a través de la interfaz del sistema?				
	¿La interfaz está orientada a facilitar el uso de las funciones del sistema por parte de los usuarios?				
Crítico	¿Los usuarios son capaces de manipular los resultados de manera que se ajusten a sus necesidades, conformando nuevos reportes?				Solo tienen permiso de conformar nuevos reportes el usuario "Especialista del Departamento de Seguridad Informática".
Crítico	¿Los procesos pueden ser analizados desde todas sus dimensiones?				
Crítico	¿El título de la gráfica se corresponde con los datos mostrados?				
Crítico	¿Los parámetros de los				

	requisitos se muestran en los reportes?				
	¿La interfaz tiene errores ortográficos?				
	¿El tipo de gráfica es el más representativo para mostrar la información?				

3.2.1.1 Resultados de las pruebas

Para evaluar el mercado desarrollado se planificaron 2 iteraciones de pruebas en las cuales se probó la solución con un alto grado de detalle. A continuación se muestran la cantidad de No Conformidades identificadas en las iteraciones realizadas.

Clasificaciones	**1ra iteración**	**2da iteración**
Alta	9	4
Media	16	7
Baja	15	3

Se muestran de forma detallada un gráfico de los errores encontrados durante las dos iteraciones realizadas:

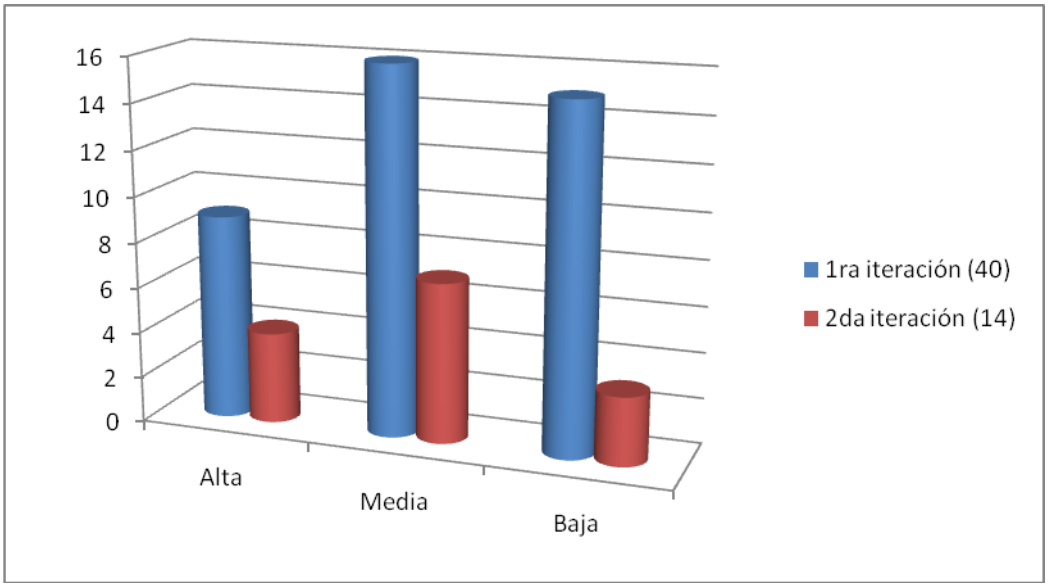


Figura 25: Resultado de pruebas.

Luego de concluida cada iteración de pruebas se analizaron, por parte del equipo de desarrollo, las No Conformidades encontradas para determinar cuáles realmente constituyeron defectos del sistema. Las pruebas se realizaron de forma iterativa e incremental, comprobando en cada iteración que hubiesen sido corregidos los errores detectados en la iteración anterior, lo que contribuyó a mejorar la calidad y funcionalidad del software, no se realizó una tercera iteración porque se corrigieron todas las no conformidades encontradas en la segunda.

3.3 Conclusiones Parciales

En el desarrollo de este capítulo se realizó la implementación del mercado de datos donde se efectuó el perfilado de datos del cual se obtuvo a través de varios reportes generados por el mismo, un conocimiento más profundo del estado de los datos fuentes. Se desarrolló la implementación de las ETL obteniendo como resultado un mercado de datos poblado y funcional. Se diseñaron los cubos multidimensionales en correspondencia con cada una de las tablas de hecho del mercado de datos y con las necesidades de información. Se realizó la implementación del subsistema de visualización, quedando conformadas las vistas de análisis de acuerdo a las necesidades de información del cliente. Se aplicó una lista de chequeo como prueba al mercado de datos.

CONCLUSIONES

Una vez concluida la implementación del mercado de datos se llegaron a las siguientes conclusiones generales:

- Se realizó el análisis y diseño del mercado de datos acorde con lo previsto en el negocio, dándose respuesta a los problemas planteados.
- Las estructuras dimensionales que fueron implementadas cumplen las condiciones necesarias para el proceso de integración de los datos.
- La implementación de los subsistemas de integración y de visualización permitió obtener como resultado un mercado de datos poblado y funcional, los reportes fueron desarrollados en su totalidad con información disponible para ser consultada por parte de los usuarios, apoyando el proceso de toma de decisiones.
- Se realizaron las con éxito las pruebas pertinentes que permitieron validar la calidad del producto, arrojándose algunas no conformidades que fueron solucionadas.

RECOMENDACIONES

Se recomienda:

- Agregarle nuevos procesos al mercado como lo son el resto de los servicios telemáticos.
- Seguir aumentando los reportes y los análisis de la solución con versiones superiores de la suite utilizada.
- Desplegar el mercado de datos en los centros de telecomunicaciones del país.
- Que se mantenga estable el personal que labora con el sistema, para lograr una evolución tanto en la mentalidad de los analistas, como en las necesidades de información.
- Desarrollar un tablero de mando para el mercado de datos implementado

REFERENCIAS BIBLIOGRÁFICAS

1. Benítez, A., *Toma de decisiones. Etapas del proceso.* 2009.
2. *La toma de decisiones (1): Conceptos básicos*, in *Pyme Activa*. 2012.
3. Antunez, I.M., *La Inteligencia de Negocio desde la perspectiva cubana: retos y tendencias*, Universidad de la Habana: La Habana, Cuba. p. 9.
4. Etecsa. 2012 [cited 2013; Available from: http://www.etecsa.cu/.
5. *Mantener un ojo en tu web proxy con el uso de Squid Gráfico.* 2008.
6. RIA-Media, *ISA Stats.* 2008.
7. Palenzuela, O.B., *Título: ISAWEB. Monitoreo de Tráfico en internet.*
8. Softonic, E.d., *Genera estadísticas a partir de logs.* 2008.
9. Computing, W., *Cyfin Reporter 8.1.0.* 2010.
10. *Una herramienta completa de Generación de Reportes MS Exchange* 2011.
11. Mendez, F., *Beneficios de Visual Studio 2010 Professional.* Hablemos de Tecnoligia.
12. Umaloop, *Umaloop MailStat.* 2011.
13. e-project, *SurfControl web Filter - Filtro de contenidos web*
14. Sánchez, L., *Spamina lanza una nueva versión de Cloud Email Firewall.* 2011.
15. Domínguez, A.B., *Data Warehousing.* 2003: p. 1.
16. Bernabeu, R.D. (2010) *DATA WAREHOUSING: Investigación y Sistematización de Conceptos - HEFESTO: Metodología propia para la Construcción de un Data Warehouse*. **V2.0**.
17. Orol, A.M., *OLAP y el diseño de cubos.* 2007.
18. Nieve, M.E.y.M.C.d., *Business Intelligence*. 2011.
19. Tamargo, L.C. *Diseño Físico del Data Warehouse*.
20. Dominguez, A.B., *Data Warehousing.* Programación en Castellano., 2003.
21. Ross, R.K.a.M., *Relentlessly Practical Tools for Data Warehousing and Business Intelligence*. 2010: Indianapolis, Indiana
22. *La plataforma Pentaho Open Source Business Intelligence.* 2008.
23. *Kettle Pentaho Data Integration.* PENTAHO 2009; Available from: http://kettle.Pentaho.org/.
24. *Pentaho Open Source Business Intelligence.* PENTAHO, 2005.
25. *Pentaho BI Platform Server.* SUMMAN, 2006.
26. *The PostgreSQL Global Development Group.* PostgreSQL, 2005.

27. Erlijman Piwen, G.F., *Problemas y soluciones en la Implementación de Extreme Programming* in *Universidad Católica*. 2011: Uruguay
28. *Squid.* Ecured.
29. Imon, W.H., *"Building the Data Warehouse" Cuarta edición*. 2005.
30. Dario, I.B.R., *Datawarehouse manager.* 2009.
31. Díaz, J.C., *Data quality: definiciones.* 2011.
32. Dario., I.B.R., *Manual del Usuario de Spoon.* 2010.
33. Yoraima Peña Hinojosa, F.G.P., *Desarrollo de un Mercado de Datos para el Área de Ingreso.* 2012, UCI.Universidad de las ciencias informaticas: La Habana, Cuba. p. 71.