Universidad de las Ciencias Informáticas

Facultad 3

Título: Implementación del proceso de extracción, transformación y carga en un almacén de datos operacional para CIMEX.

Trabajo de Diploma para optar por el título de Ingeniero en Ciencias Informáticas

Autoras:

Yohanlena Hartman Díaz

Dailen Ramón Zequeira

Tutor:

Ing. Yonelbys Iznaga González

Ciudad de la Habana

Mayo, 2009

"Los que se enamoran de la práctica sin la teoría son como los pilotos sin timón ni brújula, que nunca podrán saber a dónde van."

Leonardo Da Vinci

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 firmo la presente a los _____ días del mes de _____ del año 2009.

Yohanlena Hartman Diaz

_____

Firma del Autor

Dailen Ramon Zequeira

_____

Firma del Autor

Yonelbys Iznaga Gonzalez

_____

Firma del Tutor

Agradezco a todos los que facilitaron mi estancia en la universidad: los que enseñaron, los que criticaron, los que dejaron huellas a seguir. A mis padres, por el apoyo incondicional, por el infinito amor que a diario cultivan, por dejarme trazar mi camino, por confiar, por permitirme venir desde tan lejos a crecer. A las embajadoras (Maylen, Mayen, Marisol, Dailen y Yudith), por dejarme compartir con ellas estos años, por prestarme sus experiencias, por volar juntas con la libertad desconocida. A los profesores que me ayudaron a descubrir lecciones, y a los demás que hicieron su mejor esfuerzo. A Mailen, por seguir siendo mí mejor amiga aun en la distancia. A Tony, por regalarme sanas ironías que iluminaron el camino. A Ivet, por darle vida a mi buzón, por las muchas razones que me diste para seguir. A Ariel, por la ayuda invaluable y los sabios consejos. A Posada, por los mejores y peores momentos. A Yonelbys, porque sin ti no estaría escribiendo agradecimientos y esta tesis aún seria un manojo de archivos desordenados. Y a todos los que no menciono, pero de quienes guardo alguna historia. Gracias por la colaboración, por aun seguir aquí...

Yohanlena

La gratitud es la memoria del corazón, y son muchas las memorias a las cuales les debo agradecer por ser hoy lo que soy. Dice un viejo proverbio chino que cuando bebas agua, debes recordar siempre la fuente, hoy, es realmente difícil agradecer a todas aquellas fuentes que cuando he tenido sed han estado ahí para mí y me han bridado sus cristalinas aguas, para poder seguir adelante y llegar a realizar mi gran sueño. De todas formas y siguiendo el proverbio haré el intento.

En la vida uno tiene muchos tesoros, para mí, los más valiosos han sido mis abuelitos, quienes me guiaron en mis primeros años de vida y me dieron todo su amor, es por ello, que mi primer agradecimiento esta vez será para mima, donde quiera que estés y a pipo por saber darme las alas para volar. Para mi papito y mi mamita mi eterno agradecimiento por darme la vida, por su amor, cariño, comprensión, por estar ahí cuando los he necesitado, por apoyarme y por estar orgullosos de mí, arma fundamental para seguir adelante cada día. A mi tía Isora por todo su amor, por cuidarme y preocuparse por mí, por acogerme y hacerme sentir como una hija más. Al resto de mi familia por ser tan unida, por demostrarme que juntos podemos soportarlo todo, por su apoyo y por confiar siempre en mí.

Los amigos son un regalo de dios que pasan por la vida dejando huellas imborrables; a todos ellos mi agradecimiento por dejarme formar parte de sus vidas, en especial a las "embajadoras" por compartir nuevas experiencias, por las alegrías y las tristezas, a Mary por apoyarme siempre y abrirme su corazón, a Laira por nuestra eterna amistad y a Mailin, Yadira y Mayu porque estoy segura que siempre estaremos ahí cuando nos necesitemos. El amor es abono que toda planta necesita para seguir adelante, gracias le quiero dar en esta ocasión a Darián por su amor infinito, por ser mi abono, por soportar mis malacrianzas por estar a mi lado cuando lo necesito, por apoyarme y por ser quien es.

Gracias a todos aquellos que me abrieron las puertas y me dejaron entrar a este maravilloso mundo, a todos los que han confiando en mí, a los que me han brindado un sorbo de agua para seguir en pie…

Dailen

A mi mama, por ser la luz de mi vida, por el amor que nunca me cobras, por la sencillez y ternura que profesas, porque no imagino mi vida sin ti.

A mi papa, por el apoyo incondicional, por ser el amigo que no juzga, por ser una parte inseparable de mí.

A mi hermana, quien siempre me muestra la cara de la moneda que no alcanzo a ver, por ser el espejo que no me permite caer.

A Julio Alberto, por las muestras de afecto de todos estos años, porque siempre serás mi familia.

A mi sobrino, por sembrar sonrisas en mis días, ojala mis huellas algún día te sirvan para construir un mejor camino.

Yohanlena

A mima, mi ángel, donde estés, por tu infinito amor y entrega.

A pipo por estar tan orgulloso de mí, por todo su cariño, dedicación, por ser mi inspiración para lograr lo que soy.

A mi papito por ser mi sostén, mi amigo, por su eterno amor, por todo lo que representas para mí.

A mi mamita por su gran amor, por ser mi confidente y amiga, por confiar y apoyarme, por estar siempre junto a mí.

A mi tía Isora, por darme la posibilidad de tener una madre más.

Dailen

El proceso de Extracción, Transformación y Carga permite mover datos desde múltiples fuentes, reformatearlos, limpiarlos, y cargarlos en otra base de datos para analizar, o en otro sistema operacional para apoyar algún proceso de negocio. A grandes rasgos consiste en extraer los datos desde los sistemas de origen, convertir los datos a un formato preparado para iniciar el proceso de transformación, donde se aplican una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados en el sistema de destino.

En este trabajo, se implementará un proceso de extracción, transformación y carga a los datos ubicados en una base de datos que pertenece a la corporación CIMEX; para la creación y carga a un almacén de datos operacional, que facilitará una adecuada gestión del conocimiento y toma de decisiones acerca del negocio mayorista de esta empresa.

INDICE

INTRODUCCIÓN

El mundo actual evoluciona a pasos agigantados, diariamente la tecnología progresa, dando enormes avances al desarrollo científico-técnico. Este crecimiento tecnológico provoca una feroz competencia en el mercado, así las empresas se esfuerzan en elevar su nivel profesional en aras de ganar prestigio y demanda entre los clientes. Debido a esta alta competitividad, la información que se genera en las empresas crece en volumen, provocando que la toma de decisiones sea más difícil y que el acceso a la información almacenada requiera mayor rapidez y precisión. Una buena solución para lograr que la información sea fiel, rápida, útil y precisa lo constituye el uso de Inteligencia de Negocios[1], como proceso para analizar los bienes o datos acumulados en la empresa y extraer una cierta inteligencia o conocimiento de ellas. En inteligencia de negocios se engloban muchas funciones como; multidimensionalidad, data mining, y data warehouse. El motor impulsor de esta inteligencia de negocios lo constituye el proceso de Extracción, Transformación y Carga[2] para la integración de datos, garantizando que estos sean precisos, completos, creíbles, accesibles, rigurosos en el tiempo y con gran calidad.

La corporación CIMEX[3] es una empresa mayorista y minorista que tiene gran prestigio en Cuba. CIMEX utiliza el sistema automatizado Sentai para la gestión empresarial de su comercio mayorista por excelencia, este software ha presentado ciertos inconvenientes; pues a pesar del buen funcionamiento desde el punto de vista transaccional, no es posible realizar sobre este sistema procesamiento analítico de gran envergadura, las consultas muy complejas se ejecutan lentamente debido a la forma en que están estructurados sus datos, por lo que no satisfacen las necesidades de eficiencia que solicitan los analistas, además los datos que gestiona son inconsistentes, no están organizados ni integrados de la forma más óptima. Por esta razón la empresa pretende desarrollar un Almacén de Datos Operacional[4], para ello necesita que los datos se encuentren limpios, organizados, integrados y preparados para luego ser cargados en el ODS propuesto.

[1] Conocido este término como BI (Business Intelligence) por sus siglas en inglés.

[2] Proceso conocido en el mundo comercial como ETL (Extract, Transform and Load) por sus siglas en inglés.

[3] Corporación Importadora y Exportadora.

[4] También llamado ODS (del inglés Operational Data Store) es un contenedor de datos activos, es decir operacionales que ayudan al soporte de decisiones y a la operación.

Debido a esto se identifica el siguiente **problema científico**: ¿Cómo lograr la preparación, limpieza y organización de los datos a integrar en un almacén de datos operacional para CIMEX?

Así, este trabajo se encargará de realizar un acercamiento al proceso ETL, aportando los métodos y herramientas necesarias para reformatear, limpiar y cargar los datos desde múltiples fuentes a un almacén de datos operacional vinculado a la corporación CIMEX. Definiendo como **objeto de estudio**: Proceso de extracción, transformación y carga de datos. Y a partir del análisis de este se determina como **campo de acción**: Proceso de extracción, transformación y carga de datos en un almacén de datos operacional para CIMEX.

Para certificar que los datos a cargar en el almacén tengan la claridad y organización necesaria, garantizando con esto una mejor toma de decisiones se ha propuesto como **objetivo general**: Implementar el proceso de extracción, transformación y carga en un almacén de datos operacional para CIMEX. Concretándose en los siguientes **objetivos específicos**:

1. Realizar un estudio del estado del arte del proceso de extracción, transformación y carga de datos.
2. Seleccionar y estudiar las herramientas de código abierto que brinden los servicios y las funcionalidades necesarias para realizar la extracción, transformación y carga de datos.
3. Implementar el proceso de extracción, transformación y carga de datos hacia un almacén de datos operacional.

A partir de lo anteriormente expuesto se plantea la siguiente **hipótesis**: Si se realiza el proceso de extracción, transformación y carga en un almacén de datos operacional, se logrará la preparación, limpieza y organización de estos datos para dicho sistema.

Teniendo como **variable independiente**: Proceso de extracción, transformación y carga en un almacén de datos operacional. Y como **variable dependiente**: Preparación, limpieza y organización de los datos.

Para lograr el objetivo se plantean las siguientes **tareas investigativas**:

1. Realizar un estudio del estado del arte del proceso de extracción, transformación y carga de datos.

2. Realizar un estudio de las herramientas de código abierto para el diseño e implementación del proceso de extracción, transformación y carga de datos.
3. Realizar un estudio de los principales procesos del negocio que existen en CIMEX.
4. Realizar un estudio de los orígenes de datos de CIMEX.
5. Implementar el proceso de extracción, transformación y limpieza de los orígenes de datos de CIMEX.
6. Implementar el proceso de carga de datos al almacén de datos operacional.
7. Evaluar la eficiencia del proceso de extracción, transformación y limpieza, así como la calidad de los datos cargados.

De esta manera el **posible resultado** que se obtendrá será: La implementación del proceso de extracción, transformación y carga de datos en un almacén de datos operacional.

Para lograrlo se usan algunos métodos teóricos que brindan un carácter científico a este trabajo; entre ellos se encuentran: análisis - síntesis, que permite la división mental del fenómeno en sus múltiples relaciones para facilitar el estudio y la síntesis mental de la unión entre las partes previamente analizadas, posibilitando descubrir las características generales y las relaciones esenciales entre los elementos de la investigación; otro método es la inducción – deducción, procedimiento que facilita el estudio de hechos aislados arribando a proposiciones generales, o a partir de un hecho general inferir casos particulares por un razonamiento lógico. Otro método que se utilizó para obtener un buen desempeño de la investigación e implementación eficiente, fue la observación, como instrumento universal del científico.

Garantizando cumplir con lo antes expuesto, el presente trabajo se ha conformado de la siguiente forma; en el capítulo 1 se presentará un estudio del arte del proceso ETL, sus características, la importancia de emplear este proceso, y los desafíos que presenta su utilización, así como los roles y responsabilidades de cada uno de los implicados en el desarrollo de la solución. Brindando además la posibilidad de conocer las herramientas a manejar en la implementación del proceso ETL.

Una vez sumergidos en este mundo, conociendo sus particularidades y las herramientas con la que se trabajará para el desarrollo de este proyecto, la actividad fundamental se centra en la implementación del proceso ETL para CIMEX. Exponiendo estas ideas en el capítulo 2, para ello se realizará un estudio de los principales procesos del negocio que existen en CIMEX con el fin de garantizar un buen entendimiento entre el equipo de desarrollo y los clientes, se expondrán los orígenes de datos de CIMEX analizando detalladamente su estructura, para luego implementar el proceso de extracción, transformación y finalmente la carga de estos datos al almacén de datos operacional.

En el capítulo 3 se realizará una propuesta de validación de la calidad de los datos cargados. Evaluando la efectividad de la integración de los datos en el despliegue del almacén de datos operacional a quien el proceso de extracción, transformación y carga provee de datos limpios y en conformidad con la necesidad del CIMEX, permitiendo la toma de decisiones administrativas.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA

"En este nuevo mundo, la información reina…" afirma Geoffrey A. Moore[5]. Vivimos en una época donde la información es clave para obtener ventajas comerciales en el mundo de los negocios. Para mantenerse competitiva una empresa, sus gerentes requieren de un acceso rápido y fácil a información útil y valiosa de la organización. Una forma de solucionar este problema es mediante el uso de herramientas de análisis y aplicaciones analíticas, quienes se encargan de examinar los bienes o datos acumulados en la empresa y extraer una cierta inteligencia o conocimiento de ellos. Siendo el proceso de extracción, transformación, y carga, la base de esta inteligencia de negocios, pues un proceso ETL bien diseñado permite extraer datos del sistema fuente, asegura la calidad de los datos, hace cumplir las normas de coherencia, ajusta los datos a fin de que fuentes independientes se pueden utilizar de forma integrada, y ofrece a los datos un formato listo para que los desarrolladores pueden crear aplicaciones y los usuarios finales puedan tomar decisiones. Este capítulo se encargará de brindar un acercamiento al proceso ETL, como procedimiento que añade un valor significativo a los datos.

¿Qué es la integración de datos?

Los procesos de integración de datos están basados en la necesidad de aunar los datos pertenecientes a múltiples fuentes de datos con el fin de obtener de forma centralizada, una mirada única e integrada al problema en cuestión. Como principal enigma a la hora de integrar es que los datos son heterogéneos, se encuentran distribuidos, dispersos, y en la mayoría de los casos no estandarizados. Existen de esta forma numerosas islas de información inconsistentes que imposibilitan una comprensión unificada en cuanto a los términos, cantidades, unidades de medida, etc. Las entidades generadoras de datos, provocan que la tarea de unificar estos datos sea sumamente compleja y costosa. Es por ello que al hablar de integración de datos aparece el término calidad de datos, como una rama o subproceso a tener en cuenta. La información proveniente de diferentes sistemas es entonces inconsistente y con baja calidad, la gran mayoría de las veces al intentar compatibilizarla con otros sistemas, incluso dentro de las mismas empresas o entidades, es común encontrar múltiples sistemas, tecnologías, canales de comunicación y transporte cuando de integración de datos se trata. Tomando en consideración las diferencias entre la

[5] Director gerente en TCG Asesores, autor muy vendido en temas de asesoramiento y consulta de estrategias de transformación y modelos de negocio.

calidad de datos, los procedimientos para el manejo de datos de los expertos, los formatos, lenguajes y muchas otras inconsistencias, es preciso señalar que la plena y armoniosa integración de los recursos de información es un problema difícil de solventar. Es por esto que la integración de datos es necesaria, tomando fortaleza en las técnicas y tecnologías que maneja. (Kimball 1996)

Proceso ETL. Principales características.

El proceso de extracción, transformación y carga es el encargado de impulsar el flujo de datos haciendo transformaciones intermedias y permitiendo una integración de datos exitosa. Es por esto que cada paso, desde su diseño de acuerdo a los requerimientos de cada negocio, hasta su eficiente puesta en marcha, precisa de la mayor atención y esfuerzo posible. El proceso ETL permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos, limpiarlos, y cargarlos en otra base de datos, protegiendo el linaje de los datos. Encontrándose este proceso enfocado a la integración de datos, tanto por lote, como a tiempo real hacia almacenes, logrando un alto grado de transformaciones para la consolidación de la información. Sincronizando datos desde diversas aplicaciones e involucrando procesos de manipulación que van más allá de un simple movimiento desde el punto A hasta el punto B, donde este proceso no está separado de los sistemas operacionales, sino que está integrado con los demás procesos de la empresa. El proceso ETL fortalece datos para la construcción de bases de datos permanentes, dedicadas para el análisis o la generación de informes, utilizándose para migrar datos de una o más bases de datos a terceros, para formar repositorios de datos, data marts[6], almacenes de datos[7] y también para convertir bases de datos de un tipo o formato a otro, entre otras funcionalidades que hacen de este proceso imprescindible a la hora de integrar datos. (Kimball 2004)

Entre las características que describen las tecnologías de ETL se tienen:

- Es un mecanismo de carga muy eficiente y efectivo orientado a los almacenes de datos.
- Enfocado a migrar y mezclar datos.

[6] Versión especial de almacén de datos. Son subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones.

[7] Del inglés Data Warehouse, es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza.

- Reduce la exposición a desarrollos manuales (codificación) producto de la existencia en el mercado de herramientas potenciales para la implementación visual, con manejo de excepciones, gestión y planificación de tareas.
- Necesita pocos servicios de administración y mantenimiento.
- Gran capacidad para llevar a cabo transformaciones.
- Tecnología enfocada a la integración de datos en bases de datos versátiles hacia los almacenes de datos.

Como se puede dilucidar es un proceso complejo, pues precisa de un alto nivel de detalle, y en caso de ser mal diseñado puede provocar serios problemas operativos, por lo se debe regir por su arquitectura, donde la que se define para el tipo de solución en cuestión se presenta en la siguiente figura:

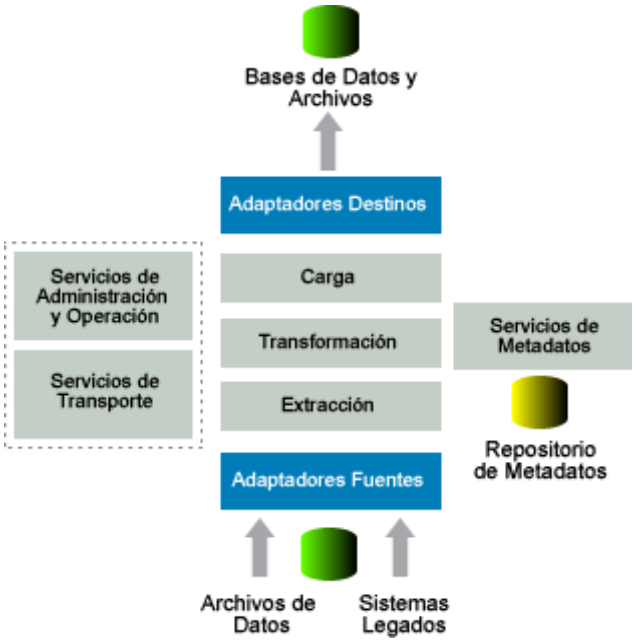


Fig. 1. Arquitectura de la solución ETL.

Las componentes de la arquitectura mostrada son los siguientes:

- Servicios de administración y operaciones: aseguran la utilización efectiva de los recursos en el ambiente de sincronización y una administración idónea mediante la planificación, seguimiento de tareas, gestión de metadatos[8] y recuperación de errores.
- Servicios de transportación: garantizan el movimiento de la información cruda o transformada desde una fuente hasta un repositorio destino.
- Servicios de metadatos: los metadatos son información descriptiva sobre los datos y otras estructuras, como objetos, reglas de negocio, y procesos que manipulan los datos. Los metadatos pueden ser agrupados en dos categorías:
    - Metadatos mecánicos: enfocado a los diseñadores, desarrolladores y administradores durante el desarrollo, y mantenimiento del proceso. Este es el punto técnico que agrupa las herramientas, aplicaciones y sistemas, para que juntos constituyan la solución.
    - Metadatos del negocio: brindan una imagen clara del servicio del ambiente de trabajo a los usuarios finales.

Teniendo el proceso ETL, tres subprocesos fundamentales, que permiten la división y entendimiento efectivo de este arduo trabajo. Los cuales se explicarán a continuación.

Extracción

Consiste en extraer los datos desde los sistemas de origen, estas fuentes primarias pueden encontrarse sobre arquitecturas, o estructuras heterogéneas, cada sistema separado puede usar una organización diferente de los datos o formatos distintos. Los formatos de las fuentes normalmente se encuentran en bases de datos relacionales o ficheros planos, pero pueden incluir bases de datos no relacionales u otras estructuras diferentes. La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación. Una parte intrínseca del proceso de extracción es verificar los datos extraídos, de lo que resulta un chequeo que comprueba si los datos cumplen lo que se espera, de no ser así los datos son rechazados. Un requerimiento importante que debe exigir la tarea de extracción es que ésta cause un mínimo impacto en el sistema origen, pues si los datos a extraer son muchos, el sistema de origen se podría colapsar, provocando que no pueda utilizarse con normalidad para su uso cotidiano. Por esta razón, en sistemas grandes las operaciones de extracción suelen programarse en horarios o días donde

[8] La definición más difundida de metadatos es que son «datos sobre datos».

este impacto sea nulo o mínimo. Las herramientas utilizadas en la extracción deben ser adaptables, extensibles y capaces de filtrar los datos relevantes a extraer de las fuentes, permitiendo la compresión, descompresión, y encriptación de datos.

Limpieza y Transformación

Etapa central del proceso, donde los datos extraídos son convertidos de su estado original a un formato consistente con el repositorio destino, sin perder su exactitud o veracidad con respecto a las fuentes. Para esta etapa medular, los datos deben ser limpiados, pues en el mundo real son sucios; se muestran incompletos, donde faltan valores de los atributos, o contienen solo agregados de datos; se presentan con ruido, conteniendo errores o valores fuera de límites, manteniendo discrepancias en nombre o códigos. Por ello se realiza un proceso de limpieza que elimina errores e inconsistencias en los datos y resuelve el problema de identidad de los objetos. Teniendo este proceso como tareas fundamentales: llenar valores ausentes, identificar valores fuera de límite, eliminar el ruido en los datos, corregir las inconsistencias de los datos, e integrarlos.

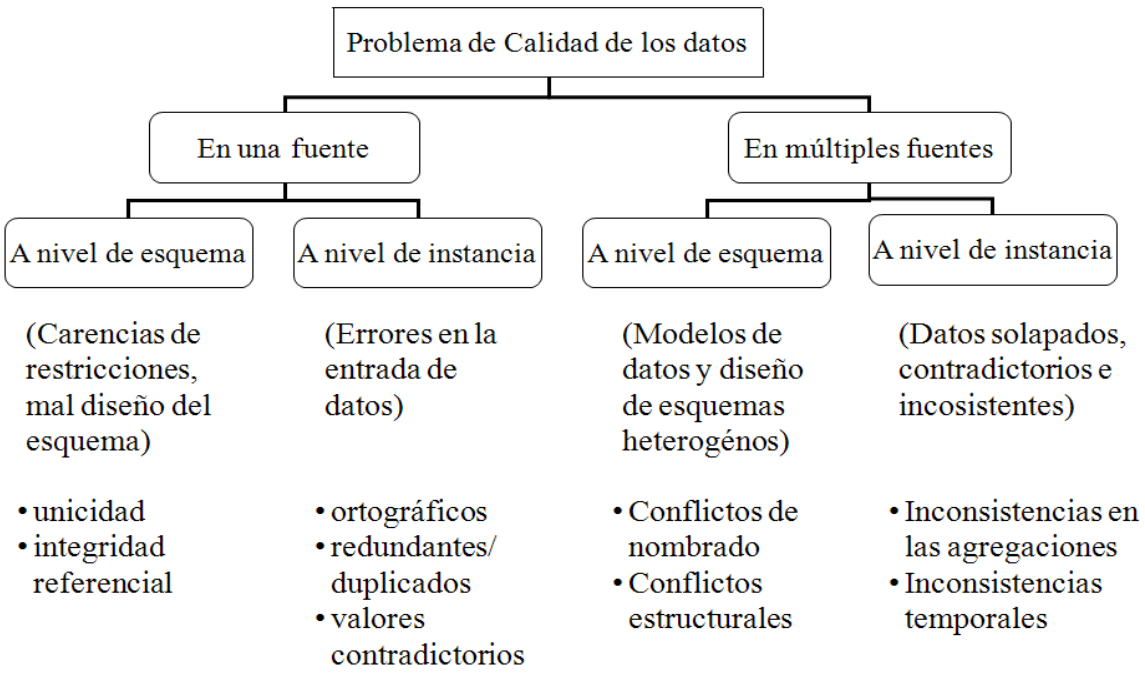


Fig. 2. Problemas a resolver por la limpieza de datos.

Una vez que los datos se encuentren limpios, se procede a realizar las transformaciones, que pueden ir desde simples conversiones de formato, hasta las más complejas operaciones de integración. Aunque lo ideal es seguir los siguientes pasos:

- Análisis de datos.
- Definición del flujo de trabajo de las transformaciones y las reglas de correspondencia.
- Verificación
- Transformación
- Flujo inverso de datos limpios.

Comenzando por el análisis de datos, obteniendo metadatos que ayuden a determinar las características de los datos y patrones de valores no usuales, usando principalmente los enfoques de perfil de datos[9] y minería de datos[10]. Luego que los datos se encuentran limpios, homogeneizados, y se cuenta con metadatos fiable. Se generan las reglas de transformación que va desde realizar tratamientos a valores nulos, aplicar reglas del negocio, combinar los datos de las distintas fuentes o realizar búsquedas de valores en diversas tablas, hasta implementar agregaciones que aceleran los tiempos de análisis, ocultan complejidad de los datos, y proveen múltiples vistas del mismo conjunto de datos. Culminando esta fase con la retroalimentación del flujo inverso de datos limpios.

Carga

Es la fase donde se realiza la transferencia del conjunto resultado de las transformaciones a su destino, ya sean sistemas o ficheros con cierto formato, dependiendo de los requerimientos de la organización. Este subproceso puede abarcar una amplia variedad de acciones diferentes. En algunas bases de datos se sobrescribe la información antigua con nuevos datos, en los almacenes de datos al mantener un historial de los registros se puede hacer una auditoría de los mismos y disponer de un rastro de toda la historia de un valor a lo largo del tiempo. Independientemente de la acción a tomar para la carga, al realizar esta

[9] Del inglés (Data Profiling)), es una técnica que utiliza perfiles de heurística donde se establecen reglas, para medir los datos, y corregirlos.

[10] Del inglés (Data Mining), prepara, sondea y explora los datos para sacar la información oculta en ellos.

operación, se aplicarán todas las restricciones y triggers[11] que se hayan definido, las cuales contribuyen a que se garantice la calidad de los datos en el proceso ETL.

Orígenes de datos. Categorías.

Según la necesidad de información y el problema a responder, son las fuentes de datos que se van a integrar, por ello la importancia de saber identificarlas. Para poder realizar el proceso de integración es necesario clasificar las plataformas técnicas y demás condiciones físicas de las fuentes de información, para en base a la clasificación de cada una, planificar y desarrollar los mecanismos de integración en cada caso. Las categorías propuestas son las siguientes:

- No cooperativas: exportan archivos de intercambio o permiten consultas directas con SQL o a través de servicios web, no se garantiza que el destino de la información la integre.
    - Snapshots: Copia completa de la Información congelada en un instante de tiempo.
    - Fuentes específicas: Fuentes que brindan información a partir de archivos intermedios, u otros mecanismos que no implican funcionalidades internas puntuales en base al receptor de la información.
    - Fuentes consultables: Suministra interfaces para consultas (SQL, servicios web, etc.).
- Cooperativas: que a través de mecanismos de replicación, u otros mecanismos, se establecen intercambios mucho más fiables, seguros y responsables.
    - Fuentes de replicación: Mecanismos de publicación/suscripción.
    - Fuentes callback: Se invocan códigos externos de ETL cuando ocurren cambios en la información.
    - Fuentes de cambios internos: Se activan acciones internas cuando ocurren los cambios (Triggers).

Luego de identificadas las fuentes, se debe tener suficiente experiencia para entender los mecanismos de transferencia de información, analizándose si es necesario cambiar de categoría alguna fuente dada.

[11] Llamados también disparadores, es un procedimiento que se ejecuta cuando se cumple una condición establecida al realizar una operación de inserción (INSERT), actualización (UPDATE) o borrado (DELETE).

Fig. 3. Categorías de fuentes.

Puede incluso pensarse en tener resultados con latencia mínima.

Fig. 4. Integración de datos con latencia mínima

Por ello clasificar las fuentes de datos es un paso importante para definir las características del proceso de integración. Donde debe quedar claro las condiciones que la información de la fuente origen a de cumplir: voluntad política de las autoridades locales y nacionales, de utilizar la información para la toma de decisiones; coordinación con las instituciones - fuentes de información; y el uso de tecnología apropiada para el análisis de información, y evaluación de estrategias. La ausencia de algunas de estas condiciones influye de forma negativa en el éxito de una solución ETL y atenta contra el cumplimiento de los objetivos básicos de la solución.

Roles y Responsabilidades del equipo ETL.

En el nivel más elemental, el equipo ETL es responsable de la extracción de datos de las fuentes origen del sistema, la limpieza y transformación, así como también la carga de los datos al almacén destino. Tener una persona por rol sería idóneo para un cómodo desarrollo, pero teniendo en cuenta la extensión del proyecto, y situaciones que se pueden presentar a lo largo de la vida del proyecto, se tiene que estar preparado para asumir en un momento determinado varios roles por especialista. Más concretamente, y para lograr el resultado óptimo en el proceso ETL las siguientes tareas son responsabilidades del equipo ETL:

- Definir el ámbito de aplicación del proceso ETL.
- Realizar un análisis de datos del sistema fuente.
- Definir una estrategia para lograr calidad en los datos.
- Trabajar con usuarios del negocio a fin de reunir y documentar las reglas de negocio.
- Desarrollar e implementar el código físico ETL.
- Crear y ejecutar subsistemas de control de calidad y planes de prueba.
- Realizar el mantenimiento del sistema.

Pero estas tareas se desglosan, formando roles que asumen responsabilidades fundamentales, que dan al traste con la construcción de un óptimo equipo ETL, facilitando la calidad del proceso en cuestión:

- Gerente ETL: encargado de la gerencia del equipo y del mantenimiento del almacén de datos operacional en todo lo que se refiere al proceso ETL. Responsable de la gestión de los datos a extraer, transformar, y los procesos de carga en el almacén de datos operacional, supervisa los ensayos y su calidad. Y desarrolla normas para el ambiente ETL, incluyendo convenciones de nomenclatura, y buenas prácticas de diseño.
- Arquitecto ETL: las responsabilidades de este rol incluyen el diseño de la arquitectura, la infraestructura y los mapas lógicos de datos[12] para el equipo de desarrollo ETL. Este arquitecto debe tener una fuerte comprensión de los requerimientos del negocio y de los sistemas fuente.

[12] Del inglés Logical Data Map, instrumento de utilidad en la enseñanza de las ciencias y en la investigación didáctica de las ciencias.

- Desarrolladores ETL: responsables de la construcción de los procesos físicos ETL. Este rol trabaja en estrecha colaboración con el arquitecto para resolver cualquier ambigüedad en las especificaciones de la codificación real. El desarrollador es encargado de crear rutinas funcionales ETL y probar su fiabilidad para garantizar que se ajusten con las necesidades del negocio.
- Especialista de calidad de datos: la calidad del almacén de datos operacional incluye la calidad del contenido y de la estructura de información dentro del almacén. El especialista en calidad de datos trabaja principalmente con el arquitecto ETL para garantizar que las reglas comerciales y las definiciones de datos sean propagadas a lo largo del proceso ETL.
- Administrador de bases de datos: principal responsable de traducir el diseño lógico de la base de datos a una estructura física. Por otra parte trabaja muy cerca del equipo ETL para garantizar que los nuevos procesos no corrompan los datos existentes. En algunos ambientes, el administrador de base de datos es propietario de los procesos ETL una vez que se haya migrado a la producción.
- Administrador de dimensión: encargado de la definición, construcción, y publicación de una o más dimensiones conformadas para la comunidad de datos extendidos. Asegurándose que las dimensiones configuradas se reproducen de manera simultánea a todas las tablas de hechos para todos los clientes proveedores.
- Proveedor de la tabla de hechos: posee tablas de hechos en un entorno de dimensiones configuradas. Recibiendo actualizaciones periódicas de las dimensiones enviadas por el administrador de dimensiones. Convirtiendo la llave natural en la llave sustituta para exponer las tablas de hechos de forma adecuada para el usuario.

Proceso ETL. Importancia y Desafíos.

Un buen proceso ETL permite una integración consistente de los datos. Al culminar este proceso, los datos serán; precisos, completos, creíbles, rigurosos en el tiempo, interpretables, accesibles y con valor añadido. Este proceso no es una simple migración, pues asegura la calidad de los datos, los prepara, conforma, limpia, trasforma, y organiza, brindando a quienes utilicen estos datos un conocimiento que facilitará la toma de decisiones. El proceso ETL suele ser muy extenso y precisa de un alto nivel de escaneo, por ello presenta a lo largo de su desarrollo ciertos inconvenientes que deben ser depurados. Entre los principales desafíos encontramos; que el rango de valores de los datos o la calidad de éstos

pueden no coincidir con las expectativas de los diseñadores a la hora de especificarse las reglas de validación o transformación. Por lo que es recomendable realizar un examen completo de la validez de los datos del sistema de origen durante el análisis, para identificar las condiciones necesarias para que los datos puedan ser tratados adecuadamente por las reglas de transformación especificadas. Otro aspecto a tener en cuenta es la escalabilidad, esto incluye la comprensión de los volúmenes de datos que tendrán que ser procesados según los acuerdos de nivel de servicio[13], el tiempo disponible para realizar la extracción de los sistemas de origen podría cambiar, lo que implicaría que la misma cantidad de datos tendría que ser procesada en menos tiempo. Estos son algunos retos que presenta el desarrollo de este proceso, aunque no son los únicos, también podemos encontrar problemas con los volúmenes de datos que crecen de forma exponencial, entonces se tendría que procesar grandes cantidades de datos granulares. Así los sistemas de información crecen en complejidad, aumentando la disparidad de las fuentes, para ello el proceso ETL necesitaría una extensa conectividad a las aplicaciones en paquetes, bases de datos, mainframes[14], archivos, servicios web, etc. Otra característica que dificulta este proceso, son las transformaciones que llegan a ser muy engorrosas, pues los datos necesitan agregarse, analizarse, computarse, y procesarse estadísticamente. Estos son algunos retos a los que normalmente se enfrenta un proceso ETL y su equipo, debido a ello en la actualidad las soluciones de integración de datos están cada vez más optimizadas para lograr una adecuada calidad empresarial, siguiendo de forma especial las siguientes características que son críticas para el diseño, desarrollo, ejecución y mantenimiento de los procesos ETL:

- Modelación de procesos orientada al negocio que implica a las partes interesadas en el negocio y garantiza una comunicación óptima en las líneas de negocio.
- Entorno de desarrollo gráfico que mejora en gran medida la productividad y facilita el mantenimiento.
- Amplia conectividad para admitir todos los sistemas.

[13] Del inglés (Service Level Agreement), conocido también por las siglas SLA, es un contrato escrito entre un proveedor de servicio y su cliente con objeto de fijar el nivel acordado para la calidad del servicio.

[14] Computadora grande, potente y costosa usada principalmente por una gran compañía para el procesamiento de una gran cantidad de datos.

- Componentes avanzados ETL, incluidas manipulaciones de cadenas, dimensiones lentamente cambiantes[15], soporte para cargas masivas, etc.

Que si bien estas consideraciones no eliminan los muchos desafíos que afronta este proceso en su desarrollo, sí los minimiza y controla.

¿Qué parámetros tener en cuenta para la selección de la herramienta ETL a utilizar?

La decisión de qué herramienta ETL utilizar para el desarrollo del proceso, es una de las de mayor peso. Pues la selección debe estar respaldada por un serio estudio, ya que migrar de una herramienta a otra es una ardua tarea, debido a que cada una tiene sus particularidades a la hora de reflejar los pasos del proceso, esta selección se apoya en una serie de parámetros que pueden facilitar el establecimiento de una comparación entre las distintas variantes:

- Multiplataforma: la herramienta debe ser capaz de funcionar en cualquier plataforma, aunque, de acuerdo a las exigencias del entorno donde se desarrolle el proceso ETL, basta con que la herramienta sea compatible con la plataforma previamente seleccionada.
- Requerimientos de hardware y software: analizar si se pueden cumplir los requerimientos de la herramienta en cuanto a hardware y software.
- Independencia del tipo de fuente o destino: debe ser capaz de leer y escribir directamente desde y hacia las fuentes y los destinos de los datos, independientemente de su tipo, dando conectividad con distintos gestores de bases de datos.
- Soporte para metadatos: una de las mayores ventajas que debe poseer una herramienta es tener disponible la información sobre los datos durante el desarrollo y ejecución de los procesos. La generación de metadatos, que incluyen la definición de los tipos de datos de las fuentes, de las transformaciones y destinos, debe ser un proceso automático.
- Soporte funcional: debe ser posible la realización eficiente de operaciones para la limpieza de los datos, transformaciones, agregaciones, reorganización y carga.

[15] Del inglés (Slowly Changing Dimensions), este concepto se aplica a como las dimensiones deben tener en cuenta los cambios históricos.

- Utilidades del sistema operativo y servicios de transporte: debe brindar la posibilidad de interactuar con el sistema operativo, para poder ejecutar distintos procesos y tareas en el sistema de ficheros. Ofreciendo servicios de transporte, por ejemplo, a través de la red o usando el Protocolo para la Transferencia de Ficheros[16].
- Soporte al modelo dimensional: la herramienta debe tener incorporado soporte para la creación de tareas de dimensiones lentamente cambiantes, generación de llaves sustitutas y construcción de dimensiones agregadas.
- Facilidad de uso: debe ser una herramienta de propósito general y amigable para el desarrollador, de modo que pueda identificarse rápidamente con la misma.
- Paralelismo: debe ser posible la ejecución de operaciones en paralelo de manera que una tarea pueda aprovechar el paralelismo inherente de la plataforma sobre la que corre.
- Corrección de errores y registro de eventos: debe ser posible rastrear los errores en las transformaciones en tiempo de ejecución, así como ver los datos antes y después de cada transformación. Además, se necesita controle el proceso registrando los distintos eventos durante la ejecución.
- Planificación de la ejecución: ofrece una manera de planificar la ejecución de los trabajos de manera automática, o sea sin necesidad de intervención humana para completar su funcionamiento.
- Puesta en funcionamiento: debe ser posible agrupar varios objetos ETL y ponerlos en funcionamiento en ambientes de testeo o producción.
- Reusabilidad: aporta la posibilidad de aprovechar parte de la lógica de las distintas tareas, de manera que el desarrollador no tenga que hacer repetidas veces una misma transformación.
- Documentación, preparación y soporte técnico: se requiere incorpore documentación básica. Se deberá tener en cuenta si ofrecen servicios de preparación para los especialistas, así como soporte técnico.
- Extensibilidad: debe permitir al usuario la definición de nuevas funciones y utilizarlas igual que las que incluye la herramienta.

[16] FTP (siglas en inglés de File Transfer Protocol) es un protocolo de red para la transferencia de archivos entre sistemas conectados a una red TCP, basado en la arquitectura cliente-servidor.

- Gestión de las dimensiones lentamente cambiantes: la herramienta debe ser capaz de manipular las dimensiones lentamente cambiantes.
- Gestión de la calidad de datos: se requiere permita al usuario realizar acciones como la limpieza, el filtrado y la validación de los datos para garantizar su calidad.
- Gestión de la sustitución de claves: se refiere al aporte de la herramienta para facilitar la sustitución de las llaves del negocio por las llaves de la dimensión.
- Perfil de datos: se refiere a si la herramienta permite realizar perfilado de datos a las fuentes.

Muchas veces resulta difícil el proceso de selección, aún más cuando la mayoría de los vendedores tratan de desarrollar un producto que cumpla todos o al menos gran parte de los parámetros descritos. Actualmente las herramientas se centran en brindar la mayor conectividad posible y ofrecer funciones de integración casi en tiempo real. En un futuro, sin duda alguna, se sumarán nuevos parámetros. No obstante, el orden o prioridad entre cada uno de ellos debe ser determinado de acuerdo a los requisitos o las más urgentes necesidades que tengan los clientes.

Herramientas ETL a tener en cuenta para selección. Principales características.

Varios productos pueden ser los candidatos a tener en cuenta en el momento de realizar una selección, pues actualmente son diversas y numerosas las herramientas de las cuales resultaría provechoso un estudio. Entre las herramientas más conocidas se encuentran; PowerCenter, una plataforma de integración de datos lanzada por Informatica Corporation, que incluye la construcción de almacenes de datos de terabytes, la migración desde los sistemas heredados, la sincronización de los almacenes de datos operacionales, la consolidación de aplicaciones o la obtención de una visión integral de todos los datos relevantes, presentando avances en tres áreas fundamentales: el acceso a los datos empresariales, el despliegue en toda la empresa y en la productividad del desarrollador. También podemos hacer referencia a otras herramientas como; Information Server de IBM[17], revolucionaria plataforma que facilita obtener valor de la heterogénea y compleja información difundida a través de los sistemas de su empresa anfitriona; otro producto es SQL Server 2008 Integration Services, bajo la tutela de Microsoft, que muestra

[17] International Business Machines o IBM (conocida coloquialmente como el Gigante Azul) es una empresa que fabrica y comercializa herramientas, programas y servicios relacionados con la informática.

altas capacidades de integración en forma fácil y asequible, permitiendo gestionar transformaciones de alto rendimiento en una amplia gama de fuentes de datos. Entre las herramientas más novedosas a tener en cuenta se encuentran; Oracle Data Integrator y Oracle Warehouse Builder, productos dentro de la familia de Oracle Fusion Middleware, quienes facilitan la optimización de inteligencia de negocios, almacén de datos y gestión de datos. Pero en este trabajo solo tendremos en cuenta la comparación de herramientas en el ambiente de código abierto[18], pues aporta la posibilidad de manejar directamente el código fuente y hasta modificarlo o adaptarlo a necesidades individuales, aunque no se tenga tanta seguridad y más que seguridad, un prestigio como el de los productos de los grandes monopolios. Se ha querido analizar dos herramientas código abierto. Las cuáles son Spoon de Pentaho Data Integration, y Talend Open Studio de Talend.

Spoon de Pentaho Data Integration

Spoon de Pentaho Data Integration es una de las herramientas ETL de código abierto más populares, la cual permite diseñar de forma gráfica la transformación ETL. Reúne un conjunto de componentes que permiten modelar y ejecutar transformaciones sobre flujos de datos. Puede funcionar sobre varias plataformas:

- Microsoft Windows (todas las plataformas, desde Windows 95 hasta Vista)
- OSX de Apple (para máquinas PowerPC e Intel)
- Linux (Gtk-PPC/x86/x86_64)
- AIX (usando una interfaz Motif)
- HPUX (usando una interfaz Motif (GTK opcional))
- Solaris (usando una interfaz Motif (GTK opcional))

Requiere un sistema que soporte Java 1.4 Runtime Environment o una versión más nueva, y de hardware exige alrededor de 128 MB de RAM. Actualmente soporta diversos tipos de fuentes. Tiene conectividad con cualquier base de datos usando ODBC[19] en Windows, Oracle, MySQL, AS/400, MS Access, MS SQL

[18] En inglés open source es el término con el que se conoce al software distribuido y desarrollado libremente.

[19] Open Database Connectivity (ODBC), es un estándar de acceso a bases de datos desarrollado por Microsoft Corporation, el objetivo de *ODBC* es hacer posible el acceder a cualquier dato desde cualquier aplicación, sin importar qué Sistema Gestor de Bases de Datos (DBMS por sus siglas en inglés) se utilice.

Server, IBM DB2, PostgresSQL, Intersystems Caché, Informix, Sybase, dBase, Firebird SQL, MaxDB (SAP DB), Hypersonic, CA Ingress, SAP R/3 System (usando el plugin ProSAPCONN), Teradata. Se integra con ficheros de Microsoft Office, web services y cubos MOLAP[20]. Incluye procesamiento optimizado de los ficheros planos y provee un JDBC[21], que permite la conexión con cualquier base de datos sin tener que instalar un cliente adicional. Esta herramienta brinda soporte para metadatos. Incorpora operaciones de transformación, así como funciones que permiten operar con los campos en el flujo de datos, renombrando, calculando campos en función de otros, correlacionando valores y realizando búsquedas auxiliares en bases de datos. Se debe destacar, que su rendimiento se puede ver afectado cuando se realizan operaciones de join[22] con grandes volúmenes de datos, pues maneja pequeñas cantidades de información en el flujo. Aunque por otra parte ofrece, soporte para operaciones de dimensiones lentamente cambiantes, permite ejecutar código javascript dentro de las transformaciones e incorpora un evaluador de expresiones regulares. Entre las tareas que se pueden incorporar están la copia y eliminación de ficheros, así como su transportación usando FTP, además se pueden descompactar ficheros. Esta herramienta es fácil de usar: es model-driven[23], por lo que no es necesario prácticamente incorporar código, aunque el usuario tiene la posibilidad de programar sus propias transformaciones y desarrollar un proceso que consiste en declarar el flujo en una interfaz gráfica. Brindando la posibilidad de copiar y leer del mismo fichero en paralelo, permitiendo maximizar la capacidad de entrada/salida en el entorno ETL. Esta herramienta también soporta la ejecución de transformaciones en clusters[24], aunque el procedimiento de particionar los datos no es automático, por lo que no se logra un procesamiento en paralelo completo. Añade un debugger integrado, diseñado para

[20] Multidimensional Online Analytical Processing (MOLAP), también se conoce como procesamiento analítico multidimensional en línea. Permite el almacenamiento de datos en una matriz de almacenamiento multidimensional optimizada.

[21] Java Database Connectivity (JDBC), permite la ejecución de operaciones sobre bases de datos desde el lenguaje de programación Java.

[22] Sentencia en SQL, que permite combinar registros de dos o más tablas en una base de datos relacional.

[23] Model-Driven Architecture (MDA) o Arquitectura dirigida por modelos, proporciona un conjunto de guías para estructurar especificaciones expresadas como modelos.

[24] Se aplica a los conjuntos o conglomerados de computadoras construidos mediante la utilización de componentes de hardware comunes y que se comportan como si fuesen una única computadora.

mejorar la productividad del desarrollador, ya que se pueden agregar puntos de ruptura condicionales en la ejecución de las transformaciones, dando la posibilidad de pausar y resumir la ejecución de la transformación, así como especificar el número de filas que se van a usar en las ejecuciones de prueba. Además, se pueden añadir registros personalizados.

Otras aplicaciones de la suite Pentaho Data Integration, son:

- PAN ejecuta las transformaciones diseñadas con SPOON.
- CHEF permite mediante una interfaz gráfica diseñar la carga de datos incluyendo un control de estado de los trabajos.
- KITCHEN permite ejecutar los trabajos diseñados con CHEF.

Mientras el Spoon de Pentaho Data Integration tiene entre sus ventajas, que es una de las más antiguas herramientas ETL de código abierto, cuenta con una gran comunidad de usuarios, su interfaz gráfica permite un aumento de la productividad. Aunque a la hora de la selección, entre sus desventajas podemos encontrar; que no automatiza el proceso de separación y redistribución de datos para el procesamiento paralelo y que no cuenta con un componente de calidad de datos especializada o una asociación con un proveedor de calidad de los datos, además que para realizar búsquedas de mayores volúmenes necesita utilizar una base de datos de búsqueda donde se ejecutan un gran número de sentencias SQL que frenan el rendimiento de ETL. Para un mejor enfoque del funcionamiento de esta herramienta, presentamos la arquitectura que utiliza.(Pentaho 2009)

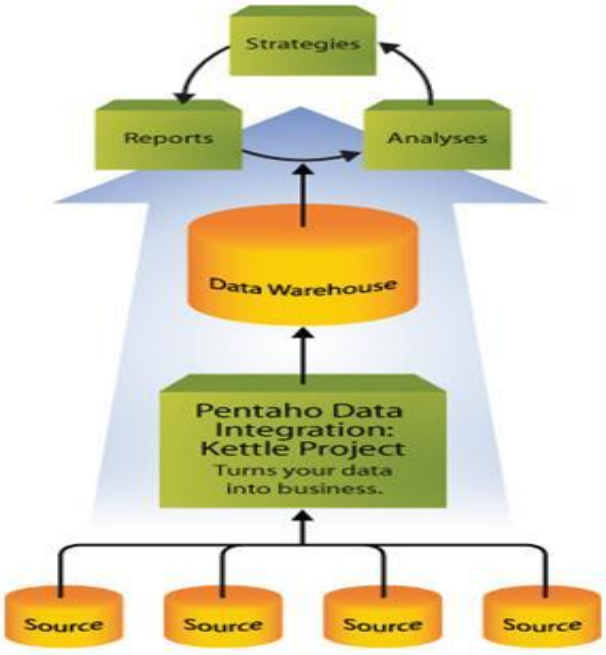


Fig. 5. Arquitectura de Pentaho Data Integration

Talend Open Studio

Talend es un proveedor de productos para la integración de datos, compañía de propiedad privada y respaldada por capital arriesgado. El producto insignia de la compañía es Talend Open Studio. El cuál está disponible bajo la licencia sin costo GPL[25], según lo dispuesto por esta licencia, Talend Open Studio no puede ser incorporado como parte de otro software sin su permiso. Esta herramienta es un producto generador de código, que permite ser analizado cada vez que se requiera, esto significa que evita la concepción de ser una caja negra seguida por otros desarrolladores. Soporta las siguientes plataformas:

- Solaris
- MAC
- Windows
- Red Hat Enterprise Linux

[25] General Public License (GNU por sus siglas en inglés), orientada principalmente a proteger la libre distribución, modificación y uso de software.

- Linux (permite arquitecturas de 32 y de 64 bits)

Los scripts generados por la herramienta se pueden ejecutar en cualquier sistema operativo que soporte, en cuanto a requerimientos de software, es necesario instalar Java o Perl en Windows o en Linux. Para realizar las transformaciones ETL, Talend tiene conectividad con: AS400, Access, DB Generic, DB2, Firebird, HSQLDb, Infomix, Ingres, Interbase, JavaDB, JDBC, LDAP, Microsoft SQLServer, MySQL, Oracle, PostgresSQL, SQLite, Sybase, Teradata y Vertica. Pero se debe señalar que el acceso a las fuentes requiere un driver JDBC. Soporta distintos formatos de ficheros y metadatos. Estos son almacenados y administrados en un repositorio, compartido por todos los módulos. Este repositorio centraliza la información de los proyectos y asegura la consistencia de todos los procesos de integración, facilitando la reutilización de los objetos y del código creado. Contiene una librería de más de doscientos componentes y conectores, la cual proporciona funciones básicas como: operaciones de correlación, búsquedas, filtrado de datos, transformaciones, y algunas facilidades para cargar hacia almacenes de datos. La librería de componentes puede ser extendida usando lenguajes como Java, Perl o SQL, pues los desarrolladores pueden introducir nuevas funcionalidades. La herramienta Talend incorpora operaciones para el manejo de ficheros, así como servicio FTP, ofrece además una vista gráfica intuitiva. Los procesos de integración son construidos arrastrando componentes y conectores al diagrama, dibujando conexiones y relaciones entre ellos y especificando sus propiedades. Talend Open Studio incluye un debugger potente y características sincronizadas que permiten seguir el flujo de datos en tiempo real a través de todo el proceso de transformación. Cuando un trabajo de integración es ejecutado a través de la interfaz de diseño las estadísticas son presentadas, mostrando el número de filas procesadas y rechazadas, así como el rendimiento, permitiendo localizar cualquier embotellamiento en el proceso de forma inmediata. Es también posible activar el modo rastreo, el cual muestra el comportamiento fila por fila, así como el resultado de las transformaciones. Los puntos de ruptura y las variables también están disponibles. Es posible la planificación de la ejecución de tareas realizadas mediante una interfaz centralizada que ofrece Talend, así como la ejecución remota de los trabajos para probarlos sin tener que utilizar procedimientos complejos para su puesta en funcionamiento. Entre sus principales ventajas, se encuentra el foro de la compañía que proporciona soporte de manera gratuita, además la herramienta permite equilibrar la carga entre el servidor de procesamiento de Talend, grupo o red, y el origen o destino de las bases de datos en escena. Otra ventaja es que cuenta con una interfaz ETL para la importación de los metadatos, la configuración, vinculación de los componentes y generación

de código, lo cual proporciona ganancias de productividad para los desarrolladores. Las desventajas que ofrece esta herramienta no son muchas, la más significativa es el no aseguramiento de la calidad de los datos , aunque esta función pueden ser cumplida por una tercera parte.(Talend 2009)

Valoraciones finales de las herramientas ETL. Comparación.

Como se ha podido dilucidar, las herramientas evaluadas intentan satisfacer o cumplir los parámetros analizados. Ambas ofrecen amplia conectividad con varias fuentes, brindan un abundante conjunto de transformaciones y soporte para metadatos. Permiten, que el usuario pueda definir sus propias funciones y brindan medios con los que se pueda corregir los errores y registrar eventos durante el flujo de datos. Para un mejor análisis se comparan las herramientas, para así obtener una mejor vista a la hora de elegir y respaldar el porqué de la herramienta seleccionada.(License 2008)

	Pentaho Data Integration	Talend Open Studio
Facilidad de uso	x	
Extensible	x	x
Reusabilidad	x	x
Ficheros planos	x	x
Ficheros XML	x	x
Ficheros EXCEL	x	x
Bases de datos Microsoft Access	x	
Bases de datos DB2	x	x
Bases de datos Oracle	x	x
Bases de datos SQL Server	x	x
Sistemas ERP (SAP)	x	x
Ficheros planos	x	x
Ficheros XML	x	x
Bases de datos Oracle	x	x
Bases de datos SQL Server	x	x
Motor OLAP		
Gestión de dimensiones	x	x

lentamente cambiantes		
Gestión de la calidad de datos	x	x
Gestión de la sustitución de claves	x	
Perfil de datos		
Manipulación de errores	x	x
Registro de eventos	x	x
Planificación de tareas	x	x
Ejecución automáticas de tareas	x	x
Gestión de metadatos	x	x
Soporte para ejecución paralela de tareas	x	x

Fig. 6. Comparación entre Spoon y Talend

En la figura 6 se observa que existen ciertas diferencias, por ejemplo, en las plataformas sobre las que corren, y los requerimientos en software y hardware, siendo la suite Pentaho Data Integration la herramienta más exigente. También en el soporte que brindan al modelo dimensional, siendo Talend la más pobre en este sentido. Aunque estas herramientas tienen requerimientos que cumplen de forma bastante igualitaria, ambas no realizan perfilado de datos, aunque el Spoon de Pentaho Data Integration tiene como ventaja que es más fácil su uso, y que además permite la gestión de llaves sustitutas, mientras Talend prescinde de este servicio. Debido a este análisis y teniendo como primer parámetro y decisivo el dictamen del cliente, se determina realizar el desarrollo del proceso ETL en la herramienta Spoon de Pentaho Data Integration.

Herramienta de modelado a utilizar. Principales características.

Existen diversas herramientas que potencian el modelado, entre las más conocidas se puede mencionar Rational Rose, la cual potencia prácticas modernas de ingeniería de software, proponiendo la utilización de diferentes tipos de modelo para realizar un diseño del sistema, utilizando una vista estática y otra dinámica de los modelos, uno lógico y otro físico, además permite crear y refinar estas vistas creando un modelo completo que representa el dominio del problema y el sistema de software. Otra famosa herramienta es el Power Designer, la cual permite de manera fácil, visualizar, analizar y manipular

metadatos, brindando un enfoque basado en modelos, alineando al negocio con la tecnología de información, facilitando la implementación de arquitecturas efectivas de información empresarial. Otra herramienta profesional y efectiva es Enterprise Architect, la cual aporta un alto rendimiento, es flexible, completa y proporciona un potente modelado en UML, provee lo más nuevo en desarrollo de sistemas, administración de proyectos y análisis de negocio. A partir del estudio de las principales herramientas de modelado, y teniendo en cuenta que la universidad utiliza de forma periódica la herramienta elegida, el equipo de desarrollo decidió modelar en Visual Paradigm para UML[26], herramienta visual que ayuda a construir aplicaciones rápidamente, y soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. Esta herramienta permite dibujar todos los tipos de diagramas de clases, código inverso, generar código desde los diagramas y generar documentación, proporcionando abundantes tutoriales de UML, demostraciones interactivas de UML y proyectos UML. Además es colaborativa, o sea soporta múltiples usuarios trabajando sobre el mismo proyecto; genera la documentación del proyecto automáticamente en varios formatos y permite control de versiones. Cabe destacar su robustez, usabilidad y portabilidad. Entre las principales características de la herramienta, encontramos: que permite la creación de diagramas de procesos de negocio, cubriendo además todo el ciclo de vida, facilitando un modelado colaborativo con Subversion[27] e interoperabilidad con modelos UML. Además el Visual Paradigm, proporciona el diseño de ingeniería inversa, código a modelo, código a diagrama, permitiendo la realización de diagramas de flujo de datos, generación de bases de datos y la ingeniería inversa de bases de datos, siendo un potente generador de informes.(Headquarters 2009)

Herramienta de perfilado de datos a utilizar. Principales características.

Actualmente el perfilado de datos, no cuenta con muchas herramientas que apoyen esta rama, la más conocida es la herramienta de perfilado de GCC gprof; para su uso, es necesario compilar los programas a perfilar, al ejecutarlos se crea un archivo llamado gmon.out que puede convertirse a un formato legible mediante gprof. Una desventaja es que se necesita recompilar el programa y enlazarlo estáticamente. En

[26] Lenguaje Unificado de Modelado (UML, por sus siglas en inglés, Unified Modeling Language) es el lenguaje de modelado de sistemas de software más conocido y utilizado en la actualidad.

[27] Software de sistema de control de versiones.

esta herramienta la usabilidad se dificulta, por lo que se eligió para el paso de análisis de datos en el proceso ETL, la novedosa herramienta Talend Open Profiler, la cual cuenta con una interfaz amigable que permite explorar las bases de datos y analizar los datos, definiendo parámetros o indicadores y recopilando estadísticas sobre sus datos en unos pocos clics. Viene con un conjunto de expresiones regulares que ayuda a identificar los datos erróneos. Además es posible crear expresiones regulares propias y usarlas en el análisis de los datos. Esta herramienta presenta diversas opciones para cada uno de los indicadores, que cambian el comportamiento del indicador a fin de obtener más información pertinente, y opciones de calidad de los datos. Otra funcionalidad que ofrece, son las estadísticas de resumen, que ayuda a descubrir el porcentaje de los datos de mala calidad, además de que los análisis que se crean se guardan automáticamente para que pueda ser ejecutado varias veces y ver cómo evoluciona la calidad. De forma general esta herramienta es fácil de usar, extensible y reusable, permite un serio perfilado de datos y manipulación de errores, así como la gestión de metadatos. Encontrándose en un ambiente código abierto, lo cual es ideal para el desarrollo del proceso ETL en cuestión.

Conclusiones del Capítulo 1.

En este primer capítulo se ha abordado de manera descriptiva y enfocada las principales características, definiciones y aspectos relacionados con el proceso de extracción, transformación y carga, teniendo en cuenta las herramientas apropiadas para la implementación de este proceso y modelado del mismo. Subrayando los roles y responsabilidades del equipo ETL, así como en los principales desafíos que los desarrolladores pueden encontrar a su paso, tomando especial hincapié en la importancia del recorrido arduo y detallado de la transformación, limpieza y carga de datos.

CAPÍTULO 2: IMPLEMENTACIÓN DEL PROCESO ETL PARA CIMEX

La corporación CIMEX es una entidad cubana encargada de comprar disímiles productos a suministradores nacionales y extranjeros, con el fin de comercializarlos a través de una red de tiendas distribuidas por todo el país. El manejo de los datos correspondientes al comercio mayorista que ocurre diariamente, se realiza haciendo uso de Sentai, un sistema automatizado orientado a la gestión empresarial. A pesar del buen funcionamiento desde el punto de vista transaccional, no es posible realizar sobre este sistema procesamiento analítico de gran envergadura, pues múltiples usuarios alteran el estado de Sentai cada minuto, y muchos otros sistemas se nutren de sus datos, por lo cual en ocasiones se sobrecarga de tareas. Las consultas muy complejas se ejecutan lentamente debido a la forma en que están estructurados sus datos, por lo que no satisfacen las necesidades de eficiencia que solicitan los analistas. A esto se le suma el hecho de que la información está localizada en varias bases de datos Sentai, incluso con distintas versiones del sistema, por lo que un análisis global del negocio es prácticamente imposible en este entorno. Debido a estos inconvenientes y las necesidades de información que existen, la corporación CIMEX decidió realizar un almacén de datos operacional[28], que permite realizar procesos de análisis a niveles gerenciales, proporcionando una visión general del comportamiento comercial mayorista. Para ello es necesario implementar el proceso ETL, que permite la integración y carga de los datos que tienen como origen las bases de datos utilizadas por el software Sentai, donde se limpiarán los datos de acuerdo a las necesidades y estructuras definidas para el almacén de datos operacional de CIMEX. Siendo el objetivo de este capítulo detallar, desarrollar, implementar, y documentar este proceso de extracción, transformación y carga de datos para CIMEX.

Principales procesos del negocio que existen en CIMEX.

En la actividad comercial mayorista se pueden identificar distintos procesos, entre ellos las compras, tanto las efectuadas como las pendientes, las ventas, y el inventario son las más significativas. Estos procesos involucran objetos del negocio como son los productos, los clientes, los proveedores, las localidades, por citar algunos. CIMEX cuenta con un flujo amplio de negocio, pero el que interesa en este trabajo es el proceso mayorista, el cual se encuentra estructurado por módulos, cada uno encargado de manejar los

[28] También llamado ODS (del inglés Operational Data Store) es un contenedor de datos activos, es decir operacionales que ayudan al soporte de decisiones y a la operación.

distintos tipos de transacciones que tienen lugar. Una acción en el sistema puede generar transacciones en uno o varios módulos. Desde el punto de vista comercial, toda negociación implica adquisición de mercancías o servicios que generan inventarios en almacenes. Su posterior venta, distribución a otras sucursales o transferencia a otros almacenes; se materializa mediante un contrato. Cuando el suministro de los proveedores es estable con la entidad y cuando las compras son eventuales o puntuales con un suministrador, se requiere de una autorización de compra u orden de compra específica para materializar la negociación. Aunque las entregas de mercancías o productos previstas en los contratos comerciales pueden ejecutarse totalmente de una vez o en forma parcial, suele confeccionarse una autorización de compra por cada entrega de suministro. Luego, la orden de compra, es por excelencia el documento primario que se utiliza para formalizar las negociaciones por concepto de adquisiciones de mercancías entre empresas. Por tanto, dicho documento, es una solicitud de las cantidades ordenadas a adquirir, expresadas en las unidades de medida acordadas, los precios unitarios pactados y el importe o monto de las operaciones que se generan por cada producto a comprar. La orden de compra, por su naturaleza, no ejecuta contabilización alguna en los registros contables del sistema, hasta tanto sea recibida en una localidad o área de inventario prevista, por las cantidades e importes realmente recibidos. Por otra parte las ventas están dirigidas a satisfacer todas las necesidades de la gestión de ventas de una organización. Por medio del mismo se permite controlar las cotizaciones a los clientes, generar órdenes de ventas, documentos asociados a la disposición de los inventarios, facturación, ingreso de pagos, devoluciones y las transferencias existencias entre localidades de una entidad.

En el proceso de negocio del CIMEX, los módulos de órdenes de compras, órdenes de ventas, inventarios, administración de la distribución y administración de almacenes, están orientados y relacionados directamente a las actividades de la gestión comercial de compras y ventas de mercancías, tanto mayoristas como minoristas, así como al almacenamiento, manipulación y entrega de las mismas en las empresas que las desarrollan. Mientras que los servicios y el inventario asociado, están orientados y relacionados directamente con la gestión comercial de prestación de servicios de reparación, mantenimiento, garantía y otras actividades afines en talleres dedicados a reparar equipos técnicos de cualquier naturaleza, talleres dedicados a reparar transportes automotores, y otros talleres similares, que laboren por órdenes de trabajo, en aquellas empresas que desarrollan estas actividades. Las órdenes de producción y el de inventario asociado, están orientados y relacionados directamente con la gestión de centros de elaboración de alimentos con fines gastronómicos, gestión de producciones materiales de

proceso continuo, así como la comercialización posterior de las mismas, en empresas, compañías y entidades dedicadas a actividades de naturaleza productiva. Teniendo como trasfondo las cuentas por cobrar, cuentas por pagar, activos fijos, el inventario propiamente, la contabilidad general y la conciliación bancaria, estos asociados por su naturaleza a la actividad económica de las empresas que explotan el sistema, y mediante los cuales pueden controlar su registro contable y obtener emisión de estados financieros, incluidos los reportes estadísticos correspondientes que se nutren de la contabilidad, donde pueden regular la actividad de planificación y presupuestos de gastos corrientes e inversiones, así como implementar su actividad financiera y bancaria, entre otros aspectos. Siendo todos estos módulos parte de la gestión comercial, de prestación de servicios o producción, en la empresa CIMEX, donde se enlazan con los módulos que reflejan la actividad económica, en aras de la integralidad de la administración empresarial en su conjunto.(CIMEX 2009)

Aspectos a tener en cuenta en el proceso ETL de CIMEX.

Llaves sustitutas.

Cuando determinados datos comparten valores pero no son interpretados como llaves, no hay problema alguno, pero cuando son identificadores, se convierte en un caso que merece especial atención. Si datos que sirven como llave comparten un mismo universo de valores, a la hora de integrar van a surgir ciertas contradicciones, pues, ¿cómo diferenciar a dos elementos diferentes con identificadores iguales? Las llaves utilizadas por los sistemas operacionales no van a ser suficientes a la hora de identificar unívocamente la información en el almacén de datos. Por ello podría pensarse como solución identificar los datos a partir de dos o más campos formando una llave compuesta, por ejemplo, el valor de la llave "natural" y la base de datos o sistema de donde provienen. Pero recordando algunas de las características del modelo dimensional, las llaves en las dimensiones no deben ser compuestas, pues esto provocaría que el tamaño de las tablas de hechos en las que participa crezca, y que las operaciones de join se tengan que hacer por más de un campo. Son deseables las llaves simples. Entonces, ¿por qué no concatenar esos valores y formar una llave simple? Porque las llaves en el modelo dimensional no deben tener significado: no debe ser posible decir algo de los datos sólo con mirar la llave. Además, la ocurrencia de modificaciones, reestructuraciones o purgas de dichos identificadores en el nivel transaccional, implicaría procesos de actualización de gran envergadura. En este caso, la solución más recomendable es

la creación de llaves sustitutas o surrogate keys. Una llave sustituta es una generalización necesaria para las llaves naturales de los sistemas operacionales. Son números enteros sin ningún significado; auto-incrementales que van a servir como identificador de cada una de las dimensiones. Son, por tanto, las que van a estar presentes en la llave compuesta de las tablas de hechos y por ellas se realizarán las operaciones de join con las dimensiones relacionadas. La utilización de llaves sustitutas implica que, cada vez que se inserten datos nuevos en las dimensiones, deberán ser generados valores que no hayan sido asignados aún, para que sirvan como identificadores de las nuevas filas. Si son modificados algunos de los campos de un elemento ya existente en una dimensión, en el proceso de carga se debe buscar la fila en cuestión, actualizarla con los cambios, pero mantener el mismo valor para la llave sustituta, pues no es objetivo del sistema mantener la historia de los datos no operacionales. A la hora de cargar hacia las tablas de hechos, primero se deberán determinar los valores sustitutos para las llaves naturales que participen, pues las llaves foráneas hacen referencia en realidad a dichos valores sustitutos.(Celko 1996)

Como se ha podido ver, la tarea de integrar no implica solamente unir los datos en un mismo lugar, sino procesarlos para brindar una imagen homogénea. Este procesamiento puede ser más complejo mientras más profundo sea el estudio de los entornos operacionales para determinar cuáles de sus datos son “iguales” o “diferentes”. Por lo tanto en la implementación de este proceso ETL, se propone:

- Identificar y dar el tratamiento adecuado a los datos que se codifican de forma diferente, que están expresados en distinta medida o son identificadores que comparten valores.
- Añadir llaves sustitutas para cada dimensión. De esta manera se cumple con lo establecido en el modelo dimensional de utilizar llaves simples sin significado.

Llaves nulas y huérfanas.

Un almacén de datos operacional, es un sistema destinado al análisis, por lo que debe ser fiable, es decir, el procesamiento que se realice sobre los datos no los debe modificar, debe ser inocuo. Las transformaciones deben estar orientadas solamente a convertir los datos en información legible para el usuario. Durante el proceso de carga, por tanto, se deberán incorporar procesos de limpieza simples. Un tratamiento bastante común es aquel que se debe realizar con los valores nulos. Pero se puede presentar de dos maneras diferentes: al ser nulo un campo que forma parte de la descripción de una dimensión o un campo que se considere en el almacén parte de la llave compuesta de alguna tabla de hechos.

Para el primer caso, lo recomendable es cambiar todos los valores nulos por alguna cadena de texto que describa la situación, como una especie de valor de inicialización, por ejemplo “Desconocido”, pues el usuario puede interpretar un valor nulo como un error que pudo haber ocurrido. Además, la intención es brindar información y un valor nulo es un concepto más computacional que informativo. En la segunda situación el tratamiento es un poco más especial. Pues se estaría en presencia de un valor nulo para uno de los elementos que conforman la llave de una tabla de hechos. Así antes de cargar los datos de dichas tablas, se deben sustituir las llaves naturales. Pero en este caso, ¿qué sustituir si no se tiene nada? Hay una solución con la que se puede constatar otra de las ventajas de asumir llaves sustitutas. En cada dimensión se podría mantener una única fila, con todos sus valores descriptivos desconocidos, a la cual se le asigne como valor para la llave sustituta, un número que nunca se vaya a otorgar. Se puede referir entonces a este récord como el desconocido general de esa dimensión. Cada vez que se encuentre un valor nulo en lo que será una llave en la tabla de hechos, simplemente se le asigna el desconocido general, manteniendo, de esta manera, la integridad referencial en la base de datos.(Vidot 2008)

Otro evento que puede ocurrir, aunque con menor frecuencia, es la presencia, en las tablas de hechos, de llaves naturales que aún no han sido registradas en las dimensiones, es decir, pudieran aparecer llaves huérfanas. Las filas en donde esto ocurra no pueden ser desechadas, pues, como ya se dijo, el proceso no debe modificar los datos. Podría pensarse entonces en la opción de asociar dichas filas con el desconocido general, pero no resulta conveniente perder, en este punto, la única información que se posee del objeto “ausente”. Es recomendable, por tanto, “obligar” a las dimensiones a incorporar estas llaves aún sin conocer los valores que toman el resto de los atributos. Se deberá generar, con tal objetivo, un valor sustituto para la llave y continuar el proceso de carga, pero también será necesario insertar en la dimensión el nuevo elemento. Como se ha visto, son varios los tratamientos que se pueden realizar sobre los datos, pero ninguno de ellos puede alterar la esencia del contenido. Por tanto, se propone:

- Incluir tratamientos a atributos de dimensiones con valores nulos, así como a llaves nulas y huérfanas.

Dimensiones lentamente cambiantes.

El modelo dimensional debe estar preparado para almacenar el estado real de cada dimensión en cada momento, por lo que debe definirse, de acuerdo a las necesidades que se tengan, el comportamiento a

seguir ante la ocurrencia de cambios. Para manejar las modificaciones en las dimensiones existen diversas técnicas, siendo una de ellas sobrescribir la información almacenada. Otras más especializadas y frecuentemente usadas para los almacenes de datos son la de crear un nuevo récord con la información duplicada, excepto por los campos que se modifiquen, o agregar un campo en la misma fila, manteniendo un valor anterior y uno actual. Estas técnicas son conocidas como dimensiones lentamente cambiantes, concepto que se aplica a una dimensión que debe tener en cuenta los cambios históricos. (Ross and Kimball 2005)

Respecto este punto de vista las dimensiones se pueden clasificar como:

- Tipo 0: No se tiene en cuenta la gestión de los cambios históricos y no se realiza esfuerzo alguno. De manera que alguna información será sobreescrita, pero otra no. Sin planificación alguna.
- Tipo 1: No se guardan cambios históricos. La nueva información sobreescribe la antigua siempre. Principalmente la sobreescritura se realiza por errores de calidad de datos. Este tipo de dimensiones es fácil mantener y son usadas cuando la información histórica no es importante.
- Tipo 2: Toda la información histórica se guarda en el almacén de datos. Cuando hay un cambio se crea una nueva entrada con su fecha y llave sustituta apropiadas. A partir de ese momento será el valor usado para las futuras entradas. Las antiguas usaran el valor anterior.
- Tipo 3: Toda la información histórica se guarda en el almacén de datos. En este caso se crean nuevas columnas con los valores antiguos y los actuales son remplazados con los nuevos.

Es conveniente comentar que algunas dimensiones pueden crecer desmesuradamente. Una buena práctica es romper la dimensión en dos tablas: una que contenga los valores estáticos y otra que contenga los valores volátiles. En este trabajo, se realizó las transformaciones para las dimensiones lentamente cambiantes, desde el punto de vista que cuando los atributos cambien, estos se actualizan. Se ejemplificará este trabajo en los anexos. (Ver Anexo 1)

Análisis de datos.

El análisis de datos es la actividad de verificar un conjunto de datos con el objetivo de ofrecerles un análisis racional. Para realizar este análisis de datos, se realiza desde dos enfoques fundamentales; minería y perfilado de datos. En este trabajo, se utiliza el perfilado de datos mediante la herramienta Talend Open Profiler, aunque se explicarán ambos para un mejor entendimiento:

Minería de Datos

Consiste en la extracción no trivial de información que reside de manera implícita en los datos. Dicha información era previamente desconocida y obtendrá un valor útil en el futuro. En otras palabras, la minería de datos prepara, sondea y explora los datos para sacar la información oculta en ellos. Para ello se usan un conjunto de técnicas encaminadas a la extracción de conocimiento procesable, implícito en las bases de datos. Un proceso típico de minería de datos consta de los siguientes pasos generales:

1. Selección del conjunto de datos, tanto en lo que se refiere a las variables dependientes, como a las variables objetivo, como posiblemente al muestreo de los registros disponibles.
2. Análisis de las propiedades de los datos, en especial los histogramas, diagramas de dispersión, presencia de valores atípicos y ausencia de datos.
3. Transformación del conjunto de datos de entrada, se realizará de diversas formas en función del análisis previo, con el objetivo de prepararlo para aplicar la técnica de minería de datos que mejor se adapte a los datos y al problema.
4. Seleccionar y aplicar la técnica de minería de datos, se construye el modelo predictivo, de clasificación o segmentación.
5. Evaluar los resultados contrastándolos con un conjunto de datos previamente reservado para validar la generalidad del modelo.

Si el modelo final no superara esta evaluación, el proceso se podría repetir desde el principio, o si el experto lo considera oportuno a partir de cualquiera de los pasos anteriores. Esta retroalimentación se podrá repetir cuantas veces se considere necesario hasta obtener un modelo válido. Para ello se utilizan diversas técnicas mediante la minería de datos:

- Redes neuronales: Paradigma de aprendizaje y procesamiento automático inspirado en la forma en que funciona el sistema nervioso de los animales. Se trata de un sistema de interconexión de neuronas en una red que colabora para producir un estímulo de salida.
- Árboles de decisión: Modelo de predicción utilizado en el ámbito de la inteligencia artificial, dada una base de datos se construyen estos diagramas de construcciones lógicas, muy similares a los sistemas de predicción basados en reglas, que sirven para representar y categorizar una serie de condiciones, para la resolución de un problema.

- Modelos estadísticos: Expresión simbólica en forma de igualdad o ecuación que se emplea en todos los diseños experimentales y en la regresión para indicar los diferentes factores que modifican la variable de respuesta.

Perfilado de Datos

Se analizan los datos mediante perfiles de heurística donde se establecen reglas, que se utilizan para medir los datos, y corregirlos. La tabla siguiente es un ejemplo del uso de esta técnica:

Problemas	**Metadatos**	**Ejemplo/heuristica**
Valores Ilegales	Cardinalidad	Ej. Si la cardinalidad del sexo>2 indica problemas.
	Max, Min	No deben exceder el rango de los valores permitidos.
	Desviación y varianza	Los valores estadísticos de la desviación y la varianza no deben ser mayores que un umbral.
Valores escritos incorrectamente	Valores de los atributos	El ordenamiento de los valores ayuda a detectar valores mal escritos. Contar los valores puede ayudar a detectar valores incorrectos
Valores ausentes	Valores nulos	Porciento/ cantidad de valores nulos.
	Valores de atributos + Valores por defecto	La presencia de valores por defecto puede indicar valores que existen que están ausentes.
Diferentes representaciones del mismo valor	Valores de atributos	Comparando el conjunto de valores de una columna de una tabla contra los valores de otra columna de otra tabla.

Duplicados	Cardinalidad+ Unicidad	La cardinalidad de un atributo identificador debe corresponderse con la cantidad de filas de la tabla.
	Valores de atributos	Ordenando los valores por el número de ocurrencias, más de una ocurrencia puede indicar valores duplicados.

Fig. 7. Análisis de datos desde el enfoque del perfil de datos.

En este trabajo, este paso de análisis de datos, se realizó utilizando la herramienta Talend Open Profiler, donde se tomaron como indicadores; el conteo de todas las tuplas dado una tabla, el conteo de los atributos nulos, los atributos distintos, los duplicados, así como una estadística de la calidad y frecuencia de los datos que se analizan. Esta herramienta gestiona otros parámetros, pero estos fueron los principales que permitieron un efectivo análisis, que implicó emitir resultados cuantificables de todos los datos extraídos, así como la comprensión de los resultados e interpretación de los mismos. Por la extensión de este paso, solo se ejemplificará. (Ver Anexo 2)

Metadatos

El término metadatos no tiene una definición única. Según la definición más difundida, metadatos son «datos sobre datos». Debido a que muchas veces no se tiene en cuenta la diferencia entre datos e informaciones también hay muchas declaraciones como «informaciones sobre datos», «datos sobre informaciones» e «informaciones sobre informaciones». Otra clase de definiciones trata de precisar el término como «descripciones estructuradas y opcionales que están disponibles de forma pública para ayudar a localizar objetos»[29] o «datos estructurados y codificados que describen características de instancias conteniendo informaciones para ayudar a identificar, descubrir, valorar y administrar las instancias descritas»[30]. Esta clase de definiciones hace mayor hincapié en los metadatos en relación con la recuperación de información, y surgió de la crítica de que las declaraciones más simples son tan difusas

[29] D. C. A. Bultermann. Is It Time for a Moratorium on Metadata? IEEE Multimedia, IEEE Computer Society Press. 2004.

[30] W. R. Durrell, McGraw-Hill. Data Administration. A Practical Guide to Data Administration. 1985.

y generales que dificultan la tarea de acordarse de estándares, pero estas definiciones no son muy comunes.(Kimball 2004)

Los metadatos pueden describir colecciones de objetos y también los procesos en los que están involucrados, describiendo cada uno de los eventos, sus componentes y cada una de las restricciones que se les aplican. Definen las relaciones entre los objetos, como las tuplas en una base de datos o clases en orientación a objetos, generando estructuras. Los beneficios derivados de la utilización de metadatos son diversos y dependen del área en que se utilicen. Los metadatos adhieren contenido, contexto y estructura a los objetos de información, asistiendo de esta forma al proceso de recuperación de conocimiento desde colecciones de objetos. Teniendo entre sus principales potencialidades las siguientes:

- Permiten generar distintos puntos de vista conceptuales para usuarios o sistemas, y liberan a estos últimos de tener conocimientos avanzados sobre la existencia o características del objeto que describen.
- Permiten el intercambio de información sin la necesidad de que implique el intercambio de los propios recursos.
- En cada proceso productivo, o en cada etapa del ciclo de vida de un objeto de información, se van generando metadatos para describirlos y metadatos para describir dichos metadatos (manual o automáticamente) generando de esta forma valor añadido a los recursos.
- Facilitan un acceso a los recursos en forma controlada ya que se conoce con precisión el objeto descrito.
- Preservan los objetos de información, permitiendo migrar sucesivamente éstos, para su posible uso por parte de las futuras generaciones.
- Coordinan búsquedas, integración y recuperación del conocimiento desde un mayor número de fuentes heterogéneas.

Los diferentes modelos de metadatos que se utilizan para codificar la información y que permiten la recuperación de información se clasifican en:

Tipo	**Objetivo**

Metadatos descriptivos	Descripción e identificación de recursos de información En el nivel (sistema) local facilitan la búsqueda y recuperación. En el nivel web, permite a los usuarios descubrir recursos.
Metadatos estructurales	Facilitan la navegación y presentación de recursos electrónicos Proporcionan información sobre la estructura interna de los recursos, incluyendo página, sección, capítulo, numeración, índices, y tabla de contenidos Describen la relación entre los materiales. Unen los archivos y los textos relacionados.
Metadatos administrativos	Facilitan la gestión y procesamiento de las colecciones digitales tanto a corto como a largo plazo Incluyen datos técnicos sobre la creación y el control de calidad Incluyen gestión de derechos y requisitos de control de acceso y utilización Información sobre acción de preservación.

Fig. 8. Modelos de Metadatos.

El mayor problema que surge con el uso de los metadatos es que estos no se gestionan de una manera automática, es decir, es el propio creador es quien tiene que preocuparse de crearlos y gestionarlos, lo cual es un gran problema si los contenidos cambian constantemente, o si estos son muy grandes. Por ello en este trabajo para la gestión de los metadatos, se creó una tarea que carga esta información desde el FTP de Sentai al almacén de datos operacional por rango de días. Lo cual resuelve el problema de tener los datos necesarios acerca de los datos utilizados por los gerentes del CIMEX.

Seguridad de los datos.

La seguridad de los datos tiene como fin la protección de la información y de los sistemas de información, manteniendo controlado el acceso, uso, divulgación, interrupción o destrucción no autorizada de los datos. Persiguiendo la protección de la confidencialidad, integridad y disponibilidad de la información, independientemente de la forma que los datos puedan tener: electrónicos, impresos, audio u otras. Para mantener una adecuada seguridad de los datos a trasformar, se utiliza un nivel de seguridad de acceso adecuado, donde a la herramienta que consigue una conexión con los datos de origen solo tendrá acceso los desarrolladores ETL, personas con capacidades para el manejo de los datos, que emplearán las claves de acceso. Además antes de su uso se verificó que el Spoon no contiene bombas lógicas, contando el equipo de desarrollo con su última actualización, documentación técnica y operativa. El Spoon de Pentaho Data Integration proporciona la capacidad de almacenar archivos de trabajo y transformaciones en archivos locales o en el repositorio. El repositorio puede ser ubicado en cualquier base de datos relacional común. Esto significa que para cargar una transformación de una base de datos del repositorio, usted necesita conectarse a este. Para ello, es necesario definir una conexión de base de datos a este repositorio. En este trabajo se utiliza el repositorio del Spoon con el objetivo de que el proceso ETL a desarrollar se beneficie con la seguridad planteada por la herramienta para los datos que serán manipulados. La información relativa a los depósitos se almacena en un archivo llamado "repositories.xml". Este archivo se encuentra en el directorio oculto “.kettle” en su directorio de inicio por defecto. En las ventanas esto es C:\Documents and Settings\<username>\.kettle. La contraseña predeterminada para el usuario admin también es admin. Pero con el objetivo de lograr que las personas idóneas conozcan y tengan los permisos necesarios, se cambia esta contraseña. (Pentaho 2007)

Extracción, transformación, y limpieza de los datos.

Para desarrollar el proceso ETL para los datos almacenados en la corporación CIMEX se realiza la extracción, transformación, validación, el proceso cualitativo, filtración y carga de los datos, donde estos pasarían a ser disponibles para el análisis por los usuarios. Durante el proceso ETL se accederá a un FTP, siendo este el origen de los datos que van a ser extraídos y luego trasformados. Para saber cuáles son los ficheros pendientes por extraer, se utiliza un registro de las extracciones realizadas junto a la fecha de completitud del sistema, permitiendo saber por día que se debe extraer, para luego auditar las tareas que se ejecuten a lo largo del proceso. El primer paso que se realizará será la obtención de los

datos que se encuentran ubicados en el FTP. Mientras dichos ficheros son descompactados, se irá verificando que de cada base de datos de la que se extrae están presentes todos los ficheros requeridos, de no estar todos, se frena la carga. Comenzando este proceso de extracción por conocer la disponibilidad de los datos y objetos del negocio que se encuentran distribuidos en múltiples tablas en varios de los módulos Sentai, y verificando que los ficheros solicitados se corresponden con cada una de las tablas que contienen datos necesarios para la carga, es decir, que se reciban tantos ficheros como tablas hubiese sido necesario recorrer para obtener la información. Esto se hace con el objetivo de no sobrecargar a Sentai con este procesamiento. El nombre de dichos ficheros está estructurado de la siguiente manera:

<aliasBaseDatos>.<fechaInicio>.<fechaFin>.<nombreTabla>.txt

Siendo fechaInicio y fechaFin los límites inferior y superior respectivamente del intervalo de tiempo a cargar. Las fechas tienen el formato AAAAMMDD. Cada uno de estos ficheros de texto es compactado en ficheros .gz con una estructura similar en el nombre, y luego agrupados en un único fichero .tar correspondiente a cada base de datos de la que se extrae. Los compactados finales son colocados de manera automática en el FTP creado con ese propósito, del cual el proceso ETL obtiene los datos. El formato general de los ficheros es el siguiente

- Las filas están separadas por los caracteres ASCII #13 y #10 (retorno e inicio de línea).
- Las columnas están separadas por la cadena “|~|”.
- Todos los ficheros tienen un encabezado con el nombre de las columnas (1ra fila).

Las cargas al FTP por la parte Sentai son realizadas por las madrugadas, por lo que la generación de los ficheros ocurre momentos antes de la hora en que comienza la ejecución del proceso ETL.(Vidot 2008)

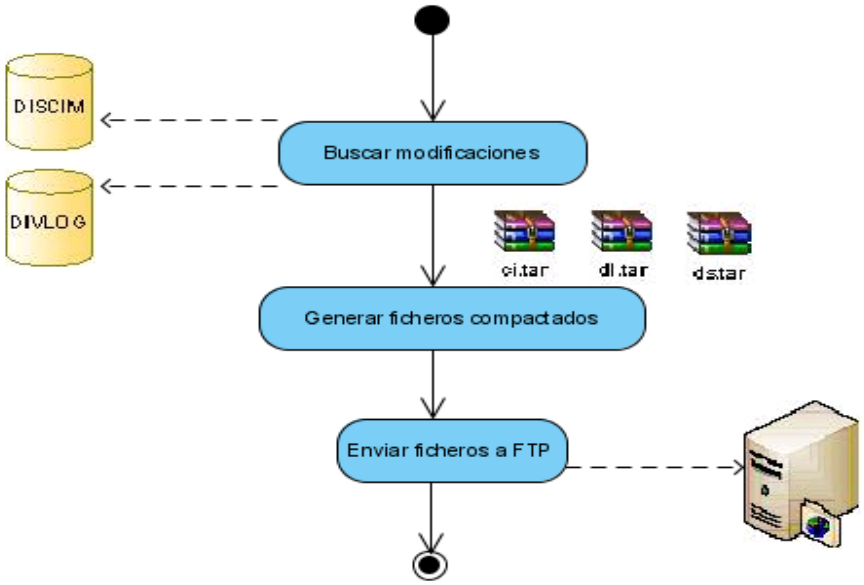


Fig. 9. Orígenes de datos, Generación de ficheros.

Una vez que se tienen los ficheros, comienza un proceso durante el cual es necesario adecuar los datos al modelo dimensional elaborado. Debido a que estos llegan estructurados similarmente como se encuentran en Sentai, se deberán enlazar distintas piezas de información hasta ir completando todos los atributos descriptivos de cada dimensión y todos los campos. Cuando concluye esta etapa es necesario notificar cuáles fueron los datos recibidos por el proceso ETL. Para esto se enviarán al mismo FTP acordado un fichero de confirmación por cada base de datos.

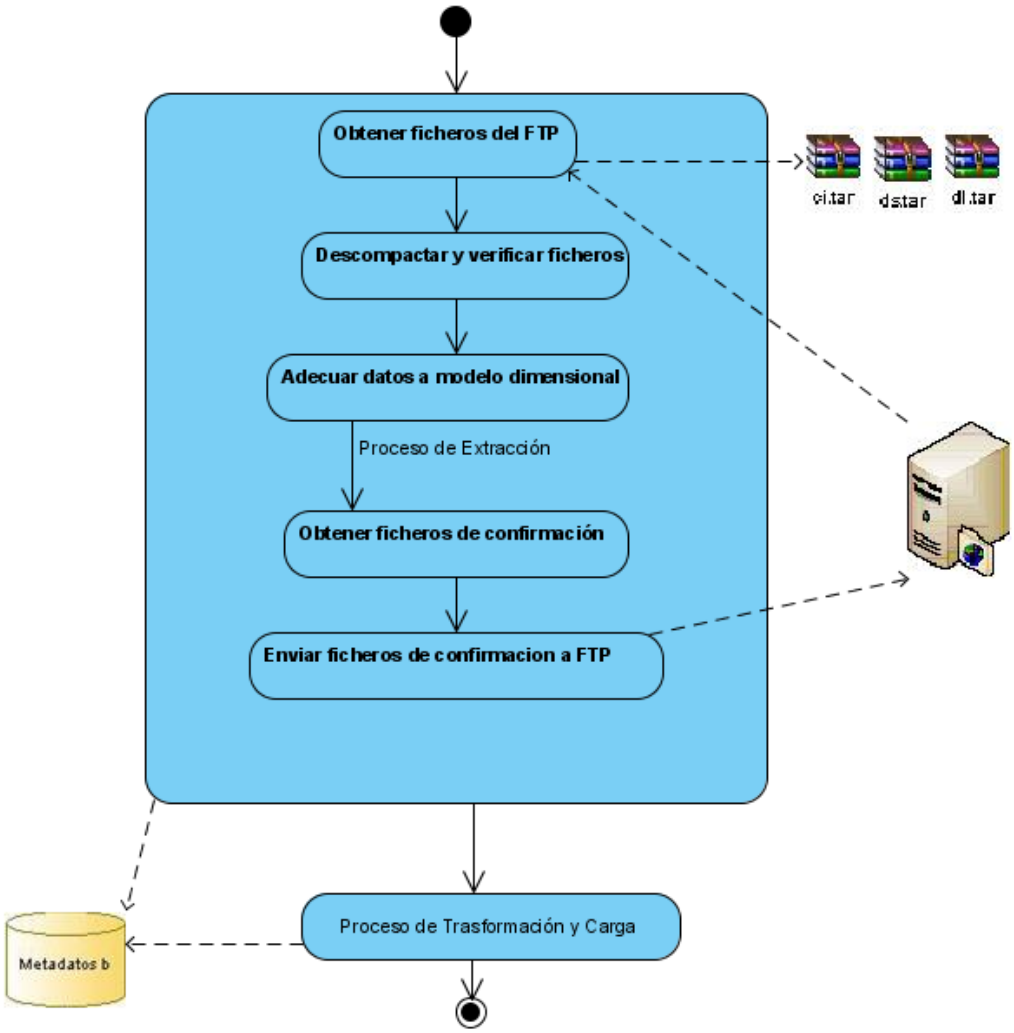


Fig. 10. Extracción en el proceso ETL – CIMEX.

Este fichero se va a identificar con el alias de la base de datos a la que se refiere y va a contener el nombre de todos los ficheros pertenecientes al compactado recibido que fueron extraídos sin problemas. El alias sería:

Nombre	Alias	Descripción
DISCIM	ds	Nomencladores centrales

Circulares	ci	Precio de venta de los productos
División Logística(Divlog)	dl	Transacciones diarias y nomencladores locales

Fig. 11. Especificación de alias por base de datos.

Sentai dispone de dos sistemas para almacenar los datos de los nomencladores centrales, Discim y Divlog. Del sistema Discim se extraen los datos asociados a producto, proveedor, cliente, entre otros. Mientras que en la fuente Divlog se extraen los datos de las transacciones, así como los de las localidades que son particulares para cada sucursal o división. De Sentai también se requieren los datos de las circulares, con las que es posible obtener la información de los precios de cada producto. Cada base de datos de las que se extraen los ficheros se define por un alias que es puesto como nombre de estos ficheros en aras de organizar el proceso y conocer en todo momento en que sistema de almacenamiento se está trabajando. (Ver Anexo 3)

Por cada descarga normal se recibirá un fichero con extensión .tar que tendrá el nombre del alias de la base de datos, y que según la base de datos contiene:

- Base de datos DISCIM (DS):
    - Tablas relacionadas con producto, incluyendo los datos de la jerarquía mayorista
    - Tablas de atributos de productos, es decir, im-attribute e im-attr-val
    - Tablas relacionadas con proveedor
    - Tablas relacionadas con cliente
    - Tabla country de países
    - Tabla unit de unidades de medida
    - Tabla im-group de grupos de inventario
    - Tabla cost-factor de factores de costo

    - Tablas relacionadas con niveles y grupos de precio

- Base de datos Circulares (CI):
    - Tablas relacionadas con circular (circular y circular-d)
- Base de datos DivLog (DL):
    - Tablas relacionadas con las transacciones del módulo de inventario (IM): im-trans,
    - im-trans-x, im-trans-value (incluyendo im-class, y exceptuando im-attribute y el resto de las que son cargadas de DISCIM)
    - Tablas relacionadas con transacciones del módulo de compra: po-trans, po-trans-d, potrans-x
    - Tablas relacionadas con transacciones del módulo de venta: so-pack, so-pack-d, sotrans, so-trans-d, so-pick, so-pick-d, so-trans-x, so-source
    - Tablas relacionadas con entidad.
    - Tablas relacionadas con localidad, incluyendo location_param
    - Tabla s_param

Es aquí donde concluye el proceso de extracción y comienza, por tanto, la etapa de transformaciones y carga. Donde se procede a limpiar los datos que pertenecen al intervalo de extracción que fue determinado, sobre los cuales se aplican transformaciones básicas definidas durante la etapa de identificación de los requerimientos del negocio. Limpieza y ajuste son los principales pasos donde el sistema ETL añade valor. Los demás pasos de la extracción y entrega son evidentemente necesarias, pero sólo se mueven y reformatear los datos. Así la limpieza y ajuste actualmente cambia datos y proporciona orientación si los datos pueden ser utilizados para los propósitos destinados.

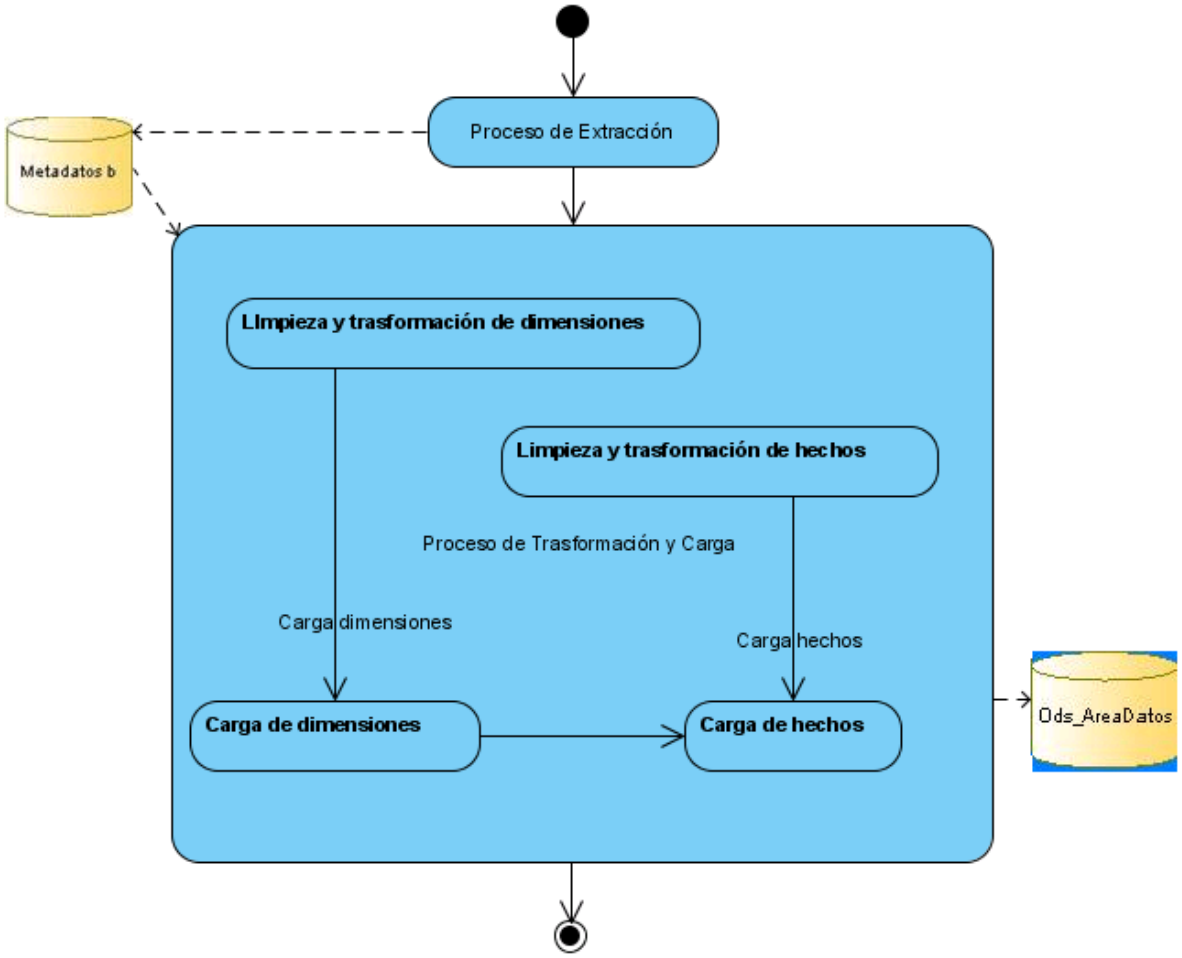


Fig. 12. Transformación y Carga en el proceso ETL – CIMEX.

Entre las transformaciones que se realizan a las dimensiones se encuentran la interpretación de códigos utilizados en el negocio, tratamientos a valores nulos y la generación de llaves sustitutas para las llaves naturales. A la hora de cargar, primero se realiza la carga correspondiente a las dimensiones y luego a los hechos, para evitar la aparición de llaves huérfanas en medio del proceso. A continuación se representan los procesos de extracción y transformación que se utilizaron, donde se puede observar un flujo de distintos pasos, que han sido implementados sobre la herramienta Spoon de la suite Pentaho, la cual está siendo muy utilizada a nivel mundial y por los desarrolladores de almacenes de datos.

Tablas dimensionales.

Las tablas dimensiones son un complemento integral de las tablas de hechos, poseen en si un conjunto de atributos que sirven como fuente primaria para las restricciones de consultas y etiquetas de reportes jugando un papel fundamental dentro de la arquitectura del ODS y de cualquier almacén de datos, ya que hacen el depósito usable y entendible. Donde la claridad y profundidad de los atributos de las dimensiones es directamente proporcional al poder del almacén de datos. (Ross and Kimball 2005)

Por esto se realiza las transformaciones establecidas a las siguientes dimensiones:

Cliente

La dimensión cliente contiene todos los datos referentes a los clientes de la corporación, contiene propiedades como el teléfono, el fax y el email, y otras características que los describen aún más. Así como la información referente al pago que este pueda efectuar en alguna entidad de CIMEX. Presenta en su estructura una jerarquía en su organización, pues un país, agrupa a varias provincias, y estas a su vez a varias ciudades que tienen varios clientes. La arquitectura propuesta para realizarle el proceso ETL a esta dimensión es la siguiente:

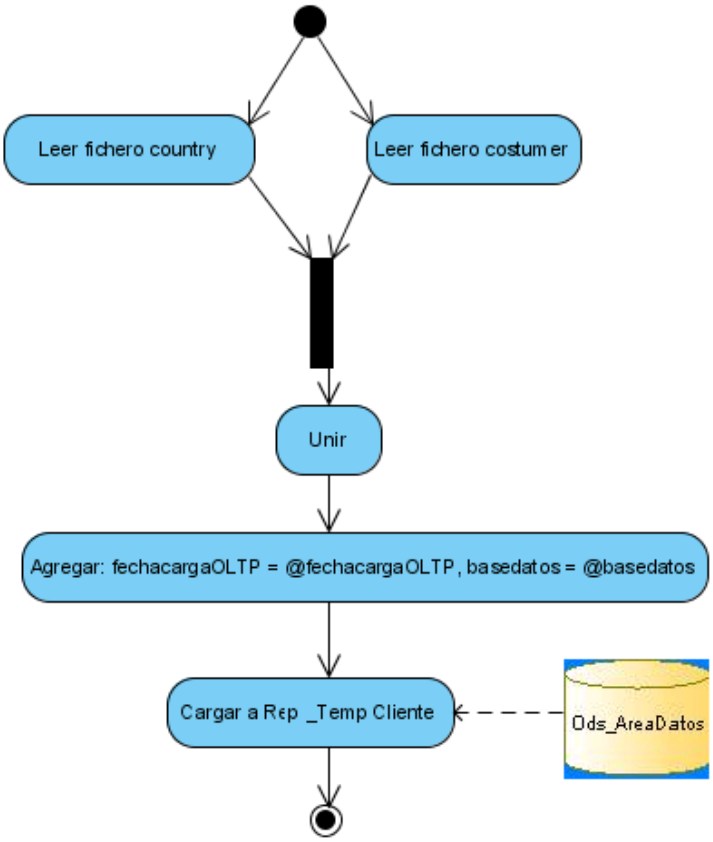


Fig. 13. Arquitectura del proceso ETL- Cliente.

Se comienza realizando dos trasformaciones, donde se ejecutan el script para la carga de los ficheros customer y country, cada transformación sobre uno de estos ficheros, extrayendo los datos y renombrando los atributos para un mejor entendimiento:

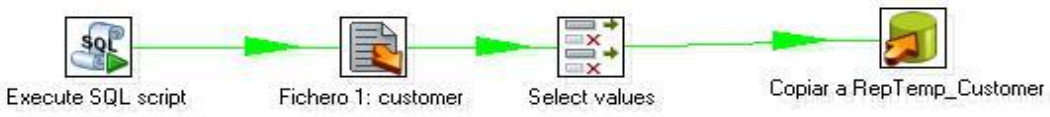


Fig. 14. Transformación al fichero costumer de la dimensión Cliente.

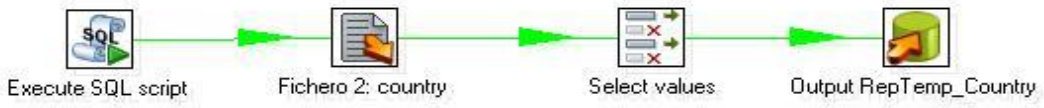


Fig. 15. Transformación al fichero country de la dimensión Cliente.

Luego de la extracción de los datos, se filtran y añaden las contantes BaseDatos y FechaCargaOLTP, aplicando las reglas convenidas a los valores necesarios, cargando luego en la tabla temporal RepTemp_Cliente. Realizando esta transformación de la siguiente forma:

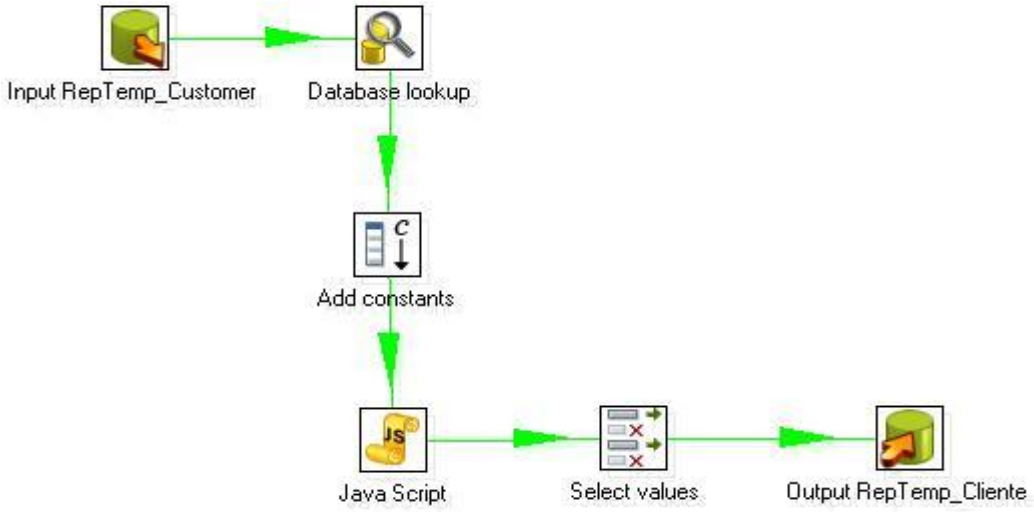


Fig. 16. Cargando en la tabla temporal RepTemp_Cliente.

Unificando estos pasos mediante el trabajo:

Fig. 17. Trabajo o Job que unifica las trasformaciones realizadas a la dimensión Cliente.

Siendo estas las columnas a mapear, y que se cargan como datos finales a la tabla temporal:

Origen	**Destino**
cust-id	Cliente_IdOLTP
name	Cliente_Nombre
search-name	Cliente_NombreBusqueda
address[1]	Cliente_Direccion
city	Cliente_Ciudad
prov	Cliente_Provincia
country-id	Cliente_CodigoPais
country-name	Cliente_Pais
telephone	Cliente_Telefono
fax	Cliente_Fax
mail-cust	Cliente_Email
postal-code	Cliente_CodigoPostal
level-id	Cliente_NivelPrecio
currency-id	Cliente_Moneda
ar-gl	Cliente_Cuenta
charge-cust	Cliente_Cargo
status-code	MT_BanderaActivo
FechaCargaOLTP	MT_FechaCargaOLTP
BaseDatos	MT_BaseDatos

Fig. 18. Mapeo de columnas del proceso ETL a la dimensión Cliente.

Quedando los atributos de la dimensión cliente cargados en las tablas temporales de la siguiente forma:

Field	Type
Cliente_IdODS	integer
Cliente_IdOLTP	varchar(20)
Cliente_Nombre	varchar(130)
Cliente_NombreBusq	varchar(100)
Cliente_Direccion	varchar(200)
Cliente_Ciudad	varchar(30)
Cliente_Provincia	varchar(30)
Cliente_CodigoPais	varchar(60)
Cliente_Pais	varchar(30)
Cliente_Telefono	varchar(40)
Cliente_Fax	varchar(40)
Cliente_Email	varchar(60)
Cliente_CodigoPostal	varchar(30)
Cliente_NivelPrecio	varchar(20)
Cliente_Moneda	varchar(30)
Cliente_Cuenta	varchar(20)
Cliente_Cargo	varchar(20)
Cliente_Publico	varchar(20)
MT_BanderaActivo	varchar(20)
MT_BaseDatos	varchar(50)
MT_FechaCargaOLT	timestamp
MT_FechaCargaODS	timestamp
MT_IdCicloCarga	integer

Fig. 19. Tabla temporal cargada en el ODS_CIMEX.

Proveedor

La dimensión proveedor es bastante similar a la de cliente, pero con la intención de tener documentado información gerencial, aparecen nuevos niveles como el organismo al que pertenece y el tipo de proveedor. Otras propiedades como la moneda que utiliza, su cuenta y el cargo que tiene, ampliando un tanto su esquema semántico. Para esta dimensión, el diseño sería:

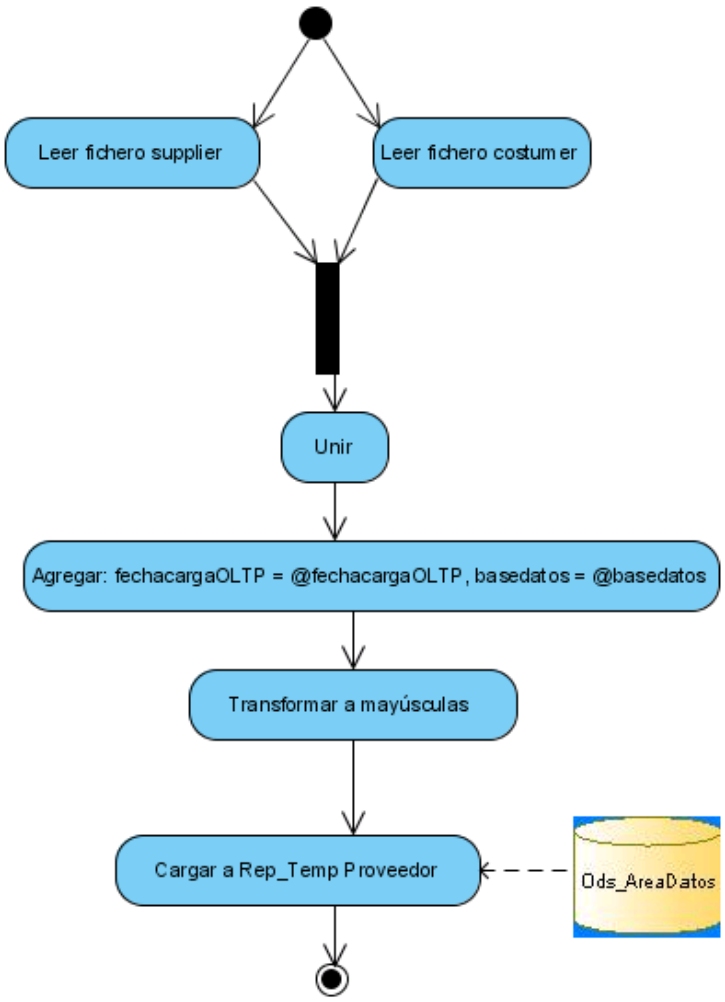


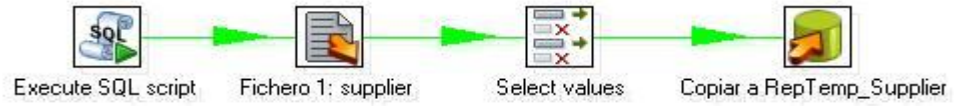
Fig. 20. Arquitectura del proceso ETL para la dimensión Proveedor.

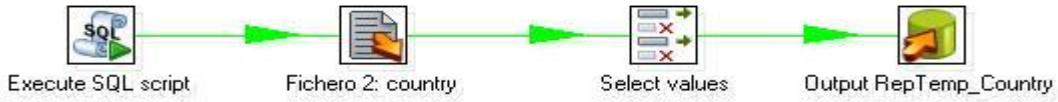
Donde los valores a mapear son los que se muestran a continuación:

Origen	Destino
supp-id	Prov_IdOLTP
name	Prov_Nombre
udf-1	Prov_CodigoPanamericano
search-name	Prov_NombreBusqueda
address[1]	Prov_Direccion
city	Prov_Ciudad
prov	Prov_Provincia
country-id	Prov_CodigoPais
country-name	Prov_Pais
telephone	Prov_Telefono
fax	Prov_Fax
postal-code	Prov_CodigoPostal
currency-id	Prov_Moneda
ap-gl	Prov_Cuenta
remit-supp	Prov_Cargo
udf-2	Prov_Organismo
status-code	MT_BanderaActivo
FechaCargaOLTP	MT_FechaCargaOLTP
BaseDatos	MT_BaseDatos

Fig. 21. Mapeo de columnas del proceso ETL para la dimensión Proveedor.

Comenzando por extraer los ficheros supplier y costumer, renombrando, filtrando y cargando luego en la tabla temporal RepTemp_Proveedor a través de las siguientes trasformaciones:





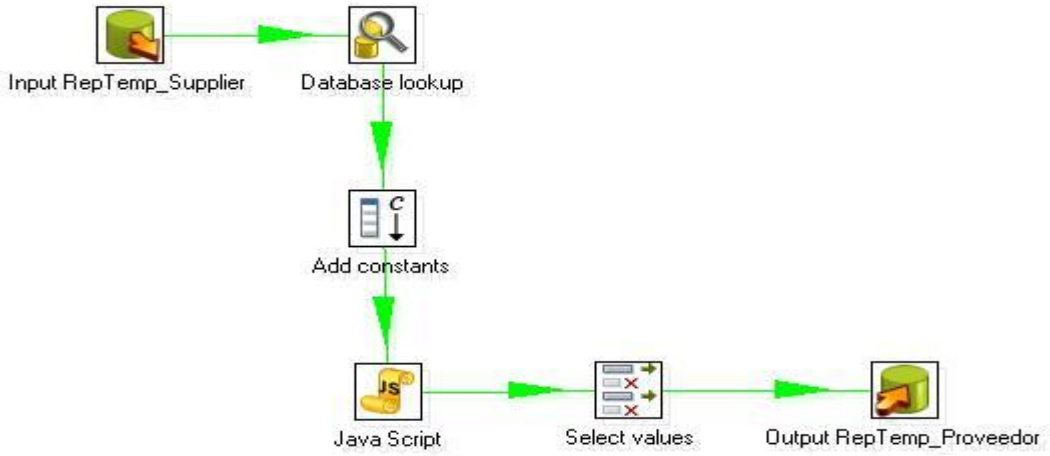


Fig. 22. Trasformaciones realizadas sobre la dimensión Proveedor.

Fig. 23. Trabajo que unifica las trasformaciones realizadas sobre la dimensión Proveedor.

Cargando en el almacén de datos:

Field	Type
Prov_IdODS	integer
Prov_IdOLTP	varchar(20)
Prov_Nombre	varchar(100)
Prov_CodigoPaname	varchar(15)
Prov_NombreBusque	varchar(30)
Prov_Direccion	varchar(150)
Prov_Ciudad	varchar(30)
Prov_Provincia	varchar(30)
Prov_CodigoPais	varchar(60)
Prov_Pais	varchar(30)
Prov_Telefono	varchar(20)
Prov_Fax	varchar(20)
Prov_Email	varchar(60)
Prov_CodigoPostal	varchar(30)
Prov_Moneda	varchar(15)
Prov_Cuenta	varchar(15)
Prov_Cargo	varchar(20)
Prov_Organismo	varchar(100)
Prov_TipoProveedor	varchar(30)
MT_BanderaActivo	varchar(20)
MT_BaseDatos	varchar(50)
MT_FechaCargaOLT	timestamp
MT_FechaCargaODS	timestamp
MT_IdCicloCarga	integer

Fig. 24. Tabla temporal RepTemp_Proveedor que se cargó en el ODS_CIMEX.

Producto

En la dimensión producto, como en las anteriores, también es visible la jerarquía que se establece en los atributos Prod_Origen, Prod_Familia y Prod_Id, desde el nivel superior al inferior. Por ello el proceso ETL de esta dimensión se complejiza, ya que no solo es preciso renombrar los atributos, y adicionar las constantes, sino se filtran alguna filas con el objetivo de controlar los productos validos. El diseño de este proceso es:

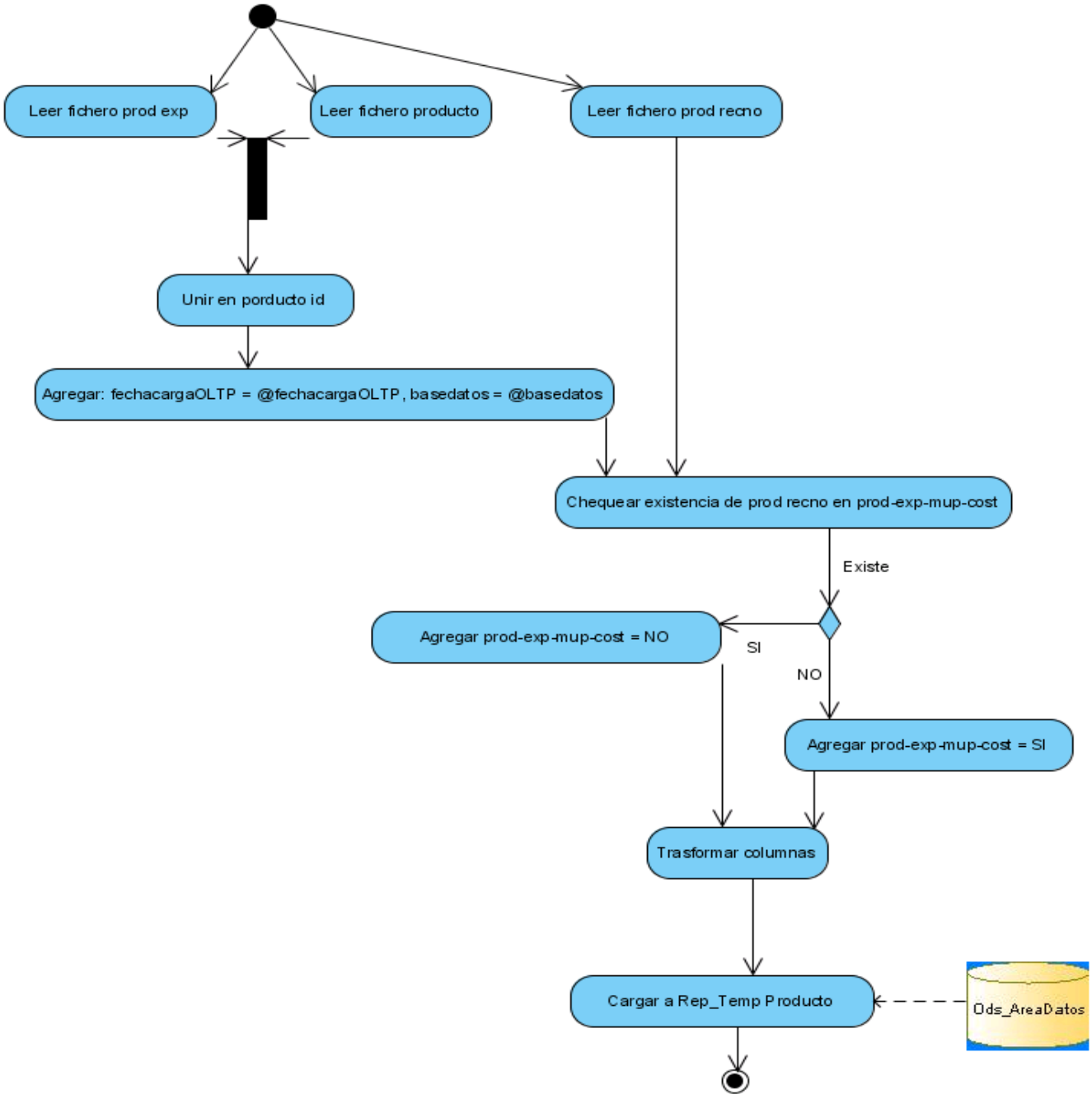


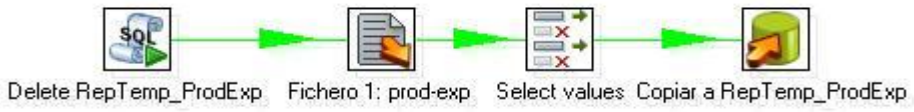
Fig. 25. Arquitectura del proceso ETL de la dimensión Producto.

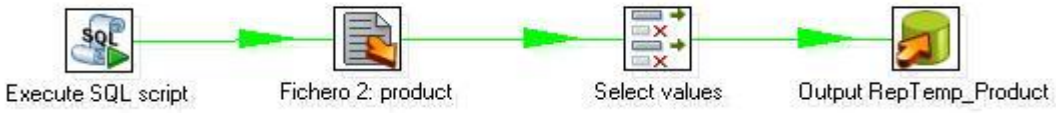
Donde los atributos a mapear son los siguientes:

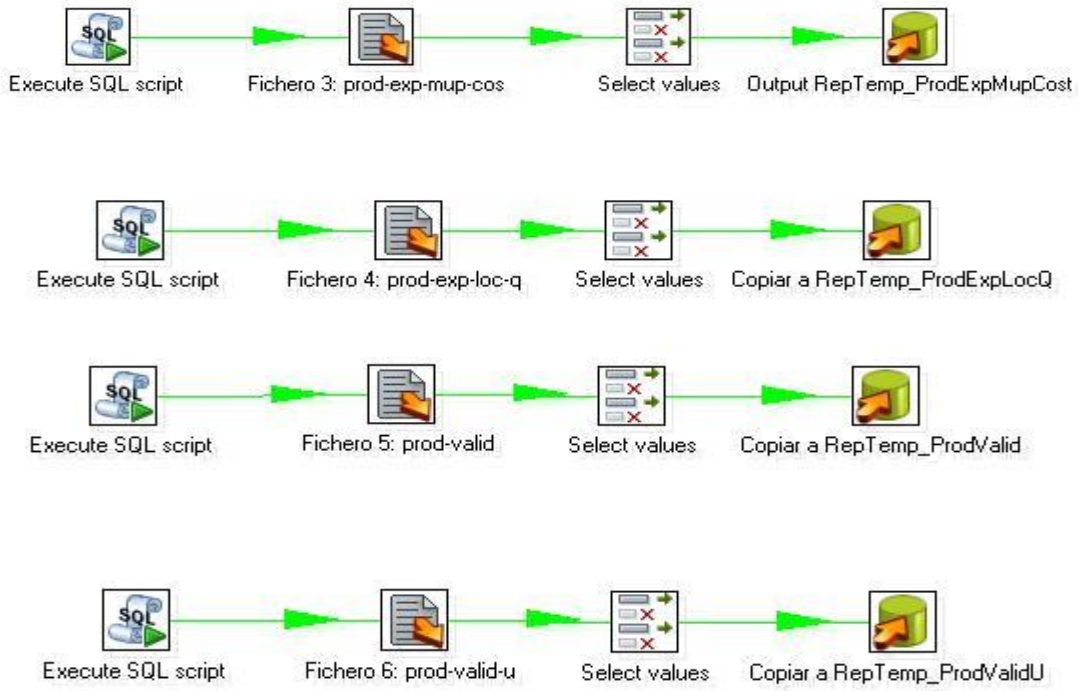
Origen	Destino
prod-recno	Prod_RecNo
report-piece	Prod_IdSentai
bar-code-syn	Prod_CodigoBarra
description	Prod_Descripcion
product-id	Prod_Familia
udf-1	Prod_Origen
smallest-unit	Prod_UnidadMenor
stock-unit	Prod_UnidadExistencia
storage-unit	Prod_UnidadAlmacenaje
purch-unit	Prod_UnidadCompra
sell-unit	Prod_UnidadVenta
bm-unit	Prod_UnidadExplosionMateriales
calc-attr	Prod_Atribuido
no-of-attr	Prod_CantAtributos
group-id	Prod_GrupoCuenta
p-m	Prod_Manufacturado
packing-group	Prod_GrupoEmpaque
r-f	Prod_MateriaPrima
org-code	Prod_FormaTraza
prod-exp-mup-cost	Prod_Minorista
perpetual	Prod_Perpetuo
status-code	MT_BanderaActivo
FechaCargaOLTP	MT_FechaCargaOLTP
BaseDatos	MT_BaseDatos

Fig. 26. Mapeo de columnas del proceso ETL-Producto.

Realizando cada transformación convenida en el diseño, teniendo como salidas al almacén de datos, las tablas temporales RepTemp_Producto y RepTemp_ConversionUm, quienes contienen los datos limpios y listos para la carga en el ODS_CIMEX. A continuación se muestran las trasformaciones realizadas:






Execute SQL script
Fichero 3: prod-exp-mup-cos
Select values
Output RepTemp_ProdExpMupCost
Execute SQL script
Fichero 4: prod-exp-loc-q
Select values
Copiar a RepTemp_ProdExpLocQ
Execute SQL script
Fichero 5: prod-valid
Select values
Copiar a RepTemp_ProdValid
Execute SQL script
Fichero 6: prod-valid-u
Select values
Copiar a RepTemp_ProdValidU

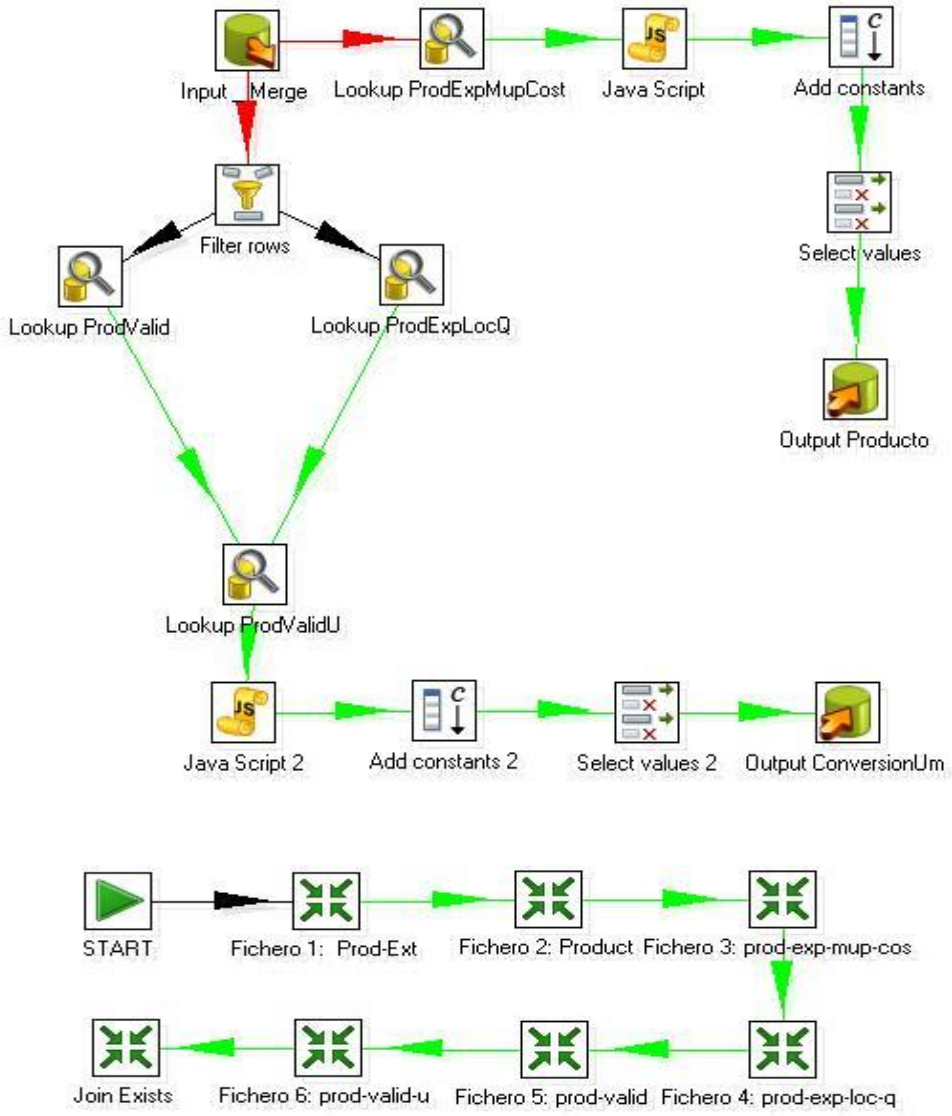
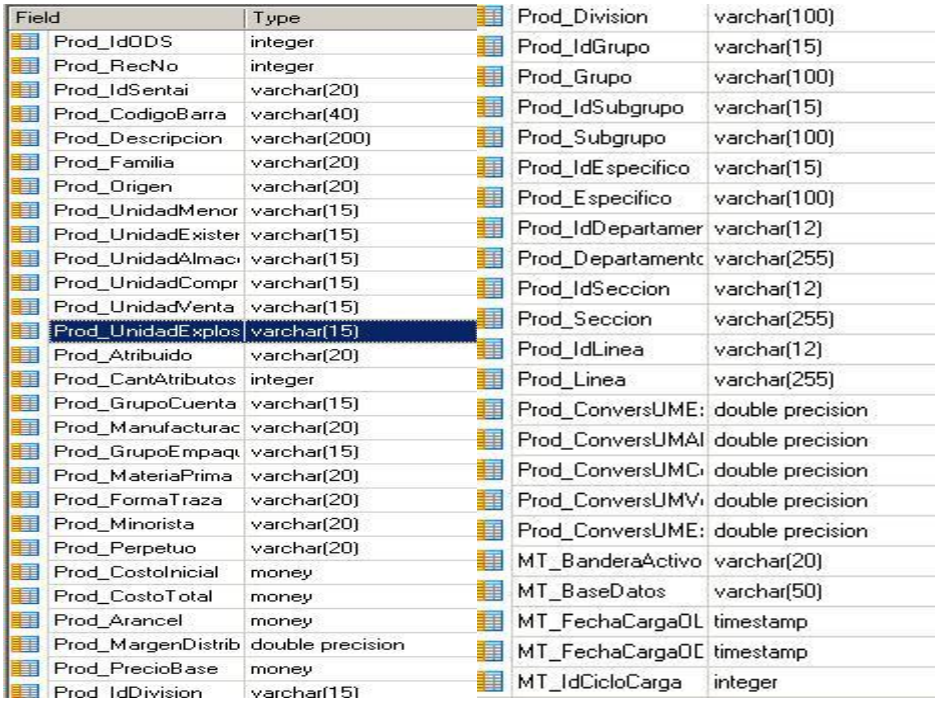


Fig. 27. Extracción y trasformaciones realizadas en la dimensión Producto.

Siendo esta la tabla temporal RepTemp_Producto que se cargó:



Field	Type
Prod_IdODS	integer
Prod_RecNo	integer
Prod_IdSentai	varchar(20)
Prod_CodigoBarra	varchar(40)
Prod_Descripcion	varchar(200)
Prod_Familia	varchar(20)
Prod_Origen	varchar(20)
Prod_UnidadMenor	varchar(15)
Prod_UnidadExister	varchar(15)
Prod_UnidadAlmac	varchar(15)
Prod_UnidadCompr	varchar(15)
Prod_UnidadVenta	varchar(15)
Prod_UnidadExplos	varchar(15)
Prod_Atribuido	varchar(20)
Prod_CantAtributos	integer
Prod_GrupoCuenta	varchar(15)
Prod_Manufacturac	varchar(20)
Prod_GrupoEmpaqu	varchar(15)
Prod_MateriaPrima	varchar(20)
Prod_FormaTraza	varchar(20)
Prod_Minorista	varchar(20)
Prod_Perpetuo	varchar(20)
Prod_CostoInicial	money
Prod_CostoTotal	money
Prod_Arancel	money
Prod_MargenDistrib	double precision
Prod_PrecioBase	money
Prod_IdDivision	varchar(15)
Prod_Division	varchar(100)
Prod_IdGrupo	varchar(15)
Prod_Grupo	varchar(100)
Prod_IdSubgrupo	varchar(15)
Prod_Subgrupo	varchar(100)
Prod_IdEspecifico	varchar(15)
Prod_Especifico	varchar(100)
Prod_IdDepartamer	varchar(12)
Prod_Departamentc	varchar(255)
Prod_IdSeccion	varchar(12)
Prod_Seccion	varchar(255)
Prod_IdLinea	varchar(12)
Prod_Linea	varchar(255)
Prod_ConversUME:	double precision
Prod_ConversUMAl	double precision
Prod_ConversUMC	double precision
Prod_ConversUMV	double precision
Prod_ConversUME:	double precision
MT_BanderaActivo	varchar(20)
MT_BaseDatos	varchar(50)
MT_FechaCargaOL	timestamp
MT_FechaCargaOL	timestamp
MT_IdCicloCarga	integer

Fig. 28. Tabla temporal de la dimensión Producto en el ODS_CIMEX.

Nivel Precio

Esta dimensión contiene campos referentes al identificador y nombre del nivel de precio. La arquitectura propuesta para realizarle el proceso ETL es la siguiente:

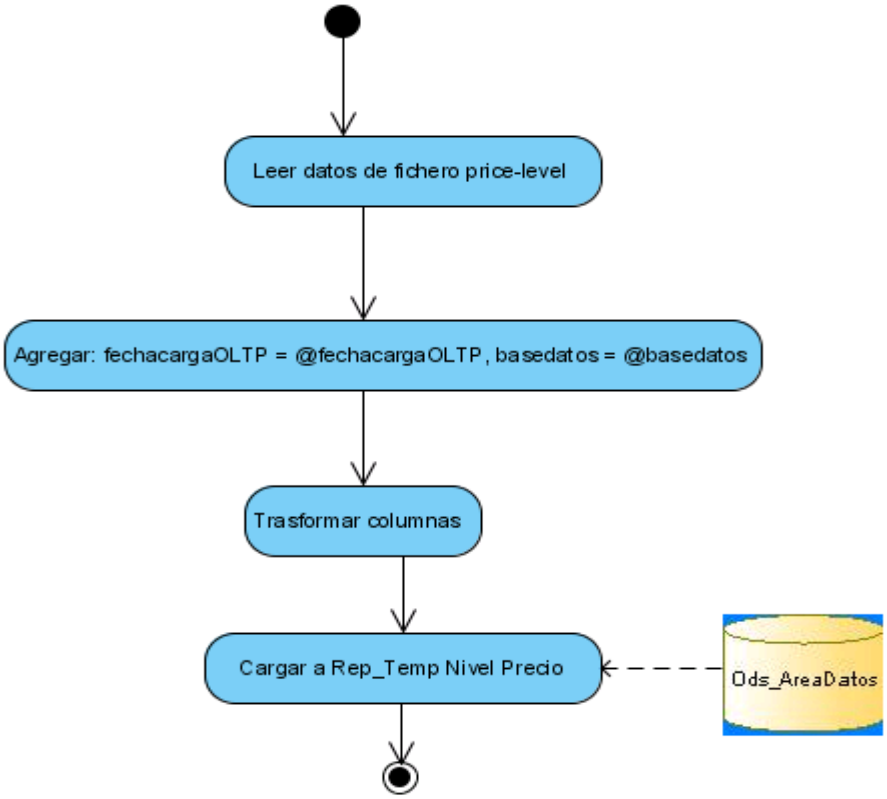


Fig. 29. Arquitectura de la dimensión Nivel precio.

Donde el fichero price-level que está en el FTP se extrae, y luego de ser trasformado será temporalmente insertados en la tabla Rep Temp_NivelPrecio:

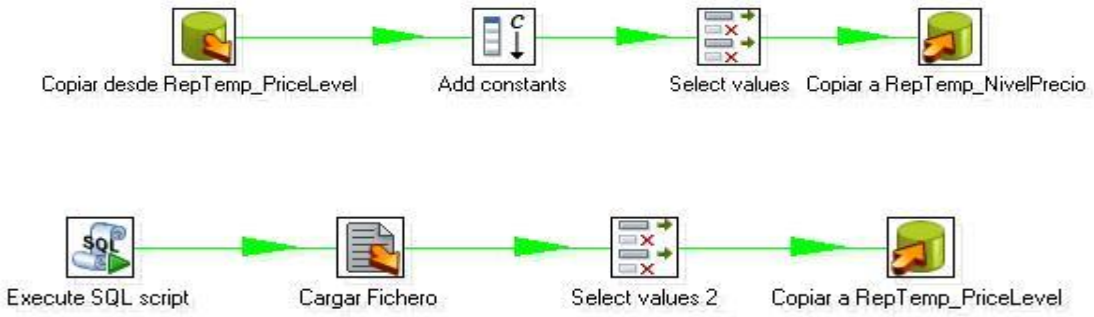


Fig. 30. Trasformaciones realizadas sobre la dimensión Nivel Precio.

Unificando estos pasos mediante el siguiente job:

Fig. 35. Trabajo para unificar los procesos ETL.

Para cargar los datos finales en el almacén de datos, se mapean las siguientes columnas:

Mapeo de Columnas

Origen	Destino
level-id	NivelPrecio_IdOLTP
description	NivelPrecio_Nombre
BanderaActivo	MT_BanderaActivo
FechaCargaOLTP	MT_FechaCargaOLTP
BaseDatos	MT_BaseDatos

Fig. 31. Mapeo de columnas en el proceso ETL de la dimensión Nivel precio.

Quedando finalmente los datos cargados de la siguiente forma en la tabla temporal Rep Temp_NivelPrecio:

Field Name	Field Type
NivelPrecio_IdODS	integer
NivelPrecio_IdOLTP	varchar(20)
NivelPrecio_Nombre	varchar(60)
MT_BanderaActivo	varchar(20)
MT_BaseDatos	varchar(50)
MT_FechaCargaOLTP	timestamp
MT_FechaCargaODS	timestamp
MT_IdCicloCarga	integer

Fig. 32. Atributos cargados en la tabla temporal Rep Temp_NivelPrecio.

Atributo valor

La dimensión atributo valor contiene los campos correspondientes al identificador del atributo, identificador del valor y descripción del mismo. La arquitectura propuesta para esta dimensión es la siguiente:

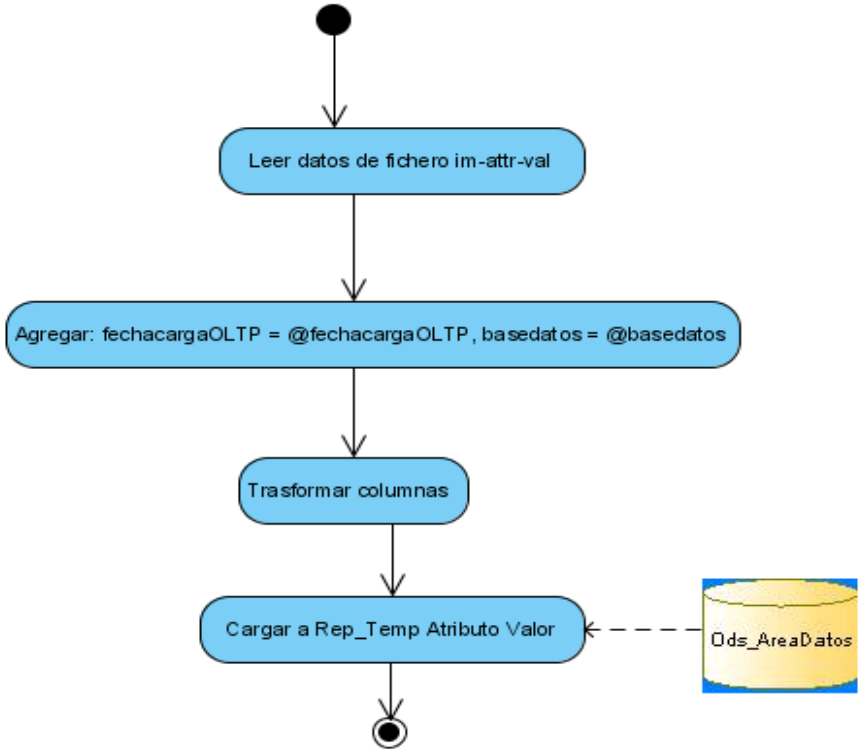
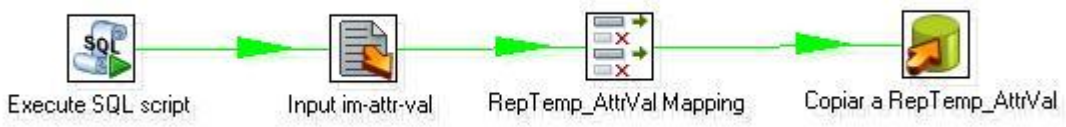


Fig. 33. Arquitectura de la dimensión Atributo Valor.

La transformación de esta dimensión es muy sencilla, pues se extrae del FTP el fichero im-attr-val, se realiza un mapeo de sus atributos y se le agregan las columnas BanderaActivo, FechaCargaOLTP y BaseDatos y luego es cargada en la tabla temporal Rep Temp_AtributoValor, este proceso se realiza de la siguiente forma:



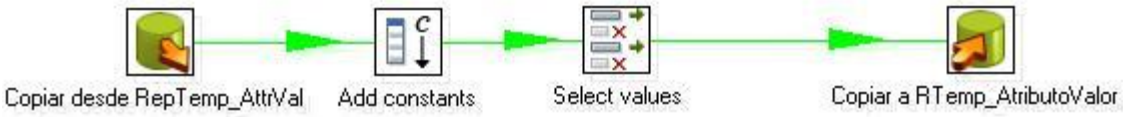


Fig. 34. Trasformaciones realizadas sobre la dimensión Atributo Valor

Unificando el proceso por el siguiente trabajo o job:

Fig. 35. Trabajo o job para el proceso ETL en la dimensión Atributo Valor

Para realizar la carga de los datos en las tablas temporales se realiza el siguiente mapeo de columnas:

Mapeo de Columnas

Origen	Destino
attr-id	Atr_IdOLTP
value-id	Valor_IdOLTP
description	Valor_Descripcion
BanderaActivo	MT_BanderaActivo
FechaCargaOLTP	MT_FechaCargaOLTP
BaseDatos	MT_BaseDatos

Fig. 36. Mapeo de columnas en la dimensión Atributo Valor

Quedando finalmente la tabla temporal Rep Temp_AtributoValor de la siguiente forma:

Field Name	Field Type
Atr_IdOLTP	varchar(20)
Valor_IdOLTP	varchar(20)
Valor_Descripcion	varchar(40)
MT_BaseDatos	varchar(50)
MT_BanderaActivo	varchar(20)
MT_FechaCargaOLTP	timestamp
MT_FechaCargaODS	timestamp
MT_IdCicloCarga	integer

Fig. 37. Tabla temporal de la dimensión Atributo Valor en el ODS_CIMEX.

Atributo

Otra dimensión a trasformar es atributo, por la similitud con la dimensión atributo valor, solo mostraremos en este documento la arquitectura propuesta para esta dimensión, con el fin de no explicar dimensiones similares:

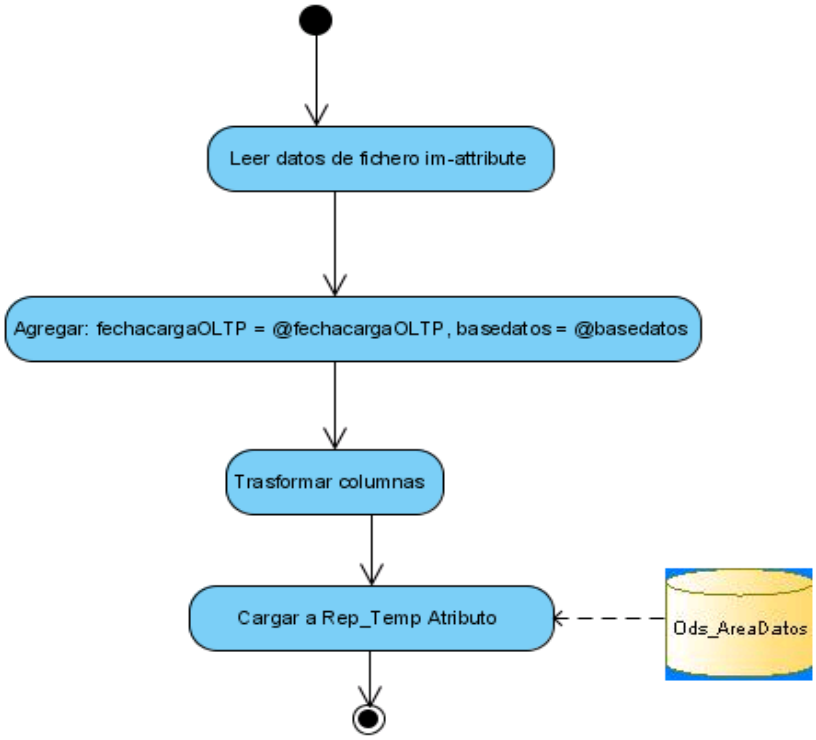


Fig. 38. Arquitectura de la dimensión Atributo.

En esta dimensión al igual que en la anterior, se comienza por extraer los datos del fichero im-attribute, posteriormente son renombrado sus valores y se adicionan las constantes necesarias con el fin de lograr la transformación adecuada para ser cargados estos datos en la tabla Rep Temp _Atributo:

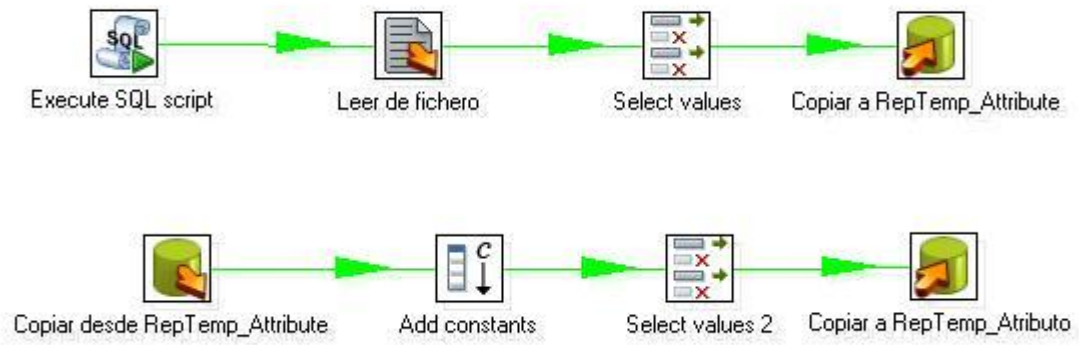


Fig. 39. Trasformaciones sobre la dimensión Atributo.

Fig. 40. Trabajo del proceso ETL para la dimensión Atributo.

Quedando posteriormente tabla Rep Temp_Atributo de la siguiente forma en las tablas temporales:

Atr_IdOLTP	varchar(20)
Atr_Descripcion	varchar(40)
Atr_ValorPorDefecto	varchar(20)
MT_BaseDatos	varchar(50)
MT_BanderaActivo	varchar(20)
MT_FechaCargaOLTP	timestamp
MT_FechaCargaODS	timestamp
MT_IdCicloCarga	integer

Fig. 41. Tabla temporal de la dimensión Atributo en el ODS_CIMEX.

Grupo Inventario

La dimensión grupo inventario contiene los campos correspondientes al identificador que es utilizado en el negocio para clasificar los productos y la descripción del grupo inventario. La arquitectura propuesta para esta dimensión es la siguiente:

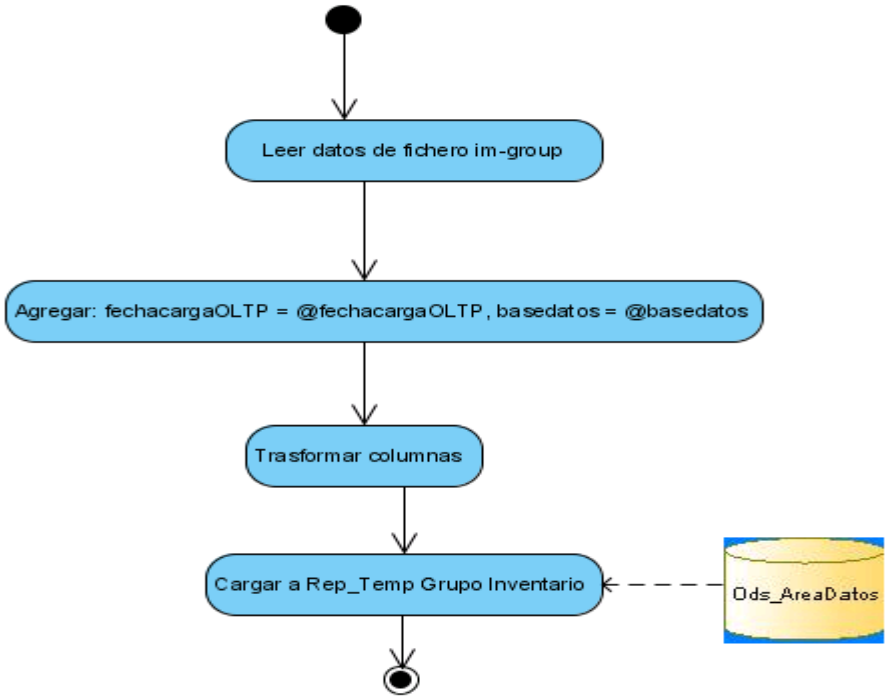


Fig. 42. Arquitectura para la dimensión Grupo Inventario.

Esta dimensión es tratada de forma semejante a las dimensiones atributo y atributo valor, quedando cargado en la tabla temporal Rep Temp_GrupoInvetario, los atributos extraídos y trasformados del fichero im-group:

.

GrupoInv_IdODS	integer
GrupoInv_IdOLTP	varchar(15)
GrupoInv_Nombre	varchar(60)
MT_BanderaActivo	varchar(20)
MT_BaseDatos	varchar(50)
MT_FechaCargaOLTP	timestamp
MT_FechaCargaODS	timestamp
MT_IdCicloCarga	integer

Fig. 43. Tabla temporal de la dimensión Grupo Inventario en el ODS_CIMEX

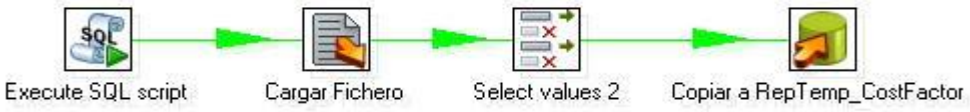
Además de las dimensiones ya explicadas en el proceso ETL para CIMEX, existen algunas otras que no se detallaran, debido a que en complejidad son similares a las vistas hasta el momento, entre ellas se encuentran:

Tiempo

Dimensión que está virtualmente garantizada a estar presente en cada proceso de negocio del CIMEX, porque virtualmente cada transacción, compra o venta es una serie de tiempo. El tiempo es usualmente la primera dimensión en el orden subyacente de organización en el almacén de datos, donde las cargas sucesivas de intervalos de tiempos de datos cargarán a estos en un territorio virgen en el disco. En tal sentido se define una dimensión tiempo con varias categorías o niveles para su mejor organización.

Factor costo

Esta dimensión contiene la configuración de todos los posibles factores de costos que se puede usar**,** en el proceso ETL de esta dimensión solo es necesario renombrar los valores y adicionar las constantes. Siendo estas las trasformaciones realizadas sobre esta dimensión:



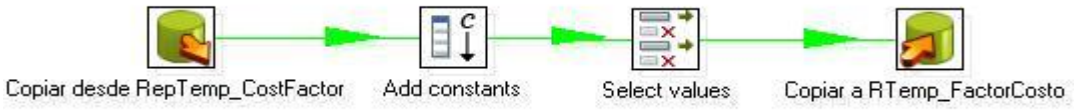


Fig. 44. Trasformaciones de la dimensión Factor costo.

Logrando cargar en el almacén los siguientes atributos:

Field	Type
Factor_IdODS	integer
Factor_IdOLTP	integer
Factor_Descripcion	varchar(60)
Factor_Tipo	varchar(30)
Factor_Indice	smallint
MT_BanderaActivo	varchar(20)
MT_BaseDatos	varchar(50)
MT_FechaCargaOLT	timestamp
MT_FechaCargaODS	timestamp
MT_IdCicloCarga	integer

Fig. 45. Tabla temporal de la dimensión Factor costo.

Específico

Contiene los campos correspondientes al identificador de la división, del grupo y del subgrupo, así como el nombre y código específico. En esta dimensión como en algunas de las anteriores los son datos leídos desde del fichero prod-specify y transformados por el siguiente proceso:

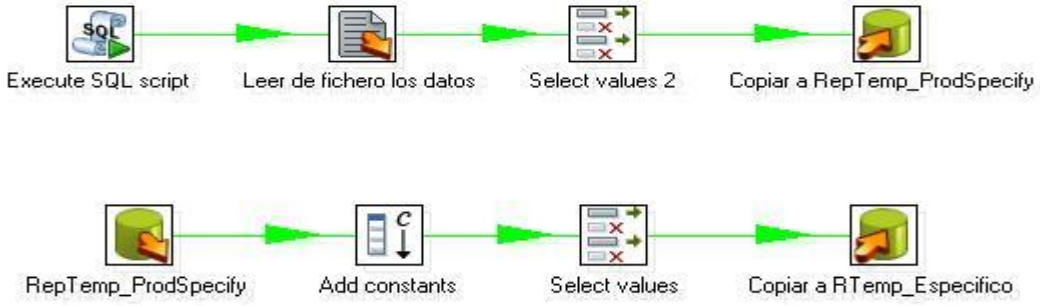


Fig. 46. Trasformaciones sobre la dimensión Específico.

Chequeando que los identificadores de grupo, subgrupo, división y específico sean distintos de NULL, siendo luego cargados en la tabla temporal Rep Temp_Específico.

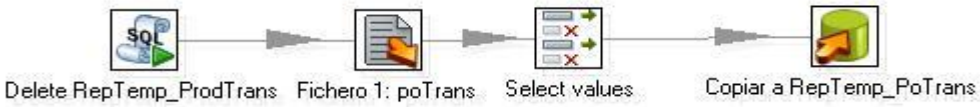
Tablas hechos.

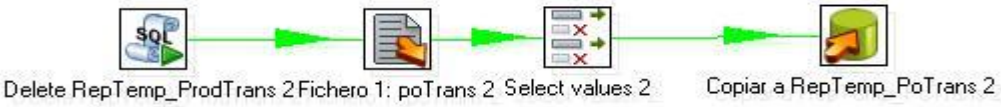
Las tablas de hechos, son tablas primarias en el modelo dimensional, donde es almacenado el rendimiento de las dimensiones numéricas del negocio.(Ross and Kimball 2005)

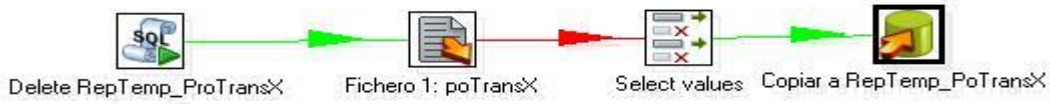
Así cada tabla de hecho define un departamento determinado, que hace referencia a la medida del negocio. Teniendo en cuenta que la condición principal que debe cumplir una tabla de hecho es que lo que se almacene ha de ser medible, es decir numérico, y a su vez aditivo para que se puedan realizar operaciones sobre él. En este trabajo se le realiza transformaciones a los hechos compra, ventas, y ajuste de precio.

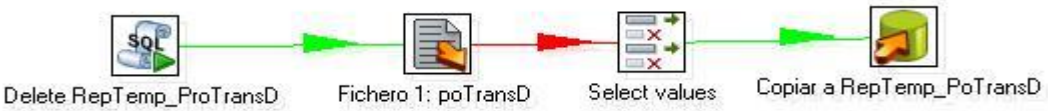
Compra

En cuanto a los hechos, se detalla la tabla compra en algunas medidas como la cantidad, el importe, la unidad de medida, entre otros. En dicha tabla se encuentran los identificadores de las dimensiones con las cuales se relaciona, además de los datos referentes a las compras efectuadas por la corporación: como productos que se compran, áreas de compra, y otras muchas que facilitan la toma de decisiones de los empresarios del CIMEX. El proceso ETL realizado para este hecho, se encarga de renombrar los atributos, y filtrar los campos necesarios, pues en el origen de datos, existen campos que no dan al traste con lo que el usuario final necesita, cargando luego en la tabla temporal RepTemp_OCompra.


Delete RepTemp_ProdTrans
Fichero 1: poTrans
Select values
Copiar a RepTemp_PoTrans


Delete RepTemp_ProdTrans 2
Fichero 1: poTrans 2
Select values 2
Copiar a RepTemp_PoTrans 2


Delete RepTemp_ProTransX
Fichero 1: poTransX
Select values
Copiar a RepTemp_PoTransX


Delete RepTemp_ProTransD
Fichero 1: poTransD
Select values
Copiar a RepTemp_PoTransD

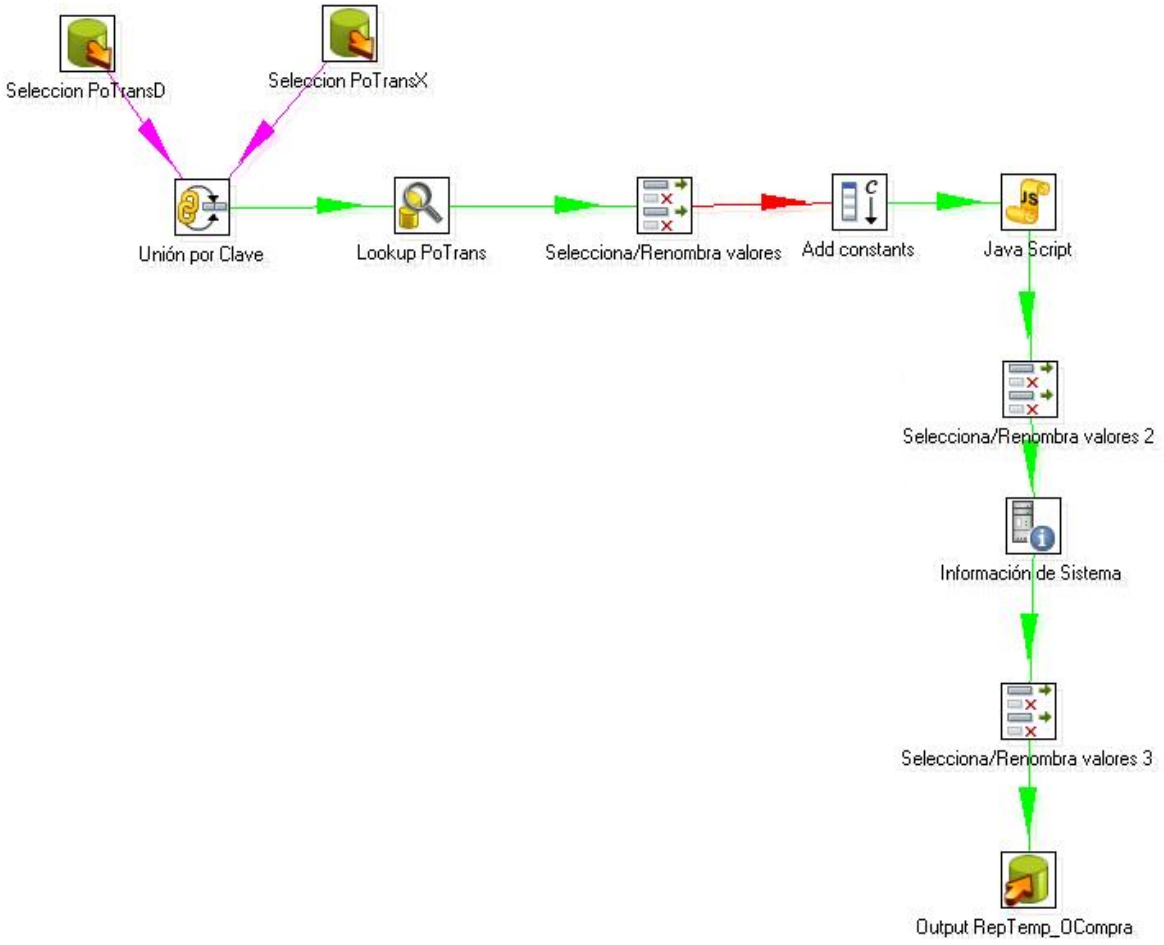
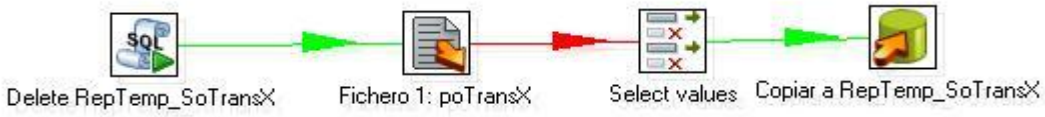


Fig. 47. Proceso ETL realizado al hecho Compra.

Venta

Muy similar al hecho compra, los principales atributos implicados en la venta son algunas medidas como el importe, la cantidad, el costo, la utilidad, además de la fecha, área y producto que se vende. Realizando prácticamente un proceso ETL igual al del hecho compra:



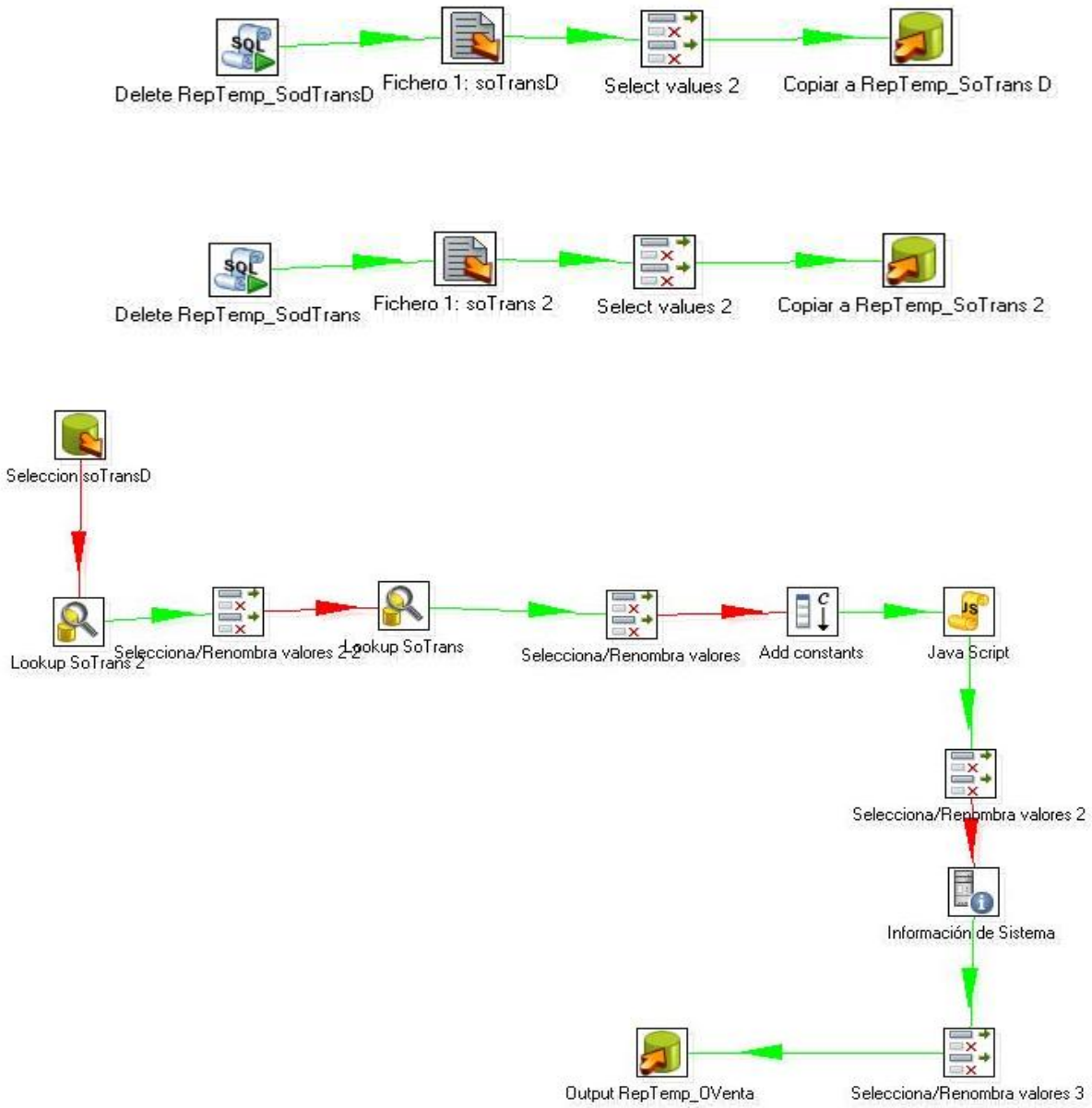


Fig. 48. Proceso ETL realizado al hecho Venta.

Obteniendo en el almacén cargado los datos, siendo la siguiente tabla temporal RmpTemp_OVenta muestra de esto:

Field	Type
Prod_IdODS	integer
Area_IdODS	integer
Fecha_IdODS	integer
Cliente_IdODS	integer
Prov_IdODS	integer
FuenteV_IdODS	integer
Trans_IdODS	integer
OV_IdODS	integer
Doc_IdODS	integer
OV_Numero	integer
Venta_CodigoTrans	varchar(20)
Venta_Cantidad	double precision
Venta_Importe	money
Venta_Costo	money
Venta_Utilidad	money
Venta_EntidadControl	varchar(50)
Venta_UMMenor	varchar(15)
Venta_UMVenta	varchar(15)
Venta_CantUMV	double precision
MT_BaseDatos	varchar(50)
Calidad_PesoConfianza	double precision
MT_IdCicloCarga	integer

Fig. 49. Tabla temporal del hecho Venta en el ODS_CIMEX.

Ajuste precio

El hecho ajuste de precio contiene los datos del identificador del tipo de ajuste y la descripción del mismo. La arquitectura propuesta para realizar los procesos de ETL es la siguiente.

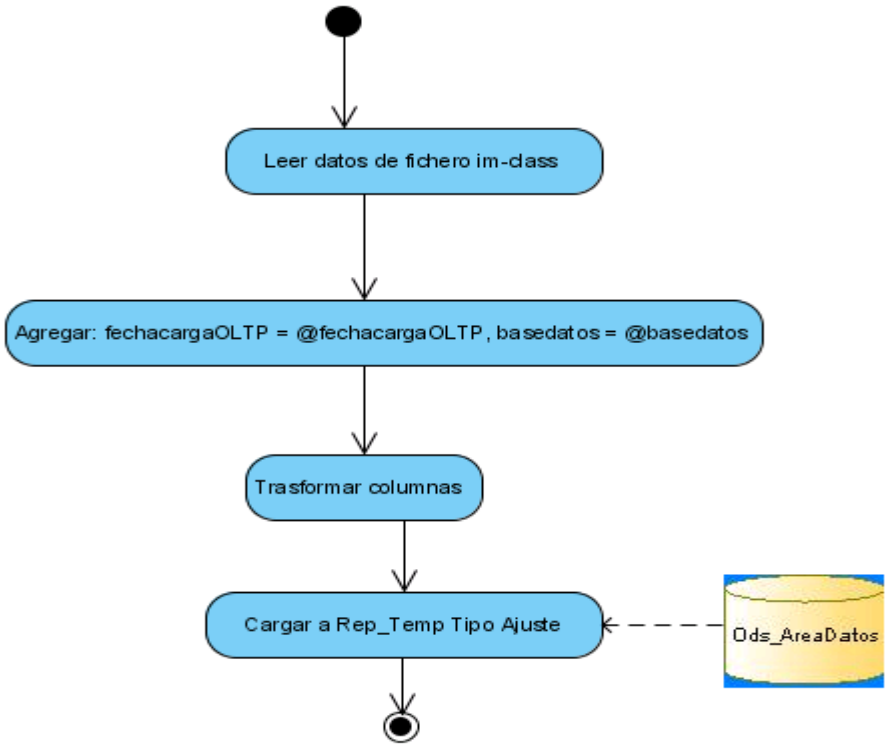
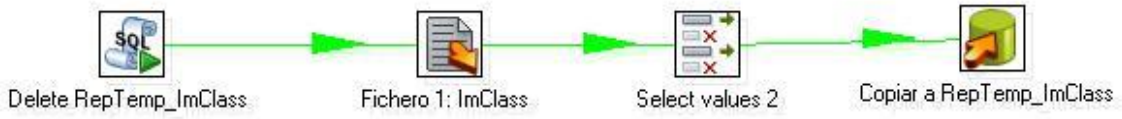


Fig. 50. Arquitectura del hecho Ajuste precio.

Primeramente se borra todo el contenido almacenado en la tabla temporal RepTem_ImClass, con el objetivo de que la tabla quede limpia para almacenar los nuevos datos, logrando que esos sean los únicos que estén almacenados y listos para realizar las transformaciones necesarias. Luego se leen los datos del archivo de origen im-class donde los datos se encuentran crudos, se renombran los atributos, se filtran y añaden las contantes BaseDatos y status-code, cargando luego en la tabla temporal RepTemp_TipoAjuste donde estarían los datos listos para ser almacenados:



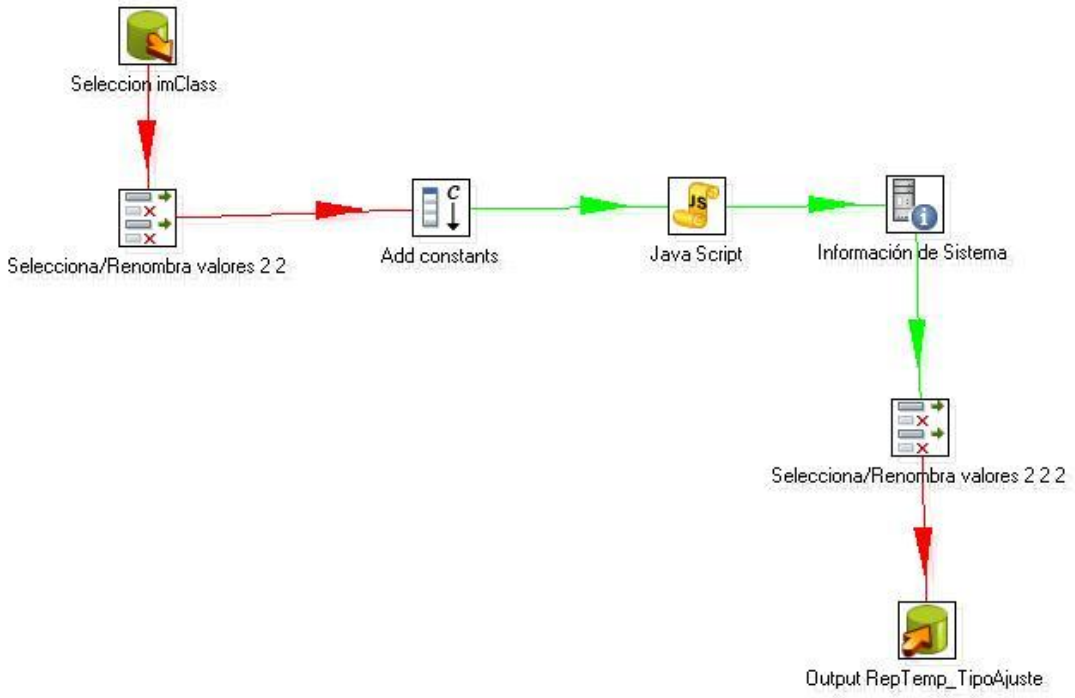


Fig. 51. Proceso ETL del hecho Ajuste precio.

Quedando los datos almacenados en la tabla temporal RepTemp_TipoAjuste de la siguiente forma:

Field Name	Field Type
TipoAj_IdOLTP	varchar(15)
TipoAj_Descripcion	varchar(60)
MT_BanderaActivo	varchar(20)
MT_BaseDatos	varchar(50)
MT_FechaCargaOLTP	timestamp

Fig. 52. Tabla temporal del hecho Ajuste precio cargada.

Carga de los datos al almacén de datos operacional CIMEX.

Luego del completamiento de la extracción, transformación, y carga de los atributos a las tablas temporales, se procede a realizar la carga de los datos al almacén de datos operacional del CIMEX, ejecutando la transferencia del conjunto de transformaciones realizadas al destino final, quedando la información lista para ser consultada. Se decidió confeccionar tablas temporales con características

similares a las tablas finales del ODS donde se almacenaron momentáneamente los resultados de cada transformación realizada a los diferentes hechos y dimensiones existentes, con el fin de tener las tablas listas para almacenar la información y garantizar un menor esfuerzo en la búsqueda de los datos. Para cargar los datos al almacén, se extrajeron los datos de las tablas temporales, según el hecho o dimensión, pues ellas contienen almacenados los datos extraídos desde el origen con las trasformaciones requeridas para cada hecho o dimensión que se trate. Luego para lograr que los datos se encuentren lo más limpios posibles a la hora de integrar se le agregan constantes y renombran algunos valores, además de realizar una validación de los valores nulos. Para finalmente ser cargados o actualizados en la tabla respectiva al nombre del hecho o dimensión, donde los datos se encuentran listos para ser consultados, y ofrecer información fiable y entendible. (Ver Anexo 4)

A grandes rasgos para la carga de los datos de las tablas temporales al ODS se tuvo en cuenta las características de cada tabla temporal que el negocio requiera, además, en dependencia de la dimensión o el hecho que se trató, se realizaron nuevas transformaciones logrando que la información almacenada en las tablas finales del ODS sea la que se utilice por los usuarios finales, donde los datos se encuentran claros y precisos. Potenciando que las consultas se logren en el menor tiempo posible y los gerentes del CIMEX pueden tomar decisiones basadas en esta información.

Conclusiones del capítulo 2

En este capítulo se explicó el proceso del negocio mayorista que se lleva a cabo actualmente en el CIMEX, a partir de esto, se realizó el proceso ETL convenido, extrayendo los datos desde un FTP, limpiándolos y transformándoles, prestándole especial detalle a los hechos y dimensiones involucrados en los procesos compra, venta e inventario, especificando los aspectos más importantes que se tuvieron en cuenta en el paso de transformación de los datos. Realizando luego la carga de los datos al almacén de datos operacional desde las tablas temporales donde los datos se encontraban limpios e integrados. Quedando en este capítulo ejemplificado todo el proceso ETL especificado por CIMEX.

CAPÍTULO 3: VALORACIÓN DE LOS RESULTADOS OBTENIDOS.

Una vez que se han desplegado todas las estructuras de la capa ETL, se debe dar paso a la validación de la solución, verificando que se cumplan todos los requerimientos que hacen confiable la información cargada en el almacén de datos operacional, la cual se brinda al usuario funcional del sistema. Por ello en este capítulo se presentan las pruebas y validación que se le realizó a la capa de integración de datos, logrando un acercamiento al grado de conformidad y aceptación por parte de los usuarios de los datos cargados en el almacén.

Calidad de Datos.

En cada operación diaria del negocio de la corporación CIMEX se toman decisiones, la mayor parte de esas decisiones surgen a partir de información o cierta intuición de los empleados, mientras más se usa la información y menos la intuición, se puede estar más seguro de haber tomado una buena decisión. Ahora con la ayuda del almacén de datos operacional, las decisiones se encuentran sustentadas en información fiable, donde la calidad de los datos produce a favor de una buena administración de los recursos humanos, financieros y materiales de la corporación CIMEX, creando valor y excelencia operacional, incrementando los costos, y logrando especial impacto en los niveles comerciales. Para afirmar con razones sustentadas en la teoría, se explican las características que cumplen los datos con calidad:

- Exactitud: Mide el grado en que la información refleja lo que está pasando en el negocio.
- Totalidad: Medición que refleje el grado en que las bases de datos cuentan con toda la información critica para el negocio.
- Oportunidad: Medición de que la información esté disponible cuando se requiere para tomar una decisión.
- Relevancia: Que la información le sirva a la persona que se la estas proporcionando.
- Nivel de detalle: Que la información tenga el nivel de detalle requerido, dependiendo del nivel organizacional y al tipo de decisión al cual este destinada la información.
- Consistencia: Que la información sea la misma en todas las áreas o sistemas utilizados por la corporación.

Estas características describen como se encuentran los datos, luego de haber pasado por el proceso ETL explicado en el capítulo anterior. Donde no solo se provee información de calidad, sino apoyo al CIMEX a

realizar un buen uso de la información para mejorar el aprovechamiento de los recursos y las operaciones del negocio. Para poder argumentar la certera calidad de los datos cargados en el almacén de datos operacional del CIMEX, se utiliza la estrategia de Kimball[31], que asegura una fuerte calidad de los datos, mediante los siguientes pasos:

- Definir las reglas de calidad de los datos

El líder del grupo ETL y los especialistas de calidad de datos son responsables conjuntamente de definir las reglas de calidad de los datos. Tienen la tarea de analizar los datos, que en el caso del proceso en cuestión, se realizó a través del perfilado, utilizando la herramienta Talend Open Profiler, y documentando los defectos identificados. Este análisis revela fallos en los datos de los sistemas fuentes y da la opción a los administradores o gerentes ETL de corregirlos para evitar futuros defectos en los datos. La opción de limpieza de datos elegida por los desarrolladores fue transformarlos en el proceso ETL.

- Documentar los defectos de los datos

Se realizó un documento con todos los defectos de los datos, que contiene los detalles de cada problema de limpieza y los de resolución. Realizando un plan de seguimiento.

- Prueba. Validación del control de las calidad de los datos

Para la validación del control de la calidad de los datos, se genera un informe de auditoría, asegurándose de incluir preguntas que aprueben la calidad de los datos en el almacén de datos operacional, corroborando que se ha trabajado como se espera en la transformación y limpieza de los datos. Con el fin que se demuestre de forma trasparente este importante paso en el proceso ETL, se amplía en el próximo epígrafe la estrategia que se utilizó para la validación de la calidad de los datos.

Estrategia de validación de calidad.

La estrategia a utilizar para validar la calidad de los datos cargados en el almacén del CIMEX, como se ha escrito en el epígrafe anterior, será la de Kimball. Como primer paso para la validación de calidad del trabajo efectuado, se crean los artefactos para la detección y documentación de errores: Tabla de hecho CASO DE ERROR, que captura los casos de error de la limpieza ocurridos a nivel de registros individuales en las tablas del proceso ETL, obviamente, estos casos de error pudieran no ocurrir

[31] Ralph Kimball, fundador de Kimball Group, una compañía que fomenta el modelado tridimensional y almacenes de datos críticas.

simplemente a nivel de registros individuales, por lo que se asocia a los indicadores de calidad de los datos con las tablas de hechos de los usuarios finales. Otro artefacto que permite la detección de errores es Dimensión Auditoria, que obtiene el contexto específico de calidad de datos de los registros individuales de las tablas de hecho. Realmente, esto no produce una enorme proliferación de registros en la dimensión auditoría, porque el propósito de este artefacto es describir cada caso general detectado con relación a la calidad de los datos.

Luego se debe aplicar un plan de pruebas, que incluye pruebas de unidad llevada a cabo por el desarrollador ETL y el analista de sistemas en el entorno de desarrollo, además se gestionan pruebas de aseguramiento de la calidad[32], esta prueba se ejecuta por un grupo aparte del equipo de desarrollo, en un entorno similar al de producción, este ambiente es creado y controlado por el administrador de BD y los miembros del equipo de control de calidad, este entorno se utilizará para asegurar que todos los procesos ETL se están realizando como se esperaba conociendo las reglas del negocio y el tiempo límite. Y como última prueba a desarrollar esta la de aceptación del usuario, donde se produce por un grupo de usuarios en un ambiente controlado y creado a partir de la garantía de la calidad. Esta prueba beneficia al equipo dejando a usuarios tener una apariencia de participación activa en los datos para asegurar que los procesos corren como se esperaba. Este plan de pruebas no se efectuó completo, debido que aun el almacén no cuenta con un entorno de desarrollo, por lo que no es posible implementar las pruebas de unidad. Mientras que para las pruebas de aseguramiento de la calidad y de aceptación del usuario, se creó como alternativa un ambiente idóneo en el centro de almacenamiento y análisis de datos, donde se desarrolló el proceso ETL, protagonizado por el grupo ETL y una representación de los clientes del CIMEX, aplicando pruebas a la carga de los datos en el almacén, validando los controles de calidad de datos, las reglas del negocio, y asegurando que el proceso ETL logró datos limpios e integrados, que aportan información al usuario. Obteniendo como resultado concreto un informe que especifica las características de los datos cargados, donde estos son correctos, lo que significa que existe correspondencia entre lo que el dato describe y lo que realmente contiene, son no ambiguos, o sea, que el valor que contienen los datos posee un solo significado, además se muestran consistentes y coherentes usando una convención estándar que determina su significado. En las pruebas efectuadas, todos los datos corroborados son completos, o sea no son nulos.

[32] Del inglés Quality Assurance (QA), consiste en tener y seguir un conjunto de acciones planificadas y sistemáticas, estas acciones deben ser demostrables para proporcionar la confianza adecuada a los clientes.

Además cada dimensión o hecho cargado contiene screens[33], usado para capturar los casos de error en la calidad de los datos, y definiendo las acciones que se deben ejecutar en cada caso. Logrando cumplir los cuatro factores o prioridades que constituyen los objetivos del proceso ETL para garantizar la calidad de los datos:

- Exhaustivo: El subproceso de limpieza y transformación está condicionado a ser exhaustivo en la detección, corrección, y documentación de la calidad de la información. Los usuarios finales deben percibir al almacén de datos como una fuente confiable de información.
- Veloz: Todo el proceso ETL a de procesar volúmenes de información cada vez más grandes en espacios de tiempo más pequeños.
- Correctivo: La corrección de los datos lo más cercano posible a la fuente, es la vía más factible para mejorar el valor de la información del CIMEX.
- Transparente: El almacén de datos operacional saca a la luz los defectos de los sistemas y prácticas de negocio que perjudican la calidad de los datos.

[33] La palabra screen es usada para referirse a los filtros o procedimientos que son aplicados para medir o chequear la calidad de los datos. Sinónimo de data-quality screen o data-quality check.

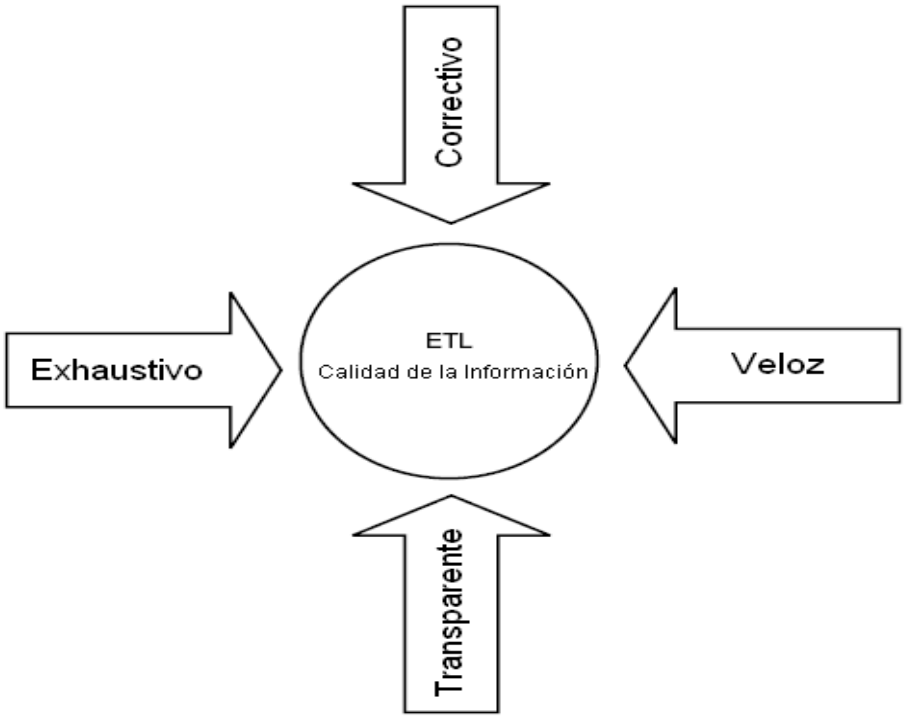


Fig. 53. Factores de influencia.

Aunque para que estos factores se logren en la práctica real debe existir un equilibrio, pues evidentemente, es imposible para el proceso ETL hacer frente, en términos absolutos, a todos estos factores simultáneamente. Lograr un equilibrio adecuado es esencial, obteniendo límites razonables entre los tipos de defectos que son corregidos frente a los que son detectados, para producir una auditoria fácil de utilizar que documente las modificaciones, estandarizaciones, las normas asumidas de la detección de errores y los componentes de reingeniería de datos.

Auditoría a los datos.

Debido al importante papel que juega el almacén de datos operacional en la toma de decisiones, es necesario garantizar que la información suministrada no cause desconfianza en los usuarios. Tener conocimiento del origen o procedencia de la información va a resultar de utilidad a la hora de enfrentar las posibles auditorías sobre el sistema o simplemente cuando al consultar los datos, se sienta curiosidad sobre las suposiciones subyacentes. Algunos de los indicadores de calidad que se pueden añadir para dar garantía de la información son el número total de elementos en la entrada, la cantidad de datos corruptos

o no válidos, el total de transformados y los cargados. Puede agregarse, además, la fecha y hora de los distintos pasos durante la extracción. Debido a esto, la información que se audita es la que se encuentra en las tablas de hechos, pues es la asociada con las operaciones que tienen lugar en la corporación CIMEX. De acuerdo con el modelo seguido, en estas tablas se encuentran las llaves de las dimensiones que participan en el proceso, además de las medidas específicas que caracterizan las acciones con las cuales se relacionan dichas dimensiones. Por ello y para asociar los datos indicadores de la calidad con el usuario final se construye una dimensión a la cual se va a llamar dimensión auditoría. La cual describe de forma completa la calidad de los datos de un registro de la tabla de hechos, la dimensión auditoría se encuentra, literalmente, adjunta a cada hecho en el registro de datos del almacén de captura y procesamiento de ETL, es un importante hito de tiempo y los errores significativos y su frecuencia de aparición son registros de auditoría creado como el último paso del tratamiento de limpieza, los cuales deben contener una descripción de las correcciones y los cambios que se han aplicado para el registro. Así en dicha dimensión van a estar ubicados los indicadores inmediatos relacionados con el hecho en específico. (Ver Anexo 5). La auditoría resulta interesante al poder aprovechar las riquezas del modelo dimensional con tales propósitos, pero sin duda la parte más trabajosa es obtener la información de auditoría en tiempo de ejecución. La mayoría de los indicadores deben construirse dentro del mismo flujo durante el proceso ETL, lo que implica manejar los errores o eventos a medida que se forma cada fila.

Conclusiones del capítulo 3.

En este capítulo se realizó un plan de prueba, que se puso en práctica para constatar la calidad del proceso ETL implementado. De ahí se puede llegar a la conclusión de que los metadatos, la dimensión auditoria y las pruebas efectuadas sobre los datos, son un arma poderosa para el sistema, pues sirven de ayuda durante el proceso ETL, además mantienen el control sobre la información que se maneja. Por lo tanto, se analizó los distintos tipos de metadatos del sistema, así como los registros obtenidos de la tabla caso error y la dimensión auditoria, sugiriendo tomar como los de importancia acentuada los asociados a las fuentes, los procesos de extracción, y las transformaciones sobre los datos. De este plan de pruebas, también se obtuvo artefactos que dan al traste con la aceptación del usuario y aseguramiento de la calidad, como un proceso continuo, enfocado al cliente y en correspondencia con los criterios de calidad a mantener en los datos.

CONCLUSIONES GENERALES

Hoy en día el mundo se mueve vertiginosamente en el paso de una tecnología a otra y en la rapidez requerida a la hora de dar respuesta a determinados problemas. Por ello para la implementación del proceso ETL para CIMEX, se utilizó la metodología orientada por el prestigioso experto Ralhp Kimball, considerado como uno de los “padres” de los almacenes de datos. Así al concluir este trabajo se puede plantear que se ha cumplido con los objetivos del mismo y con las tareas de la investigación propuestas. Planteando que se trata de un proceso de desarrollo bastante complejo y con un amplio nivel de detalle, el cual integra mucha información tanto de la entidad como de las herramientas con las que se trabajan, por lo que ha llevado varios meses de trabajo intenso, en consecuencia, los resultados son alentadores.

A partir de la fundamentación teórica se enriqueció y abundó en el conocimiento sobre todo lo que respecta y se relaciona con el proceso de Extracción, Transformación y Carga, así como toda una investigación sobre las principales herramientas a utilizar, incluyendo entrevistas a especialistas en el tema, todo lo cual sirvió de base para una futura construcción del proceso en detalle. De esta forma mediante el diseño y la implementación de este proceso ETL, se podrá proveer al almacén de datos operacional del CIMEX de datos eficientes, limpios y consistentes, facilitando el dinámico acceso a la información con la rapidez requerida en los reportes y consultas, donde la información de dicha entidad estará centralizada, ordenada e integrada, posibilitando su acceso simultáneo por varias personas. Finalmente y luego de constatar las grandes potencialidades y beneficios que brinda este proceso ETL, los clientes se encuentran ampliamente satisfechos con el resultado logrado.

RECOMENDACIONES

Con el propósito de enriquecer la propuesta realizada en este trabajo, se sugiere:

- Incentivar en la Universidad de las Ciencias Informáticas, concretamente en la facultad 3, investigaciones referentes a la integración de datos, y al proceso de extracción, transformación y carga.
- Realizar un profundo estudio acerca de técnicas de optimización, que puedan ser aplicadas al proceso de extracción, transformación y carga desarrollado.
- Establecer una comparación entre las herramientas ETL evaluadas, desarrollando un mismo proceso de carga, y confirmando en la práctica el funcionamiento de cada una.
- Realizar un estudio comparativo de las nuevas necesidades de los analistas y el nivel de satisfacción que se alcanzó, de manera que pueda comprobarse la validez de los sistemas de información en la corporación CIMEX.
- Incluir para próximas ediciones de procesos ETL todas las actividades de negocio, tanto mayoristas como minoristas, de la Corporación CIMEX.

BIBLIOGRAFÍA

Celko, J. (1996). Celko on SQL: Natural, Artificial and Surrogate Keys Explained. Revista DBMS Magazine Online United Business Media LLC

CIMEX. (2009). "Portal Corporación CIMEX." Retrieved marzo 19, 2009, from http://www.cimexweb.com/transformer.asp?event=PortalHome&inc=256&lang=sp&bw=ie.

Headquarters, C. (2009). "Visual Paradigm for UML." Retrieved abril 01, 2009, from http://www.visual-paradigm.com/product/vpuml/.

Kimball, R. (1996). The Data Warehouse Toolkit, WILEY PUBLICHING, INC.

Kimball, R. (2004). The Data Warehouse ETL Toolkit. Practical Techniques for extracting, cleaning, conforming, and delivering data, WILEY PUBLICHING, INC.

License, C. C. P. (2008). Talend Open Studio Solutions User Guide.

Pentaho, C. (2007). Pentaho Data Integration Spoon 3.0 User Guide.

Pentaho, C. (2009). "Pentaho Open Source Business Intelligence: Kettle Project." Retrieved marzo 17, 2009, from http://kettle.pentaho.org/.

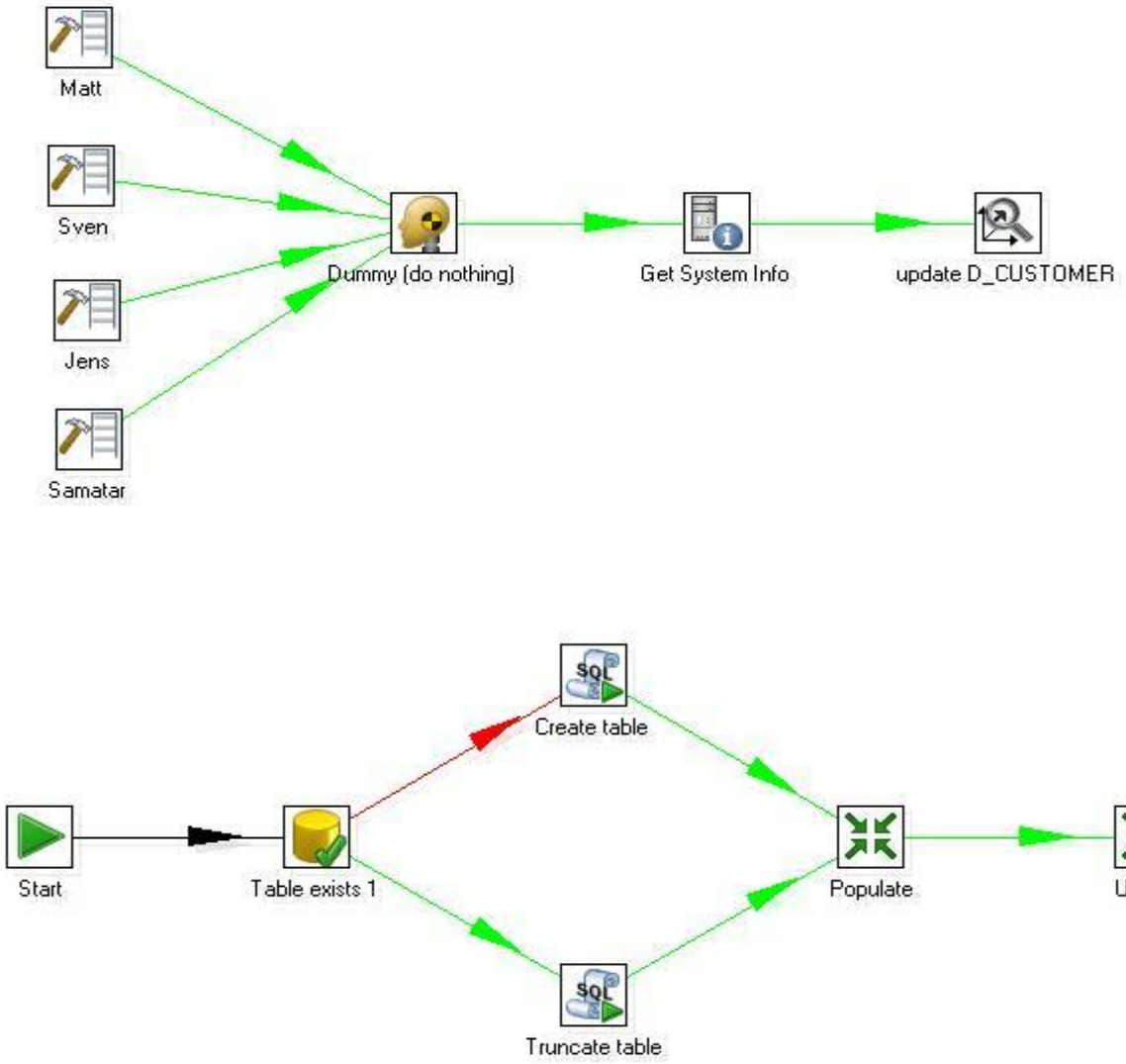
Ross, M. and R. Kimball (2005). Slowly Changing Dimensions Are Not Always as Easy as 1, 2, and 3. Revista DBMS Magazine Online United Business Media LLC

Talend, C. (2009). "Talend Open Data Solutions." Retrieved marzo 05, 2009, from http://es.talend.com/index.php.

Vidot, O. R. (2008). Almacén de Datos Operacionales: propuesta de formalización del proceso de desarrollo. Ciudad Habana, Universidad de la Habana.

ANEXOS

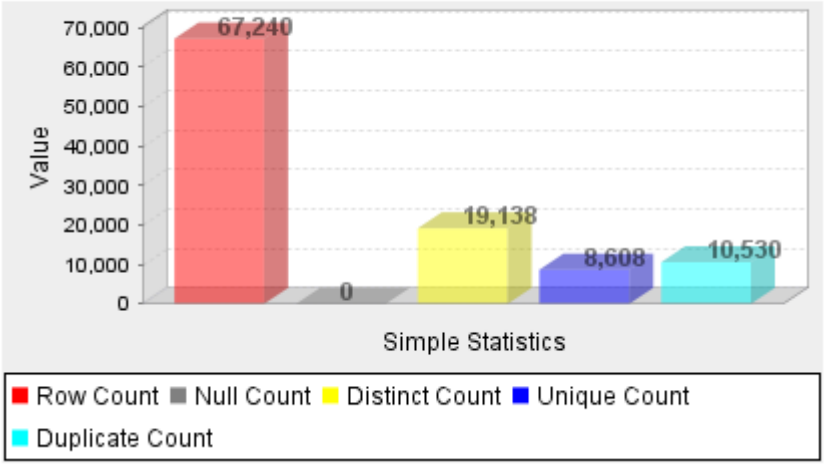
Anexo 1. Ejemplo de dimensión lentamente cambiante. Dimensión Cliente.



Anexo 2. Ejemplos de perfilado de datos.

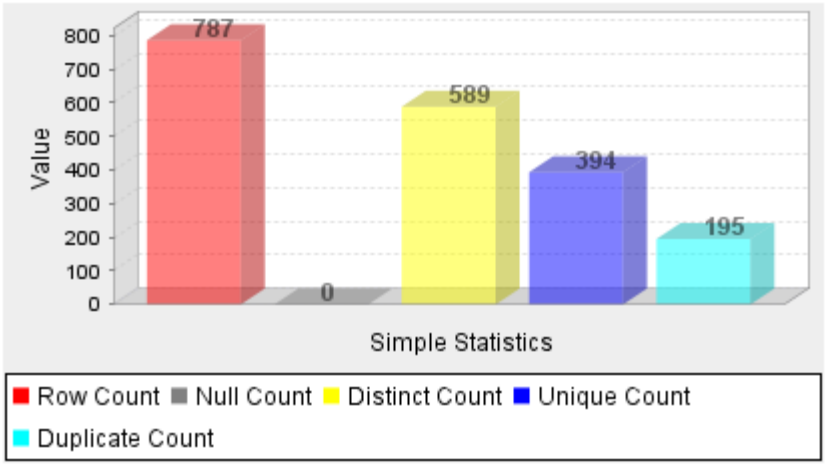
Atributo ov_trasnumero Tabla Venta.

Column: ov_transnumero



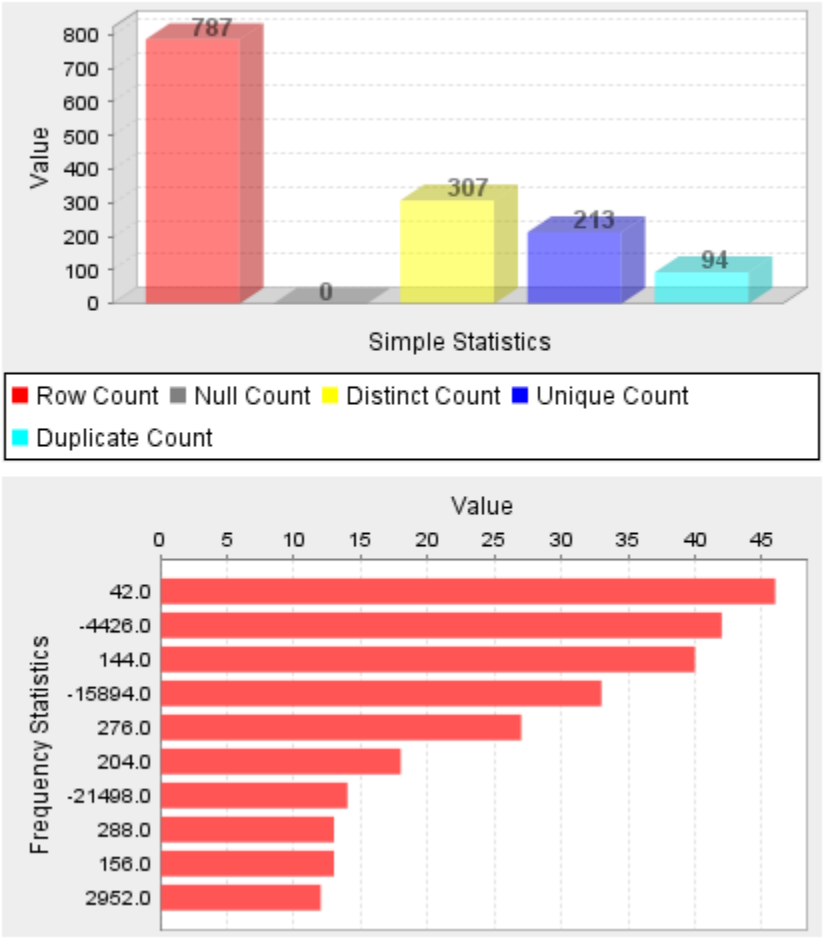
Atributo prod_recno Tabla Compra.

Column: prod_recno



Atributo factor_costo Tabla Compra.

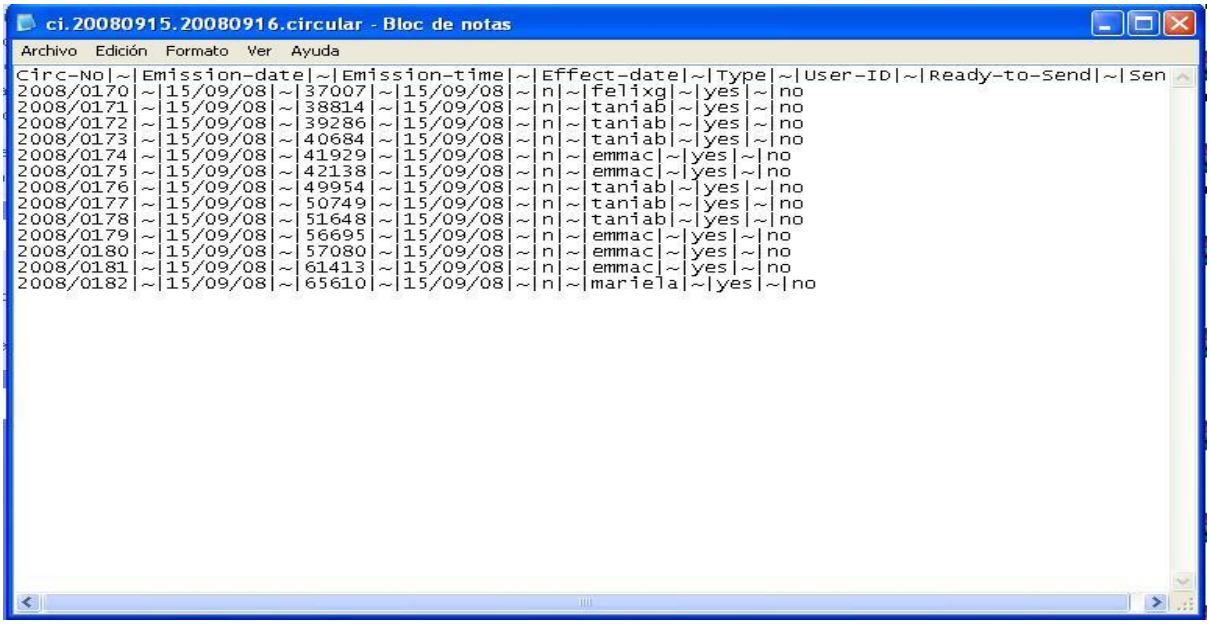
Column: oc_factorc



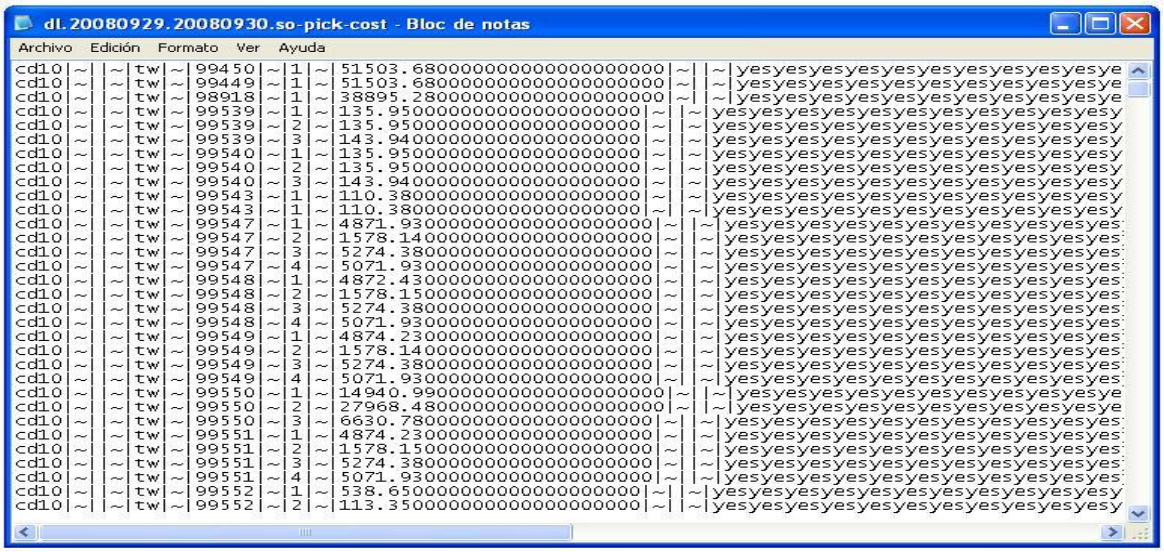
Anexo 3. Ejemplos de los ficheros extraídos.

Discim.

Circulares.



Divlog.



Anexo 4. Ejemplo de carga al almacén de datos operacional. Dimensión Proveedor.

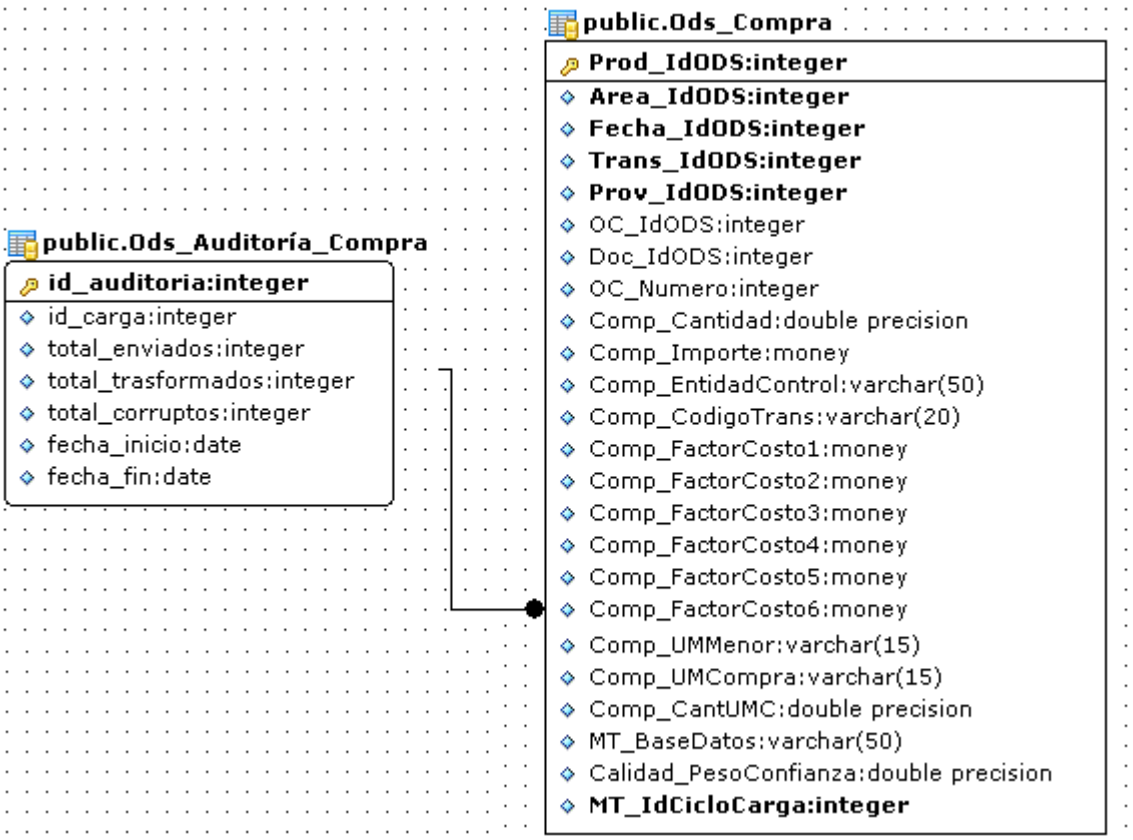
Carga de la dimensión Proveedor

Tratamiento a los valores nulos en la dimensión Proveedor

Script 1

```
//Cambiar null por cadena de texto Desconocido


if(prov_nombre.getString()==null)
prov_nombre.setValue("Desconocido");

if(prov_codigopanamericano.getString()==null)
prov_codigopanamericano.setValue("Desconocido");

if(prov_ciudad.getString()==null)
prov_ciudad.setValue("Desconocido");

if(prov_provincia.getString()==null)
prov_provincia.setValue("Desconocido");

if(prov_codigopais.getString()==null)
prov_codigopais.setValue("Desconocido");

if(prov_pais.getString()==null)
prov_pais.setValue("Desconocido");

if(prov_moneda.getString()==null)
prov_moneda.setValue("Desconocido");
```

Anexo 5. Ejemplo de dimensión auditoría. Hecho Compra.



GLOSARIO DE TÉRMINOS

Almacén de datos (Data Warehouse): Almacena datos transaccionales, específicamente estructurados para consultas y análisis.

Almacén de datos operacionales (Operational Data Store): Almacén de información detallada orientado a temas, integrado, aumentado con frecuencia dentro del almacén de datos de una empresa.

Tabla de hechos (Fact Tables): Tabla central de un esquema dimensional (en estrella[34] o en copo de nieve[35]) que contiene los valores de las medidas de negocio.

Tabla de Dimensiones (Dimension Table): Tabla que acompaña a la tabla de hechos y determina los parámetros (dimensiones) de los que dependen los hechos registrados.

Llaves Sustitutas (Surrogate Keys): Llave generada artificialmente que sustituye el campo llave natural de la dimensión.

Mercado de Datos (Data Mart): Base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento.

Minería de Datos (Data Mining): Conjunto de técnicas y tecnologías que permiten explorar grandes bases de datos, de manera automática o semiautomática, con el objetivo de encontrar patrones repetitivos, tendencias o reglas que expliquen el comportamiento de los datos en un determinado contexto.

Perfilado de Datos (Data Profiling): Proceso de examen de los datos disponibles en una fuente de datos que facilita la recogida de estadísticas e información acerca de los datos.

[34] Modelo de datos que tiene una tabla de hechos que contiene los datos para el análisis, rodeada de las tablas de dimensiones.

[35] Estructura más compleja que el esquema en estrella. Se utiliza cuando alguna de las dimensiones se implementa con más de una tabla de datos.

Integridad Referencial (Referential Integrity): Sistema de reglas que utilizan la mayoría de las bases de datos relacionales para asegurarse que los registros de tablas relacionadas son válidos y que no se borren o cambien datos relacionados de forma accidental produciendo errores de integridad.

Registro de base de datos (Database Log): Conjunto de campos que contienen los datos que pertenecen a una misma repetición de entidad. Representa un ítem único de datos implícitamente estructurados en una tabla.