Facultad 8

Centro de Tecnologías de Gestión y Almacenamiento de Datos (DATEC)

Título: Implementación de los procesos de Integración de datos de los sistemas SAIME y INTTT para el Cuerpo de Investigaciones Científicas, Penales y Criminalísticas de la República Bolivariana de Venezuela.

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autores:

Yunier Rodríguez Lucas

Héctor Alfredo Zúñiga Baldemira

Tutores:

Ing. Osniel Hernández Calvo

Ing. Yonelbys Iznaga González

Ciudad de La Habana, junio de 2010

"Año 52 de la Revolución"

Declaración de autoría

Declaramos que somos los únicos autores del trabajo "*Implementación de los procesos de Integración de datos de los sistemas SAIME y INTTT para el Cuerpo de Investigaciones Científicas, Penales y Criminalísticas de la República Bolivariana de Venezuela*" y autorizamos a la Universidad de las Ciencias Informáticas a hacer uso del mismo en su beneficio.

Para que así conste firmamos la presente a los ____ días del mes de _____ del año 2010.

Autores:

_____

Héctor Alfredo Zúñiga Baldemira

_____

Yunier Rodríguez Lucas

Tutores:

_____

Ing. Yonelbys Iznaga González

_____

Ing. Osniel Hernández Calvo

Dedicatoria

A mis padres por ser lo más grande que tengo en el mundo, por su apoyo y amor.

A mi bisabuela porque la extraño mucho y ocupa un lugar muy especial en mi corazón.

A mis abuelas porque siempre han estado pendiente de mí y por ser mis segundas madres.

A todos los que hicieron posible la realización de este sueño, a los que me brindaron su apoyo y ayuda, a los que me aconsejaron y a los que me enseñaron durante este largo camino.

Aquí les va mi mayor esfuerzo y abnegación. Muchas gracias.

Héctor Alfredo Zúñiga Baldemira.

Dedico este trabajo a todos los que de una forma u otra han estado ahí para lo que haga falta y de cierta manera han hecho que este sueño se haga realidad, especialmente a mis padres, a Rosy y Yosvier.

luk@s.

Agradecimientos

Lucas

A mi señora madre Marlen, todo lo que soy es por ti, has sido capaz de lograr dos profesionales con mucho sacrificio, valor, amor y regaños, eres mi razón de ser.

A mis hermanos y en especial a Yosvier que ha sido para mí, amigo, hermano y padre. Gran parte de lo que hoy soy es gracias a ti, has sido capaz de guiarme por el camino correcto.

A mi padre Ramón que a pesar de la distancia no has sido el padre que todos quisieran tener, pero si el que la mayoría deseara.

A mi novia Rosy que en estos 2 años me has ayudado mucho con esa paciencia, sencillez, amor incondicional y sincero que cada día me entregas. Parte de este resultado también es tuyo, gracias por demostrarme que el amor existe.

A mis tíos Maritza, Fernando, María, Reicedo, Lourdes, Milagros, Ángela y Migue por estar siempre cerca. Especialmente a Fernando que ha sido un padre para mí y a Reicedo que con sus consejos y experiencia ha hecho de mí una persona mejor.

A mis abuelos Pepa, Rubén y Victoria autores de esta maravillosa familia. Ojalá y en un futuro yo pueda hacer una familia como ustedes la han formado a pesar de los obstáculos que la vida les ha puesto.

A mis primos que han sido mucho más que primos, especialmente a Yeiner que siempre me ha entendido y la quien considero hermano.

A Nelsy y Luis por ser mi única familia aquí en la capital, por estar pendientes siempre de mis estudios y demás, por ayudarme a matar el gorrión cada vez que iba por allá.

A Carlos y Mairela por matarme el hambre y las ganas de ver a mi familia unas cuantas veces y por acogerme en su casa y llegar a ser como uno más de ustedes.

A mis amigos de siempre: Albin, Pikiri, Yoe, Coto, Felipe, Yima, Yadira, Ana, Liu 1 y 2, el papi de Moa, Dole, Lary, Maiquel. En este equipo están las personas que me han ayudado y han sido capaces de aguantarme por tantos años. Aquí tengo que destacar el papel de Albito, el cual ha sido más que un amigo para mí, el no sabe cuánto lo quiero y la cantidad de salves que me ha tirado.

A los de la vieja guardia: Dayana, Meilyn, Elvin, Goro, Isa. A ustedes que en algún momento nos vimos medio perdidos y fuimos capaces de recuperarnos y seguir adelante para cumplir nuestro objetivo.

A los del 8105, que fueron las que me ayudaron a dar mis primeros pasos en la UCI y con los cuales nació una fuerte relación. En este grupo resalta el Humbe, que desde el principio nos llevamos bastante bien y llegamos a ser como hermanos.

A los del 8202 gracias por acoger a este muchacho y brindarle todo el apoyo cuando pensaba que su vida universitaria acababa. Después que los conocí me di cuenta que repetir un año no es tan malo. Aquí un nombre sobresale y es el de mi colega y hermano Denis.

Agradecimientos

A los del 8402 pues esta es mi nueva y joven familia, aquí conocí personas imposibles de olvidar, especialmente al Nigro y a Pedruco, los cuales me han aconsejado y apoyado siempre y a mi compañero de tesis.

A mis tutores Osniel y Yonelvis aleas el Letal y el Mejor, gracias a ustedes que nos soportaron todo este año y nos ayudaron en todo momento para poder lograr nuestro sueño.

Gracias a todos por existir.

luk@s.

Agradecimientos

Héctor

A mi mamita por guiarme siempre hacia el camino correcto, por apoyarme, por sus consejos, su comprensión y por ser mi mejor amiga. Por brindarme su confianza, por formarme con todo su esfuerzo y sacrificio, por regalarme una hermanita tan especial, por querer darnos siempre lo mejor, por ser tan buena persona y la mejor madre del mundo.

A mi papá por nunca darme la espalda, por aconsejarme y ayudarme en todo lo que ha podido, por todo su sacrificio y esfuerzo para ayudar a su familia, porque a pesar de no saber lo que es un padre a dado lo mejor de sí para convertirse en el que todos quisieran tener.

A mi bisabuela Mery (Bolonga) por haberme soportado y malcriado, por ponerme por encima de sus hijos, por quererme y consentirme tanto, por brindarme su amor y cariño, porque a pesar de no poder compartir este momento conmigo sé que esté donde este se siente orgullosa de mí.

A mi abuela Magalys (Chiva) por quererme más que a sus hijos, por educarme y aconsejarme cuando no hice lo correcto, por ser mi amiga y por haberme ayudado a soportar todos estos años sin mi mamá.

A mi abuela Elisa (Mami) por ser mi cómplice, por estar presente cuando necesitaba a mi mamá, por cuidar de mi hermanita, por brindarme toda su atención y tratar siempre de darme lo mejor.

A mi abuelo Héctor por tenerme como su nieto preferido, por hacerme reír tanto, por darme su apoyo incondicional, por aceptar mis decisiones, por todo el amor que me ha brindado y por todas las enseñanzas.

A Sixto que es como si fuera un abuelo. por malcriarnos a mi primita Maricarmen y a mí, por aconsejarme, por ayudar a mi abuela, por su disposición, por formar parte de nuestra familia y ser el abuelo paterno que nunca he tenido.

A mi tía Magdalena (Nana) por compartir tantos momentos conmigo, por haberse convertido en el sustento de mi familia materna, por cuidar a mi hermanita y ayudarme en todo lo que ha podido.

A mi tía María (Tutu) por preocuparse por mi, por quererme como a su hija, por ayudarme y enseñarme a que todo en la vida es posible siempre y cuando uno se lo proponga.

A mi tía Yudi por estar pendiente de mí, por compartir su juventud conmigo y por brindarme su ayuda.

A mi hermanita Elizabeth por pedirle a Dios por mi salud, bienestar y buenos resultados en los estudios, por su dulzura, ternura y amor, por ser una mujercita y haber aguantado tanto estos años lejos de mami.

A mi Diame por brindarme su ayuda y apoyo en los momentos que más me hacían falta, por soportar mis malcriadeces, por quererme tanto jaja, respetarme y cuidarme, por aparecer en esta última etapa de la carrera porque solo me hubiese costado más trabajo, por todos

los momentos bonitos que hemos pasado, por dejarse amar, por brindarme su amor, ternura y pasión, por hacerme tan feliz, a mi chiquilina linda le doy mil gracias por estar aquí a mi lado.

A Febe y Raciel por toda la ayuda que me han dado, por haberme aguantado tanto, por matarme tanta hambre, por sus buenos consejos y trompones, por ser buenas personas y por los momentos que compartimos.

A mis amistades de HLG Margel, Yoyi, Pombo, el Vampi, Aurelio, Johnny, Portilla por ser buenos amigos, por compartir tantos momentos de alegría y por ayudarme en lo que han podido.

A mis viejas amistades de la UCI Leo, Viqui, Franklin, Felipe, Elvin, Pedro, Alex, Michel, Yasim, Jorgito, por todo lo que me han ayudado, por las jodederas que quitan el sueño, por inflar tanto, por los momentos duros que pasamos y los que compartimos, en fin por el Dota.

A mis compañeros de la UCI, el Melón, Yudita, Reynaldo, Raciel, Yisel, Adrián, Yanetsi, Lara y Lucas, por haberme ayudado en diferentes etapas de mi carrera.

A Basulto por ser un hombre con sus pantalones bien puestos, por haberme apoyado cuando muchos me dieron la espalda, por permitirme estar aquí en esta escuela, por sus métodos constructivos, en fin por haberme ayudado en el momento cumbre de mi carrera, si no fuera por él ya no estaría aquí.

A mis tutores por el apoyo que nos han dado, porque a pesar de todo nos tiraron el cabo en el momento que más hacía falta, no por gusto es que le dicen el Mejor y el Letal, son unos salvajes.

Al tribunal por corregirnos en el momento que hizo falta, por darnos las evaluación que nos correspondía en cada corte, por ayudarnos a prepararnos mejor para lograr un buen resultado, en especial a Doris por estar siempre dispuesta a ayudarnos y por convertirse en una amiga más.

A nuestro Comandante en Jefe Fidel Castro por la maravillosa idea de crear esta Universidad de excelencia, a la Revolución Cubana por permitir que tantos jóvenes de Cuba y de otros países obtengan un título universitario sin recibir nada a cambio.

A Dios por escuchar las súplicas de mi familia, por ayudar a mi mamá a estar en Cuba para este gran momento, por poner en mi camino personas tan maravillosas, por ayudarme a vencer todos los obstáculos, por proteger a los míos de todo lo malo, por haberme dado la posibilidad de tener una familia como esta y compartir mi vida con tantas personas que me aprecian y me quieren.

A todos, muchísimas gracias.

Resumen:

La Integración de datos es el proceso de unificación de los datos provenientes de múltiples fuentes. Este proceso está formado por tres subprocesos, uno de ellos es la Extracción, Transformación y Carga que permite a las organizaciones mover datos desde diversas fuentes, reformatearlos, limpiarlos y cargarlos en otras bases de datos. Estas y otras funcionalidades hacen imprescindible este proceso a la hora de integrar datos. En este trabajo, se implementará el proceso de Extracción, Transformación y Carga de los sistemas SAIME e INTTT en un Almacén Operacional de Datos para el Cuerpo de Investigaciones Científicas, Penales y Criminalísticas de la República Bolivariana de Venezuela, el cual posibilitará que los datos estén centralizados en este almacén.

Índice de contenidos

Índice de figuras

Índice de tablas

Introducción

El Ministerio del Poder Popular para Relaciones Interiores y Justicia (MPPRIJ) de la República Bolivariana de Venezuela, en aras de disminuir los altos índices de delincuencia que presenta el país, creó en 1999 el Cuerpo de Investigaciones Científicas, Penales y Criminalísticas (CICPC), institución destinada a garantizar la eficiencia en la investigación del delito mediante su determinación científica, asegurando el ejercicio de la acción penal que conduzca a una sana administración de justicia. Tiene como funciones principales colaborar con los demás órganos de seguridad ciudadana en la creación de centros de prevención del delito, y en la organización de los sistemas de control o bases de datos criminalísticas para compartir la información de los servicios de inteligencia. Una de las formas de lograr que la información se obtenga fiablemente, de manera precisa y en el momento propicio, es mediante los procesos de Integración de datos. Para que esta institución logre una eficaz manipulación de su información es necesario integrar los datos disponibles en los diferentes órganos de seguridad ciudadana. (CICPC, 2009)

Actualmente el CICPC mantiene una necesaria comunicación con los sistemas externos Servicio Administrativo Identificación Migración y Extranjería (SAIME) e Instituto Nacional de Transporte Terrestre (INTTT), pero esta no se realiza con la adecuada eficiencia, seguridad y rapidez que se requiere para lograr la interoperabilidad que necesitan estos sistemas. Estas deficiencias se reflejan en varios aspectos, entre ellos que la transportación de la información se efectúa mediante dispositivos externos o en formato duro, lo cual dificulta la inmediatez y la seguridad de los datos. En ocasiones los estados venezolanos utilizan la vía telefónica para emitir sus reportes a la capital, trayendo consigo posibles errores y problemas de entendimiento, así como la comprensión de los mismos. El correo electrónico también es usado para esta comunicación, donde se envían trámites, expedientes y tablas con datos correspondientes a estos sistemas externos, dependiendo de la seguridad y rapidez de conexión de los servicios de correos que en ocasiones no es óptima.

Basado en estos elementos el **problema investigativo** quedaría resumido en la siguiente interrogante: ¿Cómo lograr la preparación, organización y disponibilidad de los datos de los sistemas SAIME e INTTT a integrar en un Almacén Operacional de Datos, para el Cuerpo de Investigaciones Científicas, Penales y Criminalísticas de la República Bolivariana de Venezuela?

Definiendo como **objeto de estudio:** El proceso de Integración de datos, a partir del análisis de este se establece como **campo de acción:** El proceso de Extracción, Transformación y Carga de datos en un

Almacén Operacional de Datos para Cuerpo de Investigaciones Científicas, Penales y Criminalísticas de la República Bolivariana de Venezuela.

Para que los datos a cargar tengan una preparación, organización y disponibilidad necesaria se propone como **objetivo general:** Implementar el Proceso de Integración de datos de los sistemas SAIME e INTTT en un Almacén Operacional de Datos para el Cuerpo de Investigaciones Científicas, Penales y Criminalísticas de la República Bolivariana de Venezuela.

Centrándose en los siguientes **objetivos específicos:**

- Determinar el estado de los sistemas fuentes y las necesidades de SAIME e INTTT.
- Determinar las necesidades de integración en un Almacén Operacional de Datos para el CICPC.
- Integrar los datos de los sistemas SAIME e INTTT en un Almacén Operacional de Datos para el CICPC.
- Validar los resultados mediante la realización de pruebas.

Durante la investigación se sustenta la siguiente **idea a defender**: Si se realiza el Proceso de Extracción, Transformación y Carga en un Almacén Operacional de Datos, se logrará la preparación, organización y disponibilidad de los datos para el Cuerpo de Investigaciones Científicas, Penales y Criminalísticas de la República Bolivariana de Venezuela.

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

- Realizar los procesos de Perfilado de datos de los sistemas fuentes de SAIME e INTTT para determinar el estado de los mismos.
- Realizar un análisis de las necesidades del Almacén Operacional de datos para el CICPC teniendo en cuenta la definición de reglas de transformación y limpieza de los datos para establecer el mapa lógico de datos.
- Efectuar un análisis de las herramientas para definir la más adecuada para la Integración de datos del Almacén Operacional de Datos del CICPC.
- Analizar el proceso de Integración de datos del Almacén Operacional de Datos para el CICPC.
- Diseñar el proceso de Integración de datos del Almacén Operacional de Datos para el CICPC.
- Implementar el proceso de Integración de datos del Almacén Operacional de Datos para el CICPC.

- Evaluar la eficiencia del proceso de Integración de datos mediante listas de chequeo, la aplicación de casos de pruebas, perfilado de datos y auditoría de datos.

Para darle cumplimiento a las tareas de la investigación se utilizaron los siguientes métodos científicos**:**

Entre los métodos **teóricos** empleados se encuentra el Histórico-Lógico, el cual tiene como principal objetivo estudiar la evolución y desarrollo en el tiempo del objeto de estudio y el Analítico – Sintético, pues posibilita la descomposición de los componentes para estudiar por separado cada elemento y sus relaciones. Dichos métodos fueron utilizados para el estudio del estado del arte de los procesos, soluciones y herramientas de Integración de datos, además de los procesos de negocio que existen en CICPC y el estudio del estado del arte de los sistemas externos SAIME e INTTT de la República Bolivariana de Venezuela. La Modelación es otro método utilizado específicamente para el análisis y diseño del proceso de Integración de datos, permite estudiar y determinar cada uno de los procesos que integran el modelo, unir los datos, definir las relaciones entre los diferentes elementos, modelar los flujos de procesos y su interacción.

Dentro de los **empíricos** está el Experimento, que proporciona un estudio del objeto en el cual se crean condiciones y se adaptan las ya existentes para verificar el modelo creado, este método fue aplicado para la implementación del proceso de Extracción, Transformación y Carga al almacén operacional de datos. La Encuesta como método empírico se utiliza para la evaluación de la eficiencia del proceso de integración así como de la calidad de los datos cargados. La Entrevista es utilizada para obtener información sobre los requerimientos del sistema que necesita CICPC y la Observación como instrumento del investigador presente en toda la investigación.

Para darle solución al problema planteado este trabajo ha sido desglosado en tres capítulos estructurados de la siguiente manera:

Capítulo 1: Fundamentación Teórica. Contiene todos los elementos teóricos que soportan esta investigación, además de un estudio de los distintos procesos de Integración de datos, así como de las herramientas y metodologías más utilizadas en el mundo para la implementación de los procesos de Integración de datos.

Capítulo 2: Implementación del proceso de Integración de datos de SAIME e INTTT para CICPC. Se realizará un análisis de las necesidades de Integración de datos, para así centrarse en la implementación del proceso de Integración de datos para CICPC, utilizando las herramientas seleccionadas.

Capítulo 3: Validación de los resultados obtenidos. Se validan los resultados de la investigación. Se evalúa la eficiencia del proceso de Integración de datos de los sistemas SAIME e INTTT para CICPC de la República Bolivariana de Venezuela.

Capítulo 1: Fundamentación Teórica

1.1 Introducción

En el presente capítulo se abordan los conceptos teóricos que fueron necesarios investigar para la realización de este trabajo, se hace un estudio de los distintos procesos de Integración de datos haciendo énfasis en el proceso ETL, sus distintos subprocesos, y todo lo relacionado a este. Además del estudio de las principales herramientas y metodologías utilizadas.

1.2 Sistemas de información

Los sistemas de almacenamiento de información han tenido un gran auge desde sus inicios, pues se han convertido en una herramienta necesaria para el control y manejo de operaciones comerciales. El cúmulo de información en las empresas y negocios crece constantemente y dicha información en muchas ocasiones está almacenada en distintas fuentes, lo cual constituye un problema. Continuamente se necesita realizar consultas y extraer con mayor rapidez y efectividad la información, provocando que los sistemas creados fueran decayendo ante la necesidad de realizar análisis íntegros de los datos y estas operaciones eran muy costosas, atentando contra el correcto funcionamiento de los sistemas.

Para darles solución a todos estos inconvenientes surgen los sistemas de información que son un *"conjunto de componentes interrelacionados que recolectan (o recuperan), procesan, almacenan y distribuyen información para apoyar la toma de decisiones y el control en una organización"* (Laudon&Laudon, 2008).

En el mundo existen diversos tipos de sistemas de información, los más conocidos son: Sistemas de Ventas y Marketing, Sistemas de Planificación de los Recursos Empresariales y los Sistemas de Apoyo a la Toma de Decisiones (DSS).

Los DSS ayudan a los gerentes a tomar decisiones que son exclusivas, rápidamente cambiantes y no especificadas con anticipación, de carácter poco estructurada. Abordan problemas donde el procedimiento para llegar a una solución podría no estar definido con anterioridad y brindan con frecuencia información de fuentes externas.

Los Data Marts[1], Almacenes de Datos y Almacenes de Datos Operacionales son algunos de los DSS más utilizados en el mundo para brindarles servicios a gerentes, directores y ejecutivos en una empresa.

1.2.1 Almacenes de Datos

Los Almacenes de Datos según Kimball[2], considerado el principal promotor del enfoque dimensional para el diseño de Almacenes de Datos lo definen como *"una copia de los datos transaccionales específicamente estructurada para la consulta y el análisis"*.

Por su parte Inmon[3], considerado por muchos el padre del Data Warehouse, define el mismo como *un conjunto de datos orientados por temas, integrados, variantes en el tiempo y no volátiles, que tienen por objetivo dar soporte a la toma de decisiones.*

Los Almacenes de Datos o Data Warehouse, como también es conocido el término en inglés, constituyen uno de los soportes fundamentales para el proceso de toma de decisiones gerenciales. Entre sus principales funcionalidades está evitar problemas como: el de la pérdida de credibilidad o el prestigio de la organización. Posibilita medir las acciones y los resultados de una mejor forma, además los procesos empresariales pueden ser optimizados, la información incorrecta o no encontrada es eliminada.

De forma general se puede decir que un almacén de datos es una base de datos orientada al análisis de la información contenida en ella, caracterizada por la depuración e integración de la información de diferentes fuentes para después procesarla.

1.2.2 Almacenes de Datos Operacionales (ODS)

Los Almacenes de Datos Operacionales u ODS[4], como se le denomina por sus siglas, surgen como respuesta a las necesidades de constar con un sistema integrador de datos que brinde la información con un alto nivel de detalle operacional.

Kimball e Inmon, definen sus propios conceptos, los cuales son citados a continuación:

[1] Subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones.

[2] Ralph Kimball, Doctor en Filosofía, ha sido uno de los mayores visionarios en la industria del Almacén de Datos desde 1982, actualmente, reconocido conferencista, consultante y profesor.

[3] William Harvey Inmon (1945), experto reconocido mundialmente, es el creador de la llamada Corporate Information Factory.

[4] Del inglés Operational Data Store, es un contenedor de datos activos, es decir, operacionales que ayudan al soporte de decisiones y a la operación.

- Un ODS es un almacén de información detallada orientado a temas, integrado, aumentado con frecuencia, dentro del Almacén de Datos de una empresa. (Kimball, 1997)
- William Inmon plantea que un ODS es una colección de datos orientada a temas, integrada, volátil, actualizada, sólo detallada, que sustenta las necesidades de información reciente, operacional, integrada y colectiva de la organización. (Inmon, 1995)

Ambos coinciden en que estos sistemas de información contienen datos orientados a temas específicos y la información se encuentra en un nivel detallado.

Los ODS están caracterizados por el tipo de consulta que sobre ellos se hace. En este caso el análisis que se realiza es operacional, de forma más detallada, de último momento, para posibilitar que la información sea indivisible. Debido a que los datos almacenados constituyen una fuente de gran riqueza y de fácil restructuración, es más frecuente el acceso de otros sistemas al ODS. Posee una versatilidad casi inigualable en cuanto a su habilidad para adaptarse a cualquier formato de salida (Ramos, 2008).

1.3 Integración de datos

Según Ralph Kimball la Integración de datos es el proceso de unificación de los datos provenientes de múltiples fuentes. En el proceso de integración el principal problema lo constituye la diversidad de los datos, los cuales se encuentran dispersos y generalmente no estandarizados. Esto provoca que la información proveniente de los sistemas externos sea inconsistente y de baja calidad, convirtiendo este proceso en un trabajo costoso y complejo (Kimball, 1996). La integración se puede realizar de formas diferentes, las principales son:

- Replicación de Datos.
- Integración de Información Empresarial.
- Extracción, Transformación y Carga de Datos.

A continuación se explicarán las mismas.

1.3.1 Replicación de Datos

La replicación de datos es el transporte de datos entre dos o más servidores, permitiendo que ciertos datos de la base de datos estén almacenados en más de un sitio, y así aumentar la disponibilidad de éstos y mejorar el rendimiento de las consultas globales (Mustelier, 2009). El mecanismo de integración que se utiliza en esta tecnología es de bases de datos a bases de datos, basado en tablas, lo cual es una

desventaja dada la heterogeneidad de las fuentes de datos. Es un mecanismo de baja complejidad y de bajo costo, siendo además una forma simplificada de ETL.

1.3.2 Integración de Información Empresarial (EII)

Es un mecanismo transparente, optimizado de transformación y de acceso a datos para suministrar una única interfaz a lo largo de los datos de las organizaciones.

"No se basa en integrar bases de datos en sí mismas" (Morgenthal, 2005) pues la información se mantiene en las fuentes de información. En lugar de eso, se desarrolla una interfaz programática para el acceso a los datos que permita recuperarlos. Usualmente el resultado de este método es un sistema de información heterogéneo virtualmente integrado.

Por lo general este tipo de solución consiste en crear un Bróker[5] de tal forma que contengan directorios de bases de datos y que a su vez sirvan de canal de consulta y representación de la información recuperada. Es el mejor método para la Integración de datos a tiempo real, ya que la información es capturada en tiempo real de las fuentes de datos, lo cual implica que las fuentes deben tener una estructura tecnológica sólida y bien establecida. (Morgenthal, 2005)

1.3.3 Extracción, Transformación y Carga de Datos (ETL)

Es la tecnología enfocada a la Integración de datos, tanto por lote como a tiempo real hacia Almacenes de Datos (Kimball, Caserta, 2004). Estos procesos se combinan para extraer datos de bases de datos fuentes, archivos u otro sistema, y colocarlas en bases de datos destino. Los procesos ETL se utilizan para migrar datos de una o más bases de datos a terceros y también para convertir bases de datos de un tipo o formato a otro. Se utilizan además para sincronizar datos desde diversas aplicaciones.

Para lograr los objetivos del CICPC, es necesario utilizar esta tecnología ya que las fuentes de datos que se quieren integrar trabajan con diferentes nomenclaturas de los valores con los que interactúan, por lo que el proceso de integración implica aplicar reglas de transformación para depurar los datos y almacenarlos en un destino de forma tal que la información que se genere tenga preparación, organización y esté disponible para ser consultada en el momento requerido.

[5] Mediador o Intermediario

1.4 Principales características del proceso ETL

El proceso ETL permite a las organizaciones mover datos desde diversas fuentes, reformatearlos, limpiarlos y cargarlos en otras bases de datos. Fortalece los datos para la construcción de bases de datos permanentes dedicadas al análisis o generación de informes, las cuales pueden ser convertidas de un tipo o formato a otro. Es utilizado para migrar datos de una o más bases de datos a terceros. Con este proceso se pueden formar Repositorios de Datos, Mercados de Datos y Almacenes de Datos.

Éstas y otras funcionalidades hacen de este proceso imprescindible a la hora de integrar datos. Sus características más significativas son (Kimball, 2004):

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

1.4.1 Arquitectura ETL

A partir del estudio de las características del proceso ETL se puede exponer que es un proceso complejo, debido a su alto nivel de detalle, el cual se debe regir por su arquitectura para así lograr un buen diseño, la cual consta de los siguientes componentes (Syntel, 2007):

- 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[6] y recuperación de errores.

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

- 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 tres categorías: (Kimball & Caserta, 2004)
    - Metadatos técnicos: se usan a menudo por un personal más técnico, tal como los desarrolladores. Incluye temas como las definiciones de tablas y tipos de datos. Estos objetos son utilizados frecuentemente durante el diseño de la aplicación y el proceso de desarrollo.

      Ejemplos: la definición de la fuente y el destino, sus estructuras de tabla, campos y atributos, la documentación para las derivaciones de auditoría y dependencias.
    - Metadatos del negocio: son críticos para entregar un contexto para un proyecto de integración. Ayudan a definir los términos en el lenguaje cotidiano, sin reparos a la implementación técnica. Por ejemplo, el lenguaje utilizado para describir un cliente y la forma en que se categoriza, a menudo es específico de negocio, y podría diferir entre las divisiones de la compañía.

      Ejemplos: las reglas comerciales, gestión, definiciones comerciales, la terminología de auditoría, glosarios, algoritmos y linaje que utilizan el lenguaje comercial.
    - Metadatos de proceso: se refieren a los metadatos generados y capturados cuando se ejecuta un proceso. Permite que los administradores gestionen su sistema y aseguran que las cosas funcionen sin problemas. Si hay un problema con algún proceso, los metadatos operacionales también ayudan a los administradores a identificar y localizar los problemas.

      Ejemplos: información acerca de la ejecución de las aplicaciones, incluyendo la frecuencia, conteos de registro, un análisis de componente por componente y otras estadísticas con fines de auditoría.

1.5 Subprocesos ETL

El proceso ETL se divide en tres subprocesos fundamentales, que permiten la segmentación y entendimiento de este arduo trabajo, los cuales se exponen a continuación.

1.5.1 Extracción

El proceso de extracción consiste en adquirir los datos desde los sistemas de origen, estas fuentes pueden estar sobre sistemas incompatibles o de hardware diferentes. En este subproceso se convierten los datos a un formato preparado para iniciar el proceso de transformación. Aquí se verifican los datos extraídos, donde se comprueba si los datos cumplen lo que se espera y se adaptan al formato estándar diseñado, de lo contrario los datos son rechazados.

En el proceso de extracción es necesario causar un mínimo impacto en el sistema origen (Leroot, 2009), pues si se necesita extraer muchos datos el sistema origen podría ralentizar o colapsar, provocando que no pueda implementarse con normalidad para su uso cotidiano.

1.5.2 Limpieza y Transformación

En esta fase se aplican una serie de Reglas de Negocios[7] sobre los datos extraídos, con el objetivo de convertirlos en datos aptos para ser cargados (Leroot, 2007). Aquí es necesario lograr una buena calidad de los datos y para ello es necesario el control de los valores válidos, garantizar la coherencia entre los valores, la eliminación de duplicaciones y comprobar que las reglas del negocio no han sido forzadas (Kimball, 1996). Para esta fase los datos deben ser limpiados, pues estos pueden estar sucios e incompletos. 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.

Luego que los datos han sido limpiados se procede a realizar las transformaciones mediante las reglas de transformación que pueden ser: combinar los datos de distintas fuentes, realizar búsqueda de valores en distintas tablas, darle tratamiento a valores nulos, entre otras.

1.5.3 Carga

La fase de carga es el momento cuando los datos, provenientes de la fase anterior, son incluidos en el sistema de destino (Leroot, 2007), dependiendo de los requerimientos de la organización. El principal objetivo de esta fase es lograr que los datos estén listos para ser consultados (Kimball, 1996). 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

[7] Describe las políticas, normas, operaciones, definiciones y restricciones presentes en una organización y que son de vital importancia para alcanzar los objetivos misionales.

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 operación se aplicarán todas las restricciones y triggers[8] que se hayan definido, los cuales contribuyen a que se garantice la calidad de los datos en el proceso ETL.

1.6 Importancia y retos del proceso ETL

Al terminar este proceso ETL los datos serán precisos, completos, creíbles, rigurosos en el tiempo, interpretables, accesibles y con valor añadido. Esto asegura la calidad de los datos, los prepara, conforma, limpia, transforma, y organiza, brindando un conocimiento que facilitará la toma de decisiones, permitiendo una integración consistente de los datos.

En ocasiones el proceso ETL por la importancia que representa suele ser extenso, por lo que es necesario realizar un alto nivel de escaneo, debido a esto durante el desarrollo surgen algunos contratiempos que deben ser refinados. Los retos más comunes son (Hartman, Ramón, 2009):

- El rango de valores o la calidad de los datos de éstos pueden no coincidir con las esperadas de los diseñadores a la hora de especificarse las reglas de validación o transformación, en este caso lo más recomendable seria realizar un examen completo para comprobar la validez de los datos del sistema de origen durante el análisis.
- La escalabilidad incluye la comprensión de los volúmenes de datos que tendrán que ser procesados. El tiempo disponible para realizar la extracción de los sistemas de origen podría cambiar, lo que provocaría que la misma cantidad de datos tendría que ser procesada en menos tiempo.
- Las transformaciones pueden ser muy engorrosas, pues los datos necesitan agregarse, analizarse, computarse, y procesarse estadísticamente.
- El creciente volumen de datos a ser procesados. Para ello el proceso necesitaría una alta conectividad a las aplicaciones en paquete, bases de datos, archivos, servicios Web, entre otros.

Estos son algunos de los retos que enfrenta un equipo ETL. Debido a ello en la actualidad las soluciones de Integración de datos están cada vez más optimizadas para lograr una adecuada calidad empresarial,

[8] 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).

siguiendo de forma especial las siguientes características que son críticas para el diseño, desarrollo, ejecución y mantenimiento de los procesos ETL (Hartman, Ramón, 2009):

- Modelación de procesos orientada al negocio que implica 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.
- Componentes avanzados ETL, incluidas manipulaciones de cadenas, dimensiones lentamente cambiantes[9], soporte para cargas masivas, etc.

1.7 Equipo ETL. Roles y responsabilidades

Es responsabilidad del equipo ETL todo el proceso de la extracción de datos de las fuentes origen del sistema, la limpieza y transformación de los datos y la carga de ellos al almacén destino. Lo ideal sería contar con una persona en cada rol para así lograr un desarrollo más confiable y consistente, pero en muchas ocasiones esto resulta muy difícil, ya que no siempre se cuenta con el personal necesario. Además, teniendo en cuenta la extensión del proyecto y las disímiles situaciones que se pueden presentar a lo largo de la vida del mismo, el equipo tiene que estar preparado para asumir varios roles por especialista. Para lograr el resultado esperado se definen como responsabilidades del equipo ETL las siguientes tareas (Kimball, Caserta, 2004):

- 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.

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

- Realizar el mantenimiento del sistema.

Para la realización de estas tareas y proporcionar una mejor calidad del proceso es necesario que los roles asuman responsabilidades específicas (Kimball, Caserta, 2004):

- 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, 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[10] para el equipo de desarrollo ETL, este arquitecto debe tener una fuerte comprensión de los requerimientos del negocio y de los sistemas fuente.
- 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

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

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.

1.8 Fuentes de datos

Para lograr un exitoso proceso de ETL es necesario tener dominio de las fuentes de datos, identificarlas y estudiarlas para saber sus particularidades. Es necesario conocer también las clasificaciones de las fuentes de información para poder planificar y desarrollar los mecanismos de integración en cada caso, ellas se describen a continuación (Microsoft, 2007):

- **Fuentes Cooperativas:** 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).

- **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**: En este conjunto se encuentran las fuentes que brindan la información a partir de archivos intermedios u otros mecanismos que no implican funcionalidades internas puntuales sobre la base del receptor de la información. Ej. Sistemas Legados, Autónomos, etc.
    - **Fuentes Consultables**: suministran interfaces para consultas (SQL, Servicios Web, etc.).

Tabla 1: Clasificación de las fuentes de datos (Microsoft, 2007).

		Fuente 1	Fuente 2	Fuente 3
NO COOPERATIVA	Snapshot			
	Fuentes Específicas (archivos, etc.)			
	Fuentes Consultables (SQL, WS)			
COOPERATIVA	Fuentes de Replicación			
	Fuentes de Call Back			
	Fuentes de Cambios Internos			

Clasificar las fuentes de datos es un paso importante para definir las características del proceso de integración de las fuentes en cuestión. Luego de identificarlas se llega a la conclusión estas son no cooperativas.

1.9 Protocolos de comunicación

Los protocolos de comunicación son métodos estándares que permiten la comunicación entre diferentes equipos, un conjunto de reglas y procedimientos que deben respetarse para el envío y la recepción de datos a través de una red. Básicamente se tratarán aquellos protocolos que por sus características están implicados en el negocio. Estos protocolos son (Kioskea, 2008):

- HTTP
- SMTP
- FTP

1.9.1 HTTP

El protocolo HTTP es uno de los más importantes, pues se enfoca en la transferencia de hipertexto o HTTP (HyperText Transfer Protocol), utilizado básicamente en cada transacción de la Web. Es un protocolo orientado a transacciones y sigue el esquema petición-respuesta entre un cliente y un servidor. La información transmitida constituye el recurso y se identifica mediante una URL. Los recursos pueden ser archivos, el resultado de la ejecución de un programa, una consulta a una base de datos, la traducción automática de un documento, etcétera.

HTTP es un protocolo sin estado, es decir, que no guarda ninguna información sobre conexiones anteriores. El desarrollo de aplicaciones Web necesita frecuentemente mantener estado. Para esto se usan las cookies, que es información que un servidor puede almacenar en el sistema cliente.

1.9.2 SMTP

Es un protocolo simple de transferencia de correo o Simple Mail Transfer Protocol (SMTP). Constituye un protocolo de red basado en texto utilizado para el intercambio de mensajes de correo electrónico entre computadoras u otros dispositivos (PDA's[11], teléfonos móviles, etcétera.). Está definido en el RFC 2821 y es un estándar oficial de Internet. SMTP se basa en el modelo cliente-servidor, donde un cliente envía un mensaje a uno o varios receptores. La comunicación entre el cliente y el servidor consiste enteramente en líneas de texto compuestas por caracteres ASCII.

1.9.3 FTP

El protocolo de transferencia de archivos o FTP (*File Transfer Protocol*) está basado en la arquitectura cliente-servidor donde un equipo cliente se puede conectar a un servidor para descargar o enviar archivos. Este protocolo está pensado para ofrecer la máxima rapidez en la conexión, no así para la seguridad, ya que todo intercambio de información desde la autenticación en el servidor hasta la transferencia de archivos se realiza en texto plano sin ningún tipo de cifrado, por lo que resulta idóneo la utilización de este protocolo para la Integración de datos de los sistemas externos.

1.10 XML como tecnología de intercambio de datos

El metalenguaje XML (Extensible Markup Language) es un formato derivado de SGML muy simple y flexible, el cual es muy utilizado para el intercambio de datos en la Web. Es un metalenguaje que permite definir la gramática de lenguajes específicos y se propone como un estándar para el intercambio de información estructurada entre diferentes plataformas. Puede usarse en bases de datos, editores de texto, hojas de cálculo, entre otros y permite la compatibilidad entre sistemas para compartir la información de una manera segura, fiable y fácil (Web, 2010).

Según el grupo de desarrollo del sitio oficial de Adobe, entre las ventajas de usar XML sobresalen que después de diseñado y puesto en producción, es posible extender XML con la adición de nuevas etiquetas, de modo que se pueda continuar utilizando sin complicación alguna. Otra ventaja importante es que posibilita la creación de un analizador específico para cada versión de lenguaje XML, lo cual permite el empleo de los analizadores disponibles.

[11] del inglés Personal Digital Assistant (Asistente Digital Personal), es un computador de mano originalmente diseñado como agenda electrónica

Para integrar la tecnología XML en una solución deben tener en cuenta algunos aspectos como la estructura que incluye: prólogo, cuerpo, elementos y atributos. Esta tecnología es muy utilizada debido a su capacidad de recopilar una gran cantidad de información independientemente de la heterogeneidad que presente la misma o la frecuencia con que llegue a modificarse, tomándose en cuenta todos estos aspectos en el momento de seleccionar este lenguaje para la transferencia de datos.

1.11 Metodología de desarrollo de Almacenes de Datos a utilizar

Una metodología es un conjunto de procedimientos, técnicas, herramientas y un soporte documental que ayuda a los desarrolladores a realizar un nuevo software indicando quién debe hacer qué, cuándo y cómo.

Dentro de las metodologías que definen y guían todo el ciclo de vida del desarrollo se encuentran:

- Metodología SQLBI enfocada hacia las herramientas de la *suite* de *Business Inteligencie (*BI*)* de *Microsoft*. (Ferrari y otros, 2008)
- Metodología concebida a finales de 1996 llamada CRISP-DM, la cual se enfoca principalmente en la implementación de minería de datos. (Chrysler y otros, 1996)
- Metodología de Hefesto su idea principal es comprender cada paso que se realizará, para no caer en la rutina de tener que seguir un método al pie de la letra, sin saber exactamente qué se está haciendo, ni por qué. (Bernabeu, 2007)
- Metodología para el Diseño Conceptual de Almacenes de Datos presentada, tesis de Doctorado de Leopoldo Zenaido Zepeda Sánchez, se enmarca en el análisis del diseño de Almacenes de Datos y aporta la incorporación de casos de usos. (Zepeda, 2008)
- Metodologías de Kimball, la misma se basa en dividir el mundo de *Inteligencia de Negocio* entre el hecho y las dimensiones, es muy eficaz y conduce a una solución completa en una cantidad muy pequeña de tiempo. (Kimball, 1998)
- Metodología de Inmon de un almacén de datos es muy diferente de la metodología de Kimball, la estructura Inmon se basa en un complejo empresarial de bases de datos relacionales. (Inmon, 1995)

En este caso, para definir la metodología de desarrollo a utilizar en la Línea Almacenes de Datos y BI de DATEC, se tomó como base la Metodología de Ralph Kimball por los siguientes elementos:

- Crea los conceptos de Hechos y Dimensiones, lo que indudablemente es muy eficaz en el proceso de la toma de decisiones y proporciona mayor agilidad en el proceso de desarrollo.
- Propone ir construyendo el Almacén de Datos a través de la construcción de los Mercados de Datos departamentales, lo que constituye una estrategia buena y coincide con la división lógica de las empresas, entidades, organismos, etcétera.
- Existe abundante documentación sobre la misma, la respuesta a todas las dudas y preguntas que puedan surgir se pueden encontrar en la Web, a través de los servicios que brindan el grupo creador de la metodología.
- Es una metodología madura y reconocida por el resto de la comunidad dedicada al tema. Tiene bien definidas las etapas, actividades, artefactos y roles.

Como complemento a la misma y fortaleciendo la etapa del levantamiento de requisitos; se tomó lo planteado por Leopoldo Zenaido Zepeda Sánchez en su Tesis de Doctorado, orientando así el trabajo a los Casos de Uso y se logra estar más alineado con las tendencias y normas de la Universidad.

Siguiendo lo planteado en las metodologías seleccionadas como base y teniendo en cuenta las características de la UCI y DATEC se utilizará la Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio (BI) en la Línea de Almacenes de Datos e Inteligencia de Negocio (DW&BI) de DATEC. Los flujos de trabajos de esta metodología son: (Yobanis y otros, 2009)

- Estudio Preliminar o Planeación
- Requerimientos
- Arquitectura y Diseño
- Implementación
- Prueba
- Despliegue
- Soporte y Mantenimiento
- Gestión y Administración del Proyecto

En cada flujo de los antes mencionados intervienen grupos específicos los cuales son:

- El Grupo de Análisis

- El Grupo de Almacenes de Datos
- El Grupo de Extracción, Transformación y Carga (ETL)
- El Grupo de Inteligencia de Negocio (BI)
- El Grupo de Dirección

La línea tiene una estructura conformada por cinco grupos, donde cada uno realiza actividades específicas y bien delimitadas según sus responsabilidades dentro de un proyecto. Cada uno de estos grupos se especializa en un conjunto de actividades generales que contienen actividades específicas que tributan al desarrollo del proyecto, generan sus propios artefactos, se dividen por roles para darle cumplimiento a todas las actividades. A continuación se describen brevemente el grupo de ETL.

Grupo ETL: es el responsable de integrar los datos existentes en las distintas fuentes y llevarlos hasta el repositorio de datos creado por el Grupo de Almacén; la mayor carga de trabajo la tienen en los flujos de Arquitectura y Diseño, Implementación y Despliegue. Es importante señalar que es la parte más complicada de este tipo de soluciones, todo depende de la cantidad, variedad y características de las fuentes de datos a integrar. Sus principales actividades son:

- Extraer los datos de las distintas fuentes.
- Limpiar los datos.
- Transformarlos y homogeneizarlos.
- Cargar los datos al Repositorio de Datos.

Los roles correspondientes a este grupo son:

- Arquitecto de Integración.
- Desarrollador .
- Especialista de Calidad de Datos.

Los artefactos que desarrollan son:

- Diccionario de Datos.
- Registro de Sistemas Fuentes.
- Reglas de Negocio.

- Mapa Lógico de Datos.
- Perfil de los Datos.

En cada flujo de trabajo intervienen grupos específicos, cada uno con actividades y responsabilidades concretas, a continuación se describen los flujos donde el Grupo ETL es el responsable.

- *Arquitectura y Diseño:* aquí participan los tres grupos fundamentales, ETL, DWH y BI. En la definición de la arquitectura participan los arquitectos de cada uno de los grupos mencionados. En el diseño participan igualmente los tres grupos pero se incrementa considerablemente la cantidad de personas, todo depende de la complejidad de la solución. Es en ese momento donde se definen las estructuras de almacenamiento, se diseñan las reglas de extracción, transformación y carga, así como la arquitectura de información que regirá el desarrollo de la solución. Los resultados más importantes son: Arquitectura del Sistema, Modelo de Datos del Repositorio Corporativo y las Reglas de Extracción, Transformación y Carga de Datos.
- *Implementación:* participan los tres grupos de desarrollo (ETL, DWH y BI). Se lleva a cabo el diseño físico del Repositorio de Datos, se crean las estructuras de almacenamiento con las particiones y agregaciones correspondientes según la solución en desarrollo. Se crea el Área Temporal de Almacenamiento, se ejecutan las reglas de extracción, transformación y carga, haciendo los ajustes para integrar la información necesaria. Se configuran e implementan las herramientas de BI para obtener los reportes, gráficos, mapas y otros que cubran los requerimientos firmados con el cliente final. Los resultados más importantes son: Repositorio de Datos, Área Temporal de Almacenamiento, Reglas de Extracción, Transformación y Carga, configuración y personalización de las Herramientas de BI.
- *Despliegue:* este flujo consta de dos etapas, la primera es un despliegue piloto, donde se configuran los servidores necesarios y se instalan las herramientas según la arquitectura definida, se cargan una muestra de los datos en un ambiente controlado, con el fin de demostrarle al cliente final que la solución funciona. Una vez aceptada por el cliente, se realiza la carga histórica de los datos, puede ser en el mismo entorno que el despliegue piloto u otro, todo depende de las condiciones del cliente. Es aquí el momento más idóneo para llevar a cabo la capacitación y transferencia tecnológica. Participan todos los grupos y el resultado fundamental es la solución desplegada en el entorno real y en correcto funcionamiento.

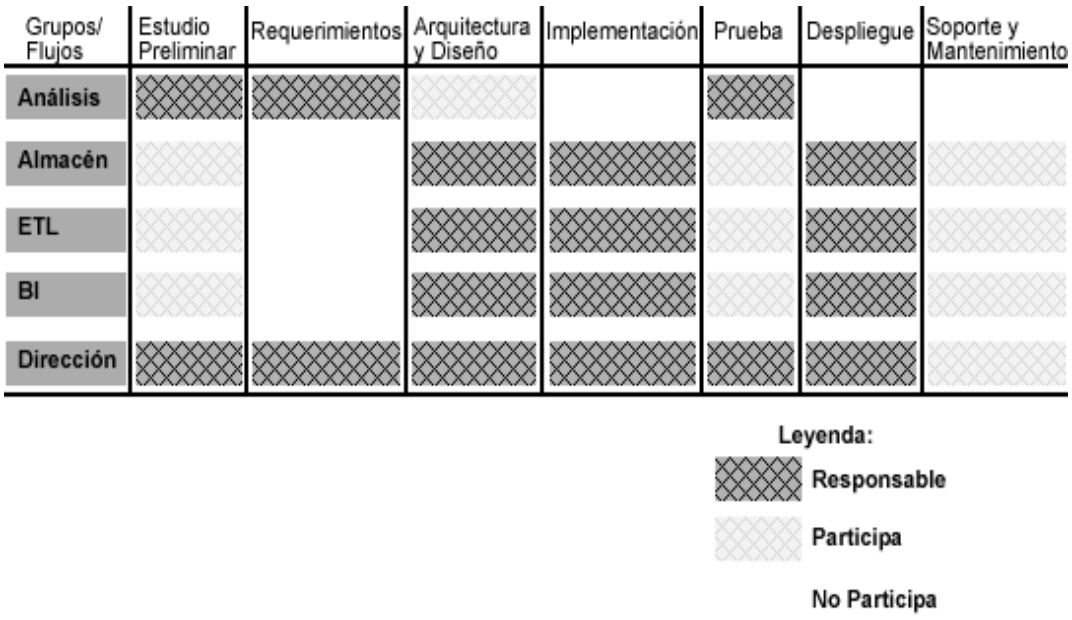

Figura 1: Relación de los Grupos con los Flujos de Trabajo.

Según la cantidad de actividades y los resultados principales de los flujos, la Figura 6 muestra una gráfica que demuestra la responsabilidad, importancia y participación de cada grupo en los distintos flujos de trabajo.

1.12 Parámetros a seguir para la elección de la herramienta ETL. Información general

Actualmente existen muchas herramientas ETL que fueron creadas para mejorar y facilitar los procesos de extracción, limpieza y transformación de datos, además de ahorrar tiempo y dinero al eliminar la necesidad de codificación manual cuando se desarrolle un nuevo almacén de datos. También se utilizan para facilitar la labor de los administradores de bases de datos que conectan las diferentes ramas de las bases de datos, así como integrar y cambiar las bases de datos existentes. Posee tres beneficios principales de las herramientas ETL, ellos son:

- Ganancias en términos de tiempo.
- Procesos automatizados.
- Fiabilidad de los datos.

Resultan esenciales para obtener, integrar y entregar datos almacenados en diversas fuentes sobre clientes, productos, transacciones y riesgos, en el momento y lugar adecuado. Esto permite impulsar otras iniciativas empresariales como favorecer el enfoque hacia el cliente, aumentar la eficiencia operacional y reducir los riesgos mediante una sólida estructura de control de riesgos y conformidad. Estas herramientas tienen facilidades como por ejemplo: clasificar, filtrar los datos de perfiles, controlar y monitorear la calidad, además de la limpieza, vigilancia, sincronización y depuración de los datos.

Los parámetros más importantes que ha de incluir una herramienta ETL son:

- Velocidad: depende en gran medida de los datos que necesita transferir por la red y la potencia de procesamiento implicados en la transformación de los datos.
- Seguimiento: debe permitir encontrar los problemas y depurar durante y después de la fase de desarrollo.
- Conectividad con BD: debe conectarse a una variedad muy amplia de bases de datos.
- Corrección de errores: debe posibilitar rastrear los errores en las transformaciones en tiempo de ejecución.
- Registro de eventos: debe controlarse el proceso registrando los eventos durante la ejecución.
- 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.
- Gestión de las dimensiones lentamente cambiantes: debe ser capaz de manipular las dimensiones lentamente cambiantes.
- Gestión de la sustitución de claves: debe facilitar la sustitución de las llaves del negocio por las llaves de la dimensión.
- Gestión de la calidad de datos: se requiere brindar al usuario realizar acciones de filtrado, limpieza y validación de los datos.
- Perfil de datos: debe realizar perfilado de datos a las fuentes.
- Facilidad de uso: debe ser amigable para el desarrollador, de manera que pueda identificarse rápidamente.
- 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.

- Exigencias de hardware y software: investigar si se pueden cumplir los requerimientos de la herramienta en cuanto a hardware y software.
- Puesta en funcionamiento: debe ser posible agrupar varios objetos ETL y ponerlos en funcionamiento en ambientes de prueba o producción.
- Utilidades del sistema operativo: debe brindar la posibilidad de interactuar con el sistema operativo, ofreciendo servicios de transporte, por ejemplo, a través de la red o usando FTP[12].
- Utilidades de los servicios de transporte: debe ofrecer servicio de transporte a través de la red o de un protocolo para la transferencia de datos.
- Reusabilidad: permite aprovechar parte de la lógica de las distintas tareas, de modo que el desarrollador no tenga que hacer repetidas veces una misma transformación.
- Documentación: se requiere incorpore documentación básica.
- Preparación: deberá ofrecer servicios de preparación para los especialistas.
- Soporte técnico: deberá ofrecer soporte técnico.
- Planificación de la ejecución: ofrece una manera de planificar la ejecución de los trabajos de manera automática.
- Extensibilidad: debe permitir al usuario la definición de nuevas funciones y utilizarlas igual que las que incluye la herramienta.
- Multiplataforma: debe ser capaz de funcionar en cualquier plataforma aunque basta con que sea compatible con la plataforma previamente seleccionada.
- Soporte para metadatos: debe generar metadatos, incluyendo 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.
- 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.

[12] 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.

Todos estos parámetros, ofrecidos por Sylvain DECLOIX, en su artículo Les ETL Open Source “Une réelle alternative aux solutions propriétaires” (Decloix, 2008) permitirán comparar las distintas herramientas, aunque se torna un poco difícil ya que todos los proveedores de software hacen productos cada vez más completos que cumplan con todos los parámetros o con la gran mayoría. En próximos años serán más parámetros, sin embargo, la prioridad entre cada uno de ellos debe ser determinado de acuerdo con los requisitos que tengan los clientes.

1.13 Herramientas ETL para la elección. Principales características

El desarrollo y la diversificación de las herramientas ETL actualmente son crecientes, y se refleja en la amplia variedad de herramientas tanto comerciales como de código abierto[13]. A continuación se mencionan algunos ejemplos de ambos grupos:

Herramientas ETL comerciales:

- IBM InfoSphere DataStage.
- Cognos Decisionstream (IBM[14]).
- Informatics Power Center.
- Oracle Warehouse Builder (OWB).
- Oracle Data Integration (ODI).
- SAS ETL Studio.
- Business objects Data Integration (Corpus).
- Microsoft SQL Server Integration Services (SSIS).
- Ab Initio.
- BI - Tool.
- Sunopsis.

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

[14] 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.

Freeware, herramientas ETL de código abierto:

- Pentaho Data Integration (Kettle).
- Talend Integration Suite.
- Scriptella Open Source ETL Tools.
- Jitterbit.
- CloverETL.
- JasperETL.

Muchas de estas herramientas pueden ser óptimas en cuanto a funcionalidades, pero realmente deben satisfacer las necesidades de integración. Entre las herramientas comerciales más utilizadas se encuentra SQL Server 2008 Integration Services, perteneciente a la empresa Microsoft Corporation, la cual permite crear soluciones de Integración de datos de alto rendimiento, incluidas la extracción, la transformación y la carga (ETL) de datos para Almacenes de Datos.

Posibilita el manejo de secuencias de comandos mediante el uso de Microsoft Visual C # y Microsoft Visual Basic. NET, contiene conectores para SAP BW, Oracle y Teradata, utiliza ADO.NET para las tareas así como para la fuente y los componentes de destino. (Knight y otros 2008)

Otras de las herramientas son Oracle Data Integration (ODI) y Oracle Warehouse Builder (OWB), productos de la familia Oracle Fusion Middleware, las cuales hacen factible la optimización de inteligencia de negocios, almacén de datos y gestión de datos. OWB se fusiona con ODI para crear un sistema unificado de datos. (Yglesias, 2008)

Entre las herramientas de código abierto, sobresalen en el mundo Pentaho Data Integration y Talend Open Studio. A continuación se abordarán las características y funcionalidades de ambas, para luego compararlas y seleccionar la más adecuada para llevar a cabo el desarrollo del trabajo.

1.13.1 Spoon de Pentaho Data Integration

Pentaho Data Integration es una de las herramientas ETL de código abierto que reúne un conjunto de componentes que permiten modelar y ejecutar transformaciones sobre flujos de datos.

Puede funcionar sobre varias plataformas a través de un sistema que soporte Java 1.4 Runtime Environment o una versión superior, y de hardware exige alrededor de 128 MB de RAM. Provee un

JDBC[15] que permite la conexión con cualquier base de datos sin tener que instalar un cliente adicional, usando ODBC [16] en Windows, Oracle, MySQL, AS/400, MS Access, MS SQL 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[17]. Esta herramienta incluye procesamiento optimizado de los ficheros planos.

Brinda soporte para metadatos e 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[18] con numerosos volúmenes de datos, pues maneja pequeñas cantidades de información en el flujo. 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 copiar, eliminar, descompactar y transportar ficheros usando FTP. Esta herramienta es fácil de usar, brinda la posibilidad de copiar y leer del mismo fichero en paralelo, permitiendo maximizar la capacidad de entrada/salida en el entorno ETL. Añade un debugger integrado diseñado para 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:

---

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

[16] 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.

[17] 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.

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

- 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.

El Spoon de Pentaho Data Integration es una de las más antiguas herramientas ETL de código abierto, cuenta con una gran comunidad de usuarios y su interfaz gráfica permite un aumento de la productividad.

Presenta algunas desventajas porque no cuenta con un componente de calidad de datos o una asociación con un proveedor de calidad de los datos, no automatiza el proceso de separación y redistribución de datos para el procesamiento paralelo, 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 entendimiento del funcionamiento de esta herramienta, se muestra la arquitectura que utiliza. (Pentaho, 2010)

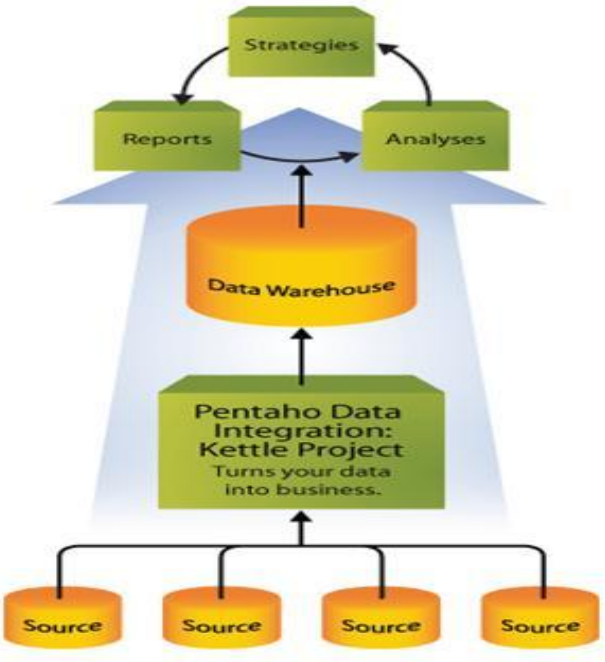

Figura 2: Arquitectura de Pentaho Data Integration.

1.13.2 Talend Open Studio

Talend Open Studio es una herramienta que permite el diseño de los procesos de integración y su seguimiento, está disponible bajo la licencia sin costo GPL[19], la misma incluye tres aplicaciones: Modelador de Negocios, Diseñador de Trabajo y Administrador de Metadatos. Es una herramienta de fácil manejo y no requiere conocimientos técnicos especiales, 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, soporta diferentes plataformas como: Solaris, MAC, Windows, Red Hat Enterprise Linux y Linux, incorpora funciones para el manejo de ficheros y servicios FTP.

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 para realizar las transformaciones ETL.

Contiene más de doscientos componentes y conectores en su librería, la cual proporciona funciones básicas como: operaciones de correlación, búsquedas, filtrado de datos, transformaciones, y algunas facilidades para cargar hacia Almacenes de Datos, la misma puede ser extendida usando lenguajes como Java, Perl o SQL.

Permite activar el modo rastreo, el cual muestra el comportamiento fila por fila, así como el resultado de las transformaciones.

Sus principales ventajas son: 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, cuenta con una interfaz ETL para la importación de los metadatos, la configuración y vinculación de los componentes, además genera scripts que se pueden ejecutar en cualquier sistema operativo que soporte. (Talend, 2009)

1.14 Valoraciones de las herramientas ETL. Comparación

Luego de haber detallado y explicado las herramientas ETL estudiadas, se procede a realizar una comparación entre ellas para tener un punto de partida para su selección. Para ello este estudio se basará en las consideraciones del prestigioso investigador Sylvain DECLOIX, el cual en su artículo Les ETL Open Source “Une réelle alternative aux solutions propriétaires” estableció una comparación extensa, donde

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

abarca casi la totalidad de las funcionalidades y características del Talend Open Studio y el Pentaho Data Integration: (Decloix, 2008)

Tabla 2: Tabla comparativa de las herramientas ETL de código abierto.

Parámetros	**Pentaho Data Integration**	**Talend Open Studio**
Velocidad	✓	✘
Seguimiento	✓	✓
Conectividad con BD	✓	✓
Corrección de errores	✓	✓
Registro de eventos	✓	✓
Ficheros planos	✓	✓
Ficheros EXCEL	✓	✓
Ficheros XML	✓	✓
Gestión de dimensiones lentamente cambiantes	✓	✓

Gestión de sustitución de claves	✓	✘
Gestión de la calidad de datos	✓	✓
Perfil de datos	✘	✓
Paralelismo	✓	✓
Ejecución automática de tareas	✓	✓
Reusabilidad	✓	✓
Planificación de tareas	✓	✓
Extensible	✓	✓
Gestión de metadatos	✓	✓
Facilidad de uso	✓	✘
Registro de eventos	✓	✓

Bases de datos Microsoft Access	✓	✘
Bases de datos Oracle	✓	✓
Bases de datos SQL Server	✓	✓
Bases de datos DB2	✓	✓
Sistemas ERP (SAP)	✓	✓
Motor OLAP	✘	✘

En la tabla anterior se muestra el desempeño en distintos parámetros de las herramientas Talend Open Studio y Pentaho Data Integration donde se puede observar cierta similitud entre la mayoría de los parámetros aunque existen algunas diferencias en cuanto a velocidad y facilidad de uso donde el Pentaho Data Integration es superior. Teniendo en cuenta esta consideración y que actualmente es la herramienta que se está utilizando en el grupo de Integración de datos del centro DATEC, en la cual existe un amplio dominio y experiencia por parte de los especialistas del grupo se propone utilizar la herramienta Pentaho Data Integration para el desarrollo del proceso.

1.15 Herramienta CASE a utilizar. Principales características

Existen disímiles herramientas CASE que consisten en diversas aplicaciones informáticas destinadas a aumentar la productividad en el desarrollo del software reduciendo el costo de las mismas en términos de tiempo y dinero, potencian el modelado y ayudan en todos los aspectos del ciclo de vida de desarrollo del software.

Entre las más distinguidas se puede nombrar Rational Rose Entreprise Edition, es el producto más completo de la familia Rational Rose. Como el resto de sus productos incluye soporte Unified Modeling

Language (UML[20]), soporta la generación de código a partir de modelos en Ada, ANSI C++, C++, CORBA, Java™/J2EE™, Visual C++ y Visual Basic. Potencia prácticas modernas de ingeniería de software, propone la utilización de diferentes tipos de modelo para realizar un diseño del sistema, proporciona un lenguaje común de modelado para el equipo de desarrollo.

Otra herramienta importante es Entreprise Architec la cual soporta la ingeniería inversa y generación de código fuente para muchos lenguajes populares, incluyendo: ActionScript, Ada, C y C++, C#, Java, Delphi, Verilog, PHP, VHDL, Python, System C, VB.Net, Visual Basic. En el ambiente de modelado para trabajo en equipo potenciado por UML 2.1, abarca el ciclo de vida completo del desarrollo de software. 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.

Aparte de las mencionadas se analiza Visual Paradigm, pues es una herramienta CASE profesional que ayuda a construir aplicaciones de forma eficiente y que soporta el ciclo de vida completo del desarrollo de software: modelado del negocio, requerimientos, análisis y diseño orientados a objetos, construcción, pruebas y despliegue. El diseño está centrado en los casos de uso y enfocado en el negocio, permitiendo que se genere un software de mayor calidad. Hay que señalar su robustez, usabilidad y portabilidad y que se integra a diferentes herramientas Java. Está diseñada para dar soporte a arquitectos de sistemas, diseñadores, desarrolladores, analistas de procesos de negocio y modeladores de datos en los procesos de desarrollo de software. Además, es colaborativa, o sea, soporta múltiples usuarios trabajando sobre el mismo proyecto; genera la documentación del proyecto automáticamente en varios formatos, permite control de versiones, proporciona el diseño de ingeniería inversa permitiendo el soporte de: Clases java, .NET (.dll y .exe), JDBC y archivos de mapeo ocioso, también permite 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. Posee generación de código en: C#, VB, .NET, Object Definition Language (ODL), Flash ActionScript, Delphi, Perl, Ruby. (Headquarters, 2009)

[20] 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.

Considerando las potencialidades ofrecidas por Visual Paradigm las cuales satisfacen las necesidades de modelado existentes en este trabajo y además que la Universidad de Ciencias Informáticas (UCI) cuenta con su licencia, se opta entonces por la utilización de la misma.

1.16 Herramienta de perfilado de datos a utilizar. Principales características

El perfilado de datos es una de las primeras tareas que se realizan en el proceso de calidad de datos, y consiste en realizar un primer análisis sobre los datos de origen, normalmente, sobre tablas, con el objetivo de empezar a conocer su estructura, formato y nivel de calidad.

Se hacen consultas a nivel de tabla, columna, relaciones entre columnas, e incluso relaciones entre tablas. La herramienta para el perfilado de datos que se decide utilizar en este trabajo es DataCleaner en su versión 1.5.3. Según su creador Kasper Sorense el sistema requiere Java Runtime Enviroment 5.0 o una versión superior y Drivers de JDBC. La misma permite la evaluación del nivel de calidad de los datos contenidos en el sistema de información. Es una aplicación muy fácil de usar, genera sofisticados informes y gráficos que permiten a los usuarios determinar de un vistazo el nivel de calidad de los datos, identificar y analizar la estructura del origen de datos y combinar resultados y gráficos, creando vistas fáciles de interpretar para evaluar la calidad de los datos.

Las características incluyen: (Sorensen, 2009)

- Los perfiles de datos se utilizan para calcular y analizar diversas medidas importantes basadas en los valores de los datos.
- Validación de datos: el validador le dará un resultado que puede ser interpretado como bueno o malo, ya que el validador valida los datos.
- Los datos de comparación.
- Diccionario de gestión.
- Análisis del modelo.
- Soporta acceso de lectura a muchos tipos de Almacenes de Datos:
    - Bases de datos compatibles con JDBC (oficialmente probadas y compatibles: Oracle, MySQL, PostgreSQL, Firebird, SQLite, HSQLDB, Derby / javadb).
    - Valores separados por comas (.csv).

- Excel (.Xls) hojas de cálculo.
- Archivos XML.
- OpenOffice Base (ODB) archivos.

1.17 Conclusiones del Capítulo 1

Después del estudio realizado en este capítulo se decide que:

- La forma de Integración de datos a utilizar será ETL para darle solución a la problemática.
- La metodología apropiada, es la Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio (BI) en la Línea de Almacenes de Datos e Inteligencia de Negocio (DW&BI) de DATEC, pues cumple con los requerimientos necesarios para la implementación del Proceso ETL.
- Las herramientas seleccionadas son:
    - Pentaho Data Integration en su versión 3.1 para realizar el proceso de Integración de datos.
    - Visual Paradigm en su versión 3.4 para el modelado del sistema.
    - DataCleaner en su versión 1.5.3 para realizar el proceso de Perfilado de datos.

CAPÍTULO 2: Implementación del Proceso de Integración de datos de SAIME e INTTT para el CICPC

2.1 Introducción

El propósito central de este capítulo es elaborar una propuesta de implementación del proceso ETL, además de detallar, desarrollar y documentar dicho proceso para el CICPC.

2.2 Propuesta arquitectura de Integración de datos

Una arquitectura en el ámbito computacional es un conjunto de estructuras o reglas que proveen un esqueleto para el diseño general de un producto o sistemas. En el proceso de Integración de datos no es recomendable iniciar el desarrollo de una solución sin haberla premeditado previamente, identificar sus fuentes, su esquema, el movimiento de los datos y determinado su enfoque de almacenamiento de datos.

Para comenzar a describir la arquitectura de este proceso, es necesario recordar algunos elementos referentes a la implementación de este sistema como son:

- Fuente de datos.
- Área temporal.
- Almacén de Datos Operacional.

Fuente de datos:

Son los ficheros o datos en bruto, que se encuentran almacenados en los sistemas fuentes o carpetas que guardan la información histórica de los sistemas. Dichos ficheros se encuentran con extensión XML. Estos sufrirán un proceso de extracción hacia un servidor local, para facilitar el trabajo con la transformación y la homogeneidad de los tipos de datos, la información y los campos de las tablas en estos ficheros.

Área temporal (Staging Area):

Constituye el área de preparación de datos para facilitar los procesos y técnicas de integración para luego ser cargados al destino. El mismo será configurado de acuerdo con las necesidades de los especialistas ETL. En este caso se utilizará el gestor PostgreSQL en su versión 8.4, creándose las tablas y atributos necesarios para dicho montaje.

Almacén Operacional de Datos (ODS):

El almacén Operacional de datos constituye el destino hacia donde se integrarán los datos a cargar. Este estará listo y montado por el diseñador y administrador de este sistema, el cual responderá a las necesidades del negocio y de integración, en correspondencia con los procesos ETL que se implementarán.

Tomando como base el nivel de detalle que requiere el proceso de Integración de datos por su complejidad, la arquitectura que se muestra en la siguiente figura se utilizó para el desarrollo de la solución.

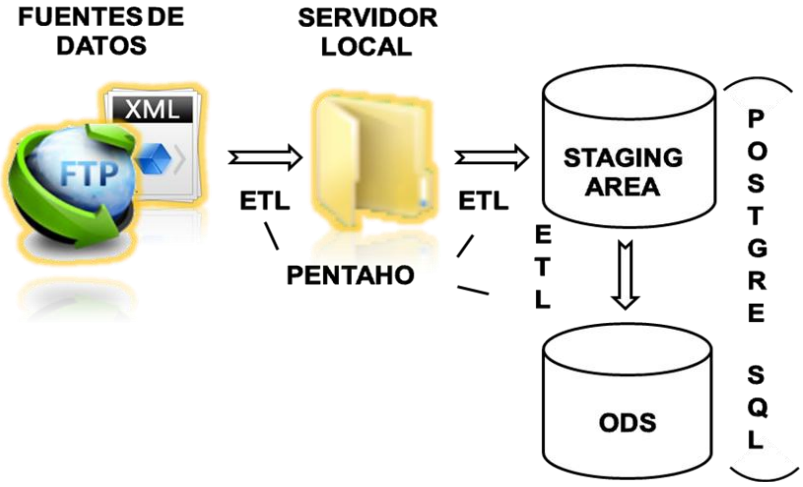

Figura 3: Arquitectura propuesta.

Como se puede ver en la figura 3, los datos extraídos de los sistemas fuentes son cargados al Staging Area mediante procesos ETL utilizándose como herramienta el Pentaho Data Integration 3.1. Así mismo se vuelven a implementar procesos ETL para integrar los datos a cargar en el ODS provenientes del área temporal usando dicha herramienta.

2.3 Aspectos generales de los sistemas fuentes

Como se había abordado anteriormente, los sistemas externos que constituyen la fuente de la integración son SAIME e INTTT. La comunicación con los mismos se realiza mediante un FTP de manera que cada entidad debe definir el esquema o estructura de la información que necesita obtener. Para este proceso,

en los sistemas es creado un buzón por el protocolo FTPS[21] con el nombre de la entidad, y dentro varias carpetas con la información solicitada. En estas carpetas se pondrán los archivos en formato ZIP o RAR, encontrándose de manera compactada los archivos en formato XML con los datos de cada notificación que haya sido solicitada.

CICPC constituye la entidad a la cual deben brindar información estos sistemas externos. En este sentido, se hace necesario definir algunas medidas de seguridad, tales como que las entidades deben acordar la dirección IP con la cual se establecerá la conexión, además se le creará un usuario a CICPC y le será emitida una confirmación digital para que pueda acceder a la información.

La información a intercambiar estará disponible en horas de la noche, pues es recomendable que la extracción de los datos de los sistemas fuentes se realice en este horario para evitar que el sistema origen colapse.

2.3.1 SAIME

Este sistema contiene información acerca de las personas registradas tanto de la República Bolivariana de Venezuela como de otros países registrados en este sistema. Entre los servicios que brinda se encuentra el de Identidad, el cual es el encargado del control del proceso de cedulación de los ciudadanos venezolanos y extranjeros así como de la emisión de pasaportes venezolanos. La Extranjería es la responsable de controlar la admisión y actividades de los ciudadanos extranjeros que ingresan al país para así garantizar que sus actos estén dentro del marco de la ley. La Migración es el servicio encargado de dirigir, controlar y supervisar las diferentes funciones que desempeñan los jefes de departamento, coordinadores regionales, jefes de secciones y oficinas de migración y fronteras, tomando en cuenta el marco jurídico legal interno y otros tratados y acuerdos internacionales emitidos por la República Bolivariana de Venezuela (SAIME, 2009).

[21] Conocido como FTP/SSL**,** nombre usado para abarcar un número de formas en las cuales el software FTP puede realizar transferencias de ficheros seguras

2.3.2 INTTT

El sistema fuente INTTT es el encargado del control de toda la información relacionada con los vehículos y los datos de cualquier operación que se haga con los mismos, debe llevar a cabo el Registro del Sistema Nacional de Tránsito y Transporte Terrestre. El mismo otorga, registra y controla tanto títulos profesionales para conducir vehículos como sus placas identificadoras, destinadas al uso público o privado. Permite el registro, expedición renovación y control de licencias para conducir en el ámbito nacional, en los diferentes grados y categorías. Entre otros servicios posibilita el otorgamiento de los permisos y registro de los servicios de transporte terrestre público y privado, así como la regulación y control del transporte terrestre público de pasajeros y de carga en el ámbito nacional.

INTTT tiene como principal misión regular, controlar y ejecutar políticas en materia de tránsito y transporte terrestre, con el fin de garantizar la comodidad, calidad, eficiencia y seguridad para los usuarios (INTTT, 2008).

2.4 Características del ODS

El Almacén Operacional de Datos ODS_CICPC constituye el sistema hacia donde se integrará la información. El mismo estará soportado físicamente en el gestor de base de datos PostgreSQL 8.4, el cual gestionará la implementación física del Modelo de Datos en el Sistema operativo Debian 5.0. Este ODS cuenta con dos esquemas: Saime e Inttt, para así organizar los datos correspondientes a los sistemas externos. El ODS de manera general cuenta con un total de 6 tablas con toda la información debidamente integrada, las cuales son: ods_pers_nat, ods_direccion, ods_relac_pers, ods_viaj_int, ods_licencia y ods_vehiculo.

2.5 Configuración y montaje del área de preparación de datos

Luego de tener definida la arquitectura a utilizar en el proceso de Integración de datos es conveniente crear las condiciones necesarias para el desarrollo. Es aquí donde se instala el hardware, el sistema operativo sobre el cual se trabajará, las herramientas relacionadas con el ODS, las conexiones necesarias así como la creación de los usuarios con los permisos requeridos.

Como se ha planteado, las herramientas utilizadas para el desarrollo del proceso de Integración de datos son:

- PostgreSQL 8.4
- Pentaho Data Integration 3.1
- DataCleaner 1.5.3

De manera tal que puedan ser aprovechadas sus potencialidades y facilidades de uso, principalmente en los procesos de almacenamiento de la información y en los de extracción, transformación y carga de datos.

La Arquitectura ETL a implantarse para facilitar los procesos de Integración de datos en el Cuerpo de Investigaciones Científicas, Penales y Criminalísticas contará con los siguientes requisitos de hardware para lograr un funcionamiento adecuado:

- 1 Servidor con 1 GB de memoria RAM, 80 GB de capacidad de disco duro, procesador a 3.0 GHz de velocidad.

La base de datos del área temporal (Stg_Area) será administrada en el servidor que aparece descrito sobre el gestor PostgreSQL 8.4, debido a que la extracción de información desde los ficheros a la base de datos del área temporal se realiza una sola vez, y automáticamente, al concluir los procesos de extracción y transformación se procederá a la carga de los datos hacia el Almacén de Datos Operacional (ODS_CICPC). Es en este servidor donde se realizarán todas las técnicas y mecanismos de limpieza e Integración de datos para lograr que estos tengan la calidad requerida. Es importante aclarar que la información en la base de datos donde se encuentra el área de preparación de datos, se eliminará mensualmente para liberar la carga de procesamiento del mismo.

2.6 Seguridad de los datos

Garantizar la seguridad de los datos durante el proceso ETL constituye una tarea de gran importancia, ya que permite garantizar la confidencialidad e integridad de los datos.

La información es protegida desde que va a ser extraída de las fuentes, ya que la comunicación se realiza mediante un protocolo seguro al cual solo podrá acceder el responsable del equipo ETL, el mismo antes de realizar la copia debe recibir una notificación de la dirección IP a la que debe conectarse y las claves de acceso.

Para la realización de cada transformación es necesario conectarse a la base de datos que se encuentra en el Stg_Area la cual se encuentra en la misma computadora que se realiza el proceso de ETL lo que posibilita la seguridad de esta conexión ya que solo el equipo ETL es el que tendrá la posibilidad de tener acceso a ella.

Para el trabajo de actualización, administración y configuración en el proceso ETL se definen usuarios y roles.

Se definió un usuario para cada sistema externo cumpliendo con las necesidades y exigencias de seguridad del negocio, el cual tiene como objetivo principal la protección y control de acceso sobre la información perteneciente a cada sistema el cual garantiza la confidencialidad de los mismos solamente para aquellos usuarios con los permisos suficientes para acceder a la misma. Estos usuarios son SAIME e INTTT.

El desarrollador ETL trabajará con un usuario de administración de la base de datos definido para la implementación del proceso de integración, el cual posee acceso de lectura y escritura sobre toda la información contenida en el gestor. Es el único rol que podrá actualizar datos sobre el ODS, de ahí que su sensibilidad sea crítica, a este rol solo pertenecerán un máximo de 2 personas, además se validará que el acceso se establezca por canales seguros y certificados para la realización del proceso.

Con el propósito de garantizar la persistencia de la información que es necesaria para la Integración de Datos se realizarán copias de seguridad completos periódicamente a los metadatos, garantizando en todo momento que exista una copia exacta de la información.

Se propone que el máximo de años a almacenar sea 10 ya que no es necesario que el tamaño histórico crezca indefinidamente, además de que a partir de su aumento la información sea almacenada anualmente en una estructura similar a las copias de respaldo.

2.7 Procesos de Integración de datos en el ODS

En este epígrafe se describirán todos los procesos de integración, detallando en cada uno de ellos sus características:

2.7.1 Extracción de datos del FTP

En este primer paso se obtendrán los ficheros compactados que se encuentran en los FTP con la información de los sistemas externos SAIME e INTTT. Luego de tener los ficheros, estos serán descompactados en el servidor local que no es más que un directorio en formato digital que se crea con el objetivo de organizar mejor el trabajo. Posteriormente se extraerán los datos XML hacia las tablas de la base de datos del área temporal (Stg_Area).

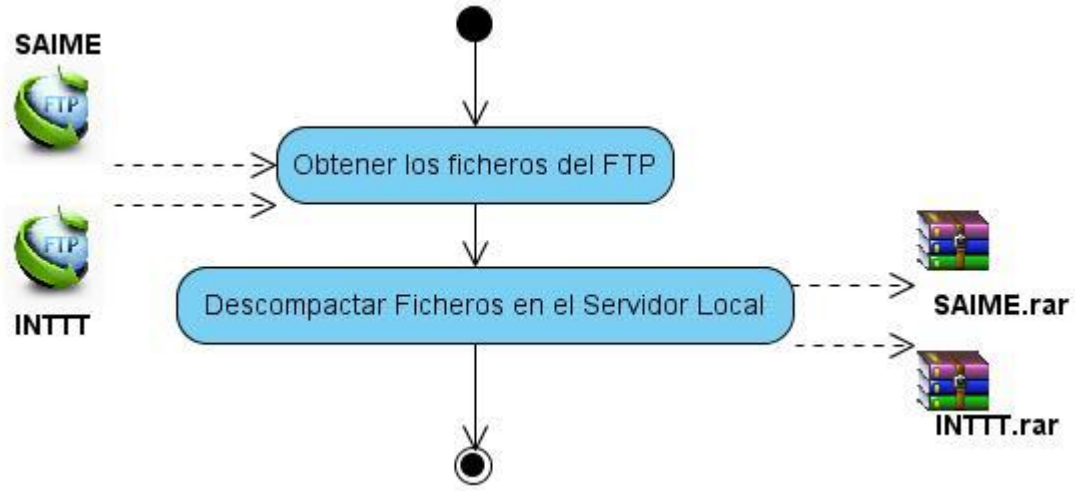

Figura 4: Orígenes de datos.

2.7.2 Extracción de datos del Servidor Local

Luego de tener los archivos XML en el Servidor Local, se implementa el proceso de ETL con la extracción de los datos provenientes del mismo hacia el área temporal, pues primeramente es necesario adecuar los datos al modelo relacional elaborado. Como estos llegan estructurados de forma similar, se deberán verificar tanto en el directorio local como en el área temporal hasta ir completando todos los atributos de cada tabla.

Este servidor dispone de dos directorios para almacenar los datos provenientes de las fuentes, estos son SAIME e INTTT. Del primero se extraen todos los datos relacionados con la persona, su o sus direcciones, los viajes internacionales que han efectuado, sus relaciones filiales, entre otras. De INTTT se obtienen los datos asociados a los vehículos terrestres, las licencias, los propietarios de los vehículos, así

como trámites realizados a los vehículos, entre otros datos de suma importancia para este sistema. Este proceso de extracción es mostrado en la figura 4.

Después de almacenar estos datos en el área temporal se procede a la eliminación de los ficheros del directorio para que se encuentre listo para la extracción de los datos siguientes.

La extracción se realizará diariamente en horas de la noche para no afectar el rendimiento de las aplicaciones, vale destacar que todo este proceso es realizado en un solo servidor. Luego de tener los datos extraídos, estos están en condiciones de ser limpiados y transformados, pero antes es recomendable realizarles un análisis o perfilado de los datos.

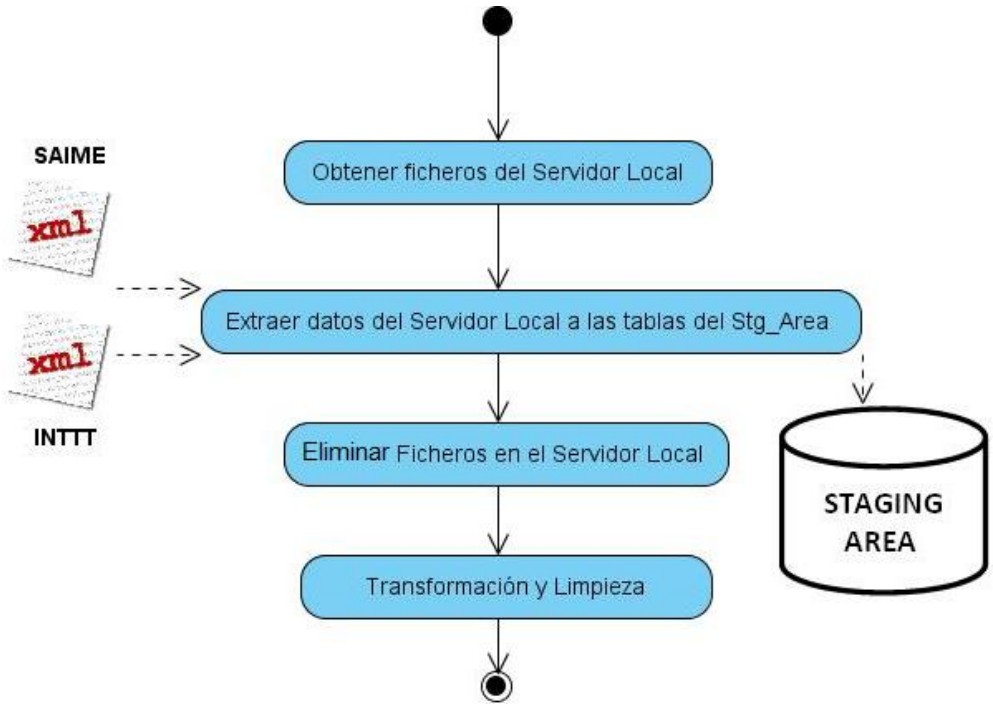

Figura 5: Extracción en el proceso ETL ODS_CICPC.

2.7.3 Perfilado de datos

El perfilado de datos es utilizado para examinar los datos existentes y obtener estadísticas e información sobre los mismos. Este proceso es muy importante pues se establecen reglas para corregir los datos que pueden presentar problemas como: valores indebidos, escritos incorrectamente, ausentes o duplicados,

ya que los datos provenientes de las fuentes externas en ocasiones no son completos o no cumplen con las normas necesarias para garantizar su disponibilidad.

En este trabajo el análisis de los datos se realizó utilizando la herramienta DataCleaner, la cual permitió generar reportes de los Estándares de medidas, Análisis de cadenas, Análisis numéricos y Tiempo de análisis. Esta herramienta genera otros reportes pero estos fueron los principales que permitieron realizar un efectivo análisis de los datos extraídos.

En los Estándares de medidas se obtiene el resultado de todas las filas, valores nulos y vacíos, así como el mejor y más bajo valor de la tabla analizada como se muestra en la siguiente tabla.

Tabla 3: Perfilado de los Estándares de medidas de la tabla stg_pers_dir.

Standard measures	idpersona	iddireccio	fecha
Row count	100	100	100
Null values	0	0	0
Empty values	0	0	0
Highest value	1199	1483	2009-10-17
Lowest value	6	15	1900-11-11

Con el reporte del Análisis de cadenas se adquiere toda la información de la cantidad, mínima, máxima y promedio de caracteres, los espacios en blanco, los caracteres no escritos, entre otros parámetros, los cuales son mostrados a continuación.

Tabla 4: Perfilado del Análisis de cadenas de la tabla stg_persona.

String analysis	estadocivil	primernombre	segundonombre	profesion	letracedula
Char count	100	659	424	863	100
Max chars	1	13	13	11	1
Min chars	1	4	4	6	1
Avg chars	1	6,59	4,28	8,63	1
Max white spaces	0	2	2	0	0
Min white spaces	0	0	0	0	0
Avg white spaces	0	0,26	0,07	0	0
Uppercase chars	53%	30%	6%	0%	85%
Lowercase chars	47%	65%	92%	100%	15%
Non-letter chars	0%	3%	1%	0%	0%
Word count	100	102	101	100	100
Max words	1	3	3	1	1
Min words	1	1	1	1	1

Como se muestra en la siguiente tabla, el resultado del reporte Análisis numérico permite conocer el valor más alto y el menor, la suma de los valores, la medida geométrica, el promedio y la desviación estándar, así como la diferencia.

Tabla 5: Perfilado del Análisis numérico de la tabla stg_direccion.

	tiporesidencia	caserio	tipodireccion	municipio
Word count	1469	1434	1455	1428
Uppercase ch...	100%	0%	100%	100%
Non-letter ch...	0%	0%	0%	0%
Min words	0	0	0	0
Min white spa...	0	0	0	0
Min chars	0	0	0	0
Max words	1	1	1	1
Max white sp...	0	0	0	0
Max chars	50	20	30	20
Lowercase ch...	0%	100%	0%	0%
Char count	36664	14806	22457	15122
Avg white spa...	0	0	0	0
Avg chars	24,44	9,87	14,97	10,08

El reporte del Tiempo de análisis está centrado específicamente en los datos de tiempo, brinda información de la mayor y menor fecha, además de la cantidad de fechas por año como se muestra a continuación.

Tabla 6: Perfilado del Tiempo de análisis de la tabla stg_prop_veh.

Time analysis	fechatramite
Highest value	2010-04-18
Lowest value	1990-04-09
Where [Year=1990]	28
Where [Year=1991]	26
Where [Year=1992]	28
Where [Year=1993]	30
Where [Year=1994]	35
Where [Year=1995]	31
Where [Year=1996]	31
Where [Year=2005]	32
Where [Year=2006]	20
Where [Year=2007]	21
Where [Year=2008]	43
Where [Year=2009]	28
Where [Year=2010]	10

2.7.4 Llaves sustitutas

En ocasiones existen algunas tablas que los valores de sus identificadores deberían ser distintos, sin embargo, hay algunos que se repiten. Para darle solución a esta situación y evitar desechar información, se aplica la técnica de Llaves sustitutas (surrogate keys), la cual consiste en añadirles a todas estas tablas un campo que asume el papel de llave primaria en el ODS. Estos campos son números enteros sin ningún valor para los usuarios finales, pero necesarios para el equipo de desarrollo. Son auto-incrementables que van a identificar a cada una de las tablas, y se utilizarán para las operaciones de unión entre tablas. Se hace necesario que cada vez que se inserten datos nuevos en las tablas, se deberán generar de forma automática nuevos valores que no hayan sido tomados para así eliminar la redundancia y disminuir la incoherencia de los datos (Vidot, 2008).

De esta manera, en este trabajo se utiliza la estrategia de llaves sustitutas para no depender de las llaves primarias que proponen los sistemas fuentes. Estas deberán ser únicas. En este caso cada vez que se inserte una fila, se incrementará automáticamente el valor de la llave primaria durante el proceso de carga al ODS_CICPC.

2.7.5 Llaves nulas y huérfanas

La información que se maneja en un ODS es muy importante para la empresa o entidad, pues brinda un alto nivel informativo y permite realizar múltiples acciones para mejorar sus servicios y negocios, la cual debe ser fiable. Por tanto, las únicas transformaciones deben estar dirigidas a convertir los datos en una información fácil de entender para el usuario. Durante el proceso de carga se incorporarán procesos de limpieza simples para los valores nulos.

Para darle tratamiento a estos valores nulos se recomienda sustituirlos por alguna cadena de texto, la cual describa la situación. En este sentido, el tratamiento a estos valores nulos puede ser realizado mediante simples transformaciones JavaScript, o se podrá utilizar una tupla con descripción Desconocido y con un valor entero que no se utilizará, a no ser para este caso.

También puede ocurrir que en alguna tabla que dependa de otra existan llaves que aún no han sido registradas en las tablas de las cuales dependa, en este caso pudieran aparecer llaves huérfanas. En las tablas donde esto ocurra no se puede desechar la información, pues en este proceso los datos no deben ser modificados. Ante esta situación es recomendable incorporar estas llaves a las tablas aun sin saber el valor que toman el resto de los atributos, para esto se generará un valor sustituto para la llave y se continuará con la carga, además de insertar el nuevo valor en la tabla o realizarle auditoría a estos datos nulos, los cuales serán guardados en un documento Excel para enviárselo a las fuentes.

En este trabajo la estrategia de llaves nulas es aplicada realizando transformaciones JavaScript en las cuales a los valores que vengan vacíos se les asignará la descripción “Desconocido” en caso de ser una cadena y con un valor 9999 en caso de ser un valor numérico. En este caso por ejemplo, al realizarse la carga de la tabla tb_vehiculo en el ODS, el atributo estado para algunos vehículos no trae ninguna información, estos datos vacíos son convertidos en desconocidos, de manera similar ocurre con algunos datos del atributo capacidad, el cual al no traer ningún valor automáticamente se le asigna el código 9999 correspondiente.

La estrategia de llaves huérfanas en este trabajo se lleva a cabo en la carga de varias tablas que tienen referencias por ejemplo a la tabla persona en el ODS. Al poblar la tabla ods_viaje_int, al encontrarse un id llave foránea de persona que no encuentra su correspondiente en la tabla persona del ODS, esto genera un conflicto, por lo cual estos datos no pueden ser cargados, es ahí cuando los datos son guardados en

un documento Excel y posteriormente son notificados a la fuente que esos datos no pudieron ser cargados.

2.7.6 Transformación y Limpieza

La etapa de transformación y limpieza es muy importante, ya que una vez culminado este proceso, los datos estarán listos para ser cargados en el ODS_CICPC.

La limpieza tiene como misión detectar y corregir los datos erróneos, sin sentido, o corruptos. Permite además detectar entradas duplicadas o incompletas y establecer algunas reglas para corregirlas. Esta etapa comienza con la limpieza de los datos que ya han sido extraídos satisfactoriamente. Principalmente en los pasos de limpieza y ajuste es donde se añade valor, en los otros pasos de transformación solo se reformatean y se mueven los datos.

Entre las tareas de transformación más comunes encontradas están: corrección de datos mal escritos, inclusión de valores por defecto para datos faltantes, eliminación de registros repetidos, combinación de datos, ordenamiento de datos, resumen de datos, asignación de claves, entre otros. Para darle cumplimiento a estas tareas se le dará tratamiento a los valores nulos y a la generación de las llaves sustitutas, además teniendo en cuanta las reglas definidas por el negocio.

Las transformaciones básicas estarán basadas en las reglas del negocio definidas para el negocio, que es donde se definen todos los posibles casos que deben de ser cambiados, por ejemplo en caso de que en la tabla stg_persona el atributo pasaporte venga con valor “vacío”, pues este valor debe transformarse en “desconocido”, otro ejemplo es que en caso de que en la tabla tb_vehiculo el atributo año tenga valor “vacío”, la transformación aquí sería convertir este valor en “0”.

2.7.7 Metadatos

Los metadatos son los encargados de describir las características de la información almacenada, ayudando a identificarla y administrarla. En este conjunto de datos se encuentran los que describen el sistema para dar una visión del buen funcionamiento del mismo y los que guían los subprocesos de extracción, transformación y carga (Vidot, 2008).

En este trabajo la herramienta Pentaho Data Integration 3.1 mantiene una gestión eficiente de los metadatos, almacenando en un repositorio todas las informaciones referentes a las transformaciones y los

trabajos, por ejemplo, el nombre, el momento de ejecución, el usuario que ejecutó el proceso, y una serie de características generales del mismo.

Además, es utilizado el metadato del negocio para realizar la transformación del atributo país (origen o destino), la cual es necesaria ya que en el origen el país viene con el nombre completo y en el destino debe aparecer este atributo con sus siglas, para realizar esta transformación fue necesaria la creación de un diccionario de correspondencia, el cual es una tabla llamada paises que cuenta con cuatro atributos: pais_orig, abreviatura_orig, pais_dest, abreviatura_dest.

Durante el proceso de integración no siempre es necesario la carga de todos los datos al ODS, a este solo se cargan los datos que no se han cargado. Para poder saber cuáles son las tablas a cargar es necesario tener un control de cuando fue que se cargó por última vez cada tabla. Para darle solución a esta problemática se utiliza el metadato técnico, el cual fue necesario para la creación de una tabla tb_metadata_fecha que tendrá un atributo que almacena el nombre de las tablas y otro atributo que guarda la fecha de la última actualización para facilitar que se realice la carga de los datos actuales. Esto se facilita mediante la verificación de la fecha de los datos a extraer, que deberá ser mayor que la fecha registrada en la tabla tb_metadata_fecha correspondiente.

2.7.8 Carga de datos

Este es el último subproceso dentro de la Integración de datos, el cual consiste en cargar todos los datos que ya han sido transformados satisfactoriamente.

Se debe tener en cuenta que para cargar todos los datos al ODS_CICPC no se debe hacer al mismo tiempo pues en caso de que ocurra algún incidente se cancelará la carga completa. Para evitar la aparición de llaves huérfanas en el proceso deben cargarse primero las tablas que no tienen llaves foráneas, las cuales son las que no dependen de la información de otras tablas, y luego se procederá a la carga de las tablas que tienen alguna dependencia de otras. Por ejemplo en la tabla ods_persona debe ser cargada primero que la tabla ods_pers_nat, pues esta última depende del identificador de la persona, el cual se encuentra en la tabla ods_persona. En la imagen siguiente se muestra todo el proceso de ETL explicado anteriormente.

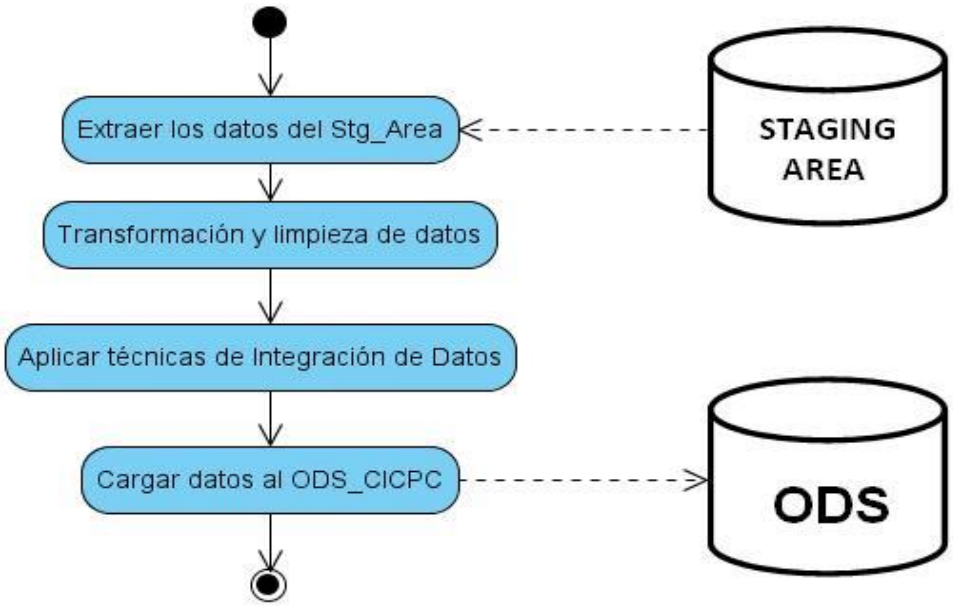

Figura 6: Proceso ETL ODS-CICPC.

2.8 Detalles del proceso de Integración de datos

A continuación se explica detalladamente el proceso de ETL realizado. Este proceso fue realizado a todas las tablas, pero por la extensión de la descripción de las mismas solo se mostrará a continuación el realizado a la tabla Persona.

2.8.1 Persona

En La tabla stg_persona se almacenan todos los datos relacionados con la persona, ya sean nacionales o extranjeros registrados en Venezuela. Ésta contiene información valiosa de los ciudadanos como el nombre, los apellidos, la letra y el número de la cédula, el pasaporte en caso de que tenga, el peso, la estatura, la fecha de nacimiento, la profesión actual del mismo, el estado civil y el color de la piel. La arquitectura propuesta para la realización del proceso ETL de esta tabla es la siguiente:

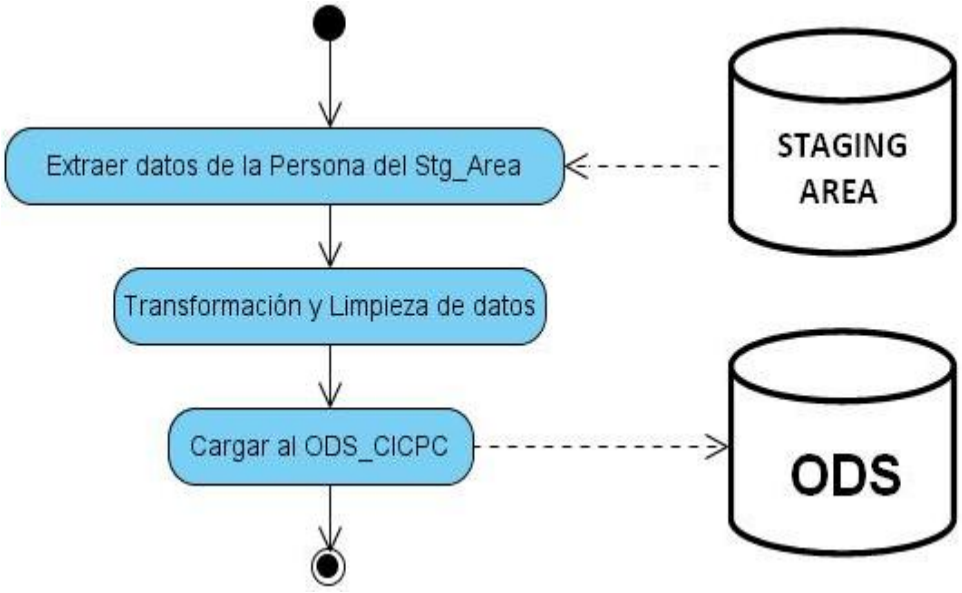

Figura 7: Arquitectura del proceso ETL de la tabla Persona.

Luego que los datos están en el área temporal o Stg_Area se procede a la realización del proceso ETL para la tabla Persona.

Este proceso comienza con la extracción de los datos de la persona que están disponibles en el área temporal mediante el componente para la entrada de tabla (Entrada Persona Stg_Area). Luego de tener los datos se procede a la selección de los datos y renombrarlos en caso de que sea necesario, ya que algunos de los datos a cargar al ODS_CICPC no se encuentran con el mismo nombre, el atributo al no tener el mismo nombre, este no lo reconocería y ese dato no sería cargado, por ejemplo aquí se utiliza el componente de Selección y Renombre de Valores para los atributos peso y estatura, pues en el ODS se llaman así y en el Stg_Area rango_peso y rango_estatura respectivamente. Seguido a este componente por medio del Mapeo de Valores son convertidos los valores de un campo a otro, por ejemplo el atributo color de la piel en el origen tiene los posibles valores: b, n, B o N y en el destino estos atributos deben tomar el valor de Blanco y Negro. El componente de filtrado de filas permite filtrar datos utilizando transformaciones sencillas y al resultado del mismo darle tratamientos diferentes, como en este caso se filtra el segundo nombre de la persona y se le da un tratamiento en caso de que el mismo venga con valor nulo y otro en caso de que esta persona tenga dos nombres. Utilizando el componente de JavaScript el cual permite convertir valores y darle tratamiento a los valores nulos, como el nombre que en la fuente

puede venir en mayúscula o minúscula pero en el destino debe tener solo la primera letra en mayúscula. En caso de que el identificador venga vacío, mediante el componente filtrado de filas, los datos correspondientes a esta persona son enviados a un documento Excel, para notificarle a la fuente que los mismos no fueron cargados. Con todo esto realizado correctamente ya los datos son cargados al OSD_CICPC usando el componente Insertar/Actualizar. Los componentes que siguen este proceso son destinados a garantizar el control de cambio del sistema, al ser cargados estos datos es actualizada la tabla tb_metadata_fecha para así, en una posterior carga realizar este proceso a los datos que deben ser actualizados o los que no se encuentran en el destino.

En la figura 8 se muestra el proceso ETL anteriormente descrito y el resto de las transformaciones realizadas podrán ser vistas en la dirección:

https://repositorio.datec.prod.uci.cu/svn/cbd_almacenes/Tesis/2009-2010/Hector-Lucas/Artefactos/img%20KTR.

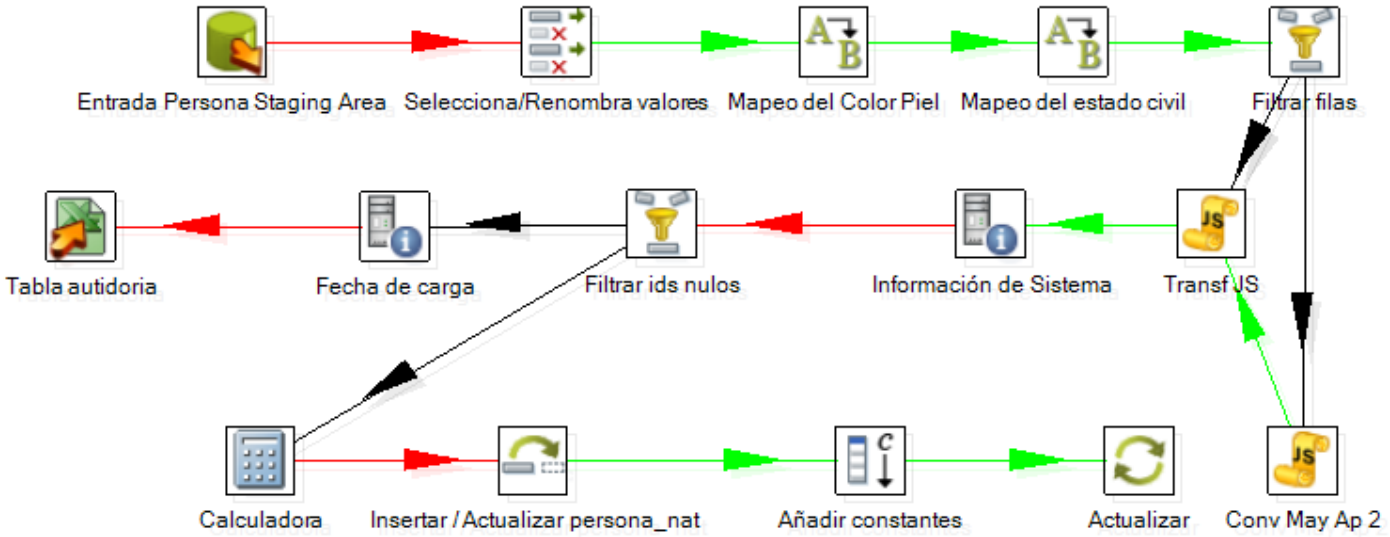

Figura 8: Carga de la tabla Persona en el ODS.

2.8.2 Job para organizar el orden de la carga

Una vez que a los datos se les ha realizado las transformaciones pertinentes, como se explica en el epígrafe 2.7.8 es necesario organizar la carga de las tablas para así evitar desechar datos con llaves nulas solo por el hecho de que esas llaves pertenecen a una tabla que aún no ha sido cargada.

Para evitar estos problemas que pueden aparecer durante la carga en el proceso de Integración de datos, se aplica la transformación Trabajo o Job como es también conocido, mediante los cuales se define el horario y frecuencia de la carga, así como el orden en que van a ser ejecutadas cada transformación para llevar a cabo la carga de los datos. En el caso de este negocio el Job es programado para ejecutarse diariamente a las 5 de la mañana, y el orden en que van a ser ejecutadas las transformaciones es el siguiente: ETL Persona, ETL Dirección, ETL Relación Personal, ETL Viaje Internacional, ETL Vehículo y por último ETL Licencia.

En la siguiente figura se muestra el Job realizado al proceso de integración de Datos anteriormente expuesto.

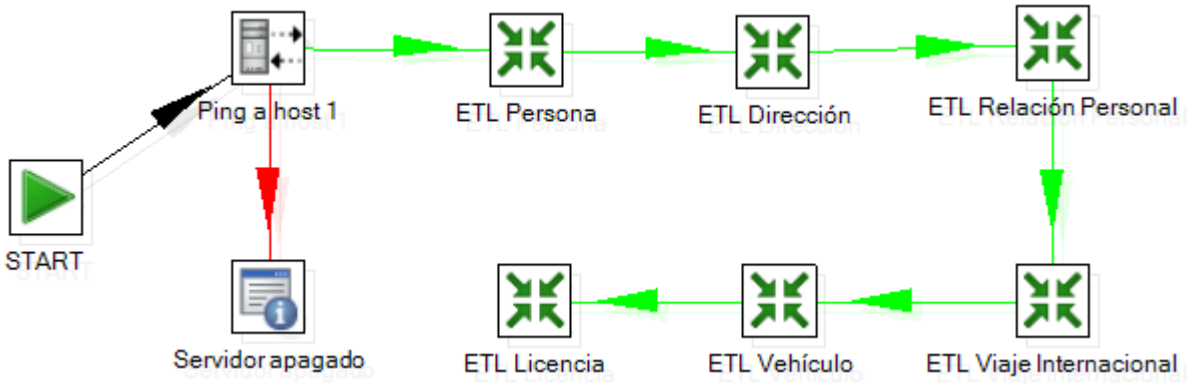

Figura 9: Trabajo o Job para el proceso ETL ODS_CICPC.

2.9 Conclusiones del Capítulo 2

Al finalizar este capítulo se puede concluir que:

- La arquitectura definida es la adecuada ya que permitió diseñar el esqueleto de todo el proceso.

- La comunicación con las fuentes de datos se realiza mediante un FTP para extraer los datos del mismo.
- Se configuró y montó el Área de Preparación de Datos para aprovechar sus potencialidades y facilidades de uso.
- Se implementó el proceso ETL para extraer los datos de las fuentes, limpiarlos, transformarlos y a partir de esto cargarlos al ODS_CICPC.

Capítulo 3: Validación de los resultados obtenidos

3.1 Introducción

En este capítulo se desarrollará el tema de calidad de datos, donde se muestra la evolución del tema, los problemas frecuentes en los datos, las pruebas y validación que se le realiza a la capa de Integración de datos, tratando de que los usuarios alcancen un acercamiento al nivel de conformidad y aceptación del producto.

3.2 Calidad de datos

En este trabajo se integran los datos en el ODS de CICPC que es un repositorio de datos único y centralizado, en el cual se tienen los datos de todos los sistemas externos. Uno de los principales problemas que surgió, fue la inmensa cantidad de datos y metadatos que llegaban al ODS desde estos sistemas, todos ellos con valores en diferentes formatos, cumpliendo diferentes estándares y teniendo diferentes significados dependiendo del contexto en el cual se originaron.

La siguiente lista de problemas en los datos representados en la siguiente tabla ayudará a comprender más la importancia de la calidad de datos y cómo los datos corruptos pueden impactar en las organizaciones (Abella y otros, 2005).

Tabla 7: Posibles problemas de los datos a cargar.

PROBLEMA	EJEMPLO
Falta de estándares	Los datos se representan en múltiples formatos. Por ejemplo: diferentes formas de escribir un nombre.
Información perdida en campos de texto	Información decisiva (identificación de entidades de negocio) es ingresada en campos de texto. Por ejemplo: se tiene un campo descripción y ahí se ingresa la cédula del cliente.
Información no consolidada	Múltiples identificadores de una misma entidad. Por ejemplo: una misma persona con distintos códigos.
Sorpresas de datos dentro de campos individuales	Valores en los datos que escapan de las descripciones de los campos y de las reglas de negocio. Por ejemplo: nombres comerciales mezclados con nombres

	personales, indicaciones de ubicación en campos de direcciones, uso inconsistente del espacio en blanco, caracteres especiales y límites de campos (cortar una palabra en un campo y continuar en el siguiente)
Errores	Dentro de las cuales se encuentran errores tipográficos, faltas de ortografías, valores fuera de rango y tipos de datos incorrectos.
Homónimos	Palabras que se escriben igual y tienen diferente significados, a veces hasta sin relación o conflictivos, y su significado correcto depende del contexto.
Datos que faltan o datos invisibles	Datos con estructura y valor apropiados pueden omitir información inadvertidamente. Por ejemplo: la dirección de una persona "J.M. Pérez 324" puede ser válida pero si es un edificio se estaría omitiendo el número de apto.
Datos fantasmas	En ocasiones se ingresan valores especiales indicando que el campo tiene valor desconocido o que no se utiliza más. Por ejemplo: en un campo de fecha se puede encontrar el valor 99/99/9999.

Se considera que para lograr erradicar dichos problemas se deben poner en práctica los procesos de: limpieza de datos y mejoramiento de la calidad de los datos, en este trabajo estos procesos se realizan con las herramientas DataCleaner en su versión 1.5.3 y el Pentaho Data Integration en su versión 3.1.

- La limpieza de datos es el proceso de llevar a un estado de calidad los datos que serán ingresados en el ODS. Este paso es necesario para evitar la entrada al ODS de datos en mal estado generados durante años por procesos erróneos.
- El mejoramiento de la calidad de los datos a diferencia de la limpieza de datos que apunta a corregir errores busca prevenirlos atacando los problemas de raíz, en la fuente de datos.

Para la correcta aplicación de estos procesos durante todo el período de ETL se deben tener en cuenta la siguiente lista de prácticas (Abella y otros, 2005).

1. Chequeo de versión

El chequeo de versión se realiza para detectar cambios en la especificación de los metadatos.

Un caso de utilidad es cuando se tiene un campo booleano codificado con valores 1 ó 0 y se modifica la codificación a TRUE o FALSE.

El chequeo se realiza apenas se extraen los datos de sus fuentes. La detección a tiempo de estos casos previene la entrada de datos que pueden ser mal interpretados posteriormente en el ODS.

2. Chequeo de uniformidad

Asegura que los valores de los datos que están siendo cargados, se encuentren dentro de límites preestablecidos. Las reglas de negocio determinan qué valores son factibles en cada campo de datos, este chequeo verifica que los datos que se ingresen al ODS cumplan dichas reglas.

3. Transformación de datos

La transformación de datos, convierte los datos a una forma adecuada para el ODS. Si bien los datos de las fuentes pueden ser correctos, por su forma, no siempre son de utilidad al ODS.

Existen dos niveles de transformaciones:

- Acondicionamiento y estandarización de datos: se modifican los datos cuando tienen error o no tiene valores (campos vacíos).
- Utilizar reglas de negocio: aplicar reglas del negocio para transformar los datos adecuadamente.

4. Integración de datos

La Integración de datos tiene como objetivo identificar y consolidar registros.

Implica la tarea de identificar entidades de datos para que no vuelvan a ser cargadas equivocadamente como entidades nuevas al ODS, y de resolver conflictos entre datos provenientes de diferentes fuentes.

5. Supervivencia y formateo de datos

Asegura que los datos más relevantes sean tenidos en cuenta y se encuentren en la forma adecuada. La rutina se encarga de varias tareas:

- Data filling: ingresar valores faltantes en registros remplazándolos con valores de registros relacionados, correspondientes a la misma entidad.
- Resolución de conflictos de datos: resolver problemas de registros múltiples que se refieren a la misma entidad con atributos conflictivos.

- Supervivencia de datos: determinar los datos apropiados a cargar en caso de múltiples posibilidades. La información se puede encontrar en más de un lugar y debe determinarse cuál seleccionar para el ODS.
- Salida de datos: adaptar la salida a los requerimientos técnicos y de negocio. El ODS es consultado utilizando distintas herramientas de consulta y análisis, por lo que sus datos deben poder ser accesibles a las mismas.

6. Chequeo de completitud

Determina que las agregaciones de los datos estén completas y correctas. Las agregaciones son útiles pero pueden ocultar datos importantes.

Un ejemplo es sacar un promedio de ventas a partir de campos con valores nulos, el promedio puede estar bien calculado pero no refleja la realidad correctamente.

7. Chequeo de conformidad

Correlaciona los datos con fuentes de datos estándares. Valida si los datos se adecuan a otras fuentes de datos y reportes. Este chequeo permite descubrir casos excepcionales, donde se debe investigar si la causa de los mismos proviene de datos erróneos o que los resultados reflejan un cambio en la realidad.

Por ejemplo, si el promedio de ventas en un departamento del país se mantiene constante durante todos los meses del año y en determinado mes aumenta al doble, los resultados pueden deberse a datos ingresados en forma errónea por los operadores de los sistemas operacionales o que las ventas en realidad subieron en dicho mes.

3.3 Estrategia de validación y prueba

Una vez realizado el proceso de Integración de datos, se da paso a la validación y prueba de la solución mediante las Listas de chequeo, los Casos de prueba y el Perfilado de datos, para así verificar que el sistema cumpla con los requerimientos necesarios que garanticen al usuario final del sistema la confiabilidad de los datos cargados en el ODS_CICPC.

3.3.1 Listas de chequeo

Se entiende por lista de chequeo a un listado de preguntas, en forma de cuestionario que sirve para verificar el grado de cumplimiento de determinas reglas establecidas a priori con un fin determinado.

El uso de estas listas está generalizado en elementos muy diversos que van desde verificar y determinar el potencial de los artefactos del proceso ETL hasta medir la confiabilidad y seguridad de los datos que se van a cargar.

En este trabajo fueron aplicadas las siguientes listas de chequeo a los artefactos correspondientes del proceso ETL con el fin de evaluar las especificaciones descritas en cada uno, estas listas de chequeo se encuentran en la siguiente dirección: https://repositorio.datec.prod.uci.cu/svn/cbd_almacenes/Tesis/2009-2010/Hector-Lucas/Artefactos/listas%20d%20chequeo/.

- Lista de chequeo de las Reglas del Negocio.
- Lista de chequeo del Mapa Lógico de Datos.
- Lista de chequeo del Diccionario de Datos.
- Lista de chequeo de Registro de Sistemas Fuentes.

3.3.2 Casos de prueba

El propósito de un caso de prueba es especificar una forma de probar el sistema, incluyendo las entradas con las que se ha de validar, los resultados esperados y las condiciones bajo las que ha de probarse. Lo que caracteriza un escrito formal de casos de prueba es que hay una entrada conocida que debe probar una precondición y una salida esperada que prueba una pos condición.

El equipo de prueba debe planificar las pruebas necesarias para validar los requerimientos del sistema. En cada iteración no se validan todos los requerimientos, por lo que hay que centrarse en los críticos o más importantes dentro del ámbito del desarrollo actual.

Se pueden realizar muchos casos de prueba para determinar que un requisito es completamente satisfactorio, uno de ellos debe realizar la prueba positiva de los requisitos y el otro debe realizar la prueba negativa.

Para validar los requerimientos del sistema de este trabajo se le realizan casos de prueba al proceso de Integración de datos, donde primeramente se toman los escenarios que se van a probar (las tablas del origen) y las distintas variables (los atributos críticos de estas tablas) las cuales tomarán los valores de válidos o inválidos en caso de que estén correctos o incorrectos. Estos resultados son los que permitirán dar la respuesta del sistema ante estos posibles valores y el resultado de la prueba.

Teniendo en cuenta que los datos a cargar son muchos, aplicarle casos de prueba a todos sería prácticamente imposible, por lo tanto, para tener una visión general de la calidad de los datos a cargar fue seleccionada una muestra representativa de los más importantes de cada tabla, para aplicarles casos de prueba los cuales arrojaron resultados satisfactorios, ya que el sistema para cada situación responde correctamente. A continuación se muestra un ejemplo de aplicación de un caso de prueba realizado al atributo primernombre de la persona, el resto de los casos de prueba se encuentra en la planilla de casos de prueba en la siguiente dirección: https://repositorio.datec.prod.uci.cu/svn/cbd_almacenes/Tesis/2009-2010/Hector-Lucas/Artefactos/casos%20de%20pruebas/.

Tabla 8: Ejemplo de un caso de prueba

Escenario	primer nombre	primer apellido	segundo apellido	letra cedula	color piel	Fecha nacimiento	pasaporte	Respuesta del sistema	Resultado de la Prueba	Flujo Central
persona	V	V	V	V	V	V	V	El sistema agrega los datos satisfactoriamente.	Satisfactorio.	El sistema extrae los datos de la tabla stg_persona y los carga al ODS_CICPC destino.
	I (YUNIER)	V	V	I (e)	V	V	V	El sistema transforma los datos inválidos.	Satisfactorio.	1-El sistema extrae los datos de la tabla stg_persona. 2-Se transforma el primernombre utilizando el componente de Java Script para convertir su valor que en la fuente puede estar en mayúscula o minúscula pero en el destino debe tener solo la primera letra en mayúscula. 3-Para la variable letracedula inválida se transforma mediante el componente Java Script tratando de que se encuentre en el destino de una única forma. 4-Se procede a cargar los datos transformados al ODS_CICPC.

3.4 Perfilado de datos al ODS_CICPC

Teniendo en cuenta lo antes estudiado acerca del tema de Perfilado de Datos se puede decir que este proceso permite examinar los datos existentes y obtener estadísticas e información sobre ellos, ya que posibilita corregir los datos que tengan problemas como: valores indebidos, escritos incorrectamente, ausentes o duplicados. Se realizó este proceso al ODS_CICPC después de haber realizado los distintos casos de prueba al proceso de Integración de datos, para saber si estos datos fueron cargados correctamente y no tengan ningún problema de los antes mencionados. Obteniendo diferentes reportes donde se evidencia un resultado satisfactorio como se muestra en la siguiente imagen. El resultado arrojado de estos reportes demostró que los datos en el ODS_CICPC destino están correctos y que el proceso de Integración de datos se realizó satisfactoriamente.

String analysis	lugarentrada	ciudadorigen	ciudaddestino	idviajeinternacional	paisorigen
Char count	2001	795	813	192	302
Max chars	22	16	16	3	4
Min chars	18	4	4	1	3
Avg chars	20,01	7,95	8,13	1,92	3,02
Max white spaces	2	2	2	0	0
Min white spaces	1	0	0	0	0
Avg white spaces	1,27	0,24	0,3	0	0
Uppercase chars	4%	12%	12%	0%	97%
Lowercase chars	88%	84%	84%	0%	2%
Non-letter chars	6%	3%	3%	100%	0%
Word count	227	121	127	100	100
Max words	3	3	3	1	1
Min words	2	1	1	1	1

Figura 10: Perfilado de Análisis de cadenas de la tabla ods_rel_fil.

3.5 Auditoría a los datos

Otra forma de garantizar que la información integrada al ODS_CICPC sea confiable es aplicando auditoría a los datos. Auditando los datos se puede tener conocimiento del número total de elementos en la entrada, la cantidad de datos corruptos o no válidos, el total de transformados y cargados, el usuario que ejecuto la transformación, el IP del cual se realizó la transformación, entre otros.

Mediante la aplicación de los metadatos de proceso, se realizó una auditoría a los datos en cada transformación, antes de insertar o actualizar los datos en el ODS_CICPC se realiza un filtrado de los identificadores con que cuenta cada tabla, en caso de que algún identificador no tenga información los datos correspondientes a esa tupla no serán cargados y se enviarán a un documento Excel en conjunto con la fecha en que se realizó la transformación, para luego notificar el incidente a la fuente.

A continuación se ilustra el resultado arrojado luego de ejecutar la transformación ETL Vehículo, en dicha transformación algunos datos no pudieron ser cargados debido a que la persona que es propietaria del vehículo no está registrada en el ODS_CICPC, por tanto, el idods_vehiculo al no encontrarlo ya que se encuentra vacío y se le debe de notificar a la fuente que los datos de esos vehículos no fueron cargados.

placa	idods_persona	cedula	fecha_sys
B1JRK4C		V0652	08/06/2010 0:16
C12X4M		V5893512297	08/06/2010 0:16
ELG5QH		V324425699	08/06/2010 0:16
JER123		V0652	08/06/2010 0:16
JKG126		V5893512297	08/06/2010 0:16
LDX789		V324425699	08/06/2010 0:16
LYQ865		7925881624	08/06/2010 0:16

Figura 11: Auditoría a la tabla ods_vehiculo.

3.6 Conclusiones del Capítulo 3

Al finalizar este capítulo se puede corroborar que se realizó el proceso de Integración de datos correctamente ya que se validó y probó el mismo mediante:

- Las listas de chequeo aplicadas a los artefactos del proceso ETL permitieron determinar la potencialidad de estos, así como la confiabilidad y seguridad de los datos que se van a cargar, la cual arrojó como resultado que las especificaciones descritas en estos artefactos cumplen con los elementos que conforman estas listas de chequeo.
- Los casos de pruebas aplicados al proceso de Integración de datos permitió evaluar si el proceso de integración fue realizado correctamente, arrojando como resultado que los datos disponibles en el destino están preparados, disponibles y organizados.
- El proceso de perfilado de datos realizado al ODS_CICPC permitió examinar los datos existentes y obtener estadísticas e información sobre los mismos, posibilitando corregir los datos que no se hayan cargado correctamente.
- La auditoría de datos permitió notificar a la fuente los datos no cargados.

Quedando en este capítulo validados todos los resultados obtenidos para que la credibilidad de los datos cargados sea un fundamental éxito del proyecto, ya que en ellos se basará el usuario final para tomar sus decisiones.

Conclusiones generales

Al concluir este trabajo se puede plantear que se cumplió con los objetivos y las tareas de la investigación propuestas ya que:

- El estudio de los sistemas fuentes permitió determinar el estado de los mismos y sus necesidades, logrando de esta forma adquirir los conocimientos necesarios para examinar los datos existentes y obtener estadísticas e información sobre estos.
- El análisis de las necesidades de integración en el ODS para el CICPC, permitió definir las reglas de transformación y limpieza de los datos para lograr que los mismos tengan la calidad requerida.
- El análisis, diseño e implementación del proceso de Integración de datos permitió tener en el ODS_CICPC datos limpios, consistentes y centralizados.
- La aplicación de casos de prueba, lista de chequeos, perfilado de datos y auditoría de datos permitió validar el sistema y proporcionó al almacén de datos operacional del CICPC datos preparados, organizados y disponibles.

Recomendaciones

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

- Realizar la integración a CICPC de otras fuentes de datos que complementen la información que éste maneja.
- Incentivar en la Universidad de las Ciencias Informáticas las 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.

Referencias bibliográficas

- *Cicpc.* (2009). Citado el 25 de 01 de 2010, de: http://www.cicpc.gob.ve/index.php
- Kimball, Ralph. *Designing the Operational Data*: Revista DM Review. *The Data Warehouse Toolkit.* s.l: WILEY PUBLICHING, 1997.
- Inmon, William. 1995. *The Operational Data Store.*
- Vidot, Osmara Ramos. 2008. *Almacén de Datos Operacionales: propuesta de formalización del proceso de desarrollo.* Ciudad Habana:
- Kimball, Ralph. *Designing the Operational Data*: Revista DM Review. *The Data Warehouse Toolkit.* s.l: WILEY PUBLICHING, 1996.
- Mustelier, Doris Medina. 2009. *Técnicas de ETL del SINSC en la República Bolivariana de las Américas.* Habana, Cuba: s.n., 2009
- Morgental. 2005. *Enterprise Information Integration: A Pragmatic Approach.* s.l.: Bk&CD, 2005.
- Kimball, Ralph y Caserta, Joe. 2004. *The DW ETL Tolkit Practical for Extracting,Cleaning,Conforming, and Delivered Data.* Canada: Wiley Publishing.
- Syntel, 2007. *EAI vs. ETL: Drawing Boundaries for Data Integration. Citado el 10 de 12 de 2009. Disponible en:*

  http://www.syntelinc.com/uploadedFiles/Syntel/Digital_Lounge/White_Papers/Syntel_EAI_vs_ETL.pdf.
- Hartman Díaz, Yohanlena y Ramón Zequeira, Dailen. 2009. *Implementación del proceso de extracción, transformación y carga en un almacén de datos operacional para CIMEX.* Habana,Cuba. Citado el 10 de 12 de 2009 de: http://bibliodoc.uci.cu/TD/TD_2180_09.pdf
- Microsoft. 2007. *Microsoft ETL Guide.*
- Kioskea. [En línea] 2008. [Citado el: 15 de 02 de 2010.]. Disponible en: http://es.kioskea.net/contents/internet/protocol.php3
- Bernabeu, D. R. (2007). *Metodología propia para la Construcción de un Data Warehouse.* Córdoba, Argentina. Citado el 15 de 02 de 2010 de: http://www.dataprix.com/files/DWH_Metodologia_HEFESTO-V1.0.pdf

- Chrysler, D., & Otros. (agosto de 2000). *CRISP-DM 1.0 Guía paso a paso de Minería de Datos.* Citado el 15 de 02 de 2010 de: http://www.dataprix.com/files/Metodologia_CRISP_DM.pdf
- Decloix, S. (2008). *Les ETL Open Source "Une réelle alternative aux solutions propriétaires".* Citado el 17 de 02 de 2010 de:

  http://www.atolcd.com/fileadmin/Publications/Atol_CD_Livre_Blanc_ETL_Open_Source_01.pdf
- Ferrari, A., & Russo, M. (octubre 1, 2008). *SQLBI METHODOLOGY AT WORK.* Citado el 15 de 02 de 2010 de: http://www.dataprix.com/files/SQLBI Methodology At Work draft.0-1.pdf
- Headquarters. (2009). *Visual Paradigm for UM*. Citado el 17 de 02 de 2010, de: http://www.visual-paradigm.com/product/vpuml
- Knight, B. (2008). *Professional Microsoft SQL Server 2008 Integration Services (Wrox Programmer to Programmer).* Citado el 6 de 02 de 2010 de:

  http://www.ebooks-space.com/ebook/1250/Professional-Microsoft-SQL-Server-2008-Integration-Services.html
- Pentaho. (2009). *Pentaho Open Source Business Intelligence: Kettle Project.* Citado el 25 de 01 de 2010, de http://kettle.pentaho.org
- Sorensen, K. (2009). *Data Cleaner.* Citado el 10 de 03 de 2010, de

  http://datacleaner.eobjects.org/resources/docs/documentation.html
- Talend. (2009). *Talend Open Data Solutions*. Citado el 10 de 02 de 2010, de

  http://es.talend.com/index.php.
- Yglesias, R. (2008). *Oracle vs Oracle.* Citado el 10 de 02 de 2010 de:

  http://www.oracle.com/technology/global/lad-es/documentation/collaterals/BI-Whitepaper-Rodolfo-Yglesias.pdf
- Yobanis Piñero, P., Limia Navarro, A., Hernández Calvo, O., Iznaga González, Y., & Otros. (2009). *Metodología para el desarrollo de Almacenes de Datos y BI.* Ciudad Habana, Cuba.
- *Saime.* [En línea] (2009). [Citado el: 28 de 03 de 2010.] .Disponible en: http://www.saime.gob.ve/index.php

- Adobe.com. [En línea] Adobe, (2010). [Citado el: 20 de 03 de 2010.].Disponible en: http://www.adobe.com/es/devnet/dreamweaver/articles/xml_overview.html
- Inmon, William. 1995. *The Operational Data Store.*
- Inttt. [En línea] 2008. [Citado el: 28 de 03 de 2010.] . disponible en: http://www.intt.gob.ve/quees.php
- Zepeda Sánchez, L. Z. (junio de 2008). *Metodología para el Diseño Conceptual de Almacenes de Datos.* Valencia, España. Citado el 15 de 02 de 2010 de:

  http://dspace.upv.es/xmlui/bitstream/handle/10251/2506/tesisUPV2841.pdf

Bibliografía

- Adobe.com. [En línea] Adobe, 2010. [Citado el: 20 de 03 de 2010.]. Disponible en: http://www.adobe.com/es/devnet/dreamweaver/articles/xml_overview.html
- Bernabeu, D. R. (2007). *Metodología propia para la Construcción de un Data Warehouse.* Códroba, Argentina. Citado el 15 de 02 de 2010. de: http://www.dataprix.com/files/DWH_Metodologia_HEFESTO-V1.0.pdf
- Colectivo de autores. 2010**.** *METODOLOGÍA PARA EL DESARROLLO DE SOLUCIONES DE ALMACENES DE DATOS E INTELIGENCIA DE NEGOCIO EN CENTALAD.* Habana, Cuba: s.n., 2010.
- Colectivo de autores Capri Software. S.L. [En línea] [Citado el: 16 de febrero de 2010.] http://www.capris.es/talend/Folleto Talend Open Profiler.pdf
- Chrysler, D., & Otros. (Agosto de 2000). *CRISP-DM 1.0 Guía paso a paso de Mineria de Datos.* Citado el 15 de 02 de 2010 de: http://www.dataprix.com/files/Metodologia_CRISP_DM.pdf
- *Cicpc.* (2009). Citado el 25 de 01 de 2010, Disponible en: http://www.cicpc.gob.ve/index.php
- Decloix, S. (2008). *Les ETL Open Source "Une réelle alternative aux solutions propriétaires".* Citado el 17 de 02 de 2010 Disponible en: http://www.atolcd.com/fileadmin/Publications/Atol_CD_Livre_Blanc_ETL_Open_Source_01.pdf
- Ferrari, A., & Russo, M. (October 1, 2008). *SQLBI METHODOLOGY AT WORK.* Citado el 15 de 02 de 2010. Disponible en: http://www.dataprix.com/files/SQLBI Methodology At Work draft.0-1.pdf
- Hartman Díaz, Yohanlena y Ramón Zequeira, Dailen. 2009. *Implementación del proceso de extracción, transformación y carga en un almacén de datos operacional para CIMEX.* Habana,Cuba. Citado el 10 de 12 de 2009. Disponible en: http://bibliodoc.uci.cu/TD/TD_2180_09.pdf
- Headquarters. (2009). *Visual Paradigm for UM*. Citado el 17 de 02 de 2010, de http://www.visual-paradigm.com/product/vpuml

- Knight, B. (October 2008). *Professional Microsoft SQL Server 2008 Integration Services (Wrox Programmer to Programmer).* Citado el 6 de 02 de 2010 de: http://www.ebooks-space.com/ebook/1250/Professional-Microsoft-SQL-Server-2008-Integration-Services.html
- *Metodología para el desarrollo de Almacenes de Datos y BI.* Ciudad Habana, Cuba.
- *Saime.* [En línea] 2009. [Citado el: 28 de 03 de 2010.]. Disponible en: http://www.saime.gob.ve/index.php
- Microsoft. 2007. *Microsoft ETL Guide.*
- Morgental. 2005. *Enterprise Information Integration: A Pragmatic Approach.* s.l.: Bk&CD, 2005.
- Pentaho. (2009). *Pentaho Open Source Business Intelligence: Kettle Project.* Citado el 25 de 01 de 2010, de http://kettle.pentaho.org
- Sorensen, K. (2009). *Data Cleaner.* Citado el 10 de 03 de 2010, de http://datacleaner.eobjects.org/resources/docs/documentation.html
- Talend. (2009). *Talend Open Data Solutions*. Citado el 10 de 02 de 2010, de http://es.talend.com/index.php.
- Vidot, Osmara Ramos. 2008. *Almacén de Datos Operacionales: propuesta de formalización del proceso de desarrollo.* Ciudad Habana.
- William. 1995. *The Operational Data Store.*
- Inttt. [En línea] 2008. [Citado el: 28 de 03 de 2010.] . Disponible en: http://www.intt.gob.ve/quees.php
- Kimball, Ralph. *Designing the Operational Data*.: Revista DM Review.
- —. 1997. *The Data Warehouse Toolkit.* s.l.: WILEY PUBLICHING, 1997.
- —. 1996. *The Data Warehouse Toolkit.* s.l.: WILEY PUBLICHING, 1996.
- Kimball, Ralph y Caserta, Joe. 2004. *The DW ETL Tolkit Practical for Extracting,Cleaning,Conforming, and Delivered Data.* Canada: Wiley Publishing.
- Yglesias, R. (Setiembre 2008). *Oracle vs Oracle.* Citado el 10 de 02 de 2010. Disponible en:

http://www.oracle.com/technology/global/lad-es/documentation/collaterals/BI-Whitepaper-Rodolfo-Yglesias.pdf

- Yobanis Piñero, P., Limia Navarro, A., Hernández Calvo, O., Iznaga González, Y., & Otros. (2009). Inmon,
- 2008. Kioskea. [En línea] 2008. [Citado el: 15 de 02 de 2010.]. Disponible en:

  http://es.kioskea.net/contents/internet/protocol.php3
- Zepeda Sánchez, L. Z. (junio de 2008). *Metodología para el Diseño Conceptual de Almacenes de Datos.* Valencia, España. Citado el 15 de 02 de 2010. Disponible en:

  http://dspace.upv.es/xmlui/bitstream/handle/10251/2506/tesisUPV2841.pdf

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.

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.

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.

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.