Universidad Autónoma del Estado de México

Centro Universitario UAEM Zumpango

Gestión de datos empresariales utilizando procesos ETL

ENSAYO

que para obtener en título de

INGENIERO EN COMPUTACIÓN

presenta:

Tonantzin Martínez Trujillo

Dr. en C. C. Asdrúbal López Chau

Asesor

Zumpango, Estado de México Septiembre, 2018

AGRADECIMIENTOS

En primera instancia agradezco a mi asesor, Dr. Asdrúbal López Chau, que siempre se mantuvo disponible el momento que lo necesitaba, por aclarar cada una de mis dudas e inquietudes y por todo el esfuerzo que puso en conjunto para tener cada sesión, revisión y/o comentarios. Y el principal motivo por confiar en mí, y motivarme para poder culminar exitosamente este proyecto. También agradezco a mis revisores, M. en C. María Guadalupe Domínguez Urban y M. en C. José Manuel Trujillo Lara, por sus acertados comentarios que ayudaron a mejorar este trabajo.

A mi madre, que siempre me impulsó a luchar por mis sueños, por darme siempre lo mejor y sacarme adelante durante el tiempo que estuve a su lado. Te amo mamá.

DEDICATORIAS PERSONALES

A mi madre Verónica.

Por haberme apoyado en todo momento, por cada uno de sus consejos, sus valores, por la motivación constante que me ha permitido hasta el momento ser una persona de bien, pero más que nada, por todo su amor.

A mis amigos.

Por su apoyo incondicional y cada uno de sus consejos.

A mis maestros.

Al Dr. Asdrúbal López Chau, por la motivación para la elaboración y conclusión de este ensayo, y a cada uno de mis maestros que marcaron cada etapa de mi camino universitario compartiendo su conocimiento día a día.

CONTENIDO

LISTA DE FIGURAS

LISTA DE TABLAS

1. INTRODUCCIÓN

Las empresas más rentables y con mayores beneficios en el mercado, son aquellas que reconocen el valor estratégico de los datos, además de que estos son utilizados para competir y ayudar. En la actualidad, la aparición de nuevas tecnologías y la explotación de datos plantean un gran desafío, todo esto genera una mayor complejidad al momento de gestionar y confirmar la calidad y claridad de los datos [1]. En mi experiencia personal, actualmente en México la mayoría de las empresas todavía no explotan los datos para mejorar su competitividad, incrementar sus ingresos o gestionar más apropiadamente sus recursos. Sin embargo, con el aumento de nueva tecnología existen posibilidades de realizar grandes cambios positivos a la organización.

Dentro del mercado mundial actual, los líderes empresariales persiguen un solo objetivo, y es tan solo convertir la información extraída en resultados importantes, es decir, que estos sirvan para el crecimiento de la empresa. Aquellos líderes con mayor éxito aplican análisis en las principales actividades de la organización para que la toma de decisiones sea rápida, sobresaliente y con esto se tenga oportunidad de optimizar resultados.

La agregación inteligente de los datos facilita la habilidad del negocio. Los miembros de una organización requieren de datos y la información necesaria para desempeñar su trabajo, por esta razón, proporcionar los datos correctos a las personas indicadas en el momento que lo requieran, permite satisfacer una necesidad básica pero importante en la empresa. Pese a que en varias partes del mundo se está haciendo cada vez más aceptada esta práctica, es todavía común encontrar una resistencia en varias empresas, sobre todo en países subdesarrollados y no desarrollados.

Hoy en día no es fácil alcanzar la excelencia en la integración de datos, las empresas manipulan un entorno difícil y variable, en donde la flexibilidad, puede también llegar a ser un contra, sin embargo, se puede trabajar para que esto no sea una problemática mayor. En mi punto de vista los líderes deben confiar en el uso de nuevas tecnologías para una buena integración de datos de manera segura y eficaz, lo cual permita el cambio de procesos que

probablemente ya sean obsoletos a aquellos procesos que además de ser seguros opten con eliminar la mayor parte de manualidades.

A continuación se enlistan algunas deficiencias que enfrentan la mayoría de las empresas modernas [2]:

- **Tienen datos, pero carecen de información** - Almacenar los datos no es suficiente para que la empresa u organización tenga mayor competencia. Para esto necesita profundizar el nivel de conocimiento de clientes y empleados, para así, tener la capacidad de encontrar motores de comportamiento, monitorear, administrar y además de tener la habilidad de responder cada una de las interrogantes que permitan aumentar el rendimiento de dicha organización. Parecería que aunque las empresas manejan sistemas de información para varias partes de sus procesos, los datos almacenados son solamente "cementerios de datos", es decir, aquellos datos que no tienen ninguna utilidad futura y que únicamente consuman recursos.
- **Fragmentación** - Todas las áreas o departamentos cuentan con aplicaciones independientes, pero carecen de una visión global de la empresa. Es común que en las empresas medianas o pequeñas usen aplicaciones comerciales para llevar la contabilidad, inventario o cartera de clientes; pero todas ellas se encuentran desarrolladas por diferentes empresas, por lo que la integración es casi imposible debido a que no existe una homologación de información.

- **Manipulación manual** – Para la mayoría de las organizaciones es de suma importancia realizar un análisis a través de reportes según el requerimiento del negocio. Dicha práctica lleva a la organización a exportar datos a distintas herramientas que resultan en procesos lentos, costosos, dualidad en trabajo, baja confiabilidad en los informes propensos a errores. Lo anterior limita a las organizaciones a seguir una línea homogénea de la información, limitar a los recursos a realizar una única actividad por tener procesos que consumen gran parte de su jornada, entro otras limitantes.

- **Baja agilidad** – Como consecuencia de la carencia de información, fragmentación y la manipulación manual, la empresa se mantiene en un nivel de rendimiento bajo.

Como conclusión a esto, se requiere de una herramienta ágil que se ajuste a las necesidades del negocio y que permita notar un crecimiento a la organización.

Los procesos ETL (Extraer, Transformar y Cargar, por sus siglas en inglés: Extract, Transform and Load) son procesos que permiten a una empresa u organización manipular datos; es decir, extraerlos desde un sistema origen, transformarlos y cargarlos en un sistema destino. En la sección 2.2 de este documento se podrá observar con más detalle la definición de un proceso ETL.

A una empresa u organización le beneficia hacer uso de procesos ETL para trasladar de un sistema origen a un sistema destino y transformar los datos que maneja, principalmente por los siguientes motivos [3] [10]:

I. Crear un *Master Data Management (MDM)*, el cual es identificado como un repositorio maestro que permite a una empresa u organización relacionar todos sus datos críticos, simplifica el intercambio de datos entre departamentos si este fuera el caso. Permite a las organizaciones a mejorar la calidad de sus procesos [4].
II. Funciona para integrar sistemas. Esto surge cuando las organizaciones crecen y por ende se van añadiendo más fuentes de datos, provocando nuevas necesidades (Sistemas *Legacy* o Sistemas heredados*)*.
III. Facilita a los directivos o analistas a tomar decisiones fundamentales basadas en el análisis de los datos que son cargados y actualizados en nuevas bases. Como un *Datamart* o un *Data Warehouse*.
IV. Obtener una visión global de todos los datos concentrados en *un Data Warehouse*. Es decir, tener la diversidad y saber que en él se encuentran las posibles áreas de mejora.

Una vez referenciado lo anterior, se hace palpable la necesidad de hacer uso de procesos ETL, considerando que son muy útiles y con grandes beneficios para las empresas u organizaciones por la gran capacidad para poder integrar grandes bases de datos, alcanzando una visión única global. Aunado a esto, surge la siguiente pregunta:

¿Porque ETL es importante para las empresas?

Las empresas han confiado en los procesos ETL durante varios años para obtener una visión consolidada de los datos que impulsa a tomar mejores decisiones comerciales. Hoy en día, este método de incorporación de datos de diversos sistemas y fuentes, sigue siendo un componente central de la caja de herramientas de integración de datos de una organización. Cabe mencionar que este es un punto de vista personal de acuerdo a la experiencia con el manejo de dichos procesos en diversos sectores laborales, además de la la investigación que se sostuvo durante la elaboración de este documento

Los procesos ETL no solo funcionan para la carga de datos hacía repositorios de un *Datamart* o *DataWarehouse,* la cual estima un 70% del trabajo del desarrollo*,* sino que es una pieza importante para el éxito cuando se habla de un proyecto BI (*Business Intelligence*, por sus siglas de Inglés) [30]. En un apartado posterior se explicará la relación de los procesos ETL y los proyectos BI.

Suponiendo que existe una base de datos a la que se quiere almacenar información y dicha información es proveniente de archivos o ficheros con un tamaño enorme (por ejemplo; la información diaria de alta de usuarios de alguna página web). Como opción sería cargar los datos de los archivos a tablas en una base de datos correspondiente y luego procesarlos y cargarlos. Si el proceso es pesado, se estarían consumiendo recursos que pueden ser útiles para otro u otros objetivos. Los procesos ETL evitan sobrecargar de procesamientos los sistemas destino de la información [31].

Aunque la aplicación de los procesos ETL es la construcción y carga de un *Data Warehouse,* se trata de procesos cada vez más comunes y más utilizados, ahora frecuentemente los procesos ETL son usados también para integrar y migrar datos. Si no existiera la posibilidad de utilizar los procesos ETL para integración de datos, no habría otra posibilidad más que realizarlo de forma manual y con una cantidad considerable de errores. Para el caso de la migración, únicamente se trata de trasladar la información de una base de datos obsoleta a una nueva.

2. DESARROLLO

2.1. Reseña histórica de los ETL

Antes que los procesos ETL existieran como concepto, y llegarán a ser más comunes y útiles para la manipulación de la información, se hacía referencia a ellos como procesos de extracción de datos, procesos de trasformación de datos y en su caso, procesos de carga de datos, todos estos por separado. Sin embargo, tiempo atrás, eran referenciados también como gestión de metadatos, los cuales describían la información de los mismos, es decir, datos que describen a otros datos. Los metadatos se caracterizaban por ser datos altamente estructurados, y como ya se mencionó anteriormente describían las características de los datos, contenido, calidad, información, además de atributos. De igual manera, estos procesos podían ser referenciados como servicios de administración u operacionales [6] [7].

En la década de 1970, ETL llegó a ser popular cuando las organizaciones comenzaron a utilizar depósitos de datos múltiples para almacenar diferentes tipos de información. ETL se convirtió en el método estándar para tomar datos de distintas fuentes y transformarlos antes de cargarlos en un destino [8].

Con el paso del tiempo, la cantidad de formatos de datos y sistemas se ha modificado y expandido gradualmente. Extraer, Transformar y Cargar ahora es solo uno de varios métodos que las organizaciones utilizan para recopilar, importar y procesar datos.

Al igual que con otras tecnologías, puedo observar una evolución natural de los procesos ETL, desde la creación de herramientas muy básicas, pasando por ser una moda usada por algunas empresas, hasta ser una necesidad, con tecnologías muy poderosas.

2.1.1. Codificación antes del uso de herramientas ETL

En los primeros días de la informática, antes de que existieran las herramientas ETL, la única manera de obtener datos de fuentes distintas y unirlos de una u otra forma, era codificando manualmente las secuencias de comandos en lenguajes, como COBOL y RPG. Aunque esto fue conocido como la primera generación de soluciones ETL, en la actualidad puede

sorprender que el 45% de todo el trabajo de ETL se realice mediante el uso de programas / scripts codificados a mano. Estos programas tenían una etiqueta de precio, sin embargo, ahora existen herramientas de código abierto y algunas licencias con costo accesible para las organizaciones [9] [10].

La segunda generación intentó superar algunas debilidades que se localizaron al codificar a mano, algunas de ellas como; ser propenso a errores, lentitud en términos de tiempo de desarrollo, manejo de errores consistentes, entre otros. Entonces fue así que se generó el código requerido en función del diseño de un flujo ETL [9].

A principios de la década de 1990, surgieron productos como Prism, Carlton y ETI, estos productos en su mayoría fueron adquiridos mucho después por proveedores de ETL. En esos primeros días ETI era el único programa con proveedor independiente y que aun ofrece una solución de generador de códigos [9]. Buscando en la Web y libros, encontré productos con nombres similares, pero no hay certeza de que se trate de los mismo softwares mencionados anteriormente.

La tercera generación de herramientas ETL surgió poco después de que los generadores de código entraran en uso y que les encontraran una desventaja mayor, en donde la mayoría de los generadores podían ser funcionales únicamente con un conjunto limitado de base de datos. Dichas herramientas se basaban en un motor donde se realizaba todo el procesamiento de datos y un conjunto de metadatos antes mencionados, los cuales aprovisionaban cada una de las reglas de conexión y transformación [9].

Después de esa fase de diseño, el esquema de destino debía relacionarse con los esquemas de origen. Todo el proceso requería de mucho tiempo y como resultado de esto, surgió una nueva generación de herramientas de *Data Warehouse* impulsadas por modelos [10].

2.2. ETL en la actualidad

Antes de hablar de cómo los procesos ETL han sido empleados radicalmente en la actualidad. Es importante plasmar la definición de un proceso ETL. Para esto, se muestran algunas definiciones que son obtenidas de diversas fuentes de información, para después encontrar un concepto general que sea claro y conciso.

La empresa de *Power Data*, especialistas en Gestión de Datos, define a los procesos ETL como un término estándar que se utiliza para referirse al movimiento y transformación de datos**.** Se trata del proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y cargarlos en otra base de datos (denominada *DataMart* o *DataWarehouse*) con el objeto de analizarlos**.** También pueden ser enviados a otro sistema operacional para apoyar un proceso de negocio [3].

Intertek, empresa que ofrece evaluaciones de software para los mercados globales, menciona que son una parte de la integración de datos, pero es un elemento importante cuya función completa el resultado de todo el desarrollo de la cohesión de aplicaciones y sistemas [29].

ETL *Tools*, define a ETL como el proceso que organiza el flujo de los datos entre diferentes sistemas en una organización y aporta los métodos y herramientas necesarias para mover datos desde múltiples fuentes a un almacén de datos, reformatearlos, limpiarlos y cargarlos en otra base de datos o almacén de datos. ETL forma parte de la Inteligencia Empresarial (*Business Intelligence*), también llamado "Gestión de los Datos" (*Data Management*) [13].

Dataprix, portal de referencia sobre software para las empresas, menciona que es un proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y cargarlos en otra base de datos, para analizar o en otro sistema operacional para apoyar un proceso de negocio.

Los procesos ETL también se pueden utilizar para la integración con sistemas heredados (aplicaciones antiguas existentes en las organizaciones que se han de integrar con los nuevos aplicativos. La tecnología utilizada en dichas aplicaciones puede hacer difícil la integración con los nuevos programas) [15].

Sin más, es momento de definir a los procesos ETL como aquellos procesos que permite a una empresa u organización manipular datos; es decir, extraerlos desde un sistema origen, transformarlos y cargarlos en un sistema destino. Esto permite eficientar tareas y obtener análisis de calidad, alcanzando objetivos empresariales y mejora en la toma de decisiones.

Aunque actualmente existe un sin fin de fuentes de información que definen y hablan de procesos ETL, para la elaboración de este documento se consideraron aquellas fuentes que personalmente mostraron definiciones interesantes y precisas.

Una vez abordada la definición, es importante conocer aspectos importantes que engloban a dichos procesos. La Inteligencia de Negocio/Empresarial o *Business Intelligence,* se menciona desde un punto de vista pragmático por su relación con estos procesos.

Se define a los *Business Intelligence* como un conjunto de aplicaciones, procesos y tecnologías que permiten reunir, depurar y transformar datos en información estructurada para su análisis, dicho análisis, dirigido a un plan o una estrategia comercial [11]. A mi punto de vista la Inteligencia de Negocio, en español, es aquella en donde las personas hacen uso de la tecnología para la buena toma de decisiones.

Aunado a esto, puedo puntualizar que los procesos ETL nutren a los sistemas BI, ya que tiene que traducir de uno o varios sistemas operacionales independientes a un único sistema desnormalizado, cuyos datos estén completamente integrados. Las organizaciones, especialmente las más robustas, utilizan varios sistemas y estos se nutren de una gran variedad de fuentes de datos, los cuales funcionan en forma de retroalimentación mutua [18].

Los sistemas BI permiten soluciones de arquitectura, dicha arquitectura se puede observar en la Figura 1.1, en donde los objetos de la izquierda representan las distintas fuentes de datos, ya sean internas o externas, el objeto que sigue representa el proceso de extracción, transformación y carga (ETL). En este proceso se definen que campos se van a utilizar de las fuentes heterogéneas, dentro de los sistemas BI se conoce como *"mapping"*. El tercer objeto representa el almacén de datos, en donde se encuentran los datos transformados, representados visualmente en modelos multidimensionales, dimensionales y tablas de datos. Por último se encuentra la interface de acceso al usuario, que permite interactuar con los datos, representar de manera gráfica resultados de consultas, entre otras acciones [2].

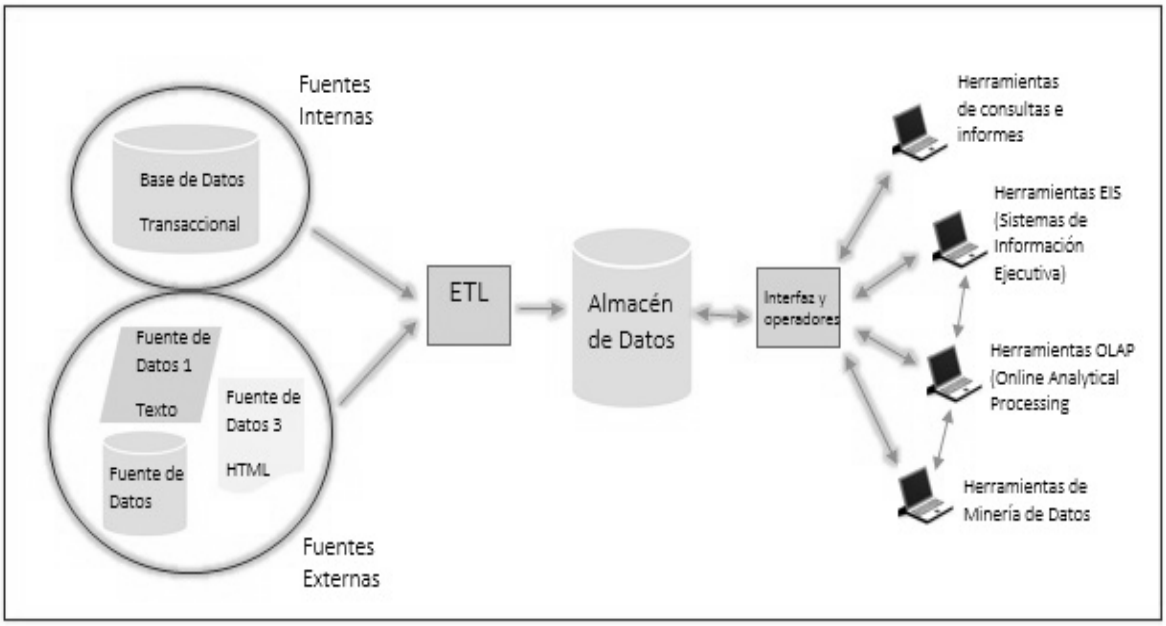

Figura 1.1 Arquitectura de inteligencia de negocios

Lo que muestra la figura anterior de manera más clara son cada uno de los niveles, desde la actualización de datos hasta la presentación y análisis de resultados.

Los datos actuales de alta velocidad se pueden capturar y analizar sobre la marcha a través de análisis de transmisión. Este enfoque presta oportunidad de actuar de inmediato, en función de lo que sucede en un momento determinado. Pero la visión histórica que ofrece ETL pone los datos en contexto. Esto a su vez, las organizaciones obtienen una compresión completa del negocio a lo largo del tiempo.

En la actualidad varias organizaciones hacen uso de procesos ETL para mover datos desde múltiples fuentes, transformarlos y cargarlos en una base de datos destino, o en otro sistema operacional para apoyar el proceso de negocio.

Una vez puntualizado la relación entre los sistemas BI y los procesos ETL, se mencionaran algunas de las características principales que identifican a los procesos ETL, así como algunas de las herramientas más comunes empleadas por las organizaciones.

2.3. Características de los ETL

Como ya se mencionó anteriormente, los procesos ETL son definidos como procesos que permiten a las organizaciones trasladar (mover) datos desde múltiples fuentes, transformarlos y cargarlos en una base de datos destino con el objetivo de analizar la información. El término proviene de las siglas en ingles *Extract-Transform-Load* (extraer, transformar y cargar). De igual manera los datos pueden ser transferidos a otro sistema operacional para apoyar el proceso del negocio [6].

Las fases o secuencias de un proceso ETL son definidas a continuación [2] [13] [10]:

- **Extracción:** de la información de uno a varios sistemas fuente. Esta fase convierte los datos a un formato homogéneo y consolidado para iniciar la fase siguiente. Uno de los requerimientos para esta secuencia, es que cause el menor impacto en el sistema origen. Si los datos a extraer son muchos, el sistema origen se podría ralentizar e incluso colapsar, provocando perdida de información. Por esta razón, en sistemas grandes las operaciones de extracción suelen programarse en horarios o días en donde el impacto sea nulo o mínimo. Para aquellas organizaciones que manejan sistemas de información con una gran cantidad de datos, es recomendable panificar ejecuciones automáticas nocturnas o en días en que la infraestructura no se encuentra sobresaturada con algunos otros sistemas departamentales. Lo anterior con la finalidad de no tener perdida de información o tener que duplicar ejecuciones por excepciones.

  Más adelante se detallaran ejemplos reales del uso de estos procesos.

  Para llevar a cabo de manera correcta el proceso de extracción, hay que seguir los siguientes pasos:

  - Extraer la información desde el/los sistemas origen.
  - Analizar la información, realizando una revisión previa.
  - Interpretar la revisión previa y verificar que la información extraída cumple con lo esperado.

- Convertir los datos a un formato requerido para iniciar con la siguiente fase.

Modos de Extracción

Según la necesidad de la organización es que se determinará la elección de una u otra forma.

Full Extract (Extracción Total)

Consiste en extraer la totalidad de los datos, para esto, se barren las tablas completas que pueden contener millones de registros.

Incremental Extract (Extracción Incremental)

El procesamiento se realiza por lotes, lo que fue agregado o modificado. De igual manera, pueden existir filas que tengan que ser eliminadas por duplicidad, tratarse de datos erróneos, entre otras.

UpdateNotification (Notificación de Actualizaciones)

Se van extrayendo los datos a medida que se produce una actualización (*insert/update/delete*).

Estos tipos de extracción son manipulados por el módulo denominado *Change Data Capture* (CDC). Este es muy utilizado en ambientes *Data Warehousing*, permitiendo el control de cambios que ocurren en una tabla y como resultado entrega los cambios de una forma rápida. Existen dos formas de implementar un módulo CDC; Síncrona y Asíncrona, la primera por cada control de cambios ejecutada, la información es capturada por *triggers* o disparadores. Y la forma Asíncrona; por cada escritura realizada, Oracle captura los cambios sobre la tabla publicadora [14].

- **Transformación:** de datos, es decir, posibilidad de reformatear y limpiar estos datos cuando sea necesario. En esta fase se aplica una serie de reglas de negocio o funciones sobre la información extraída para convertirlos en datos que después serán cargados.

Esta fase depende de lo que la organización permita y requiera, por lo cual el principal responsable de que esta fase manipule los datos de manera correcta es el desarrollador o analista del proceso ETL en acción.

A continuación enlistan algunas de las acciones más habituales dentro del proceso de transformación:

- **Reformateo de datos.**
- **Conversión de unidades.** Por ejemplo, la conversión de diferentes tipos de monedas (dólares, euros, pesos) en un valor estándar.
- **Agregación de columnas.** Agregar una columna con datos que contengan determinada información, por ejemplo, para el sector financiero existen clientes en diferentes situaciones de crédito como son; normal, extinguir, afianzar, reducir, etc. Agregar aquellos registros que contengan clientes con situación "normal".
- **Selección de columnas para su carga posterior.** Un ejemplo de esto sería que columnas con valores = "0" no se consideren para su carga posterior.
- **Dividir una columna en varias.** Es decir, si tiene un campo llamado "dirección" la cual trae información como, calle, municipio y estado. Con las instrucciones necesarias, dividir esa columna con la misma información pero en columnas independientes, teniendo "Calle", "Municipio" y "Estado".
- **Unir datos de varias fuentes.** Se pueden tener información desde un fichero hasta una base de datos local, etc.
- **Traducir códigos.** Es decir, si la fuente origen guarda un dato con un estado "seguir" y otro como "extinguir", realiza las instrucciones requeridas para que la fuente destino almacene un "1" para el primer estado y "0" para el segundo respectivamente.
- **Obtener nuevos valores calculados.** Probablemente se requieran instrucciones que realicen el cálculo de una media o cualquier otro tipo de operación.

Dentro de la fase de transformación existe un concepto llamado Calidad de Datos o *Data Quality* que más adelante se describirá.

- **Carga:** de datos en otro lugar o base de datos, con el objetivo de analizarlos o apoyar un proceso de negocio. En esta fase, la información que proviene de fase anterior son cargados en el sistema destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones. Por ejemplo, en algunas bases de datos será necesario sobrescribir la información antigua con nuevos datos, mientras que en otras, bastaría con resumir las transacciones y almacenar una magnitud considerada. En otras palabras la fase corresponde a la carga de información en los repositorios de información destino.

  Para esta tercera fase, existen dos formas básicas de desarrollar el proceso de carga:

  - Rolling: Para esta forma se almacena información resumida a distintos niveles, correspondientes a distintas agrupaciones o diferentes niveles jerárquicos en alguna o varias dimensiones.
  - Acumulación simple: consiste en hacer un resumen de todas las transacciones comprendidas en un periodo seleccionado y trasportar el resultado como una única transacción hacia el *Data Warehouse*.

  Para asegurar que la carga se realiza de forma correcta se mencionan algunas claves:
  - Considerar la calidad de la carga antes que la velocidad.
  - Asegurar la consistencia de los datos que se están cargando.
  - Utilizar la menor cantidad de recursos disponibles.

  De acuerdo a la forma de implementar la fase carga, es importante considerar que esta fase interactúa directamente con la base de datos destino como ya se ha mencionado. Por dicha razón, al realizar esta operación se aplicarán todas las restricciones que ya hayan sido definidas. Si están bien definidas, la calidad de los datos en el proceso ETL estará garantizada.

Ahora ya detalladas cada una de las fases de un proceso ETL, se hace palpable mencionar una última consideración, la cual consta en la limpieza de datos. Es una etapa previa y separada del proceso, lo que no significa que tenga menor importancia.

Para esto se enlista algunas ventajas de aplicación:

- Evita la información errónea.
- Garantiza la calidad de los datos que se van a procesar.
- Economiza costes de espacio de disco al eliminar información duplicada o no funcional.
- Agiliza las consultas por la ausencia de datos duplicados o inservibles.
- Contribuye a tomar decisiones correctas.

Fases de un proceso ETL:

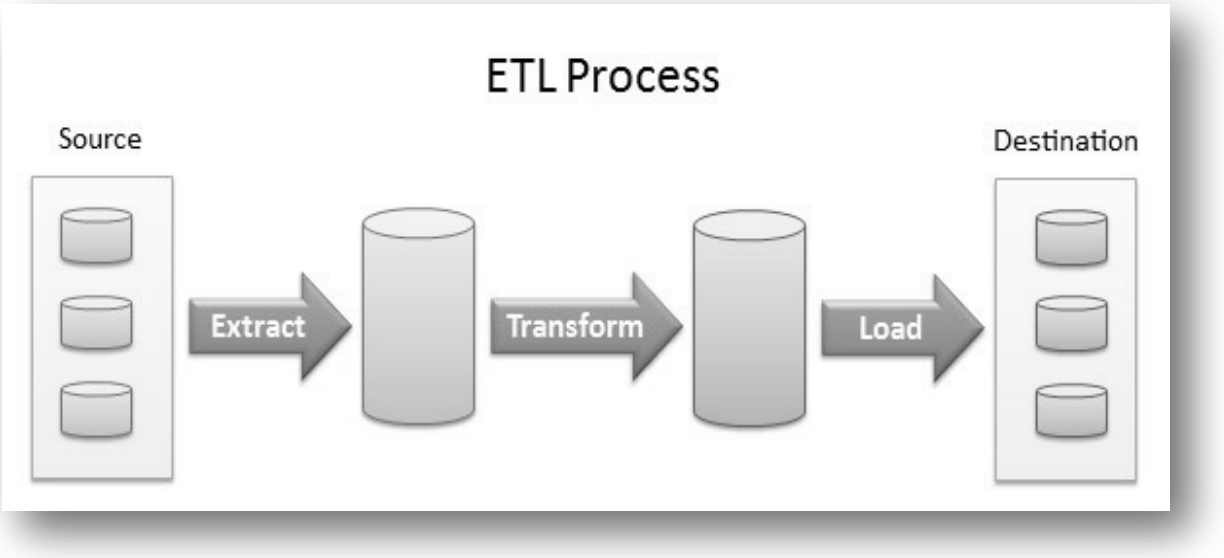

Figura 1.2 Fases de un proceso ETL.

Posibles fallas de un proceso ETL [5]:

- × **Existencia de campos o valores nulos.** En algunas ocasiones cuando se realiza el proceso de extracción, se desconoce el tipo de datos, la longitud y en cierta forma la información contenida. Este fallo no solo puede ser ocasionado en el proceso de extracción sino también en el proceso de carga. Para esto es conveniente identificar cada una de las excepciones o manejar un log de errores.
- × **Tablas de referencia inexistentes.** Puede ocurrir que durante la ejecución de los procesos no se encuentre la tabla referida por algunas circunstancias; alguien cambio el nombre de la tabla o simplemente fue eliminada accidentalmente.
- × **Cortes de energía.** Si los procesos no son monitoreados por algún especialista, este problema retrasará el trabajo planeado principalmente en jornadas nocturnas. Es importante considerar este tipo de fallos y evitar perdida de información, tiempo, etcétera.
- × **Fallos de funcionamiento en los discos de almacenamiento.** Si bien, no se tiene un control sobre los espacios disponibles en discos, surgirá esta falla recurrentemente, sin embargo se puede evitar manteniendo un control sobre la infraestructura.

Como minimizar fallas:

- ✓ Diseñar procesos robustos, es decir, en caso de fallos, los procesos sean recuperados.
- ✓ Garantizar que el diseño de los procesos, minimice los fallos considerando los puntos anteriores.

Por lo anterior, se puede observar que los fallos no siempre se pueden evitar pues muchas veces son ocasionados por circunstancias que salen de nuestras manos, como se menciona en uno de los puntos anteriores, como es un corte de energía, sin embargo, se pueden tomar medidas que logren minimizar fallos y los datos puedan ser recuperados.

Aspecto relacionado con el gerenciamiento y la recuperación

Existe un concepto que hace referencia a la recuperación de datos conocido como *Staging* o salvaguarda de los procesos ETL, para lo cual tiene como objetivo minimizar los posibles fallos en el proceso de carga, normalmente se reserva un área de disco para recuperar los datos por etapas, conocido también como gerenciamiento.

El funcionamiento del *Staging* surge de la siguiente manera:

- Como primer instancia, los datos son volcados (término utilizado para depurar un programa que ha finalizado su ejecución incorrectamente), por etapas o bloques de forma independiente en un área del disco conocido como *staging area*.
- Posteriormente los datos contenidos en el *staging area* se carga a lugar correspondiente o sistema destino como puede ser un *Data Warehouse.*

Algunas ventajas de utilizar *Staging area* [32]

- Permite tener un proceso de carga por etapas o bloques de forma independiente. Esto es muy práctico cuando se trabaja con millones de datos, ya que evita tener que reiniciar el proceso completo en caso de fallo.
- Si el *staging area* se implementa de una manera correcta, cabe la posibilidad de reiniciar por las distintas etapas del proceso ETL de manera independiente. Es decir, si falla únicamente la fase de transformación, bastaría con reiniciar dicha fase y no regresar desde la fase anterior, fase de extracción.
- Al tratarse de un almacenamiento en un disco físicamente independiente, en ningún caso afecta a otros procesos del sistema.
- Por temas de seguridad y confidencialidad de los datos el desarrollador del ETL deberá ser el único en tener acceso al *staging area,* esto para evitar complicaciones y generar alguna otra incidencia.
- El *staging* se utiliza tanto durante el proceso de extracción-transformación como el de transformación-carga.

2.4. Conceptos Importantes

2.4.1. Data Quality

La calidad de datos (*Data Quality, en inglés)* es una característica que determina la integridad de los datos para la toma de decisiones. Para las operaciones y la toma de decisiones es importante contar con la información precisa, además de que son importantes para motivos técnicos [24] [25].

En el momento del análisis de datos es esencial asegurar la calidad de los datos por distintas razones que apuntan a aspectos capitales para la organización, entre otros, el conseguir el mayor potencial de la información. Para que los datos tengan calidad, estos deben reunir los requisitos como; la consistencia (estable, coherente), precisión (ajuste completo o fidelidad de un dato), integridad (calidad) o evitar de alguna manera duplicidad. En la Figura 1.3 se pueden observar las fases que corresponden a la calidad de datos.

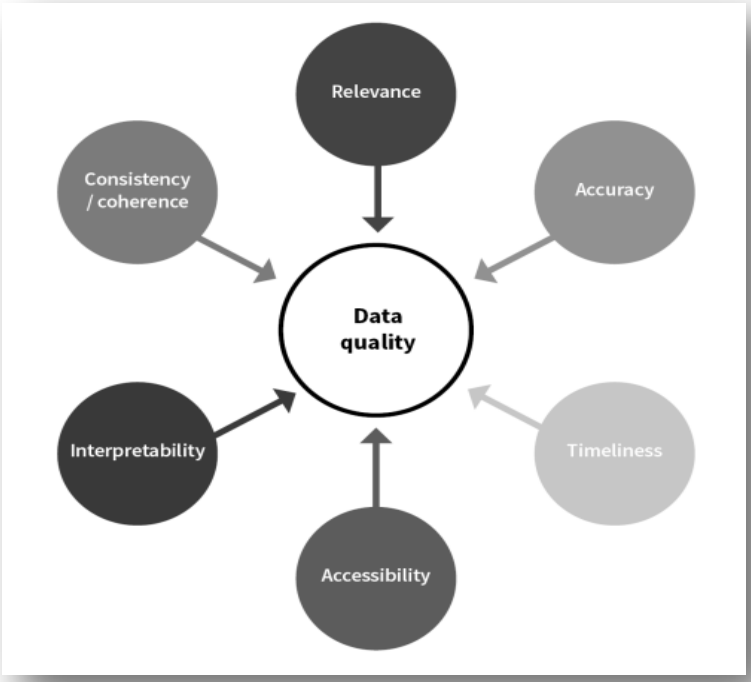

Figura 1.3 *Data Quality Process*.

En esta sección se le da el énfasis necesario a la calidad de datos, por considerarse de suma importancia para gestionar datos empresariales, para que ayuden a la toma de decisiones estratégicas. El origen de una mala calidad en los datos puede tener diversas causas, ya que los datos incorrectos y otros problemas de calidad de datos y/o procesos. La agregación de datos de diferentes fuentes que usan estándares de datos diferentes, puede dar como resultado datos inconsistentes, al igual que la sobre escritura de datos históricos. Los datos incorrectos afectan la capacidad de una empresa u organización para realizar alguna de sus funciones comerciales y proporcionar servicios a sus clientes, lo que resulta en una pérdida de credibilidad e ingresos, insatisfacción de los clientes y problemas de incumplimiento.

Por los motivos antes señalados, se considera que la calidad de datos alcanza un nivel de importancia siguiendo la línea de los procesos ETL, específicamente durante las fases de transformación y carga, en donde se busca que los datos que viajan al sistema origen sean íntegros en su totalidad.

Tener datos íntegros se refiere a que después de ser corregidos, es decir, después de haber aplicado algunas acciones como *Insert*, *Delete* o *Update* se vean modificados en la base de datos donde se encuentran almacenados, poniendo demasiada atención para poder identificar si durante esos procesos se pierden datos. Integridad en cierta forma es tener algo en su totalidad, Por lo tanto, si se tiene duplicidad o datos erróneos o no válidos, la integridad dejaría de existir [26].

En la actualidad existen empresas ofreciendo herramientas de calidad de datos para las organizaciones, como ejemplo se puede mencionar a Informática, el cual está posicionado como líder en el cuadrante Mágico de Gartner sobre herramientas de calidad de datos. Su nombre completo es Informática *Data Quality* y garantiza que los datos fiables ofrecen a una organización iniciativas de negocio. Se ha creado para líderes de negocio y de IT, y está diseñada para rendir a la escala y la velocidad que el negocio necesita tanto en la actualidad de cara al futuro [27].

El cuadrante mágico de Gartner es un ranking gráfico sobre determinadas competencias y situaciones del mercado de un producto tecnológico, elaborado por la empresa. El cuadrante es representado por dos ejes; X y Y. El Eje X representa la integridad de visión de tal grado que enseña el conocimiento del proveedor sobre cómo se puede utilizar el valor actual en el

mercado. Por otro lado, el eje Y representa la capacidad de ejecución y mide la habilidad para ejecutar con éxito su visión en el mercado [28].

Figura 1.4 Cuadrante mágico de Gartner.

La figura anterior representa el cuadrante en donde los ejes dividen al plano en cuatro cuadrantes, a cada empresa u organización se le clasifica en base a una serie de apartados (en función de su tipología y la de sus productos respectivamente) que terminan por colocarlo en uno de los cuadrantes.

Los cuadrantes son:

- Retadores u aspirantes (*challengers*). Ofrecen buenas funcionalidades, pero ofrecen menor variedad de productos al estar enfocados en un único aspecto de la demanda del mercado.

- Líderes (*leaders*). Las empresas ofertan una solución de productos amplia y completa, que evoluciona según la demanda en el mercado.

- Nichos específicos (*niche players*). Las empresas están enfocadas en diversas aéreas de tecnologías de la gestión empresarial pero sin disponer de una suite en su totalidad.

- Visionarios (*visionaries*). Estas tienen capacidades para anticiparse a las necesidades del mercado, sin embargo no pueden disponer de la capacidad para realizar implantaciones causadas por su tamaño u otras circunstancias

2.4.2. Modelos Multidimensionales

Un cubo es una estructura multidimensional que contiene información con fines analíticos; las características componen a un cubo son dimensiones y medidas. Las dimensiones como tal determinan la estructura del cubo que son utilizadas para segmentar y dividir los datos, y las medidas respectivamente proporciona valores numéricos agregados y que son importantes para el usuario final [34].

Un modelo multidimensional es un modelo de datos como conjuntos de medidas descritas por dimensiones. Este modelo es funcional para resumir y organizar datos, enfocado para trabajar sobre datos de tipo numérico, con ello hace más fácil visualizar y entender [35].

Un *Data Warehouse* es una base de datos corporativa que funciona como un repositorio para todos los datos que recogen los diferentes sistemas de una empresa u organización.

El procesamiento analítico en línea (OLAP, por sus siglas en Ingles), permite extraer de manera selectiva datos y observarlos desde diversos puntos de vista.

En la sección 2.2 se describe a un sistema BI, por lo tanto no será necesario puntualizarlo nuevamente. Las bases de datos multidimensionales son un tipo de base de datos optimizada para *Data Warehouse* que se utilizan principalmente para crear aplicaciones OLAP (*On-line Analitycal Processing*), una tecnología asociada al acceso y análisis de datos en línea, esto quiere decir que pueden verse como base de datos contenidos en una sola tabla.

Las tablas del modelo multidimensional se asimilan a un hipercubo o, si se usan herramientas OLAP, a un cubo OLAP. En ambos casos, las dimensiones de los cubos se corresponden con la de la tabla y el valor almacenado en cada celda equivale al de la métrica.

2.5. Herramientas ETL más comunes

Las empresas que manejan grandes volúmenes de datos con el fin de convertirlos en información significativa para reutilizarla en operaciones o simplemente para toma de decisiones, exigencias operacionales, análisis, grandes extracciones, así como transformaciones y carga de datos, requieren elegir la herramienta correcta.

Recientemente en el desarrollo de software para procesos ETL se aplica un procesamiento paralelo. Esto ha enriquecido el desarrollo de una serie de métodos para mejorar el rendimiento general de los procesos cuando se habla de grandes volúmenes de datos.

Los principales tipos de paralelismo actuales son los siguientes [15]:

- Paralelismo de datos: Este tipo de paralelismo consiste en particionar un solo archivo en varios más pequeños, para poder acceder a ellos de manera independiente y simultánea.
- Paralelismo de segmentación o *pipeline.* Una vez particionados los datos, permite tener de igual manera particionadas las operaciones. Permite hacer modificaciones a nivel estructura.
- Paralelismo de componente. En este tipo de paralelismo se definen componen reutilizables, permite el funcionamiento simultáneo o de múltiples procesos en diferentes flujos de datos.

El uso de herramientas ETL responde a criterios de sincronización, conectividad y actualización. Las herramientas además de ser utilizadas en entornos *Data Warehousing*, son útiles para los siguientes propósitos [6]:

- Migración entre aplicaciones, ya sea por cambio versión o cambio de aplicación.
- Tipificación entre diferentes sistemas operacionales.
- Reforzar, migrar y sincronizar bases de datos operativas.
- Interfaces de datos con sistemas externos.
- Desarrollo de procesos masivos.

Breve historia de las herramientas ETL

Cuando los procesos ETL surgieron, eran desarrollados en lenguajes de programación clásicos, SAS/BASE, Cobol, PL-SQL, sin embargo, los procesos comenzaron a estar compuestos de un número elevado de líneas de código, lo que provocaba que estos fueran difíciles de mantener. La lenta curva de aprendizaje y la dificultad de mantenimiento generaron la búsqueda de otras alternativas.

A mediados de la década de los 90, las empresas más sobresalientes del mundo de los sistemas de información decidieron invertir en implementar sus propias herramientas orientadas al diseño y desarrollo de procesos ETL sin necesidad de programarlas exclusivamente en código, de esta manera surgen herramientas como: *Informatica PowerCenter,* IBM *Datastage* y *SAS Data Integration.* Tiempo después surgieron empresas dedicadas al desarrollo de *software OpenSource* para procesos ETL, algunas herramientas son: *Talend*, KETL*, Kettle (Pentaho Data Integration,* las cuales proporcionan un entorno visual e intuitivo, mantenimiento, manejo de modelos y metadatos*,* interfaces de datos con sistemas externos, entre otros detalles [33].

Principales herramientas *ETL* [15]:

- ✓ *Pentaho Data Integration (Kettle)*
- ✓ *SAS ETL Studio*
- ✓ *Talend*

- ✓ *IBM WebsphereDataStage*
- ✓ *Microsoft IntegrationServices*
- ✓ *Oracle Warehouse Builder*
- ✓ *Informática PowerCenter*
- ✓ *IBM Cognos Data Manager*

De la lista anterior *PowerCenter, DataStage y SAS ETL* son consideradas como herramientas TOP en el mercado actual, por sus funcionalidades superiores y costo promedio por licencia. *Warehouse Builder, IntegrationServices, Data Manager,* se consideran como herramientas nivel medio, usadas principalmente para proyectos BI y por su promedio por licencia. Por último las herramientas Open Source; Pentaho Data Integration (PDI) y Talend, las cuales son libres, esto quiere decir que en ningún sentido se hace un pago de licencia para su instalación y uso [16].

Las principales herramientas fueron seleccionadas de las principales fuentes de información, dicha selección se basó en función en el *rating* que cada una mantiene actualmente para las organizaciones. Enlisto algunas que se encuentran disponibles en el mercado actual; Oracle Data Integration, SAS Data Integration, Hadoop - *Open Source* (Código abierto, en español)

A continuación se describen algunas herramientas:

2.5.1 Pentaho Data Integration (Kettle)

Como ya se mencionó, *Pentaho Data Integration (Kettle)* es una herramienta ETL de código abierto. Tiene un entorno grafico agradable y sus características y capacidades de ETL son suficientes para trabajar de forma eficiente con dichos procesos. La herramienta como tal es fácil de utilizar para comenzar a manipular datos, así como la instalación del software [17].

Pentaho tiene en el mercado dos versiones; *Pentaho DI Enterprise*, la cual tiene licencia de pago y *Pentaho DI Community*, está de versión gratuita (*Open Source*).

Pentaho permite a los usuarios extraer, transformar, limpiar y preparar diversos datos de cualquier fuente.

La compañía *Pentaho* comenzó a operar en el año 2001, hoy en día tiene una comunidad activa de usuarios, alrededor de 13,500 usuarios. Funciona utilizando Java, presentando como ventaja el ser una solución multiplataforma. Su diseñador gráfico de transformaciones se llama Spoon.

Kettle es una herramienta amigable al momento de realizar conexiones a base de datos [17]. En la Figura 1.5 se muestra la interface de dicha herramienta.

En lo particular he utilizado la herramienta en diversas organizaciones para el desarrollo de procesos ETL y me ha sido muy útil, desde su instalación, desarrollo y ejecución de procesos.

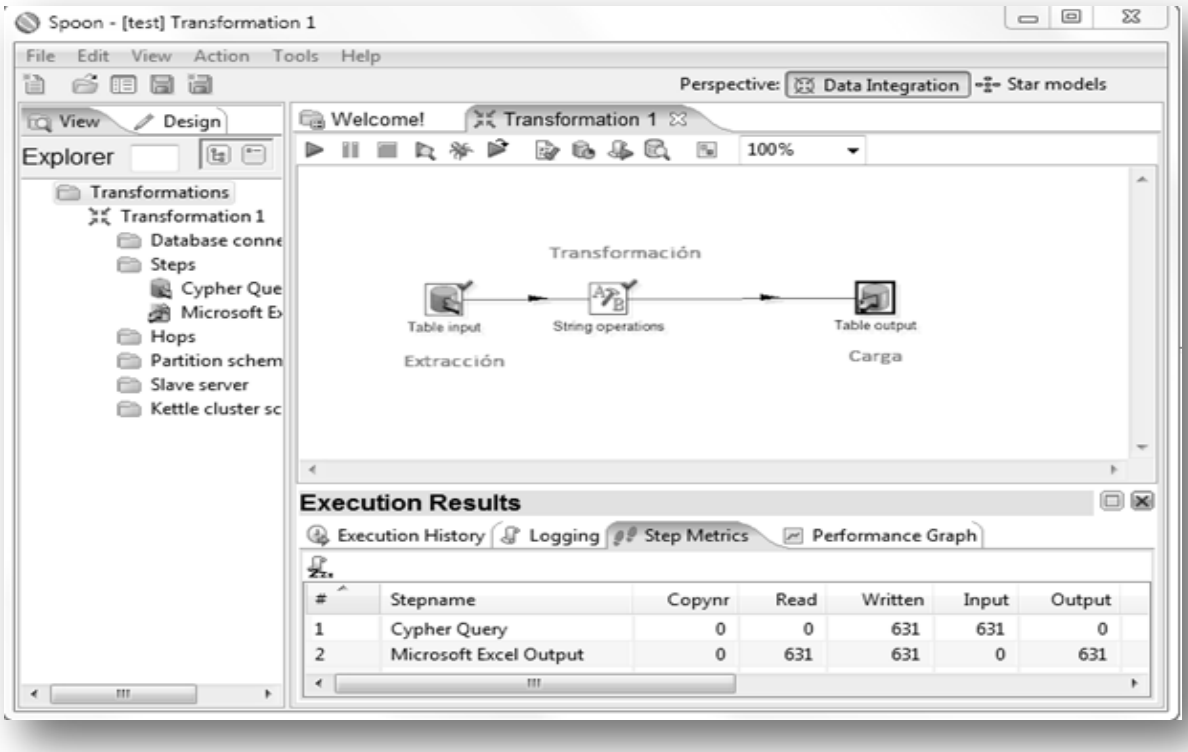

Figura 1.5 Interface *Pentaho Data Integration* Versión 5.

2.5.2. Talend

Talend al igual que *Pentaho Data Integration* (PDI) es una solución exitosa de integración de datos de código abierto, adopta nuevas tecnologías de Big Data y las integra en la infraestructura de Tecnología de la Información ya existente. En la Figura 1.6 se muestra la interface de dicha herramienta.

Talend proporciona herramientas sólidas de integración de datos para implementar procesos ETL. Permite a los desarrolladores completar trabajos de integración 10 veces más rápido que la codificación manual. En el mercado actual existen dos versiones *Talend Open Source Data Integration* y *Talend Data Mangement Platform.* Las herramientas de *Talend* hacen que desarrollar procesos ETL sea más fácil [36]

Cuenta con una interfaz y un lenguaje grafico amigable. Su primera versión fue lanzada en el año 2006. Genera código en Java o Scripts en Pearl que pueden ser implementados en servidores que lo soporten. Esta herramienta cuenta con buenos comentarios de compañías importantes, por ejemplo; DHL, BNP PARIBAS REAL ESTATE, PANASONIC, TICKETMASTER, SKY, entre otros [15] [19].

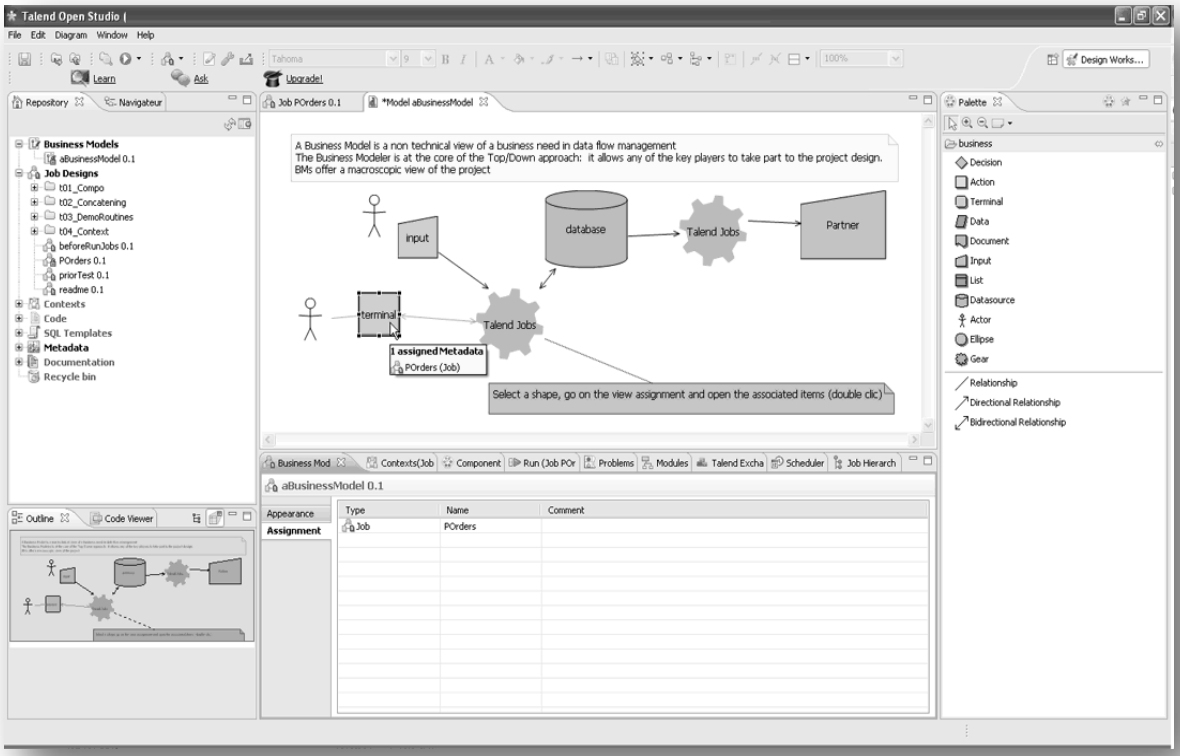

Figura 1.6 Interface *Talend*.

El diseño de Talend está basado en 3 niveles:

1. Modelos de Negocio (*Business Models*, en inglés). Este nivel ha sido diseñado para crear de forma teórica la aplicación, se realizan diagramas de flujo con actores de procesos. En la figura 1.7 se puede observar dicho nivel [19].

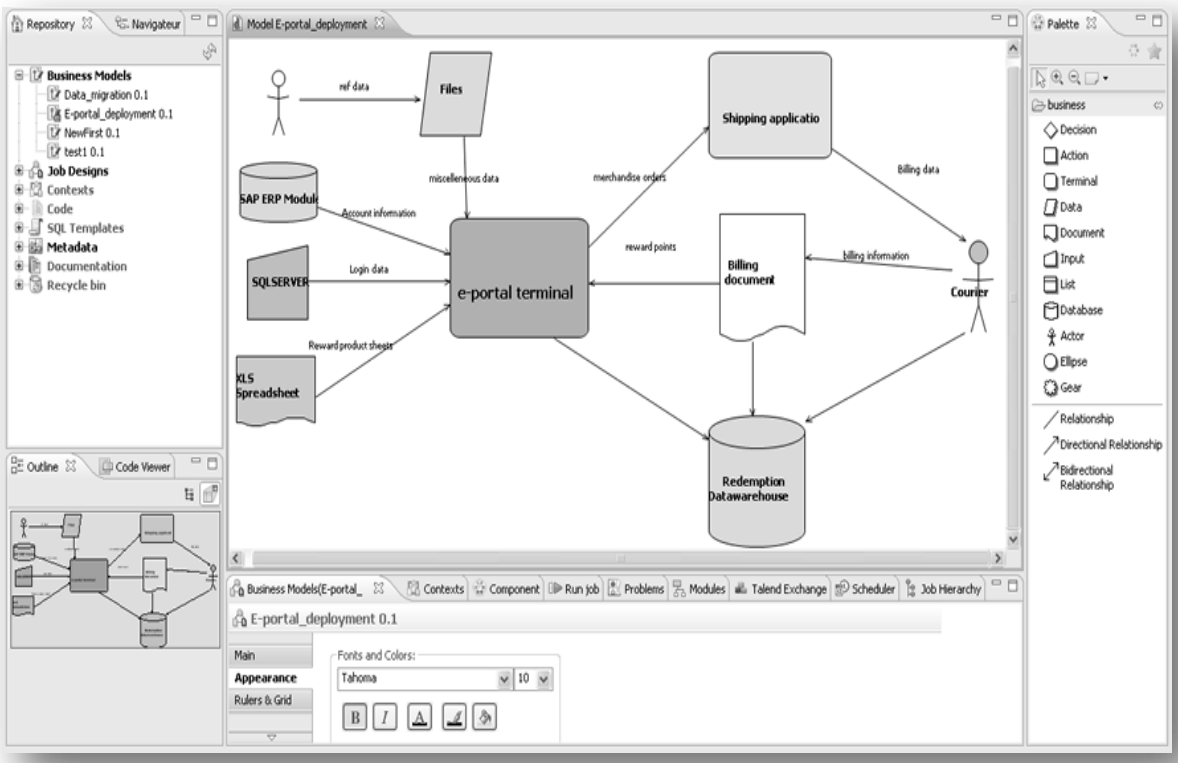

Figura 1.7 Fase modelo de negocio

2. Diseño de Trabajos (*Job Designs*, en inglés). En este nivel se realiza todo el código que será ejecutado, en realidad es el diseño del proceso [19]. Formado por el conjunto de Jobs, o tareas a realizar. En la siguiente figura se observa el diseño.

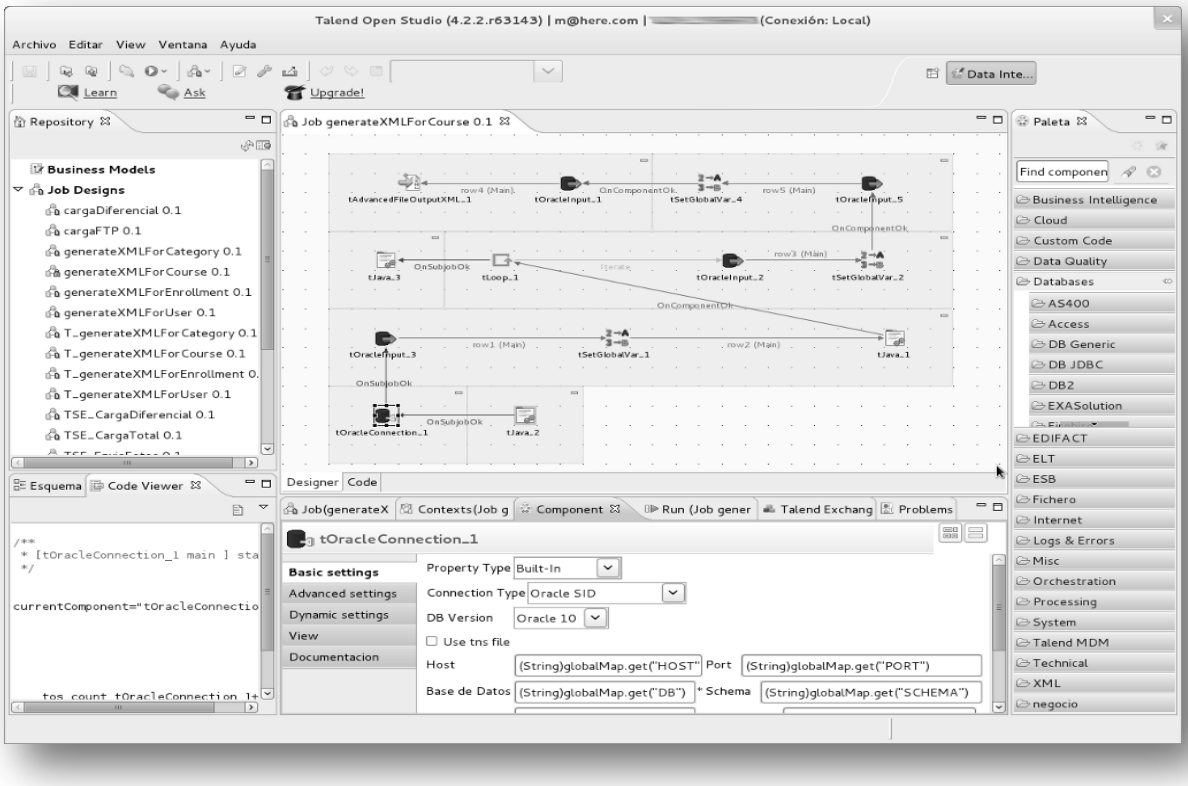

Figura 1.8 Fase diseño de trabajos.

3. Contextos (Contexts, en inglés). En este nivel se definen las variables globales de la ejecución del programa, la carpeta donde se ejecutará la aplicación final, además de que se definen las variables iniciales de entrada del programa, en la Figura 1.9 se muestra el nivel correspondiente [19].

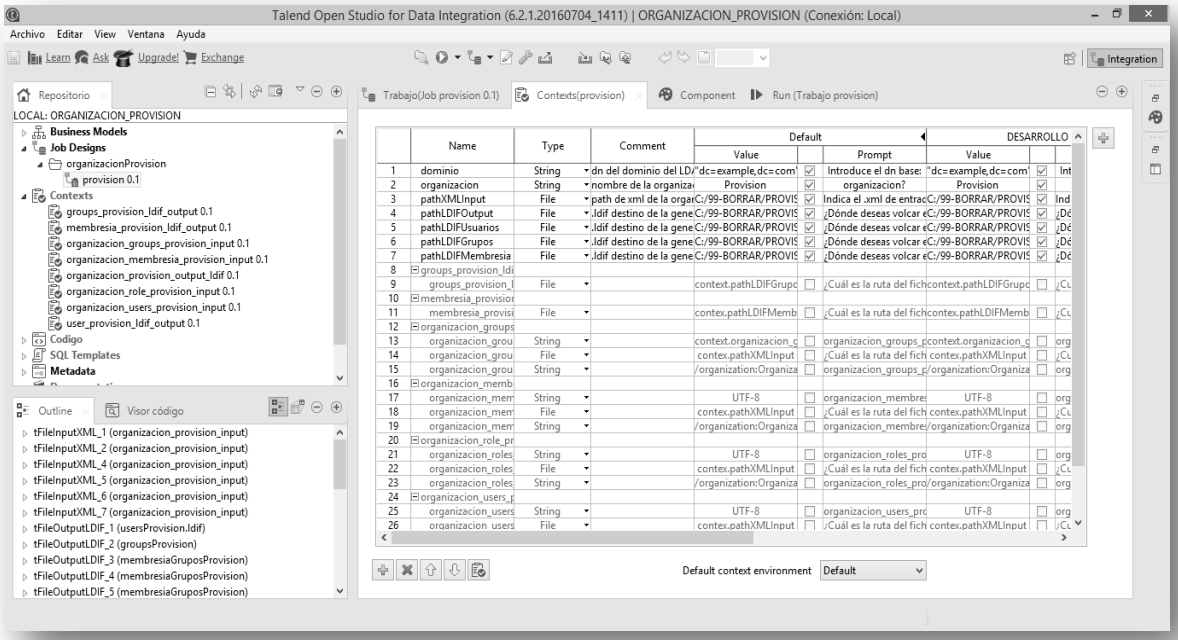

Figura 1.9 Fase de contextos.

Talend aborda todos los aspectos de integración desde la capa técnica hasta la capa de negocio y todos los productos se reagrupan en una única plataforma unificada como se observa en la figura

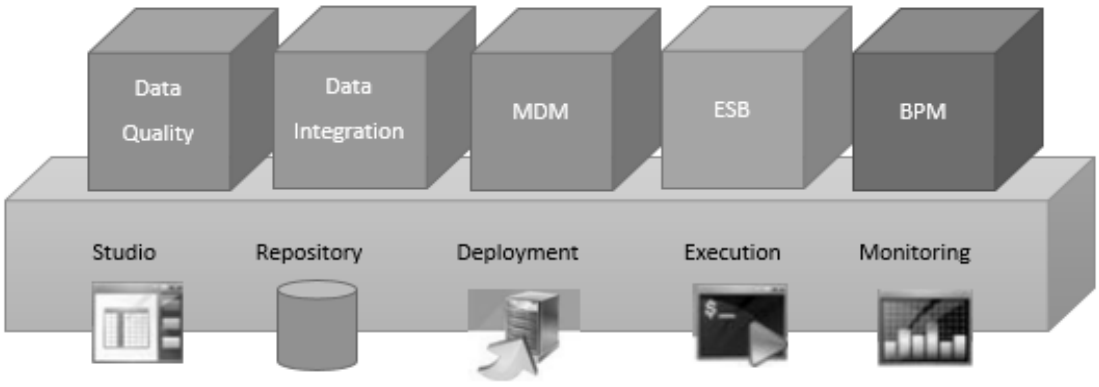

Figura 1.10 Plataforma *Talend*.

La plataforma de *Talend* ofrece un entorno único basado en Eclipse, lo que significa que los usuarios pueden saltar de un producto a otro simplemente haciendo clic en el botón sin la necesidad de cambiar las herramientas. Todos los trabajos, servicios y los activos técnicos están diseñados en el mismo entorno con la misma metodología, desplegado y ejecutado en el mismo tiempo de ejecución, monitoreado y operado en la misma consola de gestión.

2.5.3. InformaticaPowerCenter

InformaticaPowerCenter es una herramienta que ha sido bien posicionada dentro del mercado actual, además de que es considerada como segura y robusta, así como ser una herramienta líder de integración de datos según el Cuadrante Mágico de *Gartner*.

Proporciona una plataforma para la integración de datos comerciales, permite la escalabilidad para admitir grandes volúmenes de datos de diferentes orígenes, migración de datos, etc. Se fundó en el año de 1993, es considerada como líder actual del sector *Data Integration* [15] [20]. Tiene alrededor de 2600 clientes, entre los cuales figuran bancos como Grupo BBVA, organizaciones Gubernamentales, etc. En la Figura 1.11 se observa la interface de dicha herramienta.

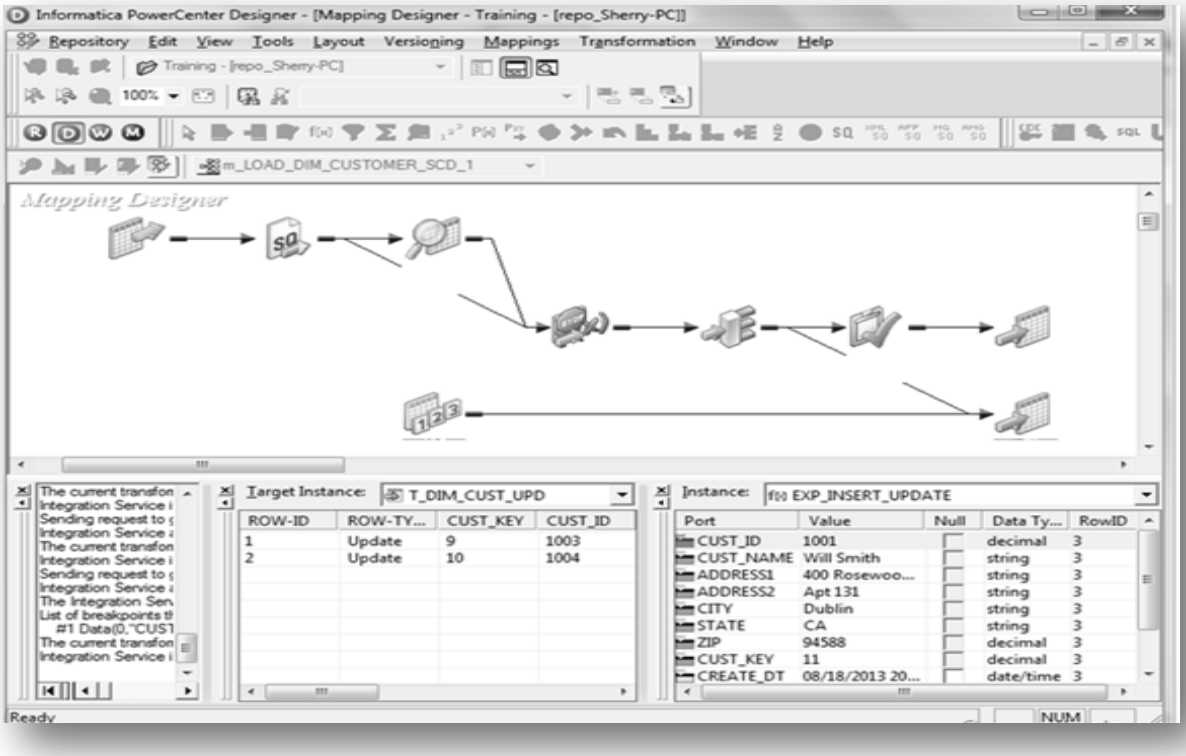

Figura 1.11 Interface *InformaticaPowerCenter.*

Algunas de las características que destacan a la herramienta, son:

- Colaboración entre el negocio y Tecnología de la Información.
- Sistematización, reutilización y facilidad de uso.
- Soporte para procesamiento distribuido, balanceo de carga, adaptabilidad y optimización.
- Supervisión de gobierno y operacional
- Información en tiempo real para el análisis y las aplicaciones.
- Desarrollo de prototipos y validación rápidos.
- Transformación avanzada de datos.

2.5.4. IBM Cognos Data Manager

Es una herramienta que permite las operaciones dimensionales de extracción, transformación y carga (ETL) para conseguir una inteligencia empresarial de alto rendimiento. Se pueden

ejecutar compilaciones y secuencias de trabajos en sistemas remotos desde un sistema de entorno de diseño de Data Manager. Esta herramienta puede ser instalada en sistema operativo Linux o Unix, contribuye a desarrollar plataformas globales de integración de datos [21]. En la Figura 1.12 se observa la interface de dicha herramienta.

Algunas de las características que destacan a la herramienta, son:

- Es capaz de soportar el análisis de alto rendimiento de datos mediante la creación de tablas en múltiples niveles.
- Brinda soporte en idiomas distintos.
- Apoya a construir una plataforma global de integración de datos.
- Automatiza distintos procesos.
- Permite que diferentes desarrolladores compartan información de componentes de la misma herramienta.

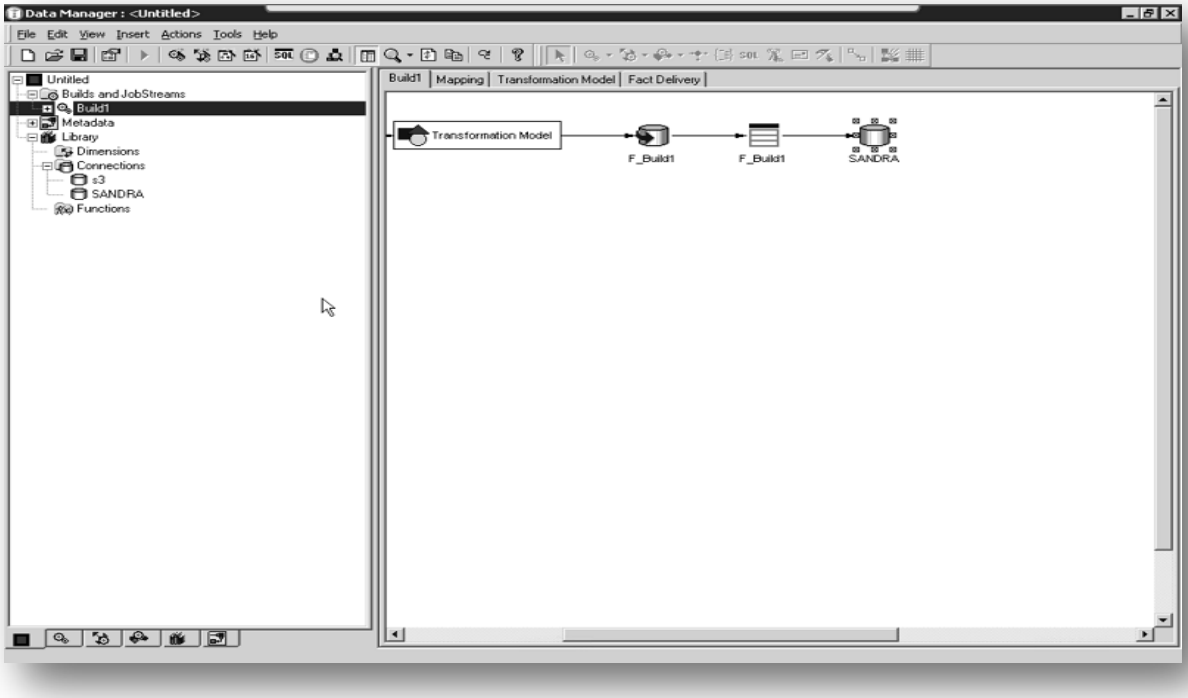

Figura 1.12 Interface IBM Cognos Data Manager.

Una vez conociendo algunas de las herramientas ETL con sus principales características, surge la siguiente pregunta:

¿Cómo elegir la herramienta para procesos ETL más adecuada?

Surgen una gran variedad de herramientas ETL disponibles en el mercado y para su elección cada empresa u organización debe elegir, todo en función de sus propias características, necesidades y objetivos. A mi punto de vista, la organización debe elegir una herramienta de acuerdo a las necesidades que tenga, de manera que la explotación y gestión de la información se realice de manera segura y que también genere los resultados esperados.

Algunas pautas para elegir una herramienta ETL [5] [22]:

- Determinar si es necesario crear un proceso ETL o comprar uno existente.
- Identificar los recursos que tiene la organización disponible.
- Considerar las iniciativas futuras de la organización.

2.6. Características de las herramientas ETL en general según *Gartner* [23]

- **Conectividad:** capacidad para poder conectar con un amplio rango de tipos de estructura de datos, por ejemplo: bases de datos relacionales y no relacionales, XML, emails, varios formatos de archivos, entre otros.
- **Capacidades de entrega de datos:** capacidad para proporcionar datos a otras aplicaciones, de igual manera se pueden programar procesos batch en tiempo real o por medio de un lanzador de eventos, entre los que figuran Gentran, Crontab, etc.
- **Capacidades de transformación de datos:** Desde una conversión de tipo de datos, agregaciones, hasta un análisis de texto en formato libre.
- **Capacidades de diseño y entorno de desarrollo:** representación gráfica de los objetos del repositorio, modelos de datos y flujo de datos, capacidades para trabajo en equipo, etc.
- **Capacidades de gestión de datos.** Calidad de datos.

- **Adaptación:** a diferentes plataformas hardware y sistemas operativos existentes o sistemas heredados.
- **Operaciones y capacidades de administración:** capacidad para gestión, monitoreo y control de los procesos de integración de datos.

2.7. Ventajas del uso de las herramientas ETL [6]

- Variedad de conectores disponibles.
- Herramientas optimizadas para los procesos de manejo de datos.
- Funcionalidades para trabajo en equipo.
- Generación de documentación.
- Detección, análisis y corrección de errores en bases de datos.
- Compatibilidad con una gran variedad de fuentes de datos.

2.8. Aplicación de ETL para gestión de datos empresariales

2.8.1 Ejemplos prácticos

Continuando con la línea de este proyecto, se describen dos aplicaciones reales de procesos ETL para la gestión de datos. La primera aplicación fue implementada para SOPHITECH una Consultora de Tecnología de la Información con experiencia en inteligencia de negocios y la segunda implementada para el sector financiero específicamente para Servicios Corporativos Scotiabank.

Primer Ejemplo Práctico.

Sistema de Análisis de las principales páginas Web de Inmuebles

En la Figura 1.13 se observa el funcionamiento lógico de la aplicación llamada "SAPPI" la cual fue desarrollada con el objetivo de comparar información extraída de las principales páginas Web de inmuebles, aplicando las tres fases de un proceso ETL.

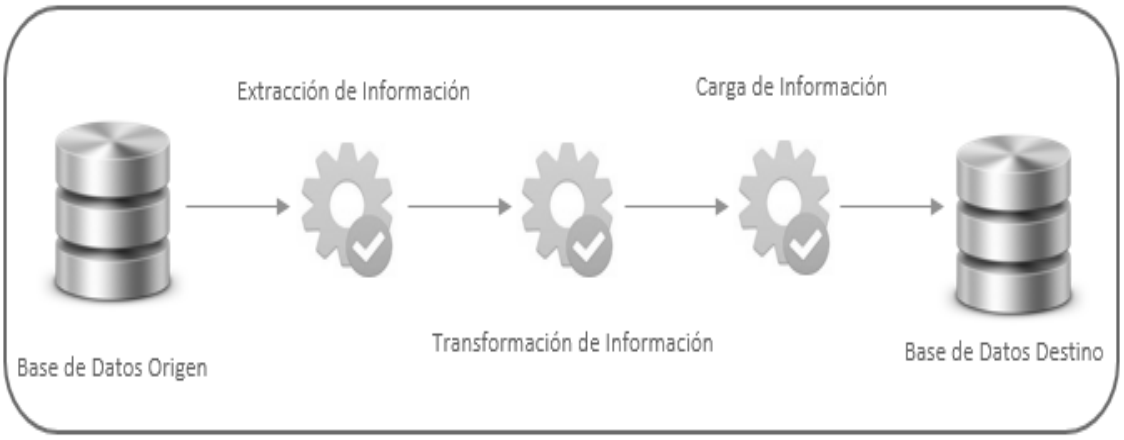

Figura 1.13 Funcionamiento lógico del aplicativo "SAPII".

Los procesos implementados para este sistema fueron desarrollados en la herramienta Pentaho Data Integration (Spoon Kettle). La figura 1.14 muestra el flujo completo del funcionamiento de "SAPPI".

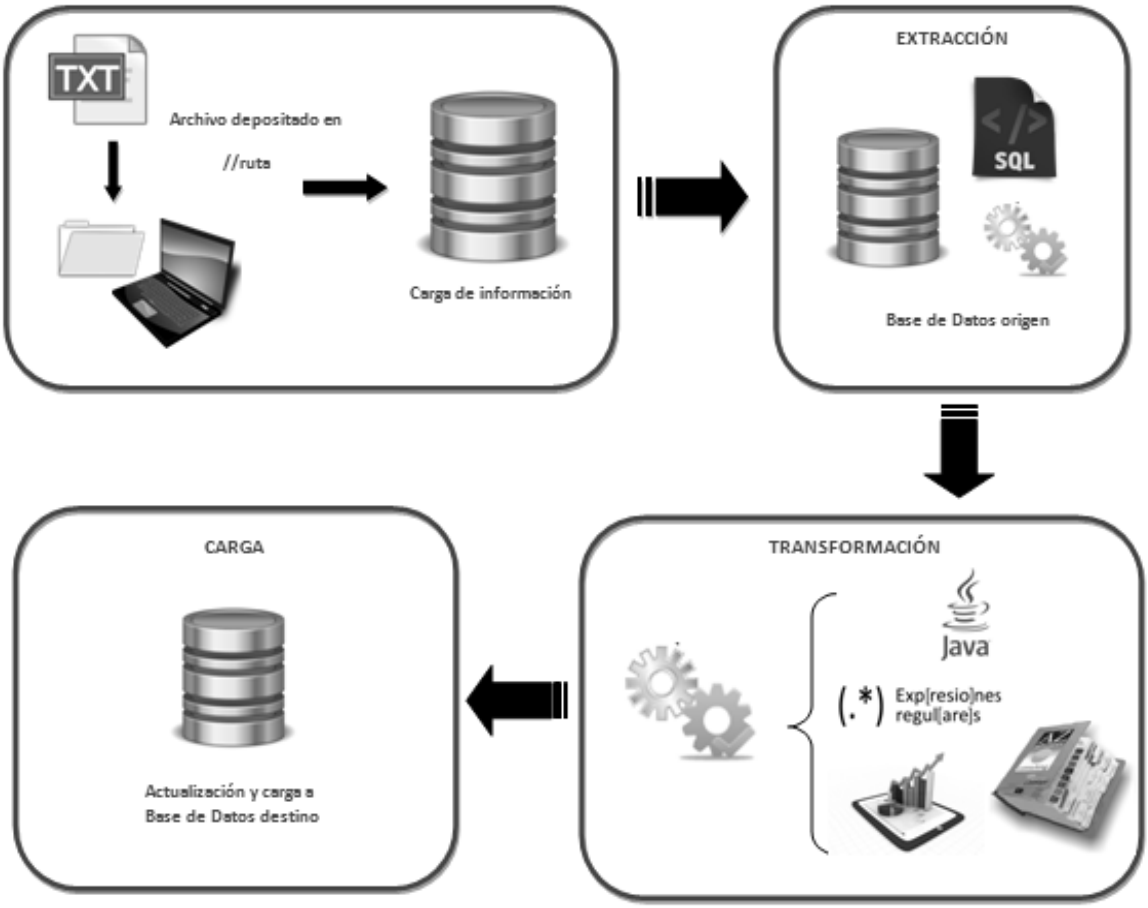

Figura 1.14 Flujo del funcionamiento de “SAPII”.

Breve descripción de cada uno de los bloques de la figura anterior:

Bloque 1. Extracción de archivo de texto y carga de información a la base de datos origen.

Bloque 2. Extracción de la información almacenada en la base de datos origen y mantenida en memoria por medio de sentencias SQL.

Bloque 3. Transformación de datos utilizando expresiones regulares y modulo Java siguiendo las reglas de negocio.

Bloque 4. Carga de información unificada a una base de datos destino, con la finalidad de obtener los datos requeridos y funcionales para la estrategia del negocio.

De acuerdo a cada uno de los bloques representado en la figura 1.14, el desarrollo del proceso ETL en la herramienta *Pentaho* quedó de manera en que se observa en la figura 1.15

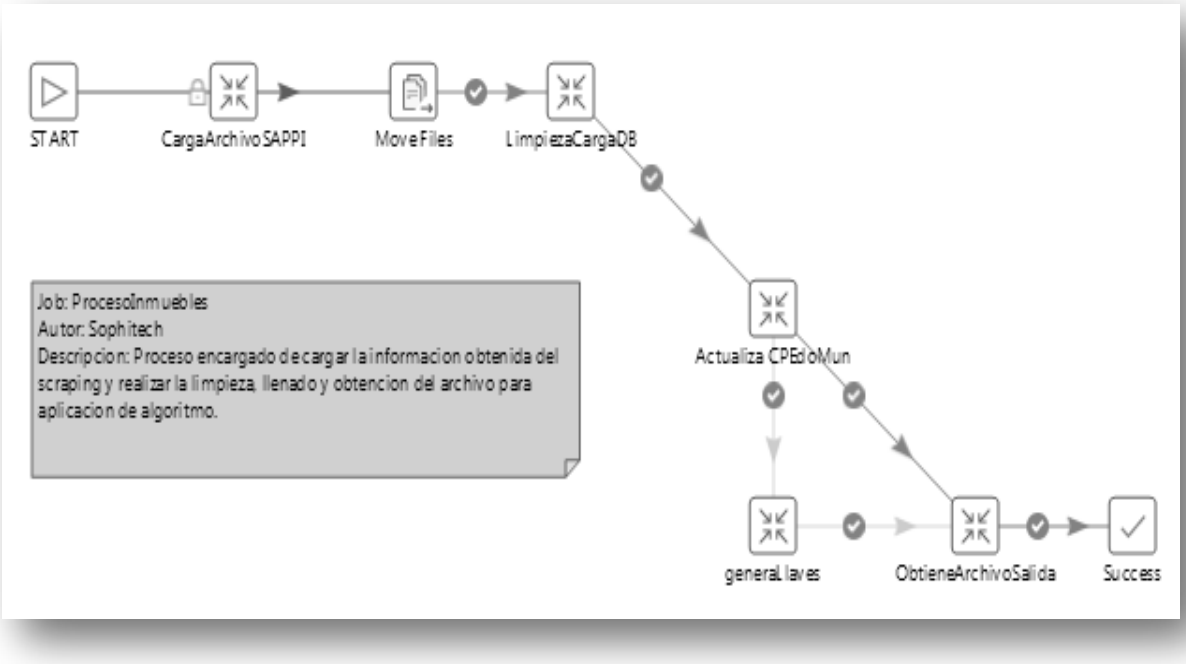

Figura 1.15 Implementación de "SAPPI" en Pentaho Data Integration.

A continuación se describen las fases del desarrollo del proceso ETL que fue implementado para el sistema:

1. El sistema extrae la información requerida de las principales páginas Web y es almacenada en un archivo con extensión .txt para después ser depositado a un directorio específico.

   El archivo de texto contiene la siguiente información:

Tabla 1.1 Información de las páginas Web que contiene el archivo de texto generado.

IDENTIFICADOR
URL
DIRECCIÓN
DATOS PRINCIPALES
VENDEDOR
CONTACTO
DESCRIPCIÓN
CARACTERÍSTICAS
FECHA

Una vez depositado el archivo en el directorio se realiza la carga de información en una base de datos origen que servirá para iniciar el proceso de extracción. La información es separada por el carácter '|' (pipe o tubo) para obtener el *layout* que se muestra en la Figura 1.16

Figura 1.16 Información del archivo de texto.

La carga de la información a la Base de Datos desde la herramienta Pentaho se observa en la Figura 1.17

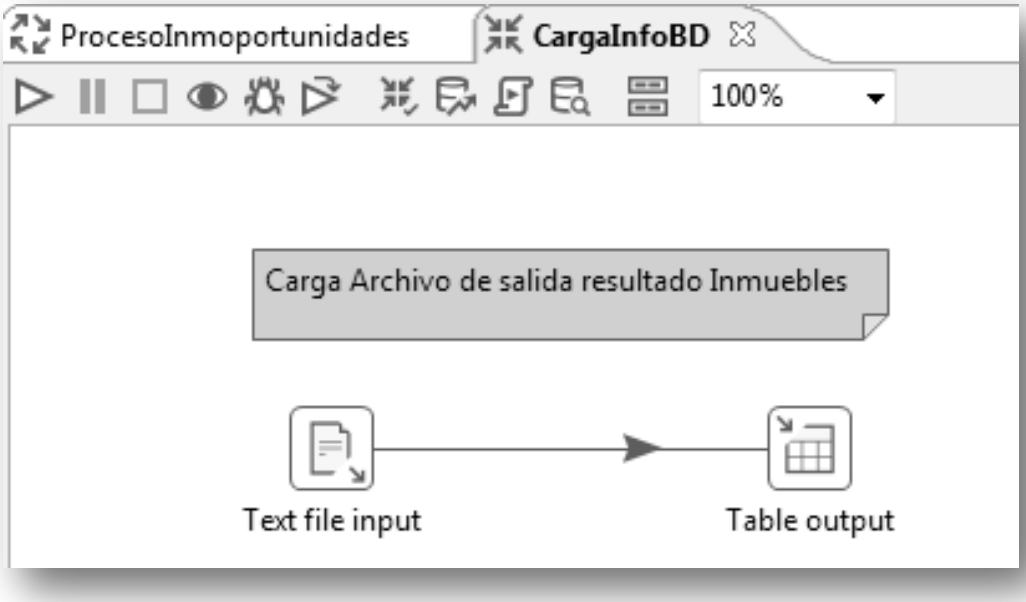

Figura 1.17 Carga de información a la Base de Datos

Los datos requieren ser reprocesados, para lo cual se realiza una limpieza de datos con la herramienta *Pentaho*, en la Figura 1.18 se muestra esta etapa.

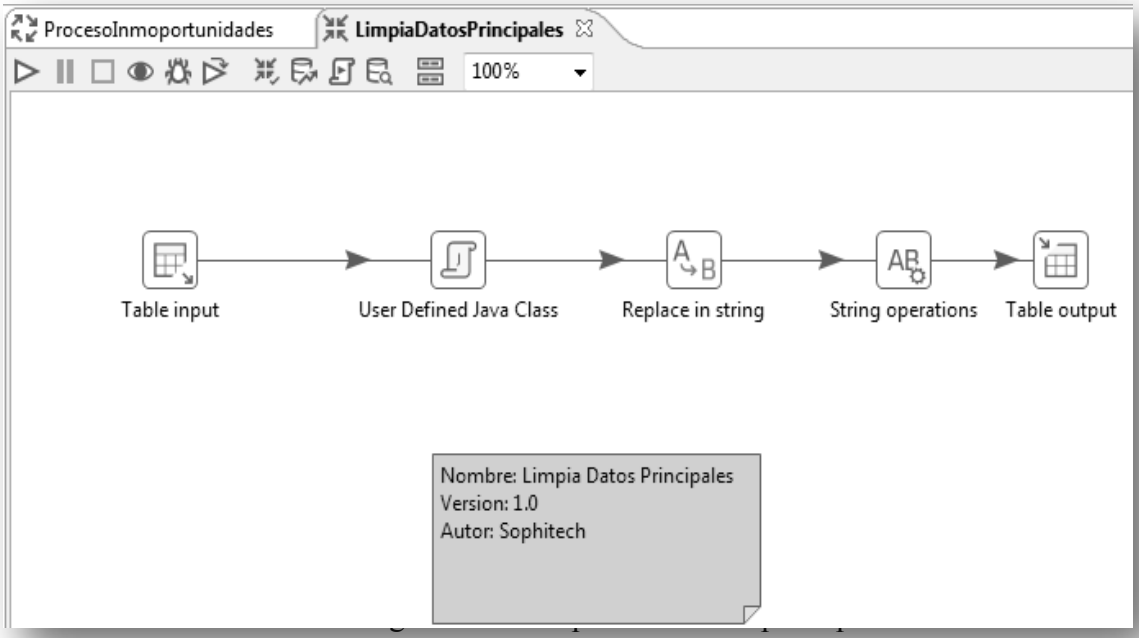

Figura 1.18 Limpieza de datos principales.

El proceso de limpieza se realiza para evitar datos erróneos y garantiza la calidad de los datos. En la sección 2.2.1 se describe detalladamente la finalidad de realizar este proceso.

2. Proceso de extracción

   a) La información que se encuentra en la base de datos origen es seleccionada, extraída por medio de SQL y mantenida en memoria para después ser procesada en la fase de transformación del ETL.

La primera fase del proceso ETL desde la herramienta Pentaho se observa en la figura 1.19

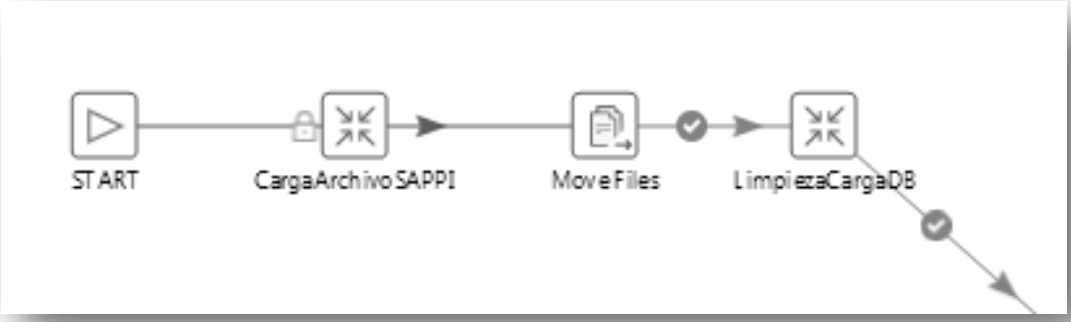

Figura 1.19 Extracción.

El proceso de extracción ejecuta después de que ha finalizado la extracción del archivo de texto origen y la carga de la información hacia la base de datos origen, con la finalidad de mantener los datos en memoria.

3. Proceso de transformación

a) Existen campos con información que debe ser fragmentada, esto se logra con un módulo JAVA y el uso de expresiones regulares.

b) Como primera instancia es segmentado el campo 'Dirección' en 'Estado', 'Municipio' y 'Colonia'. El campo es de tipo texto. Cuando se realiza la segmentación de dicho campo, se efectúa un cruce contra un catálogo de direcciones obtenido por SEPOMEX (Servicio Postal Mexicano) en la Internet. El cruce funciona de la siguiente manera:

   Por medio de los campos obtenidos en la segmentación de "Dirección" se obtiene desde el catálogo de SEPOMEX el campo "C.P." En caso de que el campo "Estado" no se encuentre informado, "Municipio" y "Colonia" servirán para complementar el campo "Estado".

   El proceso de transformación desde la herramienta Pentaho se muestra en las siguientes figuras:

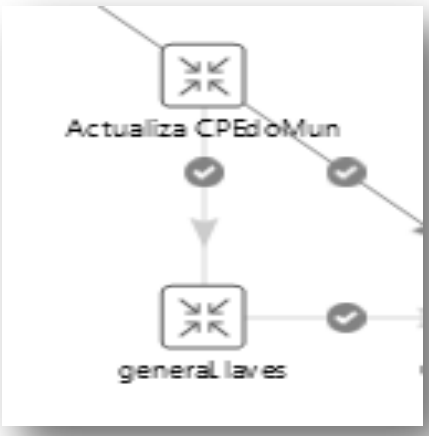

Figura 1.20 Transformación.

Figura 1.21 Actualiza información.

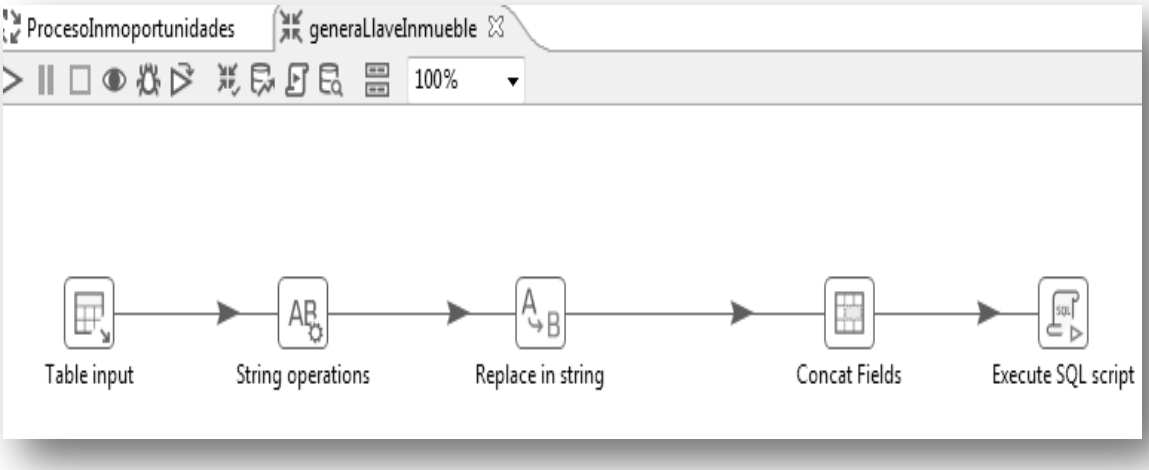

Figura 1.22 Actualiza campos vacíos.

Una vez que se tiene completa la segmentación anterior se realiza un cruce contra un catálogo de indicadores, el cual contiene un valor en escala del 0 al 10, estos muestran el nivel de servicios con los que cuenta la propiedad, siendo el CP la llave para asociarlo con la información.

El catálogo de Indicadores contiene los siguientes servicios:

- Escuelas
- Centros comerciales
- Salud
- Comercios
- Corporativos
- Energía
- Esparcimiento
- Bancos
- Transportes

La carga de indicadores desde Pentaho se muestra a continuación, en la figura 1.23

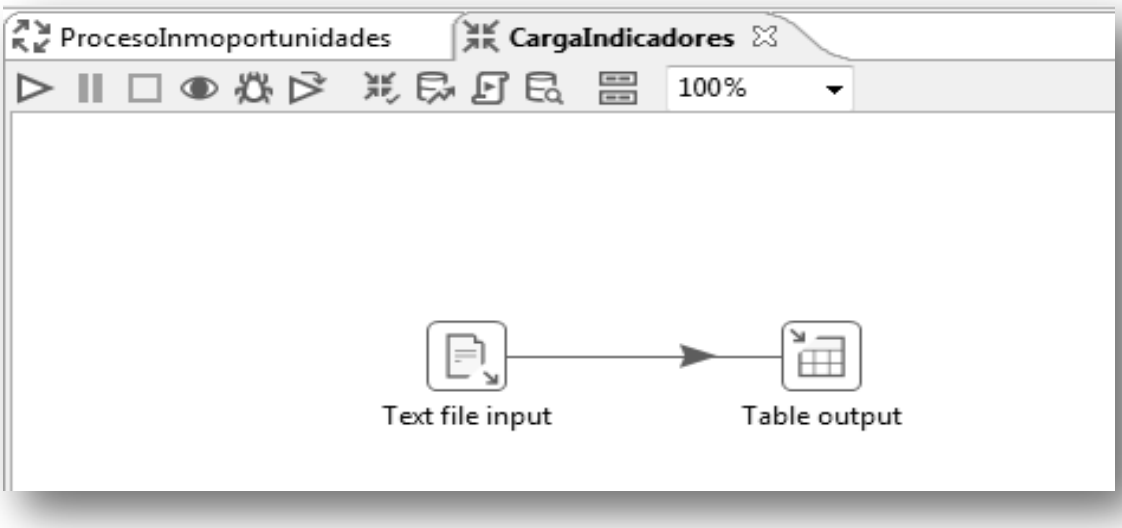

Figura 1.23 Carga de indicadores

c) El siguiente campo a segmentar es “Datos Principales”:

- Antigüedad
- Tipo inmueble
- Precio
- ^{m2} Construcción
- ^{m2} Terreno
- Recamaras
- Baños

d) Los campos siguientes no serán segmentados:

- Vendedor
- Contacto
- Descripción(Información complementaria del campo “datos principales”)

e) El campo “Descripción” será completado con un catálogo de amenidades, por ejemplo; si el espacio cuenta con gimnasio, alberca, salón de eventos, etc.

4. Proceso de carga

a) Una vez que termina el proceso de transformación, lo que sigue es la actualización y carga a una base de datos destino, en esta parte se hace un llenado automático de aquellos campos que no se encuentran informados.

El proceso finaliza con la generación de un archivo con extensión *.xls* el cual es funcional para realizar cálculos y análisis correspondientes. Ver Figura 1.24

Figura 1.24 Proceso de carga y generación de archivo final

5. La finalidad de la implementación del sistema es obtener la información necesaria y funcional para la estrategia del negocio.

Segundo Ejemplo Práctico

Durante mi instancia en Servicios Corporativos Scotiabank fui participe de la implementación de varios procesos ETL para la automatización de tareas provenientes de la operativa diaria de la venta de seguros ejecutada en el *Contact Center*.

Algunos de los procesos ETL implementados son los siguientes:

- ✓ Alarmas automáticas que monitorean los procesos batch (“Géminis”) ejecutados por Gentran (Servicio de control de procesos).

- ✓ Ejecución de reportes semanales y mensuales de atención de tickets operativos para el *Contact Center*.
- ✓ Envió de mails automáticos con resultados de domiciliación para cada una de las aseguradoras.

De manera resumida, el Contact Center de Scotiabank manejaba venta para tres aseguradoras; CARDIF, ACE y Cpp. Cada una de estas empleaba diferentes campañas, como por ejemplo; Vivienda + Segura, Efectivo + Seguro, Auto + Seguro, entre otras.

Las ventas eran procesadas por "Géminis", un sistema interno que ejecutaba los procesos batch encargado de la gestión y cobranza de seguros.

Los tickets operativos provenían de un sistema externo administrado por el área de Help Desk y con solicitud al área de desarrollo según la incidencia presentada.

Los tickets provenían de ciertas casuísticas:

a) Modificación y/o actualización de número de cuenta/tarjeta de algún cliente en específico por aseguradora.
b) Aclaración y análisis de cobranza de seguro por cliente (s).
c) Actualización de datos generales de clientes por aseguradora.

A continuación se describe el segundo ejemplo práctico implementado con la herramienta *Pentaho Data Integration* (PDI).

Reportes automáticos semanales y mensuales de atención de tickets operativos (Contact Center)

El proceso fue implementado porque la herramienta externa que gestionaba los tickets no brindaba un reporte visual, es decir, la información se veía muy simple y no decía mucho al momento de observar los datos ya que únicamente generaba un histórico en un archivo de texto. Para ello surgió la necesidad de concentrar la información de ese archivo de texto en una base de datos local y generar reportes en formato .xls, esto con la finalidad de que la información tuviera más formalidad al momento de la entrega al área correspondiente, sin necesidad de generarla manualmente.

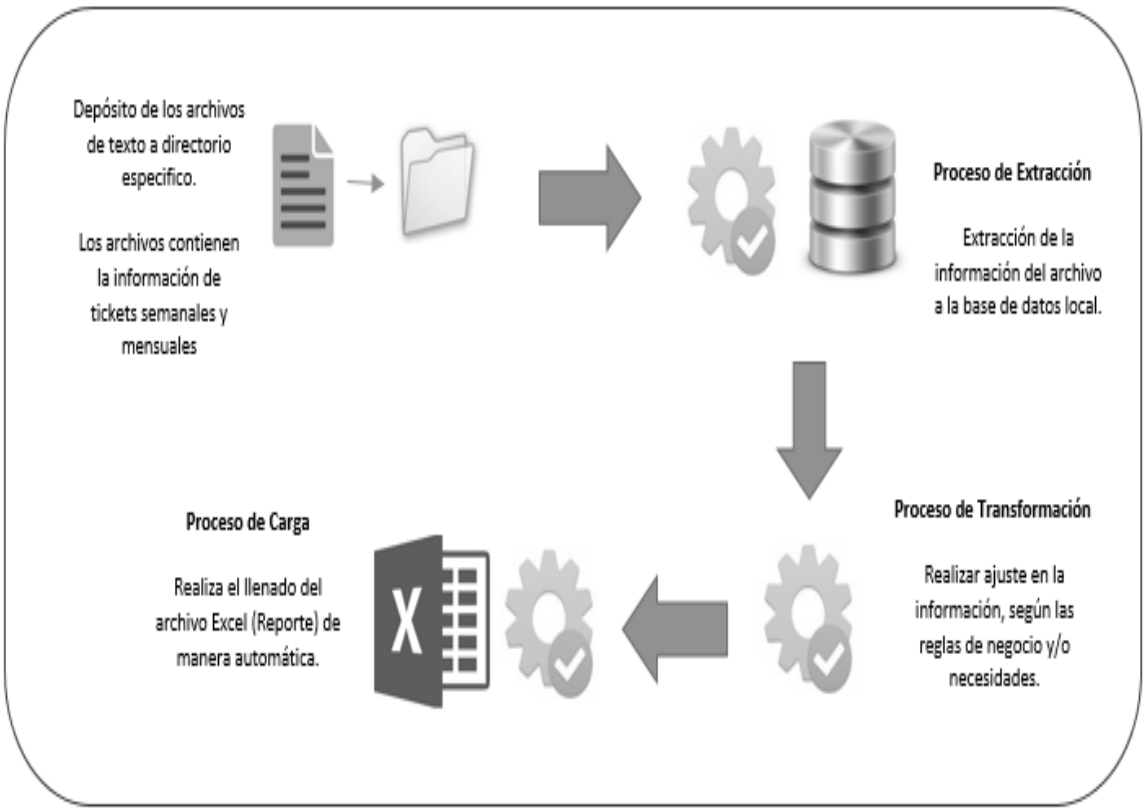

Figura 1.25 Implementación de reportes automáticos

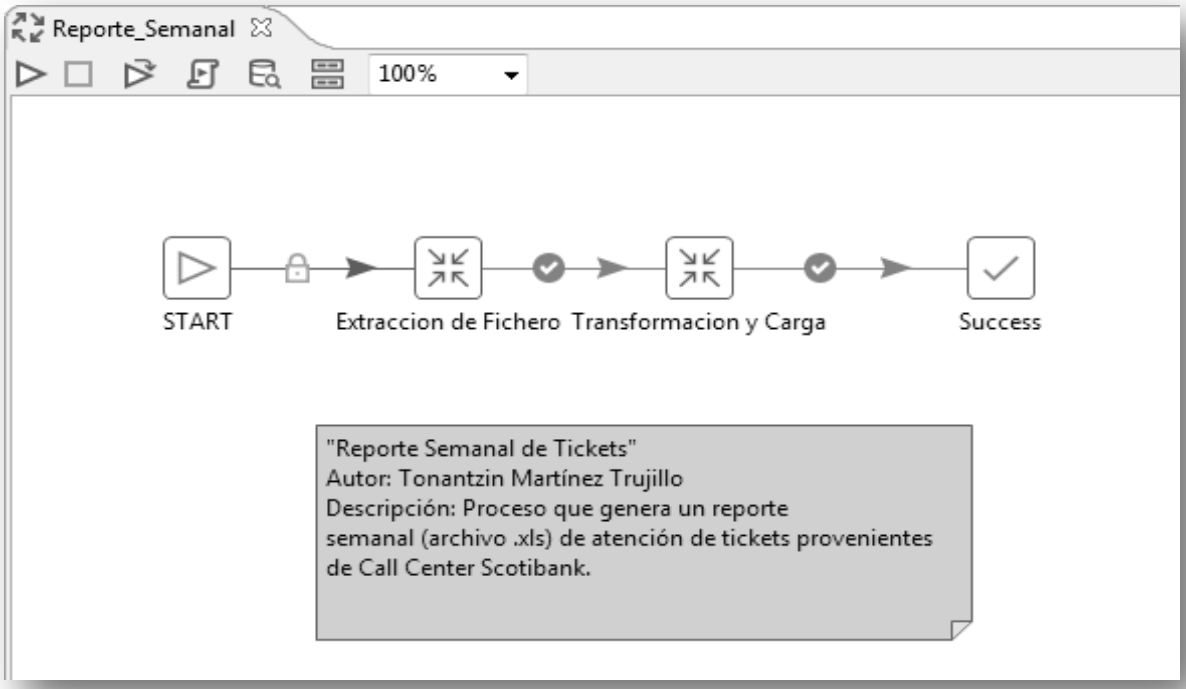

Figura 1.26 Implementación de reportes semanales en la herramienta *Pentaho Data Integration*.

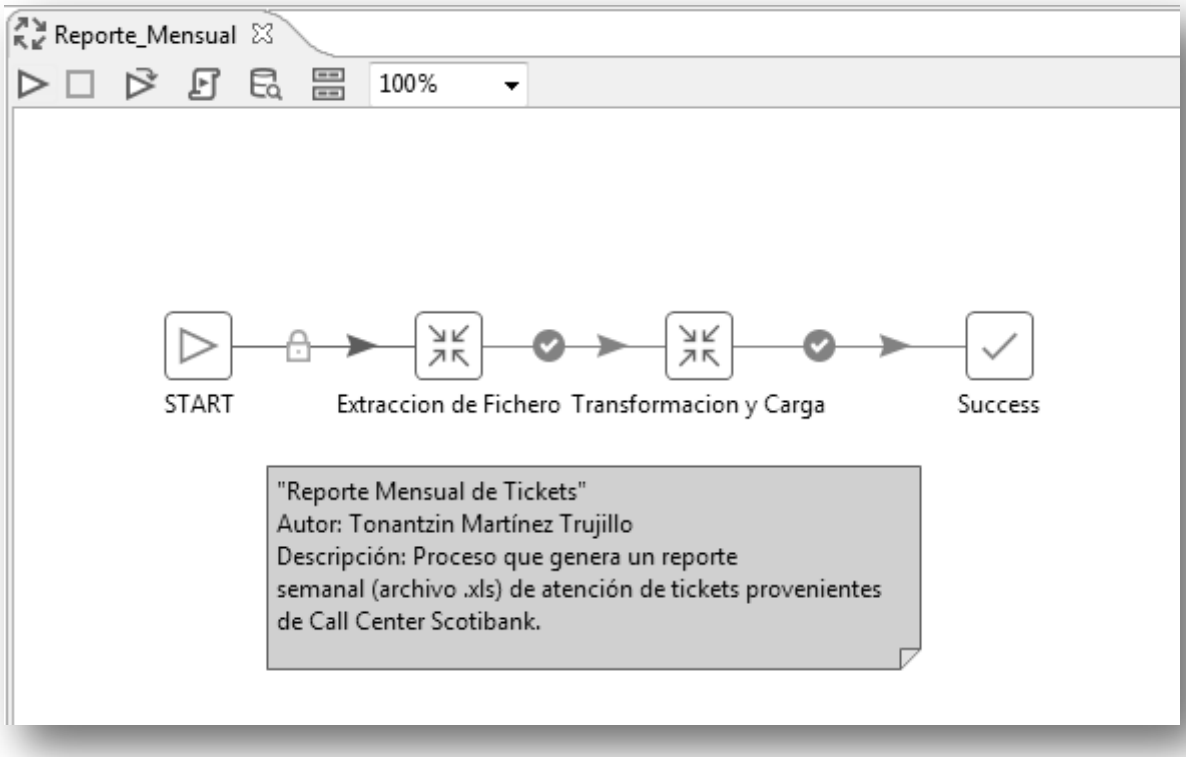

Figura 1.27 Implementación de reportes mensuales en la herramienta *Pentaho Data Integration*.

A continuación se describen cada una de las fases del proceso implementado:

1. El área de Help Desk se encargaba de extraer la información semanal y mensual de la herramienta y depositar los archivos de texto correspondientes en un directorio específico.

Tabla 1.2 Campos informados en archivo de texto.

ID_USUARIO
NOMBRE_USUARIO
ID_TICKET
RELEVANCIA
FECHA_APERTURA
FECHA_CIERRE
FECHA_LIMITE
DESCRIPCION

Descripción breve por campo:

- ID_USUARIO: Identificador de usuario
- NOMBRE_USUARIO: Nombre de la persona que atendió y resolvió el ticket
- ID_TICKET: Identificador de usuario
- RELEVANCIA: Grado de importancia (ALTA, MEDIA, BAJA)
- FECHA_APERTURA: Fecha en que el ticket es levantado
- FECHA_CIERRE: Fecha en que el ticket es atendido
- FECHA_LIMITE: Fecha límite en que se puede atender y cerrar el ticket
- DESCRIPCIÓN: Breve detalle de la incidencia.

1. Proceso de extracción

   En esta etapa se extrae el archivo de texto del directorio específico y la información es almacenada en una base de datos origen (local).

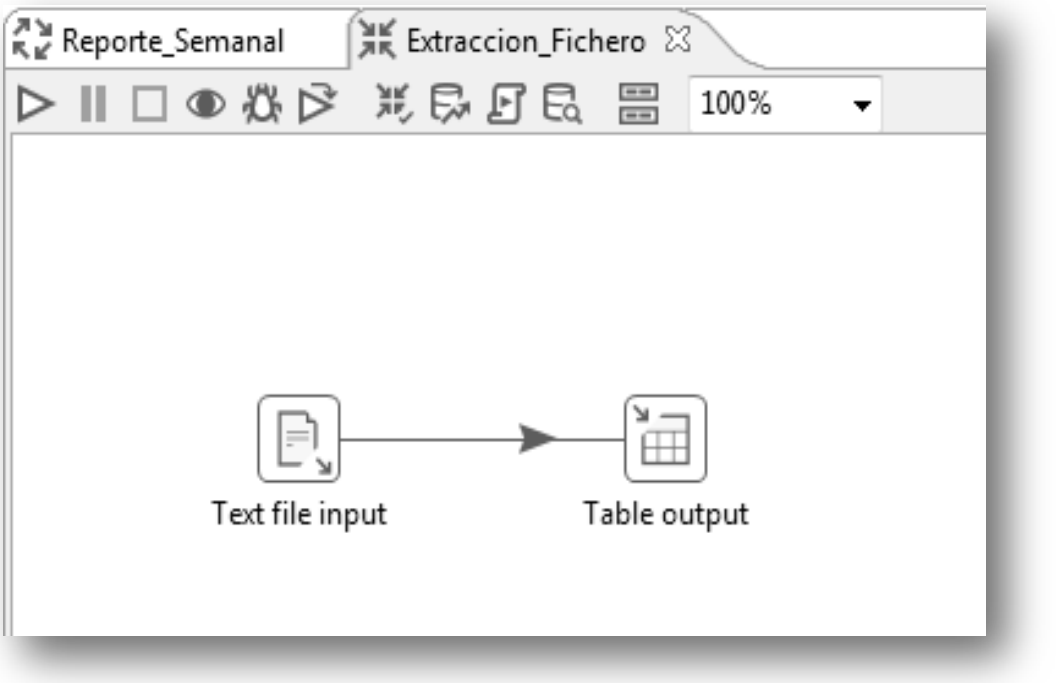

Figura 1.28 Proceso de extracción y carga a la Base de Datos local.

2. Proceso de Transformación

   Una vez que la información del archivo se texto se encuentra almacenada en la base de datos local, se realiza una serie de modificaciones a la información, específicamente a los campos FECHA_APERTURA y FECHA_CIERRE ya que de origen traen un tipo de datos *date time* y lo que se requiere es informar únicamente la fecha, por lo tanto se realiza la transformación correspondiente por medio de sentencia SQL.

Un paso previo a la fase de carga de información, se diseñó la plantilla del archivo Excel con las siguientes características:

- Logotipo empresa
- Título de reporte
- Fecha actual
- Fecha rango (semanal o mensual)
- Gráficas:
    - Total de tickets
    - Tickets por usuario
    - Tickets por relevancia (ALTA, MEDIA, BAJA)

El pre diseño de la plantilla es el siguiente:

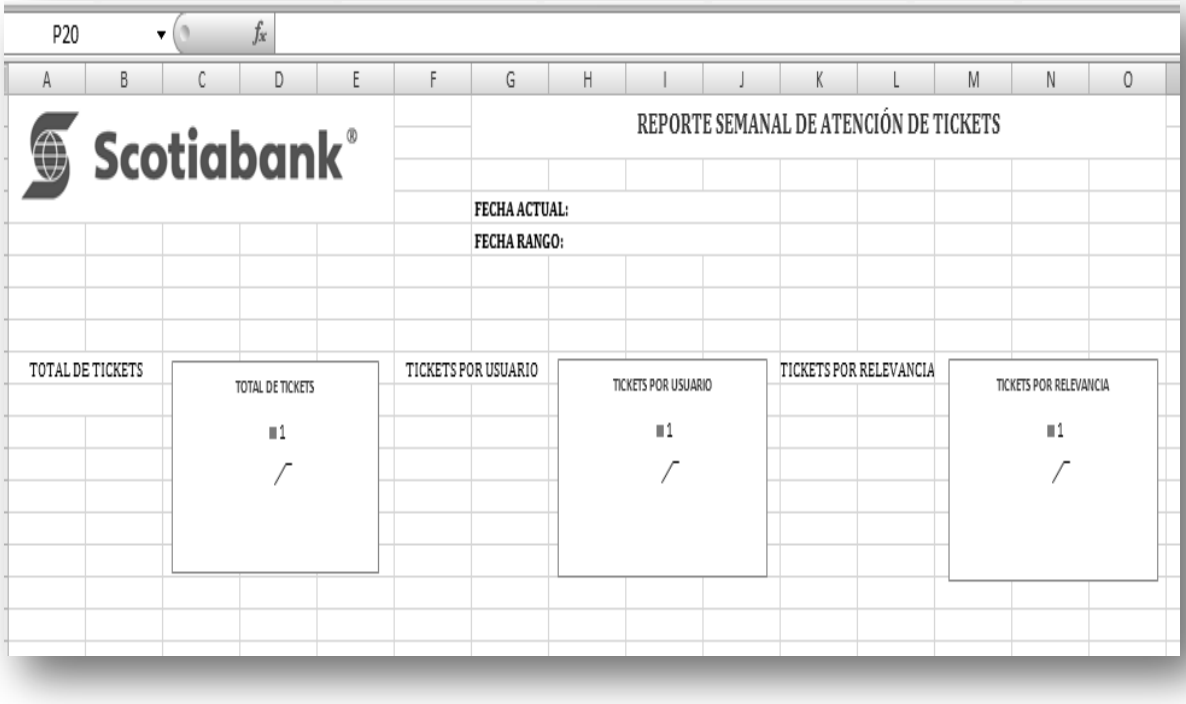

Figura 1.29 Plantilla de reporte semanal automático.

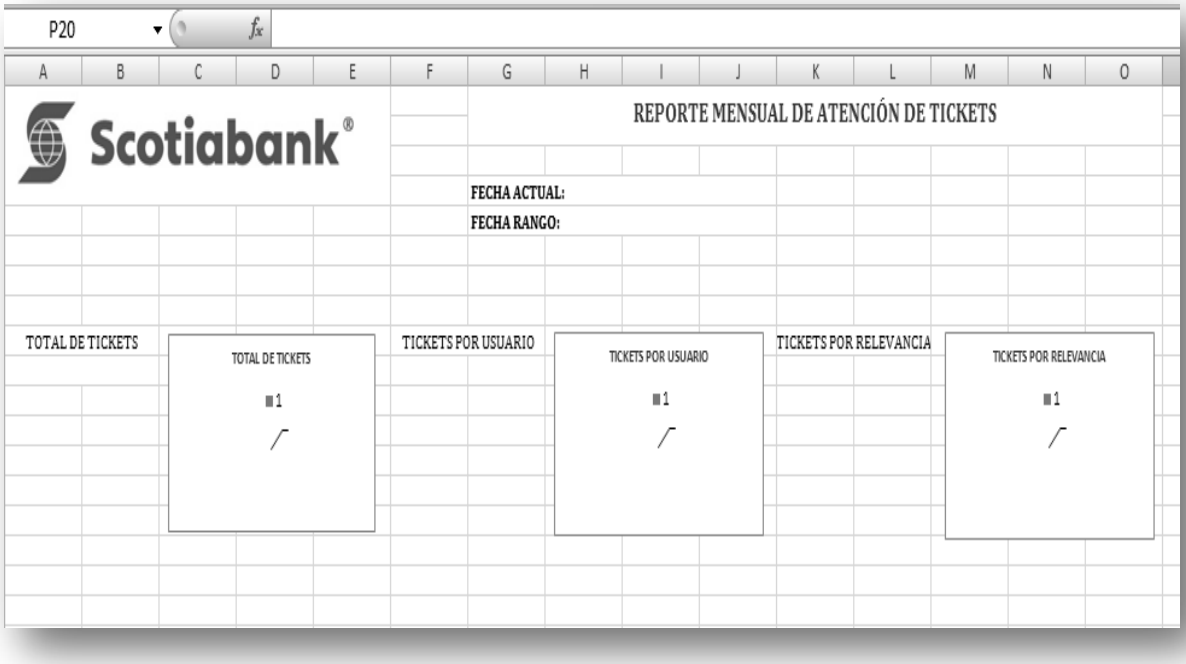

Figura 1.30 Plantilla de reporte mensual automático.

Es importante mencionar que los campos de la plantilla se actualizan automáticamente con la ejecución del proceso final.

3. Proceso de carga

   En esta última fase del proceso ETL se realiza la carga de información mediante sentencias SQL para el llenado del archivo Excel y funciones propias de la herramienta Pentaho. Los reportes se depositarán en un directorio específico para su extracción y uso correspondiente.

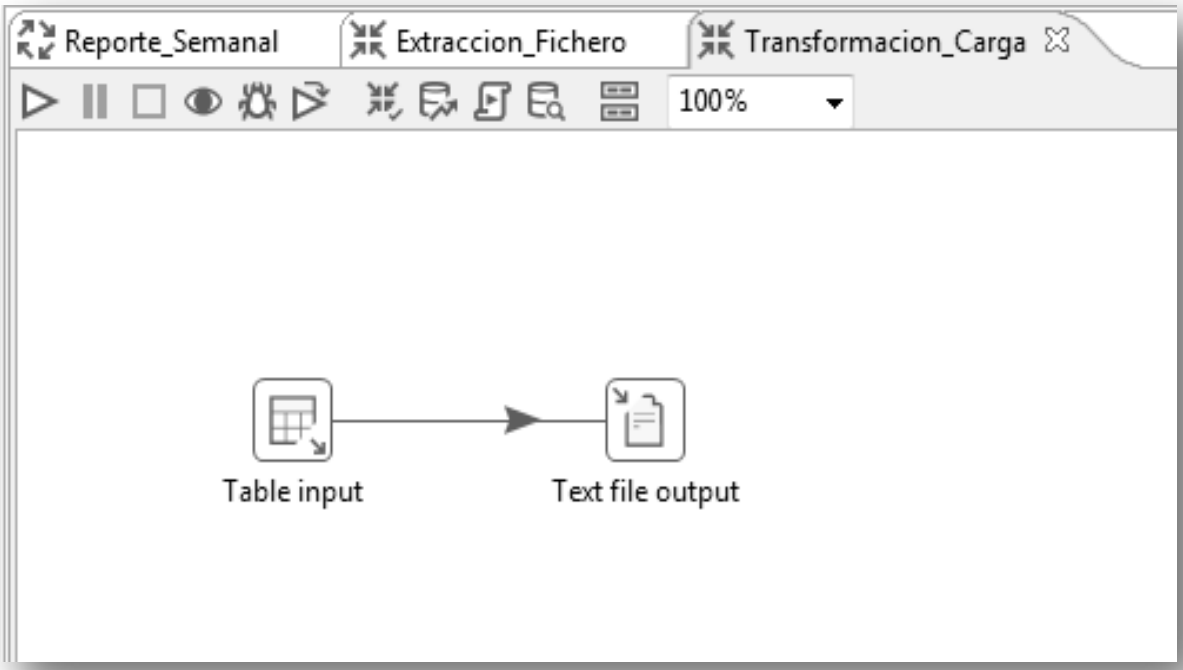

Figura 1.31 Proceso de transformación y generación de archivo .xls

Reportes automáticos finales:

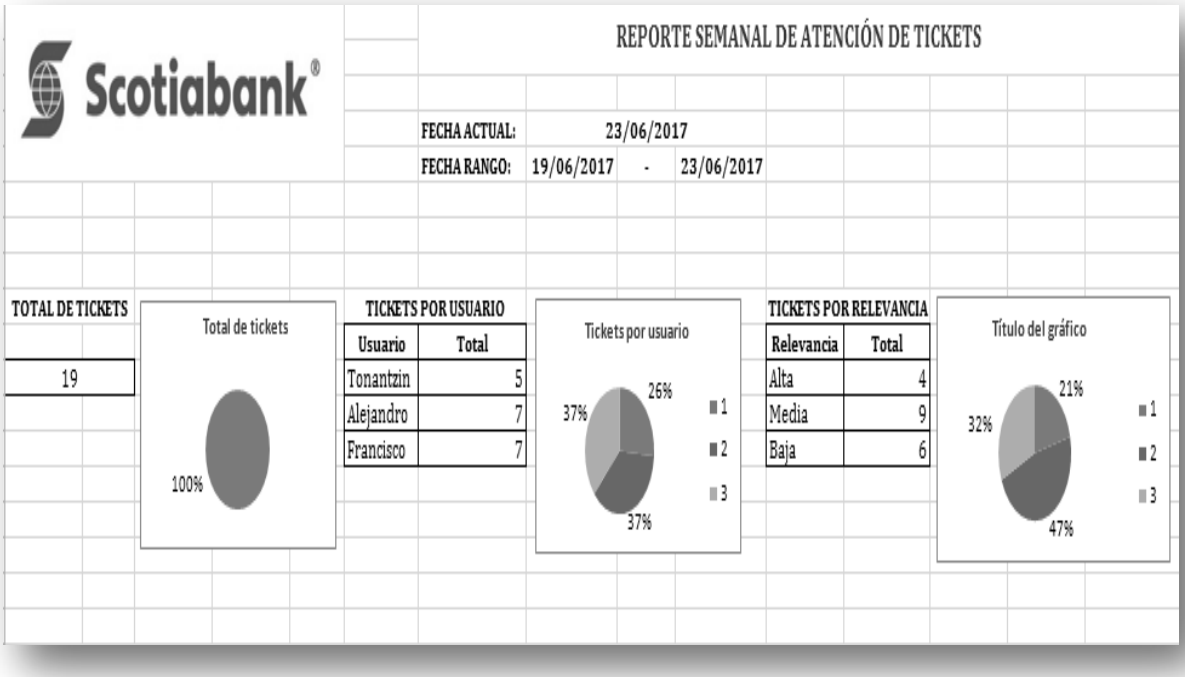

Figura 1.32 Reporte semanal final de atención de tickets.

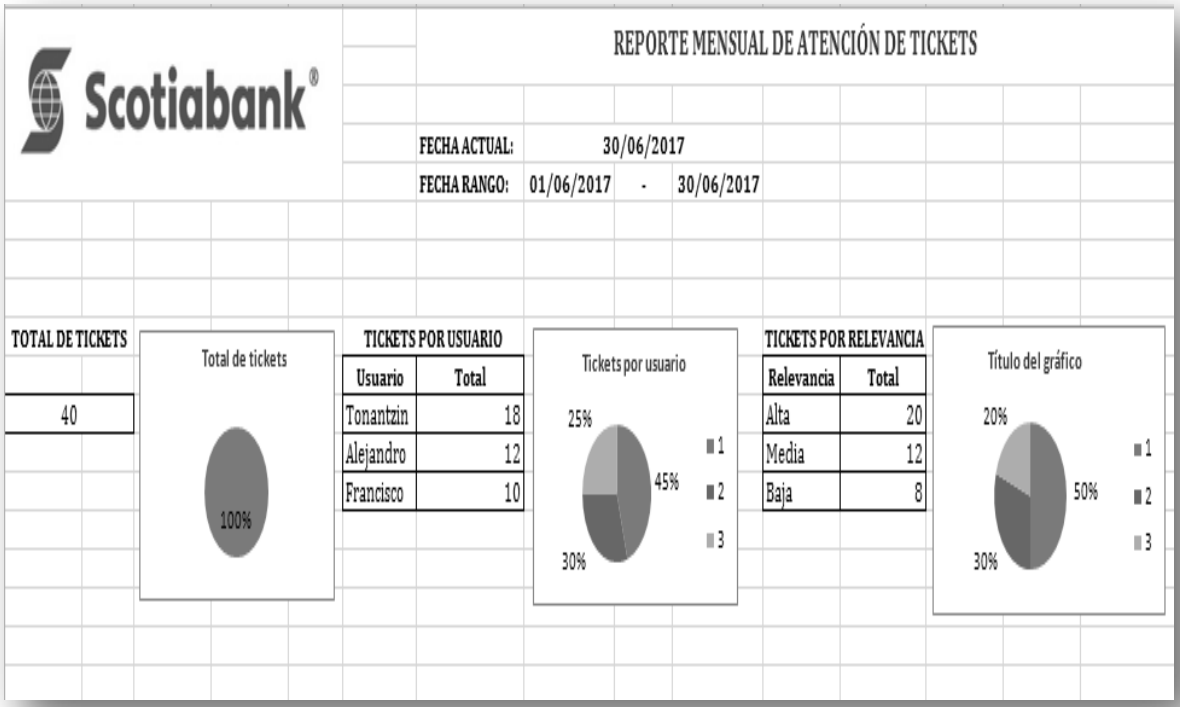

Figura 1.33 Reporte mensual final de atención de tickets.

Considerando que el llenado de los reportes se realizaba únicamente con la información correspondiente a días hábiles para el banco.

Para finalizar, cuando los reportes semanales y mensuales se han completado, son depositados en un directorio específico para después como complemento adicional ser enviados por correo electrónico de manera automática a los destinos establecidos.

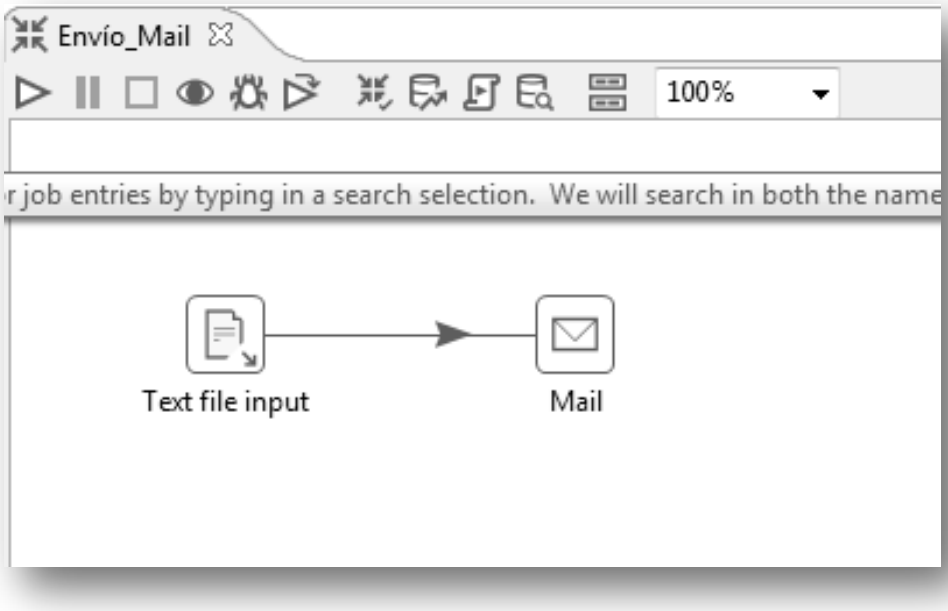

Figura 1.34 Envío automático de correo electrónico.

2.8.2. Consejos y/o consideraciones

De acuerdo con el manejo de procesos ETL se enlistan algunas consideraciones que encuentro importantes para el desarrollo de procesos ETL.

1. Conocer las necesidades y requerimientos de los procesos que se requieren implementar. Una organización que hace uso de un gran volumen de datos, siempre tendrá un fin a cometer, desde alguna exigencia operacional, análisis, extracciones, transformaciones o carga de datos a un destino correspondiente.
2. Identificar la herramienta correcta para el desarrollo de los procesos ETL que se acople a las necesidades del proyecto, presupuesto, recursos y al manejo de la información.
3. Ejecutar pruebas en tiempo y forma con la finalidad de detectar el índice de errores si estos existieran y así poder corregirlos evitando atrasos en desarrollo de los procesos.
4. Con un buen análisis e implementación es muy probable que se obtenga una buena integración de datos. Aunque no es fácil alcanzar la excelencia en este punto si se cumple lo anterior, los resultados serán significativos.

2.8.3. ETL como experiencia

Durante el tiempo que llevo trabajando con procesos ETL en el sector financiero y en otros ámbitos, puedo decir que no son procesos fáciles de aprender y desarrollar, pues se requiere de un amplio conocimiento en la herramienta con la que se va a implementar. Sin embargo para mí ha sido un gran reto y/o desafío sacar en tiempo y forma los proyectos en los que me he visto involucrada., así como desarrollar un amplio nivel lógico, análisis y amplio conocimiento para poder desarrollar procesos de mejora en tiempo y forma.

Considero que para tener una buena implementación de un proceso ETL es importante tener bien definidos y entendido los alcances y mejoras del proyecto. Si esto no sucede nos podemos encontrar con ciertas pausas, lo cual implique que la planeación se vea con cambios considerables y lleguen afectar a la organización.

El uso de los procesos ETL me ha servido para obtener una integración de datos adecuada para tener buenos resultados. También ha ayudado a comprender que los procesos no solo se utilizan en entornos de *Data Warehousing* o construcción de un *Data Warehouse,* sino que pueden ser útiles para una cantidad de propósitos, como por ejemplo; tareas de Bases de Datos, migración de datos entre diferentes aplicaciones, etc.

3. CONCLUSIONES

El presente ensayo tuvo como objetivo conocer el uso de los procesos ETL para la gestión de datos empresariales. Debido a la aparición de nuevas tecnologías en el mercado actual y que la explotación de datos plantea un gran desafío, lo anterior provoca una mayor complejidad al momento de gestionar y tener como resultado calidad y claridad de los datos.

Para conocer lo anterior, primero se realizó una investigación y análisis de los cambios que han tenido los procesos ETL desde sus inicios hasta la actualidad. También se detallaron las fases que forman parte de los procesos, así como herramientas que se encuentran en el mercado actual para implementación de procesos ETL, esto con la finalidad de elegir la herramienta de manera transparente y eficaz de acuerdo a las necesidades y recursos de las organizaciones.

El proyecto cumple con todas las exigencias tanto técnicas como ejemplos prácticos reales. Dichos ejemplos tuvieron un grado de dificultad que fue disminuyendo conforme la práctica y el conocimiento aplicado. Puedo comentar que la implementación y ejecución de procesos ETL, lograron un cambio en cada organización en la que participe. La manera en que se fueron eliminando manualidades permitió a los líderes tomar decisiones en tiempo para generar resultados positivos en la organización.

Definitivamente el uso correcto de los procesos ETL y la buena elección de la herramienta, permitirá obtener una buena gestión de datos, evitando así manipulaciones manuales y baja agilidad en las funciones que se estuvieron atendiendo, así como una buena integración de datos alcanzando un entorno adecuado.

En la actualidad las organizaciones han ido eliminando periódicamente procesos manuales, sustituyéndolos por procesos automáticos. Esto no quiere decir que el trabajo humano se está olvidando o reemplazando por maquinas, únicamente existen actividades que pueden ser programados y ejecutados en el tiempo estimado y el tiempo ganado puede servir para otras actividades que impulsen a las empresas a su crecimiento.

4. REFERENCIAS BIBLIOGRÁFICAS

[1] El valor estratégico de las empresas, Posted 2016.
http://www.datacentric.es/base-datos-clientes/ Última visita: 2017.09.21

[2] Inteligencia de Negocios. Formato PDF.
http://www.oracle.com/ocom/groups/public/@otn/documents/webcontent/317529_esa.pdf
Última visita: 2017.09.21

[3] PowerData, Especialista en Gestión de datos, Blog, Procesos ETL: Definición, Características, Beneficios y Retos.
http://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/312584/procesos-etl-definici-n-caracter-sticas-beneficios-y-retos Última visita: 2017.10.05

[4] Master Data Management (MDM)
https://www.powerdata.es/mdm Última visita: 2017.10.05

[5] PowerData, Especialista en Gestión de Datos. Guía gratuita, Los Procesos ETL en profundidad. Última visita: 2017.11.14

[6] PowerData, La Base de la Inteligencia de Negocio. Guía gratuita, Procesos ETL: Extract, Transform, Load. Última visita: 2017.11.14

[7] PowerData, Especialista en Gestión de Datos. La evolución de los procesos ETL, PostedonThu, Jun 27,2013.
https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/288890/La-evoluci-n-de-los-procesos-ETL Última visita: 2017.11.14

[8] SAS, The Power to Know. ETL.
https://www.sas.com/en_us/insights/data-management/what-is-etl.html
Última visita: 2017.12.15

[9] Pentaho Kettle Solution, Building Open Source ETL Solutions with Pentaho Data Integration, Matt Casters, Roland Bouman, Jos Van Dongen, Editorial Wiley, 2010. Última visita: 2018.01.09

[10] Pentaho Data Integration, Beginner's Guide-Open Source, Second Edition, Mariana Carina Roldan, Editorial PACKT Publishers, 2013. Última visita: 2018.01.10

[11] Sinnexus, Sinergia e Inteligencia de Negocios S.L. Postedon 2016, España. http://www.sinnexus.com/empresa/index.aspx Última visita: 2018.01.15

[12] IBM, Recopilación de datos con procesos ETL. https://www.ibm.com/support/knowledgecenter/es/SSYMRC_4.0.7/com.ibm.rational.reporting.overview.doc/topics/c_ovr_process_etl.html Última visita: 2018.01.15

[13] ETL Tools, ETL and Data Warehousing portal. http://www.etltools.org Última visita: 2018.01.15

[14] CDC (Change Data Capture) http://www.oracle.com/technetwork/es/articles/datawarehouse/oracle-change-data-capture-1545279-esa.html Última visita: 2018.01.16

[15] DATAPRIX, Knowledge Is The Goal. Herramientas ETL, ETL's Open Source, Posted on Feb, 2010. http://www.dataprix.com/blogs/respinosamilla/herramientas-etl-que-son-para-que-valen-productos-mas-conocidos-etl-s-open-sour Última visita: 2018.01.16

[16] Mundo Business Intelligence, Córdoba Argentina. https://mundobi.wordpress.com/2007/06/24/herramientas-etl-%E2%80%A6o-mundo-etl/ Última visita: 2018.01.20

[17] Pentaho Data Integration Cookbook, Second Edition, Adrian Sergio Pulvirenti, Alex Meadows, Maria Carina Roldan, Editorial PACKT Publishers, 2013. Última visita: 2018.01.20

[18] Pro Power BI Desktop, Free interactive data analysis with Microsoft Power BI, Adam Aspin, Editorial Apress, 2016. Última visita: 2018.02.04

[19] Talend for Big Data, BahaaldineAzarmi, Editorial PACKT Publishers, 2014. Última visita: 2018.02.06

[20] Learning InformaticaPowerCenter 9.x, Rahul Malewar, Editorial PACKT Publishers, 2014. Última visita: 2018.02.06

[21] IBM Cognos Data Manager, Publicado en 2014.
https://blog.es.logicalis.com/analytics/tm1-cognos-data-manager-caracteristicas-y-ventajas
Última visita: 2018.02.15

[22] Herramienta para procesos ETL
https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/312655/C-mo-Elegir-la-Herramienta-para-Procesos-ETL-m-s-Adecuada Última visita: 2018.02.15

[23] Características de las herramientas ETL, Posted o Thu, Jul 25, 2013.
https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/289586/9-Caracter-sticas-de-las-herramientas-ETL Última visita: 2018.02.15

[24] IBM Analytics, Calidad de Datos, PostedonAug, 2015.
https://www.ibm.com/analytics/es/es/technology/data-quality/ Última visita: 2018.02.20

[25] Calidad de Datos, PostedonAug, 2016.
https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/la-importancia-del-data-quality-en-los-analisis-de-datos Última visita: 2018.03.05

[26] El valor de la gestión de datos, Posted on Wed, Nov 13, 2013
https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/348870/qu-se-entiende-por-integridad-de-los-datos Última visita: 2018.03.15

[27] INFORMATICA, Informatica Data Quality
https://www.informatica.com/mx/products/data-quality.html Última visita: 2018.04.18

[28] Big Data, Cuadrante Mágico de Gartner
http://www.bigdata-social.com/informe-cuadrante-magico-gartner/
Última visita: 2018.04.29

[29] Evaluando Software (Intertek), Posted Nov 15, 2017
http://www.evaluandosoftware.com/etl-extraccion-transformacion-carga-datos/
Última visita: 2018.04.29

[30] Fundamentos de ETL, Posted 2015
https://academia.soydata.net/courses/procesos-etl-pentaho-data-integration/lectures/3037305 Última visita 2018.08.23

[31] Procesos ETL, Introducción, Posted Sep, 2014
https://rocalla.wordpress.com/2014/09/15/procesos-etl-introduccion/
Última visita: 2018.08.23

[32] Staging: la salvaduarda de los procesos ETL, Posted Aug, 2013
https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/bid/312643/staging-la-salvaguarda-de-los-procesos-etl Última visita: 2018.08.24

[33] Historia de las herramientas ETL, Posted April, 2017
https://blog.bi-geek.com/historia-las-herramientas-etl/ Última visita: 2018.08.24

[34] BI y Modelos multidimensionales, Posted Jan, 2012
https://www.businessintelligence.info/definiciones/que-es-modelo-dimensional.html
Última visita: 2018.08.24

[35] Modelos multidimensionales, Posted May, 2018
https://docs.microsoft.com/es-es/sql/analysis-services/multidimensional-models/multidimensional-models-ssas?view=sql-server-2017&viewFallbackFrom=sql-analysis-services-2016 Última visita: 2018.08.24

[36] Talend, Posted Jul, 2018
https://www.talend.com/resources/what-is-etl/ Última visita: 2018.08.24

Imágenes

[37] Rosalía Arroyo (2015). ETL, el gran coste del Big Data. [Figura 1.2]. Recuperado de https://www.channelbiz.es

[38] James Weir (2018). 2018 *Census data quality management strategy*. [Figura 1.3]. Recuperado de http://archive.stats.govt.nz/

[39] Big Data Social. Cuadrante Mágico de Gartner. [Figura 1.4]. Recuperado de http://www.bigdata-social.com

[40] Maria Carina Roldan (2013). *Pentaho Data Integration*. [Figura 1.5]. Recuperado de *Pentaho Data Integration Cookbook, Second Edition.*

[41] SoftLandMark (2015). *Talend Open Studio* 5.5.0. [Figura 1.6]. Recuperado de http://business.softlandmark.com/

[42] BahaaldineAzarmi (2014). Fase de Modelo de Negocio. [Figura 1.7]. Recuperado de *Talend for Big Data*, Editorial *PACKT Publishers*

[43] BahaaldineAzarmi (2014). Fase de Diseño de Trabajos. [Figura 1.8]. Recuperado de *Talend for Big Data*, Editorial *PACKT Publishers*

[44] BahaaldineAzarmi (2014). Fase de Contextos. [Figura 1.9]. Recuperado de *Talend for Big Data*, Editorial *PACKT Publishers*

[45] BahaaldineAzarmi (2014). Plataforma Talend. [Figura 1.10]. Recuperado de *Talend for Big Data*, Editorial *PACKT Publishers*

[46] ETL Quality Checker (2017). *Informatica PowerCenter ETL Logic*. [Figura 1.11]. Recuperado de http://www.disoln.org/

[47] Progress (2018). *IBM Cognos Data Manager*. [Figura 1.12]. Recuperado de https://www.progress.com