FACULTAD DE CIENCIAS

DEPARTAMENTO DE MATEMATICAS

EXPERIENCIA PROFESIONAL

Implementación de un nuevo modelo de consulta de información y reestructura de arquitectura de datos

Trabajo por experiencia profesional presentado por Erick Celso Zavaleta González para obtener el grado de Licenciado en Ciencias de la Computación

---

Supervisados por: Dr. Canek Peláez Valdés

Dedicado a...

Índice general

Índice de figuras

Índice de tablas

Agradecimientos

¡Muchas gracias a todos!

Capítulo 1

Introducción

El proyecto que se presenta en este trabajo fue la implementación y reestructura de la infraestructura de base de datos, así como los cambios en procesos y arquitectura de datos de una institución financiera que por razones de acuerdos de confidencialidad no puedo mencionar su nombre. Dentro de las actividades que realicé se encuentra un análisis del software a utilizar, así como la definición y desarrollo de programas para convertir los datos de diferentes bases de datos y plataformas hacia un repositorio de datos único que ayudó a eficientar los tiempos de respuesta y las consultas realizadas.

El marco teórico del proyecto se basa en la arquitecura de datos, que en tecnología de la información se compone de los modelos, políticas, reglas y/o estándares que determinan qué datos se recolectan y cómo se guardan, ordenan, integran y usan en los sistemas de datos de una institución.

Hablar de datos es entrar en un mundo complejo donde existen diferentes perspectivas de datos, dependiendo del uso que se les quiera dar o de las personas que lo utilizan. Cada grupo de personas tiene su propia perspectiva sobre el manejo de datos: manejo de grandes volúmenes, acceso al detalle de los datos de manera instantánea, manejo de la integridad, datos de acceso exclusivo, etc. La arquitectura de datos nos ayuda a que estos diferentes tipos de datos y diferentes necesidades puedan coexistir de una forma conjunta de acuerdo a las necesidades de cada área o empresa.

Actualmente no existe ningún secreto para el manejo de datos y su respectiva arquitectura; en ambos casos, es importante entender los datos en términos de su infraestructura; es decir, se requiere conocer la infraestructura que rodea a los datos para llevar a cabo un uso adecuado de los mismos.

Así mismo, hacemos notar que dentro de cualquier empresa u organi-

zación se pueden encontrar diferentes tipos de datos: estructurados y no estructurados. Los primeros son datos predecibles y que normalmente son manejados en una base de datos (SMBD): registros, atributos, llaves, índices, etc. Por su parte, los datos no estructurados no son predecibles y, como su nombre lo indica, no tienen una estructura bien definida, usualmente son de dificil acceso y generalmente se requiere una búsqueda más profunda para hacer consultas; por ejemplo, una cadena de carácteres en un texto libre.

Todos estos conceptos serán detallados y aplicados en este trabajo.

Capítulo 2

Análisis de requerimientos

En este capítulo...

2.1. Proceso de selección de software

Parte importante para llegar a la solución propuesta fue la selección del software en el cual se desarrolló el nuevo proceso ETL [1]. La selección de software requirió de varios pasos previos a la toma de la decisión. El proceso de selección de software de acuerdo a la metodología de Accenture fue la siguiente:

1. Establecer los requerimientos de negocio.
2. Establecer los requerimientos de evaluación de las herramientas de ETL.
3. Generar una lista larga de proveedores.
4. Generar una lista corta (generalmente 3 o 4 prveedores) de candidatos.
5. Realizar un RFP (Request For Approval) para evaluar a los candidatos de manera individual y detallada.
6. Realizar la recomendación final del producto.

Este proceso se datalla en la figura 2.1:
En las siguientes secciones se definirán cada uno de los pasos del proceso seguidos.

[1] Extracción, Transformación y Carga (*load* en Inglés).

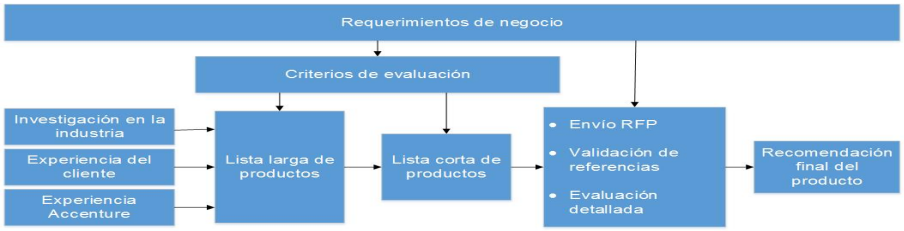

Figura 2.1: Proceso de selección de software

2.1.1. Requerimientos funcionales

Con base en las reuniones realizadas con el área de desarrollo de sistemas de la institución financiera se identificaron una serie de requerimientos funcionales detallados a continuación.

- Se requería una conexión a correo electrónico mediante le protocolo SMTP para envío de informes de termino de procesos o de error en los mismos.
- Tener un log de procesos que tuviera la información de cada uno de los procesos de extraccón/carga (inicio, fin, estatus de termino, errores (en cso de haberlos)).
- La herramienta debía de tener la capacidad de contar con un esquema de limpieza de datos que permitiera identificar y corregir los errores en nombres, direcciones, RFC's, números de teléfono, y algunos otros datos relevantes para la institución financiera.
- Permitir invocar a la herramienta ETL desde otras aplicaciones tales como PeopleSoft, aplicaciones desarrolladas en Java y .NET.
- Permitir control y detalle de los errores presentados en los procesos o en cada uno de los componentes del flujo de datos. El control se llevaría a cabo mediante el envío, por parte del ETL, de errores a tablas log o archivos de texto en dode se pudieran identificar de una manera sencilla, los errores presentados.
- Contar con un esquema de datos redundante, como puede ser un cluster, bases de datos replicadas, balanceo de cargas con tolerancia a fallos, etc.

- Contar con un esquema de control de versiones que soporte el trabajo de multiples usuarios al mismo tiempo.
- La herramienta debía tener la capacidad para entregar diferentes archivos a los sistemas destino (archivos de texto, tablas en base de datos, XML, PDF, Excel).
- Contar con un esquema de seguridad basado en roles y jerarquías de usuario que permitieran a los usuarios acceder a datos específicos de la base de datos de acuerdo a su rol.
- Contar con pantallas de configuración que permitieran a los administradores de la herramienta tener un mayor control sobre los procesos y los usuarios. Así mismo, que fuera una herramienta visualmente fácil de usar.
- Permitir a la nueva herramienta la convivencia con las tecnologías con las que contaba la institución financiera (Windows 2003, SQL Server 2005, Oracle, DB2, AS400, Windows Vista y Mainframes)
- Contar con una herramienta de fácil manejo de metadatos.
- Contar con un ambiente de diseño y desarrollo 100 % visual que permitiera a los desarrolladores implementar las soluciones de extracción y carga de una manera sencilla.
- Permitir la operación y administración de la herramienta de una forma remota.
- Generar componentes reutilizables entre aplicaciones y entre procesos.
- La herramienta debía permitir a los usuarios, la creación de funciones personalizadas que cumplieran con los estándares de la empresa y que no fueran parte de las configuraciones predefinidas de la herramienta.
- Capacidad para calendarizar procesos; es decir, la herramienta debería ser capaz de ejecutar los procesos de forma automática a la hora y día especificados.
- Soportar modelos de minería de datos que ayudarían a la institución financiera a realizar un anaálisis de sus datos y poder definir estrategias de mercado para los diferentes productos que ofecían.

- Capacidades de soporte para DataWarehouse y Business Intelligence, este requerimiento era para cumplir con el plan estratégico de la institución.

- Capacidad de la herramienta para plublicar procesos como Web Services a fin de poder acceder a ellos de manera remota o ejecutarlos a través de un portal con los permisos correspondientes.

- Capacidades de detección y correción de errores en los procesos, mediante un proceso de depuración (Debugging) que sería ejecutado paso a paso.

- Como un requerimiento básico, la herramienta debería de almacenar la información extraída en una sola fuente de datos, para la solución propuesta fue una base de datos de paso llamada base de datos stage.

- Explotación de la información de una base centralizada para realizar reportes.

- Consulta de datos históricos, en especial de transacciones, para poder ser explotados desde otro ambiente o aplicación existente en la institución financiera.

- La herramienta debería de ser capaz de obtener solamente la información que tuvo cambios entre los diferentes periodos de extracción, ayudando al rendimiento del proceso y extrayendo una cantidad menor de datos.

2.1.2. Requerimientos técnicos

Por su parte los requerimientos técnicos solicitados por la institución para seleccionar la herramienta fueron los siguientes:

- Mejorar el rendimiento de la extracción de datos de 90 minutos a 45 minutos (tiempo tentativo) con una cantidad de 6 millones de registros

- Mejorar el tiempo de transformación de datos utilizando componentes para limpieza y calidad de los datos.

- Mejorar el rendimiento de la herramienta cuando el volumen de datos exceda los 100 millones de registros.

- Que la arquitectura de la herramienta fuera compatible con la arquitectura de la institución y la futura que se planteó en el presente proyecto.
- La herramienta debía ser capaz de poder integrarse son los sistemas de la institución (PeopleSoft, Core bancario, systema de reportes, sistema de recursos humanos, pagos a terceros, etc.)
- La herramienta debía ser capaz de ejecutarse en diferentes paltaformas (Windows, Linux, Unix, Mainframes).
- Conectividad con los istemas fuentes (ODBC, OLAP, LDAP) y con diferentes sistemas manejadores de bases de datos como Oracle, DB2, SQL Server, Informix, Sybase, Teradata.
- Capacidad para crear funciones y métodos desde cualquier lenguaje de programación estructurado y que pudiera ser ejecutado desde la herramienta ETL.

Aunque el gobierno de datos no era parte del alcance del proyecto, pero si de la estrategia de crecimiento de la institución, la herramienta debería de soportar esta funcionalidad.

2.1.3. Requerimientos de limpieza de datos

Como parte de la definición de requerimientos funcionales que debía cumplir la herramienta, existe una sección que tenía que ver con la calidad y limpieza de datos; estos requerimientos solicitados por parte del área de tecnoloía fueron los siguientes:

Se requería que el proceso ETL realizara la limpieza de datos de los campos descritos a continuación:

- RFC. Se requería que el RFC se encontrara estandarizado y cumplieran con los requerimientos establecidos por el buró de crédito
- Nombres de socios. Se requería que los nombres de socios no tuvieran caracteres especiales como puntos, comas, retornos de carro, paréntesis, corchetes, porcentaje, comillas y caracteres de 16 bits.
- Fechas de morosidad. Se solicitó que las fechas para el cálculo de la morosidad se encontraran en un formato de fecha correcto yyyymmdd y que no se permitieran valores nulos o negativos para este tipo de dato.

- Direcciones de socios. Fue requerido que las direcciones de los socios tuvieran una nomenclatura estándar para nombres de calles, colonias
- Ciudades/Municipios. Los códigos y nombres de ciudades y municipios debían de contar con un estándar y estos deberian de estar corregidos de a cuerdo a la información proporcionada por SEPOMEX.
- Teléfonos. Los teléfonos de los clientes debían contar con la clave lada y todos tenían que ser de 10 dígitos para números fijos y de 13 dígitos para números celulares
- Los números telefónicos no deberían de tener guiones o caracteres especiales.

Como parte de la estandarización de direcciones era requerido que todos los códigos postales se encontraran conforme a la lista proporcionada por SEPOMEX y todos deberían de ser de 5 dígitos.

2.2. Criterios de evaluación

Como mencionamos al inicio del documento, parte importante de la selección de sofware fue la definición de los criterios de evaluación de la herramienta ETL a utilizar como parte del proyecto. Dichos criterios nos ayudaron a realizar una evaluación objetiva de las diferentes herramientas y tener un panorama general de los productos que cumplían con las caracteristicas y necesidades de la institución financiera. Tomando como base esas necesidades y la experiencia de Accenture se definieron los criterios de selección de software.

Los criterios se agruparon en conceptos genericos que describen la funcionalidad de cada área. Estos criterios se listan a continuación:

1. *Capacidades del servicio.*
2. *Opciones de integración.*
3. *Ambiente de la herramienta.*
4. *Soporte y capacitación.*
5. *Técnicas adicionales de integración de datos.*
6. *Manejo de la información.*

7. *Estrategias del producto.*
8. *Estrategias corporativas.*
9. *Costos.*
10. *Convenios con otros proveedores.*
11. *Finanzas de la compañia.*

2.2.1. Lista de proveedores

Una vez que se definieron los criterios de evaluación y se realizó el análisis de requerimientos con el equipo de la institución financiera, el siguiente paso fue dar una lista larga de proveedores candidatos para ser evaluados y que en ese momento eran las mejores opciones dentro del mercado. La lista contenía a 6 proveedores de software: Oracle (Oracle Warehouse Builder), Informática (Power Center), IBM (Infosphere Information Server), Microsoft (Integration Services (SSIS), SAP (SAP-Business Objects), SAS Institute (SAS).

2.2.2. Evaluación de proveedores

Con la lista larga de proveedores definida, la siguiente tarea que se llevó a cabo fue la evaluación de cada uno de los proveedores con el fin de establecer un grupo reducido de candidatos que representaría la lista final de selección de la herramienta ETL a utilizar dentro del proyecto. De acuerdo a las necesidades de la Institución financiera se definieron los siguientes porcentajes para los criterios de evaluación

La prioridad de cada uno de los criterios de evaluación se realizó con base en las necesidades de la institución y basados también en la funcionalidad de cada uno de ellos.

Con toda la información que se recolectó, realicé un análisis de cada proveedor y su herramientas presentadas, así como sus factores diferenciales y las características principales soportadas. Esta evaluación se presenta en la siguiente tabla con los resultados finales:

2.3. Lista corta de proveedores

Con base en los criterios de selección y al análisis de las capacidades de cada una de las herramientas, se obtuvo la lista corta de los proveedores. Los proveedores seleccionados fueron los siguientes:

Criterio	Porcentaje de evaluación	Prioridad
Capacidades del servicio	14 %	1
Manejo de la información	14 %	2
Opciones de integración	11 %	3
Ambiente de la herramienta	11 %	4
Costos	11 %	5
Soporte y capacitación	9 %	6
Estrategias del producto	6 %	7
Convenios con otros proveedores	6 %	8
Técnicas adicionales de integración de datos	6 %	9
Estrategias corporativas	6 %	10
Finanzas de la compañia proveedora	6 %	11

Tabla 2.1: Porcentajes y prioridades de evaluación.

- IBM
- Informática
- Oracle
- Microsoft

Si bien SAP-Business objects tenía una calificación mayor, el negocio decidió no incluirlo en la lista final debido a que no se contaba con capacitación suficiente de parte del proveedor. Así mismo, Microsoft se incluyó a petición explicita de la institución.

La recomendación proporcionada por Accenture fue Informática debido a que al ser una empres independiente enfocada a la integración de datos, no estaba casado con ninguna base de datos específica, como si lo están el resto de los participantes , y esto permitiría una mejor integración a las necesidades de la institución. Basados en esta recomendación la institución financiera se decidió por esta solución para implementar sus nuevos procesos ETL.

		Microsoft SSIS	Oracle OWB	Informática Power center	IBM IIS	SAPBusiness Objects	SAS
	Porcentaje						
Capacidades del servicio	**14.29 %**	**3**	**3.5**	**4.5**	**4**	**3.75**	**4**
Escalabilidad y rendimiento		4	4	5	5	4	4
Alta disponibilidad		4	3	4	4	4	3
Seguridad		3	4	5	3	4	5
Plataformas de ejecución soportadas		1	3	4	4	3	4
Opciones de integración	**11.43 %**	**3**	**3**	**4**	**4**	**3.4**	**3.6**
Conectividad con sistemas fuente		4	3	5	5	5	5
Conectividad para la carga		4	3	5	5	5	5
Servicios Web		4	4	5	5	3	3
Conexión a correo electrónico		1	1	1	1	0	1
Reusabilidad		2	4	4	4	4	4
Ambientes de la herramienta	**11.43 %**	**3.2**	**4**	**4.4**	**4.2**	**4.4**	**3.6**
Visualización del ambiente de diseño y desarrollo		4	4	4	4	4	4
Manejo de errores		4	4	5	4	5	4
Ambiente de colaboración		2	4	4	4	4	3
Manejo y modelado de datos y metadatos		2	4	4	4	4	4
Administración		4	4	5	5	5	3
Soporte y capacitación	**8.57 %**	**4.25**	**4.75**	**4.25**	**4.5**	**3.75**	**4.75**
Soporte		4	5	5	5	5	4
Capacitación		4	5	3	5	3	5
Documentación		4	4	5	5	3	5
Soporte a diferentes lenguajes		5	5	4	3	4	5
Técnicas adicionales de integración de datos	**5.71 %**	**4**	**2.5**	**4**	**3.5**	**3.5**	**1**
EII		3	3	4	4	3	0
Cambio en la captura de datos		5	2	4	3	4	2
Manejo de la información	**14.29 %**	**3.6**	**3.8**	**4.4**	**4**	**3.4**	**3.2**
Reglas de transformación		3	3	4	4	3	3
Perfiles de datos		5	5	4	5	4	4
Calidad de datos		4	3	4	5	4	4
Visualización de datos		2	4	5	2	4	4
Contenido no estructurado		4	4	5	4	2	1
Estrategias de producto	**5.71 %**	**4**	**4**	**4**	**5**	**5**	**3**
strategias corporativas	**5.71 %**	**3**	**3**	**4**	**3.5**	**3**	**3**
Contibución del producto		3	3	5	3	3	3
Porcentaje de ganancias		3	3	3	4	3	3
Costos	**11.43 %**	**4.2**	**3.8**	**2.4**	**2.4**	**3.2**	**2.4**
Promedio del percio de venta		5	5	1	1	2	2
Estructura de precios		3	3	1	1	3	1
Modularidad de precios		5	5	5	5	5	5
Pruebas de concepto		3	3	3	3	3	3
Esquema de licenciamiento		5	3	2	2	3	1
Convenios con otros proveedores	**5.71 %**	**2.66**	**3.33**	**3.33**	**4.33**	**2.66**	**1.66**
Licenciamiento de terceros		1	2	4	5	2	1
Venta del producto por terceros		5	3	3	3	3	2
Integradores de sistemas		2	5	3	5	3	2
Finanzas de la compañia	**5.71 %**	**4**	**3**	**4.66**	**4.33**	**4.33**	**5**
Ganancias		3	2	5	5	3	5
Crecimiento de ganancias		4	2	4	3	5	5
Estatus del procveedor		5	5	5	5	5	5
Total	100.00 %	168.91	171.18	190.45	186.76	172.65	158.21
Calificación total		**3.54**	**3.52**	**4.0**	**3.98**	**3.67**	**3.20**

Tabla 2.2: Evaluación de proveedores.

2.4. Análisis de fortalezas y debilidades

2.5. Definición de los requerimientos

Capítulo 3

Diseño funcional y técnico

En este capítulo...

3.1. Arquitectura de aplicaciones (alto nivel)

La arquitectura definida para las aplicaciones destino y fuente que se consideraron dentro del alcance del proyecto se muestran en la siguiente figura; dentro de esta arquitectura de aplicaciones se consideró el proceso ETL que se construyó.

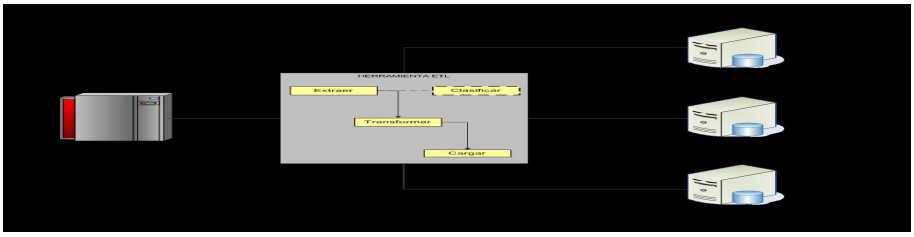

Figura 3.1: Arquitectura de aplicaciones.

La arquitectura definida contenía al core bancario como sistema fuente de información y enviaba la información a la herramienta ETL seleccionada para realizar los procesos de extracción, transformación, clasificación y carga hacia los sistemas destino y hacia cada uno de los tipos de datos que necesitaban los equipos y bases de datos destino.

3.2. Arquitectura de aplicaciones (arquitectura detallada)

La arquitectura diseñada se encontraba dividida en cuatro capas y representan el flujo de información entre cada una de dichas capas. La capa de interfaz de usuario donde se realizaron todas las tareas de monitoreo, manejo de la herramienta y visualización de datos en el sistema destino. La capa de presentación únicamente contenía el servidor de correo que envía la información a la capa de interfaz de usuario. Por su parte la capa de integración contenía los procesos de la herramienta ETL; así como la base de datos stage y la base de datos de catálogos; esta capa también incluía a las aplicaciones que envíaban información complementaria a los sistemas destino. Finalmente contabamos con la capa de datos que contenía a los sistemas destino, el sistema fuente y al servidor de control de usuarios a través del directorio activo. La funcionalidad de cada capa así como el flujo de datos se describen a continuación.

3.2.1. Capa de datos

La capa de datos contenía diferentes funcionalidades dependiendo del servidor de datos del que estemos hablando. El core bancario tenía la funcionalidad de enviar datos a la herramienta ETL a través del protocolo TCPIP. Por sus parte el directorio activo al manejar información de los usuarios de la herramienta tenía la funcionalidad de autenticar a cada usuario que accedía a la herramienta otorgándole los permisos necesarioa para ejecutar las tareas dentro de la herramienta ETL, la comunicación entre la herramienta ETL y el directorio activo se realizó mediante el protocolo LDAP.

Por su parte los sistemas destino deberían de tener la capacidad de recibir los datos de parte de la herramienta ETL; así como de parte de los ETL's actuales de la institución financiera y que complementaban la información del core bancario; los sistemas destino reciben los datos mediante el protocolo TCP/IP independientemente de la herramienta ETL que los envíe.

3.2.2. Capa de integración

La capa de integración tiene la funcionalidad de integrar los datos entre las diferentes aplicaciones; la herramienta ETL contiene todos los servicios para la extracción de datos del core bancario, transformarlos y cargarlos en los diferentes sistemas destinode la capa de datos. La herramienta ETL tiene

la funcionalidad de enviar los datos a la base de datos stage [1] así como a la base de datos de catálogos. El proceso debía enviar vía FTP los archivos de texto a una dirección específica dentro de un servidor para que el ETL de lsistema antilavado tome los archivos y complemente la información que será enviada vía TCP/IP a la aplicación de usuario final. A su vez, el proceso ETL también envía información a uno de los ETL's del sistema de reportes para que sea complementada por parte del sistema de conciliación bancaria y sea enviada a la base de datos final de reportes.

3.2.3. Capa de presentación

La capa de presentación de la institución financiera contenía sólamente un servidor de correo; la principal funcionalidad del servidor de correo era proporcionar a los usuarios la información de los procesos; es decir, recibía información del estatus de los procesos ETL y a su vez enviaría la información de este estatus al usuario final.

3.2.4. Capa de interfaz de usuario

La funcionalidad de la capa de interfaz de usuario era administrar la información que se recibe del servidor de correo así como de las diferentes aplicciones con las que contaba la institución financiera. Esta capa era capaz de soportar un ambiente gráfico que permitía a los usuarios administrar, construir y diseñar sus propios elementos.

3.3. Tecnología conceptual

A continuación definiremos el diseño conceptual de la arquitectura usada para el proyecto, este diseño conceptual está dividido en cinco grandes áreas: Ambiente de desarrollo, herramienta ETL, ambiente de operación, aplicaciones de la institución financiera y la capa de infraestructura; de cada una de ellas hablaremos durante este capitulo.

3.3.1. Ambiente de desarrollo

Dentro del ambiente de desarrollo se encuentran definidas todas las actividades que se requieren tener para la construcción de los procesos ETL,

[1] Las bases de datos llamadas stage, son bases de datos temporales o de paso antes de llevar la información a su destino final.

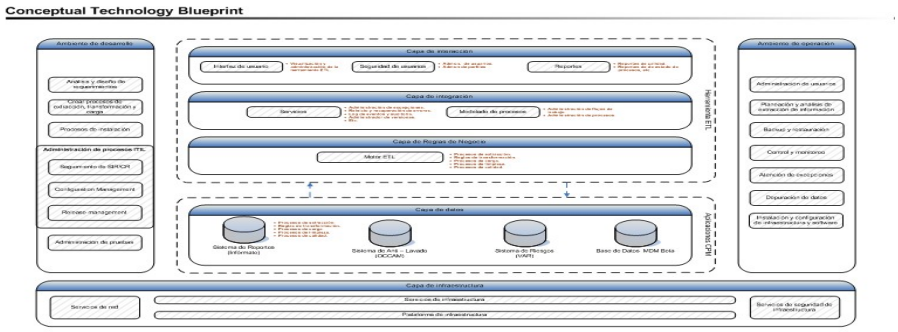

Figura 3.2: Tecnología conceptual.

así como la administración de los procesos y las pruebas. El ambiente de desarrollo tiene los siguientes componentes:

- *Análisis y diseño de requerimientos*: Dentro de la etapa de análisis y diseño de requerimientos encontramos la documentación de las fuentes y destinos de datos que serán procesadas mediante el ETL; el análisis de requerimientos abarca los requerimientos funcionales y técnicos que requiere la institución financiera para poder generar los datos en los sistemas destino. Por su parte el diseño de los requerimientos incluyó el diseño funcional y técnico de todos los procesos a implementar dentro del ETL así como la definición de las reglas de programación y estándares que se tenían que seguir para un mejor desarrollo del proceso.

- *Creación de procesos de extracción, transformación y carga*: Una vez realizado el análisis y diseño de los procesos, el siguiente paso es construir los procesos ETL dentro de la herramienta seleccionada; para realizar este desarrollo se deberá de basar en el diseño funcional, en las reglas de transformación y en las reglas de negocio definidas dentro del propio diseño. El proceso de extracción para la primera iteración fue solamente del Core Bancario ICBS, mientras que las transformaciones y la carga se realizó de acuerdo al sistema al que se cargaron los datos.

- *Procesos de instalación*: Los procesos de instalación se refieren a los procesos a seguir para instalar el ETL dentro de los diferentes ambientes (desarrollo, pruebas y producción); este proceso también incluyó

la instalación de la herramienta de ETL en la que se desarrollaron los procesos. Parte importante del proceso dentro del ambiente de desarrollo fue la administración de los procesos de ITIL, como parte de este proceso se tenía el seguimiento de reportes de incidencias (SIR) y controles de cambio (CR), la administración de configuraciones y la administración de versiones.

- *Reporte de incidencias y controles de cambio*: Se llevó a cabo un control de incidencias y control de cambios para todos los elementos desarrollados dentro del ETL.

- *Administración de configuraciones*. La administración de configuraciones fue parte importante dentro del ambiente de desarrollo por lo que fue necesario tener un registro de todas las configuraciones que se realizaron dentro de la herramienta.

- *Administración de versiones*. Toda herramienta ETL debe tener una administración de versiones para llevar el control de todos los cambios hechos en el desarrollo.

- *Administración de pruebas*: Como todo proyecto de implementación, fue necesario llevar una administración de las pruebas; las pruebas que se llevaron a cabo dentro de este proceso fueron: pruebas unitarias, pruebas de integración, pruebas de performance y pruebas de usuarios

3.4. Diseño de los programas ETL

Como parte de la herramienta ETL se integraron tres áreas que tenían relación con los requerimientos de negocio: la capa de interacción, la capa de integración y la capa que conteía las reglas de negocio.

- Capa de interacción. Esta capa se refería a la interacción que tenían los usuarios con la herramienta ETL; así como la explotación de todos los elementos de la misma. Las principales funciones que se ejecutaron dentro de esta capa fueron la interfaz de usuario, la administración de protocolos de seguridad de usuarios y los reportes.

- Capa de integración. En esta capa hicimos referencia a todos los servicios y modelado de procesos necesarios para el buen funcionamiento de nuestra herramienta ETL. Los principales procesos que se tomaron en cuenta fueron los siguientes:

- **Servicios**. Servicios necesarios para el manejo de errores, excepciones, versiones, así como servicios adicionales necesarios para los procesos de extracción, transformación y carga. Tales servicios se dividieron en: Administración de excepciones, reinicio y recuperación de errores, registros de eventos y auditoria y administración de versiones.
- **Modelado de procesos**. Fueron principalmente los procesos necesarios para la creación de los programas ETL.

- Reglas de negocio. Dentro de este proceso definimos las reglas para la transformación y carga correcta de los datos en la base de datos, así como la funcionalidad correcta que requería la entidad financiera. La principal herramienta que se tuvo en esta capa, fue el motor ETL y sus principales tareas fueron las siguientes:

  - **Procesos de extracción.** Definición de todos los procesos utilizados para realizar la extracción de datos
  - **Reglas de transformación.** Definición de todas las transformaciones utilizadas desde la extracción de datos hasta la carga de los mismos en la nueva estructura de base de datos.
  - **Proceso de carga de datos.** DEscripción de los procesos de carga de datos a cada uno de los sistemas destino.
  - **Proceso de limpieza de datos.** Descripción de los procesos de limpieza de datos como: direcciones, teléfonos, nombres, RFCs y fechas.
  - **Procesos de calidad.** Descripción de los procesos a seguir de acuerdo a los estándares de datos para tener una mejor calidad en la información.

3.5. Capa de datos

La capa de datos estaba conformada por los diferentes sistemas desde donde se extrajo la información así como los sistemas donde se almacenarían los datos. Como sistema para extracción de datos solo se consideró un sistema, mientras que la carga se realizó hacia tres diferentes sistemas de la institución financiera. A continuación describo el tipo de información que se procesaba en cada uno de los sistemas destino:

- Se tenía el sistema de detección de fraudes o posible mal manejo de una cuenta; esta detección se realizaba mediante el monitoreo de las cuentas que se enviaban a cada uno de los archivos generados por el procesador principal de transacciones bancarias.
- El sistema de manejo de riesgos contenía la información de las cuentas con pagos vencidos o que presentaban algún grado de morosidad. Esta información también se obtenía del procesador principal de transacciones bancariasy sería procesada por la herramienta ETL que se implementó.
- Finalmente, existía el sistema para el manejo de los reportes operativos que requería la organización. La principal tarea de este sistema era generar reportes regulatorios previamente definidos, así como extracción de información bajo demanda.

Capítulo 4

Diseño detallado

En este capítulo...

4.1. Dependencias y frecuencia de las extracciones de datos

Como primer paso dentro dle proceos de extracción se identificaron las dependencias de los sistemas origen con otros procesos, tareas o reglas de negocio dentro de la organizción, así como la frecuencia esperada o deseada de la extracción de datos. La extracción de datos se realizó con una periodicidad diaria; esto debido a que al tratarse del sistema principal de la institución financiera la información se actualizaba de manera diaria. La herramienta ETL nos permitió identificar aquellas tablas y campos que sufrian modificaciones, de tal forma que se extrajo solo la información que sufría cambios a lo largo del día. La extracción de datos se realizó en un horario no operativo, a fin de no afectar la operación de la institución; además este proceso dependía del termino del proceso de cierre diario de operaciones. Solamente uan vez que el proceso hubiera terminado comenzaba la extracción por parte de la herramienta ETL; este proceso contaba con una ventana de procesamiento de 5 horas. La extracci"on fue almacenada en la base de datos temporal dentro de la herramienta ETL.

4.1.1. Extracción inicial

En la extracción inicial se realizaron las siguientes transformaciones:

- Transformación de fecha jualiana a fecha gregoriana.

- Cambio en los tipos de dato al momento de almacenarlos en la tabla temporal; principalmente transformación del tipo de dato *char* al tipo de dato *varchar* y los datos númericos a decimal.

Al momento de que se terminaba la extracción de datos, se cerraba la conexión a la base de datos del procesador principal de transacciones bancarias a fin de liberar la carga de trabajo del mismo para continuar con su operación diaria.

4.1.2. Extracciones subsecuentes

Después de realizada la extracción inicial desde el sistema fuente y con los datos almacenados en el repositorio temporal, las siguientes extracciones se realizaron mediante extracciones incrementales con base en la adición, actualización o eliminación de datos de las tablas seleccionadas. Se consideró realizar la extracción total de los registros solo en caso de que las modificaciones o actualizaciones en la base de datos fueran muchas; las condiciones para realizar la extraccióncompleta fueron las siguientes:

1. Existían cambios en más del 50 % de los registros.
2. Los registros extraídos no se encontraban actualizados o contenían errores.
3. Los registros extraídos no cuentan con la calidad necesaria.

4.1.3. Relación entre las entidades de datos

En el siguiente diagrama se puede visualizar la relación entre las diferentes entidades y los datos que comparten cada uno de ellos.

Figura 4.1: Datos relacionados en las entidades.

En la imagen podemos observar que los diferentes sistemas manejan el mismo tipo de información; socios, cuentas, créditos y saldos.

4.2. Diseño para la transformación de datos

La transformación de los datos tenía varios procesos que se deberían seguir, entre los cuales se encuentra la limpieza de datos, los estándares y procedimientos utilizados, así como las reglas de transformación definidas para cada uno de los sistemas destino. La transformación de los datos se realizó dentro de la misma base de datos de paso y se almacenó en otra instancia de esta misma base de datos. El siguiente diagrama muestra como se realizarón las transformaciones requeridas.

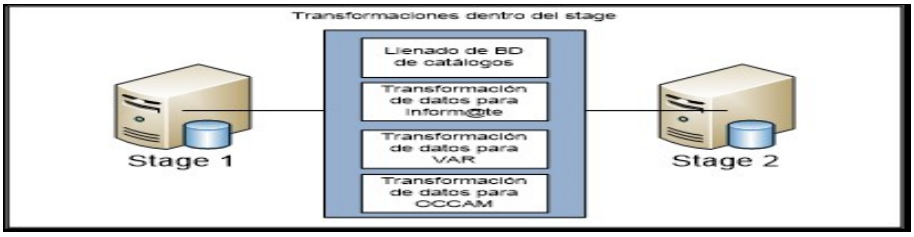

Figura 4.2: Transformaciones en base de datos temporal.

A continuación se detallan las reglas utilizadas para realizar la transfocmacón de datos.

4.2.1. Limpieza de la fuente de datos

La limpieza de la fuente de datos **NO** estaba dentro del alcance del proyecto; sin embargo se realizaron las transformaciones de la fuente de datos para realizar la limpieza de datos hacia los sistemas destino. Como parte de un proceso de calidad de la información y limpieza de datos se definierón los siguientes criterios:

1. Los nombres de los socios debían contener solamente caracteres alfabéticos; no se permitieron caracteres especiales como los siguientes: “\”, “.”, “#”, “$”, “%”, “&”.

2. Los números telefónicos debían contener solamente caracteres numéricos y debían ser de 10 posisciones para teléfonos fijos y 13 posiciones para teléfonos celulares.

    - Si *Tipo_Telefono* = Casa y Longitud (Telefono) =10 Entonces Telefono_Casa SiNo Error

- Si *Celular* = Casa y Longitud (Telefono) = 13 Entonces Telefono_Celular SiNo Error

3. El RFC de los socios debería ser de 10 o 13 posiciones para las personas físicas y 12 posiciones para las personas morales. En caso de no contar con un RFC valido, el registro era rechazado y enviado a una tabla de auditoria para su corrección dentro del sistema fuente de parte del área de tecnología de la compañia.

4. Todos los códigos postales debían ser de cinco dígitos, en caso de existir algún registro con mayor o menor número, estos debían de ser enviados a una tabla de auditoria para su corrección dentro del sistema fuente. Se consideró una homologación de datos para los códigos de ciudades y estados. Todos los códigos de estados se homologaron a tres dígitos que corresponden a las tres primeras letras de cada estado; se realizó unaexcepción para el caso de Chiapas (CHS) que podría confundirse con Chihuahua (CHI).

5. Las fechas se homologaron al formato yyyymmdd y no debían permitir valores nulos o fechas inválidas.

4.2.2. Estándares y procedimientos utilizados para la interrelación de los datos

Para realizar una estandarización de los datos se creó una base de datos temporal que contenría la información de todos los campos con los mismos datos pero que su sistema origen era diferente. La creación de esta tabla temporal fue crear la base para generar una tabla de referencias cruzadas que representara un gobierno de datos o MDM (Master Data Management) por sus siglas en Inglés

4.2.3. Reglas utilizadas para la transformación de datos

Como regla general para todos los procesos de extracción se realizaron las siguientes transformaciones.

- Transformación de fechas julianas a fechas gregorianas en formato yyyy/mm/dd.

- Transformación de los tipos de dato origen a los tipos de la base de datos stage generada en SQL Server.

Los procesos de transformación para cada destino fueron diferentes debido a las características de cada uno de los archivos de salida. Dichas transformaciones se realizaron tomando como principal fuente de datos, la base de datos temporal.

Sistema de reportes

- Todos los campos de fecha que tengan un sufijo _d deberían de ser convertidas a formato de fecha (yyyymmdd) para todas las tablas de este sistema. Se pobló la fecha de vencimiento _D, tomando como base el campo fecha de vencimiento de la tabla de facturación. La regla de transformación que se definió fue la siguiente:

```
FEC_VENCIMIENTO_5_D = CAST(RIGHT(LTRIM(RTRIM(FEC_VENCIMIENTO_5)),2) +
        LEFT(RIGHT(LTRIM(RTRIM(FEC_VENCIMIENTO_5)),4), 2) +
        REPLICATE('0', 2-len(LEFT(LTRIM(RTRIM(FEC_VENCIMIENTO_5)),
        ABS(4-len(LTRIM(RTRIM(FEC_VENCIMIENTO_5)))))) +
        LEFT(LTRIM(RTRIM(FEC_VENCIMIENTO_5)),
        ABS(4-len(LTRIM(RTRIM(FEC_VENCIMIENTO_5))))) AS DATETIME)
        WHERE  FEC_VENCIMIENTO_5 > 0
```

- Para el campo Fecha de último saldo se tomó la fecha de vencimiento más antigua que se tenía.
- Para obtener el número de créditos de un socio se realizó la siguiente regla.

```
NO_CREDITOS =  NO_CREDITOS
WHERE  NO_SOCIO IN NO_SOCIO de la tabla CIF_CREDITOS.
Si NO_CREDITOS = NULL Entonces NO_CREDITOS = 0
```

- La descripción de la información correspondiente a las garantías debía de estar vacio en la tabla destino.
- Para almacenar la fecha de último monto con saldo se tomó la fecha de vencimiento más antigua que se tenía para cada registro dentro de la tabla de saldos *LN_SALDOS.*
- Se realizó una validación sobre el tipo de cuenta para confirmar su valor = 1 (Personas físicas) o 6; en estos casos se colocó un valor 20 dentro del campo Tipo_producto de la tabla TA_SALDOS.

- El número de docio para poblar la tabla TA_SALDOS se obtuvo de la tabla de cuentas por socio, pero solamente de aquellas cuentas cuyo tipo de producto era 4 u 8 y que tuvieran un saldo en la tabla TA_SALDOS.
- Se calculó el número de créditos de cada socio y este valor se grabó en la tabla CIF_INFGEN; en caso de que el número de créditos fuera nulo se colocó un valor por default 0.
- Se insertó el número de socio dentro de la tabla TA_CTAXSOCIO para aquellos números de cuenta menores a 20000000000 y cuya clave de relación fuera 'SOW' u 'OWN'. El número de socio eran los 10 dígitos de la derecha de cada número de cuenta.
- Se definió que para el campo descripción perteneciente a la tabla LN_GARANTIAS debería poblarse con un valor NULL a pesar de tener mapeado un campo del cuál se extraía la información.
- El campo Periodo de la tabla LN_SALDOS se asignó con un valor fijo = 'M'.
- El campo frecuencia de la tabla LN_SALDOS se asignó con un valor fijo = 30.

La regla de transformación que se utilizó durante la extracción.

Reglas para el sistema para prevención de lavado de dinero

Las reglas definidas para este sistema son las siguientes:

- Se definieron algunos valores estáticos para los tres primeros campos de cada uno de los archivos generados; estos valores son los siguientes:
  1. Valor fijo 'CMPS' para el primer campo.
  2. Valor fijo 'T' para el segundo campo.
  3. Valor fijo 'F' para el tercer campo.
- Para la generación del archivo de cuentas se tomaron solo aquellas cuentas que tuvieron movimientos durante el día.
- Solamente se tomaron en cuenta los siguientes tipos de cuenta: Parte social, cuenta mexicana, servicuenta y cuentamiga.

- Si el tipo CIF es una persona moral (CUTYP = 5), entonces el campo CompanyLegalId se pobló con dicho valor, en caso contrario se asignó un valor constante vacio “”.

4.3. Arquitectura aplicativa

4.4. Arquitectura de negocio

4.5. Arquitectura de datos

Como se explicó en los capitulos anteriores, la institución financiera contaba con diferentes procesos ETL que extraían información del sistema principal. Estos sistemas trabajaban de manera independiente por lo que cada sistema destino tenía su base de datos temporal antes de enviarla a su destino. La arquitectura se encuentra detallada en el siguiente diagrama:

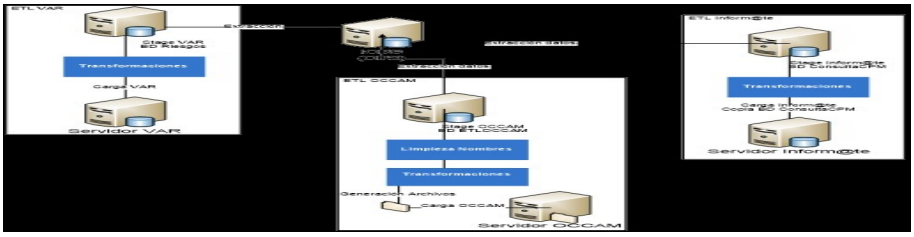

Figura 4.3: Arquitectura de datos actual.

Como se observa en la imagen anterior, existían muchas bases de datos intermedias entre la fuente de datos y los sistemas destino; la arquitectura definida dentro de este proyecto, nos permite tener una sola base de datos temporal, así como una sola capa de integración de datos. El siguiente diagrama muestra el diseño de la arquitectura desarrollada:

Arquitectura_final

4.6. Arquitectura de infraestructura

“Diseño de arquitectura aplicativa”, “Arquitectura de negocio”, “Arquitectura de datos” y “Arquitectura de infraestructura”

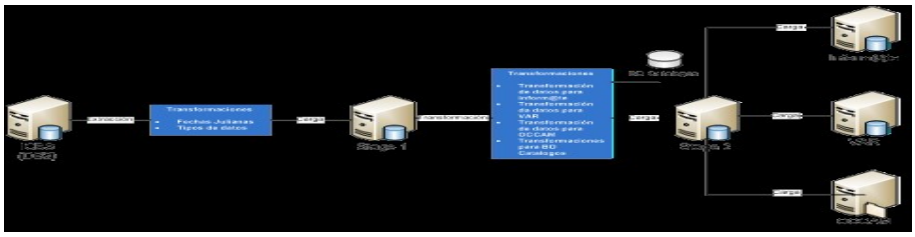

Figura 4.4: Arquitectura de datos final.

Capítulo 5

Modelo de datos

5.1. Modelo de datos lógico

5.2. Modelo de datos físico

Capítulo 6

Estándares de programación

dalsdkla

Capítulo 7

Conclusiones

Las bases en desarrollo de software, bases de datos, estructuras de datos, lenguajes de programación y, por encima de todo, análisis y solución de problemas que obtuve dentro de la carrera de Ciencias de la Computación, me ayudarón a que el proyecto que presento en este trabajo se realizara de manera éxitosa.