



Universidad Cenfotec

Programa de Maestría en Tecnologías de Bases de Datos



Memoria de Proyecto de Graduación:

“Diseño y especificación de la

Fábrica de Información Corporativa del Banco de Costa Rica”

Para optar por el grado de Máster en Tecnologías de Bases de Datos.

Lic. Edius Maicol Vargas Paniagua

Prof. Tutor: Carlos Gonzalez Alvarado, PhD.

I Cuatrimestre, 2013

"Somos lo que hacemos constantemente. La excelencia,
por tanto, no es un acto sino una costumbre".

Aristóteles

TABLA DE CONTENIDOS

TABLA DE CONTENIDOS	3
RESUMEN	6
INTRODUCCIÓN	8
OBJETIVO GENERAL Y OBJETIVOS ESPECÍFICOS DEL PROYECTO	10
Objetivo general	10
Objetivos específicos.....	10
ALCANCES Y DEFINICIÓN DEL PROBLEMA POR RESOLVER	11
Enunciado del problema.	11
Alcance del problema.....	12
PROCESO DE DESARROLLO	14
DESARROLLO DEL PROYECTO.....	17
1. Revisión bibliográfica.	18
1.1. Consideraciones iniciales	18
1.2. Caracterización de opciones	19
1.3. Selección de una arquitectura para el BCR.....	38
1.4. La Fábrica de Información Corporativa del BCR, especificación de componentes.....	39
1.5. Anotaciones sobre la bibliografía utilizada	42
2. Reflexión sobre situación previa y antecedentes.	43
2.1. Situación previa del BCR	43
3. Definición de instrumentos para el análisis.....	46
3.1. Instrumento para el mapeo de datos.	47
3.2. Instrumento para el análisis de calidad de datos.	52
4. Análisis y diseño de componentes de la CIF.	61
Data Warehouse Corporativo	62
Área de trabajo (<i>Staging Area</i>)	62
Repositorio de Metadatos	82
5. Definición de metodología de desarrollo.	87
5.1 Propuesta metodológica para el desarrollo de proyectos de Inteligencia de Negocios en el BCR.	87

6. Modelo para la generación de interfaces de consulta	103
JUSTIFICACIÓN DE DECISIONES DE DISEÑO	108
ANÁLISIS Y DISCUSIÓN DE RESULTADOS.....	114
CONCLUSIONES Y TRABAJOS FUTUROS	117
PLAN DE INVERSIÓN DEL PROYECTO	120
BIBLIOGRAFÍA.....	127
DEDICATORIA	128

RESUMEN

En toda empresa la necesidad de información confiable y oportuna es crucial para poder tomar decisiones que permitan la mejora continua y el alcance de los objetivos estratégicos. A mayor el tamaño de la empresa mayor importancia tiene el disponer de información de este tipo, lamentablemente el tamaño de la empresa también hace que esta información sea más difícil de obtener y mantener.

El Banco de Costa Rica (BCR), como una institución que brinda servicios bancarios a nivel nacional e internacional, busca la mejora continua basándose en las decisiones más acertadas posibles, y en la mejora continua en sus productos y servicios, buscando cada vez más, el cumplimiento oportuno tanto de sus objetivos estratégicos como de las expectativas de los usuarios de sus servicios. Sin embargo este proceso históricamente se ha valido de información que aunque es ampliamente validada y revisada, no es obtenida de la mejor manera, y presenta algunas deficiencias tales como la falta de mecanismos para acumular información de forma histórica y la falta de trazabilidad de la información que soporta la toma de decisiones.

Es así, como a partir de esta necesidad se comenzó a forjar en el BCR la idea de generar un repositorio de información que pudiera suplir la necesidad de integrador de información y de proveer al negocio de una fuente fidedigna de datos para apoyar procesos de toma de decisiones. Aunque se habían hecho algunos esfuerzos para lograr esta meta en el pasado, algunas circunstancias particulares minaron la posibilidad de que estos resultaran fructíferos, con las consecuencias negativas para el negocio que acarrea el seguir sin disponer de este tipo de herramientas.

Con la elección del tema de este proyecto, y la consiguiente ejecución, se buscaron los motivos de fallo de los procesos anteriores y se diseñó una arquitectura inicial, basada en las recomendaciones de diseño dadas por algunos de los más importantes autores en Inteligencia de Negocios y *Data Warehouses*, con el fin de que el esfuerzo por construir este repositorio fuera esta vez bien aprovechado y diera los frutos que el negocio tanto esperó obtener en intentos anteriores.

A través del proyecto se han diseñado y utilizado diversos elementos metodológicos, construcciones de modelado de datos, elementos lógicos, infraestructura, con el consecuente resultado final de que con un mínimo de recursos se logró llevar a producción un repositorio de información integrada con algunos de los productos más significativos para el banco, y se ha realizado una tarea de evangelización a través de la organización que ya ha comenzado a permear a los niveles más altos con el fin de impulsar el crecimiento del Data Warehouse Corporativo y por consiguiente de la Fábrica de Información Corporativa del BCR.

Los elementos desarrollados dentro del contexto del proyecto, incluyen además estándares e instrumentos de apoyo para los analistas y guías para que estos puedan ejecutar, en el marco

de una metodología de desarrollo para proyectos de inteligencia de negocios, un proceso completo que permita al negocio obtener de la información integrada en el *Data Warehouse* para los propósitos de toma de decisiones que el propio negocio plantee.

El mayor aporte del proyecto es el haber permitido trazar la línea de acción inicial sobre la que se deberá trabajar para continuar el desarrollo de la arquitectura, y ofrecer al BCR de un conjunto de instrumentos básicos que le permitirán madurar paulatinamente en el proceso de desarrollo de aplicaciones de inteligencia de negocios.

Entre los demás aportes del proyecto se pueden enumerar, la organización de los modelos de datos que deberán administrarse, la definición de los elementos de alto y bajo nivel que permitirán la administración de estructuras históricas, la definición de un estándar para darle seguimiento y trazabilidad a las cargas de datos, la definición de los conceptos básicos y método para la evaluación de la calidad de datos, el aprovechamiento de las herramientas y recursos disponibles en el banco, el establecimiento de los procesos de análisis de información y de la metodología base de desarrollo y un plan de reducción de costos que a un final logrará reducir significativamente la inversión anual a la vez que logrará mejorar el desempeño en el desarrollo y en la consulta de información desde la Fábrica de Información Corporativa.

Como consecuencia directa del trabajo realizado, y dados los resultados obtenidos que aunque por el momento han representado un pequeño éxito han logrado llamar la atención de la alta gerencia y han provocado que ha este momento se haya comenzado a trabajar en conjunto con el negocio en necesidades que provienen desde la línea gerencial, relacionadas a temas como Productividad, Riesgo, y Profundidad de Línea y se visualizan a corto plazo la necesidad de incluir algunas otras como Mercadeo y Segmentación de clientes.

INTRODUCCIÓN

En la presente era de la información, el conocimiento no solo es poder, es el motor que mueve los negocios y que permite la subsistencia de las grandes empresas. La información se vuelve ahora más que nunca el activo más importante de una organización.

Uno de los mayores retos que implica esta dependencia de la información es que esta no es un insumo directo de los negocios donde nuestras organizaciones se desenvuelven, esta característica es específica a los datos, los cuales definidos como una característica o pieza de conocimiento sobre un determinado hecho o entidad, no tienen el mismo valor que tiene la información.

Los datos son la materia prima, y son ciertamente abundantes en cualquier organización que tenga alguna trayectoria, pero los datos han probado no ser suficientes para nada más que para el día a día, para apoyar los procesos productivos más básicos de la organización; más para apoyar otro tipo de procesos, de mayor nivel, como la toma de decisiones, la creación de proyecciones a futuro, la visualización de nuevas estrategias de negocio, es necesario contar con información.

Y es que la información se obtiene a partir de dar un valor a los datos más allá de lo que son como una entidad separada, la información se obtiene a partir de llevar los datos a través de un proceso razonado y bien estructurado para darles contexto, dentro de la organización y sus procesos.

Cuando por fin obtenemos esta suma entre datos y contexto obtenemos información, más para obtener la mejor calidad, debemos de preocuparnos por las características que los datos heredan a esta información: ¿está completa?, ¿es confiable?, ¿es la verdadera y única versión de la verdad?, ¿es información actual?

Cuando logramos contestar positivamente a estas preguntas es cuando sabemos que estamos listos para poder utilizar la información como un insumo para la generación de conocimiento y para apoyar los procesos de toma de decisiones.

Más aún, en el mar de información que podamos formar dentro de nuestras empresas a través de numerosos y complejos procesos de integración y calidad de datos, falta un elemento clave, que nos permita visualizar la información de manera que pueda ser explotada para los fines que la requerimos, así hay que agregar que aún cuando podamos tener información útil y confiable, es necesario que los medios de visualización y consulta de esta información sean también eficientes.

Finalmente es necesario que todo el proceso que hemos llevado a cabo sea repetible, controlado, autosuficiente, bien estructurado, pensado para ser sostenible en el tiempo y que sea continuamente revisado y mejorado para garantizar su permanencia a través del tiempo.

Es todo esto el fin último de la Fábrica de Información Corporativa (o CIF por sus siglas en inglés), tal como su nombre lo indica, ser el medio a través del cual la organización puede crear información, centralizada, útil según sus necesidades, confiable, y ajustada a su propia realidad, y que pueda ser accesible según se necesite para apoyar toda clase de procesos operativos, de análisis y de toma de decisiones.

Esto es lo que queremos lograr, queremos hacer que el Banco de Costa Rica, cuente con todo lo que necesita para constituir la CIF como un proceso más dentro de su operativa tal que le permita reforzar sus procesos de toma de decisiones y consolidar aún más su posición dentro del mercado nacional.

OBJETIVO GENERAL Y OBJETIVOS ESPECÍFICOS DEL PROYECTO

Objetivo general

Definir y especificar un modelo para la Fábrica de Información Corporativa (CIF) del BCR, con el fin de favorecer la implementación de nuevas y mejores aplicaciones de Inteligencia de Negocios en la institución.

Objetivos específicos

1. Definir los componentes que conformarán la Fábrica de Información Corporativa del BCR y sus características.
2. Definir las funciones que cumplirán las diferentes herramientas y repositorios de Inteligencia de Negocios ya adquiridas y desarrolladas por el BCR dentro de la arquitectura propuesta.
3. Proponer una metodología de desarrollo estandarizada en el contexto de la Fábrica de Información Corporativa.
4. Desarrollar y documentar estándares para los procesos de integración de datos que deban ser incluidos dentro de la Fábrica de Información Corporativa del BCR.
5. Definir un método y un conjunto de instrumentos para la evaluación de calidad de datos en el contexto de la Fábrica de Información Corporativa.
6. Establecer un estándar para la creación de interfaces de consulta de datos desde la Fábrica de Información Corporativa del BCR.

ALCANCES Y DEFINICIÓN DEL PROBLEMA POR RESOLVER

Enunciado del problema.

El Banco de Costa Rica (BCR), es una institución financiera estatal, fundada el 22 de abril de 1877 bajo el nombre de Banco de la Unión, y posteriormente nacionalizado en 1948 por decreto de la Junta Fundadora de la Segunda República. Actualmente la institución cuenta con una extensa presencia a través de todo el país, contabilizando cientos de oficinas distribuidas a través de todos los cantones.

Su extensa presencia le ha permitido también extender ampliamente sus servicios, pudiendo así brindar productos que van desde cuentas corrientes y de ahorros, hasta una sólida operadora de pensiones, administradora de fondos de inversión, corredora de seguros y un puesto de bolsa que en conjunto reciben una amplia aceptación de parte de los costarricenses.

Es por motivo de este extenso crecimiento, tanto geográfico como en diversificación de sus productos y servicios y a la importante competencia que representan los otros bancos estatales y la banca privada en Costa Rica, que los procesos de toma de decisiones, y de generación de información para servir de insumo a los mismos, toman un papel de especial importancia en las operaciones de esta entidad bancaria.

La problemática actual, que es lo que nos lleva al desarrollo del tema de esta proyecto, es que estos procesos de generación de insumos para la toma de decisiones, en muchos casos han surgido como silos aislados, esfuerzos independientes o en algunos casos poco integrados dentro de cada departamento, la mayoría de veces comandados por personas del negocio con diferentes niveles de lo que podríamos llamar “alfabetización tecnológica”, y utilizando diferentes medios con diferentes niveles de automatización y formalidad.

Existen un conjunto variado de herramientas y procedimientos de procesamiento de información que se mantienen lejos de la División de Tecnología e inclusive del propio negocio, por ejemplo, un área de negocio usa sus propias reglas para procesar un determinado conjunto de información y otra área hace lo mismo con otro conjunto de reglas, siendo estos conjuntos de información que se intersecan entre sí, se produce duplicidad de procesos, esfuerzo y variados puntos de falla lo que da como resultado contradicción entre los resultados obtenidos entre una u otra área.

A nivel de la División de Tecnología, se pueden encontrar a lo interno esfuerzos que se realizan todos los días para integración datos y mejorar de alguna forma la calidad de los mismos, pero estos esfuerzos generalmente se encuentran asociados a necesidades inmediatas de carácter regulatorio, a compromisos públicos tales como promociones o solicitudes de alto nivel externas o internas a la institución, o necesidades espontáneas de visualización de datos poco planificadas de los gerentes medios y altos.

Es así como desde hace algunos años se han comenzado a formar pequeños equipos de trabajo en torno a la necesidad de generar un repositorio de información centralizado que ayude a dar respuesta a las necesidades de información para la toma de decisiones que demanda el entorno cada vez más creciente del banco y sus gerentes, lo anterior lamentablemente, con poco éxito.

Con el precedente anteriormente desarrollado la propuesta de esta investigación consiste en el estudio a fondo de la situación actual del banco, y la formulación de una guía que permita el desarrollo de un conjunto de procesos para la integración de datos en un conjunto de repositorios que cumpla con todas las características necesarias para dar a los usuarios de negocio los insumos necesarios para tomar decisiones; a la vez que se ayude a implementar una arquitectura y modo de trabajo que permita dar sostenibilidad a los esfuerzos de desarrollo y la continuidad de un equipo dedicado a Inteligencia de Negocios dentro del Banco de Costa Rica.

Alcance del problema

El problema a resolver se encuentra supeditado a la utilización de la teoría y buenas prácticas para la creación de la Fábrica de Información Corporativa, *Data Warehouse*, procesos de integración y de calidad de datos que puedan ser respaldados por la investigación bibliográfica y por la experiencia empírica comprobable.

Por ser este un proyecto de investigación aplicado para optar por un grado de maestría en el área de bases de datos, los límites disciplinarios guardaran una frontera algo difusa a esta disciplina, permitiendo enriquecer, claro está, la misma con aportes hechos por compañeros de otras áreas cuyos criterios interdisciplinarios se constituyan en un aporte importante desde el punto de vista de la investigación.

La investigación se limitará exclusivamente al Banco de Costa Rica (en adelante BCR) y sus subsidiarias (BCR Pensiones, BCR Sociedad Administradora de Fondos de Inversión, BCR Corredora de Seguros y BCR Puesto de Bolsa). El desarrollo de cualquier entregable desarrollado durante la investigación deberá tomara en cuenta principalmente los sistemas del BCR, sus negocios, sus estándares, procesos, normativa y entorno organizacional.

El lugar de ejecución de la investigación y recolección de insumos se limitará a oficinas centrales y la oficina de Aranjuez que es donde se encuentran la mayoría de los usuarios interesados y a la vez estas son las dos oficinas en las que se distribuye la División de Tecnología.

El alcance temático de la investigación se limita a la definición de los procedimientos que permitan al Banco de Costa Rica crear una plataforma de Inteligencia de Negocios, basada en la construcción de la Fábrica de Información Corporativa, en su propio contexto de negocios y según sus posibilidades actuales, posibilitando el crecimiento y mejora de estos procesos a

futuro mediante la consolidación de una propuesta que sea integral y extensible según dicten las necesidades del negocio y las posibilidades lo permitan.

PROCESO DE DESARROLLO

Es de esperarse que antes de comenzar en la construcción de la CIF primero definamos un método ordenado para poder garantizar que su construcción obedecerá a los fines que requiere el banco, es así como para la ejecución del proyecto se realizaron algunas sesiones con el grupo de trabajo que estaba conformado en ese momento y se consideró necesario el generar un proceso de desarrollo del proyecto en el que se diera el espacio requerido para entender:

- ¿por qué se consideró necesario el proyecto?
- ¿qué es lo que se desea obtener?
- ¿qué insumos necesitaremos para poder lograr lo que nos proponemos?
- ¿en qué entregables enfocaremos primero nuestros esfuerzos y cómo serán estos entregables?
- Dadas las características de este tipo de proyectos, ¿de qué forma deberemos proceder para generar nuevos proyectos en el futuro?
- ¿Cómo podemos aprovechar la plataforma tecnológica actual del BCR dentro de este nuevo sistema?

Con estas consultas en mente se definió proceder mediante las siguientes etapas para la realización del proyecto y la consiguiente construcción de las primeras etapas de la CIF del BCR:

1. Revisión bibliográfica.
2. Reflexión sobre situación actual y antecedentes.
3. Definición de instrumentos para el análisis.
4. Análisis y diseño de primeros componentes de la CIF.
5. Definición de metodología de desarrollo.
6. Adaptación al entorno tecnológico del BCR.
7. Ejecución de primeras etapas de desarrollo.

Las secciones que siguen a continuación brindan una explicación más detallada de cada paso con el fin de conocer qué alcance tuvo cada etapa dentro del proyecto realizado.

1. Revisión bibliográfica.

Debido a la falta de un proceso formal implementado en el BCR para el desarrollo de aplicaciones de Inteligencia de Negocios, el proceso de desarrollo a seguir debió comenzar con una revisión bibliográfica que nos diera la posibilidad de considerar todos los diferentes aspectos involucrados en la creación de la Fábrica de Información Corporativa. La idea de este estudio fue el de darnos un panorama completo acerca de los componentes que deberían desarrollarse y de qué forma debíamos proceder para la generación de información integrada dentro de la misma.

Esta etapa se hizo antes de la evaluación del propio entorno debido a que se estimó que la misma permitiría al equipo ser más crítico con respecto a las necesidades y a lo que se debe observar permitiendo enfocar los esfuerzos en aquello que va directamente relacionado.

Así el estudio bibliográfico nos da la perspectiva completa y nos permite enfocar los esfuerzos subsiguientes en la especificación detallada de cada componente.

2. Reflexión sobre situación actual y antecedentes.

Se consideró de gran importancia poder establecer los antecedentes que dieron origen a la necesidad actual y las razones por las que los esfuerzos similares resultaron infructuosos en el pasado, de manera que pudimos reconocer nuestras debilidades y buscar la manera de superarlas en esta oportunidad.

Esta fase permitió generar en los miembros de la organización involucrados una cultura más crítica ante el manejo de la información para la toma de decisiones y ayudó a hacer entender la importancia de implementación del proyecto y de destinar un primer conjunto de recursos a la generación de las primeras etapas.

3. Definición de instrumentos para análisis.

Una parte importante de los objetivos planteados inicialmente para el proyecto fue el de dotar a la organización de un conjunto de instrumentos que permitieran facilitar la ejecución de este tipo de desarrollos, aún después de la ejecución de la primera fase; el resultado de esta etapa incluye instrumentos tales como algunas plantillas básicas para análisis de integración y calidad de datos, las cuales permiten guiar a los analistas para la generación de un conjunto de información adecuado para responder tanto a las necesidades de los desarrolladores como a la recolección primaria de metadatos.

4. Análisis y diseño de primeros componentes de la CIF.

A partir del estudio bibliográfico se pudo determinar cuáles eran las necesidades de análisis y diseño que teníamos para realizar las primeras versiones de los elementos de la Fábrica de Información Corporativa y pudimos priorizar el orden de ejecución de los elementos que la conformaban, así como el nivel de profundización de cada uno. Esto bajo el entendido que los recursos disponibles para la ejecución eran escasos y que se requería una base sólida sobre la que se pudieran integrar nuevos colaboradores, los cuales puedan continuar dándole forma a la CIF.

5. Definición de metodología de desarrollo.

Igualmente, a partir de la información obtenida de las fuentes bibliográficas revisadas y del entorno organizativo, se pudo definir para el BCR una primera versión de una metodología de desarrollo específica para Inteligencia de Negocios, la cual, basada en las características particulares de esta clase de proyectos, permitirá continuar desarrollando la CIF conforme se

vayan integrando recursos y nuevos temas; la metodología queda abierta para ser complementada con nuevos pasos o instrumentos conforme la madurez de los involucrados y de la organización lo permitan y requieran.

6. Adaptación al entorno tecnológico del BCR.

Una vez que se hubo definido la estructura interna necesaria de cada componente y la forma en la que estos interactuarían entre sí, correspondió proceder a la adaptación de la arquitectura definida a las herramientas y entorno tecnológico disponible dentro del entorno del BCR, este proceso nos permitió el realizar un primer estándar de desarrollo que facilitó la aplicación de las buenas prácticas y recomendaciones, obtenidas de la revisión bibliográfica y el consecuente análisis y diseño interno de los componentes de la CIF.

7. Ejecución de primeras etapas de desarrollo.

Finalmente se procedió a utilizar los insumos y el procedimiento definido a través del proceso de ejecución del proyecto para comenzar con los primeros procesos de integración de datos hacia el *Data Warehouse* Corporativo, el cual se ha posicionado con éxito como primer repositorio de la CIF del BCR. Este ejercicio nos permitió hacer los ajustes necesarios a la metodología y diseños y comprobar que estos son efectivos, demostrando así la validez del trabajo realizado y garantizando su aplicación futura dentro de la organización.

DESARROLLO DEL PROYECTO

Tal como fue definido en la sección anterior en esta se desarrolla el contenido correspondiente a cada paso definido en la metodología, con el fin de lograr el cumplimiento de los objetivos planeados para el proyecto. En cada paso se hará énfasis al proceso realizado y se enriquecerá el contenido aportado con experiencias obtenidas durante la ejecución de las mismas en el ambiente del banco.

1. Revisión bibliográfica.

1.1. Consideraciones iniciales

Esta sección comprende los conocimientos usados como base para la conformación del proyecto, mediante esta se permite vislumbrar el estado ideal hacia el que se desea hacer evolucionar al banco y es a partir de lo detallado en esta etapa que comenzamos a dar forma a los elementos que conformarían posteriormente la Fábrica de Información Corporativa (CIF) del Banco de Costa Rica.

La Fábrica de Información Corporativa y su implementación en el entorno del Banco de Costa Rica, es el concepto primario sobre el que se basa la investigación, a partir de él es que emanan otros temas como el *Data Warehouse*, los *Data Marts*, la estructuración de los procesos de carga de datos y los procedimientos para la medición y del mejoramiento de la calidad de datos.

La Fábrica de Información Corporativa es un concepto presentado por William Inmon en 1997 el cual engloba el conjunto de componentes que forma la plataforma para la implementación de procesos de *Data Warehousing* en toda la organización.

El concepto original de la CIF [INMON, 1997] reconoce la existencia de dos tipos diferentes de datos que deben convivir dentro de su arquitectura, los datos operacionales y los datos de soporte a la toma de decisiones, así la teoría de Inmon nos propone la creación de un conjunto de sistemas y de repositorios con propósitos específicos que en conjunto conforman la llamada Fábrica de Información Corporativa.

Los datos operacionales por lo general cumplen con un conjunto específico de características:

1. Son detallados: Podemos decir que equivalen a los datos transaccionales, con ciertas diferencias que son necesarias para incorporarlos dentro de la CIF.
2. Su valor para la toma de decisiones se mantiene de un corto a mediano plazo.
3. Su voluminosidad los hace difíciles de mantener en el tiempo dentro del mismo repositorio único.
4. Son más susceptibles a cambios en los sistemas transaccionales.

Por otra parte los datos de soporte a la toma de decisiones son caracterizados por cosas tales como:

1. Son generalmente resumidos, porque se constituyen generalmente a partir de agregaciones de datos operacionales.
2. Su estructuración está más ligada a la lógica tras el negocio y no a la estructura de la transaccionalidad, por lo tanto son menos susceptibles a cambios en los sistemas origen.
3. Son relativamente poco voluminosos.
4. Su valor histórico y para la toma de decisiones se mantiene de mediano a largo plazo.

Debido a las diferencias antagónicas entre ambos tipos de datos es que la propuesta de Inmon los ubica en diferentes áreas dentro de la CIF, así los datos operacionales son estructurados dentro del repositorio que se conoce como ODS (por su siglas en inglés *Operational Data Store* o bien Almacén de Datos Operativos), y los datos de soporte a la toma de decisiones por lo general encuentran su lugar dentro del *Data Warehouse*.

Más recientemente, Inmon en su propuesta *Data Warehouse 2.0* [INMON, 2008], nos presenta nuevos elementos dentro de la CIF, reconociendo el papel de los Metadatos, introduciendo conceptos de otras disciplinas relacionadas, como el ILM (*Information Lifecycle Management* o Gestión del Ciclo de Vida de la Información) y agregando nuevas secciones para el soporte a datos no estructurados. Estos elementos igualmente deberán encontrar su lugar dentro de la plataforma de la Fábrica de Información Corporativa del BCR, pero no como un “mega proyecto” sino como un conjunto de relativamente pequeños esfuerzos que permitan tanto al negocio como a la División de Tecnología madurar en el tema y encontrar su propio camino, forjado por su propio contexto interno y externo hacia una toma de decisiones basada en información real producida por procesos de nivel institucional.

1.2. Caracterización de opciones

Todo sistema de información tiene asociada una estructura lógica y una física, estas definen la forma en que interactúan los componentes del sistema para responder a sus fines concretos. Existen principalmente dos corrientes de pensamiento sobre cómo debe ser diseñada la arquitectura lógica de un *Data Warehouse*, esta sección brinda un vistazo sobre las propuestas de arquitectura y metodología más ampliamente difundidas y utilizadas en el mundo, de manera que se puedan establecer los criterios suficientes que promovieron la elección posterior de una de ellas.

La definición de una arquitectura y la metodología de construcción del *Data Warehouse* dependerá directamente de cuál de las dos corrientes se escoja, así cada tipo de *Data Warehouse* difiere en el público meta al que se dirige y reconoce diferentes tipos de usuarios y una cantidad de esfuerzo diferente, para poder hacer una elección entre las distintas posibilidades de arquitectura lógica tuvimos que conocerlas a fondo, reconocer sus características, sus diferencias, sus enfoques, sus ventajas y desventajas y posteriormente elegir entre ellas basándonos en la realidad del BCR y en lo que estimamos que necesitaríamos en el futuro.

Las dos corrientes de pensamiento que se evaluarán se enumeran a continuación, posteriormente pasaremos a hacer un repaso de su enfoque y características, y finalmente las compararemos a través de las mismas para definir cuál se ajusta de la mejor manera a nuestras necesidades.

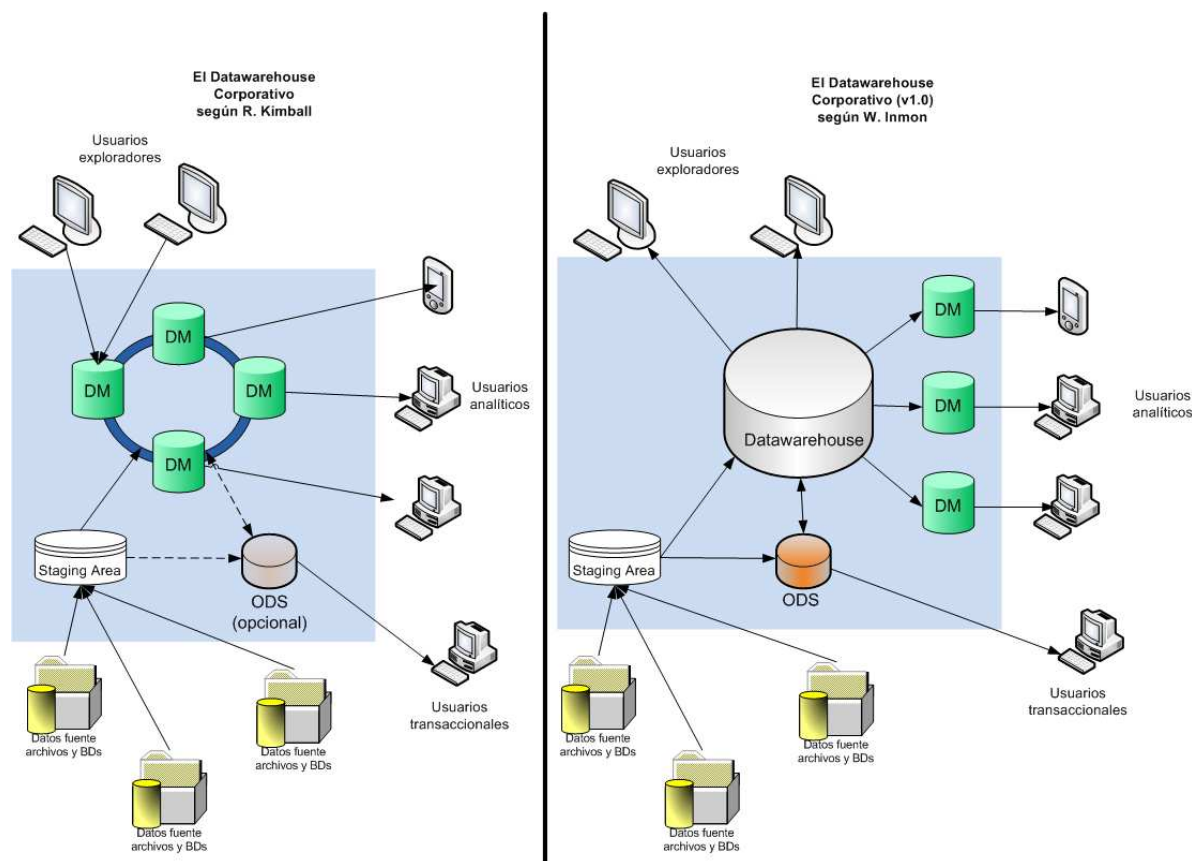


Ilustración 1. Opciones de arquitecturas lógicas para Data Warehouses Corporativos

Data Warehouse según Ralph. Kimball: En este esquema, el *data warehouse* no es más que la suma de los *data marts*, está orientado enteramente a la creación, explotación y mantenimiento de la interrelación de estructuras de datos basados en los modelos estrella y/o modelos multidimensionales.

Data Warehouse de William Inmon: Esta otra opción, reconoce al *data warehouse* (o por abreviar DW) como un repositorio centralizado de propósito general, semi-normalizado y dividido en temas, del cual irradian los *data marts*, más específicos, que pueden ser construidos de forma relacional o dimensional (usando modelos de estrella o multidimensionales tal como en la arquitectura de Kimball). Existe además una segunda versión de la arquitectura de Inmon que se conoce como DW 2.0 y que es, en sí, un cambio evolutivo de su arquitectura original.

Si nos limitamos a analizar los esquemas de cada arquitectura en sus versiones originales, tal como aparecen en el diagrama de la ilustración 1, salta a la vista solamente una diferencia, la existencia de la pieza central conocida como Data Warehouse en el modelo de Inmon, es el equivalente a la amalgama que mantiene las relaciones entre los datamarts (como veremos más adelante, las llamadas dimensiones conformadas) en el modelo de Kimball, una diferencia menos obvia es la relación entre el ODS (Depósito de datos operativos u *Operational Data Store* por sus siglas en inglés) y el *data warehouse*, en el modelo de Kimball el ODS es un atributo deseable, pero en el de Inmon se ve como una parte completamente necesaria dentro de la arquitectura.

Es importante además hacer notar la existencia de diferentes tipos de usuarios llamados de manera genérica en la ilustración como usuarios exploradores, analíticos y transaccionales, y el reconocer la existencia de diferentes usuarios del *data warehouse* y el hecho de que estos tienen a su vez una clase diferente de necesidades de información, así un explorador realmente no busca un dato específico sino que estudia la información almacenada para encontrar oportunidades nuevas que no hayan sido visualizadas, por otra parte los usuarios analíticos trabajan sobre información específicamente diseñada para alimentar indicadores de alto nivel, y los transaccionales buscan soportar procesos operativos a partir de información consolidada.

En forma general ambos esquemas utilizan un *staging area*, y tienen *data marts* y ODS (opcional o no), así que existe un acuerdo entre ambos en el papel de estos elementos dentro de la estructura de la arquitectura del *Data Warehouse* Corporativo, el significado de cada uno de estos elementos se revisará más adelante por el momento solamente se quiere dejar en claro su existencia en ambos modelos.

Con el aumento de la aceptación de las tecnologías de Inteligencia de Negocios y el *data warehousing* como tecnologías probadas y efectivas cuando se aplican adecuadamente, el modelo de Inmon además ha sufrido una evolución natural hacia un modelo aún más complejo, este modelo reconoce la existencia de repositorios de metadatos (información sobre los datos que se cargan al *data warehouse*), de un ciclo de vida (validez o actualidad) de los datos dentro del DW (*Data Warehouse*) y de la necesidad de un sistema de administración de datos maestros conocida como MDM (*Master Data Management*). Esta variante del modelo de Inmon conocida como *Data Warehouse 2.0* corresponde a la afinación del modelo antes propuesto y no debe ser visto como una nueva propuesta, sino como la mejora de su anterior versión en función de las necesidades reales de las empresas que durante los últimos 20 años han venido eligiendo el modelo de Inmon como base para la construcción de sus *data warehouses*.

Pasaremos ahora a detallar cada uno de manera que podamos reconocer sus características, ventajas y desventajas particulares y las implicaciones asociadas a la elección de cada uno.

Arquitectura de Bus de Data Warehouse (Propuesta de Ralph Kimball)

Para dar un resumen del enfoque de Kimball basada en [KIMBALL, 2008] y en pocas palabras podríamos decir que según esta visión:

“El Data Warehouse es una copia de los datos transaccionales estructurados específicamente para la consulta y el análisis; el Data Warehouse no es más que la suma de los Datamarts que lo componen, relacionados entre sí por medio de la utilización de las llamadas dimensiones conformadas.”

Hay que recordar que según el mismo Kimball lo define en [KIMBALL, 2008], un *data mart* es una sección del DW que está orientada para la explotación de la información de un tópico específico, por ejemplo: el *data mart* de ventas, el *data mart* de cuentas corrientes y ahorros, el *data mart* de finanzas, etc. Bajo esta premisa, Kimball sostiene que un DW puede obtenerse a partir de la creación de diferentes *data marts* relacionados únicamente por las llamadas dimensiones conformadas.

Las dimensiones conformadas consisten en tablas que residen en diferentes *data marts* pero que constituyen un subconjunto de estructura y contenido del mismo conjunto de datos. Por ejemplo, el catálogo de los productos utilizados en una empresa de manufactura, es necesario tanto en el *data mart* de producción como en el de ventas y en el de finanzas, y aunque pueden variar en niveles de detalle siguen siendo los mismos datos, para lograr esto el diseñador del *data warehouse* prepara un conjunto de datos con la unión de todas las necesidades de los *data marts* en el *staging area* y luego distribuye copias de partes de este conjunto a los mismos, garantizando así unidad entre todos los *data marts*.

La visión de Kimball de los *data marts* y las dimensiones conformadas y en si del proceso de creación de un *data mart* se agrupan en lo que se conoce como la Arquitectura de Bus del *Data Warehouse* (*Data Warehouse Bus Architecture*), que constituye la metodología de desarrollo propuesta por el mismo para lograr obtener el *Data Warehouse*, dicha metodología se ilustra a continuación.

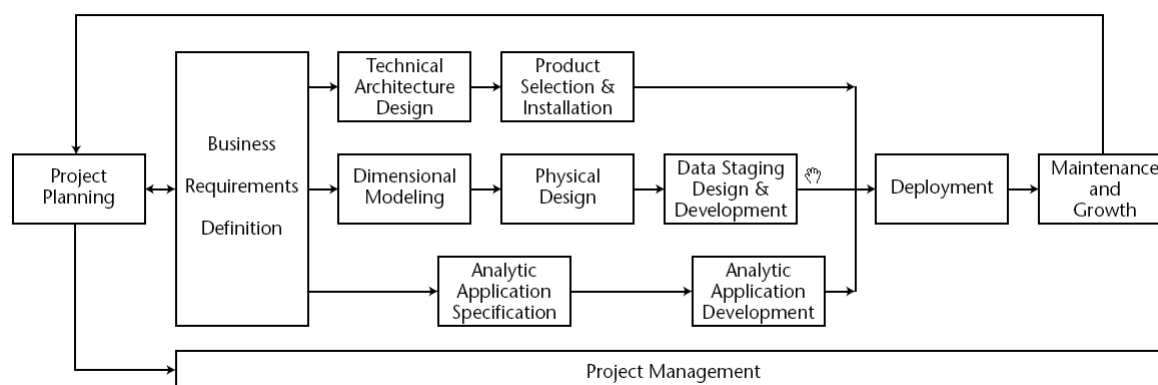


Ilustración 2. Arquitectura de Bus de Data Warehouse. Tomada de [KIMBALL, 2008]

La ilustración 2 resume la metodología propuesta por Kimball, este es un modelo sencillo que permite la elaboración de productos de inteligencia de negocios rápidamente y que se basa en la aplicación de tres fases paralelas que definen los aspectos tecnológicos, de diseño y de desarrollo para una herramienta de análisis, como se puede apreciar es un método iterativo (al llegar a su final vuelve a comenzar nuevamente), y, aunque no conste en la ilustración, incluye además una enfática e interesante recomendación de no pasarse de tres meses entre cada pasada del ciclo, acompañado todo el tiempo de un proceso de administración de proyectos continuo que lo supervise.

Los pasos de esta metodología se resumen de la siguiente manera:

1. **Planeación del proyecto (*Project Planning*):** Implica la validación de las condiciones óptimas para la generación del proyecto, la existencia de patrocinadores, cultura analítica, viabilidad de los datos, calidad de los mismos y el alcance deseado son algunas de las consideraciones más importantes a tomar en esta etapa.
2. **Definición de requerimientos de negocio (*Business Requirement Definition*):** Esta etapa requiere la ejecución de tres pasos básicos:
 - a. **Pre-planeación de los requerimientos:** Entender la necesidad del usuario, qué hace cómo lo hace y porqué lo hace, cuál es la razón por la que requiere la información que está solicitando.
 - b. **Recolección de requerimientos:** Consiste en el proceso tradicional de recibir del usuario los requerimientos para la nueva plataforma.
 - c. **Documentación y seguimiento post-recolección:** Finalmente corresponde realizar la consolidación en un documento final de análisis para los requerimientos y el seguimiento de los mismos conforme se dé el avance el proyecto.

En esta etapa es crucial que además se haga una correcta priorización de los requerimientos, basándose en un análisis de la factibilidad y el impacto de los requerimientos, cuando estos son numerosos y complejos lo mejor es dar mayor prioridad a los que sean más factibles y que representen el mayor impacto primero.

3. **Línea tecnológica:** Implica el reconocimiento y construcción de la plataforma que soportará el *Data Warehouse*, y se subdivide en una etapa de diseño (*Technical Architecture Design*) y una de selección e instalación (*Product Selection and Installation*). La idea de esta etapa es que conforme las necesidades definidas hayan dictado, la plataforma tecnológica deberá ajustarse correspondientemente para cumplir con los requerimientos de información.
4. **Línea de diseño de datos:** Implica el diseño de las estructuras de datos y procesos de integración adecuados para la formación de un *data mart* o la mejora de uno ya implementado. Incluye tres etapas principalmente:

- a. **Modelado dimensional (*Dimensional Modelling*)**: Que a su vez incluye tareas como identificar el proceso de negocio, definir la granularidad de los datos a utilizar, definir las estructuras de dimensiones y de hechos.
 - b. **Diseño físico (*Physical Design*)**: Se define cómo se almacenarán los datos en los elementos físicos de almacenamiento y si se deberán utilizar técnicas de particionamiento o similares para mantenerlos.
 - c. **Diseño y desarrollo de Staging (*Data Staging Design & Development*)**: Esta etapa consiste en la definición de las estructuras de apoyo y de dimensiones conformadas en el área de trabajo para poder cumplir con lo que requiere el modelo de datos definido.
5. **Línea de aplicación**: Consiste en los pasos necesarios para la creación del “Front-End” del data mart y se subdivide en estos subprocesos:
- a. **Especificación de la aplicación (*Analytic Application Specification*)**: Existen dos conjuntos diferentes de requerimientos, uno es de los datos a integrar a la plataforma y otro es el conjunto de salidas que se espera obtener de estos datos, en esta etapa se define qué se requerirá explotar del repositorio o data mart.
 - b. **Desarrollo de la aplicación (*Analytic Application Development*)**: Una vez definido lo que se va a requerir para ser explotado del repositorio se procede a desarrollar las interfaces que permitirán a los usuarios interactuar con los datos de la manera que lo necesitan.
6. **Implementación (*Deployment*)**: consiste simplemente en colocar en los ambientes de producción lo desarrollado.
7. **Mantenimiento y Crecimiento (*Maintenance and Growth*)**: en esta última etapa se deberá velar por el correcto funcionamiento del repositorio y de que los volúmenes de datos ingresados no afecten el rendimiento o la estabilidad de la plataforma, además se retoma los requerimientos priorizados anteriormente que fueron descalificados para considerar su inclusión en la siguiente etapa.

Con la definición de este proceso, Kimball nos propone un conjunto de pasos sencillos para trabajar en la construcción de modelos dimensionales como la base para *Data Warehouse*, sin embargo fue de nuestra consideración que la visión y el procedimiento presentados no se adaptaría bien al entorno del BCR, por razones tales como:

1. Los *data marts* son instrumentos de propósito específico, y el análisis realizado por las diferentes áreas de negocio se nutre generalmente de una visión muy específica de un conjunto de variables que pertenecen a los productos principales de captación y colocación. Si la estrategia de implementación de *Data Warehouse* se orienta a la construcción de *data marts*, pronto tendríamos

muchas versiones diferentes de la misma información y eventualmente se podría tornar muy costoso administrar modificaciones en este tipo de plataforma.

2. Con múltiples *data marts* consumiendo la misma fuente de información todos los días, los procesos de integración de datos que tomen información de *Staging Area* requerirían poder escribir en cada uno de los modelos de propósito específico, esto haría cada vez más complejo el proceso de ingreso de datos a *data warehouse*, ya que cada nuevo *data mart* que requiera información de un producto bancario se deberá integrar al flujo existente, aumentando el tiempo de carga, reduciendo las “ventanas de tiempo”¹ destinadas a la carga de información y creando nuevos y cada vez más complejos posibles puntos de falla.
3. La creación de *data marts* únicamente, ocasionaría que se perdiera la oportunidad de desarrollar un modelo unificado de propósito específico para el negocio, un instrumento único para brindar una visión completa de nuestros negocios y clientes de manera sencilla.
4. La complicación inherente a realizar consultas cruzadas entre *data marts* haría tender a un esquema de federación de datos que no se considera deseable, dadas las complicaciones que este tipo de esquemas ha demostrado tener en experiencias anteriores.
5. Las consultas cruzadas entre *data marts* podrían no siempre ser posibles o deseables ya que los modelos dimensionales están hechos con propósito específico para un área y no necesariamente facilitan el procesamiento de información para análisis cruzados debido a características como la des-normalización personalizada o a la granularidad inadecuada de los datos, esto podría provocar la necesidad de crear *data marts* que sean de propósito general, complicando aún más la administración de la plataforma, y aumentando el uso de recursos de almacenamiento, carga y monitoreo.
6. Si se utiliza este modelo y se generan modelos de estrella solamente, se reduce significativamente la capacidad de encontrar patrones no relacionados o correlaciones entre los datos originales, ya que el modelo de datos en estrella limitará significativamente el modelo a las relaciones utilizadas por el negocio para el propósito específico que dio origen al modelo.

¹ Una ventana de tiempo es un periodo de tiempo definido durante el cual es permitido que se ejecuten los procesos de carga de datos de Data Warehouse, ya que la consulta de los sistemas de producción puede afectar negativamente el rendimiento de los sistemas y esto podría afectar la continuidad del servicio.

La Fábrica de Información Corporativa (Propuesta de Bill Inmon)

Una vez examinada la arquitectura y metodología para la construcción del “*Data Warehouse Dimensional*” (como podríamos referirnos a la variante de Kimball) es importante observar detenidamente la propuesta realizada por Bill Inmon a fin de poder contrastarlas entre sí y poder contar con criterio suficiente para seleccionar entre ellas.

Para dar una definición que esté orientada a la propuesta de Inmon podríamos decir que un *Data Warehouse* es una colección de datos orientado a temas, integrada, variable en el tiempo y no volátil, usada para la toma de decisiones estratégicas, la cual provee una visión común de los datos empresariales, que está formada por información histórica, que es constante, consistente y confiable para cualquier consumidor, que debe ser diseñado para soportar crecimiento y expansión constante y que debe constituirse como el repositorio centralizado de donde fluye constantemente la información hacia los *data marts*.

Aunque el concepto de Fábrica de Información Corporativa (CIF) es más amplio que el de *Data Warehouse* y lo incluye, es acostumbrado llamar a la CIF, *Data Warehouse*, especialmente porque el término es generalmente aceptado por la industria por fuera del contexto de la arquitectura de Inmon y toma valor específico solo dentro de esta, por ello, salvo aclaración que denote lo contrario ambos términos se usan indistintamente, en este documento.

La metodología para la construcción de un *Data Warehouse* basado en la propuesta de Inmon, conocida en inglés como “*Corporate Information Factory*” o simplemente por las siglas CIF, comprende el proceso de creación de 4 niveles de modelos los cuales guían el proceso de desarrollo, siendo estos:

1. Modelo de Áreas Temáticas (*Subject Area Model*).
2. Modelo de Datos de Negocio (*Business Area Model*).
3. Modelo de Sistema de Data Warehouse (*Data Warehouse System Model*).
4. Modelo Tecnológico (*Technology Model*).

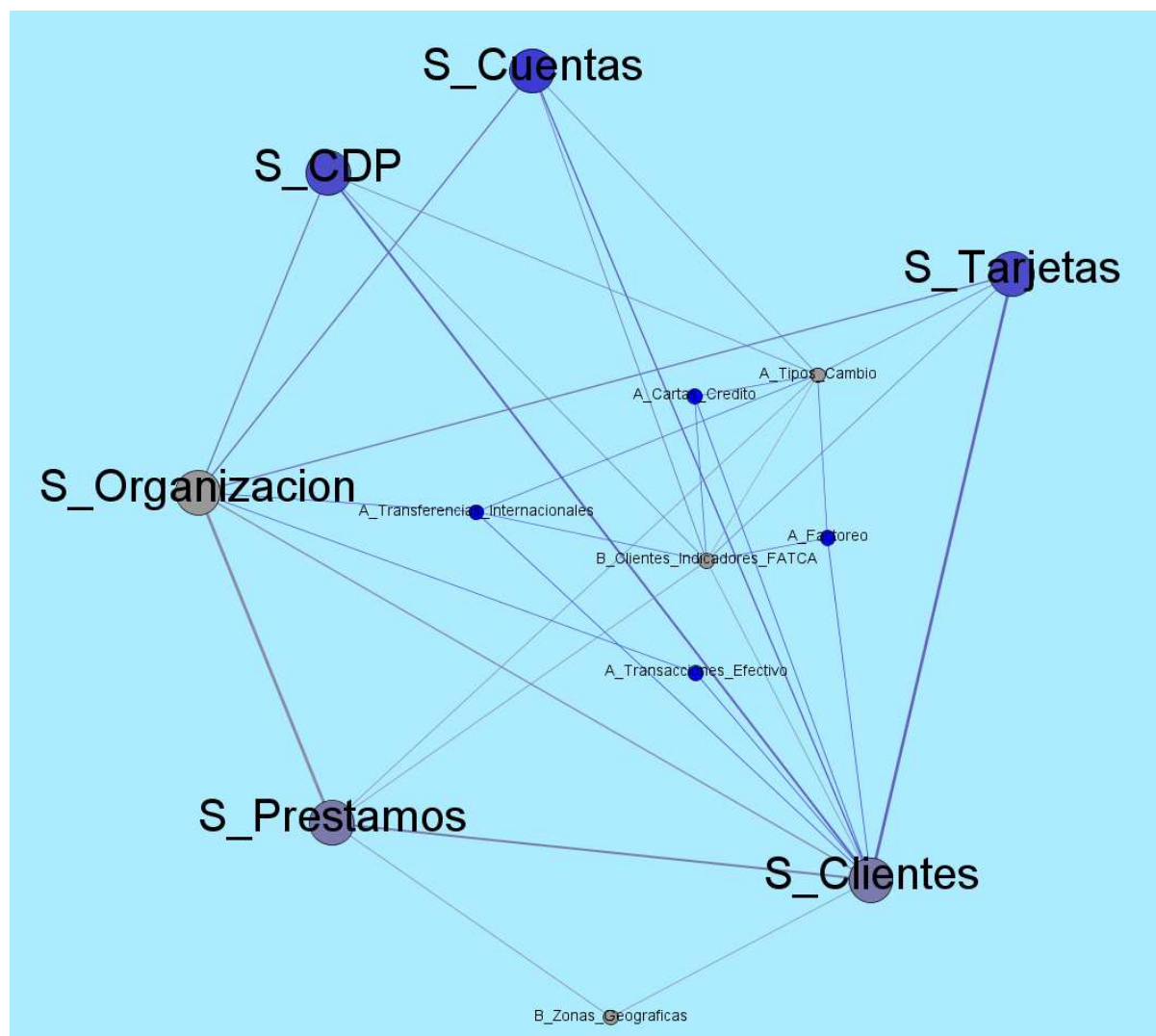
Ahora veremos cada uno de estos especificando sus propiedades y las recomendaciones a seguir para su construcción.

Modelo de Áreas Temáticas

Las áreas temáticas son conceptos de alto nivel que representan temas de interés para la organización, algunas veces estas podrían definirse como una sola tabla a nivel de base de datos o como un conjunto con un número significativo de entidades íntimamente relacionadas, para nuestros efectos este modelo se ha detallado mediante la creación de un grafo no dirigido que representa la relación entre los temas e incluso subtemas que la organización quiere llevar a la CIF.

La metodología de propuesta para la construcción de la CIF plantea la necesidad de definir inicialmente un primer subconjunto de áreas temáticas en compañía del usuario pero la construcción de este modelo no es necesariamente una lista exhaustiva de todos los temas manejados por el negocio sino que es un compendio que puede ir creciendo y mejorándose conforme se visualiza la necesidad de expandir o incluir nuevos temas.

A manera de ejemplo se muestra el modelo de áreas temáticas que se proyecta haber incluido en *Data Warehouse Corporativo* para finales del año 2013, para su mejor entendimiento se han modelado los productos principales (temas primarios) con nodos en color azul y se ha etiquetado los mismos con el prefijo "A_", los demás nodos son estructuras complementarias (temas secundarios) y se han colocado en color gris y con etiquetas antecedidas por el prefijo "B_"; los nodos de mayor tamaño cuya etiqueta se incluye el prefijo "S_" son áreas temáticas compuestas las cuales agrupan otros elementos; las aristas simples son azules o grises dependiendo de si dentro de la relación existe un nodo que sea de un tema primario o si es una relación entre temas secundarios, conforme se vaya perfeccionando el modelo se podrán utilizar otros tipos de notaciones como diferentes tamaños o colores en los nodos para representar un mayor volumen de datos o un peso sobre las aristas para representar relaciones de diferentes clases entre los temas.



El modelo de áreas temáticas se asemeja a una ciudad vista desde una gran altura podemos ver las relaciones entre las autopistas más importantes y la ubicación de algunos puntos de mayor tamaño (importancia) pero no podemos ver el detalle de cada uno hasta que nos acerquemos a él. La idea de este tipo de modelo es ser la primera de múltiples capas de abstracción que permitan llegar de poco a poco a especificar las subdivisiones de los temas, los objetos lógicos y finalmente los objetos físicos que conformarán la CIF.

Modelo de Datos de Negocio

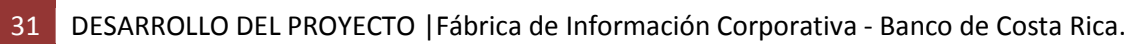
El modelo de datos de negocio tiene como finalidad ser un segundo nivel de detalle de la información que se incluirá dentro del *Data Warehouse*, este modelo debe ser específico pero no llegar a definir aspectos físicos o reflejar estructuras de datos finales, sino que debe representar los diferentes subtemas o secciones que conforman cada uno de las áreas temáticas.

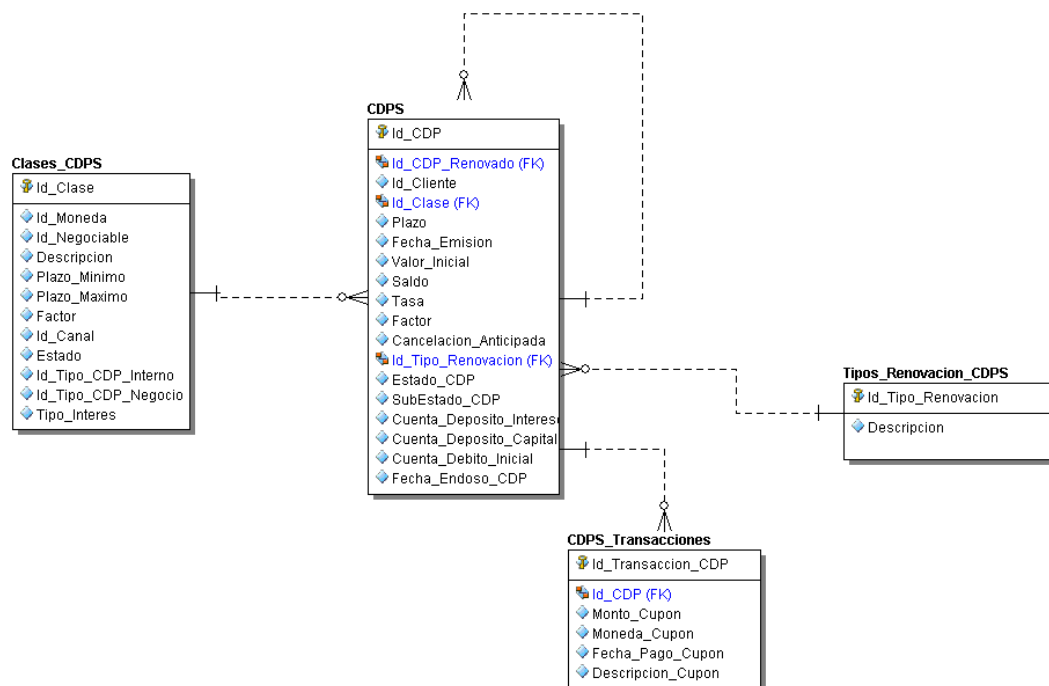
La metodología nos dicta que se debe seguir un proceso de seis pasos para la elaboración del modelo de negocio, estos pasos se resumen brevemente a continuación (tomados de [IMHOFF, 2003]):

1. **Identificar las áreas temáticas de interés cuyos datos se requerirán para el proyecto:** Este paso pretende dar prioridad a ciertos elementos más críticos que otros, se da por entendido que el modelo de áreas temáticas puede ser demasiado denso para ser abarcado por completo en la primera iteración del proceso de desarrollo, por esto es necesario establecer el orden de los desarrollos de acuerdo a las posibilidades y a los beneficios a obtener con la implementación de una nueva área temática o una mejora de la misma, se recomienda encarecidamente que se sea preferiblemente conservador en la cantidad y tamaño de áreas temáticas abarcadas, ya que de esta manera se generará valor de manera más inmediata y se podrá administrar el proceso de desarrollo de mejor manera.
2. **Identificar las entidades de interés dentro de las áreas temáticas afectadas y establecer sus nombres:** El concepto de entidad puede fácilmente confundirse con el término entidad en un modelo de base de datos E-R (Entidad – Relación), sin embargo en este contexto la entidad equivale a conceptos de negocio relacionados al área temática en particular. Por ejemplo, un préstamo tiene ciertos componentes asociados, como los planes de inversión, las garantías, las pólizas asociadas, los arreglos de pago establecidos con el cliente y cada uno de estos conceptos constituye una entidad desde el punto de vista del modelo de negocio. Aunque las entidades pueden tener correspondencia directa con estructuras de datos de los sistemas fuente es importante poder delimitar estos conceptos de manera independiente a la fuente de datos para enfocar el desarrollo hacia “conceptos universales” y no a estructuras de datos particulares, esto garantizará una mejor adaptabilidad del modelo a los cambios del mundo exterior, aunque se considera un insumo importante las estructuras usadas en las fuentes de datos, no se deben tomar estas como el modelo de negocio final. Se aconseja además en este paso generar un inventario de entidades que incluya una descripción concisa pero completa de cada entidad y sus particularidades.
3. **Determinar las relaciones entre cada par de entidades:** En este paso se definen las relaciones existentes entre las entidades de manera que se pueda determinar la forma en que interactúan y cómo juntas conforman el área temática particular, para esto se recomienda la utilización de una herramienta de modelado de datos que permita establecer relaciones y preferiblemente agrupar entidades para reflejar dentro de la misma las áreas temáticas a las que pertenecen.
4. **Agregar atributos a las entidades:** En esta etapa se deben establecer los posibles campos que se requerirá incluir en las entidades cuando estas se conviertan en objetos dentro del *Data Warehouse*.

5. **Confirmar la estructura del modelo:** Este paso consiste en la verificación de la aplicación de la tercera forma normal sobre el modelo desarrollado, el modelo de negocio no se implementará como tal en el *Data Warehouse*, pero es la base para modelos subsecuentes en este, para un sistema transaccional o para un *data mart* y la validación de su cumplimiento de la tercera forma normal, es importante para garantizar que los modelos subsecuentes tengan la calidad adecuada.
6. **Confirmar el contenido del modelo:** Finalmente corresponde verificar que el modelo contiene aquello que el negocio considera necesario para cubrir sus necesidades, en este caso es necesario que el analista o encargado del modelo lo valide en conjunto con los usuarios de negocio para garantizar que todos los elementos han sido incluidos y que la estructura y relaciones elegidas son apropiadas, en algunos casos la validación puede apoyarse en la propia representación gráfica del modelo pero a veces será necesario expresar el mismo en reglas de negocio para poder validarlo con el usuario.

Así por ejemplo podríamos detallar el modelo de negocio para un tema particular, o bien detallar el modelo de áreas temáticas de manera que se muestren relaciones entre los temas a nivel de los elementos más específicos que los componen. La implementación hecha en el BCR divide el modelo de negocio en dos modelos separados uno representado como un grafo no dirigido más específico que el de áreas temáticas y otro, de carácter temporal diseñado a nivel de un diagrama lógico hecho en una herramienta de modelado de datos, a continuación se muestra un ejemplo de ambos diagramas (simplificados) a modo de ilustración.





Los componentes de las áreas temáticas complejas se detallan en el grafo de modelo de negocio como un conjunto de nodos relacionados mediante una forma geométrica donde los vértices representan, cada uno, un elemento de la correspondiente área temática.

En cuanto al modelo lógico mostrado, primeramente hay que aclarar que este ejemplo está simplificado del real de modo que sirva simplemente como ilustración del concepto; para cada nuevo tema a introducir se genera un diagrama como este donde se puede reflejar los elementos que serán tomados en cuenta, esto permite una comunicación más fluida con los usuarios, propone una base sobre la que se desarrollarán los modelos posteriores y permite una visualización sencilla del alcance que incluirá cada tema antes de proceder a modelos más refinados y específicos para *Data Warehouse*.

Modelo de Sistema de Data Warehouse

Los modelos en tercera forma normal generados como modelos de negocio son la base sobre la cual se generará el modelo de sistema o de *Data Warehouse*, este tipo de modelo se basa en las características dinámicas o estáticas de los datos y propone la creación de nuevos tipos de estructuras semi-normalizadas para contener la información histórica, los datos derivados, las sumalizaciones pre calculadas y la variable tiempo. Para realizar esta tarea la metodología

especifica un conjunto de ocho pasos para transformar el Modelo de Datos de Negocio en el Modelo de Sistema de *Data Warehouse*.

Según [IMHOFF, 2008] los ocho pasos se pueden analizar divididos en dos grupos; los primeros cuatro pasos tienen que ver principalmente con aspectos de negocio y los pasos del quinto al octavo tienen que ver principalmente con aspectos de mejora en el desempeño.

1. Seleccionar los datos de interés.
2. Agregar tiempo a la llave de la entidad.
3. Agregar los datos derivados.
4. Determinar el nivel de granularidad.
5. Sumarizar datos.
6. Mezclar entidades.
7. Crear arreglos.
8. Segregar datos.

1. Seleccionar los datos de interés: Consiste en la escogencia de datos a partir del modelo de negocio que deberán ser llevados a una implementación según lo defina el negocio, este paso garantiza que haya alineamiento a los objetivos y el propósito del negocio y confina el alcance del proyecto a justo lo que es necesario en este momento. Para la realización de esta fase se considera necesario considerar los modelos de datos de negocio, el documento de alcance del proyecto, requerimientos, reportes ya implementados en otras plataformas, prototipos (de modelos dimensionales por ejemplo) y los modelos de datos de los sistemas que se consideran se convertirán en las fuentes de datos. Algunas consideraciones importantes para este punto son:

- a. Los datos transaccionales que no forman parte de los datos principales de cada tema requieren una atención especial para determinar si tienen valor analítico y por tanto deben ser seleccionados para incluirse.
- b. Si el usuario solicita un dato calculado pero solicita solo el resultado y no la información requerida para calcularlo, se deberán incluir igualmente cualquier información necesaria para el cálculo, ya que la fórmula de cálculo podría variar o bien el usuario puede desear revisar de forma manual posteriormente.
- c. Cuando se trata de datos transaccionales, si se tiene duda que se necesitará, y tenemos la posibilidad es mejor incluirlo desde el principio, dado que: las transacciones podrían ser eliminadas de la fuente, recuperar datos faltantes de transacciones es complejo y la integración de datos transaccionales es relativamente simple, por otra parte los datos de referencia que no se sepa con seguridad si deben incluirse, pueden ser excluidos de manera segura dado su relativa facilidad para ser incorporados posteriormente.
- d. La ejecución de este primer paso es fundamental ya que determina en gran medida el alcance, la duración, el costo y el desempeño del proyecto.

2. **Agregar tiempo a la llave de la entidad:** Este paso consiste en la inclusión de la variable de tiempo al modelo de datos de negocio, con el fin de formar entidades que puedan tener naturaleza histórica, la simple inclusión de un atributo referente al tiempo dentro de la llave primaria puede ser efectivo en algunos casos, pero en casos donde esta llave es utilizada para servir de referencia a una llave foránea con otra entidad deben considerarse situaciones derivadas del cambio, la metodología presenta la utilización de una de 5 técnicas para representar el modelo con información sobre el tiempo pero recomienda sólo el uso de algunas y de estas hemos elegido dos para que sean utilizadas de forma estándar en el BCR por facilidad se han nombrado de la siguiente manera:
 - a. **Datos Dinámicos Segregados (DDS):** En este esquema la tabla que requiere manejar historia se divide en dos, la primera o tabla principal, contiene la información de la llave primaria de la entidad tal como se definió en el modelo de negocio y mantiene una copia de aquellos atributos de la entidad original que no sean cambiantes o cuyo cambio no requiera ser registrado, esta tabla es a la que se le asocian las llaves foráneas que otras tablas requieran; la segunda tabla o tabla DDS tiene la misma llave primaria de la principal más un atributo de tiempo (pueden ser dos para fecha y hora), además en esta estructura se encuentran los atributos de la entidad original que son dinámicos y que se requieren almacenar de forma histórica.
 - b. **Entidad Histórica Separada (EHS):** En este caso se genera igualmente una tabla principal, sin la fecha en su llave primaria, a la que se le ligan las llaves foráneas de cada entidad relacionada, en esta tabla se almacenará siempre la última versión del registro, por otra parte, se generará una tabla secundaria o tabla EHS con la misma estructura de la principal, pero con la variable de tiempo en su llave primaria, esto permitirá que en ella se almacenen todas las versiones anteriores del registro.
3. **Agregar los datos derivados:** Este paso tiene más un fin de negocio que un fin de desempeño, aunque actúa en ambos sentidos, al almacenar un dato calculado en *Data Warehouse* se garantiza que todos los que lo consuman están utilizando la misma forma de calcularlo y lo entienden de la misma manera, en esta etapa es crucial que el diseñador establezca al menos la necesidad del dato derivado, ya que es posible que múltiples áreas de negocio tengan que ponerse de acuerdo en la forma del cálculo, y esto no es crucial en este momento ya que no tiene impacto en el modelo más allá que en la definición del modelo de datos tecnológico, que llevará el modelo a una implementación en una sistema gestor de base de datos particular, además la inclusión de datos derivados en el *Data Warehouse* asegura una definición y algoritmo de derivación estandarizado a cualquier *data mart* que haga uso del dato posteriormente.

4. **Determinar el nivel de granularidad:** Este punto define el nivel de detalle que deberá almacenarse y estar disponible dentro del *Data Warehouse*, por tanto, tiene implicaciones de costo y administración importantes. Múltiples factores afectan la decisión del nivel de granularidad de los datos:

- a. La necesidad actual y proyectada del negocio.
- b. El tipo de industria particular donde se está trabajando.
- c. Las necesidades futuras de proyectos de minería de datos.
- d. La presencia de datos derivados.
- e. La granularidad de los sistemas transaccionales fuente.
- f. El desempeño de los procesos de adquisición de datos.
- g. Costo de almacenamiento.
- h. Costo de administración de la plataforma.

En implementaciones de *Data Warehouses* pequeños, es a partir de esta etapa donde se podrían comenzar a ejecutar proyectos paralelos para creaciones de *Data Marts*, ya que se tiene noción del contenido que podrá ser consumido por estos y de esta manera el negocio puede irse haciendo una idea de lo que podrá tener disponible y se obtendrán mayores insumos para los pasos posteriores.

5. **Sumarizar datos:** Esta etapa consiste en la generación de estructuras de datos que sumarizan información de entidades más generales con el propósito de brindar mecanismos más eficientes para extraer información del *Data Warehouse*, en esta etapa se busca generalmente la agrupación a través de variables como el tiempo o un producto particular. Existen principalmente 5 tipos diferentes de agregación dentro de *Data Warehouse*:

a. **Para datos referentes a un periodo de tiempo:**

- i. **Sumarización acumulativa simple:** Se enfoca en el conjunto completo de datos y agrupa por diversos criterios la transaccionalidad. Por ejemplo, agrupar las ventas por fecha y producto.
- ii. **Sumarización rotativa:** Se genera para un periodo de tiempo definido y se sustituye con otro periodo de tiempo cuando este pierde interés. Por ejemplo mantener los resultados del mes anterior en una estructura y suplantarlos con los del mes actual al final de mes.

b. **Para datos referentes a un punto en el tiempo:**

- i. **Sumarización directa simple:** Provee una vista de los datos sumariados a intervalos regulares. Por ejemplo, tener un registro por mes de la transaccionalidad para cada cliente.
- ii. **Sumarización continua:** Crea un nuevo registro sumariado solo cuando la información es modificada. Por ejemplo una tabla que almacene el monto transado al final de la semana de las tarjetas de crédito que hayan tenido movimientos.

c. **Para ambos tipos de datos.**

- i. **Sumarización vertical:** Su uso es aceptable a nivel de *Data Marts* principalmente debido a que ambos tipos de datos tienen granularidad diferente y no deberían encontrarse en la misma entidad de *Data Warehouse*. Este tipo de sumarización permite hacer análisis comparativo entre sumarizaciones a un punto en el tiempo, con sumarizaciones al final de un periodo. Por ejemplo, un caso implementado, sería la estructura la desviación entre lo movido y declarado por un cliente al final de mes.
- 6. **Mezclar entidades:** Este proceso consiste en unificar en una sola dos o más entidades con el objetivo de reducir el número de uniones o “*JOINS*” y mejorar la consistencia entre los datos, las entidades originales pueden ser igualmente mantenidas, esta técnica equivale a desnormalizar datos y se practica como un medio para generar dimensiones conformadas, que posteriormente serán utilizadas en los *Data Marts*. Se requiere que las entidades involucradas cumplan con al menos 3 criterios para ser mezcladas:
 - a. Las entidades deben compartir una llave común.
 - b. Los datos entre las entidades se usarán juntos frecuentemente.
 - c. Los datos comparten un patrón de inserción similar.
- 7. **Crear arreglos:** Este paso no es muy comúnmente utilizado pero puede ayudar significativamente al desempeño y el llenado de *Data Marts*, consiste en la creación de una tabla con grupos repetitivos u “ocurrencias” (contrario a lo que indica la primera forma normal) otra manera de verlo es que será una estructura “*pivoteada*” de filas con un criterio común. Esta técnica sólo es recomendable si:
 - a. La lista de ocurrencias de los grupos repetitivos es relativamente pequeña, hasta 15 sería un límite recomendable.
 - b. Estas ocurrencias son frecuentemente usadas juntas.
 - c. El número de ocurrencias es predecible y no variará.
 - d. El patrón de inserción y eliminación de datos es estable (todos los datos del arreglo estarán disponibles al mismo momento), así un arreglo para mantener los totales por mes no sería aceptable, pero sí podría crear un arreglo para almacenar totales de movimientos por provincia.
- 8. **Segregar datos:** Este último paso consiste en el examen de las estructuras resultantes desde el punto de vista de la estabilidad y utilización de los datos, esto se logra mediante el estudio de los patrones de inserción de los datos (estabilidad), y la frecuencia y destino hacia el que se moverán los mismos a los *Data Mart* (utilización), de esta manera pueden tomarse decisiones tales como segregar (dividir) estructuras de manera que los diferentes patrones de inserción y consumo puedan trabajar sobre estructuras separadas, este tipo de consideraciones podría generar decisiones de diseño tales como separar estructuras DDS en dos o más partes, dados patrones de

inserción diferentes para subconjuntos de datos que hayan sido incluidos en la estructura originalmente definida.

El resultado final de la aplicación de este proceso es un conjunto de estructuras originalmente derivadas de los modelos de negocio pero que ahora contemplan los aspectos únicos que rodean un *Data Warehouse*, y que conforman el conjunto de entidades del Modelo de Sistema de *Data Warehouse*.

En el caso del BCR el Modelo de Sistema de *Data Warehouse* se encuentra documentado en una herramienta de modelado como un diagrama lógico de base de datos y se mantiene en un repositorio centralizado para la manipulación y consulta por parte del equipo de trabajo.

Modelo Tecnológico

Este constituye el último de los modelos relacionados a la construcción de la CIF y equivale a la representación física de un modelo de base de datos, en el cual se deberán especificar los tipos de datos, tipos de relaciones, restricciones, nulidad y otros atributos propios de un sistema gestor de base de datos seleccionado, para la elaboración de este modelo se utiliza un modelo físico de datos implementado en una herramienta de modelado de datos, este modelo al igual que el de sistema se encuentra en un repositorio centralizado con el fin de poder ser consultado y modificado por el equipo de trabajo de *Data Warehouse*.

Consideraciones finales para la comparación entre arquitecturas.

En la CIF es reconocido que no se puede pretender abarcar todo el alcance visualizado en una sola iteración, así el proceso de definición de los modelos de áreas temáticas, negocios, de sistema y tecnológico de la CIF, se encierra en un proceso iterativo, igual que en el caso de la metodología de Kimball.

Aunque este ciclo de desarrollo continuo es una similitud con la metodología anteriormente revisada, no se puede asumir que las metodologías sean intercambiables, existe una importante diferencia que a este punto debería ser evidente, los modelos sobre los que se basa la CIF tienen un enfoque integral de negocio y el enfoque multidimensional tiene un enfoque al producto, esto provoca que las estructuras finales de la CIF se construyan de manera que pueden corresponder con cualquier consulta que el negocio requiera, en vez de enfocarse específicamente a una necesidad particular, pero esto viene a un costo mayor de desarrollo y de planeación, al menos en los primeros proyectos.

Por un lado, la propuesta del Ralph Kimball, busca la máxima eficiencia en el desarrollo y la obtención de resultados efectivos en el menor tiempo posible y podemos decir que se sostiene sobre la premisa de que las necesidades de análisis son la guía primaria del desarrollo, en vez de ser obtenidas como la consecuencia de la creación de un producto con visión integral del negocio, con esto queremos decir, que el enfoque de Kimball nos permite empezar desde lo que queremos obtener y a partir de esto comenzar a trabajar hacia atrás para obtener un medio

de conseguir ese producto que requerimos. Este esquema de trabajo nos permite mantener un acercamiento constante en la necesidad del negocio y centrar nuestros esfuerzos en lo realmente necesario y urgente.

El enfoque multidimensional es lo que comúnmente se llama una estrategia *Up-Down*, identificar que tan alto requiero llegar para luego buscar la forma de “construir una escalera” para llegar hasta esa altura, no importa cómo sea la escalera mientras podamos llegar a la altura que deseamos, una de sus principales ventajas entonces se refiere a la relativa rapidez con la que se pueden obtener resultados, pero este enfoque tiene algunas dificultades importantes a largo plazo.

Por otra parte la propuesta de William Inmon, es una propuesta que podríamos decir es más un proceso de reflexión y de mediano a largo plazo, la misma nos propone la centralización y auto conocimiento como base para la expansión de los modelos de inteligencia de negocios. Así, el modelo de la Fábrica de Información Corporativa (CIF) de Inmon, propone comenzar por la creación de un modelo del negocio, o al menos una parte de él y comenzar a expandirlo conforme nuestras necesidades, este modelo deberá reflejar la estructura que intrínsecamente forma la información en la organización mediante la cooperación de sus sistemas de información, siendo el proceso de crear un *Data Warehouse* un proceso más amplio, iterativo y de descubrimiento que un esfuerzo focalizado, explosivo y de producción, tal como podría verse en el enfoque de Kimball.

1.3. Selección de una arquitectura para el BCR

Está claro para todos los que trabajamos en este proyecto dentro del Banco, que no es deseable crear una arquitectura lógica o definir una metodología de *Data Warehouse* completamente nueva, como ya lo han hecho Kimball e Inmon, en esta etapa entonces nos dimos a la tarea de elegir una de las mismas y a partir de esta estudiar cada uno de sus componentes. Fue así como se determinó desde el nivel más alto hasta el nivel más bajo, y tal vez lo más importante, de acuerdo a las posibilidades y el entorno actual del BCR, la que consideramos es la mejor forma de implementar las propuestas hechas en el modelo teórico.

Al final de un proceso de reflexión en conjunto con los compañeros involucrados en el proyecto y de evaluar las opciones y recursos que teníamos disponibles se decidió optar por el modelo de Inmon, y comenzar a dirigir nuestros esfuerzos a la creación de la figura del *Data Warehouse* Corporativo centralizado, como la piedra angular sobre la que posteriormente se desarrollarían los *Data Marts* y demás repositorios de información.

Para poder comenzar con este tema entonces tuvimos que hacer un repaso exhaustivo del modelo de Inmon para la Fábrica de Información Corporativa y tomar cada componente para detallar cómo sería su implementación, tanto a nivel lógico como a nivel físico.

Un aspecto interesante de la arquitectura que propone Inmon, es el hecho de que es una arquitectura exclusivamente de alto nivel, en su propuesta, no se pueden observar buenas

prácticas sobre la mejor forma de implementar las estructuras de datos que requiere el modelo de cada repositorio y no se hacen propuestas sobre estructuras particulares. Simplemente se presentan las definiciones y características que cada repositorio debe tener, sus interrelaciones y la forma en que se deben operar. Como se observó en las secciones anteriores las recomendaciones en este sentido no eran basadas en la bibliografía de Inmon.

Lo anterior podría ser visto como una debilidad importante y en nuestra opinión es una de las razones por las cuales el modelo de Inmon no es tan ampliamente utilizado como el de Kimball, sus componentes son estudiados pero no son generalmente implementados por las grandes compañías productoras de software dentro de sus paquetes de inteligencia de negocios, dado probablemente a la inherente complejidad inicial de su construcción y a que su implementación, al consistir en una estrategia “*bottom-up*”, requiere de mayor esfuerzo y mayor tiempo de implementación y puede derivar en estructuras de datos mucho más especializadas que las acostumbradas estrellas del modelo dimensional.

Sin embargo en nuestra opinión la teoría de Inmon es posiblemente justo lo que le corresponde ser, una guía conceptual de alto nivel que es lo suficientemente flexible para adaptarse a cualquier organización y que permite a esta ajustar su cultura y forma de desarrollo gradualmente, hasta lograr la madurez necesaria para sacar el máximo provecho de estas herramientas.

Ahora bien, como ya se ha demostrado no es necesario que nos limitemos específicamente a la propuesta de un solo autor, así que enriquecimos nuestra propuesta con ideas de autores que dan recomendaciones de diseño de menor nivel, tal como se ha mostrado, nuestro avance para generar una propuesta para la CIF del BCR, se basó en el estudio de la arquitectura elegida y el entorno desde el más alto nivel y desciende hasta la implementación física para dar los insumos de bajo nivel que deberán constituir los estándares de desarrollo que deberán seguirse para garantizar la continuidad del servicio y en el desarrollo.

Comenzamos entonces enumerando los componentes y características más importantes de la CIF de Inmon a la vez que detallamos algunos aspectos relacionados con la realidad del BCR que fueron tomados en cuenta para establecer algunas decisiones de diseño.

1.4. La Fábrica de Información Corporativa del BCR, especificación de componentes.

Para el desarrollo de la propuesta, se considera el modelo de Fábrica de Información Corporativa presentado por Inmon en [INMON, 2008] el cual es normalmente conocido por el nombre de *Data Warehouse 2.0*²(DW 2.0), es importante hacer notar que este nombre no se refiere a una actualización que deba aplicarse a los *Data Warehouse* antiguos, sino a la segunda propuesta presentada por Inmon para la creación de la CIF, la cual toma en cuenta la

² Propuesta de arquitectura para Datawarehouse Corporativo, presentada por W. Inmon en 2008, como respuesta a las necesidades analíticas cambiantes de las industrias sobre las que él mismo desarrolló proyectos como consultor y que incluyen conceptos de Administración de Ciclo de Vida de la Información, Administración de Metadatos y Datos no estructurados, entre sus mayores aportes.

experiencia obtenida a través de los años en la implementación de la primera versión de su propuesta y se enriquece con la experiencia, observaciones y aportes de las organizaciones en donde se ha implementado su arquitectura, con esto la propuesta busca mantener vigente la arquitectura de *Data Warehouse* Corporativo a las necesidades de la industria actual manteniendo vigente su propósito inicial.

Asimismo algunos de los elementos que presenta el *Data Warehouse* 2.0 no pueden ser inmediatamente aplicados a la arquitectura del BCR, dado principalmente la falta de recursos internos para su desarrollo, aunque la arquitectura se mantendrá abierta lo suficiente para permitir su introducción en un futuro.

La arquitectura de DW 2.0 se basa en un conjunto de componentes tanto físicos como lógicos que se pueden resumir en el siguiente esquema, a continuación además se incluye una breve reseña de cada uno para ubicar al lector en el contexto:

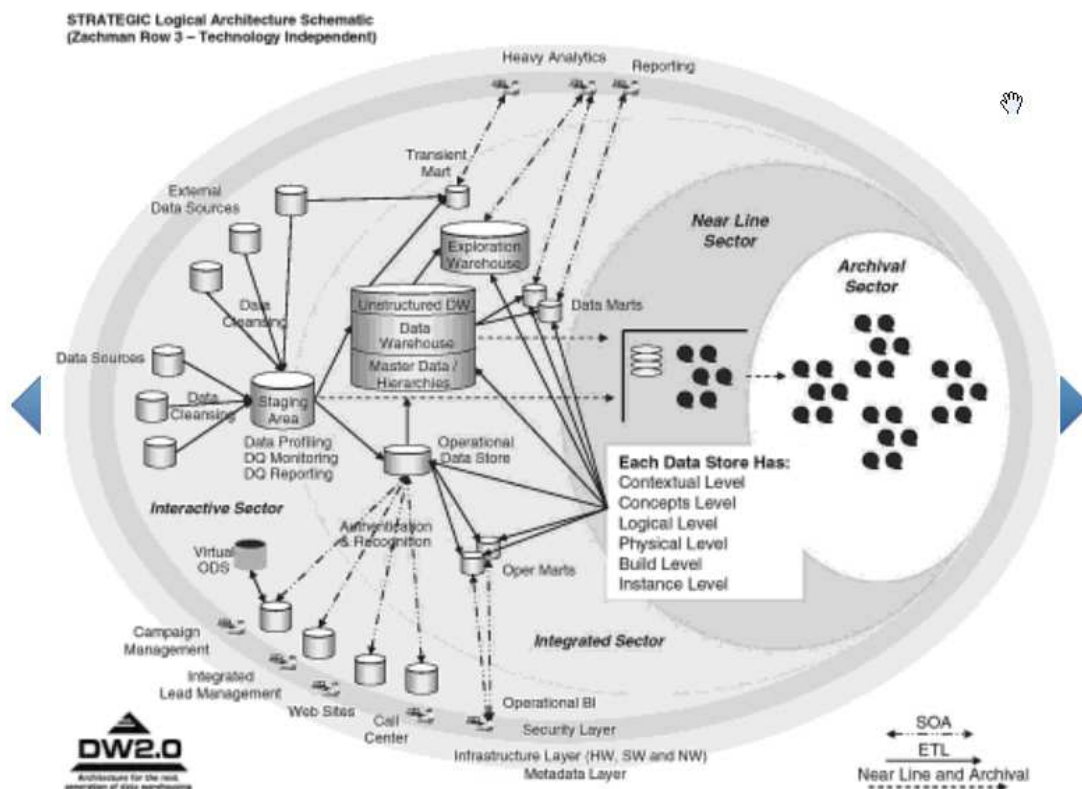


Ilustración 3. Modelo lógico de Data Warehouse 2.0 según W.Inmon. Tomado de [INMON, 2008]

1.4.1. Data Warehouse (Bodega de datos)

El *Data Warehouse* es el corazón de la Fábrica de Información Corporativa, constituye el repositorio que engloba el modelo de negocio de bajo nivel y los datos asociados al mismo, su

presencia garantiza una fuente única de información integrada para los demás repositorios, a través del desarrollo del mismo los demás repositorios relacionados podrán mejorar o incrementar la información que requieren, este repositorio fue incluido como primer gran objetivo en el BCR, su construcción se realiza mediante procesos iterativos e incrementales que permiten cubrir los más grandes temas y centrarse en la información verdaderamente crucial para la toma de decisiones.

1.4.2. Master Data / Hierarchies (Datos maestros / Jerarquías)

Esta sección del *Data Warehouse*, se considera generalmente como una sección dentro del repositorio centralizado que almacena y administra la información de los datos maestros, estos que conforman entidades que están presentes transversalmente a través del negocio y que por tanto son piezas clave para realizar análisis de cualquier información que exista en *Data Warehouse*, estos datos tienden comúnmente a convertirse en el modelo de *Data Warehouse* dimensional, como dimensiones conformadas, comunes a múltiples *data marts* para garantizar la consistencia entre datos de diferentes áreas con estos conceptos transversales. Las jerarquías son un tipo especial de datos maestros que especifican la relación entre elementos de naturaleza común, de manera que se pueden establecer relaciones de herencia entre elementos que permiten generar esquemas de agregación específicos al negocio. Existe una tendencia de mercado a generar un repositorio aparte para esta sección y a realizar sobre ella un conjunto específico de operaciones para tratarla.

1.4.3. Unstructured DW (Data Warehouse no estructurado)

El *Data Warehouse* no estructurado es el repositorio de información que contiene información cuya estructura o volumen es demasiado atípico para ser incluido en *Data Warehouse* (siendo este un repositorio que requiere una estructura relacional particular), y sin embargo, siempre mantiene algún tipo de relación con el mismo, principalmente a nivel de datos maestros.

Basándonos en las tendencias de tecnología actuales, este tipo de repositorios se pueden generar a través de herramientas diseñadas para “*Big Data*”, como HBase y Apache Hadoop, Cassandra y otros semejantes.

1.4.4. Exploration Warehouse (Bodega de datos de exploración)

El *Warehouse* de exploración es un repositorio generalmente réplica parcial o total del *Data Warehouse* central, donde se permite ejecutar procesos de consulta mucho más elaborados que los que se aplican regularmente a *Data Warehouse*, este repositorio facilita el análisis de cantidades masivas de datos por parte por ejemplo de algoritmos de minería de datos. Con las nuevas tendencias sobre “*Big Data*”, se visualiza el uso de tecnologías como Bases de Datos Columnares y repositorios que utilizan algoritmos de “*MapReduce*” para ejecutar minería de datos sobre cantidades importantes de datos, en vez de realizar una réplica de los datos para consulta por medios tradicionales.

1.4.5. Staging Area (Área de Trabajo)

El área de trabajo de *Data Warehouse* es un repositorio de información temporal que contiene datos obtenidos de las fuentes y se utiliza para apoyar el proceso de transformación e integración de los datos que ingresarán a *Data Warehouse*, muchos proveedores de equipos y

productos de *software* para Inteligencia de Negocios presentan como una ventaja el que no exista un área de trabajo previa, dado que su uso se basa en almacenamiento físico y este produce ventanas de tiempo más largas. Aún así la presencia de un *Staging Area*, como veremos más adelante, permite ejercer un control adecuado sobre los procesos de carga, recuperación en caso de errores, mejora en procesos de diferenciación de datos por mencionar algunas ventajas importantes. En el esquema de Fábrica de Información Corporativa del BCR se mantiene y recomienda el uso de *Staging Area* para poder aprovechar los beneficios ya mencionados.

1.4.6. *Data Mart, Transient Mart y Operative Mart (Mercado de datos, Mercado de datos transitorio y Mercado de datos operativo)*

Los mercados de datos, son modelos de datos derivados de *Data Warehouse* con estructura generalmente dimensional, y que comúnmente se modelan como estrellas de datos y se explotan a través de cubos de datos OLAP, para facilitar su consulta, permiten una vista específica a un propósito o área de negocios de la información de *Data Warehouse* y eliminan la necesidad de consulta directa del *Data Warehouse* por parte de usuarios analíticos que requieren consultas ad hoc. Dentro de los “*Marts*” pueden haber tipos especiales, tales como los transitorios, los cuales son construidos con naturaleza temporal para una necesidad y los operativos que contienen información poco agregada para soportar un proceso operativo que requiera información integrada.

1.4.7. *ODS: Operational Data Store (Almacén de Datos Operativos)*

El ODS es un repositorio especializado en la arquitectura de la Fábrica de Información Corporativa que almacena un amplio detalle de la transaccionalidad realizada sobre los elementos integrados del *Data Warehouse*, su existencia es requerida generalmente para la realización de consultas especializadas y específicas pero que requieren del nivel de confianza e integración que tendría el *Data Warehouse*, generalmente sus datos subsisten por un tiempo menor del que tiene el *Data Warehouse*, pero a un mayor nivel de detalle y apoyan la toma de decisiones operativas en vez de las tácticas o gerenciales.

1.5. *Anotaciones sobre la bibliografía utilizada*

Una vez expuesto el entorno que propone la creación de una CIF en la organización es importante reconocer que la teoría publicada por Inmon, no aporta mucho más allá de los conceptos básicos y caracterización del *Data Warehouse*, así la propuesta de Inmon es desde el punto de vista de arquitectura y utilizarla de manera aislada no ayudará en todos los niveles de la construcción de una CIF. Su propuesta entonces debe ser complementada con ideas de otro autor o de diseño propio que permitan especificar y estandarizar técnicas que son adecuadas para la construcción de cada elemento.

Cada repositorio tiene características únicas que deben ser consideradas para aplicar el mejor diseño y las mejores prácticas posibles en su construcción y explotación, asimismo cada

repositorio es ampliamente afectado por las condiciones que se encuentran dentro del entorno de la organización, en este caso el banco; la forma de extraer la información, los formatos, las ventanas de tiempo, los procedimientos internos, la calidad de los datos, son algunos de los aspectos únicos del BCR que afectan cada diseño y que nos llevan a la necesidad de “tropicalizar” las propuestas de los autores para hacerlas propias.

Las prácticas presentadas en esta investigación han sido puestas a prueba y mejoradas desde el diseño de la primera iteración del *Data Warehouse*, el Banco de Costa Rica a través de nuestra supervisión optó por seguir el modelo de desarrollo de bajo nivel para el *Data Warehouse* Corporativo basado en las recomendaciones para el almacenamiento de datos históricos que propone Claudia Imhoff [IMHOFF, 2003] y el proceso de generación de modelos de datos de múltiples niveles que se explicó anteriormente, esta metodología detalla a nivel de estructuras de datos la forma adecuada de modelado que requiere el *Data Warehouse* Corporativo. Aunque la propuesta de Imhoff es bastante completa, existen ciertas necesidades que no se encuentran reflejadas en su trabajo y que han tenido que ser adaptadas de otros autores o bien han sido agregadas por los analistas encargados de la implementación del proceso.

Por ejemplo, para el diseño de los *data marts*, se aconseja utilizar como es de esperarse mucha de la teoría propuesta por Ralph Kimball [KIMBALL, 2002] para el diseño de modelos de estrella y de copo de nieve, los cuales decidimos complementar con las recomendaciones de [ADAMSON, 2010]. Juntos conformarán la base para la construcción de modelos dimensionales y cubos de datos mediante el uso de las herramientas con las que dispone el Banco de Costa Rica, si fuera del interés del banco comenzar a trabajar con este tipo de repositorios.

Los estándares para los procesos de integración de datos, se generaron siguiendo los principios dados para la creación de procesos de integración de datos recomendadas por [GIORDANO, 2011] con ciertos agregados hechos por los analistas dado el entorno particular del BCR. En conjunto estas recomendaciones dieron origen a la estandarización del llamado “*Staging Area*” o Área de trabajo, tal como veremos más adelante.

Finalmente los apartados referentes a la metodología que se presentarán fueron principalmente referenciados al trabajo de Moss y Atre [MOSS, 2003], de las cuales se tomó la estrategia de desarrollo para proyectos de Inteligencia de Negocios.

2. Reflexión sobre situación previa y antecedentes.

2.1. Situación previa del BCR

Es importante a este punto justificar nuestra decisión sobre la arquitectura a desarrollar para el Banco de Costa Rica (BCR) mediante un breve desarrollo de algunos aspectos propios del negocio en cuanto a lo que se refiere a la generación y explotación de información analítica.

Un importante grupo de representantes del negocio son actualmente conscientes de la importancia y la necesidad que existe de desarrollar un punto central e integrado de consulta que soporte la toma de decisiones, y la mayoría de ellos buscan respuestas rápidas a sus necesidades de análisis, sin embargo ninguno de ellos ha mostrado un conocimiento previo en cómo formular requerimientos analíticos, la mayoría no tienen una noción perfectamente clara de cómo sus necesidades de análisis se entrecruzan con las de otros y aún no existe un foro común donde se puedan definir las prioridades de la organización como un todo.

En nuestro análisis hemos encontrado algunos casos de áreas de negocio que invierten semanas consolidando información de manera casi manual, con herramientas limitadas como Microsoft Excel o Monarch³, estas áreas muchas veces generan la misma información que otras, pero como es una necesidad diferente y no hay relación entre ellas, se repiten los procesos de integración (de forma manual) de la información mediante otros medios o herramientas, generalmente esto se hace con el objetivo de dar respuesta a instancias superiores, pero dado que son logradas a través de procesos independientes, desconocidos entre sí y en general desarrollados empíricamente se cae en la duplicación de esfuerzos, contradicción de la información y en que al final no se puede determinar a ciencia cierta de donde proviene la información o qué tan confiable es.

Hablamos en este caso que la ejecución de los procesos está a cargo de personas que si bien tienen una profunda experiencia en los procesos y en su área de especialidad (administradores, contadores, financistas) no cuentan con nociones de temas como administración de datos, diseño de modelos de datos, integración, administración de la información y calidad de datos, las cuales resultan cruciales para poder realizar un adecuado tratamiento de los datos para producir información consolidada para la toma de decisiones gerenciales. Tal vez uno de los aspectos más preocupantes, es que estos esfuerzos se utilizan para cumplir con una necesidad inmediata, y la mayoría de las veces no es posible, o no se cuenta con un medio, para mantener la información de manera histórica de forma adecuada, de manera que se pierde la visión de la evolución del negocio y la información analítica se vuelve solamente en la “foto” al día de hoy.

A primera vista la situación actual del BCR nos podría empujar a decidarnos por una metodología que nos dé resultados de la manera más rápida posible, tal como podría ser la arquitectura de *Data Warehouse* de Ralph Kimball, sin embargo los intentos anteriores por desarrollar repositorios de información consolidada bajo esta filosofía han desembocado en procesos poco mantenibles, poco administrables y poco organizados, en algunos casos demasiado enfocados en necesidades específicas para poder volverse generales y obtener un lugar a nivel institucional, y en otros casos los desarrollos han sido demasiado generales para ser útiles a necesidades específicas.

³ Sistema informático creado por Datawatch Corporation que comparte algunas funcionalidades de una hoja de cálculo con algunas facilidades que permiten realizar transformaciones como búsquedas *lookup*, evaluar expresiones y realizar uniones entre conjuntos de datos de fuentes diversas, a fin de generar un conjunto de datos derivados que se puede utilizar para un propósito específico.

Al dar una mirada a los esfuerzos anteriores nos dimos cuenta de que a pesar del importante esfuerzo realizado, las principales razones del poco éxito de estos esfuerzos se debieron a razones que obedecían a procesos y modelos poco organizados y poco adaptables.

Entre las conclusiones más importantes del porqué los esfuerzos anteriores no dieron los resultados esperados se encuentran las siguientes, basados en el análisis y observación que se hizo de estos esfuerzos.

- La premura por generar modelos de datos para entregar resultados específicos y poco trascendentales jugó en contra de la creación de un modelo de datos que facilitara la mantenibilidad en el tiempo.
- Cada proceso de integración se manejó como un caso individual y particular en el que no se llegaron a reconocer los elementos comunes que permiten la estandarización de los procesos de integración y de registro de datos históricos.
- Un exceso de dependencia de las estructuras de datos de los sistemas fuente y la falta de la generación de un modelo desde el punto de vista de negocio hizo que las cargas de datos fueran más réplicas, que procesos de integración de datos y generación de información.
- La falta de un proceso formal para la evaluación y corrección de calidad de datos hizo que los esfuerzos anteriores vieran reducido su impacto y la confianza de los usuarios.
- La falta de una interfaz de visualización eficaz y dinámica para la explotación de la información debilitó el esfuerzo de integración realizado y produjo el abandono de los repositorios de datos y la plataforma que se había generado.
- Las estructuras de *Staging Area* no estaban correctamente definidas o no existían, de manera que se utilizaban en muchos casos simplemente una copia de la información en producción que no contaba con trazabilidad o con procesos de diferenciación de datos adecuados.
- No se mantenían metadatos de ningún tipo de la información que se cargaba a estos modelos de datos.

A través de las observaciones realizadas la propuesta seleccionada como mejor alternativa para la construcción del *Data Warehouse* Corporativo (como se le conoce internamente al proyecto) se encontraron los siguientes detonantes para respaldar la decisión de basar en la arquitectura propuesta por W. Inmon:

- a) El BCR no contaba con un repositorio centralizado en funcionamiento al momento de comenzar con la iniciativa, esto nos permitía comenzar desde cero con una propuesta que incluyera todos los elementos que se encontraban ausentes en las propuestas anteriores, maximizando así nuestra posibilidad de generar una arquitectura propia ajustada a nuestras necesidades y que pudiera contar con los medios necesarios para ser mantenible a través del tiempo.
- b) La propuesta de Inmon si bien se basa en un repositorio centralizado de datos normalizados, incluye explícitamente la posibilidad de trabajar con *data marts*, cuyo

desarrollo puede ser guiado con las características y buenas prácticas de los modelos propuestos por Kimball, mismos que dan cabida a su vez a la implementación de repositorios OLAP, los cuales han demostrado su eficacia como medio de consulta para la generación de valor para el negocio.

- c) La Fábrica de Información Corporativa permite una amplia gama de otras aplicaciones a partir de sus repositorios, como la información operativa a través del ODS, y la minería de datos a través de *Mining Warehouse*; siendo el BCR una empresa de tan importantes proporciones la constitución de esta arquitectura le permitirá el extender saludablemente sus aplicaciones y expandir su utilización para múltiples propósitos.
- d) Los procesos de introducción de nueva información en la CIF son fácilmente orientados a las necesidades del negocio, sin permitir que sean tan específicos que no permitan satisfacer más que la necesidad inicial para la que se diseñaron. La visión de la CIF del negocio mediante un único modelo, contrapuesta a la visión orientada al proceso de Kimball, permite a la empresa descubrirse a sí misma, reconocer sus interrelaciones y consolidar su conocimiento a la vez que facilita la creación de repositorios y herramientas que irradian del *Data Warehouse*.

3. Definición de instrumentos para el análisis.

Debido a que los procesos de recolección de requerimientos implementados en el banco se encontraban específicamente orientados a la creación de aplicaciones transaccionales, inicialmente saltó a la vista que los instrumentos utilizados para la formulación de requerimientos y análisis carecían de los elementos suficientes para poder apoyar un proceso de análisis y diseño orientado a *Datawarehousing* e Inteligencia de Negocios, así nos dimos a la tarea de diseñar y con el tiempo ir mejorando algunos instrumentos que facilitaban la organización de la información requerida por el negocio al punto que pudimos establecer que se requerían específicamente dos instrumentos como mínimo para poder soportar el proceso de manera eficaz, uno para apoyar el proceso de análisis y perfilado de calidad de datos y otro para representar las relaciones entre los datos de los orígenes y destino para apoyar el trabajo del desarrollador de procesos de integración.

Con la introducción de estos instrumentos se trató de mantener al mínimo la documentación y ofrecer una guía para apoyar el proceso calidad e integración de datos delimitando específicamente los elementos a valorar y permitiendo así un marco de referencia ordenado y completo para ambos procesos. Estos pueden ser refinados posteriormente para incluir nueva información que se considere necesario o modificar su funcionamiento para adaptarse de mejor manera al entorno del BCR.

3.1. Instrumento para el mapeo de datos.

El primero de los instrumentos desarrollados, se orienta a cubrir la necesidad de documentar la relación entre los datos origen y los datos del destino, buscando establecer la forma como deben convertirse los datos en el proceso de integración y estableciendo para cada campo destino un origen de datos apropiado.

La fuente que deberá ser preferida para este instrumento es el modelo de datos de sistema de *Data Warehouse*, sin embargo dado la naturaleza variada de las necesidades de integración de datos, puede ser alimentado con una lista de campos solicitados por un área usuaria, en este último caso, la solicitud del usuario por información puede requerir el mapeo de múltiples campos de origen de múltiples sistemas en uno o varios campos del destino.

Hay que tener presente que este instrumento debe limitarse al mapeo de datos a un solo repositorio destino y múltiples destinos requerirán de múltiples instrumentos, esto porque cada instrumento apoyará la creación de un proceso de integración de datos diferente.

3.1.1. Estructura del documento

El documento usado por el instrumento de mapeo de datos se estructura de la siguiente manera:

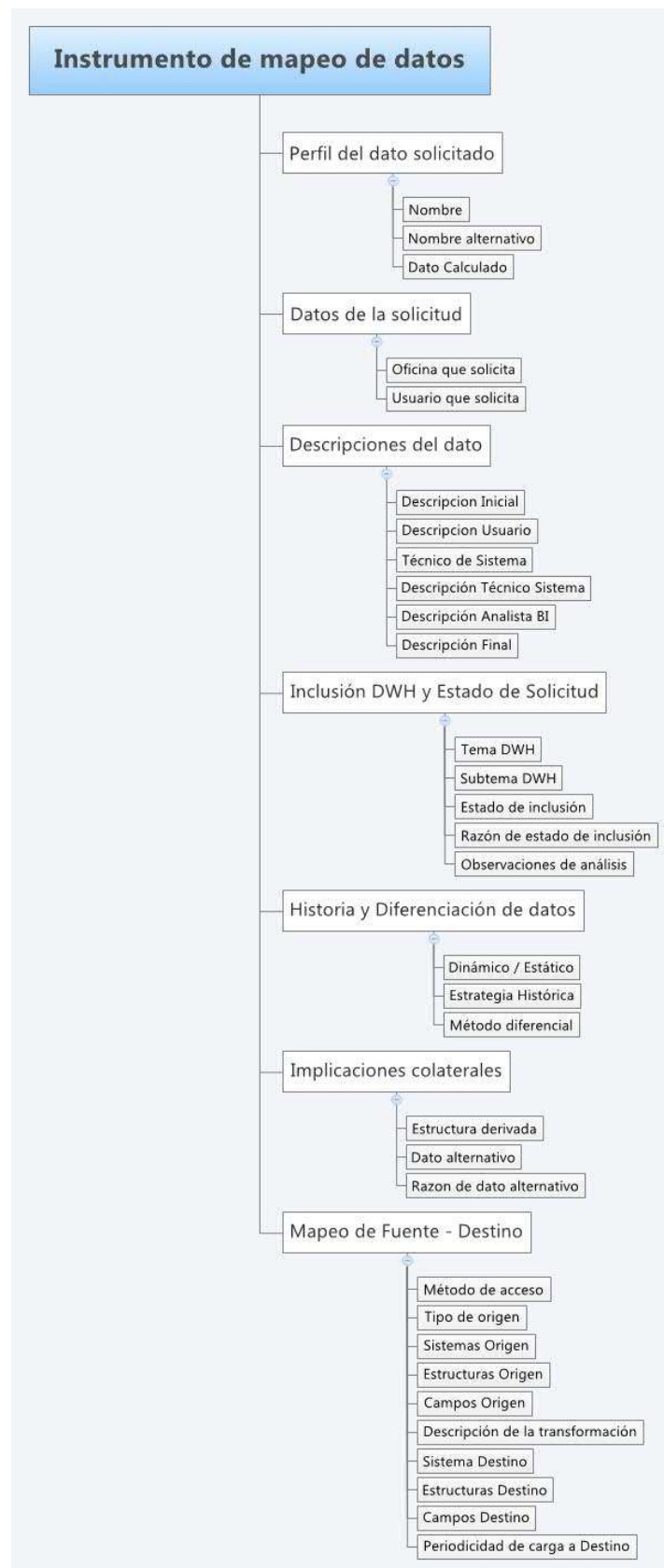


Ilustración 4. Estructura del instrumento de mapeo de datos

La estructura original del documento solo contemplaba una parte de lo que se incluye en la última sección, si bien es cierto algunos campos pueden ser de naturaleza opcional esta estructura en la que se ha evolucionado permite una mejor documentación y un proceso más adaptable a cada iteración en el proceso de análisis.

3.1.2. Elementos del instrumento

a. Perfil del dato solicitado

En esta sección se especifica una manera de identificar la información que solicita el usuario, algunas veces los nombres de los campos que el usuario conoce distan de lo que hay en los sistemas fuente o de lo que realmente significa la información que requieren, este primer paso establece cual es el nombre por el que el usuario reconoce la información y el nombre que el analista en conjunto con el usuario han definido como el adecuado para referirse al concepto que está en el campo que han solicitado, asimismo se incluye un campo para determinar si el dato es un cálculo o si no lo es y por tanto debe ser obtenido de un origen de datos. Los cálculos a veces pueden ser integrados como datos derivados dentro del *Data Warehouse*, o bien pueden ser calculados por la consulta que extrae la información para revisarla posteriormente, esto dependerá de la naturaleza del dato, y de si es importante el momento en que se ejecuta el cálculo, por ejemplo, un indicador de estado de morosidad de un préstamo a fin de mes puede ser pre-calculado, pero el monto dolarizado de una operación al momento de la consulta no puede ser almacenado como dato derivado porque cambiará de un momento a otro conforme cambie el tipo de cambio.

b. Datos de la solicitud

Esta sección determina de donde proviene la solicitud de información, ya que en algunas ocasiones las listas de campos a incluir provienen de áreas diversas es importante saber de qué área y a qué persona pueden pedirse aclaraciones posteriores sobre la información solicitada.

c. Descripciones del dato

Aunque lo ideal es contar con una única descripción sobre la naturaleza del campo, es común que las definiciones del campo dadas originalmente en el requerimiento, la dada por el usuario, el experto del sistema origen y el analista no sean iguales, esta sección se encarga de documentar todas ellas y de establecer una única descripción final que consolida lo aprendido sobre el campo cuando se ha logrado llegar a consenso.

d. Inclusión en DWH y Estado de la Solicitud.

Cuando la solicitud de información es solicitada en un requerimiento y este no proviene de un subconjunto de información obtenida del Modelo de Sistema de *Data Warehouse* se pueden dar casos donde el usuario solicite información que no puede ser obtenida de un sistema origen o puede que también solicite información múltiples veces en la misma lista, inclusive puede ser que solicite información que por el esfuerzo que represente deba ser pospuesta ya que pertenece a un área temática que aún no ha sido reflejada en los modelos de datos.

Así aunque se hace la revisión de esta información en este punto se puede definir si la información es Aceptada para ser incluida en *Data Warehouse*, es pospuesta, es rechazada, o ya está incluida por medio de un proceso de integración anterior.

En esta etapa se establecen:

- La clasificación de los campos solicitados en (Aceptado, Rechazado, Pospuesto, Ya Implementado), y en caso de ser necesario se justifica la decisión tomada.
- El área temática a la que pertenece el campo según el modelo de áreas temáticas.
- La “sub-área” específica dentro de esta área temática, que estará representada en el modelo de negocio.
- Una observación de análisis que corresponde a aclaraciones sobre la naturaleza del campo que puedan ser de utilidad a la hora de generar un proceso de integración.

e. Historia y Diferenciación de Datos.

Esta etapa determina las características del campo que lo hacen tener que ser considerado en el modelo de sistema como un factor que puede generar segregación de datos entre entidades del modelo de negocio. Aunque estos factores no determinan la decisión última de diseño ayudan al analista a sugerir al modelador algunas decisiones de diseño que son visualizadas desde esta etapa.

Los siguientes campos se solicitan en esta sección:

- Dinámico / Estático: Si el analista prevé que este dato tiene importancia histórica deberá marcar el dato como dinámico de lo contrario el dato se considerará estático. Esto permite hacer notar que el analista supone que deberá existir una estructura que almacene cambios en este campo dentro del modelo de sistema de *data warehouse*.
- Estrategia Histórica: Si el analista tiene la visibilidad completa de la entidad que deberá ser modelada, puede proponer una de dos estrategias para mantener datos históricos, mismas que se vieron en una sección anterior: EHS (Entidad Histórica Separada) o DDS (Datos Dinámicos Segregados).
- Método diferencial: Hacer referencia a cómo visualiza el analista que se deberá determinar que ocurrieron cambios en el sistema origen para que estos cambios sean propagados a *Data Warehouse*, esto podría ser por fecha de modificación

del registro, en el caso de que esta información se almacene en el origen o bien mediante la comparación de los valores de una entidad con una copia de una versión anterior de los mismos en *Staging Area* o mediante el cálculo de un *Hash* y una estructura de puente externo (de los puentes externos se comenta con más detalle en una sección posterior).

f. Implicaciones Colaterales.

En esta sección se representan algunos casos particulares que deben ser tomados en cuenta por el modelador, en algunas ocasiones los datos solicitados son en realidad un concepto de datos más detallados que requerirá una estructura aparte o que mediante el análisis se determina que se prefiere un campo diferente de mayor calidad, cuando esto se da se explica en estos campos los efectos colaterales de colocar este campo y podría definirse un dato alternativo a ser cargado en vez del solicitado. Esto se requiere reflejar acá porque es importante que el requerimiento original sea cumplido o justificado de manera adecuada y no podría simplemente cambiarse un dato por otro debido a la necesidad posterior de validar el desarrollo contra lo solicitado.

g. Mapeo de Fuente - Destino.

La parte final de este instrumento establece la manera en la que se relacionan los datos origen y destino y cómo son transformados los unos en los otros, en esta sección se maneja información tal como:

- **Método de acceso:** Consiste en la forma en que los datos serán obtenidos para ser cargados a DWH, aquí hay dos tipos de estrategias “*push*” en la que los datos llegan al alcance de *Staging Area*, como la copia de archivos a un FTP o la réplica de datos desde un sistema origen posterior a un corte diario, o “*pull*” en la que el propio *Staging Area* obtiene los datos, como por ejemplo mediante una consulta directa al sistema de producción.
- **Tipo de origen:** Consiste en tipos de orígenes como archivos planos, bases de datos, colas de mensajería, conexión a un *web service*, entre otros.
- **Sistemas origen:** Nombre que el negocio da al sistema de donde se toman los datos.
- **Estructuras origen:** Nombre de las estructuras origen o archivo de donde se toma la información.
- **Campos Origen:** Nombre de las columnas que serán leídas de la estructura origen con el propósito de crear un campo en el destino.
- **Descripción de la transformación:** Una descripción breve pero lo más completa posible de qué pasos deben llevarse a cabo para convertir los datos de los campos origen y depositarlos en los campos destino tal como los requiere el negocio.
- **Sistema destino:** Sistema hacia donde se cargarán los datos obtenidos del origen después de aplicar la transformación

- **Estructuras destino:** Estructura donde se almacenarán los datos destino.
- **Campos destino:** Campo o campos donde se guardará la información obtenida del proceso de integración
- **Periodicidad de carga a destino:** Definición de cada cuanto deberá obtenerse este dato desde el origen, con el fin de contar con criterios suficientes para proceder a la segregación de datos según los patrones de inserción de los datos.

Este instrumento tiene como finalidad última proveer al desarrollador de los procesos de integración y al diseñador del modelo de sistema los insumos básicos para generar un proceso de carga y un modelo que cumplan con las necesidades del negocio.

3.2. Instrumento para el análisis de calidad de datos.

El segundo de los instrumentos desarrollados está hecho para ser utilizado por TI en colaboración con las áreas de negocio, y es un instrumento facilitador que se debe implementar preferiblemente en un ciclo continuo de calidad, la entrada que requiere este documento es el listado de entidades y campos identificados como orígenes de datos en el documento de mapeo de datos.

Este instrumento podría ser la entrada primaria para la realimentación de un comité de calidad de datos, que agrupe a usuarios de negocio con el conocimiento y la autoridad suficiente de gestionar cambios en los sistemas en producción y en los datos que los mismos almacenan.

En una organización madura en Inteligencia de Negocios este comité está formalmente establecido y se conoce como “*Data Quality Management Comitee*”, entre sus funciones está el revisar las estadísticas de las evaluaciones de calidad de datos e implementar mecanismos de mejora de los mismos de forma continua.

Este tipo de estrategia se desarrolla generalmente dentro del marco de un Programa de Inteligencia de Negocios con el fin de ofrecer dirección a los esfuerzos de calidad de datos, de manera que el comité conoce cuales son los datos que se pretende integrar y les da prioridad. Al existir documentación y estar predefinido el contenido que se va a ingresar a la CIF se le facilita al comité de calidad enfocar sus esfuerzos en los datos más críticos para la toma de decisiones.

3.2.1 Estructura del documento

El instrumento en cuestión, esta implementado en formato de hoja de Microsoft Excel e incluye instrucciones para analizar un repositorio de datos desde tres diferentes puntos de vista, el aspecto técnico, la naturaleza de negocio que tienen los datos y una evaluación desde el punto de vista de siete dimensiones de calidad.

Así dados los datos solicitados en el instrumento se pueden generar gráficos que muestran el estado actual de cada entidad dentro del repositorio desde el punto de vista de las 7 dimensiones de datos, esto permite enfocarse en los temas más problemáticos y trazar líneas de acción para mejorar los aspectos de calidad de datos que más afectan a la organización.

La estructura original del documento es como se muestra a continuación:



Ilustración 5. Estructura del instrumento de análisis de calidad de datos.

Como puede observarse el documento solicita información básica sobre cada campo, para servir a propósitos de documentación, análisis y de administración de la información, y también incluye campos opcionales que facilitan la introducción de explicaciones técnicas y consideraciones a tomar en cuenta para la limpieza de datos, tales como reglas de negocio que hayan sido identificadas durante su elaboración.

Cada entidad a consumir de un sistema origen es representada en el instrumento y este es completado por el analista mediante las observaciones realizadas en la información, este documento puede además ser transcrito para ser almacenado en el repositorio de metadatos de manera que eventuales nuevas aplicaciones del mismo sobre los mismos datos puedan facilitar el seguimiento de los niveles de calidad de datos en la organización.

A continuación un ejemplo de cómo se utiliza el instrumento para calificar una estructura origen en particular:

Ilustración 6. Vista parcial del contenido del instrumento de calidad de datos.

Secciones:

3.2.2. Caracterización técnica de datos

La caracterización técnica de los datos permite a los analistas establecer una línea base de cómo se interpretaba la información que estaba siendo leída desde los sistemas origen cuando se realizó el estudio, esto permite que se puedan visualizar situaciones tales como cambios en las estructuras que dieron origen a un proceso de integración y reaccionar a ellas la próxima vez que se haga el estudio de calidad, esta sección se subdivide en dos partes que se explican a continuación:

- a. **Perfilado estructural:** Corresponden principalmente a las características de forma de los datos que impone el medio de donde estos son obtenidos, en el caso de bases de datos relacionales, la mayoría de estos datos existen en los diccionarios de datos y pueden ser fácilmente obtenidos, en otros están definidos por un conjunto de documentación externa, o una interpretación de los mismos que hace el analista, por ejemplo en el caso de archivos planos la estructura de los datos se define por medio del estudio de los mismos datos y no porque haya una definición explícita del dato que pueda ser consultada.

Dentro del perfilado estructural se documenta:

- i. **Tipo:** Según el medio donde se almacene el dato, tal como int, number, float, char, varchar o varchar2.
 - ii. **Clasificación de datos:** Tales como numérico, alfabético, alfanumérico, complejo, todos ligados al tipo naturalmente.
 - iii. **Longitud:** Cantidad de caracteres o dígitos que soporta el tipo de dato.
 - iv. **Precisión:** Cantidad de dígitos decimales para los tipos numéricos.
 - v. **Permite nulos:** Un indicador con valores Si o No.
 - vi. **Es llave primaria:** Un indicador con valores Si o No que indica si el campo es parte de la llave primaria implementada en la entidad en el sistema origen.
 - vii. **Es identificador:** Un indicador con valores Si o No que indica si el campo puede usarse para identificar de manera única un registro en el origen de datos (no necesariamente parte de la llave primaria).
 - viii. **Implementa llave foránea:** Determina si el campo es una llave foránea de otra entidad.
 - ix. **Rango posible:** Valores máximo y mínimo que puede tomar el campo según el tipo de dato particular del sistema fuente.
- b. **Perfilado de contenido:** Este perfilado permite la valoración de los datos almacenados en cada registro, representa la información sobre el mismo que está relacionada al contexto del dato y no solo a sus características físicas, por ejemplo puede que un campo tenga un rango posible de valores mayor al que tendrá en la práctica o que guarde alguna relación tácita con otros datos, estas

características se documentan en esta sección formada por los siguientes campos:

- i. **Valores permitidos:** Conjunto de valores que se consideran válidos para el campo dada su naturaleza, independientemente del rango de valores que pueda contener.
- ii. **Unicidad:** Porcentaje de la muestra total de datos en que son únicos los valores del campo, un valor que sea identificador o llave primaria no compuesta debería tener un valor del 100%, por otra parte un valor que indica una característica reflejada en una tabla catálogo pequeña, podría tener un valor de unicidad muy pequeño menor al 0.01%
- iii. **Porcentaje de nulidad:** En qué porcentaje los valores del campo en cuestión son nulos o tienen un valor que se puede considerar que es equivalente a nulo.
- iv. **Dependencias con otros campos y tablas:** Esta es una descripción que aclara la relación de un campo con otros elementos de datos, esta puede ser desde una relación de llave foránea o una correspondencia particular como que el campo debe tener un conjunto de valores dependiendo del valor de otro campo en la misma estructura.
- v. **Formato:** Este valor representa el formato del campo de acuerdo a como lo especifica el negocio por ejemplo una cédula nacional podría tener el formato 9-9999-9999, un código postal el formato 9-99-99 y una fecha podría tener el formato DD/MM/YYYY o YYYYMMDD.

3.2.3. Dimensiones de datos de negocio

La sección de dimensiones de datos del negocio termina de enumerar las características de los datos que son relacionadas a su papel dentro de la organización, así esta sección está compuesta los siguientes elementos:

a. Semántica del campo:

- i. **Significado:** Explica ampliamente cómo se interpreta la información contenida en el campo, desde el punto de vista de su utilidad para el negocio.
- ii. **Reglas de negocio aplicables al campo:** Esta sección documenta la forma en que deben interpretarse algunos datos, o cómo deben limpiarse en caso de tener problemas de calidad, donde se pueda aplicar directamente limpieza.

b. Consideraciones adicionales:

- i. **Comentarios:** Este apartado permite agregar información relacionada con el campo que por su naturaleza es importante pero no cae dentro de ninguno de los campos anteriores, por ejemplo una referencia a una fuente externa, a

una normativa particular con el que se relaciona o un documento de análisis adicional que deba ser consultado.

3.2.4. Dimensiones de calidad de datos

Para completar esta sección el analista debe valerse de la información sobre los datos que pueda determinar por sus medios a partir de la muestra de datos de que disponga y de su valoración objetiva con el fin de establecer calificaciones sobre cada una de las siete dimensiones de datos incluidas en el documento. Al realizar este ejercicio sobre toda la información que deberá ser consumida por los procesos de integración se puede obtener una calificación muy importante sobre la calidad de los datos que provienen una fuente particular.

Dado que el tema de dimensiones de calidad de datos no es ampliamente conocido, se incluyeron criterios por escrito que puedan ser evaluados por el analista para estandarizar el análisis. Así damos uno o varios criterios que deben ser calificados y estas calificaciones son ponderadas mediante un peso para calificar cada campo desde el punto de vista de las 7 dimensiones. A continuación se muestra un ejemplo de cómo funciona esta ponderación y se presentan los criterios que se dan en el documento para consideración del analista. Estas dimensiones obtienen un puntaje mediante observación y muestreo de los datos y se complementan con información obtenida del negocio y de TI.

Validez				
Grado en el que los datos cumplen con las reglas aplicables para ser considerados útiles.				
Una columna tiene datos válidos si:				
A. Cumple con las reglas de negocio definidas(16)	B. No tiene valores fuera del conjunto de valores permitidos (10)	C. Tiene un formato que corresponde adecuadamente al definido(14)	TOTAL	
60%	20%	20%	100%	
Columna19	Columna20	Columna21	Columna22	
100.00%	100.00%	100.00%	100.00%	
30.09%	30.09%	30.09%	30.09%	
30.79%	30.79%	30.79%	30.79%	
53.63%	53.63%	53.63%	53.63%	

Ilustración 7. Vista parcial de los resultados de las mediciones en el instrumento de calidad de datos

Dimensión de calidad de datos	Definición de la dimensión	Criterio	Ponderación
Validez	Grado en el que los datos cumplen con las reglas aplicables para ser considerados útiles. Una columna tiene datos válidos si:	Cumple con las reglas de negocio definidas.	60
		No tiene valores fuera del conjunto de valores permitidos.	20
		Tiene un formato que corresponde adecuadamente al definido.	20
Unicidad	Existencia de datos repetidos dentro de una misma columna. Una columna tiene unicidad adecuada si:	Si el campo no requiere ser único este indicador es 100% de lo contrario se pone el porcentaje de unicidad	100
Compleitud	Existencia y cumplimiento de los criterios de integridad de un dato que es requerido. Una columna tiene datos completos si:	(Desde la perspectiva del registro) Existe para cada registro un valor para el campo. (Un nulo puede ser adecuado si el contexto lo permite de esta forma, completitud es diferente y no es el complemento de Nulidad)	50
		(Desde la perspectiva de la celda/valor) Los valores de datos tienen toda la información que se requiere (Ejemplos de incompleitud: Teléfono de 7 dígitos, cédula de 8 dígitos,etc.)	50
Actualidad	Propiedad de un dato que lo define como la versión más actualizada del mismo. Un campo es actual si:	Para todas las fuentes disponibles desde donde se puede obtener el valor del dato, no se puede encontrar una versión más actualizada que represente el valor del mismo para el momento en que la información se capturó. (En el caso de catálogos de fuentes externas se puede apoyar en análisis de tablas cruzadas para verificar la actualidad de un dato)	100
Exactitud	Grado en que los datos se aproximan a lo que representan en la realidad. Un campo es exacto si:	Porcentaje de datos encontrados que corresponden con datos reales según se pueda verificar con una referencia adicional o externa al sistema.	100
Precisión	Grado en que el dato aporta el nivel de detalle requerido. Un campo es preciso si:	El nivel de detalle aportado por el dato es suficiente para cumplir de la mejor manera con el propósito para el que se definió.	100

3.2.5. Salidas del instrumento

Al finalizar el análisis basado en el instrumento la segunda hoja del documento nos muestra los resultados del análisis de forma gráfica para poder utilizarlo para priorizar las acciones correctivas necesarias.

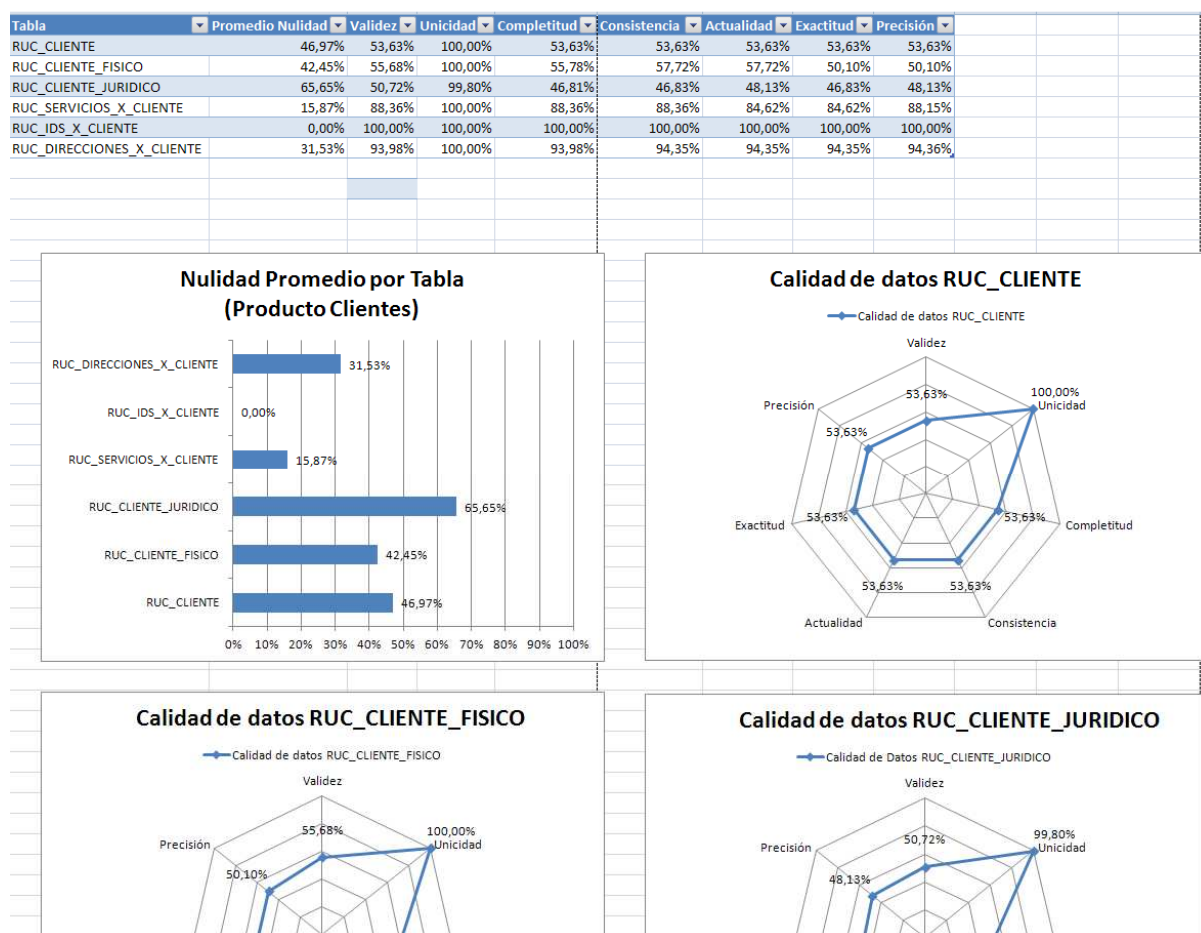


Ilustración 8. Resultados obtenidos de la aplicación del instrumento de calidad de datos.

Dado que algunas de las observaciones para la medición de las dimensiones de calidad pueden basarse en apreciaciones y en datos cualitativos o no documentados, es recomendable que las calificaciones puedan ser además complementadas con el

conocimiento que las originó, de manera que se justifique el motivo de cada una cuando este no esté suficientemente explícito.

La utilidad de este instrumento no solo radica en la posibilidad de hacer las mediciones sino que brinda otros muchos beneficios tales como:

- Mejora el desarrollo de procesos de integración que requieren conocer los aspectos de calidad para hacer ajustes adicionales.
- Permite al negocio ajustar sus definiciones de calidad según su comprensión de calidad de datos y su nivel de madurez.
- Permite enfocar específicamente los esfuerzos de calidad de datos a los problemas y áreas más relevantes.
- Permite hacer comparativas entre estudios hechos en momentos distintos, con lo cual se puede dar seguimiento al avance de los procesos de mejora de la calidad de datos.

Los instrumentos revisados hasta este punto son producción del equipo de *Data Warehouse* del BCR y son utilizados como un paso más en el proceso previo a la generación de los mapas de integración de datos que llevan información a *Data Warehouse*, estos han ido evolucionando con cada vez que se utilizan y deberán continuar haciéndolo en favor de mayor practicidad y mayor recolección de conocimiento. Se consideran pasos adicionales para el futuro la formalización de estos como procedimientos de análisis normados y la inclusión de la información que contienen dentro del repositorio de Metadatos.

4. Análisis y diseño de componentes de la CIF.

El proceso de consolidación del *Data Warehouse* Corporativo del BCR comenzó prácticamente desde cero, aunque muchos de los conocimientos brindados por los compañeros que habían trabajado en algunos procesos de integración de información fueron insumos imprescindibles para el desarrollo adecuado de las primeras versiones.

Aunque contábamos con algún conocimiento teórico previo sobre los procedimientos, los primeros pasos en calidad e integración de datos generaron muchísimas necesidades de estandarización y requirieron de un tiempo considerable para realizar un análisis adecuado, conforme se fue avanzando muchas de estas necesidades dieron origen a un conjunto de técnicas y metodologías propias las cuales son ahora estándares para la generación de procesos de integración y calidad de datos en el *Data Warehouse* Corporativo, tales como los instrumentos de calidad de datos y de mapeo que vimos anteriormente.

Para poder comprender las razones de la implementación de estas técnicas dentro de los procesos de *Data Warehouse* Corporativo, se requiere caracterizar en un nivel más detallado algunos de los componentes a los que ya nos hemos referido, así, esta sección del documento

se estructura en componentes, los cuales corresponden a los elementos actualmente implementados.

Al momento de redacción de este documento la Fábrica de Información Corporativa del BCR consta de tres elementos principales:

- *Data Warehouse* Corporativo.
- Área de trabajo (*Staging Area*).
- Repositorio de Metadatos.

Data Warehouse Corporativo

Este es el componente central de la Fábrica de Información Corporativa, los demás repositorios irradian de este repositorio o sirven para apoyar el proceso de llenado del mismo. En la sección correspondiente a el Modelo de Sistema, ya se han definido las estructuras que conforman este repositorio, y las consideraciones necesarias para definir cuáles de estas estructuras son necesarias para cada fin particular ya han sido revisadas dentro del análisis de los modelos de datos hecho en la sección de Caracterización de Opciones referente a la Fábrica de Información Corporativa, no las repasaremos en esta sección sin embargo es importante recordar qué tipo de estructuras existen dentro de este repositorio:

las cuales incluyen:

1. Estructuras de datos maestros o tablas principales.
2. Estructuras de catálogos simples.
3. Estructuras de datos dinámicos (DDS y EHS).
4. Estructuras de arreglos de datos.
5. Estructuras de sumarización.
6. Dimensiones conformadas para el consumo de los *data marts*.
7. Estructuras mezcladas para facilitar la consulta.

Los modelos de datos implementados en el repositorio de *Data Warehouse* Corporativo tendrán que hacer uso de ciertas de estas estructuras para soportar los procesos de integración para cada área temática según lo dicten los requerimientos del negocio.

Área de trabajo (*Staging Area*)

Toda fábrica tiene necesidad de insumos para poder producir, no es un error entonces suponer que la Fábrica de Información Corporativa requiera de materia prima con la cual generar productos de valor. El punto de entrada por excelencia para la CIF es el área de trabajo o más comúnmente conocida por su nombre en inglés "*Staging Area*", aunque podría tenderse a confundir con un simple medio o repositorio para recepción y acumulación de información, la experiencia que hemos tenido nos muestra que invertir tiempo y recursos en un diseño adecuado de *Staging Area* es invertir tiempo en consolidación de la CIF, después de nuestros

análisis y la generación de los primeros procesos de integración, podemos definir entre las contribuciones más importantes del *Staging Area* las siguientes:

1. Las estructuras de datos de *Staging Area* permiten orientar el diseño de los procesos de integración a través de los conceptos de negocio.
2. Permite resumir cargas que por alguna razón hayan tenido una falla sin necesidad de repetir la misma consulta contra ambientes de producción.
3. Permite llevar seguimiento adecuado del flujo de los datos: por ejemplo, obtener estadísticas de cuántos registros han sido cargados a una determinada sección de la CIF.
4. Permite unificar en una sola estructura de datos el concepto del negocio que se representará de forma unificada al obtener de múltiples sistemas heterogéneos información sobre el mismo tipo de entidad.
5. Permite separar adecuadamente la lógica entre extracción y carga de datos facilitando a los procesos de integración de datos una característica de modularidad que permite realizar mejoras e inclusiones de nuevos procesos de manera más sencilla.
6. Centraliza componentes facilitando la reutilización de los mismos a través de varios procesos de integración de datos.

Un *Staging Area* adecuadamente diseñado permite canalizar la información hacia un concepto de negocio tan pronto esta hace su ingreso a la CIF.

Según [GIORDANO, 2011] la clave para tener procesos de integración flexibles, es dividir la funcionalidad en componentes reusables y discretos. Uno de los principios fundamentales utilizados para incrementar la modularidad y flexibilidad en el diseño es encapsular ambos, datos y funciones, en las capas de *Staging* usando formatos de archivos comunes guiados por las áreas temáticas del modelo de datos destino. Estos formatos consisten en un archivo o tabla basada en un concepto de negocio dentro de cada sección de las capas de *Staging*, mediante estos “archivos comunes” se garantiza que los datos sean modelados por los conceptos de negocio en vez de por los sistemas que originan los datos.

La utilización de un *Staging Area* basado en copias de las estructuras de datos de los sistemas de producción (fuentes de datos) se encontró presente en las primeras versiones de *Data Warehouse* que se hicieron en el BCR. Algunos de los problemas más importantes identificados al seguir esta estrategia fueron:

- **Inflexibilidad para adaptarse a los cambios:** Un cambio en una tabla, en el tipo de datos de una columna, el nombre de una columna o tabla, la cantidad de campos que tenga la tabla, implicará irremediamente cambios en el *Staging Area* para poder reflejar los cambios del exterior, por lo que generalmente, un proceso de carga que sea generado para alimentar estas estructuras, deberá ser afectado consecuentemente para reflejar los cambios.
- **Dependencia de la fuente:** Es normal que con el tiempo los sistemas sean cambiados por otros más modernos, generalmente aunque los procesos de negocio que soportan los nuevos sistemas son los mismos las estructuras de datos utilizadas raramente son las mismas, así un proceso de carga de datos que “copie” la estructura de las tablas del sistema de producción está condenado a desaparecer y tener que ser reinventado para ajustarse al nuevo sistema, el problema se vuelve más complejo al tener que coordinar los pases simultáneos que implica la aplicación del cambio en el ambiente de la CIF y la entrada del nuevo sistema en producción, sobretodo en ambientes tan controlados como el que representa la banca estatal.
- **Dificultad de administración:** Muchos diferentes sistemas que almacenan un mismo concepto de negocios implicarían muchas más estructuras que administrar a nivel de *Staging Area*.

Principios de Diseño del Modelo de Staging Area

Vemos hasta acá que la dependencia con las fuentes de datos es un problema en sí, y que nuestra propia estrategia debe buscar el menor acoplamiento posible entre los componentes de la CIF y las estructuras del exterior. Fue nuestra decisión al revisar el tema que en el caso del BCR las estructuras de datos que reciban la información de las fuentes deberán de guardar poca o ninguna relación con la estructura de las mismas.

Esto puede sonar contradictorio en un principio, pero según [GIORDANO, 2011] uno de los más fundamentales principios de diseño para un proceso de integración de datos es el diseñar para el destino de los datos; si además agregamos a esto el que los destinos de datos son generalmente repositorios como el *Data Warehouse* Corporativo, salta a la luz la línea base sobre la que debemos diseñar el área de trabajo, deberemos trabajar hacia atrás desde el destino hasta el origen de los datos para asegurar que la influencia de las estructuras origen no guíe la construcción del área de trabajo.

El modelo de datos de sistema de *Data Warehouse* sobre el que trabajamos anteriormente tiene poca influencia del sistema origen y una fuerte influencia del negocio, es por esto que consideramos como lo más recomendable el generar a partir del modelo de sistema, un modelo de *Staging Area* que permita acumular la información de los sistemas fuente en

estructuras que se encuentren alineadas a el concepto de negocio que busca el modelo de sistema.

El resultado de este proceso es que obtuvimos estructuras de trabajo que dependen poco o nada de la fuente de datos pero que están íntimamente ligadas a un concepto o proceso de negocios y que por tanto son más durables, y menos susceptibles a los cambios del exterior transaccional.

Podríamos entonces proponer el establecimiento de una regla general, “el diseño del área de trabajo nunca precede a la del repositorio destino”, y dado que el diseño de este repositorio no se hace con base en el contenido de un sistema fuente, sino bajo la comprensión adecuada de un concepto o proceso de negocios y sus relaciones con otros procesos de negocios dentro del modelo de sistema, se asegura que toda “la maquinaria” del *Data Warehouse* se mantenga alineada a la necesidad de negocio.

Debido a que estamos siguiendo la metodología de Inmon para la construcción de *Data Warehouses*, puede que nuestro modelo de datos destino sea demasiado normalizado para ser práctico a nivel de área de trabajo, así que es permitido y deseable el utilizar estructuras menos normalizadas dentro de *Staging Area*, que permitan un movimiento más fluido de los datos a través del proceso de carga, de forma que los datos vayan mutando de manera más natural desde el origen, dentro de *Staging Area* y finalmente en el *Data Warehouse*.

Por ejemplo, un registro de clientes cuya información histórica haya sido segregada dentro del modelo de sistema en una estructura principal y una DDS, no requerirá de ambas estructuras a nivel de *Staging Area*, pero si requerirá de información que indique si se está cargando el registro por un cambio en los datos de la estructura principal o en la DDS, para que el proceso de integración pueda afectar adecuadamente las estructuras correspondientes.

En el ejemplo siguiente se muestra un caso sencillo de un proceso de integración basado en la estructura del sistema origen. En esta caso se tiene un modelo de sistema de *Data Warehouse*, que ha determinado la necesidad de una estructura de clientes, la cual está conformada por datos simples como la identificación del cliente, nombre y apellidos, fecha de nacimiento, oficina a la que está asignado, y la dirección principal.

Por otra parte el sistema de Cuentas Corrientes (origen) tiene dos estructuras en las que se encuentra aproximadamente la misma información pero dividida en dos tablas, dado que pueden haber varias direcciones, y solo una de ellas es la principal.

El *Staging Area* se ha diseñado basada en la estructura del sistema de Cuentas Corrientes por esto utiliza estructuras muy semejantes a las del sistema origen, la única variación es que la estructura de direcciones no incluye el campo *Es_Principal* porque aplica un filtro a los datos para solo traer a *Staging Area* las direcciones principales.

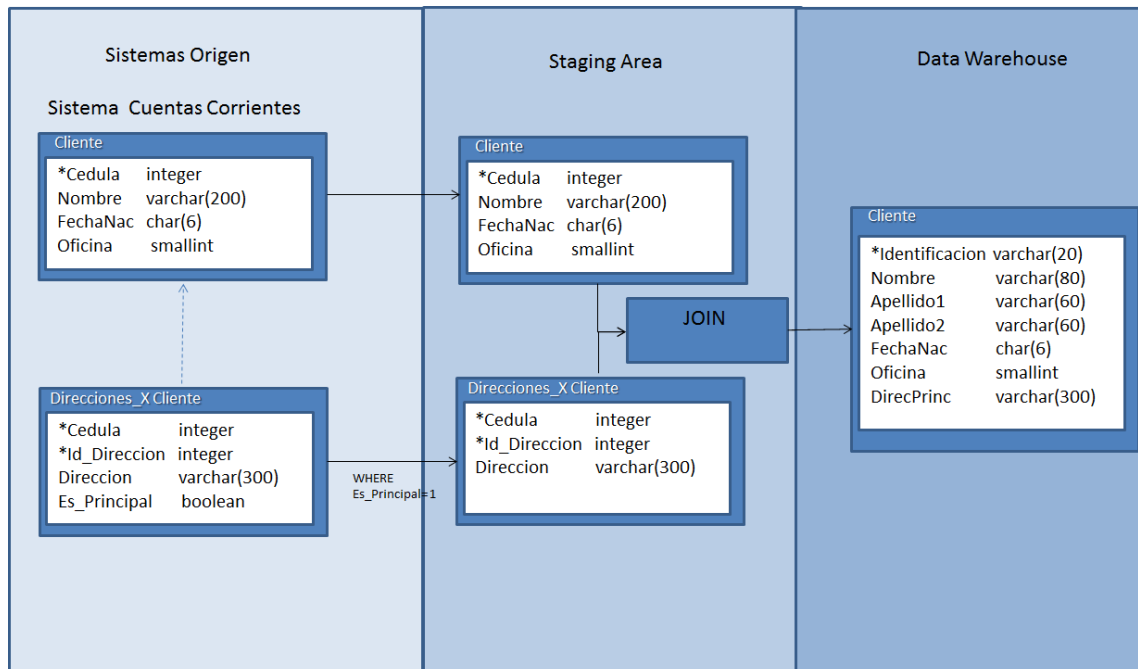
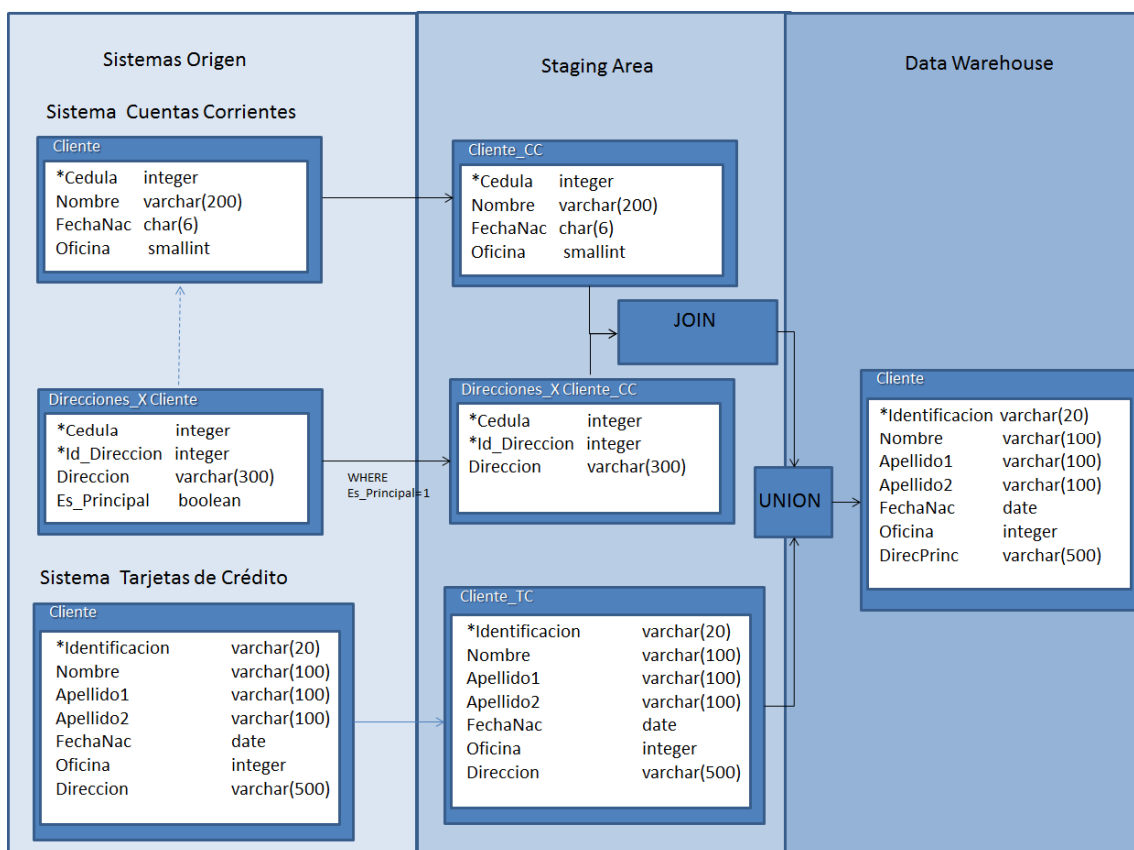


Ilustración 9. Ejemplo de estructuras de *Staging Area* diseñadas con base en los sistemas origen.

Aunque el ejemplo está simplificado, este esquema funciona adecuadamente mientras se sigan dando las condiciones actuales, aunque podríamos hacer algunas observaciones en este momento desde el punto de vista del nivel de recursos utilizados y el rendimiento, se han ejecutado dos consultas diferentes con lo que se han abierto dos conexiones a base de datos y se han hecho pasar dos flujos de datos por la red, se han incluido campos innecesarios como la cédula de la tabla de direcciones (que se requiere para el JOIN) y el id_direccion (que no es necesario del todo), se han definido como llaves primarias en *Staging Area* los mismos campos que estaban en el origen, por lo tanto se han actualizado índices en *Staging Area* al momento de inserción, finalmente se han llenado dos estructuras diferentes para realizar el proceso de carga y realizado dos COMMIT, el JOIN entre ambas para producir el resultado que necesita el *Data Warehouse* se hizo a nivel del repositorio de *Staging*.

El problema con la aproximación usada va más allá de rendimiento, el verdadero problema es lo que sucede cuando entra en el proceso un nuevo sistema y trata de aportar información al flujo que anteriormente había sido definido basado en la estructura del sistema de Cuentas corrientes.



Siguiendo la misma estrategia, de basar nuestro diseño de *Staging Area* en las estructuras de los sistemas origen, han ocurrido algunos cambios y se han agregado nuevos pasos al proceso, primeramente fue necesario renombrar las tablas de *Staging Area*, para poder diferenciar entre los Clientes de uno y otro sistema, la estructura de Cliente en *Data Warehouse* ha tenido que alterarse para poder soportar la información incluida en el nuevo sistema, que tiene mayor espacio para nombres apellidos y dirección, en el sistema de tarjetas no hay necesidad de hacer un JOIN para obtener la dirección principal, pero debido a que ahora tenemos un sistema nuevo, deberemos realizar una operación de UNION para poder incorporar juntos datos de ambos sistemas en *Staging Area*, adicionalmente dado que los tipos de datos son distintos esta unión implicará algunas conversiones de datos por ejemplo para convertir la fecha de nacimiento de un char(6) a un tipo fecha, así mismo dado que nuestra estructura de *Data Warehouse* tuvo que alterarse nos veremos en la necesidad de hacer una migración de los datos de la estructura anterior a una nueva o de hacer las modificaciones a la tabla original, ambas operaciones costosas y riesgosas desde el punto de vista de recursos computacionales.

Ahora, para poder establecer la comparación veamos el caso de un *Staging Area* diseñado en base a un modelo de sistema de *Data Warehouse*, y orientado a un concepto de negocio reflejado en un destino de datos.

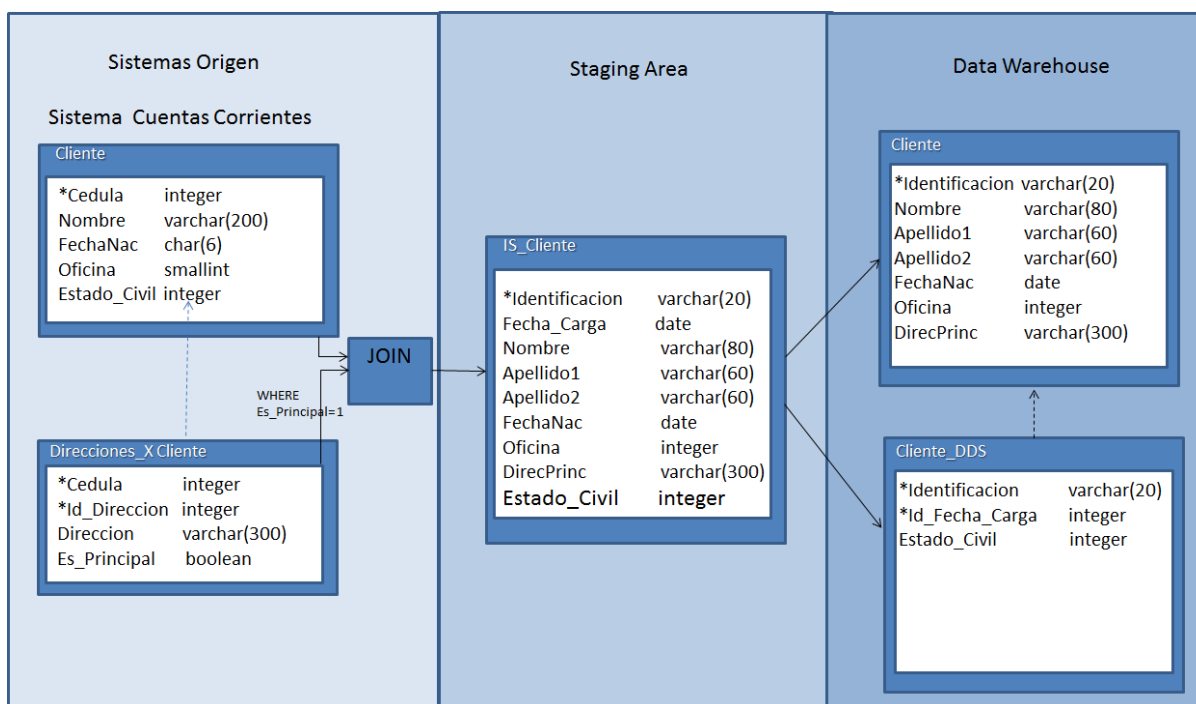
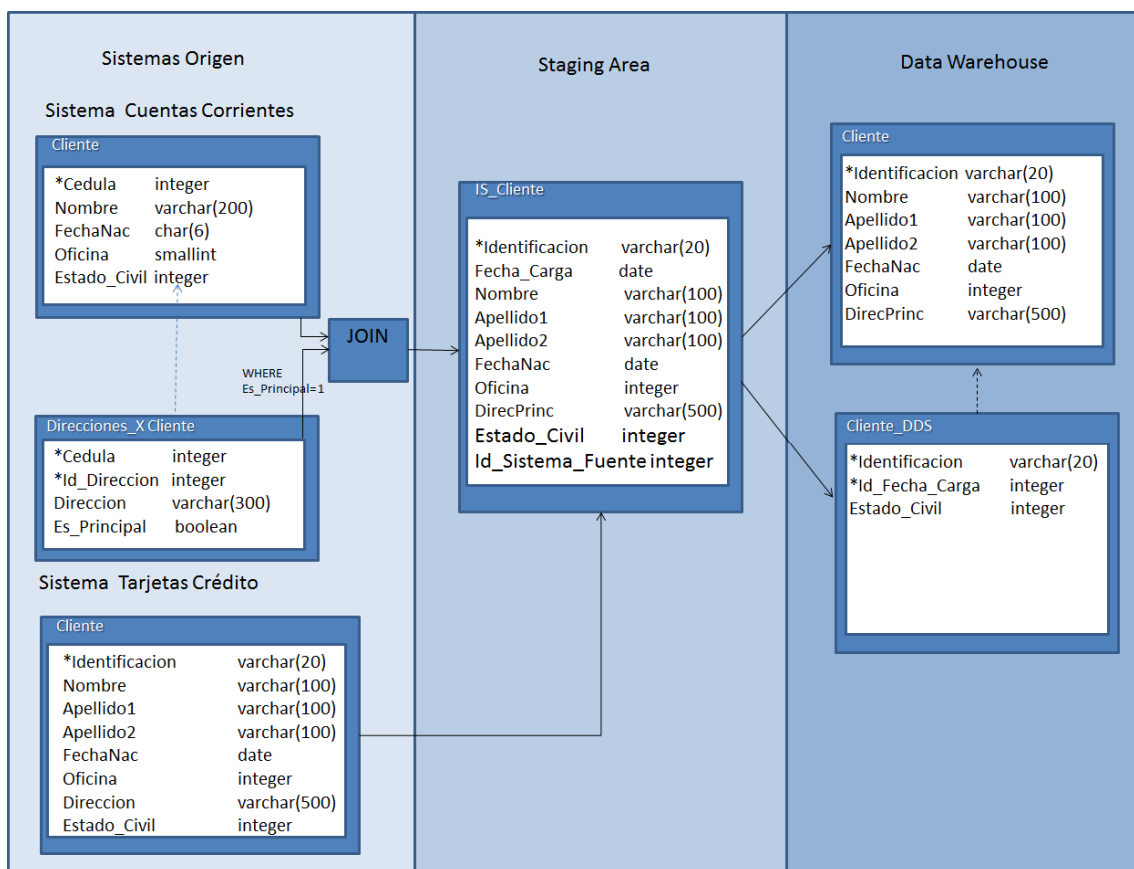


Ilustración 10. Ejemplo de estructuras de *Staging Area* diseñadas orientadas a un concepto de negocio reflejado en un destino de datos.

En este caso para poder demostrar que aunque se orienta al destino, esta aproximación puede generar en *Staging Area* modelos de datos diferentes a los de *Data Warehouse* hemos incluido un dato más dentro del flujo, el estado civil, el cual requiere ser almacenado de manera histórica y por tanto generó dentro de *Data Warehouse* una estructura tipo DDS, el proceso de carga deberá determinar si hubo un cambio en esta característica y de ser así almacenar la nueva versión en la tabla DDS.

La única estructura que se mantiene en *Staging Area* es antecedida por el prefijo "IS", este indica que la tabla es una estructura de tipo "Initial Staging", este término se revisa en la próxima sección.

Antes de proceder a analizar el ejemplo, veamos cómo sería la situación cuando entre en el proceso un nuevo sistema origen.



Podemos observar que el JOIN original entre las tablas de Clientes y Direcciones de Cuenta Corriente, ahora corre por cuenta del sistema origen, al igual que la condición WHERE, ya no se lleva información innecesaria a *Staging Area*, parte del trabajo es delegado a los sistemas origen (aunque hay que tener cuidado de bloqueos a tablas o registros si se realiza integración de datos en horas laborales) y las consultas que se realizan sobre ellos estandarizan la información al requerimiento de *Staging Area*, cualquier otro sistema que requiera participar del proceso, deberá crear un proceso independiente que escriba de manera estándar en la tabla IS.

En síntesis la responsabilidad de cómo se deben obtener los datos para pasar a *Staging* es delegada a los sistemas origen, o más bien a la consulta que extrae de ellos la información, de esta manera si esta forma cambia no afectará la lógica de carga a *Data Warehouse* y no habrá afectación en los demás sistemas. En esta aproximación cada sistema se mantiene aislado y la lógica de integración de *Staging Area* a *Data Warehouse* se limita a los cambios mínimos, mismos que con un adecuado proceso de análisis aplicado a los modelos de sistema debe mantenerse en muy bajos niveles de impacto.

Cabe aclarar acá que en el esquema de múltiples sistemas de entrada con un *Staging Area* orientado al destino, a veces se requiere proceder de manera diferente para un registro dependiendo de dónde se obtiene, para esto se agrega un indicador de *Id_Sistema_Fuente*

que permita mantener la lógica de *Staging Area* lo más general posible y que aun así pueda responder a las particularidades de cada origen de datos.

Vemos así como el orientar los procesos de integración al destino, es la aproximación más mantenible y flexible para el diseño del modelo de *Staging Area* y consecuentemente también para los procesos de integración de datos.

Capas de Sistema de Integración de *Staging Area*

IBM ha sido siempre una compañía referente en temas de arquitectura tecnológica, sus propuestas han sido ampliamente probadas y son reconocidas por muchos de los más importantes productores de software a nivel mundial, es así como hemos tomado la propuesta de [GEORDIANO, 2011], la cual está basada en la arquitectura de referencia de inteligencia de negocios de IBM y hemos rescatado de la misma un conjunto de buenas prácticas para ser aplicadas en los procesos de integración de datos de *Data Warehouse* Corporativo del BCR.

La arquitectura de referencia de inteligencia de negocios de IBM, toma en cuenta todos los elementos que intervienen en la implementación de un proceso de inteligencia de negocios en la organización, desde el acceso a los datos, las fuentes de datos, los repositorios especializados, las aplicaciones y procesos analíticos y los procesos de integración de datos. Estos últimos son a su vez subdivididos en subelementos que según su propuesta deben estar presentes de una u otra manera en cualquier organización que trabaje con integración de datos.

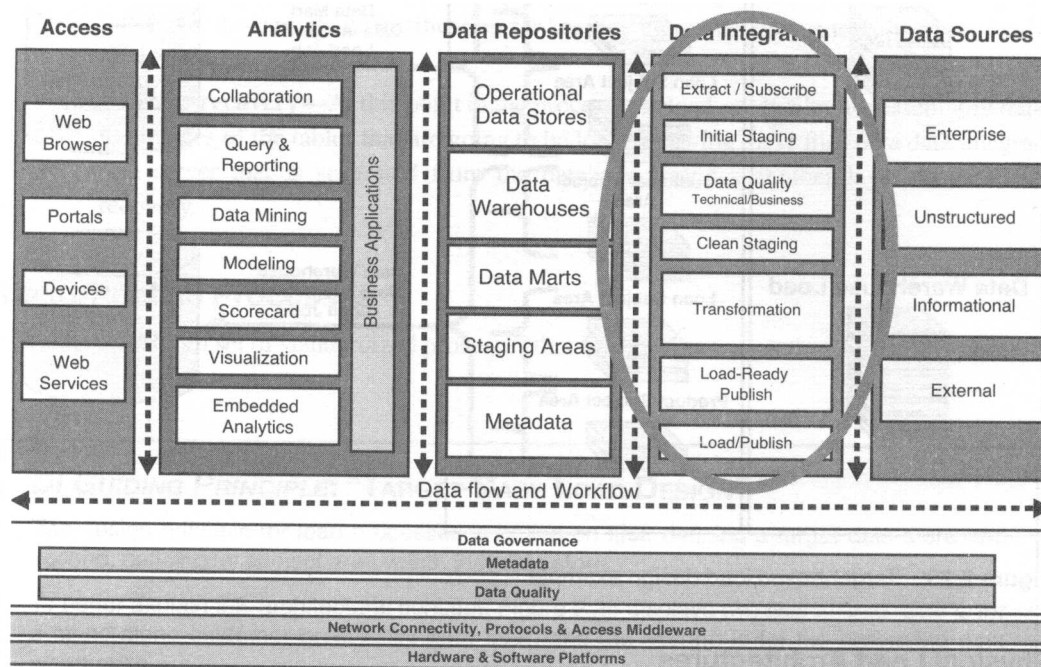
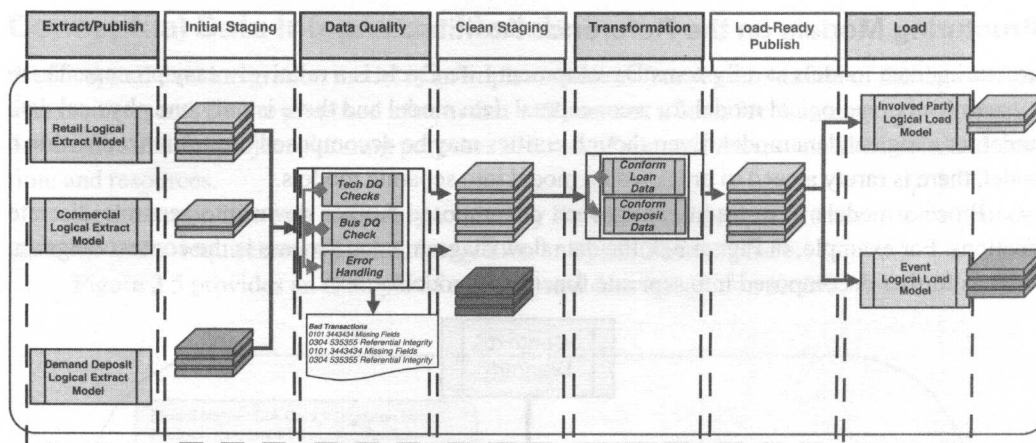


Ilustración 11. Estrategia de Inteligencia de Negocios de IBM. Tomada de [GIORDIANO, 2011] .

En la propuesta de [GEORDIANO, 2011] el *Staging Area* tiene una arquitectura en capas que refleja los elementos de la arquitectura de referencia de inteligencia de negocios de IBM, así existen 7 capas en el *Staging Area* que deberían considerarse en procesos de integración maduros. Es importante hacer notar que de las siete capas, tres representan capas de almacenamiento de datos y cuatro son procesos.



1. **Extract/Subscribe (Proceso de extracción o suscripción):** Es un conjunto de procesos que toman información de una o múltiples fuentes o están suscritos a un servicio que les provee los datos, estos procesos toman la información y los escriben en las tablas de entrada de *Staging Area*, estos procesos son análogos a los que se representaron en la sección anterior y para nuestros efectos están formados por uno o varios procesos de carga que alimentan el mismo destino dentro de *Staging Area*.
2. **Initial Staging (Almacenamiento Inicial):** Las estructuras iniciales donde se almacena la información obtenida de los orígenes de datos son las tablas de "*Initial Staging*", en un ejemplo anterior nos referimos a ellas, y se enfatizó que su definición se hace desde lo definido en las tablas de destino.
3. **Data Quality (Proceso de Calidad de Datos):** Este proceso se encarga de realizar dos conjuntos de validaciones de calidad, una desde el punto de vista técnico, y otra desde el punto de vista de reglas de negocio, este proceso incluye la lógica necesaria para el proceso pueda manejar adecuadamente errores, incluye instrucciones para limpieza de datos donde es posible y genera tres salidas, los datos limpios, los datos rechazados y un reporte de rechazos. La entrada propuesta para la generación de este proceso es el instrumento de calidad de datos que se revisó en la anteriormente.

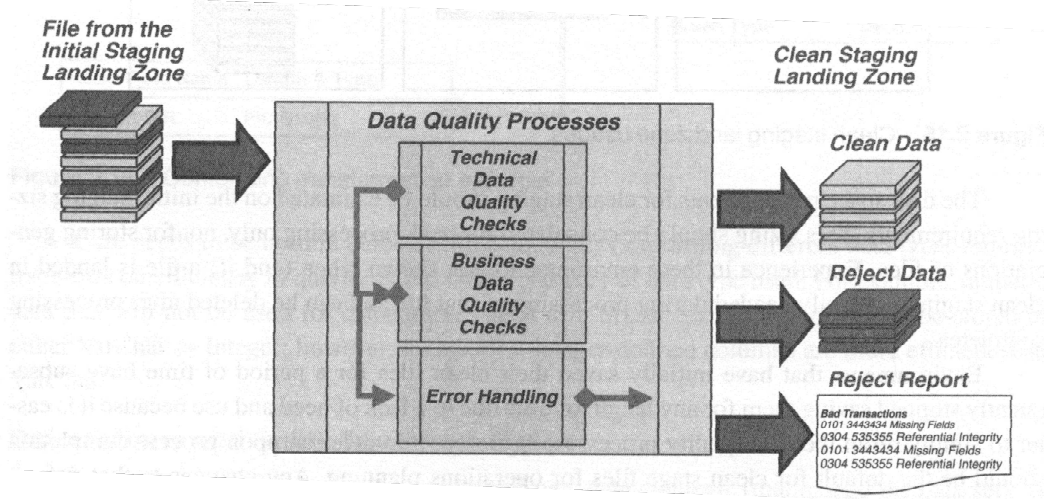


Ilustración 12. Proceso de calidad de datos dentro de los procesos de integración. Tomado de [GIORDIANO, 2011].

4. **Clean Staging (Almacenamiento de limpieza):** Es un espacio de almacenamiento en donde se almacena la información de los datos limpios, los rechazados y los reportes de rechazos, algunos datos no saldrán de esta etapa hasta tener en claro la forma de cómo se procederá con ellos. En caso de que no se requieran transformaciones adicionales para la información, los datos podrían ser escritos directamente a la capa 6 (Almacenamiento / Publicación de Datos Listos para Cargar) o pasar de manera normal a la siguiente etapa.

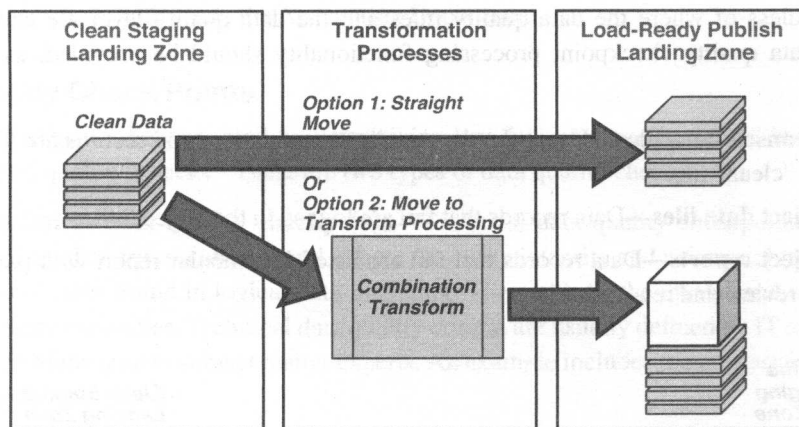


Ilustración 13. Movimiento de datos entre áreas de limpieza y la zona de publicación. Tomada de [GIORDIANO, 2011]

5. **Transformation (Proceso de Transformación de datos):** El proceso de transformación de datos tomará la información grabada como datos limpios del área de almacenamiento de limpieza y le aplicará las transformaciones necesarias para convertir estos datos a lo que se requiere en el destino de datos. Al igual que en el caso anterior puede que los datos transformados estén listos para cargarse al destino, en ese caso, se puede proceder directamente a la capa 7 sin pasar al almacenamiento final.

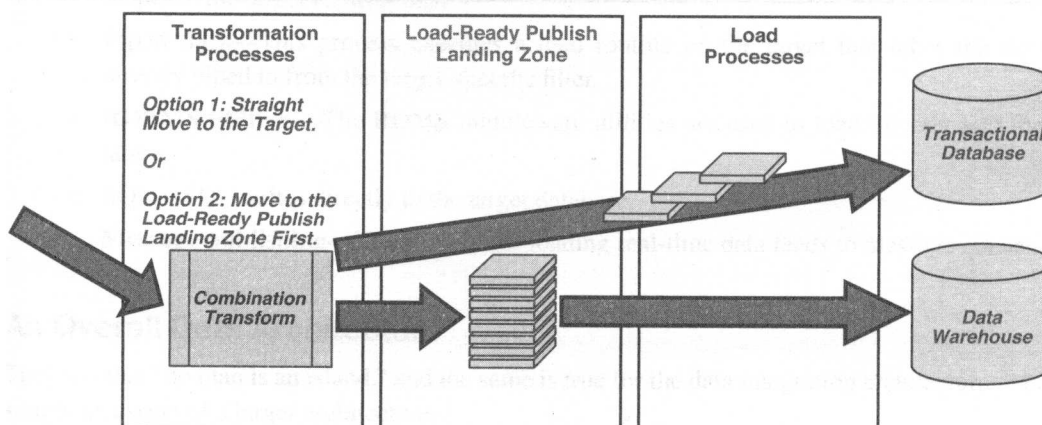


Ilustración 14. Movimiento de datos entre los procesos de transformación, la zona de publicación y los procesos de carga finales. Tomada de [GIORDIANO, 2011]

6. **Load-Ready Publish (Almacenamiento / Publicación de Datos Listos para Cargar):** La información transformada y lista para ser incluida en los sistemas destino se almacena en este repositorio hasta que sea posible integrarla, es posible que esta etapa se descarte algunas veces debido a que se pueda proceder inmediatamente a cargarla al destino, pero en algunas ocasiones será preferible incluir este tipo de *Staging*, sobre todo cuando la información integrada debe ser complementada con otra información, cuando haya ventanas de tiempo definidas para las cargas o cuando sea necesario distribuir la información entre múltiples destinos, como por ejemplo cuando se genere una dimensión conformada que deba ser distribuida a los *data marts*.
7. **Load (Proceso de Carga de Datos):** El proceso final toma la información de la última área de almacenamiento y la carga en los sistemas destino, esta etapa también deberá garantizar que la escritura de la información se hace correctamente.

Las siete capas del modelo propuesto por [GEORDIANO, 2011] obedecen a algunas reglas básicas y mientras podamos cumplir con estas reglas se nos permite cierta flexibilidad, a manera de resumen podríamos enumerar las siguientes:

- **“Read Once, Write Many”** (Lea una vez, escriba muchas): A mayor cantidad de lecturas de los sistemas de producción mayor será el costo, tanto en tiempo, como en recursos, debe evitarse a toda costa leer la misma información más de una vez, ya que esto atenta contra el rendimiento de todo el sistema de integración y la mantenibilidad de los procesos implementados.
- **“Grab Everything”** (Obtenga todo): Este principio apunta a que es mejor leer toda la información disponible y almacenarla en *Staging*, para posteriormente leer de esta estructura solamente lo que sea necesario, en este caso debemos establecer un límite a la regla ya que podríamos caer en el esquema de diseñar el *Staging Area* con base en los orígenes, así consideramos que esta regla debe tratarse con cautela, más como una

invitación a hacer procesos de integración con datos que sirvan a múltiples propósitos para que no tengamos que modificar constantemente los flujos de datos.

Es importante aquí resaltar que el modelo de *Staging Area* que se ha elegido para el BCR toma en cuenta las recomendaciones de este autor pero las ajusta para su realidad así, el diseño de *Staging Area* del BCR está representado por la siguiente figura, en ella se detallan algunos componentes a los que se hará referencia posteriormente:

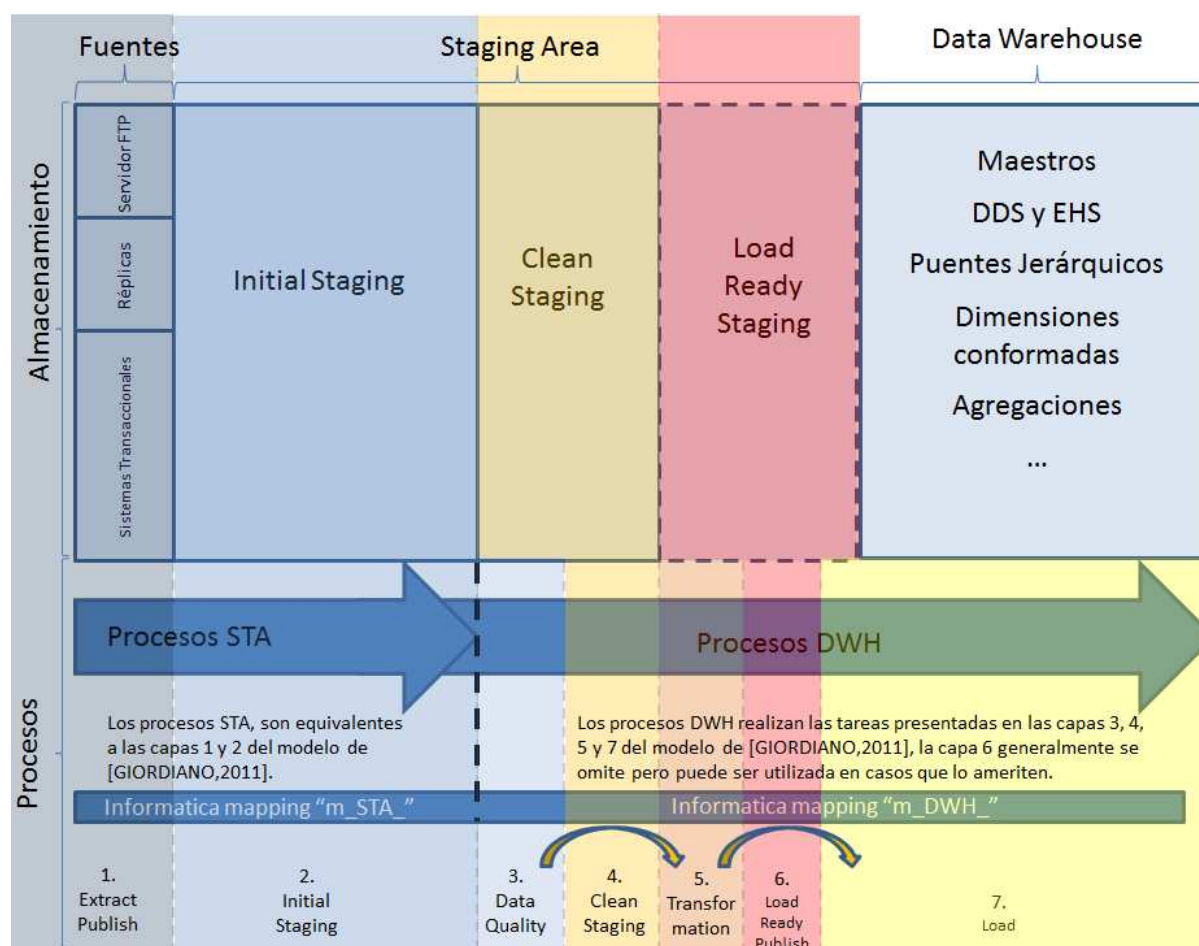


Ilustración 15. Correspondencia lógica entre la propuesta de desarrollo de procesos de integración de [GIORDIANO, 2011] y la propuesta implementada en el BCR.

Tal como lo explica la figura el *Staging Area* del BCR está formado principalmente por elementos que forman parte del *Initial Staging* o del *Clean Staging* los cuales son homólogos a los propuestos por [GIORDIANO, 2011], en cuanto al *Load Ready Staging*, es equivalente a la capa 6 pero su utilización es raramente necesaria, así que generalmente los datos pasan directamente a la capa 7.

Algunas de las capas se implementan en almacenamiento, y otras a nivel de procesos, en general se utilizan dos procesos para cada carga de información a *Data Warehouse*, el primero de los procesos es el llamado "Proceso STA", este incluye la lógica para obtener datos del

origen y “aterrizarlos” en las tablas de *Initial Staging*. El segundo proceso, llamado “Proceso DWH” tiene la lógica de verificación calidad de datos, en caso de encontrar problemas con la información este se encarga de ingresar los registros necesarios a “*Clean Staging*” pero a diferencia de lo propuesto en la metodología, los datos que si pasan las verificaciones de calidad se transmiten inmediatamente a la capa de Transformación y en caso de no haberse considerado necesario almacenarlos en *Load-Ready Staging*, pasarán directamente al repositorio destino, estas posibilidades se representan en el diagrama con la flechas color naranja que saltan las capas 4 y 6.

Las aclaraciones hechas en la barra superior a las etiquetas de capas, se refieren al estándar de nomenclatura diseñado para ser utilizado con la herramienta de integración de datos estándar del BCR para la carga de *Data Warehouse* Corporativo, esta herramienta es conocida como Informatica PowerCenter, y divide cada movimiento de datos entre fuentes y destinos en elementos denominados “Mappings”, así cada “Proceso STA” y “Proceso DWH” estarán representados por un mapping que utilizará correspondientemente para su nombre, los prefijos “m_STA_” o “m_DWH_”.

Estructuras elementales de Staging Area

Resolución de identidades por puentes externos

Dado que se desea mantener separados los sistemas de producción de las estructuras de *Data Warehouse* (principio de bajo acoplamiento), con el fin de que las particularidades de estos sistemas no intervengan de manera frecuente con las estructuras y procesos definidos en *Data Warehouse*, es común el que dentro de este repositorio se asignen llaves sustitutas a los elementos de datos. Así, aunque en el sistema de clientes o de cuentas corrientes se use como llave para el cliente la identificación y tipo de identificación o un número de identificación interno, estos identificadores, aunque se carguen a *Data Warehouse*, no serán los que conformarán las llaves primarias, esto según [IMHOFF, 2001] proporciona varias ventajas, entre las que podríamos mencionar:

- Llaves primarias más pequeñas
- Se elimina la necesidad de llaves compuestas, con la consiguiente mejora de rendimiento en operaciones de JOIN y mantenimiento de índices.
- La llave se mantiene estable a pesar de cambios en el origen de datos.

El utilizar llaves sustitutas en vez de las “naturales” si bien presenta las ventajas anteriores supone la necesidad de un proceso de “traducción” de la llave natural a la llave sustituta, o tal como podríamos llamarla, llave interna o de *Data Warehouse*.

Para poder traducir una llave natural a una interna, existen algunas posibilidades, una de ellas sería el almacenar ambos valores dentro de la tabla destino, así cuando se tenga que determinar si ya se ha cargado un registro particular, o si se desea saber si hubo

modificaciones a una entidad, se puede hacer una búsqueda en la tabla destino, y en ella ubicar la llave natural para obtener la llave sustituta que le corresponde.

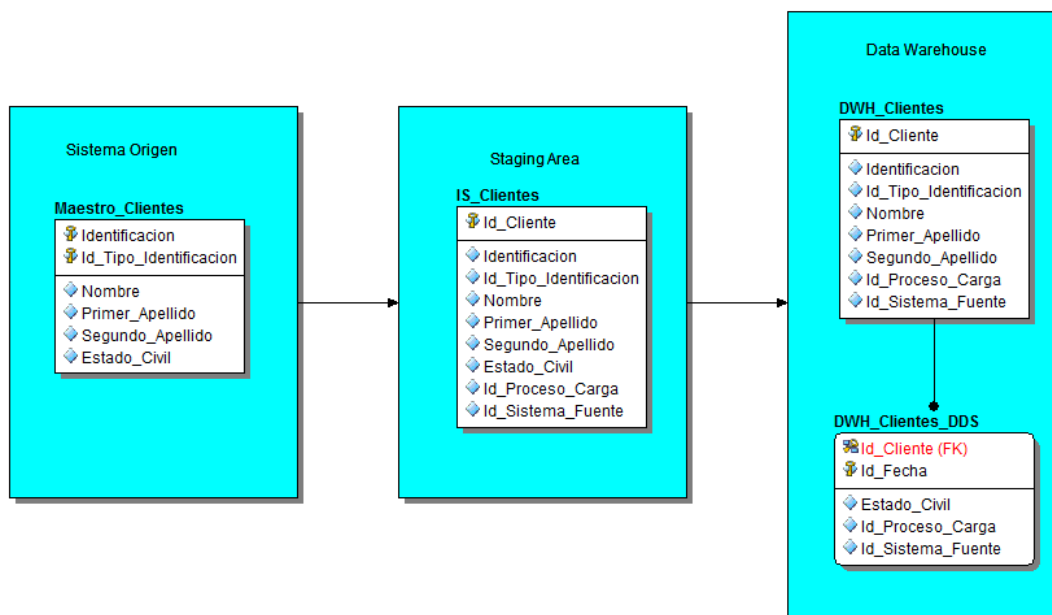


Ilustración 16. Esquema de estructuras para proceso de integración de datos sin la utilización de puentes externos.

La ventaja de esta opción es que se almacena poca información redundante y es posible verificar el resto de datos para ver si hay necesidad de actualizar información que haya sido modificada en el origen, las desventajas que tiene es que en caso de ser necesario revisar si hubieron cambios se requerirá comparar de uno por uno cada uno de los elementos, lo cual producirá un movimiento de datos a través de la red entre el servidor de base de datos de *Data Warehouse*, y el servidor de Integración que ejecuta la herramienta de ETL, en casos donde se habla de tablas de gran tamaño esto puede ser desfavorable para el rendimiento.

Viendo estas desventajas en el BCR se optó por un modelo que bautizamos como “Resolución de identidades por puentes externos”. Desde esta perspectiva el puente externo mantiene los siguientes elementos:

1. La relación entre una llave natural y una interna (de ahí su nombre al conectar los identificadores externos con el mundo exterior).
2. Uno o varios códigos *Hash* resultantes de la aplicación del algoritmo MD5 sobre la concatenación de un conjunto de valores de la tabla destino que son supervisados y que podrían disparar que se dé una actualización, sobre una tabla principal o sobre una tabla histórica (por ejemplo DDS o EHS).

3. Un identificador de sistema fuente, que define el sistema de donde proviene la llave natural.
4. Un identificador de proceso de carga, que es único para cada vez que se ejecuta el proceso que actualiza un conjunto de información en particular.

Así el modelo presentado anteriormente modificado para utilizar un puente externo se vería como el siguiente:

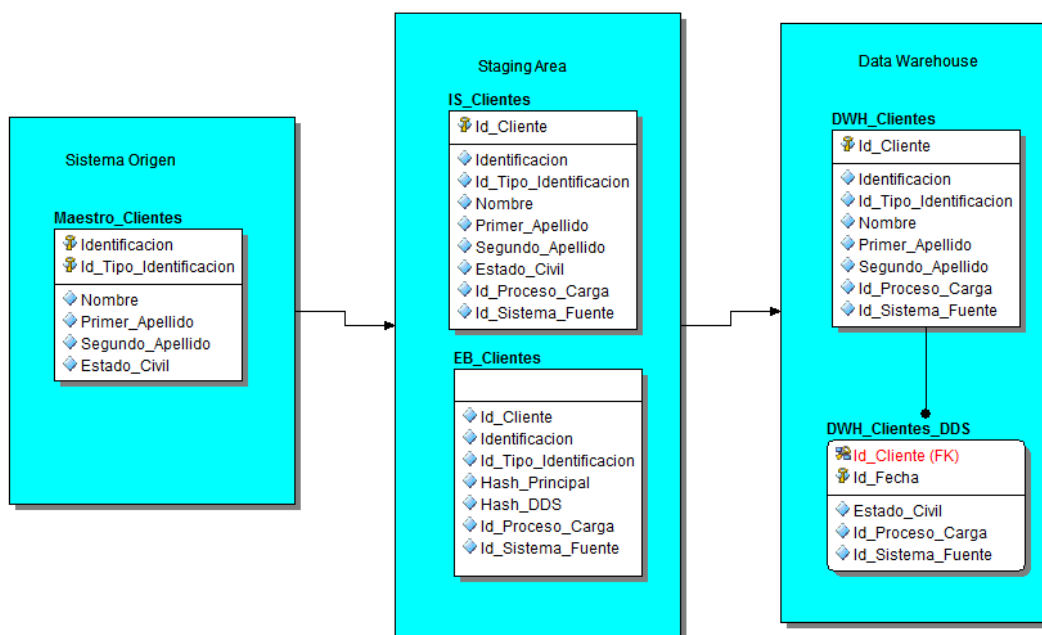


Ilustración 17. Esquema de estructuras para proceso de integración de datos que implementa puentes externos.

En este modelo el proceso de carga se da como sigue:

1. Se toma el registro del origen y se pasa a *Staging Area* a la tabla *IS_Clientes*, en este momento se especifica el *Id_Sistema_Fuente* y el *Id_Proceso_Carga* que corresponden.
2. El registro almacenado en “*Initial Staging*” se compara contra los datos del puente, se toma la información del nombre y apellidos y se computa un *Hash*, se toma la información del estado civil y se computa otro *Hash* (aunque en este caso el *Hash* es mayor en tamaño que el indicador de estado civil, la ventaja de este método es que el agregar más datos a la estructura no requerirá modificar el puente externo, solo la operación de cálculo de *hash*).

3. Se comparan los hash resultantes con los almacenados en el Puente Externo y con base en la comparación se determina si es necesario ejecutar una modificación en la tabla DWH_Clientes o una inserción de un nuevo registro en DWH_Clientes_DDS.
4. Al alterar cualquiera de las dos tablas sea para modificar o insertar, el Id_Proceso_Carga de la tabla EB_Clientes se actualiza al valor actual.
5. Si se insertan datos nuevos en la tabla DDS se actualiza el Hash_DDS, idéntica situación se dará cuando se modifique o inserte en la tabla de DWH_Clientes.

La implementación de puentes externos además nos ofrece un valor agregado adicional, dentro del puente externo podemos asociar a dos llaves naturales distintas el mismo identificador interno. Esto nos permite ayudar a simplificar en gran medida el proceso de resolución de identidades entre varios sistemas, los puentes pueden ser manipulados de forma manual o automática para que se entienda que dos registros de sistemas diferentes o del mismo, son en realidad la misma entidad.

Para ilustrar esta última ventaja podemos ver por ejemplo lo que pasaría si requiriéramos generar un proceso de integración de datos que implique leer información de más de un sistema y en cada sistema el cliente se representara con un identificador de cliente diferente:

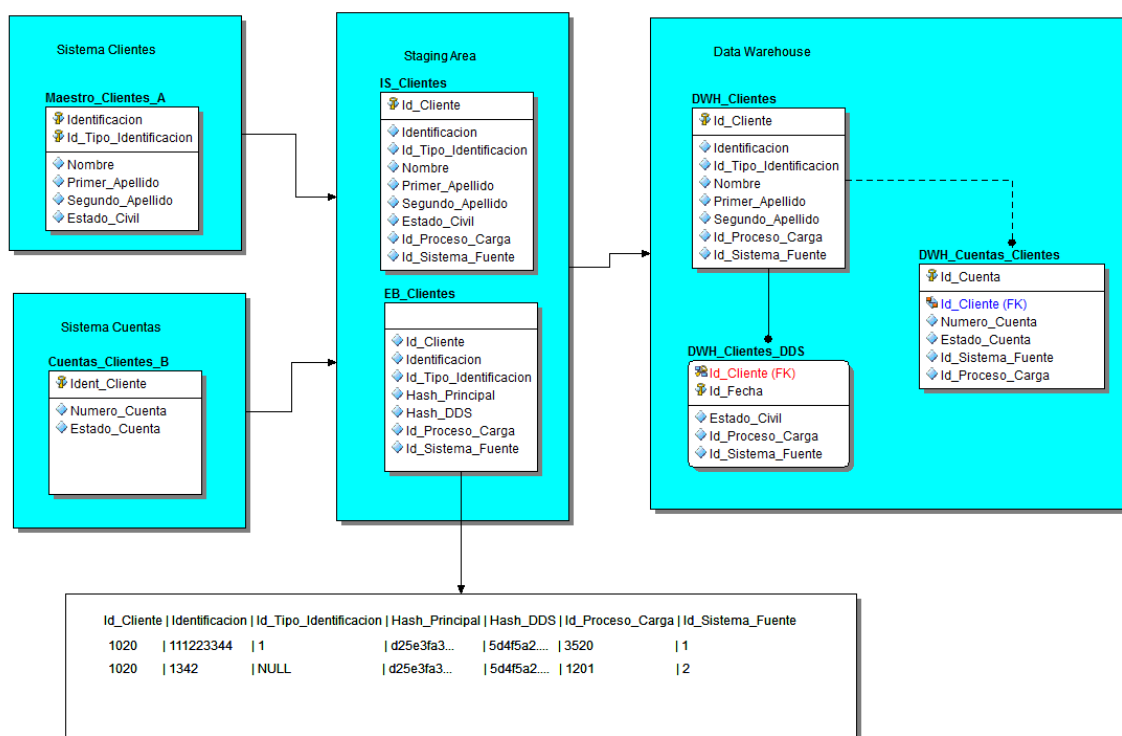


Ilustración 18.Utilización de puentes externos para resolver identidades a partir de información obtenida desde múltiples sistemas origen.

Como se puede apreciar, cuando en el Sistema de Cuentas (cuyo Id_Sistema_Fuente es 2) se extrae información sobre un cliente con identificación 1342 el Puente Externo permite determinar que el ID interno de *Data Warehouse* que le corresponde a este cliente es el 1020, mismo que se asignó para el cliente con identificación 111223344 y tipo de identificación 1 del Sistema de Clientes, el cual se identifica con un Id_Sistema_Fuente de 1.

El propio proceso de relacionar ambos identificadores externos al mismo Id_Cliente interno variará dependiendo de la información disponible en la empresa y del caso particular, a pesar de esto se puede observar la ventaja de la utilización de estas estructuras para unificar la información de las diferentes bases de datos.

Un caso especial de estos puentes externos son los que llamamos “Catálogos de Resolución” estos son estructuras que mantienen la misma estructura de un puente externo con la salvedad que se refieren a conceptos más genéricos como zonas geográficas, o estados civiles, los cuales se pueden representar en diferentes sistemas de diferentes formas y por tanto se usa un “catálogo de resolución” para estandarizar los identificadores que les corresponde según de donde provengan los datos.

Estructuras de limpieza (calidad) de datos

Tal como se presentó en la sección referente a las capas del sistema de integración de *Staging Area*, el *Staging Area* contiene una sección de “*Clean Staging*”, en esta se mantienen un conjunto de tablas que en estructura son muy similares a las de “*Initial Staging*” con algunas diferencias importantes.

Recordemos que según el modelo de integración definido por [GIORDIANO, 2011], en conjunto las capas 3. Calidad de Datos y 4. *Clean Staging*, generan tres diferentes salidas, datos limpios, datos rechazados y un reporte de rechazos. En el caso del BCR se hizo una adaptación de esta teoría para aprovechar la plataforma tecnológica disponible.

En primer lugar, los datos limpios no son escritos a *Clean Staging*, sino que pasan directamente a las otras capas del modelo.

Seguidamente los datos rechazados son escritos en las tablas de *Clean Staging*, pero estas tablas son manipuladas para alinearse a un formato específico, el cual requiere la herramienta Informatica Data Analyst (Parte de la suite de Informatica Data Quality).

Esta herramienta permite brindar una interfase web para que los usuarios expertos de negocio puedan evaluar los problemas de calidad de un conjunto de datos y proveer al negocio de enfoque y priorización en los problemas de calidad más comunes y de mayor impacto, o bien dar realimentación al equipo de *Data Warehouse*, para mejorar los procesos de integración de datos, ajustarlos o rediseñarlos según sea necesario.

Esta herramienta permite evaluar ya sea datos con problemas de calidad o resolver registros duplicados para convertirlos en uno solo, actualmente sólo se utiliza la característica de evaluación y corrección de problemas de calidad.

Cuando los datos son corregidos es posible reincorporarlos a la corriente de los procesos de carga de *Data Warehouse*, pasando la información corregida de las tablas de *Clean Staging* a las de *Initial Staging*.

Cada tabla de *Initial Staging* tiene tablas de *Clean Staging* asociadas, para poder crear un *Clean Staging* compatible con Informatica Analyst, se construyen dos tablas que inicialmente tienen la misma estructura que una tabla *Initial Staging* con las siguientes diferencias, usando como ejemplo una tabla de *Initial Staging* llamada IS_Clientes:

1. Las tablas de *Clean Staging* que le corresponden a IS_Clientes deberán llamarse: CS_Clientes y CS_Clientes_ISSUE y tener todos los campos que tiene IS_Clientes

- (el sufijo ISSUE es exigido por la herramienta, el prefijo CS es por motivos de estandarización).
2. A la tabla CS_Clientes se le debe agregar (para el caso de Oracle) los campos :
 - o ROW_IDENTIFIER de tipo NUMBER(38,0)
 - o RECORD_STATUS de tipo VARCHAR2(20 CHAR)
 - o UPDATED_STATUS de tipo VARCHAR2(20 CHAR)
 - o DESCRIPCION_CALIDAD de tipo VARCHAR(200 CHAR)
 3. A la tabla CS_Clientes_ISSUE se le deberá agregar el campo:
 - o ROW_IDENTIFIER de tipo NUMBER(38,0)
 - o DESCRIPCION_CALIDAD de tipo VARCHAR(200 CHAR)
 4. Cada una de las tablas deberá incluir todas las columnas de la tabla IS_Clientes, y cuando se encuentre un registro con problemas todas las columnas que se tomaron de IS_Clientes deberán almacenarse en las dos tablas.
 5. El ROW_IDENTIFIER deberá ser definido por el proceso de carga de ambas tablas y será el mismo en ambas cuando se escriba en ellas un registro con problemas.
 6. El RECORD_STATUS, deberá inicializarse con el valor 'INVALID'.
 7. Además como parte del estándar propuesto para el BCR, cada regla de calidad que haya hecho que el registro fuera rechazado, deberá concatenar una corta descripción de la razón del fallo en el campo DESCRIPCION_CALIDAD.

Cuando el usuario utilice Informatica Analyst para examinar los datos con problemas, la herramienta utilizará ambas tablas para guardar la información de cómo era el registro antes y después de la modificación del usuario, si el cambio es aprobado, el registro podrá volver a colocarse en *Initial Staging* para ser integrado en el *Data Warehouse* con la próxima carga de datos, o bien la observación de la frecuencia de aparición de errores podrá impulsar la creación de un proceso de limpieza en las fuentes de datos o la generación de una nueva regla de negocio dentro de los procesos de integración. El usuario podrá utilizar además el campo DESCRIPCION_CALIDAD para poder encontrar más fácilmente los campos con problemas y poder proveer una solución más adecuada a los problemas de calidad de datos.

El ejemplo a continuación muestra la relación entre las tablas de *Initial Staging* y de *Clean Staging* para el tema CDPs, que en este momento se está desarrollando para introducirlo a *Data Warehouse*, a modo de aclaración, el orden de los campos responde a reglas internas del BCR para especificar el orden de los campos de una base de datos relacional.

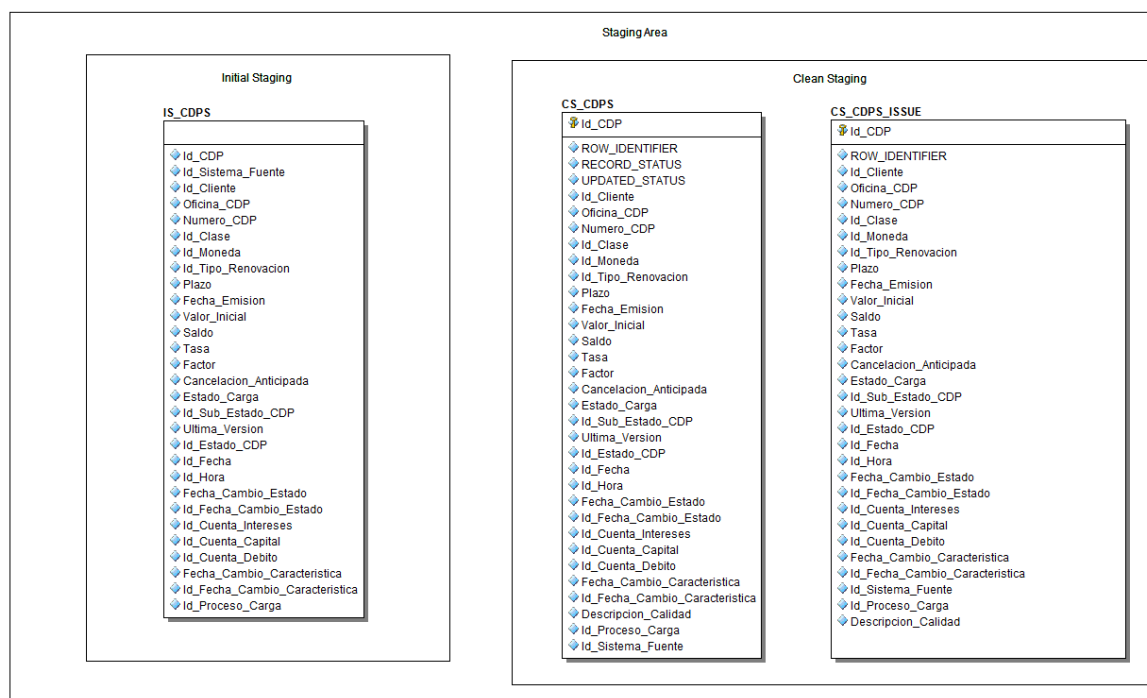


Ilustración 19. Ejemplo de un conjunto de tablas Initial Staging y Clean Staging correspondientes al mismo tema.

Repositorio de Metadatos

El repositorio de metadatos cumple variadas funciones, este contiene la información referente a los datos ingresados, constituye el “centro de control” de los procesos de integración de *Data Warehouse* y la biblioteca de información referente a cada elemento incluido en la CIF.

Inmon reconoce que la existencia de este tipo de repositorios es un factor crítico de éxito en cualquier proyecto de *Datawarehousing*, y establece su papel como el “pegamento que mantiene unida la CIF”.

Este repositorio, cumple dentro de este momento un papel tanto de fuente de referencia para la información contenida en el mismo, como de coordinador de las cargas realizadas hacia el interior de la CIF, este último papel se refleja en los procesos de seguimiento, limpieza y reversión de cargas que se verán más adelante.

Datos informativos del repositorio de metadatos

La información de referencia incluida dentro del repositorio de Metadatos, incluye la información básica de cada proceso de carga implementado, de los desarrolladores que lo realizaron, el usuario que lo solicitó y la designación de dueño de los datos para cada columna en cada estructura del repositorio, con fines de seguridad y concesión de permisos de lectura.

Así mismo, se incluye una descripción textual de cada campo integrado en la CIF y la relación a sistema o sistemas que originan la información incluida en cada campo.

Además de la información mencionada, es importante agregar que el repositorio de metadatos incluye además la información recopilada en los instrumentos de calidad y de mapeo de datos, con lo que esta información se mantiene disponible y fácilmente accesible a los usuarios y desarrolladores que la requieran e íntimamente ligada al repositorio al que se refieren.

Procesos de seguimiento, limpieza y reversión de cargas fallidas

Dado a que es imperativo la creación de un procedimiento que elimine la información almacenada temporalmente dentro de *Staging Area* cuando se ha realizado con éxito la carga de la información y que realice procesos de limpieza que garanticen que en caso de producirse problemas durante las cargas de datos realizadas hacia la CIF se pueda volver a un estado consistente de la información, se diseñó un proceso de limpieza que se ejecuta sobre una parte de la información almacenada en el repositorio de metadatos que refleja los estados de las cargas realizadas y la estructura de los procesos de carga, dándonos la posibilidad de implementar un proceso automático que es calendarizado para ejecutarse al terminar con todas las cargas del día y una vez durante la noche como contingencia. El modelo implementado se representa mediante el siguiente diagrama el cual se ilustra con un ejemplo a lo largo de esta sección.

Es importante aclarar que el movimiento a través de este flujo de estados se da en tres momentos diferentes que se ejecutan de manera secuencial para cada proceso de integración de datos, las dos primeras etapas: la carga de *Staging Area* y la carga de *Data Warehouse*, se ejecutan una después de la otra y la ejecución del Proceso de Limpieza, se deja para el final de la ejecución de todas las cargas con el fin de aprovechar al máximo la ventana de tiempo para las cargas y los recursos de procesamiento, de los servidores de base de datos e integración.

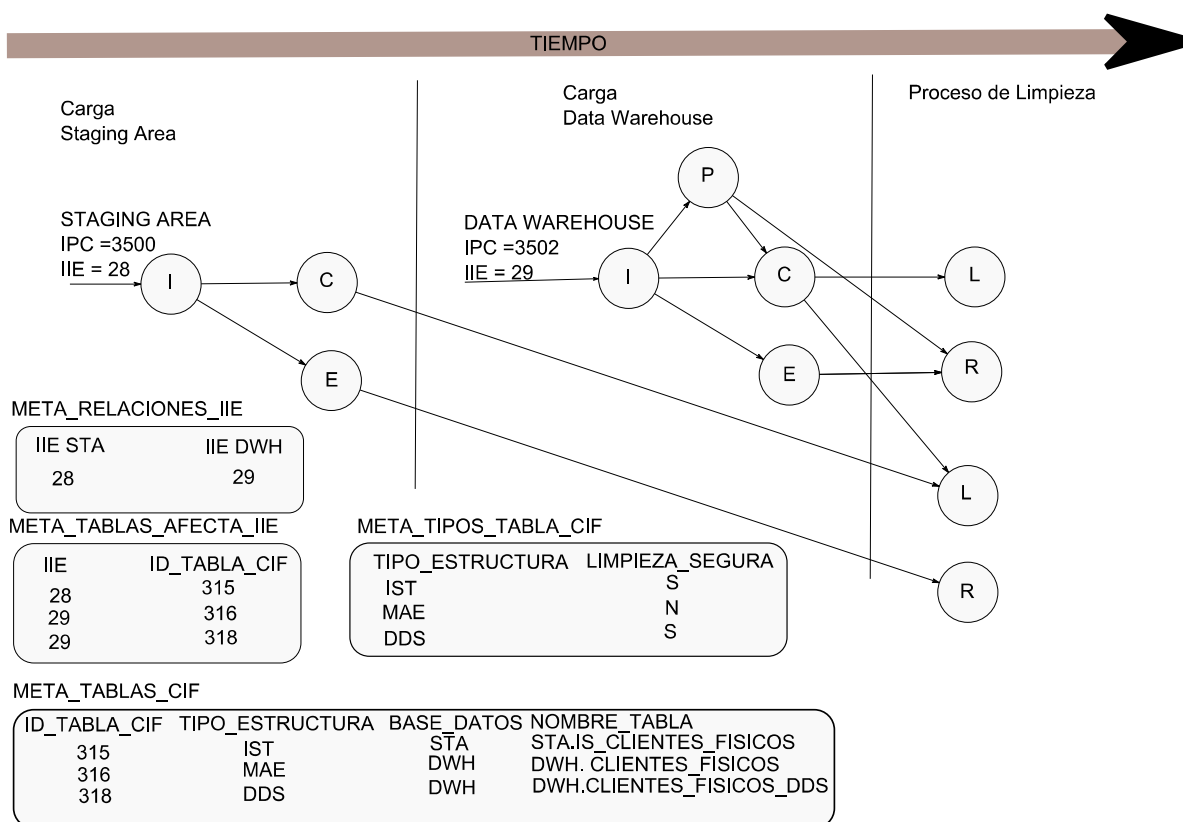


Ilustración 20. Evolución de los estados de procesos de carga a través de la ejecución en Staging Area y Data Warehouse.

Con el fin de poder dar un seguimiento adecuado a cada carga de datos hecha hacia Data Warehouse Corporativo, cada vez que se hace un movimiento de datos dentro de la CIF la herramienta de integración de datos ejecuta una subrutina que solicita a la base de datos del repositorio de metadatos, la asignación de un “Identificador de Proceso de Carga” (IPC), estos son administrados dentro del repositorio de metadatos en la estructura “META_Procesos_Carga”.

Los identificadores de procesos de carga son únicos para cada ejecución de un proceso de integración de datos y son almacenados como una columna más en todas las estructuras modificadas por este proceso dentro de la Fábrica de Información Corporativa, sean estas de *Staging Area*, de *Data Warehouse* o de un *Data Mart*.

Por otra parte cada proceso de integración de datos (o ETL) que se incorpora a la CIF tiene su propio identificador, al que llamamos “Identificador de Implementación de ETL” (IIE), estos identificadores sólo varían cuando se hace alguna modificación importante sobre un proceso de integración, de manera que cada nueva versión de cada proceso de integración tendrá un *Id_Implementacion_ETL* único.

Cuando se solicita la asignación de un IPC, por ejemplo para un proceso de carga a *Staging Area*, un procedimiento almacenado se encarga de grabar en una bitácora el IIE del proceso que hizo la solicitud, asociarlo al IPC designado y grabar la fecha y hora de inicio de la ejecución del proceso, así como algunos parámetros básicos propios de la ejecución, como el periodo de tiempo que se esperaba extraer de los sistemas origen con la ejecución del proceso de integración; además establece el valor de una columna llamada Estado_Carga en "I" con el fin de indicar que el proceso ha sido Iniciado y que se encuentra en ejecución actualmente.

El proceso de integración entonces procede a ejecutar su lógica y en cada registro insertado o modificado establece el valor de la columna Id_Proceso_Carga (en cada tabla afectada) al valor del IPC recién asignado. Esto nos permite saber exactamente qué registros son afectados por cada ejecución de un proceso de integración.

Los procesos de carga de *Data Warehouse*, a diferencia de los de *Staging Area*, además pueden adoptar el estado "P" el cual es utilizado cuando los procesos de carga de los puentes externos se dan en un proceso de integración de manera separada a la carga de las tablas de datos de *Data Warehouse*, esta situación requiere una lógica de limpieza distinta así que se requiere determinar el momento en que se llega a este estado para poder proceder adecuadamente durante cualquier tipo de eventualidad.

Al finalizar correctamente, el proceso de integración ejecuta nuevamente una rutina a la que proporciona el IPC que se le había asignado y solicita el "cierre" del proceso de carga que estaba en la bitácora, esto lanza un procedimiento almacenado que actualiza el registro de bitácora con los siguientes datos:

- la fecha y hora de finalización del proceso,
- la cantidad de filas leídas correctamente,
- la cantidad de filas leídas que produjeron un error en la entrada,
- la cantidad de filas escritas correctamente,
- la cantidad de filas que al intentar escribirse produjeron un error
- con base en la cantidad de errores presentados establece el valor de Estado_Carga en 'C' o en 'E' para indicar correspondientemente que el proceso ha terminado de manera Correcta o con Errores.
- una descripción del último error encontrado en caso de haberse dado alguno o un resumen formateado en texto de las cantidades de filas escritas.

La implementación de este método de "etiquetado" de los datos alterados por cada proceso, nos brinda una forma muy sencilla de hacer seguimiento de la ejecución de los procesos de carga y nos permite además implementar dos importantes procesos:

1. La limpieza de los datos introducidos en *Staging Area*.
2. La reversión de procesos de carga fallidos, cuando estos son interrumpidos antes de terminar.

Ambos procesos están implementados como procedimientos internos de base de datos y son ejecutados mediante procesos de integración de datos que se calendarizan para ejecutar al final del día o al terminar con todas las cargas calendarizadas.

El repositorio de meta datos, además de almacenar la información de la bitácora de las cargas, almacena los nombres de cada tabla modificada por cada implementación de ETL (asocia los nombres de las tablas a un IIE particular) como se ve en la figura en la tabla META_TABLAS_IIE, además define la relación de dependencia entre procesos que cargan información a *Staging Area* y los que la cargan a *Data Warehouse* (tabla META_RELACIONES_IIE) y especifica la naturaleza de cada tabla (tabla META_TIPOS_TABLA_CIF) para determinar si en caso de reversión de las cargas la tabla se pueden eliminar datos o no.

Cuando llega el momento de ejecutar la limpieza de *Data Warehouse*, los procedimientos de base de datos utilizan la información sobre los procesos de carga, implementaciones de ETL y las relaciones entre procesos para determinar a partir de los estados de carga de la estructura de bitácora de cargas, si es adecuado proceder a limpiar las tablas de *Staging Area* en caso de una carga adecuada a *Data Warehouse*, o si se requiere mantener esta información ahí y reversar una carga parcial debido a que ocurrió un problema, cuando se procede a una limpieza de *Staging Area*, los procesos de carga que insertaron información en estas estructuras establecen su estado de carga en “L” para indicar que se realizó la Limpieza de los datos ingresados por este proceso, cuando se hizo la reversión de algún proceso de carga se establece en “R” para indicar que este proceso de carga fue Reversado.

Al mantener la información dentro de *Staging Area* cuando sucede algún problema, y reversar la carga parcialmente realizada, la siguiente ejecución del proceso de carga saltará la sección correspondiente a *Staging Area*, y procederá a ejecutar la sección correspondiente a *Data Warehouse* de manera que podrá actualizar la información adecuadamente y así mantener la integridad de la información dentro del *Data Warehouse* Corporativo.

La limpieza de una tabla depende de la naturaleza de la tabla afectada por el proceso de carga, tal como se muestra en el diagrama de esta sección las tablas especifican una base de datos donde se encuentran (STA y DWH) y a través del Tipo de Estructura podemos determinar si es seguro aplicar a este tipo de tabla una limpieza, con esta información se utiliza el siguiente conjunto de reglas para aplicar la limpieza:

¿Limpieza Segura?	¿Base de datos?	Proceso de limpieza a aplicar
No	STA	No generar instrucciones para limpieza.
No	DWH	No generar instrucciones para limpieza.
Si	STA	Aplicar TRUNCATE sobre la tabla afectada por el proceso

		de carga
Si	DWH	Aplicar un DELETE con un filtro sobre el campo ID_PROCESO_CARGA que haga que se eliminen solo los registros correspondientes al IPC que falló en la tabla afectada.

El criterio de limpieza segura se determina según si la tabla tiene o no posibilidad de alterar filas de datos ya existentes, cuando una carga puede actualizar información no es seguro eliminar los registros marcados con el IPC que falló, sin embargo dada la característica atómica de las transacciones de bases de datos un registro que pertenezca a un IPC fallido es aún válido, por lo que no es necesario reversar su carga.

Las instrucciones de limpieza se definen cada vez que un proceso de carga de datos finaliza su ejecución pero solamente se almacenan en una tabla, hasta que llegue el momento de proceder a la limpieza general, en este momento se aplican todas las instrucciones de limpieza de manera secuencial ya sin tener que preocuparse de la ventana de carga o de los efectos adversos de esta operación sobre las cargas de datos.

5. Definición de metodología de desarrollo.

5.1 Propuesta metodológica para el desarrollo de proyectos de Inteligencia de Negocios en el BCR.

El desarrollo de un *Data Warehouse* difiere significativamente de una aplicación tradicional. Para dar una idea más próxima a la realidad que muestre el nivel de complejidad asociado a este tipo de proyecto podemos mencionar el caso de una de las áreas interesadas en el proyecto, esta área solicitó información de productos de crédito, entiéndase por esto Préstamos y Tarjetas de Crédito. Fue muy evidente para nosotros después de una primera reunión que la información que esta área necesitaba requería de un proceso de integración de datos que representaba un esfuerzo significativo, y que hacía falta una gestión adecuada del requerimiento antes de proceder a las fases de análisis y desarrollo, el área en cuestión presentó originalmente un requerimiento consolidado y exhaustivo que incluía más de 500 valores a obtener desde diversos sistemas, e inclusive algunos que aún no estaban en sistemas transaccionales.

El primer paso fue reconocer que los recursos disponibles no serían capaces de desarrollar el requerimiento completo para todo el *Data Warehouse*, en un tiempo razonablemente corto, así que fue necesario cuestionar y priorizar sobre el contenido del requerimiento y optar por una metodología de desarrollo diferente a la tradicional.

Tomando como base las recomendaciones de [INMON, 2008] bajo un proceso de desarrollo tan complejo como el de un *Data Warehouse*, no nos era práctico ejecutar un proceso de desarrollo convencional de cascada, sino que decidimos presentar al usuario el esquema de desarrollo mucho más incremental e iterativo, por lo que se optó por el modelo en espiral, mismo que recomienda Inmon, mediante el cual podemos obtener rápidamente una “iteración” que consiste en un subconjunto finito de un conjunto de necesidades que es, o se visualiza, como demasiado grande o que tiene una frontera difusa, esta iteración es un producto utilizable y visible inmediatamente y sirve como principal insumo para la generación de una nueva iteración, cada iteración está compuesta a su vez de etapas y cada una constituyen un proceso completo de desarrollo.

Si examinamos en detalle la propuesta de Inmon nos deja algunas dudas sobre cómo deben estar estructuradas estas fases, pero como ya se había aclarado, la guía de Inmon viene a ser únicamente de alto nivel y debemos saber buscar información proveniente de autores complementarios, por ejemplo en este caso [MOSS , 2003] nos presenta la estructura interna de un ciclo de desarrollo para proyectos de Inteligencia de Negocios, en general la propuesta de estas autoras nos ubica primeramente en un modelo de desarrollo en espiral como el que se ilustra a continuación:

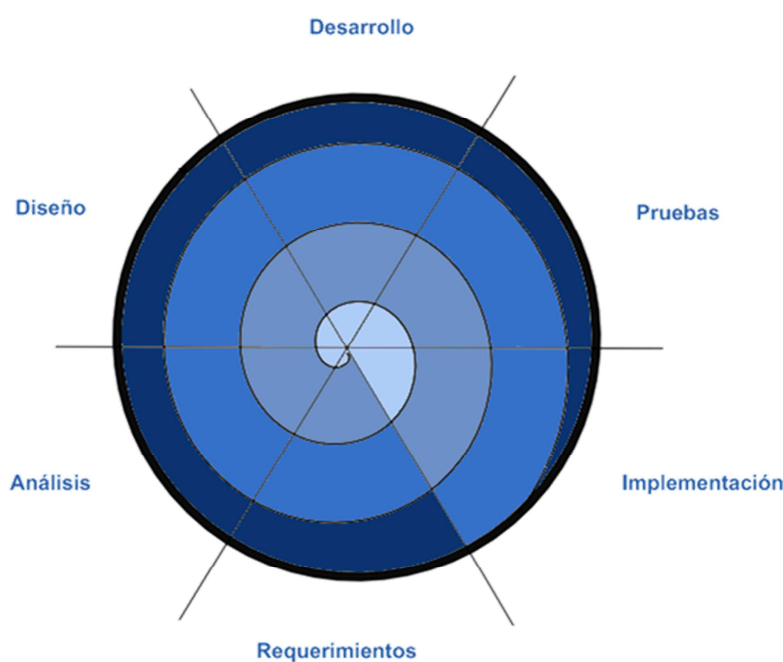


Ilustración 21. Modelo de un ciclo de desarrollo en espiral clásico.

La metodología de desarrollo en espiral se usa especialmente en proyectos de desarrollo que son difíciles de delimitar inicialmente, también en aquellos donde se ha determinado que es necesario un importante control del riesgo o en proyectos donde se requiere de ajustes constantes en los requerimientos durante la marcha, todas estas características minan la efectividad de un proceso de desarrollo tradicional (por ejemplo, en cascada), y son elementos que son fácilmente reconocibles en un proyecto de Inteligencia de Negocios, así, la propuesta

de [MOSS , 2003] nos sugiere el uso de esta forma de desarrollo, donde cada vuelta de la espiral corresponde a una iteración, y está formada por un ciclo completo que va desde los requerimientos hasta la implementación de un producto funcional, al menos visto desde el punto de vista de la ilustración anterior.

La única diferencia con el diagrama anterior es que las iteraciones de [MOSS , 2003] establecen etapas más especializadas para el desarrollo de proyectos de Inteligencia de Negocios, siendo así que se requieren 16 etapas diferentes que en conjunto forman un proceso de desarrollo especializado y que da como resultado 3 productos diferentes pero complementarios:

1. **Los ETL:** o procesos de extracción, transformación y carga, que es el significado de estas siglas en español, y que son los encargados de hacer llegar la información a la CIF de la manera que la requiere el negocio.
2. **Aplicaciones:** consistentes en interfaces de consulta o de generación de información, como herramientas de minería de datos.
3. **Elementos del Repositorio de Metadatos:** Información sobre la implementación de los procesos de carga, los datos cargados, los orígenes y destinos de información y la información de negocio relacionada a los datos cargados.

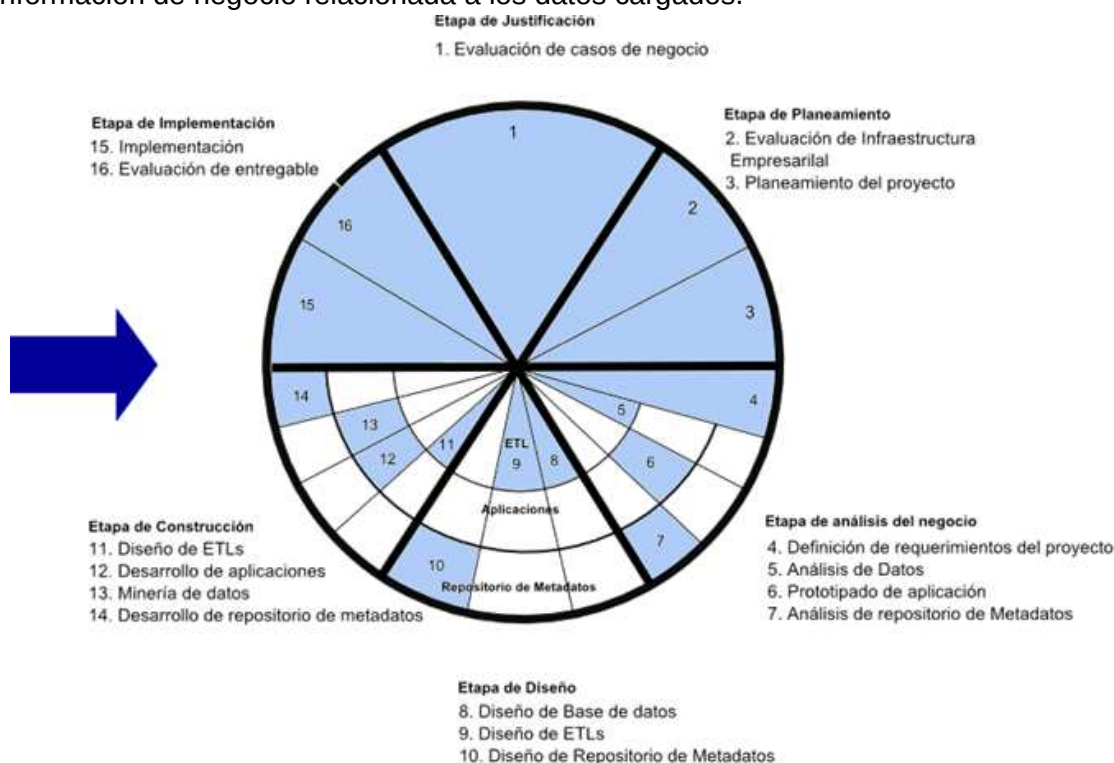


Ilustración 22. Modelo de desarrollo de una iteración de la espiral para un proyecto de Inteligencia de Negocios según [MOSS, 2003] .

Los pasos propuestos en la metodología de [MOSS, 2003] se pueden apreciar en la ilustración anterior. La sección del círculo con partes en blanco se refiere a las partes de la metodología en las que hay pasos exclusivos para el análisis, diseño y construcción de los tres productos arriba mencionados. Estos productos deben ser desarrollados en paralelo para poder facilitar el éxito del proyecto de Inteligencia de Negocios.

Esta metodología de desarrollo permite administrar los riesgos de manera más efectiva, mejora la planeación y la visualización del trabajo y permite organizar los recursos para dedicarlos a tareas mucho más enfocadas, además que permite “forzar” al usuario a centrarse primordialmente en aquello que es imprescindible.

A continuación se hace un breve resumen de cada uno de estos pasos basándonos para ello en el primer capítulo de [MOSS, 2003].

1. Etapa de Justificación

- 1.1. **Evaluación de casos de negocio:** Esta etapa consiste en la definición de un problema o una oportunidad de negocio, que será cubierta con un proyecto de inteligencia de negocios. Se debe justificar mediante un análisis costo-beneficio, que deje claros los beneficios de resolver el problema de negocio o de aprovechar la oportunidad.

2. Etapa de Planeamiento

- 2.1. **Evaluación de Infraestructura Empresarial:** Dado que algunos componentes de Infraestructura pueden no estar disponibles desde el inicio de cada proyecto de BI, es posible que la incorporación de estos elementos tenga que hacerse como parte de los proyectos, así puede que sea necesario desplegar nuevos elementos de infraestructura técnica, tales como hardware, software, repositorios, bases de datos, equipos de red, etc. o de infraestructura no técnica, tales como estándares de nomenclatura, estándares de metadatos, el modelo lógico de negocio (modelo de áreas temáticas de Inmon), metodologías, guías, procedimientos de pruebas, etc.

2.2. Planeamiento del Proyecto:

Los proyectos de inteligencia de negocios son extremadamente dinámicos, los cambios en las variables que los gobiernan, (tiempo, alcance, patrocinios) en el mismo pueden impactar severamente la consecución del objetivo, por esto es necesario llevar un seguimiento estrecho de los mismos. En esta etapa se planean detalladamente las actividades necesarias para la ejecución del proyecto.

3. Etapa de Análisis de Negocio

- 3.1. **Definición de requerimientos del proyecto:** Es importante delimitar desde el principio los alcances de cada proyecto, ya que es difícil luchar contra la necesidad de tener todo lo más pronto posible, la claridad en el alcance siempre será un factor

clave de éxito. Esta etapa se evoca a delimitar estos requerimientos a favor de la consecución de los objetivos.

- 3.2. **Análisis de Datos:** Esta etapa equivale al estudio de los datos del origen para determinar qué tanto estos datos son utilizables para cumplir con los requerimientos de negocio. En esta etapa deben de realizarse el mapeo de datos origen y el análisis de calidad de datos y la calidad de estos determinará en gran medida la calidad de los productos finales, por esto por lo que es muy posible que esta etapa tome una significativa cantidad del tiempo total del proyecto.
- 3.3. **Prototipado de aplicación:** Consiste en la creación de un prototipo de las interfaces de consulta que se utilizarán para cubrir la necesidad o el problema para el que se desarrolla el proyecto, esto permite a los usuarios de negocio proponer mejoras y establecer sus necesidades de mejor manera desde un principio, reduciendo la necesidad de cambios en las aplicaciones finales.
- 3.4. **Análisis de repositorio de metadatos:** Este paso consiste en la revisión de los requerimientos de información sobre los datos que serán cargados al repositorio donde se integrará la información, la importancia de su realización radica en que no se puede brindar seguridad al usuario de cómo se trataron los datos si no se le da trazabilidad completa a cada campo cargado, incluyendo información como su origen, las transformaciones y reglas de negocio aplicadas, entre otra información de importancia.

4. Etapa de Diseño

- 4.1. **Diseño de base de datos:** Consiste en la generación de los modelos lógicos y físicos donde se almacenará la información a integrar y la resultante, no siempre relacional y no siempre multidimensional, el tipo o tipos de modelos deberá ser diseñado de acuerdo a la necesidad del negocio.
- 4.2. **Diseño de ETLs:** En esta etapa, el diseño de los ETLs comprende la generación de modelos previos al desarrollo que estén específicamente diseñados para cumplir con los requerimientos que dicta el negocio, generalmente este proceso de diseño tiene que enfrentar modificaciones posteriores debido a las reducidas ventanas de tiempo y a que la pobre calidad de los datos hace más complejos y por lo tanto más largos los procesos de integración de información.
- 4.3. **Diseño de repositorio de metadatos:** Este paso consiste en la creación de un modelo lógico y físico, en caso de que no haya uno previo, o bien de la modificación del existente para ajustarse a las nuevas necesidades de información sobre los datos cargados a través de la ejecución del alcance del proyecto.

5. Etapa de Construcción

- 5.1. **Desarrollo de ETLs:** Consiste en la creación de los procesos de ETL correspondientes al diseño y al análisis de datos realizado, estos generalmente se modelan en una herramienta ETL, pero algunas veces se requiere extender la herramienta o bien, realizar procesos previos a la integración para lograr los fines definidos, así que la herramienta de ETL por sí sola no siempre será la única necesaria en esta etapa.
- 5.2. **Desarrollo de aplicaciones:** Esta etapa se refiere a la implementación final de lo establecido a nivel del prototipo y puede requerir esfuerzos diversos. Generalmente puede realizarse en paralelo con otras tareas de desarrollo ligadas a los demás entregables
- 5.3. **Minería de datos:** Esta etapa comprende la definición de un modelo analítico que permita ejecutar sobre él distintas técnicas de minería de datos, para luego analizar los resultados obtenidos, presentarlos de manera entendible, realizar validación de los resultados y monitorear el modelo analítico a través del tiempo.
- 5.4. **Desarrollo de repositorio de metadatos:** Cuando se decide desarrollar internamente el repositorio de metadatos este proceso equivale en sí a un sub-proyecto dentro de la construcción de la plataforma de Inteligencia de Negocios.

6. Etapa de Implementación

- 6.1. **Implementación:** Este paso se ejecuta después de haber finalizado y probado los tres entregables del proceso, en este momento se llevan a producción lo desarrollado y comienzan a darse las funciones de soporte, la mesa de servicio, mantenimiento de las bases de datos de BI, calendarización y ejecución de ETLs, monitoreo del desempeño y optimización de bases de datos.
- 6.2. **Evaluación de entregable:** Consiste en la revisión de lo sucedido durante la ejecución del proyecto, las lecciones aprendidas, problemas presentados y cómo se solucionaron, las técnicas y herramientas utilizadas, y los procesos que no fueron de provecho deben ser revisados y mejorados.

Aunque la metodología presentada por [MOSS,2003] es por mucho, más detallada que lo presentado en esta sección, el fin de presentarla de manera corta en memoria de proyecto, es poder dar a entender el alcance y la visión que engloba, dada la poca madurez en este tipo de proyectos dentro del BCR y a que existen ya normados procesos específicos para la implementación de aplicaciones dentro de la organización, la aplicación desde un inicio de este tipo de metodología no ha sido totalmente posible, sin embargo, los principios y consideraciones ofrecidas por la metodología si han sido incorporados en los instrumentos y los

procesos de análisis que se aplican actualmente. El objetivo que se persigue ahora es el de generar un conjunto adicional de instrumentos de análisis y documentación que permita poco a poco ir alineando el proceso de creación de repositorios de inteligencia de negocios hacia un proceso mucho más formal y estructurado como el que proponen Moss y Atre.

En este momento los esfuerzos por implementación de la metodología han sido parcialmente incorporados dentro de los proyectos de BI, de manera que podemos hacer una comparativa entre el modelo objetivo y las estrategias que actualmente se están llevando a cabo para cumplir con los mismos objetivos. Es importante rescatar que cada nuevo proyecto va mejorando o incluyendo nuevos procesos o insumos a la metodología practicada en el banco y estos son alineados a la propuesta de Moss y Atre para ir completando su alcance conforme el nivel de madurez nos lo permite.

Etapa según propuesta de Moss y Atre	Paso metodológico según propuesta de Moss y Atre	Medida, Procedimiento o Instrumento practicado actualmente en el BCR.	Procesos pendientes de implementar para alinear al modelo meta.
Etapa de Justificación	Evaluación de casos de negocio	Realización de reuniones previas para valoración de la necesidad al solicitar el inicio de un desarrollo sobre <i>Data Warehouse</i> Corporativo.	Generación de un documento que facilite el proceso de evaluación, y que incluya: <ul style="list-style-type: none"> Definición de objetivos del proyecto y su alineación a objetivos estratégicos del negocio. Problemas que presenta la solución actual. Estandarización de análisis costo-beneficio (obtención de ROI) Análisis de riesgos.
Etapa de Planeación	Evaluación de la Infraestructura empresarial	Actualmente este paso no se ejecuta dentro del BCR. Se utilizan los recursos disponibles que se heredaron de otros proyectos.	La generación de un proceso continuo de evaluación de la infraestructura técnica y no técnica para definir prioridades de atención y establecer cursos de acción. Se requiere una manera de determinar la necesidad de expandir la plataforma en cuando a equipos se refiere y utilizar las reuniones de evaluación de las etapas para determinar mejoras en los estándares y el proceso.

	Planeación del proyecto	Cada nueva iteración sobre un producto existente o para incluir un producto nuevo incluye un cronograma que incluye las tareas a ejecutar responsables y tiempos de entrega, estos son pactados e incluidos como órdenes de trabajo a los colaboradores, se hacen revisiones periódicas del cronograma y se hacen ajustes según sea necesario.	<p>Se requiere agregar instrumentos como bitácoras de cambios, que permitan examinar las razones por las que se modificó el cronograma.</p> <p>Se requiere mejorar las estimaciones previas mediante métricas tomadas durante los procesos de desarrollo, a fin de obtener estimaciones más realistas.</p> <p>Se requiere formalizar y realizar un proceso de análisis de datos inicial a pequeña escala que permita visualizar algunos posibles problemas desde el inicio.</p>
	Definición de requerimientos del proyecto	Cada nueva iteración sobre un producto existente o para incluir un producto nuevo tiene asociado un documento cuyo formato está dentro de la normativa del banco que representa la solicitud de información hecha por el cliente.	<p>Incorporar requerimientos de infraestructura técnica y no técnica en esta etapa.</p> <p>Definir necesidades de reportería aparte de necesidades extracción de información y de metadatos.</p> <p>Generar un documento complementario al oficial que facilite la incorporación de los requerimientos para cada entregable de la metodología.</p>
	Análisis de datos	<p>Se implementaron los instrumentos de calidad y de mapeo de datos para contar con una guía que permita hacer un análisis de datos apropiado, estos son constantemente revisados y mejorados, mediante las recomendaciones de los analistas que los utilizan.</p> <p>Los hallazgos encontrados en esta</p>	<p>Este paso se encuentra en concordancia directa con las recomendaciones de Moss y Atre, ya que implícitamente incluye todos los pasos requeridos por las autoras para la realización de esta etapa.</p> <p>Aun cuando esta etapa se encuentra en correspondencia con lo propuesto en la metodología que se desea implementar es deseable modificar los instrumentos</p>

	<p>etapa se aplican al modelo de datos de sistema de <i>Data Warehouse</i>, el cual se mantiene en relación directa con el modelo de negocio y de áreas temáticas.</p> <p>Cada nuevo análisis contempla los posibles problemas de limpieza que implica la carga de datos desde las fuentes que se han mapeado y la definición de un proceso de limpieza apropiado definido en conjunto con el negocio.</p>	<p>realizados para mejorar aún más la alineación de los mismos a las técnicas complementarias de análisis de datos recomendadas por Moss y Atre, el modelado de datos lógicos Top-Down y el análisis de datos origen Bottom-Up.</p>
Prototipado de aplicación.	<p>Para cada nueva área temática introducida en <i>Data Warehouse</i> Corporativo se crea una capa semántica (conocido como Universo) y un reporte base utilizando la herramienta SAP Business Objects XI, sobre el que posteriormente los usuarios solicitan mejoras y agregados, estos cumplen en estas etapas iniciales la función de prototipos de aplicación.</p>	<p>Definir un método de prototipado menos elaborado para cumplir con el objetivo de establecer el alcance de la aplicación, sin llegar a tener que desarrollar toda la capa semántica.</p> <p>Facilitar los ajustes al diseño de datos establecido según se defina con más exactitud los prototipos de aplicaciones, a través de anotaciones posteriores al análisis.</p> <p>Formalizar el proceso de creación de prototipos para que sea ordenado y metódico, más que un ejercicio previo solamente.</p>

	Análisis de repositorio de metadatos.	Cada nueva área temática o datos introducidos a la CIF incluyen dentro de su análisis la necesidad de hacer ajustes a los repositorios de metadatos, además de que introducen la información básica requerida al modelo de metadatos ya implementados.	Considerar dentro de esta etapa los requerimientos para las aplicaciones de acceso a metadatos, utilizar para esto los estándares de requerimientos para aplicaciones definidos dentro del banco. Agregar la capa de meta-meta datos dentro de la arquitectura de <i>Data Warehouse</i> Corporativo.
Etapas de Diseño	Diseño de base de datos	El diseño de base de datos se hace utilizando una herramienta de modelado de datos (ER/Studio Data Architect), esta facilita la aplicación de las recomendaciones de modelado de Imhoff y facilita la alineación a los niveles de modelado de datos que recomienda Inmon en la metodología de la CIF.	Es importante fortalecer constantemente la utilización de las mejores prácticas en modelado de parte de los diseñadores o analistas encargados de su construcción, las lecciones aprendidas deberían estar disponibles y distribuirse entre los participantes de la construcción de la CIF. Se requiere mejorar la integración entre estos modelos de datos y los definidos a nivel de arquitectura empresarial, con el fin de proveer a la organización con los medios adecuados para poder dar seguimiento a la información desde un concepto de negocio hasta un elemento dentro de los repositorios de la CIF.
	Diseño de ETLs	El diseño de los ETLs se realiza de manera informal, ya que no se contaba con una aplicación para realizar esta antes de proceder, para el diseño se realizan algunas reuniones entre analistas para determinar la mejor manera de proceder posteriormente con el desarrollo de los ETL.	Se debe estandarizar la utilización de una herramienta que permita generar un diseño preliminar de los ETL. Se debe garantizar que los procesos de ETL responden a la estructura de reformatear, reconciliar y limpiar la información. Se debe estandarizar formalmente la

		nomenclatura utilizada para nombrar cada elemento dentro de los modelos de datos, basándose preferiblemente en los estándares definidos por el fabricante de la herramienta y los elementos de arquitectura definidos a través de esta investigación.	
	Diseño de Repositorio de Metadatos	El diseño de repositorio de metadatos se realiza dentro del mismo proceso de Diseño de base de datos.	Incluir dentro de los modelos de metadatos la posibilidad de integrarse con los modelos de arquitectura empresarial definidos a nivel corporativo por la oficina de Arquitectura, con el fin de mantener la sincronización entre la información referente a dueños de datos y la información que se almacena en <i>Data Warehouse</i> Corporativo.
Etapas de Construcción	Desarrollo de ETLs	<p>Para el desarrollo de ETLs se usa una herramienta de integración de datos estándar (Informatica PowerCenter), la construcción de estos flujos se complementa con algunas utilidades de sistema operativo o de transferencia de archivos para poder alimentar los procesos de manera adecuada.</p> <p>La creación de los procesos de carga mantiene un subconjunto de buenas prácticas definidas por el fabricante para la creación de procesos de integración de datos.</p> <p>Las pruebas de los procesos de carga son realizadas únicamente</p>	<p>Implementación de revisiones en pareja para cada proceso de carga desarrollado, antes de proceder con las pruebas.</p> <p>Introducir nuevos y más tipos de pruebas antes de considerar aprobado un proceso de carga: Unitarias, de Regresión, de Aseguramiento de Calidad, de Aceptación, de Desempeño y de Integración, según las recomendaciones de Moss y Atre.</p> <p>Realizar acercamientos con el área de aseguramiento de calidad para ofrecerles el conocimiento suficiente para hacer pruebas sobre este tipo de desarrollos.</p>

	por el desarrollador de ETL en presencia de un recurso de aseguramiento de calidad que no necesariamente tiene conocimientos sobre integración de datos.	
Desarrollo de aplicaciones	El desarrollo de aplicaciones orientadas a la reportería y análisis de información se realiza mediante la creación de universos, reportería simple y asistida, que se diseña e implementa en la plataforma de SAP Business Objects XI.	<p>Promover el uso de herramientas OLAP dentro de la arquitectura de CIF definida.</p> <p>Especializar secciones del <i>Data Warehouse</i> Corporativo en <i>Data Marts</i> para darle a la información un contexto más cercano a cada área de negocio y utilizar estos repositorios para generar nuevas interfaces de usuario.</p> <p>Incorporar un proceso de pruebas unitarias más formal sobre las aplicaciones.</p> <p>Gestionar capacitación constante a los usuarios de las aplicaciones, para mejorar y estimular el uso de las interfaces de acceso.</p> <p>Gestionar capacitación en procesos de análisis basados en los datos publicados para mejorar el aprovechamiento de la información.</p> <p>Promover el uso de herramientas de análisis de datos adicionales que faciliten la aplicación de procesos avanzados de análisis en la organización (por ejemplo SPSS).</p>
Minería de datos	En este momento no existen procesos de minería de datos dentro de la organización.	Generar un primer proceso de desarrollo en el tema de minería de datos y capacitar al personal en la aplicación de las técnicas de minería de datos.

			<p>Presentar a los usuarios el concepto de minería de datos, fomentar el aprendizaje y utilización de estas técnicas en la organización.</p>
	Desarrollo de repositorio de metadatos	<p>El desarrollo y población del repositorio de metadatos se realiza en conjunto con el desarrollo de los ETLs, en esta etapa el desarrollador genera las instrucciones necesarias para llenar el repositorio y hace los agregados necesarios al modelo de metadatos. Actualmente la introducción de datos al repositorio de metadatos se hace de forma manual a la base de datos, no se cuenta con una aplicación que permita administrar esta información de manera automática.</p>	<p>Formalizar el proceso de desarrollo de metadatos mediante un proceso definido.</p> <p>Generar los procesos de migración de metadatos necesarios para completar los modelos de metadatos actuales.</p> <p>Fomentar el uso y mantenimiento de una utilidad para la administración de metadatos.</p> <p>Establecer el uso de pruebas unitarias para la prueba de los desarrollos de integración de metadatos.</p> <p>Generar entrenamiento en la utilización de metadatos para los usuarios técnicos y de negocio.</p>
Etapas de Implementación	Implementación	<p>En la etapa de implementación se ejecutan los procedimientos de pase a pruebas y producción que rigen en el BCR.</p>	<p>Aplicar el concepto de despliegue incremental dentro de los pases a producción, conforme vaya incrementándose la comunidad de usuarios de la plataforma.</p> <p>Establecer un mecanismo para aplicar y mantener los esquemas de seguridad de acceso a los datos en producción.</p> <p>Generar una política de respaldos más acorde a la naturaleza de los datos almacenados.</p> <p>Establecer un enfoque de</p>

monitoreo de los recursos adecuado a la naturaleza de los repositorios manejados.		
Evaluación del entregable	Dado que los usuarios a este momento no hacen un uso masivo de la herramienta, las evaluaciones de entregables se hacen como sesiones a nivel técnico tomando en cuenta las observaciones dadas por los clientes.	La evaluación de los entregables en el BCR está poco normada y requiere de un mecanismo más estructurado sobre el cual apoyar sesiones de análisis, posteriores a la implementación de cada desarrollo. Las sesiones de revisión post-implementación deben ser obligatorias y se debe obtener la participación activa del negocio en todas ellas.

Teniendo ahora claro el nivel de adaptación de la metodología objetivo que se está llevando a cabo es importante hacer notar que el desarrollo de la CIF se está dividiendo en Etapas, divididas a su vez, a nivel administrativo, en sub-proyectos los cuales basados en una necesidad de negocio definen entregables, que representan una iteración sobre la información de un nueva área temática o la construcción de una nueva iteración sobre un área temática ya implementada. Por ejemplo, la ejecución de la Etapa 3 de *Data Warehouse* Corporativo, incluyó un proyecto para la inclusión de Préstamos, Tarjetas y CDPs, este proyecto implicó la planeación de generación de la primera iteración del producto de Préstamos, pero a la vez requirió de la planeación de la segunda iteración de Clientes, de la tercera iteración de Cuentas Corrientes y de Ahorros y la primera iteración de Información de Deudores (un subconjunto de Clientes) y de Estructura Organizacional, esto se muestra con detalle en la figura siguiente.

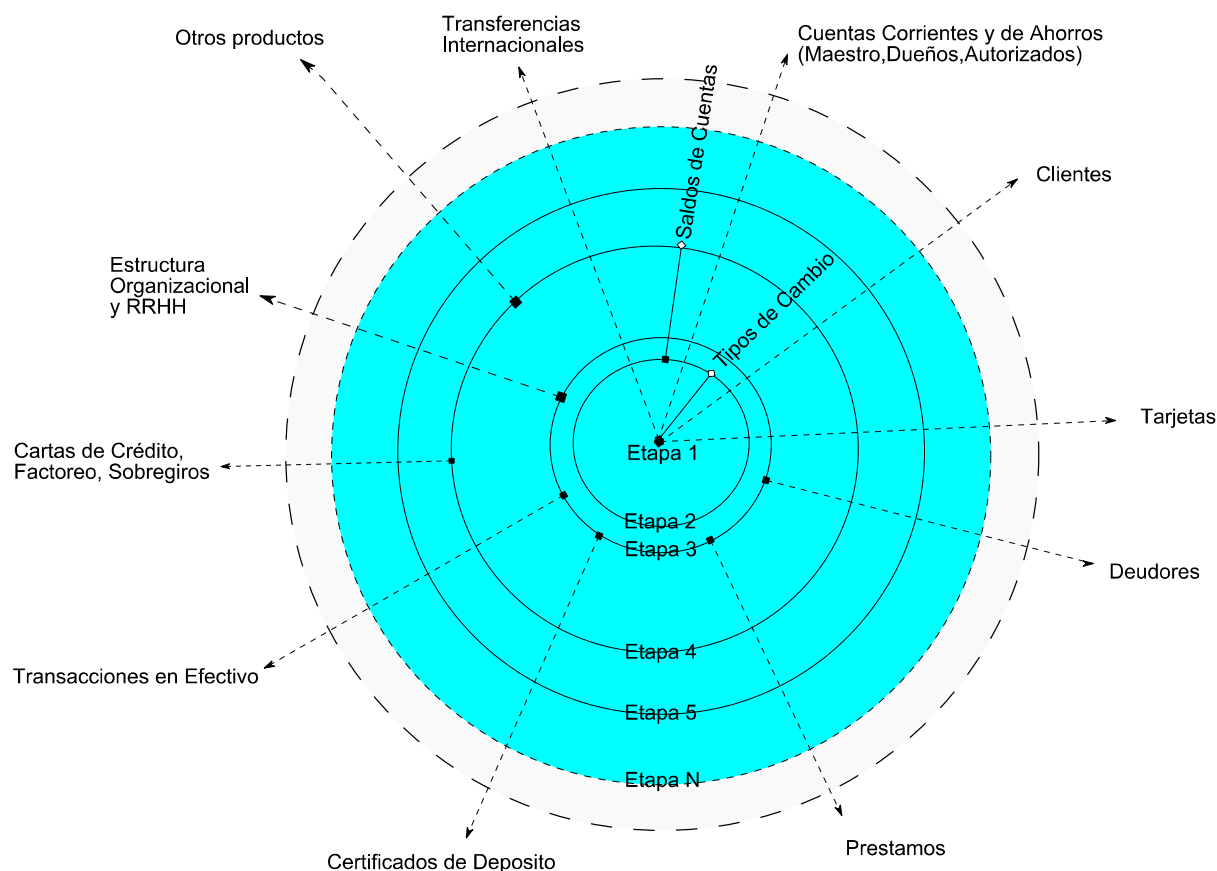
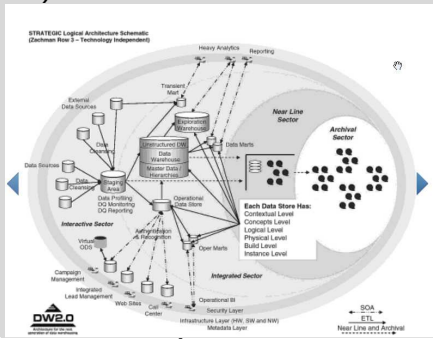
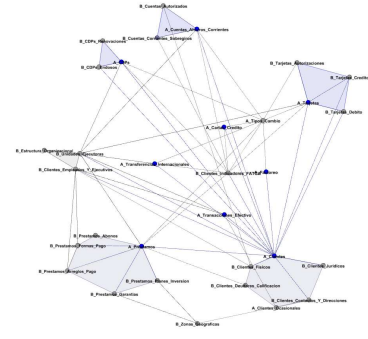
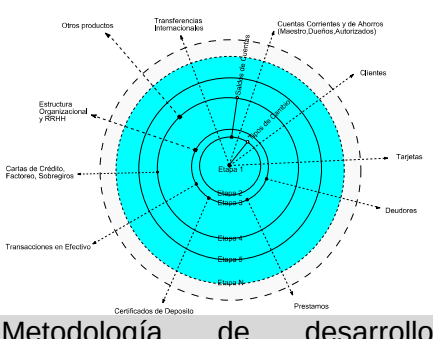


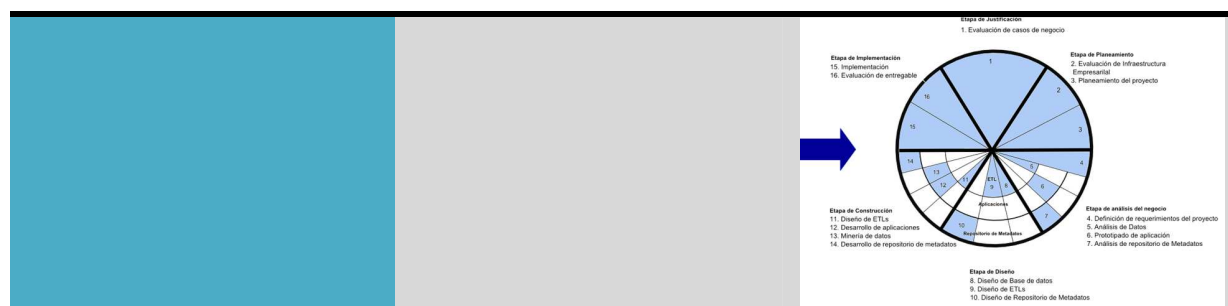
Ilustración 23. Evolución del programa de Inteligencia de Negocios del BCR, en fases por temas basado en el enfoque de desarrollo de espiral con fases para inteligencia de negocios de [MOSS,2003]

La implementación de la Fábrica de Información Corporativa del BCR, se ha desarrollado dividida en etapas, cada nueva etapa dota al equipo y al negocio de una mayor visibilidad de hacia adonde debe guiar el esfuerzo de desarrollo de la misma. Así el diagrama anterior refleja el orden de implementación de las áreas temáticas dentro del Data Warehouse Corporativo, el punto central y los círculos concéntricos representan las etapas de desarrollo, donde se acuerdan los entregables a realizar en un periodo y estos se llevan a producción a través de la aplicación de los pasos de la metodología de desarrollo, según la capacidad de los recursos humanos disponibles lo va permitiendo.

Los círculos que tienen un borde continuo representan etapas que se han reconocido como necesarias a este momento, estas han sido visualizadas por el negocio y los analistas durante la ejecución, los círculos con borde discontinuo representan las etapas subsecuentes y el límite difuso del alcance total del Data Warehouse Corporativo, una frontera que no se debe considerar como alcanzable dado que es muy probable que siempre existan nuevos requerimientos que implementar dentro de la fábrica.

En el BCR la relación entre el proyecto general (o como podría llamarse programa de proyectos), el proyecto específico, las áreas temáticas y los modelos propuestos por Inmon, así como las iteraciones sobre los productos se pueden entender de mejor manera a través del resumen que presenta la tabla siguiente:

Nivel en la Arquitectura	Nivel de administración y control	Representación
Fábrica de Información Corporativa	Estrategia de Inteligencia de Negocios	Modelo de Fábrica de Información Corporativa (DW 2.0) de Inmon: 
Áreas temáticas y Modelos de negocio	Programa de proyectos de Inteligencia de Negocios	Modelo de Áreas Temáticas y Modelo de Negocio: 
Modelos de sistema, Modelos tecnológicos, Iteraciones sobre productos.	Proyecto de Inteligencia de Negocios	Metodología de desarrollo basado en desarrollo de tres entregables: 



En las primeras etapas para un trabajo como este, las iteraciones eran repartidas entre los recursos disponibles, tomando en cuenta el tamaño del área temática y su experiencia previa, sin embargo gracias al apoyo obtenido por la administración de alto nivel que resultó en la asignación de nuevos recursos para el desarrollo de la CIF, actualmente es posible contar con un analista desarrollador para cada iteración o dependiendo de su complejidad podrán asignarse inclusive a dos para trabajar en pareja.

De esta manera, los procesos de desarrollo de productos de inteligencia de negocios quedan dentro de un marco administrativo y dentro de una metodología de desarrollo que permite un proceso adecuado que contempla los controles y pasos apropiados para lograr un producto adecuado para el negocio y para proporcionar a la organización de un esquema de trabajo adecuado que puede ir mejorando constantemente, mientras obtiene el nivel de madurez deseado en desarrollo e implementación de sistemas de inteligencia de negocios.

6. Modelo para la generación de interfaces de consulta

Tal como se vio en la sección referente a la metodología una parte importante de cada proyecto de Inteligencia de Negocios se refiere a la creación de una interface o aplicación que permita visualizar la información que fue originalmente solicitada por el negocio dentro de los repositorios de *Data Warehouse*.

Para poder proceder a entregar resultados sin embargo existen algunas consideraciones a tomar en cuenta.

El modelo de datos de *Data Warehouse* que se fundamenta en la arquitectura de Inmon, representa conceptos de negocio que son reconocidos a través de un proceso ordenado de análisis, los cuales permiten a la información ser lo suficientemente flexible para ser utilizada con cualquier finalidad. Al contrario que en un modelo de estrella que responde a una necesidad particular los modelos de datos de la Fábrica de Información Corporativa son de propósito general, esta característica es una ventaja en el mediano y largo plazo ya que comúnmente este modo de diseño permite evitar el re-trabajo.

No obstante lo anterior, el modelo de datos de *Data Warehouse* es el de una base de datos relacional, y es complicado para un usuario sin conocimiento técnico o del proceso realizado,

lidiar con las estructuras y elementos propios de los repositorios de datos implementados dentro de la CIF; más no lo sería tanto si los el usuario comprende los conceptos de negocio y la información se le presenta de manera similar a como se entiende el concepto de negocio.

El proceso de abstraer al usuario las estructuras de datos incluidas dentro de la CIF para que las reconozca como elementos de negocio que puedan ser utilizados por él o ella para generar reportes y hacer análisis para la toma de decisiones, se logra a través de la implementación de una capa semántica, la cual, es diseñada, desarrollada y puede ser explotada a través de una herramienta de reportería analítica para inteligencia de negocios; esta herramienta en el caso del BCR corresponde a SAP Business Objects XI, esta herramienta puede generar capas semánticas tanto sobre bases de datos relacionales como sobre bases de datos OLAP.

La herramienta en cuestión establece una capa semántica entre los repositorios de datos y los clientes a través del concepto de universo, el cual es un conjunto de conexiones a bases de datos, consultas SQL, definiciones de estructuras, operaciones predefinidas y relaciones entre los datos, que son presentadas al usuario en forma de “objetos” (elementos de datos extraídos directamente de una estructura o a través de un cálculo) y “clases” (agrupaciones de objetos que pertenecen a un mismo concepto de negocio). El usuario luego utiliza los universos, clases y objetos que le son visibles y a través de ellos puede definir, sin conocimiento directo de las estructuras subyacentes o de un lenguaje de consulta, cómo desea que se le presenten los datos que requiere y realizar sobre ellos algunas operaciones adicionales.

Existen algunas consideraciones importantes a la hora de diseñar un universo, dado que durante su creación se puede “ignorar” consciente o inconscientemente el esquema de datos implementado y definir uno propio a nivel semántico, es importante que se eviten algunas situaciones anómalas, como la creación de bucles al definir relaciones entre las tablas y definir contextos en la información para limitar la forma en que se generan automáticamente las consultas en respuesta a una solicitud de un cliente.

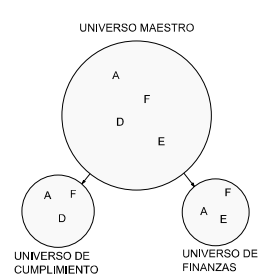
Al margen de este tipo de consideraciones que se refieren a buenas prácticas de modelado de universos, uno de los problemas de arquitectura más importantes a definir es cómo darle mantenibilidad a los universos, cómo evitar el re-trabajo, y cómo gestionar adecuadamente los desarrollos para que los cambios en los modelos de datos afecten lo menos posible los modelos semánticos (universos), y para que estos últimos puedan ser alterados de manera efectiva y transparente cada vez que se presentan cambios en un repositorio que esté siendo consumido como el *Data Warehouse Corporativo*.

Para entender la escogencia de arquitectura de universos escogida por el BCR, se debe primero entender el concepto de “*linked universe*” o universo vinculado. Cuando se crea un universo es posible definir que este debe “heredar” algunos elementos de otro universo ya existente; así el concepto de universo vinculado es análogo al de herencia en lenguajes de programación orientada a objetos.

Un vínculo entre universos tiene la ventaja de que al alterarse el universo “padre” las modificaciones en este serán percibidas por el “hijo” con lo que se logra cierta centralización de los cambios.

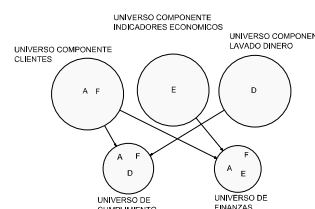
Un universo puede heredar de otros muchos, esto quiere decir, que un universo puede tener varios universos padre vinculados, pero por haber establecido esta relación él nunca podrá a su vez ser padre para otro universo.

Con el fin de seleccionar una estrategia para la implementación de universos en el BCR y de que la estructura determinada para generar los universos sea acorde al resto de la arquitectura, se sopesaron tres posibilidades diferentes de vínculos entre universos, y para cada una se determinaron las ventajas y desventajas, vistas desde el punto de vista de la arquitectura elegida a nivel de *Data Warehouse Corporativo*.

Vínculos a un Universo Maestro			
Descripción	Ventajas	Desventajas	Ejemplo
<p>→En esta estrategia se implementa un único universo maestro, que contiene todas las clases y objetos definidos en el modelo semántico a implementar.</p> <p>→Posteriormente cada necesidad de negocio a implementar, genera un universo particular y establece un ligamen al universo maestro, heredando para sí solamente las clases y objetos de su interés.</p>	<p>→Menor cantidad de universos a administrar.</p> <p>→Mayor centralización de los cambios, siempre se debe alterar solo el universo maestro.</p>	<p>→Un cambio en el universo maestro puede impactar a todos los universos que se ligán a él, no hay posibilidad de seguimiento cuando la cantidad de universos ligados crece. Todos podrían ser afectados por un cambio en el maestro.</p> <p>→El universo maestro puede volverse demasiado complejo para ser utilizado en proyectos muy extensos.</p>	<p>El universo maestro contiene las clases A,D,E,F, el universo de cumplimiento tiene solo las clases A,F y D y el universo de finanzas solamente las clases A,F y E.</p> 
Vínculos a uno o múltiples Universos Kernel			
Descripción	Ventajas	Desventajas	Ejemplo
<p>→En la estrategia de universos Kernel se deben determinar los elementos que son comunes a dos o más universos e implementarlos en uno o un</p>	<p>→Al consolidarse universos más pequeños que en el caso de universos maestro su modificación es más sencilla.</p> <p>→Se mantiene la posibilidad de que</p>	<p>→Algunas clases implementadas en universos que ya están hechos y son utilizados podrían migrarse de universos básicos a universos kernel para mantener la arquitectura de</p>	<p>El universo kernel contiene las clases A y ,F, el universo de cumplimiento tiene solo las clases A y D y el universo de finanzas solamente las clases A,E y F.</p>

<p>conjunto de universos Kernel, los cuales contienen los elementos compartidos. → Cada nuevo universo que requiera de un elemento de un universo Kernel deberá ligar el universo Kernel a si, y si dos universos requieren el mismo elemento este deberá estar en un universo Kernel para que todos accedan al mismo elemento.</p>	<p>al alterar un elemento en un universo kernel este cambio se propague a sus hijos. → Hay mayor facilidad para determinar si un cambio en un universo padre va a afectar a otros universos.</p>	<p>kernel. → Se debe de ser consciente de las clases implementadas para poder definir en qué momento una clase deberá ser introducida en un universo cliente. → Esta forma de arquitectura generará muchos más universos de los que podría ser útil administrar, algunos universos podrían estar conformados hasta por una sola tabla, mientras otros podrían tener cientos y ser utilizados para un propósito específico.</p>	
---	--	--	--

Vínculos de Universos diseñados por Componentes			
Descripción	Ventajas	Desventajas	Ejemplo
<p>→ Este enfoque, es en sí un caso especial del enfoque de Kernel, con un enfoque diferente, consiste en la mezcla de uno o más universos en uno solo. Así cada universo padre se refiere a un tema particular y agrupa las clases que pertenecen a este tema, cuando se hace un cambio a nivel de un tema, todos los universos derivados que consuman este universo componente se verán afectados.</p>	<p>→ Se mantiene la ventaja de aplicar cambios en cascada desde los universos padre o de referencia. → Como cada universo padre es especializado en un tema, es mucho más sencillo identificar cuales universos deberán modificarse al ocurrir un cambio en la arquitectura. → Se mantienen dos niveles de universos, los de referencia o padres y los especializados o hijos. → Se facilita la</p>	<p>→ Se generan más universos que en el esquema de universo maestro, pero menos que en el de Kernel y se construyen de forma más ordenada. → Se requiere un mayor control para planificar los universos padre y establecer su relación directa a las áreas temáticas.</p>	<p>Los universos de componentes contienen cada uno solo las clases que les competen al tema que representan, cuando los universos derivados requieren esta información lo heredan de los universos de componentes y cualquier cambio sobre estos últimos afectará en cascada los universos relacionados.</p>



alineación de los
universos al
concepto de
áreas temáticas.

Debido a que el proceso de desarrollo de la Fábrica de Información Corporativa del BCR se basó en un proceso gobernado por modelos los cuales en su más alto nivel dividen el contenido de la CIF en áreas temáticas, y a que nos es imperativo el que exista una capa semántica que pueda ser extendida con relativa facilidad al ir agregando nuevas estructuras a nivel de los repositorios de información, se decidió finalmente por optar por privilegiar el diseño de universos basándonos en la vinculación de universos basados en componentes, y establecer una relación directa entre los universos de referencia (padres de las relaciones) y las áreas definidas en el modelo de áreas temáticas desarrollado, así, el contenido de los universos de referencia estará muy relacionado al modelo de negocios definido, facilitando el uso de estos instrumentos para guiar a los usuarios en la utilización de la capa semántica y abstrayendo de ellos todos los elementos que se encuentran entre este modelo y los medios de explotación de la información que se hayan definido.

Es de esta manera como finalmente se cierra el círculo completo de arquitectura y metodología para brindar a la Fábrica de Información Corporativa del BCR todos los insumos necesarios para continuar desarrollándose en pro de una mejora significativa en los procesos de toma de decisiones dentro de la organización y un proceso mucho más estandarizado y ordenados para el desarrollo de aplicaciones de inteligencia de negocios dentro del departamento de tecnologías de información.

JUSTIFICACIÓN DE DECISIONES DE DISEÑO

Las decisiones de diseño tomadas durante la ejecución del proyecto han sido ampliamente justificadas en las secciones correspondientes sin embargo es muy importante dar un vistazo general a cada una de las más importantes decisiones de diseño con el fin de dar a estas el peso suficiente para convertirse en los pilares que sostienen la arquitectura de Fábrica de Información Corporativa definida para el BCR.

Arquitectura base

La arquitectura base sobre la que se despliega la implementación de todas las aplicaciones de Inteligencia de Negocios de la organización es la piedra angular sobre la que debería partir un proyecto como la construcción del *Data Warehouse* Corporativo, las escogencias en este tema son limitadas si queremos utilizar arquitecturas estándar y reconocidas, fue así como la decisión de utilización de arquitectura de la Fábrica de Información Corporativa (CIF) de Inmon fue seleccionada por sobre la propuesta del *Data Warehouse* Multidimensional de Kimball.

Las principales razones para esta decisión fueron el que las necesidades inicialmente encontradas a través del análisis de la situación actual, dejaron ver que la arquitectura de Inmon facilitaría el crecimiento del *Data Warehouse* Corporativo en el mediano y largo plazo, dando origen a una organización más orientada a un proceso integral para el bien común de toda la organización.

Se visualizó desde un principio que esta orientación hacia un repositorio único orientado a conceptos de negocios, en vez de la creación de repositorios separados orientados a fines específicos, fomentaría la creación de una cultura organizacional de desarrollo de procesos de integración para Inteligencia de Negocios mucho mejor estructurados y al fomento de la formación de una cultura de análisis eventualmente mucho más madura dentro de la organización.

Coordinación de los proyectos de BI mediante los niveles de modelado de Inmon (Programa de proyectos de Inteligencia de Negocios).

Cuando se hubo seleccionado una arquitectura base el siguiente paso fue el determinar cómo se administraría a alto nivel el conjunto de proyectos que se empezarían a gestar a nivel de la organización, para resolver este problema, el enfoque elegido consistió en la implementación de los modelos de datos propuestos por Inmon, las razones de haber elegido estos modelos fueron principalmente el nivel de enfoque y la visibilidad sobre la complejidad de la CIF que pueden proveer tanto a usuarios técnicos como a los usuarios de negocio.

La coordinación de los esfuerzos de desarrollo en concordancia constante con estos modelos mantiene alineados todos los demás elementos de la arquitectura de la CIF y permite mantener los objetivos de los proyectos alineados con los grandes objetivos estratégicos del negocio.

En general estos modelos y los subsecuentes modelos de sistema y tecnológico que se derivan de los mismos, constituyen el enlace lógico entre el negocio, los conceptos que maneja y la estructura física de los repositorios de *Data Warehouse* Corporativo.

Definición del núcleo de la Fábrica de Información Corporativa (CIF)

Tal como fue revisado en secciones anteriores el concepto de CIF va mucho más allá de solamente el *Data Warehouse*, la especialización de los datos en estructuras y repositorios específicamente diseñados permite satisfacer las necesidades de todos los diferentes tipos de usuarios de naturaleza analítica y con necesidad de información operativa integrada de la organización.

Las etapas realizadas hasta el momento se han enfocado específicamente en tres de los componentes de la CIF, el *Staging Area*, el *Data Warehouse*, y el repositorio de Metadatos.

Es importante justificar el hecho de que los recursos inicialmente asignados al proyecto eran escasos, el presupuesto reducido y la experiencia y apoyo organizacional relativamente pocos, esto nos forzó a priorizar la utilización de los recursos limitados para generar un primer insumo de valor, el cual tendría que ser la base sobre la que se deberían continuar desarrollando el resto de los elementos de la arquitectura.

Con esto en mente se hizo la elección de los elementos que formaron el núcleo de la CIF. Con el tiempo la situación de recursos y apoyo asignados a la iniciativa han comenzado a reforzarse, se reconoce apoyo de la organización a varios y altos niveles gerenciales, se han destinado nuevos recursos humanos y capacitación para los mismos con el fin de fortalecer la plataforma, y se han abierto las posibilidades de inversión en infraestructura y en herramientas, así como relaciones con los proveedores y mejoras en los estándares definidos.

Esta nueva situación facilitará que en el futuro las demás secciones de la Fábrica de Información Corporativa vayan tomando su lugar, conforme la organización madure conforme más información vaya siendo generada y las necesidades de esta se vayan canalizando de manera preferente a través de la CIF.

Definición de estrategias de modelado de datos bajo la arquitectura de *Data Warehouse* definida por la Fábrica de Información Corporativa.

Dado que la arquitectura seleccionada no incluye (al menos a nivel de la bibliografía presentada por su principal exponente) una propuesta propia de estrategias de modelado de datos a nivel de los repositorios que mantienen la información de la CIF, se debió optar por adoptar las recomendaciones dadas por otro autor, que complementara su teoría, en este caso, las dadas por Claudia Imhoff se adaptaron perfectamente a la propuesta de alto nivel de Inmon y permitieron definir muchas de las estructuras base sobre las que se fundamenta el modelo de datos de sistema de *Data Warehouse*.

La elección de este conjunto de estrategias de modelado se hizo necesaria ya que el modelo de datos clásico en tercera forma normal o en una forma normal superior a esta, no brindaba las capacidades de almacenamiento y modelado de historia que implica la creación del repositorio central de *Data Warehouse*.

Es por esto que usando como objetivo la obtención de un modelo de datos inspirado en las recomendaciones de Imhoff y en concordancia de la arquitectura de Inmon, se diseñaron estructuras auxiliares como los Puentes Externos las estructuras de *Initial Staging* y las tablas de *Clean Staging*, las cuales complementan la arquitectura con recomendaciones dadas por otros autores que presentan las necesidades del *Data Warehouse* desde perspectivas de integración y calidad de datos.

Selección de un enfoque para el análisis de la calidad de datos.

La presencia de los esfuerzos de calidad de datos dentro de la CIF es un garante de la sostenibilidad de la misma a través del tiempo y de la satisfacción de las necesidades del negocio en el corto, mediano y largo plazo.

La calidad de datos es un concepto bastante mal comprendido por la mayoría de profesionales en sistemas de información, generalmente se tiene una idea básica sobre conceptos como completitud y coherencia de la información pero la calidad de los datos trasciende estos aspectos superficiales y debe ser complementada con conceptos y mediciones que originalmente pueden parecer subjetivos, y que representan un verdadero reto cuando tratan de ser representados en un modelo de medición cuantitativo.

Es por esto que el primer paso para poder atacar la calidad de los datos en la organización fue definir calidad de datos mediante la implementación de un modelo de dimensiones de datos del negocio, técnicas y dimensiones de calidad de datos.

En tanto que el negocio conozca y refine su comprensión sobre estos conceptos podrá convertir los modelos más subjetivos y cualitativos por observaciones objetivas y cuantitativas que le permitan orientar los recursos para resolución de problemas de calidad de datos hacia los problemas más cruciales que afectan la información para la toma de decisiones.

Las dimensiones de datos seleccionadas podrán a su vez ser alineadas a la visión de calidad del negocio permitiendo así especializar los análisis de calidad de datos a la realidad actual de la institución.

Estandarización de la estructura de los procesos de integración de datos.

La estructura interna de los procesos de integración de datos debía ser estandarizada para que estuviera acorde a los elementos de arquitectura que se habían definido. Dada la naturaleza incremental del proceso de desarrollo de estos procesos y la importancia de su mantenibilidad, fue necesario diseñar su estructura base bajo las recomendaciones en un marco teórico que

definiera formalmente los elementos mínimos que debían tomarse en cuenta dentro de cada tipo de proceso de carga.

Fue así como nos basamos en la teoría de procesos de integración de datos definidos por Giordiano, como un marco de referencia para la estructura base de los procesos de integración de datos de la Fábrica de Información Corporativa. A través de nuestra alineación con los elementos de esta propuesta, fueron evidentes las necesidades de estructuras especializadas para soportar aspectos como los procesos de diferenciación y estrategias de actualización de información, y se fueron quedando en evidencia los pasos algorítmicos necesarios para la ejecución de los procesos de integración de datos.

La elección de orientar la estructura interna de los procesos de integración a través la propuesta de Giordiano, nos proporcionó una organización y visibilidad mucho más clara de la relación entre los procesos de carga y los demás elementos de la CIF.

Alineación de la propuesta de estructuras de limpieza de datos al estándar de Informatica DataQuality

Aunque el modelo de procesos de integración seleccionado no menciona la utilización de herramientas de calidad en la etapa de calidad de datos y el área de limpieza, la implementación de este modelo nos permitió explotar de mejor manera las capacidades de la plataforma actualmente disponible dentro del BCR.

La razón de haber tomado esta decisión de diseño, consistió principalmente en el interés por explotar al máximo las herramientas disponibles en el BCR y de establecer un vínculo directo entre las cargas de datos y las decisiones del negocio sobre las mayores prioridades en calidad de datos.

Al brindar a los usuarios y a los técnicos acceso a los problemas más relevantes de calidad de datos dentro de la organización y a la posibilidad de tomar decisiones sobre las estrategias a aplicar para la mejora de los datos, brindamos un medio de comunicación directa entre la información, los procesos técnicos y el negocio, a la vez que dimos al negocio un importante insumo para que canalice los esfuerzos de calidad de los datos y aprovechemos la infraestructura existente en el BCR.

Generación de un proceso automatizado basado en metadatos para el seguimiento, limpieza y mantenimiento de la consistencia de los repositorios de la CIF.

La creación de este proceso, se realizó principalmente debido a la necesidad de brindar control y contexto temporal a los procesos de carga y la información. Era imperativo que existiera un control muy bien definido en el movimiento de la información, tanto para fines de aseguramiento de calidad como para dar seguimiento a la ejecución de cargas de datos.

Agregado a esto, también era necesario definir un proceso de limpieza del *Staging Area* y brindar capacidades de corrección de errores, para poder mantener la consistencia de los datos en caso de que se interrumpiera una carga de datos.

La utilización de un identificador para el proceso de carga que alteró un registro de datos es parte de la propuesta de Imhoff para modelado de datos dentro del *Data Warehouse*, en nuestro caso se hizo una adaptación de la misma para poder darle contexto a cada ejecución dentro de un conjunto de datos que hacen de alguna manera “consciente” a la CIF de cómo actúan los movimientos de los datos sobre sus repositorios, de manera que los procesos de limpieza pueden utilizar este conocimiento para mantener automáticamente la consistencia de la información con excelentes resultados hasta el momento.

Elección e implementación de una metodología de desarrollo de proyectos de Inteligencia de Negocios.

La iniciativa de la CIF inició en el Banco de Costa Rica, como un esfuerzo separado de la oficina de Control y Seguimiento de Proyectos, la cual siendo consciente de la necesidad de generar un mayor nivel de madurez de procesos de información para la toma de decisiones decidió alinear los esfuerzos que ya se habían realizado con un proceso mucho más ordenado para el desarrollo de la Inteligencia de Negocios dentro del BCR.

La oficina de Control y Seguimiento de Proyectos está dedicada a la coordinación y realización de los proyectos de Tecnología de Información dentro del BCR, sin embargo los proyectos desarrollados hasta el momento, han sido principalmente de creación o adquisición de herramientas e infraestructura para el manejo de las acciones operativas, y un proyecto de soporte a la toma de decisiones, tal como lo hemos demostrado a través de este proyecto, comprende un conjunto amplio y complejo de consideraciones que son particulares a este tipo de proyectos y que por tanto requieren de un tratamiento diferente y de una metodología de administración de proyectos adecuada a las particularidades de este entorno.

Habiendo tenido claro las necesidades particulares de este tipo de proyecto nos fue evidente que los enfoques de administración de proyectos clásicos no serían adecuados para ejecutar los proyectos de Inteligencia de Negocios, con esto en mente, se seleccionó la propuesta metodológica de Moss y Atre como una guía para generar una metodología propia para el BCR.

Si bien es cierto la metodología actualmente practicada no es exactamente correspondiente con la propuesta de Moss y Atre, esto se debe principalmente a la necesidad de formalizar los elementos e instrumentos que la metodología requiere y a la necesidad de generar a través de la experiencia las adaptaciones a la misma que son necesarias para que esta pueda ser aplicada dentro del marco de normativa que rige los procesos de administración de proyectos del BCR.

Diseño y administración de la capa semántica de consulta

La capa de aplicación es la que permite a los usuarios el utilizar los datos almacenados en la CIF para la generación de información útil para la toma de decisiones.

Esta capa es un reto por sí misma, porque implica la reorganización de la información en formatos y agrupaciones que estén en concordancia con los conceptos de negocios que son familiares a los usuarios de negocio, aún cuando las estructuras que derivaron del proceso de diseño e integración de datos puedan ser sustantivamente diferentes de los conceptos de negocio definidos inicialmente.

Para poder normar la construcción y crecimiento de las aplicaciones de consulta, se decidió utilizar la herramienta de SAP Business Objects XI, y a través de su característica de universos implementar una capa semántica que permitiera brindar a los usuarios una abstracción de la información en la CIF más cercana a los conceptos de negocio que acostumbran manejar.

La capa semántica se organiza en un conjunto de universos relacionados en una arquitectura por componentes, ya que se estima que este esquema de organización permite una alineación directa con las áreas temáticas, genera mayor mantenibilidad de la plataforma y permitirá realizar tanto universos específicamente diseñados para un fin como universos de referencia que se refieren a un tema particular.

ANÁLISIS Y DISCUSIÓN DE RESULTADOS

Los resultados obtenidos a través de la ejecución del proyecto han permitido brindar al BCR de una plataforma base lo suficientemente flexible para permitirle seguir extendiendo su Fábrica de Información Corporativa. Estamos seguros que si se continúa por este camino en muy poco tiempo los repositorios de información que se están alimentando en este momento se llegarán a convertir en un insumo imprescindible para la toma de decisiones en la organización.

Los resultados obtenidos, aunque satisfacen y en algunos casos hasta exceden los objetivos iniciales planteados, no pueden ser vistos como el objetivo final a nivel de la organización. Se han cubierto a satisfacción los aspectos centrales que permitirán continuar madurando en el tema y obtener experiencia en beneficio de la mejora en la interacción con los usuarios, la integración de los datos, y el entendimiento del negocio, sin embargo esto solo es la ruta a seguir, hay un largo trecho por recorrer antes de poder sacar el máximo provecho de esta clase de esfuerzos y antes de que la organización aprenda cuál es la mejor forma de llevar a cabo el desarrollo, ya que tal como pudimos apreciar durante la ejecución la metodología no puede ser totalmente impuesta sino que debe ajustarse al entorno en el que se implementa.

Los resultados obtenidos pueden resumirse en cuatro grandes áreas, arquitectura, integración, calidad de datos y metodología.

Enfoque de arquitectura.

Con respecto al enfoque de la arquitectura, se desarrolló ampliamente la orientación de la misma hacia un modelo de datos especialmente orientado a conceptos de negocio más que a estructuras de datos técnicas, independiente de la plataforma y de los orígenes de datos y que está conformada por áreas bien definidas, tales como *Staging Area* y sus componentes (*Initial Staging*, *Clean Staging*, etc.), el *Data Warehouse* y el repositorio de Metadatos.

Las relaciones entre los elementos de arquitectura quedan claramente establecidas y su papel dentro de la Fábrica de Información Corporativa mantiene concordancia con la propuesta de algunos de los autores más reconocidos en la materia.

La arquitectura se ha mantenido inmutable durante el tiempo que ha sido utilizada dentro del BCR, al momento de presentar esta memoria ya se está formando personal de algunas otras áreas para que sean partícipes de la aplicación y mejora de esta arquitectura y su utilización se convertido en un estándar dentro de la organización en espera de ser eventualmente introducida por normativa.

La arquitectura propuesta por el modelo de referencia (CIF de Inmon) no es aún desarrollada en su totalidad debido a la carencia inicial de los recursos suficientes y a la falta de madurez del negocio y la infraestructura para poder llevarla a un nivel superior, sin embargo dado el empuje del proyecto y la forma como se está proyectando a nivel de la organización estas

situaciones ya han sido o están siendo atendidas para que estas limitantes dejen de estar presentes.

Perspectiva de procesos de integración

Bajo la perspectiva del establecimiento de insumos y estandarización para la creación de los procesos de integración de datos dentro, hacia y desde la CIF, se ha cubierto un muy importante conjunto de elementos.

Los estándares establecidos se han basado en recomendaciones teóricas y de alto nivel propuestas por los mismos autores de la arquitectura, han sido complementados con técnicas de modelado específicamente diseñadas y probadas para soportar las características históricas de un *Data Warehouse*, se han alineado a las mejores prácticas reconocidas para la generación de procesos de integración de datos y se han enriquecido con las particularidades inherentes al ambiente productivo del BCR. Con lo cual hemos logrado establecer un orden y una mecánica adecuada para la integración de la información dentro de la organización con fines de soporte a la toma de decisiones.

Aún hoy quedan variadas oportunidades de mejora que debemos explorar, posibles mejoras en rendimiento y en tiempo de desarrollo, posibilidades aún no exploradas en las herramientas disponibles y nuevos conceptos y procedimientos para integración de datos que se han considerado deseables, sin embargo, lo implementado hasta ahora muestra ser lo suficientemente flexible y dinámico para brindarnos lo necesario para satisfacer nuestras necesidades actuales y las que visualizamos en un futuro cercano.

Aunque se ha hecho un esfuerzo importante por intentar establecer además una estrategia adecuada para cada tipo de proceso de carga, cada proceso tiene sus propias complicaciones y requiere de múltiples decisiones de diseño que tienen que ser continuamente retroalimentadas por el equipo de desarrollo de ETLs y alineadas a nuestros objetivos, la arquitectura y las capas que conforman los procesos de integración de datos.

Perspectiva de procesos de calidad de datos

La calidad de los datos como un reconocido factor crítico de éxito para cualquier proyecto de Inteligencia de Negocios, *Data Warehousing* y en general para sistemas de soporte a la toma de decisiones, no podía ser dejada en un segundo plano durante la ejecución del proyecto.

Los resultados obtenidos en esta perspectiva si bien de naturaleza primordialmente técnica han permitido acercar el concepto de Calidad de Datos al negocio, han establecido una forma estandarizada para evaluar la calidad y han colocado sobre la mesa el tema para que sea contemplado como una parte esencial del proyecto.

No podemos decir que se esté realizando a este momento todo lo necesario para mejorar y garantizar la mejor calidad de los datos posibles, pero lo realizado ha permitido clarificar el papel protagónico que la calidad de datos tendrá para garantizar la sostenibilidad de todo

esfuerzo realizado a través de la creación de la Fábrica de Información Corporativa, y de esta manera estamos impulsando la aparición de un proceso paralelo de mejora continua en los datos que a través de su inclusión en el *Data Warehouse* han demostrado ser el núcleo de información de mayor importancia dentro del BCR.

Enfoque metodológico.

Dado que es imposible que se pueda coordinar un trabajo de estas dimensiones sin un conjunto mínimo de herramientas metodológicas, y un proceso definido de pasos para lograr el desarrollo, en el futuro una de las mayores tareas que tendrá el equipo de Inteligencia de Negocios del BCR será la consolidación y mejora continua de la propuesta metodológica presentada en este proyecto.

Es así como la propuesta metodológica presentada es en sí mismo un proceso de desarrollo en espiral, donde conforme nuestra madurez y conocimiento lo van permitiendo complementamos nuestros procesos metodológicos con mejores instrumentos de análisis, con pasos apropiadamente ajustados a nuestra realidad y con nuevos aportes provenientes de las personas que se encuentran involucradas en cada proceso.

Es a través de la extensión y mejora de la metodología, en un proceso que no agregue solo complejidad, sino más bien completitud y sintetización de la misma a un entorno propio, que finalmente el BCR conseguirá un proceso de Inteligencia de Negocios maduro y completo con una fuerte base teórica y estándar que le permita evitar el re-trabajo y conseguir los mejores entregables posibles en la cantidad adecuada de tiempo.

Consideramos así, que los resultados obtenidos, más que satisfactorios para la consecución de los objetivos originalmente planteados, fueron la semilla sobre la que a partir del día de hoy se forjará una plataforma de Inteligencia de Negocios que apoyará al BCR en la consecución de los objetivos estratégicos, tácticos y operativos que se plantee y por tanto en la consecución exitosa de la misión y visión definidas por la organización, en beneficio tanto propio como del país, al mejorar sustantivamente su capacidad para reconocer en su información lo necesario para tomar las mejores decisiones posibles.

CONCLUSIONES Y TRABAJOS FUTUROS

Conclusiones:

1. La implementación del modelo de Fábrica de Información Corporativa dentro del BCR ha permitido dar orientación, a los esfuerzos de integración de datos y reportería analítica y para la toma de decisiones. La organización ha reconocido la necesidad de establecer un proceso más estructurado para soportar los procesos que antes se realizaban sólo de forma manual.
2. La propuesta realizada ha permitido caracterizar los componentes de la arquitectura y ha permitido brindar al negocio y a TI la comprensión adecuada que necesitan para coordinar esfuerzos en pro de la mejora continua de la plataforma de Inteligencia de Negocios.
3. La metodología propuesta ha ayudado efectivamente a realizar dos de las etapas de desarrollo definidas hasta el momento para la construcción CIF, cada utilización de la misma nos ha permitido mejorar los procesos que se realizan y ha facilitado el incremento en la madurez del personal técnico y del negocio en temas de integración y calidad de datos.
4. El éxito obtenido hasta el momento ha motivado al negocio a gestionar diversas formas de apoyo para el proyecto, entre las que se pueden mencionar la inclusión de nuevos recursos humanos dedicados, la designación de un dueño del producto y un patrocinador de alto nivel en el negocio y la visualización de la necesidad de nuevos recursos tecnológicos y de infraestructura para un futuro cercano.
5. Los estándares aplicados para la generación de procesos de integración de datos son han permitido construir efectiva y eficientemente más de sesenta procesos ETL, los cuales se encuentran ya implementados en ambiente de producción y en total han acumulado más de 17.000 cargas de datos correctas hacia los repositorios de la CIF.
6. El *Data Warehouse* es la fuente primaria de la que se alimentan o alimentarán otros sistemas de gran importancia en la organización, dado la visión central e integrada del negocio que este puede proveer.
7. Los instrumentos definidos para la evaluación de la calidad de datos han producido mejoras significativas en la información que es integrada a nivel del *Data Warehouse*, con tan solo el análisis realizado se han podido establecer muchas de las reglas de negocio que permiten que el nivel de calidad de la información sea mejor del que manejan las fuentes de datos. A la vez que nos han permitido señalar algunas situaciones anómalas en los sistemas orígenes con la consiguiente corrección de las causas raíz de los problemas de calidad y la consecuencia positiva de estas correcciones sobre los datos que llegan al *Data Warehouse*.
8. Los lineamientos definidos en cuanto a la forma de construir aplicaciones y la capa semántica que las soporta han permitido implementar un modelo adecuadamente estructurado y mantenible que permitirá extender la utilización de la CIF mientras se

desarrollan nuevas áreas temáticas y se mejora el contenido incluido en las ya existentes.

9. La realización de este proyecto ha brindado al BCR los insumos básicos iniciales para formar un Programa de Proyectos de Inteligencia de Negocios, que pueda responder efectivamente a las necesidades de generar información para análisis del negocio y toma de decisiones, y ha facilitado a la División de Tecnología un conjunto de elementos e instrumentos que podrán ser constantemente mejorados en beneficio de un mejor y más efectivo manejo de los recursos destinados a este tipo de proyectos, y de un mejor manejo de la información a nivel de todo el conglomerado BCR.

Trabajos futuros

1. Se requiere fortalecer mucho aún la parte de análisis de calidad de datos y explotar las herramientas de las que dispone el Banco en beneficio de la mejora de calidad de datos, se debe establecer una metodología de calidad de datos como TQDM (*Total Quality Data Management*) o una similar que al igual que la metodología de proyectos de Inteligencia de Negocios guíe la aplicación de un proceso corporativo de calidad de datos. En este sentido se ha conseguido una partida de presupuesto para ejecutar una contratación que será utilizada para desarrollar un conjunto de indicadores y medios automatizados para la medición de los niveles de calidad de datos de las fuentes de datos consumidas por *Data Warehouse*, y de los repositorios de la CIF, la ejecución de este presupuesto y de esta contratación forma parte de los trabajos futuros propuestos.
2. Es muy importante que la alineación a la metodología de proyectos de Inteligencia de Negocios continúe con el proceso continuo de mejora y formalización que ya ha comenzado, conforme esta vaya madurando deberá de asemejarse cada vez más a la metodología propuesta por Moss y Atre, esto permitirá que los procesos se hagan de manera cada vez más ordenada y controlada, mejorando los resultados obtenidos.
3. El modelo de áreas temáticas debe continuar su expansión hasta cubrir totalmente los principales productos del banco, siendo estos los ubicados en las categorías de captación y colocación, a la vez que deben de ejecutarse proyectos paralelos para integrar a las sociedades del banco dentro del *Data Warehouse*, este proceso ya ha sido iniciado y se ha obtenido el compromiso de estas a desarrollar por su cuenta, bajo nuestra tutela los procesos de carga y modelos de datos necesarios para que todo el conglomerado BCR se encuentre dentro del *Data Warehouse* Corporativo. Será nuestra responsabilidad el entrenar y mantener alineados a la arquitectura propuesta en este proyecto a estas áreas de negocio externas al Banco como tal.
4. Se debe invertir una cantidad importante de recursos y atención en desarrollar las aplicaciones para el acceso a la información de la CIF, ya que conforme estas aplicaciones puedan responder a las necesidades de los usuarios se estimulará el uso de la CIF dentro del conglomerado.
5. Se ha visualizado la necesidad de hacer mejoras en la infraestructura tecnológica que sostiene la CIF, para este fin se ha comenzado a coordinar con los departamentos de Redes y Servidores para hacer migraciones de las herramientas de integración hacia

una arquitectura virtualizada y se han comenzado las conversaciones con proveedores nacionales para adquirir un equipo específicamente diseñado para *Datawarehousing* que sustituya la base de datos relacional tradicional donde se encuentra en este momento el ambiente de producción, con el fin de poder brindar un menor tiempo de respuesta y una mayor capacidad de almacenamiento.

6. Se debe continuar formando al equipo de Inteligencia de Negocios del BCR, fortaleciendo sus habilidades técnicas y personales, instruyéndoles acerca de los aspectos más representativos de la arquitectura propuesta, convirtiéndolos en agentes de cambio y mejora continua y en analistas críticos que aporten al banco con su labor y conocimiento una mejor y más eficiente Fábrica de Información Corporativa.
7. Se debe promover un acercamiento cada vez más profundo con el negocio, con los dueños de la información, con las otras áreas de la División de Tecnología, y hacerlos conscientes de la importancia de ser partícipes activos de este esfuerzo, el cual representará para el BCR una muy importante ventaja competitiva y un muy importante esfuerzo en pro de la mejora en la efectividad de nuestros servicios y en beneficio de los clientes.
8. Se deben formalizar todos los instrumentos y metodología desarrollados, preferiblemente mediante la inclusión de estos dentro de la normativa del BCR, con el fin de hacerlos de acatamiento obligatorio y normado y garantizar que los mismos se encuentran disponibles para cualquier interesado en aplicarlos o mejorarlos de alguna manera a través del tiempo.

PLAN DE INVERSIÓN DEL PROYECTO

La inversión para la realización del proyecto puede incluir costos muy variados y complejos, los costos directos e indirectos aplicables a un proyecto de esta envergadura no son fácilmente estimables, sin embargo hemos realizado una aproximación con el fin de ilustrar las necesidades que han sido y deberían ser cubiertas para la realización del proyecto en los próximos años.

Dado que algunas de las fases ya se han implementado, estos costos no están pensados para ser aplicados inmediatamente, además la infraestructura y licenciamientos han sido reutilizados de proyectos anteriores, de alguna manera estos son recursos que se han adquirido y reutilizado por nosotros, para poder establecer este hecho se ha colocado junto al costo de estos elementos una indicación que explica la situación actual de este recurso.

Para realizar una estimación aproximada de los costos del proyecto para los próximos años, nos enfocamos en cuatro áreas principales:

- Inversión en capacitación.
- Licenciamiento y soporte de *Software*.
- Infraestructura de *Hardware* dedicada.
- Recursos humanos asignados y contratados al proyecto.
- Otras inversiones visualizadas a futuro.

Inversión en capacitación

Se han realizado importantes inversiones en capacitación para que los recursos involucrados puedan sacar el mayor provecho a las herramientas con que cuenta el Banco, para hacer más claro el detalle de costos se asocia entre paréntesis un número de referencia de manera que se entienda de donde se toma cada total, así podríamos detallar la inversión realizada en capacitación de la siguiente manera:

Ítem	Costo	Detalle
Capacitación en Informatica PowerCenter e Informatica DataQuality	\$86.000 (1) en tres temporadas (Capacitación 1: \$44.000 (1a) Capacitación 2: \$32.000 (1b) Capacitación 3: \$10.000 (1c))	Realizadas en tres tratos incluyó capacitación para un total de 20 personas, distribuidos en calidad de datos, integración básica e integración avanzada. De estos recursos asignados 4 lo fueron directamente al proyecto, y 5 se capacitaron para apoyar otras tareas.
Capacitación en SAP Business Objects XI 4 SP2	\$12.000 una única vez (2)	4 asistentes de los cuales 3 están asociados al proyecto directamente.
Otras capacitaciones recibidas	\$1.500 una única vez. (3)	4 asistentes de los cuales 2

están asociados al proyecto
directamente

La razón de la asignación de que los recursos asistentes a las capacitaciones sean más que los asignados al proyecto se debe a varias razones. En algunos casos se debe a la movilización del personal y en otras en las que las herramientas se visualizaron al momento de la capacitación para lograr otros fines.

Licenciamiento y soporte de *Software*

Ítem	Costo	Detalle
Licenciamiento Informatica PowerCenter e Informatica Data Quality.	\$600.000 inversión inicial (4) \$160.000 mantenimiento anual (5)	Incluye soporte y derecho a nuevas versiones, licenciamiento de ambientes de producción, pruebas y desarrollo y características adicionales para soportar alta disponibilidad en producción. Esta herramienta ya se encontraba dentro del inventario de licencias antes de iniciar con el proyecto.
Licenciamiento SAP Business Objects XI 4 SP2	\$160.000 inversión inicial (6) \$40.000 mantenimiento anual (7)	Incluye soporte y derecho a nuevas versiones, licenciamiento de ambientes de producción y desarrollo. Esta herramienta ya se encontraba dentro del inventario de licencias antes de iniciar con el proyecto.
Licenciamiento Embarcadero ER/Studio Data Architect XE2	\$35.000 una única vez.(8)	Se licenció y se mantuvo soporte por un año. El soporte se discontinuó posteriormente, pero las licencias siguen vigentes. Esta licencia se adquirió al inicio de la Etapa 1.
Oracle Database Standard Edition	\$70.000 inversión inicial.(9) \$15.400 mantenimiento anual.(10)	Incluye soporte y derecho a nuevas versiones, licenciamiento de ambientes de producción y desarrollo. Incluye además opción de Oracle RAC (Real Application Cluster) en ambos ambientes. Esta herramienta ya se encontraba dentro del inventario de licencias al iniciar con el proyecto.

El motivo para discontinuar el soporte a un licenciamiento Embarcadero ER/Studio se sostuvo en la mala evaluación del servicio prestado por el contratista que se hizo antes del vencimiento de la vigencia del contrato.

Infraestructura de *Hardware* dedicada.

Por otra parte la infraestructura de *hardware* que se dedica exclusivamente para el proyecto representan los siguientes costos para la organización.

Ítem	Costo	Detalle
Servidores de integración y calidad de datos (2 conectados en Clúster)	\$80.000 adquisición (11) \$7.500 mantenimiento anual (12)	Licenciamiento de Sun Solaris y adquisición y mantenimiento de equipo.
Servidores para base de datos (2 conectados en RAC)	\$100.000 adquisición (13) \$10.000 mantenimiento anual (14)	Licenciamiento de Sun Solaris y adquisición y mantenimiento de equipo.
Servidores para interfaces de consulta.	\$5.000 adquisición (15) \$4.000 mantenimiento anual (16)	Parte de una solución de virtualización empresarial.
Servidores para repositorio de modelos de datos	\$5.000 adquisición (17) \$2.000 mantenimiento anual (18)	Parte de una solución de virtualización empresarial.
Servidor de archivos FTP	\$5.000 adquisición (19) \$200 mantenimiento anual (20)	Parte de una solución de virtualización empresarial, que se comparte con otros proyectos.
Espacio en Infraestructura de Almacenamiento	\$5000 adquisición (21) \$4.000 mantenimiento anual (22)	Parte de una solución de almacenamiento empresarial SAN.

Recursos humanos asignados y contratados al proyecto.

Los recursos asignados al proyecto actualmente tienen este costo por concepto de salarios devengados anualmente.

Ítem	Costo	Detalle
Analista de Sistemas Junior	\$40.000 anuales (23)	Incluye el costo de 2 analistas asignados, en el año 2012 solo era uno de estos recursos.
Analista de Sistemas Senior	\$34.500 anuales (24)	Incluye el costo de 1.5 analistas asignados. En el año 2012 solo se dedicó un 50% de uno de estos recursos.
Coordinador de Proyectos Senior	\$6.400 anuales (25)	Incluye un coordinador de proyectos asignado 1/6 parte de tiempo aproximadamente.

Desarrolladores externos subcontratados	\$96.000 anuales (26)	Incluye una contratación por 1200 horas anuales a un costo de \$80 por hora, para generar procesos de evaluación y supervisión de calidad de datos en la plataforma. Solo aplica para el año 2013 en adelante.
Administradores de Bases de Datos Senior/Junior	\$4.000 anuales (27)	Incluye dos administradores de bases de datos que trabajan en algún aspecto del proyecto bajo solicitud aproximadamente dos horas por semana.

Otras inversiones visualizadas a futuro.

Dado que se han visualizado algunas importantes mejoras a realizar en el futuro, se incluye en este plan los costos asociados a la adquisición de cierto software y cierto hardware especializados que permitan mejorar el desempeño de la plataforma para ajustarse a las necesidades futuras.

Ítem	Costo	Detalle
Drivers de integración de alto desempeño.	\$40.000 anuales (28)	Estos serán requeridos para la optimización de las cargas que se realicen hacia un posible <i>Data Warehouse Appliance</i> .
Equipo Data Warehouse Appliance	\$200.000 inversión inicial (29) \$30.000 mantenimiento anual (30)	Los costos de iniciales incluyen soporte, instalación y acompañamiento inicial, los costos anuales son en soporte y derecho a nuevas versiones.

Usando como base los costos presentados para cada rubro se puede hacer un resumen de la inversión realizada y proyectada para los próximos años, que nos permite mostrar fácilmente la efectividad que tendrán las inversiones planeadas para los próximos años en la reducción del costo de operación de la Fábrica de Información Corporativa del BCR.

Año	Rubro	Monto inversión
Inversiones iniciales	Inversión en capacitación.	\$76.000 (1a,1b)
	Licenciamiento y soporte de <i>Software</i> .	\$830.000 (4,5,9)
	Infraestructura de <i>Hardware</i> dedicada.	\$200.000 (11,13,15,17,19,21)
	Recursos humanos asignados y contratados al proyecto.	\$0

	Otras inversiones visualizadas a futuro.	\$0
	TOTAL INICIAL:	\$1.074.000
2012	Inversión en capacitación.	\$42.000 (1b,2)
	Licenciamiento y soporte de <i>Software</i> .	\$250.400 (5,7,8,10)
	Infraestructura de <i>Hardware</i> dedicada.	\$27.700 (12,14,16,18,20,22)
	Recursos humanos asignados y contratados al proyecto.	\$41.900 (23,24,25,27)
	Otras inversiones visualizadas a futuro.	\$0
	TOTAL 2012:	\$362.000
2013	Inversión en capacitación.	\$10.000 (1c)
	Licenciamiento y soporte de <i>Software</i> .	\$215.400 (5,7,10)
	Infraestructura de <i>Hardware</i> dedicada.	\$27.700 (12,14,16,18,20,22)
	Recursos humanos asignados y contratados al proyecto.	\$180.400 (23,24,25,26,27)
	Otras inversiones visualizadas a futuro.	\$0
	TOTAL 2013:	\$433.500
2014	Inversión en capacitación.	\$0
	Licenciamiento y soporte de <i>Software</i> .	\$215.400 (5,7,10)
	Infraestructura de <i>Hardware</i> dedicada.	\$27.700 (12,14,16,18,20,22)
	Recursos humanos asignados y contratados al proyecto.	\$180.400 (23,24,25,26,27)
	Otras inversiones visualizadas a futuro.	\$240.000 (28,29)
	TOTAL 2014:	\$663.500
2015	Inversión en capacitación.	\$0
	Licenciamiento y soporte de <i>Software</i> .	\$255.400 (5,7,10,28)
	Infraestructura de <i>Hardware</i> dedicada.	\$43.700 (12,16,18,20,28)
	Recursos humanos asignados y contratados al proyecto.	\$180.400 (23,24,25,26,27)
	Otras inversiones visualizadas a futuro.	\$0
	TOTAL 2015:	\$479.500

Aunque los costos de operación actuales son aceptables para el banco es muy importante para nosotros el garantizar que los recursos son aprovechados de la mejor manera posible dada la naturaleza pública de los fondos, por esto es que a través de la interpretación del cuadro podemos ver que se planea hacer algunos cambios a nivel de la forma en que se ejecutará el presupuesto para el año 2014 y que hacia el año 2015 las medidas tomadas redundarán en un costo mayor del que se maneja actualmente, este aumento en la inversión se deberá especialmente a la implementación de un equipo de *Data Warehouse Appliance* el cual vendrá a sustituir la plataforma de Oracle actual, elimina algunos costos del rubro de mantenimiento de hardware, el cual podría usarse para otros proyectos, y se elimina el costo de almacenamiento en SAN a favor de almacenamiento independiente. Con la implementación del *Data Warehouse Appliance* obtendremos la ventaja adicional de una mejora significativa en el desempeño en gran medida la velocidad de respuesta de la plataforma, facilita en gran medida la administración y proporciona amplias posibilidades de extensibilidad en el futuro.

La principal razón de la adquisición de la plataforma de *Data Warehouse Appliance* se debe a que actualmente con los recursos disponibles se ha previsto la necesidad de hacer una mejora en la plataforma de desarrollo y producción, si decidiéramos mantener los equipos actuales y sólo cambiar de licenciamiento de Standard a Enterprise aplicarían los siguientes costos:

Item	Costo unitario	Cantidad	Costo Total
Adquisición	\$47.500	8 cores por equipo.	\$1.520.000
Licenciamiento Oracle Enterprise (se cobra por core)		4 equipos	
Adquisición de opción Real Application Clusters para Oracle Enterprise	\$23.000	2 RACs (uno para desarrollo otro para producción).	\$46.000
Adquisición	\$11.400	8 cores por equipo.	\$364.800
Licenciamiento Opción de Particionamiento (se cobra por core)		4 equipos	
TOTAL ADQUISICIONES			\$1.930.800
MANTENIMIENTO ANUAL (22%)			\$ 424.776

Estos costos se vuelven prohibitivos en contraposición a los \$200.000 de adquisición del *Data Warehouse Appliance* y los \$70.000 que implica esta adquisición en los rubros de mantenimiento anual, máxime si se considera que el *Appliance* más básico incluye redundancia en todos sus componentes, 10TB de almacenamiento, 16 núcleos y resolución de consultas a través de hardware.

Con esto en mente se puede concluir que el plan de costos asociados al proyecto se mantiene en un nivel aceptable y que las proyecciones del mismo al futuro permiten garantizar una

inversión adecuada de recursos en pro de la mejora de la infraestructura, el servicio y el desarrollo de la Fábrica de Información Corporativa del BCR.

BIBLIOGRAFÍA

1. [IMHOFF, 2003] IMHOFF, Claudia; GALEMMO, Nicholas; GEIGER, Jonathan. *Mastering Data Warehouse Design*. New York, NY: Wiley Publishing, 2003. 438 p. ISBN:0-471-32421-3
2. [INMON, 2001] INMON, William; IMHOFF, Claudia; SOUSA, Ryan. *Corporate Information Factory*. New York, NY: Wiley Publishing, 2001. 400 p. ISBN: 978-0-47-139961-2
3. [INMON, 2008] INMON, William; STRAUSS, Derek; NEUSHLOSS, Genia. *DW 2.0 The Architecture for the Next Generation of Data Warehousing*. Burlington,MA: Morgan Kaufmann Publishing, 2008. 371 p. ISBN: 978-0-12-374319-0
4. [KIMBALL, 2002] KIMBALL, Ralph; ROSS, Margy. *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling*. New York, NY: Wiley Publishing, 2002. 464 p. ISBN: 978-0-47-120024-6
5. [KIMBALL, 2008] KIMBALL, Ralph; ROSS, Margy; THORNTHWAITTE, Warren. *The Data Warehouse Lifecycle Toolkit*. New York, NY: Wiley Publishing, 2008. 672 p. ISBN: 978-0-47-014977-5
6. [MOSS , 2003] MOSS, Larissa; ATRE, Shaku. *Business Intelligence RoadMap The Complete Project Lifecycle for Decision-Support Applications*. Boston, MA: Addison Wesley, 543 p. ISBN:0-201-78420-3.
7. [GIORDANO, 2011] GIORDANO, Anthony David. *Data Integration Blueprint and Modeling Techniques for a Scalable and Sustainable Architecture*. IBM Press. 384p. ISBN: 978-0-13-708493-7

DEDICATORIA

“A mi familia que de una u otra forma se mantiene siempre conmigo.”

“A mi profesor tutor, que además de serlo en esta oportunidad, fue la persona que, con su conocimiento y su amistad, sembró en mi la admiración hacia esta disciplina.”

“A todas esas personas que me apoyaron de una u otra manera ya sea con su conocimiento, sus recomendaciones, su trabajo, o por sus palabras de apoyo, que por ser muchas no puedo mencionar, pero para las que deseo lo mejor en su vida.”

“Al Banco de Costa Rica y sus colaboradores, quienes no dejan de mostrarme que es la excelencia no es un acto sino una costumbre.”

“A mi Dios por darme la vida y por permitirme encontrar un trabajo y una pasión que me mantiene despierto aún en estas altas horas de la noche.”