CONSTRUCCION DE UN DATA MART PARA EL CALCULO DE INDICADORES DE CALIDAD DEL SERVICIO EN EL ÁREA DE GESTION OPERATIVA DE LA CHEC S.A. E.S.P.

OSCAR MAURICIO TRUJILLO PULECIO

TRABAJO DIRIGIDO POR:

JAIRO IVÁN VÉLEZ BEDOYA

UNIVERSIDAD AUTÓNOMA DE MANIZALES

FACULTAD DE INGENIERIA

MAESTRÍA EN GESTIÓN Y DESARROLLO DE PROYECTOS DE SOFTWARE

MANIZALES

2018

CONSTRUCCION DE UN DATA MART PARA EL CALCULO DE INDICADORES DE CALIDAD DEL SERVICIO EN EL ÁREA DE GESTION OPERATIVA DE LA CHEC S.A. E.S.P.

OSCAR MAURICIO TRUJILLO PULECIO

TRABAJO DIRIGIDO POR:

JAIRO IVÁN VÉLEZ BEDOYA

UNIVERSIDAD AUTÓNOMA DE MANIZALES

FACULTAD DE INGENIERIA

MAESTRÍA EN GESTIÓN Y DESARROLLO DE PROYECTOS DE SOFTWARE

MANIZALES

2018

RESUMEN

La acelerada evolución organizacional a nivel mundial, ha traído consigo que se deban almacenar y manejar grandes volúmenes de datos, por lo que cada vez se hace más costoso mantenerlos en los sistemas de información transaccionales. Para el caso de las empresas del sector público, los entes reguladores han elevado los estándares de calidad del servicio buscando que las compensaciones a los usuarios finales sean cada vez menores y que los incentivos brindados por el gobierno sean justificados con datos y hechos históricos. Estas exigencias se hacen efectivas a través de indicadores y reportes que demuestren la gestión realizada.

Pero el procesamiento de estos datos no puede interrumpir la operación normal de las empresas, por lo que se requiere tener sistemas alternos que permitan procesar grandes volúmenes de datos con tiempos de respuesta bajos que además apoyen la toma de decisiones a nivel organizacional.

Es por estos que los almacenes de datos han jugado un papel importante en los últimos años, pues permiten que la información, requerida para el cálculo de indicadores y la toma de decisiones empresariales, esté consolidada en un mismo repositorio de acuerdo a los datos recolectados en los diferentes sistemas transaccionales.

En este trabajo se expone la creación de un *Data Mart* para realizar el cálculo de los indicadores de calidad del servicio en el área de Gestión Operativa de la CHEC, el cual surge por las necesidades regulatorias y por la preocupación frente al impacto reflejado en los ingresos y/o egresos por concepto del desempeño de la calidad del servicio de energía.

Se utilizará como marco de trabajo la metodología de Ralph Kimball con apoyo, en algunas etapas, de la metodología HEFESTO propuesta por el Ing. Ricardo Darío Bernabeu, la cual permite obtener resultados más tempranos al aplicar buenas prácticas descritas en esta última metodología.

Palabras Clave: Almacén de datos, Metodología Kimball, Metodología Hefesto, Procesos ETL

ABSTRACT

The intensive organizational evolution worldwide has brought the necessity of storing and handling big volumes of information, by what every time it becomes more expensive to support them in the transactional information systems. For the case of public sector companies, the regulatory entities have raised the quality standards of the service looking that the compensations to the final users are every time minors and that the incentives offered by the government are justified by information and historical facts. These requirements become effective across indicators and reports that demostrate the management realized.

But the processing of this information cannot interrupt the normal operation of the companies, by what it is needed to have alternate systems that allow to process big volumes of information with low times of response that in addition support the organizational decision-making process.

It is for this reason that the stores of information have played an important role in the last years, since they allow the information needed for the calculation of indicators and the capture of managerial decisions to be consolidated in the same repository of agreement to the information gathered in the different transactional systems.

This work shows the creation of a Data Mart to deal with the calculation of the service quality indicators in the Operative Management area from CHEC, which arises due to the regulative needs and the concern over the impact reflected in the income and/or expenditures for concept of the performance of the quality of the energy service.

Ralph Kimball's methodology will be in use as frame of work with support, in some stages, of the methodology HEFESTO proposed by the Ing. Ricardo Dario Bernabeu, which allows to obtain earlier results on having applied good practices described in the latter methodology.

Key Words: Data mart, Kimball Methodology, Hefesto Methodology, ETL processes

TABLA DE CONTENIDO

LISTA DE TABLAS

Lista de figuras

1 INTRODUCCIÓN

En el presente documento se estructura el proyecto de creación de un *data mart* que permita realizar el cálculo de los indicadores de calidad del servicio en el área de Gestión Operativa de la CHEC.

La construcción de un *data mart* con estas características, surge por las necesidades regulatorias y por la preocupación frente al impacto reflejado en los ingresos y/o egresos por concepto del desempeño de la calidad del servicio de energía. Esto impacta directamente en el esquema de incentivos y compensaciones indicado en la resolución CREG 097 del 2008 (Comisión de Regulación de Energia y Gas - CREG, 2008).

Para la construcción del *data mart* se utilizará como marco de referencia la metodología "Kimball" la cual es utilizada para la construcción de Data Warehouses. Esta metodología provee un enfoque botton-up, es decir, de menor a mayor, lo cual permite realizar entregas pequeñas incrementales, simplificando la complejidad de implementar el *data mart.*

Este marco de referencia se complementará con la metodología HEFESTO, que se basa en la comprensión de cada paso que se ejecutará, y no de seguir un método al pie de la letra sin saber exactamente qué se está haciendo, ni por qué.

Con estas metodologías se crearán artefactos, los cuales se podrían incorporar en el proceso para la construcción de futuros desarrollos de *data marts* en CHEC.

2 ÁREA PROBLEMÁTICA

La Central Hidroeléctrica de Caldas “CHEC” es una empresa que presta el servicio público de energía eléctrica mediante los negocios de generación, distribución y comercialización en los departamentos de Caldas y Risaralda, exceptuando Pereira.

Los entes reguladores como la Superintendencia de Servicios Públicos Domiciliarios (SSPD) y la Comisión de Regulación de Energía y Gas (CREG), a través de diferentes resoluciones, definen indicadores que miden la calidad del servicio prestada por las empresas del sector eléctrico, por lo cual se hace necesario disponer de herramientas informáticas que faciliten el almacenamiento de los datos, el cálculo de los indicadores y la generación de reportes que aporten a la gestión de la operación diaria.

De acuerdo a la Resolución 097 (Comisión de Regulación de Energia y Gas - CREG, 2008) se establecen los principios generales y la metodología para el cálculo de los cargos por uso y las reglas que deben cumplir los Operadores de Red (OR) en cuanto a la calidad en la prestación del servicio de distribución de energía evaluando la Calidad Media brindada por el OR a sus usuarios y comparándola contra una Calidad Media de Referencia, estableciendo un esquema de incentivos y compensaciones.

El esquema de incentivos se aplicará a cada OR basado en la cantidad de Energía No Suministrada (ENS) durante un trimestre específico, de manera respectiva disminuirá su cargo por uso en el correspondiente nivel de tensión o lo aumentará durante el trimestre inmediatamente siguiente a la evaluación.

Los sistemas de Transmisión Regional y Distribución Local se clasifican por niveles de tensión, en función de la tensión nominal de operación, según la siguiente definición:

- Nivel 4: Sistemas con tensión nominal mayor o igual a 57,5 kV y menor a 220 kV.
- Nivel 3: Sistemas con tensión nominal mayor o igual a 30 kV y menor de 57,5 kV.
- Nivel 2: Sistemas con tensión nominal mayor o igual a 1 kV y menor de 30 kV.
- Nivel 1: Sistemas con tensión nominal menor a 1 kV.

El esquema de incentivos se complementará con el esquema de compensaciones a los usuarios “peor servidos”, el cual busca disminuir la dispersión de la calidad prestada por el OR en torno a la calidad media, garantizando así un nivel mínimo de calidad a los usuarios.

En esta misma resolución se incrementan las exigencias en calidad del servicio hasta el punto de presentar afectaciones al reconocimiento de los ingresos del Operador de Red por efecto de la Administración, Operación y Mantenimiento (AOM) de los activos de la red eléctrica, obligando a que los Operadores de Red revalúen los lineamientos, políticas, reglas y procedimientos de ingeniería, análisis, mantenimiento, operación y reposición que le permitan ser más eficientes y puedan cumplir con las exigencias regulatorias.

A finales del 2012, se realizó un trabajo con los profesionales de la Subgerencia de Trasmisión y Distribución (T&D) para la construcción de un árbol de problemas e identificación de las causas que afectan la calidad del servicio. Este trabajo se basó en la experiencia de la operación del servicio y en la información que en ese momento arrojaban los sistemas de información. Dentro de los resultados se detectaron los siguientes problemas:

- ÍNDICES-INFORMACIÓN: Se detectó que los indicadores regulatorios no daban las señales del comportamiento del sistema eléctrico.
- RURALIDAD: El alto componente rural que dificulta las labores de mantenimiento y reparación.
- HERRAMIENTAS INFORMÁTICAS: Necesidad de herramientas informáticas que permitan realizar simulaciones, costeo de eventos, localización de fallas, ubicación de cuadrillas, etc.
- EQUIPOS: Automatización de la red insuficiente para disminuir las indisponibilidades.
- PROCESOS: Falta de capacitación constante, y revisión de procedimientos.

Del 2012 a la fecha, CHEC ha invertido recursos tanto económicos como humanos para poder implementar un plan cuyo objetivo es el mejoramiento de la calidad del servicio. Sin

embargo, uno de los puntos que se han evidenciado como críticos es que no se cuentan con suficientes herramientas informáticas que faciliten el análisis de los volúmenes de información con los que actualmente cuenta la empresa en torno a esta necesidad.

A la fecha, las herramientas existentes carecen de un lenguaje y almacenamiento común que permita tener disponibles todos los datos de la operación y mantenimiento de los equipos de la red eléctrica. A esto se suma que los datos se encuentran dispersos en diferentes fuentes de información, y como consecuencia se pueden encontrar datos duplicados y en ocasiones inconsistentes, por lo que no es fácil mantener su integridad. Tal diversidad se convierte en un obstáculo de cara al usuario final frente a la disponibilidad y oportunidad de la información.

El proyecto a realizar busca facilitar la toma de decisiones respecto a la calidad del servicio con la implementación de una herramienta informática en la cual se incorporarán técnicas para el análisis de datos, lo cual apalanca los siguientes objetivos estratégicos de CHEC que se mencionan en la Figura 1:

- Optimizar procesos
- Optimizar y consolidar los sistemas de información para la toma de decisiones en el Grupo

Figura 1 Mapa de Objetivos Estratégicos de CHEC (Central Hidroeléctrica de Caldas - CHEC, 2015)

3 JUSTIFICACIÓN

Este trabajo será un aporte importante a la tesis de doctorado de la Ingeniera Mónica Rosa López de la Universidad Nacional de Colombia sede Manizales (Lopez Guayasamín, Castrillón, & Cano, 2016), quien propone un modelo basado en minería de datos para medir los costos de administración, operación y mantenimiento en los transformadores de distribución de acuerdo a las fallas presentadas en estos. La implementación del proyecto se hará en la CHEC. El componente que se propone desarrollar en la presente tesis permitirá disponer de una base de datos multidimensional creada a partir de los datos recientes e históricos del mantenimiento y operación de la red eléctrica que facilitará la extracción de datos a través de reportes e indicadores, además de la generación de cubos de datos para apoyar la toma de decisiones en la calidad del servicio. Este último proceso será desarrollado en la tesis de doctorado.

Con este trabajo también se busca dar las herramientas suficientes a los usuarios del Área de Gestión Operativa para que dispongan de la información requerida en su gestión diaria y en la toma de decisiones, de tal manera que contribuya a la mejora en la calidad del servicio optimizando la programación de los mantenimientos y atendiendo oportunamente los daños y reparaciones con el fin de garantizar la continuidad del servicio de energía eléctrica a los clientes de CHEC.

En el Área de Gestión Operativa se encuentra el proceso de Gestión de Información, en donde se realizan actividades relacionadas con la medición de la calidad del servicio, y que son de alto impacto en la toma de decisiones para la operación diaria y que apalancan el objetivo estratégico de “Optimización procesos” (Central Hidroeléctrica de Caldas - CHEC, 2015). La información que entrega el proceso debe cumplir con las características de confidencialidad, integridad, disponibilidad, trazabilidad y no repudio, lo que implica que los sistemas de información que la aportan deben estar preparados para cubrir las necesidades que puedan surgir.

La disponibilidad de la información es un factor determinante en la toma de decisiones, por lo cual se deben proporcionar herramientas e implementar procedimientos suficientes para que el negocio pueda obtener y disponer de los datos para su gestión.

4 OBJETIVOS

4.1 OBJETIVO GENERAL

Construir un data mart que le permita al área de Gestión Operativa de CHEC centralizar los datos para el cálculo de indicadores y la toma de decisiones en la calidad del servicio de energía eléctrica.

4.2 OBJETIVOS ESPECÍFICOS

- Utilizar las metodologías Kimball y Hefesto para la construcción del data mart basado en las buenas prácticas que ofrecen estas metodologías.
- Determinar cuáles son los datos requeridos para el cálculo de indicadores, la gestión y toma de decisiones en la calidad del servicio, a través de los formatos para el levantamiento de requerimientos de la metodología Hefesto.
- Diseñar la Arquitectura de datos para el data mart, tomando como referencia el estado del arte de las herramientas Microsoft SQL Server Integration Services (SSIS), Oracle Warehouse Builder, Pentaho Data Integration.
- Diseñar el proceso de extracción, transformación y carga de datos que aportan en la medición de la calidad del servicio.
- Generar los indicadores que permitan medir la calidad de los datos en el data mart para cumplir con lo estipulado en la Resolución 097 de 2008[1].

[1] Comisión de Regulación de Energia y Gas - CREG. (2008). *Resolución 097.* Bogotá.

5 REFERENTE TEÓRICO

5.1 MARCO CONTEXTUAL

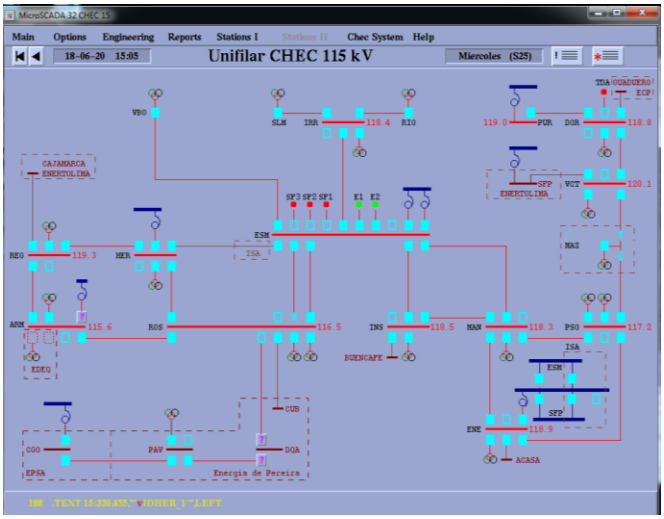
La medición de la calidad del servicio en CHEC es realizada por el Área de Gestión Operativa, en el equipo de Gestión de la Información, quienes toman los datos registrados en la operación diaria para generar los reportes e indicadores que se envían a los entes reguladores.

Los datos actuales salen de distintos sistemas de información transaccionales que apoyan los procesos del equipo de Operación Integrada. Al tener distintas fuentes de dato se dificulta consolidar la información y hacer análisis sobre ellos. A esto hay que sumarle que se manejan diferentes motores de bases de datos como Oracle y Postgres por lo que se deben extraer reportes y consolidarlos en Excel en algunas ocasiones. Los principales sistemas de información de los cuales se extrae información en calidad del servicio son:

SCADA

Supervisory Control And Data Acquisition (Control de Supervisión y Adquisición de Datos). Este sistema integra los IED (equipos de control y protección numéricos) recibiendo señales de campo y ejecutando comandos sobre ellos. Permite hacer la supervisión, coordinación y ejecución de la operación y es el mayor generador de información del sistema eléctrico de la empresa.

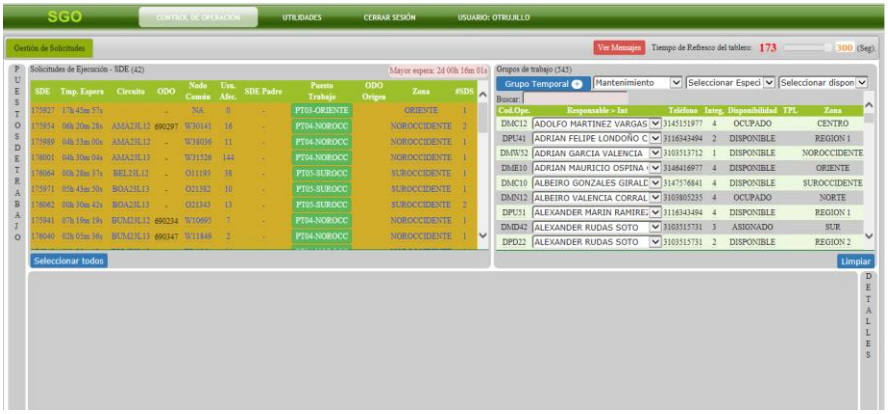
Figura 2 Interfaz Gráfica de una Subestación en el sistema SCADA



SGO

El Sistema de Gestión de la Operación SGO, permite realizar la operación diaria de acuerdo a los eventos que se ingresan en el sistema SCADA y las maniobras que se registran manualmente y se gestionan desde campo. Acá se ingresan los tiempos de las órdenes de operación ejecutadas por los grupos de trabajo incluyendo las novedades reportadas sobre la red eléctrica, además de las llamadas que entran por el Contact Center para reportar daños.

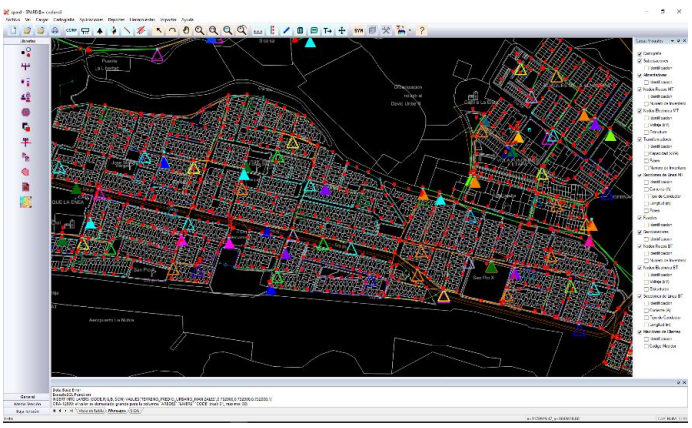
Figura 3 Tablero de Control SGO



SPARD

El Sistema Para Administración de Redes de Distribución SPARD, permite almacenar información de los activos de la red, sus atributos y relaciones topológicas. En este sistema los equipos y sus características son clasificados como tablas de inventarios entre las cuales están transformadores, interruptores y/o seccionadores, alimentadores, tramos de línea de media y baja tensión, clientes y subestaciones.

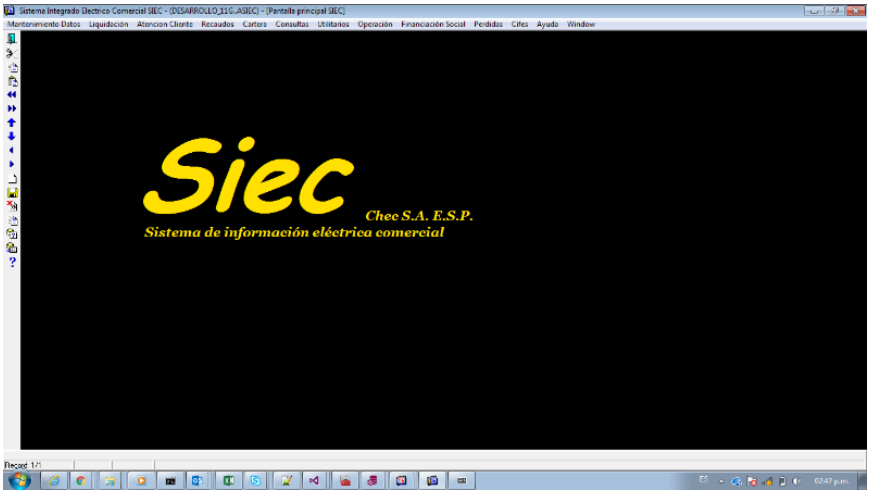
Figura 4 - Interfaz Gráfica SPARD



SIEC

El Sistema de Información Eléctrica Comercial SIEC, permite facturar los consumos de energía y el cobro por cuotas de conceptos asociados a energía (materiales). Recibe todas las novedades de cambios de los datos básicos de los clientes, lecturas de contadores, costos de prestación del servicio (tarifas), interrupciones del servicio y recaudos. Calcula y liquida consumos y otros conceptos, liquida compensaciones asociadas a la Calidad del Servicio y aplica recaudos y beneficios. Como salida del sistema está la factura de venta del servicio, reportes para la suspensión del servicio, información estadística para uso interno y para entes externos (SUI, CREG, Ministerio de Minas, SSPD).

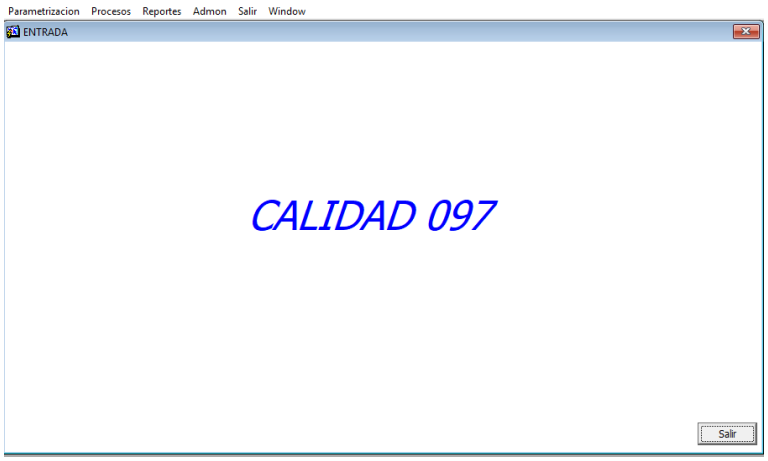
Figura 5 - Interfaz SIEC



CALIDAD097

En este sistema se generan los reportes y formatos de calidad del servicio que se deben cargar al sistema Liquidador y Administrador de Cuenta (LAC) y a la SSPD (aplicativo SUI). También se consolida la información de eventos, maniobras que se generan en la operación diaria ejecutando el algoritmo aguas abajo que calcula las indisponibilidades del servicio a usuarios finales.

Figura 6 - Interfaz CALIDAD097



Data Warehouse y Data Mart en el sistema eléctrico

La incorporación de almacenes de datos en el sector eléctrico, ha tenido como principales promotores a las entidades reguladoras de servicios públicos domiciliarios de cada país. El cálculo de indicadores y la generación de informes a éstos entes ha demandado el movimiento de grandes volúmenes de datos para demostrar que históricamente se ha hecho alguna gestión en mejorar la calidad del servicio de energía eléctrica.

A su vez, ha surgido una necesidad en común por obtener un conocimiento profundo de los clientes, el cual permita definir estrategias que faciliten la prestación del servicio a casi la totalidad de la población objetivo, teniendo en cuenta sus capacidades económicas y necesidades básicas.

A continuación, se exponen algunos trabajos realizados en el sector eléctrico que ayudan a comprender mejor el enfoque de los almacenes de datos:

TITULO: “Uso de DataMart en la distribución de electricidad”

AUTORES: Ignacio Gabriel Hanesman

CONCLUSIONES: Este texto muestra los diferentes usos que se le pueden dar a un Data Mart en una empresa de servicios públicos de energía eléctrica. Se mencionan varios ejemplos en donde se identifican posibles modelos a desarrollar, generando valor apoyando procesos de análisis y toma de decisiones al interior de la empresa. Estos modelos propuestos brindan a los usuarios la posibilidad de estudiar íntegramente el negocio, sin depender de los tiempos y estructuras del área de sistemas. Además, se menciona la posibilidad de utilizar esta herramienta combinándola con herramientas ofimáticas como Excel.

TITULO: "SISTEMA DE MONITOREO DE MERCADO ELÉCTRICO"

AUTORES: Juan Francisco González Obreque

AÑO: 2013

CONCLUSIONES: Este proyecto se desarrolló en Chile, con la finalidad de consolidar información del sistema de monitoreo de mercado eléctrico que dispone el Ministerio de energía de ese país. Dado que la información de proyectos de energía eléctrica se mostraba de forma muy detallada y plana, se promovió una oportunidad para desarrollar nuevas formas de visualizar gráficamente esta información por parte de las empresas del sector brindándoles un aporte en la comprensión del estado del sistema energético en el país. A través de procesos ETLs se transmiten los datos desde el sistema del Ministerior a una base de datos local. Finalmente se definen unos indicadores bases que se utilizarán para mostrar la información obtenida y almacenada en la base de datos local.

TITULO: "Desarrollo de un sistema de información ejecutivo e implementación de un data warehouse para la gestión de indicadores en una empresa eléctrica distribuidora"

AUTORES: Francisco Joseph Bolaños Burgos

Margarita del Rocío Filián Gómez

Gonzalo Patricio Maldonado Asanza

AÑO: 2009

CONCLUSIONES: Este trabajo es desarrollado en Ecuador, y plantea la elaboración de dos sistemas: Un Sistema de Información Ejecutivo y un Data Warehouse que es usado por el primer sistema para enviar información a organismos de control. No se utiliza una metodología estándar para la creación del Data Warehouse, por lo tanto, se hizo un desarrollo por prototipos.

TITULO: "Análisis y diseño del modelo de información del sector eléctrico ecuatoriano relacionado con la distribución y comercialización de energía eléctrica"

AUTORES: Danny Patricio Andrade Cárdenas

Francisco Javier Carrasco Astudillo

Fernando Patricio Espinoza Encalada

AÑO: 2007

CONCLUSIONES: Este trabajo muestra el análisis y diseño del modelo de información del sector eléctrico ecuatoriano relacionado con la distribución y comercialización de energía eléctrica. Se realizaron entrevistas con funcionarios de tres empresas del sector para conocer el modelo de negocio desde las perspectivas estratégicas, tácticas y operativas para detectar requerimientos de información para la toma de decisiones. Con este resultado se crearon cinco Data Marts, los cuales compondrán el Data Warehouse corporativo.

Los trabajos expuestos anteriormente se fundamentan en mejorar la calidad del servicio del sistema eléctrico en diferentes países, ya sea buscando resultados de la operación que se reflejan en los indicadores internos de cada empresa o en el ámbito regulatorio, en donde cada país genera leyes que controlan el accionar de las empresas prestadoras de este servicio público domiciliario. Este último componente pretende potencializar las empresas

prestadoras del servicio de energía para que puedan compararse con empresas de otros países, midiendo su calidad con indicadores estándares internacionales y de gran acogida a nivel mundial.

5.2 MARCO NORMATIVO

La transformación del Sector Eléctrico Colombiano dirigida por las Leyes 142 y 143 de 1994, cambia la orientación de la disponibilidad del servicio de energía como factor principal en la prestación del mismo y realza criterios de eficiencia y calidad como parte fundamental del desarrollo del sector, reflejándose en beneficio para los usuarios: *"Garantizar la calidad del bien objeto del servicio público y su disposición final para asegurar el mejoramiento de la calidad de vida de los usuarios."*[1]

La regulación colombiana a través de la Comisión de Regulación de Energía y Gas (CREG), entidad encargada de regular las actividades de los servicios públicos en Colombia, ha venido emitiendo diversas resoluciones con el fin de que se cumpla de manera confiable y eficiente la prestación de este servicio para los usuarios. Uno de los aspectos en los cuales se ha hecho mayor énfasis, es en la continuidad del suministro de energía eléctrica, que afecta muchísimo a los usuarios particularmente a los industriales por el alto impacto económico que las interrupciones de mucha o poca duración representan.

En la actualidad, la resolución vigente en Colombia que regula la calidad del servicio de energía eléctrica es la CREG 097 de 2008 (Esquema de Incentivos y compensaciones de calidad del servicio), sin embargo, la agenda regulatoria para el año 2019 tiene estimado la entrada en vigencia de la resolución CREG 015 de 2018, la cual modifica el esquema de incentivos y compensaciones y los indicadores de la calidad del servicio.

[1] Ley 142 de 1994

El esquema de incentivos y compensaciones por calidad del servicio de energía eléctrica según la CREG 097 de 2008, funciona de la siguiente manera:

- Sí el operador de red cumple con los indicadores de calidad del servicio de energía eléctrica durante el trimestre de evaluación, es incentivado con la posibilidad de obtener mayores ganancias por su servicio de calidad a través del cargo D.
- Sí el operador de red no cumple con los indicadores de calidad del servicio de energía eléctrica durante el trimestre de evaluación, tendrá que compensar a los usuarios o disminuir el cargo de Distribución (componente D).

La evaluación de los indicadores de calidad del servicio se hace trimestralmente, para el cálculo de cada uno de estos indicadores de calidad, el Operador de Red (OR) debe reportar al Sistema Único de Información (SUI) la cantidad de interrupciones y sus respectivas duraciones por cada transformador o circuito, al igual que la cantidad de usuarios conectados a la red eléctrica y la información comercial de cada uno de ellos.

Con la información suministrada por el OR se realiza el cálculo del Índice Trimestral Agrupado de la Discontinuidad (ITAD), el cual se calcula según un promedio entre la energía que dejaron de consumir los usuarios por las interrupciones del servicio (demanda interrumpida) respecto a la energía que consumieron los usuarios de las compañías trimestralmente (demanda suministrada).

El ITAD es un indicador que reemplazó y unificó dos términos que se manejaban antes, el DES (Duración de las Interrupciones en Horas) y el FES (Frecuencias de las Interrupciones en número de Veces), y de acuerdo con sus resultados permitía establecer la calidad del servicio.

La Resolución CREG 015 del 2018 modifica nuevamente el cálculo de indicadores de calidad incorporando dos índices internacionales como lo son SAIDI y SAIFI.

El indicador SAIDI representa la duración total en horas de los eventos que en promedio percibe cada usuario de un OR, hayan sido o no afectados por un evento, en un período

anual. El indicador SAIFI representa la cantidad total de los eventos que en promedio perciben todos los usuarios del SDL de un OR, hayan sido o no afectados por un evento, en un período anual [1].

A continuación, se listan las principales resoluciones que rigen la calidad de la energía eléctrica en Colombia:

Tabla 1 Principales resoluciones que rigen la calidad del servicio de energía en Colombia (autoría propia)

RESOLUCION	SINTESIS	AUTOR	FECHA
Resolución 070-1998	Se adopta el reglamento de distribución eléctrica que trata un poco sobre las definiciones, plan de expansión, condiciones de conexión, operación, calidad, propiedad de los sistemas de transmisión regional, medida y alumbrado público.	ORLANDO CABRALES MARTINEZ (MINISTRO DE MINAS Y ENERGIA) - JORGE PINTO MOLLA (DIRECTOR EJECUTIVO)	28 DE MAYO - 1998
Resolución 025-1999	Modifica el numeral 6.3.2.1 de la resolución CREG 070-1998, el cual trata sobre los indicadores para el periodo de transición, que se medirá al nivel de circuito con base a los siguientes indicadores: Duración equivalente de interrupciones del servicio (DES). Frecuencia de interrupciones de servicio (FES).	LUIS CARLOS VALENZUELA (MINISTRO DE MINAS Y ENERGIA) – JOSE CAMILO MAZUR (DIRECTOR EJECUTIVO)	9 DE JUNIO - 1999

[1] Resolución CREG 015 de 2018

Resolución 058-2000	Aplica a los comercializadores de energía eléctrica y a los distribuidores-comercializadores de gas combustible que atienden usuarios finales de este servicio. Estos publicarán de forma simple y comprensible, las tarifas que aplicaran a los usuarios, en un periodo de amplia circulación en los municipios donde prestan los servicios, o en una circulación nacional.	JUAN MANUEL ROJAS PAYAN (VICEMINISTRO DE ENERGIA) - CARMENZA CHAHIN (DIRECTORA EJECUTIVA)	17 DE AGOSTO - 2000
Resolución 096-2000	Se dictan normas relacionadas con el periodo de transmisión de que trata el reglamento de distribución de energía eléctrica, y se complementan algunas disposiciones de dichas resoluciones.	CARLOS CABALLERO ARGAEZ (MINISTRO DE MINAS Y ENERGIA) - CARMENZA CHAHIN (DIRECTORA EJECUTIVA)	30 DE NOVIEMBRE - 2000
Resolución 159-2001	Propone la primera etapa de una opción de varios costos, a la que podrán acogerse las empresas prestadoras del servicio público domiciliario de electricidad a usuarios reguladores y se establecen otras disposiciones en cuanto a las compensaciones por incumplimiento en los estándares de calidad por el servicio prestado en los STR y/o SDL del SIN.	LUISA FERNANDA LAFAURIE (MINISTRA DE MINAS Y ENERGIA) - DAVID REINSTEIN (DIRECTOR EJECUTIVO)	27 DE DICIEMBRE - 2001
Resolución 082-2002	Aprueban los principios generales y la metodología para el establecimiento de los cargos por uso de los sistemas de transmisión regional y distribución local.	LUIS ERNESTO MEJIA CASTRO (MINISTRO DE MINAS Y ENERGIA) - JAIME ALBERTO BLANDO (DIRECTOR EJECUTIVO)	17 DE DICIEMBRE - 2002
Resolución 084-2002	Se dictan normas en materia de calidad del servicio de energía eléctrica prestado en el sistema de interconectado nacional,	JUAN MANUEL GERS (VICEMINISTRO DE ENERGIA) - JAIME	30 DE DICIEMBRE - 2002

	relacionado con la resolución Creg159-2001 y con el primer año de periodo de transición, que trata el reglamento de distribución de energía eléctrica.	ALBERTO BLANDO (DIRECTOR EJECUTIVO)	
Resolución 065-2003	Presenta solicitudes revocatorias contra la resolución Creg084-2002, La cual da razones que justifiquen la revocatoria directa.	MANUEL MAIGUASHCA OLANO (VICEMINISTRO DE MINAS Y ENERGIA) - JAIME BLANDON (DIRECTOR EJECUTIVO)	9 DE JULIO - 2003
Resolución 074-2004	Modificación de la definición de "consumo", en el artículo de Creg108-1997, y por defecto en la presente resolución adoptan la definición de consumo de energía reactiva.	LUIS ERNESTO MEJIA CASTRO (MINISTRO DE MINAS Y ENERGIA) - SANDRA STELLA FOSENCA (DIRECTORA EJECUTIVA)	4 DE JUNIO - 2004
Resolución 097-2008	Se definen los indicadores Trimestrales de medición de la calidad del servicio que remplazan los anteriores indicadores DES y FES. Además se nombran las condiciones para compensaciones a usuarios finales y los incentivos a los ORs, y la metodología para el establecimiento de los cargos por uso de los Sistemas de Transmisión Regional y Distribución Local.	MANUEL MAIGUASHCA OLANO (VICEMINISTRO DE MINAS Y ENERGIA) – HERNAN MOLINA VALENCIA (DIRECTOR EJECUTIVO)	26 DE SEPTIEMBRE – 2008
Resolución 043-2010	Se aclaran disposiciones de la resolución CREG 097 de 2008 relacionadas con la regulación de calidad del servicio en los sistemas de distribución local y se adoptan disposiciones complementarias a dicha resolución	SILVANA GIAIMO CHAVEZ (VICEMINISTRA DE MINAS Y ENERGIA) – JAVIER AUGUSTO DIAZ VELAZCO (DIRECTOR EJECUTIVO)	16 DE MARZO – 2010
Resolución 094-2012	Se establece el reglamento para el reporte de eventos y el	MAURICIO CARDENAS SANTAMARIA	24 DE AGOSTO – 2012

	procedimiento para el cálculo de la Energía No Suministrada, y se precisan otras disposiciones relacionadas con la calidad del servicio en los Sistemas de Transmisión Regional.	(VICEMINISTRA DE MINAS Y ENERGIA) – GERMAN CASTRO FERREIRA (DIRECTOR EJECUTIVO)	
Resolución 015-2018	Se establece la metodología para la remuneración de la actividad de distribución de energía eléctrica en el Sistema Interconectado Nacional. Se definen los nuevos indicadores de referencia de calidad mínima garantizada SAIDI y SAIFI	GERMAN ARCE ZAPATA (MINISTRO DE MINAS Y ENERGIA) – GERMAN CASTRO FERREIRA (DIRECTOR EJECUTIVO)	29 DE ENERO – 2018

5.3 MARCO CONCEPTUAL

Aportes de los almacenes de datos

Los almacenes de datos se han convertido en soluciones robustas que se constituyen en la solución de soporte a la información histórica de las organizaciones, especialmente en las áreas autónomas que buscan extraer conocimientos de sus datos para soportan las decisiones empresariales. A su vez, los data marts se han convertido en soluciones de primera mano para las áreas de tecnología quienes buscan almacenar grandes volúmenes de datos y dan acceso a ellos permitiendo a los usuarios extraer el conocimiento de los datos históricos para realizar análisis, formulación e implementación de estrategias y alcanzar los objetivos de la organización.

Un DW debe permitir que la información almacenada allí sea accesible, correcta, uniforme y actualizada. Estas características son requeridas en un almacén de datos y ayudan a obtener las siguientes ventajas para la organización (Universidad del Cauca, s.f.):

Menor coste en la toma de decisiones: Disminuye los tiempos que se podían producir al intentar ejecutar consultas de datos largas y complejas con bases de datos que estaban diseñadas específicamente para transacciones más cortas y sencillas.

Mayor flexibilidad ante el entorno: El DW convierte los datos operacionales en información relacionada y estructurada, que genera el "conocimiento" necesario para la toma de decisiones. Esto permite establecer una base única del modelo de información de la organización, que puede dar lugar a una visión global de la información basado en los conceptos de negocio que tratan los usuarios. Además, aporta una mejor calidad y flexibilidad en el análisis del mercado, y del entorno en general.

Mejor servicio al cliente: Todo lo que se indicó en el punto anterior implica una importante mejora en la calidad de gestión, lo que también repercute en la relación con el cliente, que es, como sabemos, uno de los pilares básicos en los que descansa cualquier organización ajustada. De hecho, el que un DW implique una mayor flexibilidad ante el entorno tiene una consecuencia directa en una mayor capacidad para responder a las necesidades de los clientes.

Rediseño de procesos: Ofrecer a los usuarios una capacidad de análisis de la información de su negocio que tiende a ser ilimitada y permite con frecuencia obtener una visión más profunda y clara de los procesos de negocio propiamente dichos, lo que a su vez permite obtener ideas renovadoras para el rediseño de los mismos.

Alineamiento con los objetivos: Se distribuye cada vez más en toda la organización la responsabilidad en la toma de decisiones. Esta capacidad de decisiones distribuidas es cada vez más necesaria para el dimensionamiento correcto de las empresas, y es uno de los aspectos en los que el DW puede aportar una contribución esencial.

En conclusión, el concepto de DW abarca mucho más que simplemente copiar datos operacionales a una base de datos informacional distinta. El sistema deberá ofrecer una solución completa para gestionar y controlar el flujo de información desde bases de datos

corporativas y fuentes externas a sistemas de soporte de decisiones de usuarios finales. Además, debe permitir a los usuarios conocer qué información existe en el almacén de datos, y cómo poder acceder a ella y manipularla

A continuación, se presenta un marco teórico referencial que presenta breves introducciones de conceptos que serán de utilidad para comprender los objetivos del presente proyecto de tesis.

DATA WAREHOUSE (DW)

Probablemente la definición más conocida de Data Warehouse es la de William Inmon (Inmon, 2005), quien indica que es "una colección de datos orientada a un determinado ámbito (empresa, organización, etc), integrada, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad".

Otra definición es la de Ralph Kimball (Kimball & Ross, 2002) quien indica que "la Bodega de Datos es un colección de datos en forma de una base de datos que guarda y ordena información que se extrae directamente de los sistemas operacionales (ventas, producción, finanzas, marketing, etc.) y de datos externos".

La primera definición permite identificar las principales características de un DW. Estas características se exponen a continuación:

Orientada a temas:

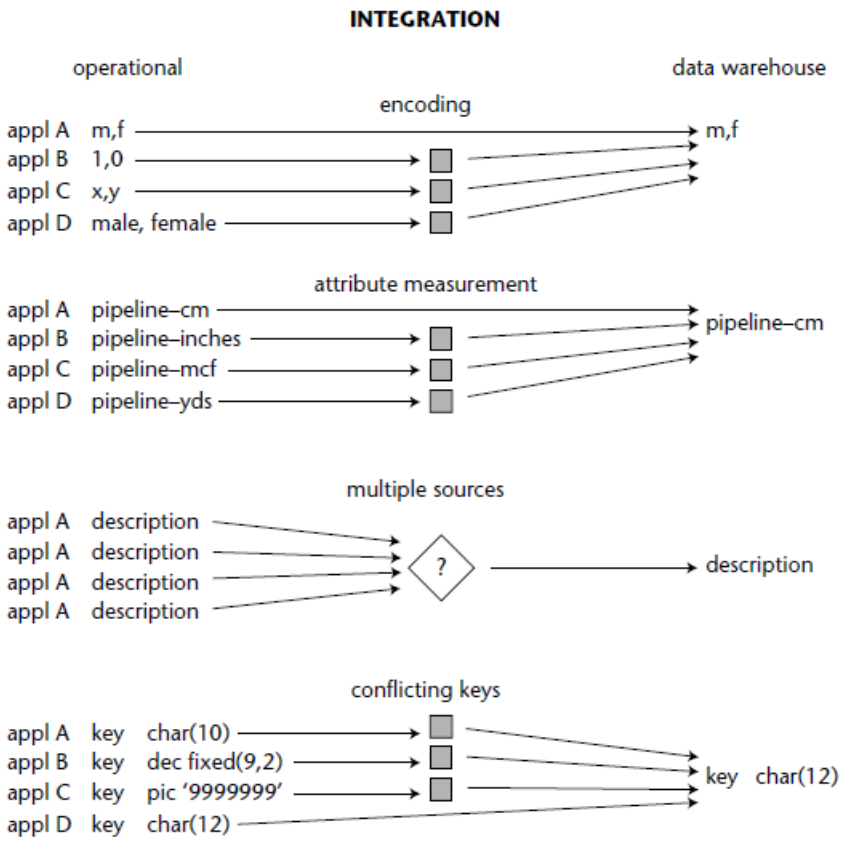
En un DW la información se clasifica de acuerdo a los aspectos que son de interés para la empresa, esto difiere de los procesos orientados a aplicaciones, la cual se centra en los datos necesarios para la operación. Por ejemplo, en un proceso orientado a aplicaciones de una entidad financiera importarán los datos de préstamos, ahorros, tarjeta bancaria y depósitos, mientras que un proceso orientado a temas que son de interés para a empresa los datos que se pueden obtener son de cliente, vendedor, producto y actividad.

Integración:

Esta característica es la más importante del DW, ya que los datos provienen de diferentes fuentes transaccionales y no transaccionales, por lo que se deben limpiar, formatear, resumir y completar de tal forma que lleguen uniformes al almacén de datos.

En la Figura 7 se muestra las diferentes formas de almacenar datos en las bases de datos operacionales y como se realiza la integración para llevarlas a un DW.

Figura 7 - El problema de la integración (Inmon, 2005)



No volátil:

La información consignada en un almacén de datos es cargada una sola vez y no se modifica, mientras que una base de datos operacional está en constantes cambios de sus

datos (actualización, inserción, y borrado). Por lo tanto se tienen dos tipos de operaciones en la manipulación de datos de un DW: la carga inicial de los datos y el acceso a los mismos. La actualización del data warehouse se da cuando se agregan nuevos valores a las variables que se definieron en el diseño. De este modo se conservan los valores históricos.

De tiempo variante:

La variación del tiempo implica que cada registro en el DW que se registró en un momento del tiempo se conserva intacto, es decir no se modifica y por el contrario se almacenan los diferentes valores que se presentan para una variable en el tiempo manteniendo un registro histórico de los datos y marcándolos con una estampa de tiempo. En los sistemas operacionales, por el contrario, los datos obtenidos siempre reflejarán el estado de la actividad del negocio en el momento presente, por lo que los datos históricos son de poco uso en estos sistemas.

En un almacén de datos, los registros históricos pueden ser utilizados para identificación y evaluación de tendencias, y permitir comparaciones.

DATA MART (DM)

La definición de Data Mart que propone Kimball (Kimball et al, 1998) es un repositorio de información, similar a un DW, pero orientado a un área o departamento específico de la organización, a diferencia del DW que cubre toda la organización, es decir la diferencia fundamental es su alcance. Mientras un DW proporciona una visión global de la organización, común e integrada de los datos de la organización, un DM responde al análisis, función o necesidad de un grupo de trabajo o departamento dentro de la organización.

Muchos almacenes de datos comienzan siendo data marts, buscando almacenamiento de datos comunes de un área, o realizar una primera entrega en tiempos razonables, y hasta minimizar riesgos de una implementación a más grande escala. Una vez que se

implementan exitosamente, se busca ampliar su alcance gradualmente hasta conseguir un DW.

Los DM tienen las mismas características de orientación a temas, integración, no volatilidad y de tiempo variable que tiene un DW.

Un DM se alimenta de diferentes fuentes de información como bases de datos transaccionales, información histórica, plantillas en Excel, y hasta otros data marts. Para consolidar la información que se encuentra en fuentes heterogéneas es necesario recurrir a procesos de extracción, transformación y carga (ETLs) que permitan mover los datos desde diferentes fuentes, limpiarlos y darles un formato homogéneo para luego cargarlos a un repositorio común.

PROCESOS DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA DE LOS DATOS (ETL)

Los procesos ETL (por su sigla en inglés de Extract-Transform-Load) son un estándar que se utiliza para referirse al movimiento y transformación de datos desde múltiples fuentes hasta un repositorio final, como un DW o DM, realizando actividades intermedias de limpieza, reformateo e integración de los datos.

El éxito de un buen DW o DM está en el correcto diseño de los ETLs ya que estos deben garantizar la extracción de datos de los sistemas de origen, ya sean legados o sistemas distribuidos y de gran tamaño (por ejemplo ERP o CRM), asegurar la calidad de los datos y su consistencia permitiendo ajustar los datos que no cumplen con todas las características requeridas, y proporcionar un formato que permita presentar los datos de forma que los usuarios finales puedan usarlos y tomar decisiones con ellos.

La construcción del sistema ETL puede consumir el 70% de los recursos necesarios para la implementación y mantenimiento de un DW (kimball & Caserta, 2004). Esto es porque añaden un valor significativo a los datos:

- Elimina errores y corrige los datos faltantes
- Proporciona medidas documentadas sobre la confiabilidad de los datos
- Captura el flujo de datos transaccionales para su custodia
- Ajusta los datos de múltiples fuentes para ser utilizados conjuntamente
- Estructura datos que puedan utilizar las herramientas del usuario final

Todos los procesos ETL deben cumplir al menos con las siguientes fases:

Extracción de datos

Esta etapa se encarga de obtener los datos desde los sistemas origen, que en la mayoría de casos son sistemas de procesamiento de transacciones en línea u OLTP (por sus siglas en inglés de On-line Transaction Processing). Sin embargo, muchos almacenes de datos también incorporan datos de otros sistemas que no son OLTP como archivos de texto, sistemas heredados y hojas de cálculo (Shaker H, Abdeltawab M., & Ali Hamed, 2011). Cada fuente de datos tiene características propias del sistema al que pertenece, y la configuración de estos puede diferir de un sistema a otro, por lo cual este proceso debe encargarse de extraer eficazmente los datos y prepararlos para el proceso de transformación.

Una parte específica del proceso de extracción es la de analizar los datos obtenidos, por lo cual se requiere realizar un chequeo para verificar si los datos cumplen la con la estructura que se espera construir en el DW o DM, de no ser así los datos deben ser rechazados.

También es importante tener en cuenta que, al extraer los datos de la fuente, el sistema no se debe ver afectado. Por ejemplo, si el volumen de datos es grande esto puede ralentizar o colapsar el sistema, provocando que este no pueda utilizarse con normalidad en la operación diaria. Por lo tanto, se debe tener en cuenta la programación de la extracción, pues en ocasiones se debe hacer en horarios o días no laborales, donde el impacto sea mínimo.

Transformación

La etapa de transformación aplica una serie de funciones sobre los datos extraídos que permiten homogenizarlos para que sean cargados en el DW o DM. Algunos de los casos que se deben tener en cuenta en esta fase son (Power Data, 2013):

- Seleccionar sólo ciertas columnas para su carga (por ejemplo, que las columnas con valores nulos no se carguen).
- Traducir códigos (por ejemplo, si la fuente almacena una "H" para Hombre y "M" para Mujer, pero el destino tiene que guardar "1″ para Hombre y "2″ para Mujer).
- Codificar valores libres (por ejemplo, convertir "Hombre" en "H" o "Sr" en "1″).
- Obtener nuevos valores calculados (por ejemplo, total_venta = cantidad * precio).
- Unir datos de múltiples fuentes (por ejemplo, búsquedas, combinaciones, etc.).
- Calcular totales de múltiples filas de datos (por ejemplo, ventas totales de cada región).
- Generar campos clave en el destino.
- Transponer o pivotar (girando múltiples columnas en filas o viceversa).
- Dividir una columna en varias (por ejemplo, columna "Nombre: García, Miguel"; pasar a dos columnas "Nombre: Miguel" y "Apellido: García").
- Aplicar para formas simples o complejas, la acción que en cada caso se requiera, como por ejemplo:
    - Datos OK: entregar datos a la siguiente etapa (fase de carga).
    - Datos erróneos: ejecutar políticas de tratamiento de excepciones (por ejemplo, rechazar el registro completo, dar al campo erróneo un valor nulo o un valor centinela).

La etapa de transformación tiende a hacer un poco de limpieza y ajustes a los datos de entrada para obtener datos precisos, correctos, completos, consistentes y sin ambigüedades.

En esta fase también se define la granularidad de las tablas de hechos, las tablas de dimensiones, el esquema del DW (estrella o copo de nieve), los hechos derivados, las

dimensiones lentamente cambiantes, y la factibilidad de las tablas de hechos (Shaker H, Abdeltawab M., & Ali Hamed, 2011).

Carga

La carga de datos es el paso final de los procesos ETL. El objetivo es tomar los datos procedentes de la fase de transformación y cargarlos en una estructura objetivo multidimensional. En este paso, los datos extraídos y transformados son escritos en las estructuras dimensionales que serán accedidas por los usuarios finales. En la carga se incluyen tanto las tablas de dimensiones como las tablas de hechos.

El principal reto de la fase de carga es el manejo de un gran volumen de datos. Para esto se dan unas claves que pueden minimizar el impacto en la caga de datos (kimball & Caserta, 2004):

- Separar Inserciones de actualizaciones: Muchas herramientas ETL (y algunas bases de datos) ofrecen la funcionalidad de actualización además de la inserción. Esta funcionalidad es muy conveniente y simplifica el flujo de datos lógico, pero es notoriamente lento. Los procesos ETL que requieren actualización a datos existentes deberían incluir la lógica que separa los registros nuevos de los que ya existen en la tabla de hechos. Siempre que se tenga una cantidad sustancial de datos se requerirá realizar una carga masiva en el DW o DM.
  Lamentablemente, muchas herramientas de carga masiva no pueden actualizar registros existentes. Tratando por separado la actualización de la inserción de datos, primero debería ejecutar las actualizaciones y luego realizar la carga masiva para balancear el registro de la información, buscando un óptimo desempeño de la carga.
- Utilice una herramienta de carga masiva: La utilización de una herramienta de carga masiva en vez de sentencias SQL de INSERT para cargar datos, disminuye sustancialmente la sobrecarga de la base de datos y mejora drásticamente el rendimiento de la carga.

- Carga en Paralelo: Al cargar volúmenes de datos, los datos se dividen físicamente en segmentos lógicos. Si se cargan cinco años de datos, tal vez se puedan hacer cinco *data files* que contengan un año cada uno. Algunas herramientas ETL permiten dividir dinámicamente los datos basados en rangos de valores. Una vez que los datos se dividen en segmentos iguales, se puede ejecutar el proceso de ETL para cargar los segmentos en paralelo.
- Minimizar los cambios físicos: La actualización de registros en una tabla requiere grandes cantidades de sobrecarga en el DBMS, la mayoría de las cuales se debe al log de *rollback* al poblar la base de datos. Para minimizar la escritura en el log de *rollback* se debe realizar carga masiva de datos en la base de datos.
  En muchos casos, es mejor eliminar los registros que se actualizarían y luego cargar masivamente las nuevas versiones de dichos registros, junto con los registros nuevos que se van a insertar en el almacén de datos. Dado que la proporción de los registros que se actualizan contra el número de filas existentes juega un factor importante en la selección de la técnica óptima, generalmente se requiere realizar algunas pruebas de ensayo y error para ver si este enfoque es la estrategia de carga que en últimas se utilizará para su situación particular.
- Construya las cláusulas fuera de la base de datos: El ordenamiento, la unión y la construcción de las cláusulas fuera de la base de datos pueden ser más eficientes que el uso de SQL con funciones de conteo (COUNT) y suma (SUM) y palabras clave como GROUP BY y ORDER BY en el DBMS. Los procesos ETL que requieren el reordenamiento y/o la unión de grandes volúmenes de datos debe llevar a cabo estas funciones antes de entrar en la preparación de la base de datos relacional. Muchas herramientas ETL están adecuadas para ejecutar estas funciones, pero las herramientas dedicadas a realizar ordenamiento y/o unión a nivel de sistema operativo son una buena inversión para el procesamiento de grandes conjuntos de datos.

MODELO DIMENSIONAL

Históricamente los analistas y diseñadores se han apoyado en herramientas como el modelo entidad relación (MER) para diseñar bases de datos tradicionales. Sin embargo, para realizar el diseño de una bodega de datos, la herramienta más utilizada es el modelo dimensional ya que permite organizar la información de forma más flexible para lograr mayor desempeño y optimizar la recuperación de la información desde un punto de vista más cercano al usuario final.

Este modelo divide "las particularidades de los procesos que ocurren en una organización, entre mediciones y entorno. Las medidas son en su mayoría, medidas numéricas, y se les denomina hechos. Alrededor de estos hechos existe un contexto que describe en qué condiciones y en qué momento se registró este hecho. Aunque el entorno se ve como un todo, existen registros lógicos de diferentes características que describen un hecho, por ejemplo, si el hecho referido, es la venta de un producto en una cadena de tiendas, se podría dividir el entorno que rodea al hecho de la cantidad vendida, en el producto vendido, el cliente que lo compró, la tienda y la fecha en que se realizó la venta. A estas divisiones se le denomina dimensiones y a diferencia de los hechos que son numéricos, estos son fundamentalmente textos descriptivos". (Cedeño Trujillo, 2006) .

Según el mismo autor, existen tablas de hechos y dimensiones como se describen a continuación:

Tablas de Hechos

Las tablas de hechos, representan los procesos que ocurren en la organización, son independientes entre sí (no se relacionan unas con otras). En estas, se almacenan las medidas numéricas de la organización. Cada medida, se corresponde con una intersección de valores de las dimensiones y generalmente se trata de cantidades numéricas, continuamente evaluadas y aditivas. La razón de estas características, es que facilita que los miles de registros que involucran una consulta, sean comprimidos más fácilmente y se pueda dar respuesta con rapidez, a una solicitud que abarque gran cantidad de información.

La llave de la tabla de hechos, es una llave compuesta, debido a que se forma de la composición de las llaves primarias de las tablas dimensionales a las que está unida. Se pueden distinguir dos tipos de columnas en una tabla de hechos, columnas de hechos y columnas llaves. Las columnas de hechos almacenan las medidas del negocio que se quieren controlar y las columnas llaves, forman parte de la llave de la tabla.

Existen tablas de hechos que no contienen medidas, a estas tablas se les denomina tablas de hechos sin hechos. La semántica de la relación entre las dimensiones que definen la llave de esta tabla de hechos, implica por sí sola la ocurrencia de un evento, por ejemplo, si se quiere representar el hecho de que un estudiante matriculó en una universidad, la combinación de las siguientes dimensiones definiría este suceso: el estudiante matriculado, la carrera en que matriculó, la fecha de matrícula, el tipo de curso que va a cursar, etcétera.

Tablas de Dimensiones

Una tabla de dimensión contiene, por lo general, una llave simple y un conjunto de atributos que describen la dimensión. En dependencia del esquema multidimensional que se siga, pueden existir atributos que representen llaves foráneas de otras tablas de dimensión, es decir, que establecen una relación de esta tabla con otra dimensión.

Las tablas de dimensión, son las que alimentan a las tablas de hechos, como se expresó anteriormente, la llave de un hecho es la composición de las llaves de las dimensiones que están conectados a esta, por tanto, los atributos que conforman las tablas de dimensiones también describen el hecho.

Los atributos dimensionales son fundamentalmente textos descriptivos, estos desempeñan un papel determinante, son la fuente de gran parte de todas las necesidades que deben cubrirse, además, sirven de restricciones en la mayoría de las consultas que realizan los usuarios. Esto significa, que la calidad del modelo multidimensional, dependerá en gran parte de cuán descriptivos y manejables sean los atributos dimensionales escogidos.

Las tablas de dimensión en general, son mucho más pequeñas que las tablas de hechos en cuanto a cantidad de registros. En cuanto a cantidad de atributos, una tabla de hechos bien descriptiva puede tener un gran número de estos.

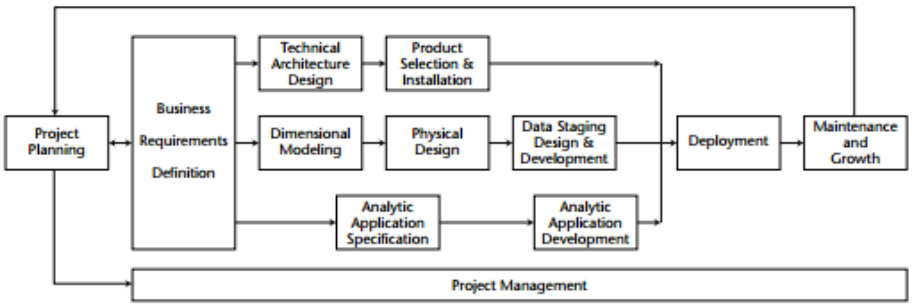
6 ESTRATEGIA METODOLÓGICA

En el campo de los almacenes de datos existen múltiples metodologías para su construcción con diferencias significativas en el ciclo de vida de su desarrollo. Algunos autores han creado sus propias metodologías y otros han modificado y adaptado las existentes. Sin embargo, no existe una metodología que defina los elementos que llenen todas las características particulares del desarrollo de un almacén de datos.

Una de las metodologías más utilizadas para el desarrollo de Data WareHouse es la de Ralph Kimball, la cual se basa en un desarrollo iterativo e incremental, en donde se hacen entregas parciales de acuerdo a cada iteración.

6.1 METODOLOGÍA KIMBALL

Figura 8 - Tareas metodología Kimball, denominada Business Dimensional Lifecycle (Kimball and Ross, 2002)



De la figura anterior se puede decir que la base del Data Mart está en la definición de requerimientos, ya que son el soporte inicial de las tareas subsiguientes. Los requerimientos también tienen influencia en la planeación del proyecto (nótese el doble sentido de la flecha en entre las cajas de definición de requerimientos y planificación del proyecto). Igualmente, se pueden ver tres rutas o caminos que se enfocan en diferentes áreas (Rivadera, 2014):

- Tecnología (Camino Superior). Implica tareas relacionadas con software específico, por ejemplo, Microsoft SQL Analysis Services.

- Datos (Camino del medio). En la misma diseñaremos e implementaremos el modelo dimensional, y desarrollaremos el subsistema de Extracción, Transformación y Carga (Extract, Transformation, and Load - ETL) para cargar el DW.
- Aplicaciones de Inteligencia de Negocios (Camino Inferior). En esta ruta se encuentran tareas en las que diseñamos y desarrollamos las aplicaciones de negocios para los usuarios finales.

A continuación, se explican las tareas de la metodología de acuerdo al ciclo de vida de Kimball:

Planeación del proyecto: Este proceso busca definir el alcance del proyecto, los objetivos específicos, el alcance, recursos, riesgos e interesados.

Rivadera afirma que "En la visión de programas y proyectos de Kimball, Proyecto, se refiere a una iteración simple del KLC (Kimball Life Cycle), desde el lanzamiento hasta el despliegue. Esta tarea incluye las siguientes acciones típicas de un plan de proyecto:

- Definir el alcance (entender los requerimientos del negocio).
- Identificar las tareas
- Programar las tareas
- Planificar el uso de los recursos.
- Asignar la carga de trabajo a los recursos
- Elaboración de un documento final que representa un plan del proyecto". (Rivadera, 2014)

Definición de Requerimientos del Negocio: Este proceso busca conocer las expectativas del proyecto de cada una de las personas y/o grupos involucrados en el proyecto. Se deben preparar las entrevistas, cuestionarios y demás herramientas que permitan identificar detalladamente los requerimientos de cada individuo dentro del proyecto. A partir del análisis realizado en esta etapa, se puede construir una matriz preliminar de dimensiones y medidas (hechos).

Según Rivadera (Rivadera, 2014) parte del proceso de preparación es averiguar a quién se debe realmente entrevistar. Hay cuatro grupos de personas con las que se debe hablar desde el principio:

- Directivos: Son los responsables de tomar las decisiones estratégicas.
- Administradores intermedios y de negocio: Son los responsables de explorar alternativas estratégicas y aplicar decisiones.
- Personal de sistemas: Son las personas que realmente sabe qué tipos de problemas informáticos y de datos existen.
- Otros: Personas que se deben entrevistar por razones políticas.

Diseño de la Arquitectura Técnica: Acá se deben escoger las tecnologías a utilizar. Se deben tener en cuenta tres factores: los requerimientos del negocio, los actuales ambientes técnicos y las directrices técnicas estratégicas.

También se deben tener en cuenta algunas consideraciones del sistema como por ejemplo el volumen de datos que se va a almacenar, el número de usuarios simultáneos y los recursos del sistema (memoria, sistema de almacenamiento, capacidad de procesamiento).

En un entorno de bodega de datos se requiere de la integración de diferentes tecnologías. Por lo tanto, el diseño de la arquitectura técnica funciona como un marco de trabajo donde se encuentran los elementos y servicios técnicos que harán parte de la bodega de datos y permite identificar el orden en el que los componentes podrán ser instalados.

Selección de Productos e Implementación: A partir de la arquitectura técnica, es necesario evaluar y seleccionar componentes específicos de la arquitectura como la plataforma de hardware, el motor de base de datos, la herramienta de ETL o el desarrollo pertinente, herramientas de acceso, etc.

Modelado dimensional: Acá se determinan los datos requeridos para cumplir con las necesidades de los usuarios. Se debe definir el nivel de detalle que se desea manejar para definir las dimensiones y los hechos.

Rivadera (Rivadera, 2014) indica que este proceso es iterativo y tiene cuatro pasos principales:

1. Elegir el proceso de negocio

Se debe elegir el área a que se va a modelar. Esta es una decisión de la dirección, y depende fundamentalmente del análisis de requerimientos de la etapa anterior.

2. Establecer el nivel de granularidad

La granularidad tiene que ver con el nivel de detalle que se va a especificar. La elección de la granularidad depende de los requerimientos del negocio, partiendo de los datos actuales. La sugerencia general es comenzar a diseñar el almacén de datos al mayor nivel de detalle posible, ya que se podría luego realizar agrupamientos al nivel deseado.

3. Elegir las dimensiones

Las dimensiones surgen de las discusiones del equipo, y son facilitadas por la elección del nivel de granularidad y de la matriz de procesos/dimensiones. Las tablas de dimensiones tienen un conjunto de atributos (generalmente textuales) que brindan una perspectiva o forma de análisis sobre una medida en una tabla hechos. Una forma de identificar las tablas de dimensiones es que sus atributos son posibles candidatos para ser encabezado en los informes, tablas pivot, cubos, o cualquier forma de visualización, unidimensional o multidimensional.

4. Identificar medidas y las tablas de hechos

El último paso consiste en identificar las medidas que surgen de los procesos de negocios. Una medida es un atributo (campo) de una tabla que se desea analizar, totalizando o agrupando sus datos, usando los criterios de corte conocidos como dimensiones. Las medidas habitualmente se vinculan con el nivel de granularidad, y se encuentran en tablas que denominamos tablas de hechos (fact en inglés). Cada tabla de hechos tiene como

atributos una o más medidas de un proceso organizacional, de acuerdo a los requerimientos. Un registro contiene una medida expresada en números, como cantidad, tiempo, dinero, etc., sobre la cual se desea realizar una operación de agregación (promedio, conteo, suma, etc.) en función de una o más dimensiones. La granularidad es el nivel de detalle que posee cada registro de una tabla de hechos.

Diseño Físico: Este proceso define las características físicas para soportar el diseño lógico de la solución.

Las decisiones tomadas durante ésta fase están dirigidas principalmente a optimizar el rendimiento en las consultas y el mantenimiento de la bodega de datos. Durante el proceso de transformación del modelo lógico al diseño físico las entidades definidas previamente se convierten en tablas y los atributos pasan a ser columnas de dichas tablas. En este punto se deben tener en cuenta las siguientes consideraciones (Moreno Ocampo, 2012):

- Garantizar las características físicas de las columnas (uso de dominios o asignación de tipos de datos, longitud, nulabilidad)
- Garantizar que los nombres propios de la base de datos sean usados (seguir estándares para los nombres físicos)
- Agregar las restricciones necesarias y las reglas del negocio
- Crear llaves sustitutas
- Resolver la implementación de subtipos (decidir cómo será implementada la jerarquía)
- Agregar estructuras de indexación

Diseño e Implementación de la presentación de Datos: Este proceso permite definir los ETLs (Extracción, Transformación y Carga) a utilizar y las reglas que se definirán para alimentar el Data Mart.

Estos procesos permiten mover grandes volúmenes de datos desde varios orígenes a un destino específico, en este caso el data mart. La calidad de los datos que se envíen al data mart dependerá del éxito del modelo a generar.

Especificación de Aplicaciones de Análisis: Acá se deben especificar las formas de acceso a los datos almacenados para todos los tipos de usuarios. Puede ser a través de informes estándar o de herramientas analíticas dependiendo de la finalidad del usuario. Esto incluye la definición de roles y perfiles por usuario.

Desarrollo de Aplicaciones de Análisis: Este proceso ejecuta la definición hecha en el proceso anterior. En esta fase se pueden desarrollar una amplia gama de aplicaciones de inteligencia de negocios incluyendo aplicaciones de minería de datos, tableros de mando, modelos analíticos, consultas parametrizadas que le proporcionan a los usuarios las herramientas necesarias para realizar diferentes tipos de análisis.

Despliegue: El despliegue representa la convergencia de la tecnología, los datos y las aplicaciones de usuarios finales accesible desde el escritorio del usuario del negocio. Hay varios factores extras que aseguran el correcto funcionamiento de todas estas piezas, entre ellos se encuentran la capacitación, el soporte técnico, la comunicación y las estrategias de retroalimentación. Se debe tener en cuenta en esta etapa la documentación y capacitación para usuarios finales.

Mantenimiento y Crecimiento: Al terminar el proyecto, éste se debe volver un proceso dentro de la organización. Por lo tanto, éste debe evolucionar de acuerdo a la dinámica del negocio.

Debido a que los almacenes de datos son proyectos a largo plazo que acompañan la evolución de una organización, una vez realizada la implementación, se deben seguir realizando tareas de soporte y capacitación para asegurar el mantenimiento y crecimiento del data mart. Durante esta evolución, el almacén de datos es actualizado con nuevos datos,

nuevos requerimientos, nuevos usuarios y nuevas mejoras en las aplicaciones existentes que permiten asegurar su crecimiento en procura de conseguir los objetivos propuestos.

Administración del Proyecto: La gestión del proyecto asegura que las actividades del ciclo de vida del negocio se lleven en forma sincronizada, por eso se debe hacer durante todo el ciclo. Entre sus actividades principales se encuentra el monitoreo del estado del proyecto y la comunicación entre los requerimientos del negocio y las restricciones de información para poder manejar correctamente las expectativas en ambos sentidos.

Aunque la metodología de Kimball es de las más utilizadas en el desarrollo de almacenes de datos, sólo indica qué se debe hacer, pero no cómo hacerlo, y por ende puede provocar demoras en los resultados esperados. Hace falta el detalle en el diseño de los modelos de datos y la forma de obtener las variables para lograr la correspondencia con los datos fuente.

Dado que existen otras metodologías que permiten ir más al detalle, es factible utilizar como marco de referencia la metodología Kimball, y apoyarse en algunas etapas con otras metodologías que permitan obtener resultados más tempranos con artefactos específicos del detalle.

6.2 METODOLOGÍA HEFESTO

Es una metodología creada por el Ingeniero Bernabeu Ricardo Darío, su última actualización es la versión 1.2 en julio del 2010 y disponible bajo licencia GNU FDL.

De acuerdo al autor (Bernabeu, 2010), esta metodología está fundamentada en una muy amplia investigación, comparación de metodologías existentes, experiencias propias en procesos de confección de almacenes de datos. Cabe destacar que HEFESTO está en continua evolución, y se han tenido en cuenta, como gran valor agregado, todos los feedbacks que han aportado quienes han utilizado esta metodología en diversos países y con diversos fines.

La idea principal, es comprender cada paso que se realizará, para no caer en el tedio de tener que seguir un método al pie de la letra sin saber exactamente qué se está haciendo, ni por qué.

Esta metodología se puede utilizar en cualquier ciclo de vida que no requieran fases extensas de reunión de requerimientos y análisis, fases de desarrollo monolítico que conlleve demasiado tiempo y fases de despliegue muy largas. Lo que se busca es entregar una primera implementación que satisfaga una parte de las necesidades, para demostrar las ventajas del almacén de datos y motivar a los usuarios.

Esta metodología se puede resumir en la siguiente figura:

Figura 9 – Metodología HEFESTO, pasos (Bernabeu, 2010)

En la figura anterior, se puede visualizar el flujo que sigue la metodología comenzando por la recolección de las necesidades de información de los usuarios hasta la integración de los

datos que se realiza con procesos ETL. A continuación, se explican cada uno de los pasos de la metodología:

Análisis de Requerimientos: Se identifican los requerimientos del usuario con el fin de entender los objetivos de la organización, haciendo uso de técnicas y herramientas, como la entrevista, la encuesta, el cuestionario, la observación, el diagrama de flujo y el diccionario de datos, obteniendo como resultado una serie de preguntas que se deberán analizar con el fin de establecer cuáles serán los indicadores y perspectivas que serán tomadas en cuenta para la construcción del almacén de datos. Finalmente se realizará un modelo conceptual en donde se podrá visualizar el resultado obtenido en este primer paso.

Identificar preguntas

El primer paso comienza con el acopio de las necesidades de información, el cual puede llevarse a cabo a través de muy variadas y diferentes técnicas, cada una de las cuales poseen características inherentes y específicas, como por ejemplo entrevistas, cuestionarios, observaciones, etc.

El análisis de los requerimientos de los diferentes usuarios, es el punto de partida de esta metodología, ya que ellos son los que deben, en cierto modo, guiar la investigación hacia un desarrollo que refleje claramente lo que se espera del depósito de datos, en relación a sus funciones y cualidades.

El objetivo principal de esta fase, es la de obtener e identificar las necesidades de información clave de alto nivel, que es esencial para llevar a cabo las metas y estrategias de la empresa, y que facilitará una eficaz y eficiente toma de decisiones.

Identificar indicadores y perspectivas

Una vez que se han establecido las preguntas de negocio, se debe proceder a su descomposición para descubrir los indicadores que se utilizarán y las perspectivas de análisis que intervendrán.

Para ello, se debe tener en cuenta que los indicadores, para que sean realmente efectivos son, en general, valores numéricos y representan lo que se desea analizar concretamente, por ejemplo: saldos, promedios, cantidades, sumatorias, fórmulas, etc.

En cambio, las perspectivas se refieren a los objetos mediante los cuales se quiere

examinar los indicadores, con el fin de responder a las preguntas planteadas, por ejemplo: clientes, proveedores, sucursales, países, productos, rubros, etc. Cabe destacar, que el Tiempo es muy comúnmente una perspectiva.

Modelo conceptual

En esta etapa, se construirá un modelo conceptual1 a partir de los indicadores y perspectivas obtenidas en el paso anterior.

A través de este modelo, se podrá observar con claridad cuáles son los alcances del

proyecto, para luego poder trabajar sobre ellos, además al poseer un alto nivel definición de los datos, permite que pueda ser presentado ante los usuarios y explicado con facilidad.

Análisis de los OLTP: Tomando en cuenta el resultado obtenido en el paso anterior se analizarán las fuentes OLTP para determinar cómo serán calculados los indicadores con el objetivo de establecer las respectivas correspondencias entre el modelo conceptual y las fuentes de datos. Luego, se definirán qué campos se incluirán en cada perspectiva y finalmente, se ampliará el modelo conceptual con la información obtenida en este paso.

Conformar indicadores

En este paso se deberán explicitar cómo se calcularán los indicadores, definiendo los siguientes conceptos para cada uno de ellos:

Hecho/s que lo componen, con su respectiva fórmula de cálculo. Por ejemplo: Hecho1+ Hecho2.

Función de sumarización que se utilizará para su agregación. Por ejemplo: SUM, AVG, COUNT, etc.

Establecer correspondencias

El objetivo de este paso, es el de examinar los OLTP disponibles que contengan la

información requerida, como así también sus características, para poder identificar las correspondencias entre el modelo conceptual y las fuentes de datos.

La idea es, que todos los elementos del modelo conceptual estén correspondidos en los OLTP.

Nivel de granularidad

Una vez que se han establecido las relaciones con los OLTP, se deben seleccionar los campos que contendrá cada perspectiva, ya que será a través de estos por los que se examinarán y filtrarán los indicadores.

Para ello, basándose en las correspondencias establecidas en el paso anterior, se debe presentar a los usuarios los datos de análisis disponibles para cada perspectiva. Es muy importante conocer en detalle que significa cada campo y/o valor de los datos encontrados en los OLTP, por lo cual, es conveniente investigar su sentido, ya sea a través de diccionarios de datos, reuniones con los encargados del sistema, análisis de los datos propiamente dichos, etc.

Modelo conceptual ampliado

En este paso, y con el fin de graficar los resultados obtenidos en los pasos anteriores, se ampliará el modelo conceptual, colocando bajo cada perspectiva los campos seleccionados y bajo cada indicador su respectiva fórmula de cálculo.

Modelo lógico del Almacén de Datos: Como tercer paso, se realizará el modelo lógico de la estructura del almacén de datos, teniendo como base el modelo conceptual. Para esto, se debe definir el tipo de representación de un almacén de datos que será utilizado, posteriormente se llevarán a cabo las acciones propias al proceso, para diseñar las tablas de dimensiones y de hechos. Por último, se realizarán las uniones pertinentes entre estas tablas.

Tipo de modelo lógico del DW

Se debe seleccionar cuál será el tipo de esquema que se utilizará para contener la estructura del depósito de datos, que se adapte mejor a los requerimientos y necesidades de los usuarios. Es muy importante definir objetivamente si se empleará un esquema en estrella, constelación o copo de nieve, ya que esta decisión afectará considerablemente la elaboración del modelo lógico.

Tablas de dimensiones

En este paso se deben diseñar las tablas de dimensiones que formaran parte del DW. Para los tres tipos de esquemas, cada perspectiva definida en en modelo conceptual constituirá una tabla de dimensión. Para ello deberá tomarse cada perspectiva con sus campos relacionados y realizarse el siguiente proceso:

- Se elegirá un nombre que identifique la tabla de dimensión.
- Se añadirá un campo que represente su clave principal.
- Se redefinirán los nombres de los campos si es que no son lo suficientemente intuitivos.

Tablas de hechos

En este paso, se definirán las tablas de hechos, que son las que contendrán los hechos a través de los cuales se construirán los indicadores de estudio.

Para los esquemas en estrella y copo de nieve, se realizará lo siguiente:

- Se le deberá asignar un nombre a la tabla de hechos que represente la información analizada, área de investigación, negocio enfocado, etc.
- Se definirá su clave primaria, que se compone de la combinación de las claves primarias de cada tabla de dimensión relacionada.
- Se crearán tantos campos de hechos como indicadores se hayan definido en el modelo conceptual y se les asignará los mismos nombres que estos. En caso que se prefiera, podrán ser nombrados de cualquier otro modo.

Uniones

Para los tres tipos de esquemas, se realizarán las uniones correspondientes entre sus tablas de dimensiones y sus tablas de hechos.

Integración de datos: Se prueban los datos a través de procesos ETL. Para realizar la compleja actividad de extraer datos de diferentes fuentes, luego integrarlos, filtrarlos y depurarlos, se podrá hacer uso de software que facilita dichas tareas, por lo cual este paso se centrará solo en la generación de las sentencias SQL que contendrán los datos que serán de interés.

Carga inicial

Debemos en este paso realizar la Carga Inicial4 al DW, poblando el modelo de datos que hemos construido anteriormente. Para lo cual debemos llevar adelante una serie de tareas básicas, tales como limpieza de datos, calidad de datos, procesos ETL, etc.

La realización de estas tareas, pueden contener una lógica realmente compleja en algunos casos. Afortunadamente, en la actualidad existen muchos softwares que se pueden emplear a tal fin, y que nos facilitarán el trabajo.

Se debe evitar que el DW sea cargado con valores faltantes o anómalos, así como también se deben establecer condiciones y restricciones para asegurar que solo se utilicen los datos de interés.

Cuando se trabaja con un esquema constelación, hay que tener presente que varias tablas de dimensiones serán compartidas con diferentes tablas de hechos, ya que puede darse el caso de que algunas restricciones aplicadas sobre una tabla de dimensión en particular para analizar una tabla de hechos, se puedan contraponer con otras restricciones o condiciones de análisis de otras tablas de hechos.

Primero se cargarán los datos de las dimensiones y luego los de las tablas de hechos, teniendo en cuenta siempre, la correcta correspondencia entre cada elemento. En el caso en que se esté utilizando un esquema copo de nieve, cada vez que existan jerarquías de dimensiones, se comenzarán cargando las tablas de dimensiones del nivel más general al más detallado.

Actualización

Cuando se haya cargado en su totalidad el DW, se deben establecer sus políticas y estrategias de actualización o refresco de datos.

Una vez realizado esto, se tendrán que llevar a cabo las siguientes acciones:

- Especificar las tareas de limpieza de datos, calidad de datos, procesos ETL, etc., que deberán realizarse para actualizar los datos del DW.
- Especificar de forma general y detallada las acciones que deberá realizar cada software.

7 DESARROLLO DEL PROYECTO

Antes de entrar a desarrollar la metodología, debemos revisar las herramientas ETLs existentes en el mercado para saber con cuál de ellas se desarrollará el proyecto.

7.1 HERRAMIENTAS ETLS

Las herramientas ETLs permiten mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data mart o data warehouse para analizar, o en otro sistema operacional para apoyar un proceso de negocio.

Existen varios fabricantes que ofrecen soluciones robustas que permiten conectar diferentes marcas y tecnologías de bases de datos. Igualmente existen productos de uso libre que pueden suplir muchas de las funcionalidades que se buscan al migrar o concentrar información.

Gartner (Gartner, 2016) define las siguientes capacidades que los proveedores de herramientas de integración de datos deben proporcionar para dar un valor óptimo a las organizaciones:

Capacidades de Conectividad / Adaptación (con soporte a orígenes y destinos de datos): habilidad para conectar con un amplio rango de tipos de estructura de datos, que incluyen bases de datos relacionales y no relacionales, variados formatos de ficheros, XML, aplicaciones ERP, CRM o SCM, formatos de mensajes estándar (EDI, SWIFT o HL7), colas de mensajes, correos electrónicos, sitios web, repositorios de contenido o herramientas de ofimática.

Capacidades de entrega de datos: habilidad para proporcionar datos a otras aplicaciones, procesos o bases de datos en varias formas, con capacidades para programación de procesos batch, en tiempo real o mediante lanzamiento de eventos.

Capacidades de transformación de datos: habilidad para la transformación de los datos, desde transformaciones básicas (conversión de tipos, manipulación de cadenas o cálculos

simples), transformaciones intermedias (agregaciones, resúmenes, búsquedas) hasta transformaciones complejas como análisis de texto en formato libre o texto enriquecido.

Soporte de Metadatos y Modelado de Datos: recuperación de los modelos de datos desde los orígenes de datos o aplicaciones, creación y mantenimiento de modelos de datos, mapeo de modelo físico a lógico, repositorio de metadatos abierto (con posibilidad de interactuar con otras herramientas), sincronización automatizada de los metadatos en varias instancias de la herramienta.

Capacidades de diseño y desarrollo del entorno: representación gráfica de los objetos del repositorio, modelos de datos y flujos de datos, capacidades para trabajo en equipo, gestión de flujos de trabajo de los procesos de desarrollo, funcionalidad para apoyar la reutilización entre desarrolladores y proyectos para facilitar la identificación de redundancias, soporte para pruebas y depuración.

Capacidades de soporte al gobierno de la información (A través de la interoperación con capacidades de calidad de datos, perfilado y minería con las herramientas del proveedor o de un tercero): Mecanismos para trabajar con capacidades relacionadas para ayudar a comprender y asegurar la calidad de los datos a lo largo del tiempo, incluyendo la interoperabilidad con herramientas de perfilamiento de datos, herramientas de minería de datos, herramientas de calidad de datos, puntuación y evaluación de los datos que se mueven a través de procesos en línea.

Capacidades de plataforma en tiempo de ejecución y opciones de despliegue: Mainframes (IBM Z/OS), AS/400, HP Tandem, Unix, Wintel, Linux, Servidores Virtualizados, etc.

Capacidades en operaciones y administración: habilidades para gestión, monitorización y control de los procesos de integración de datos, como gestión de errores, recolección de estadísticas de ejecución, controles de seguridad, etc.

Capacidades de arquitectura e integración: grado de compactación, consistencia e interoperabilidad de los diferentes componentes que forman la herramienta de integración

de datos (con un deseable mínimo número de productos, un único repositorio, un entorno de desarrollo común, interoperabilidad con otras herramientas o vía API), etc.

Capacidades de habilitación de servicios: las herramientas de integración de datos deben exhibir características orientadas al servicio y proporcionar soporte para SOA, como La capacidad de implementar todos los aspectos de la funcionalidad de tiempo de ejecución como servicios de datos, gestión de la publicación y pruebas de servicios de datos, interacción con los repositorios y registros de servicios.

A continuación, se revisarán tres herramientas licenciadas que aportan características para la creación de los procesos ETL.

7.1.1 Microsoft Integration Services

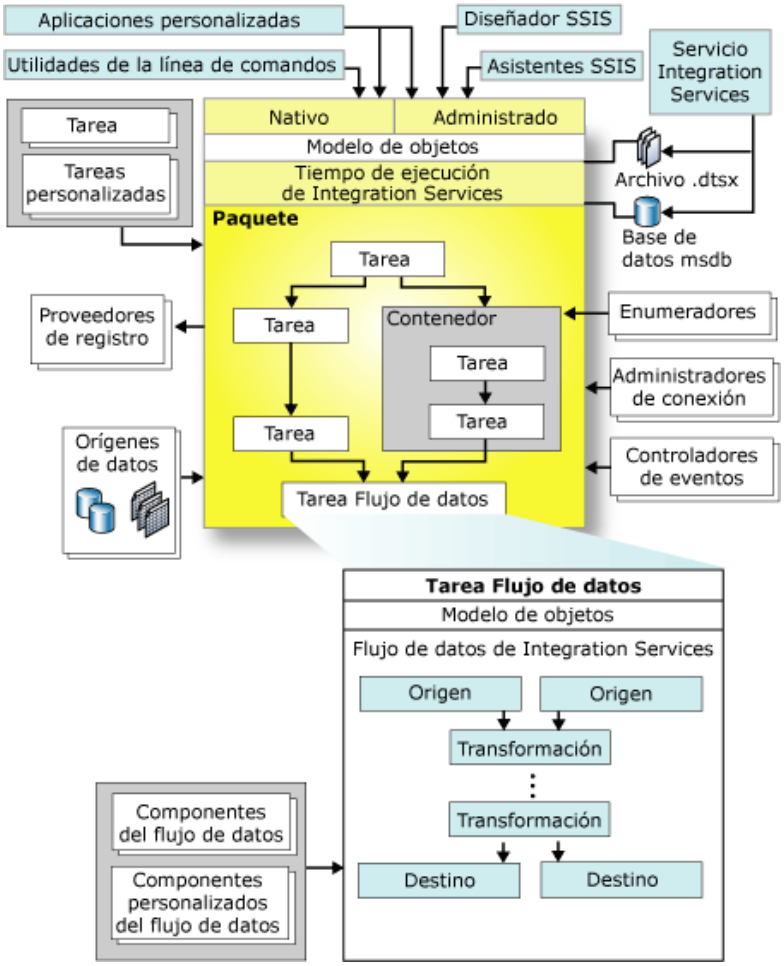
"Microsoft Integration Services es una plataforma para compilar soluciones de transformaciones de datos e integración de datos de nivel empresarial. Integration Services sirve para resolver complejos problemas empresariales mediante la copia o descarga de archivos, el envío de mensajes de correo electrónico como respuesta a eventos, la actualización de almacenamientos de datos, la limpieza y minería de datos, y la administración de objetos y datos de SQL Server. Los paquetes pueden funcionar en solitario o junto con otros paquetes para hacer frente a las complejas necesidades de la empresa. Integration Services puede extraer y transformar datos de diversos orígenes como archivos de datos XML, archivos planos y orígenes de datos relacionales y, después, cargar los datos en uno o varios destinos.

Integration Services incluye una amplia gama de tareas integradas y transformaciones; herramientas de creación de paquetes y el Integration Services service para ejecutar y administrar paquetes. Las herramientas gráficas de Integration Services se pueden usar para crear soluciones sin escribir una sola línea de código. También se puede programar el amplio modelo de objetos de Integration Services para crear paquetes mediante

programación y codificar tareas personalizadas y otros objetos de paquete". (Microsoft, 2016)

De acuerdo al sitio de Microsoft (Microsoft, 2016) la herramienta Microsoft Integration Servicies define una arquitectura que se describe a continuación:

Figura 10 - Arquitectura de Integration Services (Microsoft, 2016)



Contiene un componente llamado *Paquete*, el cual es la unidad de trabajo que se recupera, ejecuta y guarda, y además es el objeto de Integration Services más importante.

Los elementos de flujo de control (*tareas* y *contenedores*) que generan el flujo de control en un paquete. Los elementos de flujo de control preparan o copian datos, interactúan con

otros procesos o implementan flujo de trabajo repetido. Las restricciones de precedencia ordenan en una secuencia los elementos de flujo de control en un flujo de control ordenado y especifican las condiciones para ejecutar tareas o contenedores.

Los *componentes de flujo de datos* (orígenes, transformaciones y destinos) generan flujos de datos en un paquete que extrae, transforma y carga datos. Las rutas ordenan los componentes de flujo de datos en un flujo de datos ordenado.

Los *administradores de conexión* se conectan a diferentes tipos de orígenes de datos para extraer y cargar datos.

Las variables se pueden usar en expresiones para actualizar de forma dinámica los valores de columna y expresiones de propiedad, controlar la ejecución de flujos de control repetidos y definir las condiciones a las que se aplican las restricciones de precedencia.

Los *controladores de eventos* se ejecutan como respuesta ante los eventos de tiempo de ejecución que activan los paquetes, tareas y contenedores.

Los *proveedores de registro* admiten el registro de información de tiempo de ejecución de paquetes, como la hora de inicio y de detención del paquete y sus tareas y contenedores.

7.1.2 Oracle Warehouse Builder

Oracle Warehouse Builder (OWB) es una herramienta de inteligencia de negocios que proporciona una solución integrada para el diseño y despliegue de bodegas de datos, data marts y aplicaciones de BI (Oracle, 2007).

Esta herramienta permite crear almacenes de datos, migrar los datos desde los sistemas heredados, consolidar datos de fuentes de datos dispares, limpiar y transformar datos para proporcionar información de calidad, y la administración de metadatos corporativos. (Oracle, 2016)

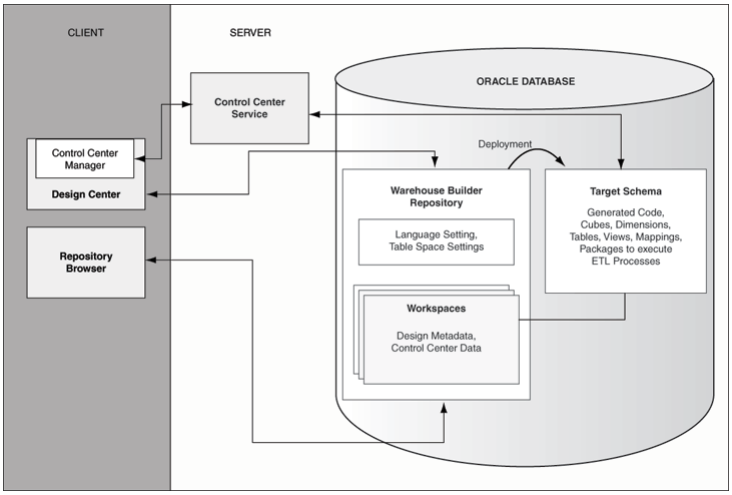
OWB proporciona un entorno gráfico para la construcción, administración y mantenimiento de un repositorio de metadatos que mantiene información detallada sobre el diseño de las bases de datos. El repositorio es implementado como un conjunto de tablas en una base de datos Oracle.

Esta herramienta posee componentes que permiten crear el diseño lógico del data mart, mapear el diseño lógico al diseño físico, generar código para crear los objetos para el data mart, crear un flujo de proceso para la población del data mart y ejecutar el flujo de proceso para poblar el data mart.

En otro documento de Oracle (Yglesias, 2008) se indica que "Oracle Warehouse Builder es una herramienta orientada no solamente a realizar el proceso de ETL, sino también la definición, administración y mantenimiento de un data warehouse. Está concebida para trabajar integrada con la tecnología de Base de Datos Oracle y ha mejorado su desempeño como herramienta de ETL, convirtiéndose en una muy buena opción cuando la base de datos destino sea la Base de Datos Oracle, tomando en cuenta que el producto ya viene incluido sin costo adicional con sus características básicas".

De acuerdo a Oracle (Oracle, 2017), en la siguiente figura se muestran los componentes de la herramienta Warehouse Builder.

Figura 11 - Componentes de Warehouse Builder (Oracle, 2017)



7.1.3 Pentaho Data Integration

Es una herramienta Open Source alternativa a las soluciones propietarias tradicionales como Mcrosoft y Oracle, por lo que incluye los mismos componentes que se encuentran en estas soluciones de BI propietarias.

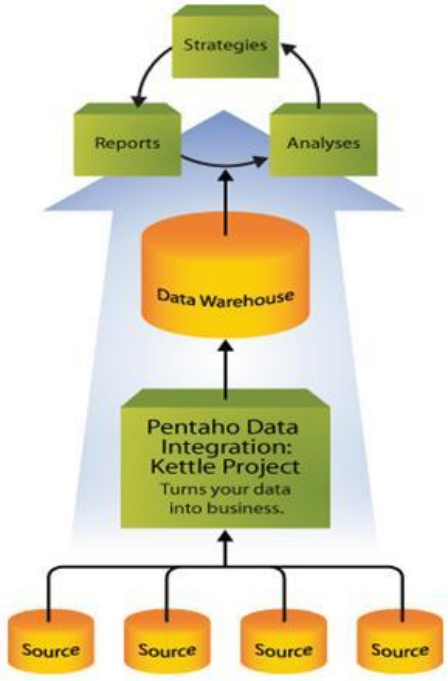
De acuerdo al sitio de Pentaho (Oseguera, 2012) la herramienta Pentaho Data Integration abre, limpia e integra la información y la pone en manos del usuario. Provee una consistencia, una sola versión de todos los recursos de información, que es uno de los más grandes desafíos para las organizaciones TI hoy en día.

A parte de ser open source y sin costes de licencia, las características básicas de esta herramienta son:

- Entorno gráfico de desarrollo
- Uso de tecnologías estándar: Java, XML, JavaScript
- Fácil de instalar y configurar
- Multiplataforma: windows, macintosh, Linux
- Basado en dos tipos de objetos: Transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones)
- Incluye cuatro herramientas:
    - Spoon: para diseñar transformaciones ETL usando el entorno gráfico
    - PAN: para ejecutar transformaciones diseñadas con spoon
    - CHEF: para crear trabajos
    - Kitchen: para ejecutar trabajos.

La arquitectura de Pentaho Data Integration viene representada por el siguiente esquema:

Figura 12 - Arquitectura Pentaho Data Integration (Oseguera, 2012)



7.1.4 Análisis de las Herramientas

Para escoger una herramienta no sólo se deben comparar sus características y capacidades, sino también el entorno de la empresa donde se va a utilizar ya que no siempre la herramienta más costosa es la mejor.

A continuación, se presenta un cuadro comparativo con las principales características de las tres herramientas evaluadas:

Tabla 2 - Cuadro comparativo de características de las herramientas ETL (autoría propia)

	Microsoft Integration Services	Oracle Warehouse Builder	Pentaho Data Integration

Facilidad de Uso		Se pueden realizar tareas de migración fácilmente usando tareas visuales	Fácil cuando se trata de información almacenada en bases de datos Oracle, debido a las herramientas Data Pump y transportable tablespaces, pero no ofrece mucha compatibilidad con otras bases de datos	Tiene la GUI más fácil de utilizar dentro de las alternativas OpenSource
Soporte		Soporte vía plataforma TechNet	Vía soporte local Oracle Latinoamérica	Soporte en Estados Unidos, reino Unido y consultorías asociadas
Implementación	Plataforma	Windows Server	Oracle, Linux	Cualquiera compatible con Java
	RAM	2 GB	2 GB	512 MB
	CPU	2.2 GHZ 2 Cores	Varía	1 GHZ
	Extra			Puede utilizar Slave Servers
Velocidad		La velocidad es proporcional al servicio MSSQL en el cual esté trabajando	La velocidad es proporcional al servicio Oracle en el cual esté trabajando	Al requerir Java Database Connector disminuye la velocidad de transacciones
Calidad de datos		Requiere del software SQL Server Data Quality Services para ofrecer herramientas DQ	Permite DQ mediante el uso de Oracle Warehouse Builder Data Profiling Features	Ofrece herraminetas para SQL dentro de su GUI, sentencias SQL personalizadas así como herramientas Javascript y REGEX para la depuración de la información
Monitoreo		Tiene herramientas prácticas y extensivas de monitoreo y registro histórico	Tiene herramientas prácticas y extensivas de monitoreo y registro histórico	Tiene herramientas prácticas de monitoreo y registro histórico
Conectividad		Bases de datos SQL Server, Access, ADO .NET	Solamente compatible con bases de datos Oracle mismas que la	Varias bases de datos, archivos planos, XML, Excel, servicios Web.

		instalada en el DataWarehouse	

La información principal para este trabajo se obtendrá de bases de datos PostgresSQL y Oracle. La herramienta Oracle Warehouse Builder se descarta por no ofrecer una compatibilidad con otros gestores de bases de datos diferente a Oracle (Bustillos, 2014).

De las otras dos herramientas que se están evaluando (Microsoft Integration Services y Pentaho Data integration) se puede decir que tienen características similares y que su principal diferencia es el costo de licenciamiento. Sin embargo, la característica de velocidad es importante debido a que se manejará un volumen de datos alto, ya que se tienen datos históricos de hace 6 años que deben utilizarse en el cálculo de indicadores de calidad del servicio. Por lo tanto, y dado que CHEC ya cuenta con licencias de Microsoft Integration Services, se opta por trabajar con esta herramienta. El uso de esta herramienta implica adicionalmente que el repositorio del datamart deberá estar en SQL Server, buscando compatibilidad entre el gestor de base de datos y la herramienta de ETL a usar. Cabe aclarar que la empresa también cuenta con licencias de SQL Server 2008 y que estas se pueden usar para el desarrollo a realizar.

7.2 DESARROLLO DE LA METODOLOGÍA

Como se había indicado, el marco de referencia será la metodología Kimball. Al ser una guía que pretende llevar a buen término la creación de un almacén de datos, se desarrollaron las tareas más relevantes que son necesarias para el desarrollo de un proyecto de data mart o data warehouse:

7.2.1 Planeación del Proyecto

Alcance del proyecto

Desarrollar una solución informática que permita al Área de Gestión Operativa de CHEC unificar los datos necesarios para el cálculo de los indicadores de calidad SAIDI – SAIFI,

con el fin de que se pueda visualizar y analizar para la toma de decisiones, utilizando herramientas para tal fin.

7.2.2 Definición de Requerimientos del Negocio:

En la definición de los requerimientos se hace de vital importancia acudir a las personas que trabajan en el día a día con los datos que permiten calcular los indicadores de calidad del servicio.

La metodología Kimball indica que se debe recurrir a artefactos como entrevistas, encuestas o cuestionarios, entre otros, que permiten registrar de forma ordenada y concisa las necesidades de cada uno de ellos. Sin embargo, esta metodología no indica cómo llegar a la identificación de los requisitos del negocio, por lo que se hace necesario recurrir a otras metodologías como Hefesto para su consecución.

En este apartado se hace una adaptación de la metodología Hefesto para la identificación y el análisis de los requisitos. Para la consecución de este punto se realizó lo siguiente:

Identificación de preguntas

Como instrumento de recolección de datos se utiliza una encuesta, la cual tiene como objetivo recoger información de las variables que requieren los funcionarios para generar los indicadores de calidad del servicio y apoyar la toma de decisiones a nivel operativo, ejecutivo y direccional. La encuesta fue dirigida a profesionales y asistentes del área de Gestión Operativa de CHEC.

Para identificar las necesidades de los usuarios, hay que partir de lo conocido, es decir, de los indicadores con los que cuenta el negocio para medir su calidad y las variables utilizadas para su cálculo. Esta acción permite evaluar el nivel de conocimiento de las personas que trabajan para el negocio y su claridad frente a los objetivos estratégicos de la empresa.

Es importante saber si los indicadores actuales aportan información valiosa al negocio en la toma de decisiones, o si por el contrario hace falta tener en cuenta otro tipo de variables que requieren ser medidas y que aporten a la mejora de la calidad del servicio.

Por lo anterior, se diseñó la encuesta para recoger estos elementos y se aplicó a los usuarios del Área de Gestión Operativa que cumplen con los roles operacional y ejecutivo (Ver Anexo 1).

La aplicación de este instrumento permite identificar los valores cuantificables que realmente se desean analizar (*indicadores*); y las variables mediante los cuales se requiere examinar los indicadores (*perspectivas*).

Identificación de indicadores y perspectivas

De acuerdo a los resultados de la encuesta, se identifican los siguientes indicadores y perspectivas:

Indicadores:

- Indicador SAIDI: Índice de Duración Promedio de las Interrupciones del Sistema de Distribución (System Average Interruption Duration Index).
  Es el promedio de duración en horas en las cuales un usuario del sistema estuvo sin servicio de energía eléctrica. Tiene la siguiente fórmula:

$$\boldsymbol{SAIDI = \frac{\sum_{i=1}^{n} Ca(i) * t(i)\, A}{Cs}}$$

  Donde:

  Ca(i) es el número de clientes afectados por la interrupción i.

  t(i) es el tiempo de duración de la interrupción (i).

  Cs es el número total de clientes conectados al sistema.

n es el número total de interrupciones contabilizadas en el sistema.

- Indicador SAIFI: Índice de Frecuencia Promedio de Interrupciones del Sistema de Distribución (System Average Interruption Frequency Index).
  Es el número de interrupciones promedio en las cuales un usuario del sistema estuvo sin servicio de energía eléctrica. Tiene la siguiente fórmula:

$$\boldsymbol{SAIFI = \frac{\sum_{i=1}^{n} Ca(i)}{Cs}}$$

  Donde:

  Ca(i) es el número de clientes afectados por la interrupción i.

  Cs es el número total de clientes conectados al sistema.

  n es el número total de interrupciones contabilizadas en el sistema.

- Duración en horas de la afectación del servicio: Es la duración en horas del transformador afectado.
  Número de usuarios afectados: Es la cantidad de usuarios asociados a un transformador afectado por la ocurrencia de un evento.
  Demanda no atendida aproximada: Se calcula para cada transformador afectado dividiendo su consumo mensual entre 720 horas del mes para obtener un consumo aproximado por horas, este resultado se multiplica por la duración del transformador afectado en horas:

$$\frac{Consumo\ de\ energía\ *\ Duración\ en\ horas\ de\ la\ afectación\ del\ servicio}{720\ horas\ del\ mes}$$

- Aporte a los indicadores en términos de porcentaje el cual se obtiene de la siguiente formula:

$$\frac{Valor\ SAIDI\ o\ SAIFI\ por\ transformador}{Valor\ SAIDI\ o\ SAIFI\ del\ sistema}$$

Perspectivas:

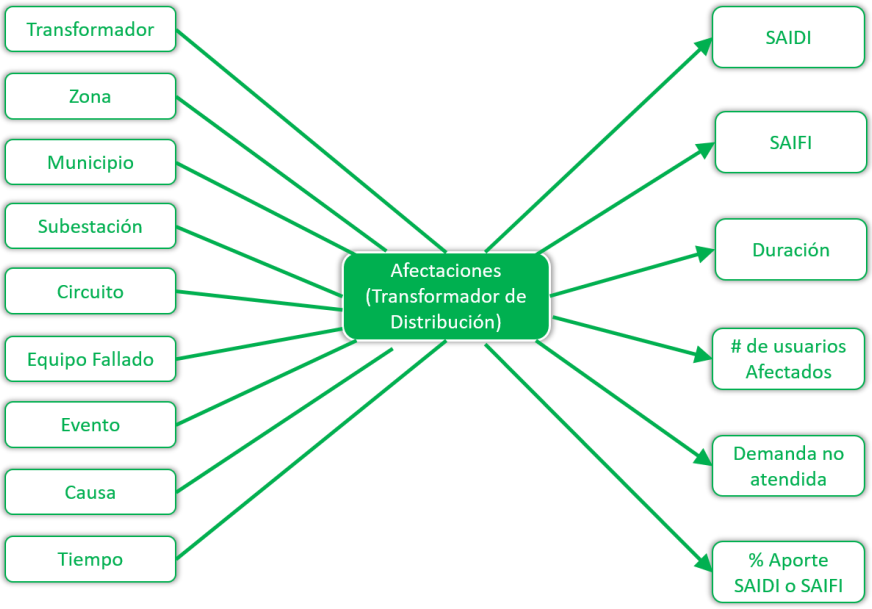
- Transformador
- Zona
- Municipio
- Subestación
- Circuito
- Equipo fallado
- Evento
- Causa
- Tiempo

Modelo Conceptual

A continuación, se muestra un modelo conceptual de los indicadores y perspectivas identificados, los cuales serán el insumo para el modelo dimensional con sus hechos y dimensiones. Como se puede ver en la Figura 13, a la izquierda se listan las perspectivas identificadas y a la derecha los indicadores que se tendrán en el modelo.

La relación entre las perspectivas y los indicadores está dada por las afectaciones sobre los transformadores de distribución. Esto ocurre porque la calidad del servicio se debe medir basados en el número de interrupciones y la duración de éstas en cada transformador de la red eléctrica, ya que cada usuario identificado en CHEC está conectado a un transformador, el cual provee el servicio de energía.

Figura 13 – Mapa conceptual de requerimientos



7.2.3 Modelo Dimensional

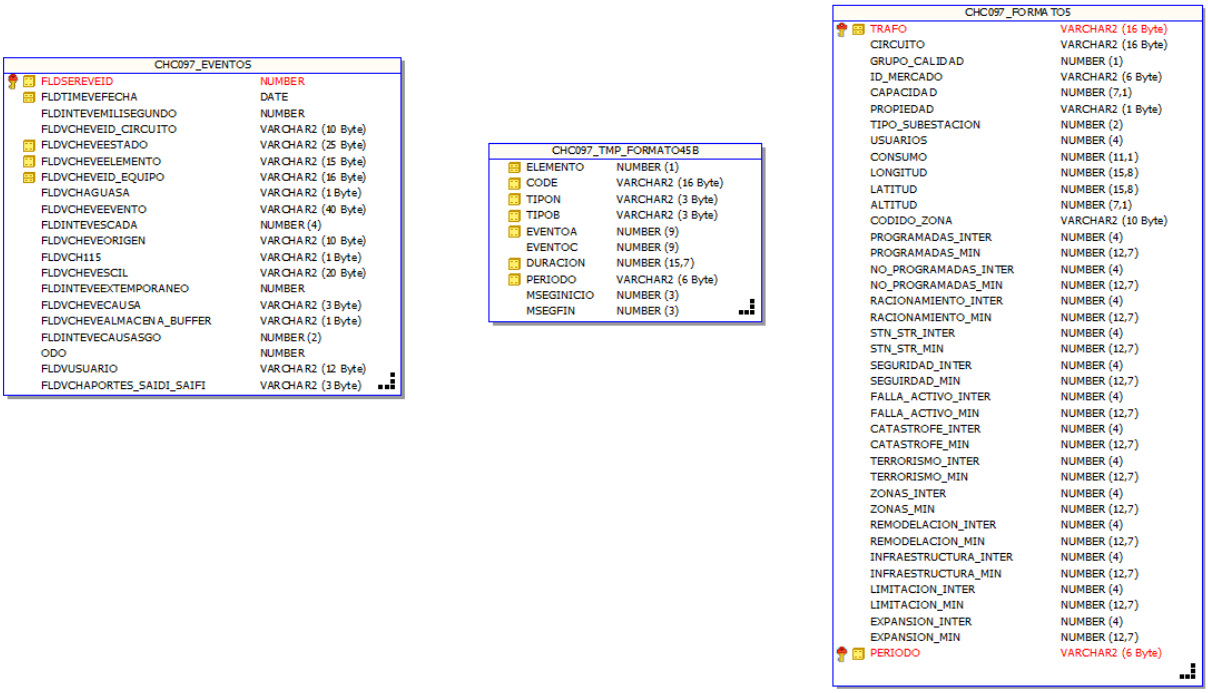
A partir de este modelo se deben explorar los diferentes sistemas OLTP disponibles, que contengan la información requerida para poder identificar las correspondencias entre el modelo conceptual y las fuentes de datos.

Se identifican tres bases de datos donde se puede generar la información requerida:

CALIDAD097: Es una base de datos Oracle donde se encuentra almacenada la información relacionada con calidad del servicio y la afectación de los transformadores por la ocurrencia de eventos, al igual que la información de dichos eventos generados por la operación o falla de alguno de los elementos del sistema de distribución de la empresa y que causan afectación del servicio a los usuarios.

Se identificaron los siguientes objetos de donde se puede obtener información relevante para la construcción del modelo dimensional.

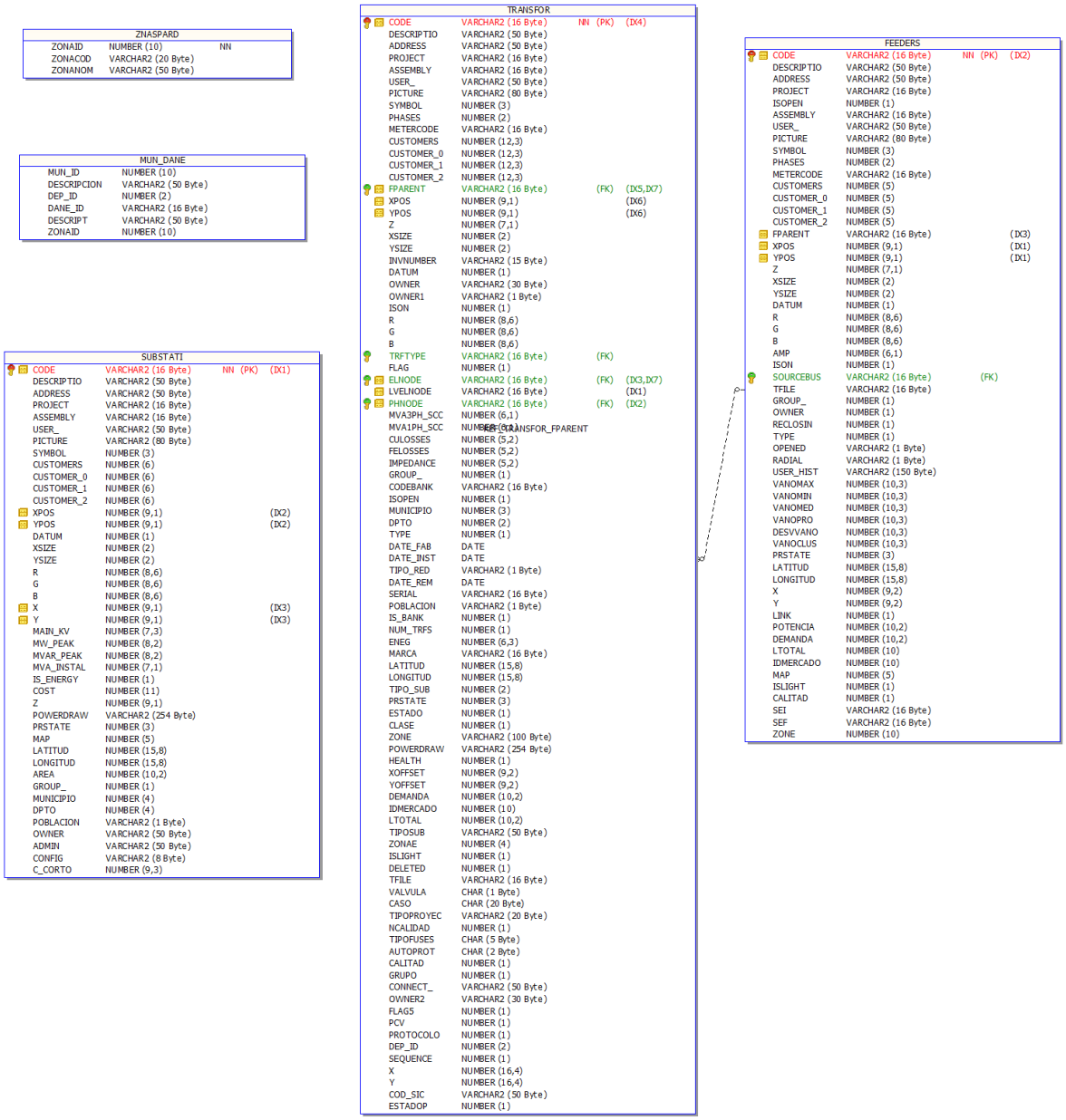
Figura 14 – Tablas esquema CALIDAD097



AREDES: Es una base de datos Oracle donde se encuentra almacenada la información relacionada con la topología de la red eléctrica de la empresa. Dicha topología está compuesta por transformadores de distribución, usuarios asociados, cables, equipos de corte, subestaciones, entre otros. La información comprende características de equipos, tipos de equipos, ubicación geográfica y todos aquellos datos requeridos para conocer de la infraestructura eléctrica de la empresa.

Se identificaron los siguientes objetos de donde se puede obtener información relevante para la construcción del modelo dimensional.

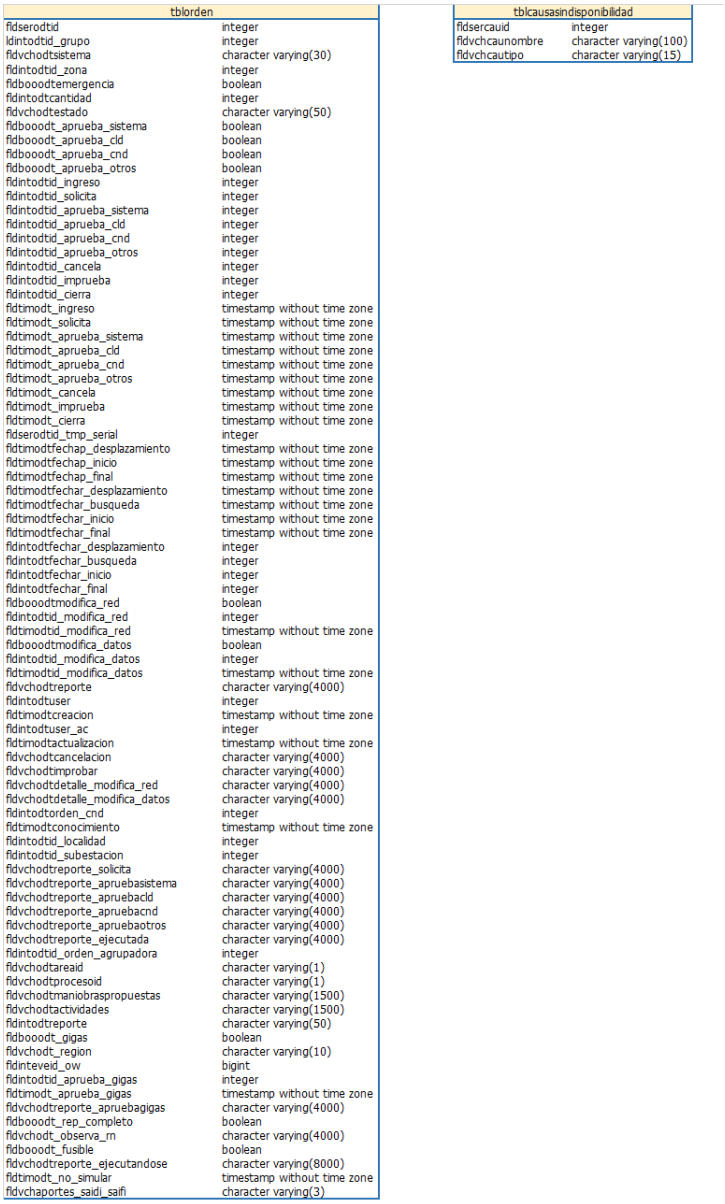
Figura 15 – Tablas esquema AREDES



SGC: Es una base de datos Postgres donde se encuentra almacenada la información relacionada con las órdenes de operación y los eventos y maniobras que se registran sobre la red eléctrica.

Se identificaron los siguientes objetos de donde se puede obtener información relevante para la construcción del modelo dimensional.

Figura 16 – Tablas esquema SGC

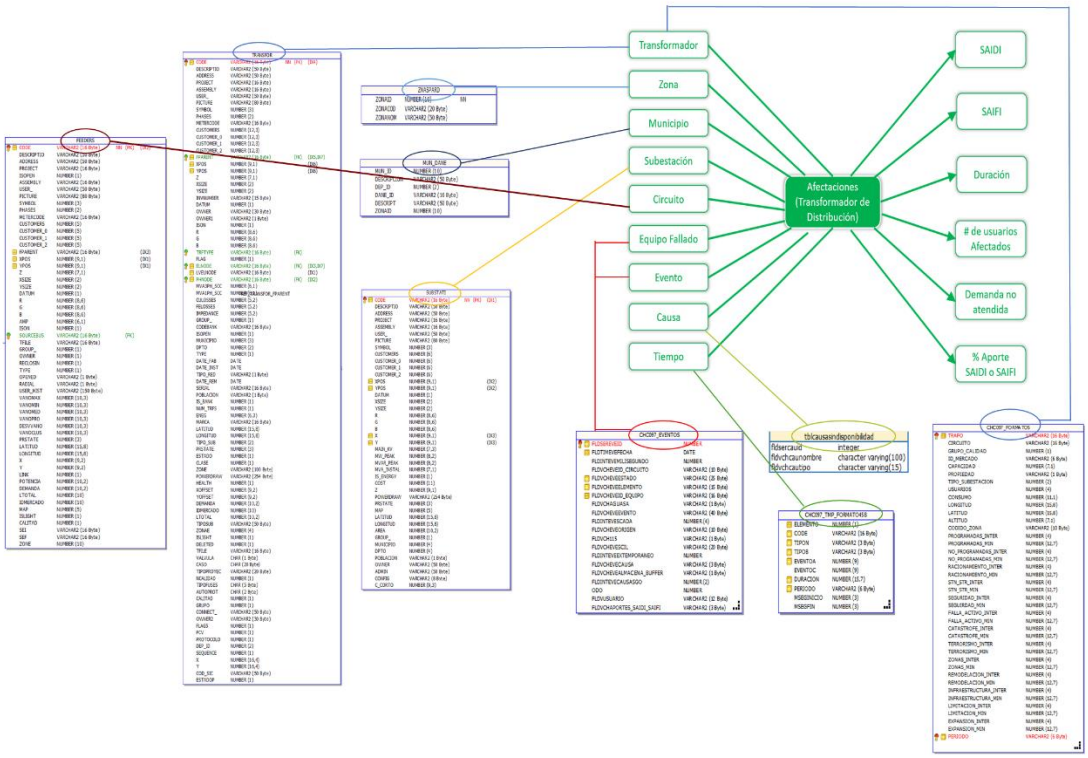


tblorden	
fldserodtid	integer
ldintodtid_grupo	integer
fldvchodtsistema	character varying(30)
fldintodtid_zona	integer
fldbooodtemergencia	boolean
fldintodtcantidad	integer
fldvchodtestado	character varying(50)
fldbooodt_aprueba_sistema	boolean
fldbooodt_aprueba_cld	boolean
fldbooodt_aprueba_cnd	boolean
fldbooodt_aprueba_otros	boolean
fldintodtid_ingreso	integer
fldintodtid_solicita	integer
fldintodtid_aprueba_sistema	integer
fldintodtid_aprueba_cld	integer
fldintodtid_aprueba_cnd	integer
fldintodtid_aprueba_otros	integer
fldintodtid_cancela	integer
fldintodtid_imprueba	integer
fldintodtid_cierra	integer
fldtimodt_ingreso	timestamp without time zone
fldtimodt_solicita	timestamp without time zone
fldtimodt_aprueba_sistema	timestamp without time zone
fldtimodt_aprueba_cld	timestamp without time zone
fldtimodt_aprueba_cnd	timestamp without time zone
fldtimodt_aprueba_otros	timestamp without time zone
fldtimodt_cancela	timestamp without time zone
fldtimodt_imprueba	timestamp without time zone
fldtimodt_cierra	timestamp without time zone
fldserodtid_tmp_serial	integer
fldtimodtfechap_desplazamiento	timestamp without time zone
fldtimodtfechap_inicio	timestamp without time zone
fldtimodtfechap_final	timestamp without time zone
fldtimodtfechar_desplazamiento	timestamp without time zone
fldtimodtfechar_busqueda	timestamp without time zone
fldtimodtfechar_inicio	timestamp without time zone
fldtimodtfechar_final	timestamp without time zone
fldintodtfechar_desplazamiento	integer
fldintodtfechar_busqueda	integer
fldintodtfechar_inicio	integer
fldintodtfechar_final	integer
fldbooodtmodifica_red	boolean
fldintodtid_modifica_red	integer
fldtimodtid_modifica_red	timestamp without time zone
fldbooodtmodifica_datos	boolean
fldintodtid_modifica_datos	integer
fldtimodtid_modifica_datos	timestamp without time zone
fldvchodtreporte	character varying(4000)
fldintodtuser	integer
fldtimodtcreacion	timestamp without time zone
fldintodtuser_ac	integer
fldtimodtactualizacion	timestamp without time zone
fldvchodtcancelacion	character varying(4000)
fldvchodtimprobar	character varying(4000)
fldvchodtdetalle_modifica_red	character varying(4000)
fldvchodtdetalle_modifica_datos	character varying(4000)
fldintodtorden_cnd	integer
fldtimodtconocimiento	timestamp without time zone
fldintodtid_localidad	integer
fldintodtid_subestacion	integer
fldvchodtreporte_solicita	character varying(4000)
fldvchodtreporte_apruebasistema	character varying(4000)
fldvchodtreporte_apruebacld	character varying(4000)
fldvchodtreporte_apruebacnd	character varying(4000)
fldvchodtreporte_apruebaotros	character varying(4000)
fldvchodtreporte_ejecutada	character varying(4000)
fldintodtid_orden_agrupadora	integer
fldvchodtareaid	character varying(1)
fldvchodtprocesoid	character varying(1)
fldvchodtmaniobraspropuestas	character varying(1500)
fldvchodtactividades	character varying(1500)
fldintodtreporte	character varying(50)
fldbooodt_gigas	boolean
fldvchodt_region	character varying(10)
fldinteveid_ow	bigint
fldintodtid_aprueba_gigas	integer
fldtimodt_aprueba_gigas	timestamp without time zone
fldvchodtreporte_apruebagigas	character varying(4000)
fldbooodt_rep_completo	boolean
fldvchodt_observa_rn	character varying(4000)
fldbooodt_fusible	boolean
fldvchodtreporte_ejecutandose	character varying(8000)
fldtimodt_no_simular	timestamp without time zone
fldvchaportes_saidi_saifi	character varying(3)

tblcausasindisponibilidad	
fldsercauid	integer
fldvchcaunombre	character varying(100)
fldvchcautipo	character varying(15)

Después de realizar la identificación de los objetos en las bases de datos OLTP se hacen los cruces de correspondencia con las perspectivas y los indicadores identificados en el modelo conceptual:

Figura 17 – Correspondencia perspectivas e indicadores con objetos de BD OLTP



Como se puede ver en la Figura 17 se tienen correspondencias en las bases de datos OLTP con las perspectivas. Sin embargo, no hay una relación directa entre los indicadores y las bases de datos, pues son cálculos que se deben realizar una vez se carguen los datos al *data mart*. A continuación, se nombran las correspondencias identificadas:

- La perspectiva “Transformador” se relaciona con las tablas TRANSFOR y CHC097_FORMATO5.
- La perspectiva “Zona” se relaciona con la tabla ZNASPARD.
- La perspectiva “Municipio” se relaciona con la tabla MUNDANE.
- La perspectiva “Subestación” se relaciona con la tabla SUBSTATI.
- La perspectiva “Circuito” se relaciona con la tabla FEEDERS.
- La perspectiva “Equipo fallado” se relaciona con la tabla “CHC097.EVENTOS”.
- La perspectiva “Evento” se relaciona con la tabla CHC097.EVENTOS.

- La perspectiva “Causa” se relaciona con la tabla TBLCAUSASINDISPONIBILIDAD.
- La perspectiva “Tiempo” se relaciona con la tabla CHC097.TEMP_FORMATO45B por medio del campo “Periodo”, donde se observa el año y el mes en formato “yyyymm”.
- El indicador “SAIDI” se relaciona con la tabla CHC097.TEMP_FORMATO45B por medio del campo “Duración” y con la tabla “CHC097.FORMATO5” por medio del campo “Usuarios”.
- El indicador “SAIFI” se relaciona con la tabla CHC097.FORMATO5 por medio del campo “Usuarios”.
- El indicador “Duración” se relaciona con la tabla CHC097.TEMP_FORMATO45B por medio del campo “Duración”.
- El indicador “# de usuarios afectados” se relaciona con la tabla CHC097.FORMATO5 por medio del campo “Usuarios”.
- El indicador “Demanda No Atendida” se relaciona con la tabla CHC097.TEMP_FORMATO45B por medio del campo “Duración”, y con la tabla CHC097.FORMATO5 por medio del campo “Consumo”
- El indicador “% Aporte a SAIDI o SAIFI” se relaciona con la tabla CHC097.TEMP_FORMATO45B por medio del campo “Duración” y con la tabla CHC097.FORMATO5 por medio del campo “Usuarios”.

7.2.4 Diseño Físico

Se utilizará un modelo en estrella ya que, según el análisis del modelo conceptual y los orígenes de datos, las dimensiones no requieren ser desagregadas para obtener la información requerida.

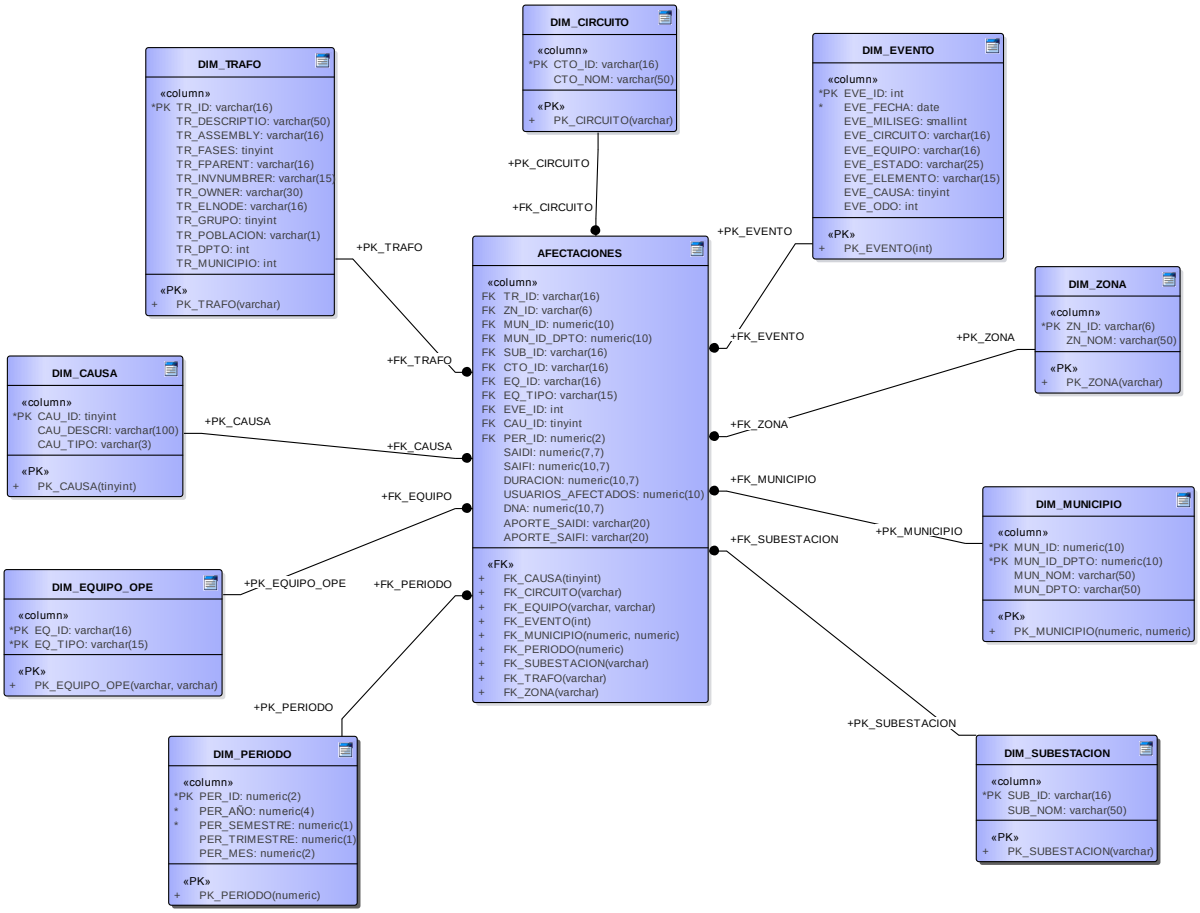
A continuación, se describen algunas ventajas de este modelo:

- Este modelo brinda gran velocidad en los tiempos de respuesta al consultar la información.

- Fácil de entender y de optimizar, no se requerirían muchos cambios ante un nuevo requerimiento.
- Soportado por la mayoría de herramientas de consulta y análisis.
- Fácil de implementar después del análisis.

En la siguiente figura se visualiza un bosquejo del modelo en estrella:

Figura 18 – Modelo en Estrella



Tablas de Dimensiones

En este paso cada perspectiva planteada en el modelo conceptual se convertirá en una tabla de dimensión con sus respectivos campos. Para ello es necesario seguir los siguientes pasos:

- A cada tabla de dimensión se le dará un nombre que la identifique.
- Cada tabla de dimensión debe tener al menos un campo como llave principal.
- Se debe dar nombre a los campos que sean lo suficientemente claros e intuitivos.

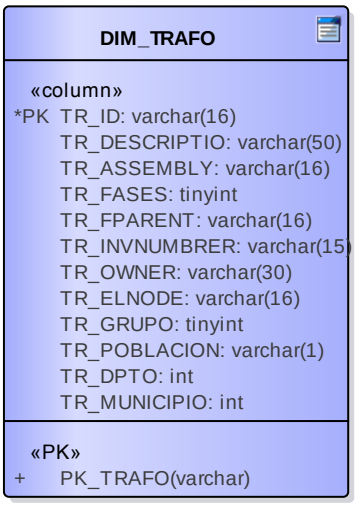
A continuación, se definen las tablas de dimensiones:

Perspectiva Transformador:

Se tendrá una tabla de dimensión que tendrá por nombre "DIM_TRAFO". Los campos que la componen serán los siguientes:

- TR_ID: Campo tipo alfanumérico de 16 caracteres. Será la llave primaria de la tabla y tendrá el código nemotécnico del transformador que asigna la empresa.
- TR_DESCRIPTIO: Campo tipo alfanumérico de 50 caracteres. En este campo se guardará el nombre del transformador.
- TR_ASSEMBLY: Campo tipo alfanumérico de 16 caracteres. Acá se guardará el código que lleva el inventario del transformador.
- TR_FASES: Campo tipo entero pequeño. Campo para identificar las fases que tiene el transformador.
- TR_FPARENT: Campo tipo alfanumérico de 16 caracteres. Identifica el código del alimentador del transformador.
- TR_INVNUMBER: Campo tipo alfanumérico de 16 caracteres. Identifica el número en el inventario de la empresa.
- TR_OWNER: Campo tipo alfanumérico de 30 caracteres. Identifica el nombre del propietario del transformador.
- TR_ELNODE: Campo tipo alfanumérico de 16 caracteres. Identifica el código del nodo eléctrico asociado al transformador.
- TR_GRUPO: Campo tipo entero pequeño. Campo para identificar el grupo de calidad al que pertenece el transformador.
- TR_POBLACION: Campo tipo alfanumérico de 16 caracteres. Identifica el tipo de población. (U: Urbana; R: Rural)

- TR_DPTO: Campo tipo entero. Identifica el código del departamento donde está ubicado físicamente el transformador
- TR_MUNICIPIO: Campo tipo entero. Identifica el código del municipio donde está ubicado físicamente el transformador

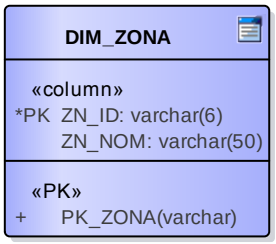


Perspectiva Zona:

Se tendrá una tabla de dimensión que tendrá por nombre “DIM_ZONA”. Los campos que la componen serán los siguientes:

- ZN_ID: Campo tipo alfanumérico de 6 caracteres. Será la llave primaria de la tabla.

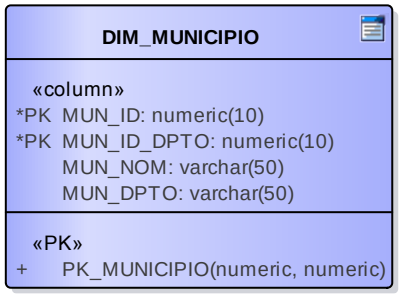
ZN_NOM: Campo tipo alfanumérico de 50 caracteres. En este campo se guardará el nombre de la zona.



Perspectiva Municipio:

Se tendrá una tabla de dimensión que tendrá por nombre “DIM_MUNICIPIO”. Los campos que la componen serán los siguientes:

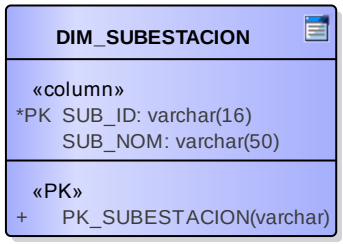
- MUN_ID: Campo tipo numérico de 10 dígitos. Será parte de la llave primaria de la tabla.
- MUN_ID_DPTO: Campo tipo numérico de 10 dígitos. Será parte de la llave primaria de la tabla.
- MUN_NOM: Campo tipo alfanumérico de 50 caracteres. En este campo se guardará el nombre del municipio.
- MUN_DPTO: Campo tipo alfanumérico de 50 caracteres. En este campo se guardará el nombre del departamento.



Perspectiva Subestación:

Se tendrá una tabla de dimensión que tendrá por nombre “DIM_SUBESTACION”. Los campos que la componen serán los siguientes:

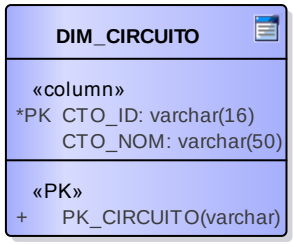
- SUB_ID: Campo tipo alfanumérico de 16 caracteres. Será la llave primaria de la tabla.
- SUB_NOM: Campo tipo alfanumérico de 50 caracteres. En este campo se guardará el nombre de la subestación.



Perspectiva Circuito:

Se tendrá una tabla de dimensión que tendrá por nombre "DIM_CIRCUITO". Los campos que la componen serán los siguientes:

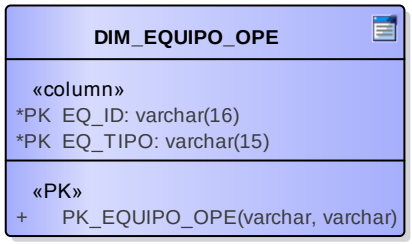
- CTO_ID: Campo tipo alfanumérico de 16 caracteres. Será la llave primaria de la tabla.
- CTO_NOM: Campo tipo alfanumérico de 50 caracteres. En este campo se guardará el nombre del circuito.



Perspectiva Equipo fallado:

Se tendrá una tabla de dimensión que tendrá por nombre "DIM_EQUIPO_OPE". Los campos que la componen serán los siguientes:

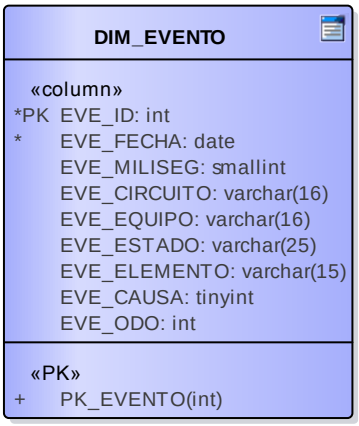
- EQ_ID: Campo tipo alfanumérico de 16 caracteres. Será parte de la llave primaria de la tabla.
- EQ_TIPO: Campo tipo alfanumérico de 15 caracteres. En este campo se guardará el tipo de equipo (interruptor, transformador). Será parte de la llave primaria de la tabla.



Perspectiva Evento:

Se tendrá una tabla de dimensión que tendrá por nombre “DIM_EVENTO”. Los campos que la componen serán los siguientes:

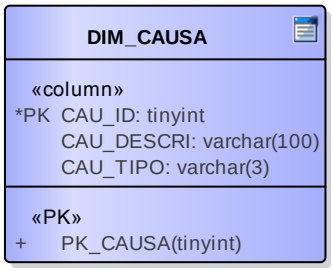
- EVE_ID: Campo tipo entero. Será la llave primaria de la tabla. Identifica un evento o maniobra en el sistema.
- EVE_FECHA: Campo tipo fecha. Identifica la fecha y hora de ocurrencia de un evento o maniobra.
- EVE_MILISEG: Campo tipo entero pequeño. Identifica los milisegundos de la hora y fecha de ocurrencia del evento o maniobra.
- EVE_CIRCUITO: Campo tipo alfanumérico de 16 caracteres. Identifica el código nemotécnico del circuito.
- EVE_EQUIPO: Campo tipo alfanumérico de 16 caracteres. Identifica el código nemotécnico del equipo.
- EVE_ESTADO: Campo tipo alfanumérico de 25 caracteres. Identifica estado del evento o maniobra (APERTURA / CIERRE).
- EVE_ELEMENTO: Campo tipo alfanumérico de 15 caracteres. Identifica el tipo de equipo operado (interruptor / transformador / tramo de línea / barraje).
- EVE_CAUSA: Campo tipo entero pequeño. Identifica el código de la causa operativa asociada al evento o maniobra.
- EVE_ODO: Campo tipo entero. Identifica el número de la orden de operación con el cual se atendió el evento o maniobra.



Perspectiva Causa:

Se tendrá una tabla de dimensión que tendrá por nombre "DIM_CAUSA". Los campos que la componen serán los siguientes:

- CAU_ID: Campo tipo entero pequeño. Será la llave primaria de la tabla.
- CAU_DESCRI: Campo tipo alfanumérico de 100 caracteres. En este campo se guardará la descripción de la causa.
- CAU_TIPO: Campo tipo alfanumérico de 3 caracteres. En este campo se guardará el tipo de causa (PRO; NPR; EXC).

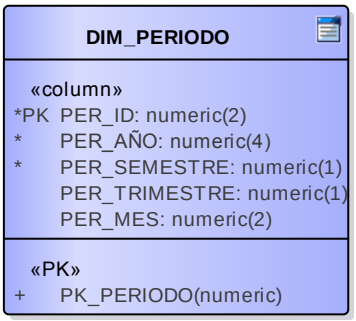


Perspectiva Tiempo:

Se tendrá una tabla de dimensión que tendrá por nombre "DIM_PERIODO". Los campos que la componen serán los siguientes:

- PER_ID: Campo tipo numérico de 2 dígitos. Será la llave primaria de la tabla.

- PER_AÑO: Campo tipo numérico de 4 dígitos.
- PER_SEMESTRE: Campo tipo numérico de un dígito.
- PER_TRIMESTRE: Campo tipo numérico de un dígito.
- PER_MES: Campo tipo numérico de 2 dígitos.

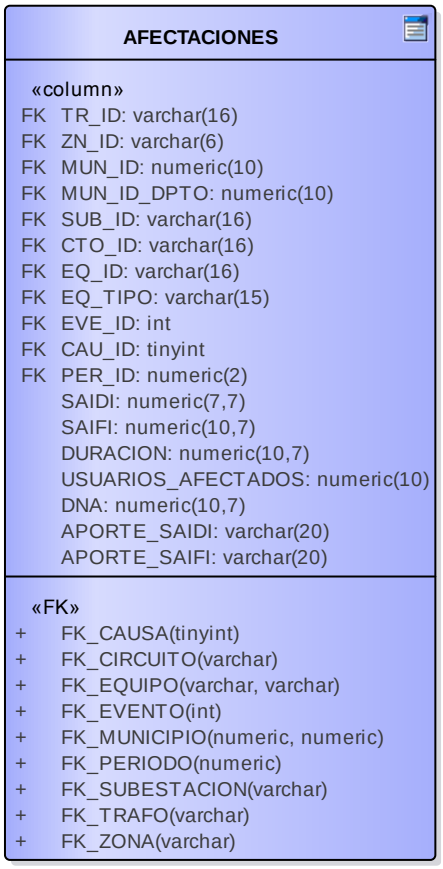


Tablas de Hechos

Se contará con una tabla de hechos que tendrá por nombre "AFECTACIONES". Los campos que la componen serán los siguientes:

- TR_ID: Campo tipo alfanumérico de 16 caracteres. Es llave foránea y se relaciona con la tabla DIM_TRAFO.
- ZN_ID: Campo tipo alfanumérico de 6 caracteres. Es llave foránea y se relaciona con la tabla DIM_ZONA.
- MUN_ID: Campo tipo numérico de 10 dígitos. Hace parte de la llave foránea que se relaciona con la tabla DIM_MUNICIPIO.
- MUN_ID_DPTO: Campo tipo numérico de 10 dígitos. Hace parte de la llave foránea que se relaciona con la tabla DIM_MUNICIPIO.
- SUB_ID: Campo tipo alfanumérico de 16 caracteres. Es llave foránea y se relaciona con la tabla DIM_SUBESTACION.
- CTO_ID: Campo tipo alfanumérico de 16 caracteres. Es llave foránea y se relaciona con la tabla DIM_CIRCUITO.
- EQ_ID: Campo tipo alfanumérico de 16 caracteres. Hace parte de la llave foránea que se relaciona con la tabla DIM_EQUIPO_OPE.

- EQ_TIPO: Campo tipo alfanumérico de 15 caracteres. Hace parte de la llave foránea que se relaciona con la tabla DIM_EQUIPO_OPE.
- EVE_ID: Campo tipo entero. Es llave foránea y se relaciona con la tabla DIM_EVENTO.
- CAU_ID: Campo tipo entero pequeño. Es llave foránea y se relaciona con la tabla DIM_CAUSA.
- PER_ID: Campo tipo numérico de 2 dígitos. Es llave foránea y se relaciona con la tabla DIM_PERIODO.
- SAIDI: Campo tipo numérico de 10 dígitos hasta con 7 decimales. Campo para el indicador SAIDI que mide el tiempo promedio de interrupción por usuario.
- SAIFI: Campo tipo numérico de 10 dígitos hasta con 7 decimales. Campo para el indicador SAIFI que mide la frecuencia media de interrupción por usuario.
- DURACION: Campo tipo numérico de 10 dígitos hasta con 7 decimales. Campo para guardar la duración en minutos por interrupción.
- USUARIOS_AFECTADOS: Campo tipo numérico de 10 dígitos. Campo para guardar el número de usuarios afectados por interrupción.
- DNA: Campo tipo numérico de 10 dígitos hasta con 7 decimales. Campo para guardar el cálculo de demanda no atendida en KV/h durante el tiempo de interrupción.
- APORTE_SAIDI: Campo tipo alfanumérico de 20 caracteres. Campo para guardar el porcentaje de aporte que un equipo hizo al indicador general del SAIDI.
- APORTE_SAIFI: Campo tipo alfanumérico de 20 caracteres. Campo para guardar el porcentaje de aporte que un equipo hizo al indicador general del SAIFI.



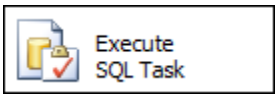
7.2.5 Integración de Datos

Una vez construido el modelo lógico, se procede a poblarlo utilizando procesos de extracción, transformación y/o limpieza y cargue de datos o ETLs. Los procesos ETLs se desarrollan con la herramienta Microsoft Integration Services, la cual permite realizar la carga de información a través de componentes propios muy intuitivos, por lo que se facilita su uso. A continuación, se describen los componentes utilizados:

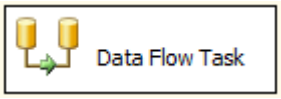
Ejecutar Tarea de SQL: Esta opción puede ser usada para ejecutar cualquier comando SQL o procedimiento almacenado. La tarea puede contener una sola instrucción SQL o múltiples instrucciones SQL que se ejecutarán de forma secuencial. Se puede usar la tarea Ejecutar SQL para los siguientes fines:

- Truncar una tabla o vista en preparación para insertar datos.

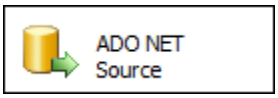
- Crear, modificar y quitar objetos de base de datos, como tablas y vistas.
- Volver a crear tablas de hechos y tablas de dimensiones antes de cargar datos en ellas.
- Ejecutar procedimientos almacenados. Si la instrucción SQL invoca un procedimiento almacenado que devuelve resultados de una tabla temporal, use la opción WITH RESULT SETS para definir los metadatos del conjunto de resultados.
- Guardar en una variable el conjunto de filas devuelto por una consulta.



Tarea Flujo de datos: La tarea Flujo de datos encapsula el flujo que mueve datos entre orígenes y destinos, y permite al usuario transformar, limpiar y modificar datos a medida que se mueven.



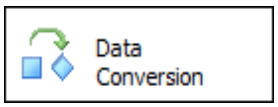
Origen ADO NET: Permite realizar una conexión a base de datos compatible con el componente ADO NET y habilita los datos obtenidos para el flujo de datos.



Conversión de datos: Convierte los datos de una columna de entrada a otro tipo de datos diferente y después los copia a una nueva columna de salida. Por ejemplo, un paquete puede extraer los datos de diferentes orígenes y después usar esta transformación para convertir las columnas al tipo de datos necesario para el almacén de datos de destino. Puede aplicar múltiples conversiones a una sola columna de entrada.

Un paquete puede utilizar esta transformación para realizar las siguientes conversiones de tipos de datos:

- Cambiar el tipo de datos.
- Establecer la longitud de la columna de los datos de cadena, así como la precisión y la escala de los datos numéricos.



Destino ADO NET: Permite realizar una conexión a base de datos compatible con ADO NET. Permite cargar los datos en una tabla o vista existente, o bien puede crear una nueva tabla y cargar los datos en ella.

Con estos elementos se configuran los siguientes ETLs:

ETL_DIM_1

Se configuró este ETL para cargar los datos de las tablas DIM_TRAFO y DIM_EVENTO. Se separó de las demás tablas porque estas son las más grandes y su procesamiento puede tardar más tiempo que el resto de tablas.

Figura 19 – ETLs Transformadores y eventos

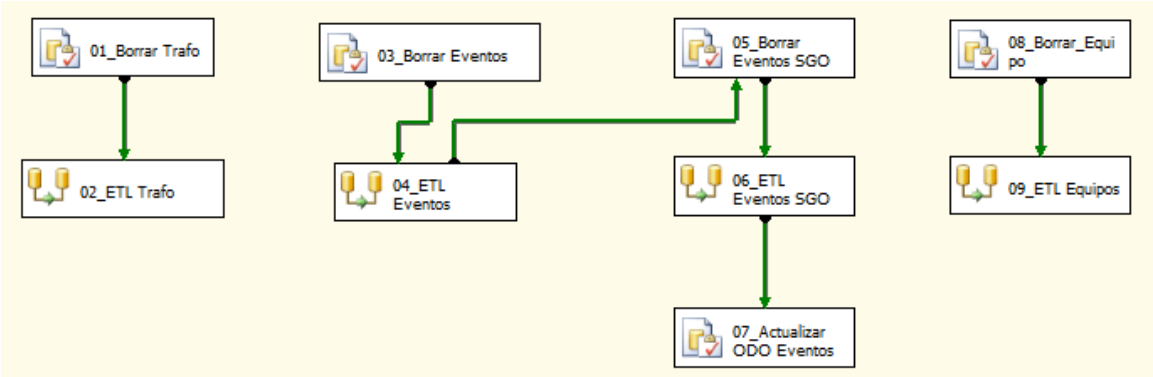
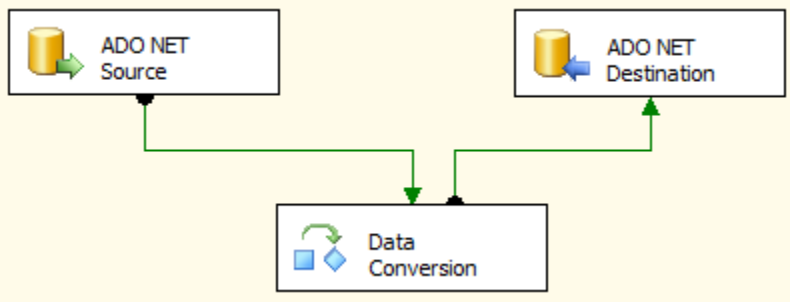


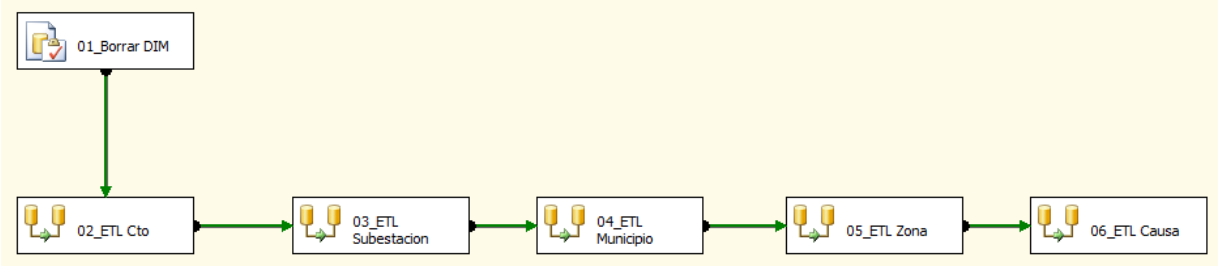
Figura 20 – Detalle ETL Eventos



ETL_DIM_2

Este ETL permite cargar la información de en las tablas DIM_CIRCUITO, DIM_SUBESTACION, DIM_MUNICIPIO, DIM_ZONA y DIM_CAUSA

Figura 21 – ETLs Dimensiones pequeñas



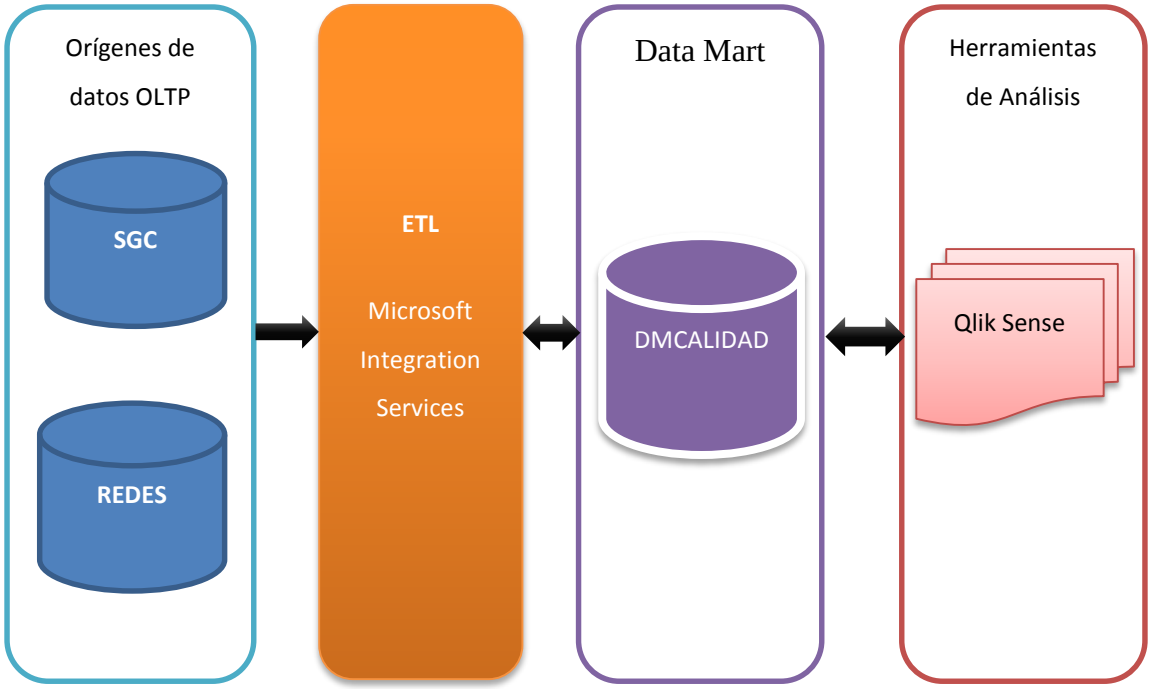
7.2.6 Diseño de la Arquitectura Técnica

El entorno tecnológico para inteligencia de negocios (BI) en CHEC, se ha desarrollado bajo herramientas Microsoft como Microsoft Integration Services y SQL Server 2008. Las directrices de arquitectura definidas en el Grupo EPM tienen tendencia a utilizar productos de este fabricante, por lo que se facilita la elección de las herramientas a utilizar.

Adicionalmente, se ha tenido experiencia en otros proyectos en la creación de ETLs tomando la información origen de bases de datos Oracle y Postgres.

En la siguiente figura se muestra un bosquejo de la arquitectura utilizada para el *data mart:*

Figura 22 – Diseño Arquitectura Tecnológica



7.2.7 Especificación de Aplicaciones de BI

Las aplicaciones de BI proporcionan información útil a los usuarios, e incluyen un amplio espectro de tipos de informes y herramientas de análisis, que van desde informes simples de formato fijo a sofisticadas aplicaciones analíticas que usan complejos algoritmos e información del dominio.

Los informes proporcionan a los usuarios un conjunto básico de información acerca de lo que está sucediendo en un área determinada de la empresa. Este tipo de aplicaciones son el caballo de batalla de la BI de la empresa. Son informes que los usuarios usan día a día. La mayor parte de lo que piden las personas durante el proceso de definición de requisitos se puede clasificar como informes.

Para este caso se desarrollan reporten en la herramienta Qlik Sense, la cual es una plataforma que permite generar informes empresariales cuyo contenido se puede extraer de diversas bases de datos y se pueden ver en distintos formatos.

8 RESULTADOS

De acuerdo al trabajo realizado, se logran los siguientes resultados:

- Se siguió un marco de referencia como la metodología de Ralph Kimball apoyado en la identificación de requerimientos en la metodología Hefesto, para desarrollar el Data mart de calidad del servicio para el área de Gestión Operativa de CHEC. Como experiencia en la creación de un data mart, sirve como referencia el paso a paso utilizado para futuros proyectos de almacenes de datos.
- Con la guía de la metodología Hefesto, se pudieron identificar las necesidades de indicadores y datos clave para el negocio que aporten al análisis de la calidad del servicio prestado, pues se logró evidenciar que los usuarios no contaban con herramientas y datos manipulables para procesar información que sirviera en la toma de decisiones del proceso en la calidad del servicio.
- Se obtuvo una arquitectura técnica basada en herramientas Microsoft y los datos de los diferentes sistemas transaccionales que intervienen en la captura de datos de calidad del servicio. A su vez sirve para documentar el proceso buscando que se tenga información histórica del modelo definido y los cambios que se puedan realizar a futuro.
- Con la ayuda de la herramienta Microsoft Integration Services, se pudo desarrollar el proceso de extracción, transformación y carga de datos desde las bases de datos transaccionales hacia una base de datos unificada en el motor SQL Server 2008.
- Se generaron los indicadores relevantes para el negocio como SAIDI, SAIFI, duración de las afectaciones y número de usuarios afectados, entre otros, lo cual aporta al negocio de Gestión Operativa en la oportunidad e integridad de la información para la toma de decisiones.
- Se creó un almacén de datos unificado para la calidad del servicio, el cual se complementa con reportes creados en la herramienta Qlik Sense para que los usuarios puedan disponer de la información en el momento que lo requieran.

9 CONCLUSIONES

- Al finalizar el proyecto, se puede decir que el desarrollo de un Sistema de almacén de datos DW es un proceso complejo, pues abarca recursos como personas (usuarios, directivos, desarrolladores, personal del área de tecnología de información, entre otras), recursos económicos y físicos de la empresa, pues se debe contar con una infraestructura física que permita soportar el volumen de datos que se mueve a diario en una organización.
- Sin duda, la utilización de una metodología de desarrollo de DW aporta conocimiento y un orden a la consecución del logro. Sin embargo, no se debe perder la visión de que una metodología es una guía para llegar a un fin, y que los pasos que se describen no siempre son necesarios ejecutarlos al pie de la letra, pues existen buenas prácticas que se recogen con la experiencia de personas involucradas en el proceso, que aportan a que el objetivo se logre en menor tiempo o en mejores condiciones, como es el caso de haber utilizado apartes de la metodología Hefesto.
- También es importante mencionar que el desarrollo de un sistema de DW es diferente al desarrollo de un sistema operacional. A grandes rasgos, uno apoya al negocio en sus operaciones diarias transaccional y el otro aporta información para la toma de decisiones.
- Es claro que el trabajo no finaliza aquí, se debe seguir alimentando el modelo generado de acuerdo a los hallazgos que se realicen con la información resultante, de tal forma que se pueda ir mejorando el proceso de toma de decisiones en el negocio.

10 REFERENCIAS

Bernabeu, R. D. (2010). *HEFESTO, Metodología para la Construcción de un DataWarehouse.* Córdoba, Argentina.

Bustillos, J. (2014). *slideshare.net.* Recuperado el 13 de 12 de 2016, de https://es.slideshare.net/JorgeCarlos3/comparativa-herramientas-etl

Cedeño Trujillo, A. (2006). MODELO MULTIDIMENSIONAL. *Ingeniería Industrial*(1), 15-18. Recuperado el 24 de 08 de 2016, de http://www.redalyc.org/articulo.oa?id=360433560009

Central Hidroeléctrica de Caldas - CHEC. (2015). *Direccionamiento Estratégico.* Manizales.

Comisión de Regulación de Energia y Gas - CREG. (2008). *Resolución 097.* Bogotá.

Gartner. (8 de agosto de 2016). *Magic Quadrant for Data Integration Tools.* Recuperado el enero de 2017, de https://www.gartner.com/doc/reprints?id=1-2W22JWE&ct=160112&st=sb&mkt_tok=eyJpIjoiT0RNNFpHRmpOVEZqTWprMSIsInQiOiJWOE9QR0lYUk43RUF0NXpXTzVYZjd2WGNpUWo3TWh3XC8zUUMzRjhQbmtSSStGRHU2eXAwaUkzRGt3RDVjNktNaVhFTVJ3UUlrcG5ibmxTSDdxS05lMmhUSXJqT0Q0UlBDTGRoVUwxbG43a

Inmon, W. (2005). *Building the Data Warehouse* (Fourth ed.). Indianapolis: Wiley.

Kimball et al. (1998). *The Data Warehouse Lifecycle Toolkit.* New York: Wiley.

kimball, R., & Caserta, J. (2004). *The DataWarehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* Indianapolis: Wiley.

Kimball, R., & Ross, M. (2002). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling* (Second ed.). New York: Wiley.

Lopez Guayasamín, M. R., Castrillón, O. D., & Cano, E. (2016). ANÁLISIS DE EVENTOS SOBRE TRANSFORMADORES DE DISTRIBUCIÓN EN UNA EMPRESA DEL SECTOR ELÉCTRICO EN COLOMBIA. (U. d. A., Ed.) *Revista Colombiana de Tecnologías de Avanzada, 1*(27), 112-117. Recuperado el 26 de agosto de 2016, de http://www.unipamplona.edu.co/unipamplona/portalIG/home_40/recursos/05_v25_30/revista_27/16052016/18.pdf

Microsoft. (2016). Recuperado el 29 de septiembre de 2016, de https://msdn.microsoft.com/es-es/library/ms141026.aspx

Microsoft. (2016). Recuperado el 28 de noviembre de 2016, de https://technet.microsoft.com/es-es/library/bb522498(v=sql.105).aspx

Moreno Ocampo, R. (2012). *Guía metodológica para el estudio y utilización de la plataforma de inteligencia de negocios Oracle Business Intelligence Standard Edition One.* Pereira. Recuperado el 19 de abril de 2017, de http://repositorio.utp.edu.co/dspace/bitstream/handle/11059/2689/0057565M843g.pdf?sequence=1

Oracle. (2007). Oracle Business Intelligence Standard Edition One. Recuperado el 28 de noviembre de 2016, de http://docs.oracle.com/cd/E10352_01/doc/bi.1013/e10312.pdf

Oracle. (2016). *doc.oracle.com*. Recuperado el 14 de octubre de 2016, de http://docs.oracle.com/cd/B28359_01/owb.111/b31278/concept_overview.htm#WBDOD10100

Oracle. (2017). *Getting Started with Oracle Warehouse Builder.* Recuperado el 3 de enero de 2017, de http://docs.oracle.com/cd/B28359_01/owb.111/b31278/concept_basics.htm#BABEGICH

Oseguera, P. (4 de febrero de 2012). *Pentaho Business Intelligence*. Recuperado el 3 de enero de 2017, de https://sites.google.com/site/pentahobisuite/home

Power Data. (23 de 07 de 2013). Procesos ETL: La Base de la Inteligencia de Negocio. Bogotá D.C., Cundinamarca, Colombia. Recuperado el 22 de 08 de 2016, de http://cdn2.hubspot.net/hub/239039/file-44151143-pdf/docs/PowerData_-

Rivadera, G. R. (2014). La metodología de Kimball para el diseño de almacenes de datos (Data warehouses). *Universidad Católica de Salta*, 56-71.

Shaker H, A.-S., Abdeltawab M., A., & Ali Hamed, E. (2011). A proposed model for data warehouse ETL processes. *Journal of King Saud University - Computer and Information Sciences, 23*, 91-104.

Universidad del Cauca. (s.f.). *DATA WAREHOUSE, ASPECTOS TÉCNICOS, CARACTERÍSTICAS, USOS, BENEFICIOS, COMPONENTES, HERRAMIENTAS OLAP*. Recuperado el 2 de marzo de 2017, de http://fccea.unicauca.edu.co/old/datawarehouse.htm

Yglesias, R. (septiembre de 2008). Recuperado el 2017 de enero de 3, de oracle: http://www.oracle.com/technetwork/es/documentation/317445-esa.pdf

11 ANEXOS

Anexo 1 - Formato de encuesta

Fecha: _____

Nombre del encuestado: _____

Cargo: _____

PERFIL DE LA ENCUESTA

La presente encuesta está dirigida a profesionales y asistentes del área de Gestión Operativa de CHEC. Tiene como objetivo recoger información de las variables que requieren los funcionarios para generar los indicadores de calidad del servicio y apoyar la toma de decisiones a nivel operativo, ejecutivo y direccional.

I. INDICADORES ACTUALES DEL NEGOCIO

1.- ¿Cuáles son los indicadores que miden la gestión del proceso que apoya? Descríbalos brevemente

2.- ¿Cuáles son las variables que utiliza para evaluar sus indicadores?

3.- ¿Qué desea analizar con estos indicadores?

4.- Los indicadores que tiene actualmente apoyan la toma de decisiones de su proceso

☐ Si ☐ No

¿Por qué?

5.- Indique las fuentes requeridas en la generación de los indicadores de su proceso

☐ Aplicaciones de CHEC ☐ Macros ☐ Información física ☐ Aplicaciones externas

☐ Otra (por favor, especifique)

II. INFORMACIÓN SIN INDEXAR

6.- ¿Qué información le gustaría conocer acerca de su proceso?

7.- ¿Cuáles son las variables que deben tenerse en cuenta para poder tomar decisiones basadas en la información anterior?

Muchas gracias por su amabilidad y por el tiempo dedicado a contestar esta encuesta