

UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA
FACULTAD DE CIENCIAS ADMINISTRATIVAS Y SOCIALES



**IMPLEMENTACIÓN DE UN REPOSITORIO UNIFICADO DE DATOS Y
SU CONSULTA MEDIANTE TABLEROS DE CONTROL CON
SOFTWARE DE INTELIGENCIA DE NEGOCIOS**

**TRABAJO TERMINAL PARA OBTENER EL GRADO DE:
MAESTRO EN GESTIÓN EN TECNOLOGÍAS DE LA INFORMACIÓN Y
LA COMUNICACIÓN**

PRESENTA
JOSUÉ GUTIÉRREZ SALAZAR

Ensenada, B.C.

Octubre del 2019

CONSTANCIA DE APROBACIÓN

Director del trabajo terminal: _____


Dr. José Ángel González Fraga

Aprobado por los Integrantes del Sínodo:

1.- _____


Sinodal Dr. Omar Álvarez Xochihua

2.- _____


Sinodal M.C. Evello Martínez Martínez



**Constancia de Presentación de Trabajo Terminal
de la Maestría en Gestión de Tecnologías de Información y Comunicación**

En la Facultad de Ciencias Administrativas y Sociales, a las 16:00 horas del día 31 de Octubre de 2019, se reunieron:

Dr. José Ángel González Fraga director del trabajo terminal, **Dr. Omar Álvarez Xochihua** sinodal y **M.C. Evelio Martínez Martínez** sinodal, para examinar a:

Josué Gutiérrez Salazar

Quien presenta su trabajo terminal con el tema:

“IMPLEMENTACIÓN DE UN REPOSITORIO UNIFICADO DE DATOS Y SU CONSULTA MEDIANTE TABLEROS DE CONTROL CON SOFTWARE DE INTELIGENCIA DE NEGOCIOS”

Después de presentar su trabajo terminal, los miembros del Comité de Trabajo Terminal resolvieron:

Aprobarlo

“POR LA REALIZACIÓN PLENA DEL HOMBRE”



Dr. José Ángel González Fraga
Director



Dr. Omar Álvarez Xochihua
Sinodal



M.C. Evelio Martínez Martínez
Sinodal



Josué Gutiérrez Salazar
Sustentante



UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ciencias Administrativas y Sociales
Maestría en Gestión de las Tecnologías de Información y Comunicación

DRA. EUNICE VARGAS CONTRERAS
COORDINADORA DE POSGRADO E INVESTIGACIÓN
FACULTAD DE CIENCIAS ADMINISTRATIVAS Y SOCIALES
Presente.

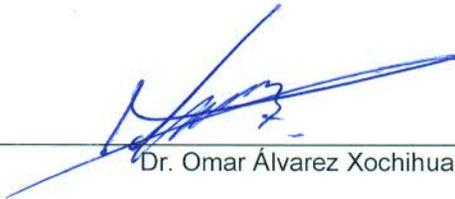
Asunto: Voto aprobatorio sobre trabajo
Terminal de grado de Maestro

Por este medio hago de su conocimiento que una vez revisado el trabajo terminal “IMPLEMENTACIÓN DE UN REPOSITORIO UNIFICADO DE DATOS Y SU CONSULTA MEDIANTE TABLEROS DE CONTROL CON SOFTWARE DE INTELIGENCIA DE NEGOCIOS” que presenta **Josué Gutiérrez Salazar**, considero que satisface los requerimientos teóricos y metodológicos exigidos para obtener el Grado de Maestro en Gestión de las Tecnologías de Información y Comunicación.

Por lo anterior emito MI VOTO **APROBATORIO** para que continúe con los trámites requeridos por el reglamento respectivo, para obtener el Grado correspondiente.

Agradeciendo su consideración, aprovecho la ocasión para enviarle un cordial saludo.

ATENTAMENTE
Ensenada, B.C., 31 de Octubre del 2019


Dr. Omar Álvarez Xochihua



UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ciencias Administrativas y Sociales
Maestría en Gestión de las Tecnologías de Información y Comunicación

DRA. EUNICE VARGAS CONTRERAS
COORDINADORA DE POSGRADO E INVESTIGACIÓN
FACULTAD DE CIENCIAS ADMINISTRATIVAS Y SOCIALES
Presente.

Asunto: Voto aprobatorio sobre trabajo
Terminal de grado de Maestro

Por este medio hago de su conocimiento que una vez revisado el trabajo terminal "IMPLEMENTACIÓN DE UN REPOSITORIO UNIFICADO DE DATOS Y SU CONSULTA MEDIANTE TABLEROS DE CONTROL CON SOFTWARE DE INTELIGENCIA DE NEGOCIOS" que presenta **Josué Gutiérrez Salazar**, considero que satisface los requerimientos teóricos y metodológicos exigidos para obtener el Grado de Maestro en Gestión de las Tecnologías de Información y Comunicación.

Por lo anterior emito MI VOTO **APROBATORIO** para que continúe con los trámites requeridos por el reglamento respectivo, para obtener el Grado correspondiente.

Agradeciendo su consideración, aprovecho la ocasión para enviarle un cordial saludo.

ATENTAMENTE
Ensenada, B.C., 31 de Octubre del 2019



M.C. Evelio Martínez Martínez



UNIVERSIDAD AUTÓNOMA DE BAJA CALIFORNIA

Facultad de Ciencias Administrativas y Sociales
Maestría en Gestión de las Tecnologías de Información y Comunicación

DRA. EUNICE VARGAS CONTRERAS
COORDINADORA DE POSGRADO E INVESTIGACIÓN
FACULTAD DE CIENCIAS ADMINISTRATIVAS Y SOCIALES
Presente.

Asunto: Voto aprobatorio sobre trabajo
Terminal de grado de Maestro

Por este medio hago de su conocimiento que una vez revisado el trabajo terminal "IMPLEMENTACIÓN DE UN REPOSITORIO UNIFICADO DE DATOS Y SU CONSULTA MEDIANTE TABLEROS DE CONTROL CON SOFTWARE DE INTELIGENCIA DE NEGOCIOS" que presenta **Josué Gutiérrez Salazar**, considero que satisface los requerimientos teóricos y metodológicos exigidos para obtener el Grado de Maestro en Gestión de las Tecnologías de Información y Comunicación.

Por lo anterior emito MI VOTO **APROBATORIO** para que continúe con los trámites requeridos por el reglamento respectivo, para obtener el Grado correspondiente.

Agradeciendo su consideración, aprovecho la ocasión para enviarle un cordial saludo.

ATENTAMENTE
Ensenada, B.C., 31 de Octubre del 2019

A handwritten signature in blue ink, appearing to read "José Ángel González Fraga", written over a horizontal line.

Dr. José Ángel González Fraga



CENTRO DE INVESTIGACIÓN CIENTÍFICA Y DE EDUCACIÓN
SUPERIOR DE ENSENADA

Dra. Karina Caro Corrales
Coordinadora de la Maestría en Gestión de Tecnologías de la Información y
la Comunicación,
UABC, Sede Ensenada
Presente. -

A través de la presente se hace constar que el alumno **Josue Gutiérrez Salazar** con número de matrícula **308502** realizó su Proyecto Tecnológico de Trabajo Terminal denominado **“Implementación de un repositorio unificado de datos y su consulta mediante tableros de control con software de inteligencia de negocios”**, bajo la dirección del Dr. José Ángel González Fraga, como parte de los procesos de vinculación entre la Maestría en Gestión de Tecnologías de la Información y la Comunicación y el Centro de Investigación Científica y de Educación Superior de Ensenada, durante el período comprendido de Agosto 2017 a Junio 2019. Consideramos que el proyecto tecnológico tiene el potencial para ser implementado/implantado como parte de las actividades de la empresa/centro/institución en un futuro cercano, dado que el proyecto deja sentadas las bases para su ejecución a largo plazo. El objetivo principal del proyecto será la realización de un análisis que nos ayude a determinar los tiempos que toma actualmente el acceso a esta información, y que tanto se logran mejorar estos tiempos con la implementación de un Data Warehouse y el uso del Business Intelligence de Oracle.

Se emite la presente para los fines legales que al interesado(a) convengan en la ciudad de Ensenada, a los 7 días del mes de junio de 2019.

Sin más por el momento quedo de usted.


M.T.I.C. Diana Pérez Perea
Supervisor del proyecto



Dedicatoria:

Dedico este trabajo a todos los profesores por ayudarme en mi formación académica; también lo dedico a mi familia, por estar siempre apoyándome en las diferentes etapas de este proceso.

Agradecimiento:

Agradezco a los docentes de la Maestría en Gestión de las Tecnologías de Información y de la Comunicación de la Universidad Autónoma de Baja California, en especial a mi director de trabajo terminal, el Dr. José Ángel González Fraga por guiar esta investigación y formar parte de este proceso.

Resumen

La información es hoy en día uno de los activos más importantes de una organización, implementar una simple base de datos relacional no será suficiente, se crean y manipulan millones de registros todos los días y eso es un problema para las personas a cargo de analizar esos registros.

La generación de consultas a la base de datos puede llevar bastante tiempo, evitando o retrasando la toma de decisiones. La implementación de un repositorio unificado de datos es una de las estrategias más utilizadas para la mejora en la recopilación y análisis de información, existen muchas herramientas y software para implementarlo, una de esas herramientas es la base de datos Oracle.

Oracle tiene todas las herramientas necesarias para crear, mantener y consultar un repositorio de datos unificado. El problema es cómo implementarlo, qué tipo de licencia se necesita y la pregunta más importante es si realmente mejora el rendimiento y los tiempos de respuesta al consultar la información.

Palabras clave: Repositorio unificado de datos, inteligencia de negocios, base de datos Oracle.

Tabla de contenidos

1. Introducción	8
1.1. Planteamiento del problema	10
1.2. Hipótesis	10
1.3. Objetivos de investigación	10
1.3.1. Objetivo general	10
1.3.2. Objetivos específicos	10
1.4. Justificación	11
1.5. Viabilidad de la investigación	11
2. Marco teórico	13
2.1. Bases de datos OLTP	13
2.2. Almacén de datos unificado	14
2.2.1. Proceso ETL	15
2.2.2. Bases de datos multidimensionales	17
2.2.3. Lo que no es un almacén de datos unificado	19
2.3. Software de inteligencia de negocios	21
2.4. Tecnologías Oracle para Almacén de datos y BI	22
3. Metodología	24
3.1. Análisis de requerimientos	25
3.2. Análisis de datos sources	26
3.3. Modelo lógico del DW o repositorio unificado	26
3.4. Integración de datos	26
4. Implementación	27
4.1. Implementación de la metodología HEFESTO	27
4.1.1. Análisis de requerimientos:	27
4.1.2. Indicadores y perspectivas	28
4.1.3. Modelo Conceptual	29
4.1.4. Análisis de las estructuras de datos	30
4.1.5. Modelo Conceptual ampliado	31
4.1.6. Modelo lógico del almacén de datos	32

4.1.7. Proceso ETL	35
4.1.8. Actualización de información	36
4.2. Resumen en la implementación de la metodología	37
5. Resultados	38
5.1. Estado actual en las consultas a la información contable presupuestal	38
5.2. Resultados obtenidos con la implementación del repositorio unificado de datos y las consultas a través de tableros de control del BI	40
5.2.1. Ventajas al usar un software de inteligencia de negocios	41
5.2.2. Experiencia de usuario	42
6. Conclusiones y trabajo a futuro	43
6.1. Trabajo a futuro	43
7. Referencias	44

Tabla de figuras

1.	Figura con la representación de una estructura tipo estrella	18
2.	Figura con un ejemplo de diagrama de copo de nieve	19
3.	Figura del modelo conceptual.	29
4.	Figura del modelo conceptual ampliado	31
5.	Figura de las tres dimensiones generadas.	33
6.	Figura con el diagrama del cubo con sus tres dimensiones.	34
7.	Figura del diagrama estrella en la base de datos.	34
8.	Figura con ejemplo para el mapeo de la dimensión de gastos.	36
9.	Figura que muestra el estado actual de las consultas a la base de datos.	38
10.	Figura con el diagrama de la nueva estructura de consultas a la base de datos	40

1. Introducción

El Centro de Investigación Científica y de Educación Superior de Ensenada (CICESE), con el paso de los años ha ido creciendo exponencialmente y conjuntamente la necesidad de administrar y controlar una gran cantidad de información. Tal es el caso de las áreas administrativas de Contabilidad y Presupuestos, las cuales tienen la tarea y responsabilidad de manejar la información financiera del centro.

Estas dos áreas controlan toda la información contable/presupuestal a través de sistemas propios, desarrollados dentro del mismo centro por personal del departamento de informática. Conforme han pasado los años esta información ha ido creciendo en gran escala y volumen, y como consecuencia las bases de datos, lo cual ha impedido la toma de decisiones asertivas y oportunas. Debido al tiempo que tarda el análisis de información mediante reportes directos de las bases de datos relacionales, en algunas ocasiones se ha perdido la oportunidad de tomar decisiones asertivas. Información que constantemente debe ser tratada y manipulada para dar el formato y resultados deseados. Así mismo, es importante considerar los tiempos muertos que los sistemas invierten al realizar consultas complejas.

Actualmente las consultas que realizan los departamentos de contabilidad y presupuestos para la toma de decisiones son sobre la misma base de datos donde se realizan todas las operaciones diarias del centro, utilizando reportes hechos a la medida, lo cual provoca que los requerimientos del servidor sean mayores.

Como una posible solución a esta problemática, en este trabajo se plantea el análisis e implementación de un repositorio unificado de datos (Data Warehouse por su nombre en inglés) y tableros de control con software de inteligencia de negocios (BI por sus siglas en inglés) para realizar las consultas y determinar el grado de mejora y optimización que nos brinda esta propuesta en el manejo de información y así agilizar las operaciones del centro.

Toda la información resguardada se encuentra en base de datos Oracle 11g por lo cual se explorarán las diversas paqueterías y funciones con las que ya cuenta esta tecnología para sacar provecho de ellas y lograr así una mejor implementación del repositorio unificado.

Según *William H. (Bill) Inmon* (Wiley & Sons. 2002) (considerado el padre de los Data Warehouse) define un repositorio unificado de datos o también conocido como almacén de datos unificado de la siguiente manera:

- Orientado al sujeto: datos que brindan información sobre un tema en particular en lugar de las operaciones en curso de una empresa.
- Integrado: Los datos se recopilan en el almacén de datos desde una variedad de fuentes y se fusionan en un todo coherente.
- Variante de tiempo: todos los datos en el almacén de datos se identifican con un período de tiempo particular.
- No volátil: los datos son estables en un almacén de datos. Se agregan más datos, pero los datos nunca se eliminan. Esto permite a la gerencia obtener una imagen coherente del negocio.

Es por eso que un repositorio unificado de datos mejorar la entrega de información, información completa, correcta, consistente, oportuna y accesible. Información que es requerida, en el tiempo que se necesita y en el formato que la necesita. Facilita el proceso de toma de decisiones. Cuando se cuenta con un mayor soporte de información se obtienen decisiones más rápidas.

Por otra parte, un software de inteligencia de negocios es una plataforma que permite al usuario final transformar los datos en información de valor que le permite tomar de manera correcta y oportuna las decisiones de la empresa, además de poder predecir posibles errores en la información. Todo esto se logra a través de tableros control, tablas y graficas que el mismo usuario final puede definir y construir.

1.1. Planteamiento del problema

Las áreas administrativas de contabilidad y presupuestos son las responsables de manejar y controlar el presupuesto de CICESE, lo cual requiere que estén analizando continuamente la información generada por los distintos sistemas administrativos con los que cuentan, y con esto tomar decisiones oportunas y precisas en tiempo real. Con el paso de los años y el crecimiento de la información contenida en la base de datos, esto se ha convertido en un problema debido a las largas esperas en la generación de los reportes.

Por lo antes expuesto, la presente investigación se desarrolla bajo la siguiente pregunta de investigación: ¿Cuáles son las mejoras logradas con la consulta de información contable presupuestal al implementar un repositorio unificado de datos con modelado multidimensional en una base de datos Oracle 11g y tableros de control del software de inteligencia de negocios de Oracle?

1.2. Hipótesis

Siempre y cuando se sigan los lineamientos correctos en su implementación, el uso de un repositorio unificado de datos y su consulta a través de un software de inteligencia de negocios mejorará la toma de decisiones de los departamentos de contabilidad y presupuestos de CICESE.

1.3. Objetivos de investigación

1.3.1. Objetivo general

Evaluar las mejoras logradas al consultar la información contable presupuestal de CICESE a través de un almacén de datos unificado en Oracle 11g y tableros de control de un BI de Oracle.

1.3.2. Objetivos específicos

- Analizar el estado actual de las consultas de la información contable presupuestal.

- Diseñar e implementar la arquitectura física y lógica del repositorio unificado de datos.
- Evaluar los resultados obtenidos con la implementación del repositorio unificado de datos y las consultas a través de tableros de control del BI para determinar las mejoras.

1.4. Justificación

En un repositorio unificado de datos se encuentra integrada toda o una parte de la información de la Institución la cual es utilizada en el proceso de toma de decisiones y análisis de la información.

Reunir los elementos de datos apropiados desde diversas fuentes en un ambiente integral centralizado, ayuda y simplifica el problema de acceso a la información y, por lo tanto, acelera el proceso de análisis y consultas.

Implementar un repositorio unificado en el centro no solo ayudará con la operación actual y la toma de decisiones de las distintas áreas estratégicas, también servirá como base de nuevos y robustos desarrollos que el centro necesite en un futuro.

Es común que muchas instituciones o centros implementen un software para comenzar con la automatización de sus procesos contables/presupuestales, sin tomar en cuenta o saber el crecimiento exponencial de información que tendrán con el pasar de los años, la implementación de un repositorio unificado de datos como el que se expondrá es esta investigación les será de ayuda si ya tienen el mismo problema, o prever con tiempo una solución.

1.5. Viabilidad de la investigación

El tema de investigación principal cuenta con suficiente acceso de información primaria a través de distintas fuentes.

En el centro ya se cuenta con los recursos necesarios para la implementación de un repositorio unificado de datos, tanto en hardware como en el licenciamiento del software

que es requerido, aunque parte de la investigación propuesta definirá si es necesario aumentar los recursos de hardware de los servidores, no serán necesarios para la realización de las pruebas, por lo tanto, no se requiere de una inversión financiera para el proceso de la investigación.

Existen muchas tecnologías para implementar un almacén de datos unificado, para esta investigación se utilizará la base de datos Oracle, ya que es la tecnología con la que cuenta el centro, Oracle cuenta con una gran cantidad de herramientas que facilitan la implementación de la arquitectura de un almacén de datos, pero hay que considerar que muchas de ellas se manejan con licenciamientos independientes, el departamento de informática de CICESE no cuenta con el presupuesto para conseguir nuevos licenciamientos, debido a eso solo se utilizaran las herramientas de Oracle que se incluyan en el licenciamiento Oracle Enterprise Edition.

Se cuenta con acceso controlado a los servidores de producción, con lo cual se podrán hacer las pruebas necesarias para obtener las mediciones iniciales, también se cuenta con un servidor de pruebas para la realización de la implementación. Algunas actividades de prueba tendrán que ser aprobadas y ejecutadas por el administrador de base de datos del departamento el cual conoce el propósito de la investigación y está de acuerdo en apoyar.

La subdirección de programación, presupuestación y estadística de CICESE es un actor clave para el desarrollo de esta investigación, ya que es donde más se generan consultas a la base de datos de gran escala durante todo el año, es por eso que los requerimientos iniciales se obtendrán de esa subdirección, la cual está de acuerdo con la implementación y dispuesta a colaborar.

2. Marco teórico

El conocimiento y uso de las bases de datos relacionales data ya de muchos años, su funcionamiento e implementación han sido de gran ayuda en el manejo de los recursos de las empresas, sin embargo, cuando hablamos de la implementación de una base de datos, por lo general se trata de la instalación de un solo servidor de base de datos en el cual se almacenará toda la información, además de manejar transacciones tanto para el área operativa como para las áreas analíticas.

Si bien esto funciona de una manera adecuada para empresas pequeñas que no manejan grandes cantidades de información y su servidor de base de datos cuenta con el hardware adecuado. Con los avances en tecnología y la creación de nuevos algoritmos que permiten analizar grandes cantidades de información, las organizaciones pueden tomar decisiones más asertivas, prediciendo inclusive lo que pasará en un futuro.

Es por esto que las necesidades de implementación de una base de datos han cambiado, ahora es necesario separar la operación de la base de datos, una para servir las necesidades de las operaciones diarias y otra para llevar a cabo las acciones analíticas.

2.1. Bases de datos OLTP

Para poder comprender el uso de un repositorio unificado de datos primero debemos comprender su origen o el motivo por el cual es mejor implementar un repositorio unificado y no una simple base de datos operativa relacional separada, exclusivamente para el uso de consultas analíticas.

Si bien el hecho de tener dos servidores con una base de datos relacional, uno para las actividades operativas diarias de la empresa y otra con una réplica de la información del primer servidor y usado exclusivamente para las consultas analíticas ayudaría a mejorar el rendimiento de los servidores, no se le estaría dando la importancia adecuada a las necesidades de la empresa en cuestión del análisis de información.

Una base de datos operacional o OLTP (por sus siglas en inglés online transaction processing) (Kimball & Ross, 2003) es una base de datos especialmente diseñada para

llevar el control de la operación diaria de una organización, están optimizadas para procesar las transacciones de manera ágil y rápida.

Este tipo de sistemas normalmente no guardan un histórico de la información, la misma información se mantiene actualizada con los nuevos registros. Los datos se manejan de uno en uno y las consultas se limitan solo a la información más actual. Todo esto difiere de las necesidades de un sistema de consultas analíticas que requiere un histórico en la información y la consulta de grandes cantidades de información.

2.2. Almacén de datos unificado

Un almacén de datos (data warehouse, DW) según Inmon (Wiley & Sons, 2002) es una colección de datos orientada a un determinado ámbito (empresa, organización, etc.), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. Es un historial completo de la organización, más allá de la información transaccional y operacional, almacenado en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos. Por otra parte, Kimball (Kimball & Ross, 2003) la define como “una copia de los datos transaccionales estructurados específicamente para consultas y análisis”.

Un almacén de datos (Corporation, 2012) es una base de datos relacional o multidimensional especialmente diseñada para el análisis de información, es decir que está optimizada para generar grandes consultas de información a la base de datos. Un almacén de datos no está optimizado para generar transacciones normales de una base de datos OLTP, en un almacén de datos casi nunca se elimina o modifica la información, solo se va agregando más.

Una arquitectura de almacén de datos recopila información de diversas fuentes de datos externas como archivos de texto, archivos de hojas de cálculo, partes de una base de datos sin importar la tecnología entre muchas otras.

Las características principales de un almacén de datos son las siguientes:

- Los datos almacenados en el almacén de datos se desnormaliza antes de ser procesados, esto para simplificar la lectura de los datos, ya que la información

será manipulada y consultada por usuarios que no tienen experiencia en el uso de una base de datos, además que mejora los tiempos de respuesta en las consultas.

- Toda la información se queda resguardada en el almacén de datos, solo se borra algún dato si fue insertado por error o la información estaba incorrecta, es decir que se tiene una gran cantidad de datos históricos.
- Las consultas al almacén de datos normalmente regresan grandes cantidades de datos.
- Todo el proceso de carga al almacén de datos es controlado.

2.2.1. Proceso ETL

Una vez recopilada la información debe de pasar por un proceso llamado ETL (por su nombre en inglés extract, transform, load) en el cual los datos son procesados para poder ser almacenados en la base de datos donde se encuentra el almacén de datos.

El proceso ETL es uno de los pasos más importantes en la creación del repositorio unificado, ya que será el responsable de reunión toda la información de las distintas fuentes de datos, estas pueden llegar a ser documentos de texto, archivos XML, documentos EXCEL, partes de una base de una base de datos, toda una base de datos completa, todo depende de los requerimientos de la empresa.

Un ejemplo claro de este proceso es una compañía que posee tiendas de autoservicio alrededor del mundo, cada tienda maneja un sistema que controla las ventas diarias, algunas de esas tiendas pueden llevar registros específicos de productos que solo se manejan en ciertas tiendas, inclusive debido a la región del mundo donde se encuentra una tienda, se podrían manejar bases de datos distintas.

Esta compañía tiene una oficina central donde se encargan de reunir toda la información de las ventas generadas en todo el mundo y procesarla para llevar a cabo la toma de decisiones importantes de la empresa.

Para llevar a cabo esta tarea será necesario generar una conexión con cada una de las bases de datos de las tiendas alrededor del mundo, recabar los archivos con datos importante que no se almacenan en la base de datos.

Debido a que los datos de las tiendas se encuentran normalizados ya que se trata de una base de datos OLTP, y en almacén de datos se requiere desnormalizada para mejorar el desempeño de las consultas y que los datos sean lo suficientemente digeribles o entendibles para los usuarios que realizaran las consultas al almacén de datos, será necesario realizar las tareas dentro de este proceso para llevar a cabo esa conversión de datos antes de ser almacenados en el repositorio unificado. También se debe tomar en cuenta que las bases de datos no son todas iguales o que ni siquiera provienen de una base datos.

También se deben de llevar a cabo tareas de limpieza sobre los datos recopilados, por ejemplo, remover los datos que no son importantes para la toma de decisiones, convertir o remover los datos vacíos, cambiar las cadenas de texto a mayúsculas, limpiar los espacios vacíos, convertir de números a cadenas, cambiar el formato de fecha, etc.

Una vez que se tienen listos los datos ya se pueden almacenar en el almacén de datos para que el personal responsable de llevar a cabo las consultas y toma de decisiones de la empresa puedan hacer su trabajo.

Otro factor importante sobre la implementación de la arquitectura del repositorio unificado es el tiempo en que se actualiza su información desde las fuentes de datos operativas, si la cantidad de información procesada es mucha, será casi imposible mantenerla en tiempo real debido al tiempo de procesamiento de la información y los recursos físicos que serán consumidos por ambos servidores.

En resumen, el proceso de ETL debe llevar a cabo las siguientes tareas (Kimball, R., & Caserta, J. 2009)

- Extraer datos de las fuentes originales.
- Entregar datos de manera más efectiva a las herramientas del usuario final.
- Asegurar la calidad y limpieza de los datos.
- Agregar valor a los datos en los pasos de limpieza y conformidad.

- Proteger y documentar el linaje de datos
- Conformar las etiquetas y las medidas en los datos para lograr la coherencia entre las fuentes originales.
- Entrega de datos en un formato físico que puede ser utilizado por herramientas de consulta, redactores de informes y tableros de control.

2.2.2. Bases de datos multidimensionales

Los datos son almacenados y gestionados por uno o más servidores OLAP (On-Line Analytical Processing) por sus siglas en inglés. OLAP (Tamayo & Moreno, 2006) es una tecnología o grupo de algoritmos que facilita las consultas sobre un repositorio unificado de datos, mejorando el tiempo de consulta.

Existen varios tipos de OLAP según el tipo de almacenamiento que se utilice, las principales son:

ROLAP: Toda la información es almacenada en base de datos en tablas relacionales y el motor OLAP es el encargado de generar las consultas.

Este tipo de implementación tiende a ser el más lento en cuestiones de consultas ya que se base en sentencias SQL.

MOLAP: Toda la información es almacenada multidimensionalmente en objetos de la base de datos.

Está optimizada para mejorar los tiempos de respuesta ya que le resumen de la información es usualmente calculado por adelantado.

HOLAP: Es una combinación de ROLAP y MOLAP, algunos datos son almacenados en tablas relacionales y otras en objetos multidimensionales.

Cada una de estas implantaciones tiene sus ventajas y desventajas, las cuales varían en opciones dependiendo los distintos proveedores de los servicios, en general ROLAP consume más espacio en disco debido a la forma en que se almacena la información, a diferencia de MOLAP que se encuentra almacenado en archivos más ligeros.

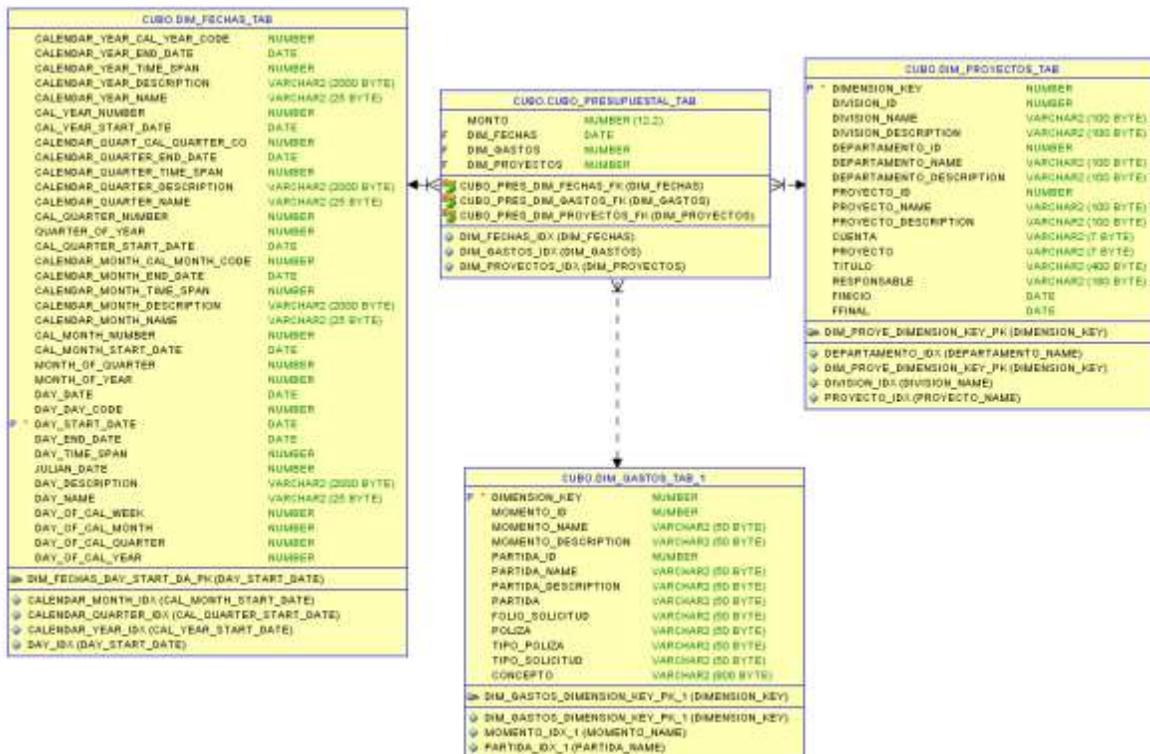
Otra de las desventajas es que para la implementación de algunos de los algoritmos de MOLAP requieren licenciamientos extras dependiendo el cliente de base de datos que se utilice.

Pero cualquiera de estas tecnologías puede ser consultadas a través de muchos de los sistemas disponibles en el mercado que permiten de manera visual y sencilla generar reportes y análisis de la información, incluyendo la minería de datos.

Otra de las características importantes sobre un OLAP es su relación o estructura dentro de la base de datos (Kimball & Ross, 2003), los dos tipos más comunes son:

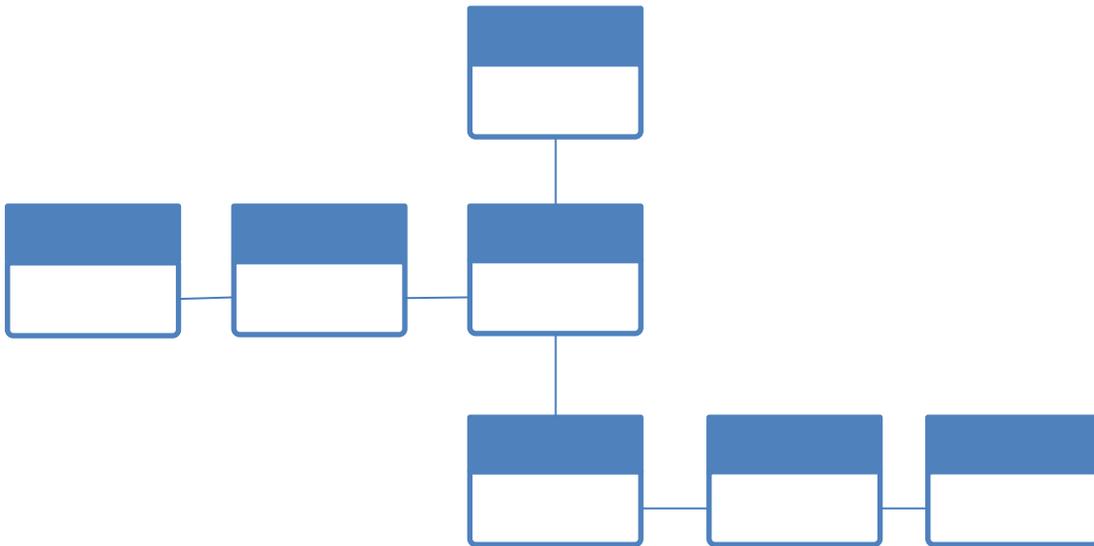
- Esquema tipo Estrella: Es una estructura de dimensión en una base de datos relacional, consiste en una tabla central llamada tabla de hechos que contiene la relación o llaves primarias de todas las dimensiones del cubo. Un ejemplo de este tipo de estructura se puede observar en la figura 1.

1. Figura con la representación de una estructura tipo estrella



- Esquema de copo de nieve: Es similar a la estructura de estrella, pero en este caso cada dimensión puede contener otro nivel de jerarquía, extendiendo otro nivel a la dimensión. Este tipo de esquemas es un poco menos recomendable ya que puede resultar muy confuso al usuario final si existen múltiples niveles o ramificaciones en cada dimensión del cubo. En la figura 2 se muestra un ejemplo de este tipo de estructura.

2. Figura con un ejemplo de diagrama de copo de nieve



2.2.3. Lo que no es un almacén de datos unificado

Puede llegar a existir mucha confusión en lo que es un almacén de datos unificado debido a los múltiples pasos para su implementación y las distintas maneras en que se puede implementar cada una de ellos.

Es importante que todos los involucrados en el proceso de construcción del almacén de datos estén sincronizados en lo que es y no es un almacén de datos y los pasos que se deben de seguir para su implementación.

A continuación, se listan las cinco principales confusiones acerca de lo que es un almacén de datos (Kimball, R., & Caserta, J. 2009):

1. Un producto: Un almacén de datos no se puede comprar, si bien existen distintas paqueterías o herramientas que te ayudan con su implementación, estas herramientas no pueden llevar el proceso completo de la construcción del almacén de datos.
2. Un lenguaje de programación: No es posible aprender a programar un almacén de datos como se haría con Java, SQL o algún lenguaje de este estilo, lo más probable es que para cada etapa de la construcción del almacén se requiera la programación o uso de herramientas distintas.
3. Un proyecto: Un almacén de datos bien construido, consiste en muchos proyectos distintos, uno para cada etapa de la implementación, un almacén de datos que está definido como un solo proyecto, tiene muchas probabilidades de fracasar en una de las etapas.
4. Un modelo de datos: Si bien para la construcción de un almacén de datos se requiere un modelo de datos, este no por su cuenta no hace a un almacén de datos, aún se requiere el proceso ETL para la construcción completa del almacén de datos.
5. Una copia de la base de datos transaccional: Un error común es creer que el almacén de datos es solo una copia de la base de datos transaccional utilizada para el uso específico de reportes o análisis. No toda la información de la base

de datos transaccional es utilizada en el almacén de datos además que requiere primero una limpieza y transformación.

2.3. Software de inteligencia de negocios

Como ya se mencionó anteriormente, un sistema operacional ayuda en el manejo diario de la empresa o intuición, es donde se llevan a cabo los registros diarios de los datos, (Kimball & Ross, 2003) donde se crean los clientes, las operaciones que se llevaron a cabo, el flujo de efectivo. En resumen, son los sistemas que alimentan la base de datos con datos.

También se han mencionado los sistemas de consulta analítica o también conocidos como software de inteligencia de negocios. Estos sistemas están especialmente desarrollados para la consulta de grandes cantidades de datos, datos que fueron creados desde los sistemas operacionales, también se menciona que son los sistemas que convierten los datos en información, información que los responsables en la toma de decisiones pueden utilizar para llevar a cabo estadísticas de lo que está pasando actualmente es la empresa o institución, inclusive algunos de estos sistemas tienen la capacidad de implementar algoritmos avanzados de minería de datos para predecir a futuro.

Hoy en día existe una gran cantidad de software de inteligencia de negocios, algunos de ellos de libre distribución y algunos otros de licenciamiento.Cuál utilizar dependerá mucho de las necesidades de cada empresa o institución, de la cantidad de información que se maneje y de las herramientas que provea dicho sistema.

2.4. Tecnologías Oracle para Almacén de datos y BI

Oracle Database 11g (Lumpkin, 2013) es una plataforma integral de base de datos ideal para la implementación de un almacén de datos que combina escalabilidad y desempeño. Oracle Database ofrece funcionalidad para almacén de datos y data marts, con escalabilidad probada en cientos de terabytes y desempeño líder en el mercado. También ofrece una plataforma exclusivamente integrada para análisis. Al incorporar OLAP, Data Mining y capacidades estadísticas directamente en la base de datos, Oracle brinda toda la funcionalidad de motores analíticos autónomos con la escalabilidad, seguridad y confiabilidad empresarial de una base de datos Oracle.

Si bien Oracle ofrece software con muchas opciones y configuraciones para lograr replicar la información en tiempo real (Corporation, 2008), como lo es Golden Gate, este software ocupa un licenciamiento independiente, licencia que puede llegar ser muy costosa. Para un centro del gobierno federal es complicado depender de licenciamiento de software, ya que podrían no tener el suficiente ingreso al año siguiente para volver a contratar este licenciamiento, por lo cual es necesario buscar otras opciones que sean más económicas o que ya se encuentren incluidas en el licenciamiento de la base de datos Oracle Enterprise 11g.

Dentro de estas opciones las que nos pueden servir son: replicación de información a través de vistas materializadas. Esta opción utiliza la creación de vistas materializadas para guardar la información de una o varias tablas en otra base de datos y no en el disco duro como lo hace normalmente, la actualización de la información pueden ser a través de “commits”, cada cierto tiempo, o de manera manual. La información se puede mandar de manera incremental, completa o una mezcla de las dos.

La bibliografía de Oracle recomienda este tipo de implementación cuando la conexión puede ser un problema, es decir cuando la base de datos cliente (el almacén de datos en este caso) debe de funcionar sin red o cuando no se requiere actualizar la información en tiempo real. Cuando no se tiene problemas de conexión o se requiere estar actualizando la base de datos en tiempo real (o casi en tiempo real), la bibliografía recomienda el uso de Oracle Streams. Oracle Streams ofrece la replicación de

información en tiempo casi real como dice la bibliografía y distintas configuraciones según sea necesario, una vez configurado en las dos bases de datos estas se pueden comunicar y enviar sus cambios de manera automática.

Si bien esta opción mantiene la información actualizada casi en tiempo real, hay que considerar que se requerirá de una cantidad considerable de procesamiento de ambos servidores, ya que se estará ejecutando el respaldo cada vez que exista un cambio, es decir que el hardware del servidor tendrá que estar a la altura para este tipo de replicación.

Otra alternativa que se maneja dentro del licenciamiento Enterprise Edition de Oracle es el software Oracle Warehouse Builder, esta herramienta cuenta con todo lo necesario para el modelado y la integración de datos del almacén de datos.

Esta herramienta permite de manera gráfica crear la integración de distintas fuentes de datos, no solo de bases de datos Oracle, tiene además componentes que permite crear los procesos ETL de manera más rápida.

Otra de las características importantes de esta herramienta es la posibilidad de calendarizar el mapeo automático de información de las distintas fuentes de datos al repositorio unificado de datos.

3. Metodología

Hoy en día existen varias propuestas metodológicas para la creación de un almacén de datos unificado con el propósito de facilitar su implementación. Cada una de estas metodologías cuenta con características específicas que las hacen distintas, por lo cual la elección depende de la implementación que se requiera.

Para este estudio se utilizará la metodología HEFESTO (Castelán & Ocharán, 2012) la cual es una metodología creada por el Ing. Bernabeu Ricardo Darío, se fundamenta en una amplia investigación, comparación de metodologías existentes y experiencias propias en procesos de confección de almacenes de datos. Puede ser utilizada en cualquier ciclo de vida que no requiera fases extensas de requerimientos y análisis. Una de sus ventajas principales es que documenta los pasos a seguir para cada fase a diferencia de otras metodologías disponibles para la construcción de un almacén de datos.

La metodología HEFESTO inicia con la recolección y análisis de los requerimientos del negocio, después se hace un análisis de la información de la base de datos operativa que nos permitirá construir nuestro modelo lógico, por último, se deben definir los pasos que se utilizaran para la carga de información hacia nuestro almacén de datos.

HEFESTO está compuesto por los siguientes pasos:

1) ANÁLISIS DE REQUERIMIENTOS

- 1.1) Preguntas del Negocio.
- 1.2) Indicadores y Perspectivas.
- 1.3) Modelo Conceptual.

2) ANÁLISIS DE DATA SOURCES

- 2.1) Hechos e Indicadores.
- 2.2) Mapeo.
- 2.3) Granularidad.
- 2.4) Modelo Conceptual Ampliado.

3) MODELO LÓGICO DEL DW

3.1) Tipología.

3.2) Tablas de Dimensiones.

3.3) Tablas de Hechos.

3.4) Uniones.

4) INTEGRACIÓN DE DATOS

4.1) Carga Inicial.

4.2) Actualización.

3.1. Análisis de requerimientos

Esta etapa consiste en identificar las necesidades de los usuarios que llevarán a cabo el análisis de la información, esto se puede realizar a través de preguntas específicas, las respuestas a estas preguntas nos deberán permitir identificar de manera sencilla el identificador y perspectivas.

Los identificadores son en general el dato número que finalmente queremos obtener, es decir la cantidad de ventas por año en cierto país, o la cantidad de productos más vendidos.

Las perspectivas por otra parte son los datos que nos van a permitir identificar los indicadores, por ejemplo, la información del país o la descripción a detalle del producto más vendido.

Obteniendo esta información se construye el modelo conceptual, el cual es un diagrama en el cual se detallan los identificadores y las perspectivas y las conexiones que existen entre ellos.

3.2. Análisis de datos sources

En este paso se analizan los datos que se generan en los sistemas OLTP, es decir la estructura actual de la base de datos, para poder determinar en detalle los datos que se deberán utilizar para llegar a los identificadores y perspectivas definidos en el paso anterior. También se deberá analizar las conexiones que se requieren para obtener estos datos. Con esta información se actualiza el modelo conceptual incluyendo los datos para obtener ya sean los identificadores o perspectivas.

3.3. Modelo lógico del DW o repositorio unificado

Es esta etapa ya se tiene la información necesaria para la construcción de nuestro modelo lógico, lo primero que se tiene que definir es el tipo de esquema que se va a utilizar dependiendo de los niveles que se hayan obtenido en los pasos anteriores. Estos pueden ser esquema de estrella o esquema de copo de nieve.

Una vez definido el tipo de esquema hay que definir las tablas que contendrán las dimensiones, para cada perspectiva definida en el modelo conceptual corresponderá una tabla, y los datos de cada tabla corresponderán a los elegidos en el modelo conceptual ampliado.

En esta etapa también se define la tabla de hechos, la cual contendrá las llaves primarias de cada uno de las dimensiones y también contendrá los identificadores que se definieron.

3.4. Integración de datos

Una vez que se tiene la estructura de base de datos que contendrá al almacén de datos, hay que definir el proceso ETL que será necesario para limpiar y desnormalizar los datos de la base de datos operativa.

Una vez definido el proceso ETL hay que llenar el almacén de datos con la información y también definir el proceso que lo estará alimentando, es decir cómo y cada cuando se estará actualizando.

4. Implementación

4.1. Implementación de la metodología HEFESTO

A continuación, se explicará cómo se llevó a cabo el seguimiento paso por paso de la metodología HEFESTO para la construcción del repositorio unificado en CICESE.

4.1.1. Análisis de requerimientos:

Como ya se explicó anteriormente, los actores principales para nuestro análisis, son los departamentos de contabilidad y presupuestos, los cuales como parte de su operación diaria deben ejecutar reportes sobre la situación actual del centro, así como reportes históricos.

El levantamiento de requerimientos consistió en entrevistas presenciales con los subdirectores de cada departamento, cuestionando acerca de los reportes con los que cuentan actualmente. Las preguntas fueron:

1. ¿Cuántos reportes tienen actualmente?
2. ¿Cuáles son los reportes que más utilizan?
3. ¿Cuáles son los reportes que más toman tiempo en ejecutarse?
4. ¿En promedio cuánto tiempo les toma ejecutarlos?
5. ¿Cuáles son los reportes que tienen en común los dos departamentos?
6. ¿La información mostrada es la adecuada?
7. ¿Los formatos resultantes son suficientes?
8. ¿Requieren de más reportes?
9. ¿Requieren que la información esté en tiempo real?
10. ¿Normalmente en que horarios ejecutan los reportes?
11. ¿Los ejecutan varias veces al día?

Para este caso práctico elegiremos los reportes más importantes que tienen en común ambos departamentos, que tienen una gran cantidad de información y que además su tiempo de ejecución es muy elevado. Si intentamos medir reportes que se ejecuten en

segundos nos será más difícil hacer las comparaciones, ya que la diferencia será mínima debido a los tiempos cortos de ejecución.

Por lo tanto y en base a las respuestas obtenidas en las preguntas 1, 2 y 4, la elección de los reportes es la siguiente:

- **Disponible presupuestal multianual por partida, momento contable y tipo de cuenta.** Debido a los largos tiempos de ejecución, se ven obligados a ejecutar por partes este reporte y posteriormente unirlos usando la paquetería de Microsoft, no han logrado ejecutar este reporte para todas las cuentas de gastos que existen en el centro.
- **Lista de asignaciones multianual por tipo de cuenta.** Debido a los largos tiempos de ejecución, se ven obligados a ejecutar por partes este reporte y posteriormente unirlos usando la paquetería de Microsoft, tampoco han logrado ejecutar este reporte para todas las cuentas de gastos que existen en el centro, sin embargo, los tiempos de ejecución son menores al reporte anterior.
- **Disponible presupuestal detallado multianual por cuenta.** De los tres reportes este es el que toma menos tiempo debido a que es menos información, pero aún así los tiempos de respuesta se eleva para cuentas que llevan activas varios años.

Otro de los datos importantes resultantes del cuestionario es el requerimiento de consultas en tiempo real y ejecuciones en las horas más operativas del centro, por lo cual se deberá de implementar un mecanismo que permita la actualización del almacén de datos de manera constante, pero para la realización de esta investigación, la actualización solo se realizará una vez al día en horarios no operativos del centro.

4.1.2. Indicadores y perspectivas

A continuación, de la selección de reportes se analizará la información que arrojan y con ello se detallarán cuáles son sus Indicadores y Perspectivas.

- Monto disponible por tipo de cuenta en un rango de fechas dividido por momentos, tipo de trámite y el reservado.

- Monto del disponible en un rango de fechas dividido por cuenta, tipo de solicitud, momento y el reservado.

Por lo tanto, los indicadores son:

- Monto del disponible

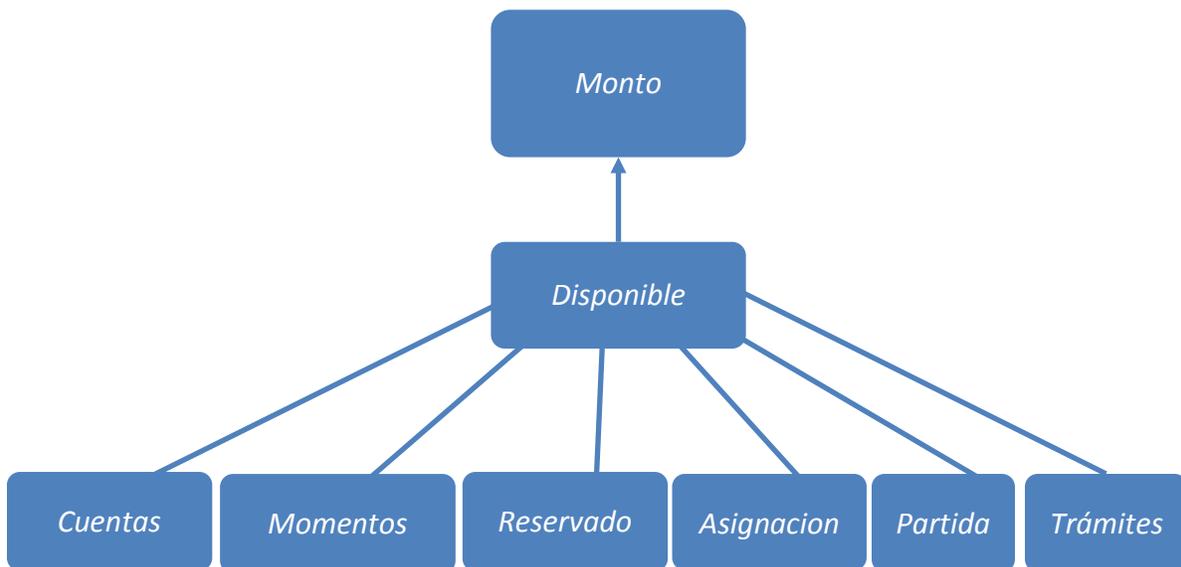
Y las perspectivas de análisis son:

- Cuentas de gastos.
- Momentos
- Reservado
- Asignaciones
- Partidas
- Trámites

4.1.3. Modelo Conceptual

En la figura 3 se muestra el diagrama del modelo conceptual a partir de los indicadores y perspectivas elegidos.

3. Figura del modelo conceptual.



4.1.4. Análisis de las estructuras de datos

En este paso analizaremos las estructuras de datos de la base de datos para determinar cómo serán calculados los Indicadores y poder establecer el mapeo entre el modelo conceptual creado y los datos de la base de datos, se definirán qué campos se incluirán en cada perspectiva y se ampliará el modelo conceptual con la información.

Granularidad

A continuación, se enlistan los nombres de los campos seleccionados para cada perspectiva

Cuentas

- Cuenta: número de cuenta asignado a los proyectos del centro.
- Proyecto: clave alfanumérica asignada a los proyectos del centro.
- Título: nombre de proyecto.
- Fecha inicio: fecha de inicio del proyecto.
- Fecha final: fecha final del proyecto.
- Responsable: persona responsable del proyecto.
- División: división al que pertenece el proyecto.
- Departamento: departamento al que pertenece el proyecto.

Momentos

- Momento: tipo de momento contable.
- Fecha: fecha del movimiento.
- Póliza: número de póliza asignado al movimiento.
- Tipo póliza: tipo de póliza del movimiento.
- Concepto: descripción del movimiento.

Trámites

- Trámite: nombre del trámite.
- Folio: folio generado para el trámite.

Partidas

- Partida: número de partida.

Asignaciones

- Descripción: descripción de la asignación.
- Fecha: fecha de la asignación.

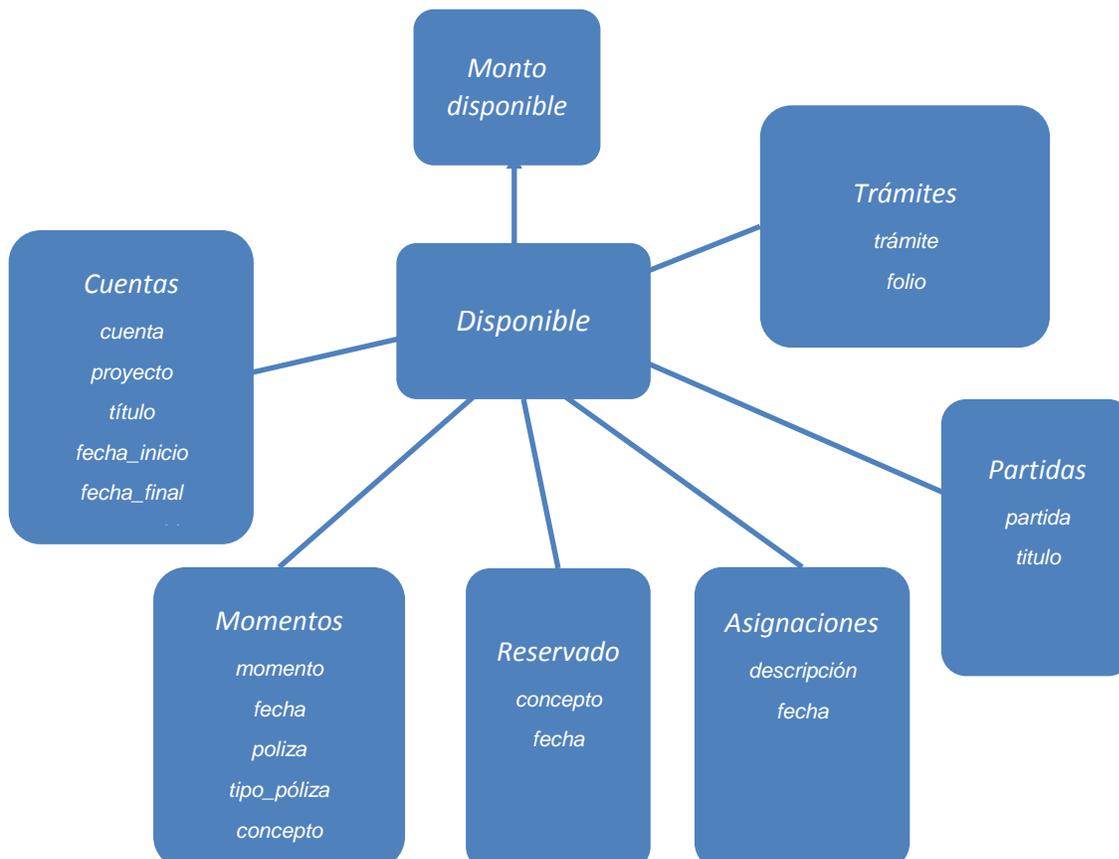
Reservado

- Concepto: descripción del reservado generado.
- Fecha: fecha del reservado.

4.1.5. Modelo Conceptual ampliado

En base a la gradualidad detectada en la figura 4 se muestra el diagrama del modelo conceptual ampliado.

4. Figura del modelo conceptual ampliado



4.1.6. Modelo lógico del almacén de datos

Para la implementación del modelo lógico del almacén de datos se ha seleccionado el Esquema en Estrella ROLAP ya que cumple con todos los requerimientos planteados y además es simple de implementar y comprender ya que todo es almacenado en tablas relacionales.

Si bien una implementación totalmente multidimensional MOLAP, mejora el rendimiento en las consultas ya que está compuesto por objetos de la base de datos y no por tablas relacionales, este requiere de configuraciones extras en la base de datos que pueden incluir licenciamientos extras.

Pero una de las ventajas de la implementación ROLAP es que nos permite migrar en un futuro a una implementación de tipo MOLAP.

Para construir el modelo lógico del almacén de datos se utilizó la herramienta de Oracle Data Warehouse Builder (OWBD por sus siglas en inglés), la cual nos permite de manera gráfica ir generando paso a paso cada una de las dimensiones que conforman el cubo del almacén de datos.

Una vez que están generadas todas las dimensiones y el cubo que las contendrá, la herramienta genera automáticamente toda los metadatos y estructuras relacionales que contendrán la información del almacén de datos.

Ya que los registros que componen al modelo conceptual de las asignaciones y los reservados comparten casi la misma información que los momentos contables, todo se unió en una solo estructura tipificando el tipo de movimiento. Por lo tanto, el cubo se compone de tres dimensiones:

- Cuentas o proyectos: Incluye la información general del proyecto.
- Gastos: Lista todos los movimientos que se generaron en las cuentas o proyectos. En estos se incluyen los momentos presupuestales, las asignaciones y los reservados.
- Fechas: Esta dimensión es construida automáticamente por la herramienta, incluye una serie de información fija relacionada con las fechas, que ayuda a

generar las búsquedas por fecha a un nivel muy general o muy a detalle según sea requerido.

En la figura 5 se muestra el diagrama con las tres dimensiones y sus campos, en la figura 6 se muestra el diagrama del cubo con la unión de las tres dimensiones.

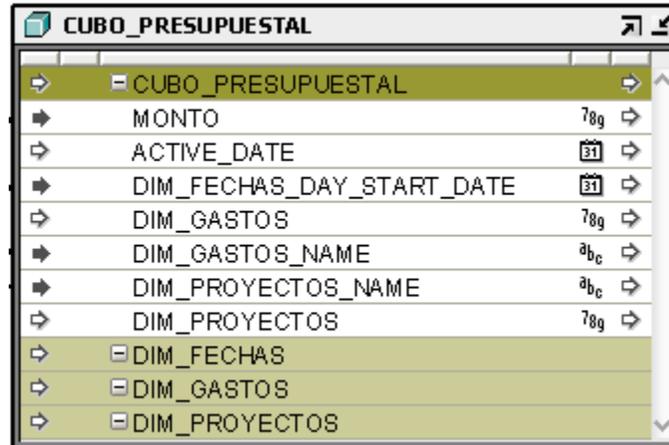
5. Figura de las tres dimensiones generadas.

DIM_FECHAS		
CALENDAR_YEAR		
CAL_YEAR_CODE	789	
END_DATE	31	
TIME_SPAN	789	
DESCRIPTION	abc	
NAME	abc	
CAL_YEAR_NUMBER	789	
CAL_YEAR_START_DATE	31	
CALENDAR_QUARTER		
CAL_QUARTER_CODE	789	
END_DATE	31	
TIME_SPAN	789	
DESCRIPTION	abc	
NAME	abc	
CAL_QUARTER_NUMBER	789	
QUARTER_OF_YEAR	789	
CAL_QUARTER_START_DATE	31	
CALENDAR_Y_CAL_YEAR_START_DATE	31	
CALENDAR_MONTH		
CAL_MONTH_CODE	789	
END_DATE	31	
TIME_SPAN	789	
DESCRIPTION	abc	
NAME	abc	
CAL_MONTH_NUMBER	789	
CAL_MONTH_START_DATE	31	
MONTH_OF_QUARTER	789	
MONTH_OF_YEAR	789	
CALENDAR_Y_CAL_YEAR_START_DATE	31	
DAY		
DAY_DATE	31	
DAY_CODE	789	
DAY_START_DATE	31	
END_DATE	31	
TIME_SPAN	789	
JULIAN_DATE	789	
DESCRIPTION	abc	
NAME	abc	
DAY_OF_CAL_WEEK	789	
DAY_OF_CAL_MONTH	789	
DAY_OF_CAL_QUARTER	789	
DAY_OF_CAL_YEAR	789	
CALENDAR_CAL_MONTH_START_DATE	31	

DIM_GASTOS		
MOMENTO		
ID	789	
NAME	abc	
DESCRIPTION	abc	
PARTIDA		
ID	789	
NAME	abc	
DESCRIPTION	abc	
PARTIDA	abc	
FOLIO_SOLICITUD	abc	
POLIZA	abc	
TIPO_POLIZA	abc	
TIPO_SOLICITUD	abc	
CONCEPTO	abc	
MOMENTO_ID	789	
MOMENTO_NAME	abc	

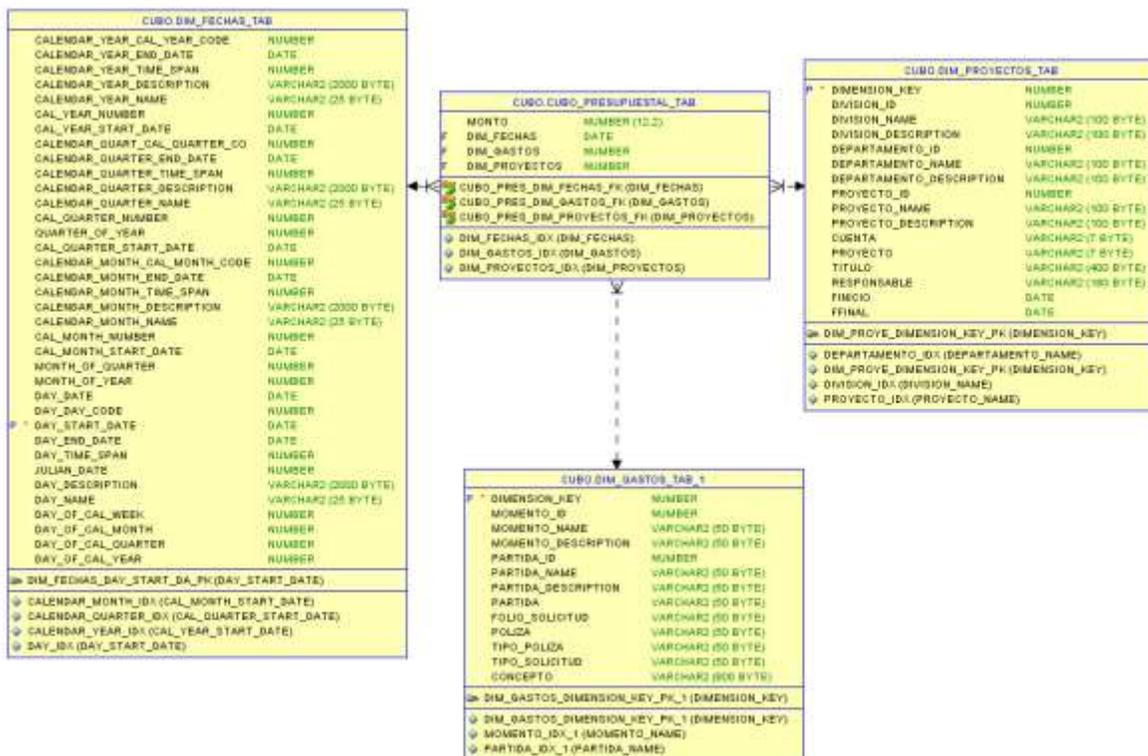
DIM_PROYECTOS		
DIVISION		
ID	789	
NAME	abc	
DESCRIPTION	abc	
DEPARTAMENTO		
ID	789	
NAME	abc	
DESCRIPTION	abc	
DIVISION_NAME	abc	
DIVISION_ID	789	
PROYECTO		
ID	789	
NAME	abc	
DESCRIPTION	abc	
CUENTA	abc	
PROYECTO	abc	
TITULO	abc	
RESPONSABLE	abc	
FINICIO	31	
FFINAL	31	
DEPARTAMENTO_ID	789	
DEPARTAMENTO_NAME	abc	

6. Figura con el diagrama del cubo con sus tres dimensiones.



Una vez definidos los campos y dimensiones del cubo se generaron las estructuras de base de datos a través de la misma herramienta, la estructura resultante se muestra en la figura 7.

7. Figura del diagrama estrella en la base de datos.



4.1.7. Proceso ETL

Con la estructura de la base datos del repositorio unificado generada el siguiente paso fue definir el proceso ETL, proceso en el cual se obtiene la información de la o las bases de datos operativas origen y procesada para ingresarla en el esquema destino del repositorio unificado.

Las bases de datos operativas están construidas de manera normalizada ya que es la mejor manera de obtener un mejor rendimiento en el proceso diario de inserción actualización o borrado de la información que se genera de los distintos sistemas con los que se cuenta, por otro lado en un repositorio unificado en el cual solo se van a generar consultas, se recomienda una estructura lo más desnormalizada posible para mejorar el proceso de consultas, es por eso que se requiere un proceso ETL que genere todas las operaciones necesarias para preparar la información antes de ingresarla al repositorio unificado.

Si bien uno de los requisitos de los responsables de las áreas de contabilidad y presupuestos es actualizar la información en tiempo real o lo más cercanos posible, el hardware de los servidores actuales deberá ser actualizado para poder llevar este requerimiento, para cuestiones de llevar a cabo esta investigación la actualización de información se hará una solo vez en horario nocturno para no afectar la operación del centro, este proceso se llevará a cabo utilizando el software de Oracle Datawarehouse Builder, incluido en la licencia de la base de datos Oracle Enterprise Edition, licencia con la que ya cuenta el centro.

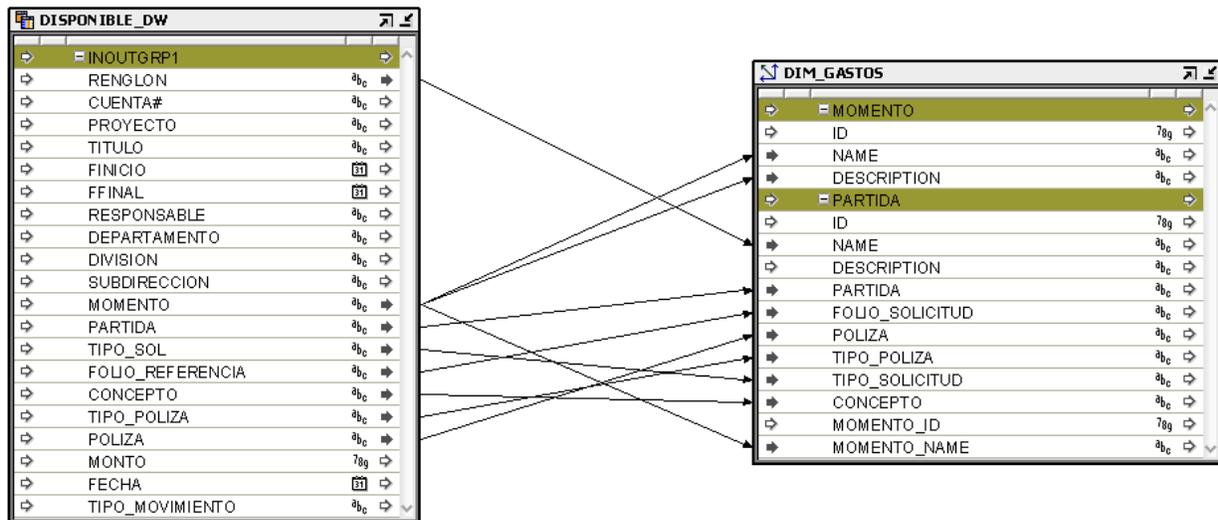
La herramienta de ODWB también permite utilizar de manera gráfica una serie de funciones ya establecidas para lograr estos resultados. Para este estudio todo el proceso ETL se generó en la construcción de una vista en la base de datos origen que une toda la información que va a ser requerida, en esta misma vista toda la información se desnormaliza para que pueda ser insertada directamente en el repositorio unificado, todo esto fue posible porque toda la información que se requiere se encuentra en una sola base de datos Oracle 11g.

Si bien todo este proceso se puede generar de manera manual, la herramienta ODWB cuenta con la opción para generar el mapeo de manera automática, generando el código

necesario en la base de datos origen que permitirá actualizar la información automatizada de la base de datos origen al almacén de datos.

Para este caso fue necesario la construcción de cuatro procedimientos para generar el mapeo de cada dimensión y otro para la unión de cada uno de ellos, en la figura 8 se muestra un ejemplo de cómo se realiza el mapeo de manera gráfica en la herramienta.

8. Figura con ejemplo para el mapeo de la dimensión de gastos.



4.1.8. Actualización de información

El último paso en la creación del repositorio unificado consistió en la generación de información para nuestro almacén de datos, para ello se utilizó la opción de la herramienta ODWB que ejecuta el código creado a través de los diagramas construidos en el paso anterior, esta actividad se puede generar de manera manual o definiendo un proceso automatizado por fecha y hora que lo genere de manera regular.

Para este paso hay que tomar en cuenta que les costará a las bases de datos un gran consumo de procesamiento y memoria dependiendo la cantidad de información contenida y las distintas conexiones que deba generar, para este caso y para no causar problemas en la base de datos operativa, cada actualización de información se realiza una vez por día en horarios no operativos del centro.

4.2. Resumen en la implementación de la metodología

La metodología HEFESTO nos va llevando de la mano paso a paso para poder construir nuestro almacén de datos sin la necesidad de tantos requerimientos, esto está bien para almacenes de datos que no son tan grandes y que no cuentan con tantos niveles de abstracción, de otra manera lo más conveniente sería seguir otra metodología que sí lo permita.

Si bien el seguimiento de la metodología ayudó en la definición de los datos necesarios para la creación del repositorio unificado, la herramienta ODWB no solo nos facilita la construcción de todo el código necesario para la construcción de las estructuras y funciones para la actualización, sino que también nos ayuda a definir las estructuras siguiendo las mejores prácticas, las cuales han sido perfeccionadas con el paso del tiempo. Si bien se puede generar todo este proceso con scripts propios, el mantenimiento de estos podría causar problemas con las actualizaciones continuas de la información.

5. Resultados

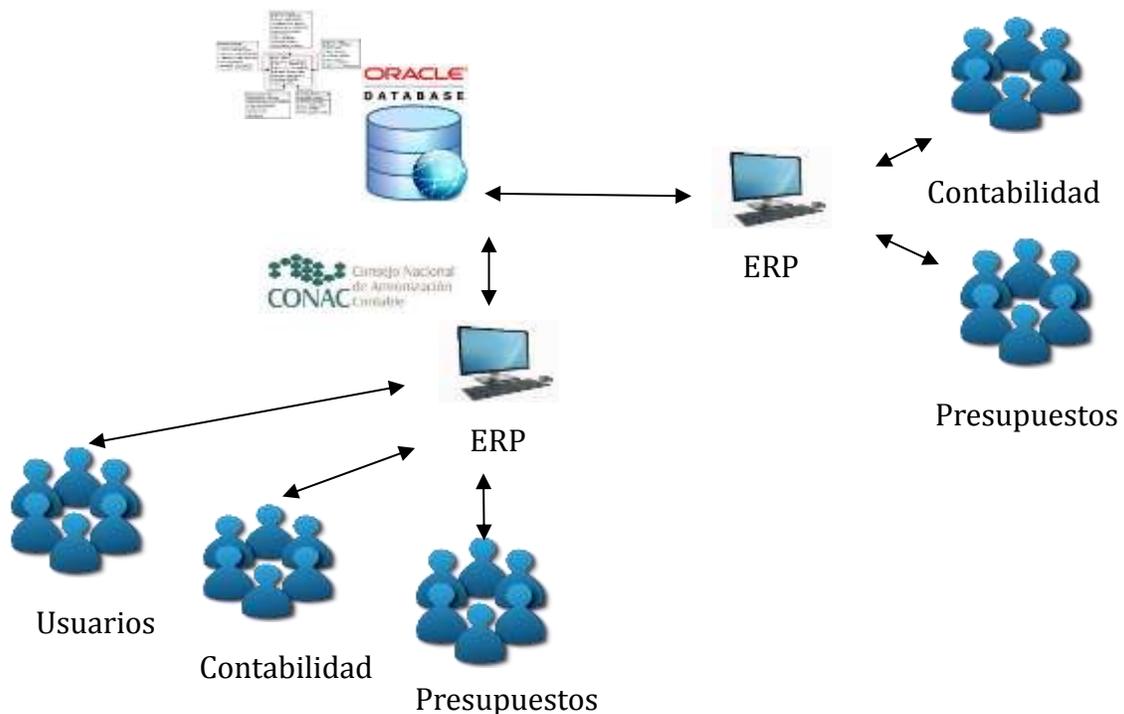
5.1. Estado actual en las consultas a la información contable presupuestal

Actualmente CICESE cuenta con sistemas de gestión administrativa (ERP) desarrollados por el departamento de Informática, en los cuales se lleva el control de todos los servicios administrativos del centro.

Se cuenta aproximadamente con más de 50 sistemas, en los cuales se generan alrededor de unos 2000 trámites por mes, aproximadamente unos 200 usuarios usando los sistemas todos los días.

Actualmente la parte operacional y analítica se llevan a cabo en la misma base de datos relacional, lo cual provoca que el servidor se vea afectado al no contar con los recursos suficientes para mantener activas tantas consultas. En la figura 9 se muestra un diagrama que representa la situación actual de los sistemas y las consultas a la base de datos.

9. Figura que muestra el estado actual de las consultas a la base de datos.



En el caso más específico de las consultas analíticas del departamento de presupuestos, cuentan con más de 30 reportes variados, programados según se requiera el formato de la información. Cada que se requiere algún dato distinto o algún formato con un cambio por más sencillo que sea, es necesario la programación de un programa nuevo o la modificación de alguno existente.

La mayoría de estos reportes se generan en formato Microsoft Excel, ya que para ellos es más sencillo generar uno o más reportes con toda la información necesaria y después trabajar sobre un documento Excel, para generar otro tipo de información o agrupación de información que los reportes no arrojan.

Si bien la información que consultan se encuentra en la misma base de datos Oracle, ésta se encuentra dividida en diferentes esquemas de base de datos, por lo cual muchos de los reportes salen de vistas de vistas que unen la información de los diferentes esquemas. Lo cual provoca que la generación de varios reportes tome una cantidad considerable de tiempo, por lo cual tienen que generar los reportes por secciones, es decir filtrando la información en pequeñas partes y trabajando esa pequeña sección.

Es importante mencionar que ninguno de estos reportes tiene la capacidad de generar gráficas basadas en la información, para lograr este tienen que trabajar sobre archivos Excel y generar los gráficos ellos mismos. La mayoría de ellos tienen un nivel alto de conocimiento de la herramienta de Excel, pero nunca han utilizado un software de inteligencia de negocios.

Se generan reportes casi todos los días, pero a finales de mes tienen que ejecutar un mayor número de reportes ya que deben cerrar la contabilidad mensual, si bien la base de datos no deja de funcionar si se ve reflejado en el desempeño de todos los sistemas la sobrecarga de procesamiento del servidor de base de datos. Pero si se han registrado fallas en el servidor de base de datos debido a esta sobrecarga, la cual se tiene que solucionar reiniciando el servidor.

Cada uno de los reportes con los que cuentan varían en la cantidad de datos que consultan por lo cual varía mucho el tiempo de ejecución de cada uno de ellos, pero los reportes que se eligieron para el desarrollo de esta investigación en promedio consultado

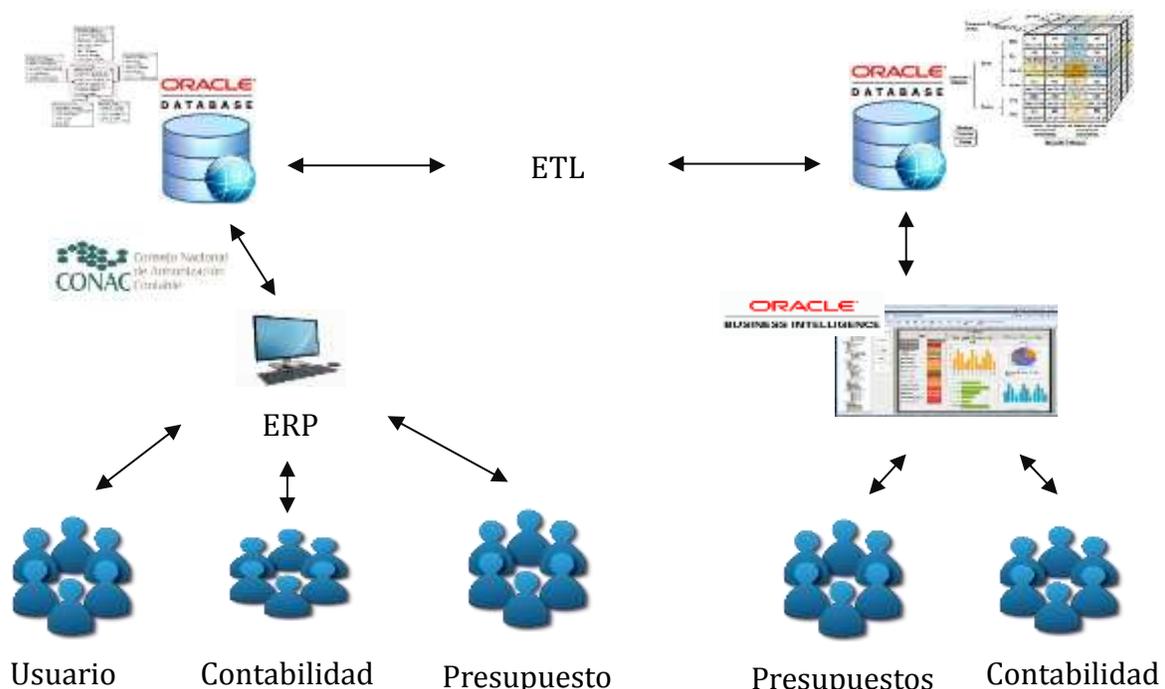
información a partir del 2012 y para todos los proyectos que existen, le toma un poco menos de 60 minutos ejecutarlo.

5.2. Resultados obtenidos con la implementación del repositorio unificado de datos y las consultas a través de tableros de control del BI

Para la realización de esta investigación solo se implementó el repositorio unificado para hacer algunas pruebas y poder recabar información de las mejoras en su implementación, solo fue usado por dos personas del departamento de presupuestos para tomar sus impresiones acerca del uso de esta nueva herramienta.

Es decir que en esta etapa no se logró probar las mejoras del rendimiento del servidor operativo al generar las consultas analíticas únicamente en el repositorio unificado de datos, pero si tomamos en cuenta que la mayor carga de procesamiento del servidor se debe a las consultas analíticas, dividiendo estas dos actividades en dos servidores distintos disminuirá la carga del servidor operativo de base de datos, en la figura 10 se muestra esta separación.

10. Figura con el diagrama de la nueva estructura de consultas a la base de datos



5.2.1. Ventajas al usar un software de inteligencia de negocios

Como ya se mencionó anteriormente, un software de inteligencia de negocios le permite al usuario final generar una gran cantidad de reportes gráficos y no gráficos a través de tableros de control con una interfaz hasta cierto punto amigable si se está familiarizado con este tipo de sistemas.

Con esto se logra que los mismos usuarios puedan generar sus reportes a través de esta herramienta sin la necesidad de pedir al departamento de informática que genere otro reporte de manera programática con información estática. La única responsabilidad del departamento de informática sería el mantenimiento y actualización de información del repositorio unificado.

En las pruebas que se generaron para la realización de esta investigación se logró determinar que con la información que se generó en el repositorio unificado de datos, se pueden obtener no solo los reportes que fueron elegidos para la investigación, también se pueden generar la mayoría de los reportes con los que ya cuentan, sin contar con los diversos gráficos que se pueden generar desde el software de inteligencia de negocios, es decir que ya no habría necesidad de construir reportes desde Excel, ya que los pueden generar desde esta herramienta y después importarlos en los distintos formatos disponibles.

Otra de las ventajas de la integración del repositorio unificado y el software de inteligencia de negocios, es que toda la información contenida en el almacén de datos está en un formato lo suficientemente claro para que cualquier persona pueda entenderla y no tenga que ser un administrador de base de datos para armar una consulta compleja, como se tiene que hacer en las bases de datos normalizadas.

En cuestión de tiempos de respuestas en las consultas desde el software de inteligencia de negocios al almacén de datos, en promedio un reporte sencillo sin gráficos y contemplando la misma información de los reportes de pruebas toma de 20 a 30 minutos, es decir la mitad o menos de la mitad que los reportes actuales.

Hay que tomar en cuenta que un reporte en el software de inteligencia de negocios que muestra una tabla con información y un par de gráficos, subirá el consumo de

procesamiento por lo tanto incrementara significativamente los tiempos de ejecución, es por eso que se deben de tomar consideraciones en las consultas, es decir que si se requiere un reporte con varios gráficos hay que filtrar la cantidad de información lo más posible como lo recomienda las bibliografía de Oracle, es decir que aunque los tiempos de respuesta se disminuyen con la implementación del repositorio unificado, hay que intentar en lo mayor posible filtrar la información en porciones más pequeñas.

5.2.2. Experiencia de usuario

En las pruebas que se hicieron con los usuarios del departamento de presupuestos, su reacción fue favorable ya que las consultas se asemejan un poco a las tablas dinámicas que se utilizan en Excel, además de que conocen la información contenida y digerida del almacén de datos, sin embargo, sí expresaron la necesidad de tener una capacitación para poder utilizar la herramienta.

En resumen, mostraron interés en su implementación ya que de tener el conocimiento completo de la herramienta no tendría que estar mezclando información de varios reportes y manipulándolos en archivos de Excel que eventualmente pueden llegar a perder. Además de tener la opción de generar reportes según sea la necesidad del momento, de tener el conocimiento completo y experiencia en el uso creen que si mejoraría la toma de decisiones.

6. Conclusiones y trabajo a futuro

Si bien se pueden cambiar los reportes estáticos que se utilizan actualmente para que consulten la información del almacén de datos para mejorar los tiempos de respuesta, se perdería la libertad que ofrece un software de inteligencia de negocios para crear distintos tipos de reportes y gráficas según serán requeridos.

También se podría utilizar el software de inteligencia de negocios con la base de datos relacional operativa y dejar de utilizar los reportes fijos, esto afectaría aún más el procesamiento del servidor de base de datos, ya que las consultas del software de inteligencia de negocios por ser gráficas requieren un mayor procesamiento.

Es por eso que la combinación perfecta es consultar un repositorio unificado de datos a través de un software de inteligencia de negocios para sacar el mayor provecho a toda la información que está disponible.

Oracle cuenta con mucha paquetería para llevar a cabo las mismas tareas de esta investigación, la elección de qué software utilizar depende de las necesidades del almacén de datos y del licenciamiento con el que se cuenta, para el caso de esta investigación la herramienta de ODWB cumplió con las necesidades.

6.1. Trabajo a futuro

- Buscar una capacitación del software de Oracle BI para que el personal de las áreas pueda sacar un mejor provecho de la herramienta.
- Si bien todo apunta a que no sería necesario una actualización del repositorio unificado de datos en tiempo real, se investigara más a fondo las opciones y requerimiento para lograr esta implementación.
- Para esta investigación se implementó un repositorio unificado relacional (ROLAP) por ser más fácil de entender e implementar, pero también se continuará con la investigación para implementar un repositorio multidimensional (MOLAP) y determinar las mejoras.

7. Referencias

- Castelán, L., & Ocharán, J. (2012). Diseño de un Almacén de datos basado en Data Warehouse Engineering Process (DWEP) y HEFESTO. *Universidad Veracruzana*, 10. Retrieved from <http://www.uv.mx/mis/files/2012/11/Diseno-de-un-Almacen-de-datos.pdf>
- Corporation, O. (2008). *Oracle 11g (11.1) DBA*
- Corporation, O. (2012). *Oracle® Database 2 Day + Data Warehousing Guide 11g Release 2 (11.2)*.
- Kimball, R., & Ross, M. (2003). The Data Warehouse Toolkit. In *ACM SIGMOD Record* (Third Edit, Vol. 32). <https://doi.org/10.1097/00005721-200605000-00016>
- Lumpkin, G. (2013). Oracle Database 11g para Data Warehousing e Inteligencia de Negocios. Retrieved from Informe Ejecutivo website: www.oracle.com/database
- Wiley & Sons. (2002). *W. Building the data warehouse*.