Universidad de las Ciencias Informáticas

Facultad 6

Sistema de Información de Gobierno. Mercado de datos Turismo.

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas.

Autores:

Yulier De la Cruz Olivares.

Ariel Manresa Sánchez.

Tutores:

Ing. Yosvany Arrastia Machín.

Ing. Yamila Mateu Romero.

"Ciudad de la Habana"

Junio, 2011

"...aquí está una de las tareas de la juventud: empujar, dirigir con el ejemplo la producción del hombre de mañana. Y en esta producción, en esta dirección, está comprendida la producción de sí mismos..."

DECLARACIÓN DE AUTORÍA

Declaro ser autor de la presente tesis y reconozco a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

Para que así conste firmo la presente a los ____ días del mes de _____ del año _____.

_____ _____

Yulier De la Cruz Olivares

Autor

Ariel Manresa Sánchez

Autor

_____ _____

Ing. Yamila Mateu Romero

Tutor

Ing. Yosvany Arrastia Machin

Tutor

Datos de contacto

Tutora: Ing. Yamila Mateu Romero

Email: ymateu@uci.cu

Graduada en la Universidad de las Ciencias Informáticas en 2008.

Tutor: Ing. Yosvany Arrastia Machin

Email: yarrastia@uci.cu

Graduado en la Universidad de las Ciencias Informáticas en 2010.

A mi madre, por brindarme todo su amor y apoyo, por ser la luz de mi vida, por ser lo más grande que tengo en el mundo.

A mi tío Lorenzo.

Yulier.

A toda mi familia, especialmente a mis padres y hermanos.

Ariel.

A mi familia por su confianza y apoyo.

A mis padres y hermanos por impulsarme y ayudarme a ser alguien en la vida.

A Orelmis, Yoendy, Ransel, Yuniel, Leonel, Fabián, Minato, Eugenio, Herryman y a todos los amigos que hice en estos cinco años y que siempre estuvieron presentes cuando los necesité.

A mi compañero de tesis Yulio por ser de los dos el que mejor supo guiar el trabajo.

A los tutores por todo el apoyo brindado.

Ariel.

En primer lugar tengo que agradecer a mi mamá que me ha guiado y apoyado en todo momento, quien siempre ha estado presente para mí y ha sacrificado todo lo necesario por mí y a quien debo todo lo que soy, gracias por confiar en mí.

A toda mi familia por estar siempre conmigo, por todo el cariño, dedicación y ternura, por simplemente estar ahí, y en especial a mis hermanas Neima Lien y Nelia, a Rodi, a mi querida abuela Nellibe, a mis tíos Lorenzo, Lay, Nilvia, Rosa Iris, Iliuva, Diego, Irael, Marilín, Victor. A todos mis primos sobre todo a Yudelkis.

A Enrique por guiarme y apoyarme en todo momento, por tratarme como un hijo más.

A Mauresa, mi compañero de tesis.

A mis tutores por apoyarme y estar siempre disponibles.

A Yeilen por estar conmigo, por soportar mis pesadeces y ayudarme todo el tiempo. A todas mis amistades, A Yusliel, Lari, Patry, Lisandra, Daynelis, Liset, Yohana, Yanier, Yuneidy, David.

A Doris por enseñarme, prepararme y apoyarme desde el momento en que pasé a formar parte del proyecto.

Resumen

El departamento de Turismo es una de las áreas de la Oficina Nacional de Estadísticas (ONE), que tiene la finalidad de presentar un conjunto de indicadores y cifras estadísticas que reflejan el desarrollo alcanzado en la actividad turística de nuestro país. Muchos de los análisis deseados sobre esta área no pueden ser realizados, pues no cuentan con el soporte para llevarlos a cabo. En este trabajo se propone una solución basada en el desarrollo de un mercado de datos que contribuya en el proceso de toma de decisiones de los directivos de la ONE y de Cuba. Para ello se realizó el análisis, diseño, implementación y prueba de la aplicación que constituye el resultado principal de este trabajo de diploma. Conjuntamente se hizo uso de varias herramientas y de una metodología de desarrollo para guiar el proceso de implementación. Se obtuvo un mercado de datos poblado y funcional, con una capa de inteligencia de negocio que brinda vistas de análisis actualizadas, permitiéndole a los especialistas de Turismo un mejor estudio de la información.

Palabras claves: Turismo, Mercado de datos, Almacén de datos e Inteligencia de Negocios.

TABLA DE CONTENIDOS.

INTRODUCCIÓN 1

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA DE LA INVESTIGACIÓN 4

1.1 Almacenes de Datos. 4

1.1.1 Definiciones de almacenes de datos 4

1.1.1 Tipos de Almacenamiento OLAP 6

1.1.2 Modelo Multidimensional 7

1.1.3 Soluciones estadísticas en el mundo 8

1.2 Integración de datos. 9

1.3 Inteligencia de Negocios. 14

1.4 Metodología de desarrollo. 15

1.5 Herramientas de Desarrollo. 18

Conclusiones 26

CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS TURISMO. 27

Introducción. 27

2.1 Caracterización de las áreas de la organización 27

2.2 Reglas del negocio 27

2.3 Especificación de requerimientos 28

2.3.1 Requisitos de información 28

2.3.2 Requisitos funcionales 31

2.3.3 Requisitos no funcionales 32

2.4 Modelo de casos de usos del sistema 33

2.4.1 Actores del sistema 33

2.4.2 Diagrama de casos de uso del sistema 33

2.4.3 Especificación de casos de uso del sistema 34

2.5 Especificación del modelo dimensional 34

2.5.1 Tablas de hechos 35

2.5.2 Tablas de dimensiones 36

2.5.3 Matriz Bus 36

2.5.4 Modelo dimensional 37

2.6 Especificación del modelo físico 39

2.7 Perfilado de datos 42

Conclusiones 43

CAPÍTULO 3: IMPLEMENTACIÓN DEL MERCADO DE DATOS TURISMO. 44

Introducción 44

3.1 Implementación del modelo de datos 44

3.2 Arquitectura de integración 45

3.3 Implementación de los procesos ETL 46

3.3.1 Extracción de los datos 46

3.3.2 Transformación y carga de los datos. 46

3.4 Implementación de subsistema de visualización 50

3.4.1 Cubos OLAP 50

3.4.2 Arquitectura de información 53

3.4.3 Creación y administración de los reportes. 54

3.4.4 Configuración de la seguridad de los usuarios 55

Conclusiones 56

CAPÍTULO 4: VALIDACIÓN DEL MERCADO DE DATOS TURISMO 57

4.1 Listas de chequeo. 57

4.2 Casos de prueba. 59

4.3 Pruebas de aceptación 61

Conclusiones. 62

Conclusiones generales. 63

Recomendaciones 64

Bibliografía 65

Referencias Bibliográficas 67

INTRODUCCIÓN

Con el progreso de los años la informática evoluciona cada vez más, diariamente surgen nuevos avances científico-técnicos, provocando una gran competencia entre empresas y en el mercado. Debido a esto y al crecimiento del volumen de la información que se genera frecuentemente, las empresas se mantienen en la búsqueda de nuevas tecnologías que faciliten que la información almacenada sea precisa, accesible de forma rápida y fiable.

Unas de las tecnologías más usadas en estas situaciones es la Inteligencia de Negocios (BI), que no es más que el conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización o empresa. Abarca técnicas como minería de datos y almacenes de datos. Como parte del ciclo de vida de una implementación de Inteligencia de Negocios está el proceso Extracción, Transformación y Carga (ETL) para lograr la integridad de los datos.

Cuba se preocupa por el uso de estas tecnologías con el fin de optimizar el trabajo en cada una de sus empresas como es el caso de la Oficina Nacional de Estadísticas (ONE), encargada en el país de garantizar la producción de estadísticas de calidad a través del Sistema Estadístico Nacional (SEN). Dicha entidad es el órgano rector de la estadística y tiene como objetivo fundamental captar, analizar y difundir los datos recogidos a lo largo y ancho de todo el país.

El departamento de Turismo es una de las áreas de la ONE, que tiene la finalidad de presentar un conjunto de indicadores y cifras estadísticas que reflejen el desarrollo alcanzado en la actividad turística de nuestro país. Muchos de los análisis deseados sobre esta área no pueden ser realizados, pues no cuentan con el soporte para llevarlos a cabo. Estos datos son almacenados en formatos de difícil acceso para su consulta por lo que se hace muy complejo el proceso de acceder y divulgar dicha información.

Debido a esto se identifica el siguiente **problema de la investigación**: ¿Cómo contribuir a la toma de decisiones en el área Turismo del Sistema de Información de Gobierno?

Definiendo como **objeto de estudio:** los almacenes de datos e inteligencia de negocios, centrando su **campo de acción** en el mercado de datos y capa de visualización para el área Turismo del Sistema de Información de Gobierno.

Para dar respuesta al problema planteado se determina como **objetivo general** desarrollar el mercado de datos Turismo del Sistema de Información de Gobierno que contribuya a la toma de decisiones. Centrándose en los siguientes **objetivos específicos:**

1. Realizar el análisis y diseño del mercado de datos del área de Turismo.
2. Implementar el mercado de datos del área de Turismo.
3. Implementar la capa de visualización de datos del área de Turismo.
4. Validar el mercado de datos del área de Turismo.

Para cumplir el objetivo y darle una solución eficiente al problema planteado se trazan las siguientes **tareas de la investigación**:

1. Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
2. Levantamiento de requisitos.
3. Descripción de los casos de uso del mercado de datos.
4. Definición de los hechos, las medidas y las dimensiones del mercado de datos.
5. Diseño del modelo de datos.
6. Definición de la arquitectura del mercado de datos.
7. Diseño del subsistema de integración.
8. Diseño del subsistema de visualización.
9. Diseño de los casos de pruebas.
10. Implementación del subsistema de integración.
11. Implementación del subsistema de visualización.
12. Aplicación de las listas de chequeo.
13. Aplicación de los casos de pruebas.

Así se obtendrá como **posible resultado** un mercado de datos poblado.

Para cumplir lo antes expuesto el presente Trabajo de Diploma se ha estructurado en cuatro capítulos:

Capítulo 1: Fundamentación teórica de la investigación.

En el capítulo se definen temas relacionados con el desarrollo de los almacenes de datos, se describen los conceptos fundamentales que serán tratados a lo largo de la investigación, así como la selección de la metodología, técnicas y herramientas que serán utilizadas.

Capítulo 2: Análisis y diseño del mercado de datos Turismo

En este capítulo se hará un estudio preliminar del negocio con el objetivo de definir necesidades de información, reglas del negocio, casos de uso y su descripción, identificación de dimensiones, hechos y medidas, desarrollo de la matriz BUS, modelo de datos.

Capítulo 3: Implementación del mercado de datos Turismo

En este capítulo se procederá a la implementación de la solución propuesta, se implementará el modelo de datos, el modelo de despliegue, los flujos de transformación y los trabajos. Se implementará la base de datos y se realizará el montaje de los clasificadores para el mercado de datos.

Capítulo 4: Validación del mercado de datos Turismo.

Una vez realizado el análisis, diseño e implementación del mercado de datos Turismo, se da paso a la validación y prueba de la solución mediante las listas de chequeo, los casos de prueba y las pruebas de aceptación para así verificar que el sistema cumpla con los requerimientos necesarios que garanticen al usuario final del sistema la confiabilidad de los datos cargados en el almacén de datos.

CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA DE LA INVESTIGACIÓN

Introducción

En el capítulo se definen temas relacionados con el desarrollo de los almacenes de datos, se describen los conceptos fundamentales que serán tratados a lo largo de la investigación, así como la selección de la metodología, técnicas y herramientas que serán utilizadas.

1.1 Almacenes de Datos.

El desarrollo en las últimas décadas ha conducido a una situación en la que el volumen de datos históricos almacenados en las organizaciones crece significativamente, creando la necesidad de disponer de un sistema que les permita gestionar la información y ayude a tomar decisiones eficaces. Así surge el punto de partida de los Almacenes de Datos.

1.1.1 Definiciones de almacenes de datos

Definición de Bill Inmon

Bill Inmon [1] a principios de los noventa definió el tema de los Almacenes de Datos (Data Warehouse, DW): "*Un almacén de datos es una colección de datos orientados por temas, integrados, variables en el tiempo y no volátiles para el apoyo de la toma de decisiones*". Orientados por tema porque los datos son estructurados según los temas de interés para facilitar su análisis. Integrado debido a que los datos que se introducen en el almacén de datos se obtienen de diversas fuentes de origen y distintos formatos, y es tarea del proceso ETL aplicarles las transformaciones necesarias para integrarlos. Variables en el tiempo porque los datos siempre tienen que estar ligados a un instante de tiempo. Y no volátil porque en el almacén los datos pueden ser consultados pero no modificados, por tanto, la información es permanente y la actualización consiste exclusivamente en la incorporación de nuevos datos.

Definición de Ralph Kimball

Ralph Kimball es reconocido a nivel mundial en el diseño de los almacenes de datos y creador del enfoque multidimensional. Se ha dedicado al desarrollo de su metodología para que éste concepto sea bien aplicado en las organizaciones y se asegure la calidad en el desarrollo de estos proyectos. Define

almacén de datos como la unión de todos los mercados de datos constituyentes. Un almacén de datos se alimenta desde el área de preparación de datos.

En el Centro de Tecnologías de Gestión de Datos (DATEC) los almacenes de datos se forman a partir de los mercados de datos, sin embargo cumplen con las características que hace referencia Inmon en su definición, por tanto en este trabajo de investigación se define que un almacén de datos es una colección de datos orientados por temas, integrados, variables en el tiempo y no volátiles para el apoyo de la toma de decisiones, que se compone por la unión de todos los mercados de datos constituyentes.

Definición de Mercado de Datos

Un Mercado de Datos es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la estructura óptima de datos para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento. [2]

Arquitectura

La arquitectura de un almacén de datos se suele representar como varias capas a través de las cuales circulan datos, de modo que los datos de una capa se obtienen a partir de los datos de la capa previa. A partir de esta arquitectura se considera que el desarrollo de un almacén de datos se puede estructurar en un marco integrado por tres niveles que definen los diferentes diagramas empleados para modelar un almacén de datos:

- **Conceptual:** define el almacén de datos desde un punto de vista conceptual, o sea, desde el mayor nivel de abstracción y contiene únicamente los objetos y relaciones más importantes.
- **Lógico:** abarca aspectos lógicos del diseño del almacén de datos, como la definición de las tablas y claves, la definición de los procesos ETL y otros.
- **Físico:** define los aspectos físicos del almacén de datos, como el almacenamiento de las estructuras lógicas o la configuración de los servidores que mantienen el almacén de datos.

Principales aportes de un almacén de datos

- Proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global del negocio.
- Facilita la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén; obteniendo un valor añadido para el negocio de dicha información.
- Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.
- Simplifica dentro de la empresa la implantación de sistemas de gestión integral de la relación con el cliente.
- Supone una optimización tecnológica y económica en entornos de Centro de Información, estadística o de generación de informes con retornos de la inversión espectaculares.

1.1.1 Tipos de Almacenamiento OLAP

Los sistemas OLAP (Procesamiento Analítico en Línea) son bases de datos orientadas al procesamiento analítico. Este análisis suele implicar, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejos. Existen diferentes tipos de almacenamientos OLAP entre ellos MOLAP, HOLAP y ROLAP. [3]

Procesamiento Analítico Multidimensional (MOLAP)

La arquitectura MOLAP usa unas bases de datos multidimensionales para proporcionar el análisis, su principal premisa es que el OLAP está mejor implantado almacenando los datos multidimensionalmente. Utiliza una arquitectura de dos niveles: Las bases de datos multidimensionales y el motor analítico donde la base de datos multidimensional es la encargada del manejo, acceso y obtención del dato.

La arquitectura MOLAP requiere unos cálculos intensivos de compilación. Lee datos precompilados, y tiene capacidades limitadas de crear agregaciones dinámicamente o de hallar ratios que no se hayan precalculados y almacenados previamente.

Procesamiento Analítico en Línea Relacional (ROLAP)

La arquitectura ROLAP, accede a los datos almacenados en un almacén de datos para proporcionar los análisis OLAP. La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan mejor contra las bases de datos relacionales.

La arquitectura ROLAP es capaz de usar datos precalculados si estos están disponibles, o de generar dinámicamente los resultados desde los datos elementales si es preciso. Esta arquitectura accede directamente a los datos del almacén de datos, y soporta técnicas de optimización de accesos para acelerar las consultas. Estas optimizaciones son, entre otras, particionado de los datos a nivel de aplicación, soporte a la desnormalización y uniones múltiples.

Procesamiento Analítico Híbrido en Línea (HOLAP)

Combina las arquitecturas ROLAP y MOLAP para brindar una solución con las mejores características de ambas: desempeño superior y gran escalabilidad. Un tipo de HOLAP mantiene los registros de detalle (los volúmenes más grandes) en la base de datos relacional, mientras que mantiene las agregaciones en un almacén MOLAP separado.

Después de realizarse un análisis de los modos de almacenamiento de datos mencionados anteriormente, se llegó a la conclusión de que el más adecuado a la solución es ROLAP, además que soporta Postgres, siendo este el sistema gestor de base de datos seleccionado.

1.1.2 Modelo Multidimensional

A diferencia de los sistemas de bases de datos convencionales, la estructura de las tablas y sus relaciones son representadas mediante un modelo multidimensional, es decir, almacenan la misma información que el diagrama entidad relación pero organizada de manera diferente, de modo que se garantice velocidad y eficiencia en la recuperación de dicha información. Las principales ventajas están dadas por:

- Su enfoque al negocio y sus actividades.
- Búsquedas a gran velocidad.
- Adición de dimensiones y hechos que no se habían previsto, sin que esto implique volver a cargar los datos ya almacenados.

- Adición de nuevos atributos a las dimensiones.

El modelo multidimensional incluye tres variantes de modelación, que está determinado por la complejidad del sistema:

Esquema estrella: Es la técnica más común. En este esquema existe un único elemento central (tabla de hechos) conectado radialmente con las tablas de dimensiones.

Esquema copo de nieve (Snowflake): Es un esquema derivado del esquema de estrella, las tablas de dimensiones se ramifican en más puntas.

Esquema constelación: Está conformado por una serie de esquemas de estrella, o sea, una tabla de hechos central con otras auxiliares y sus respectivas tablas de dimensiones.

Para una adecuada comprensión de los distintos tipos de modelado existentes es necesario dominar algunos conceptos básicos referentes al tema, que a continuación se exponen.

Hecho

Es el objeto a analizar, posee atributos llamados de hechos o de síntesis, y son de tipo cuantitativo. Sus valores que se denominan medidas, se obtienen generalmente por la aplicación de una función estadística que resume un conjunto de valores en un único valor.

Dimensiones

Representan cada uno de los ejes en un espacio multidimensional. Suministran el contexto en el que se obtienen las medidas de un hecho. Las dimensiones se utilizan para seleccionar y agrupar los datos en un nivel de detalle deseado. Los componentes de una dimensión se denominan niveles y se organizan en jerarquías.

Es importante acotar que los hechos se guardan en tablas de hechos y las dimensiones en tablas de dimensiones.

1.1.3 Soluciones estadísticas en el mundo

La estadística es una ciencia de aplicación práctica casi universal en todos los campos científicos. Aunque comúnmente se asocie a estudios demográficos, económicos y sociológicos, gran parte de los

logros de la estadística se derivan del interés de los científicos por desarrollar modelos que expliquen su comportamiento. Varios estudios estadísticos comunes que aparecen con frecuencia en los medios de comunicación son los siguientes:

- **Encuesta de Población Activa** (EPA), elaborada por el Instituto Nacional de Estadística (INE) con periodicidad trimestral, según recomendaciones de la Organización Internacional del Trabajo (OIT), para obtener y clasificar datos sobre la actividad de la población.
- **Índice de Precios al Consumo** (IPC), que mide por medios estadísticos la evolución experimentada por los precios de los bienes y servicios consumidos por la población española. Se basa en la Encuesta de Presupuestos Familiares (EPF), y selecciona varios centenares de artículos, clasificados en ocho grupos, que se consideran representativos de la evolución de los precios.
- **Producto Interior Bruto** (PIB), que registra la producción nacional de un país en bienes y servicios asociados a procesos considerados productivos.
- **Poder adquisitivo**, que maneja combinadamente datos del Salario Mínimo Interprofesional (SMI) y el IPC.

Los almacenes de datos además de ser una herramienta que facilita la aplicación de técnicas estadísticas y que apoyan la toma de decisiones en cualquier área funcional no es la única solución estadística para las empresas a nivel mundial. Existen diferentes aplicaciones que facilitan el control estadístico, entre ellas se pueden mencionar la aplicación Movimiento Natural de la Población, versión 1 (*MNPv1*), es una herramienta destinada a la gestión administrativa e informática de los datos del Movimiento Natural de la Población en Andalucía. Dicha gestión se inscribe en el marco de lo dispuesto en el convenio de colaboración del Instituto Nacional de Estadísticas (INE) y el Instituto de Estadística de Andalucía (IEA). Otro sistema para el control estadístico es el *openGis-EIEL* que ha sido diseñado para la gestión de Encuestas de Infraestructuras y Equitaciones Locales (EIEL).

1.2 Integración de datos.

Las empresas luchan aún con la necesidad vital de entregar los datos correctos a las personas correctas y en el momento adecuado, de mezclar las diversas necesidades de información presente en diferentes fuentes (bases de datos, ficheros), y de proveer una vista unificada de esta información para

el usuario. Precisamente este es el objetivo de la integración de datos, permitir el desarrollo rápido de nuevas aplicaciones que requieren información de múltiples fuentes. Este simple objetivo oculta muchos desafíos, desde la identificación de las mejores fuentes de datos a utilizar, a la creación de una interfaz adecuada para los datos integrados. Muchas investigaciones se han centrado en la forma de mejorar la integración en sí, por ejemplo: ¿Cómo consultar diversas fuentes con diferentes capacidades y cómo optimizar las consultas o los planes de ejecución?

Definición de Laura Haas

De acuerdo con lo descrito por Laura Haas [4] la integración de datos es un proceso con cuatro tareas principales: la comprensión, la normalización, especificación y ejecución.

Comprensión: La primera tarea de la integración de datos es analizar y entender la fuente. Durante esta tarea, el integrador puede buscar relaciones entre los datos y sus significados.

Normalización: Esta tarea aprovecha el trabajo de la tarea Comprensión para determinar el mejor método de integración, la forma de limpiar o reparar los datos. En esta etapa surgen cuestiones tales: ¿Cómo manejar los datos inconsistentes o incompletos? ¿Cómo identificar los datos que hace referencia a los mismos?

Especificación: En esta tarea se producen los artefactos que controlarán la ejecución. Las técnicas y tecnologías utilizadas para obtener las especificaciones están íntimamente vinculadas a la elección del motor de ejecución. La especificación es parte de la configuración de un motor de integración para hacer la integración deseada.

Ejecución: Aquí es donde realmente sucede la integración de datos y puede ser lograda a través de la Materialización, Federación o Indexación. La Materialización crea y almacena los datos integrados, una de las técnicas que utiliza es Extracción, Transformación y Carga (ETL).

Aún siendo invisibles por los usuarios de la plataforma BI, la exactitud de dicha plataforma depende del proceso ETL. La idea es que una aplicación ETL lea los datos desde diferentes fuentes, realice transformación, validación, el proceso cualitativo, filtración y al final los inserte en un almacén de datos listos para ser analizados por los usuarios.

Métodos básicos de integración de datos

1. Un almacén de datos tradicional, se actualiza periódicamente desde orígenes de datos de producción.
2. Un almacén de datos en tiempo real, se actualiza constantemente mediante recogida uniforme de datos desde orígenes de datos de producción.
3. Acceso a datos operativos, donde los usuarios obtienen una vista en tiempo real de la actividad del negocio desde las aplicaciones y los datos.
4. Integración de información empresarial (EII), donde los usuarios de BI pueden agregar en tiempo real datos corporativos en diversos orígenes de datos.
5. Integración de procesos, que permite entregar información en tiempo real basada en un evento empresarial o como parte de un proceso de negocios.
6. Tecnología de búsqueda, que permite buscar rápidamente en contenido indexado para crear resultados (al estilo Google) a partir de orígenes de datos de la empresa.
7. Servicios Web, que pueden mostrar o extraer datos de varios orígenes de información, independientemente del sistema operativo, aplicación o base de datos.[5]

Principales técnicas

Existen tres técnicas principales usadas para la integración de datos: consolidación, federación y propagación. La consolidación de datos captura los datos de múltiples fuentes y los integra en un único almacén de datos. La federación de datos proporciona una vista virtual de uno o más archivos de datos de origen, Integración de Información Empresarial (EII) es una tecnología que sirve como ejemplo para esta técnica. Y la propagación no es más que copiar los datos desde una fuente hacia otra, Integración de Aplicaciones Empresariales (EAI) y Empresa de Replicación de Datos (EDR) son tecnologías que sirven de apoyo para esta técnica. Las técnicas utilizadas por las aplicaciones de integración de datos dependerán de los requerimientos del negocio y la tecnología. Es muy común para una aplicación de integración de datos utilizar un enfoque híbrido que incluye varias técnicas de integración de datos.[6]

Principales tecnologías

Entre las principales tecnologías para la integración de datos se pueden encontrar Extracción, Transformación y Carga (ETL), Integración de Información Empresarial (EII), Integración de Aplicaciones Empresariales (EAI), y la Empresa de Replicación de Datos (EDR).

ETL: Como su nombre lo indica, esta tecnología permite a las organizaciones mover datos desde diversas fuentes, reformatearlos, limpiarlos y cargarlos en otras bases de datos. Fortalece los datos para la construcción de bases de datos permanentes dedicadas al análisis o generación de informes, las cuales pueden ser convertidas de un tipo o formato a otro. Es utilizado para migrar datos de una o más bases de datos a terceros. Con este proceso se pueden formar Repositorios de Datos, Mercados de Datos y Almacenes de Datos.

Éstas y otras funcionalidades hacen de este proceso imprescindible a la hora de integrar datos. Sus características más significativas son:

- Es un mecanismo de carga muy eficiente y efectivo orientado a los Almacenes de Datos.
- Enfocado a migrar y mezclar datos.
- Reduce la exposición a desarrollos manuales (codificación) producto de la existencia en el mercado de herramientas potenciales para la implementación visual, con manejo de excepciones, gestión y planificación de tareas.
- Necesita pocos servicios de administración y mantenimiento.
- Gran capacidad para llevar a cabo transformaciones.
- Tecnología enfocada a la Integración de datos en bases de datos versátiles hacia los Almacenes de Datos.

Subprocesos ETL

El proceso ETL se divide en tres subprocesos fundamentales, que permiten la segmentación y entendimiento de este arduo trabajo, los cuales se exponen a continuación:

Extracción

El proceso de extracción consiste en adquirir los datos desde los sistemas de origen, estas fuentes pueden estar sobre sistemas incompatibles o de hardware diferentes. En este subproceso se convierten los datos a un formato preparado para iniciar el proceso de transformación. Aquí se verifican los datos extraídos, donde se comprueba si los datos cumplen lo que se espera y se adaptan al formato estándar diseñado, de lo contrario los datos son rechazados.

En el proceso de extracción es necesario causar un mínimo impacto en el sistema origen, pues si se necesita extraer muchos datos el sistema origen podría ralentizar o colapsar, provocando que no pueda implementarse con normalidad para su uso cotidiano.

Limpieza y Transformación

En esta fase se aplican reglas de negocios[1] sobre los datos extraídos, con el objetivo de convertirlos en datos aptos para ser cargados. Aquí es necesario lograr una buena calidad de los datos y para ello es necesario el control de los valores válidos, garantizar la coherencia entre los valores, la eliminación de duplicaciones y comprobar que las reglas del negocio no han sido forzadas [7]. Para esta fase los datos deben ser limpiados, pues estos pueden estar sucios e incompletos. Por ello se realiza un proceso de limpieza que elimina errores e inconsistencias en los datos y resuelve el problema de identidad de los objetos.

Luego que los datos han sido limpiados se procede a realizar las transformaciones mediante las reglas de transformación que pueden ser: combinar los datos de distintas fuentes, realizar búsqueda de valores en distintas tablas, darle tratamiento a valores nulos, entre otras.

Carga

La fase de carga es el momento cuando los datos, provenientes de la fase anterior, son incluidos en el sistema de destino dependiendo de los requerimientos de la organización. El principal objetivo de esta fase es lograr que los datos estén listos para ser consultados. Este subproceso puede abarcar una amplia variedad de acciones diferentes. En algunas bases de datos se sobrescribe la información antigua con nuevos datos. En los Almacenes de Datos al mantener un historial de los registros se puede hacer una auditoría de los mismos y disponer de un rastro de toda la historia de un valor a lo largo del tiempo, independientemente de la acción a tomar para la carga, al realizar esta operación se aplicarán todas las restricciones que se hayan definido, lo que contribuye a que se garantice la calidad de la información en el proceso ETL.

[1] Describe las políticas, normas, operaciones, definiciones y restricciones presentes en una organización y que son de vital importancia para alcanzar los objetivos misionales.

1.3 Inteligencia de Negocios.

Uno de los activos más valiosos de cualquier empresa son los datos ya que contienen información histórica, pero mientras no se transformen en información útil para el negocio de forma tal que permita realizar el mejor análisis posible su valor es poco significativo.

Definición

Para lograr tener la correcta utilización de los datos se utiliza la Inteligencia de Negocios con el fin de transformar los datos en información, la información en conocimiento y luego utilizar ese conocimiento para la toma correcta de decisiones. Se puede definir como el proceso de analizar los datos acumulados con el tiempo por una empresa y extraer conocimientos de ellos, lo que posibilita conocer cómo su negocio se ha comportado a lo largo del tiempo, cómo se comporta en el presente y cómo se estima se comportará en el futuro.

También se puede definir como Inteligencia de Negocios al conjunto de estrategias y herramientas enfocadas a la administración y creación de conocimiento mediante el análisis de datos existentes en una organización. Estas estrategias y herramientas tienen en común las siguientes características:

- **Accesibilidad a la información:** Los datos son la fuente principal de este concepto. Lo primero que deben garantizar este tipo de herramientas y técnicas.
- **Facilidad de acceso por los usuarios** a los datos con independencia de la procedencia de estos.
- **Apoyo en la toma de decisiones:** Se busca ir más allá en la presentación de la información, de manera que los usuarios tengan acceso a herramientas de apoyo a la toma de decisiones.
- **Análisis** que les permitan seleccionar y manipular sólo aquellos datos que les interesen.
- **Con orientación al usuario final:** Se busca independencia entre los conocimientos técnicos de los usuarios y su capacidad para utilizar estas herramientas.

Clasificación de las soluciones BI

Las soluciones de BI se pueden clasificar en:

- Consultas e informes (Queries y reports)

- Cubos OLAP (On Line Analytic Processing)
- Minería de datos (Data Mining)
- Sistemas de previsión empresarial
- Sistemas de Gestión de la Performance Corporativa.

Beneficios

Son múltiples los beneficios que puede obtener una empresa con la implementación de un Sistema de Inteligencia de Negocios, entre ellos:

- Mayor capacidad de análisis.
- Reducción de costos.
- Reducción de tiempo de los procesos.
- Búsqueda de patrones que solo aparecen cuando se analizan los datos.
- Generación de pronósticos.
- Planificación empresarial.

1.4 Metodología de desarrollo.

Las metodologías de desarrollo no son más que una colección de documentación formal referente a los procesos, las políticas y los procedimientos que intervienen en el desarrollo del software y tienen la finalidad de garantizar la eficacia y la eficiencia en el proceso de generación de software.

1.4.1 Metodologías para los almacenes de datos

Existen diferentes metodologías para el desarrollo de los almacenes de datos entre las que se destacan la metodología de Inmon y la metodología Kimball. Ralph Kimball y Bill Inmon son dos de las personalidades referentes y más influyentes en el área de los Almacenes de Datos. El primero es un especialista reconocido a nivel mundial en el diseño de los almacenes de datos y creador del enfoque Multidimensional; mientras que el segundo es el creador del término almacenes de datos y considerado como el padre de la disciplina.

La principal diferencia que existe entre ambas metodologías está basada en la forma de enfrentar el problema. La visión de Inmon se basa en un enfoque descendente (top-down), propone construir primero el almacén de datos y a partir de este los mercados de datos, plantea la creación de un repositorio de datos corporativo como fuente de información consolidada, persistente, histórica y de

calidad. Al ser construido descendentemente los mercados de datos se nutren del almacén de datos corporativo, convirtiéndose en un complejo empresarial de bases de datos relacionales.

Inmon afirma que la creación de una base de datos relacional con una leve normalización necesita ser la base para los mercados de datos. Por lo que no se crean los mercados de datos directamente desde los Sistemas de Procesamiento de Transacciones en Línea (OLTP) a través de un área de ensayo. En lugar de ello, se crean a partir de la arquitectura relacional de los datos corporativos.

A diferencia de la anterior, la propuesta de Kimball se basa en dividir el mundo de Inteligencia de Negocio entre los hechos y las dimensiones, esta es eficaz y conduce a una solución completa en un corto período de tiempo. Se puede empezar de cero y dar al usuario una primera información sobre sus datos en cuestión de días. Además, tiene abundante documentación y se puede encontrar una respuesta a casi todas las preguntas que se puedan tener.

Entre sus características principales es que su arquitectura es ascendente (bottom-up), plantea que se debe crear por cada departamento un conjunto de mercados de datos independientes orientados a los temas que estén relacionados con él.

Fases de la metodología Kimball

1. **Requerimientos y Gestión del Proyecto**

- Definición del Proyecto
- Planeación y gestión del Proyecto
- Definición de los requerimientos de usuario

2. **Arquitectura**

- Diseño Técnico de la Arquitectura
- Medidas Tácticas de Seguridad
- Plan Estratégico de Seguridad
- Selección e Instalación de Productos

3. **Diseño e Implementación**

- Análisis Multidimensional (Lógico y Físico)
- Análisis de Fuentes de Datos
- Diseño & Implementación del Área Temporal
- Popular & Validar Data bases

- Optimización del Rendimiento
- Especificación y Desarrollo de Aplicaciones de Usuario Final

4. **Implantación & Operaciones**

- Plan de Implantación
- Pruebas
- Implantación (Alpha, Beta & Product Iterations)
- Optimización del Rendimiento
- Mantenimiento
- Crecimiento
- Capacitación y Transferencia Tecnológica

En este caso, para definir la metodología de desarrollo a utilizar, se tomó como base la Metodología Kimball por los siguientes elementos:

- Crea los conceptos de hechos y dimensiones, lo que indudablemente es muy eficaz en el proceso de la toma de decisiones y proporciona mayor agilidad en el proceso de desarrollo.
- Propone ir construyendo el almacén de datos a través de la construcción de los mercados de datos departamentales, lo que constituye una estrategia buena y coincide con la división lógica de las empresas, entidades, organismos, etc.
- Existe abundante documentación sobre la misma, la respuesta a todas las dudas y preguntas que puedan surgir se pueden encontrar en la web, a través de los servicios que brindan el grupo creador de la metodología.
- Es una metodología madura y reconocida por el resto de la comunidad dedicada al tema. Tiene bien definidas las etapas, actividades, artefactos y roles.

Como complemento a la misma y fortaleciendo la etapa del levantamiento de requisitos; se tomó lo planteado por Leopoldo Zenaido Zepeda Sánchez en su Tesis de Doctorado, orientando así el trabajo a los Casos de Uso y se logra estar más alineado con las tendencias y normas de la Universidad, teniendo en cuenta las características de la Universidad de las Ciencias Informáticas (UCI) y DATEC.

Ciclo de vida de la Metodología para el desarrollo.

- Estudio Preliminar o Planeación

- Requerimientos
- Arquitectura y Diseño
- Implementación
- Prueba
- Despliegue
- Soporte y Mantenimiento
- Gestión y Administración del Proyecto

1.5 Herramientas de Desarrollo.

Pentaho Data Integration

Para el desarrollo del proceso ETL se propone el Pentaho Data Integration debido a que es una herramienta de código abierto adoptado por Pentaho BI. Proporciona la extracción de gran alcance, Transformación y Carga (ETL) utilizando un enfoque innovador, orientado a los metadatos. Con una interfaz intuitiva, gráfica de arrastre, una probada arquitectura escalable y basada en estándares, Pentaho Data Integration es cada vez más la elección de las organizaciones tradicionales.

Presenta un enfoque orientado a los metadatos que simplemente indica qué quiere hacer, pero no cómo desea hacerlo. Con él los desarrolladores ETL, BI y los administradores pueden crear complejas transformaciones y el empleo en un entorno gráfico, arrastrar y soltar sin tener que generar ningún código personalizado. Pentaho Data Integration es una solución de ETL con todas las funciones incluyendo:

- Rica colección de transformación con más de 150 objetos de asignación.
- Amplia fuente de datos de apoyo incluyendo aplicaciones empaquetadas, más de 30 plataformas de código abierto y propietario de base de datos, archivos planos, documentos de Excel, y más.
- Apoyo al gran análisis de datos con la integración y gestión de datos.
- Opciones avanzadas de almacenamiento de datos que sirven de apoyo para la variación lenta no deseada y dimensiones.

- Rendimiento y escalabilidad.
- La integración con la suite de Pentaho BI para Enterprise Information Integration (EII), programación avanzada, y la integración de procesos.
- Unificado de ETL, modelado y visualización de entorno de desarrollo para el diseño de aplicaciones de BI.

Usos comunes para Pentaho Data Integration:

- Población de almacenes de datos.
- Diseño ágil de aplicaciones de BI.
- Información de enriquecimiento mediante la integración de datos de diversas fuentes.
- Migración de datos entre aplicaciones.
- La importación de datos en bases de datos de archivos de texto, hojas de cálculo Excel, los sistemas relacionales y más.
- Limpieza de datos mediante la aplicación de las complejas condiciones en las transformaciones de datos.
- Exploración de los datos existentes en bases de datos.

Sistema Gestor de Base Datos.

Un Sistema Gestor de base datos (DBMS por sus siglas en inglés) es un conjunto de programas que permiten crear y mantener una base datos, asegurando su integridad, confidencialidad y seguridad. Ejemplos de gestores de bases de datos son:

- PostgreSQL
- SQLite
- DB2 Express-C
- Apache Derby
- Microsoft SQL Server
- Sybase ASE Express Edition
- Oracle Express Edition 10

PostgreSQL

Se decide utilizar como sistema gestor de base datos PostgreSQL, teniendo en cuenta que es un sistema de gestión de bases de datos objeto-relacional, distribuido bajo licencia BSD y con su código fuente disponible libremente. Es el sistema de gestión de bases de datos de código abierto más potente del mercado y en sus últimas versiones no tiene nada que envidiarle a otras bases de datos comerciales.

PostgreSQL utiliza un modelo cliente/servidor y usa multiprocesos en vez de multihilos para garantizar la estabilidad del sistema. Un fallo en uno de los procesos no afectará el resto y el sistema continuará funcionando. Funciona de manera excelente con grandes cantidades de datos y una alta concurrencia de usuarios accediendo a la vez al sistema. Entre las características generales, además de estabilidad, potencia, robustez, facilidad de administración e implementación de estándares se encuentran:

- Es una base de datos 100% ACID.
- Integridad referencial.
- Espacios de tablas.
- Transacciones anidadas.
- Replicación asíncrona.
- Copias de seguridad en caliente (Online/hot backups).
- Unicode.
- Juegos de caracteres internacionales.
- Múltiples métodos de autentificación.
- Actualización in-situ integrada (pg_upgrade).
- Completa documentación.
- Licencia BSD.
- Disponible para Linux y UNIX en todas sus variantes (AIX, BSD, HP-UX, SGI IRIX, Mac OS X, Solaris, Tru64) y Windows 32/64bit.

Ventajas:

- Diseñado para ambientes de alto volumen.
- Multiplataforma.

- Extensible.
- Ahorros considerables en costos de operación.
- Mejor soporte que los proveedores comerciales.
- Tamaño de base de dato ilimitado.

Visual Paradigm

Como herramienta de modelado se usará Visual Paradigm para UML ya que es una herramienta UML fácil de usar que soporta ingeniería inversa, generación de código, importación desde Rational Rose, exportación/importación XMI, generador de informes, editor de figuras, integración con MS Visio, plug-in, integración IDE con Visual Studio, IntelliJ IDEA, Eclipse, NetBeans y además soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. Entre sus nuevas características se incluyen el modelado colaborativo con CVS y Subversion, interoperabilidad con modelos UML2 a través de XMI.

Otras de las características que presenta son:

- Ingeniería de ida y vuelta.
- Soporte de *UML* versión 2.1.
- Diagramas de flujos de datos.
- Generación de bases de datos.
- Editor de detalles de casos de uso.
- Diagramas de procesos del negocio.
- Ingeniería inversa de bases de datos.
- Distribución automática de diagramas.

Entre las ventajas se encuentran:

- Navegación intuitiva entre el código y el modelo.
- Poderoso generador de documentación y reportes *UML PDF/HTML/MS Word*.
- Demanda en tiempo real, modelo incremental de viaje redondo y sincronización de código fuente.
- Superior entorno de modelado visual.

- Soporte completo de notaciones *UML*.

PgAdmin3

Para la gestión de la base datos se utilizará PgAdmin3, teniendo en cuenta que es una aplicación gráfica para gestionar el gestor de bases de datos PostgreSQL, siendo la más completa y popular con licencia Open Source. Está escrita en C++ usando la librería gráfica multiplataforma wxWidgets, lo que permite que se pueda usan en Linux, FreeBSD, Solaris, Mac OS X y Windows. Es capaz de gestionar versiones a partir de PostgreSQL 7.3 ejecutándose en cualquier plataforma, así como versiones comerciales de PostgreSQL como Pervasive Postgres, EnterpriseDB, Mammoth Replicator y SRA PowerGres. Incluye funcionalidades para responder a las necesidades de todos los usuarios, desde escribir consultas SQL simples hasta desarrollar bases de datos complejas. La interfaz gráfica soporta todas las características de PostgreSQL y facilita enormemente la administración.

Algunas características de pgAdmin3 son:

- Incluye una interfaz gráfica de administración, una herramienta para el trabajo con SQL, un editor de código de procedimientos y funciones, un agente para lanzar scripts programados y soporte para el motor de replicación Slony-I.
- La conexión al servidor puede hacerse mediante conexión TCP/IP o Unix Domain Sockets en el caso de plataformas *nix y puede cifrarse mediante SSL para mayor seguridad.
- Utiliza procedimientos almacenados.
- Muy rápido para la visualización y entrada de datos.

Mondrian OLAP Server

Para obtener la funcionalidad de procesamiento analítico en línea (OLAP) se utilizan otras dos aplicaciones, el servidor OLAP Mondrian, que combinado con Jpivot, permiten realizar consultas al almacén de datos y permite que los resultados sean presentados mediante un navegador de modo que el usuario pueda realizar las actividades típicas de navegación. Mondrian utiliza MDX como lenguaje de consulta, que fue un lenguaje propuesto por Microsoft. Funciona sobre las bases de datos estándar del mercado como son Oracle, DB2, SQL-Server, MySQL y otras lo cual habilita y facilita el desarrollo de negocio basado en la plataforma Pentaho.

Mondrian es una de las aplicaciones más importantes de la plataforma Pentaho BI. Es un servidor OLAP de código abierto que gestiona comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuente.

Es decir, Mondrian actúa como "JDBC para OLAP".

Funcionamiento de Mondrian OLAP Server:

1. El cliente manda una solicitud de consulta bajo la interfaz web JPivot
2. Mondrian recibe la solicitud y bajo el esquema de metadatos que definen sus conceptos multidimensionales busca si ya tiene los datos en cache respondiendo rápidamente a la petición.
3. Si los datos no se encontraron en cache ejecuta las sentencias SQL para generar los datos.
4. Se almacenan los datos recibos en cache para agilizar posteriores consultas.
5. Y finalmente se devuelve el resultado al usuario cliente a través de la interfaz.

Entre sus ventajas están:

- Agilizar la consulta de grandes cantidades de datos
- Alta velocidad de respuesta
- Permite realizar consultas al mercado de datos
- Es un motor ROLAP con cache

Pentaho BI Server

La plataforma Pentaho BI Server provee el soporte y la infraestructura necesarios para crear soluciones de inteligencia empresarial a problemas de negocios. El marco proporciona los servicios básicos, incluidos autenticación, registro, auditoría, servicios web y motor de reglas. La plataforma también incluye un motor de solución que integra reportes, análisis, tableros de comandos y componentes de minería de datos. El diseño modular y arquitectura basada en plugin permite a todos o parte de la plataforma estar inmersa en aplicaciones de terceros por los usuarios finales, así como fabricantes de equipos originales.

La aplicación Pentaho BI Server funciona como un sistema basado en administración web de informes, el servidor de integración de aplicaciones y un motor de flujo de trabajo ligero (secuencias de acción.) Está diseñado para integrarse fácilmente en cualquier proceso de negocio.

Algunas de sus ventajas son:

- Integración con procesos de negocio
- Administra y programa reportes
- Administra seguridad de usuarios

Schema Workbench

Mondrian Schema Workbench es un entorno visual para el desarrollo y prueba de cubos OLAP Mondrian. Si bien la definición del XML para esquemas Mondrian no es extremadamente compleja, en la práctica resulta engorroso recordar cada uno de los elementos junto a sus atributos y sub-elementos. Con esta aplicación, se puede configurar una conexión JDBC como el modelo físico, para luego elaborar el esquema lógico de manera simple y efectiva. Permite crear y probar los cubos OLAP visualmente para que luego el motor de Mondrian procese las solicitudes MDX con los esquemas creados. Los esquemas son modelos metadatos XML que se crean en una estructura específica utilizada por el motor de Mondrian.

Ofrece las siguientes funcionalidades:

- Editor de esquemas integrados con un origen de datos subyacente para su validación.
- Prueba de consultas MDX contra el esquema y la base de datos.
- Examinar la estructura subyacente de bases de datos.

Apache Tomcat.

Apache Tomcat es una implementación de software de código abierto de Java Servlet y tecnologías JavaServer Pages. Es desarrollado en un entorno abierto y participativo, publicado bajo la licencia Apache versión 2. Tiene la intención de ser una colaboración de los mejores desarrolladores de su clase de todo el mundo. Funciona en cualquier sistema operativo que disponga de una máquina virtual de java. Tomcat puede utilizarse como un contenedor solitario o como plugin para un servidor web existente.

En la versión 6.0.x se pueden encontrar nuevas características como:

- Conformidad de mantenimiento.
- Requiere un mínimo de JDK 1.5.
- web.xml ya no es necesario.

- Anotaciones de apoyo.
- Alternativas a algunas entradas XML en web.xml.
- Nuevo lenguaje de expresión unificada.
- Tiene su propia especificación.
- Nueva manipulación TLD.
- Inyección de recursos.
- API basado en j.u.concurrent.Executor.
- Nuevo conector Java NIO.
- Permite múltiples url-pattern en servlet-mapping.

DataCleaner

DataCleaner es una aplicación de código abierto para el perfilado, la validación y comparación de datos. Estas actividades ayudan a administrar y supervisar la calidad de los datos con el fin de garantizar que la información es útil y aplicable a su situación de negocio. Es una de las aplicaciones más fáciles de usar para la calidad de los datos. Normalmente se utiliza antes, durante y después del proceso ETL

- Antes, para profundizar en los orígenes de los datos que serán usados en el trabajo.
- Durante, en caso de existir cualquier desajustes inesperados durante el proceso de ETL.
- Después de asegurar la coherencia y la calidad en la fuente de datos que han poblado.

DataCleaner puede acceder y analizar prácticamente cualquier almacén de datos, incluyendo:

- Base datos como Oracle, Microsoft SQL Server, PostgreSQL, MySQL, OpenOffice (ODB) y más
- Archivos .csv y .tsv
- Hojas de cálculo Excel
- Archivos XML

Conclusiones

Partiendo del análisis realizado del estado del arte y de la importancia que tiene la utilización de los almacenes de datos para apoyar la toma de decisiones a partir de la información recogida de cualquier sector o empresa, se eligieron en este capítulo las herramientas y la metodología a utilizar en la solución del problema, teniendo en cuentas las características y las ventajas que cada una de ellas ofrece.

Durante la investigación detallada de las tecnologías informáticas existentes referente a los almacenes de datos, se decidió seleccionar como gestor de base de datos PostgreSQL 8.4, como herramienta CASE para el modelado se utilizará Visual Paradigm 6.4 y como metodología de desarrollo se toma como base la metodología Kimball y como complemento a la misma lo planteado por Leopoldo Zenaido Zepeda Sánchez en su Tesis de Doctorado.

CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS TURISMO.

Introducción.

En este capítulo se hará un estudio preliminar del negocio con el objetivo de definir necesidades de información, reglas del negocio, casos de uso y su descripción, identificación de dimensiones, hechos y medidas, desarrollo de la matriz BUS, modelo de datos.

2.1 Caracterización de las áreas de la organización

En Cuba la Oficina Nacional de Estadísticas (ONE) tiene como objetivo general garantizar la producción de estadísticas de calidad a través del Sistema Estadístico Nacional ejerciendo una adecuada dirección, ejecución y control de la captación de las cifras económicas y sociales, así como su adecuada difusión de acuerdo con las necesidades de la economía y las demás necesidades del país en información estadística.

La ONE está estructurada en sucursales esparcidas por todo el país, encargadas de enviar informaciones periódicamente referente al Turismo, precisamente el envío de estos reportes es uno de los principales problemas que presenta la ONE. Al no existir una forma adecuada de almacenar, recuperar y presentar la información proveniente de estas sucursales dificulta el análisis de los principales indicadores en la toma de decisiones.

Para un mejor análisis del área de Turismo ésta se divide en otras dos áreas: "*Turismo extranjero*" y "*Turismo nacional*" para poder analizar los indicadores de cada uno de ellas por separados.

2.2 Reglas del negocio

Las reglas de negocio describen políticas o condiciones que deben cumplirse. El proceso de especificación implica que hay que identificarlas dentro del negocio y aplicarlas a la solución propuesta.

1. Los identificadores de los indicadores no pueden estar repetidos.
2. Los identificadores de los indicadores no pueden ser nulos.
3. Los identificadores de las dimensiones no pueden ser nulos.

4. El total de ingresos debe comprender el total de ingresos de cada cadena y de cada actividad y con un valor mayor que cero.

5. Los tipos de cadena estarán definidos de la siguiente manera: sociedad mercantil, empresa mixta, campismo y poder popular.

6. Los tipos de actividad estarán definidos de la siguiente manera: hotelera, extra hotelera y apoyo al turismo.

7. En una misma dimensión no pueden existir identificadores repetidos.

2.3 Especificación de requerimientos

Las necesidades de información para los especialistas de la ONE son especificaciones que ellos precisan para darle cumplimiento a sus tareas internas. Con el objetivo de controlar estas necesidades se han definido una serie de medidas e indicadores que muestran el comportamiento de estas. Se identifican las necesidades de la organización y los requisitos a través del análisis de los objetivos y de los indicadores del área.

Entregable: Especificación de requerimientos.

2.3.1 Requisitos de información

Los datos estadísticos del área de Turismo se recogen en dos modelos, Modelo 1398 y Modelo 1394, y para un mayor entendimiento de los pedidos de información se identifican los requisitos de información por separado para cada modelo.

Modelo 1398

1. R1. Obtener el total de Ingresos por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
2. R2. Obtener el total de Ingresos turísticos por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
3. R3. Obtener el total de Ingresos financiados por entidades por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.

4. R4. Obtener el total de Ingresos autofinanciados por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
5. R5. Obtener el total de Costos y Gastos por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
6. R6. Obtener el total de Costos y Gastos en luz, agua y fuerza por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
7. R7. Obtener el total de Costos y Gastos en salarios por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
8. R8. Obtener el total de Costos y Gastos en depreciación por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
9. R9. Obtener el total de Utilidad antes del impuesto por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
10. R10. Obtener el total de Ventas de paquetes por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
11. R11. Obtener el total de Ventas de opcionales por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
12. R12. Obtener el total de Cuentas por Cobrar por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
13. R13. Obtener el total de Cuentas por Cobrar vencidas por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
14. R14. Obtener el total de Cuentas por Pagar por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
15. R15. Obtener el total de Inventarios por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
16. R16. Obtener el total de Salario devengado por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
17. R17. Obtener el Promedio de trabajadores por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
18. R18. Obtener el total de Turistas físicos por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
19. R19. Obtener el total de Turistas físicos extranjeros por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.

20. R20. Obtener el total de Turistas físicos nacionales por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
21. R21. Obtener el total de Turistas por días por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
22. R22. Obtener el total de Turistas por días extranjeros por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
23. R23. Obtener el total de Turistas por días extranjeros por países en el tiempo
24. R24. Obtener el total de Turistas por días nacionales por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
25. R25. Obtener el total de Habitaciones existentes por días por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
26. R26. Obtener el total de Habitaciones ocupadas por días por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
27. R27. Obtener el total de Habitaciones diarias ocupadas por extranjeros por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
28. R28. Obtener el total de Habitaciones diarias ocupadas por turismo nacional por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
29. R29. Obtener el total de Habitaciones diarias disponibles por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
30. R30. Obtener el total de turistas financiados por entidades por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
31. R31. Obtener el total de eventos de interés nacional por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
32. R32. Obtener el total de cubanos acreditados a medios de prensa extranjera por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
33. R33. Obtener el total de cubanos autofinanciados moneda nacional (CUP) por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
34. R34. Obtener el total de cubanos autofinanciados moneda convertible (CUC) por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.
35. R35. Obtener el total de Habitaciones físicas por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.

36. R36. Obtener el total de Plazas camas por actividad, tipo de entidad, cadena, centro informante y DPA en el tiempo.

Modelo 1394

1. R1. Obtener el total de ingresos provenientes del servicio de Alojamiento por cadena en el tiempo.
2. R2. Obtener el total de ingresos a través de Comercio Minorista por cadena en el tiempo.
3. R3. Obtener el total de ingresos provenientes de Gastronomía por cadena en el tiempo.
4. R4. Obtener el total de ingresos por medio de Transporte por cadena en el tiempo.
5. R5. Obtener el total de ingresos por medio de Recreación por cadena en el tiempo.
6. R6. Obtener el total de ingresos provenientes de otras vías por cadena en el tiempo.
7. R7. Obtener el total de Plazas diarias ocupadas por cadena en el tiempo.
8. R8. Obtener el total de Plazas diarias ocupadas por turismo nacional por cadena en el tiempo.
9. R9. Obtener el total de Habitaciones existentes diarias por cadena en el tiempo.
10. R10. Obtener el total de Habitaciones diarias nacionales existentes diarias por cadena en el tiempo.
11. R11. Obtener el total de Habitaciones ocupadas por cadena en el tiempo.
12. R12. Obtener el total de Habitaciones ocupadas por turismo nacional por cadena en el tiempo.
13. R13. Obtener el total de Habitaciones disponibles por cadena en el tiempo.
14. R14. Obtener el total de Habitaciones físicas por cadena en el tiempo.
15. R15. Obtener el total de Huéspedes por cadena en el tiempo.
16. R16. Obtener el total de Huéspedes nacionales por cadena en el tiempo.
17. R17. Obtener el total de Plazas camas por cadena en el tiempo.
18. R18. Obtener la Suma de control por cadena en el tiempo.

2.3.2 Requisitos funcionales

- RF1. Extraer datos del DBF del modelo 1398 y del Excel del 1394 (fuente de datos).
- RF2. Transformar y Cargar datos del DBF del modelo 1398 y del Excel del 1394 (fuente de datos).
- RF3. Autenticar usuario.
- RF4. Insertar usuario.
- RF5. Eliminar usuario.

- RF6. Insertar roles.
- RF7. Eliminar roles.
- RF8. Insertar reportes.
- RF9. Visualizar reportes.
- RF10. Eliminar reportes.
- RF11. Visualizar gráfica.
- RF12. Configurar gráfico.
- RF13. Exportar reporte a pdf (formato de documento portátil).
- RF14. Exportar reporte a excel.

2.3.3 Requisitos no funcionales

Requisitos no funcionales de usabilidad

- El tiempo de entrenamiento requerido para que usuarios normales y avanzados sean productivos operando el sistema será de una semana.
- El sistema solo será manejado por usuarios autorizados (2 a 5 Usuarios).

Requisitos no funcionales de fiabilidad

- El sistema será accedido para su mantenimiento una vez al mes.
- Debe estar disponible el 100% del tiempo mientras no se le estén aplicando cambios y mientras esté fuera de mantenimiento.
- La cantidad de errores en el proceso de integración define la calidad de los datos que se están almacenando, es por ello que es crítico, definiéndose así 0 errores/puntos de función.
- El tiempo medio entre fallos es de aproximadamente 6 días.
- El tiempo medio de reparación depende fundamentalmente de la magnitud del fallo pero se estima que como promedio sea de 24 horas.

Requisitos no funcionales de eficiencia

- El sistema deberá permitir 5 usuarios conectados sincrónicamente sin que se afecte el tiempo de respuesta.
- En el proceso de integración solo tendrá conectado un usuario que tendrá la tarea de vigilar el proceso de integración de datos.
- El lenguaje de programación del proceso de integración de la base de datos será SQL, desarrollado en PostgreSQL 8.4.

Requisitos no funcionales de hardware

- La computadora donde será instalado el servidor debe tener como mínimo 60 gigabytes de disco duro, 512 megabytes y un micro de 2.20 GHz.

2.4 Modelo de casos de usos del sistema

2.4.1 Actores del sistema

El sistema contará con tres actores:

1. Analista, quien será el encargado de analizar las necesidades de información.
2. Administrador de ETL, quien llevará a cabo las funciones ETL.
3. Administrador del sistema, que incluye las funciones del analista y además es el encargado de gestionar los usuarios, roles, permisos y reportes.

2.4.2 Diagrama de casos de uso del sistema

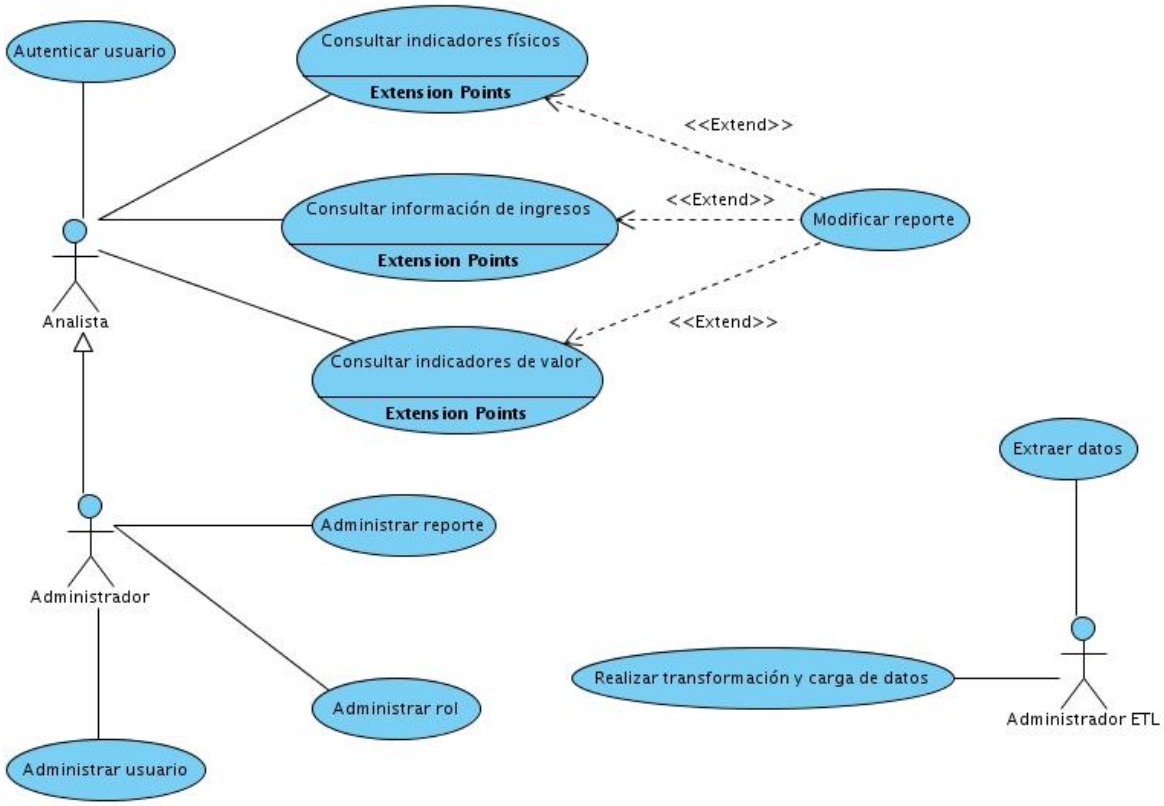


Figura 1: Diagrama de Casos de Uso

2.4.3 Especificación de casos de uso del sistema

Se identificaron 10 casos de uso:

1. Autenticar usuario: Realiza la autenticación de los usuarios en el sistema.
2. Administrar reporte: Elimina, inserta y modifica los reportes que se visualizan.
3. Administrar rol: Elimina, inserta y modifica los roles que interactúan con el sistema.
4. Administrar usuario: Gestiona los usuarios del sistema.
5. Realizar transformación y carga de datos: Realiza la carga y transformación de los datos.
6. Extraer datos: Realiza la extracción de los datos.
7. Consultar indicadores físicos: Analiza la información referente a los indicadores físicos.
8. Consultar indicadores de valor: Analiza la información referente a los indicadores de valor.
9. Consultar información de ingresos: Analiza la información referente a los ingresos.
10. Modificar reporte: Modifica los reportes.

2.5 Especificación del modelo dimensional

Dentro del Modelo de Datos se encuentra el "*modelo dimensional*", que tiene como elementos fundamentales los hechos y dimensiones. Los hechos serían aquellos datos que proporcionan una información cuantitativa sobre los indicadores del negocio en análisis. Su finalidad es proporcionar información necesaria para la gestión, facilitando el conocimiento del negocio o proceso a modelar, y fundamentar la toma de decisiones. Las dimensiones buscan determinar un contexto para el análisis de los Hechos. Se trata de grupos homogéneos de elementos, en muchas ocasiones, jerarquizados. Su papel es promocionar la información contenida en los hechos. [8]

2.5.1 Tablas de hechos

Al analizar el modelo se definieron tres tablas de hechos:

1. Hecho indicadores preliminares del Turismo (hecho_indicadores_preliminares_turismo): en él se recogen los datos de los indicadores del turismo.

2. Hecho turistas por países (hecho_turistas_paises): en él se recoge la cantidad de turistas por países.

3. Hecho ingresos, costos y gastos (hecho_ingresos_costos_y_gastos): en él se recogen los datos de los indicadores seleccionados de la actividad turística en el territorio.

Medidas

- Real_anno_actual_mes: Se refiere al valor real en el mes que se realiza el análisis.
- Real_anno_actual_acumulado: Se refiere al acumulado total de todo el año.
- Real_anno_anterior: Se refiere al acumulado total de todo el año anterior.
- Cantidad_turistas: Se refiere al total de turistas que entra al país en la fecha en cuestión.
- Plan: Se refiere al total de turistas que se esperaba para la fecha en cuestión.
- Anno_anterior: Se refiere al total de turistas que entró al país en el año anterior.
- Real_anno_actual: Indica el ingreso total en ese año en moneda nacional.
- Real_anno_actual_divisa: Indica el ingreso total en ese año en divisa.
- Real_anno_anterior: Indica el ingreso total en el año anterior en moneda nacional.
- Real_anno_anterior_divisa: Indica el ingreso total en el año anterior en divisa.
- Plan_anno_actual. Indica el ingreso total previsto en el año en moneda nacional.
- Plan_anno_actual_divisa: Indica el ingreso total previsto en el año en divisa.

2.5.2 Tablas de dimensiones

Se definieron ocho tablas de dimensiones:

1. Dimensión país (dim_pais): en esta se registran los países al que pertenecen los turistas extranjeros.

2. Dimensión DPA (dim_dpa): en ella se registra la división política administrativa.

3. Dimensión tipo de actividad (dim_tipo_actividad): en ella se registran los tipos de actividad turística.

4. Dimensión tipo de entidad (dim_tipo_entidad): en ella se registran los tipos de entidades.

5. Dimensión entidad (dim_entidad): en ella se registran los centros informantes de los cuales pertenecen los datos.

6. Dimensión cadena (dim_cadena_turistica): en ella se registran las cadenas de los cuales pertenecen los datos.

7. Dimensión temporal mes (dim_temporal_mes): Es una dimensión temporal donde se registran los meses

8. Dimensión indicador (dim_indicador_turismo): Es la dimensión que contiene todos los indicadores del Turismo.

2.5.3 Matriz Bus

La Matriz Bus es la representación de la relación que existe entre los hechos y las dimensiones. Se define como la habilidad para describir y seguir la vida tanto de una dimensión como de un hecho, la cual permite determinar el impacto que provocaría un cambio durante el desarrollo del sistema.

Hechos/Dimensiones	temporal_mes	pais	cadena_turistica	tipo_entidad	entidad	tipo_actividad	dpa	indicador_turismo
Hech_indicadores_preliminares_turismo	X		X					X
Hech_ingresos_costos_gastos	X		X	X	X	X	X	X
Hech_turistas_paises	X	X	X	X	X	X	X	

Figura 2: Matriz Bus

2.5.4 Modelo dimensional

El modelo está compuesto por tres tablas de hechos y ocho tablas de dimensiones, las dimensiones no se ramifican y los hechos solo se relacionan con las dimensiones que intervienen en sus respectivos temas de análisis, dimensiones que en algunos casos son comunes o compartidas. Por lo que el modelo dimensional es una constelación de hechos.

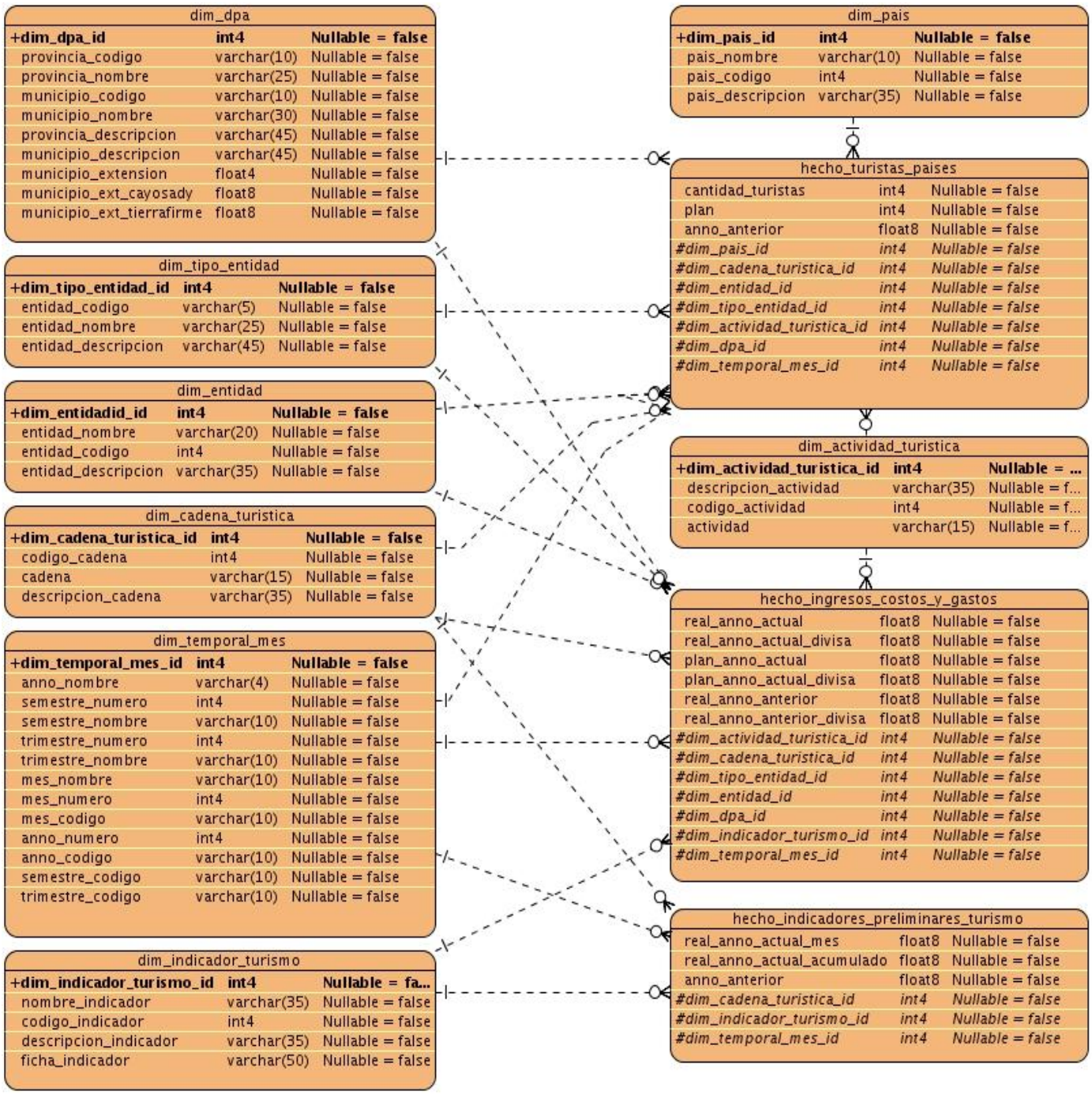


Figura 3: Modelo dimensional.

2.6 Especificación del modelo físico

El modelo físico contiene once tablas dentro de las cuales ocho son dimensiones y tres son tablas de hecho.

Para la creación de las llaves primarias se utilizó el estándar de definición dim_DIMENSION_id. El tipo de datos de la llave primaria de cada tabla es ENTERO, siendo este no nulo. Para una mejor distribución de las tablas se tienen dos esquemas:

1. **dimensiones:** en este esquema se encuentran todas las dimensiones identificadas para la realización del mercado de datos.

2. **hech_turismo:** este esquema recoge todos lo hechos identificados para la construcción del mercado de datos.

Descripción de la estructura Física

No.	Nombre de la Tabla	Descripción	Llave Primaria
1	dim_temporal_mes	Dimensión temporal mes	dim_temporal_mes_id
2	dim_pais	Dimensión país	dim_pais_id
3	dim_cadena_turistica	Dimensión cadena	dim_cadena_turistica _id
4	dim_tipo_entidad	Dimensión tipo de entidad	dim_tipo_entidad_id
5	dim_entidad	Dimensión centro informante	dim_entidad_id
6	dim_actividad_turistica	Dimensión actividad	dim_actividad_turistica _id
7	dim_dpa	Dimensión DPA	dim_dpa_id

8	dim_indicador_turismo	Dimensión indicador	dim_indicador_turismo _id
9	hech_indicadores_preliminares_turismo	Hecho indicadores preliminares del turismo	hech_indicadores_preliminares_turismo_id
10	hech_turistas_paises	Hecho turistas por países	hech_turistas_paises_id
11	hech_ingresos_costos_y_gastos	Hecho ingresos, costos y gastos	hech_ingresos_costos_y_gastos_id

Modelo físico.

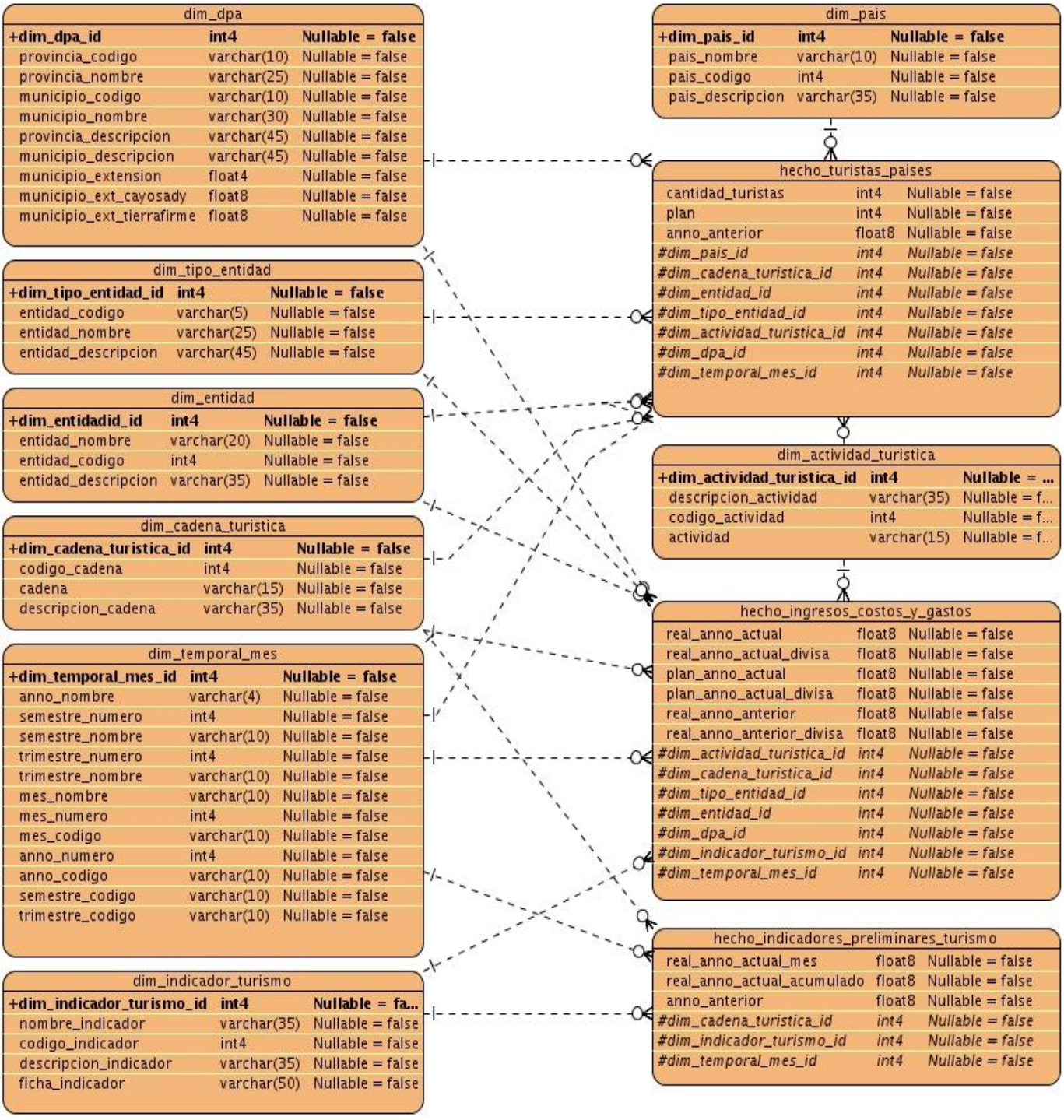
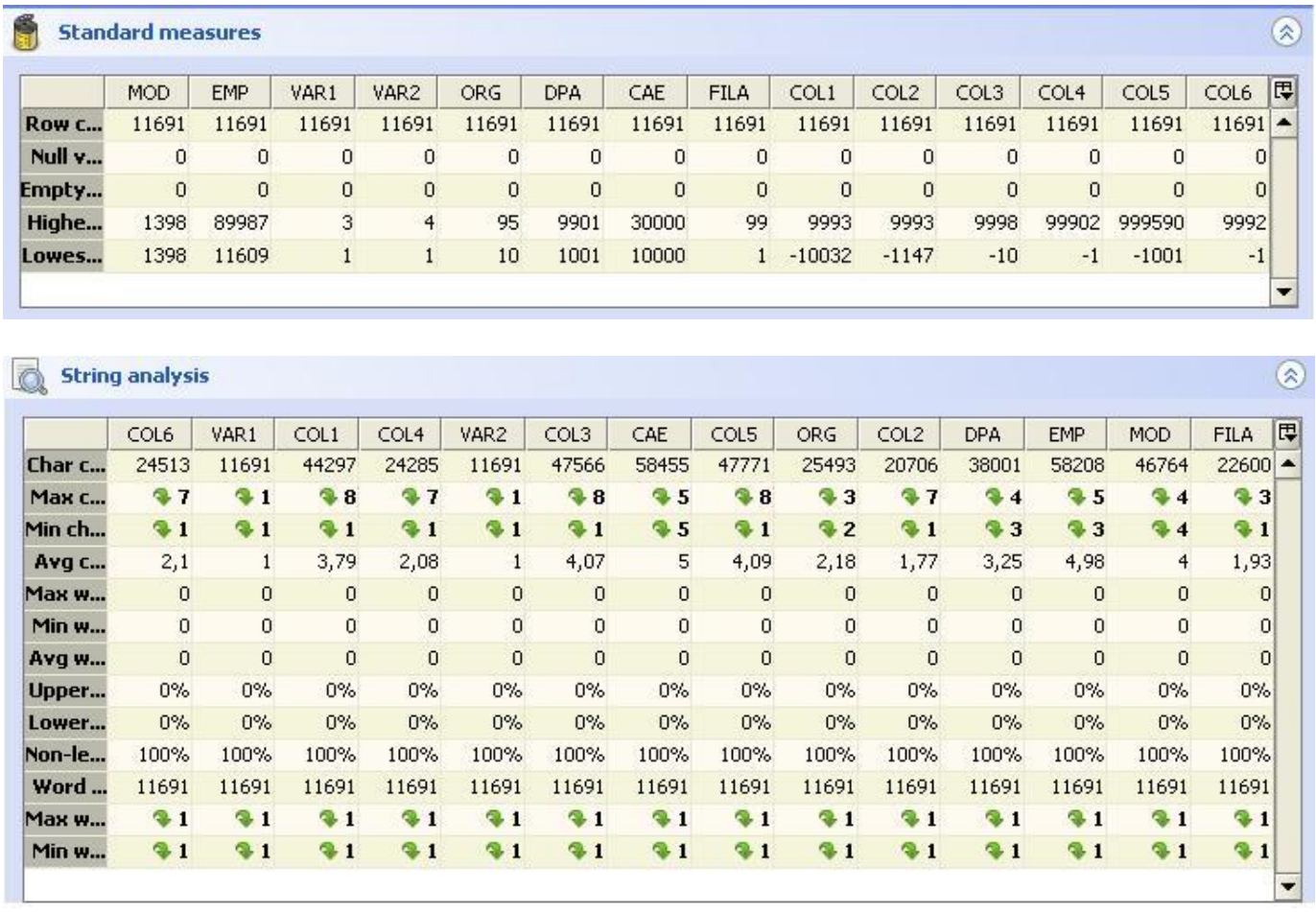


Figura 4: Modelo físico.

2.7 Perfilado de datos

El perfilado de datos se realiza con el objetivo de conocer el estado en que se encuentran los datos que próximamente se extraerán en el proceso ETL, así como administrar y supervisar la calidad de los mismos con el fin de garantizar que la información es útil y aplicable a su situación de negocio. En este caso para el análisis de los datos se generaron los reportes Estándares de medidas, Análisis de cadenas y Análisis numéricos, de los cuáles se arribó a la conclusión de que no se encontraron valores nulos, los tipos de datos encontrados fueron varchar y entero, y existen valores duplicados y únicos.



Standard measures

	MOD	EMP	VAR1	VAR2	ORG	DPA	CAE	FILA	COL1	COL2	COL3	COL4	COL5	COL6
Row c...	11691	11691	11691	11691	11691	11691	11691	11691	11691	11691	11691	11691	11691	11691
Null v...	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Empty...	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Highe...	1398	89987	3	4	95	9901	30000	99	9993	9993	9998	99902	999590	9992
Lowes...	1398	11609	1	1	10	1001	10000	1	-10032	-1147	-10	-1	-1001	-1

String analysis

	COL6	VAR1	COL1	COL4	VAR2	COL3	CAE	COL5	ORG	COL2	DPA	EMP	MOD	FILA
Char c...	24513	11691	44297	24285	11691	47566	58455	47771	25493	20706	38001	58208	46764	22600
Max c...	7	1	8	7	1	8	5	8	3	7	4	5	4	3
Min ch...	1	1	1	1	1	1	5	1	2	1	3	3	4	1
Avg c...	2,1	1	3,79	2,08	1	4,07	5	4,09	2,18	1,77	3,25	4,98	4	1,93
Max w...	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Min w...	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Avg w...	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Upper...	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Lower...	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%	0%
Non-le...	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Word ...	11691	11691	11691	11691	11691	11691	11691	11691	11691	11691	11691	11691	11691	11691
Max w...	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Min w...	1	1	1	1	1	1	1	1	1	1	1	1	1	1

Figura 5: Imágenes del perfilado de datos

Conclusiones

En este capítulo se definieron los principales procesos que se llevan a cabo en el desarrollo del análisis y diseño del mercado de datos de comunicaciones en los cuales se determinaron y justificaron 54 requisitos de información, 14 funcionales y 10 no funcionales. Se diseñó el modelo de casos de uso del sistema donde se identificaron los actores y casos de uso, así como los roles y permisos. También se especificaron 8 dimensiones y 3 hechos por los cuales está compuesto el modelo dimensional.

CAPÍTULO 3: IMPLEMENTACIÓN DEL MERCADO DE DATOS TURISMO.

Introducción

En este capítulo se procederá a la implementación de la solución propuesta, se implementará el modelo de datos, el modelo de despliegue, los flujos de transformación y los trabajos. Se implementará la base de datos y se realizará el montaje de los clasificadores para el mercado de datos.

3.1 Implementación del modelo de datos

Para el desarrollo del mercado de datos Turismo y para lograr una mejor organización del mismo se definieron dos esquemas: ***dimensiones*** que contiene todas las tablas de dimensiones comunes del almacén y ***mart_turismo*** que está integrado por las dimensiones y hechos que son particulares del mercado de datos Turismo. El modelo de datos cuenta en total con ocho tablas de dimensiones y tres tablas de hechos. La siguiente tabla describe dicha estructura.

Esquema	Tabla
dimensiones	dim_dpa
dimensiones	dim_pais
dimensiones	dim_entidad
dimensiones	dim_temporal_mes
mart_turismo	dim_tipo_actividad
mart_turismo	dim_cadena_turistica
mart_turismo	dim_actividad_turistica
mart_turismo	dim_indicador_turismo
mart_turismo	hecho_turistas_paises
mart_turismo	hecho_ingresos_costos_y_gastos
mart_turismo	hecho_indicadores_preliminares_turismo

3.2 Arquitectura de integración

Una arquitectura en el ámbito computacional es un conjunto de estructuras o reglas que proveen un esqueleto para el diseño general de un producto o sistemas. En el proceso de Integración de datos no es recomendable iniciar el desarrollo de una solución sin haberla premeditado previamente, identificar sus fuentes, su esquema, el movimiento de los datos y determinado su enfoque de almacenamiento de datos.

La integración de sistemas básicos puede dividirse en tres grupos: origen de datos, gestión de datos y presentación. (9) Origen de datos es la arquitectura estándar para el proceso ETL y fue el que se tomó como base para la integración de los datos, se basa en la periódica extracción de los datos de origen, realizarle a éstos las transformaciones que requieran e integrarlos en la fuente de destino. El proceso de integración de datos resuelve el problema que representan los datos no válidos y nulos, permitiendo proporcionar al usuario una vista unificada de los datos. (10)

Tomando como base el nivel de detalle que requiere el proceso de Integración de datos por su complejidad, la arquitectura que se muestra en la siguiente figura se utilizó para el desarrollo de la solución.

Figura 6: Arquitectura de integración

3.3 Implementación de los procesos ETL

En este epígrafe se describirá el proceso de Extracción, Transformación y Carga, detallando en cada subproceso sus características.

3.3.1 Extracción de los datos

La extracción de los datos es el punto de partida del Proceso de Extracción, Transformación y Carga, inicia al extraer los datos de los sistemas de origen. El formato de los datos correspondientes al área de Turismo se encuentran en formato DBF y Excel. Los datos en formatos Excel fueron estandarizados para iniciar el proceso de transformación. Para esto se tomaron como base los datos provenientes de la Oficina Nacional de Estadísticas, en el caso de las dimensiones se tomó como fuente los clasificadores establecidos por dicha entidad y para los hechos se utilizaron los datos referentes al tema de análisis enviados por los analistas.

3.3.2 Transformación y carga de los datos.

Transformaciones para cargar las tablas de dimensiones.

No fue necesario aplicarles transformaciones a los datos, el proceso ETL fue extraer y cargar los datos de los clasificadores que se encontraban en formato Excel.

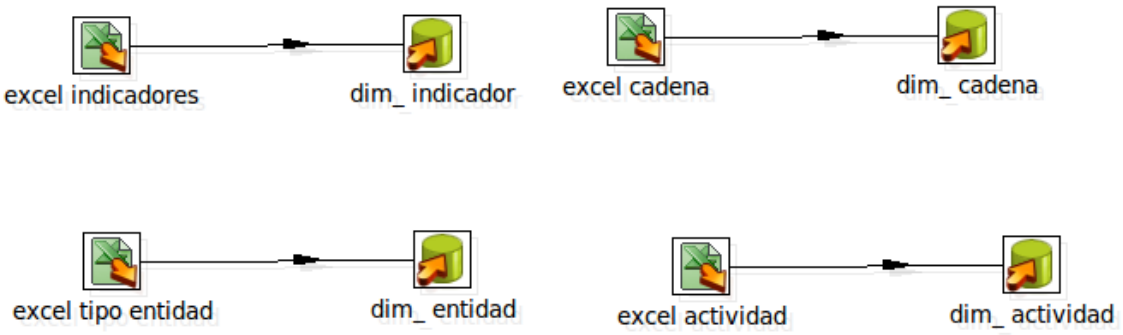


Figura 7: Transformación para cargar las tablas de dimensiones.

Para poder cargar éstas dimensiones de forma programada se realiza el trabajo: **trabajo_dimensiones.kjb.**

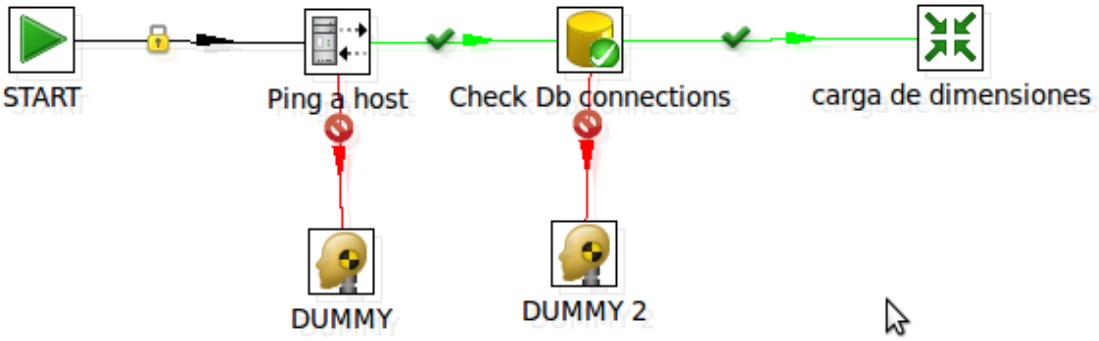


Figura 8: Trabajo para cargar las dimensiones.

Transformación para cargar el hecho hecho_indicadores_preliminares_turismo.

Para poblar el hecho hecho_indicadores_preliminares_turismo, a diferencia de las tablas de dimensiones, fue necesario realizarle transformaciones a los datos, inicialmente se validaron los valores no posibles, que no son más que errores en los datos de origen, luego se seleccionaron y renombraron aquellos que serán utilizados más adelante en el proceso de trasformación. Se generan

los valores necesarios para cargar la fecha a la que pertenecen los datos, se hace una búsqueda en base de datos de los identificadores de las dimensiones que componen el hecho a partir de los códigos provenientes de la fuente, se analizan los valores nulos, siendo éstos enviados junto con los valores no posibles para un Excel para luego ser analizados por especialistas del tema y los datos correctos se cargan en el hecho.

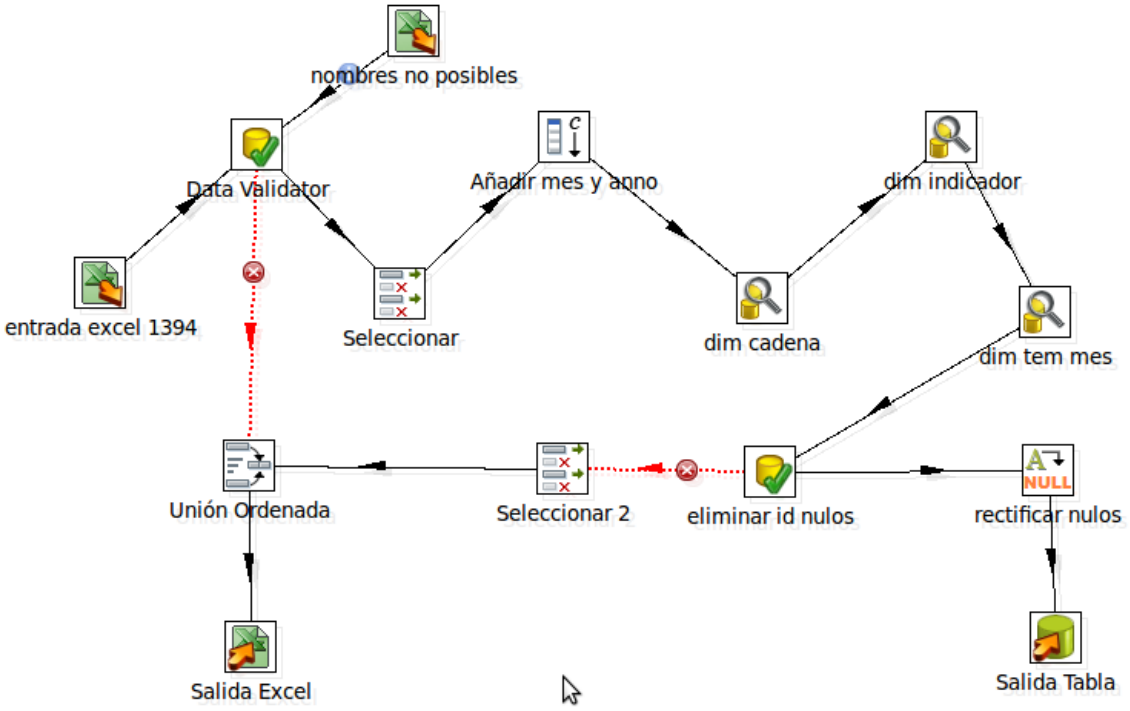


Figura 9: Transformación para cargar el hecho hecho_indicadores_preliminares_turismo.

Transformaciones para cargar los hechos hecho_ingresos_costos_y_gastos y hecho_turistas_paises.

El proceso ETL para estos dos hechos es similar, se modifican los códigos de las dimensiones dim_dpa y dim_entidad, se realiza la búsqueda en base de datos para obtener los identificadores de las dimensiones, se seleccionan y validan los datos que se van a cargar y finalmente se realiza la carga de los hechos.

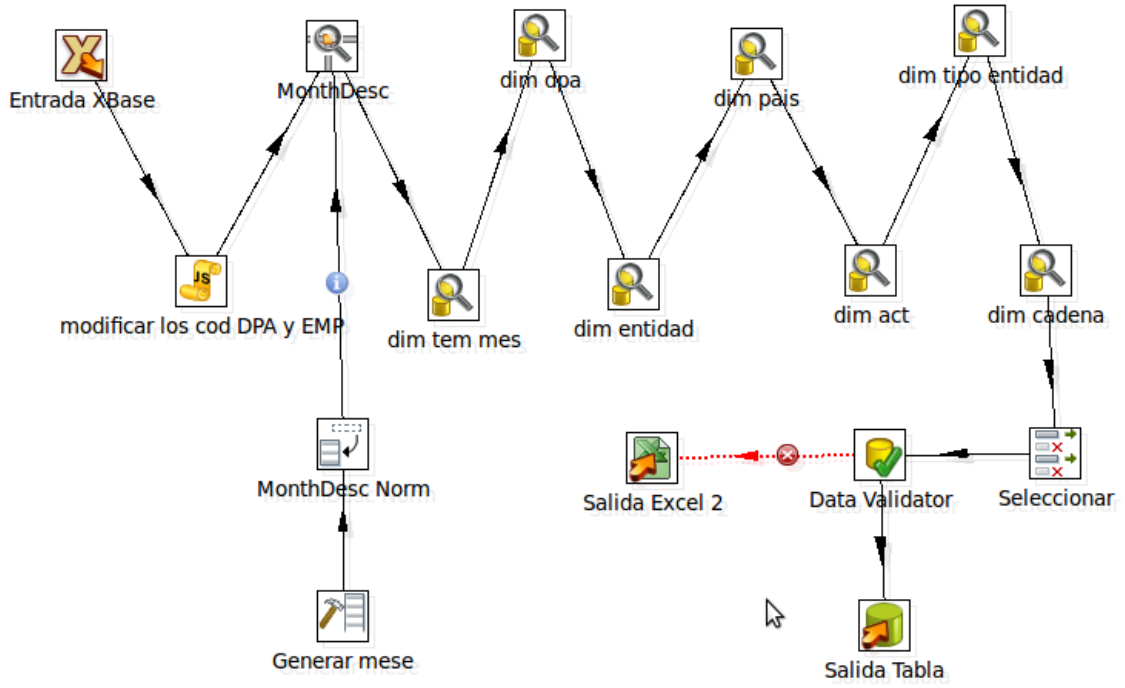


Figura 10: Transformación para cargar el hecho hecho_turistas_paises.

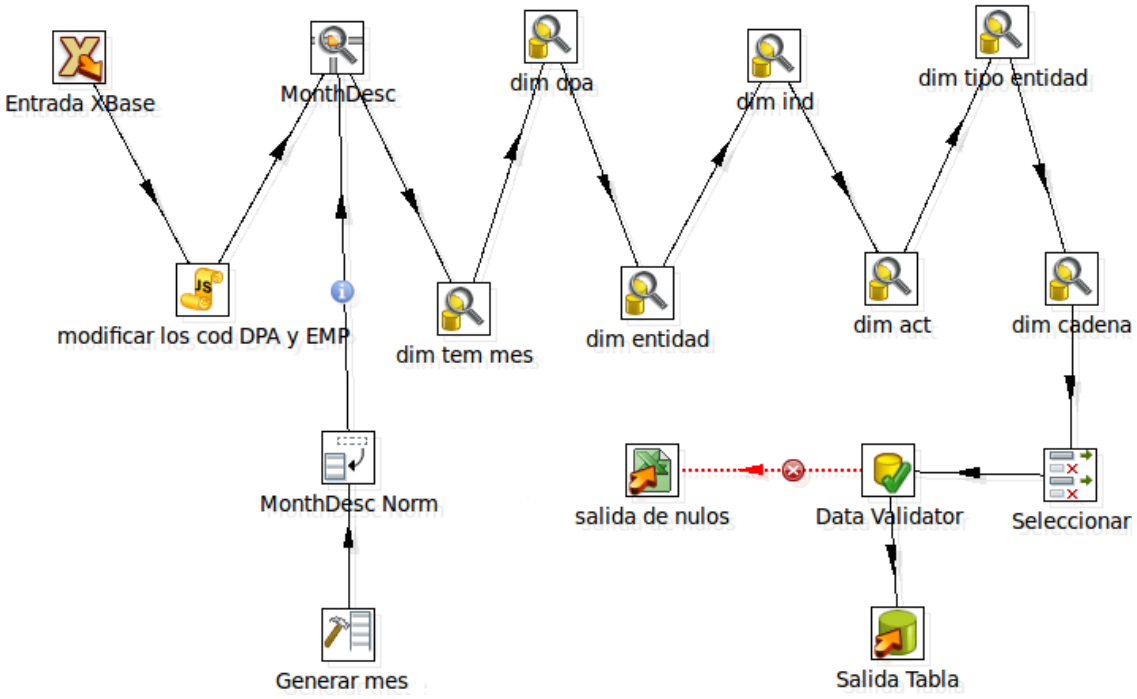


Figura 11: Transformación para cargar el hecho hecho_ingresos_costos_y_gastos.

La carga de los tres hechos se unifican a través del trabajo: **carga_hechos_turismo.**

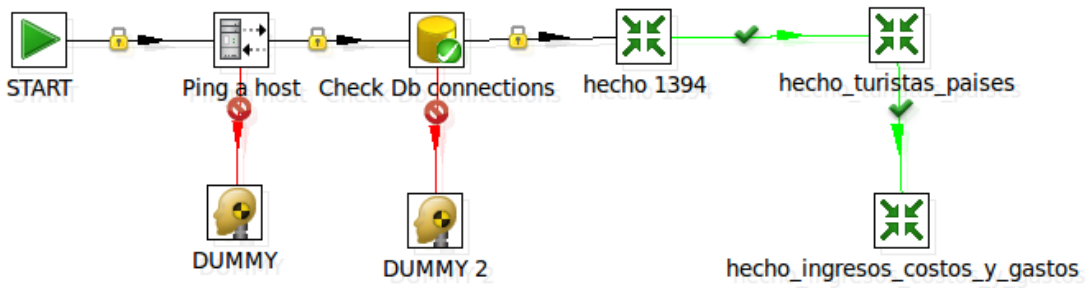


Figura 12: Trabajo para cargar el hecho hecho_ingresos_costos_y_gastos

3.4 Implementación de subsistema de visualización

3.4.1 Cubos OLAP

Para implementar el subsistema de visualización se hizo necesaria la creación de los cubos OLAP en cada uno de los cuales se definieron las dimensiones, las medidas y los niveles de jerarquía de cada dimensión. Se utilizó la herramienta Pentaho Schema Workbench en la cual se creó el esquema Turismo. Se definió crear un cubo por cada tabla de hecho y en cada uno se incluyeron las medidas y dimensiones relacionadas con dicha tabla.

Diseño de los cubos utilizando Pentaho Schema Workbench

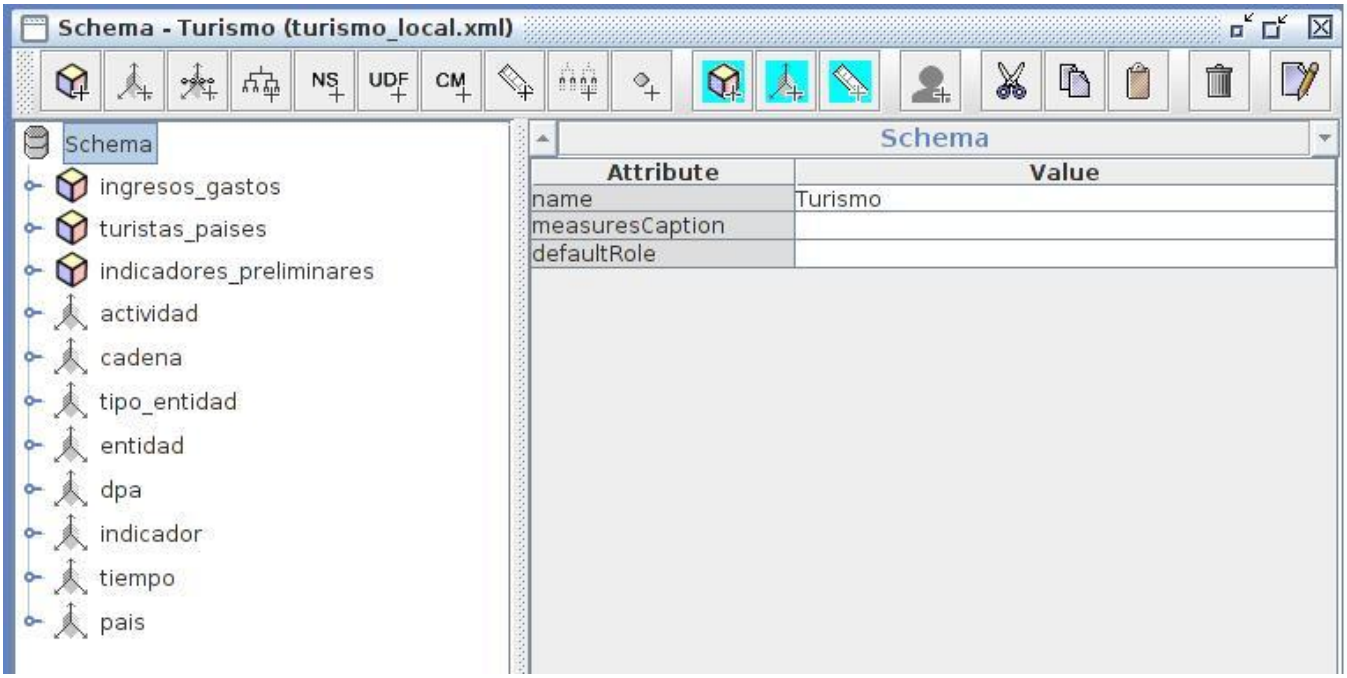


Figura 13: Diseño general del esquema.

Elementos que componen el cubo ingresos_gastos

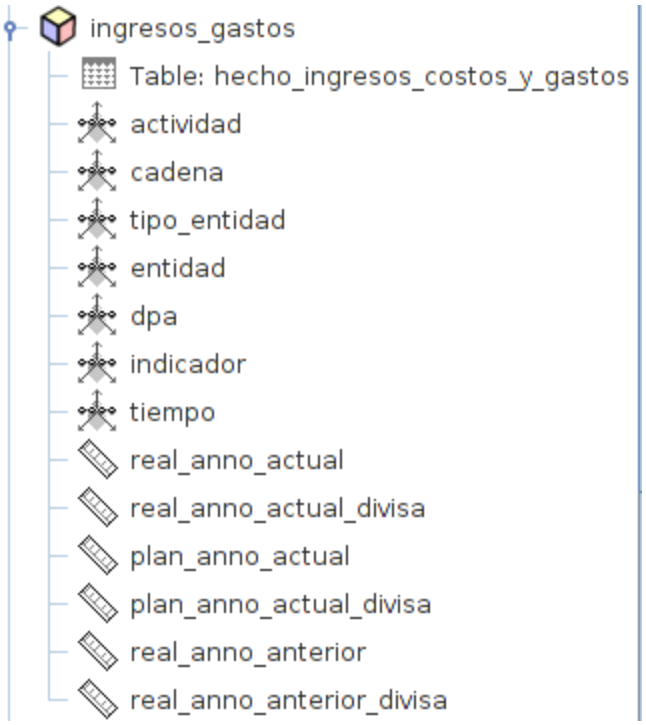


Figura 14: Diseño del cubo ingresos_gastos.

Elementos que componen el cubo indicadores_preliminares

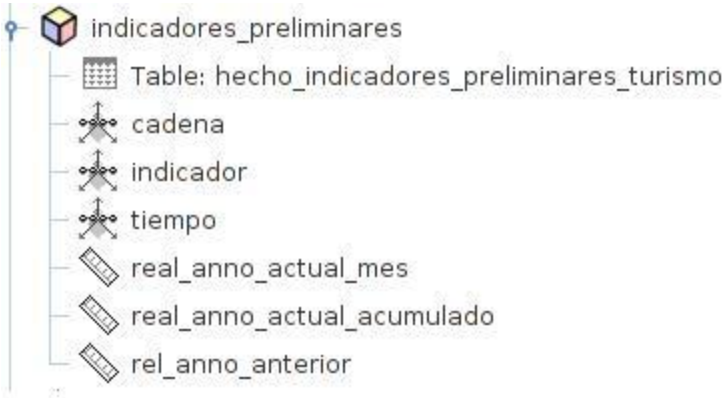


Figura 15: Diseño del cubo indicadores_preliminares.

Elementos que componen el cubo turistas_paises

Figura 16: Diseño del cubo turistas_paises.

3.4.2 Arquitectura de información

La estructura de navegación de la capa de visualización está organizada de la siguiente manera: un área de análisis general, un área de análisis específica para el área de Turismo y tres libros de trabajo incluidos en esta última área de análisis.

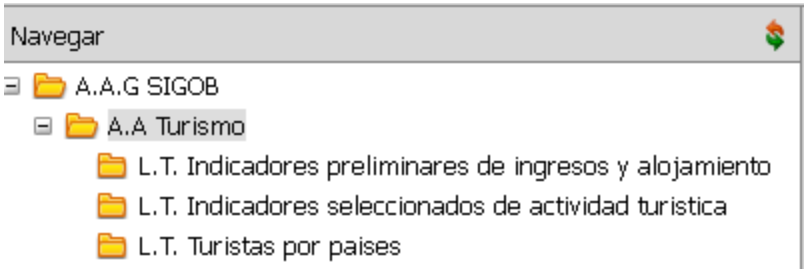


Figura 17: Mapa de navegación.

Descripción del Área de General (AAG)

- **AAG SIGOB:** Agrupa toda la información de todos los mercados de datos de la ONE, formando el almacén SIGOB.

Descripción del Área de Análisis (AA)

- **AA Turismo:** Agrupa toda la información referente a los indicadores del área de Turismo y contiene los libros de trabajo donde se encuentran los reportes.

Descripción de los Libros de Trabajo (LT)

- **LT Turistas por países:** Agrupa todos los reportes referentes a los turistas, como cantidad de turistas por países.

- **LT Indicadores preliminares de ingresos y alojamiento**: Agrupa todos lo reportes referentes a la información de los indicadores preliminares.

- **LT Indicadores seleccionados de actividad turística:** Agrupa todos los reportes referentes a la información del resto de los indicadores, como son datos de habitaciones, salarios, huéspedes nacionales y extranjeros.

3.4.3 Creación y administración de los reportes.

La visualización de los reportes se realiza a través de consultas mdx y son administradas por el Pentaho BI Server. Se crearon 83 reportes en total. A continuación se muestra el entorno de trabajo de esta herramienta y como se muestran los datos después de ejecutado un reporte.

		Tiempo		
		2009		
		Medidas		
Indicadores	DPA	Plan año actual	Real año actual	Real año anterior
Total de turistas fisicos	todas	56.895.950	54.432.680	60.276.660
Total de turistas fisicos extranjeros	todas	35.693.050	32.045.580	32.300.870
Total de turistas fisicos nacionales	todas	21.129.900	22.095.040	27.684.710
Total de turistas por dias	todas	240.574.610	235.653.520	271.830.680
Total de turistas por dias extranjeros	todas	171.103.560	161.501.330	162.465.130
Total de turistas por dias nacionales	todas	69.395.830	73.465.760	108.834.190

Figura 18: Imagen de los reportes.

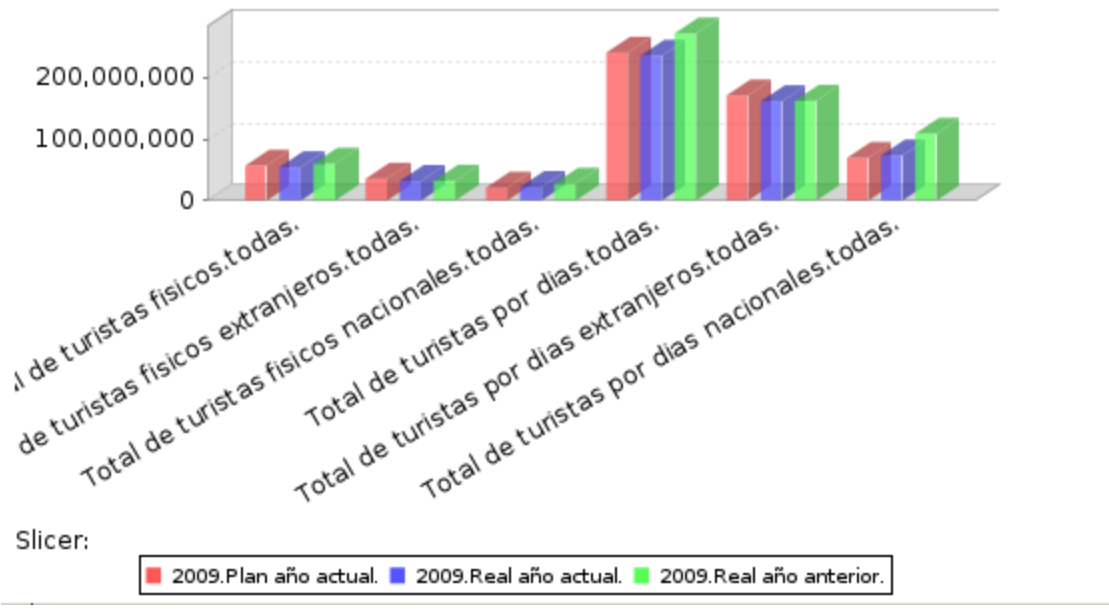


Figura 19: Reporte en forma de gráfica.

3.4.4 Configuración de la seguridad de los usuarios

Durante la implementación del subsistema de visualización del mercado de datos Turismo se crearon dos usuarios y roles los cuales tienen diferentes permisos de acceso a la información, proporcionando una mayor seguridad al sistema.

- El rol de administrador tiene todos los permisos de la aplicación y posee el usuario administrador del sistema.
- El rol de analista tiene permiso de solo lectura y posee el usuario analista del sistema.

Conclusiones

En el presente capítulo se describieron los elementos de implementación para la construcción del mercado de datos Turismo. Quedó definida la estructura de los datos a partir del modelo físico, contando con dos esquemas: dimensiones y mart_turismo, se realizó el proceso ETL, se implementaron los cubos OLAP, quedando definidos tres cubos y ocho dimensiones, se desarrollaron los subsistemas de visualización determinándose dos libros de trabajos y finalmente se implementaron todos los requisitos de información.

CAPÍTULO 4: VALIDACIÓN DEL MERCADO DE DATOS TURISMO

Introducción

Una vez realizado el análisis, diseño e implementación del mercado de datos Turismo, se da paso a la validación y prueba de la solución mediante las listas de chequeo, los casos de prueba y las pruebas de aceptación para así verificar que el sistema cumpla con los requerimientos necesarios que garanticen al usuario final del sistema la confiabilidad de los datos cargados en el almacén de datos.

4.1 Listas de chequeo.

Se entiende por lista de chequeo a un listado de preguntas, en forma de cuestionario que sirve para verificar el grado de cumplimiento de determinas reglas establecidas a priori con un fin determinado.

El uso de estas listas está generalizado en elementos muy diversos que van desde verificar y determinar el potencial de los artefactos del proceso ETL hasta medir la confiabilidad y seguridad de los datos que han sido cargados. La lista de chequeo consta de tres elementos fundamentales:

- Estructura del documento: abarca todos los aspectos definidos por el expediente de proyecto o el formato establecido por el proyecto.
- Indicadores definidos: abarca todos los indicadores a evaluar durante la etapa de desarrollo del mercado.
- Semántica del documento**:** contempla todos los indicadores a evaluar respecto a la ortografía, redacción y demás.

La lista de chequeo que se utilizó fue:

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**

crítico	¿Los entregables contienen las secciones obligatorias de la plantilla estándar definidas para un expediente de proyecto? (portada, control de versiones, reglas de confidencialidad, tabla de contenidos y contenido) (Ver expediente de proyecto)				
Indicadores definidos en el desarrollo					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
	¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?				
crítico	¿Los reportes son configurables a través de la interfaz del sistema?				
	¿La interfaz está orientada a facilitar el uso de las funciones del sistema por parte de los usuarios?				
crítico	¿No existen restricciones para construir cubos OLAP con dimensiones y niveles de agregación ilimitados?				
crítico	¿Los usuarios son capaces de manipular los resultados de manera que se ajusten a sus necesidades, conformando nuevos reportes?				
	¿El sistema responde de una forma rápida a la información que le sea solicitada por el usuario?				
	¿El sistema refleja cualquier lógica del negocio para poder responder a preguntas específicas?				
crítico	¿El sistema garantiza la confidencialidad y seguridad de acceso a los datos por rol de los usuarios?				

	¿Los datos e información derivados del proceso de análisis realizado mediante la aplicación, apoyan la toma de decisiones en la Institución?				
crítico	¿Los cambios en los datos se reflejan automáticamente en los reportes de forma instantánea?				
Semántica del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	¿Se han identificado errores ortográficos en los entregables?				
crítico	¿Se entiende claramente lo que se ha especificado en el documento?				
	¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?				

4.2 Casos de prueba.

El propósito de un caso de prueba es especificar una forma de probar el sistema, incluyendo las entradas con las que se ha de validar, los resultados esperados y las condiciones bajo las que ha de probarse. Un escrito formal se caracteriza por la existencia de una entrada conocida que debe probar una precondición y una salida esperada que prueba una pos condición. Se realizó un caso de prueba por caso de uso de información, a continuación se muestra un ejemplo para el caso de uso “Consultar indicadores físicos”:

Escenario	Descripción	Perfiles de análisis	Indicadores a medir	Respuesta del sistema	Flujo central
EC 1.1: Turistas por países	Permite visualizar el reporte	*Tiempo* *País*	*plan* *cantidad de turistas* *cantidad del*	Se muestra la tabla con los valores	Se autentifica. Se entra al sistema. Se despliega hacia la derecha el componente ubicado en el lateral izquierdo que contiene el

	con las variables		*año anterior*	correspondientes a cada escenario.	navegador. Se selecciona el área de análisis de **AA.Turismo.** Se selecciona el libro de trabajo **LT. Turistas por países.** En la parte inferior izquierda se selecciona el reporte deseado. En el área de trabajo se visualiza la tabla correspondiente al reporte. Se visualiza el gráfico correspondiente a la información de la tabla. Se visualiza el gráfico en correspondencia con el tipo de gráfico que haya determinado el cliente.
EC 1.2: Turistas por países y por DPA	presentes en el mismo.	*Tiempo* *DPA* *País*	*plan* *cantidad de turistas* *cantidad del año anterior*		
EC 1.3: Turistas por países y por cadena		*Tiempo* *Cadena* *País*	*plan* *cantidad de turistas* *cantidad del año anterior*		
EC 1.4: Turistas por países y por entidad		*Tiempo* *País* *Entidad*	*plan* *cantidad de turistas* *cantidad del año anterior*		
EC 1.5: Turistas por países y por tipo actividad		*Tiempo.* *Actividad* *País*	*plan* *cantidad de turistas* *cantidad del año anterior*		
EC 1.6: Turistas por países y por tipo de entidad		*Tiempo* *País* *Tipo entidad*	*plan* *cantidad de turistas* *cantidad del año anterior*		

4.3 Pruebas de aceptación

El objetivo de las pruebas de aceptación es comprobar que un sistema cumple con el funcionamiento esperado y permitir al cliente que determine su aceptación. Para validar la solución se realizó un encuentro con el cliente, donde el mismo realizó las pruebas de validación que demostraron la conformidad con los requisitos.

Conclusiones.

En este capítulo se realizó la validación del mercado de datos Turismo, se aplicaron los casos de pruebas, las listas de chequeo, y se revisó la aplicación por el cliente, donde se encontraron 47 no conformidades, las mismas fueron resueltas y finalmente se aceptó la solución por el cliente.

Conclusiones generales.

Al concluir este trabajo se puede plantear que se cumplió con los objetivos específicos y las tareas de investigación propuestas ya que:

- Se analizaron los principales conceptos relacionados con los almacenes de datos.
- Se identificaron la metodología y herramientas de desarrollo.
- Se realizó el levantamiento de los requisitos.
- Se definieron las reglas del negocio.
- Se diseñó el Modelo de Caso de Uso y el Modelo de Datos.
- Se diseñaron los flujos de transformaciones.
- Se implementó el modelo de datos.
- Se implementaron los flujos de transformaciones.
- Se implementaron los trabajos.
- Se implementó la capa de visualización.
- Se validó el Mercado de datos Turismo.

Recomendaciones

Con el propósito de mejorar la propuesta realizada en este trabajo, se sugiere:

- Realizar la integración de otras fuentes de datos que complementen la información del Mercado de datos.
- Incentivar en la Universidad de las Ciencias Informáticas las investigaciones referentes al proceso de extracción, transformación y carga e Inteligencia de Negocios.
- Realizar un profundo estudio acerca de técnicas de optimización, que puedan ser aplicadas al proceso de extracción, transformación y carga e Inteligencia de Negocios desarrollado.

Bibliografía

1. Almacenes de datos (datawarehouse), Recuperado el 20 de 10 de 2010, disponible en http://www.rhernando.net/modules/tutorials/doc/bd/dw.html.

2. Apache Tomcat - Welcome. (s.f.). Recuperado el 20 de 10 de 2010, disponible en http://tomcat.apache.org/.

3. Business Intelligence. Recuperado el 5 de 5 de 2011, disponible en http://www.productive.com.ar/productos_bi.php.

4. Business Intelligence fácil: DSS: Tipos de decisiones empresariales. Recuperado el 5 de 5 de 2011, disponible en http://www.businessintelligence.info/dss/toma-decisiones-business-intelligence.html.

5. Data Integration: Using ETL. EAI and EII Tools to Create an Integrate Enterprise. (s.f.). Recuperado el 20 de 11 de 2010, de http://www.bi-bestpractices.com/view-articles/4737.

6. Datawarehouse. (s.f.). Recuperado el 20 de 11 de 2010, disponible en http://www.sinnexus.com/business_intelligence/datawarehouse.aspx.

7. Definiciones Conceptos Mercado de datos, disponible en http://www.mitecnologico.com/Main/DefinicionesConceptosMercadosDatos, consultado el 8/05/2011.

8. Hanik, F. (2007). INTRODUCTION TO APACHE TOMCAT 6.

9. http://www.sinnexus.com/business_intelligence/datamart.aspx © Copyright 2007 - 2011 - Sinnexus - Ronda de Outeiro nº 116 - 15008 (A Coruña) - Tel/Fax: 881 884 859.

10. Kimball, Ralph. Designing the Operational Data: Revista DM Review. The Data Warehouse Toolkit. s.l: WILEY PUBLICHING, 1996.

11. Laura Haas. Beauty and the beast: The theory and practice of information integration. ICDT 2007: 28-43.

12. Maurizio Lenzerini," Data Integration: a Theoretical Perspective", Universita di Roma "La Sapienza". ACM PODS 2002.

13. Modelo Dimensional-Almacenando Datos. Recuperado el 10 de 2 de 2011, disponible en http://proyectopentahodw.wordpress.com/2010/05/04/modelo-dimensional/.

14. MOLAP, ROLAP, HOLAP. (2007-2010). Recuperado el 20 de 11 de 2010, disponible en http://www.sinnexus.com/business_intelligence/olap_avanzado.aspx.

15. Mondrian - El servidor OLAP Open Source. (s.f.). Recuperado el 27 de 10 de 2010, disponible en http://pentaho.almacen-datos.com/mondrian.html.

16. OLAP: Definición << Portal BI, Recuperado el 8/05/2011, disponible en http://portal-bi.cl/site/noticias/2011/olap-definicion

17. Open source data quality, data profiling and master data management | DataCleaner. (s.f.). Recuperado el 33 de 11 de 2010, de http://datacleaner.eobjects.org/.

18. Pentaho Mondrian Documentación. Recuperado el 30 de 11 de 2010, disponible en http://mondrian.pentaho.com/documentation/index.php.

19. Pentaho.com Sitemap. (s.f.). Recuperado el 30 de 11 de 2010, disponible en http://www.pentaho.com/products/data_integration/r.

20. Ponniah, P. (2001). Data Warehousing Fundamentals (1a. ed.). New York: John Wiley.

21. Prof. Lauro Soto, Ensenada, BC, Mexico http://www.mitecnologico.com/Main/DefinicionesConceptosMercadosDatos.

22. Siete Formas de integración de datos. (2008). Recuperado el 16 de 11 de 2010, disponible en http://www.muycomputerpro.com/centro-de-conocimientowhite-papers7-formas-de-integracion-de-datos_we9erk2xxdazdqyeebujg7or6shmcwyoivks1jrr48ecaalib5aehoygizgw2zmr/.

23. Sobre PostgresSQL. (s.f.). Recuperado el 10 de 11 de 2010, de http://www.postgresql-es.org/sobre_postgresql.

24. The Data Integration Architecture, 360DegreeView, LLC.

25. Todotecnologia.com:datamart, disponible en http://todotecnology.blogspot.com/2009/09/datamart.html, consultado el 8/05/2011.

26. W. H Inmon. Building the Data Warehouse. QED Press/John Wiley, 1992.

Referencias Bibliográficas

1 W. H Inmon. Building the Data Warehouse. QED Press/John Wiley, 1992

2 http://www.sinnexus.com/business_intelligence/datamart.aspx © Copyright 2007 - 2011 - Sinnexus - Ronda de Outeiro nº 116 - 15008 (A Coruña) - Tel/Fax: 881 884 859

3 OLAP: Definición << Portal BI, Recuperado el 8/05/2011, disponible en http://portal-bi.cl/site/noticias/2011/olap-definicion

4 Laura Haas. Beauty and the beast: The theory and practice of information integration.

ICDT 2007: 28-43.

5 Siete Formas de integración de datos. (2008). Recuperado el 16 de 11 de 2010, de http://www.muycomputerpro.com/centro-de-conocimientowhite-papers7-formas-de-integracion-de-datos_we9erk2xxdazdqyeebujg7or6shmcwyoivks1jrr48ecaalib5aehoygizgw2zmr/

6 *Data Integration: Using ETL. EAI and EII Tools to Create an Integrate Enterprise*. (s.f.). Recuperado el 20 de 11 de 2010, de http://www.bi-bestpractices.com/view-articles/4737

7 Kimball, Ralph. *Designing the Operational Data*: Revista DM Review. *The Data Warehouse Toolkit.* s.l: WILEY PUBLICHING, 1996.

8 Modelo Dimensional-Almacenando Datos. Recuperado el 10 de 2 de 2011, de http://proyectopentahodw.wordpress.com/2010/05/04/modelo-dimensional/

9 Maurizio Lenzerini," Data Integration: a Theoretical Perspective", Universita di Roma "La Sapienza". ACM PODS 2002

10 The Data Integration Architecture, 360DegreeView, LLC