

Universidad de las Ciencias Informáticas

Facultad 6



Título: Mercado de datos Series históricas de turismo para el Sistema de Información de Gobierno

Trabajo de Diploma para optar por el título de
Ingeniero en Ciencias Informáticas

Autora:

Claudia Daniela Cadahía Fernández

Tutores:

Ing. Yuneimy Tellez Pérez

Ing. José Salvador Bermúdez Rodríguez

La Habana, Junio de 2012
“Año 54 de la Revolución”



“Ser bueno es el único modo de ser dichoso.

Ser culto es el único modo de ser libre.

Pero, en lo común de la naturaleza humana, se necesita ser próspero para ser bueno.”

José Julián Martí Pérez.

Declaración de autoría

Declaro ser autora 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 _____.

Claudia Daniela Cadahía Fernández

Firma de la autora

Ing. Yuneimy Tellez Pérez

Firma de la tutora

Ing. José Salvador Bermúdez Rodríguez

Firma del tutor

Datos de contacto

Tutores:

Ing. Yuneimy Tellez Pérez
Especialidad de graduación: Ingeniería Informática
Categoría docente: Instructora
Categoría científica: Ingeniera
Años de experiencia en el tema: 2
Años de graduada: 3
Correo electrónico: ytellez@uci.cu

Ing. José Salvador Bermúdez Rodríguez
Especialidad de graduación: Ingeniería Informática
Categoría docente: Recién Graduado en Adiestramiento
Categoría científica: Ingeniero
Años de experiencia en el tema: 2
Años de graduado: 1
Correo electrónico: jsbermudez@uci.cu

Agradecimientos

A mis padres por amarme, guiarme, cuidarme y estar siempre a mi lado. Por impulsarme cada día a ser mejor persona y mejor profesional. Por acompañarme incansablemente en mis sueños, sobre todo el de ser ingeniera.

A mi familia, incluyéndote a ti Lisy, por estar presente en cada momento de mi vida, aconsejándome y queriéndome, por sobre todas las cosas. Por ser tan bella y unida, y ser lo más importante para mí.

A Salvi, por ser un excelente tutor, un ejemplo a seguir. Por ser mi amigo, por apoyarme y guiarme siempre, por creer y confiar en mí desde el primer día. Por todos aquellos días, tardes, noches y madrugadas que te dedicaste a enseñarme lo que he aprendido a lo largo de esta investigación. Por quererme.

A Mari, por tu eterna paciencia, y por todos y cada uno de los instantes que estuve a tu lado aprendiendo a ser mejor profesional.

A Yune, por ser una magnífica tutora, por su comprensión, apoyo y optimismo en cada momento.

A Osvi, mi eterno y gran amigo, por demostrarme siempre que podía y puedo contar contigo, por quererme y estar ahí en cada instante que te necesité.

A todos los profesores que contribuyeron en mi formación como ingeniera, especialmente a Eddy, Geydi, Yaquelin, Isleny, Abelito, Mario y Maky; por ser ejemplos a seguir y por los consejos brindados.

A mis compañeros del B103-B403, por compartir el sueño que ya hemos alcanzado. Por aportar en cada instante que pasamos juntos, conocimientos, alegría, cariño y amistad a mi vida.

A mis eternos amigos de la FRArtemisa que siempre estuvieron a mi lado, Eve, Neybis, Yisel, Mirni, Dayli, Carlos Feyt y Viti.

A todas aquellas personas que han contribuido de una forma u otra a que mi tránsito por los cinco años de carrera fuesen lo mejor posible.

A Dios y a la vida, por permitir que este momento fuese real y pudiese compartirlo con las personas que amo.

Dedicatoria

A mis padres por amarme, por darme la vida, por ser mi orgullo y ejemplo a seguir en cada paso que doy. Por guiarme a lo largo de los años con su eterno amor, paciencia y comprensión. Por estar a mi lado siempre. Por todo su apoyo, esfuerzo y dedicación en cada instante de mi vida, sobre todo en el camino a ser ingeniera. A ustedes va dedicado todo mi empeño a lograrlo, simplemente porque son los mejores padres del mundo.

A Mima, por preocuparte cada día por mi y tenerme siempre presente. Por tus consejos y todo el amor que me has dado. Por crear, cuidar y guiar a la familia que tengo. Por ser tan especial.

A mis tía Marisol, mi segunda mamá, por el eterno amor que desde niña me has dado, por tus consejos y alientos en cada instante, por tus correos llenos de amor que fueron y son muy importantes en mi vida. Por ser siempre tu princesita.

A mi Titi y a Yurito, que aunque no han podido estar a mi lado como quisieran, me han cuidado, aconsejado y amado en donde quiera que estén, y han sabido ser los mejores hermanos del mundo.

A mis tíos y primos, que han estado conmigo, en los buenos y malos momentos, ya sea de lejos o de cerca, pero siempre ahí para mí: Marisol, Eduar, Juany, Ade, Esther, Ale, Erme, Saito y Claudy.

Especialmente a Juany, por ser mi ejemplo a seguir como profesional, estoy muy orgullosa de tener un tío como tú; y a mi niña Claudy, nunca olvidaré que dedicaste un día con mucho amor a formar parte de mi vida universitaria, que a pasar de los trabajos, siempre permaneciste con aquella sonrisa que alumbraba mi día, te adoro mi niña.

A Lisy, por ser más que mi amiga, mi hermana. Por quererme como soy, por estar siempre dispuesta a todo en cualquier momento que te he necesitado. Porque te quiero.

A mi machi, por compartir mis sueños, por ayudarme y apoyarme en todo lo que hago, por creer en mí. Por estar a mi lado y quererme, como solo tú sabes hacerlo.

A Salvi y a Mari, que más que mis tutores o profesores, se han convertido en mis amigos y han ocupado un lugar muy especial en mi corazón. Por estar ahí, siempre. Por su eterna paciencia, por sus consejos, apoyos y dedicación en cada momento que pasé a su lado, destinados a que fuese mejor profesional. Por el cariño y el amor que les tengo y les tendré siempre.

A Osvi, por ser mi amigo y demostrármelo siempre. Por ser tan especial. Por quererme incondicionalmente.

Resumen

La presente investigación surge como parte de la colaboración que existe entre la Universidad de la Ciencias Informáticas y la Oficina Nacional de Estadísticas e Información. Esta última, guarda cúmulos de datos históricos de diferentes esferas de la vida social, entre ellas el turismo. La forma en que se manipula la información en dicha área, dificulta el análisis de diferentes variables relacionadas con los datos pertenecientes a las series históricas. Debido a esto, se identifica como objetivo fundamental de la investigación: desarrollar el Mercado de Datos Series históricas de turismo para el Sistema de Información de Gobierno. Para cumplir con el objetivo propuesto, se fundamentaron las herramientas, técnicas y metodologías a utilizar durante la investigación. De igual forma, se realizó el análisis y diseño de los principales componentes del Mercado de Datos, sirviendo como base para su implementación. Por último, se implementaron y probaron los diferentes subsistemas que componen la arquitectura de la solución (almacenamiento, integración y visualización). Como resultado final, se obtuvo el Mercado de Datos poblado y una capa de presentación que contiene todos los reportes que satisfacen las solicitudes de información hechas por los especialistas del área de turismo de la Oficina Nacional de Estadísticas e Información.

Palabras claves:

Almacenes de Datos, Mercado de Datos, Oficina Nacional de Estadísticas e Información, turismo.

Tabla de contenidos

INTRODUCCIÓN.....	8
CAPÍTULO 1: FUNDAMENTOS TEÓRICOS SOBRE EL DESARROLLO DE UN ALMACÉN DE DATOS	12
1.1 Introducción	12
1.2 Proceso de digitalización de los datos de las series históricas de turismo que se almacenan en la Oficina Nacional de Estadísticas e Información	12
1.3 Almacenes de Datos	12
1.4 Mercado de Datos	14
1.5 Almacenamiento de la información.....	14
1.5.1 Modos de almacenamiento.....	14
1.5.2 Topología de esquemas.....	18
1.6 Experiencias en el mundo sobre el uso de los Almacenes de Datos	18
1.7 Metodologías para el desarrollo de Almacenes de Datos	20
1.7.1 Ciclo de vida Kimball	21
1.7.2 Metodología para el desarrollo de soluciones de Almacenes de Datos e Inteligencia de Negocios en DATEC	22
1.8 Procesos de integración y visualización de datos.....	24
1.8.1 Extracción, Transformación y Carga.....	24
1.8.2 Inteligencia de Negocio	25
1.9 Técnicas de captura de requisitos	25
1.10 Herramienta de modelado	26
1.11 Sistema Gestor de Bases de Datos	27
1.11.1 PostgreSQL.....	27
1.11.2 PgAdmin	29
1.12 Herramientas informáticas para el proceso de Extracción, Transformación y Carga	29
1.12.1 DataCleaner	29
1.12.2 Pentaho Data Integration	30
1.13 Herramientas informáticas para la Inteligencia de Negocio	31
1.13.1 Schema Workbench	31
1.13.2 Mondrian OLAP Server	31
1.13.3 Pentaho BI Server	31
1.13.4 Apache Tomcat	32
1.14 Conclusiones del capítulo	32
CAPÍTULO II: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS SERIES HISTÓRICAS DE TURISMO PARA EL SISTEMA DE INFORMACIÓN DE GOBIERNO	34
2.1 Introducción	34
2.2 Caracterización de las áreas de la organización.....	34
2.3 Necesidades de los usuarios.....	35
2.4 Reglas del negocio	35
2.5 Especificación de requerimientos	36
2.5.1 Requerimientos de información	36

2.5.2	Requerimientos funcionales	37
2.5.3	Requerimientos no funcionales.....	37
2.6	Casos de uso del sistema	38
2.6.1	Actores del sistema.....	38
2.6.2	Diagrama de Casos de Uso del Sistema	39
2.6.3	Especificación de Casos de Uso del Sistema	39
2.7	Definición de la arquitectura del Mercado de Datos	41
2.8	Diseño de la solución	42
2.8.1	Diseño del subsistema de almacenamiento	42
2.8.2	Diseño del subsistema de integración	46
2.8.3	Diseño del subsistema de visualización	49
2.9	Política de respaldo y recuperación	51
2.10	Esquema de seguridad.....	51
2.11	Conclusiones del capítulo	52
CAPÍTULO III: IMPLEMENTACIÓN Y PRUEBAS DEL MERCADO DE DATOS SERIES HISTÓRICAS DE TURISMO PARA EL SISTEMA DE INFORMACIÓN DE GOBIERNO.....		53
3.1	Introducción	53
3.2	Implementación del subsistema de almacenamiento.....	53
3.2.1	Estructura de los datos.....	53
3.2.2	Estándares de codificación.....	54
3.3	Implementación del subsistema de integración de datos	55
3.3.1	Subsistema de extracción.....	55
3.3.2	Limpieza y transformación de datos	56
3.3.3	Transformaciones y trabajos	56
3.3.4	Carga de datos	58
3.3.5	Gestión del cambio en las dimensiones	59
3.3.6	Gestión de los metadatos del proceso de integración	60
3.4	Implementación del subsistema de visualización de datos	62
3.4.1	Implementación de la capa de visualización	62
3.4.3	Configurar la seguridad de los usuarios y roles	63
3.5	Pruebas.....	64
3.5.1	Pruebas de software	64
3.5.2	Diseño de los Casos de Prueba	65
3.5.3	Listas de chequeo	66
3.5.4	Calidad de datos.....	68
3.6	Evaluación del resultado de las pruebas	69
3.6.1	Aplicación de los Casos de Pruebas.....	69
3.6.2	Aplicación de las Listas de Chequeo	70
3.7	Conclusiones del capítulo	71
CONCLUSIONES GENERALES		72
RECOMENDACIONES		73
REFERENCIAS BIBLIOGRÁFICAS.....		74
BIBLIOGRAFÍA.....		76
GLOSARIO DE TÉRMINOS.....		78

INTRODUCCIÓN

A lo largo de la historia el hombre ha necesitado transmitir y tratar información de forma continua. Producto a esto y apoyándose en el avance de la ciencia y la tecnología la humanidad no ha parado de crear herramientas, programas y diversas técnicas destinadas al procesamiento de la información. Con este fin surge la informática, ciencia que estudia el tratamiento automático y racional de la información.⁽¹⁾ Esta ciencia se encarga del estudio de técnicas, métodos y procesos destinados a almacenar y procesar datos de manera digital. Actualmente, dicha disciplina es aplicada en numerosos sectores de la vida social, y es difícil concebir un área que no utilice, en forma alguna, el apoyo de la informática.

Entre las áreas que la utiliza se encuentra la estadística, eslabón importante para el desarrollo de un país, la cual depende de un elevado volumen de información que necesita ser tratado con herramientas que permitan la gestión y consulta de los datos. Debido a esto, la industria del software juega su papel fundamental en la creación de dichas herramientas para el apoyo a este sector y con el propósito de extraer estadísticas acertadas en todas las esferas de la vida social que se necesiten.

Cuba no está exenta de tales necesidades de información estadística, y con el fin de resolverlas se crea la Oficina Nacional de Estadísticas (ONE) el 21 de abril de 1994, actualmente Oficina Nacional de Estadísticas e Información (ONEI). Esta entidad tiene como principal objetivo captar y difundir cifras económicas y sociales de acuerdo a las necesidades del país. Producto a esto, dicho centro guarda cúmulos de datos históricos de diferentes esferas de la vida social cubana, que deben ser analizados manualmente para la adquisición de la información, dificultándose la inmediata obtención de los reportes para el apoyo a la toma de decisiones.

Enfocada a fomentar principalmente las necesidades de la industria de software en el país, se crea la Universidad de las Ciencias Informáticas (UCI), centro que vincula la docencia y la producción para formar profesionales capaces de fortalecer la industria de software en Cuba. Dentro de los centros de desarrollo que fomentan la línea productiva de la universidad está el Centro de Tecnologías y Gestión de Datos (DATEC), el cual tiene bajo su responsabilidad diversos proyectos. Uno de ellos, tiene como misión fundamental desarrollar el Almacén de Datos (AD)¹ Sistema de Información de Gobierno (SIGOB) con el objetivo de integrar la información que se procesa en la ONEI. Este proyecto liberó, como parte de un Trabajo de Diploma precedente, el Mercado de Datos (MD)² correspondiente al área de turismo, sin embargo no le fueron incluidas las series históricas que se analizan en dicha área de la entidad.

¹ Hace referencia a los términos Almacén de Datos y Almacenes de Datos

² Hace referencia a los términos Mercado de Datos y Mercados de Datos

Dentro de las principales deficiencias que posee la ONEI, con el manejo de la información referente a las series históricas de turismo, se encuentran las siguientes:

- A pesar de utilizar la herramienta Microsoft Excel para el almacenamiento de los datos y de las ventajas que ella posee, los especialistas deben realizar el análisis estadístico de la información de forma manual, lo cual trae consigo la existencia de errores humanos y la pérdida de información útil para la entidad.
- Se genera un gran número de datos anuales obstaculizando su análisis.
- Poseen variadas versiones de los datos, lo que origina la no integración de los mismos.
- La recuperación y creación de los informes se torna engorroso y a veces costoso en cuanto a tiempo y esfuerzo.

Todo esto deteriora la calidad de la información, en lo que se refiere a seguridad, disponibilidad e integridad, surgiendo dificultades para almacenar, recuperar y presentar la información proveniente de los organismos, tales como: principales reportes, cruces de variables, indicadores, porcentajes y demás aspectos de interés; dificultando así la toma de decisiones.

Por todo lo antes expuesto se define como **problema de la investigación**: ¿Cómo contribuir a la toma de decisiones en el área de turismo del Sistema de Información de Gobierno, partiendo de los datos históricos pertenecientes a esta área?

La presente investigación tiene como **objeto de estudio**: los Almacenes de Datos, enmarcado en el **campo de acción**: Mercado de Datos Series históricas de turismo para el Sistema de Información de Gobierno.

Para darle solución al problema científico de la investigación se define como **objetivo general**: desarrollar el Mercado de Datos Series históricas de turismo para el Sistema de Información de Gobierno, que contribuya a la toma de decisiones.

A partir de un análisis realizado al objetivo general se desglosan los siguientes **objetivos específicos**:

1. Fundamentar la selección de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de los Almacenes de Datos.
2. Realizar análisis y diseño del Mercado de Datos Series históricas de turismo para el Sistema de Información de Gobierno.

3. Realizar implementación y pruebas del Mercado de Datos Series históricas de turismo para el Sistema de Información de Gobierno.

Para cumplir con los objetivos planteados se definen las siguientes **tareas de investigación**:

1. Análisis de los principales conceptos, metodologías, herramientas y tecnologías a utilizar en el desarrollo de los Almacenes de Datos, lo que contribuirá a determinar cuáles se utilizarán durante la investigación.
2. Levantamiento de requisitos para determinar las necesidades de información.
3. Descripción de los casos de uso del Mercado de Datos para especificar cada una de las funcionalidades del sistema.
4. Definición de la arquitectura del Mercado de Datos, lo cual permitirá identificar los principales subsistemas que la componen.
5. Definición de los hechos, las medidas y las dimensiones del Mercado de Datos para determinar los elementos que forman parte del modelo lógico de datos.
6. Diseño del modelo lógico de datos para así determinar los elementos que componen el modelo físico de los datos.
7. Diseño del subsistema de integración como guía para la implementación de dicho subsistema.
8. Diseño del subsistema de visualización, permitiendo la definición de la capa de presentación y realizando el diseño de los cubos OLAP, así como los reportes candidatos.
9. Diseño de los casos de pruebas para identificar los elementos que tiene que estar disponibles en el Mercado de Datos una vez culminada la implementación.
10. Implementación del modelo de datos para que queden disponibles las estructuras de la base de datos a la hora de realizar la carga al Mercado de Datos.
11. Implementación del subsistema de integración para que quede poblado el Mercado de Datos, cargando los hechos y las dimensiones correspondientes.
12. Implementación del subsistema de visualización con el objetivo de obtener los reportes para los usuarios finales.
13. Aplicación de las listas de chequeo para determinar que la estructura de los artefactos que corresponden a los procesos de Extracción, Transformación y Carga, tengan la calidad requerida.
14. Aplicación de los casos de prueba para validar los reportes realizados.

Para respaldar el cumplimiento de todos los elementos planteados anteriormente, el presente trabajo de diploma estará compuesto por tres capítulos:

Capítulo 1: Fundamentos teóricos para el desarrollo de un Almacén de Datos

Está referido al análisis del estado del arte de los Almacenes de Datos y de los Mercados de Datos, con sus principales características, metas y elementos que los componen. Se define cómo se realiza el proceso de gestión de la información en el área de turismo y se caracterizan las metodologías, técnicas y herramientas a utilizar para el desarrollo de la solución.

Capítulo 2: Análisis y diseño del Mercado de Datos Series históricas de turismo para el Sistema de Información de Gobierno

En este capítulo se realiza un estudio preliminar del negocio y de la organización, con el fin de obtener las reglas del negocio, identificar los requerimientos y los casos de uso con sus relaciones, describiendo brevemente los actores que van a interactuar con el sistema. Se define la arquitectura del Mercado de Datos y se diseñan los subsistemas de almacenamiento, de integración y visualización. Además se establece el esquema de seguridad a la hora de interactuar con la Base de Datos y la aplicación.

Capítulo 3: Implementación y pruebas del Mercado de Datos Series históricas de turismo para el Sistema de Información de Gobierno

En este capítulo se hace referencia a la implementación de la solución, abordando específicamente cómo se realiza la implementación del subsistema de almacenamiento, de integración y de visualización para las series históricas del área de turismo del Sistema de Información de Gobierno, teniendo en cuenta los requerimientos y necesidades del negocio. De igual forma, hace referencia a las pruebas, mediante la utilización de las listas de chequeo, para determinar que los artefactos de documentación de los procesos de Extracción, Transformación y Carga tengan la calidad requerida, y de los casos de pruebas, basados en Casos de usos y reglas de transformación, para validar los reportes del Mercado de Datos y el cumplimiento de las reglas del negocio respectivamente.

CAPÍTULO 1: FUNDAMENTOS TEÓRICOS SOBRE EL DESARROLLO DE UN ALMACÉN DE DATOS

1.1 Introducción

Está referido al análisis del estado del arte de los Almacenes de Datos y de los Mercados de Datos, con sus principales características, metas y elementos que los componen. Se define cómo se realiza el proceso de gestión de la información en el área de turismo y se caracterizan las metodologías, técnicas y herramientas a utilizar para el desarrollo de la solución.

1.2 Proceso de digitalización de los datos de las series históricas de turismo que se almacenan en la Oficina Nacional de Estadísticas e Información

Las series históricas de turismo se almacenan en tablas, las cuales se encuentran en formato excel. Estos datos dependen de la información turística recibida en un período de tiempo. Para analizarlos es necesaria la participación de especialistas que realizan el análisis de los datos de forma manual. Este es un trabajo minucioso donde se debe revisar tabla por tabla para detectar incongruencias en los datos calculados. Todo este proceso descrito anteriormente hace que se dificulte la manera de realizar los análisis estadísticos sobre el turismo; corriéndose el riesgo de que se pierda información útil al no contar con una herramienta informática que contribuya a mejorar el análisis de la información. Por todas estas razones, es objetivo del presente trabajo de investigación, desarrollar un MD con el fin de permitirle a los especialistas una mejor visualización de los datos y apoyar la toma de decisiones.

1.3 Almacenes de Datos

Un AD es una Base de Datos (BD) ³ corporativa, que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas. Se puede caracterizar un AD haciendo una comparación de cómo los datos de un negocio almacenados en este, se diferencian de los datos operacionales usados por las aplicaciones de producción.

Base de Datos Operacional	Almacén de Datos
datos operacionales	datos del negocio para información
orientado a la aplicación	orientado al sujeto
actual	actual + histórico
detallada	detallada + más resumida
cambia continuamente	estable

Tabla 1: Comparación entre las bases de datos operacionales y los AD

³ Hace referencia a los términos Base de Datos y Bases de Datos

Características de un AD: (2)

El término AD fue acuñado por primera vez por Bill Inmon y según lo definió está caracterizado por ser:

Integrado: los datos almacenados deben integrarse en una estructura consistente, por lo que las inconsistencias existentes entre los diversos sistemas operacionales deben ser eliminadas. La información suele estructurarse también en distintos niveles de detalle para adecuarse a las necesidades de los usuarios.

Temático: sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales. De esta forma, las peticiones de información sobre clientes serán más fáciles de responder dado que toda la información reside en el mismo lugar.

Histórico: el tiempo es parte implícita de la información contenida en un AD. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada en el AD sirve, entre otras cosas, para realizar análisis de tendencias. Por lo tanto, el AD se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

No volátil: el AD existe para ser leído, pero no modificado. La información es por tanto permanente, significando la actualización del AD y la incorporación de los últimos valores que tomaron las distintas variables contenidas en él, sin ningún tipo de acción sobre lo que ya existía.

Otra de las características de los AD es que posee metadatos, es decir, datos sobre los datos, permitiendo simplificar y automatizar la obtención de la información desde los sistemas operacionales a los sistemas informacionales.

Ventajas y desventajas del uso de un AD

➤ **Ventajas**

1. Proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global del negocio.
2. Facilita la integración de sistemas de aplicación no integrados.
3. Organiza y almacena los datos que se necesitan para el procesamiento de la información sobre una amplia perspectiva de tiempo.
4. Hace más fácil el acceso a una gran variedad de datos a los usuarios finales.
5. Permite la limpieza de los datos, es decir, eliminan los datos duplicados, erróneos o incompletos que suelen existir en las bases de datos operacionales.
6. Posibilita el ajuste de los datos para posibles combinaciones que se necesiten realizar.
7. Permite el almacenamiento de los datos históricos.

8. Posee mayor rendimiento, pues se tarda mucho menos en acceder a los datos del repositorio del AD que en hacer una consulta a bases de datos distintas.
9. Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.
10. Permite que los usuarios accedan a la información en línea, contribuyendo a la efectividad para operar en las tareas.

➤ **Desventajas**

1. A lo largo de su vida se elevan los costos de mantenimiento.
2. Pueden quedarse obsoletos relativamente pronto.

1.4 Mercado de Datos

Concepto
Un MD es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. (3)

Tabla 2: Concepto de Mercado de Datos

A pesar de que un MD también puede ser un AD, por tanto posee las mismas características que estos, los MD se centran solamente en los requerimientos de usuarios asociados a un departamento determinado, no contienen datos operacionales detallados y contienen menos información permitiendo que estos tengan mejor entendimiento y navegabilidad.

1.5 Almacenamiento de la información

1.5.1 Modos de almacenamiento

Cuando se toma la decisión de crear un MD de un área funcional de la empresa es preciso encontrar la estructura óptima para el análisis de su información, estructura que puede estar montada sobre una base de datos de Procesamiento Transaccional en Línea (OLTP), como el propio AD, o sobre una base de datos de Procesamiento Analítico en Línea (OLAP). La designación de una u otra dependerá de los datos, los requisitos y las características específicas de cada departamento.

Características de las bases de datos OLTP y OLAP

Los sistemas OLTP son bases de datos orientadas al procesamiento de transacciones. Una transacción genera un proceso atómico (que debe ser validado o invalidado), y que puede involucrar operaciones de inserción, modificación y borrado de datos. El proceso transaccional es típico de las bases de datos operacionales.

A continuación se presentan algunas de las características de estos sistemas:

- El acceso a los datos está optimizado para tareas frecuentes de lectura y escritura. (por ejemplo: la enorme cantidad de transacciones que tienen que soportar las BD de bancos o hipermercados diariamente).
- Los datos se estructuran según el nivel de aplicación.
- Los formatos de los datos no son necesariamente uniformes en los diferentes departamentos (es común la falta de compatibilidad y la existencia de islas de datos).
- El historial de datos suele limitarse a los datos actuales o recientes.

Los sistemas OLAP 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, entre otros. Este sistema es típico de los MD.

A continuación se presentan algunas características de estos sistemas:

- El acceso a los datos suele ser de sólo lectura. La acción más común es la consulta, con muy pocas inserciones, actualizaciones o eliminaciones.
- Los datos se estructuran según las áreas de negocio, y los formatos de los datos están integrados de manera uniforme en toda la organización.
- El historial de datos es a largo plazo, normalmente de dos a cinco años.
- Las bases de datos OLAP se suelen alimentar de información procedente de los sistemas operacionales existentes, mediante un proceso de ETL. (4)

La mayoría de las compañías asumían que la única solución para una aplicación OLAP era un modelo de almacenamiento no relacional. Posteriormente, con la utilización de la topología estrella, copo de nieve y constelación de hechos, de los índices y el almacenamiento de agregados, se podrían utilizar sistemas de administración de bases de datos relacionales (conocido en su término en inglés como Relational Data Base Management System (RDBMS)) para el OLAP. Esta tecnología acogió el nombre de OLAP relacional (ROLAP).

Las primeras compañías adoptaron entonces el término OLAP multidimensional (MOLAP). Las implementaciones MOLAP normalmente se desempeñan mejor que la tecnología ROLAP, pero presentan problemas de escalabilidad. Por otro lado, las implementaciones ROLAP son más escalables y son frecuentemente atractivos a los clientes debido a que aprovechan las inversiones en tecnologías de bases de datos relacionales preexistentes. Por su parte la tecnología HOLAP es una combinación de estas dos últimas. Para un mejor entendimiento a continuación se abordarán las características que poseen estos tipos de sistemas.

Clasificación de los sistemas OLAP

➤ **Sistemas MOLAP (5)**

La arquitectura MOLAP usa bases de datos multidimensionales para proporcionar el análisis; su principal premisa es que el OLAP está mejor implantado almacenando los datos multidimensionalmente. Por el contrario, la arquitectura ROLAP cree que las capacidades OLAP están perfectamente implantadas sobre bases de datos relacionales. Un sistema MOLAP usa una base de datos propietaria multidimensional, en la que la información se almacena multidimensionalmente, para ser visualizada en varias dimensiones de análisis.

El sistema MOLAP utiliza una arquitectura de dos niveles: la bases de datos multidimensionales y el motor analítico. La base de datos multidimensional es la encargada del manejo, acceso y obtención del dato. El nivel de aplicación es el responsable de la ejecución de los requerimientos OLAP. El nivel de presentación se integra con el de aplicación y proporciona una interfaz a través de la cual los usuarios finales visualizan los análisis OLAP. Una arquitectura cliente/servidor permite a varios usuarios acceder a la misma base de datos multidimensional.

La información procedente de los sistemas operacionales, se carga en el sistema MOLAP, mediante una serie de rutinas por lotes. Una vez cargado el dato elemental en la base de datos multidimensional (conocido como Multidimensional Database (MDB)), se realizan una serie de cálculos por lotes, para calcular los datos agregados, a través de las dimensiones de negocio, rellenando la estructura MDB. Una vez que el proceso de compilación se ha acabado, la MDB está lista para su uso. Los usuarios solicitan informes a través de la interfaz y la lógica de aplicación de la MDB obtiene el dato. La arquitectura MOLAP requiere unos cálculos intensivos de compilación. Tiene capacidades limitadas para crear agregaciones dinámicamente.

➤ **Sistemas ROLAP**

La arquitectura ROLAP, accede a los datos almacenados en un AD 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. El sistema ROLAP utiliza una arquitectura de tres niveles. La base de datos relacional maneja los requerimientos de almacenamiento de datos, mientras el motor ROLAP proporciona la funcionalidad analítica. El nivel de base de datos usa bases de datos relacionales para el manejo, acceso y obtención del dato. El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios.

El motor ROLAP se integra con niveles de presentación, a través de los cuales los usuarios realizan los análisis OLAP. Después que se ha definido el modelo de datos para el AD, los datos se cargan desde el sistema operacional. Se ejecutan rutinas de bases de datos para agregar el dato,

si así lo requiere el modelo de datos. Se crean entonces los índices para optimizar los tiempos de acceso a las consultas.

Los usuarios finales ejecutan sus análisis multidimensionales, a través del motor ROLAP, que transforma dinámicamente sus consultas a consultas SQL. Se ejecutan estas consultas SQL en las bases de datos relacionales, y sus resultados se relacionan mediante tablas cruzadas y conjuntos multidimensionales para devolver los resultados a los usuarios.

La arquitectura ROLAP es capaz de usar datos pre calculados, 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 AD, 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 *joins* múltiples.

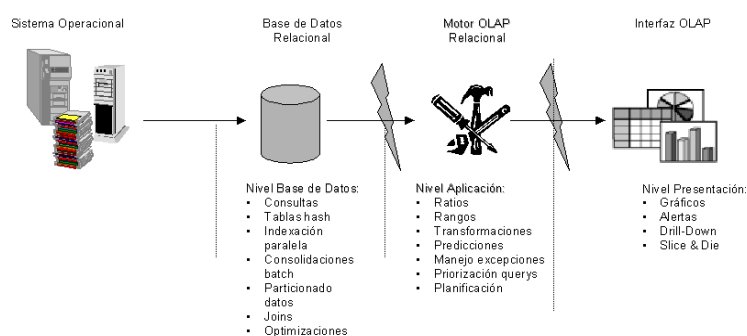


Figura 1: Estructura de los sistemas ROLAP

➤ Sistemas HOLAP

Un desarrollo un poco más reciente ha sido la solución OLAP híbrida (HOLAP), la cual 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.

En la presente investigación se utiliza como modo de almacenamiento de los datos el sistema ROLAP, debido a las características anteriormente descritas, y principalmente, porque el Sistema Gestor de Base de Datos (SGBD) que se utiliza es PostgreSQL, el cual no soporta las estructuras MOLAP ni HOLAP.

1.5.2 Topología de esquemas

Para realizar el modelo de datos dimensional y físico se debe tener presente el tipo de topología que se empleará, analizando cuál de los tres tipos de esquemas existentes es el más adecuado y el que se adapta a los requerimientos y necesidades del cliente. Ellos son: (6)

- Esquema en estrella: el diseño consiste en una tabla de hechos (conocido como *fact table*) en el centro, como objeto de análisis y una o varias tablas de dimensión (conocido como *dimension table*) por cada dimensión de análisis que participa de la descripción de ese hecho, las cuales no se relacionan entre sí. Las tablas de dimensiones tendrán siempre una clave primaria simple, mientras que en la tabla de hechos, la clave principal estará compuesta por las claves principales de las tablas dimensionales. Por tanto, cada tupla de la tabla de hechos incluye las medidas y una referencia a cada dimensión. Las tablas de dimensiones se encuentran además totalmente desnormalizadas, es decir, toda la información referente a una dimensión se almacena en la misma tabla.
- Esquema bola de nieve o copo de nieve: es una variedad más compleja del esquema estrella. Lo que diferencia a la arquitectura en copo de nieve del esquema estrella es que las tablas de dimensión se normalizan en múltiples tablas, es decir, las tablas de dimensiones pueden estar relacionadas entre sí. Por tanto, la tabla de hechos deja de ser la única que se relaciona con otras tablas del esquema. Existen dos tipos de esquemas en copo de nieve: Copo de nieve Completo, donde todas las tablas de dimensión en el esquema en estrella aparecen normalizadas, y el Copo de nieve Parcial, donde sólo se lleva a cabo la normalización de algunas de ellas.
- Esquema constelación de hechos: es una generalización de los esquemas en estrella y copo de nieve, que se obtiene con la inclusión de distintas tablas de hechos que comparten todas o algunas de las dimensiones.

1.6 Experiencias en el mundo sobre el uso de los Almacenes de Datos

Una de las características fundamentales e importantes que poseen los AD es que son un soporte vital para la toma de decisiones de las empresas u organizaciones que gestionan grandes cúmulos de información. Muchas han sido las empresas que han tenido grandes éxitos en cuanto a la implementación de un AD.

Una de ellas es la compañía Wall-Mart, principal tienda de descuentos del mundo, la cual recurre frecuentemente a su gran AD (una de las mayores bases de datos en el mundo), en la que diariamente se registra cada una de las 10 millones de transacciones que realiza con sus clientes. Su AD les permite agrupar la información en las formas más diversas: por tienda, por región, por horas del día, o

cualquier otro parámetro y con este conocimiento la firma puede determinar con exactitud casi matemática el surtido adecuado de mercancía para cada una de sus tiendas. Haciendo que cada tienda esté perfectamente surtida con los productos en ella demandados. (7)

También, Twentieth Century Fox utiliza los AD para filtrar millones de recipientes de zonas postales y predecir qué actores, argumentos y filmes serán populares en cada vecindario. Otra de estas empresas es Jhon Deere, manufacturera de equipos para agricultura que mejora su negocio dando a los clientes una gran diversidad de opciones en los productos que ellos pueden requerir, resultando en millones de permutaciones para cada opción. (7)

En el campo del transporte de carga aparecen empresas como Cornrail, Union Pacific, Norfolk Southern, American President Lines, Delta, Lufthansa, QANTAS, British Airways, Air France, American Airlines, Canadian Airlines y SNFC que utilizan los AD para el análisis de los datos históricos referentes al estudio sobre ventas, clientes, transportaciones, monitoreo de ganancias y futuras proyecciones.(8)

En la esfera de las telecomunicaciones se encuentran Bouygues Telecom, la tercera compañía operadora más grande en telecomunicaciones inalámbricas en Francia, Jazztel, Vodafone, France Telecom y CRTVG la Compañía de Radio Televisión de Galicia, que han utilizado los AD para la monitorización de clientes, la prestación de servicios, los cobros y pagos, el marketing, en fin, para la realización de estudios necesarios para mantenerse en la preferencia de los clientes. (8)

El tema estadístico también utiliza este tipo de tecnología, en países como México, específicamente en el Instituto Nacional de Estadísticas e Informática (INEGI), se tiene la información almacenada en AD y de esta forma es servido para la toma de decisiones a nivel gubernamental.

Además el Instituto Nacional de Estadística de Venezuela posee un conjunto de MD para realizar el análisis de las operaciones estadísticas más importantes que utilizan. El Instituto Nacional de Estadística de España es otro ejemplo de entidades de este tipo que actualmente están utilizando esta tecnología. (8)

En la esfera de la medicina se encuentra la Congregación de Hermanas Hospitalarias, organización internacional de asistencia médica, activa en 24 países en Europa, América, África y Asia, que utiliza un AD dentro de su infraestructura tecnológica con el fin de realizar estudios de patrones de comportamiento en pacientes con diferentes patologías, y para dar seguimiento a los pacientes con sus tratamientos. (8)

Otras organizaciones como Bacardí Martini (distribución de bebidas) recurren a su AD para lograr el máximo de ventas con un coste preestablecido de antemano. (8)

El diario El Mundo también cuenta con un AD con el objetivo de obtener la información completa sobre la contratación de publicidad en sus medios. (8)

Cuba tampoco está exenta de estas nuevas tecnologías, y aunque todavía faltan muchos aspectos por mejorar ya se han visto algunos ejemplos que han dado pasos firmes dentro de esta rama. Un ejemplo de esto lo constituye el grupo empresarial CIMEX el cual utiliza un AD para la gestión de inventarios permitiendo una gestión de compra-venta eficiente, con la finalidad de disminuir los costos, sin afectar al cliente. (9) Además, la Universidad de las Ciencias Informáticas utiliza un AD para la toma de decisiones respecto al consumo energético y desarrolló uno para el Centro de Inmunología Molecular orientado al análisis de los ensayos clínicos que allí se gestionan. (9) Existen otras entidades como UNION CUPET y Copextel que se encuentran en el desarrollo de sus almacenes. (8) En el área del turismo, en específico en la empresa FINTUR S.A Sucursal de Villa Clara, se creó un AD que brinda a sus clientes la obtención de los saldos de sus cuentas bancarias, emisión de estados de cuentas y flujo de caja; además, brinda la posibilidad a la gerencia de consultar analíticamente la información financiera de sus clientes.

1.7 Metodologías para el desarrollo de Almacenes de Datos

Una metodología es el conjunto de métodos por los cuales se regirá una investigación científica. Por lo tanto lo que hace la metodología es estudiar los métodos para luego determinar cuál es el más adecuado a aplicar o sistematizar en una investigación o trabajo.

No existe una única metodología en la cual basarse para la construcción de un AD, sino que dependiendo del contexto que se encuentre la empresa y los objetivos que persiga, se puede emplear una u otra. Estas metodologías se engloban dentro de dos grandes enfoques: ascendente (top-down) y descendente (bottom-up) que se corresponden con las metodologías propuestas por Bill Inmon y Ralph Kimball respectivamente. El primero considerado como el padre de la disciplina y creador del término AD, el segundo, especialista reconocido a nivel mundial en el desarrollo de los AD y creador del enfoque multidimensional. La principal diferencia que existe entre estas metodologías es la forma de enfrentar el problema. El enfoque de Inmon es un enfoque descendente, por lo que al ser construidos descendentemente los MD se nutren del AD. Este enfoque es utilizado cuando existe un conocimiento previo de la tecnología y los problemas del negocio. En él, los datos son extraídos por los procesos de ETL y cargados en el área temporal (conocido por su término en inglés como *staging area*), donde son validados y consolidados. Una vez realizado este proceso, los procesos de actualización de los MD departamentales obtienen la información de él, y con las consiguientes transformaciones, organizan los datos en las estructuras particulares requeridas por cada uno de ellos, actualizando su contenido. Es

un enfoque sistémico, que minimiza los problemas de integración, pero es costoso, debido a la gran cantidad de datos y su poca flexibilidad. Por el contrario, el enfoque que propone Kimball es ascendente, pues al final el AD no es más que la unión de los diferentes MD. Esta característica le hace más flexible y sencillo de implementar, pues se puede construir un MD como primer elemento del sistema, y luego ir añadiendo otros que comparten las dimensiones ya definidas o incluyen otras nuevas. En este sistema, los procesos ETL extraen la información y los procesan igualmente en el área temporal realizando luego el llenado de cada uno de los MD de una forma individual, y siempre respetando la estandarización de las dimensiones. Es una metodología rápida que se basa en experimentos y prototipos, que permite a la organización ir más lejos con menores costos. Este enfoque parte de los requisitos del negocio, mientras que el enfoque descendente propone la validación de los requisitos una vez que se tiene el sistema.

1.7.1 Ciclo de vida Kimball

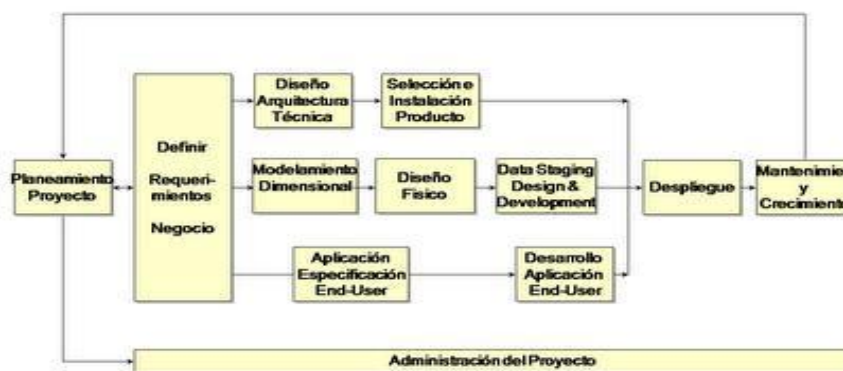


Figura 2: Ciclo de vida Kimball

El Ciclo de vida Kimball comienza con una planificación de proyecto, donde se define el alcance, se identifican y programan las tareas, se planifica el uso de los recursos, conformando con todo esto el plan de proyecto. En la segunda etapa se definen los requerimientos del negocio. Luego de esto el proyecto se enfoca en tres líneas concurrentes: tecnología, datos y aplicaciones de BI. El ciclo de vida culmina con el despliegue y mantenimiento del producto.

Para definir la metodología de desarrollo a utilizar en el Departamento de Almacenes de Datos de DATEC, se tomó como base la Metodología de Kimball por los siguientes elementos:(10)

- Crea los conceptos de hechos y dimensiones, lo que es importante en el proceso de la toma de decisiones.

- Propone ir construyendo el Almacén de Datos a través de la construcción de los Mercados de Datos departamentales, lo cual coincide con la división lógica de las empresas, entidades u organismos, y constituye una buena estrategia pues permite ir presentando resultados parciales a los clientes en cortos plazos.
- Existe abundante documentación sobre la misma y se puede consultar la web a través de los servicios que brindan el grupo creador de la metodología.

A pesar de todas las ventajas que ofrece la utilización de la Metodología de Kimball, esta no era totalmente adaptable a las características del centro y de la producción en la UCI, por lo que solo se decidió utilizarla como guía en el proceso de confección de metodología de desarrollo del Departamento de Almacenes de Datos. Entre sus principales desventajas se encuentran:

- No tiene definido un criterio que permita estimar los costos de desarrollo de un Almacén de Datos, basándose en las características de la construcción del mismo.
- Presenta un grupo de roles, pero no explica claramente cuáles son las competencias y responsabilidades de cada uno dentro del proyecto. Por la cantidad de roles que propone se necesita de grupos grandes para su desarrollo.
- Propone un gran número de actividades y artefactos que pueden extender los tiempos de desarrollo si se cuenta con pocos recursos humanos, además no se especifica cómo deben realizarse estos artefactos.
- Está estructurada para el desarrollo de proyectos – productos, donde un proyecto desarrolla un producto determinado.
- No establece el análisis de diferentes criterios de diseño en el levantamiento de requisitos que permita la construcción más adecuada del almacén, teniendo en cuenta las metas de la organización, las necesidades de los usuarios y la disponibilidad de las fuentes operaciones.

Por tales motivos se definió una metodología que permite mitigar las desventajas identificadas en la Metodología de Kimball, que se ajusta a las condiciones y características de producción de la UCI. (10)

1.7.2 Metodología para el desarrollo de soluciones de Almacenes de Datos e Inteligencia de Negocios en DATEC

La metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio en DATEC, se basa en el ciclo de vida Kimball y en la propuesta realizada por Leopoldo Zenaido Zepeda Sánchez en su tesis de doctorado, en la cual plantea incluir los casos de uso para guiar el proceso de desarrollo (11). Está adaptada a las necesidades de la UCI, y cubre las fases por las que pasa la

construcción de un AD. Las particularidades que presenta este modelo de adaptación es la identificación de requerimientos de información y a su vez, la trazabilidad que tienen estos en todo el ciclo de desarrollo del MD. También la inclusión de una etapa de pruebas que fortalece en gran medida la calidad con que se despliegue la solución propuesta. Además, ajusta las fases, actividades y artefactos a la propuesta de programa de mejoras que lleva a cabo el Centro Nacional de Calidad de Software (CALISOFT) con el objetivo de alcanzar el nivel 2 del Modelo de Integración de Capacidades de Madurez (conocido como Capability Maturity Model Integration, CMMI) en los centros productivos de la UCI.

Ciclo de vida de la metodología:

El ciclo de vida de la metodología está organizado por fases, las cuales podrán ser implementadas de forma paralela según el componente que se está desarrollando para integrarse al final de la solución. A continuación se describen las fases del ciclo de vida de la metodología: (10)

- **Estudio preliminar y planeación:** se realiza el estudio de la entidad cliente para determinar lo que se desea construir y las condiciones que existen para el desarrollo de la misma. En la planeación del proyecto se definen los objetivos, el alcance preliminar, los costos estimados y otras actividades.
- **Levantamiento de requisitos:** se realiza en tres direcciones; una, identificando las metas y objetivos de la organización; dos, identificando las necesidades de información y reglas del negocio; y tres, realizando un levantamiento detallado de las fuentes de datos a integrar. Es aquí donde se definen los requerimientos a través de la comparación de las necesidades y las reglas del negocio.
- **Arquitectura:** se define la arquitectura de la solución según los requisitos no funcionales obtenidos. Se definen aspectos como: la seguridad del sistema, la comunicación entre los subsistemas, la tecnología a utilizar, hardware y software, entre otros aspectos de gran importancia.
- **Diseño e implementación:** se define el diseño de las estructuras de almacenamiento, se diseñan los procesos de integración de datos como, las reglas de ETL, se diseñan los cubos para la presentación de los datos, así como el diseño visual de la aplicación definido por el cliente. Después se implementan cada uno de los subsistemas (almacenamiento, integración y visualización). Se lleva a cabo el diseño físico del Repositorio de Datos, se crean las estructuras de almacenamiento con las particiones y agregaciones correspondientes. Se crea el área temporal de almacenamiento, se ejecutan las reglas de ETL, haciendo los ajustes para integrar la información. Se configuran e implementan las herramientas de Inteligencia de

Negocio o Business Intelligence (BI) para obtener los reportes, gráficos, mapas y otros que cubran los requerimientos firmados con el cliente final.

- Prueba: se realizan las pruebas de unidad, luego las pruebas de integración y sistema, hasta llegar a las pruebas de aceptación con el cliente final.
- Despliegue: consta de dos etapas, despliegue piloto en el cual se configuran los servidores, se instalan las herramientas según la arquitectura definida y se carga una muestra de los datos para demostrar al cliente que el sistema funciona. Posterior a la aceptación del cliente, se realiza la carga histórica de los datos, la capacitación y la transferencia tecnológica. El resultado fundamental es la solución desplegada en el entorno real y en correcto funcionamiento.
- Soporte y mantenimiento: después de haber implantado la solución, se brindan los servicios de soporte en línea, vía telefónica, web u otros, hasta el acompañamiento junto al cliente según el contrato firmado y las condiciones de soporte establecidas.
- Gestión y administración del proyecto: se lleva durante todo el ciclo de vida, es aquí donde se controla, gestiona y chequea todo el desarrollo, los gastos, las utilidades, los recursos, las adquisiciones, y demás actividades relacionadas con la gestión del proyecto.

En el presente Trabajo de Diploma se decidió utilizar la Metodología para el Desarrollo de Soluciones de AD e Inteligencia de Negocio en DATEC debido a que se considera la apropiada para dar solución al problema de investigación planteado, pues cubre las fases por la que pasa la construcción de un AD y brinda diversas ventajas que facilitan el desarrollo del MD, tales como:

- La solución completa se puede implementar en poco tiempo.
- Los productos son más comprensibles para los usuarios.
- Es resistente y tolerante ante los cambios.

1.8 Procesos de integración y visualización de datos

1.8.1 Extracción, Transformación y Carga

ETL -Extracción, Transformación y Carga, del inglés Extract, Transform y Load. Se define como el proceso a través del cual se gestionan datos obtenidos de múltiples fuentes, con el fin de extraerlos, transformarlos y cargarlos en bases de datos especializadas, denominadas MD, para analizar y apoyar una determinada línea de producto o unidad de negocios. (12)

Extracción: consiste en sustraer los datos brutos desde las fuentes de origen, integrando en una misma metodología de negocios, toda la información empresarial proveniente de diferentes fuentes.

Por lo general los datos brutos se ubican en bases de datos relacionales o ficheros planos, igual pueden incluir bases de datos no relacionales y demás estructuras de datos diferentes.

Transformación: una vez terminada la fase de extracción se realiza un chequeo que verifique si los datos cumplen con las pautas estipuladas, en caso de que los datos no cumplan se aplican una serie de procedimientos donde estos quedarían listos para ser cargados.

Carga: la fase de carga interactúa directamente con la base de datos destino, debido a que los datos son incluidos en el sistema dependiendo de los requerimientos de la organización. En algunas bases de datos, se sobrescribe la información antigua con los nuevos datos, pero en el caso de los AD, estos mantienen un historial de los registros de manera que se pueda hacer una auditoría de los mismos y disponer de un rastro del comportamiento de un determinado valor a lo largo del tiempo. (12)

1.8.2 Inteligencia de Negocio

La Inteligencia de Negocio es la habilidad para *transformar los datos en información, y la información en conocimiento*, de forma que ayude en el proceso de toma de decisiones en los negocios. Asociándolo directamente con las tecnologías de la información, se puede definir BI como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la organización) en información estructurada, para su explotación directa o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio.

BI actúa como un factor estratégico para una organización, generando una potencial ventaja competitiva, pues proporciona información privilegiada para responder a los problemas de negocio: entrada a nuevos mercados, promociones u ofertas de productos, control financiero, optimización de costes, planificación de la producción, análisis de perfiles de clientes, rentabilidad de un producto concreto. (13)

En definitiva, una solución BI completa permite: (14)

Observar ¿qué está ocurriendo?

Comprender ¿por qué ocurre?

Predecir ¿qué ocurriría?

Colaborar ¿qué debería hacer el equipo?

Decidir ¿qué camino se debe seguir?

1.9 Técnicas de captura de requisitos

La captura de los requisitos es una pieza fundamental en el desarrollo de los sistemas informáticos, pues sirven de base para verificar si se cumplieron los objetivos propuestos en el proyecto. Desde sus

inicios, los ingenieros han presentado dificultades para obtener los requisitos debido a que la información a veces proviene de diferentes personas y los datos pueden variar. Con el objetivo de darle solución a dichas dificultades se han desarrollado técnicas que permiten realizar el proceso de manera más segura. Algunas de las más utilizadas son: entrevistas, tormenta de ideas, cuestionarios, observaciones, discusiones, análisis de protocolo y casos de uso.

En la presente investigación se utilizaron las entrevistas, permitiendo tener una mejor comprensión del problema y los objetivos de la solución, posibilitando de esta forma obtener una amplia visión del trabajo y de las necesidades del usuario. Otras de las técnicas que se utilizaron fueron las discusiones con los especialistas, para aclarar dudas y poder realizar una correcta captura de los requisitos.

1.10 Herramienta de modelado

Con el fin de desarrollar programas, utilizando técnicas de diseño y metodologías bien definidas, soportadas por herramientas automatizadas, existen hoy en día diversas herramientas que han sido creadas para el desarrollo de la Ingeniería de Software. Ejemplo de esto, son las herramientas CASE (Computer Aided Software Engineering), las cuales constituyen un conjunto de programas y ayudas que dan asistencia a los analistas, ingenieros de software y desarrolladores, durante todos los pasos del Ciclo de Vida de desarrollo de un Software. (15)

Estas herramientas pueden proveer muchos beneficios en todas las etapas del proceso de desarrollo de software, algunas de ellas son:

- Verificar el uso de todos los elementos en el sistema diseñado.
- Automatizar el dibujo de diagramas.
- Ayudar en la documentación del sistema.
- Ayudar en la creación de relaciones en la base de datos.
- Generar estructuras de código. (15)

UML: es un lenguaje para el desarrollo de software orientado a objetos, teniendo como propósito el de visualizar, especificar, construir y documentar proyectos de software. (16)

Entre las herramientas CASE orientadas a UML se encuentran: (16)

- Rational Rose
- ArgoUML
- Poseidón
- Visual Paradigm
- MagicDraw UML
- Borland Together

En la presente investigación se decidió utilizar Visual Paradigm para UML en su versión 8.0 debido a que posee las siguientes características: (17)

- Es multiplataforma y permite su uso en cualquier sistema operativo.
- Utiliza UML como lenguaje de modelado.
- Posibilita una rápida construcción de las aplicaciones con alta calidad.
- Es factible a la hora de dibujar diagramas de clases y generar script para diferentes SGBD.
- Permite una integración con sistemas de control de versiones que almacenan centralmente los artefactos y realizan un seguimiento de los cambios realizados sobre un proyecto.
- Se integra con las siguientes herramientas Java:
 - Eclipse/IBM WebSphere
 - JBuilder
 - NetBeans IDE
 - Oracle JDeveloper
 - BEA Weblogic
- Está disponible en varias ediciones, cada una destinada a distintas necesidades: empresarial, profesional, comunidad, estándar, modelador y personal.

1.11 Sistema Gestor de Bases de Datos

Los Sistemas Gestores de Bases de Datos son un conjunto de programas, que permiten crear y mantener una base de datos, asegurando su integridad, confidencialidad y seguridad. Se componen de un lenguaje de definición de datos, uno de manipulación de datos y uno de consulta. Por su término en inglés son conocidos como DataBase Management System (DBMS). (18)

Características de un Sistema Gestor de Bases de Datos:

- Capaces de controlar la concurrencia y las operaciones que implican la recuperación de fallos.
- Definen usuarios y sus restricciones de acceso.
- Respetan la integridad y seguridad de los datos.
- Toleran definiciones de esquemas y vistas.

1.11.1 PostgreSQL

Es un Sistema Gestor de Bases de Datos objeto-relacional, con su código fuente disponible libremente. Utiliza una arquitectura cliente/servidor y usa multiprocesos para garantizar la estabilidad del sistema, es decir, un fallo en uno de los procesos no afectará el resto y el sistema continuará funcionando. Además es un sistema que funciona de manera excelente con grandes cantidades de datos y una gran

cantidad de usuarios accediendo a la vez al sistema.

A continuación se muestran algunas de las ventajas de utilizar PostgreSQL: (9)

- Soporta distintos lenguajes: PHP, C, C++, Perl y Python.
- Drivers: ODBC, JDBC y .Net.
- Soporta: *triggers*, procedimientos almacenados, funciones, secuencias, relaciones, reglas, tipos de datos definidos por el usuario, vistas y vistas materializadas.
- Soporte de tipos de datos de SQL92, SQL99 y SQL2003.
- Soporte de protocolo de comunicación encriptado por SSL.
- El máximo de bases de datos que posee es ilimitado.
- Posee un máximo de tamaño de tabla de 32 TB.
- El máximo de tamaño de registro es de 1.6 TB.
- El máximo de tamaño de campo es de 1GB.
- El máximo de registros por tabla es ilimitado.
- El máximo de campos por tabla es de 250 a 1600 (depende de los tipos usados).
- El máximo de índices por tabla es ilimitado.
- Número de lenguajes en los que se puede programar funciones: aproximadamente 10 (pl/pgsql, pl/java, pl/perl, pl/python, tcl, pl/php, C, C++ y Ruby).

En la actualidad Cuba se encuentra en proceso de migración hacia software libre, como parte fundamental de este proceso, la UCI, ha decidido migrar para lograr una independencia tecnológica. Actualmente la universidad está trabajando en plataforma PostgreSQL, por lo que este queda definido como el SGBD a utilizar. La versión a utilizar de la plataforma será 9.1 debido a las siguientes características:

- Es estable y seguro.
- Replicación asincrónica.
- Posee copia de seguridad en línea.
- Debido a que soporta operadores, funciones, métodos de acceso y tipos de datos definidos por el usuario hace que sea extensible.
- Soporta la integridad referencial, la cual es utilizada para garantizar la validez de los datos en una base de datos.
- Se encuentra disponible en varios lenguajes y para varios sistemas operativos como Microsoft Windows, Linux, FreeBSD, Mac OSX y Solaris.

1.11.2 PgAdmin

PgAdmin es una plataforma de desarrollo y administración de código abierto para el SGBD PostgreSQL. Es multiplataforma debido a que puede ser usada en Linux, FreeBSD, Solaris, Mac OSX y Windows para gestionar dicho SGBD, así como las versiones derivadas y comerciales como Postgres Plus Advanced Server y Greenplumdatabase.

En la presente investigación se escoge PgAdmin III en su versión 1.14 como la herramienta de administración, pues está diseñada para responder, desde simples consultas SQL hasta el desarrollo de complejas bases de datos; y la interfaz gráfica que posee es compatible con todas las características de PostgreSQL.

Entre algunas de las funcionalidades que incluye la aplicación se encuentra un editor de sobresaltado de sintaxis SQL (*syntaxhighlighting SQL editor*) y un editor de código seguro del lado del servidor (*server-sidecode editor*). También, puede realizarse la conexión mediante TCP/IP o *Unix Domain Sockets*, y puede ser encriptada mediante SSL para mayor seguridad.

La aplicación es desarrollada por una comunidad de expertos de PostgreSQL de todo el mundo, por lo que es un software libre lanzado bajo la licencia de PostgreSQL y está disponible en más de una docena de idiomas. Además tiene acceso a todos los objetos de PostgreSQL los cuales son mostrados con su definición SQL y una lista de propiedades. También se pueden explorar los objetos dependientes, las dependencias y las estadísticas de los objetos consultados en cada caso.

1.12 Herramientas informáticas para el proceso de Extracción, Transformación y Carga

1.12.1 DataCleaner

Se decide usar la herramienta DataCleaner en su versión 1.5.3, utilizada para el perfilado, validación y comparación de los datos. Dentro de las principales características por lo cual se decidió utilizar dicha herramienta se encuentran las siguientes:

- Es una aplicación de código abierto.
- Es muy fácil de utilizar.
- Genera sofisticados informes y gráficos que permiten a los usuarios determinar el nivel de calidad de los datos e identificar y analizar la estructura del origen de los datos.
- Se considera una alternativa libre para la metodología de administración de datos, para proyectos de AD, búsquedas estadísticas, para actividades de preparación de ETL y otras.

1.12.2 Pentaho Data Integration

En la presente investigación se decidió utilizar Pentaho Data Integration en su versión 4.2.1 para el desarrollo de los procesos de ETL, debido a que es una herramienta libre, muy potente, antigua y una de las más utilizadas por los usuarios, considerándola la más completa por la gran cantidad de conectores que posee y la posibilidad de crear flujos de trabajo integrados con transformaciones de datos de manera muy sencilla y funcional. Dentro de las principales características que posee dicha herramienta se encuentran las siguientes: (19)

- Posee un entorno gráfico de desarrollo.
- Uso de tecnologías estándar: Java, XML, JavaScript.
- Es fácil de instalar y de configurar.
- Es multiplataforma: Windows, Macintosh, Linux.
- Basado en dos tipos de objetos: transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones).
- Es un software de código abierto.
- Sin costes de licencia.

Está formado por un conjunto de herramientas, cada una con un propósito específico:

Spoon: herramienta gráfica que permite el diseño de las transformaciones y trabajos. Incluye opciones para pre visualizar y testear los elementos desarrollados. Es la principal herramienta de trabajo de Pentaho Data Integration y con la que se construyen y validan los procesos de ETL.

PAN: herramienta que permite la ejecución de las transformaciones diseñadas en spoon (bien desde un fichero o desde el repositorio). Permite desde la línea de comandos preparar la ejecución mediante scripts.

CHEF: para crear trabajos.

Kitchen: permite ejecutar los trabajos *batch* diseñados con CHEF. (19)

Ventajas de Pentaho Data Integration:

- Funciona en Windows, Unix y Linux.
- Tiene una interfaz gráfica con indicadores de las transformaciones.
- Es una aplicación implementada en Java con algunas características avanzadas en JavaScript.
- Ofrece una licencia pública GPL (del inglés General Public License).
- Basada en metadatos.
- Como soporte se encuentran los foros de Pentaho y la comunidad Pentaho.
- Soporta Oracle, DB2.SQL Server, Sybase así como MySQL y Postgres.

- Soporta la arquitectura de procesamiento en paralelo para distribuir las tareas de ETL a través de múltiples servidores, basado en dos tipos de objetos: transformaciones y trabajos.

1.13 Herramientas informáticas para la Inteligencia de Negocio

1.13.1 Schema Workbench

Se decidió utilizar en la presente investigación Schema Workbench en su versión 3.2.1, debido a que es una herramienta de análisis caracterizada por su potencia gráfica y su capacidad multitarea. Puede utilizarse para ingeniería inversa a una base de datos o a un modelo para visualizar mejor o realizar su mantenimiento. El esquema de Mondrian Workbench es una interfaz de diseño que permite crear y probar los cubos OLAP visualmente. El motor de Mondrian procesa las solicitudes de MDX con el ROLAP (Relational OLAP). Estos archivos son los modelos de esquemas de metadatos XML creados en una estructura específica que utiliza el motor de Mondrian. Estos modelos XML pueden ser considerados como cubos. Con esta aplicación, se puede configurar una conexión JDBC como el modelo físico, para luego elaborar el esquema lógico de manera simple y efectiva. (20)

1.13.2 Mondrian OLAP Server

Mondrian es una de las aplicaciones más importantes de la plataforma Pentaho BI. En el presente trabajo de diploma se decidió utilizar en su versión 3.0.4 debido a que es un servidor OLAP de código abierto que gestiona la comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuente. Además permite crear cubos OLAP para el análisis multidimensional, y combinado con Jpivot, el servidor OLAP Mondrian, permite realizar consultas al AD, lo que posibilita que los resultados sean presentados a través del navegador.

Entre sus ventajas se encuentran:

- Hace posible la agilización de consultas de grandes cantidades de datos.
- Posee una alta velocidad de respuesta.
- Permite realizar consultas al MD.
- Es un motor ROLAP con caché.

1.13.3 Pentaho BI Server

La plataforma Pentaho BI proporciona la arquitectura y la infraestructura necesaria para crear soluciones de BI. Ofrece los servicios básicos incluyendo la autenticación, registro, auditoría, servicios web y motores de reglas. La plataforma también incluye una solución que integra la presentación de informes, análisis, cuadros de mando y los componentes de minería de datos. La aplicación más conocida de la Plataforma Pentaho BI es BI Server, que funciona como una Web, basado en un

sistema de gestión de reportes, servidores de integración de aplicaciones y en el motor de flujo de trabajo ligero (secuencias de acción). Está diseñado para integrarse fácilmente en cualquier proceso de negocio.

Algunas de sus ventajas son:

- Se integra con procesos de negocio.
- Permite la administración y programación de reportes.
- Posibilita la administración de seguridad de usuarios (21).

Por todo lo antes mencionado en la presente investigación se decidió utilizar Pentaho BI en su versión 3.8.

1.13.4 Apache Tomcat

Tomcat es un servidor de código abierto, un contenedor de aplicaciones web basadas en Java que fue creado para ejecutar Servlets y Java Server Page, (JSP por sus siglas en inglés), de aplicaciones web. Existe en el marco del subproyecto Apache Jakarta; es un proceso abierto y participativo, desarrollado en el medio ambiente y publicado bajo la licencia Apache versión 2. Apache Tomcat está destinado a ser una colaboración de los mejores desarrolladores de su clase de todo el mundo. En la presente investigación se decidió utilizar el servidor web Apache Tomcat en su versión 5.5.

1.14 Conclusiones del capítulo

Este capítulo abarca una panorámica general del proceso de desarrollo de los Almacenes de Datos, así como el estudio de las tendencias actuales en cuanto a la metodología de desarrollo, tecnologías y herramientas que se utilizarán en este Trabajo de Diploma. Luego de esta investigación se arribaron a las siguientes conclusiones:

- La metodología seleccionada: Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio en DATEC, cubre las fases por las que pasa la construcción de un Almacén de Datos y brinda diversas ventajas que facilitan su desarrollo; además está adaptada a las necesidades de la universidad, teniendo como base el Ciclo de vida Kimball.
- Las técnicas de captura de requisitos seleccionadas (entrevistas, discusiones y observaciones) permitirán la obtención de las necesidades del cliente, así como la definición de los requisitos del Mercado de Datos.
- El Lenguaje de Modelado Unificado en su versión 2.0 y el Visual Paradigm en su versión 8.0 (versión 3.4 de la suite del producto) permitirán la elaboración de los principales diagramas que formarán parte de la solución.

- El uso de PostgreSQL en su versión 9.1 como Sistema Gestor de Bases de Datos, así como el PgAdmin III en su versión 1.14 como herramienta de interfaz gráfica para la administración de los datos, posibilitarán la disponibilidad de las estructuras físicas para un correcto almacenamiento de la información.
- Las herramientas seleccionadas para los procesos de ETL, DataCleaner y Pentaho Data Integration en sus versiones 1.5.3 y 4.0.1 respectivamente, contribuirán a que los datos tengan la calidad requerida para ser cargados al Mercado de Datos.
- Las herramientas de BI escogidas (Pentaho Schema Workbench, Pentaho BI Server, Mondrian OLAP y Apache Jakarta Tomcat en sus versiones 3.2.1, 3.8, 3.0.4 y 5.5 respectivamente) permitirán la implementación de una capa de visualización al Mercado de Datos para mostrar los reportes que servirán de apoyo a la toma de decisiones.

CAPÍTULO II: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS SERIES HISTÓRICAS DE TURISMO PARA EL SISTEMA DE INFORMACIÓN DE GOBIERNO

2.1 Introducción

En este capítulo se realiza un estudio preliminar del negocio y de la organización, con el fin de obtener las reglas del negocio, identificar los requerimientos y los casos de uso con sus relaciones, describiendo brevemente los actores que van a interactuar con el sistema. Se define la arquitectura del Mercado de Datos y se diseñan los subsistemas de almacenamiento, de integración y visualización. Además se establece el esquema de seguridad a la hora de interactuar con la Base de Datos y la aplicación.

2.2 Caracterización de las áreas de la organización

La ONEI, organismo rector de las estadísticas en Cuba, tiene como principal objetivo analizar, almacenar y gestionar toda la información proveniente de diferentes áreas de la vida social con el fin de apoyar a la toma de decisiones en los principales sectores socioeconómicos. Debido a la gran cantidad de información que se procesa en dicha entidad, esta es agrupada en diversos departamentos o áreas, relacionado con un sector del país. Ejemplo de ello el Turismo, donde se obtienen y analizan, entre otras cosas, las series históricas de turismo, las cuales dependen de la información censada en un período de tiempo y son almacenadas en 12 archivos excel. A continuación se hace una descripción de los principales indicadores que se analizan en estos archivos:

- **Visitantes:** personas que visitan a un país diferente de aquel en el que tienen su lugar de residencia habitual por un período no superior a un año, cuyo motivo principal de la visita no es el de ejercer una actividad remunerada en el país visitado. Comprende dos categorías: turistas y excursionistas.
- **Llegadas:** los datos se refieren al número de llegadas de visitantes y no al número de personas. La misma persona, que puede efectuar varios viajes con destino a un país durante determinado período, será contada cada vez como una nueva llegada.
- **Viaje al extranjero (salidas):** se refiere al número de salidas al extranjero, es decir, a los desplazamientos que toda persona efectúa de su país de residencia habitual hacia otro país, por cualquier motivo que no sea el de ejercer una actividad remunerada en el país visitado.
- **Pernoctaciones:** es el número de noches que las personas pasan en los establecimientos de alojamiento.
- **Establecimiento:** es aquel que ofrece habitaciones amuebladas y donde se facilita el servicio de comidas y bebidas, además del alojamiento que es la actividad fundamental.

- **Habitaciones:** se refiere a la capacidad de alojamiento, representa el número total de habitaciones con que cuenta un centro de alojamiento, estén disponibles para su uso o no.
- **Plazas:** es la capacidad de alojamiento que tiene cada instalación, determinada por el número de camas con que cuentan las habitaciones en existencia.
- **Polo turístico:** lugar geográfico diseñado para exponer un gran conjunto de actividades coherentes, que permitan caracterizar en una agrupación de atractivos turísticos, bellezas y cuidados del entorno, infraestructura, equipamiento, servicios y organización, orientados a producir actividades en un ámbito turístico recreativo, para lograr la satisfacción al cliente.
- **Tasa de ocupación:** a través de este cálculo se obtiene un índice de ocupación o de utilización de la instalación.
- **Ingresos en divisas asociados al turismo:** se definen como los gastos efectuados en el país de acogida por los visitantes internacionales, incluido el pago de sus transportes internacionales a las compañías nacionales de transportación.
- **Bases de campismo:** es el lugar donde se brinda a los participantes las condiciones materiales y organizativas mínimas para acampar y desarrollar sus actividades.

2.3 Necesidades de los usuarios

Debido a que es de esencial importancia conocer qué desea y necesitan los clientes para el buen desarrollo de la solución, se realizaron reuniones con la especialista del área de turismo en la ONEI con el fin de analizar la información referente a dicha área. Las necesidades de información se obtuvieron mediante un estudio donde se definieron los objetivos que persigue la organización y los indicadores relacionados con los datos de las series históricas de turismo, tanto del año en curso como de los años anteriores. La información que se analizará está relacionada con los Visitantes, los Turistas, las Salidas, los Alojamientos, los Polos Turísticos, la Capacidad Hotelera, la Tasa de Ocupación, los Ingresos en Divisas y las Bases de Campismo.

2.4 Reglas del negocio

El principal objetivo de las reglas del negocio es especificar las políticas o condiciones que deben ser cumplidas a lo largo del desarrollo de la solución para lograr el adecuado funcionamiento del sistema. Debido a que regulan aspectos del sistema, la identificación de estas es de vital importancia. A continuación se muestran una de las 17 reglas del negocio identificadas en los datos de las series históricas de turismo de la ONEI, con el fin de homogenizar y estructurar la información para facilitar su análisis:

RN17: Los indicadores cuya unidad de medida sea unidad (U) o Miles, se mostrarán con decimal igual a cero, excepto las medidas *tasa_ocupacion_media*, *total_tasa_ocupacion*, *cant_ingresos_divisas*, que se mostrarán con un lugar decimal.

Las otras reglas del negocio se encuentran detalladas en el artefacto Reglas del Negocio del Expediente de Proyecto.

2.5 Especificación de requerimientos

2.5.1 Requerimientos de información

Los requerimientos de información deben estar disponibles para el usuario final a la hora de este realizar las consultas necesarias para analizar los datos, con el objetivo de apoyar la toma de decisiones en la empresa. A continuación se especifican los requerimientos de información que fueron identificados después de haber realizado un análisis minucioso sobre las series históricas del área de turismo de la ONEI:

RI1- Obtener la cantidad de llegadas de visitantes internacionales por tipo de visitante, región, meses, países y tiempo.

RI2- Obtener la cantidad de llegadas de turistas por medio de transporte, motivo de visita y tiempo.

RI3- Obtener la cantidad de salidas por tiempo.

RI4- Obtener la cantidad de llegadas de turistas internacionales por establecimiento y tiempo.

RI5- Obtener la cantidad de pernoctaciones de turistas internacionales por establecimiento y tiempo.

RI6- Obtener la cantidad de pernoctaciones de turistas nacionales por establecimiento y tiempo.

RI7- Obtener el total de establecimientos por establecimiento de servicio de alojamiento y tiempo.

RI8- Obtener el total de habitaciones por establecimiento de servicio de alojamiento y tiempo.

RI9- Obtener el total de plazas camas por establecimiento de servicio de alojamiento y tiempo.

RI10- Obtener el total de establecimientos de los polos por polo turístico y tiempo.

RI11- Obtener el total de habitaciones de los polos por polo turístico y tiempo.

RI12- Obtener el total de plazas camas de los polos por polo turístico y tiempo.

RI13- Obtener el número de establecimientos por establecimiento seleccionado y tiempo.

RI14- Obtener el número de habitaciones por establecimiento seleccionado y tiempo.

RI15- Obtener el número de plazas camas por establecimiento seleccionado y tiempo.

RI16- Obtener el total de tasa de ocupación media anual por tiempo.

RI17- Obtener la tasa de ocupación media anual por hotel y tiempo.

RI18- Obtener la cantidad de ingresos en divisas por ingreso y tiempo.

RI19- Obtener la cantidad de indicadores de campismo por indicador de campismo y tiempo.

2.5.2 Requerimientos funcionales

Los requerimientos funcionales representan las capacidades y condiciones que el sistema debe cumplir para dar respuesta a los requerimientos de información. A continuación se muestran los requerimientos funcionales que fueron identificados:

RF1 - Autenticar usuario.

RF2 - Adicionar roles.

RF3 - Eliminar roles.

RF4 - Adicionar usuarios.

RF5 - Eliminar usuarios.

RF6 - Insertar reportes.

RF7 - Modificar reportes.

RF8 - Eliminar reportes.

RF9 - Extraer información.

RF10 - Realizar transformación y carga.

RF11 - Abrir navegador OLAP.

RF12 - Mostrar editor MDX.

RF13 - Mostrar padres.

RF14 - Ocultar repeticiones.

RF15 - Intercambiar ejes.

RF16 - Mostrar gráfico.

RF17 - Configurar gráfico.

RF18 - Configurar impresión.

RF19 - Exportar a PDF.

RF20 - Exportar a excel.

RF21 - Mostrar propiedades.

RF22 - Suprimir filas.

RF23 - Detallar miembros.

RF24 - Entrar en detalles.

RF25 - Mostrar datos de origen.

2.5.3 Requerimientos no funcionales

Los requerimientos no funcionales son las propiedades o cualidades que el sistema debe de cumplir. Estos determinan como debe comportarse el producto. En la presente investigación fueron definidos

17 requerimientos no funcionales los cuales se encuentran contemplados en el artefacto DATEC-SIGOBMDSeries_Turismo-0113_ERSv1.1, ubicado en el Expediente de Proyecto. A continuación se muestran los requerimientos no funcionales de usabilidad:

➤ **Usabilidad**

- RNF 1.** Cumplir con las pautas de diseño de las interfaces.
- RNF 2.** Mostrar los mensajes, títulos y demás textos que aparezcan en la interfaz del sistema en idioma español e inglés.
- RNF 3.** Establecer tiempo de entrenamiento requerido para que usuarios normales sean productivos operando el sistema.
- RNF 4.** Asegurar la disponibilidad del sistema y la recuperación ante un fallo.
- RNF 5.** Garantizar la conexión de múltiples usuarios al mismo tiempo.

2.6 Casos de uso del sistema

Luego de definidos los requerimientos de información y funcionales se agruparon en los Casos de usos (CU) correspondientes, identificando además los actores encargados de inicializar tales CU. Todo esto permitió elaborar el Diagrama de Casos de Uso del Sistema (DCUS) compuesto por diez Casos de Uso de Información (CUI), siete Casos de Uso Funcionales (CUF) y tres actores (ver figura 3). A continuación se realiza una descripción de las responsabilidades de cada uno de los actores que forman parte del DCUS.

2.6.1 Actores del sistema

Actor	Responsabilidad
Administrador	Inserta y elimina los usuarios del sistema. Inserta y elimina los roles asignados a los usuarios del sistema. Inserta, modifica y elimina los reportes disponibles en el sistema. Consulta información de los reportes.
Administrador de ETL	Realiza la extracción, transformación y carga de los datos de los ficheros fuentes.
Especialista	Consulta la información de los reportes.

Tabla 3: Actores del sistema

2.6.2 Diagrama de Casos de Uso del Sistema

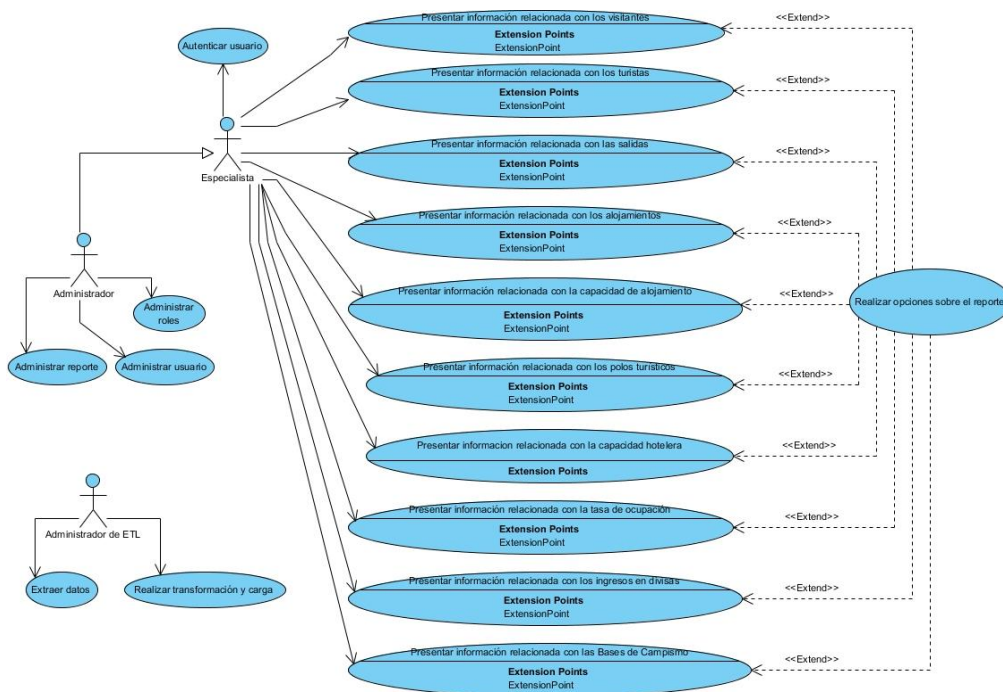


Figura 3: Diagrama de Casos de Uso del Sistema

2.6.3 Especificación de Casos de Uso del Sistema

CUI 1. Presentar información relacionada con los visitantes

Objetivo	Presentar información relacionada con los visitantes
Actores	Especialista, Administrador
Resumen	El caso de uso inicia cuando el Especialista o el Administrador desean hacer un análisis de la información relacionada con los visitantes desde diferentes perspectivas. El Especialista o el Administrador seleccionan el reporte que desea ver, el sistema muestra la información contenida en él y las opciones de los posibles cambios que le puede hacer al reporte. El caso de uso finaliza cuando el Especialista o el Administrador terminan el análisis de la información relacionada con los visitantes.
Complejidad	Alta
Prioridad	Media
Precondiciones	El usuario se autenticó correctamente. Los datos correspondientes fueron cargados en el mercado de datos. Los reportes relacionados con los visitantes fueron creados.

Postcondiciones	Los reportes correspondientes al caso de uso fueron consultados.	
Flujo de eventos		
Flujo básico Presentar información relacionada con los visitantes		
	Actor	Sistema
1	Selecciona el Área de Análisis General A.A.G SIGOB	
2		Muestra las Área de Análisis contenidas en el A.A.G SIGOB.
3	Selecciona el Área de Análisis A.A. Series de turismo	
4		Muestra los Libros de Trabajo contenidos en el A.A. Series de turismo.
5	Selecciona el Libros de Trabajo L.T 01-Visitante	
6		Muestra los reportes contenidos en el L.T 01- Visitante.
7	Selecciona el reporte que desea analizar	
8		Muestra la información contenida en el reporte seleccionado y brinda opciones al actor para hacer cambios al reporte durante su análisis. Ir al CU Realizar opciones sobre el reporte. Finaliza el CU.
Opciones del reportes Presentar información relacionada con los visitantes		
	Perspectivas de análisis	Posibles resultados
		Medidas Periodicidad
	Variables de entrada relacionadas con el CU Presentar información relacionada con los visitantes: <ul style="list-style-type: none"> • visitante • dim_temporal_mes 	Variables de salida disponibles en el hecho Visitantes: <ul style="list-style-type: none"> • cant_llegadas_visitantes_int Rango de tiempo en que se solicitan las variables de salida: <ul style="list-style-type: none"> • Anual

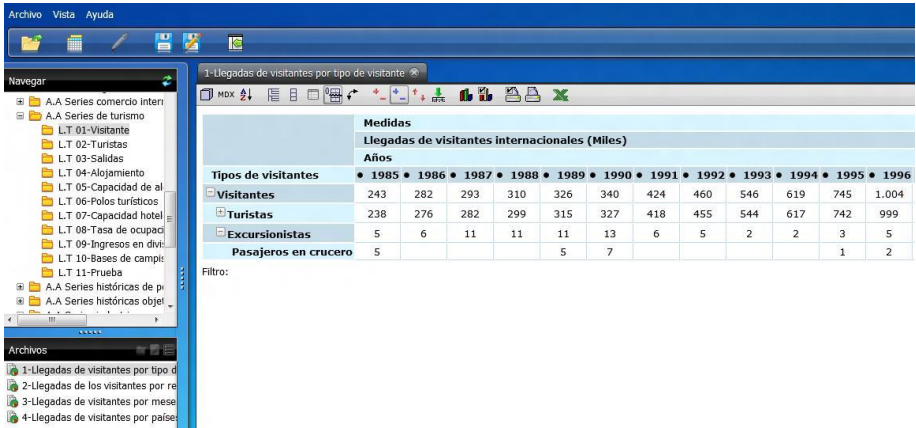
<p>Prototipo de interfaz de usuario:</p>		
<p>Relaciones</p>	<p>CU Incluidos</p>	<p>No aplica.</p>
	<p>CU Extendidos</p>	<p>Realizar opciones sobre el reporte: Paso 8 del Flujo Básico. Realizar opciones sobre el reporte en el CU Presentar información relacionada con los visitantes.</p>
<p>Requisitos no funcionales</p>	<p>RNF1, RNF2, RNF3, RNF4, RNF5, RNF6, RNF7, RNF8, RNF9, RNF10, RNF11, RNF12, RNF13, RNF14, RNF15.</p>	
<p>Asuntos pendientes</p>	<p>No aplica</p>	

Tabla 4: Especificación del Caso de Uso Presentar información relacionada con los visitantes

El resto de las descripciones de los Casos de Uso se encuentran en el artefacto DATEC-SIGOBMDSeries_Turismo-0114_ECU, en la sección “3.2 Requisitos no funcionales”.

2.7 Definición de la arquitectura del Mercado de Datos

La arquitectura general del MD Series históricas de turismo está compuesta por las fuentes de datos (archivos excel) y por tres subsistemas fundamentales: el subsistema de integración, el subsistema de almacenamiento y el subsistema de visualización como se muestra en la figura siguiente:

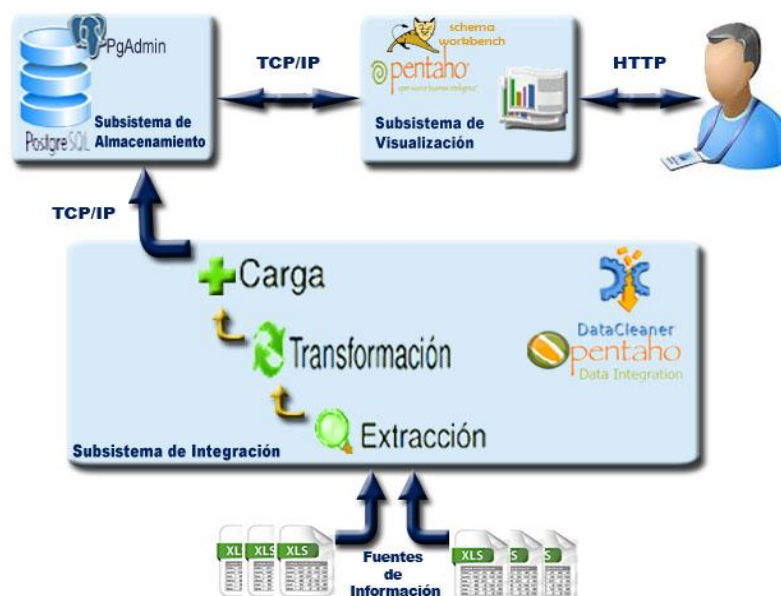


Figura 4: Arquitectura general de MD Series históricas de turismo

A continuación se explicarán cada uno de los subsistemas que definen la arquitectura del MD:

- Subsistema de integración: se encarga de integrar, estandarizar y limpiar la información de la fuente de datos, con el fin de cargarla hacia el almacén, a través de los procesos de extracción, transformación y carga de los datos.
- Subsistema de almacenamiento: es el encargado de almacenar toda la información correspondiente al área específica del MD. El almacén estará compuesto por dimensiones y tablas de hechos, que a su vez contendrán los datos que describirán un hecho, así como las medidas.
- Subsistema de visualización: tiene como objetivo principal consultar la información almacenada en el AD permitiendo mostrar los reportes que necesitan los clientes.

2.8 Diseño de la solución

2.8.1 Diseño del subsistema de almacenamiento

Para el desarrollo y el correcto funcionamiento del MD, en esta etapa se realiza el modelo dimensional el cual contiene las tablas de hechos identificadas en el negocio, las dimensiones seleccionadas para la solución y las relaciones que existen entre estas.

Dimensiones

Las dimensiones representan las características de un hecho, que permite su análisis posterior en el proceso de toma de decisiones y brinda una perspectiva adicional a un hecho dado. Una de las características principales que poseen las dimensiones es la definición de jerarquías entre sus atributos, las cuales tienen como fin plasmar la forma en que se puede consolidar la realización del proceso de análisis, ya sea mediante el uso de sumas, porcentos, máximos, mínimos, entre otros. La definición correcta de los atributos de las dimensiones es de vital importancia para la realización del MD, debido a que estos sirven como fuente primaria de las restricciones de las consultas, agrupaciones y las etiquetas de los reportes. A continuación se muestra un ejemplo de una de las dimensiones que contiene la solución, las otras se encuentran especificadas en el artefacto del Expediente de Proyecto de nombre Especificaciones del Modelo de Datos Dimensional.

Dimensión **dim_visitante**: dimensión que describe los tipos de visitantes que se analizan en las series históricas de turismo.

mart_turismo_series.dim_visitante	
dim_visitante_id	int4
tipo_visitante	varchar(255)
tipo_excursionista	varchar(255)
tipo_excursionista_cod	int4
tipo_visitante_cod	int4

Figura 5: Dimensión dim_visitante

Tablas de hechos

Las tablas de hechos pueden representar un objeto o evento del negocio que es utilizado por los analistas de la información para analizar el comportamiento de los datos de la empresa. Además, estas tablas contendrán las medidas numéricas las cuales simbolizan las variables de salida del almacén. A continuación se muestra un ejemplo de una de las tablas de hechos que contiene la solución, las otras se encuentran especificadas en el artefacto Especificaciones del Modelo de Datos Dimensional.

Hecho **hech_visitantes**: este hecho recoge la información relacionada con la cantidad de llegadas internacionales de los visitantes al país.

mart_turismo_series.hech_visitantes	
dim_visitante_id	int4
dim_temporal_anno_id	int4
cant_llegadas_visitantes_int	int4

Figura 6: Tabla de hecho hech_visitantes

Matriz bus o matriz dimensional

La Matriz bus es una herramienta potente para la planificación y la comunicación, pues permite definir la arquitectura general de los datos para el MD, y determinar el impacto que provocaría durante el desarrollo del sistema un cambio a la solución. En ella se describen las relaciones que existen entre las tablas de hechos y las dimensiones. Las columnas de la matriz representan las dimensiones utilizadas en el MD y las filas, los hechos identificados. Las celdas sombreadas con una X indican que la columna de dimensión está relacionada con la fila del proceso de negocio. Esto posibilita ver de inmediato cuáles son las dimensiones que merecen una atención especial debido a su participación o relación con múltiples hechos. En la siguiente tabla se muestra la Matriz bus para el MD Series históricas de turismo:

Hechos/ Dimensiones	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14
H1		X	X											
H2		X			X									
H3	X	X												
H4		X		X										
H5		X												
H6		X				X								
H7		X					X							
H8		X						X						
H9		X							X					
H10		X								X				
H11		X									X			
H12		X										X		
H13		X											X	
H14		X												X

Tabla 5: Matriz bus

Leyenda:

➤ **Hechos**

H1: hech_visitantes

H2: hech_visitantes_region

H3: hech_visitantes_meses

H4: hech_visitantes_pais

H5: hech_salidas
H6: hech_turistas_medio_transporte
H7: hech_turistas_motivo_visita
H8: hech_alojamiento
H9: hech_capacidad_alojamiento
H10: hech_polos_turisticos
H11: hech_capacidad_hotelera
H12: hech_ocupacion_media
H13: hech_ingresos_divisas
H14: hech_bases_campismos

➤ **Dimensiones**

D1: dim_temporal_mes
D2: dim_temporal_anno
D3: dim_visitante
D4: dim_país
D5: dim_región
D6: dim_medio_transporte
D7: dim_motivo_visita
D8: dim_establecimiento
D9: dim_establecimiento_cap_aloj
D10: dim_polo_turístico
D11: dim_establecimiento_cap_hot
D12: dim_hotel
D13: dim_ingreso
D14: dim_indicador_campismo

Modelo de datos

El principal objetivo de realizar el modelo dimensional es para que los datos del negocio queden representados en una estructura lógica, evidenciando las relaciones que existen entre las tablas de hechos y las dimensiones. En la figura 7, se muestra el modelo dimensional propuesto para el desarrollo del MD Series históricas de turismo en el cual se observa la relación existente entre las tablas de hechos propuestas y las dimensiones definidas para la solución, evidenciándose como topología de esquema Constelación de hechos.



Figura 7: Modelo dimensional

2.8.2 Diseño del subsistema de integración

Registro de sistemas fuentes

En el Expediente de Proyecto se generó un artefacto llamado Registro de Sistema Fuente con el fin de describir las fuentes de datos correspondientes a las Series históricas de turismo, para de esta forma lograr una mayor comprensión de la información contenida en las mismas. Este documento tiene como objetivo contribuir a la documentación de los sistemas fuentes, enfocado a la documentación del estado físico de los datos y de los responsables de los sistemas fuentes en cuestión. La información de las series históricas se encuentra en 12 ficheros excel, debido a que la entidad no cuenta con un SGBD que almacene y gestione dicha información. Estos son utilizados por el departamento de

Turismo de la ONEI, en la cual labora diariamente una persona que es la única que puede consultar las informaciones manejadas por las empresas referentes al turismo en el país. El tamaño bruto de la fuente de datos es de 384kb.

Diccionario de datos

El diccionario de datos sirve de apoyo para los procesos de ETL, debido a que este contiene la información necesaria para un correcto entendimiento de los sistemas fuentes, enfocado principalmente en la documentación de las variables con las que interactúa la fuente de datos. En él se describen cada una de estas variables especificando el significado que tienen en el negocio y los posibles valores que pueden tomar. Todo esto es recogido en el artefacto Diccionario de datos que se encuentra en el Expediente de Proyecto. De forma general se identificaron 14 variables con las que interactúa la fuente de datos, que corresponden a cada una de las dimensiones del MD.

Perfilado de datos

El perfilado de los datos es el proceso que se encarga de analizar las fuentes de datos con el objetivo de entender su contenido, estructura, calidad y dependencia, para así poder definir las transformaciones que se le deben realizar a los mismos. Todo esto se evidencia en el artefacto Perfil de los datos del Expediente de Proyecto, en el cual se hace referencia a la ubicación física de la fuente de datos, el tipo de formato que tiene, y se explica qué datos se analizan en cada una de las tablas implicadas en dicha fuente. Además, en este documento se realiza una descripción específica de las tablas, mostrando importantes resultados con los que se establecen reglas, como son los valores nulos, los distintos, los únicos, el mínimo y el máximo entre otros, los cuales fueron obtenidos mediante la herramienta DataCleaner. En la figura 8 se muestra el resultado del perfilado de los datos realizado al fichero excel 15.5.



Figura 8: Perfilado de datos del fichero excel 15.5

Diseño general de las transformaciones

Las transformaciones en los procesos de ETL son una colección de pasos, que al ser ejecutadas permiten que los datos sean cargados correctamente en las tablas que se encuentran en la BD del MD. Después de haber realizado el perfilado de los datos se definieron las transformaciones que se llevarán a cabo para poblar el MD Series históricas de turismo. En el diseño general que se siguió para implementar cada una de ellas, lo primero que se realiza es la extracción de los datos desde la fuente de datos (excel), obteniéndose diferentes indicadores por años y los valores numéricos de estos indicadores. Posteriormente se procede a la limpieza y transformación de los datos, aplicando las reglas de transformación necesarias. Luego se busca el identificador correspondiente en las tablas de dimensiones. Además, con el objetivo de que los datos tengan la calidad requerida, se validan los tipos de datos; si existen problemas con la validación se envían, hacia un excel de errores, la columna que fue validada y una descripción del error que se produjo; sino, se insertan o actualizan los datos en la tabla de hecho correspondiente. También, se obtiene información del sistema, necesaria para generar los metadatos del proceso: nombre del fichero fuente, nombre de la transformación y fecha de ejecución de la transformación. Por último se inserta o actualiza dicha información en la tabla md_hist del MD Series históricas de turismo. (Ver figura 9)

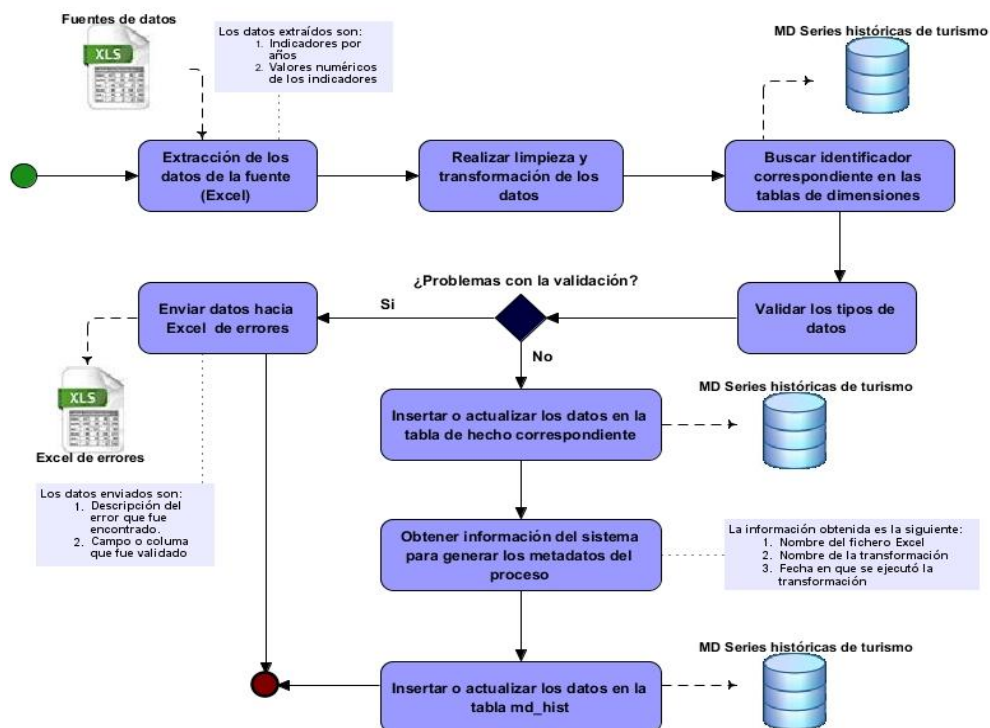


Figura 9: Diseño general de la implementación de las transformaciones

Mapa lógico de datos

Para lograr un correcto entendimiento del contenido, estructura y dependencia de los datos se encuentra en el Expediente de Proyecto el artefacto Mapa lógico de datos. Este documento tiene como principal objetivo describir los datos correspondientes a los hechos y dimensiones que han sido identificadas en el MD Series históricas de turismo.

2.8.3 Diseño del subsistema de visualización

Arquitectura de información

La Arquitectura de información permite tener una mejor visualización respecto a los elementos que estructuran el sistema, así como los que serán mostrados en la capa del MD. En el Expediente de Proyecto se encuentra el artefacto Arquitectura de información, el cual tiene como objetivo fundamental ofrecer un entorno de análisis, monitoreo y control de la información de las Series históricas de turismo de la ONEI para el apoyo al proceso de toma de decisiones. En este documento se identificó un Área de Análisis (A.A), que contiene diez Libros de Trabajo (LT) y 16 reportes agrupados en los libros de trabajo. En la figura 10, se detallan los elementos que componen las estructuras de navegación de la información presentada en la capa de visualización:

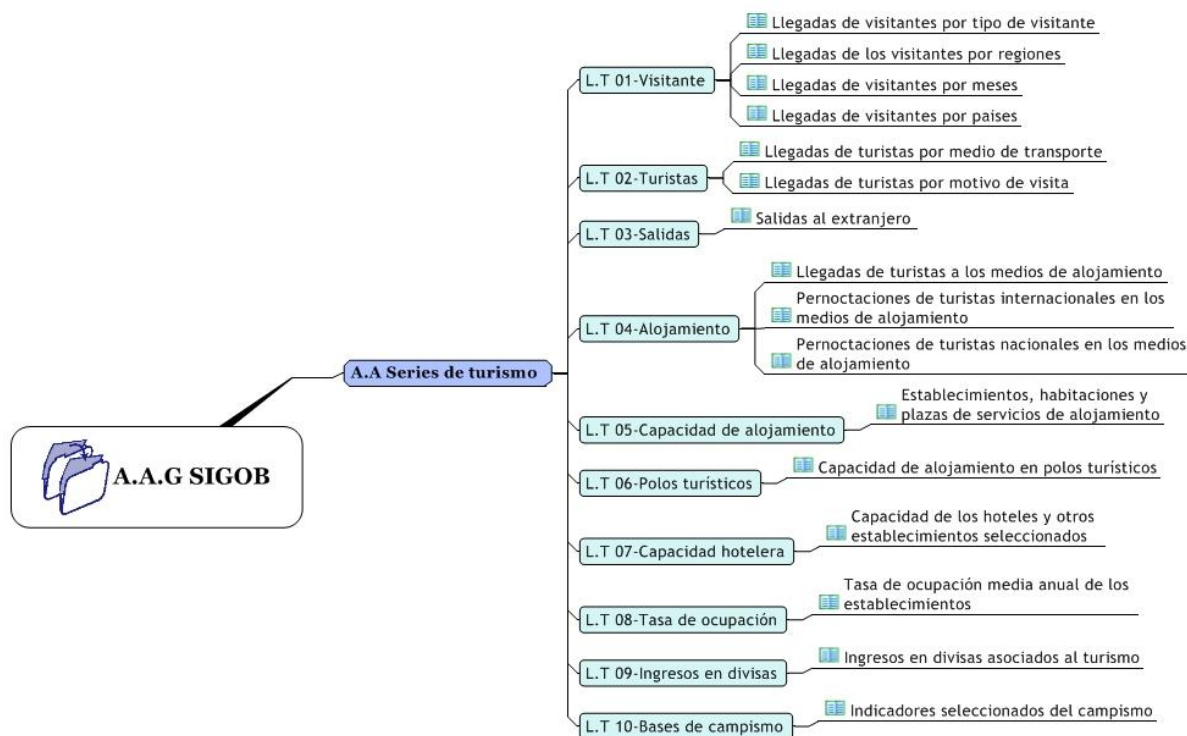


Figura 10: Estructura de navegación del MD Series históricas de turismo

Diseño de los reportes candidatos

En el Expediente de Proyecto existe el artefacto Reportes candidatos en el cual se describen los reportes candidatos que fueron identificados en el desarrollo del MD Series históricas de turismo, para así tener una mejor comprensión de los elementos que lo componen. A continuación se muestra la descripción de uno de los reportes:

Área de análisis (AA)	Series de turismo
Libro de Trabajo (LT)	LT1- Visitante
Reporte (Tabla de Salida – TS)	TS1- Llegadas de visitantes por tipo de visitante
Descripción	El reporte muestra las llegadas de los visitantes internacionales por tipo de visitante y años.
Elementos del reporte	<ul style="list-style-type: none"> • Años • Tipos de visitantes
Frecuencia de emisión	Anual

Tabla 6: Descripción del reporte Llegadas de visitante por tipo de visitante

Diseño de los cubos OLAP:

El diseño de los cubos OLAP se realizó a través de la herramienta Pentaho Schema Workbench, en la cual se definen los hechos identificados, las dimensiones que corresponden a cada uno de estos hechos con sus niveles de jerarquía, así como la conexión a la BD que contiene los datos para el cubo multidimensional. En la presente investigación se modelaron 14 cubos y 14 dimensiones, correspondientes a cada una de las tablas de hechos y dimensiones respectivamente. A continuación se muestra, en la figura 11, un ejemplo del diseño de uno de los cubos modelados (Visitante) con sus dimensiones (dim_visitante y dim_temporal_anno) y medida:



Figura 11: Diseño de los cubos OLAP a través de la herramienta Schema Workbench

2.9 Política de respaldo y recuperación

En el MD Series históricas de turismo se utiliza una política de respaldo y recuperación sólida, midiéndose en dos aspectos fundamentales:

- Periodicidad de las salvas: las salvas de toda la información contenida en la BD se realizan anualmente, verificando en todo momento que exista una copia escrita de la información almacenada en el servidor. Las tablas que se involucran en este proceso son las catorce tablas de hechos y las 12 tablas de dimensiones identificadas en el proceso de análisis.
- Salvas existentes: a pesar de que actualmente no existen salvas en esta área, se prevee la realización de reemplazos de estas anualmente.

2.10 Esquema de seguridad

La seguridad en el MD Series históricas de turismo está dada por los niveles de acceso al sistema, regida fundamentalmente por los permisos y roles que los usuarios tienen a la hora de interactuar con la BD y la aplicación (ver tabla 7). Para la seguridad en la BD se definió el rol Administrador de BD el cual posee total acceso a la BD del sistema y tiene la responsabilidad de llevar a cabo la política de respaldo y recuperación definida en el epígrafe anterior. De igual manera, se definió el rol Administrador de ETL, el cual tiene permisos de lectura y escritura sobre la BD, pues es el encargado de realizar los procesos de integración. Para la seguridad en la aplicación se definieron los roles Administrador y Analista, el primero tiene acceso total al A.A.G SIGOB y administra los usuarios, roles y reportes; el segundo tiene acceso de solo lectura al A.A Series de turismo.

Roles	Permisos
Administrador de BD	Total acceso a la BD. Lleva a cabo la política de respaldo y recuperación.
Administrador de ETL	Permisos de lectura y escritura sobre la BD.
Administrador	Acceso total al A.A.G SIGOB. Administra los usuarios, roles y reportes.
Analista	Tiene acceso de sólo lectura al A.A Series de turismo.

Tabla 7: Roles y permisos

Elementos de aplicación	Roles con acceso
A.A.G SIGOB	Administrador
A.A Series de turismo	Administrador y Analista

Tabla 8: Nivel de acceso a los elementos de aplicación

2.11 Conclusiones del capítulo

Después de haber realizado el análisis y diseño del Mercado de Datos Series históricas de turismo se arribaron a las siguientes conclusiones:

- El levantamiento de requisitos arrojó como principal resultado la identificación de 19 requerimientos de información, 25 requisitos funcionales y 17 no funcionales, sirviendo de base para elaborar el diagrama de Casos de Uso del Sistema. De igual manera, se identificaron 17 reglas del negocio, permitiendo definir dos reglas de transformación que serán utilizadas en los procesos de integración. Por último, se realizó la descripción de los casos de uso del Mercado de Datos para especificar cada una de las funcionalidades del sistema.
- La arquitectura base del Mercado de Datos definida, permitió identificar los elementos y subsistemas que están implicados en el desarrollo de la solución.
- El modelo dimensional diseñado, representa las relaciones entre las 14 tablas de hechos y las 14 dimensiones identificadas para el Mercado de Datos.
- Durante el diseño de los subsistemas de integración y visualización, quedó definido el diseño general para las transformaciones, la arquitectura de información del Mercado de Datos, el diseño de los reportes candidatos y de los cubos OLAP, todo esto servirá de guía para la implementación de dichos subsistemas.
- A través de las políticas de respaldo y recuperación definidas, se definió que la periodicidad que van a tener las salvadas es anual.
- Los roles y permisos definidos en el Mercado de Datos, contribuirán a la seguridad de la aplicación.

CAPÍTULO III: IMPLEMENTACIÓN Y PRUEBAS DEL MERCADO DE DATOS SERIES HISTÓRICAS DE TURISMO PARA EL SISTEMA DE INFORMACIÓN DE GOBIERNO

3.1 Introducción

En este capítulo se hace referencia a la implementación de la solución, abordando específicamente cómo se realiza la implementación del subsistema de almacenamiento, de integración y de visualización para las series históricas del área de turismo del Sistema de Información de Gobierno, teniendo en cuenta los requerimientos y necesidades del negocio. De igual forma, hace referencia a las pruebas, mediante la utilización de las listas de chequeo, para determinar que los artefactos de documentación de los procesos de Extracción, Transformación y Carga tengan la calidad requerida, y de los casos de pruebas, basados en Casos de usos y reglas de transformación, para validar los reportes del Mercado de Datos y el cumplimiento de las reglas del negocio respectivamente.

3.2 Implementación del subsistema de almacenamiento

3.2.1 Estructura de los datos

En la base de datos, los datos se encuentran organizados en estructuras lógicas que facilitan la correcta manipulación de los mismos. Estas estructuras son denominadas esquemas y tablas.

Esquemas

Los esquemas en una base de datos representan una forma de organizar la información contenida en la misma. Dentro de los esquemas se pueden encontrar funciones, operadores y tipos de datos que facilitarán su implementación. En el presente trabajo se definieron tres esquemas los cuales serán explicados a continuación:

Esquema dimensiones: contiene las tablas de las dimensiones generales del AD, y de ellas se utilizaron las necesarias para implementar el MD.

Esquema mart_turismo_series: contiene las tablas de hechos y las tablas de dimensiones propias del MD Series históricas de turismo.

Esquema metadatos: contiene las tablas para la captura de los metadatos de los procesos de ETL.

Tablas

Con el diseño del modelo físico se genera el script de la base de datos, y a partir de este se concluye que la solución propuesta tiene 31 tablas en total, de ellas 26 contenidas en el esquema mart_turismo_series, dos contenidas en el esquema dimensiones y tres en el esquema metadatos. En la figura 12 se muestra la estructura física del MD.

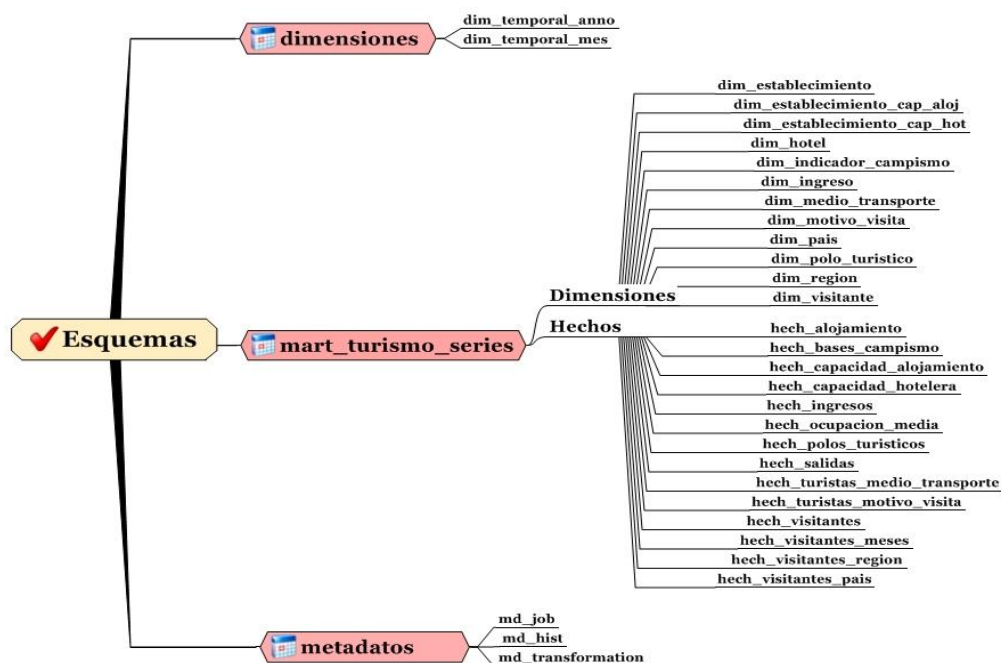


Figura 12: Estructura física del MD Series históricas de turismo

3.2.2 Estándares de codificación

Con el propósito de lograr un entendimiento entre todas las partes implicadas en un proyecto, son utilizados los estándares de codificación. Es por esto, que en el departamento de Almacenes de Datos del centro DATEC, se creó un artefacto que refleja los estándares a seguir para el desarrollo de un producto. En el MD Series históricas de turismo fueron utilizados algunos de esos estándares, los cuales son explicados en la siguiente tabla:

Tipo de objeto	Función	Nomenclatura	Descripción
Esquema	Dimensiones compartidas	Dimensiones	Esquema donde se organizan las dimensiones compartidas por varios MD (tablas dimensionales y secuencias).
	Esquemas de datos	mart_[temática]	Esquema donde se almacenan las tablas de hechos y vistas materializadas definidas para gestionar los datos asociados a cada área temática.

Tablas	Dimensiones	dim_[nombre]	Tablas de dimensiones utilizadas.
	Hechos	hech_[nombre]	Tablas de hechos que definen las principales medidas requeridas para calcular indicadores y otras medidas derivadas.
Constraints	Claves Primarias	pk_[tabla]_id	Clave primaria.
	Clave Foráneas	fk_[tablaFuente]_id	Clave foránea.

Tabla 9: Estándares de codificación

3.2.3 Configuración de la seguridad de la base de datos

Con el objetivo de restringir el acceso a la BD y de este modo tener control sobre la seguridad de la misma, fue definido el rol Administrador de la BD mediante la herramienta de interfaz gráfica PgAdmin. Este rol posee todos los permisos sobre la BD y tiene definido una contraseña para acceder a ella. A continuación se muestra, en la figura 13, cómo se adicionó dicho rol.

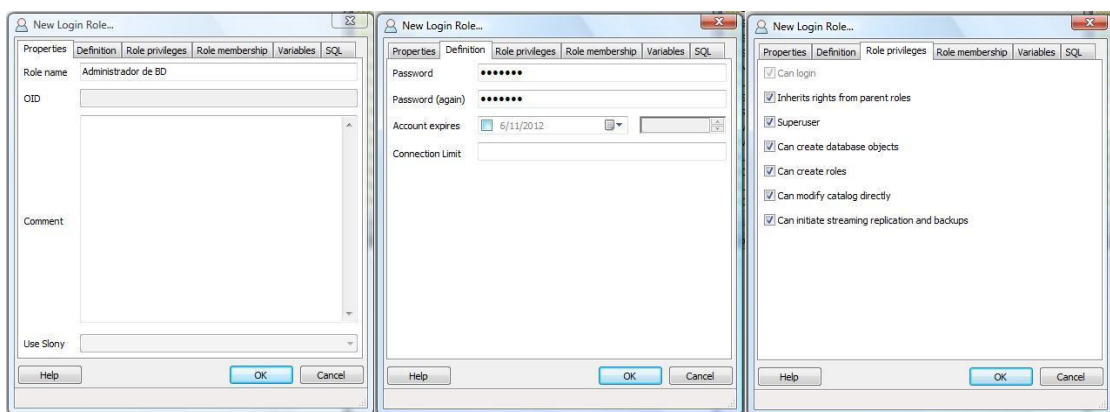


Figura 13: Adicionar el rol Administrador de la BD

3.3 Implementación del subsistema de integración de datos

3.3.1 Subsistema de extracción

En el subsistema de extracción se obtienen todos los datos que poblarán el MD Series históricas de turismo, a través de las fuentes de información que proporcionó el cliente. Estos datos fueron extraídos de 12 ficheros excel con el objetivo de adaptarlos al modelo relacional establecido anteriormente. Cada uno de los ficheros contiene la información histórica referente a las series de turismo analizadas por los especialistas de la ONEI desde el año 1985 hasta el año 2010.

3.3.2 Limpieza y transformación de datos

Posterior a la extracción de los datos se realizó la limpieza de los mismos. En este proceso se detectaron posibles errores e incoherencias que poseía la fuente dándole el tratamiento adecuado para corregirlos, con el fin de que los datos que se cargaron a la BD fuesen precisos, completos y consistentes, para que de esta manera se pudieran interpretar de forma correcta a la hora de ser presentados al cliente. Luego de la limpieza, se realizaron transformaciones a los datos aplicando las reglas del negocio. Con esto se logró que la información tuviese la calidad requerida por el usuario final, pues se eliminaron datos duplicados, se detectaron y corrigieron errores ortográficos y se trataron los posibles valores nulos, quedando los datos listos para la carga.

3.3.3 Transformaciones y trabajos

En el MD Series históricas de turismo se implementaron un total de 26 transformaciones, 12 para la carga de las dimensiones y 14 para la carga de los hechos. La figura 14, mostrada a continuación, refleja los pasos lógicos que se siguieron para la carga de la tabla `hech_visitantes`. Primeramente se hizo la extracción de los datos a partir del fichero de formato excel con nombre 15.1, el cual recoge la información relacionada con la cantidad de llegadas de visitantes internacionales al país. Luego, utilizando el componente String operations, se realizó la limpieza de los datos, donde fueron eliminados los espacios a ambos lados de los valores de cada fila en el flujo. Además, a través de los componentes Normalizar filas y Des-normalizar se transformó el flujo de entrada de los datos, de manera que posibilitara la resta de valores numéricos utilizando el componente Calculadora, permitiendo calcular un nuevo campo que era necesario para la sumatoria total. En los pasos posteriores, se descartaron las filas que no se deseaban en el flujo con el componente Filtrar filas, se realizó el tratamiento a los valores nulos utilizando el componente Mapeo de valores y se hizo la búsqueda en las tablas de dimensiones `dim_visitante` y `dim_temporal_anno` para obtener los identificadores de los valores de cada una de ellas. Después que el flujo estuvo preparado con la estructura correspondiente a la tabla de hecho en la BD, fueron validados los tipos de datos, y de haber errores en la validación, se capturaron estos en un archivo excel, con la finalidad de que puedan ser revisados por el cliente. Luego de comprobar que los datos estuviesen correctos, se procedió a realizar la carga de los mismos hacia la tabla `hech_visitantes`. Por último, se utilizaron pasos para obtener datos referentes a las fuentes, como por ejemplo, el nombre del fichero desde el cual se extraen los datos, el nombre de la transformación que se ejecutó y la fecha en la que se hizo la ejecución, con el fin de cargarlos hacia una tabla de metadatos en la BD, que garantizará la gestión histórica de dichos datos.

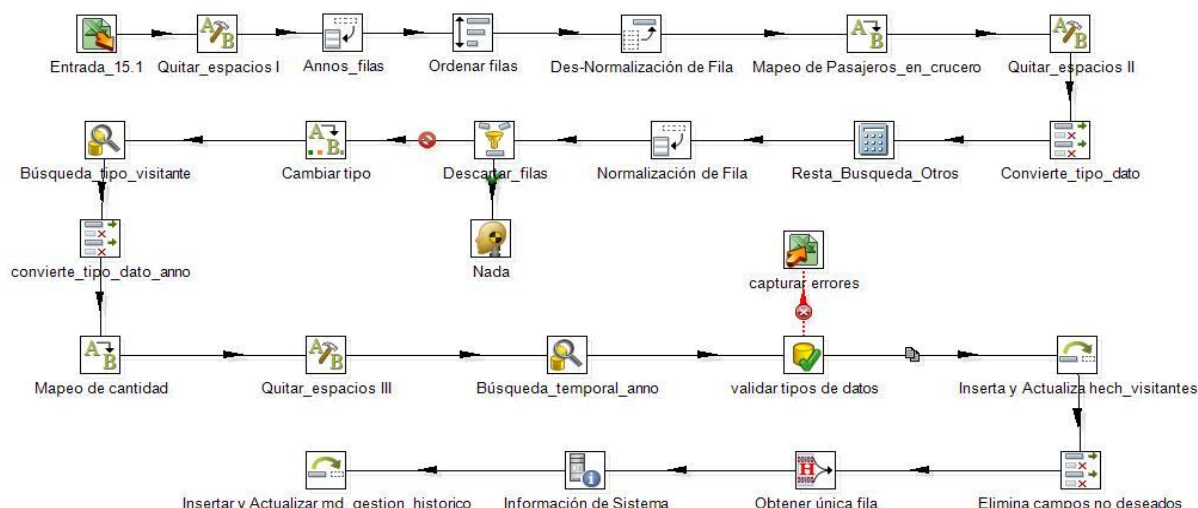


Figura 14: Transformación “transf_hech_visitante” para cargar la tabla hech_visitantes

Con el objetivo de lograr una estandarización en el desarrollo de las transformaciones se utilizó la siguiente nomenclatura:

Cuando se agregó una nueva transformación, en el proceso de integración de datos, para cargar los hechos se nombró de la siguiente manera:

- transf_hech_<nombre>

Ejemplo: transf_hech_visitantes.ktr

Cuando se agregó una nueva transformación, en el proceso de integración de datos, para cargar las dimensiones se nombró de la siguiente manera:

- transf_dim_<nombre>

Ejemplo: transf_dim_visitante.ktr

De igual manera se implementaron tres trabajos, uno que ejecuta todas las transformaciones que hacen posible la carga de los hechos (trabajo_cargar_hechos), uno para la carga de todas las dimensiones (trabajo_cargar_dimensiones) y el trabajo principal del MD (trabajo_principal_DMSeries_Turismo) el cual ejecuta los dos trabajos explicados anteriormente. En este último, debido a la dependencia que poseen las tablas de hechos con las tablas de dimensiones, se hizo necesario ejecutar primero el trabajo **trabajo_cargar_dimensiones** y posteriormente el trabajo **trabajo_cargar_hechos**. También existe el trabajo llamado Main_job que carga las variables de entorno, las cuales establecen la configuración para la conexión a la BD. A continuación se muestra este último en la figura 15, y en la figura 16 y 17, el trabajo principal del MD y el trabajo que ejecuta todas las transformaciones que cargan los hechos, respectivamente.



Figura 15: Trabajo principal Main_job

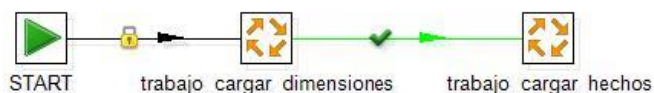


Figura 16: Trabajo principal del MD Series históricas de turismo

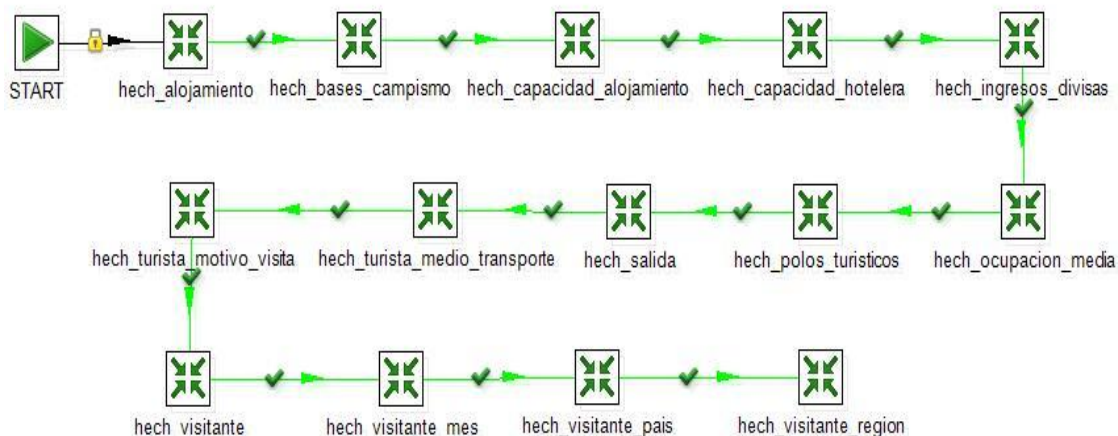


Figura 16: Trabajo “trabajo_cargar_hechos” que carga todos los hechos

Cuando se agregó un nuevo flujo de transformaciones (trabajo), en el proceso de integración de datos, se nombró de la siguiente manera:

- trabajo_<nombre>

Ejemplo: trabajo_cargar_hechos.kjb

3.3.4 Carga de datos

El proceso de carga fue el paso final del desarrollo del almacén, donde los datos que fueron limpiados y transformados en los procesos anteriores, se cargaron hacia las tablas de hechos y dimensiones correspondientes que conforman el modelo dimensional del MD Series históricas de turismo, posibilitando que puedan ser consultados por los usuarios finales.

3.3.5 Gestión del cambio en las dimensiones

Las dimensiones lentamente cambiantes o SCD (Slowly Changing Dimensions) son dimensiones en las cuales sus datos tienden a modificarse a través del tiempo, ya sea de forma ocasional o constante, o que implique a un solo registro o la tabla completa. Cuando ocurren estos cambios, se puede optar por seguir alguna de estas dos opciones:

- Registrar el historial de cambios.
- Reemplazar los valores que sean necesarios.

Inicialmente Ralph Kimball planteó tres estrategias a seguir cuando se tratan las SCD: tipo 1, tipo 2 y tipo 3; pero a través de los años la comunidad de personas que se encargan de modelar BD profundizaron en las definiciones iniciales e incluyeron varios tipos de SCD más como son el tipo 0, tipo 4 y tipo 6. A continuación se detallará cada tipo de estrategia SCD (22):

SCD Tipo 0: este es un enfoque pasivo, es decir no se hace nada al respecto. Los valores permanecen como estaba la dimensión cuando los registros fueron creados.

SCD Tipo 1: este tipo es el más básico y sencillo de implementar. En este caso cuando un registro presente un cambio en alguno de los valores de sus campos, se procede simplemente a actualizar el dato en cuestión, sobrescribiendo el antiguo. Usualmente este tipo es utilizado en casos en donde la información histórica no sea importante de mantener, tal como sucede, por ejemplo, cuando se debe modificar el valor de un registro porque tiene errores ortográficos.

SCD Tipo 2: en este enfoque se inserta un nuevo registro cada vez que existe un cambio en la dimensión. Se agrega un campo de versión u opcionalmente se agregan dos columnas para capturar la fecha de inicio y final de ese valor. Con este método se puede relacionar fácilmente el período de tiempo para el cual es válido cierto dato en la dimensión. Esta técnica permite guardar ilimitada información de cambios.

SCD Tipo 3: este método da seguimiento al cambio agregando nuevas columnas. Una columna mantendría el dato actual, y otra el dato nuevo por el que se quiere cambiar el actual, así como una columna de fecha efectiva del cambio. Este enfoque sólo puede mantener un cambio histórico, a diferencia del Tipo 2 que puede mantener cambios ilimitados en la historia.

SCD Tipo 4: este método mantiene una tabla histórica para todos los cambios y una tabla con el valor actual de la dimensión. Esta tabla histórica indicará, por ejemplo, que tipo de operación se ha realizado (Insertar, Modificar, Eliminar), sobre qué campo y en qué fecha. El objetivo de mantener esta tabla es el de contar con un detalle de todos los cambios, para luego analizarlos y poder tomar decisiones acerca de cuál técnica SCD podría aplicarse mejor.

SCD Tipo 6: este método, conocido también como Híbrido, es una combinación de los Tipos 1, 2 y 3 ($1 + 2 + 3 = 6$). El enfoque es usar una Dimensión Tipo 1 (escribiendo el dato actual), pero agregar un par adicional de columnas con las fechas de validez (Tipo 2).

Debido a que los indicadores de las Series históricas de turismo han sido recogidos y publicados en internet durante años de la misma forma, es decir, que no han sufrido cambio alguno, se decidió con previo consentimiento del cliente, que para la gestión del cambio en las dimensiones, en el MD Series históricas de turismo, la estrategia de SCD a utilizar fuese de Tipo 0, donde los valores van a permanecer como estaban en la dimensión cuando fueron cargados. Si existiera en el futuro la necesidad de cambiar un valor en alguna de las dimensiones o incluir otros nuevos, el MD brinda esta posibilidad, pues en la implementación de las transformaciones que cargan las dimensiones se utilizó el componente Insertar/Actualizar el cual se configuró para permitir estos cambios.

3.3.6 Gestión de los metadatos del proceso de integración

Los metadatos son datos sobre los datos, o información descriptiva sobre los datos y otras estructuras, como objetos, reglas de negocio y procesos que manipulan los datos. Esta información puede ser sobre cuándo se creó un archivo, quién lo creó, cuándo fue actualizado la última vez, su tamaño y su extensión, entre otros. Dado que los metadatos han sido utilizados en varios campos, existen modelos especializados y aceptados en su agrupación para especificar los tipos de metadatos. A continuación se explica brevemente cada uno de ellos:

Bretheron y Singley distinguen dos clases de metadatos distintos: metadatos estructurales/control y metadatos de guía. Los metadatos estructurales/control se utilizan para describir la estructura de sistemas de computación tales como tablas, columnas e índices. Los metadatos de guía se utilizan para ayudar a los seres humanos encontrar a objetos específicos, y normalmente se expresa con un conjunto de palabras claves en lenguaje natural.

Por otro lado, se encuentran otros tipos como son los metadatos descriptivos, estructurales y administrativos. Los metadatos descriptivos incluyen la información utilizada para buscar y ubicar un

objeto tal como el título, el autor, los temas, las palabras claves, la casa editorial. Los estructurales dan la descripción de cómo los componentes del objeto están organizados, y los administrativos, hacen referencia a la información técnica incluyendo el tipo de archivo.

Según Ralph Kimball, los metadatos se pueden dividir en tres categorías (23):

- Metadatos técnicos: se usan a menudo por un personal más técnico, tal como los desarrolladores. Incluye temas como las definiciones de tablas y tipos de datos. Estos objetos son utilizados frecuentemente durante el diseño de la aplicación y el proceso de desarrollo.

Ejemplos: la definición de la fuente y el destino, sus estructuras de tabla, campos y atributos, la documentación para las derivaciones de auditoría y dependencias.

- Metadatos del negocio: ayudan a definir los términos en el lenguaje cotidiano, sin reparos a la implementación técnica. Por ejemplo, el lenguaje utilizado para describir un cliente y la forma en que se categoriza, a menudo es específico de negocio, y podría diferir entre las divisiones de la compañía.

Ejemplos: las reglas comerciales, gestión, definiciones comerciales, la terminología de auditoría, glosarios, algoritmos y linaje que utilizan el lenguaje comercial.

- Metadatos de proceso: se refieren a los metadatos generados y capturados cuando se ejecuta un proceso. Permite que los administradores gestionen su sistema y aseguran que los procesos funcionen sin problemas. Si hay un problema con alguno de ellos, los metadatos operacionales también ayudan a los administradores a identificar y localizar los problemas.

Ejemplos: información acerca de la ejecución de las aplicaciones, incluyendo la frecuencia, conteos de registro, un análisis de componente por componente y otras estadísticas con fines de auditoría.

En la presente investigación se utilizaron los metadatos de procesos. A través de la herramienta Pentaho Data Integration 4.2.1, que permite mantener la gestión de ellos, se almacenó en la BD todas las informaciones referentes a las transformaciones y los trabajos, como por ejemplo, el nombre, la fecha de inicio y fin de la ejecución, los errores que se produjeron, y otras series de características. También se almacenó información para mantener y gestionar el histórico de las fuentes de datos, como el nombre de la transformación, el nombre de la fuente de datos y la fecha exacta de la ejecución. Todos estos datos guiaron los procesos de ETL, permitiendo describir el sistema para dar una visión del funcionamiento del mismo.

3.4 Implementación del subsistema de visualización de datos

3.4.1 Implementación de la capa de visualización

Mediante el mapa de navegación se puede tener una mejor visualización de cómo se muestra la información y de la forma en que se encuentra organizada. El MD Series históricas de turismo está conformado por un Área de Análisis General (A.A.G), un A.A y diez L.T dentro de los cuales se encuentran las 16 tablas de salidas (TS) o reportes que fueron implementados. A continuación se muestra la estructura que posee la capa de visualización:

Descripción del Área de Análisis General (A.A.G)

A.A.G SIGOB: agrupa toda la información referente a los diferentes MD realizados para las distintas áreas de la ONEI que forman el AD SIGOB.

Descripción del Área de Análisis (A.A)

A.A Series de turismo: agrupa toda la información referente a las series históricas de turismo.

Descripción de los Libros de Trabajo (L.T)

L.T 01-Visitante: contiene cuatro reportes relacionados con los visitantes.

L.T 02-Turistas: contiene dos reportes relacionados con los turistas.

L.T 03-Salidas: contiene el reporte relacionado con las salidas al extranjero.

L.T 04-Alojamiento: contiene tres reportes relacionados con las llegadas y las pernoctaciones de los turistas a los medios de alojamiento.

L.T 05-Capacidad de alojamiento: contiene el reporte relacionado con la capacidad de alojamiento,

L.T 06-Polos turísticos: contiene el reporte relacionado con la capacidad de alojamiento en los polos turísticos.

L.T 07-Capacidad hotelera: contiene el reporte relacionado con la capacidad de los hoteles y otros establecimientos seleccionados.

L.T 08-Tasa de ocupación: contiene el reporte relacionado con la tasa de ocupación media anual de los establecimientos de alojamiento.

L.T 09-Ingresos en divisas: contiene el reporte relacionado con los ingresos en divisas asociados al turismo.

L.T 10-Bases de campismo: contiene el reporte relacionado con los indicadores seleccionados del campismo.

A continuación se muestra la TS o reporte “Llegadas de visitantes por tipo de visitante”, el cual se encuentra dentro del L.T 01-Visitante:

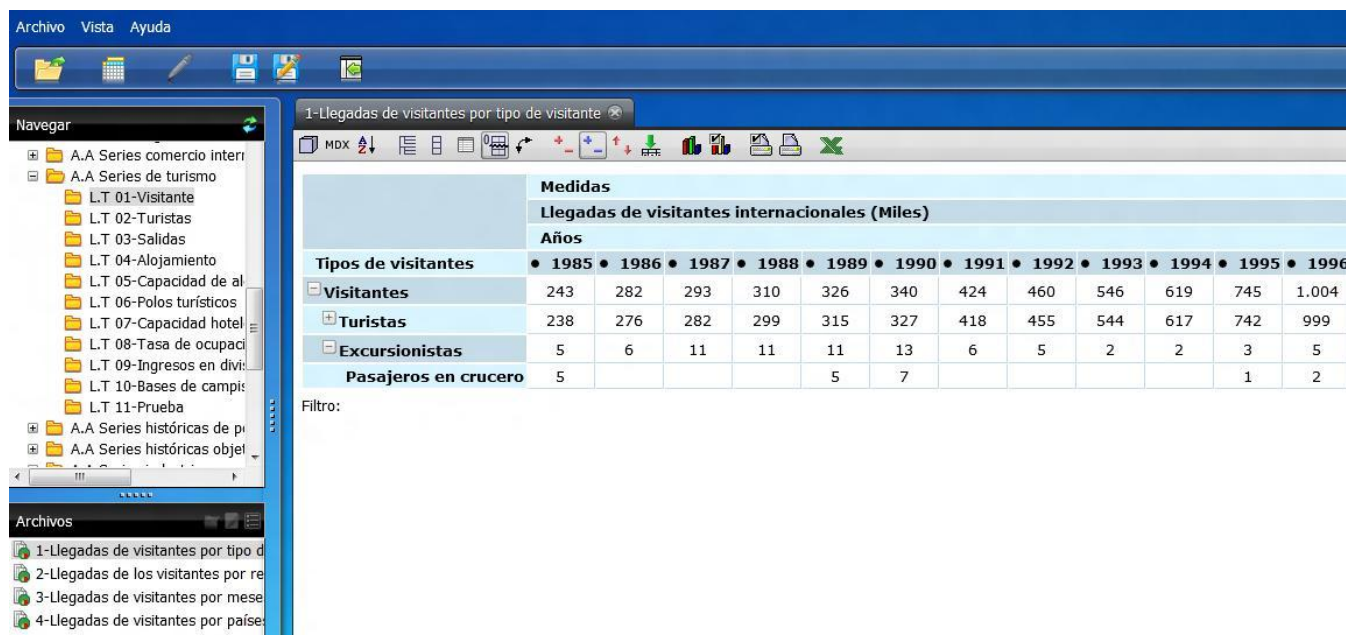


Figura 18: Reporte Llegadas de visitantes por tipo de visitante

3.4.3 Configurar la seguridad de los usuarios y roles

Con el objetivo de proporcionar una mayor seguridad al sistema, durante la implementación del subsistema de visualización del MD Series históricas de turismo se crearon dos roles y dos usuarios que poseen diferentes permisos para el acceso a la información. A continuación se describen cada uno de ellos:

- Rol de administrador: tiene asignados todos los permisos sobre la aplicación y posee el usuario administrador del sistema.
- Rol de analista: tiene el permiso de sólo lectura y posee el usuario analista del sistema.



Figura 19: Creación de los roles Analista y Administrador

3.5 Pruebas

3.5.1 Pruebas de software

La elaboración de un software es un proceso complejo y voluminoso que está propenso a errores. Debido a esto es necesario que su desarrollo esté acompañado de una actividad que permita identificar posibles fallos en la implementación, calidad o usabilidad del mismo. Dentro de las diferentes fases del ciclo de vida del software se integran las pruebas de software, mediante las cuales se podrá verificar y revelar la calidad de un producto de software y asegurar que este cumpla con los requisitos del cliente. A continuación se muestran algunas de las pruebas que pueden ser utilizadas para la validación de un software (24):

Pruebas unitarias: esta prueba centra el proceso de verificación en la menor unidad del diseño del software: el componente de software o módulo.

Pruebas de integración: consiste en construir el sistema a partir de los distintos componentes y probarlo con todos integrados. Estas pruebas deben realizarse progresivamente.

Pruebas de regresión: consiste en volver a ejecutar un subconjunto de pruebas que han sido llevadas a cabo anteriormente, para asegurarse que los cambios que se hayan realizado no introduzcan un comportamiento no deseado o errores adicionales.

Pruebas del sistema: se refiere al comportamiento del sistema integrado. Estas se aplican generalmente para probar los requerimientos de la solución.

Pruebas de aceptación: se realizan para probar que el sistema cumpla con los requerimientos y expectativas del cliente. Estas se pueden distinguir entre dos pruebas:

- Pruebas alfa: las realiza el usuario en presencia del personal de desarrollo del proyecto haciendo uso de una máquina preparada para tal fin.
- Pruebas beta: las realiza el usuario después de que el equipo de desarrollo les entregue una versión casi definitiva del producto.

En la presente investigación se decidió utilizar el Modelo V, propuesto por el Centro Nacional de Calidad de Software (CALISOFT) y empleado por el centro DATEC, para garantizar la calidad del producto final. A continuación se muestra una representación gráfica del ciclo de vida del software propuesto en el Modelo V, donde a la izquierda se encuentran las diferentes etapas de desarrollo y a la derecha las pruebas correspondientes a cada una de ellas. (Ver figura 20).

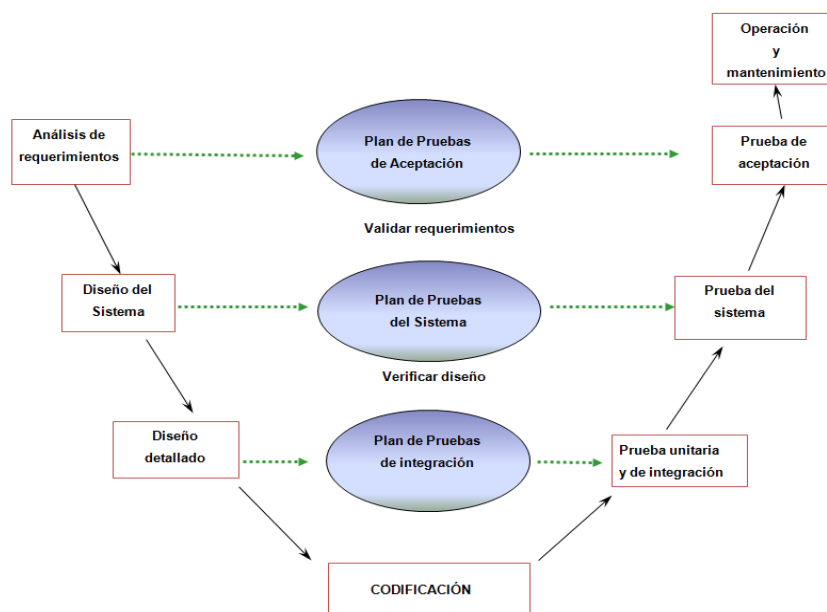


Figura 20: Modelo V

3.5.2 Diseño de los Casos de Prueba

Los Casos de Prueba son un conjunto de condiciones o variables bajo las cuales el analista podrá determinar si el requisito de una aplicación es parcial o completamente satisfactorio. Con el propósito de verificar los requerimientos de información definidos durante la etapa de análisis, y para validar el MD Series históricas de turismo, se diseñaron diez Casos de Pruebas basados en los diez CU informativos y dos Casos de Pruebas basados en reglas de transformación, los cuales están recogidos en el Expediente de Proyecto. En la siguiente figura se muestra el caso de prueba correspondiente al CU Presentar información relacionada con la capacidad de alojamiento.

Escenario	Descripción	Variable de entrada	Variabes de salida	Respuesta del sistema	Flujo central
EC 1.1 L.T 05- Capacidad de alojamiento_Establecimientos, habitaciones y plazas de servicios de alojamiento	Muestra la cantidad de establecimientos, habitaciones y plazas camas de servicios de alojamiento por años y servicios de alojamiento	V Años Servicios de alojamiento	V Total de establecimientos (U) Total de habitaciones (U) Total de plazas camas (U)	Se muestra la información correspondiente al escenario.	1. En la parte superior izquierda se selecciona el Área de Análisis General A.A.G SIGOB. 2. Se selecciona el Área de Análisis A.A Series de turismo. 3. Se selecciona el Libro de Trabajo L.T 11-Prueba. 4. En la parte inferior izquierda, se selecciona el reporte que corresponde al escenario especificado. 5. Se visualiza el reporte en el área de trabajo.

Figura 21: Caso de prueba basado en el CU Presentar información relacionada con la capacidad de alojamiento

3.5.3 Listas de chequeo

En todas partes del mundo se ha hecho necesario crear diferentes mecanismos con el fin de asegurar la calidad y obtener productos satisfactorios. En la UCI se ha establecido por CALISOFT, estándares para definir un modelo propio de aseguramiento de la calidad de software. Uno de los mecanismos más utilizados para la evaluación y control de los productos que se desarrollan en la universidad son las listas de chequeo. Estas no son más que un conjunto o listado de preguntas sobre un aspecto determinado, en forma de cuestionario, que sirven para verificar el grado de cumplimiento de determinadas reglas. Cada una de las preguntas tiene asociada diversos aspectos los cuales posibilitan determinar el grado de cumplimiento y disponibilidad del indicador evaluado.

Para elaborar la lista de chequeo, se tuvieron en cuenta los elementos de evaluación que no deben faltar una vez que se realice el proceso de ETL del MD Series históricas de turismo, permitiendo obtener los puntos satisfactorios o insatisfactorios que posee dicho proceso. Esta lista de chequeo contiene diversos indicadores a evaluar, los cuales se encuentran distribuidos en tres secciones fundamentales:

- ✓ **Estructura del documento:** abarca todos los aspectos definidos por el expediente de proyecto o el formato establecido por el proyecto.
- ✓ **Indicadores definidos:** contiene los indicadores a evaluar de los artefactos generados en los procesos de ETL.
- ✓ **Semántica del documento:** contempla todos los indicadores a evaluar respecto a la ortografía, redacción y otros aspectos de forma y estilo.

Elementos que forman parte de la estructura de la lista de chequeo:

- ✓ **Peso:** define si el indicador a evaluar es crítico o no.
- ✓ **Indicadores a evaluar:** son los indicadores a evaluar en las secciones **Estructura del documento, Semántica del documento e Indicadores definidos**.
- ✓ **Evaluación (Eval):** es la forma de evaluar el indicador en cuestión. El mismo se evalúa de 1 en caso de que exista alguna dificultad sobre el indicador y 0 en caso de que el indicador revisado no presente problemas.
- ✓ **N.P. (No Procede):** se usa para especificar que el indicador no es necesario evaluarlo en ese caso.
- ✓ **Cantidad de elementos afectados:** especifica la cantidad de errores encontrados sobre el mismo indicador.
- ✓ **Comentario:** especifica los señalamientos o sugerencias que quiera incluir la persona que aplica la lista de chequeo. Pueden o no existir señalamientos o sugerencias.

La tabla que se muestra a continuación refleja cómo quedó elaborada la lista de chequeo para evaluar el documento Diccionario de datos. El resto de las listas de chequeo se encuentran en el Expediente de Proyecto:

Estructura del documento					
Peso	Indicadores a evaluar	Eval	(NP)	Cantidad de elementos afectados	Comentarios
crítico	1. ¿El entregable contiene las secciones obligatorias de la plantilla estándar definida para el expediente de proyecto?				
crítico	2. ¿El alcance del proyecto describe correctamente los datos de las dimensiones y hechos del mercado de datos?				
crítico	3. ¿El objetivo expresa correctamente el propósito del documento?				
	4. ¿Se hace un uso adecuado del control del documento?				
	5. ¿En la sección de acrónimos se definen todos los acrónimos utilizados en el documento?				
	6. ¿En el entregable, la definición de las variables se hace correctamente?				
	7. ¿Existe una adecuada correspondencia entre las variables definidas y las descripciones que tienen estas variables?				
	8. ¿En el entregable se crea una hoja por cada variable definida?				
	9. ¿Queda registrado en el entregable todos los posibles valores que van a tener las variables definidas?				

Indicadores definidos en el desarrollo					
Peso	Indicadores a evaluar	Eval	(NP)	Cantidad de elementos afectados	Comentarios
	1. ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?				
Semántica del documento					
Peso	Indicadores a evaluar	Eval	(NP)	Cantidad de elementos afectados	Comentarios
crítico	1. ¿Se han identificado errores ortográficos en el entregable?				
crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?				
	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?				

Tabla 10: Lista de chequeo para el documento Diccionario de datos

3.5.4 Calidad de datos

La calidad de los datos es un aspecto importante a tener en cuenta para el desarrollo del MD, debido a que evita que la información almacenada posea errores. En la presente investigación, después de haber realizado el proceso de integración de datos, se realizó el Perfilado de los datos a los valores que fueron cargados, con el objetivo de verificar que estos tuviesen la calidad requerida. El proceso de perfilado permite obtener, entre otras cosas, estadísticas e información sobre los datos, que posibilitan corregir problemas tales como: valores escritos incorrectamente, duplicados o ausentes. Con la utilización de la herramienta DataCleaner para el perfilado de los datos, se obtuvieron diferentes reportes que evidenciaron un resultado satisfactorio para el MD, mostrando que los datos fueron correctamente cargados hacia la BD. A continuación se muestra, en la figura 22, el perfilado realizado a la tabla de hecho hech_visitantes.

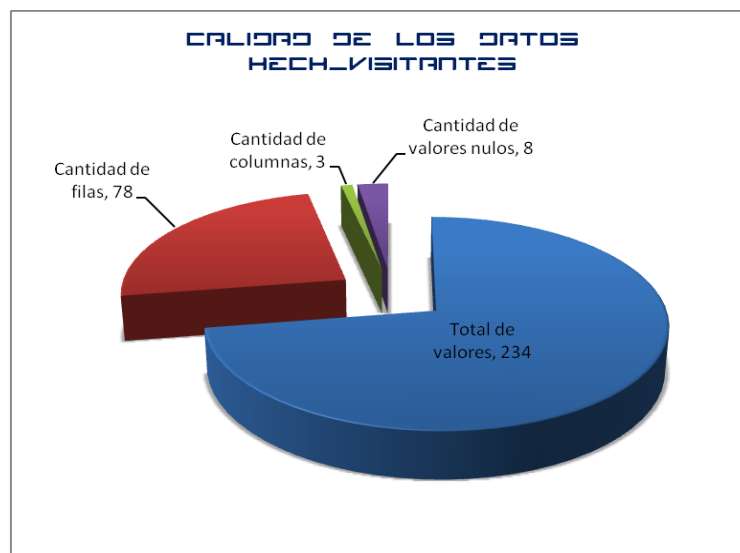


Figura 22: Perfilado realizado a la tabla de hecho hech_visitantes

Además del perfilado, se realizaron al MD pruebas de aceptación con el cliente para comprobar la calidad de los datos, verificando primeramente que se cumplieran los requisitos de información y que los Libros de Trabajo correspondieran al objeto o evento del negocio que es analizado en las series históricas de turismo. Para esto se compararon los 16 reportes con las fuentes de información (12 excel), comprobando que las cifras o datos mostrados sean exactamente los mismos que se recogen en los excel. De esta manera se logró que la información visualizada fuese la correcta, emitiéndose, por parte del cliente, el acta de aceptación del producto.

3.6 Evaluación del resultado de las pruebas

3.6.1 Aplicación de los Casos de Pruebas

Los Casos de Pruebas fueron aplicados con el objetivo de comprobar que el producto cumpla con los requerimientos y necesidades del cliente. Para ello, al MD Series históricas de turismo se le realizó una revisión de calidad interna, hecha en el departamento de Almacenes de Datos del centro DATEC, en la cual fueron detectadas, en los diez Casos de Pruebas basados en CU, cuatro no conformidades. Estas se resolvieron satisfactoriamente. De igual manera fueron aplicados los dos Casos de Pruebas basados en reglas de transformación y no se detectaron no conformidades. Por último, CALISOFT realizó las pruebas finales al MD emitiendo la carta de liberación del producto. Las no conformidades detectadas en estas pruebas quedaron resueltas y el MD quedó liberado en la segunda iteración.

3.6.2 Aplicación de las Listas de Chequeo

Para la evaluación del resultado de las listas de chequeo, se tuvieron en cuenta los siguientes elementos:

Se aborta el proceso de aplicación de la lista en caso de:

- ✓ Existen al menos dos indicadores críticos evaluados de mal.
- ✓ Más del 50 % de los indicadores a evaluar están evaluados de mal.
- ✓ Se mantienen las no conformidades de una revisión a otra.

Se evalúa de regular la calidad del artefacto de ETL revisado, en caso de:

- ✓ Incumple con los indicadores críticos a evaluar de las secciones **Estructura del documento** y **Semántica del documento** que posee la lista de chequeo.
- ✓ Existe al menos un indicador crítico evaluado de mal.
- ✓ Existen al menos cinco indicadores no críticos evaluados de mal de la sección **Indicadores evaluados por la etapa** que posee la lista de chequeo.

El resultado de la evaluación del artefacto es bien si no cumple ninguno de los criterios anteriores.

En el siguiente gráfico de barras aparece el comportamiento de los indicadores evaluados luego de aplicar la lista de chequeo del artefacto Diccionario de Datos. En dicha lista, se identificaron un total de trece indicadores, de ellos cinco críticos, y luego de aplicada no se generó ninguna **no conformidad**.

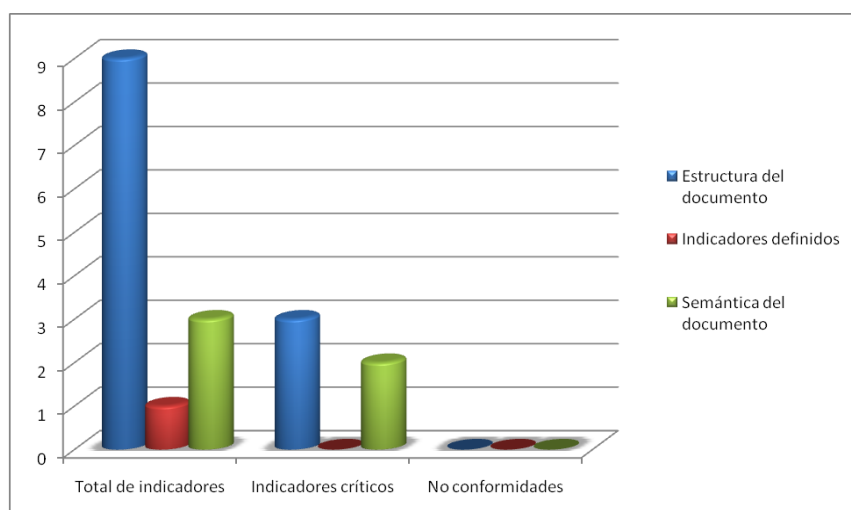


Figura 23: Comportamiento de los indicadores por secciones

Es importante destacar que de esta misma forma se evaluaron el resto de los artefactos correspondientes a los procesos de ETL, y ninguna de las listas de chequeo aplicadas generaron no conformidades.

3.7 Conclusiones del capítulo

Después de haber realizado la implementación y pruebas del Mercado de Datos Series históricas de turismo se arribaron a las siguientes conclusiones:

- Durante la implementación del subsistema de almacenamiento quedaron definidas las estructuras físicas del Mercado de Datos con 31 tablas y tres esquemas.
- En la implementación del subsistema de integración, se ejecutaron un total de 26 transformaciones: 12 para la carga de las dimensiones y 14 para la carga de los hechos, de esta manera se logró poblar el Mercado de Datos satisfactoriamente. Además, se definió que el tipo de SCD a utilizar fuese de Tipo 0, para la gestión del cambio en las dimensiones; y que los metadatos a utilizar fuesen los de procesos, permitiendo que se gestionaran y capturaran los datos cada vez que se ejecute un proceso, y asegurando que los mismos funcionaran sin problemas.
- A través de la implementación del subsistema de visualización quedó definida un Área de Análisis, diez Libros de Trabajos y 16 reportes, los cuales fueron implementados, permitiendo la visualización de la información. Además, se establecieron los roles y permisos de los usuarios que accederán a dicha información, con el objetivo de proporcionar la seguridad al sistema.
- Con la utilización de las listas de chequeo para comprobar que los artefactos del proceso de ETL tuviesen la estructura correcta, y de los casos de pruebas realizados por reglas de transformación y por casos de uso para validar los procesos de ETL y de BI respectivamente, se logró probar el Mercado de Datos, y así determinar que el sistema cumpliera con los requisitos del cliente.

Conclusiones generales

En la Oficina Nacional de Estadísticas e Información, específicamente en el área del turismo, fueron encontradas diversas deficiencias con respecto al manejo de la información, lo cual dificultaba el proceso de toma de decisiones. Para mitigar estas deficiencias, se desarrolló el Mercado de Datos Series históricas de turismo, el cual arrojó resultados satisfactorios y cumple con los objetivos propuestos en la investigación:

1. Se fundamentó la selección de las metodologías, herramientas y tecnologías a utilizar en el desarrollo del Mercado de Datos, quedando, en dicha selección, el Modelo para el Desarrollo de Almacenes de Datos e Inteligencia de Negocio en DATEC como metodología utilizada, y diferentes herramientas que permitieron el correcto desarrollo del Mercado de Datos.
2. Se realizó el análisis y diseño del Mercado de Datos Series históricas de turismo, permitiendo identificar los requerimientos informativos, los funcionales y no funcionales. Además, se diseñaron los tres subsistemas fundamentales (almacenamiento, integración y visualización), arrojando como elementos principales el modelo dimensional, el perfilado de los datos, el diseño general de las transformaciones, la arquitectura de información del sistema y los reportes candidatos, los cuales sirvieron como base para la implementación de la solución.
3. Se implementó y probó el MD Series históricas de turismo, lo que contribuyó, a través de la implementación de los subsistemas, que los datos estuviesen organizados en una estructura física (esquemas y tablas), que fuesen cargados satisfactoriamente hacia dicha estructura y que se visualizara la información para su correcto análisis. Además, se comprobó, utilizando los casos de pruebas y las listas de chequeo, que el sistema cumple con las necesidades del cliente.

Recomendaciones

- Desplegar el Mercado de datos Series históricas de turismo, en la Oficina Nacional de Estadísticas e Información, específicamente en el área de Turismo.
- Realizar capacitaciones en las distintas áreas de la Oficina Nacional de Estadísticas e Información, tomando como referencia la presente investigación, con el objetivo de darle a conocer a los especialistas de dicha entidad, las ventajas que tiene la utilización de los Almacenes de Datos en su labor diaria, en particular el Sistema de Información de Gobierno.
- Realizar pruebas de rendimiento a la Base de Datos correspondiente al Mercado de Datos Series históricas de turismo, teniendo en cuenta que se realizan cargas anuales y el número de tuplas debe aumentar considerablemente.

Referencias bibliográficas

1. **Informática.** [En línea] <http://ifdmmginformatica.blogspot.com/2008/10/definicion-y-origen-del-termino.html>.
2. **Sinnexus.** [En línea] http://www.sinnexus.com/business_intelligence/datawarehouse.aspx.
3. **Definiciones Conceptos Mercados de Datos.**
[En línea] <http://www.mitecnologico.com/Main/DefinicionesConceptosMercadosDatos>.
4. **Sinnexus.** [En línea] http://www.sinnexus.com/business_intelligence/olap_vs_oltp.aspx.
5. **Sinnexus.** [En línea] http://www.sinnexus.com/business_intelligence/olap_avanzado.aspx.
6. **Curto, Josep.** Information Management. *Diseño de un data warehouse: estrella y copo de nieve.* [En línea] 19 de noviembre de 2007.
<http://informationmanagement.wordpress.com/2007/11/19/disenio-de-un-data-warehouse-estrella-y-copo-de-nieve/>
7. **Carrasco, Roberto Clemente Navarrete.** *Business Intelligence: la necesidad actual*
[En línea] <http://www.gestiopolis.com/recursos/documentos/fulldocs/ger/bintna.htm>.
8. **Ortiz, Julio Ernesto.** *Diseño e implementación de un Mercado de datos para la Oficina Nacional de Estadísticas.*
9. **Morales, Jose Salvador Bermudez Rodriguez y Themis Patricia Diaz.** *Diseño de un Datawarehouse para los Ensayos Clínicos que se gestionan en el Centro de Inmunología Molecular.*
10. **Hernández, Yanisbel González.** *PROPUESTA DE METODOLOGIA PARA EL DASARROLLO DE ALMACENES DE DATOS EN DATEC.* Habana : s.n., 2011.
11. **DATEC, Especialistas de.** *Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio en DATEC.* Habana : s.n., 2010.
12. **Leroot Inteligencia de Negocio.** [En línea]
http://www.leroot.com/site/index.php?option=com_content&view=article&id=63&Itemid=69.
13. **Inteligencias de Negocios para empresas.** *Qué es Business Intelligence?.* [En línea]
<http://www.bi-argentina.com.ar/que-es-business-intelligence/>.
14. **Sinnexus.** [En línea] http://www.sinnexus.com/business_intelligence/index.aspx.
15. **Informáticas, Instituto Nacional de Estadísticas.** [En línea] <http://www.onei.gov.ve>
16. **Zapata, Luis Giraldo y Yuliana.** [En línea] Septiembre de 2005.
17. [En línea] <http://www.slideshare.net/vanquishdarkenigma/visual-paradigm-for-uml>.
18. **Glosario.net.** [En línea] <http://cultura.glosario.net/terminos-bibliotecarios/sgbd-12441.html>.
19. **Gravitar Información sin límites.** *El proyecto Pentaho BI.* [En línea]
<http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/>.

20. **Mondrian Documentation.** [En línea] <http://mondrian.pentaho.com/documentation/workbench.php>.
21. **Pentaho Open Sources Business Intelligence.** [En línea] <http://www.pentaho.com>.
22. **Diaz, Josep Curto;** *Introducción al business intelligence (2012)*.
23. **Kimball, Ralph.** *The Data Warehouse ETL Toolkit.* Indianapolis: Wiley Publishing, Inc, 2004. ISBN: 0-764-57923-1.
24. **Sommerville, Ian.** *Ingeniería del Software. s.l: Prentice Hall, 2005.* ISBN: 8478290745.

Bibliografía

1. **Carrasco, Roberto Clemente Navarrete.** *Business Intelligence: la necesidad actual* [En línea] <http://www.gestiopolis.com/recursos/documentos/fulldocs/ger/bintna.htm>.
2. **DATEC, Especialistas de.** *Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocio en DATEC.* Habana : s.n., 2010.
3. **Definiciones Conceptos Mercados de Datos.** [En línea] <http://www.mitecnologico.com/Main/DefinicionesConceptosMercadosDatos>.
4. **Diaz, Josep Curto;** *Introducción al business intelligence* (2012). Barcelona: UOC 2010.
5. **El Salvador.** *¿Porqué quebró Kmart?* [En línea] <http://www.elsalvador.com/noticias/2002/4/25/negocios/negoc2.html>.
6. **Glosario.net.** [En línea] <http://cultura.glosario.net/terminos-bibliotecarios/sgbd-12441.html>.
7. **Gravitar Información sin límites.** *El proyecto Pentaho BI.* [En línea] <http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/>.
8. **Informáticas, Instituto Nacional de Estadísticas.** [En línea] <http://www.ine.gov.ve>
9. **Inteligencias de Negocios para empresas.** *Qué es Business Intelligence?.* [En línea] <http://www.bi-argentina.com.ar/que-es-business-intelligence/>.
10. **Kimball, Ralph.** *The Data Warehouse ETL Toolkit.* Indianapolis: Wiley Publishing, Inc, 2004. ISBN: 0-764-57923-1.
11. **Leroot Inteligencia de Negocio.** [En línea] http://www.leroot.com/site/index.php?option=com_content&view=article&id=63&Itemid=69.
12. **Milanés, Yelena Islen San Juan y Judith Recio.** *Mercado de datos Agricultura, silvicultura y pesca para el Sistema de información de Gobierno.*
13. **Mondrian Documentation.**[En línea] <http://mondrian.penthao.com/documentation/workbench.php>
14. **Morales, Jose Salvador Bermudez Rodriguez y Themis Patricia Diaz.** *Diseño de un Datawarehouse para los Ensayos Clínicos que se gestionan en el Centro de Inmunología Molecular.*
15. **Ortiz, Julio Ernesto.** *Diseño e implementación de un Mercado de datos para la Oficina Nacional de Estadísticas.*
16. **Pentaho Open Sources Business Intelligence.** [En línea] <http://www.pentaho.com>
17. **Sinnexus.** [En línea] http://www.sinnexus.com/business_intelligence/datawarehouse.aspx.
18. **Sinnexus.** [En línea] http://www.sinnexus.com/business_intelligence/index.aspx.
19. **Sinnexus.** [En línea] http://www.sinnexus.com/business_intelligence/olap_avanzado.aspx.

20. **Sinnexus**. [En línea] http://www.sinnexus.com/business_intelligence/olap_vs_oltp.aspx.
21. **Sommerville, Ian**. *Ingeniería del Software*. s.l: Prentice Hall, 2005. ISBN: 8478290745.
22. **Zapata, Luis Giraldo y Yuliana**. [En línea] Septiembre de 2005.
23. [En línea] <http://www.slideshare.net/vanquishdarkenigma/visual-paradigm-for-uml>.

Glosario de términos

Almacén de Datos: es una Base de Datos corporativa caracterizada por integrar y depurar información de una o más fuentes distintas, permitiendo su análisis desde diversas perspectivas.

Área de Análisis: agrupación de información según su propósito, aunque el criterio depende de las necesidades de la institución o empresa donde se aplica el sistema. Permite restringir el número de usuarios que acceden a los datos.

Base de datos relacional: es una base de datos que cumple con el modelo relacional, el cual es el modelo más utilizado en la actualidad para implementar bases de datos ya planificadas. Permiten establecer relaciones entre los datos (que están guardados en tablas), y a través de ellas relacionar los datos de ambas tablas, de ahí proviene el nombre de: "Modelo Relacional".

BI: conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la organización) en información estructurada, para su explotación directa o para su análisis y conversión en conocimiento, dando así soporte a la toma de decisiones sobre el negocio.

Cubo: colección de dimensiones y medidas en un área temática particular.

Casos de Uso del sistema (CUS): proceso dentro del negocio que se estudia, por lo que se corresponde con una secuencia de acciones con un orden lógico, y que producen un resultado observable para ciertos actores del negocio.

Dimensión: característica de un hecho que permite su análisis posterior en el proceso de toma de decisiones y brinda una perspectiva adicional a un hecho dado.

ELT: proceso a través del cual se gestionan datos obtenidos de múltiples fuentes, con el fin de extraerlos, transformarlos y cargarlos en bases de datos especializadas, denominadas Mercado de Datos, para analizar y apoyar una determinada línea de producto o unidad de negocios.

Granularidad: nivel de detalle al que se desea almacenar la información y se define en dependencia del negocio que se esté analizando.

Hecho: operación que se realiza en el negocio la cual está estrechamente relacionada con el tiempo y es objeto de análisis para la toma de decisiones.

Herramientas CASE: conjunto de aplicaciones informáticas orientadas al incremento de la productividad en el desarrollo de software; las siglas CASE vienen dadas por su nombre en inglés Computer Aided Software Engineering que se conoce como Ingeniería de Software Asistida por Computadoras.

Hipermercados: establecimiento de venta al por menor, que tiene una superficie de venta de más de 2 500 m², realiza sus operaciones comerciales en régimen de autoservicio y pago de un solo acto en las cajas de salida y que dispone de un gran espacio de aparcamiento.

Jerarquía: implica una organización de niveles dentro de una dimensión, donde cada nivel representa el total agregado de los datos del nivel inferior.

Libro de Trabajo: estructura organizativa que agrupa los reportes generados dentro de las Áreas de Análisis. Puede ser creado teniendo en cuenta criterios que permitan organizar la información: emisor de los reportes, receptor del reporte, contenido, entre otros.

Mercado de Datos: es una base de datos departamental, que se especializa en el almacenamiento de los datos de un área específica, brindando una estructura óptima para analizar los procesos que tienen lugar dentro del departamento. Están orientados a temas específicos y contienen datos de solo una línea del negocio.

ONEI: Oficina Nacional de Estadísticas e Información.

RN: regla del negocio.

SIGOB: Sistema de Información de Gobierno.

SGBD (Sistema Gestor de Bases de Datos): conjunto de programas que permiten crear y mantener una base de datos, asegurando su integridad, confidencialidad y seguridad.

TCP/IP: son las siglas de Protocolo de Control de Transmisión/Protocolo de Internet (en inglés Transmission Control Protocol/Internet Protocol), un sistema de protocolos que hacen posibles servicios Telnet, FTP, E-mail, y otros, entre ordenadores que no pertenecen a la misma red. TCP garantiza la entrega de datos y que los paquetes sean entregados en el mismo orden en el cual fueron enviados. IP utiliza direcciones que son series de cuatro números octetos (byte) con un formato de punto decimal.