Universidad de las Ciencias Informáticas

Facultad 8

Título: Mercado de Datos Estadístico de Inmigración y Extranjería para el Departamento de Turismo y Comercio de la Oficina Nacional de Estadísticas.

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

Autores:

Yolanda Falcón Rodríguez

Reinaldo Leyva Osorio

Tutores:

Ing. Madelys Cuesta Villa

Ing. Daulemys Rigó Portillo

Ciudad Habana, junio del 2010

"Si el presente es de lucha, el futuro es nuestro."

"Tenemos que ir sobre nuestros errores, machacar sobre ellos, analizarlos y que no se repitan. "

"Lo fundamental es que seamos capaces de hacer cada día algo que perfeccione lo que hicimos el día anterior. "

"Podrán morir las personas, pero jamás sus ideas."

Ernesto Che Guevara

Declaración de Autoría

Declaramos ser los autores de la presente tesis y otorgamos 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 2010.

Firma del autor	**Firma del autor**
Firma del tutor	**Firma del tutor**

Resumen

La solución del presente trabajo de diploma se enmarca en el tema de los almacenes de datos, los mercados de datos, y su utilización para los cálculos y análisis estadísticos de la información. Abarca una investigación, en la cual se detallan las metodologías, herramientas, tendencias actuales y las mejores prácticas para el desarrollo de este tipo de soluciones. Se efectúa un levantamiento de requisitos al cliente, para recoger las necesidades del usuario y pasar al proceso de definición del diseño de la solución.

Como resultado se obtiene la estructura del modelo dimensional que comprende: las dimensiones, las jerarquías, la tabla de hechos, la vista materializada y las medidas necesarias para proceder con los cálculos y análisis estadísticos. Se precisan las reglas del negocio utilizadas y se detalla el proceso de carga de los datos de la fuente de la Dirección de Inmigración y Extranjería al mercado de datos. De igual manera, la solución incluye las estrategias de seguridad, respaldo y recuperación de los datos, así como el proceso de validación. Como característica especial, incluye la utilización de herramientas libres en el proceso de implementación.

Palabras claves: Almacenes de datos, mercados de datos y Dirección de Inmigración y Extranjería (DIE).

Contenido

Introducción

Tanto es el desarrollo de las tecnologías en el mundo actual, que provoca que la mayoría de las empresas del planeta estén en una competencia constante, por lo que cada una de ellas hace su mayor esfuerzo por prevalecer por encima de la otra, presentando los parámetros y requisitos que se destacan en esta rama de la informática, recalcando, la existencia de un envidiable sistema de gestión.

La posesión de estos sistemas de gestión para el funcionamiento de la propia empresa es esencial, y no basta con registrar los datos, hay que extraer información significativa de los mismos. Con este objetivo están pensadas nuevas soluciones basadas en el ciclo de inteligencia del negocio, que cuenten con múltiples miradas, múltiples enfoques, a un mismo punto. Se utilizan como medio para describir un tipo de procesos orientados a la toma de decisiones más acertadas y estratégicas para el desarrollo de un negocio.

Cuba, país que procura mantener en su sociedad una educación íntegra, se propone continuar avanzando en el desarrollo de nuevas tecnologías. Ejemplo claro de ello, es el desarrollo en que se encuentra inmersa la Oficina Nacional de Estadística (ONE).

La ONE fue creada como resultado de la reorganización de los organismos de la Administración Central del Estado, a partir de la desaparición del Comité Estatal de Estadísticas. Dicha entidad es una unidad presupuestada, creada para proponer, organizar y ejecutar, según corresponda, la aplicación de la política estatal en materia de estadística del país.

Según el Acuerdo N^{o} 3552 del Comité Ejecutivo del Consejo de Ministros del 5 de octubre de 1999, la ONE se encarga de (García., 2010)**:**

- **Organizar y aprobar** la producción de las estadísticas centralizadas y territoriales.
- **Dirigir metodológicamente** la actividad estadística de los órganos, organismos, instituciones y entidades estatales, así como normar, velar y dictaminar acerca del funcionamiento de las estadísticas complementarias.
- **Definir las atribuciones y responsabilidades** de otros productores de estadísticas de interés nacional y su lugar en el sistema estadístico del país.
- **Aprobar las normas metodológicas y de clasificación** que se utilizan en las estadísticas centralizadas, territoriales y complementarias

- **Identificar** las unidades de observación estadística **y captar**, a través de la red territorial, la información correspondiente a las estadísticas centralizadas.
- **Dirigir los procesos y ejecutar**, según corresponda, los censos económicos y de población y encuestas económicas o sociales de carácter nacional. Aprobar la realización de este tipo de investigaciones estadísticas en el país.
- **Llevar los registros** estatales de empresas, unidades presupuestadas y otras entidades.
- **Supervisar** el trabajo estadístico de organismos y entidades, organizar la auditoría y comprobación estadísticas velando por la autenticidad de la información.
- **Centralizar y emitir la estadística oficial del país.**

La ONE cuenta con diversas direcciones y departamentos que rigen los datos por diferentes criterios, dentro de las cuales están (ONE, 2006):

- Centro de Estudio de Población y Desarrollo.
  - ✓ Demografía.
  - ✓ Desarrollo Social.
  - ✓ Encuestas.
  - ✓ Matemática Aplicada.
  - ✓ Sistemas Geográficos.
  - ✓ Centro Documentación.
- Cuadros, Capacitación y Relaciones internacionales.
- Sistemas Estadísticos.
  - ✓ Metodología.
  - ✓ Registros y Clasificadores.
  - ✓ Supervisión y Control.
- Cuentas nacionales.
  - ✓ Sectores Institucionales.
  - ✓ Bienes y Consumo.
- Industria y medioambiente.
  - ✓ Industria.
  - ✓ Energía.
- Agropecuario.

- Comercio, Turismo y Servicios.
  - ✓ Turismo y Comercio.
- Estadísticas Sociales.
  - ✓ Estadísticas Sociales.
  - ✓ IPC.
- Informática.
  - ✓ Diseño de Sistemas.
  - ✓ Proceso.
- Información.
  - ✓ Comunicación.
  - ✓ Organizaciones Internacionales.
  - ✓ Sitio Web ONE.
  - ✓ Sitio Web Gobierno.
- Administración.
- Economía y Recursos Humanos.
- Auditoria.

El objetivo fundamental de este trabajo está directamente relacionado con la Dirección de Comercio, Turismo y Servicios, específicamente con el Departamento de Turismo y Comercio de la ONE, que se encarga de llevar el control de las personas que entran a la isla por cualquier motivo, ya sean: visitantes, turistas o excursionistas.

Uno de los problemas que se evidencia en este departamento, es que la información que se recoge, se recibe de la Dirección de Inmigración y Extranjería (DIE) del Ministerio del Interior (MININT) y se traslada hacia la sede nacional en disquetes 3 1/2 y en CD-ROM. A medida que pasa el tiempo esa información se incrementa debido a la propia gestión estadística y se hace casi imposible manipularla, lo que provoca una gestión compleja de los datos. Además, los resultados obtenidos no siempre se alcanzan en el tiempo y con la calidad requerida, lo que provoca que el Departamento de Turismo y Comercio presente los siguientes problemas:

- Ineficiente manejo de datos de la institución para la obtención de información valiosa.
- Inconsistencia de los datos almacenados en las estructuras propias de la ONE.
- Carencia de reportes flexibles con la información actualizada.

Estos elementos influyen negativamente en el proceso estadístico de la ONE para la ayuda a la toma de decisiones del país.

Debido a lo antes mencionado, se identifica como **problema científico**: ¿Cómo estructurar los datos de Inmigración y Extranjería, en un repositorio central, en el Departamento de Turismo y Comercio de la ONE? Definiendo como **objeto de estudio**: los Almacenes de Datos Estadísticos. Por tanto, se ha propuesto como **objetivo general**: desarrollar el análisis y el diseño del Mercado de Datos de Inmigración y Extranjería, para el Departamento de Turismo y Comercio de la ONE. A partir del análisis realizado, se desglosan como **objetivos específicos**:

- Elaborar el marco teórico de la investigación acerca de las principales tendencias de implementación de los almacenes de datos y los mercados de datos.
- Realizar el diseño del Mercado de Datos de Inmigración y extranjería para el Departamento de Turismo y Comercio de la ONE.
- Validar la solución desarrollada mediante la realización de pruebas.

Mediante el análisis del objeto de estudio se puede determinar como **campo de acción**: Mercado de Datos de Inmigración y Extranjería para el Departamento de Turismo y Comercio de la ONE. Teniendo en cuenta que para lograr todos los objetivos se plantean las siguientes **tareas de Investigación**:

- Seleccionar de las metodologías existentes para el desarrollo de mercados de datos, la más apropiada según los criterios de éxito.
- Caracterizar los temas relacionados a las mejores prácticas en el desarrollo de mercados de datos de las herramientas de código abierto, para el diseño de la solución.
- Efectuar las entrevistas al Departamento de Turismo y Comercio de la ONE para conformar el levantamiento de requisitos del negocio.
- Establecer las posibles fuentes de datos de la ONE para realizar el proceso de análisis.
- Elegir la granularidad del proceso del negocio para definir el nivel de análisis de la información.
- Definir la arquitectura del sistema para establecer las bases del desarrollo del mercado de datos.
- Definir las dimensiones y los hechos del mercado de datos para realizar el diseño de la solución.
- Estructurar el modelo dimensional.
- Implementar el proceso de carga de los nomencladores al Mercado de Datos de Inmigración y Extranjería del Departamento de Turismo y Comercio de la ONE.
- Evaluar la eficiencia del proceso a través de las listas de chequeo y la carta de aceptación del cliente, para garantizar la calidad del desarrollo de la solución.
- Documentar el proceso de desarrollo del sistema que conformará el expediente de proyecto.

De tal manera el **posible resultado** que se obtendrá será: Mercado de Datos de Inmigración y Extranjería para el Departamento de Turismo y Comercio de la Oficina Nacional de Estadísticas.

Para lograr complementar este resultado se definen los **métodos científicos**, que se dividen en dos categorías los métodos teóricos y los métodos empíricos. Los **métodos teóricos** usados fueron: el **analítico – sintético**, que centrándose en el análisis de los documentos y materiales, permite seleccionar, de las metodologías existentes para el desarrollo de mercados de datos, la más apropiada según los criterios de éxito; elegir la granularidad del proceso del negocio para definir el nivel de análisis de la información, y evaluar la eficiencia del proceso a través de las listas de chequeo y la carta de aceptación del cliente, para garantizar la calidad del desarrollo de la solución. Estos elementos permiten la extracción de los aspectos más importantes para la modelación del diseño de un sistema que facilite el almacenamiento de todos los datos estadísticos sobre la llegada de visitantes a Cuba.

Se utiliza además el método **histórico – lógico**, el cual define cómo ha evolucionado y se ha desarrollado el diseño del mercado de datos. Para ello, se caracterizan los temas relacionados con las mejores prácticas en el desarrollo de mercados de datos de las herramientas de código abierto, se establecen las posibles fuentes de datos de la ONE para realizar el proceso de análisis, y se documentará el proceso de desarrollo del sistema que conformará el expediente de proyecto.

Como otro de los métodos utilizados en esta categoría, se encuentra el **modelado**, a través del cual se establece la necesidad práctica para la cual se usa la modelación y la posible solución del problema de investigación, y que da la medida en que se logra la relación entre el modelo y el objeto. Para conformarlo, se define la arquitectura del sistema para establecer las bases del desarrollo del mercado de datos, se definen las dimensiones y los hechos del mismo, para realizar el diseño de la solución.

Así mismo, se utiliza como **método empírico** la entrevista, de la cual se obtiene información sobre los requerimientos que plantea la ONE sobre el sistema.

Garantizando cumplir con todos los elementos anteriormente planteados, el presente trabajo estará compuesto por **3 capítulos**, de los cuales el primero abordará los temas relacionados con la fundamentación teórica; necesarios para el desarrollo del objetivo, el segundo describirá el proceso de desarrollo de la solución implementada y el último estará orientado al análisis y la valoración de los resultados.

El **capítulo 1**, estará referido al estudio sobre los sistemas de almacenes de datos y los mercados de datos, sus principales metas, características, y los elementos fundamentales que los componen. Se

realizará también, una investigación acerca de las metodologías existentes y las principales herramientas para el desarrollo de los almacenes de datos, tanto en su desarrollo a nivel mundial como nacional.

El **capítulo 2**, abordará aspectos concernientes a la descripción e implementación de la solución, específicamente, a las características de las áreas de análisis, el ciclo completo del proceso del negocio, el análisis de los datos, el modelo dimensional propuesto, tablas de hechos, y las políticas de respaldo a utilizar en el sistema.

Por último, en el **capítulo 3**, se describirá el proceso de carga de los datos a las dimensiones de la solución. Se detallarán las pruebas de implantación, se validará la solución a través del empleo de las listas de chequeo y la carta de aceptación del cliente.

Capítulo 1: Fundamentación Teórica

Introducción

Este capítulo aborda el estudio sobre los sistemas de almacenes de datos y los mercados de datos así como sus características, metas y los componentes que los integran. Se hace un estudio de las metodologías existentes y las principales herramientas para el desarrollo de los almacenes de datos, tanto en su desarrollo a nivel mundial como nacional.

1.1. Almacenes de datos. Características

Existen diversas tendencias sobre los temas de almacenes de datos y formas de conceptualizar, y aunque se diferencian en algunos aspectos, todas giran sobre el mismo eje central. A continuación se enuncian dos de ellas:

“Son un conjunto de datos integrados, temáticos, históricos y no volátiles usados en la estrategia de toma de decisiones administrativas” (Bill Inmon, 2003 - Business Intelligence Galicia, 2007).

“…los almacenes de datos son una copia de los datos de la transacción estructurados específicamente para la pregunta y el análisis” (Kimball, 2002).

Un almacén de datos es un repositorio de datos de muy fácil acceso, alimentado de numerosas fuentes, transformadas en grupos de información sobre temas específicos de negocios, para permitir nuevas consultas y análisis. Es una base de datos 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 y con grandes velocidades de respuesta. La creación de un almacén de datos representa en la mayoría de las ocasiones el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de inteligencia de negocios (Business Intelligence Galicia, 2007).

El almacenamiento de los datos no debe efectuarse con datos que estén en uso. Los almacenes de datos contienen a menudo grandes cantidades de información que se subdividen a veces en unidades lógicas más pequeñas, en dependencia del subsistema de la entidad del que procedan, o para el que sean necesarias. Todos los autores de investigaciones referentes a almacenes de datos coinciden en sus trabajos en los mismos argumentos. Brindan, una percepción cualitativa sobre el tema, desde su punto de vista aunque se expresan de forma diferente. Queda claro que los almacenes de datos son estructuras que se definen en función de temas específicos, en que la información histórica debe estar

integrada y robusta ante los cambios que puedan afectar a la organización. Su objetivo principal, es apoyar el proceso de toma de decisiones empresariales.

Entre los principales aportes de un almacén de datos (Business Intelligence Galicia, 2007):

- Proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global del negocio.
- Facilita la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén; de modo que se obtiene de dicha información, un valor añadido para el negocio.
- Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios de cualquier esfera.
- Simplifica, dentro de la empresa, la implantación de sistemas de gestión, lo que facilita que el trabajo de la información se efectúe de forma integral.
- Supone una optimización tecnológica y económica, en entornos de centros de información estadística o de generación de informes.

En la actualidad se puede afirmar que los avances alcanzados en el desarrollo de los almacenes de datos, confirman que ya es una tecnología madura, estable y que soluciona las problemáticas presentadas, lo que no impide su constante evolución (Evelia, 2009).

Los almacenes de datos reúnen características especiales, las cuales ayudan a la toma de decisiones en la entidad en la que se utilizan. A continuación se mencionan cuatro de ellas, clasificadas como primarias:

- Orientado al tema: Tiene en cuenta los procesos de negocio de una empresa que se deseen priorizar.
- Integrado: Agrupa todos los sistemas operacionales que se generan en el proceso de negocio en un sistema de información, al proveerlos de formatos y códigos consistentes.
- Variable en el tiempo: Los datos se organizan y almacenan en jerarquías en el tiempo, lo que permite análisis comparativos de estados actuales y de períodos anteriores.
- No volátil: Se usa principalmente para operaciones de carga, recuperación de información y no para actualizaciones.
- Otra característica del almacén de datos es que contiene metadatos, es decir, "información sobre información" o "datos sobre los datos". Los metadatos permiten saber la procedencia de

la información, su periodicidad de actualización, su fiabilidad y forma de cálculo. Son datos altamente estructurados que describen la información, el contenido, la calidad, la condición y otras características de los datos (Business Intelligence Galicia, 2007).

Los metadatos serán los que permiten simplificar y automatizar la obtención de la información desde los sistemas operacionales a los sistemas informacionales.

Las metas definidas por Kimball para los almacenes de datos son adaptables totalmente a los mercados de datos, debido a que según el mismo autor referencia, los mercados de datos son la unidad básica de los almacenes de datos corporativos. Cada mercado de datos que se diseñe debe cumplir los objetivos planteados, pero adaptados específicamente a un proceso del negocio, lo que significa, mirándolo de este punto de vista, que el universo de información útil y necesaria para el soporte a la toma de decisiones empresariales estaría enmarcada en un departamento determinado.

1.1.1. Mercados de datos. Características

Los mercados de datos, son un subconjunto de datos de un almacén de datos donde se almacenan la mayoría de las actividades de análisis que se llevarán a cabo, en el entorno de la inteligencia de negocio (Bill Inmon, 2003 - Business Intelligence Galicia, 2007).

La visión de Inmon se basa en un enfoque descendente, propone construir primero el almacén de datos, y a partir de este los mercados de datos. Plantea la creación de un repositorio de datos corporativo como fuente de información consolidada, persistente, histórica y de calidad. Al ser construidos descendentemente, los mercados de datos se nutren del mercado de datos corporativo, y así se convierten en un complejo empresarial de bases de datos relacionales.

A diferencia de la anterior, la propuesta de Kimball se basa en el hecho de poseer una arquitectura ascendente. Plantea que se debe crear por cada departamento un conjunto de mercados de datos independientes, orientados a los temas que estén relacionados con aquel. Considera adecuado, además, dividir el mundo de la inteligencia de negocio en hechos y dimensiones. Esta idea conduce a una solución completa en un corto período de tiempo. Entre sus características principales está la definición de que: “El almacén de datos es la unión de todos los mercados de datos de una entidad” (Curto, 2008).

Siendo así, se puede decir que un mercado de datos es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por

disponer la estructura óptima de los datos, para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento. Son almacenes de datos orientados a temas específicos o aplicaciones específicas y contienen datos de sólo una línea del negocio. La mayor diferencia entre los almacenes y los mercados de datos, es el ámbito de la información que contienen, debido a que en los mercados de datos es más pequeño y los datos se obtienen de un menor número de fuentes, por tanto, provoca que el tiempo de desarrollo sea menor. Los datos existentes en este contexto pueden ser agrupados, explorados y propagados de múltiples formas, para que diversos grupos de usuarios realicen una explotación de la forma más conveniente, según sus necesidades. Son una alternativa de solución, al igual que los almacenes de datos a los problemas planteados anteriormente, pues el diseño y la construcción son similares, a ello sumado a que poseen una secuencia común.

1.1.2. Ventajas de la utilización de los almacenes de datos

Una de las principales ventajas que trae este tipo de solución, radica en las estructuras en las que se almacena la información. Estos tipos de persistencia de la información son homogéneos y fiables, pues permite su consulta y tratamiento jerarquizado, siempre en un entorno diferente de los sistemas operacionales.

Entre las ventajas que caracterizan los almacenes de datos, se destaca que la organización de la información debe ser (Corporate Executive Board, 2008):

- Accesible: Los contenidos del almacén de datos son entendibles y navegables, y el acceso a ellos se caracteriza por su rápido desempeño. Estos requerimientos no tienen ni fronteras, ni límites fijos.
- Consistente: La información de una parte de la organización puede hacerse coincidir con la información de la otra parte de la organización. Si dos medidas de la organización tienen el mismo nombre, entonces deben significar lo mismo y a la inversa, si dos medidas no tienen el mismo significado, entonces son etiquetados diferentes. Información consistente significa, información de alta calidad, toda ella contabilizada y completada. Todo lo demás es un compromiso y por consiguiente algo que se quiere mejorar.
- Adaptable: El almacén de datos está diseñado para cambios continuos. Cuando se hacen nuevas preguntas al almacén de datos, los datos existentes y las tecnologías no cambian, ni se corrompen cuando se agregan datos nuevos al almacén de datos.

1.2. Componentes de los almacenes de datos

1.2.1. Sistemas de fuentes operacionales

Los sistemas de fuentes operacionales son los que poseen las compañías o empresas para la gestión de sus transacciones diarias. Estas operaciones son almacenadas en los más diversos formatos, desde una base de datos relacional hasta otros tipos de ficheros, en los cuales se puedan hacer cualquier tipo de consultas. Se encuentran localizados fuera del repositorio debido a que se tiene poco o ningún control sobre el volumen y formato de los datos de estas fuentes. Las prioridades principales de este componente son el procesamiento, el rendimiento y la disponibilidad. Generalmente realizan salvas de la información que gestionan y sólo trabajan con los datos generados en un corto período de tiempo para hacer las recuperaciones de forma óptima. Puede existir la posibilidad de que sean fuentes creadas manualmente, debido a que no posean un sistema que las procese. La principal función de los sistemas fuentes es capturar las transacciones del negocio.

Estas fuentes, están agrupadas en cuatro categorías principales (Ponniah, 2001):

- Los datos de producción, son los datos de interés para el almacén de datos: se encuentran archivados en los diferentes sistemas operacionales y son utilizados dentro de la organización en sus funciones diarias.
- Los datos internos, son los que posee cada departamento dentro de la organización, almacenados en archivos o bases de datos internas para auxiliarse en sus actividades. Esta información es generalmente útil para el almacén de datos.
- Los datos archivados, son los provenientes de sistemas operacionales y se almacenan con el objetivo de llevar un control histórico de la información de la organización.
- Los datos externos, son los que provienen de fuentes ajenas a la organización. Generalmente son informaciones compartidas entre competidores o entre proveedores y clientes.

1.2.2. Área de procesamiento

El área de procesamiento es el componente donde se invierte la mayor cantidad de tiempo y esfuerzo durante el desarrollo del almacén de datos. Es donde se realiza el proceso de extracción de los datos de las diversas fuentes operacionales que se deseen integrar, teniendo como principal tarea la de almacenar toda esa información en bases de datos relacionales, generalmente, para realizar el análisis y procesamiento de los datos (Ponniah, 2001).

1.2.3. Área de presentación

El área de presentación es el componente donde los datos se encuentran listos para ser consultados, reportados o analizados por los usuarios finales. En ella se encuentra la información diseñada mediante esquemas dimensionales, definidos por los usuarios como útiles para la toma de decisiones. Generalmente esta área es referenciada como una serie de mercados de datos integrados, donde cada uno representativo de un proceso específico del negocio (Ponniah, 2001).

1.2.4. Herramientas de acceso a datos

Su actividad principal es consultar el área de presentación del almacén de datos. Puede abarcar desde una simple o personalizada herramienta de consulta hasta una compleja y sofisticada aplicación de modelado o de minería de datos (Ponniah, 2001).

1.3. Modelos de almacenamiento de datos

1.3.1. Modelos

En el transcurso del desarrollo de las bases de datos, se han empleado varias formas de modelado para su diseño. Dentro de los modelos más importantes se tienen (Technologies, 2010):

- Modelo relacional: El diagrama o modelo entidad-relación (DER), es un lenguaje para el modelado de datos de sistemas de información. Estos modelos expresan entidades más relevantes para el sistema, sus inter-relaciones y propiedades. Trabajan dividiendo los datos en muchas entidades discretas, cada una de las cuales se convierte en una tabla física en la base de datos operacional. Es el modelo más utilizado para el diseño de bases de datos, pues está formado por un conjunto de conceptos que permiten describir la realidad mediante un grupo de representaciones gráficas y lingüísticas.
- Modelo dimensional: A diferencia de los clásicos sistemas de bases de datos que presentan sus estructuras diseñadas mediante el modelo Entidad-Relación (DER), los almacenes de datos se diseñan mediante un modelo dimensional. Poseen la misma información que el DER, pero la organiza de manera diferente para garantizar la velocidad y eficiencia en su recuperación. Una de sus características principales es que no necesita una predefinición de los reportes, debido a que se diseñan de forma tal que cubran el universo de variantes que los usuarios necesiten para consultar la información almacenada.

En un modelo de datos dimensional los datos se organizan alrededor de los temas de la organización. Este modelo está compuesto por dimensiones y hechos. Las dimensiones no son más que la representación de cada uno de los ejes dimensionales. Suministran el contexto de cómo se obtienen las medidas de los hechos; se puede decir que se utilizan para seleccionar y agrupar por niveles. Por su parte, los hechos no son más que los objetos que se van a analizar y son de tipo cuantitativo; sus datos se obtienen generalmente por la aplicación de una función estadística que resume el conjunto de valores en un único valor.

1.4. Modos de almacenamiento de datos

Hoy en día para lograr el almacenamiento de los datos en las grandes bases de datos y especialmente en los almacenes de datos, se utilizan varios modos de procesamiento analítico en línea (OLAP), dentro de los cuales se destacan (Technologies, 2010):

- ROLAP: Procesamiento Analítico Relacional en Línea.
- MOLAP: Procesamiento Analítico Multidimensional en Línea.
- HOLAP: Procesamiento Analítico Híbrido en Línea.

1.4.1. ROLAP

En el procesamiento analítico relacional en línea, los datos son almacenados en filas y columnas de forma relacional. Este modelo presenta los datos a los usuarios en forma de dimensiones de negocio. La semántica de las etiquetas de los metadatos es creada con el fin de ocultar las estructuras de almacenamiento y presentar los datos dimensionalmente; ellas toleran el mapeo de las dimensiones a las tablas relacionales. Estos metadatos también son almacenados en tablas relacionales. El modelo ROLAP es usado fundamentalmente sobre información que no se consulta frecuentemente, debido a que no es óptimo en este sentido. Por ejemplo: la información histórica de muchos años de antigüedad.

1.4.2. MOLAP

Un sistema de procesamiento analítico multidimensional en línea, utiliza una base de datos multidimensional en la que la información se almacena dimensionalmente. Sus sistemas utilizan una arquitectura de dos niveles: la base de datos de modelos dimensional y el motor analítico. El nivel de presentación se integra con el de aplicación y proporciona una interfaz a través de la cual los usuarios finales visualizan las operaciones OLAP. Otro aspecto a destacar es que MOLAP, a diferencia del ROLAP, almacena los datos dimensionalmente. Aquí las estructuras de los datos están fijas para que

la lógica, al procesar la información, pueda estar basada en métodos bien definidos para establecer las coordenadas del almacenamiento de los datos.

1.4.3. HOLAP

En los sistemas de procesamiento analítico híbrido en línea, se mantienen los registros detallados en la base de datos relacional, mientras que los datos resumidos o agregados se almacenan en una base de datos multidimensional separada. Estos sistemas se conocen como híbridos por mantener características de los modelos anteriores. Este modelo posee dos tipos de particionamiento:

- Particionamiento Vertical: Almacena las agregaciones como un MOLAP para mejorar la velocidad de las consultas, y los datos se detallan en ROLAP para optimizar el tiempo en que se procesa el cubo.
- Particionamiento Horizontal: En HOLAP se almacena una sección de los datos, normalmente, los más recientes en modo MOLAP, para mejorar la velocidad de las consultas y los datos.

1.5. Comparación entre MOLAP y ROLAP

Se han originado debates alrededor de dos tipos de almacenamiento MOLAP y ROLAP. Por lo general, las implementaciones de MOLAP presentan el mejor rendimiento de la tecnología relacional, pero tienen problemas de escalabilidad, por ejemplo en cuanto a la adición de dimensiones a un esquema ya existente. Por otra parte, las implementaciones de ROLAP son más escalables y a menudo condescendientes, debido a que se aprovechan de las inversiones de la tecnología de las bases de datos relacionales. En la siguiente tabla se pueden apreciar algunas diferencias entre los modelos MOLAP y ROLAP:

	MOLAP **Multidimensional**	**ROLAP** **Relacional**
Datos	Detalle y pre calculados	Detalle y agregado
Estructura	Matrices comprimidas	Tablas relacionales
Administración	Especialista en BDMD	Administrador de BD
Acceso	Lenguaje especializado	SQL

En la solución de este sistema, se usará el modelo ROLAP.

1.6. Evolución de los almacenes de datos

1.6.1. Evolución de los almacenes de datos en el mundo

La tecnología en el mundo avanza y cada día las empresas tienen mayor número de aplicaciones automatizadas, almacenan la información diaria en grandes bases de datos, y necesitan conocer al momento, su estado en inventarios, su volumen de ventas del día, tener sus precios actualizados. Con el transcurso del tiempo las empresas han ido almacenando un gran número de información en diferentes fuentes de datos, y sus directivos se dieron cuenta de que estos datos podrían ser de utilidad para un mejor desempeño de las instituciones, pues reflejaba la mayoría de las operaciones diarias del negocio. En 1994 el 90% de las empresas, según la revista Fortune 2000, planeaban implementar un almacén de datos. Ya en 1996 el 90% de las grandes corporaciones consideraba adoptar esta tecnología. A raíz de los acontecimientos anteriores, se desencadenó una revolución en esta esfera, por lo que Hill Hostian, de la empresa Gartner, estimó que para el 2007, el 50% de los proyectos de inteligencia de negocio requerirían de alguien que les proveyera los servicios, para librarse de los obstáculos debidos a falta de personal capacitado y recursos (Durán, 2007).

En los mercados empresariales existe mayor competencia, por lo que las empresas requieren mayor rapidez y eficiencia de sus procesos, así como precisión en la información para tomar decisiones adecuadas. Debido a ello, se pensó que lo ideal sería unificar las diferentes fuentes de información de las cuales se disponía, almacenándolas en un único lugar, de tal forma que sólo se incorporara información relevante. Esta nueva herramienta para el almacenaje de la información tendría como nombre ¨repositorio de datos¨, el cual debería tener una estructura organizada, integrada, lógica y dinámica, y además ser de fácil explotación.

A partir de esta problemática, un gran número de compañías se dieron a la tarea de implementar la tecnología de almacenamiento de datos para convertir sus datos en información útil.

En la actualidad, en América Latina existen empresas como: Telefónica de Argentina, Visa y Arcor, todas ellas de Argentina; en México existen algunas como Wallmart, Procter & Gamble, Whirpol, Tv Azteca, Baxter, GNP, Warner Lambert y Sabre, que también han venido incorporando el uso de los almacenes de datos para la toma de decisiones a nivel gerencial. Se pueden mencionar igualmente a

American Stores (Estados Unidos), Canadian Tyre (Canadá), Owens Corning Glass (Estados Unidos), y Karsten Ping Golf Clubs que han obtenido grandes avances en este sentido. En Europa, pioneras en este campo, existen empresas tales como Carrefour (España), WH Smith Books, (Gran Bretaña), BonPreu (España), Corte Inglés (Francia), Supermercados Casino (Francia). Del mismo modo, grandes transnacionales como, Coca Cola, Walt Disney, Nike, Maybelline, Adidas, 3M, Bosh Siemens, se han incorporado a la utilización de los almacenes de datos para la realización de estudios de mercado y de inteligencia de negocio.

Las esferas bancarias y de seguros, también han dado pasos concretos en este sentido, con entidades como Banco Galicia en Argentina, Banco de México, Banorte y Banamex en México, Banco París en Francia y el European Central Bank perteneciente a la Unión Europea. Todos ellos son ejemplos palpables del avance, y exhiben estudios sobre inflación, población, monetarios y tendencias. El tema estadístico no ha estado ajeno a estas necesidades. En países como México (específicamente en el INEGI (Instituto Nacional de Estadísticas e Informática)), se tiene la información en almacenes de datos, y de esta forma es utilizada para la toma de decisiones a nivel gubernamental. De la misma manera el Instituto Nacional de Estadística de España es otro ejemplo de entidades que actualmente están utilizando la tecnología.

1.6.2. Evolución de los almacenes de datos en Cuba

En Cuba se ha incentivado una cultura tecnológica sobre el uso de los almacenes de datos. El ejemplo más elocuente, es el almacén comercial de la Corporación CIMEX, la cual se dedica fundamentalmente a la exportación e importación de mercancías. Forman parte de ella, un conjunto de empresas que se encuentran enfocadas en diversos negocios, de las que se puede citar la red de Comercio Minorista y la Dirección de Logística, esta última dedicada al comercio mayorista.
En el XIII Concurso Nacional de Computación y en la Feria de Informática del 2002, se presentó un almacén de datos para CUBACEL, desarrollado sobre plataforma Oracle con grandes resultados obtenidos a partir de su implantación. Existen otras entidades como UNIÓN CUPET, COPEXTEL y por supuesto la ONE, que en la actualidad se encuentran en el proceso de diseño e implementación de sus respectivos almacenes.

1.7. Metodologías para el desarrollo

Las soluciones de almacenes de datos e inteligencia de negocio (DW&BI) surgen a principios de la década del 90 del siglo pasado; desde entonces ha venido madurando y alcanzando un lugar

sobresaliente entre los sistemas destinados para el análisis de información histórica y apoyo a la toma de decisiones.

Al mismo tiempo, se han venido desarrollando las metodologías para el perfeccionamiento e implantación de este tipo de soluciones. En este sentido, se destacan un conjunto de metodologías que definen y guían todo el ciclo de vida del proceso. Estas tendencias son conocidas como Metodología de Kimball y Metodología de Inmon, en honor a sus creadores Ralph Kimball y William H. Inmon, personalidades más influyentes en el área de los almacenes de datos. La principal diferencia que existe entre ambas tendencias está basada en la forma de enfrentar el problema.

Basados en estas propuestas, se han desarrollado un conjunto de metodologías que realizan una selección de lo mejor de cada una de las anteriores y definen sus propias características. A continuación se describen brevemente algunas de ellas (Sánchez, 2008):

- Metodología SQLBI: Avalada por Microsoft y orientada totalmente a sus herramientas: Microsoft SQL Server, SQL Server Analysis Services y su oferta más completa en este campo, que es Microsoft Suite for Business Intelligence.
- Metodología Hefesto: En una segunda parte de su publicación, se propone una metodología propia para la construcción de un almacén de datos, que parte de la recolección de requerimientos y necesidades de información del usuario, y concluye en la confección de un esquema lógico y sus respectivos procesos de extracción, transformación y carga de datos. Además, se ejemplifica cada etapa de la metodología a través de su aplicación a una empresa real, que sirve de guía para visualizar los resultados que se esperan de cada paso y para clarificar los conceptos enunciados.
- CRISP-DM (CRoss-Industry Standard Process for Data Mining): Propuesta en 1996 como herramienta industrial y de aplicación neutral. Está descrita partiendo de un modelo de proceso jerárquico, consistente en un conjunto de tareas en cuatro niveles de abstracción (de lo general a lo específico): fase, tarea genérica, tarea especializada e instancia de procesos.
- Metodología para el diseño conceptual de almacenes de datos: Aporta como aspecto novedoso con respecto a las anteriores la incorporación de los casos de uso para guiar el proceso de desarrollo, al mismo tiempo que define una serie de transformaciones para llevar desde un diagrama relacional a uno dimensional y así obtener las estructuras que conformarán el repositorio de datos (DATEC, 2010).

1.7.1 Metodología utilizada en el centro DATEC

La metodología que se propone utilizar en el presente trabajo, es desarrollada por la línea de soluciones de almacenes de datos e inteligencia de negocios, del Centro de Tecnologías y Análisis de Datos (DATEC), de la Universidad de las Ciencias Informáticas (UCI). Dicha metodología se caracteriza por cubrir todas las fases del ciclo de vida por las que pasa la construcción de un almacén de datos, desde el levantamiento de información inicial hasta la implementación de la herramienta de inteligencia de negocio. Es una metodología mixta que reúne elementos de varias metodologías de desarrollo de proyectos de integración de datos.

En la primera fase resalta el levantamiento de información a nivel de negocio para identificar los posibles indicadores y aspectos a medir en los análisis, que luego de algunas transformaciones se convierten en los requerimientos de información de entrada y de salida para la solución de integración. De forma paralela a esta actividad se lleva a cabo un estudio de las fuentes de datos que soportan los datos a cargar. Finalizadas estas dos tareas, se corrobora que la información levantada sobre las necesidades de los clientes esté realmente almacenada en las fuentes correspondientes, para posteriormente, teniendo los requerimientos informativos correctamente definidos, proceder a diseñar la solución de bases de datos. Una vez diseñada la estructura de la base de datos, se realiza la carga de los datos desde las fuentes, y posteriormente se implementan los requerimientos de inteligencia de negocio identificados en el levantamiento de información inicial. Las actividades y artefactos de la solución son realizados por 4 grupos que conforman la línea, especializados en componentes específicos de la solución (Villa, 2009).

Los grupos de desarrollo son 4:

- Grupo de Análisis.
- Grupos de Almacenes.
- Grupo de Integración de Datos.
- Grupo de Inteligencia de Negocio.

Cada uno de los grupos tiene sus objetivos específicos, artefactos y herramientas; asi mismo, las personas que trabajan en ellos, protagonizan roles y responsabilidades diferentes. En el proceso de desarrollo, se tiene en cuenta el flujo de artefactos por fases del proyecto, la distribución de artefactos por fases y etapas de desarrollo del proyecto, y el cronograma genérico gerencial y detallado de los proyectos de soluciones de almacenes de datos e inteligencia de negocio. Vale aclarar, que todos los

pasos y las fases que se aplican en esta metodología marcan pautas importantes en este proceso. Se destacan las siguientes:

- Estudio preliminar.
- Fase requerimientos.
- Fase de Arquitectura y Diseño.
- Fase de Implementación.
- Fase de prueba.
- Fase de despliegue.
- Fase de soporte (Con esta fase se le da cierre al proyecto desarrollado).

La organización para la cual va dirigido el producto es la ONE, órgano rector en materia estadística en Cuba. Ello amerita la utilización de una metodología robusta y madura que garantice el éxito de la integración de la información de que actualmente disponen. De todo el conjunto de metodologías existentes para enfrentar el desarrollo del mercado de datos, se escoge esta metodología por política del proyecto, pues el centro ajustó la mundialmente conocida como "metodología de Kimball"a sus necesidades, para un mejor desarrollo en tiempo real de sus productos.

1.8. Gestores de bases de datos

Sistema Gestor de Bases de Datos (SGBD): Conjunto de programas que permite a los usuarios crear y mantener una base de datos, por lo tanto, el SGBD es un software de propósito general que facilita el proceso de definir, construir y manipular la base de datos para diversas aplicaciones. Pueden ser de propósito general o específico (Castillo, 2008).

Hoy día existen varias herramientas para la creación de bases de datos, algunas libres y otras propietarias. Dentro de estos grupos aparecen:

- **Microsoft SQL Server**: Propiedad de Microsoft, pertenece a la familia de los sistemas de administración de base de datos, opera en una arquitectura cliente / servidor de gran rendimiento. Su desarrollo fue orientado para manejar grandes volúmenes de información, y un elevado número de transacciones. SQL Server es una aplicación completa que realiza toda la gestión relacionada con los datos. El servidor sólo tiene que enviar una cadena de caracteres y esperar a que le devuelvan los datos.

- **PostgreSQL:** PostgreSQL ejecuta procedimientos almacenados en más de una docena de lenguajes de programación, como Java, Perl, Python, Ruby, Tcl, C / C + +, y su propio PL / pgsql, que es similar al de Oracle PL / SQL.

- **Oracle:** Oracle Corporation es la primera compañía mundial proveedora de soluciones de software al mundo de la empresa. Con unos ingresos de 10.900 millones de dólares, la compañía ofrece su base de datos, herramientas y aplicaciones de gestión, junto con los correspondientes servicios de consultoría, formación y soporte, en más de 145 países de todo el mundo (Oracle Corporation, 2005-2010).

1.8.1. PostgreSQL

Fig. 1.2. PostgreSQL.

PostgreSQL: Es un SGBD, que no pertenece a ninguna compañía, sino que es dirigido por una comunidad de desarrolladores, que se hace llamar “PostgreSQL Global Development Group” y organizaciones comerciales que están a cargo de su desarrollo. Tiene más de 15 años de desarrollo activo y se ha ganado la reputación de ser confiable y mantener la integridad de los datos. Existen compañías que aseguran haberlo empleado durante varios años y con altas tasas de actividad, sin experimentado problemas de ningún tipo. El PostgreSQL se ejecuta en la mayoría de los sistemas operativos más utilizados en el mundo, incluido Linux, versiones de UNIX y, por supuesto, Microsoft Windows XP o superior.

Debido a sus características técnicas sobresalientes, el PostgreSQL se ha ganado la admiración y el respeto de sus usuarios, así como el reconocimiento de la industria, ha sido ganador del Linux New Media Award for Best Database System y tres veces ganador del The Linux Journal Editors' Choice Award for best DBMS (Corporate Executive Board, 2008).

Las características que presenta PostgreSQL son:

- La cantidad máxima de bases de datos que permite es ilimitada.
- El tamaño máximo de las tablas es de 32 Tb.
- El tamaño máximo de registro es de 1.6 Tb.
- El máximo de tamaño del campo es de 1 Gb.
- El máximo de registros por tablas es ilimitado.
- El máximo de campos por tabla es 250 a 1600 en dependencia de los tipos de datos usados.
- El máximo de índices por tablas es ilimitado.

El código fuente de PostgreSQL está disponible bajo los más liberales términos de licencia de código abierto: la licencia BSD (Berkeley Software Distribution), por tanto pueden hacerse todas las modificaciones, mejoras o cambios que se estimen convenientes.

1.8.2. PgAdmin

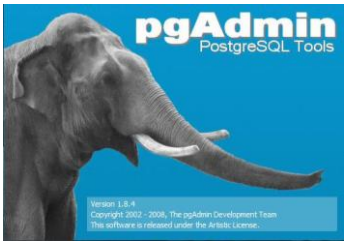

Fig. 1.3. PgAdmin.

PgAdmin III 1.10: Es una herramienta de código abierto para la administración de bases de datos PostgreSQL y derivados, incluye:

- Interfaz administrativa gráfica.
- Herramienta de consulta SQL.
- Editor de código procedural.
- Agente de planificación SQL/shell/batch.
- Administración de Slony-I.

PgAdmin se diseña para responder a las necesidades de la mayoría de los usuarios: desde escribir simples consultas SQL hasta desarrollar bases de datos complejas. La interfaz gráfica soporta todas las características de PostgreSQL y hace simple la administración. Está disponible en más de una docena de lenguajes y para varios sistemas operativos, incluyendo Microsoft Windows, Linux,

FreeBSD, Mac OSX y Solaris. Resiste versiones de servidores 7.3 y superiores. Versiones anteriores a 7.3 deben usar el PgAdmin II (ArPug, 2009).

1.9. Herramientas case de modelado

En la actualidad existen muchas herramientas de modelado que tienen resultados satisfactorios. Entre las más destacadas están:

- **CA Erwin Data Modeler**: Es una herramienta de modelado de datos, líder en este sector, pues ha sido de confianza en el modelado por más de 20 años, con la funcionalidad de clase empresarial a un precio factible. Entre sus principales características se pueden destacar:

  - ✓ Diseño de la capa de arquitectura: Alinea los modelos de datos con los requerimientos del negocio en el nivel lógico con el diseño de bases de datos a nivel físico.
  - ✓ Comparación completa: Automatiza la sincronización bidireccional de los modelos, scripts y bases de datos. En él se comparan los elementos, muestra sus diferencias y permite seleccionar qué diferencias se mueven y en qué dirección.
  - ✓ Normas definición: Apoya la definición y mantenimiento de las normas a través de su diccionario de dominio, nombres, editor de normas y editor de tipos de datos estándares.
  - ✓ Presentación de informes y publicación: Ofertas flexibles, presentación de informes personalizables y capacidades de impresión que generan en una variedad de formatos, incluyendo HTML y PDF (Kornspan, 2008).

- **Visual Paradigm for UML:** Es un Lenguaje de Modelado Unificado (UML), una herramienta profesional que implementa el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue (Technologies, 2010).

- **CASE Studio**: Potente utilidad de modelado para varias bases de datos. Es una herramienta profesional con la que se podrán diseñar bases de datos propias y facilita herramientas para la creación de diagramas de relación, modelado de datos y gestión de estructuras. Tiene soporte para trabajar con una amplia variedad de formatos de base de datos, entre ellos: Oracle, SQL, MySQL, PostgreSQL, Access) (Technologies, 2010).

1.9.1. Visual Paradigm for UML

Fig. 1.4. Visual Paradigm for UML.

Visual Paradigm for UML: Ayuda a la construcción de aplicaciones más rápidas, con la calidad requerida y a un menor costo. Permite representar todos los tipos de diagramas de clases, código inverso, generar código desde diagramas y generar documentación. Esta herramienta proporciona además, abundantes tutoriales, demostraciones interactivas y proyectos UML.

Características:

- Soporta aplicaciones web.
- Generación de código para Java y exportación como HTML.
- Fácil de instalar y actualizar.
- Compatibilidad entre ediciones.
- Diagramas de procesos de negocio: proceso, decisión, actor de negocio, documento.
- Modelado colaborativo con CVS y Subversion.
- Interoperabilidad con modelos UML2 (metamodelos UML 2.x para plataforma Eclipse) a través de XMI.
- Ingeniería inversa - código a modelo, código a diagrama.
- Generación de código - Modelo a código, diagrama a código.
- Editor de detalles de casos de uso, entorno todo en uno para la especificación de los detalles de los casos de uso, incluyendo la especificación del modelo general y de las descripciones de los casos de usos.
- Diagramas de flujo de datos.
- Generación de bases de datos. Transformación de diagramas de entidad-relación en tablas de base de datos.
- Ingeniería inversa de bases de datos. Desde SGBD existentes a diagramas de entidad-relación.

- Generador de informes para generación de documentación.
- Distribución automática de diagramas - reorganización de las figuras y conectores de los diagramas UML.
- Importación y exportación de ficheros XMI.
- Integración con Visio - dibujo de diagramas UML con plantillas de MS Viso Editor de figuras. (Technologies, 2010).

1.10. Herramientas de control de versiones

Actualmente en el mercado existen varias herramientas para realizar el control de versiones. Entre ellas se encuentran:

- **CVS**: Es el sistema de control de versiones más popular en la actualidad. Es robusto, probado mundialmente, y de software libre. Pero debido a diversas dificultades en mantener su código estable después de muchos años, y por la necesidad de elaborar nuevos paradigmas de herramientas de control de versión, fue creado el Subversion (CollabNet, Inc., 2001-2009).
- **Subversion**: El objetivo de este proyecto es construir un sistema de control de versiones que sería la sustitución de CVS en la comunidad de software libre, bajo una licencia de tipo Apache/BSD y se le conoce también como SVN, por ser ese el nombre de la herramienta de línea de comandos. Una característica importante de Subversion es que, a diferencia de lo que ocurre en CVS, los archivos versionados no tienen cada uno un número de revisión independiente. En cambio, todo el repositorio tiene un único número de versión que identifica un estado común de todos los archivos del repositorio en cierto punto del tiempo (CollabNet, Inc., 2001-2009).
- **TortoiseSVN**: Es un cliente de Subversion, implementado como una extensión shell de Windows. TortoiseSVN es realmente fácil de usar para el control de revisiones / control de versiones / control de código para Windows. Dado que no es una herramienta para un entorno integrado de desarrollo (IDE) específico, puede ser usada con cualquier herramienta de desarrollo a gusto. TortoiseSVN es de utilización libre (CollabNet, Inc., 2001-2009).
- **SVK**: Es un sistema de control de versiones descentralizado, construido con el robusto sistema de ficheros de Subversion. Soporta repositorio en espejo, desconectado de la operación, posee merge sensitivo a la historia y se integra con otros sistemas de control de versiones, así como las herramientas merge visuales más populares (CollabNet, Inc., 2001-2009).

1.10.1. TortoiseSVN

Fig. 1.5. SubVersion.

TortoiseSVN: Es un cliente gratuito de código abierto para el sistema de control de versiones Subversion. TortoiseSVN maneja ficheros y directorios a lo largo del tiempo y los ficheros se almacenan en un repositorio central. El repositorio es prácticamente lo mismo que un servidor de ficheros ordinario, salvo que recuerda todos los cambios que se hayan hecho a sus ficheros y directorios. Por ello puede recuperar versiones antiguas y examinar la historia de cuándo y cómo cambiaron sus datos, así como quién hizo el cambio. Algunos sistemas de control de versiones, son también sistemas de manejo de configuración del software (SCM) (Stefan Küng, 2006/10/13).

A continuación, se muestran algunas de sus características:

- Integración con el shell de Windows: TortoiseSVN se integra perfectamente en el shell de Windows (por ejemplo, el explorador). Esto significa que puede seguir trabajando con las herramientas que ya conoce. No tiene que cambiar a una aplicación diferente cada vez que necesite las funciones del control de versiones.
- Íconos sobreimpresionados: El estado de cada carpeta y fichero versionado, se indica por pequeños íconos sobreimpresionados. De esta forma, se puede ver fácilmente el estado en el que se encuentra su copia de trabajo.
- Fácil acceso a los comandos de Subversion: Todos los comandos de Subversion están disponibles desde el menú contextual del explorador. TortoiseSVN añade su propio submenú.

Conclusiones

Mediante el estudio realizado sobre los almacenes de datos en el mundo y en Cuba, se puede plantear que ninguna de las soluciones estudiadas cubre aún las necesidades que tiene la Oficina Nacional de Estadística de Cuba. Por tanto, se decide implementar una solución que dé respuesta al problema existente. Para ello:

- Se concilia que la metodología apropiada para resolver la problemática en cuestión, es **una adaptación de la metodología de Kimball**, desarrollada por DATEC.

- Se diseñará la solución mediante el **modelo de estrella**.
- Se utilizará el modo de almacenamiento **ROLAP**.
- Se desarrollará el sistema mediante las herramientas: **Visual Paradigm 3.4, PostgreSQL 8.4, PgAdmin III 1.10 y TortoiseSVN 2.0**.

Capítulo 2: Descripción de la solución

Introducción

Este capítulo aborda aspectos concernientes a la descripción e implementación de la solución, específicamente, a las características de las áreas de análisis, análisis de datos, la arquitectura, los metadatos, y el modelo de datos propuesto. De manera general, aborda el resultado del análisis y el diseño del Mercado de Datos de Inmigración y Extranjería, para el Departamento de Turismo y Comercio de la ONE.

2.1. Análisis

El análisis, es la distinción y la separación de las partes de un todo hasta llegar a conocer sus principios o elementos. Es el estudio de los límites, las características y las posibles soluciones de un problema al que se aplica un tratamiento; en este caso, por computadora.

2.1.1. Definición del negocio

La ONE, debido al papel que desempeña como entidad rectora y coordinadora de los temas estadísticos en Cuba, funciona como repositorio central, donde actúan un conjunto de procesos que ejecutan y supervisan la gestión estadística del país. En este sentido, el mecanismo de captación que se encuentra en vigor, está compuesto por diferentes procesos orientados a los distintos tipos de modelos estadísticos existentes. Los modelos de Estadísticas Continuas y Encuestas Periódicas son los principales afluentes de la ramificación de procesos definidos, aunque en algunos casos, como el objeto de este estudio, se utilizan los Registros Primarios de los organismos o instituciones que los originan. Las diferencias entre unos y otros, se basan principalmente en los períodos de captura (mensual, trimestral, semestral, nonestral y anual), y en características específicas de la recolección de información, así como el soporte (digital o impreso), de ella.

Una vez descrito brevemente el negocio de la ONE, para esta área, objeto de estudio, se puede resaltar que su proceso principal, es la captación de la información estadística, a todos los niveles, almacenándola desde el mayor nivel de detalle hasta los consolidados más densos y complejos, con el objetivo de permitir su disponibilidad para la consulta con la velocidad y eficacia requerida.

Las fuentes que se identifican en el presente trabajo están divididas en dos grupos: la información histórica que se recoge (pertenece a la categoría de datos archivados) y el conjunto de clasificadores

establecidos por la ONE para la clasificación de los datos con vistas a los diferentes análisis estadísticos (enmarcados en la categoría de datos internos).

Fuente: Información histórica 1994-2010

La ONE posee la fuente de información histórica, necesaria para la confección de los distintos análisis y reportes que se les solicitan. Esta fuente viene directamente del Departamento de Turismo y Comercio de la propia organización, y está compuesta por un conjunto de archivos, en formato DBF.

Los datos son suministrados por la DIE, a partir de sus Registros Primarios de control de la entrada y salida de visitantes al país; estos se obtienen una vez al mes, generalmente a partir del día 15 y se reciben grabados en un CD.

Una vez recogida la información, la Dirección de Informática de la ONE, revisa la consistencia y estructura de los datos, de no reunir estos requisitos se devuelve y se hace una nueva captación. Los cuales, se comprueban a través de un análisis manual. Ya revisada, se transforma y se hacen los reportes que se utilizan para los cálculos y análisis estadísticos, tanto por el área especializada de la ONE, como por el Ministerio de Turismo (MINTUR) y otras entidades relacionadas con el sector, a las cuales se les brinda servicio estadístico. Los datos provenientes de la DIE, en algunos casos como el clasificador de país, no corresponden con los utilizados por la ONE y el registro internacional, por lo que se realiza una correlación entre ambos clasificadores para llevar a cabo el proceso.

Ejemplo de la estructura que posee un fichero DBF con la información de entrada para la realización de los cálculos estadísticos:

TURISMO - Microsoft Excel

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	CONS	SEXO	CIU	MOT_VIAJE	PAIS_EMB	FECHA_ENT	F_NAC	FECHA_SAL	VUE_ENT	VUE_SAL	PTO_ENT	PTO_SAL	
2	1	M	535	34	408	070320	870927		AFR474		AJ		
3	1	F	535	34	408	070320	890804		AFR474		AJ		
4	1	F	535	34	408	070320	881114		AFR474		AJ		
5	1	F	535	34	408	070320	860601		AFR474		AJ		
6	1	M	504	34	428	071028	860523		IBE6621		AJ		
7	1	M	104	22	201	071225	911228		CBE7574		AJ		
8	1	M	757	34	428	071028	860609		IBE6621		AJ		
9	1	M	417	22	428	070618	400706		AEA051		AJ		
10	1	M	247	05	209	071213	910408		CMP438		AJ		
11	1	F	247	05	431	070415	720502		NOS730		AA		
12	1	M	418	22	408	071222	420317		AFR474		AJ		
13	1	F	247	05	209	071223	950819		CMP246		AJ		
14	1	F	247	05	209	071223	900501		CMP246		AJ		
15	1	F	205	34	202	071019	810215		LRC652		AJ		
16	1	F	247	05	431	071124	750515		VLE2918		AJ		
17	1	M	247	05	103	070614	031101		ACA964		AJ		
18	1	F	103	22	103	070228	900617		CUB191		AN		
19	1	F	103	22	103	070808	911105		SWG731		AN		
20	1	F	103	22	103	070808	560907		SWG731		AN		
21	1	M	247	05	201	071019	721023		CUB153		AJ		
22	1	F	247	05	420	071019	781029		CFG5196		AA		
23	1	F	247	05	420	071019	630316		CFG5196		AA		
24	1	F	305	31	305	070424	610912		CUB313		AJ		
25	1	M	419	22	201	071213	661215		CUB153		AJ		

Fig. 2.1 Estructura de la información en DBF.

Además, el Departamento de Turismo y Servicios de la ONE realiza algunos cálculos y procesos con estos datos para sus análisis particulares, los cuales los trabaja en la aplicación: Corel Paradox.

2.1.2. Temas de análisis

Las áreas de análisis son un punto importante en el desarrollo de los mercados de datos, pues marcan la diferencia entre sus estructuras, buscando fiabilidad, nuevas perspectivas y buen uso. En la solución, se proponen las áreas de acuerdo con la situación actual de la organización y la manera en que trabajan los datos.

El área de análisis que se identifica es:

- Departamento de Turismo y Comercio.

2.1.3. Roles y permisos

En la ONE sólo el personal que atiende el Departamento de Turismo y Comercio tiene acceso a los datos, con excepción de los especialistas de informatización, pues son los que realizan los reportes que se utilizan para los cálculos estadísticos. Por lo que se definen como roles y permisos:

- Los roles se representan por: el analista y el administrador. El analista es la persona que trabaja con el manejo de la información y hace uso de ella para los cálculos estadísticos; de igual manera, se encarga de administrar y ejecutar los reportes, actualizar los datos, y funciona como el usuario frecuente del sistema. El administrador, trabaja en el proceso de extracción, limpieza, transformación y carga de los datos, es quien trabaja directamente con la base de datos, así como en todo lo referente a su configuración y administración.
- Los permisos de acceso y el trabajo con el sistema están definidos como se informa a continuación:

  - ✓ El analista: Permiso de lectura, escritura y actualización.
  - ✓ El administrador: Permiso de lectura y escritura.

Ningún usuario fuera de esta clasificación podrá tener permisos de acceso al sistema. El tiempo de entrenamiento a los usuarios normales y avanzados está comprendido de 2 a 3 semanas.

2.1.4. Reglas del negocio

Las reglas del negocio describen las políticas, normas, operaciones, definiciones y restricciones presentes en una organización, y son de vital importancia para alcanzar los objetivos misionales, pues actúan como un medio por el cual la estrategia es implementada. Especifican en un nivel adecuado de

detalle lo que una organización debe hacer. Las reglas del negocio deben ser expresadas en lenguaje natural y orientadas al negocio.

Las organizaciones funcionan siguiendo múltiples reglas del negocio, explícitas o tácitas, que están divididas en procesos, aplicaciones informáticas y/o documentos. Pueden ser parte del conocimiento de las personas o estar presentes en el código fuente de programas informáticos. En los últimos años se viene observando una tendencia a gestionar de forma sistemática y centralizada las reglas de negocio, de modo que sea fácil y sencillo consultarlas, entenderlas, utilizarlas y cambiarlas (ScriBD, 2010).

Para poder llevar a cabo los cálculos y análisis estadísticos en el Departamento de Turismo y Comercio de la ONE, es necesario organizar la información, por lo que se hace más factible el trabajo clasificando los diferentes datos por códigos. En el caso de la situación geográfica, la DIE tiene una codificación que no es la misma que utiliza el Departamento de Turismo y Comercio de la ONE, por lo que la DIE debe utilizar el clasificador de país internacional, puesto en vigor por la ONE. (Ver **Anexo 1**).

En la actualidad, el Departamento de Turismo y Comercio, tiene establecidas algunas reglas a seguir para la realización de los cálculos estadísticos, las cuales se mantienen en la implementación de la solución. Dentro de las reglas implantadas están:

- Variación absoluta = Al resto del valor total del período actual - el valor total del período anterior con el que se está comparando.
- Variación relativa = A la división del valor total del período actual, entre el valor total del período anterior con el que se está comparando, todo ello, multiplicado por 100 para que salga en por ciento.
- Acumulado = A la suma por trimestre.
- Edad = Año actual – Año de nacimiento.
- Estancia:

Una regla importante para la elaboración de estos cálculos estadísticos, es la que se refiere a la **estancia media**, la cual actualmente no está incluida en los reportes que se realizan y que se considera primordial para el análisis, por lo que se propone incluirla en la solución y establecerla como otra regla elemental para el negocio.

Para calcular la **estancia**, se debe tener en cuenta que en el fichero existen dos campos: la fecha de entrada y la fecha de salida, pero este último no siempre tiene datos, por lo que para calcular los **días**

de estancia se debe restar a la fecha de salida la fecha de entrada, y en el caso de que no existan datos en la fecha de salida, se toma la fecha de cierre de la información. Para hallar la **estancia media**, que debe calcularse por país de ciudadanía, se suma la cantidad de turistas por cada país y la cantidad de días de estancia de los turistas de cada país y se divide la cantidad de días entre la cantidad de turistas. Debe calcularse también una **estancia media total**, o sea, la sumatoria de todos los días de estancia de todos los turistas, entre el total de turistas del fichero.

2.2. Necesidades de los usuarios

Con la solución actual se gestionan las necesidades informáticas de los usuarios del Departamento de Turismo y Comercio de la ONE, lo cual se efectúa en función del análisis que se les realiza a todos los visitantes que entran a la isla, a través de reportes por sexo, edad, ciudadanía, motivos de viaje, puntos aeroportuarios y país de embarque.

Como problemática actual, se puede apreciar que para los cálculos y análisis estadísticos, se utilizan los datos de la ciudadanía, cuando en realidad lo que necesitan es realizar esta operación como se hace internacionalmente, a través de los datos de la residencia.

2.2.1. Requisitos de información

Es la información que debe estar disponible, la entrada fundamental para todo el proceso de inteligencia del negocio y los futuros reportes bases.

Los reportes que se efectúan con dicha información son:

- Visitantes:
  - ✓ Obtener la serie de llegadas de visitantes.
  - ✓ Calcular los principales emisores de visitantes a Cuba.
  - ✓ Calcular el arribo de visitantes por áreas geográficas.
  - ✓ Obtener la serie de llegadas de visitantes por sexo, edad, ciudadanía, motivos de viaje, puntos aeroportuarios y país de embarque (Mes acumulado).
- Turistas:
  - ✓ Obtener la serie de llegadas de turistas.
  - ✓ Calcular los principales emisores de turistas a Cuba.
  - ✓ Calcular el arribo de turistas por áreas geográficas.

- ✓ Obtener la serie de llegadas de turistas por sexo, edad, ciudadanía, motivos de viaje, puntos aeroportuarios y país de embarque (Mes acumulado).

Existen reportes que no están informatizados. El departamento no tiene soluciones para ello, pues la aplicación que utilizan no los proporciona; esto trae como consecuencia que haya que hacer este trabajo manual. Estos reportes se pueden sacar de los datos que son recogidos de la DIE:

- Promedio de estancia por ciudadanía.
- Promedio de estancia por provincia.

Con la información recogida del Departamento de Turismo y Comercio, se calcula el por ciento de visitantes y turistas que entraron al país, lo cual se da a conocer a todos los interesados como noticia, y es una información que más tarde es publicada en el sitio de la organización.

Debido a las necesidades anteriores, se propone informatizar y generar en la solución todos los reportes que en este momento son manuales, así como todos los que se necesiten, para resolver los problemas actuales del Departamento de Turismo y Comercio de la ONE.

2.2.2. Requisitos Multidimensionales

Los requisitos multidimensionales constituyen la entrada fundamental para el diseño de las estructuras del almacén. Parámetros de entrada y salida de las solicitudes de información de los clientes.

- **Arribo de Visitantes:**

VE: Variables de entrada:

- Ciudadanía.
- Sexo.
- Puntos Aeroportuarios.
- Motivos de Viaje.
- País de Embarque.
- Tiempo.
- Edad.

VS: Variables de salida:

- Cantidad de arribos por motivos de viaje (por ciudadanía).

- Acumulado de arribos por motivos de viaje (por ciudadanía).
- Cantidad de llegadas de visitantes.
- Acumulado por principales emisores.
- Variación absoluta de los principales emisores.
- Cantidad de arribos por área geográfica.
- Estancia media de los visitantes.
- Cantidad de días de estancia de los visitantes.

- **Arribo de Turistas:**

VE: Variables de entrada:

- Ciudadanía.
- Sexo.
- Puntos Aeroportuarios.
- Motivos de Viaje.
- País de Embarque.
- Tiempo.
- Edad.

VS: Variable de salida:

- Cantidad de arribos por motivos de viaje (por ciudadanía).
- Acumulado de arribos por motivos de viaje (por ciudadanía).
- Cantidad de llegadas de turistas.
- Acumulado por principales emisores.
- Variación absoluta de los principales emisores.
- Cantidad de arribos por área geográfica.
- Estancia media de los turistas.
- Cantidad de días de estancia de los turistas.

2.2.3. Requisitos funcionales

Los requerimientos funcionales son capacidades o condiciones que el sistema debe cumplir. En la realización de los casos de uso del negocio, se obtienen las actividades que serán objeto de automatización. Estas actividades no son exactamente los requerimientos funcionales, pero sí son el

punto de partida para identificar qué debe hacer el sistema. Los requerimientos funcionales no alteran la funcionalidad del producto: esto quiere decir que se mantienen invariables, sin importar con qué propiedades o cualidades se relacionen.

Los requisitos funcionales de la solución están compuestos por:

- Extracción de los datos de la fuente de la DIE: Consiste en sustraer los datos brutos desde los sistemas de origen, los datos de estas fuentes primarias pueden encontrarse sobre arquitecturas, cada sistema separado puede usar una organización diferente de los datos o formatos distintos.
- Transformación de los datos de la fuente de la DIE: En esta fase se aplican una serie de procedimientos de negocios sobre los datos extraídos, con el objeto de convertirlos en datos aptos para ser cargados.
- Limpieza de los datos de la fuente de la DIE: En esta fase del proceso, las características fundamentales están enfocadas en identificar los datos redundantes, los valores atípicos, los valores perdidos, además de corregir, estandarizar y completar los que estén en uso.
- Carga de los datos de la fuente de la DIE: La fase de carga es el momento cuando los datos, provenientes de la fase anterior, son incluidos en el sistema de destino.
- *Backup* de los datos: Es la forma de crear salvas de la información importante y necesaria en la organización, para dar continuidad al trabajo.
- Administración de reportes: Controlar la entrada y salida de reportes del mercado de datos, crearlos, configurarlos y administrarlos.

2.2.4. Requisitos no funcionales

Los requerimientos no funcionales son propiedades o cualidades que el producto debe tener. Debe pensarse en estas propiedades como las características que hacen al producto atractivo, usable, rápido o confiable. Los requerimientos no funcionales forman una parte significativa de la especificación. Son importantes para que clientes y usuarios puedan valorar las características no funcionales del producto.

- Software: Se deberá disponer de un sistema operativo Windows XP o superior, Ubuntu 9.10 o Debian 6.0, aplicaciones como Java Virtual Machine 6 Update 2 o superior y un servidor de base de datos (se recomienda un servidor PostgreSQL v 8.4 o superior).
- Soporte: Una vez terminado el software, se asistirá a los clientes por un período de tiempo de 6 meses, con motivo de lograr su mejoramiento progresivo y evolución en el tiempo. Esta

asistencia incluye: pruebas de chequeo, extensibilidad, adaptabilidad, mantenimiento, compatibilidad, configuración, servicios, instalación, internacionalización y requerimientos de portabilidad.

- Usabilidad: Estos requerimientos describen los niveles apropiados de usabilidad, dados los usuarios finales del producto, para ello se tienen en cuenta las especificaciones de los perfiles de usuarios (definidos en roles y permisos) y las clasificaciones de sus niveles de experiencia. Se exige que los usuarios tengan un conocimiento básico de informática, para poder complementar su utilización.
- Eficiencia: La eficiencia del producto se mide por la calidad de las funcionalidades solicitadas por los clientes y su mantenimiento preventivo. Deberá efectuarse mediante una visita semestral durante el período de vigencia de la garantía técnica. El mantenimiento correctivo deberá efectuarse anualmente, a requerimiento de la ONE con tiempos de respuesta por soporte técnico de 4 horas y de solución de 48 horas los 365 días del año, de fácil ubicación vía e-mail o telefónica. Se deben considerar actualizaciones de software: parches y sistema operativo, los mismos que deben ser recomendados por los fabricantes de los productos entregados.
- Seguridad: Este es, quizás, el tipo de requerimiento más difícil, que provoca los mayores riesgos si no se maneja correctamente. Por tanto la seguridad de la solución será tratada en tres aspectos diferentes:

    - ✓ Confiabilidad: La información manejada por el sistema estará protegida de acceso no autorizado y divulgación.
    - ✓ Integridad: La información manejada por el sistema será objeto de cuidadosa protección contra la corrupción y estados inconsistentes, de igual manera será considerada la fuente o autoridad de los datos. Pueden incluir también mecanismos de chequeo de integridad y realización de auditorías.
    - ✓ Disponibilidad: A Los usuarios autorizados se les garantizará el acceso a la información y los dispositivos o mecanismos utilizados para lograr la seguridad, no ocultarán o retrasarán a los usuarios la obtención de los datos deseados en un momento dado. El sistema debe estar disponible siempre entre las 8:00 am y las 5:00 pm de lunes a viernes y los sábados de 8:00 am a 12:00 pm. Se debe tener en

cuenta que de acuerdo con las necesidades de trabajo, se podrá extender el uso del sistema.

- Rendimiento de las consultas: La mayoría de las consultas del almacenamiento de datos están diseñadas para seguir un esquema de estrella y pueden procesar centenares de millones de filas en una única consulta. De manera predeterminada, el optimizador detectará las consultas en los esquemas de estrella y creará planes eficaces para ellos (Microsoft Corporation, 2009).
  Esta aplicación utiliza el modelo de estrella, permitiendo que el rendimiento de las consultas sea elevado para dar respuesta a los pedidos de trabajo.

- Rendimiento de la carga: Es el comportamiento de la aplicación bajo una cantidad de peticiones esperada. Esta carga puede ser el número esperado de usuarios concurrentes que utilizan la aplicación y que realizan un número específico de transacciones durante el tiempo que dura la carga (Microsoft Corporation, 2009).
  Los tiempos de respuesta de todas las transacciones importantes de la aplicación, están aproximadamente de 1 a 3 minutos.

2.2.5. Casos de uso del sistema

Los casos de uso del sistema están especificados por:

Casos de uso de información:

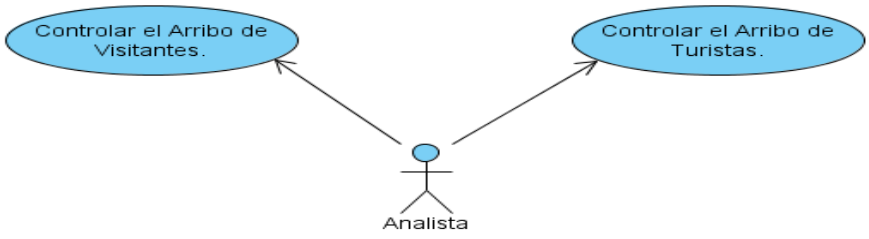

Fig. 2.2 Esquema de casos de uso de información.

Caso de uso Controlar el Arribo de Visitantes: En este caso uso se va a manejar toda la información referente al arribo de los visitantes a Cuba.

Caso de uso Controlar el Arribo de Turistas: En este caso uso se va a manejar toda la información referente al arribo de los turistas a Cuba.

Casos de uso funcionales:

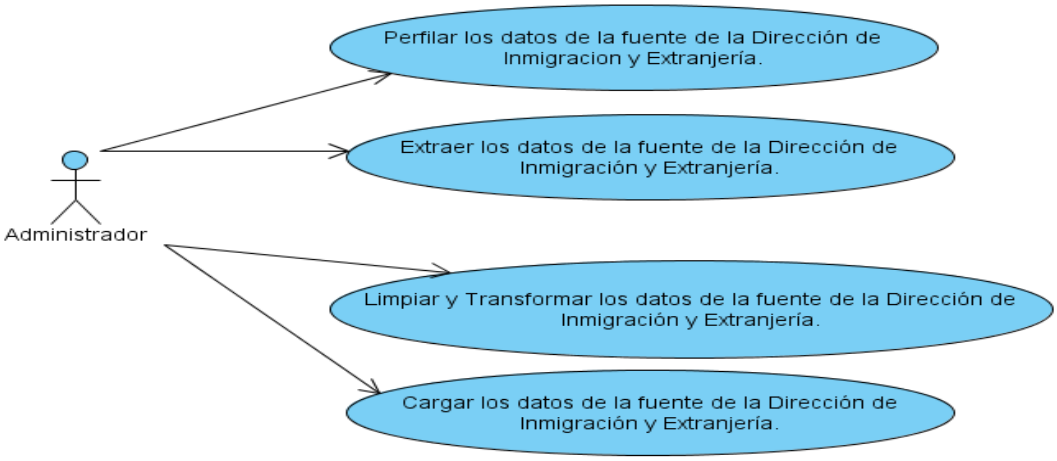

Fig. 2.3 Esquema de casos de uso funcionales.

Caso de uso: Perfilar los de datos de la fuente de la DIE: En este caso uso la actividad fundamental es el perfilado de datos de la fuente, que consiste en la revisión de los datos de la fuente para la detección de errores.

Casos de uso: Extraer los datos de la fuente de la DIE: En este caso de uso la actividad fundamental es la extracción de los datos de la fuente.

Caso de uso: Limpiar y Transformar los datos de la fuente de la DIE: En este caso de uso hay dos actividades fundamentales, la limpieza de los datos que se extraen de la fuente, para evitar inconsistencia y la transformación de los datos, para adaptarlos a las necesidades del cliente y de esta manera asegurar su disponibilidad para el uso posterior.

Caso de uso: Cargar los datos de la fuente de la DIE: En este caso de uso la actividad fundamental es la carga de los datos, así como los nomencladores definidos, al mercado de datos.

2.3. Diseño

En la solución se ha definido un mercado de datos donde convergen todas las dimensiones propuestas: Sexo, Edad, Ciudadanía, Motivos de Viaje, Puntos Aeroportuarios, País de Embarque y Tiempo. Los hechos mensurables del mercado de datos, son los aspectos que se evalúan en el modelo estadístico: Arribo de Visitantes y Arribo de Turistas, también se mide la estancia media. Una de las razones por las cuales se propuso como metodología de desarrollo la de Kimball, es por la posibilidad que brinda de ir desarrollando un proceso a la vez.

2.3.1. Matriz BUS

La matriz BUS, tiene como propósito obtener un modelo lógico inicial y no es más que la relación que existe entre las dimensiones y los hechos del almacén de datos.

Tablas de Hechos.

- Arribo de Visitantes.
- Arribo de Turistas.

Tablas de Dimensiones.

- Dimensión Ciudadanía.
- Dimensión Sexo.
- Dimensión Puntos Aeroportuarios.
- Dimensión Motivos de Viaje.
- Dimensión País de Embarque.
- Dimensión Tiempo.
- Dimensión Edad.

Matriz BUS:

Dimensiones /Tabla de Hechos.	Arribo de Visitantes	Arribo de Turistas
dim_ciudadania	x	x
dim_sexo	x	x
dim_puntos_aeroportuarios	x	x
dim_motivos_de_viaje	x	Turistas
dim_pais_de_embarque	x	x
dim_tiempo	x	x
dim_edad	x	x

2.4. Modelo de datos

El modelo de datos lo integra el modelo dimensional. Su enfoque viene precedido por el esclarecimiento de las dimensiones, medidas, granularidad y hechos identificados. De esta manera se procede a la estructuración del modelo dimensional que tendrá el sistema. Se destacan que, por las

necesidades actuales del negocio existen varios modelos que relacionan las dimensiones definidas y las medidas que se han detallado hasta el momento.

En la solución propuesta se escogió el modelo tipo estrella para su desarrollo, el cual tendrá una tabla de hechos llamada hec_arribo_de_visitantes, que se relacionará con las 7 dimensiones propuestas, al igual que una tabla de una vista materializada nombrada vm_arribo_de_turistas. Todas las dimensiones identificadas, tienen una llave primaria, que se encarga de mantener la integridad referencial entre ellas y la tabla de hechos. Esta llave no conserva ningún tipo de significado dentro del negocio. Simplemente, es un número que responde a la unicidad.

Ejemplo del modelo de datos lógico elaborado para el sistema:

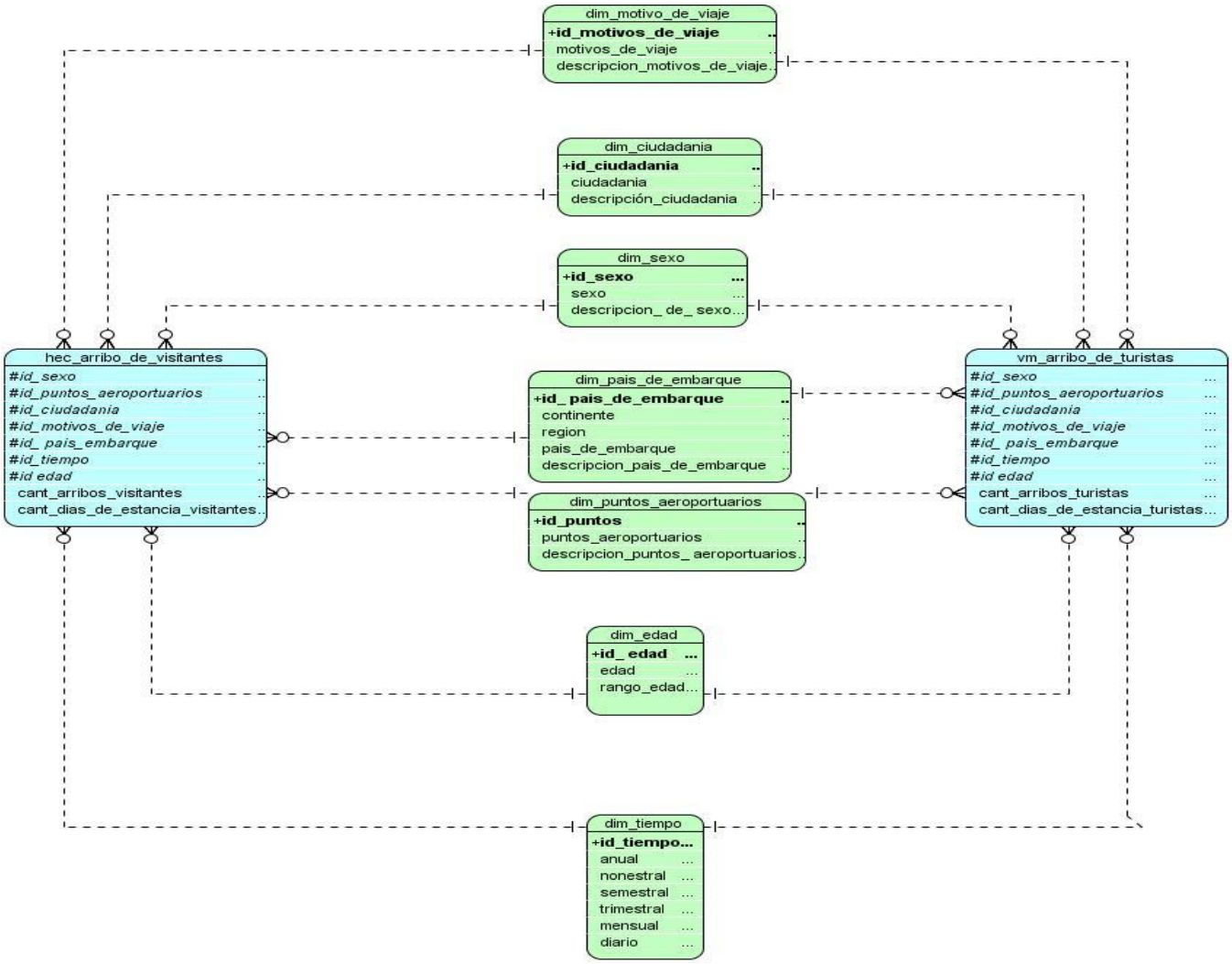

Fig 2.4. Modelo dimensional de la solución.

2.4.1. Dimensiones

Después de haber declarado el modelo de datos del proceso, se detallan las dimensiones, las cuales poseen características esenciales como: la definición de jerarquías entre sus atributos, con el objetivo de plasmar explícitamente la forma en que se pueden consolidar los datos.

A continuación se describen las dimensiones y jerarquías que están relacionadas con el repositorio principal donde se va a almacenar la información:

Dimensión Ciudadanía: Describe el universo de valores bajo los cuales puede clasificarse la información atendiendo al nomenclador de ciudadanía.

Jerarquía:
Continente.
Continente->Región.
Continente->Región->Ciudadanía.

Nombre del atributo	Descripción	Cardinalidad
id_ciudadania	Es la llave primaria de la dimensión. No posee significado para el negocio.	10
ciudadania	Almacena los códigos de las ciudades.	10000
descripcion_ciudadania	Descripción de las ciudades.	10

Dimensión Sexo: Describe el universo de valores bajo los cuales puede clasificarse el sexo.

Jerarquía: Sexo.

Nombre del atributo	Descripción	Cardinalidad
id_sexo	Es la llave primaria de la dimensión. No posee significado para el negocio.	10
sexo	Almacena los códigos de los sexos.	255

descripcion_sexo	Descripción de los sexos.	

Dimensión Puntos Aeroportuarios: Describe el universo de valores bajo los cuales pueden clasificarse los puntos de entrada de los visitantes.

Jerarquía: Puntos Aeroportuarios.

Nombre del atributo	Descripción	Cardinalidad
id_puntos_aeroportuarios	Es la llave primaria de la dimensión. No posee significado para el negocio.	10
puntos_aeroportuarios	Almacena los códigos de los puntos de entrada y de salida de los visitantes.	10000
descripcion_puntos_aeroportuarios	Almacena los nombres de los puntos.	1000

Dimensión Motivos de Viaje: Describe el universo de valores bajo los cuales puede clasificarse la información atendiendo al nomenclador de los motivos de viaje.

Jerarquía: Motivos de Viaje.

Nombre del atributo	Descripción	Cardinalidad
id_motivos_de_viaje	Es la llave primaria de la dimensión. No posee significado para el negocio.	10
motivos_de_viaje	Almacena los códigos de los motivos de viajes.	10000
descripcion_motivos_de_viaje	Descripción de los motivos de viajes.	10000

Dimensión País de Embarque: Describe el universo de valores bajo los cuales puede clasificarse la información atendiendo al nomenclador de país de embarque.

Jerarquía:

Continente.

Continente->Región.

Continente->Región->País de Embarque.

Nombre del atributo	Descripción	Cardinalidad
id_pais_de_embarque	Es la llave primaria de la dimensión. No posee significado para el negocio.	10
pais_de_embarque	Almacena los códigos de los países de embarque.	10000
continente	Almacena los códigos de los continentes.	10000
region	Almacena los códigos de las regiones.	10000
descripcion_pais_de_embarque	Descripción de los países de embarque.	10

Dimensión Tiempo: Es la más común e importante en los diseños de mercados de datos ya que define una línea de tiempo específica.

Jerarquía:

Anual.

Anual ->Nonestral.

Anual ->Nonestral->Semestral.

Anual ->Nonestral->Semestral->Trimestral.

Anual ->Nonestral->Semestral->Trimestral->Mensual.

Anual ->Nonestral->Semestral->Trimestral->Mensual-> Diario.

Nombre del atributo	Descripción	Cardinalidad
id_tiempo	Es la llave primaria de la dimensión. No posee significado para el negocio.	10

anual	Almacena los datos anuales	10000
nonestral	Almacena los datos nonestrales.	10000
semestral	Almacena los datos semestrales.	10000
trimestral	Almacena los datos trimestrales.	10000
mensual	Almacena los datos mensuales.	10000
diario	Almacena los datos diarios.	10000

Dimensión Edad: Describe la edad de los visitantes.

Jerarquía: Edad.

Nombre del atributo	**Descripción**	**Cardinalidad**
id_edad	Llave primaria	10
edad	Comprende la edad puntual de la persona.	10000
rango_edad	Comprende la edad de las todos los rangos que se pueden formar, con respecto a la edad de las personas.	10000

2.4.2. Tablas de hechos

Las tablas de hechos son las que almacenan las medidas numéricas. En este caso, se definieron como medidas numéricas los 7 valores que se captan en los ficheros de la DIE. Las tablas de hechos identificadas se describen a continuación:

- Tabla de Hechos Arribo de Visitantes:

En esta tabla es donde va a residir, toda la información existente del modelo estadístico. Es la tabla que servirá como fuente de información principal para la realización de las estructuras que soporten los reportes más comunes de la institución enmarcada en el Departamento de Turismo y Comercio, y específicamente en el arribo de visitantes a Cuba.

- Tabla de Hechos de Arribo de Turistas:

En esta tabla es donde se va a almacenar la información concerniente a los visitantes con motivos de viaje para el turismo. Servirá como fuente de información principal para la realización de las estructuras que soporten los reportes más comunes de la institución, que contengan información de los arribos de los turistas a Cuba. (Esta tabla es una vista materializada que se comporta como un hecho).

La granularidad en la tabla de hechos se determina después de identificar las columnas que existirán en dichas tablas. La granularidad, como concepto, es una medida del nivel de detalle enfocada a cada ocurrencia que exista en la tabla de hechos. Por ello, se puede inferir la estrecha relación existente entre las dimensiones y la granularidad. Es recomendable no mezclar varias granularidades en una misma tabla de hechos, ni almacenar, en dicha tabla, sumas, promedios, porcientos o resúmenes, debido a que contradicen la filosofía de almacenar el mínimo detalle de la información. En casos como esos, se deben almacenar dichos resúmenes o agregados en tablas separadas con sus respectivos niveles de granularidad. Además de la importancia que reviste el mantenimiento de la mínima granularidad dentro del diseño de los almacenes de datos, en ocasiones también es aconsejable almacenar la información en niveles de detalles intermedios, es decir, con altos niveles de granularidad debido a que podría ser beneficioso para empresas que no requieran altos niveles de detalles para el análisis de su información.

Después del análisis anterior se puede concluir que la granularidad del repositorio central de la solución propuesta está dada por el registro del dato estadístico captado en un mes determinado.

2.4.3. Medidas

Las medidas están implícitas en las Tablas de Hechos:

- Cantidad de arribos de visitantes.
  - ✓ (cant_arribos_visitantes).
- Cantidad de días de estancia de los visitantes.
  - ✓ (cant_días_de_estancia_visitantes).
- Cantidad de arribos de turistas.
  - ✓ (cant_arribos_turistas).
- Cantidad de días de estancia de los turistas.
  - ✓ (cant_días_de_estancia_turistas).

2.5. Esquema de seguridad

El esquema de seguridad está respaldado por los niveles de acceso al sistema, específicamente por los roles definidos. La solución del sistema de seguridad y alta disponibilidad debe ser de arquitectura de 3 capas:

- Funcionamiento: Dispositivos de seguridad (firewall, inspectores de contenido, sensores).
- Servidores: Servidor de gestión de administración y de base de datos.
- Presentación: Consolas de administración.

La solución debe tener una interfaz de administración gráfica en la consola de administración, la cual proporciona una visualización de la topología de red en el editor de políticas y la relación entre estos objetos y la red. Para la auditoría del tráfico entrante y saliente, así como para la generación de informes de eventos e intentos de ataque, se debe utilizar un registro del tipo *log file* como un evento que deba ser registrado. El formato de las entradas del registro, pedidas por una regla, es determinado por el tipo del registro especificado en la regla, todo esto sin alterar el rendimiento de acceso a los equipos protegidos por la solución. (Mafla, 21/03/2005)

La configuración y pruebas de todos los equipos deberán efectuarse en las instalaciones de la ONE en Ciudad de La Habana, conjuntamente con personal técnico de Informática, de acuerdo con las especificaciones que serán elaboradas por la Dirección de Informática de la ONE.

El mantenimiento preventivo deberá efectuarse mediante una visita semestral durante el período de vigencia de la garantía técnica y el mantenimiento correctivo deberá efectuarse anualmente, a requerimiento de la ONE con tiempos de respuesta por soporte técnico de 4 horas y de solución de 48 horas los 365 días del año, de fácil ubicación vía correo electrónico o telefónica. Se deben considerar actualizaciones de software, parches y sistema operativo, de igual manera deben ser recomendados por los fabricantes de los productos entregados.

2.6. Política de respaldo y recuperación

La política de respaldo y recuperación que utiliza la solución es sencilla pero a la vez sólida, por ello se miden 3 puntos esenciales:

- Periodicidad de las salvas: Las salvas se realizan cada 6 meses mediante una cruzada de servidores: se hacen 2 copias, una se almacena en la ONE y la otra fuera de la organización. El Banco de Datos es el área que se encarga de este trabajo, certificando en todo momento que exista una copia estricta de la información que está presente en el servidor.

- Tablas involucradas: Las tablas que se involucran en las salvas son tabla de hechos Arribo de Visitantes, la vista materializada Arribo de Turistas y las tablas de las dimensiones: Sexo, Edad, Ciudadanía, Motivos de Viaje, Puntos Aeroportuarios y País de Embarque.
- Backups existentes: Actualmente no existen backups en el Departamento de Turismo y Comercio de la ONE.

    - ✓ Periodicidad de reemplazo de los backups: Se realizan los reemplazos de backups cada 6 meses.
    - ✓ Periodicidad de las pruebas a los backups: El estado de los backups se verifica mensualmente por un grupo de chequeo interno de la propia organización, mediante pruebas de rendimiento y flexibilidad, y cada 6 meses por una comisión de auditoría, primero en marzo y más tarde en septiembre.

Conclusiones

A partir del estudio realizado, se concluye del análisis y diseño de la solución, se destacan los siguientes aspectos:

- Del estudio preliminar realizado en el proceso del negocio se identificó como fuente de datos la **información histórica de 1994-2010** y como **área de análisis el Departamento de Turismo y Comercio de la ONE**.
- Partiendo de las normas y definiciones presentes en el Departamento de Turismo y Comercio, se conformaron **5 reglas del negocio** a medir.
- Para darle cumpliendo a las políticas de seguridad con que cuenta el Departamento de Turismo y Comercio se identificaron **2 roles con sus permisos** correspondientes, los cuales aseguran que cada usuario accederá a la información que le corresponde en el sistema.
- A partir del análisis realizado se conformaron **7 dimensiones**, **una tabla de hechos** y **una vista materializada**, los cuales integraron el **diseño del modelo lógico** que estructura el mercado de datos.
- Teniendo en cuenta los niveles de acceso analizados, se establecieron **4 medidas** y **3 políticas de seguridad** a seguir en el sistema.

Capítulo 3: Implementación y Prueba

Introducción

En el presente capítulo, se describe el proceso de carga de los datos a las dimensiones de la solución. Se detallan las pruebas de implantación, se valida la solución a través del empleo de las listas de chequeo y la carta de aceptación del cliente.

3.1. Modelo de datos físico

Los modelos de datos se pueden clasificar dependiendo de los tipos de conceptos que ofrecen para describir la estructura de la base de datos. Los modelos de datos de alto nivel, o modelos conceptuales, disponen de conceptos muy cercanos al modo en que la mayoría de los usuarios percibe los datos, mientras que los modelos de datos de bajo nivel, o modelos físicos, proporcionan conceptos que describen los detalles de cómo se almacenan los datos en el computador. Los conceptos de los modelos físicos están dirigidos al personal informático, no a los usuarios finales. Entre estos dos extremos se encuentran los modelos lógicos, cuyos conceptos pueden ser entendidos por los usuarios finales, aunque no están demasiado alejados de la forma en que los datos se organizan físicamente (Universitat Jaume * I, 2001).

Los modelos físicos, poseen tablas de mantenimiento que habitualmente no se incluyen en el modelo lógico. Difieren mayormente en la especificación absoluta y puntualizada de las características físicas de la base de datos, comenzando por los tipos de datos hasta las tablas de segmentación, parámetros de almacenamiento de tablas y bandas de discos. En este trabajo se diseñó el modelo dimensional de la solución, que recoge toda la información de interés para el cliente.

3.2. Estructuras de datos

Inferior a las estructuras de datos, hay un nivel, en el cual se sitúan los archivos, discos, particiones y espacios de tablas. El uso adecuado de dichos elementos y su dominio, inciden representativamente en el éxito de la solución. Los elementos que se van a priorizar, en tanto se efectúa el desarrollo, son el particionamiento de las tablas, en función de lograr una mayor organización de la información y velocidad en su recuperación, y estructuras de control de cambios con el fin de minimizar la utilización de recursos físicos cuando se refresquen las agregaciones.

3.2.1. Esquemas y tablas

A la descripción de una base de datos mediante un modelo de datos, se le denomina esquema de la base de datos. Este esquema se especifica durante el diseño, y no es de esperar que se modifique a menudo. Sin embargo, los datos que se almacenan en la base de datos pueden cambiar con mucha frecuencia pues se insertan continuamente.

La distinción entre el esquema y el estado de la base de datos es elemental. Cuando se define una nueva base de datos, sólo se especifica su esquema al SGBD. En ese momento, el estado de la base de datos es el ¨estado vacío¨, sin datos. Cuando se cargan datos por primera vez, la base datos pasa al ¨estado inicial¨. De ahí en adelante, siempre que se realice una operación de actualización, se tendrá un nuevo estado. El SGBD se encarga, en parte, de garantizar que todos los estados sean estados válidos, que satisfagan la estructura y las restricciones especificadas en el esquema. Por lo tanto, es muy importante que el esquema que se especifique al SGBD sea correcto y se debe tener muchísimo cuidado al diseñarlo. El SGBD almacena el esquema en su catálogo o diccionario de datos, de modo que se pueda consultar siempre que sea necesario (Universitat Jaume * I, 2001).

Esquemas:

Para garantizar las estructuras de datos, se definieron 3 esquemas, en los que se almacenan las tablas:

- Esquema Dimensiones: comprende las tablas de Ciudadanía, Sexo, Motivos de Viaje, Tiempo, Edad, Puntos Aeroportuarios y País de Embarque, las cuales aparecen implementadas en el script DDL_InmEx.sql del expediente de proyecto.

  Script para la creación del esquema Dimensiones:
  ESQUEMA: Dimensiones
  PROPIETARIO: postgres
  CODIGO FUENTE: CREATE SCHEMA "dimensiones" AUTHORIZATION "postgres";
  COMMENT ON SCHEMA "dimensiones"
  IS 'standard public schema';

- Esquema Hechos: comprende la tabla arribo de visitantes, la cual se implementa en el script DDL_ InmEx.sql del expediente de proyecto.

  Script para la creación del esquema Hechos:

ESQUEMA: Hechos.

PROPIETARIO: postgres

CODIGO FUENTE: CREATE SCHEMA "hechos" AUTHORIZATION "postgres";

- Esquema Vistas materializadas: comprende la tabla arribo de turistas, la cual se implementa en el script DDL_ InmEx.sql del expediente de proyecto.

Script para la creación del esquema Vistas materializadas:

ESQUEMA: Vistas_materializadas.

PROPIETARIO: postgres

CREATE SCHEMA "Vistas_materializadas" AUTHORIZATION "Admnistrador";

Tablas:

Cada una de las entidades modeladas en el análisis representa una tabla en el diseño. Por tratarse de la necesidad de almacenamiento, en una tabla se recolectan muchos registros.

- **Tabla Edad:**

ESQUEMA: Dimensiones.

PROPIETARIO: postgres.

DESCRIPCIÓN: Esta tabla corresponde a la dimensión Edad.

PK	Nombre	Tipo de Dato	Descripción
x	id_edad	integer	Llave primaria
	edad	integer	Comprende la edad puntual de la persona.
	rango_edad	varchar(50)	Comprende la edad de las todos los rangos que se pueden formar, con respecto a la edad de las personas.

- **Tabla Tiempo:**

ESQUEMA: dimensiones

PROPIETARIO: postgres

DESCRIPCIÓN: Esta tabla corresponde a la dimensión Tiempo.

	Nombre	Tipo de Dato	Descripción
x	id_tiempo	integer	Llave primaria
	diaria	integer	Diaria
	mensual	varchar(50)	Mensual
	trimestral	varchar(50)	Cada tres meses
	semestral	text	Cada seis meses
	nonestral	text	Cada nueve meses
	anual	text	Anual

- **Tabla Ciudadanía:**

ESQUEMA: Dimensiones.

PROPIETARIO: postgres.

DESCRIPCIÓN: Esta tabla corresponde a la dimensión Ciudadanía.

PK	Nombre	Tipo de Dato	Descripción
x	id_ciudadania	serial	Llave primaria
	continente	varchar(50)	Refleja los continentes
	region	varchar(50)	Refleja las regiones
	ciudadania	varchar(50)	Refleja la ciudadanía
	descripcion_ciudadania	varchar(50)	Descripción de la ciudadanía

- **Tabla Motivos de Viaje:**

ESQUEMA: dimensiones

PROPIETARIO: postgres

DESCRIPCIÓN: Esta tabla corresponde a la dimensión Motivos de viaje.

PK	Nombre	Tipo de Dato	Descripción
x	id_motivos_de_viaje	integer	Llave primaria

	motivos_de_viaje	integer	Refleja los motivos de viaje.
	descripcion_motivos_de_viaje	varchar(50)	Refleja la descripción de los motivos de viaje.

- **Tabla País de Embarque:**

ESQUEMA: Dimensiones.

PROPIETARIO: postgres.

DESCRIPCIÓN: Esta tabla corresponde a la dimensión País de Embarque.

PK	Nombre	Tipo de Dato	Descripción
x	id_pais_de_embarque	serial	Llave primaria
	continente	varchar(50)	Refleja los continentes
	region	varchar(50)	Refleja las regiones geográficas
	pais_de_embarque	integer	Refleja los países de embarque.
	descripcion_pais_de_embarque.	varchar(50)	Descripción de los países de embarque.

- **Tabla Puntos Aeroportuarios:**

ESQUEMA: Dimensiones.

PROPIETARIO: postgres.

DESCRIPCIÓN: Esta tabla corresponde a la dimensión Puntos Aeroportuarios.

PK	Nombre	Tipo de Dato	Descripción
x	id_puntos_aeroportuarios	serial	Llave primaria
	puntos_aeroportuarios	integer	Refleja los puntos de entrada y de salida de los visitantes y turistas.

	descripcion_puntos_aeroportuarios.	varchar(50)	Nombres de los puntos aeroportuarios

- **Tabla Sexo:**

ESQUEMA: Dimensiones.

PROPIETARIO: postgres.

DESCRIPCIÓN: Esta tabla corresponde a la dimensión Sexo.

PK	Nombre	Tipo de Dato	Descripción
x	id_sexo	serial	Llave primaria
	sexo	varchar(50)	Refleja los diferentes sexos.
	descripcion_sexo.	varchar(50)	Significado de cada sexo.

- **Tabla Arribo de Visitantes:**

ESQUEMA: Hechos.

PROPIETARIO: postgres.

DESCRIPCIÓN: Esta tabla corresponde al hecho Arribo de Visitantes.

PK	FK	Nombre	Tipo de Dato	Descripción
x	x	id_ciudadania	serial	Llave foránea
x	x	id_sexo	serial	Llave foránea
x	x	id_motivos_de_viaje	serial	Llave foránea
x	x	id_pais_de_embarque	serial	Llave foránea
x	x	id_tiempo	serial	Llave foránea
x	x	id_puntos_aeroportuarios	serial	Llave foránea
x	x	id_edad	serial	Llave foránea
		cant_arribos_visitantes	integer	Indicador para los cálculos de la cantidad de arribos de

				visitantes.
x	x	cant_dias_de_estancia_visitantes	integer	Indicador para los cálculos de la cantidad de días de estancia de los visitantes.

➢ **Tabla Arribo de Turistas:**

ESQUEMA: Vista_materializada.
PROPIETARIO: postgres.
DESCRIPCIÓN: Esta tabla corresponde a la vista materializada Arribo de Turistas.

PK	FK	Nombre	Tipo de Dato	Descripción
x	x	id_ciudadania	serial	Llave foránea
x	x	id_sexo	serial	Llave foránea
x	x	id_motivos_de_viaje	serial	Llave foránea
x	x	id_pais_de_embarque	serial	Llave foránea
x	x	id_tiempo	serial	Llave foránea
x	x	id_puntos_aeroportuarios	serial	Llave foránea
x	x	id_edad	serial	Llave foránea
		cant_arribos_turistas	integer	Indicador para los cálculos de la cantidad de arribos de turistas.
x	x	cant_dias_de_estancia_turistas	integer	Indicador para los cálculos de la cantidad de días de estancia de los turistas.

3.2.2. Restricciones y secuencias

Restricciones:

Cuando se diseña una base de datos se debe reflejar fielmente el universo del discurso que se está tratando, o mejor, las restricciones existentes en el mundo real. Los componentes de una restricción son los siguientes (Casares, 1999-2010):

- La operación de actualización (inserción, borrado o eliminación) cuya ejecución ha de dar lugar a la comprobación del cumplimiento de la restricción.
- La condición que debe cumplirse, la cual es en general una proposición lógica, definida sobre uno o varios elementos del esquema, que puede tomar uno de los valores de verdad (cierto o falso).
- La acción que debe llevarse a cabo dependiendo del resultado de la condición.

Se puede decir que existen varios tipos de integridad. En la solución se aplican:

- Integridad de dominio: restringimos los valores que puede tomar un atributo respecto a su dominio, por ejemplo: edad >= 60.
- Integridad de entidad: la clave primaria de una entidad no puede tener valores nulos y siempre deberá ser única, por ejemplo: las llaves primarias de cada dimensión.
- Integridad referencial: las claves ajenas de una tabla hija se tienen que corresponder con la clave primaria de la tabla padre con la que se relaciona, por ejemplo: las llaves foráneas de las tablas de hechos.

Las restricciones pueden ser representadas por las llaves foráneas, pues cada una de ellas tiene relación con una dimensión específica, ejemplo:

Llaves Foráneas	**Únicas**	**Tabla**
id_edad	x	dim_edad
id_tiempo	x	dim_tiempo
id_ciudadania	x	dim_ciudadania
id_motivos_de_viaje	x	dim_motivo_de_viaje
id_pais_de_embarque	x	dim_pais_de_embarque
id_puntos_aeroportuarios	x	dim_puntos_aeroportuarios
id_sexo	x	dim_sexo
id_cuidadania, id_sexo, id_motivos_de_viaje, id_pais_de_embarque, id_tiempo, id_puntos_aeroportuarios, id_edad	x	hec_arribo_de_visitantes

id_cuidadania, id_sexo, id_motivos_de_viaje, id_pais_de_embarque, id_tiempo, id_puntos_aeroportuarios, id_edad	x	vm_arribo_de_turistas

Secuencias:

Son los atributos que se van a ir incrementando secuencialmente a medida que pasa el tiempo mientras se extienda el proceso, en este caso, las llaves primarias:

Llaves primarias
pk_edad
pk_tiempo
pk_ciudadania
pk_motivos_de_viaje
pk_pais_de_embarque
pk_puntos_aeroportuarios
pk_sexo

3.2.3. Índices

Un índice es una estructura de disco asociada con una tabla o una vista que acelera la recuperación de las filas de estas. Contiene claves generadas a partir de una o varias columnas de la tabla o la vista, las cuales están almacenadas en una estructura (árbol binario) que permite que gestor busque de forma rápida y eficiente la fila o filas asociadas a los valores de cada clave (MSDN Microsoft Corporation, julio de 2009-2010).

Una tabla o una vista pueden contener los siguientes tipos de índices:

- Agrupado: Ordena y almacena las filas de datos de la tabla o vista por orden, en función de la clave del índice agrupado. Éste se implementa como una estructura de árbol binario que admite la recuperación rápida de las filas, a partir de los valores de las claves de su índice.
- No agrupado: Se pueden definir en una tabla o vista con un índice agrupado o en un montón. Cada fila de éste índice contiene un valor de clave no agrupada y un localizador de fila. Dicho localizador apunta a la fila de datos del índice agrupado o al montón que contiene el valor de

clave. Las filas del índice se almacenan en el mismo orden que los valores de la clave del índice, pero no se garantiza que las filas de datos estén en un determinado orden a menos que se cree un índice agrupado en la tabla.

- Único: Garantiza que la clave de índice no contenga valores duplicados y, por tanto, cada fila de la tabla o vista es en cierta forma única. Tanto los índices agrupados como los no agrupados pueden ser únicos.
- Vistas indizadas: Un índice en una vista materializa (ejecuta) la vista, y el conjunto de resultados se almacena de forma permanente en un índice agrupado único, del mismo modo que se almacena una tabla con un índice agrupado. Los índices no agrupados de la vista se pueden agregar una vez creado el índice agrupado.

Tanto los índices agrupados como los no agrupados pueden ser únicos. Esto significa que dos filas no pueden tener el mismo valor para la clave de índice. De lo contrario, el índice no es único y varias filas pueden compartir el mismo valor de clave. Los índices se mantienen automáticamente para una tabla o vista cuando se modifican los datos de la tabla.

Los índices se crean automáticamente cuando las restricciones PRIMARY KEY y UNIQUE se definen en las columnas de tabla. Por ejemplo, cuando cree una tabla e identifique una determinada columna como la clave primaria, el motor de base de datos, crea automáticamente una restricción PRIMARY KEY y un índice en esa columna.

Índices de la solución:

ÍNDICE	Tabla	ESQUEMA	TIPO	CAMPO	PK	ÚNICO
PK9	dim_edad	dimensiones	btree	id_edad	✓	✓
PK6	dim_tiempo	dimensiones	btree	id_tiempo	✓	✓
PK2	dim_ciudadania	dimensiones	btree	id_ciudadania	✓	✓
PK3	dim_motivo_de_viaje	dimensiones	btree	id_motivos_de_viaje	✓	✓
PK4	dim_pais_de_embarque	dimensiones	btree	id_pais_de_embarque	✓	✓

PK7	dim_puntos_aeroportuarios	dimensiones	btree	id_puntos_aeroportuarios	✓	✓
PK1	dim_sexo	dimensiones	btree	id_sexo	✓	✓
PK8	hec_arribo_de_visitantes	hechos	btree	id_cuidadania, id_sexo, id_motivos_de_viaje, id_pais_de_embarque, id_tiempo, id_puntos_aeroportuarios, id_edad	✓	✓
PK10	vm_arribo_de_turistas	hechos	btree	id_cuidadania, id_sexo, id_motivos_de_viaje, id_pais_de_embarque, id_tiempo, id_puntos_aeroportuarios, id_edad	✓	✓

3.2.4. Describir artefacto DDL

El principal artefacto de la construcción de un esquema de datos o modelo del servidor (CADM) lo constituyen los scripts de creación de los objetos de la base de datos. Estos scripts se llaman DDL (lenguaje de definición de datos) y se construyen a partir de la transformación del esquema de datos, teniendo en cuenta las siguientes consideraciones (Juan Bernardo Quintero, julio 2008):

- Cada tabla se convierte en una sentencia “CREATE TABLE…” que debe tener una definición de columna, con su respectivo tipo y restricciones, por cada una de las columnas del esquema de datos.
- Cada clave primaria se convierte en una cláusula “… PRIMARY KEY…” en la creación o alteración de tabla, y los valores de esta columna no podrán repetirse. En caso de que sea compuesta, el orden de las columnas suele definir el orden de la columna en el índice asociado.
- Cada clave foránea se convierte en una cláusula “… REFERENCES…” en la creación o alteración de tabla de la clave foránea, hacia la tabla de la clave primaria que se referencia. Es

importante recalcar que sólo se pueden referenciar columnas o combinaciones de ellas que sean únicas (PRIMARY KEY o UNIQUE).

- Cada objeto creado en la base de datos debe considerar el cálculo de espacios de almacenamiento, para evitar problemas de rendimiento en el acceso a datos.

El artefacto DDL de la solución, se incluye en el expediente de proyecto.

3.3. Usuarios y privilegios

3.3.1. Usuarios

Los usuarios de la base de datos serán:

- Analista: Es quien analiza y visualiza los datos de la base de datos.
- Administrador: Es el que administra toda la base de datos, se encarga de su rendimiento y buen funcionamiento.
- Arquitecto de datos: Es quien analiza, visualiza y modifica los datos de la base de datos.

3.3.2. Privilegios

Los privilegios serán otorgados a los usuarios del sistema, según el rol que representen. En este caso:

- El analista tendrá el privilegio de actuar con los esquemas y las tablas, poseerá permisos de lectura, en este caso, la acción llamada SELECT en la base de datos.
- El administrador tendrá el privilegio de controlar la base de datos, su configuración y administración, disfrutará de todos los permisos permitidos en la base de datos.
- El arquitecto de datos tendrá el privilegio de visualizar y modificar los datos de la base de datos, por lo que poseerá permisos de lectura y escritura, las acciones nombradas SELECT, INSERT, UPDATE, DELETE en la base de datos.

El artefacto que genera dichas acciones es el DCL (lenguaje de control de datos), que se encarga de reflejar todo el proceso de acceso a la base de datos, el mismo se muestra en el expediente de proyecto.

3.3.3. Describir Artefacto DCL

Al artefacto DCL, se le conoce como lenguaje de control de datos, y se clasifica dentro de las herramientas de administración de SGBD. Está directamente relacionado con la seguridad del sistema

y constituye un elemento de apoyo para las labores de administración de la base de datos, mediante un subconjunto de instrucciones SQL (Juan Bernardo Quintero, julio 2008).

El artefacto DCL de la solución se incluye en el expediente de proyecto.

3.4. Carga de nomencladores

La fase de carga, es el momento cuando los datos, provenientes de la fuente ya sometidos al proceso de verificación, son incluidos en el sistema de destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes; en algunas bases de datos, se sobrescribe la información antigua con nuevos datos. Los almacenes de datos, por ejemplo, mantienen un historial de los registros de manera que se pueda hacer una auditoría y disponer de un rastro del comportamiento de un determinado valor a lo largo del tiempo.

La fase de carga interactúa directamente con la base de datos de destino. Al realizar esta operación se aplicarán todas las restricciones que se hayan definido en ella, por ejemplo, valores únicos, integridad referencial, campos obligatorios, rangos de valores (Kimball, R, 2002).

3.4.1. Nomencladores

En la solución actual se utilizan dos tipos de nomencladores:

- Codificación por tipo de Ciudadanía (ver **Anexo 1)**: Estos nomencladores dan el lugar de procedencia del visitante, es el mismo que se utiliza para declarar el País de Embarque.
- Codificación por Motivos de Viajes (Ver **Anexo 2**): Permite conocer los Motivos del Viaje de cualquier visitante que entre al país.

A los visitantes y viajeros, se les clasifica de igual manera.

En la clasificación de los motivos de viaje, turistas se consideran todos, menos:

26 Turismo Crucero y **73 Tripulantes de Buques**; estos dos códigos forman los excursionistas o visitantes del día, el resto son turistas y todos juntos son visitantes.

3.4.2. Describir artefacto DML

Al artefacto DML, se le conoce como lenguaje de manipulación de datos y forma parte de las herramientas de acceso a los SGBD, las cuales sirven para escribir e interpretar también, instrucciones

SQL o DDL, que se envían al motor de bases de datos para ser ejecutadas (Juan Bernardo Quintero, julio 2008).

El artefacto DML de la solución se incluye en el expediente de proyecto.

3.5. Guía de Implantación

3.5.1. Requerimientos

Los requisitos mínimos que debe tener el servidor, en cuanto a:

- Software: Sistema operativo Debian 5.0, SGBD PostgreSQL 8.4.
- Hardware: 1 gb de memoria RAM, 80 gb o más de capacidad de disco duro, procesador a 2.0 ghz o más de velocidad.

3.5.2. Secuencia de pasos

Para la instalación de la base de datos, primeramente se debe verificar un grupo de pasos, los cuales se enuncian a continuación:

- Estar instalado el gestor de bases de datos: PostgreSQL 8.4.
- Estar instalada la herramienta de administración, manejadora de bases de datos: PgAdmin III 1.10.
- Crear una base de datos nueva utilizando la herramienta de administración de bases de datos.
- Crear los esquemas de la base de datos.
- Cargar y ejecutar el script DDL_InmEx.sql, generado por la herramienta de modelado en el gestor de base de datos.
- Crear los roles que se utilizarán en la base de datos: Administrador, Analista y Arquitecto de datos.
- Cargar y ejecutar el script DCL_ InmEx.sql.
- Cargar y ejecutar el script DML_ InmEx.sql.

3.6. Validación y pruebas

3.6.1. Listas de chequeo de análisis

Para el análisis de la solución, es recomendable seguir listas de chequeo, adjuntas en el expediente de proyecto de la solución, las cuales son específicas para este sistema. Las listas son informativas y

pretenden cubrir todos los aspectos aplicables en los temas de análisis, particularmente para el Departamento de Turismo y Comercio de la ONE.

En la solución actual, las listas de chequeo que se utilizan para el análisis, son: Lista de Chequeo Evaluación de Áreas de la Organización (Ver **Anexo 3**), Lista de Chequeo de la Herramienta para la recolección y análisis de la información (Ver **Anexo 4**), así como la Lista de Chequeo Especificación de Requisitos (Ver **Anexo 5**). Del resultado de la aplicación de las listas de chequeo, se obtuvieron 66 puntos satisfactorios y dos no conformidades, clasificadas de no críticas según el proceso de chequeo establecido. Por lo antes mencionado, se establece que los resultados obtenidos son satisfactorios.

3.6.2. Validación de requisitos por el cliente

En la validación del mercado de datos, estuvieron presentes los clientes:

- Lic. Elena Leonila Fernández García. Representante de la ONE en la UCI.
- Lic. Mirtha Alarcón. Representante del Departamento de Turismo y Comercio de la ONE.

Los cuales estuvieron de acuerdo con el resultado obtenido en la implementación de la solución que se documenta.

3.6.3. Lista de chequeo de diseño

Para el diseño de la solución, es recomendable seguir una lista de chequeo, adjunta en el expediente de proyecto de la solución, la cual es específica para este sistema. La lista es informativa y pretende cubrir todos los aspectos aplicables en los temas de diseño e implementación, particularmente para el Departamento de Turismo y Comercio de la ONE.

En la solución actual, la lista de chequeo que se utilizó para el diseño, fue la Lista de Chequeo del Modelo de Datos (Ver **Anexo 6**). En ella se establecen 12 puntos, resultando todos satisfactorios, por tanto, se establece que los resultados obtenidos son favorables.

3.6.4. Pruebas de implantación

En la solución se establece un modelo de casos de pruebas para verificar la calidad del proceso de implementación, en el cual, se recogen precondiciones y poscondiciones que facilitan la identificación de resultados verídicos y eficientes.

Modelo de Casos de Prueba:

- **Prueba de Roles y Permisos:**

Precondición	La acción	Poscondición	Resultados
- Debe de estar instalado el gestor de base datos.	- Establecer los roles y permisos que van a acceder a las tablas de la base datos	-Que todos los objetos de la base datos tengan los permisos establecidos	Los resultados que se obtuvieron fueron satisfactorios.

- **Prueba de Creación de la Base de Datos:**

Precondición	La acción	Poscondición	Resultados
-Deben estar creados los roles y asignado cada permiso. -Debe de crear la base de datos.	-Ejecutar el script de la base de datos	-Que después de haber terminado con los pasos establecidos que se encuentre la base de datos estructurada con todas sus tablas y que cada tabla tenga las políticas de seguridad establecidas.	Los resultados que se obtuvieron fueron satisfactorios.

- **Prueba de carga de los Nomencladores:**

Precondición	La acción	Poscondición	Resultados
-Deben estar creados los roles y asignado cada permiso. -Debe de crear la base de datos. -Debe de estar estructuras con todas sus tablas.	-Carga de los nomencladores.	- Que la base de datos esté creada, estructurada y con todos los datos cargados.	Los resultados que se obtuvieron fueron satisfactorios.

Conclusiones

Después de analizar los resultados obtenidos en la etapa de implementación y validación se puede definir como conclusiones:

- A partir del **modelo de datos físico**, compuesto por 7 dimensiones, 1 tabla de hechos y 1 vista materializada, se crean **3 esquemas** y **9 tablas** en la base de datos.
- Mediante la definición de los grupos de **usuarios de la base de datos**, integrados por: el **analista**, el **administrador** y el **arquitecto de datos**, se establecen **permisos y privilegios**, obteniendo así, una mayor protección para el sistema.
- A través de la **carga de los nomencladores al mercado de datos**, se logra que la información de las dimensiones estén listas para ser procesadas, lo que prueba que el objetivo fue alcanzado.
- Mediante la **guía de implantación**, se logran detallar los pasos para la instalación de la base de datos, así como los requerimientos necesarios para ello.
- Con la aplicación de las **listas de chequeo** establecidas, para validar el análisis y el diseño de la solución, se obtienen como resultado **78 aspectos positivos** y **2 aspectos negativos**, de un total de 80 puntos, por lo que se considera que los resultados son satisfactorios.
- Con la **validación de los requerimientos por el cliente**, se concluye que el desarrollo de la solución es factible y viable en todos los sentidos.

Conclusiones

Al concluir la solución, se puede plantear que fueron cumplidos los objetivos trazados y las tareas de investigación propuestas. Por tanto, se llega a las siguientes conclusiones:

- A partir del estudio realizado, **se logra definir en el marco teórico de la investigación**, que los mercados de datos son la solución más idónea para la situación problémica planteada.
- Mediante el proceso de análisis, diseño e implementación, **se logra la estructura final del Mercado de Datos de Inmigración y Extranjería** para el Departamento de Turismo y Comercio de la ONE.
- A través **del proceso de validación mediante las listas de chequeo** y **la carta de aceptación del cliente**, se concluye que la solución desarrollada posee 2 no conformidades, por lo que se estima que **los resultados que se obtienen son satisfactorios**.
- Con la creación del Mercado de Datos de Inmigración y Extranjería para el Departamento de Turismo y Comercio de la ONE, **se conforma el expediente de proyecto**, el cual posee los documentos necesarios para el análisis y el incremento futuro de la solución, con el objetivo de dejar en archivos todo el proceso de desarrollo del producto final.

Recomendaciones

Con el propósito de enriquecer la propuesta plasmada en el trabajo, se sugiere:

- Implementar el proceso de ETL al mercado de datos desarrollado.
- Realizar un profundo estudio acerca de técnicas de optimización, que puedan ser aplicadas al proceso de ETL a implementar.
- La incorporación, sobre el diseño propuesto, de los modelos estadísticos venideros, según la prioridad de los especialistas de la ONE.
- La creación de la base de datos continúe con el proceso de desarrollo de la capa de visualización del mercado de datos.

Referencias Bibliográficas

Oracle Corporation. (2005-2010). ORACLE. Retrieved 1 12, 2010, from ORACLE:

Corporate Executive Board. (2008). Toolbox for IT. Recuperado el 14 de 1 de 2010, de Toolbox for IT: http://datawarehouse.ittoolbox.com/

Kornspan, Michael - Corporate Communications. (Copyright © 2010 CA, Inc). ERwin. Retrieved 2 11, 2010, from ERwin: http://www.erwin.com

Technologies, B. P. (2010). Visual Paradigm. Retrieved 2 11, 2010, from Visual Paradigm: http://www.visual-paradigm.com

CollabNet, Inc. (2001-2009). Tigris.org. Retrieved 1 14, 2010, from Tigris.org: http://tortoisesvn.net

ArPug. (2009). ArPug - Grupo de usuarios PostgreSQL de Argentina. Retrieved 2 11, 2010, from ArPug - Grupo de usuarios PostgreSQL de Argentina: http://www.arpug.com.ar

Pentaho Corporation. (Copyright © 2005 – 2010). pentaho. Retrieved 1 12, 2010, from pentaho: www.pentaho.com

ONE. (© Copyright 2006). O.N.E. Oficina Nacional de Estadísticas de Cuba. Retrieved 1 12, 2010, from O.N.E. Oficina Nacional de Estadísticas de Cuba: http://www.one.cu.

Universitat Jaume * I (UJI). (2001, 2 12). Universitat Jaume*I. (M. M. Andrés, Producer) Retrieved 04 5, 2010, from Universitat Jaume*I: http://www3.uji.es/~mmarques/f47/apun/node32.html

Juan Bernardo Quintero, D. M. (julio 2008). DIRECTRICES PARA LA CONSTRUCCIÓN DE ARTEFACTOS DE PERSISTENCIA EN EL PROCESO DE DESARROLLO DE SOFTWARE. Revista EIA, ISSN 1794-1237 Número 9, p. 77-90. , 88.

Cásares, C. (1999-2010). programatium.com. Retrieved 4 5, 2010, from programatium.com: http://www.programatium.com/manuales/sql/moddat007.htm

MSDN © 2010 Microsoft Corporation. (julio de 2009-2010). msdn. Retrieved 4 5, 2010, from msdn: http://msdn.microsoft.com/es-es/library/ms175049.aspx

Kimball, R. (2002). The Data Warehouse ETL Toolkit. Practical Techniques for extracting, cleaning, conforming, and delivering data. WILEY PUBLICHING, INC.

Kimball, R. (1996). The Data Warehouse Toolkit. s.l. : WILEY PUBLICHING, INC. , 1996.

García., E. L. (2010). Presentación ONE. La Habana.

Bibliografía

Microsoft Corporation. (2009, julio). msdn. Retrieved 2 11, 2010, from msdn: http://msdn.microsoft.com/es-es/library/bb522541.aspx

ScriBD.(2010). ScripBD. Retrieved 2 11, 2010, from ScripBD: http://www.scribd.com/

Sánchez, L. Z. (2008). Metodología para el Diseño Conceptual de Almacenes de Datos. Valencia, España: UNIVERSIDAD POLITÉCNICA DE VALENCIA.

CollabNet, Inc. (2001-2009). Tigris.org. Retrieved 1 14, 2010, from Tigris.org: http://subversion.tigris.org.

Castillo, C. (2008). Sistemas de Información II. Sistemas gestores de bases de datos. UPF.

Stefan Küng, L. O. (2006/10/13). TortoiseSVN.

Corporate Executive Board. (2008). Toolbox for IT. (http://mayita.chacharaselnido.com/rene/Respaldo%20USB%20Saira/datawarehouse.doc) Recuperado el 14 de 1 de 2010, de Toolbox for IT: http://datawarehouse.ittoolbox.com/

Durán, D. G. (2007). Introducción a los Datawarehouses. Revista de Ciencias Básicas UJAT , 37-41.

MKM. (2007-2010). mkm. Recuperado el 14 de 1 de 2010, de mkm: http://www.mkm-pi.com/mkmpi.php?article1881

Ponniah, P. (2001). Data Warehousing Fundamentals. E.E.U.U.: Wiley Publishing Inc.

Villa, M. C. (2009). Metodología de Proceso de Desarrollo Línea Soluciones de almacenes de datos e inteligencia de negocio. Ciudad Habana: UCI.

Evelia, C. C. (2009). Data Warehouse (Almacenes de Datos).

Business Intelligence Galicia. © Copyright 2007. Sinnexus. Sinnexus. [En línea] © Copyright 2007. [Citado el: 14 de 1 de 2010.] http://www.sinnexus.com/business_intelligence/datawarehouse.aspx..

Mafla, I. E. (21/03/2005). Especificaciones Generales Requeridas Para La Implementación Del Sistema De Seguridad y Alta Disponibilidad. Ecuador: Aduana del Ecuador.

(OMT), O. M. (2008). Recomendaciones Internacionales para las Estadísticas del Turismo 2008 (RIET2008).

Glosario de Términos

ONE: Oficina Nacional de Estadísticas.

DIE: Dirección de Inmigración y Extranjería.

MININT: Ministerio del Interior.

Centros Informantes: Los Centros Informantes son las empresas u organismos que suministran información a las oficinas de estadísticas en sus diferentes niveles.

Indicador: Se dice de la variable que puede tomar un valor de una determinada unidad de medida y de un determinado tipo de datos (generalmente numérico). Los indicadores de la ONE están bien definidos y tienen un código único que los identifica.

Clasificador: Es un instrumento que asigna un código a elementos ya definidos por otras vías.

El término ***viaje*** designa la actividad de los ***viajeros***. Un ***viajero*** es toda persona que se desplaza entre dos lugares geográficos distintos por cualquier motivo y duración. Designa todo desplazamiento de una persona a un lugar fuera de su lugar de residencia habitual, desde el momento de su salida hasta su regreso. Por lo tanto, se refiere a un viaje de ida y vuelta ((OMT), 2008).

Un ***visitante*** es una persona que viaja a un destino principal distinto al de su entorno habitual, por una duración inferior a un año, con cualquier finalidad principal (ocio, negocios u otro motivo personal) que no sea la de ejercer una actividad remunerada en el país o lugar visitados. Estos viajes realizados por los visitantes se consideran viajes turísticos. El turismo hace referencia a la actividad de los visitantes ((OMT), 2008).

Por lo tanto, el ***turismo*** es un subconjunto de los viajes, y los ***visitantes*** un subconjunto de los viajeros. Estas distinciones son fundamentales para la recopilación de datos sobre los movimientos de viajeros y visitantes, y para la credibilidad de las estadísticas de turismo ((OMT), 2008).

Un visitante (doméstico, receptor o emisor) se clasifica como ***turista*** (***o visitante que pernocta***), si su viaje incluye una pernoctación, o como ***visitante del día*** (***o excursionista***), en caso contrario ((OMT), 2008).

Ralph Kimball: Conocido innovador, escritor, educador y consultor en el campo de Almacenes de Datos. En la actualidad posee más de 100 artículos sobre inteligencia empresarial. Es Vicepresidente

de Metaphor Cumputer Systems, pionera en software para ayuda a la toma de decisiones y proveedora de servicios de esta índole. La asociación Ralph Kimball fue creada en 1992 para proveer consultoría y educación sobre la tecnología de almacenamiento de datos.

William H. Inmon: Se le conoce como “El padre de la tecnología de Almacenes de Datos”, creador de la metodología CIF (Corporate Information Factory) y más recientemente GIF (Government Information Factory). Tiene más de 35 años de experiencia en tecnología de administración de base de datos y diseño de almacenes de datos. Ha escrito más de 650 artículos sobre construcción, uso y mantenimiento de almacenes de datos. Es el autor de más de 46 libros de temas relacionados a tecnologías de base de datos.

BD: Base de Datos.

BDMD: Base de Datos Multidimensional.

SGBD: Sistema Gestor de Bases de Datos.

ETL: Proceso de Extracción, Limpieza, Transformación y Carga de los datos.

DATEC: Centro de Tecnologías y Análisis de Datos.