Universidad de las Ciencias Informáticas

Facultad 6

Título: *Sistema de información de gobierno Mercado de datos Inmigración y extranjería.*

Trabajo de Diploma para optar por el título de

Ingeniero en Ciencias Informáticas

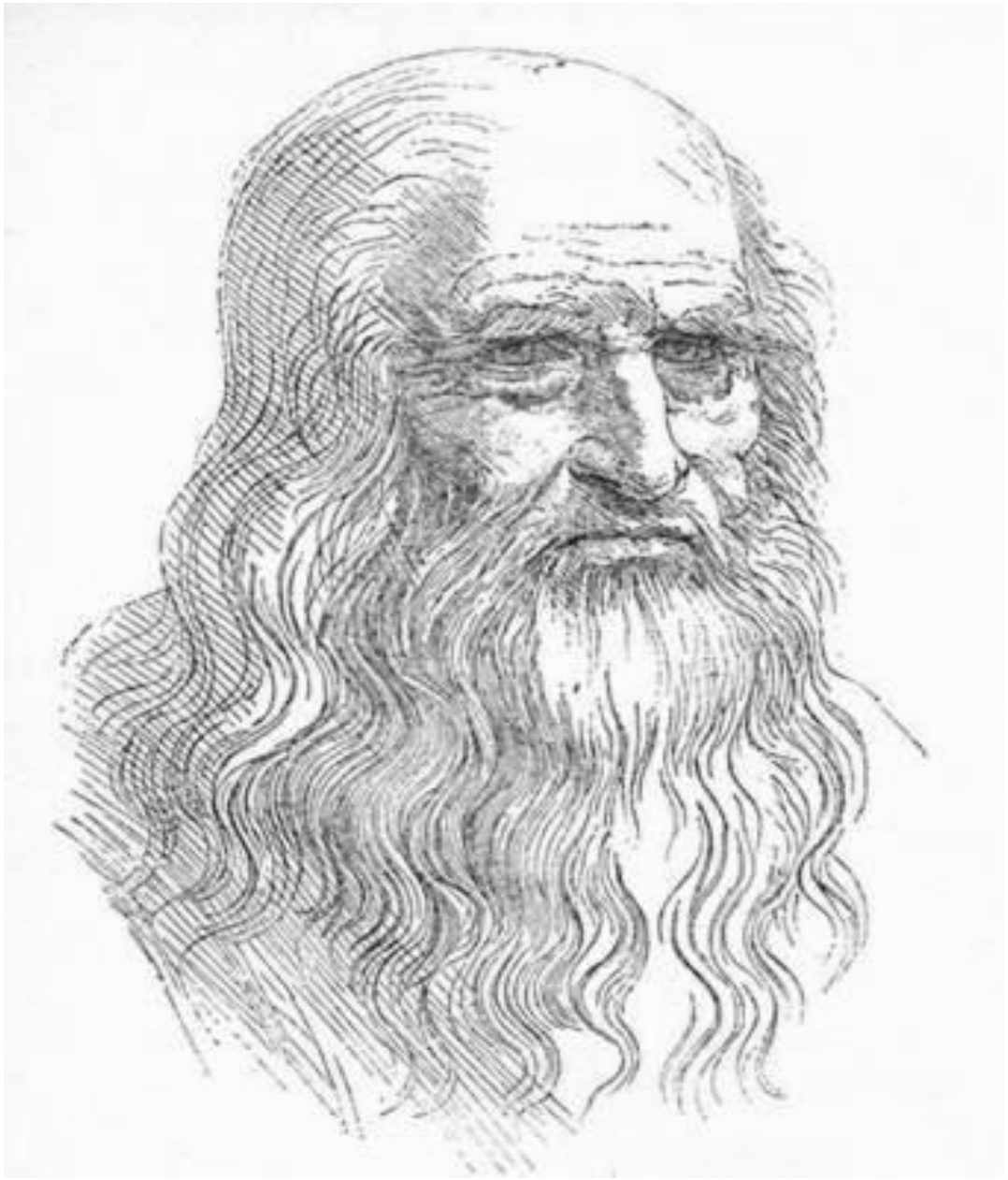
Autora:

Yaneisy Pedraza González

Tutor:

Ing. Roberto Tellez Ibarra.

Ciudad de la Habana junio 2011



La adquisición de cualquier conocimiento es siempre útil al intelecto, que sabrá descartar lo malo y conservar lo bueno.

Leonardo Da Vinci.

DECLARACIÓN DE AUTORÍA

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

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

Autora: _____

Yaneisy Pedraza González

Tutor: _____

Ing. Roberto Tellez Ibarra

Datos de Contacto

Tutor:

Tutor: Ing. Roberto Tellez Ibarra

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: Instructor Recién graduado

Categoría Científica: no

Años de experiencia en el tema: 1

Años de graduado: 1

Correo Electrónico: rtibarra@uci.cu

Autora: Yaneisy Pedraza González.

Correo Electrónico: ypedraza@estudiantes.uci.cu

AGRADECIMIENTOS

Agradezco:

A la revolución por hacer mi sueño realidad: graduarme.

A mi mamá por su esfuerzo, amor y dedicación durante todos los años de la carrera.

A mi papá por apoyarme siempre.

A mis abuelas Olga, Silvia y Aida por su infinito amor.

A mi esposo, que nunca me ha dejado sola y siempre estuvo a mi lado en todo momento.

A mi tutor por ayudarme.

A mis amigas Ailyn, Aliannis, Ana N., Dainelis, Ana G., Themis, Salvador, Marisel, el ruso, por apoyarme siempre.

A mi amiga Nela que tanto me ayudó cuidando a mi bebé para yo poder estudiar e ir a las reuniones.

A mi amigo Yosvani por todo el tiempo que me dedicó.

A mi tía Carmen por cuidarme a mi hijito para yo poder entrar a la escuela.

A Rosado y Katia, por ayudarme en los momentos en que los necesité.

A Yamila por haber estudiado conmigo cada vez que le preguntaba algo.

A todos aquellos que de una forma u otra me ayudaron con el desarrollo de este trabajo.

DEDICATORIA

Este trabajo se lo dedico a las personas que me dieron la vida, a mi mamá y a mi papá.

A mi hijito querido, porque este esfuerzo es por él, por su futuro.

A mi abuela Olga, que aunque no me dio la vida, me crió.

Y a mi esposo que es parte de mi familia.

RESUMEN

El presente trabajo de diploma se enmarca en el tema de los almacenes de datos, los mercados de datos, y su utilización para los análisis estadísticos de la información. En la investigación se detalla la metodología, topología, herramientas, especificación de requerimientos, junto con las necesidades del cliente, para lograr un buen diseño e implementación de los procesos de integración y análisis de datos de la solución. Como resultado se obtiene un mercado de datos poblado disponible para hacer análisis OLAP, además se tiene la estructura del modelo dimensional que comprende: las dimensiones, las jerarquías, las tablas de hechos, 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 al mercado de datos Inmigración y extranjería. De igual manera, la solución incluye las políticas de seguridad, respaldo y recuperación de los datos, así como las pruebas para la validación del mercado.

Palabras claves: Almacenes de datos, mercados de datos.

TABLA DE CONTENIDO

TABLA DE CONTENIDO

TABLA DE CONTENIDO

ÍNDICE DE FIGURAS

ÍNDICE DE TABLAS

INTRODUCCIÓN

El aumento, desarrollo y explotación de las Tecnologías de la Información y las Comunicaciones (TIC) en la sociedad aparejado a la creciente evolución y desarrollo de las ciencias de la información, le imponen al mundo una nueva forma de conceptualizar las soluciones a los problemas que hoy se presentan, debido que para controlar los procesos se generan grandes cantidades de datos.

El control de los datos estadísticos dentro de la infraestructura de un país constituye uno de los principales eslabones para la toma de decisiones en los diferentes sectores socioeconómicos. La entidad rectora del Sistema de Información de Gobierno en Cuba, es la Oficina Nacional de Estadísticas e Información (ONEI) la cual mediante su Sistema estadístico nacional (SEN), organiza, dirige, controla y regula la información de todos los sectores del país.

La ONEI cuenta con diversas direcciones y departamentos que organizan los datos por diferentes criterios, dentro de las cuales se encuentra la Dirección de Comercio, Turismo y Servicios, específicamente este trabajo está relacionado con el Departamento de Turismo y Comercio de la ONEI, que se encarga de controlar los datos estadísticos de las personas que entran a la isla por cualquier motivo, ya sean visitantes, turistas o excursionistas.

Dentro de este departamento se evidencian algunos problemas como son: la información se almacena en varios formatos como excel, archivos de texto, archivos DBF, formato duro (papeles), documentos Word, entre otros. La información solo puede ser consultada por un especialista de la informática y de la información con alto conocimiento del negocio, porque algunas de las fuentes están codificadas de forma que si no se es conocedor de la terminología usada, no puede comprenderse el contenido de estas, y por otra parte, los datos de las fuentes pueden extraerse de forma manual o por un informático. Otro de los problemas, es que se generan ficheros anuales con los cuales se hace muy difícil el proceso de obtención de información porque esta, va acumulándose año tras año y es cada vez más complicado analizarla. Además los datos no están integrados, porque hay referencias a la misma información que usan diferente codificación, o diferente cantidad de caracteres y atenta contra la calidad de estos porque existen múltiples versiones de los mismos datos. Se carece de una aplicación informática que brinde reportes flexibles con información actualizada para apoyar el proceso de toma de decisiones. Actualmente hay aplicaciones que procesan información, pero algunas son muy antiguas y para usarlas hay que simular el entorno MS-DOS, estas son las que pueden exportar

los datos al formato DBF; sin embargo, no hay una aplicación que permita generar reportes "ad hoc" con los datos actuales, que permita analizarlos.

De las situaciones anteriormente mencionadas surge el problema de la investigación: ¿cómo contribuir a la toma de decisiones en el área de Inmigración y extranjería del Sistema de Información de Gobierno? Definiendo como **objeto de estudio**: Los almacenes de datos, enmarcado en el **campo** de **acción**: Mercado de datos para el área de Inmigración y extranjería del Sistema de Información de Gobierno. Para dar solución a la problemática que dió surgimiento al presente trabajo se ha propuesto como **objetivo general**: Desarrollar el mercado de datos de Inmigración y extranjería del Sistema de Información de Gobierno, que contribuya a la toma de decisiones. Una vez realizado el análisis general, los **objetivos específicos** quedan desglosados en:

- Refinar el análisis y diseño del mercado de datos del área Inmigración y extranjería.
- Implementar el mercado de datos del área Inmigración y extranjería.
- Validar el mercado de datos del área Inmigración y extranjería.

Para alcanzar los objetivos propuestos se plantean las siguientes **tareas**:

- Caracterización de las metodologías y herramientas a utilizar en el desarrollo de almacenes de datos.
- Refinamiento del levantamiento de requisitos.
- Refinamiento de la descripción de los casos de uso del mercado de datos.
- Refinamiento de la definición de los hechos, las medidas y las dimensiones del mercado de datos.
- Refinamiento del diseño del modelo de datos.
- Definición de la arquitectura del mercado de datos.
- Realización del diseño del subsistema de integración.
- Realización del diseño del subsistema de visualización.
- Realización del diseño de los casos de pruebas.

- Implementación de la base de datos.
- Implementación del subsistema de integración.
- Implementación del subsistema de visualización.
- Aplicación de las listas de chequeo en el subsistema de visualización.
- Aplicación de los casos de pruebas.

Para cumplir con todos los elementos planteados, el presente trabajo está estructurado en cuatro capítulos:

Capítulo 1: Fundamentación teórica de los almacenes de datos: En este capítulo se realiza un estudio del negocio, de las principales definiciones, metodologías y herramientas que se utilizan para desarrollar un mercado de datos para el área Inmigración y extranjería.

Capítulo 2: Análisis y diseño del mercado de datos del área Inmigración y extranjería: Se hace referencia al flujo de trabajo de análisis y diseño, en el cual se definen los requerimientos del sistema, se diseña el diagrama de casos de uso del sistema; así como los subsistemas de integración y visualización.

Capítulo 3: Implementación del mercado de datos del área Inmigración y extranjería: Se implementan los subsistemas de implementación y visualización, con el objetivo de darle solución a los requisitos del sistema.

Capitulo 4: Validación del mercado de datos del área Inmigración y extranjería: Es donde se valida el mercado de datos a través de las listas de chequeo y los casos de prueba.

Capítulo 1. Fundamentación Teórica. Almacenes de Datos.

Introducción.

Este capítulo aborda el estudio sobre los almacenes de datos (AD) y los mercados de datos (MD) así como sus características, metas y los componentes que los integran. Se plasma el resultado de las metodologías existentes y las principales herramientas para el desarrollo de los almacenes de datos.

1.1Almacenes de datos.

Con la aparición de la computación se presenta el primer problema para el hombre, que consistía en cómo poder almacenar la información que se generaba día tras día. Con el transcurso de los años y el avance de las tecnologías la información que se almacenaba, comenzó a aumentar a tal punto que en ocasiones resultaba demasiado complejo analizarla. Estos tipos de problemas trajeron como consecuencia que se desarrollaran nuevas herramientas para el almacenamiento de la información como por ejemplo, los AD.

En cualquier revisión que se realice sobre lo que se entiende por AD, es difícil encontrar una concepción acabada y compartida por los autores, por el contrario, existen diversas aproximaciones teóricas; lo que demuestra que se trata de una herramienta en evolución y de compleja concepción.

Un AD 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, para luego procesarla permitiendo su análisis desde diferentes perspectivas y con grandes velocidades de respuesta. La creación de un AD 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 (este término se abordará en el epígrafe 1.5) (1).

Los AD son estructuras que se definen en función de temas específicos, donde la información histórica debe estar integrada y robusta ante los cambios que puedan afectar a la organización. Su objetivo principal, es servir de ayuda a la toma de decisiones empresariales (1).

Se puede decir entonces que los AD son una base de datos donde se almacena una gran cantidad de datos integrados, disponibles para ser usados en análisis OLAP por usuarios especializados en el tema.

1.1.1 Principales aportes de un almacén de datos.

Un AD aporta grandes beneficios a una empresa, ayudando significativamente al proceso de toma de decisiones, dentro de sus principales aportes están: proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global del negocio. Facilita la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén; obteniendo un valor añadido para el negocio de dicha información. Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios de cualquier esfera. Además, el mismo simplifica dentro de la empresa, la implantación de sistemas de gestión integral de la relación con el cliente, facilitando así la gestión de la información de forma integral; y supone una optimización tecnológica y económica en entornos de Centro de Información, estadística o de generación de informes. (1)

1.1.2 Metas de los almacenes de datos.

Con la necesidad de que la información cada día sea de mayor utilidad para los directivos de las organizaciones, y con la complejidad de las soluciones empresariales, se obliga a la realización de sistemas más abiertos y dinámicos que permitan su adaptación ante los cambios que se imponen en la actualidad. Ante esta disyuntiva, con la maduración de las técnicas de desarrollo de almacenes de datos y mercados de datos se han definido un conjunto de metas que deben cumplir estos sistemas (2).

- Deben hacer fácilmente accesible la información.

  La información que se almacena debe diseñarse utilizando datos que sean intuitivos y obvios, pensando siempre en las necesidades de los usuarios. El diseño de las estructuras debe soportar cualquier combinación que los usuarios estimen conveniente. La recuperación de la información almacenada debe realizarse en un tiempo mínimo de espera y ser accesible en todo momento.

- La información de la organización debe ser presentada de forma consistente.

La información almacenada debe ser creíble y consistente. Después de ensamblar o agregar los datos de las fuentes que existen alrededor de la organización, se requiere que transiten por un proceso de limpieza que garantice que los mismos tengan una alta calidad. Además, se debe garantizar que siempre se encuentren disponibles para los usuarios.

- Deben ser adaptables y resistentes al cambio.

  Cuando se realiza el diseño de los almacenes de datos debe tenerse en cuenta el inevitable cambio que proponen las condiciones del negocio, los datos y la tecnología. Los cambios deberán realizarse teniendo en cuenta que no se altere el significando de los datos y que no invaliden algunos existentes o aplicaciones empleadas.

- Deben ser un baluarte seguro que apoye los recursos de información.

  La información no siempre puede ser consultada por todos, lo que implica tener un control de acceso efectivo para la información confidencial de la organización. Se deben establecer niveles de seguridad para cada funcionalidad que va a brindar el almacén.

- Debe servir como base para mejorar la toma de decisiones.

  Los datos almacenados deben ser correctos y a la vez ser útiles para dar soporte a la toma de decisiones empresariales. La organización de la información debe ser lo suficientemente dinámica y efectiva para que pueda ser servida oportunamente a los usuarios.

- La comunidad del negocio debe aceptar el AD para poder ser juzgado como exitoso.

  No importa que tan elegante sea la solución usando los mejores productos y plataformas, si la comunidad de la organización no depende del AD y no es explotado activamente después del entrenamiento, entonces ha fallado la prueba de aceptación.

1.1.3 Características de los almacenes de datos.

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

- Orientado al tema: Tiene en cuenta los procesos de negocio de una empresa que se deseen priorizar.

- Integrado: Agrupa a todos los sistemas operacionales en un sistema de información que se generan en el proceso de negocio, proveyéndolos 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 información, describen el contenido, la calidad, la condición y otras características de los datos. (1)

Se puede afirmar que los avances alcanzados en el desarrollo de los AD, en la actualidad, confirman que el mismo, es una tecnología madura, estable y que soluciona las problemáticas presentadas, lo que no impide su constante evolución.

1.1.4 Ventajas y desventajas de usar un almacén de datos.

A pesar de las ventajas que brinda el uso de los AD se pueden observar además algunas desventajas. A continuación se evidencia lo antes planteado:

La implementación de un AD puede beneficiar a una organización porque:

- Los AD hacen más fácil el acceso a una gran variedad de datos a los usuarios finales.
- Se obtiene una base de datos histórica y clasificada por temas.
- Integración de información procedente de múltiples sistemas externos.
- Facilitan la toma de decisiones estratégicas.

Sin embargo, utilizar AD también trae consigo algunos problemas como:

- La subestimación del tiempo requerido para extraer, limpiar y cargar los datos en el almacén.

- Problemas con los sistemas de origen de los datos.
- Los datos obtenidos no son suficientes.
- Los gastos de mantenimiento son muy elevados.

1.1.5 Componentes de los almacenes de datos.

Los AD están compuestos por procesos que definen en su conjunto, el ambiente que estos poseen. Aunque cada desarrollo de AD es diferente debido a la especificidad de las organizaciones, generalmente, cumplen con la realización de los siguientes componentes:

La Ilustración 1 muestra los componentes de un AD y la relación que existe entre ellos según la metodología de Kimball. (1)

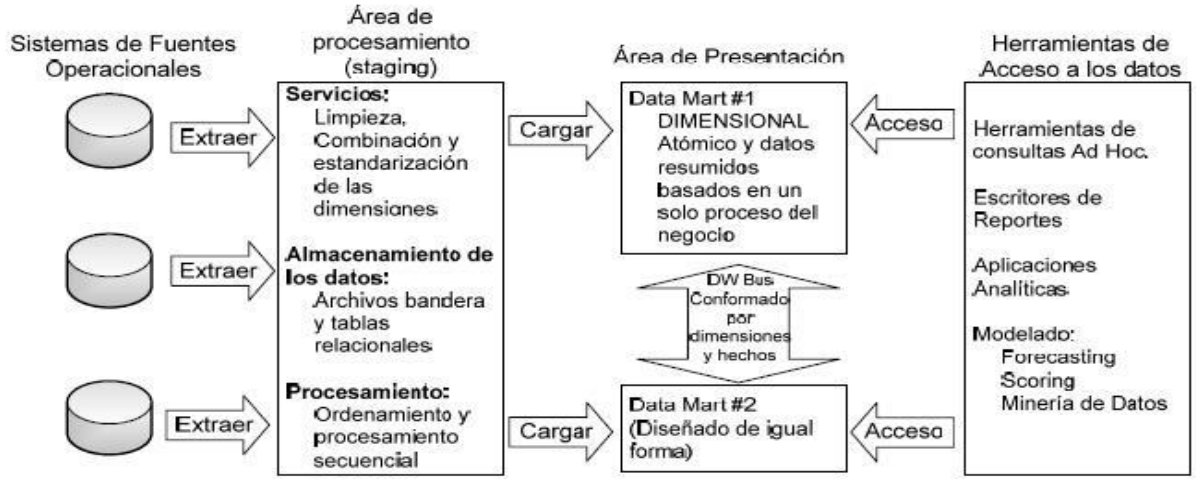


Ilustración 1.Componentes de un AD.

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. 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 (3):

- Los datos de producción, son los datos de interés para el AD: 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 AD.
- 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.

Área de Procesamiento.

El área de procesamiento es el componente donde se invierte la mayor cantidad de tiempo y esfuerzo durante el desarrollo del AD. 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 (3).

Área de Presentación.

En este componente los datos se encuentran organizados, almacenados y disponibles para ser consultados, reportados o analizados por parte de los usuarios finales. Es donde se encuentra la información, diseñada mediante esquemas dimensionales, que ha sido definida por los usuarios como útil para la toma de decisiones. Generalmente esta área es referenciada como una serie de MD integrados donde cada uno se encuentra representando a un proceso específico del negocio.

Herramientas de Acceso a Datos.

En este componente se usa la palabra herramientas para referirse a la variedad de capacidades que pueden ser provistos a los usuarios del negocio para el soporte a la toma de decisiones. Su actividad principal es consultar el área de presentación del AD. El mismo puede abarcar desde una simple o personalizada herramienta de consulta, hasta una compleja y sofisticada aplicación de modelado o minería de datos (3).

1.2 Mercado de Datos. Características.

Los MD: son un subconjunto de datos de un AD donde se almacenan la mayoría de las actividades de análisis que en el entorno de Inteligencia de Negocio se llevará a cabo.

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

A diferencia de la anterior, la propuesta de Kimball se basa en dividir el mundo de Inteligencia de Negocio entre los hechos y las dimensiones, ésta es eficaz y conduce a una solución completa en un corto período de tiempo. Además, tiene abundante documentación y se puede encontrar una respuesta a casi todas las preguntas que se puedan tener. Entre sus características principales, está el hecho de poseer una arquitectura ascendente, plantea que se debe crear por cada departamento un conjunto de MD independientes orientados a los temas que estén relacionados con él. Un AD es la unión de todos los MD existentes en una entidad.

Se puede decir que un MD es una base de datos departamental que se especializa en almacenar datos de un área específica, brindando una estructura óptima para analizar los procesos que tienen lugar dentro del departamento. Son AD orientados a temas específicos y contienen datos de solo una línea del negocio. La mayor diferencia entre ambos, es el alcance de la información que contienen, debido a que en los MD es más pequeño y los datos se obtienen de un menor número de fuentes, por tanto es menos el tiempo de desarrollo.

1.3 Modos de almacenamiento de datos.

La tecnología de Procesamiento Analítico en Línea –OLAP- (*Online Analytical Processing*) permite un uso más eficaz de los AD para el análisis de datos en línea, proporciona respuestas rápidas a consultas analíticas complejas e iterativas utilizada generalmente para sistemas de ayuda para la toma de decisiones, presenta los datos a los usuarios a través de un modelo de datos intuitivo y natural. Permitiendo a los usuarios finales ver y entender la información de sus bases de datos.

OLAP acelera la entrega de información a los usuarios finales que ven estas estructuras de datos como cubos, denominadas multidimensionales debido a que la información es vista en varias dimensiones.

Las características principales del OLAP son:

- **Rápido**: Proporciona la información al usuario a una velocidad constante. La mayoría de las peticiones se deben de responder al usuario en cinco segundos o menos.
- **Análisis**: Realiza análisis estadísticos y numéricos básicos de los datos, predefinidos por el desarrollador de la aplicación o definido “ad hoc” por el usuario.
- **Compartida**: Implementa los requerimientos de seguridad necesarios para compartir datos potencialmente confidenciales a través de una gran población de usuarios.
- **Multidimensional**: Llena la característica esencial del OLAP, que es ver la información en determinadas vistas o dimensiones.
- **Información**: Acceden a todos los datos y a la información necesaria y relevante para la aplicación, donde sea que ésta resida y no esté limitada por el volumen.

Existen tres modelos para el proceso analítico en línea (OLAP) de la información: ROLAP, MOLAP y HOLAP. El proceso de análisis se realiza de igual forma lo que varía en uno y otro caso es la metodología de almacenamiento. La forma de almacenamiento es crítica para garantizar la velocidad de recuperación de la información, las zonas de ubicación de las agregaciones y el procesamiento de los datos en general. (4)

1.3.1 ROLAP: Procesamiento Analítico Relacional en Línea.

En el Procesamiento Analítico Relacional en Línea (Relational Online Analitical Process, en inglés) 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. Con el fin de ocultar las estructuras de almacenamiento y presentar los datos dimensionalmente es creada la semántica de las etiquetas de los metadatos. Ellas soportan 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.

El sistema ROLAP utiliza una arquitectura de tres niveles. La base de datos relacional maneja los requerimientos de almacenamiento de datos, y el motor ROLAP proporciona la funcionalidad analítica. El nivel de base de datos usa bases de datos relacionales para el manejo, acceso y obtención del dato. El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios. El motor ROLAP se integra con niveles de presentación, a través de los cuales los usuarios realizan los análisis OLAP.

Después de que el modelo de datos para el AD se ha definido, los datos se cargan desde el sistema operacional. Se ejecutan rutinas de bases de datos para agregar el dato, si así es requerido por el modelo de datos. Se crean entonces los índices para optimizar los tiempos de acceso a las consultas.

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

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

1.3.2 MOLAP: Procesamiento Analítico Multidimensional en Línea.

Un sistema de Procesamiento Analítico Multidimensional en Línea (MOLAP), utiliza una base de datos multidimensional en la que la información se almacena dimensionalmente. Este sistema utiliza una arquitectura de dos niveles: Las bases de datos multidimensionales y el motor analítico. La base de datos multidimensional es la encargada del manejo, acceso y obtención del dato; y el motor analítico es el responsable de la ejecución de los requerimientos OLAP. El nivel de presentación se integra con el de aplicación y proporciona una interfaz a través de la cual los usuarios finales visualizan los análisis OLAP.

El nivel de aplicación es el responsable de la ejecución de los requerimientos OLAP. El nivel de presentación se integra con el de aplicación y proporciona una interfaz a través de la cual los usuarios finales visualizan los análisis OLAP. Una arquitectura cliente/servidor permite a varios usuarios acceder a la misma base de datos multidimensional.

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.

1.3.3 HOLAP: Procesamiento Analítico Híbrido en Línea.

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.

Últimamente, 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 tiene problema de escalabilidad, por ejemplo: la adición de dimensiones a un esquema

ya existente. Por otra parte, las implementaciones de ROLAP son más escalables y a menudo son más sociables debido a que se aprovechan de las inversiones de la tecnología de las bases de datos relacionales. En la solución de este sistema, se usará el modelo ROLAP.

1.4 Integración de datos.

1.4.1 Integración de los datos. Características.

ETL - este término viene de ingles de las siglas Extract-Transform-Load que significan Extraer, Transformar y Cargar y se refiere a los datos en una empresa. ETL es el proceso que organiza el flujo de los datos entre diferentes sistemas en una organización y aporta los métodos y herramientas necesarias para mover datos desde múltiples fuentes a un AD, limpiarlos y cargarlos en otra base de datos, Datamart ó bodega de datos (5).

1.4.2 Etapas del proceso de integración de datos.

Debido a que los datos deberán ser extraídos, transformados, limpiados y cargados desde el conjunto de archivos DBF hacia el MD, es imprescindible conocer como se realizarán cada una de estas actividades.

Extracción: Obtención de la información de las distintas fuentes tanto internas como externas.

Transformación: Luego de realizarse el proceso de extracción los datos provenientes de las diferentes fuentes pueden ser incoherentes, tener errores o estar incompletos. Con esto se busca obtener datos lo más precisos, completos, consistentes, interpretables y accesibles. Después del proceso de limpieza se lleva a cabo la integración de los datos con el propósito de eliminar problemas de redundancia e identificar las fuentes de datos más fiables. Una vez realizado el proceso de extracción y limpieza se procede a transformar los datos para de esta forma estandarizar los códigos, corregir los datos, eliminar registros duplicados, usar conversiones y combinaciones para generar nuevos campos.

Carga: Organización y actualización de los datos y los metadatos en la base de datos.

Si no se realiza un correcto proceso de ETL se pudieran obtener datos incorrectos lo que afectaría el proceso de toma de decisiones, es por eso que este proceso constituye aproximadamente un 70% del trabajo de la construcción de un AD (6).

1.5 Inteligencia de Negocios.

1.5.1 ¿Qué es la inteligencia de negocios?

Inteligencia de negocio (BI) por sus siglas en inglés, es el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la compañía) en información estructurada, para su explotación directa (reporting, vistas OLAP) o para su análisis y conversión en conocimiento soporte a la toma de decisiones sobre el negocio (7).

La inteligencia de negocio es un factor estratégico dentro de una empresa u organización, genera una potencial ventaja competitiva, proporcionando información privilegiada que responden a los problemas de negocio: entrada a nuevos mercados, promociones u ofertas de productos, eliminación de islas de información, control financiero, optimización de costes, planificación de la producción, análisis de perfiles de clientes, y rentabilidad de un producto concreto.

1.5.2 Componentes de una solución de inteligencia de negocio.

Todas las soluciones de BI tienen funciones similares, pero deben reunir al menos los siguientes componentes:

Multidimensionalidad: La información multidimensional se puede encontrar en hojas de cálculo, bases de datos, etc. Una herramienta de BI debe de ser capaz de reunir información dispersa en toda la empresa e incluso en diferentes fuentes para así proporcionar a los departamentos la accesibilidad, poder y flexibilidad que necesitan para analizar la información. Por ejemplo, un pronóstico de ventas de un nuevo producto en varias regiones no está completo si no se toma en cuenta también el comportamiento histórico de las ventas de cada región y la forma en que la introducción de nuevos productos se ha desarrollado en cada región en cuestión.

Data Mining: Las empresas suelen generar grandes cantidades de información sobre sus procesos productivos, desempeño operacional, mercados y clientes. Pero el éxito de los negocios depende por lo general de la habilidad para ver nuevas tendencias o cambios en las tendencias. Las aplicaciones de minería de datos pueden identificar tendencias y comportamientos, no sólo para extraer información, sino también para descubrir las relaciones en bases de datos que pueden identificar comportamientos que no son muy evidentes.

Agentes: Los agentes son programas que "piensan". Ellos pueden realizar tareas a un nivel muy básico sin necesidad de intervención humana. Por ejemplo, un agente puede realizar tareas un poco complejas, como elaborar documentos, establecer diagramas de flujo.

AD es la respuesta de la tecnología de información a la descentralización en la toma de decisiones. Coloca información de todas las áreas funcionales de la organización en manos de quien toma las decisiones. También proporciona herramientas para búsqueda y análisis.

1.6 Modelo de datos.

Un modelo de datos es un conjunto de estructuras que describen los datos, sus relaciones, su significado y las condiciones que los datos deben cumplir para reflejar correctamente la realidad deseada.

Características (8)

- Es el proceso de analizar los aspectos de interés para una organización y la relación que tienen unos con otros.
- Resulta en el descubrimiento y documentación de los recursos de datos del negocio.
- El modelado hace la pregunta " ¿Qué? " en lugar de " ¿Cómo? ", ésta última orientada al procesamiento de los datos.
- Es una tarea difícil, pero es una actividad necesaria cuya habilidad solo se adquiere con la experiencia.
- Registrar los requerimientos de datos de un proceso de negocio.
- Permite observar:

    Patrones de datos.

    Usos potenciales de los datos.

1.6.1 Modelo Entidad-Relación.

Los diagramas o modelos entidad-relación (denominado por sus siglas, ERD “Diagram Entity relationship”) son una herramienta para el modelado de datos de un sistema de información. Estos modelos expresan entidades relevantes para un sistema de información, sus inter-relaciones y propiedades.

El ERD está basado en una percepción del mundo real que consta de un conjunto de objetos básicos llamados entidades con sus atributos y de las interrelaciones que existen entre estos objetos. Se desarrolló para facilitar el diseño de bases de datos permitiendo la especificación de un esquema del universo de discurso que representa la estructura completa de las mismas. El modelo entidad-relación (MER) es uno de los diferentes modelos de datos semánticos que existe; el aspecto semántico del modelo reside en su intento de representar el significado de los datos. Este modelo es extremadamente útil para hacer corresponder los significados e interacciones del desarrollo del mundo real con un esquema conceptual. Los esquemas de MER usan diagramas para representar la estructura natural de los datos, que se nombran diagrama entidad relación. En esos diagramas los rectángulos representan a las entidades y los rombos representan a las interrelaciones. Las interrelaciones son enlazadas con sus entidades constitutivas por arcos, y el grado de la interrelación es indicado en el arco.

Elementos que componen el diagrama entidad-relación:

- Entidades.

  Las entidades son objetos reales o abstractos relevantes en el universo de discurso, que pueden ser identificadas unívocamente y acerca de los que se colecciona información; usualmente denotan una persona, lugar, cosa o evento de interés informacional.

- Atributos.

  Cada entidad, instancia de un conjunto de entidades, es descrita por un conjunto de atributos que representan sus cualidades, características o propiedades relevantes, y por los valores asociados a los mismos.

- Relaciones.

Las relaciones son asociaciones o conexiones que existen entre dos o más entidades. Las entidades relacionadas pueden pertenecer al mismo o a distintos conjuntos de entidades.

- Cardinalidad.

Número de posibles relaciones que una entidad determinada puede tener sobre otra y se especifica por la cantidad mínima y máxima de instancias de la asociación.

1.6.2 Modelo dimensional.

El modelado dimensional es una técnica de modelado de datos que permite la visualización de los mismos. Se utilizan para diseñar AD con la particularidad de que estos van a estar compuestos por hechos, medidas y dimensiones. (9)

Existen tres esquemas que se utilizan para modelar la estructura de un almacén de datos. Ellos son:

Esquema estrella

El esquema estrella está formado por una tabla de hechos con una única tabla para cada dimensión. Este método se basa en el “esquema en estrella”, que consiste en un modelo asimétrico con una tabla grande dominante en el centro del esquema, se encarga de conectar las otras tablas. El esquema en estrella básico tiene 4 componentes: hechos, dimensiones, atributos y jerarquías de atributo (5).

Esquema en copo de nieves

Es una variante más compleja del esquema estrella. Las tablas de las dimensiones en este modelo representan relaciones normalizadas y forman parte del modelo relacional de base de datos (5).

Constelación de hechos

La constelación de hechos es un conjunto de tablas de hechos que comparten algunas tablas de dimensiones (5).

Hechos

Un hecho es una colección de medidas relacionadas con sus dimensiones. Puede representar un objeto de negocio, una transacción o un evento que es utilizado por el analista de información. Los hechos contienen:

- ✓ Un identificador de hechos.
- ✓ Llaves de dimensión, que lo enlaza con las dimensiones.
- ✓ Medidas.
- ✓ Varios tipos de atributos, los que usualmente se derivan de otros datos en el modelo.

Dimensiones

Una dimensión es una entidad o una colección de entidades relacionadas usadas por los analistas para identificar el contexto de las medidas con las que trabajan, estas determinan el contexto para las medidas. Las dimensiones contienen:

- ✓ Entidades de dimensión.
- ✓ Atributos de dimensión.
- ✓ Jerarquías de dimensión.
- ✓ Niveles de agregación.

Medidas

Una medida es un tipo de dato cuya información es usada por los analistas (usuarios) en sus consultas para medir el rendimiento del comportamiento de un proceso o un objeto del negocio. Las medidas candidatas son los datos numéricos, pero no cada atributo numérico es una medida candidata.

Indicadores

Los indicadores son variables que pueden tomar un valor de una determinada unidad de medida y de un determinado tipo de datos.

Atributos

Son criterios utilizados para analizar los indicadores. Se basan, en los datos de referencia de las tablas de dimensiones. En un cubo, los atributos son los ejes del mismo. Son campos o criterios de análisis, pertenecientes a tablas de dimensiones.

Jerarquía

Una jerarquía representa una relación lógica entre dos o más atributos; si poseen una relación "padre-hijo".

Tienen las siguientes características:

- ✓ Existen varias en un mismo cubo.
- ✓ Tienen dos o más niveles.
- ✓ Relación "1-n" o "padre-hijo" entre atributos consecutivos de un nivel superior y uno inferior.
- ✓ Se pueden identificar cuando existen relaciones "1-n" o "padre-hijo" entre los propios atributos de un cubo.

Granularidad

La granularidad es el nivel de detalle en que se almacena la información. A mayor nivel de detalle, mayor posibilidad analítica. Los datos con granularidad fina (nivel de detalle) podrán ser resumidos hasta obtener una granularidad media o gruesa. No sucede lo mismo en sentido contrario.

1.7 Selección de topología, herramientas y metodología de desarrollo.

1.7.1 Topología usada para el desarrollo del presente mercado de datos.

La topología o esquema usado para el mercado de datos Inmigración y extranjería es la de estrella, porque está conformado por una tabla de hecho llamada hech_arribo_por_visitantes y cinco dimensiones (edad, sexo, país, motivo de viaje, tiempo), donde se relacionan todas las dimensiones con el hecho.

1.7.2 Metodología para el desarrollo.

La Oficina Nacional de Estadísticas e Información basándose en su papel como órgano rector en materia estadística en Cuba amerita la utilización de una metodología robusta y madura que garantice el éxito de la integración de la información que actualmente disponen. De todo el conjunto de metodologías existentes para enfrentar el desarrollo del MD la decisión es adecuar, la mundialmente

conocida, metodología de Kimball, adaptándola, a la realidad de la Universidad de las Ciencias Informáticas (UCI), por las siguientes razones:

- La técnica de Kimball posee una gran cantidad de documentación y generalmente se puede encontrar una respuesta a casi todas las problemáticas que puedan presentar.
- Su creador Ralph Kimball es una figura emblemática en el mundo de almacenar teniendo publicados alrededor de 100 artículos científicos proponiendo mejoras al proceso, además de innumerables libros que se han posicionado como guías de obligatoria consulta para el desarrollo, ejemplo de esto es su libro Técnicas de Diseño Dimensional que en la actualidad se ha convertido en un éxito editorial dentro del campo.
- Claridad de las actividades a realizar por cada rol propuesto.
- Esta metodología de dividir el mundo de BI entre el hecho y las dimensiones es muy eficaz y conduce a una solución completa en un tiempo razonable.
- Es iterativo, donde se construye una pieza a la vez (MD) garantizando mayor velocidad de respuesta a los clientes.
- La forma de almacenar la información es de fácil entendimiento por parte del usuario lo que permite mayor comprensión para el análisis de los datos que se encuentran integrados.
- Es una metodología resistente y adaptable ante los cambios.

De la metodología de Kimball se toma sus cuatro fases fundamentales: análisis, diseño, inteligencia de negocio, e integración de datos; y su trabajo orientado a requerimientos; y se propone el empleo de casos de uso (10).

1.7.3 Justificación de las herramientas a utilizar.

En la actualidad se han desarrollado diversas herramientas con el fin de dar un acercamiento a la automatización del diseño, construcción, implementación y mantenimiento de los almacenes de datos.

Herramientas de modelado.

Las herramientas de modelado se emplean para la creación de modelos de sistemas que ya existen o que se desarrollarán. Ofrecen gran usabilidad y generación de código.

Las herramientas CASE (Ingeniería de Software Asistida por Computadora), constituyen un conjunto de ayudas para el desarrollo de programas informáticos.

En la presente investigación se decidió utilizar **Visual Paradigm 6.4** para UML como herramienta de modelado, por ser una herramienta UML profesional que soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. El software de modelado UML ayuda a una más rápida construcción de aplicaciones de calidad a un menor coste. Permite representar todos los tipos de diagramas de clases, código inverso, generar código desde diagramas y generar documentación. Genera código para un gran número de lenguajes de programación entre los que se encuentra Java y permite integración con varias herramientas de Java. Además, brinda una versión libre para uso no comercia (11).

También brinda características como:

- Generación de bases de datos.
- Ingeniería inversa de bases de datos desde sistemas gestores de bases de datos (DBMS) existentes a diagramas de Entidad-Relación.
- Generador de informes para generación de documentación.
- Soporta aplicaciones web.
- 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.

- 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.

Gestor de Base de Datos.

PostgreSQL 8.4 es un potente Sistema de Base de Datos Relacional libre (Open Source, su código fuente está disponible) liberado bajo licencia The PostgreSQL Licence (TPL) similar a la Berkeley Software Distribución (BSD) (12)

Entre las diversas características que presenta esta herramienta, se tienen:

- Soporta sintaxis en SQL.
- Posee puntos de recuperación a un momento dado, tablespaces, replicación asincrónica, transacciones jerarquizadas (savepoints), copia de seguridad en línea.
- Alta concurrencia, lo que permite que mientras un proceso escribe en una tabla, otros accedan a la misma tabla sin necesidad de bloqueos.
- Aproxima los datos a un modelo objeto-relacional, y es capaz de manejar complejas rutinas y reglas. Ejemplos de su avanzada funcionalidad son consultas SQL declarativas, control de concurrencia multi-versión, soporte multi-usuario, transacciones, optimización de consultas, herencia, y arrays.
- Soporta operadores, funciones métodos de acceso y tipos de datos definidos por el usuario.
- Incluye características avanzadas tales como las uniones (joins) SQL92.
- Soporta integridad referencial, la cual es utilizada para garantizar la validez de los datos de la base de datos.

- La flexibilidad del API de PostgreSQL ha permitido a los vendedores proporcionar soporte al desarrollo fácilmente para el RDBMS PostgreSQL. Estas interfaces incluyen Object Pascal, Python, Perl, PHP, ODBC, Java/JDBC, Ruby, TCL, C/C++, y Pike.

- Tiene soporte para lenguajes procedurales internos, incluyendo un lenguaje nativo denominado PL/pgSQL. Este lenguaje es comparable al lenguaje procedural de Oracle, PL/SQL. Otra ventaja de PostgreSQL es su habilidad para usar Perl, Python, o TCL como lenguaje procedural embebido.

- Hace transparente al usuario los detalles del almacenamiento físico de los datos, mediante varios niveles de abstracción de la información.

- Permite la realización de cambios a la estructura de la base de datos, sin tener que modificar la aplicación que emplea.

- Provee al usuario la seguridad de que sus datos no podrán ser accedidos, ni manipulados por quien no tenga permiso para ello. Debido a esto, posee un complejo sistema que maneja grupos, usuarios y permisos para las diferentes actividades que se pueden realizar dentro del mismo.

- Mantiene la integridad de los datos.

- Proporciona una manera eficiente de realizar copias de seguridad de la información almacenada, y permite a partir de estas copias restaurar los datos.

- Controla el acceso concurrente de los usuarios.

- Facilita el manejo de grandes volúmenes de información.

PgAdmin III, 1.10: Es una herramienta de código abierto para la administración de bases de datos PostgreSQL y derivados (EnterpriseDB Postgres Plus Advanced Server y Greenplum Database). Incluye: (13)

- Interfaz administrativa gráfica.

- Herramienta de consulta SQL (con un EXPLAIN gráfico).

- 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. Soporta versiones de servidores 7.3 y superiores. Versiones anteriores a 7.3 deben usar el PgAdmin II.

Herramientas para la integración de datos.

Las herramientas para la integración de datos son muy útiles para que el proceso de ETL concluya con los resultados esperados, su uso garantiza: ganancias en términos de tiempo y total fiabilidad de los datos.

Pentaho Data Integration 4.0.1 (PDI) (14):

- Es de formato abierto y de fácil lectura para los XML que recogen transformaciones, tareas programadas y un repositorio relacional de metadatos ETL.
- Es aplicable a diversos tipos de bases de datos (SQL server, PostgreSQL, MySQL, Microsoft Access, etc.).
- Posee facilidad para la importación y exportación de datos de un formato a otro cualquiera.
- Su principal fortaleza es la posibilidad que brinda de ser extensible mediante pluggins.

DataCleaner 1.5.3

DataCleaner es una aplicación Open Source para el perfil, la validación y comparación de datos. Estas actividades ayudan a administrar y supervisar la calidad de los datos con el fin de garantizar que la información es útil y aplicable a su situación de negocio. Es la alternativa gratuita almacenamiento de datos, proyectos de investigación estadística, la preparación para la extracción, transformación y carga

(ETL) de actividades y mucho más. DataCleaner le da el poder de personalizar en el respeto de lo agradable de la sencillez, además de ser un software de código abierto y gratuito. (15).

Herramientas para la Inteligencia de Negocios.

Estas herramientas son un tipo de software de aplicaciones diseñado para colaborar con BI en los procesos de las organizaciones. Específicamente se trata de herramientas que asisten el análisis y la presentación de los datos.

Pentaho Schema Workbench 3.2.0

Es una interfaz de diseño que permite crear y probar esquemas de cubos mondrian OLAP visualmente. La Plataforma de BI de Pentaho incrusta el motor de consulta Mondrian, como parte de su arquitectura. Además, permite la ejecución de consultas MDX (16).

Pentaho BI server 3.6.0

La aplicación más conocida de la Plataforma de BI es la Pentaho BI Server que funciona como una red basada en sistema de gestión de informe, el servidor de integración de aplicaciones y un motor de flujo de trabajo ligero (secuencias de acción.) Está diseñado para integrarse fácilmente en cualquier proceso de negocio (17).

Mondrian 3.0.4

En el plano Open Source la herramienta más significativa es **Mondrian**, la cual es una de las aplicaciones más importantes de la Suite Pentaho BI. Mondrian es un servidor OLAP open source que gestiona la comunicación entre una aplicación OLAP y la base de datos con los datos fuente. Es desarrollado en Java/Servlets/JSPs que permite ser instalado en servidores de aplicaciones como JBoss. Entre sus principales características se encuentra la facilidad para el análisis de grandes volúmenes de información que se encuentren almacenados en bases de datos que soporten JDBC. Mondrian soporta el lenguaje Microsoft's Multidimensional Expressions (MDX). También soporta los APIs: Java OLAP (JOLAP) y XML for Analysis application Programming (18).

Apache Tomcat 6.0

En sus inicios el uso de Tomcat de forma autónoma era sólo recomendable para entornos de desarrollo y entornos con requisitos mínimos de velocidad y gestión de transacciones. En la actualidad ya no existe esa percepción y Tomcat es usado como servidor web autónomo en entornos con alto nivel de tráfico y alta disponibilidad.

El mismo puede funcionar como servidor web por sí mismo, y funciona en cualquier sistema operativo que disponga de la máquina virtual Java. Los usuarios disponen de libre acceso a su código fuente y a su forma binaria en los términos establecidos en la Apache Software Licence (19).

1.8 Conclusiones.

A partir de lo estudiado sobre los Almacenes de Datos se concluyó:

- La tecnología apropiada para la problemática en cuestión es el Mercado de Datos.
- La metodología de desarrollo adoptada es una adaptación a la de Kimball.
- El modo de almacenamiento propuesto es ROLAP.
- Las herramientas a utilizar son: Pentaho Schema Workbench 3.2.0, Pentaho Analysis server (Mondrian) 3.0.4, Pentaho BI Server, Apache Tomcat 6.0, Pentaho Data Integrator 4.0.1, DataCleaner 1.5.3, PostgreSQL 8.4, PgAdmin III, y Visual Paradigm 6.4.

Capítulo 2. Análisis y diseño del mercado de datos del área Inmigración y Extranjería.

2.1 Introducción.

En este capítulo se abordan aspectos concernientes a la descripción e implementación de la solución, específicamente, a las características de las fuentes a integrar, definición de las áreas de análisis, análisis de datos, la arquitectura, y modelo de datos propuesto. De manera general, aborda el resultado del análisis y el diseño del MD de Inmigración y extranjería, para el Departamento de Turismo y Comercio de la ONEI.

2.2 Caracterización de las áreas de la organización.

Con la evaluación de las áreas de la organización se tiene como propósito analizar y evaluar el área en la que está enmarcado el MD. Lo que se pretende es hacer constancia de las evaluaciones efectuadas a las áreas de la organización. Para la evaluación de las mismas a informatizar, se utilizaron las técnicas de recopilación de información, específicamente la entrevista con el cliente.

Se obtiene una descripción detallada de las entrevistas que se ha tenido con el cliente, un informe de los resultados y una fundamentación de los mismos, así como las conclusiones de dicha evaluación.

Día de la semana	Hora	Elemento a evaluar	Especialistas a participar	Nombres y Apellidos del Especialista.
12/10/2010	10:00 am	-Establecer el alcance del proyecto. -Entender el negocio	Especialista de la ONE	Elena Leonila Fernández
19/11/2010	2:00 pm	Aprobar los requisitos de información, funcionales y no funcionales	Especialista de la ONE	Elena Leonila Fernández
25/01/2011	11:20 am	Definir las dimensiones, los hechos y las	Especialista de la ONE	Elena Leonila Fernández

		medidas del mercado de datos	Ingeniero Informático	Roberto Téllez
2/02/2011	3:30pm	-Aprobación del Modelo de Datos	Especialista de la ONE	Elena Leonila Fernández

Tabla 1. Cronograma de la evaluación de las áreas a informatizar.

Informe de resultados.

Primera Entrevista:

La especialista explicó todo lo referente con el negocio y se estableció el alcance del proyecto.

Segunda Entrevista:

Se analizaron los requisitos propuestos y se agregaron nuevas funcionalidades.

Tercera Entrevista:

Se establecieron los hechos, las dimensiones y las medidas del mercado de datos.

Cuarta Entrevista:

La especialista agregó y quitó nuevas dimensiones y medidas.

Fundamentación de los resultados.

Primera Entrevista:

No se realizó señalamientos.

Segunda Entrevista:

Se decidió agregar nuevos requisitos de información.

Tercera Entrevista:

Se aprobó la propuesta de hechos, dimensiones y medidas.

Cuarta Entrevista:

Se decidió quitar dimensiones innecesarias para el MD.

Se agregaron nuevas medidas al modelo de datos.

Finalmente se aprobó el diseño del modelo de datos.

Conclusiones de la evaluación.

Primera Entrevista:

No se realizó señalamientos, por lo que fue evaluado de satisfactorio.

Segunda Entrevista:

Los resultados fueron aceptados, aunque se precisaron algunos detalles sobre los requisitos.

Tercera Entrevista:

Los resultados fueron satisfactorios pues se aprobó la propuesta de hechos, dimensiones y medidas.

Cuarta Entrevista:

Los resultados fueron aceptados, aunque se realizaron algunos reajustes en el modelo de datos.

2.3 Necesidades de los usuarios.

Se definen como necesidades de los usuarios el análisis y la difusión de los indicadores de Inmigración y extranjería, tanto del año en curso como de años anteriores. El propósito es analizar todo lo referente a la entrada al país de visitantes, por cualquier motivo. Los datos se brindarán en un período de tiempo variable: día, mensuales, trimestrales o anuales.

2.4 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 principales objetivos, 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.

Para poder llevar a cabo los cálculos y análisis estadísticos en el Departamento de Turismo y Comercio de la ONEI, 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 dirección de

Inmigración y extranjería (DIE) tiene una codificación que no es la misma que utiliza el departamento de Turismo y Comercio de la ONEI, por lo que la DIE debe utilizar el clasificador de país internacional, puesto en vigor por la ONEI.

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 = Al resto del valor total del período actual - el valor total del período anterior con el que se está comparando.
- Dinámica = 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.
- Una vez cargados los datos en el almacén, no pueden existir campos nulos.
- El código de los atributos en cada una de las dimensiones no pueden tomar valores repetidos.
- En el sexo, la letra “F” significa mujer, y la letra “M” significa hombre.
- Cada cálculo de la fuente de datos significa una persona que entra al país.
- En la fecha de nacimiento esta compuesta por seis dígitos: los dos primeros dígitos significan el año, los siguiente dos dígitos significan el mes, y los último dos dígitos el día.
- En la fecha de nacimiento, los dos primeros dígitos que significan el año, se refiere a los años 1900.

2.5 Especificación de requerimientos.

La especificación de requerimientos es una descripción completa del comportamiento del sistema que se va a desarrollar.

2.5.1 Requisitos de información.

Los requisitos de información, constituyen la información que debe estar disponible, la entrada fundamental para todo el proceso de inteligencia del negocio y los futuros reportes bases. De acuerdo con la investigación del negocio se definieron los siguientes requisitos de información:

RI1. Obtener la serie de llegadas de visitantes según motivo de viaje por mes.

RI2. Calcular los principales emisores de visitantes a Cuba según motivo de viaje por mes.

RI3. Calcular el arribo de visitantes por áreas geográficas según el motivo de viaje, por mes.

RI4. Obtener la serie de llegadas de visitantes por mes, sexo, edad, ciudadanía, motivos de viaje, y país de embarque, por mes.

2.5.2 Requisitos funcionales.

Los requisitos funcionales son capacidades o condiciones que el sistema debe cumplir, lo que es muy importante para satisfacer las expectativas del cliente. Se definió como requisitos funcionales:

RF 1 Autenticar usuario.
RF 2. Adicionar usuario.
RF 3. Eliminar usuario.
RF 4. Adicionar rol.
RF 5. Eliminar rol.
RF 6. Adicionar reporte.
RF 7. Eliminar reporte.
RF 8. Realizar la extracción de los datos.
RF 9. Realizar la transformación y carga de los datos.
RF 10. Configurar vistas de análisis OLAP.
RF 11. Editar consultas MDX.
RF 12. Suprimir filas y columnas vacías.

2.5.3 Requisitos no funcionales.

Los requerimientos no funcionales son propiedades o cualidades que el producto debe tener**.** Se identificaron 12 requisitos no funcionales dentro del MD Inmigración y extranjería, a los cuales se hace

referencia en el artefacto de análisis especificación de requerimientos, donde se realiza una explicación detalla de los mismos. Ellos son:

Usabilidad

RN1 Cumplir con las pautas de diseño de la interfaces.

RN2 Mostrar los mensajes, títulos y demás textos que aparezcan en la interfaz del sistema en idioma español.

RN3 Diseñar un reporte del Almacén de Datos de manera sencilla y ágil.

RN4 Navegar en los reportes del Almacén de Datos de manera ágil.

Fiabilidad

RN5 Garantizar la persistencia de la información.

Restricciones de diseño

RN6 Utilizar el lenguaje de programación definido durante la investigación.

RN7 Lograr que los elementos definidos en el almacén tengan una estructura homogénea.

Requisitos para la documentación de usuarios y ayuda del sistema

RN8 Confeccionar manual de usuario

Requisitos de seguridad

RN9 Sesión de usuarios

Requisitos de software

RN10 Instalar en las estaciones de trabajo el software necesario para el correcto funcionamiento del sistema

Requisitos de hardware

RN11 Proporcionar características mínimas de hardware a las estaciones de trabajo.

RN12 Proporcionar características mínimas de hardware a los servidores.

2.6 Casos de uso del sistema.

Los diagramas de caso de uso son una representación de la relación que existe entre los casos de uso. El mismo es una técnica para la captura de requisitos y los actores que intervienen en el sistema. Se define en el MD Inmigración y extranjería, el siguiente diagrama de caso de uso de sistema que se muestra en la Ilustración 2, el cual está compuesto por tres actores, un analista, un administrador, un administrador de ETL, y las actividades que realizan cada uno de ellos.

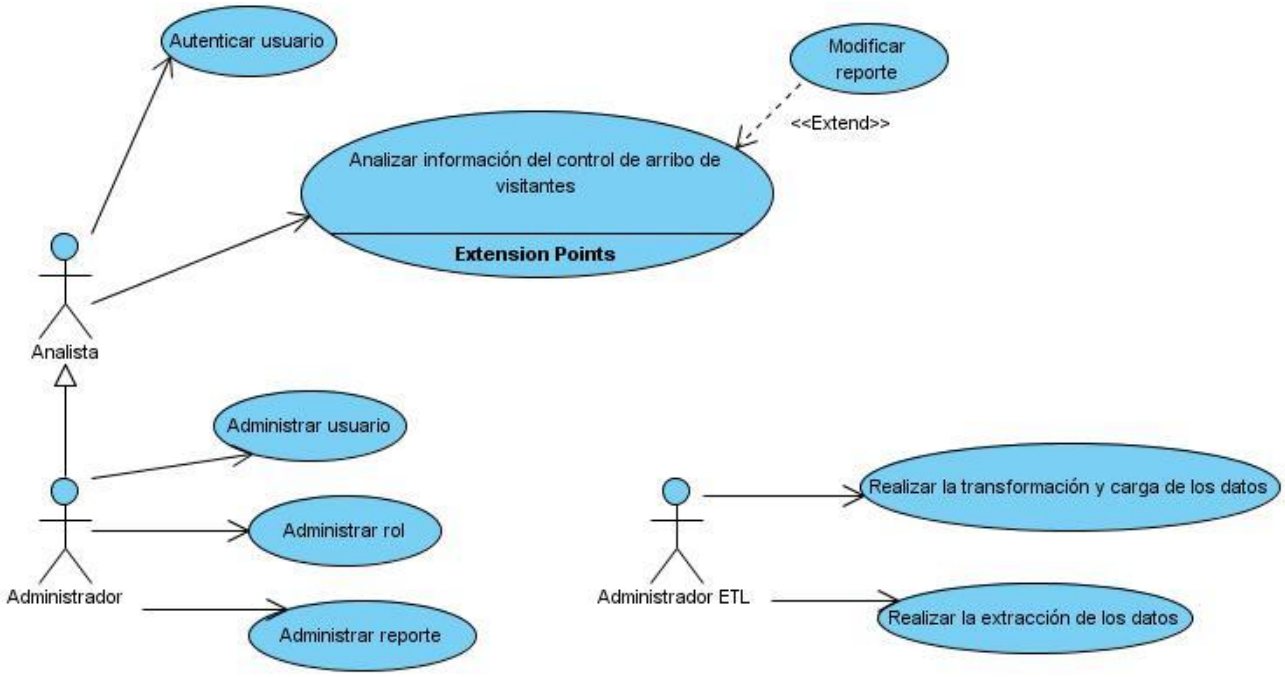


Ilustración 2. Diagrama de CU.

Actores	Descripción
Analista	Es el responsable de analizar la información de los diferentes indicadores de Inmigración y extranjería.
Administrador	Es el responsable de la administración de los usuarios y rol, además de la administración de los reportes.

Administrador ETL	Es el responsable de la extracción, transformación y carga de los datos.

Tabla 2. Actores del diagrama de Casos de uso.

Caso de Uso	Descripción
Autenticar Usuario	Se autentica el usuario en la aplicación.
Analizar información del control de arribo de visitantes.	Visualiza todos los reportes del control de arribo de visitantes a la isla.
Realizar la transformación y carga de los datos.	Realiza la transformación y carga de los datos.
Realizar la extracción de los datos.	Realiza la extracción de los datos.
Administrar rol	Se insertan roles nuevos y se eliminar otros ya existentes.
Administrar usuario	Se insertan usuarios nuevos y se eliminar otros ya existentes.
Administrar reporte	Se insertan, actualizan y eliminan reportes.
Modificar reporte	Se muestran los reportes existentes y te permite interactuar con ellos.

Tabla 3. Casos de uso.

2.7 Diseño de la Solución.

El diseño es un proceso mediante el que se traducen los requisitos en una representación del software. En la solución se ha definido un MD donde convergen todas las dimensiones propuestas: sexo, edad, motivos de viaje, país y tiempo.

2.7.1 Matriz bus o matriz dimensional.

La matriz bus, es la relación que existe entre las dimensiones y los hechos del MD. Se define como la habilidad para describir y seguir la vida tanto de una dimensión como de un hecho, la cual permite determinar el impacto que provocaría un cambio durante el desarrollo del sistema. En la tabla 4 que se muestra a continuación se muestra la relación que existe entre el hecho arribo por visitantes y las dimensiones país, sexo, motivo de viaje, temporal mes y edad; aquí vemos que todas estas dimensiones se relacionan con el hecho.

Dimensiones /Tabla de Hecho.	**Arribo por Visitantes**
dim_pais	X
dim_sexo	X
dim_motivo_viaje	X
dim_temporal_mes	x
dim_edad	X

Tabla 4.Matriz bus.

2.7.2 Modelo de datos.

El modelo de datos está conformado por las dimensiones, medidas, y hechos identificados para el área de Inmigración y extranjería. Por las necesidades del negocio existe un modelo que relaciona las dimensiones definidas y las medidas que se han detallado hasta el momento. En la Ilustración 3, se

muestra una vista del modelo de datos para el área, en el cual se ve la relación que existe entre las dimensiones y la tabla de hechos.

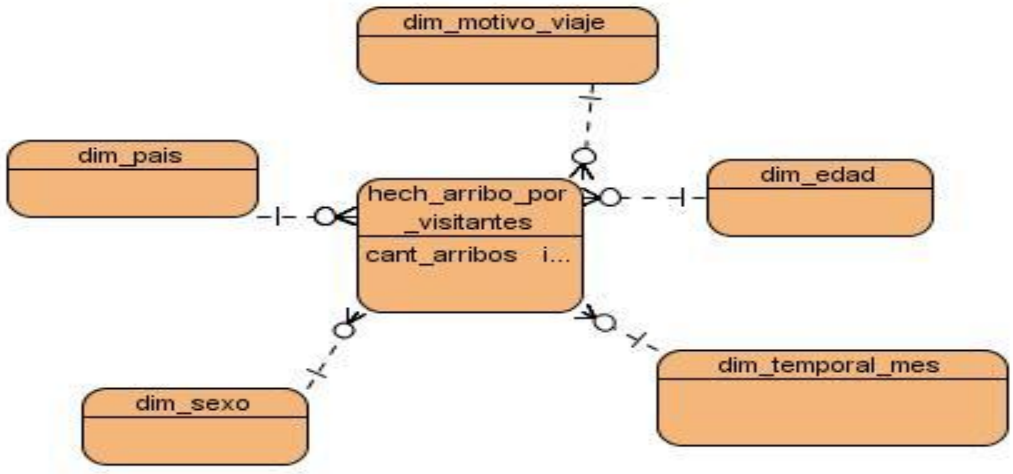


Ilustración 3. Modelo de datos.

2.7.3 Dimensiones.

Las dimensiones poseen algunas características que prevalecen, entre ellas, por ejemplo la definición de jerarquías entre sus atributos, con el objetivo de plasmar explícitamente la forma en que se puede consolidar.

Dimensión país:

Esta dimensión contiene el país al que pertenece cada visitante, y su país de embarque, además describe el universo de valores bajo los cuales puede clasificarse la información atendiendo el clasificador de país.

Jerarquía:

Continente.

Continente->Área geográfica.

Continente->Área geográfica->país.

Dimensión Sexo:

Esta dimensión describe el universo de valores bajo los cuales puede clasificarse el sexo.

Jerarquía: Sexo.

Dimensión Motivo de Viaje:

Esta dimensión contiene los diferentes motivos de viaje que pueden tener los visitantes y describe el universo de valores bajo los cuales puede clasificarse la información atendiendo al clasificador de motivo de viajes.

Jerarquía: Motivos de Viaje.

Dimensión Tiempo:

Esta es la dimensión más común e importante en los diseños de MD ya que define una línea de tiempo específica.

Jerarquía:

Anual.

Anual -> Trimestral.

Anual -> Trimestral->Mensual.

Dimensión Edad:

Esta dimensión describe la edad de los visitantes.

Jerarquía:

Grupo quinquenal

Grupo quinquenal->edad

2.7.4 Tablas de hechos.

Las tablas de hechos son las que almacenan las medidas numéricas. En este caso, se definió como medida numérica un valor, que se capta en los ficheros de la Dirección de Inmigración y extranjería (DIE). Se identificó una sola tabla de hecho:

Tabla de Hecho: Arribo por Visitantes.

En esta tabla es donde va a residir, toda la información existente. 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 enmarcado en el departamento de Turismo y Comercio, y específicamente en el arribo de visitantes a Cuba.

Medidas

Las medidas están implícitas en las Tablas de Hechos:

Cantidad de arribos de visitantes.

(cant_arribos).

2.8 Política de respaldo y recuperación.

En el MD Inmigración y extranjería, la política de respaldo y recuperación que se empleará está condicionada por tres puntos fundamentales:

- Periodicidad de salvas del sistema**:** Estas se realizaran constantemente en un período aproximado de 30 días a toda la información contenida en el MD, verificando siempre la existencia de una copia de toda la información almacenada.
- Tablas involucradas: Las tablas que se involucran en la realización son: hech_arribo_por_visitantes.
- Salvas existentes: A pesar de que actualmente no existen salvas en esta área se prevé la realización de reemplazos de estas cada un año, así como también el chequeo de su estado mensualmente, mediante pruebas de rendimiento y flexibilidad.

2.9 Esquema de seguridad.

La seguridad para el MD Inmigración y extranjería, está determinada mayormente por los niveles de acceso al sistema. Dicha seguridad se rige fundamentalmente por los roles y permisos que los usuarios poseen en su interacción con la base de datos y la aplicación.

Seguridad en la base de datos.

Para la interacción de los usuarios con la base de datos se definió el rol de:

Actor	Permiso
Administrador	Tiene total acceso a la base de dato del sistema.

Tabla 5. Seguridad en la aplicación.

Actualmente las aplicaciones desplegadas en el servidor de BI de Pentaho muestran un sucesivo incremento, así como los usuarios que tiene acceso a estas. Como consecuencia de esto se define los siguientes roles:

Roles	Permisos
Administrador	Tiene acceso total a todas las Áreas de Análisis General (AAG). Gestiona el Sistema de información de gobierno.
Analista	Tiene acceso de solo lectura al AA inmigración y extranjería. Visualiza los reportes.

Tabla 6. Roles y permisos.

Elemento de aplicación	Roles con acceso
AA General	Administrador

Carpeta raíz: AA inmigración y extranjería.	Administrador Analista

Tabla 7. Roles.

2.10 Conclusiones.

Al finalizar el desarrollo de este capítulo:

- Quedaron explícitas las reglas del negocio, así como los requisitos según las necesidades de los usuarios.
- Se diseñó el modelo de caso de uso del sistema donde se representa la relación de los casos de usos y los actores de la aplicación.
- Se diseñó el modelo de datos en el cual se muestra la interacción de los hechos y las dimensiones identificadas.
- Se elaboró los roles y permisos con el fin de obtener una mayor seguridad en la aplicación.

Capítulo 3. Implementación del mercado de datos del área Inmigración y extranjería.

3.1 Introducción.

En este capítulo se desarrollará la implementación de los procesos ETL y BI para el área de Inmigración y extranjería. Se muestran los cubos OLAP, las consultas MDX y la interfaz visual para que el cliente interactúe con el servidor Mondrian o con el pentaho bi-server; además se presenta el procedimiento elaborado, y aplicado para la etapa de implementación a partir de las características del negocio, dándole solución a los requisitos del sistema.

3.2 Implementación de la base de datos.

El modelado de datos es uno de los elementos más importantes a la hora de iniciar el desarrollo de cualquier proyecto. La verdadera esencia de una aplicación reside en esta estructura. Cuando se diseña el modelo dimensional, el mismo se transforma a un modelo físico, del cual se genera el script de la base de datos, y es allí donde se evidencian las relaciones que existen entre las diferentes tablas, y a la vez determina si el proyecto va a cumplir con su verdadero objetivo.

3.2.1 Estructura de los datos.

Para que la manipulación de los datos en un MD se realice de manera correcta, es necesario organizarlos en estructuras lógicas que faciliten y optimicen esta tarea. Estas estructuras son denominadas: esquemas y tablas.

Esquemas.

Los esquemas en una base de datos, son una forma de organizar la información contenida en la misma. Los usuarios solo tendrán acceso a aquellos que su rol se lo permita. Dentro de los esquemas se pueden encontrar funciones, operadores y tipos de datos que facilitarán su implementación.

En el presente trabajo se definieron dos esquemas:

Esquema dimensiones: contiene las tablas de las dimensiones generales del AD, y de ellas utilizamos las que se necesiten para implementar el MD.

Esquema mart_inmigración_extranjería: contiene todas las tablas de hechos y las dimensiones propias propuestas en el MD.

Tablas.

Del diseño del modelo de datos, se concluye que la solución propuesta consta de seis tablas en total, cinco dimensiones y un hecho, distribuidas en los dos esquemas anteriormente planteados. En la Ilustración 4 se muestra utilizando la herramienta pgAdmin la estructura física del MD, quedando definido dentro del esquema dimensiones, las tablas país, sexo, edad, temporal mes y en el esquema mart_inmigración_extran, las tablas propias del MD, motivo de viaje y el hecho arribo por visitantes.

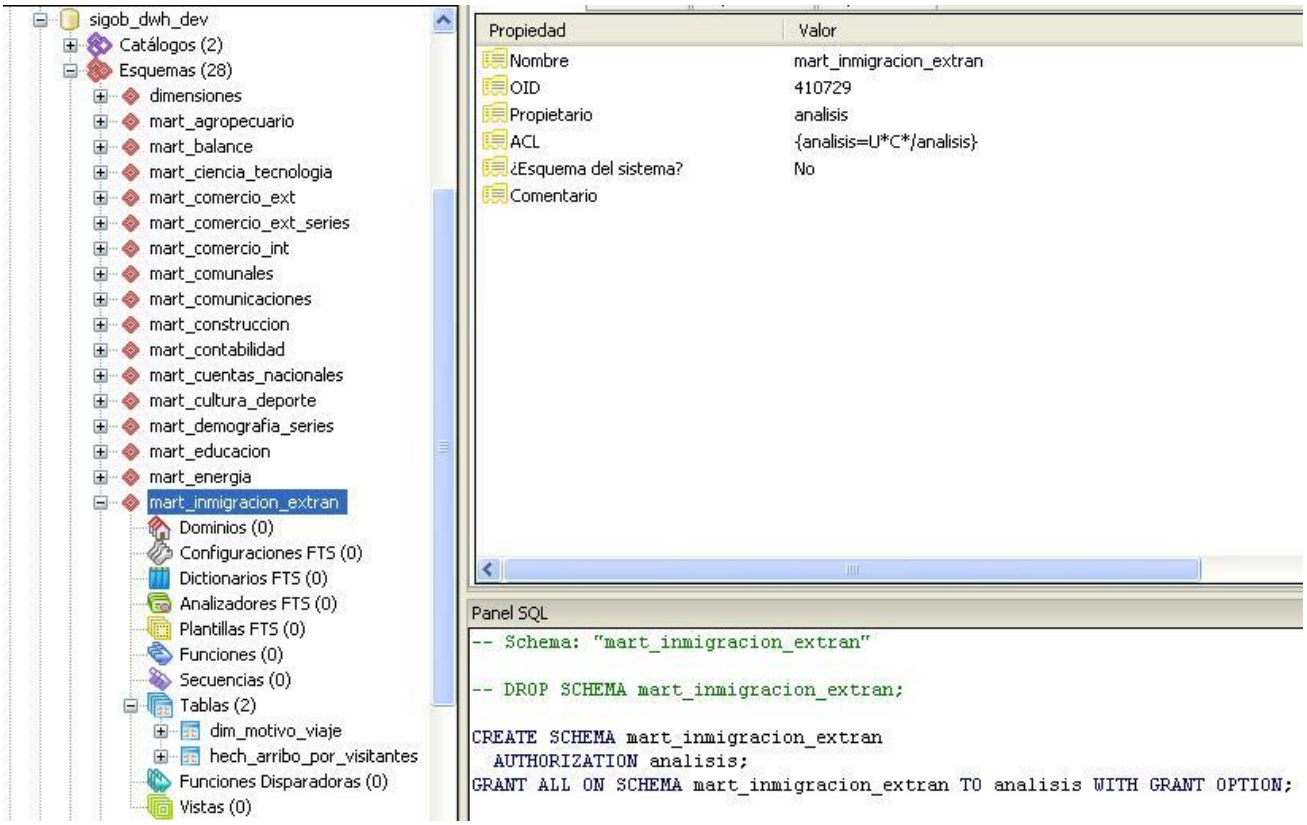


Ilustración 4. Estructura física de la BD.

3.3. Usuarios y privilegios.

Para garantizar la seguridad de la base de datos, se deben definir los usuarios y roles, para luego poder agruparlos según las necesidades de permisos y accesos que necesitaran cada uno de ellos, para con esto, poder realizar su función como trabajador del sistema.

Los roles establecidos son:

Administrador de ETL: Su función se basa en la realización de los procesos de ETL en la interacción con la BD.

Privilegios:

Los privilegios que se les asigna a los usuarios del sistema son basados en el rol que desempeñan:

Administrador de ETL: Por el rol que este usuario desempeña a la hora de interactuar con la BD, se le es asignado los privilegios de Select, Insert, Update, Delete, Refresh y Trigger de los datos almacenados en el MD.

3.4 Implementación del subsistema de integración de datos.

3.4.1 Arquitectura del subsistema de integración.

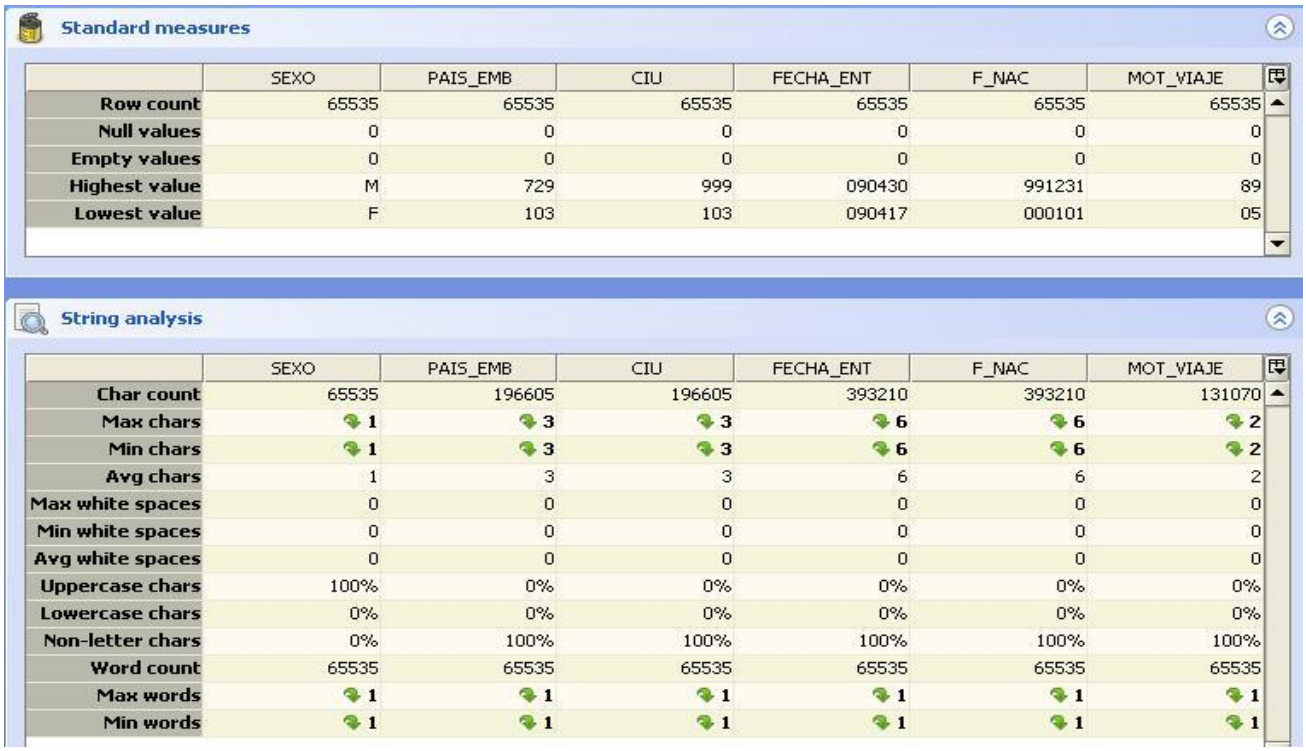
La construcción de un software no se puede comenzar hasta no tener bien definida una arquitectura. Esta no es más que un grupo de patrones que sirven de guía para la elaboración del mismo.

Para la integración de los datos, la arquitectura queda de la siguiente forma:

- **Fuente de Datos**: Son archivos de extensión dbf o xls que contienen la información.
- **Área temporal**: Es el punto intermedio entre la fuente de datos y el MD. Es donde se realiza la integración y transformación de los datos.
- **Mercado de Datos**: Donde son cargados los datos para su futuro análisis.

Perfilado de Datos.

Es el proceso que se encarga de analizar las fuentes, para conocer el estado en que se encuentran los datos, su calidad y su estructura. La herramienta que se utilizó fue DataCleaner, obteniéndose importantes resultados con los cuales se establecen reglas para luego realizar el proceso de integración de los datos.



Standard measures

	SEXO	PAIS_EMB	CIU	FECHA_ENT	F_NAC	MOT_VIAJE
Row count	65535	65535	65535	65535	65535	65535
Null values	0	0	0	0	0	0
Empty values	0	0	0	0	0	0
Highest value	M	729	999	090430	991231	89
Lowest value	F	103	103	090417	000101	05

String analysis

	SEXO	PAIS_EMB	CIU	FECHA_ENT	F_NAC	MOT_VIAJE
Char count	65535	196605	196605	393210	393210	131070
Max chars	1	3	3	6	6	2
Min chars	1	3	3	6	6	2
Avg chars	1	3	3	6	6	2
Max white spaces	0	0	0	0	0	0
Min white spaces	0	0	0	0	0	0
Avg white spaces	0	0	0	0	0	0
Uppercase chars	100%	0%	0%	0%	0%	0%
Lowercase chars	0%	0%	0%	0%	0%	0%
Non-letter chars	0%	100%	100%	100%	100%	100%
Word count	65535	65535	65535	65535	65535	65535
Max words	1	1	1	1	1	1
Min words	1	1	1	1	1	1

Ilustración 5. Perfilado de datos.

Extracción de los Datos.

Es el proceso donde se obtiene toda la información de las distintas fuentes tanto internas como externas. Se cargan los datos de los archivos dbf o excel, para el area temporal y así adaptarlos al modelo relacional que se ha establecido. Estos archivos contienen toda la información referente al MD inmigración y extranjería, la cual será almacenada en la tabla de hecho.

Transformación y Limpieza.

Luego de realizarse el proceso de extracción, se realiza la limpieza de los datos provenientes de las diferentes fuentes, porque los mismos pueden ser incoherentes, tener errores o estar incompletos. Con esto se busca obtener datos precisos, completos, consistentes, interpretables y lo más accesibles posibles. Después del proceso de limpieza se lleva a cabo la integración de los datos con el propósito de eliminar problemas de redundancia e identificar las fuentes de datos más fiables. La transformación y limpieza es de gran importancia porque en esta etapa es donde se garantiza el resultado final de cómo se van a mostrar los datos, se aplican las reglas del negocio; y se detectan otras posibles deficiencias de la fuente y se corrigen.

Carga

El paso final de este proceso lo constituye la carga y es donde los datos son cargados al MD, organizados y actualizados, para que puedan ser usados por el cliente de forma satisfactoria. En el presente trabajo se elaboró un área temporal, donde van a ser llevados los datos, y una vez estando allí serán limpiados y transformados para posteriormente ser cargados al MD.

En la ilustración 6 se muestra la carga al área temporal, la misma es un ejemplo de las transformaciones implementadas que abarca todos los procesos de la integración de datos: la extracción, limpieza, transformación y carga de los mismos.

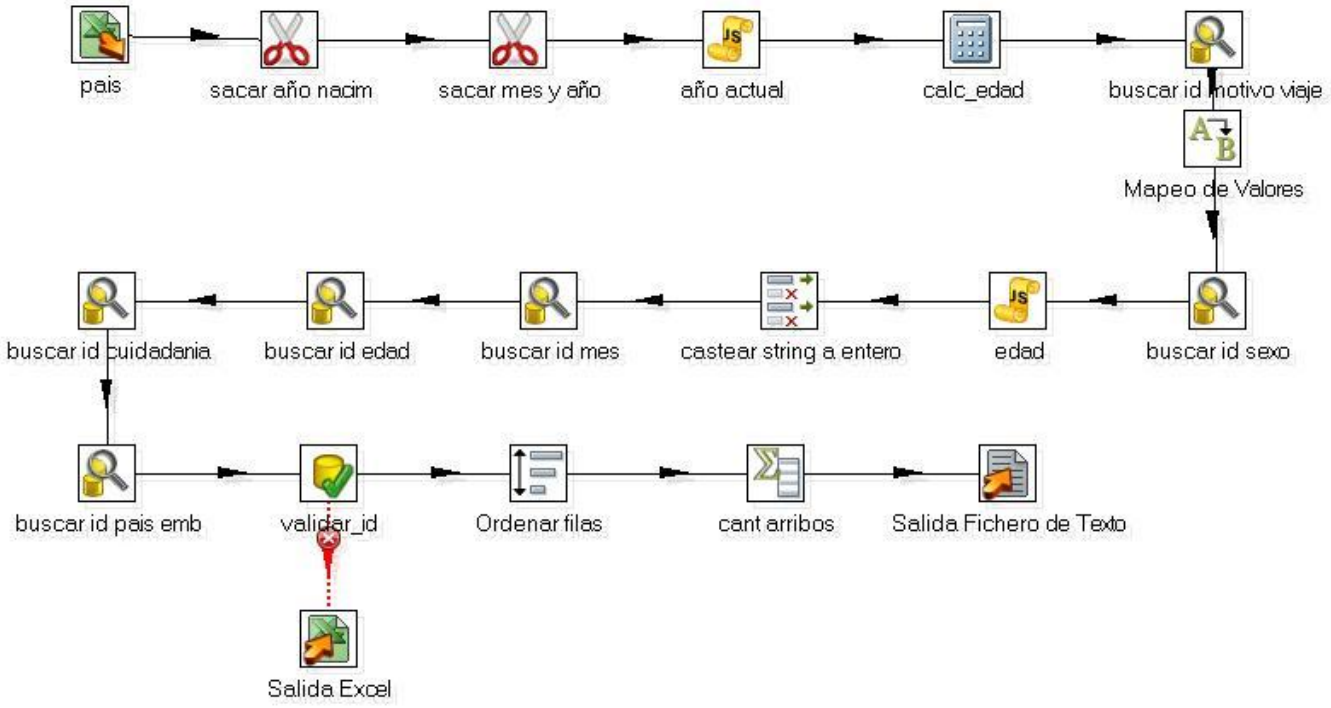


Ilustración 6. Transformación: Carga al área temporal.

Del archivo fuente se obtiene el año de nacimiento, luego se procede a obtener el mes y el año en que la persona entró al país, que sirve para una vez sacado el año actual con un componente, poder calcular la edad del visitante. Se buscan las llaves primarias en las dimensiones correspondientes y se aplican las reglas del negocio identificadas, con el propósito de calcular la medida cantidad de arribos y una vez concluida se procede a la carga de la tabla de hecho correspondiente.

3.5 Implementación de los trabajos.

Una vez que la conexión al MD se encuentra en perfecto estado, se procede a la carga del mismo, y el trabajo (job en ingles) es la forma en que se realiza la carga de los datos hacia el MD.

Para la correcta realización de un job, se debe tener bien definido cuales son las dimensiones estáticas y cuales no, pues en un job sólo se cargan las dimensiones que pueden tomar valores nuevos o cambiar los que tenían anteriormente.

En este caso solo cargaremos la dim_motivo_viaje y el hech arribo_por_visitantes, porque la carga de las demás dimensiones se realiza previamente. En caso de ocurrir algún error cuando se está realizando la cargar del job, el error se guardará en un fichero excel, el cual tendrá una ubicación establecida anteriormente y conocida por el cliente, para poder solucionar los problemas que se presenten.

Debido a que la información origen se encuentra en varios excel, se crearon dos trabajos; un trabajo general donde se carga la dimensión motivo de viaje, luego la información existente en los archivos excel es cargada al área temporal (esto lo hace el otro trabajo) y una vez estando allí, los datos son preparados y llevados al lugar destino, que es el MD.

En las siguientes Ilustraciones se muestran los trabajos que se realizaron en este proceso.

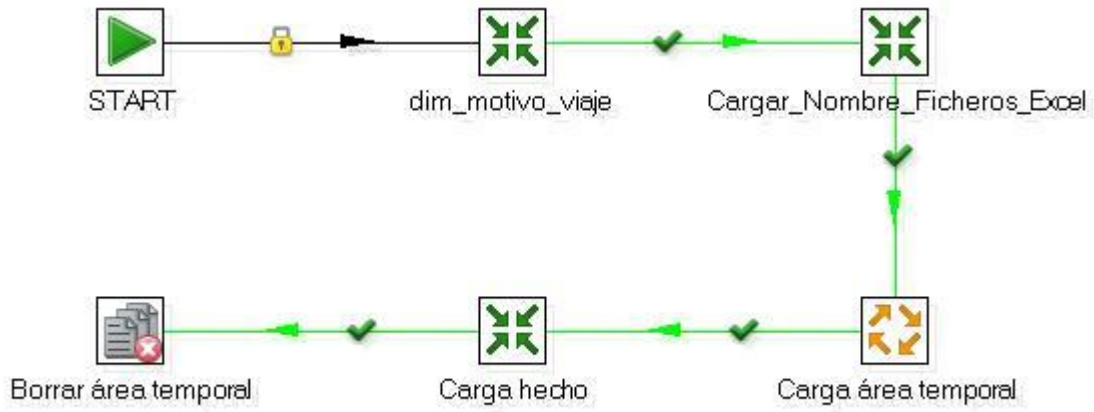


Ilustración 7. Trabajo general.

Ilustración 8. Trabajo cargar hecho.

3.6 Implementación del subsistema de visualización de datos.

3.6.1 Cubos OLAP.

El objetivo del OLAP es agrupar los datos con el propósito de facilitar su posterior análisis, este proceso consiste en obtener datos relevantes entre la gran cantidad de información contenida en el MD. Se pueden agregar múltiples dimensiones que permitirán extraer, en forma rápida y eficiente, la información que se requiere. Aprovechar las ventajas de un sistema OLAP a través de cubos permitirá realizar análisis multidimensionales a un menor costo.

Para la implementación del subsistema de visualización, es necesaria la creación de los cubos multidimensionales en los cuales se definen las dimensiones, los niveles de jerarquía de las dimensiones y las medidas.

En el diseño de los cubos OLAP debe realizarse un estudio de las necesidades del negocio identificadas con el cliente, además, deben tenerse en cuenta los indicadores y perspectivas de análisis así como los tipos de reportes identificados anteriormente.

A partir de las dimensiones y medidas descritas del cubo OLAP, se procede a crear cada uno de los cubos necesarios para obtener los reportes que se desean, usando para esto las herramientas seleccionadas.

En la realización del diseño de los cubos se modeló en el esquema inmigración un cubo OLAP: hech_arribo_por_visitantes, el cual esta formado por sus dimensiones, y medidas.

La siguiente Ilustración muestra el diseño del cubo OLAP.

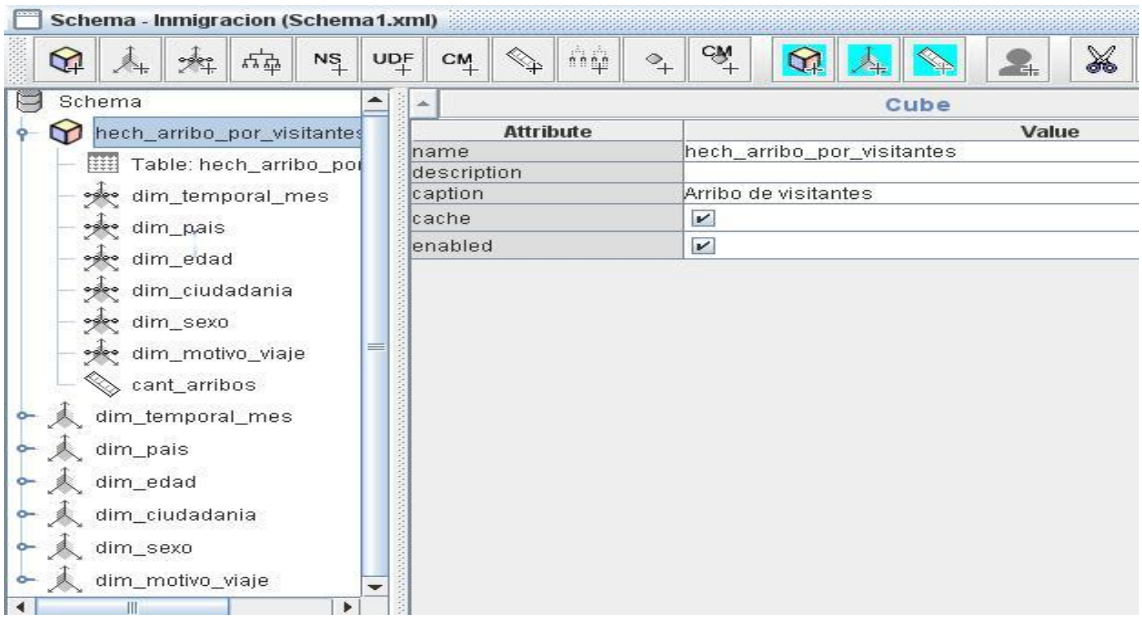


Ilustración 9. Cubo OLAP.

3.6.2 Navegación de la capa de visualización.

El mapa de navegación es una representación gráfica donde se encuentra organizada toda la información. El MD inmigración y extranjería está conformado por un Área de Análisis General (A.A.G), un Área de Análisis (A.A) y un Libro de Trabajo (L.T), dentro del cual se encuentran las tablas de salida (TS), que son cuatro con sus respectivas vistas de análisis.

A continuación se muestra la estructura en la cual se representa la capa de visualización.

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

A.A.G SIGOB: Agrupa toda la información referente a los diferentes MD presentes en la ONE, formando un AD.

Descripción del Área de Análisis (A.A).

A.A Inmigración y extranjería: Agrupa toda la información referente a los visitantes que entran a la isla por cualquier motivo de viaje.

Descripción de los Libros de Trabajo (L.T)

L.T Visitantes: Libro de trabajo contenido dentro del A.A inmigración y extranjería, contiene cuatro reportes que permiten realizar un análisis general de datos.

Descripción de las vistas de análisis OLAP del L.T inmigración y extranjería.

TS1- obtener la serie de llegada de visitantes según motivo de viaje por mes.

TS2- Calcular los principales emisores de visitantes a Cuba según motivo de viaje por mes.

TS3- Calcular el arribo de visitantes por áreas geográficas según el motivo de viaje, por mes.

TS4- Obtener la serie de llegadas de visitantes por sexo, edad, ciudadanía, motivos de viaje y país de embarque, por mes.

En la siguiente Ilustración se muestra la vista de análisis calcular el arribo de visitantes por áreas geográficas según el motivo de viaje y país de embarque, por mes. Además permite aplicar técnicas OLAP como son navegar en profundidad, rotar, suprimir filas vacías, entre otras más.

		Tiempo			
		Todos		2009	
		Medidas		Medidas	
Motivo de viaje	País	Variación	Cantidad de arribos (U)	Variación	Cantidad de arribos (U)
Todos	África	270	270	172	221
	África Central	197	197	185	191
	África Meridional	10	10	-10	
	África Occidental	12	12	-8	2
	África Oriental	47	47	9	28
	África Septentrional	4	4	-4	
	Oceanía	12	12	-12	
	América	80.229	80.229	80.229	80.229
	Asia	11	11	3	7
	Europa	36.019	36.019	35.981	36.000

Ilustración 10. Vista OLAP.

3.7 Configurar la seguridad de los usuarios.

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

El rol de administrador tiene todos los permisos de la aplicación y posee el usuario administrador del sistema.

El rol de analista tiene permiso de solo lectura y posee el usuario analista del sistema.

Ilustración 11. Roles.

3.8 Conclusiones.

- Se analizó la fuente de datos para lograr implementar el MD.
- Se mostró la estructura física del MD, logrando poblarlo satisfactoriamente.
- Se modeló el esquema multidimensional con sus respectivos cubos OLAP para agrupar los datos y así facilitar su análisis posteriormente.
- Se logró identificar el área de análisis, el libro de trabajo y los reportes candidatos.
- Se implementaron y visualizaron los reportes candidatos correctamente.

Capítulo 4. Validación del mercado de datos del área Inmigración y Extranjería.

4.1 Introducción.

En este capítulo se documenta una lista de chequeo para evaluar el proceso de análisis de datos y de ETL del MD Inmigración y extranjería, así como realizar casos de pruebas por cada caso de uso de información. Primeramente se realiza una introducción a los conceptos en cuestión y luego se elaboran dichas pruebas a partir de aspectos importantes encontrados durante todo el ciclo de vida del presente trabajo.

4.2 Concepto de evaluación.

La definición textual de evaluación es: análisis de una cosa que determina su valor, importancia o trascendencia. (20)

Su concepto es polisémico, porque este se aplica según las necesidades mismas de la evaluación y en función de las diferentes formas de concebirla. En efecto, puede significar tanto estimar y calcular como valorar o apreciar.

La evaluación es el proceso de juzgar los logros que un proyecto o programa ha conseguido, especialmente en relación con los objetivos propuestos; proporciona un panorama claro del grado en que se han realizado estos objetivos; implica un juicio de valor; y es importante para identificar los obstáculos o estancamientos que pudieran impedir al proyecto alcanzar sus objetivos. De esta manera las soluciones a los obstáculos pueden identificarse e implementarse. (21)

La evaluación puede hacerse antes, durante y después de la implementación. En este trabajo de diploma se evaluará el procedimiento después de haberlo implementado; de esta forma se podrá:

- Identificar obstáculos y problemas en la implementación del procedimiento.
- Proporcionar descripciones, a manera de solución a los problemas que se hayan presentado.
- Proporcionar un panorama real del grado de cumplimiento de determinadas reglas, basadas en la satisfacción de los objetivos de las técnicas OLAP.

4.3 ¿Qué es una lista de chequeo?

Se entiende por lista de chequeo a un listado de preguntas, en forma de cuestionario que sirve para verificar el grado de cumplimiento de determinadas reglas establecidas a priori con un fin determinado. Se utilizan básicamente en la práctica de la investigación que forma parte de un proceso de evaluación. En fin, la lista de chequeo no es más que una herramienta confiable y manipulable, que permite registrar, clasificar y organizar todo tipo de elementos para una evaluación.

Las listas de chequeo son herramientas importantes para agrupar gran cantidad de información y conocimiento, de manera concisa, evitando en su aplicación errores de omisión, creando un mecanismo fiable y reproducible, mediante evaluaciones que permiten mejorar normas de calidad.

El uso de estas listas está generalizado en áreas muy diversas, que van desde verificar y determinar el potencial de mercados extranjeros, hasta medir la confiabilidad y seguridad de sistemas informáticos. Se utilizan en áreas multivariadas dentro de perspectivas tan distintas como puede ser la sanitaria, la industrial o las ciencias.

Entre las principales características de las listas de chequeo, se encuentran:

- Facilita que personas no especialistas en el tema tratado, puedan entender y validar teorías, procedimientos o metodologías con una validez y confiabilidad estadística, que le permite extrapolar información dentro de los estándares establecidos.
- Obliga al evaluador a considerar por separado cada uno de los elementos que se evalúan la lista en cuestión.
- Son herramientas metodológicas que constan de una serie de indicadores que evalúan detalladamente una serie de elementos teóricos o prácticos sobre temáticas específicas.
- Permiten descubrir el proceso de validación de una evaluación.
- Las respuestas deben determinar de modo satisfactorio el cumplimiento o no de los procedimientos que se estén evaluando. (22)

4.4 Elaboración y evaluación de la lista de chequeo.

Para elaborar la lista de chequeo, se tuvieron en cuenta los elementos de evaluación que no deben faltar una vez que se realice el proceso de ETL y BI del MD Inmigración y extranjería permitiendo

recoger los puntos eficientes e ineficientes que posean dichos procesos. La lista de chequeo contiene diferentes indicadores a evaluar los cuales se encuentran distribuidos en tres secciones fundamentales:

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

4.4.1 Elementos que forman parte de la estructura de la lista de chequeo.

Peso: Define si el indicador a evaluar es crítico o no.

Indicadores a evaluar: Son los indicadores a evaluar en las secciones Estructura del documento, Semántica del documento e Indicadores definidos por las diferentes etapas.

Evaluación: Es la forma de evaluar el indicador en cuestión. El mismo se evalúa de uno en caso de que exista alguna dificultad sobre el indicador y cero en caso de que el indicador revisado no presente problemas.

No procede: Se usa para especificar que el indicador no es necesario evaluarlo en ese caso.

Cantidad de elementos afectados: Especifica la cantidad de errores encontrados sobre el mismo indicador.

Comentario: Especifica los señalamientos o sugerencias que quiera incluir la persona que aplica la lista de chequeo.

Una vez aplicada la lista de chequeo se detectan los indicadores evaluados de mal, con el objetivo de darles solución.

4.4.2 Evaluación del resultado de la lista de chequeo.

Se evalúa de mal la calidad del proceso de análisis de datos si:

Existen al menos dos indicadores críticos evaluados de mal en la sección Indicadores evaluados por las etapas, que posee la lista de chequeo.

- Incumple con más del 50 % de los indicadores a evaluar.
- Se mantienen las no conformidades (NC) de una revisión a otra.

Se evalúa de regular la calidad del proceso de BI y ETL si el mismo no cumple los criterios para ser evaluado de mal y:

- La cantidad de elementos afectados de un indicador evaluado de mal es superior a tres.
- Incumple con los indicadores críticos a evaluar de las secciones Estructura del documento y Semántica del documento que posee la lista de chequeo.
- Existe al menos un indicador crítico evaluado de mal.
- Existen al menos cuatro indicadores no críticos evaluados de mal de la sección Indicadores definidos por la etapa, que posee la lista de chequeo.

El proceso de análisis de datos es evaluado de bien si no cumple ninguno de los criterios anteriores y:

- No existe ningún indicador crítico evaluado de mal.
- Si la cantidad indicadores no críticos evaluados de mal de la sección Indicadores definidos por la etapa, que posee la lista de chequeo no es mayor que tres.

Lista de Chequeo para evaluar el proceso de ETL y BI realizado al MD: (Ver anexo 1).

4.5 Casos de prueba.

Las pruebas de software son un elemento crítico para garantizar la calidad del software y representan una revisión final de las especificaciones del diseño y de la codificación (23).

- Las pruebas de software son siempre necesarias.
- En algunos casos ocupan un 40% del tiempo de un proyecto informático.
- Las pruebas pretenden descubrir errores.

Un buen caso de prueba es aquel que tiene una probabilidad muy alta de descubrir un nuevo error, es una herramienta para comprobar la disponibilidad de los perfiles de análisis y los indicadores a medir, así como también verificar el cumplimiento de los requisitos de información a través de los reportes candidatos.

En el presente trabajo de diploma se le realiza un caso de prueba al caso de uso de información: Analizar información del control de arribo de visitantes. (Ver Anexo 2 y 3)

De la aplicación de los casos de prueba se obtuvieron NC que fueron analizadas y a todas se le dio el tratamiento correspondiente, obteniendo la carta de aceptación por parte del cliente.

4.6 Conclusiones.

- Se describió el desarrollo y la aplicación de una lista de chequeo y de un caso de prueba.
- Se evaluaron los resultados de los procesos de análisis e integración de datos.
- Se analizaron 24 indicadores de la lista de chequeo, que se encuentran distribuidos en tres secciones: Estructura del documento, Indicadores definidos por la etapa, y Semántica del documento, necesarios e imprescindibles para la evaluación final de los procesos, de ellos 11 indicadores con peso crítico.
- La evaluación final fue de Bien, resultado que demuestra la calidad del análisis e integración de datos realizado.

CONCLUSIONES GENERALES

En el trabajo que se concluye, como consecuencia de la investigación realizada por las necesidades de la Oficina Nacional de Estadística e Información, se logró aplicar técnicas OLAP para el análisis de datos, así como integrar toda la información de Inmigración y extranjería para el Sistema de Información de Gobierno. Se lograron los siguientes resultados:

- Se seleccionaron las herramientas necesarias para implementar el MD así como la metodología que guió todo el proceso.
- Se refinó el análisis y diseño del MD Inmigración y extranjería, para brindar una mejor solución a las necesidades del cliente, se obtuvieron seis nuevas reglas del negocio, se agregaron seis requisitos funcionales, nuevos requisitos no funcionales y de los ocho requisitos de información, sólo se dejaron cuatro.
- Se realizó la implementación del proceso de ETL, quedando poblado el MD Inmigración y extranjería.
- Se desarrolló el cubo OLAP con el cual se llevó a cabo el proceso de análisis de datos del MD Inmigración y extranjería.
- Se elaboró y aplicó una lista de chequeo y un caso de prueba, para evaluar los procesos de análisis e integración de datos, arrojando no conformidades que fueron resueltas, obteniendo la carta de aceptación por parte del cliente.

Con este trabajo de diploma se logró obtener un MD para la ONEI, que permite el análisis e integra toda la información de Inmigración y extranjería, ayudando al proceso de toma de decisiones.

RECOMENDACIONES

Para el presente trabajo de diploma se recomienda:

- Adicionar nuevas formas de visualización de la información y aplicar otras técnicas de inteligencia de negocios.
- Realizar un estudio sobre el tema de la seguridad en la base de datos, para lograr la confidencialidad de los mismos.

REFERENCIAS BIBLIOGRÁFICAS

1. Sinergia e Inteligencia de Negocio. [En línea] 18 de enero de 2007. [Citado el: 2010 de octubre de 25.] http://www.sinnexus.com/empresa/creditos.aspx.

2. **Ralph, Kimball.** The Data Warehouse Lifecycle Toolkit. EUA : Wiley Publishing Inc, 2002.

3. **Ponniah, Paulraj.** Data Warehousing Fundamentals. EUA : Wiley Publishing Inc, 2001.

4. Rolap,Molap,Holap. [En línea] [Citado el: 2 de noviembre de 2010.] http://www.csae.map.es/csi/silice/DW2251.html.

5. Business Intelligence - Almacenes de Datos - ETL. (s.f.). Recuperado el 12 de 11 de 2010, de http://etl-tools.info/es/bi/proceso_etl.htm.

6. Sinnexus. [En línea] 2007. [Citado el: 2 de octubre de 2010.] www.sinnexus.com.

7. Business Intelligence. 2007.

8. UDLAP. UDLAP. [Online] 1997. [Cited: 11 11, 2010.] http://ict.udlap.mx/people/carlos/is341/bases02.html.

9. Aran Bey Tcholakian Morales, Dr. Eng. Modelo Dimensional.

10. Zepeda Sánchez, Leopoldo. Tesis

11. Boost Productivity with innovative and intuitive technologies. (s.f.). Recuperado el 2 de 10 de 2010, de http://www.visual-paradigm.com/product/?favor=vpuml

12. PostgreSQL. [En línea] Tinysofa Copyright-grupo de desarrollo global de postgreSQL, 1996-2010. [Citado el: 6 de noviembre de 2010.] http://www.postgresql.org.

13. Programming 4 Us. (s.f.). Recuperado el 7 de 11 de 2010, de http://mscerts.programming4.us/es/16904.aspx

14. Pentaho Data Integrator. [En línea] 2005. [Citado el: 1 de noviembre de 2010.] http://www.pentaho.com/products/data_integration/.

15. DataClenaer. [En línea] [Citado el: 3 de diciembre de 2010.] http://datacleaner.eobjects.org/.

16. Inteligencia de negocios en acción. [En línea] [Citado el: 2 de 12 de 2010.]

17. Pentaho. [En línea] [Citado el: 13 de 1 de 2011.] http://jira.pentaho.com/browse/BISERVER/fixforversion/10861.

18. Mondrian. [En línea] [Citado el: 2 de 12 de 2010.] http://jasperserver.sourceforge.net/docs/3-7-0/Mondrian-3.0-Technical-Guide.pdf.

19. Apache. [En línea] [Citado el: 2 de 12 de 2010.] http://tomcat.apache.org/tomcat-6.0-doc/index.html.

20. MÁS ALLÁ DE LA SUPERVISIÓN. Evaluación de los logros. [En línea] 9 de septiembre del 2009. [Citado el: 25 de Enero del 2010.] 3.4.1HTTP://WWW.SCN.ORG/MPFC/INDEX.HTM)

21. **ROTTEMBERG, ANIJOVICH** (2005) “Cap. 4 La evaluación” en: Estrategia de enseñanza y diseño de unidades de aprendizaje, Universidad Nacional de Quilmes (Carpeta de Trabajo) [Citado el: 10 de Enero del 2010.]

22. **Scriven Michael.** The Logic and Methodology of Checklists. June 2000. [Citado el: 18 de Enero del 2010.]

23. Autores, varios. Pruebas. 2003.

BIBLIOGRAFÍA

1. Business Intelligence. 2007.

2. DataClenaer. [En línea] [Citado el: 3 de diciembre de 2010.] http://datacleaner.eobjects.org/.

3. **Escalona, Ana G. y Aguilar, Yadiurvis.** Sistema de Información de Gobierno: Mercado de Datos para la actividad de comercio. Habana.UCI : s.n., 2010.

4. **Falcón, Yolanda y Leyva, Reynaldo.** Mercado de Datos Estadístico de Inmigración y Extranjería para el Departamento de Turismo y Comercio de la Oficina Nacional de Estadísticas. Habana.UCI : s.n., 2010.

5. **Hernández Lopez, Asnioby.** DOCUMENTO DE ARQUITECTURA DEL SISTEMA.Almacén de datos para la ONE. Cuba,Habana : s.n., 2009.

6. **Inmon, Bill.** Products & Services. [En línea] 2007. [Citado el: 3 de octubre de 2010.] http://www.inmoncif.com/about/index.php.

7. **Kimball, Ralph.** The Data Warehouse Lifecycle Toolkit. EUA: Wiley Publishing Inc, 2002.

8. **Kimballa, Ralph.** The Data Warehouse Toolkit. EUA : Wiley Publishing Inc, 2002.

9. Oficina Nacional de Estadísticas. [En línea] [Citado el: 28 de septiembre de 2010.] http://www.one.cu/.

10. **Ponniah, Paulraj.** Data Warehousing Fundamentals. EUA: Wiley Publishing Inc, 2001.

11. PostgreSQL. [En línea] Tinysofa Copyright-grupo de desarrollo global de postgreSQL, 1996-2010. [Citado el: 6 de noviembre de 2010.] http://www.postgresql.org.

12. Pentaho Data Integrator. [En línea] 2005. [Citado el: 1 de noviembre de 2010.] http://www.pentaho.com/products/data_integration/.

13. Prof. Dr. Ramón García Martínez (Dir.). Servente M. “Algoritmos TDIDT aplicados a la Minería de Datos Inteligente”. Universidad de Buenos Aires, Facultad de Ingeniería. Tesis de Grado en Ingeniería

Informática, 2002. s.l.: Disponible en: http://laboratorios.fi.uba.ar/lsi/servente-tesisingenieriainformatica.pdf.

14. Portada sobre la plataforma Pentaho Open Source Business Intelligence . La plataforma Pentaho Open Source Business Intelligence. [En línea] [Consultada el: 28 de 01 de 2010.] Disponible en: http://pentaho.almacen-datos.com.

15.Qué es OLAP. OlapX. [En línea] OlapXSoftware.com, 2005. http://www.olapxsoftware.com/es/WhatIsOlap.asp.

16. Rolap,Molap,Holap. [En línea] [Citado el: 2 de noviembre de 2010.] http://www.csae.map.es/csi/silice/DW2251.html.

17. Ronda, Borja, iWorld. La empresa multidimensional: OLAP, España: Marta Cabanillas, 1/11/2002, Vols. Revista número 54, Página 62. Sección Firma invitada.

18. Rodríguez Sotolongo, Javier y Peralta Góngora, Yohan Orlando. Implementación del proceso de extracción, transformación y carga de un Datawarehouse para los Ensayos Clínicos del Centro de Inmunología Molecular. 2010.

19. Rodríguez Sotolongo, Javier y Peralta Góngora, Yohan Orlando. Implementación del proceso de extracción, transformación y carga de un Datawarehouse para los Ensayos Clínicos del Centro de Inmunología Molecular. 2010.

20. Revista: iWorld. Revista número: 54. Página: 62. Sección: Firma Invitada. Autor: Borja Ronda. Categoría del artículo: Opinión Gestión

21. Sinergia e Inteligencia de Negocio. [En línea] 18 de enero de 2007. [Citado el: 2010 de octubre de 25.] http://www.sinnexus.com/empresa/creditos.aspx.

22. Sinnexus. [En línea] 2007. [Citado el: 2 de octubre de 2010.] www.sinnexus.com.

23. Símbolo del sistema Inc (Editor), et al**.** PostgreSQL práctica (O'Reilly Unix). 2002.

24. Scriven Michael. The Logic and Methodology of Checklists. June 2000.

25. Thomsen, Erik. OLAP Solutions. Building Multidimensional Information Systems. NEW YORK , CHICHESTER , WEINHEIM, BRISBANE, SINGAPORE, TORONTO: Robert Ipsen, 2002.

26. Vallejos, Sofía J. "Minería de Datos", Trabajo de Adscripción para la Licenciatura en Sistemas de Información, Universidad Nacional del Nordeste Facultad de Ciencias 27.Exactas, Naturales y

Agrimensura, 2006 Argentina. Disponible en: http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/Mineria_Datos_Vallejos.pdf.

28. Zorrilla, Marta. 2007. Data Warehouse y OLAP. Universidad de Cantabria: s.n., 2007.

29. Zepeda Sánchez, Leopoldo. Tesis. [En línea] junio de 2008. [Citado el: 23 de noviembre de 2010.] http://tesis.com.es/autores/zepeda-sanchez-leopoldo-zenaido/.

ANEXOS

Anexo 1

Estructura del documento					
Peso	**Indicadores a evaluar**	**Evaluación**	**No procede**	**Cantidad de elementos**	**Comentarios**
Crítico	1. ¿Los entregables contienen las secciones obligatorias de la plantilla estándar definidas para un expediente de proyecto? (Portada, Control de Versiones, Reglas de Confidencialidad, Tabla de Contenidos y Contenido) (Ver Expediente de Proyecto)	0		0	
Indicadores definidos por la etapa					
Peso	**Indicadores a evaluar**	**Evaluación**	**No procede**	**Cantidad de elemento**	**Comentarios**
	1. ¿La arquitectura satisface las necesidades del proyecto?	0		0	0
	2. ¿La arquitectura soporta el incremento del proyecto?	0		0	0
	3. ¿Se utilizó el menor número de transformaciones posibles al cargar los datos hacia el *Staging area*?	0		0	0
crítico	4. ¿Se creó el Modelo Físico a partir del Modelo Lógico?	0		0	0

crítico	5. ¿Cumple la implementación del proceso de ETL con la arquitectura definida?	0		0	0
	6. ¿Se tuvo en cuenta los formatos fuentes y tipos de datos de las perspectivas de análisis?	0		0	0
	7. ¿Se realiza una limpieza de los datos antes de realizar la carga de los mismos?	0		0	0
	8. ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?	0		0	
	9. ¿Se realizó una interfaz amigable para hacer consultas?	0		0	
Crítico	10. ¿Los reportes son configurables a través de la interfaz del sistema?	0		0	
	11. ¿El rendimiento de los reportes no se afecta cuando el número de dimensiones del modelo se incrementa?	0		0	
	12. ¿Presenta la capacidad de crear todo tipo de dimensiones con funcionalidades aplicables de una dimensión a otra?	0		0	
	13. ¿La interfaz está orientada a facilitar el uso de las funciones del	0		0	

	sistema por parte de los usuarios?				
Crítico	14. ¿No existen restricciones para construir cubos OLAP con dimensiones y niveles de agregación ilimitados?	0		0	
Crítico	15. ¿Los usuarios son capaces de manipular los resultados de manera que se ajusten a sus necesidades, conformando nuevos reportes?	0		0	
	16. ¿El sistema responde de una forma rápida y veraz a la información que le sea solicitada por el usuario?	0		0	
Crítico	17. ¿El sistema refleja cualquier lógica del negocio para poder responder a preguntas específicas?	0		0	
Crítico	18. ¿El sistema garantiza la confidencialidad y seguridad de acceso a los datos por rol de los usuarios?	0		0	
	19. ¿Los datos e información derivados del proceso de análisis realizado mediante la aplicación, apoyan la toma de decisiones en la Institución?	0		0	

Crítico	20. ¿Los cambios en los datos se reflejan automáticamente en los reportes de forma instantánea?	0		0	
Semántica del documento					
Peso	**Indicadores a evaluar**	**Evaluación**	**No proce**	**Cantidad de**	**Comentarios**
Crítico	1. ¿Se han identificado errores ortográficos en los entregables?	0		0	
Crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?	0		0	
	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?	0		0	

Tabla 8. Lista de chequeo

Anexo 2 Reportes candidatos.

Escenario	Descripción	Indicadores a medir	Perfil de análisis	Respuesta del sistema	Flujo central
EC 1.1 Calcular el arribo de visitantes por áreas geográficas según el motivo de viaje, por mes.	Permite visualizar el reporte con las variables presentes en el mismo	Tiempo País Motivo de viaje	Cantidad de arribos Variación Dinámica	Se muestra la tabla con los valores correspondientes a cada escenario.	Se abre la aplicación Se autentifica Se entra al sistema Se despliega hacia la derecha el componente ubicado en

		Sexo Edad			el lateral izquierdo que contiene el navegador Se selecciona el área de análisis **A.A.G General SIGOB**
EC 1.2 Calcular los principales emisores de visitantes a Cuba según motivo de viaje por mes.	Permite visualizar el reporte con las variables presentes en el mismo			Se muestra la tabla con los valores correspondientes a cada escenario.	Se selecciona las áreas de análisis **A.A Inmigración y extranjería** Se selecciona el L.T **visitantes** En la parte inferior izquierda se selecciona el reporte deseado En el área de trabajo se visualiza la tabla correspondiente al reporte Se visualiza el gráfico correspondiente a la información de la tabla.
EC 1.3 Obtener la serie de llegada de visitantes según motivo de viaje por mes.	Permite visualizar el reporte con las variables presentes en el mismo				

EC1.4 Obtener la serie de llegadas de visitantes por sexo, ciudadanía, motivos de viaje, edad, y país de embarque, por mes.	Permite visualizar el reporte con las variables presentes en el mismo				

Tabla 9. Reportes candidatos.

Anexo 3 Descripción de las variables.

No	Nombre de campo	Clasificación	Valor Nulo	Descripción
1	Tiempo	Lista desplegable	No	Se especifican los años que van a ser mostrados en el reporte.
2	Edad	Lista desplegable	No	Se especifica el nivel territorial del reporte.
3	País	Lista desplegable	No	Se especifica la unidad de medida en la que va a estar dado el reporte.
4	Sexo	Lista desplegable	No	Se especifican los indicadores que van a ser mostrados en el reporte.

5	Motivo de viaje	Lista desplegable	No	Se especifican los indicadores que van a ser mostrados en el reporte.
6	Variación	Valores calculables	Si	Se especifican los indicadores que van a ser mostrados en el reporte.
7	Dinámica	Valores calculables	Si	Se especifican los indicadores que van a ser mostrados en el reporte.
8	Cantidad de arribos	Valores calculables	No	Se especifican los indicadores que van a ser mostrados en el reporte.

Tabla 10. Descripción de las variables.

GLOSARIO DE TÉRMINOS

Ad hoc: Se utiliza en informática para referirse a consultas en bases de datos ad hoc querying o ad hoc reporting. Esto implica que el sistema permite al usuario personalizar una consulta en tiempo real, en vez de estar atado a las consultas prediseñadas para informes. Generalmente las consultas ad hoc permiten a los usuarios con poca experiencia en SQL tener el mismo acceso a la información de la base de datos, para esto los sistemas que soportan ad hoc poseen GUIs para generarlas.

Almacén de datos: Es una estructura que se define en función de temas específicos, donde la información histórica debe estar integrada y robusta ante los cambios que puedan afectar a la organización. Su objetivo principal, es servir de ayuda a la toma de decisiones empresariales.

Array: Es un conjunto o agrupación de variables del mismo tipo cuyo acceso se realiza por índices. Los vectores o arrays de dos o más dimensiones se denominan matrices, que pueden tener tantas dimensiones como se desee; aunque lo correcto es llamarlo arreglo (de memoria).

API (Interfaz de Programación de Aplicaciones): Es un conjunto de convenciones internacionales que definen cómo debe invocarse una determinada función de un programa desde una aplicación. Cuando se intenta estandarizar una plataforma, se estipulan unos APIs comunes a los que deben ajustarse todos los desarrolladores de aplicaciones.

Base de datos históricos: Se exponen los principales componentes de las bases de datos como genuinos ejemplos de los sistemas de información histórica y las fuentes de información histórica.

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

Concurrent Versions System o simplemente **CVS**, también conocido como **Concurrent Versioning System**, es una aplicación informática que implementa un sistema de control de versiones: mantiene el registro de todo el trabajo y los cambios en los ficheros (código fuente principalmente) que forman un proyecto (de programa) y permite que distintos desarrolladores (potencialmente situados a gran distancia) colaboren. CVS se ha hecho popular en el mundo del software libre. Sus desarrolladores difunden el sistema bajo la licencia GPL

Data warehousing: Es el centro de la arquitectura para los sistemas de información. Esta arquitectura es capaz de soportar el procesamiento informático al proveer una plataforma sólida, a partir de los datos históricos para hacer el análisis.

Interoperabilidad: Característica de los ordenadores que les permite su interconexión y funcionamiento conjunto de manera compatible. Esto no siempre es posible, debido a los diferentes sistemas operativos y arquitecturas de cada sistema, pero los esfuerzos de estandarización están permitiendo que cada vez sean más los ordenadores capaces de interoperar entre sí.

Minería de datos: Es el proceso de detectar la información procesable de los conjuntos grandes de datos. Utiliza el análisis matemático para deducir los patrones y tendencias que existen en los datos. Normalmente, estos patrones no se pueden detectar mediante la exploración tradicional de los datos porque las relaciones son demasiado complejas o porque hay demasiados datos.

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

OLAP: Es el acrónimo en inglés de procesamiento analítico en línea (On-Line Analytical Processing). Es una solución utilizada en el campo de inteligencia de negocio, cuyo objetivo es agilizar la consulta de grandes cantidades de datos.

Polisémico: Se aplica a la palabra que tiene más de un significado.

Repositorio: Es un término utilizado en el dominio de las herramientas CASE. El repositorio podría definirse como la base de datos fundamental para el diseño; no sólo guarda datos, sino también algoritmos de diseño y, en general, elementos necesarios para llevar a cabo nuestro trabajo.

Sistemas externos: Son un método mas rápido y cómodo de salvaguardar datos en redes pequeñas o para trabajo en equipo sin necesidad de usar voluminosos ordenadores.

Software: Es el equipamiento lógico o soporte lógico de una computadora digital; comprende el conjunto de los componentes lógicos necesarios que hacen posible la realización de tareas específicas, en contraposición a los componentes físicos, que son llamados hardware..

Transacciones: Una transacción es una unidad de la ejecución de un programa que accede y, posiblemente, actualiza varios elementos de datos. Una Transacción está delimitada por instrucciones de inicio y fin (la transacción consiste en todas las operaciones que se ejecutan entre inicio y fin de la misma).