Universidad de las Ciencias Informáticas

Facultad 6

Título: Sistema de Información de Gobierno. Mercado de datos Ciencia e innovación tecnológica.

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

Autores:

Dainelis Cespedes Maceo

Maria Lisandra Gonzalez Hurtado

Tutores:

Ing. Yuneimy Tellez Pérez

Ing. Rayko Emilio Torres Cruz

Ciudad de la Habana, Junio del 2011

“La inteligencia consiste no sólo en el conocimiento, sino también en la destreza de aplicar los conocimientos en la práctica”

Aristóteles

DECLARACIÓN DE AUTORÍA

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

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

Autores: _____ _____
Maria Lisandra Gonzalez Hurtado Dainelis Cespedes Maceo

Tutores: _____ _____
Ing. Yuneimy Tellez Pérez Ing. Rayko Emilio Torres Cruz

Datos de Contacto

Ing. Yuneimy Tellez Pérez

Correo electrónico: ytellez@uci.cu

Ing. Rayko Emilio Torres Cruz

Correo electrónico: retorres@uci.cu

DEDICATORIA

Dedico este trabajo:

A mi mamita linda

A mis hermanas

A mis abuelos

A mi esposo

A todas las personas que confiaron en mí y me apoyaron

Muchas gracias.

Dainelis Cespedes Maceo

Dedico esta tesis:

A mis padres Salvador y María Magdalena.

A mis hermanos.

A mis primas y primos.

A mis sobrinos.

A mi novio Luis Mariano.

A mi familia y amigos en general que confiaron y me apoyaron en los momentos difíciles.

Maria Lisandra Gonzalez Hurtado

AGRADECIMIENTOS

Primero agradecerle a la Revolución y a Fidel por estar en una escuela como esta, a Dios y a la Virgen de la Caridad del Cobre por darme fuerzas en los momentos malos y buenos y lograr mi objetivo.

A mi compañera de tesis que siempre estuvo presente y ayudándonos para obtener este título.

A mi mamá que es la persona más importante en mi vida y no sé de qué forma agradecerle todo lo que ha hecho por mí, por brindarme apoyo, por preocuparse por mí en todo momento, gracias a ti hoy me hago Ing.

A mi papá, que no está físicamente pero lo llevo siempre presente y está conmigo en todas partes, él se sentiría orgulloso de mí, siempre quiso que fuese la persona que soy.

A mis hermanas y a mis abuelos por siempre estar preocupados y darme consejos.

A mi esposo que lo quiero y adoro cantidad, ha sido algo especial en mi vida, por ayudarme y apoyarme, te amo.

A Yusimi, Lurdes, Odelaisis y Katia que han sido para mí como una madre y siempre me apoyan, las quiero.

A Patricia y Laritza por ser tan buenas compañeras de cuarto y estar en cada momento junto a mí, gracias Patry fuiste quien me hizo engordar y a la vez desgraciarme la vida y a Lary por ser tan amistosa, a todas las compañeras y compañeros de apto, de aula y del proyecto que han compartido en todos estos años, a las personas que siempre me preguntaban cómo iba en la tesis.

A todos los profes que me han enseñado y han compartido conmigo a lo largo de mi carrera, a Yuraisy, a Téllez, a Esley que siempre han estado disponible cuando los necesitaba.

A una amiga y amigo que los quiero mucho Yaneisis y Yorlandis gracias por todo.

A Ana Niuska que aunque siempre vivimos fajadas la quiero.

A los tutores por su dedicación y nos ayudaron mucho, al tribunal que me hicieron hasta llorar.
A dos personas que son como hermanas para mi Ana G y Yadi, de corazón nunca las olvidaré, estuvieron conmigo en momentos malos y buenos, hoy y siempre pueden contar conmigo.
Gracias a todos.

Dainelis Cespedes Maceo

AGRADECIMIENTOS

Quiero agradecer a todas las personas que siempre estuvieron conmigo luchando y batallando día a día.

A mis padres que han sido lo mejor del mundo.

A mi papá que siempre ha estado presente para lo que necesite, por sentirse orgulloso de mi y por brindarme ese amor tan grande que me hace crecer cada día más.

A mi mamá por brindarme amor, dedicación, por enseñarme que siempre se sale adelante de la situación más difícil.

A ellos agradecerles por haber confiado en mí, ser lo más importante de mi vida y mi razón de ser.

A mi dúo de tesis que aunque siempre nos fajábamos la quiero como si fuera una hermana.

A mis hermanos Salvador, Roelmis, Yodanis, Yoandris y Yoelmis, por darme su amor, cariño y aguantarme todas mis malacrianzas y travesuras.

A mis primas Omara y Onelia y a mis primos Leudys y Yumary, por estar todo este tiempo presentes ayudándome en todo lo que me hacía falta y quererme tanto como ellos sólo saben hacer.

A mi familia en general por ser como son conmigo, a Mary, Lidia, Cira, Reina, Evaristo, Esteban, Julia, Panchón, Firi, Joaquina, Cary, Baby, Virgen, Tatún, Omaida, Yolvis, Marilú, Elizabeth, Yailín, Neilín, Mailín, Odalis, Mikito, Maikel, Liuvin, Virgen, Melkis, Juana, Shari, Ida, Liudmila, mis padrinos, en fin a toda esta familia tan grande y linda que me quiere y me mima.

A mi tía Melba, que aunque ya no está físicamente con nosotros fue como una madre para mí.

A mi novio Luisy que ha estado aguantándome mis travesuras y yo las suyas, por demostrarme que siempre se puede y brindarme mucho amor.

A mis hermanas postizas Yarena y Leyanis por preocuparse siempre por mí.

A Laritza y Yulier por ser como mis hermanitos, consejeros y confidentes en todo este tiempo.

A todas mis compañeras de cuarto que hemos estado juntas en las buenas y en las malas Aleyda, Yudelkis, Susy, Laydi y Maryin.

A todos mis vecinos por demostrar su gran interés en mi formación profesional.

A mis tutores que se han portado de una manera extraordinaria y han luchado cada día con nosotras al igual que el tribunal y todos mis profesores que de una forma u otra ayudaron en mi formación.

A mis tíos Juan B y Orlando que me enseñaron, aunque no lo crean, el verdadero valor de la Revolución que gracias a ella me puedo graduar hoy de ingeniera en informática.

Maria Lisandra Gonzalez Hurtado

RESUMEN

La Oficina Nacional de Estadísticas es la institución que se encarga de centralizar toda la información de los distintos sectores del país. En ella los datos están almacenados en formato de difícil acceso para su consulta y no cuenta con un sistema informático que permita gestionar la información y agilizar la toma de decisiones. En el presente trabajo se muestra la implementación de un Mercado de Datos para el área de Ciencia e innovación tecnológica, donde se utilizó una adaptación de la metodología Ralph Kimball y lo planteado en la Tesis de Doctorado de Leopoldo Zenaido Zepeda Sánchez.. Apoyándose en la herramienta Visual Paradigm para el modelado y PostgreSQL como Sistema Gestor de Base de Datos. Para los procesos de Extracción, Transformación y Carga el Pentaho Data Integration y el DataCleaner y para los procesos de Inteligencia de Negocio el Pentaho Schema Workbench, Pentaho BI Server, Mondrian y el Apache Tomcat. Este Mercado de Datos contribuye a que el trabajo de los especialistas sea más eficiente en el proceso para la toma de decisiones, por lo que posibilita una mejor organización en esta entidad sobre los datos que se manejan.

Palabras clave: ONE, Mercado de Datos, BI, ETL.

TABLA DE CONTENIDO

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

Introducción

Actualmente el desarrollo humano no podría existir sin los adelantos de la ciencia y la tecnología. Prácticamente ninguna actividad escapa a su influencia, incluso a su carácter determinante. Es realmente asombrosa la dependencia que la civilización actual tiene de la tecnología [2]. La ciencia y la innovación tecnológica son factores determinantes en el desarrollo económico y social para elevar la calidad de vida de la población.

En Cuba, el impacto social de la ciencia y la tecnología constituye un tema de actualidad y de particular interés, el desarrollo de esta actividad tiene como objetivo principal a la sociedad y por ende, el propio hombre. El país está en vías de desarrollo, en un profundo y novedoso proceso de transformaciones educacionales y sociales como los programas de la batalla de ideas. En estas circunstancias surge la Universidad de las Ciencias Informáticas (UCI), la misma cuenta con muchos proyectos, que en su mayoría aportan considerables ingresos a la economía del país, investigando, produciendo software, servicios informáticos para la sociedad cubana y para el mundo.

En la universidad se creó el Centro de Tecnologías de Almacenamiento de Datos (DATEC), el cual se dedica al análisis de los datos y al desarrollo de aplicaciones que automatizan el control estadístico de cualquier institución. Actualmente brinda apoyo a algunas instituciones del país, entre la que se destaca la Oficina Nacional de Estadísticas (ONE).

Esta institución tiene como misión garantizar la producción de estadísticas de calidad a través del Sistema Estadístico Nacional; ejerciendo una adecuada dirección, ejecución y control de la captación de las cifras económicas y sociales, así como su adecuada difusión de acuerdo con las necesidades de la economía y las demás necesidades del país en información estadística. Es la encargada de centralizar, emitir, organizar y aprobar las estadísticas destinadas a satisfacer los requerimientos informativos y los compromisos de los órganos de dirección del gobierno en los territorios [1].

Los sistemas de gestión que se utilizan actualmente no permiten realizar consultas a la información de manera eficiente ya que solo brindan la posibilidad de hacer reportes estáticos. Los datos no están integrados y atentan contra la calidad de estos. Están recopilados en formato de difícil acceso y solo pueden ser consultados por un especialista de la informática y de la información con alto conocimiento del negocio. Se generan ficheros anuales que deben ser procesados para obtener información y no existe una aplicación informática que brinde reportes flexibles con información actualizada. Estos elementos influyen negativamente en el proceso estadístico de la ONE para la ayuda a la toma de decisiones del país. Anteriormente la ONE en conjunto con DATEC realizó el análisis y diseño del

mercado Ciencia e innovación tecnológica, sin embargo en la actualidad han surgido nuevas necesidades, por lo que se hizo necesario realizar el refinamiento de esta área debido a los problemas existentes ya mencionados.

A partir de esta situación surge el siguiente **problema de la investigación**: ¿Cómo contribuir a la toma de decisiones en el área de Ciencia e innovación tecnológica del Sistema de Información de Gobierno de la Oficina Nacional de Estadísticas?

Se define como **objeto de estudio**: Los Almacenes de Datos (AD) y como **campo de acción**: Mercados de Datos (MD) para el área de Ciencia e innovación tecnológica del Sistema de Información de Gobierno de la Oficina Nacional de Estadísticas.

Para darle cumplimiento al problema expuesto se determina como **objetivo general:** Desarrollar el MD Ciencia e innovación tecnológica del Sistema de Información de Gobierno que contribuya a la toma de decisiones en la ONE.

De este objetivo se derivan los siguientes **objetivos específicos:**

- ✓ Refinar el análisis y diseño del Mercado de Datos.
- ✓ Implementar el Mercado de Datos.
- ✓ Validar el Mercado de Datos.

Posibles resultados:

Mercado de Datos Ciencia e innovación tecnológica.

Tareas de la investigación:

- ✓ Revisión bibliográfica de los sistemas desarrollados para el área Ciencia e innovación tecnológica.
- ✓ Selección de las herramientas, tecnologías y metodologías para el desarrollo del MD.
- ✓ Refinamiento de los requisitos del MD.
- ✓ Refinamiento de la descripción de los casos de uso del MD.
- ✓ Refinamiento de la definición de las dimensiones, los hechos y las medidas del MD.

- ✓ Refinamiento del desarrollo de la matriz dimensional.
- ✓ Refinamiento de la estructuración del modelo de datos.
- ✓ Definición de la arquitectura de integración de los datos.
- ✓ Definición de la arquitectura de información.
- ✓ Diseño de los procesos de integración de los datos.
- ✓ Implementación del modelo de datos.
- ✓ Implementación de los flujos de las transformaciones y los trabajos.
- ✓ Implementación de los cubos multidimensionales.
- ✓ Implementación de los reportes candidatos.
- ✓ Implementación de los niveles de acceso de los usuarios.
- ✓ Validación de la solución obtenida.

Estructura del trabajo

Capítulo 1: Fundamentación teórica

En este capítulo se ofrece la fundamentación teórica, donde se definen conceptos fundamentales para un buen desarrollo de un MD, así como las herramientas, metodologías y tecnologías seleccionadas que van a permitir una adecuada solución al problema.

Capítulo 2: Análisis y diseño del Mercado de Datos

En este capítulo se realiza el análisis y diseño. Se definen los requisitos funcionales, no funcionales y de información, el modelo de datos, entre otros artefactos que ayudan a entender las necesidades de información.

Capítulo 3: Implementación del Mercado de Datos

En este capítulo se realiza la implementación del MD, donde se le dará solución a los requisitos del sistema.

Capítulo 4: Validación y pruebas del Mercado de Datos

En este capítulo se realiza la validación y pruebas de la solución propuesta a través de las listas de chequeo, para así evaluar los resultados obtenidos.

Capítulo 1. Fundamentación teórica

1.1 Introducción

En el capítulo se realiza un estudio del estado del arte acerca de los indicadores relacionados con la Ciencia e innovación tecnológica. Se definen conceptos fundamentales para el buen desarrollo de un MD, así como las herramientas, metodologías y tecnologías seleccionadas que van a permitir una adecuada solución al problema.

1.2 Ciencia e innovación tecnológica

La ONE cuenta con modelos estadísticos donde se recoge información de todos los sectores de la economía y la sociedad cubana, dentro de estos modelos se encuentra el que posee información del indicador de Ciencia e innovación tecnológica.

Esta área se encuentra en el departamento de Estadísticas Sociales y recopila información brindada por el Ministerio de Ciencia, Tecnología y Medio Ambiente como organismo rector, siendo reportada por todos los centros que independientemente de la esfera en que se desarrolle su actividad principal, realizan investigaciones u otras actividades científicas y tecnológicas.

El modelo estadístico 1001-00 es donde se recoge información del indicador de Ciencia e innovación tecnológica, en el cual se conoce la cantidad de innovaciones y racionalizaciones que son presentadas, concedidas y aplicadas por los centros durante el año que se informa. También se conoce el efecto económico obtenido [2].

1.3 Soluciones existentes en el mundo

En el mundo existen diversas empresas que se dieron a la tarea de implementar la tecnología de almacenamiento de datos para convertirlos en información útil y confiable, muchas de estas compañías son: El Instituto Nacional de Estadística Geográfica e Informática de México que tiene la responsabilidad de coordinar los Sistemas Nacionales Estadístico y de Información Geográfica, además de promover y orientar el desarrollo informático en el país. También se destaca el Instituto Nacional de Estadística de España que actualmente está utilizando la tecnología, el Instituto Nacional de Estadísticas de Chile que posee un AD implementado con varios MD que contienen la información acerca de los empleos, la industria, minería, las ventas, entre otros indicadores y el Instituto Nacional de Estadística en Venezuela, el cual está desarrollando un Sistema de Información Estadística de Ciencia y Tecnología con el propósito de poseer mecanismos que posibiliten la integración de fuentes de datos diversas, dando la posibilidad de recodificar dichos datos de forma integral.

1.3.1 Soluciones existentes en Cuba

En Cuba se encuentran entidades que han implementado varios AD para la gestión de información algunos de ellos son: la Corporación CIMEX, la compañía de Unión Cuba Petróleo (CUPET), la Corporación COPEXTEL, en el XIII Concurso Nacional de Computación y en la Feria de Informática del 2002 se presentó un AD para CUBACEL. Otra organización es la ONE, donde se lleva el control estadístico y se encuentra en el proceso de diseño e implementación de sus respectivas áreas.

1.4 Tecnologías de almacenamiento de datos

Una de las tecnologías más extensas son las Bases de Datos (BD), estas evolucionan y seguirán evolucionando para satisfacer las nuevas necesidades sobre el tratamiento de datos que vayan surgiendo. En ellas se pueden solucionar mucho más aspectos, trabajar en tiempo real de forma más sencilla, eficaz y segura.

1.4.1 Base de Datos

Las BD se han convertido en un elemento imprescindible para manejar diversos volúmenes de información.

Definición de Base de Datos

Es una serie de datos organizados y relacionados entre sí, los cuales son recolectados y explotados por los sistemas de información de una empresa o negocio en particular [3].

Características de las Base de Datos

- ✓ Redundancia mínima.
- ✓ Acceso concurrente por parte de múltiples usuarios.
- ✓ Integridad de los datos.
- ✓ Respaldo y recuperación.
- ✓ Acceso a través de lenguajes de programación estándar.

1.4.2 Almacenes de Datos

Constituyen uno de los soportes fundamentales para el proceso de toma de decisiones, los cuales tienen gran importancia porque la información en ellos es confiable y con calidad. Se han convertido en una potente herramienta para la recuperación efectiva de las más complejas consultas.

Existen diversas definiciones de AD, a continuación se muestran algunas:

Un AD es una gran colección de datos que recoge información de múltiples sistemas fuentes u operacionales dispersos, y cuya actividad se centra en la toma de decisiones [4].

Según W.H. Inmon:

"Un conjunto de datos orientado a temas, integrado, no volátil, variante en el tiempo, como soporte para la toma de decisiones" [5].

Según Ralph Kimball

"…Los AD son una copia de los datos de la transacción estructurados específicamente para la pregunta y el análisis" [6].

Es una BD corporativa que se caracteriza por integrar y depurar información de una o más fuentes distintas, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta.

El término AD fue acuñado por primera vez por Bill Inmon, según este se caracteriza por ser:

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

Temático: Sólo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional. Los datos se organizan por temas para facilitar su acceso y entendimiento por parte de los usuarios finales.

Histórico: El tiempo es parte implícito de la información contenida en un AD. En los sistemas operacionales, los datos siempre reflejan el estado de la actividad del negocio en el momento presente. Por el contrario, la información almacenada sirve, entre otras cosas, para realizar análisis de

tendencias. Por lo tanto, el AD se carga con los distintos valores que toma una variable en el tiempo para permitir comparaciones.

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

Otra característica es que contiene metadatos, es decir, datos sobre los datos. Los metadatos permiten saber la procedencia de la información, su fiabilidad, forma de cálculo, entre otros. Los metadatos serán los que permiten simplificar y automatizar la obtención de la información desde los sistemas operacionales a los sistemas informacionales [7].

Ventajas del uso de Almacén de Datos

- ✓ La inversión que realiza una organización para una correcta implantación conlleva un coste muy elevado, sin embargo el retorno de la inversión es garantizado en gran medida.
- ✓ Como consecuencia de la ventaja anterior se puede alcanzar una ventaja competitiva debido a una buena toma de decisiones.
- ✓ Mejoran la productividad de los responsables en la toma de decisiones de la organización debido a que:
  - Hacen más fácil el acceso a una gran variedad de datos.
  - Se obtiene una BD histórica y clasificada por temas.
  - Integración de información procedente de múltiples sistemas externos [4].

Desventaja del uso de un Almacén de Datos

- ✓ La subestimación del tiempo requerido para extraer, limpiar y cargar los datos en el almacén.
- ✓ Problemas con los sistemas de origen de los datos.
- ✓ La construcción de un AD puede requerir de mucho tiempo [4].

Componentes de un Almacén de Datos

Están compuestos por una serie de 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 y la relación que existe en el diseño entre ellos según la metodología propuesta por Kimball.

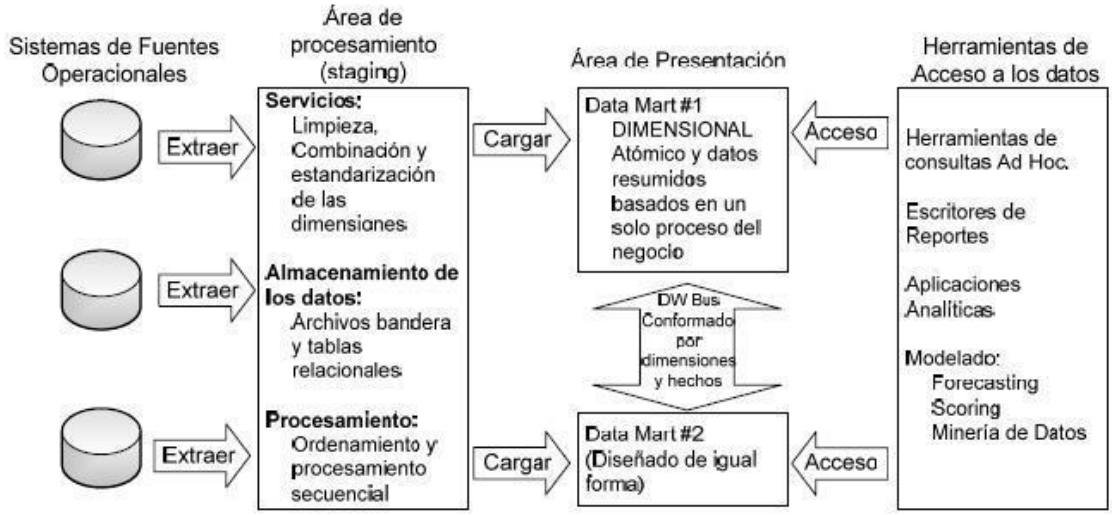


Ilustración 1: Relación entre los componentes de un AD.

Estos son los sistemas que poseen las compañías o empresas para la gestión de sus transacciones diarias. Las transacciones son almacenadas en los más diversos formatos, desde una BD relacional hasta cualquier tipo de ficheros, ya sea Excel, XML, DBF, texto plano, entre otros. Se encuentran localizados fuera del repositorio debido a que se tiene poco o ningún control sobre el volumen y formato de los datos de las 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 período corto de tiempo para hacer las recuperaciones de forma más óptima. También existe la posibilidad de que sean fuentes creadas anualmente debido a que no posean un sistema que las procese.

Área de procesamiento

Es el área que almacena los datos temporalmente y realiza un conjunto de procesos comúnmente llamados Extracción, Transformación y Carga (ETL, por sus siglas en inglés). Realiza la función de interfaz entre las fuentes operacionales y el área de presentación.

En esta área es donde se invierte la mayor cantidad de tiempo y esfuerzo durante el desarrollo del almacén. 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 la información en BD relacionales, generalmente, para realizar el análisis y procesamiento de los datos. Una vez los datos estén almacenados se procede a su limpieza donde se detectan inconsistencias, duplicaciones, errores de formato e inexistencias, estandarizándose la información almacenada en diferentes fuentes.

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 la de 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 de minería de datos[1] [8].

1.4.3 Mercados de Datos

Es una BD departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Puede ser alimentado desde los datos de un AD, o integrar por sí mismo un compendio de distintas fuentes de información [9].

Ventajas de usar un Mercado de Datos

[1]Conocido como *Data Mining* por sus siglas en inglés. Se emplea para el descubrimiento de conocimiento: es un proceso de búsqueda, a partir de los datos, de conocimientos nuevos y no anticipados.

- ✓ Son simples de implementar.
- ✓ Conllevan poco tiempo de construcción y puesta en marcha.
- ✓ Permiten manejar información confidencial.
- ✓ Reflejan rápidamente sus beneficios y cualidades.
- ✓ Reducen la demanda del depósito de datos.

Desventajas de usar un Mercado de Datos

- ✓ Se pierde capacidad de procesamiento debido al crecimiento de los datos.
- ✓ Los usuarios necesitan acceder a varios MD.
- ✓ Dificultad para construir debido al corto plazo de desarrollo.

1.4.4 Justificación de la tecnología de almacenamiento a utilizar

Los MD se han hecho necesarios para el manejo de la información en distintas organizaciones y empresas a nivel mundial, asegurando una gestión más eficiente y proyectando una visión única de la empresa. Son subconjuntos del AD los cuales integran un número de diversas fuentes y analizan la información desde disímiles perspectivas.

1.5 Metodologías para el desarrollo

Metodología

Una metodología es una guía que se sigue a fin de realizar las acciones propias de una investigación. En términos más sencillos, se trata de la guía que va indicando qué hacer y cómo actuar cuando se quiere obtener algún tipo de investigación. Es posible definir una metodología como aquel enfoque que permite observar un problema de una forma total, sistemática y disciplinada.

En dependencia de la problemática presentada es la política de selección para una u otra metodología. Según Bill Inmon, el AD es una parte del todo que conforma un sistema de inteligencia. Una entidad tiene un AD y los MD tienen como fuente de información ese almacén. Esta tendencia es conocida como “Top-Down”. Para Ralph Kimball el AD se compone por el conjunto de MD que existan en la entidad donde esté implementado y la información se modela en un modelo dimensional. Esta tendencia es conocida como “Bottom-Up” [10].

La Metodología de Inmon 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 construidos descendentemente los MD, se nutren del AD, convirtiéndose este en un complejo empresarial de bases de datos relacionales.

Kimball a diferencia de Inmon propone la elaboración del AD a través de la construcción de los MD, crea los conceptos hechos y dimensiones lo cual proporciona agilidad en el proceso de toma de decisiones. Existe una gran documentación acerca de esta lo que proporciona facilidades al equipo de desarrollo. Es una metodología madura donde se definen etapas, actividades, artefactos y roles.

La metodología propuesta es una adaptación de la metodología Kimball la cual se basa en dividir el mundo de BI entre los hechos y las dimensiones, es eficaz y conduce a una solución completa en un corto período de tiempo, se tomaron sus cuatro fases fundamentales: análisis, diseño, ETL y BI. Se le realizaron algunas modificaciones en correspondencia a las características del centro. Para la selección de la misma se utilizaron los siguientes criterios:

- ✓ Desarrollo iterativo incremental, donde se construye una pieza a la vez del MD a diferencia de Inmon.
- ✓ Mayor velocidad de respuesta al cliente.
- ✓ Comprensible para el usuario.
- ✓ Resistente y preparado para cambios.
- ✓ Documentación.
- ✓ Respaldo y soporte.

Como complemento a la metodología Kimball y fortaleciendo la etapa del levantamiento de requisitos, se tomó lo planteado por Leopoldo Zenaido Zepeda Sánchez en su Tesis de Doctorado, orientando así el trabajo a los casos de uso para guiar el proceso de desarrollo.

Dentro del ciclo de vida de esta metodología existen una serie de flujos de trabajo mencionados a continuación:

Estudio preliminar o planeación: Se realiza un estudio minucioso en la entidad cliente. Esto incluye un diagnóstico de información, de datos y de infraestructura tecnológica.

Requerimientos: Es llevado a cabo por el grupo de análisis, se realiza en dos direcciones, una, identificando las necesidades de información y reglas de negocio; y la otra haciendo un levantamiento detallado de cada una de las distintas fuentes de datos a integrar.

Arquitectura y diseño: Aquí participan los tres grupos fundamentales: ETL, AD e Inteligencia de Negocio (BI). En la definición de la arquitectura participan los arquitectos de cada uno de los grupos mencionados y en el diseño participan igualmente los tres grupos pero se incrementa considerablemente la cantidad de personas, todo depende de la complejidad de la solución.

Implementación: Participan los tres grupos de desarrollo (ETL, AD y BI). Se lleva a cabo el diseño físico del repositorio de datos, se crean las estructuras de almacenamiento con las particiones y agregaciones correspondientes según la solución en desarrollo.

Prueba: Se realizan varias pruebas, comenzando por las pruebas de unidad, llevadas a cabo por los propios desarrolladores de cada uno de los grupos, luego las pruebas de integración y sistema, hasta las pruebas de aceptación con el cliente final.

Despliegue: Este flujo consta de dos etapas, la primera es un despliegue piloto, donde se configuran los servidores necesarios y se instalan las herramientas según la arquitectura definida, se carga una muestra de los datos en un ambiente controlado, con el fin de demostrarle al cliente final que la solución funciona.

Soporte y mantenimiento: Pueden ser desde un soporte en línea, vía telefónica, web, correo u otras, hasta el acompañamiento junto al cliente.

Gestión y administración del proyecto: Es llevado a cabo por el grupo de dirección del proyecto, el cual gestiona y chequea todo el desarrollo, los gastos, las utilidades, los recursos, las adquisiciones, y demás actividades relacionadas con la gestión de proyecto.

Como se ha descrito, en cada flujo intervienen grupos específicos, cada uno con actividades y responsabilidades concretas, a continuación se mencionan los mismos:

- ✓ Grupo de análisis.
- ✓ Grupo de AD.
- ✓ Grupo de ETL.
- ✓ Grupo de BI.

- ✓ Grupo de dirección.

1.6 Modelos 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

- ✓ 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 [11].

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. Expresan entidades relevantes para un sistema de información, sus inter-relaciones y propiedades.

El modelo está compuesto por:

- ✓ Entidades
- ✓ Atributos
- ✓ Relaciones
- ✓ Cardinalidad
- ✓ Llaves

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 éstos van a estar compuestos por hechos, medidas y dimensiones [12].

Tipos de modelado de un Almacén de Datos

Esquema estrella

El esquema estrella está formado por una tabla de hecho 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.

Esquema en copo de nieves

Es una variante del esquema de estrella en el que las tablas dimensionales de este último se organizan jerárquicamente mediante su normalización.

Constelación de hechos

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

Hechos

Un hecho es una colección de medidas relacionadas con sus dimensiones, representadas por las llaves dimensión. Puede representar un objeto de negocio 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. 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[2], 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.

[2] Subconjunto de datos de un AD que es almacenado en una estructura multidimensional, contiene los valores agregados de todos los niveles de todas las dimensiones y presenta los datos de interés para el usuario.

- ✓ 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.

1.7 Herramienta de modelado

Las herramientas de modelado dan la posibilidad de realizar el diseño y una mejor construcción de la BD del nivel lógico a un nivel físico. Se encargan de plasmar los requerimientos del sistema y la funcionalidad.

La herramienta de modelado propuesta por DATEC es Visual Paradigm, porque es una herramienta de diseño que soporta todos los diagramas UML, esquemas y diagramas de entidad-relación y la universidad paga por su licencia debido a su uso y popularidad.

Es 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 ayuda a una más rápida construcción de aplicaciones de calidad y a un menor coste. Permite dibujar todos los tipos de diagramas de clases, código inverso, generar código desde diagramas y generar documentación. También proporciona tutoriales y proyectos UML [13].

Esta herramienta acelera el desarrollo de aplicaciones, sirve de puente visual entre arquitectos, analistas y diseñadores de sistemas de información, haciendo el trabajo más fácil y dinámico.

Ventajas

- ✓ La navegación intuitiva entre el código y el modelo visual.
- ✓ Poderoso generador de informes PDF/HTML.
- ✓ Tiempo real en la demanda.
- ✓ Un ambiente modelador visual superior.
- ✓ Sofisticado diseño de diagramas [14].

1.8 Gestor de Base de Datos

Los Sistema de Gestión de Base de Datos (SGBD) son un tipo de software muy específico, dedicados a servir de interfaz entre la BD, el usuario y las aplicaciones que lo utilizan. Se compone de lenguajes de definición, manipulación, consulta y seguridad de datos. El propósito general de los SGBD es el de manejar de manera clara, sencilla y ordenada un conjunto de datos.

El SGBD propuesto por DATEC es el PostgreSql 8.4, posee todos los requisitos necesarios para la implementación eficiente del MD. El código fuente está disponible bajo los más liberales términos de licencia de código abierto: la licencia BSD (Berkeley Software Distribution), por tanto pueden hacerse todas las modificaciones, mejoras o cambios que se estimen convenientes. Es un sistema de BD utilizado por miles de organizaciones alrededor del mundo, considerada muchas veces como la mejor opción de código abierto [16].

Características

- ✓ Multiplataforma.
- ✓ Permite la gestión de diferentes usuarios como también los permisos asignados a cada unos de ellos.

1.9 Modos de almacenamiento de datos

Los sistemas OLAP[3] son BD orientadas al procesamiento analítico. Este análisis suele implicar generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil. En este sistema:

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

[3]Por sus siglas en inglés On-Line Analytical Processing.

Con OLAP se puede ver un conjunto de datos del negocio de muchas y diversas formas sin mucho esfuerzo. Los archivos OLAP o cubos modelan los datos en dimensiones [7].

Características

- ✓ Permite recolectar y organizar la información analítica necesaria para los usuarios y disponer de ella en diversos formatos, tales como tablas, gráficos, reportes, tableros de control, entre otros.
- ✓ Soporta análisis complejos de grandes volúmenes de datos.
- ✓ Complementa las actividades de otras herramientas que requieran procesamiento analítico en línea.
- ✓ Presenta al usuario una visión multidimensional de los datos para cada tema de interés del negocio [15].

Existen tres modelos para el proceso analítico en línea de la información ROLAP[4], MOLAP[5] y HOLAP[6]. El proceso de análisis se realiza de igual forma lo que varía en uno y otro caso es la metodología de almacenamiento, esta influye en la velocidad de recuperación de la información.

A través de este tipo de herramienta, se puede analizar el negocio desde diferentes escenarios históricos y proyectar como se ha venido comportando y evolucionando en un ambiente multidimensional, o sea, mediante la combinación de diferentes perspectivas, temas de interés o dimensiones.

1.9.1 Procesamiento Analítico en Línea Relacional (ROLAP)

La arquitectura ROLAP, accede a los datos almacenados en un AD para proporcionar los análisis OLAP. La premisa de los sistemas es que las capacidades OLAP se soportan mejor contra las BD relacionales. La BD relacional maneja los requerimientos de almacenamiento de datos, y el motor ROLAP proporciona la funcionalidad analítica.

[4] Por sus siglas en inglés Proceso Analítico en Línea Relacional.
[5] Por sus siglas en inglés Proceso Analítico en Línea Multidimensional.
[6] Por sus siglas en inglés Proceso Analítico en Línea Híbrido.

El nivel de aplicación es el motor que ejecuta las consultas multidimensionales de los usuarios. Este motor se integra con niveles de presentación, a través de los cuales los usuarios realizan los análisis OLAP [7].

1.9.2 Procesamiento Analítico en Línea Multidimensional (MOLAP)

La arquitectura MOLAP usa BD multidimensionales para proporcionar el análisis, su principal premisa es que el OLAP está mejor implantado almacenando los datos multidimensionalmente. Este sistema usa una BD propietaria multidimensional, donde la información se almacena multidimensionalmente para ser visualizada en varias dimensiones de análisis.

Utiliza una arquitectura de dos niveles: las BD multidimensionales y el motor analítico. La BD multidimensional es la encargada del manejo, acceso y obtención del dato [7].

1.9.3 Procesamiento Analítico en Línea Híbrido (HOLAP)

Un desarrollo un poco más reciente ha sido la solución HOLAP, la cual combina las arquitecturas ROLAP y MOLAP para brindar una solución con las mejores características de ambas: desempeño superior y gran escalabilidad. Un tipo de este motor mantiene los registros de detalle (los volúmenes más grandes) en la BD relacional, mientras que mantiene las agregaciones en un almacén MOLAP separado [7].

1.9.4 Ventajas y desventajas

	Ventajas	Desventajas
MOLAP	Mejor rendimiento en los tiempos de respuesta.	Duplica el almacenamiento de datos (ocupa más espacio).
ROLAP	Ahorra espacio de almacenamiento. Útil cuando se trabaja con muy grandes conjuntos de datos.	El tiempo de respuesta a consultas es mayor.
HOLAP	Buen tiempo de respuesta sólo para información resumida.	Volúmenes de datos más grandes en la BD relacional.

Tabla 1: Ventajas y desventajas.

1.10 Herramientas para el proceso Extracción, Transformación y Carga

Las herramientas para el proceso de ETL fueron las propuestas por DATEC. A continuación se mencionan:

DataCleaner

Es una aplicación código abierto para el perfilado, la validación y comparación de datos. Estas actividades ayudan a administrar y supervisar la calidad de los datos con el fin de garantizar que la información sea útil y aplicable a su situación de negocio.

Según su creador Kasper Sorense el sistema requiere Java Runtime Enviroment 5.0 o una versión superior. La misma permite la evaluación del nivel de calidad de los datos contenidos en el sistema de información. Es una aplicación muy fácil de usar, genera sofisticados informes y gráficos que permiten a los usuarios determinar de un vistazo el nivel de calidad de los datos, identificar y analizar la estructura del origen de datos y combinar resultados y gráficos, creando vistas fáciles de interpretar para evaluar la calidad de los datos [17].

Características

- ✓ Los perfiles de datos se utilizan para calcular y analizar diversas medidas importantes basadas en los valores de los datos.
- ✓ Validación de datos: el validador le dará un resultado que puede ser interpretado como bueno o malo, ya que el mismo valida los datos.
- ✓ Análisis del modelo.
- ✓ Soporta acceso de lectura a muchos tipos de AD:
  - BD compatibles con JDBC (oficialmente probadas y compatibles: Oracle, MySql, PostgreSql, Firebird, SQLite, HSQLDB, Derby / javadb).
  - Excel (.xls) hojas de cálculo.
  - Archivos XML.
  - Open Office Base (ODB) archivos.

Pentaho Data Integration

Pentaho Data Integration (PDI, también llamada Kettle) es el componente responsable de los procesos ETL. Aunque las herramientas ETL son las más usadas en entornos de AD, Kettle también se puede utilizar para otros fines:

- ✓ Migración de datos entre las aplicaciones o BD.
- ✓ Exportación de datos desde BD para archivos planos.
- ✓ Carga de datos de forma masiva en BD.
- ✓ Limpieza de datos [18].

Características

- ✓ Cada proceso es creado con una herramienta gráfica donde se especifica qué se va hacer sin necesidad de escribir código que indique cómo hacerlo.
- ✓ Admite una amplia gama de formatos de entrada y salida, incluyendo archivos de texto, hojas de datos, archivos XML, propiedades de java y los motores de BD de código abierto y propietarios.
- ✓ Basado en repositorio, facilita la reutilización de componentes de transformación, colaboración y administración de modelos, conexiones, entre otros.

Ventajas

- ✓ Es una de las herramientas libres más antiguas que tiene una gran cantidad de usuarios y una nueva dirección por parte del soporte técnico de Pentaho.
- ✓ Los usuarios comparten muchos consejos y trucos.
- ✓ Multiplataforma.

Se utiliza ETL ya que 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, reformatearlos, limpiarlos y cargarlos en otra BD y MD. La idea es que una aplicación ETL lea los datos primarios de unas BD de sistemas principales, realice transformación, validación, el proceso cualitativo, filtración y al final escriba datos en el almacén.

Módulo encargado de:

- ✓ Obtener los datos de distintas BD.
- ✓ Unificar la información.
- ✓ Almacenar la información en un AD.

Debe garantizar:

- ✓ Integración de datos heterogéneos.
- ✓ Seguridad de la información transmitida [19].

Extracción

- ✓ Cada sistema separado puede usar una organización diferente de los datos.
- ✓ Los formatos de las fuentes normalmente se encuentran en BD relacionales o ficheros planos.
- ✓ La extracción convierte los datos a un formato preparado para iniciar el proceso de transformación.

Trasformación

- ✓ La fase de transformación aplica una serie de reglas de negocio o funciones sobre los datos extraídos para convertirlos en datos que serán cargados.
- ✓ Dividir una columna en varias.
- ✓ Calcular totales de múltiples filas de datos.
- ✓ Seleccionar sólo ciertas columnas para su carga.
- ✓ Unir datos de múltiples fuentes.

Carga

La fase de carga es el momento en el cual los datos de la fase anterior (transformación) son cargados en el sistema destino. Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes.

1.11 Herramientas para el proceso de Inteligencia del Negocio

La plataforma Open Source Pentaho Business Intelligence cubre muy amplias necesidades de análisis de los datos y de los informes empresariales. Las soluciones de Pentaho están escritas en java y tienen un ambiente de implementación también basado en java. Eso hace que Pentaho sea una solución muy flexible para cubrir una amplia gama de necesidades empresariales.

Las tecnologías de BI son herramientas de soporte de decisiones que permiten en tiempo real, acceso interactivo, análisis y manipulación de información crítica para la empresa. Intentan ayudar a las personas a entender los datos más rápido, a fin de que puedan tomar mejores y más rápidas decisiones y finalmente, mejorar sus movimientos hacia la consecución de objetivos de negocios. Los impulsores claves detrás de los objetivos de BI son incrementar la eficiencia organizacional y la efectividad. Algunas de las tecnologías de BI apuntan a crear un flujo de datos dentro de la organización más rápido y accesible.

A continuación se mencionan las herramientas definidas por DATEC:

Pentaho Schema Workbench

El esquema Mondrian Workbench es una interfaz de diseño que permite crear y probar esquemas cubo OLAP Mondrian visualmente. El motor de Mondrian procesa las solicitudes de MDX con el esquema ROLAP (OLAP Relacional). Estos archivos son los modelos de esquemas XML (lenguaje de marcas extensibles) de metadatos que se crean en una estructura específica utilizada por el motor de Mondrian. Los modelos XML se pueden considerar como las estructuras en forma de cubo que utilizan hechos existentes y tablas de dimensiones [20].

Pentaho BI Server

El BI Server de Pentaho es una aplicación que permite analizar todas las informaciones y administrar todos los recursos de BI. Cuenta con una interfaz de usuario donde se encuentran disponibles todos los informes, vistas OLAP y cuadros de mando, así como accesos a una consola de administración que permite crear roles y usuarios.

Ventajas

- ✓ Aplicación extensible, adaptable y configurable.
- ✓ Se integra con la mayoría de entornos y se puede comunicar con otras aplicaciones vía servicios web.
- ✓ Integra todos los recursos informacionales en una única plataforma de explotación.
- ✓ Proporciona mucha libertad al usuario y los desarrolladores para crear contenidos nuevos.

Desventajas

- ✓ Los cuadros de mandos son complejos de realizar y ofrecen poca flexibilidad.
- ✓ La traducción al español no es al 100% [21].

Mondrian

Mondrian es un motor Hybrid OLAP que combina la flexibilidad de los motores ROLAP con una caché que le proporciona velocidad. OLAP es la tecnología que permite organizar la información en una estructura dimensional que proporcionará la posibilidad de moverse por la información desplazándose por sus dimensiones.

Ventajas

- ✓ Es un motor ampliamente utilizado y consolidado en entornos java.
- ✓ Permite realizar consultas a MD.
- ✓ Posee alta velocidad de respuesta.

Desventajas

- ✓ No permite reescribir.

Apache Tomcat

Apache Tomcat funciona como un contenedor de servlets desarrollado bajo el proyecto Jakarta en la Apache Software Foundation. Tomcat implementa las especificaciones de los servlets y de JavaServer Pages (JSP) de Sun Microsystems. Dado que fue escrito en java, funciona en cualquier sistema operativo que disponga de una máquina virtual java. Es cada vez más utilizado por las empresas en los entornos de producción debido a su contrastada estabilidad. Es una implementación de software de código abierto de Java Servlet y tecnologías JavaServer Pages. Es desarrollado en un entorno abierto y participativo y publicado bajo la licencia Apache versión 2 [22].

1.12 Conclusiones del capítulo

Fueron analizados los principales conceptos relacionados con los AD para un mejor desarrollo del trabajo. Se decidió utilizar una adaptación de la metodología Kimball y como complemento lo planteado en la Tesis de Doctorado de Leopoldo Zenaido. Las herramientas a utilizar para los procesos de ETL y BI son las definidas por el centro, las cuales fueron seleccionadas por sus características y ventajas.

Capítulo 2: Análisis y Diseño del Mercado de Datos

2.1 Introducción

En este capítulo se describe el negocio, donde se realiza el análisis y diseño del MD. Se definen los requisitos funcionales, no funcionales, multidimensionales y de información, los casos de uso del sistema, las reglas del negocio, las necesidades de los usuarios, la matriz Bus y el modelo de datos.

2.2 Análisis de la solución

El análisis es el proceso donde se obtiene una visión del sistema. Se definen los requisitos de la organización, las reglas del negocio, las necesidades de información, entre otros; que van a conllevar a una aproximación del diseño.

2.2.1 Definición del negocio

La ONE es la institución rectora de las estadísticas en Cuba. Es la entidad encargada de recopilar, organizar y ejecutar toda la información estadística del país. Está compuesta por diversas direcciones como la de Estadísticas Sociales, la cual recoge toda la información relacionada con los indicadores y los temas de Ciencia e innovación tecnológica. Esta información se encuentra en el modelo 1001, donde se recogen datos de todas las empresas y unidades presupuestadas del país que durante el año hayan realizado actividades referidas a la ciencia e innovación tecnológica. Anteriormente se realizó el análisis y diseño encontrándose 2 requisitos de información, 5 tablas de dimensiones y 2 tablas de hechos.

2.2.2 Reglas del Negocio

Describen las políticas, normas, operaciones, definiciones y restricciones presentes en una organización y que son de vital importancia para alcanzar sus objetivos. Pueden estar formalmente definidas en manuales de procedimiento, contratos o acuerdos, se establecen algunas restricciones por los clientes para darle tratamiento a los datos. En el análisis se identificaron las siguientes reglas del negocio:

RN1[7]**:** El total de los trabajadores físicos en la actividad de ciencia y tecnología según nivel educacional se calculan de la siguiente forma: la suma de nivel superior, nivel medio, grados científicos otorgados y otros.

[7] RN: Reglas del Negocio.

RN2: El total de los trabajadores físicos en la actividad de ciencia y tecnología según categoría ocupacional está dado por la suma de los dirigentes, técnicos, administrativos, obreros, de servicios y del total de mujeres.

RN3: El gasto total en actividades de ciencia y tecnología por tipo de actividades es la suma de Investigación y desarrollo, otras actividades científicas y tecnológicas.

RN4: El total de los gastos corrientes en actividades de ciencia y tecnología por fuente de financiamiento está dado por la suma del presupuesto del estado, financiamiento empresarial y financiamiento externo.

RN5: El total de las inversiones ejecutadas en la actividad de ciencia y tecnología por componentes es la suma de construcción y montajes, equipos y otros.

RN6: El total de publicaciones seriadas de ciencia y tecnología va a estar dada por la suma de formato impreso, formato electrónico y ambos formatos.

RN7: El total de registro de patentes de invenciones presentadas en Cuba es la suma de las solicitudes extranjeras y las solicitudes nacionales.

RN8: El total de registro de patentes de modelos industriales presentadas en Cuba es la suma de las solicitudes extranjeras y las solicitudes nacionales.

RN9: El total de registro de patentes de invención por países va a estar dado por la suma de las patentes concedidas y solicitadas.

2.2.3 Necesidades de los usuarios

Las necesidades de los usuarios son de vital importancia porque así se pueden crear productos y servicios con éxito. Las necesidades de los clientes son diversas y es importante que se entienda que si se desea mejorar o crear valor, los usuarios evalúan el desempeño de los productos y servicios a través de sus características.

2.2.4 Requisitos de Información

Describen qué información debe almacenar el sistema para satisfacer las necesidades de clientes y usuarios. Identifican los conceptos relevantes sobre los que se debe almacenar información y los datos específicos que son de interés. Son los reportes que el cliente necesita visualizar.

A continuación se mencionan los requisitos de información:

RI[8]1: Obtener cantidad total de trabajadores físicos según categoría ocupacional.

RI2: Obtener cantidad total de trabajadores físicos según nivel educacional.

RI3: Obtener cantidad total de gastos corrientes por fuente de financiamiento.

RI4: Obtener cantidad total de gastos por tipo de actividad.

RI5: Obtener cantidad total de inversiones ejecutadas.

RI6: Obtener cantidad total de solicitudes de registro de patentes de invenciones.

RI7: Obtener cantidad total de solicitudes de registro de patentes de modelos industriales.

RI8: Obtener cantidad de patentes de invención por países.

RI9: Obtener cantidad de publicaciones impresas, electrónicas y en ambos formatos.

2.2.5 Requisitos Multidimensionales

Son las variables de entrada y de salida de la solución. Estos se definen a partir de los requisitos de información anteriormente descritos. La idea fundamental del modelo es que los datos del negocio puedan ser representados como un tipo de cubo de datos. Constituyen la entrada fundamental para el diseño de las estructuras del almacén.

RM[9]1: Obtener cantidad total de trabajadores físicos según categoría ocupacional.

VE: categoría ocupacional, año, división política administrativa y unidad de medida.

VS[10]: cantidad de trabajadores.

RM2: Obtener cantidad total de trabajadores físicos según y nivel educacional.

VE: nivel educacional, año, división política administrativa y unidad de medida.

VS: cantidad de trabajadores.

RM3: Obtener cantidad total de gastos corrientes por fuente de financiamiento.

VE[11]: gastos corrientes, año, división política administrativa y unidad de medida.

VS: gastos.

RM4: Obtener cantidad total de gastos por tipo de actividad.

[8] RI: Requisito de Información.
[9] RM: Requisito Multidimensional.
[10] VS: Variable de Salida.
[11] VE: Variable de Entrada.

VE: gasto total, año, división política administrativa y unidad de medida.

VS: gastos.

RM5: Obtener cantidad total de inversiones ejecutadas.

VE: componentes, año, división política administrativa y unidad de medida.

VS: inversiones ejecutadas.

RM6: Obtener cantidad total de solicitudes de registro de patentes de invenciones.

VE: solicitudes, año, división política administrativa y unidad de medida.

VS: cantidad de patentes de invenciones.

RM7: Obtener cantidad total de solicitudes de registro de patentes de modelos industriales.

VE: solicitudes, año, división política administrativa y unidad de medida.

VS: cantidad de modelos industriales.

RM8: Obtener cantidad de patentes de invención por países.

VE: país, estado legal, año y unidad de medida.

VS: cantidad de patentes de invención por países.

RM9: Obtener cantidad de publicaciones impresas, electrónicas y en ambos formatos.

VE: título, año, división política administrativa y unidad de medida.

VS: cantidad de publicaciones impresas, electrónicas y en ambos formatos.

2.2.6 Requisitos Funcionales

Son capacidades o condiciones que el sistema debe cumplir. Los requisitos funcionales permiten expresar específicamente las responsabilidades del sistema que se propone y permiten determinar de una manera clara las posibles respuestas del sistema. A continuación se mencionan:

RF[12]1: Autenticar usuario.

RF2: Adicionar usuario.

RF3: Eliminar usuario.

RF4: Adicionar rol.

RF5: Eliminar rol.

RF6: Adicionar reporte.

RF7: Eliminar reporte.

RF8: Modificar reporte.

RF9: Realizar extracción de los datos.

RF10: Realizar la transformación y carga de los datos.

RF11: Mostrar consulta MDX.

RF12: Realizar cruce de variables.

RF13: Mostrar gráfica.

RF14: Suprimir filas y columnas vacías.

RF15: Imprimir.

RF16: Exportar reporte como pdf.

RF17: Exportar reporte como excel.

2.2.7 Requisitos no funcionales

Los requisitos no funcionales son propiedades o cualidades que el producto debe tener. Describen aquellas características no funcionales que los clientes y usuarios desean que tenga el sistema a desarrollar, lo que permite hacer el producto usable, rápido y confiable.

Usabilidad

RNF1: Cumplir con las pautas de diseño de la interfaz.

El sistema debe tener una interfaz gráfica uniforme a través del mismo incluyendo pantallas, menús y opciones.

RNF2: Los textos que aparezcan en la interfaz del sistema deben ser en idioma español y en inglés.

[12] RF: Requisitos Funcionales.

Los textos que aparezcan deben de estar en idioma español e inglés y deben de ser los más explicativos posibles.

Fiabilidad

RNF5: Garantizar la persistencia de la información.

Se realizará un respaldo de los datos del almacén con frecuencia anual. Esta información se almacenará en el edificio correspondiente a la ONE y será responsabilidad del grupo de administración de redes de la misma.

Requisitos para la documentación de usuarios en línea y ayuda del sistema.

RNF8: Confeccionar manual de usuario.

El sistema debe estar acompañado de un documento que guiará la ejecución del usuario teniendo en cuenta cada funcionalidad.

2.2.8 Casos de uso del sistema

Son parte del análisis y ayudan a describir qué es lo que el sistema debe hacer. Describen un uso del sistema y cómo este interactúa con el usuario. Es un proceso que da un resultado de valor para un actor determinado y una secuencia de actividades a automatizar.

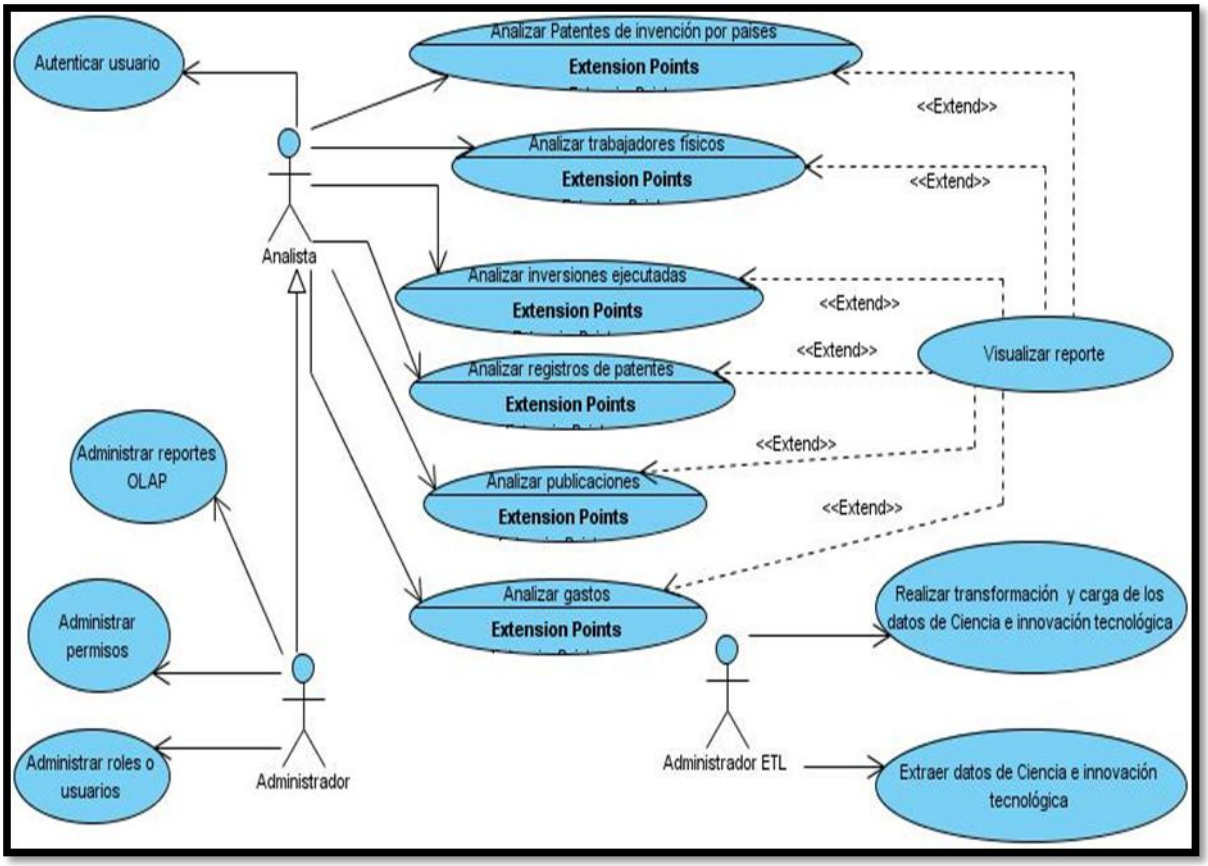


Ilustración 2: Caso de uso del sistema

A continuación se describen los actores del sistema.

Actores	Descripción
Analista	Es el encargado de analizar y consultar la información de los indicadores de ciencia y tecnología.
Administrador	Es el responsable de administrar los reportes OLAP, usuarios, roles y permisos.
Administrador de ETL	Es el responsable de la extracción, transformación y carga de los datos.

Tabla 2: Descripción de los actores del sistema.

Caso de Uso	Descripción
Analizar trabajadores físicos	Visualiza los reportes de trabajadores físicos para analizar la información
Analizar registro patentes	Visualiza los reportes de registro patentes para analizar la información
Analizar gastos	Visualiza los reportes de gastos para analizar la información
Analizar inversiones ejecutadas	Visualiza los reportes de inversiones ejecutadas para analizar la información
Analizar publicaciones	Visualiza los reportes de publicaciones para analizar la información
Analizar patentes de invención por países	Visualiza los reportes de invención por países para analizar la información
Realizar transformación y carga de los datos de Ciencia e innovación tecnológica	Realiza la transformación y carga de los datos
Extraer datos de Ciencia e innovación tecnológica	Realiza la extracción de los datos
Administrar roles	Elimina e inserta los roles
Administrar reportes	Elimina, inserta y modifica los reportes que se visualizan
Administrar usuario	Elimina e inserta los usuarios que interactúan en el sistema
Autenticar usuario	Realiza la autenticación de los usuarios en el sistema
Visualizar reportes	Visualiza todos los reportes

Tabla 3: Descripción de los casos de uso del sistema.

2.3 Diseño de la solución

En el diseño es donde se refina el análisis, en este el sistema está listo para trabajar sin ambigüedades. Se define el esquema de seguridad, la matriz bus, el modelo de datos, se establecen

las políticas de seguridad así como los permisos que tendrá el usuario a la hora de interactuar con el mercado de datos.

2.3.1 Matriz BUS o matriz dimensional

Es la representación de las relaciones existentes entre los hechos y las dimensiones del sistema.

Dimensiones/ Hechos	Gastos	Inversiones	Publicaciones	Trabajadores físicos	Registro de patentes	Patentes de invención por países
País						X
Título			X			
Gasto total	X					
Estado legal						X
Componentes		X				
Nivel categoría				X		
Solicitud estado					X	
Temporal año	X	X	X	X	X	X
Unidad de medida	X	X	X	X	X	X
División política administrativa	X	X	X	X	X	

Ilustración 3: Matriz BUS.

Hechos

Los hechos contienen las dimensiones y medidas asociadas. Luego de realizar un análisis se identificaron los siguientes hechos:

- ✓ hech_gastos
- ✓ hech_inversiones
- ✓ hech_trabajadores_fisicos
- ✓ hech_patentes_invencion_paises

- ✓ hech_publicaciones
- ✓ hech_registros_patentes_invenciones_modelos

Dimensiones

Son las características del hecho. Las dimensiones definidas son las siguientes:

- ✓ dim_temporal_anno: Es una dimensión temporal donde se registran los años.
- ✓ dim_um: Es una dimensión donde se registran todas las unidades de medidas.
- ✓ dim_dpa: Es una dimensión que registra la división política administrativa del país.
- ✓ dim_gastos_corrientes: Es una dimensión donde se registran los gastos totales por tipo de actividades y los gastos corrientes por fuente de financiamiento.
- ✓ dim_nivel_categoria: Es una dimensión donde se registran los niveles educacionales y las categorías ocupacionales.
- ✓ dim_componentes: Es una dimensión donde se registran los componentes de las inversiones ejecutadas.
- ✓ dim_titulo: Es una dimensión donde se registra el título de una publicación.
- ✓ dim_pais: Es una dimensión donde se registra el país.
- ✓ dim_estado_legal: Es una dimensión donde se registran los estados legales, ya sean solicitados o concedidos.
- ✓ dim_solicitud_estado: Es una dimensión donde se registran las solicitudes, ya sean nacionales o extranjeras.

Medidas

Son los valores numéricos analizados que representan lo que se necesita conocer. Las medidas definidas son las siguientes:

- ✓ gastos
- ✓ cant_inv_ejecutadas
- ✓ cantidad_modelos_industriales

- ✓ cant_registro_patentes_invenciones
- ✓ cant_impreso
- ✓ cant_electronico
- ✓ cant_ambos_formatos
- ✓ cant_patentes_invencion_paises

2.3.2 Modelo de datos

El modelo de datos está compuesto por los hechos, dimensiones y medidas definidas.

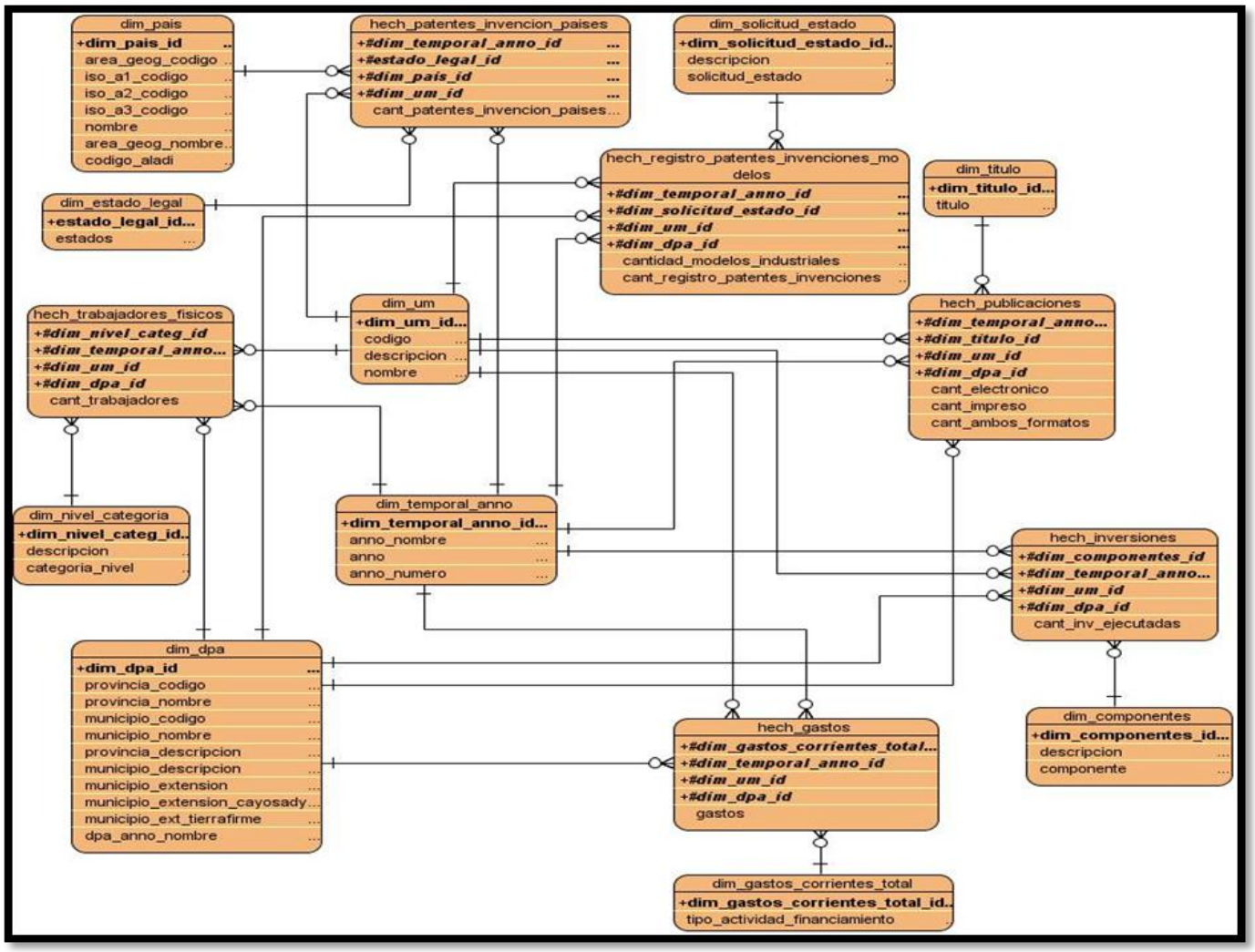


Ilustración 4: Modelo de datos.

2.3.3 Política de respaldo y recuperación

En la política de respaldo y recuperación que utiliza la solución se miden 3 puntos esenciales:

Periodicidad de las salvas: Las salvas se realizan mensualmente a toda la información contenida en el MD. La organización lo tiene definido de esta manera, certificando en todo momento la existencia de una copia escrita de la información que está presente en el servidor.

Tablas involucradas: Las tablas que se involucran en la realización son: tabla de hecho hech_gastos, hech_inversiones, hech_trabajadores_fisicos, hech_patentes_invencion_paises, hech_publicaciones y hech_registros_patentes_invenciones_modelos.

Backups existentes: En esta área existe backups.

2.3.4 Esquema de seguridad

Representa el respaldado por los niveles de acceso, específicamente por los roles definidos. Administra los permisos que tendría cada tipo de usuario a la hora de acceder a la información del sistema.

A continuación se muestran 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

Rol definido para la interacción de los usuarios con la BD:

Actor	Descripción
Administrador ETL	Realiza los procesos de ETL de los datos.
Administrador	Tiene permisos de insertar, actualizar, modificar y eliminar.

Tabla 4: Seguridad en la BD.

Seguridad en la aplicación

Las aplicaciones desplegadas en el servidor de BI de Pentaho muestran un continuo incremento, así como los usuarios que tiene acceso a estas. A continuación se definen los roles de seguridad para la

interacción de los usuarios con la aplicación para hacer sostenible el manejo de la misma en este servidor.

Roles	Permisos
Administrador	Tiene acceso total a todas las Áreas de Análisis (AA) General.
Analista	Tiene acceso de solo lectura al AA Ciencia y tecnología. Visualiza los reportes.

Tabla 5: Roles y permisos.

Elemento de aplicación	Roles con acceso
AA General	Administrador Analista
Carpeta raíz: AA Ciencia y tecnología	Administrador Analista

Tabla 6: Seguridad en la aplicación.

2.4 Conclusiones del capítulo

En este capítulo, luego de un refinamiento de los requisitos se detectaron 9 requisitos de información, 17 requisitos funcionales y 18 requisitos no funcionales. Se obtuvieron 12 casos de uso, los cuales se describieron para un mejor entendimiento de los mismos. El modelo de datos fue refinado, donde se identificaron nuevos hechos, dimensiones y medidas. Se diseñó la matriz bus, las políticas de seguridad, los roles y permisos, entre otros artefactos, los cuales constituyen la entrada a la realización del MD.

Capítulo 3: Implementación del Mercado de Datos

3.1 Introducción

Este capítulo tiene como objetivo principal desarrollar la implementación de los procesos ETL y BI para el área de Ciencia e innovación tecnológica y darle solución a los requisitos del sistema.

3.2 Implementación de la base de datos

Luego del diseño del modelo dimensional se realizó la trasformación al modelo físico, que permite describir cómo se almacenan los datos y la relación existente entre las tablas.

3.2.1 Estructura de los datos

Esquemas

Con los esquemas toda la información contenida en la BD se representa de forma organizada, estos contienen funciones, operadores y tipos de datos. Los esquemas no se encuentran separados, lo que permite que el usuario pueda tener acceso a ellos siempre que tenga los permisos necesarios.

En el presente trabajo se definieron 2 esquemas:

dimensiones: contiene las tablas de las dimensiones propuestas que son comunes con la BD de SIGOB.

mart_ciencia_tecnologia: contiene todas las tablas de hechos y las dimensiones propias del MD.

Tablas

El MD Ciencia e innovación tecnológica cuenta con 16 tablas en total, 10 dimensiones y 6 hechos, distribuidas en los 2 esquemas anteriormente mencionados como se muestra a continuación:

Esquemas	Tablas
dimensiones	dim_um
	dim_pais
	dim_dpa
	dim_temporal_anno
mart_ciencia_tecnologia	dim_titulo
	dim_componente
	dim_estado_legal
	dim_nivel_categoria
	dim_solicitud_estado
	dim_gastos_corrientes_total
	hech_ gastos
	hech_inversiones
	hech_publicaciones
	hech_trabajadores_fisicos
	hech_patentes_invencion_paises
	hech_registro_patentes_invenciones_modelos

Tabla 7: Seguridad en la aplicación.

En la siguiente imagen se muestra como queda conformada la estructura física de la BD.

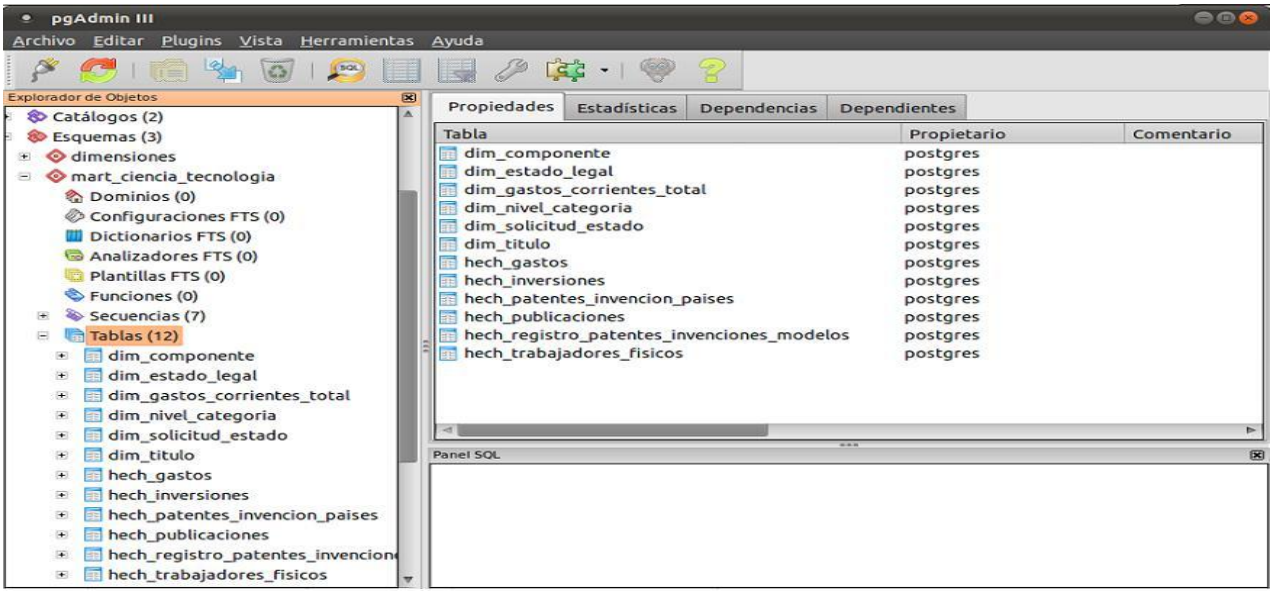


Ilustración 5: Estructura física de la BD.

3.3 Implementación del subsistema de integración de datos.

En el proceso de integración de datos no es recomendable iniciar el desarrollo de la fuente sin haberla analizado previamente.

La fuente son los datos que se encuentran almacenados en los sistemas fuentes que guardan la información histórica, estos se encuentran en formato .xls y sufrirán un proceso de cambios para facilitar el trabajo con las transformaciones.

La etapa de transformación y limpieza es muy importante, una vez que se realiza, la información está lista para ser cargada en la BD. Con la limpieza se detectan los datos erróneos, además de detectar entradas duplicadas y con las transformaciones se combinan y ordenan los datos.

3.3.1 Arquitectura del subsistema de integración

La arquitectura es un conjunto de reglas que se utilizan como guía para el diseño de un sistema.

La siguiente figura muestra la arquitectura que se utilizó para el desarrollo de la solución.

Ilustración 6: Arquitectura de la integración.

A continuación se describen los elementos que conforman la arquitectura de integración:

Fuente de datos: Son los datos que se encuentran almacenados en los sistemas fuentes que guardan información histórica de los sistemas.

Entorno de trabajo: Es donde se preparan los datos para facilitar los procesos de integración y se encuentran todas las transformaciones.

Mercado de datos: Es el destino hacia donde son cargados los datos.

3.3.2 Proceso de Extracción, Transformación y Carga

Para la realización del proceso de ETL del MD Ciencia e innovación tecnológica se realizaron 6 transformaciones para la carga de los hechos. A continuación se muestra un ejemplo de la carga correspondiente al hecho trabajadores físicos:

carga_hech_trabajadores_fisicos: En la transformación del hecho trabajadores físicos se realiza la carga de los datos correspondiente a las 2 series históricas siguientes:

- ✓ Trabajadores físicos en la actividad de ciencia y tecnología según nivel educacional.
- ✓ Trabajadores físicos en la actividad de ciencia y tecnología según categoría ocupacional.

Para registrar los datos de los trabajadores físicos en el MD se verifica que las variables que no permitan nulos y los identificadores se validen. De los identificadores de las dimensiones relacionadas con el hecho se realizan las búsquedas y finalmente se registra la información.

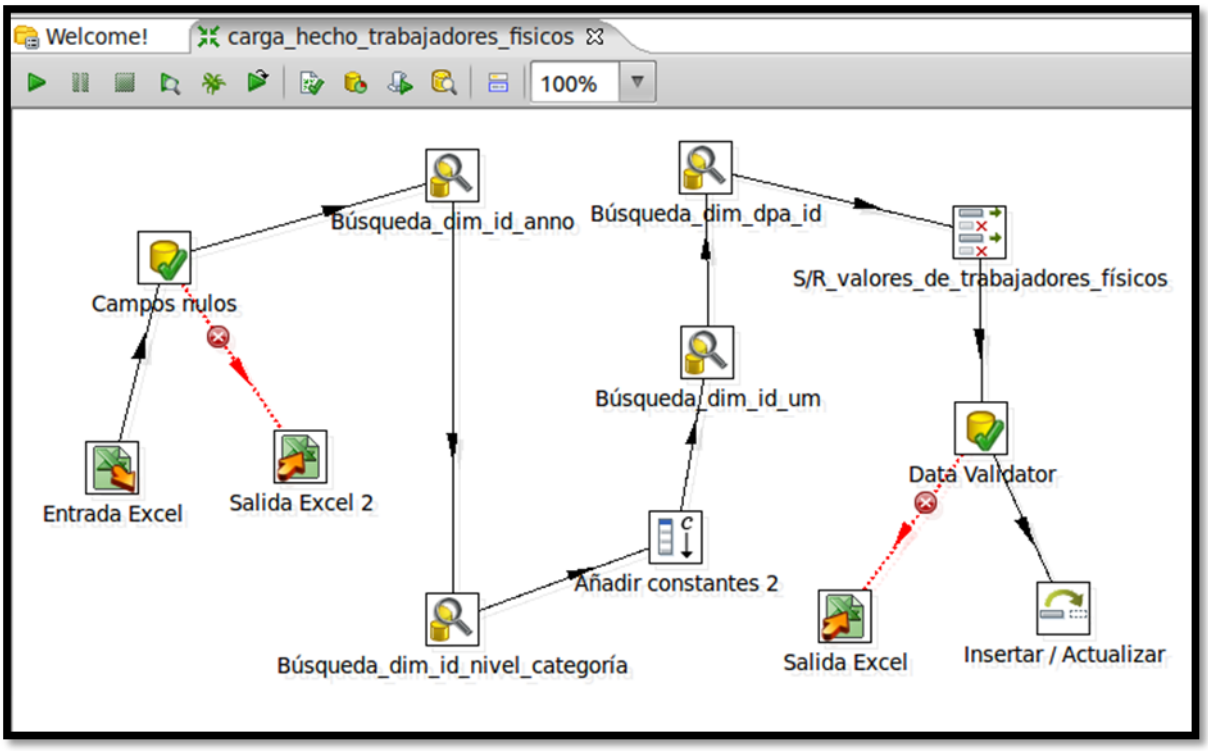


Ilustración 7: Transformación del hecho trabajadores físicos.

3.4 Implementación del trabajo

El trabajo es la forma en que se realiza la carga de los datos de una manera organizada. En la siguiente figura se muestra el trabajo que se realizó en este proceso.

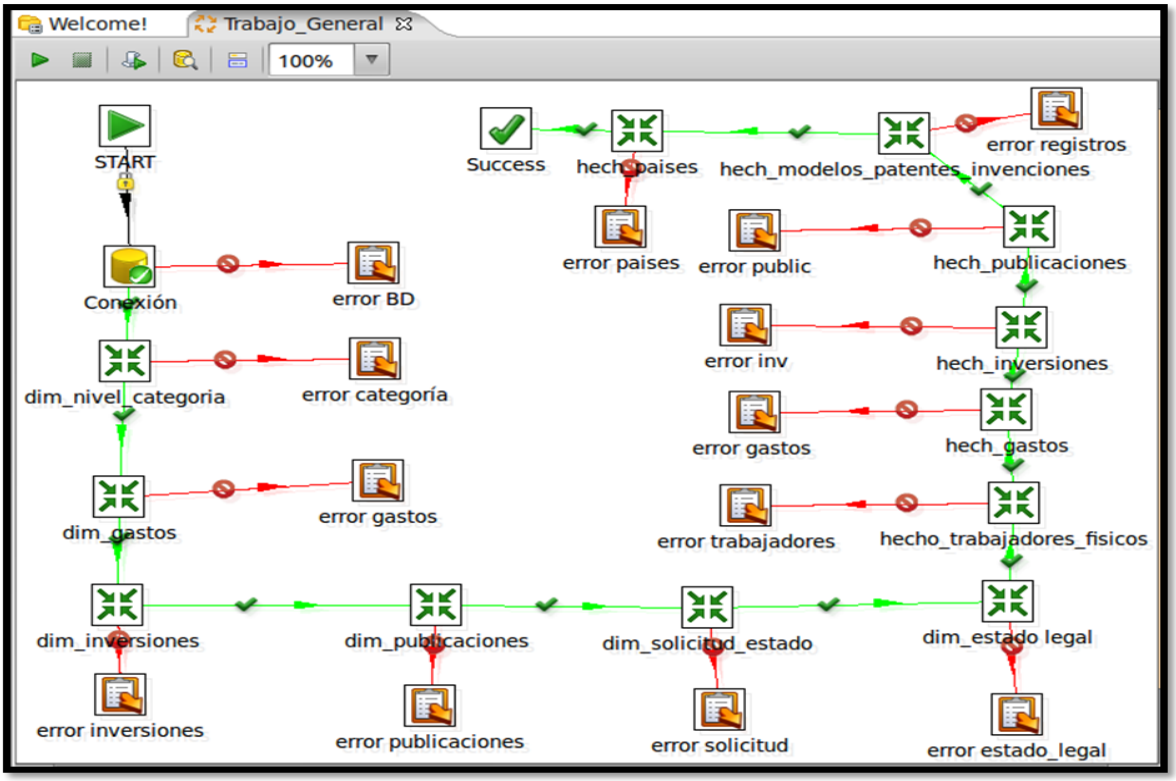


Ilustración 8: Trabajo.

3.5 Implementación del subsistema de visualización

3.5.1 Cubos OLAP

En el desarrollo del subsistema de visualización es necesario utilizar la herramienta Pentaho Schema Workbech para la creación de los cubos multidimensionales. Se genera un fichero .xml que guardará los cubos, donde quedan definidas las dimensiones, los niveles de jerarquías de las dimensiones y las medidas.

Se modelaron 6 cubos multidimensionales y se relacionaron con sus respectivas dimensiones y medidas. Para los hechos existen dos dimensiones que son comunes para todos, la dimensión temporal año y la dimensión unidad de medida, ya que los reportes se van a mostrar anualmente y con una unidad de medida.

A continuación se muestran los cubos modelados:

Ilustración 9: Diseño de los cubos en el Pentaho Schema Workbech.

A continuación se muestra el ejemplo del cubo trabajadores físicos. El mismo muestra sus dimensiones asociadas. La dimensión dim_nivel_categoría almacena información de la categoría ocupacional y nivel educacional de los trabajadores físicos. La dimensión dim_temporal_anno almacena información de cada uno de los años. La dimensión dim_um almacena información de las unidades de medidas. La dimensión dim_dpa almacena información de la división política administrativa del país y la medida cant_trabajadores muestra la cantidad de trabajadores físicos.

Ilustración 10: Diseño del cubo trabajadores físicos.

3.5.2 Navegación de la capa de visualización

Los elementos que componen las estructuras de navegación de la información que será presentada en la capa de visualización del MD Ciencia e innovación tecnológica se describen a continuación, la misma contiene 1 Área de Análisis (A.A), 1 Libro de Trabajo (L.T) y 9 reportes:

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

A.A.G SIGOB: Agrupa información de todos los MD.

Descripción de las Áreas de Análisis (A.A):

A.A Ciencia e innovación tecnológica: Agrupa información acerca de las innovaciones que son presentadas, concedidas y aplicadas por los centros durante el año que se informa. Posee reportes estadísticos que apoyan la concepción de medidas y estrategias encaminadas a realizar un mejor control en el proceso de toma de decisiones en el área Ciencia e innovación tecnológica.

Descripción de los libros de trabajo:

L.T Ciencia y tecnología**:** Libro de trabajo contenido dentro del área de análisis Ciencia e innovación tecnológica. Contiene 9 reportes que permiten realizar un análisis de los datos correspondiente a las innovaciones realizadas.

Descripción de los reportes: Reportes correspondientes al Libro de Trabajo de Ciencia y tecnología

TS1 – Gasto total en actividades de ciencia y tecnología por fuente de financiamiento.

TS2 – Gasto total en actividades de ciencia y tecnología por tipo de actividades.

TS3 – Inversiones ejecutadas en la actividad de ciencia y tecnología por componentes.

TS4 – Patentes de invención por países.

TS5 – Registro de patentes de invenciones presentadas en Cuba.

TS6– Registro de patentes de modelos industriales presentados en Cuba.

TS7 – Trabajadores físicos en la actividad de ciencia y tecnología según categoría ocupacional.

TS8 – Trabajadores físicos en la actividad de ciencia y tecnología según nivel educacional.

TS9 – Títulos de publicaciones seriadas de ciencia y tecnología.

Mapa de navegación

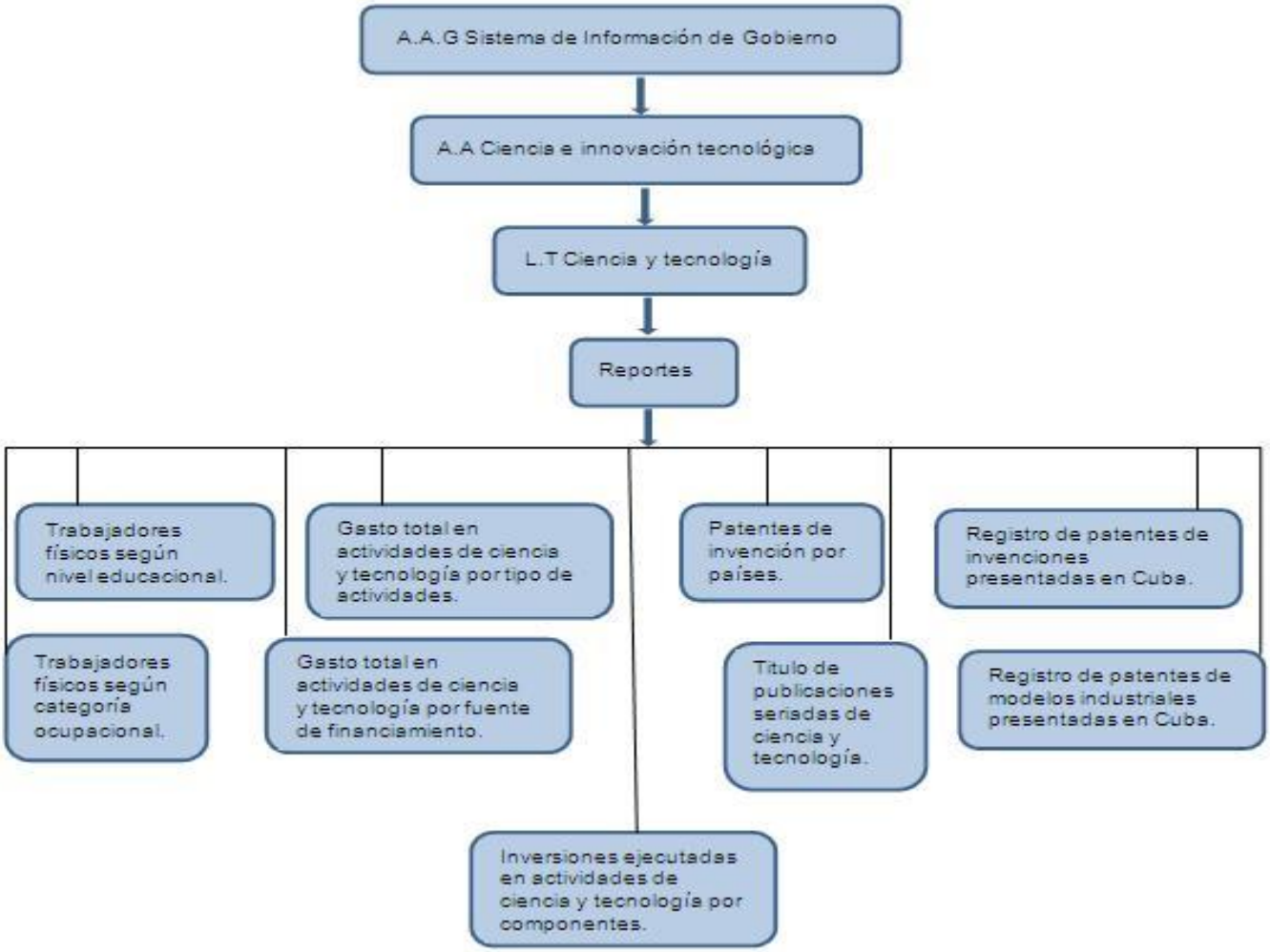


Ilustración 11: Mapa de navegación.

	Nivel territorial:													
	Nacional													
	Unidad de medida:													
	Unidad													
	Año													
Concepto	1977	1978	1979	1980	1981	1982	1983	1984	1985	1986	1987	1988	1989	1990
Total	19.301	19.868	21.489	21.521	28.813	30.863	33.506	37.696	40.183	41.284	40.459	39.009	41.784	43.445
Nivel superior	4.601	4.972	5.383	5.637	10.073	11.174	12.290	13.837	15.808	16.665	17.272	17.196	19.985	21.354
Nivel medio	6.075	6.077	6.593	6.556	8.199	8.422	8.704	9.134	9.656	9.738	9.472	9.375	9.491	10.409
Otros	8.625	8.819	9.513	9.328	10.541	11.267	12.512	14.725	14.719	14.881	13.715	12.438	12.308	11.682
Grados científicos	200	453	478	655	802	1.002	1.200	1.449	1.819	1.819	1.959	2.312	2.524	4.108

Ilustración 12 : Reporte trabajadores físicos según nivel educacional.

3.6 Conclusiones del capítulo

En este capítulo se realizó un análisis de la fuente de datos y la implementación del MD lo que permitió que se efectuara la carga de los datos a la BD satisfactoriamente quedando implementado el subsistema de integración. El esquema multidimensional quedó modelado, se identificó el área de análisis, el libro de trabajo y reportes candidatos contenidos. Una vez visualizados cada uno de los reportes van a ayudar a medir el desempeño organizacional de la entidad, alcanzando resultados satisfactorios en el objeto de la investigación, llegando a conclusiones para la proyección de un mejor trabajo en la organización, que le permita mayores niveles de eficacia, efectividad y eficiencia a la ONE.

Capítulo 4: Validación y pruebas del Mercado de Datos

4.1 Introducción

En este capítulo se describe el proceso que se realiza para la validación del MD Ciencia e innovación tecnológica. Se realizan las validaciones y la aplicación de las listas de chequeo y de los casos de pruebas, con el fin de evaluar la calidad del producto. Una vez realizado todos estos procesos se aplicaron pruebas al sistema, verificando la integridad de los datos de los reportes y el alcance del dominio informativo.

4.2 Prueba

La fase de pruebas es el último ciclo de vida antes de entregar un programa para su explotación y garantizan que el software funcione y presente la calidad requerida.

Las pruebas se realizan para determinar la rapidez con que realiza una tarea bajo condiciones particulares de trabajo. Validan y verifican la calidad del sistema, tales como la escalabilidad y fiabilidad. Verifican el correcto funcionamiento de los pasos de implantación del sistema y luego de la aplicación de las listas de chequeo para validar.

A la solución se le aplicaron varias pruebas como son: las pruebas unitarias e internas realizadas una vez terminado el desarrollo y las pruebas de integración en conjunto con los especialistas del centro DATEC. Una vez desarrolladas las pruebas por el centro, el MD es entregado a calidad UCI donde se encarga de realizar pruebas de liberación y aceptación y por último la operacional, efectuada con el cliente.

4.2.1 Casos de prueba

Los casos de prueba verifican si el producto satisface los requerimientos del usuario, tal y como se describe en la especificación de requerimientos y los casos de uso. En la solución propuesta se utilizan varios casos de prueba basados en casos de uso para realizar validaciones concretas.

4.2.2 Listas de chequeo

Las listas de chequeo se crean con el fin de concretar y propiciar un buen desarrollo en el trabajo. Son un conjunto de preguntas que sirven para verificar el cumplimiento de los objetivos.

Contienen diferentes indicadores a evaluar, los cuales se encuentran distribuidos en tres secciones:

- ✓ Estructura del documento: Contiene todos los aspectos definidos por el expediente del proyecto.
- ✓ Indicadores definidos por la etapa: Contiene todos los indicadores a evaluar durante la etapa de análisis de datos.
- ✓ Semántica del documento: Contiene todos los indicadores a evaluar respecto a la redacción y ortografía.

4.2.3 Estructura de las listas 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 1 en caso de que exista alguna dificultad sobre el indicador y 0 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 y con el objetivo de darles solución se especifican en una tabla de No Conformidades (NC), la cual presenta la siguiente estructura:

No: Es un número consecutivo que indica la cantidad de NC identificadas.

Elemento de evaluación: Se refiere a un número que identifica al elemento de evaluación para el cual se corresponden los indicadores identificados.

No Conformidad: Especifica la NC a la que se refiere.

Fase correspondiente: Especifica la fase del procedimiento a la que corresponde la NC encontrada.

Significación: Especifica si la NC es o no significativa, dependiendo si el indicador es o no crítico.

Recomendación: Especifica si la NC es una recomendación, es decir que no es de obligatorio cumplimiento que se solucione por parte de los diseñadores.

Estado NC: Especifica el estado de solución en que se encuentra la NC, puede ser pendiente o solucionada.

Respuesta del equipo de desarrollo: Si es necesario se especifica la respuesta que le da el equipo de desarrollo a la NC.

4.2.4 Aplicación de las listas de chequeo

Estructura del documento					
Peso	**Indicadores a evaluar**	**Evaluación**	**No procede**	**Cantidad de elementos afectados**	**Comentarios**
Crítico	✓ ¿Los entregables contienen las secciones obligatorias de la plantilla estándar definidas para un expediente de proyecto? (Portada, Control de Versiones, Reglas de Confidencialidad, Tabla de Contenidos y Contenido)	0		0	
Indicadores definidos por la etapa					
Peso	**Indicadores a evaluar**	**Evaluación**	**No procede**	**Cantidad de elementos afectados**	**Comentarios**
	✓ ¿La arquitectura satisface las necesidades del proyecto?	0		0	0
	✓ ¿La arquitectura soporta el incremento del proyecto?	0		0	0

	✓ ¿Se utilizó el menor número de transformaciones posibles al cargar los datos hacia el área de trabajo?	0		0	0
Crítico	✓ ¿Se creó el modelo físico a partir del modelo lógico?	0		0	0
Crítico	✓ ¿Cumple la implementación del proceso de ETL con la arquitectura definida?	0		0	0
	✓ ¿Se tuvo en cuenta los formatos fuentes y tipos de datos de las perspectivas de análisis?	0		0	0
	✓ ¿Se realiza una limpieza de los datos antes de realizar la carga de los mismos?	0		0	0
	✓ ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?	0		0	
	✓ ¿Se realizó una interfaz amigable para hacer consultas?	0		0	
Crítico	✓ ¿Los reportes son configurables a través de la interfaz del sistema?	0		0	
	✓ ¿El rendimiento de los reportes no se afecta cuando el número de dimensiones del modelo se incrementa?	0		0	

	✓ ¿Presenta la capacidad de crear todo tipo de dimensiones con funcionalidades aplicables de una dimensión a otra?	0		0	
	✓ ¿La interfaz está orientada a facilitar el uso de las funciones del sistema por parte de los usuarios?	0		0	
Crítico	✓ ¿No existen restricciones para construir cubos OLAP con dimensiones y niveles de agregación ilimitados?	0		0	
Crítico	✓ ¿Los usuarios son capaces de manipular los resultados de manera que se ajusten a sus necesidades, conformando nuevos reportes?	0		0	
	✓ ¿El sistema responde de una forma rápida y veraz a la información que le sea solicitada por el usuario?	0		0	
Crítico	✓ ¿El sistema refleja cualquier lógica del negocio para poder responder a preguntas específicas?	0		0	
Crítico	✓ ¿El sistema garantiza la confidencialidad y seguridad de acceso a los datos por rol de los usuarios?	0		0	

	✓ ¿Los datos e información derivados del proceso de análisis realizado mediante la aplicación, apoyan la toma de decisiones en la institución?	0		0	
Crítico	✓ ¿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 procede**	**Cantidad de elementos afectados**	**Comentarios**
Crítico	✓ ¿Se han identificado errores ortográficos en los entregables?	0		0	
Crítico	✓ ¿Se entiende claramente lo que se ha especificado en el documento?	0		0	
	✓ ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?	0		0	

Tabla 8: Aplicación de las listas de chequeo.

Luego de aplicar las listas de chequeo se genera un gráfico donde se visualiza el comportamiento de los 24 indicadores identificados, de los cuales 11 son críticos y en el cual no se generaron ningunas no conformidades.

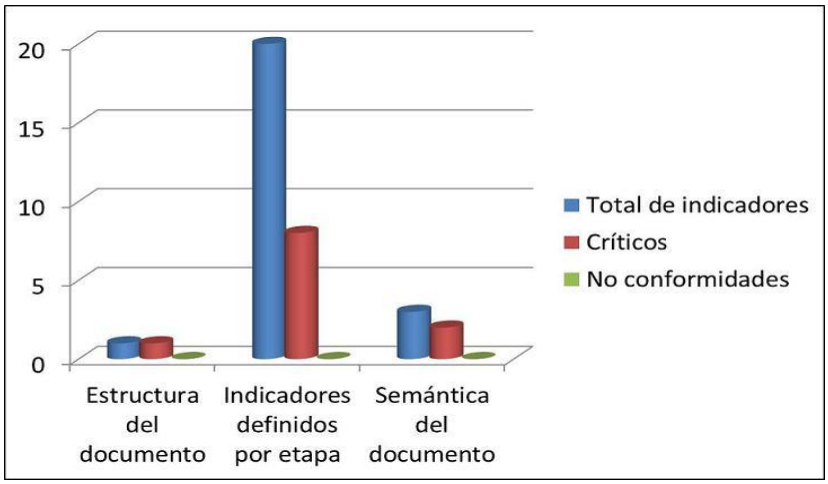


Ilustración 13: Comportamiento de los indicadores.

4.2.5 No conformidades

Una vez culminado el proceso de pruebas se registran las no conformidades. En la solución propuesta se detectaron 16 no conformidades, 14 en los artefactos y 1 error ortográfico, las cuales fueron corregidas.

4.3 Validación

La validación es la comprobación de que la solución está acorde a las necesidades y exigencias de los clientes.

4.3.1 Validación de requisitos con el cliente

Para aprobar la propuesta de solución se realizó una encuentro con la cliente Elena Fernández García, quedando conforme con la propuesta mostrada y satisfecha con el cumplimiento de los requisitos.

4.4 Conclusiones

Se realizó una descripción de la aplicación de las listas de chequeo para evaluar el MD, obteniendo resultados satisfactorios. Fueron evaluados tres aspectos fundamentales: estructura del documento, indicadores definidos por etapa y semántica del documento, en el cual se evaluaron 24 indicadores, de ellos 11 críticos que son fundamentales para el análisis de la evaluación y no se detectaron inconformidades. Todo lo planteado contribuyó a un buen funcionamiento de la aplicación, garantizando la calidad y dándole cumplimiento a los requisitos.

Conclusiones Generales

Después de haber desarrollado el MD Ciencia e innovación tecnológica se llegan a las siguientes conclusiones:

- ✓ El refinamiento del análisis y diseño del MD Ciencia e innovación tecnológica permitió detallar los requisitos los cuales dan solución a las necesidades del cliente.
- ✓ Los procesos de ETL y BI permitieron poblar el MD y desarrollar la capa de visualización, determinándose un libro de trabajo con un total de nueve vistas de análisis.
- ✓ Las pruebas y validaciones realizadas al MD Ciencia e innovación tecnológica permitieron verificar la integridad de los datos en las vistas de análisis.

Recomendaciones

Luego de la presentación del estudio realizado que culmina con la implementación de un MD que contribuye a la toma de decisiones en el área de Ciencia e innovación tecnológica del Sistema de Información de Gobierno para la ONE, se recomienda:

- ✓ Impartir cursos de capacitación para los especialistas de la ONE.
- ✓ Integrar el MD Ciencia e innovación tecnológica con el Sistema Informático de Gestión Estadística (SIGE).

Referencias Bibliográficas

1. ONE. [Online] 2006. [Cited: septiembre 20, 2010.] http://www.one.cu/.

2. Serguera, Alexis Fernández. La gestión de la actividad de ciencia e innovación tecnológica en la organización. Ciudad de la Habana : s.n.

3. Christian Van Der Henst S. Maestros del Web. [Online] Julio 19, 1997. [Cited: octubre 30, 2010.] www.maestrosdelweb.com/principiantes/%C2%BFque-son-las-bases-de-datos. .

4. Data Warehouse (Almacenes de Datos). Cabrera, María Evelia Casales. 2009.

5. Inmon. W.H. Using the Data Warehouse. 1992.

6.Sinnexus. [Online] 2007. [Cited: noviembre 5, 2010.] http://www.sinnexus.com/business_intelligence/datawarehouse.aspx.

7. Kimball, Ralph. The Data Warehouse ETL Toolkit. s.l. : Second Edition, 2002. 0-471-20024-7.

8. Kimball, Ralph. The Data Warehouse Lifecycle Toolkit. 2002.

9.Todo Tecnología . [Online] [Cited: noviembre 5, 2010.] http://todotecnology.blogspot.com/2009/09/datamart.html.

10.Transforming Knowledge Into Action! [Online] [Cited: noviembre 6, 2010.] http://kle.sisinfomanagement.com/spanish/articulo01.html.

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

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

13.Free Download Manager. [Online] 2004. [Cited: noviembre 11, 2010.] http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML_%5Bcuenta_de_Plataforma_de_Java_14715_p/.

14. Moreno, Mauro Callejas Cuervo and Baquero, Oscar Yovany. Herramientas libres para modelar software. 2005. ISSN 0121–1129.

15.Dario, Ing. Bernabeu Ricardo. Metodología propia para la la Construcción de un Data Warehouse. Argentina, Córdoba : s.n., 2009.

16. Casanova, Jaime. PosgreSQL. [Online] septiembre 1, 2009. [Cited: noviembre 30, 2010.] http://archives.postgresql.org/pgsql-es-fomento/2009-07/msg00000.php.

17. DataCleaner. [Online] [Cited: diciembre 5, 2010.] http://datacleaner.eobjects.org/.

18. Schmidt, Jose and Ferreira, Keyla. Sistema de Información de Pentaho. 2009.

19. Aguilera, Álvaro Moreno, García, Carlos Muñoz and Flores, Alberto Alvarado. Herramientas ETL de código abierto.

20. Portada sobre la plataforma Pentaho Open Source Business Intelligence. [Online] 2006. [Cited: diciembre 8, 2010.] http://pentaho.almacen-datos.com/.

21. Wood, Sherman. Pentaho Mondrian Project. [Online] 2008. [Cited: diciembre 8, 2010.] http://mondrian.pentaho.com/documentation/workbench.php.

22. Agapea. [Online] 2002. [Cited: diciembre 10, 2010.] http://www.agapea.com/libros/Tomcat-6-0-La-guia-definitiva-isbn-8441524319-i.htm.

Bibliografía

ONE. [En línea] 2006. [Citado el: 20 de septiembre de 2010.] http://www.one.cu/.

Serguera, Alexis Fernández. *La gestión de la actividad de ciencia e innovación tecnológica en la organización.* Ciudad de la Habana : s.n.

Christian Van Der Henst S. Maestros del Web. [En línea] 19 de Julio de 1997. [Citado el: 30 de octubre de 2010.] www.maestrosdelweb.com/principiantes/%C2%BFque-son-las-bases-de-datos. .

Data Warehouse (Almacenes de Datos). **Cabrera, María Evelia Casales.** 2009.

Inmon. *W.H. Using the Data Warehouse.* 1992.

Sinnexus. [En línea] 2007. [Citado el: 5 de noviembre de 2010.] http://www.sinnexus.com/business_intelligence/datawarehouse.aspx.

Kimball, Ralph. *The Data Warehouse ETL Toolkit.* s.l. : Second Edition, 2002. 0-471-20024-7.

—. *The Data Warehouse Lifecycle Toolkit.* 2002.

Todo Tecnología . [En línea] [Citado el: 5 de noviembre de 2010.] http://todotecnology.blogspot.com/2009/09/datamart.html.

Transforming Knowledge Into Action! [En línea] [Citado el: 6 de noviembre de 2010.] http://kle.sisinfomanagement.com/spanish/articulo01.html.

UDLAP. [En línea] 1997. [Citado el: 11 de noviembre de 2010.] http://ict.udlap.mx/people/carlos/is341/bases02.html.

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

Free Download Manager. [En línea] 2004. [Citado el: 11 de noviembre de 2010.] http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML_%5Bcuenta_de_Plataforma_de_Java_14715_p/.

Moreno, Mauro Callejas Cuervo y Baquero, Oscar Yovany. *Herramientas libres para modelar software.* 2005. ISSN 0121–1129.

Dario, Ing. Bernabeu Ricardo. *Metodología propia para la la Construcción de un Data Warehouse.* Argentina, Córdoba : s.n., 2009.

Casanova, Jaime. PosgreSQL. [En línea] 1 de septiembre de 2009. [Citado el: 30 de noviembre de 2010.] http://archives.postgresql.org/pgsql-es-fomento/2009-07/msg00000.php.

DataCleaner. [En línea] [Citado el: 5 de diciembre de 2010.] http://datacleaner.eobjects.org/.

Schmidt, Jose y Ferreira, Keyla. *Sistema de Información de Pentaho.* 2009.

Aguilera, Álvaro Moreno, García, Carlos Muñoz y Flores, Alberto Alvarado. *Herramientas ETL de código abierto.*

Portada sobre la plataforma Pentaho Open Source Business Intelligence. [En línea] 2006. [Citado el: 8 de diciembre de 2010.] http://pentaho.almacen-datos.com/.

Wood, Sherman. Pentaho Mondrian Project. [En línea] 2008. [Citado el: 8 de diciembre de 2010.] http://mondrian.pentaho.com/documentation/workbench.php.

Agapea. [En línea] 2002. [Citado el: 10 de diciembre de 2010.] http://www.agapea.com/libros/Tomcat-6-0-La-guia-definitiva-isbn-8441524319-i.htm.

Alfaro, Félix Murrillo. *Manual para la Construcción de un Data Warehouse.* 1997.

Buenastareas. [En línea] [Citado el: 30 de noviembre de 2010.] http://www.buenastareas.com/ensayos/Etl-Extraer-Transformar-Y-Cargar/452354.html.

CENTALAD. *Extracción, Tranformación y Carga.*

Dataprix. [En línea] [Citado el: 30 de noviembre de 2010.] http://www.dataprix.com/ca/node/1931.

Centro de Tecnologías de Almacenamiento y Análisis de Datos. *Introducción a Data Warehouse y Data Marts.*

Introducción al Proceso Analítico en Línea. Modelamiento Multidimensional.

Minería de Datos.

Centro de Tecnologías de Almacenamiento y Análisis Inteligente de Datos. *Inteligencia de Negocios.*

QDiario. [En línea] [Citado el: 15 de octubre de 2010.] http://www.aplicacionesempresariales.com/postgresql-84.html.

Díaz, Luis Rojas. *Desarrollo de sistemas de información estadísticas de ciencia y tecnología.* 2008.

ECURED. [En línea] [Citado el: 15 de octubre de 2010.] http://www.ecured.cu/index.php/Almac%C3%A9n_de_Datos.

Espinosa, Jaime Oyarzo. *Bases de datos.*

GESTEC. *Sistema de Ciencia e Innovacion Tecnologica en Cuba, desarrollo y desafíos.* 2004.

GestioPolis. [En línea] 2008. [Citado el: 16 de octubre de 2010.] http://www.gestiopolis.com/canales8/ger/olap-online-analytic-processing.htm.

González, Carlos Caballero. *Botanical Database: Retrieval of species using leaf images.* Málaga : s.n., 2007.

Gravitar. [En línea] [Citado el: 2 de diciembre de 2010.] http://www.gravitar.biz/index.php/herramientas-bi/pentaho/.

Hoy es un buen día para empezar. [En línea] [Citado el: 6 de diciembre de 2010.] http://www.hugoaguilar.mex.tl/218384_ETL---DataStage.html.

Ibarra, María de los Angeles. *Procesamiento Analítico en línea.* 2005.

Informática Hoy. [En línea] 2007. [Citado el: 3 de diciembre de 2010.] http://www.informatica-hoy.com.ar/telefonos-celulares/Cubo-OLAP-una-base-de-datos-multidimensional.php.

msdn. [En línea] [Citado el: 3 de diciembre de 2010.] http://msdn.microsoft.com/es-es/library/ms166352%28SQL.90%29.aspx.

Oficina Nacional de Gobierno Electrónico en Perú. [En línea] [Citado el: 29 de septiembre de 2010.] http://www.ongei.gob.pe/publica/metodologias/Lib5084/14.HTM.

OLAPX. [En línea] 2005. [Citado el: 29 de septiembre de 2010.] http://www.olapxsoftware.com/es/WhatIsOlap.asp.

ONE República Dominicana. [En línea] [Citado el: 29 de septiembre de 2010.] http://www.one.gob.do/index.php?module=articles&func=display&aid=1377.

Portada sobre la plataforma Pentaho Open Source Business Intelligence. [En línea] 2006. [Citado el: 4 de diciembre de 2010.] http://pentaho.almacen-datos.com/.

Prieto, José Abásolo. *Integración de Datos en la Organización: Necesidades y Soluciones.* Bogotá, Colombia : s.n., 2005.

Riunet. [En línea] [Citado el: 12 de diciembre de 2010.] http://riunet.upv.es/manakin/handle/10251/2506.

Rivera, Javier Fernandez. *Modelo de datos.*

Salazar, Ricardo Luján. *Datawarehouse para la prestación del servicio público de información estadísticas.* México : s.n.

Segundo, Ricardi Prieto. *DATAMART CAPACITACIÓN.* 2009.

slideshare. [En línea] [Citado el: 6 de diciembre de 2010.] http://www.slideshare.net/dannoblack/datawarehouse-y-datamining-parte-i.

Tandrón, Iván M. Cárdenas. *Metodología de Desarrollo de Soluciones BI y Warehousing (BI&W).* Ciudad de la Habana : s.n.

Thonon, Lic.Henry. *Diseño de un repositorio de datos para la gestión en la cadena de suministros en una empresa de consumo masivo.*

TopBits.com. [En línea] [Citado el: 1 de diciembre de 2010.] http://www.tech-faq.com/es/cubo-olap.html.

Verastegui, Hazbleydi C. *Modelado dimensional de datos.* 2007.

Villanueva, Wladimiro Díaz. *Almacenes de datos.* Valencia : s.n.

Worldlingo. [En línea] [Citado el: 29 de septiembre de 2010.] http://www.worldlingo.com/ma/enwiki/es/Ralph_Kimball.

Yglesias, Rodolfo. *Oracle vs Oracle.* 2008.

Zorrilla, Marta. *Datawarehouse y OLAP.* 2007/2008.

García, Joaquín. Webestilo. [En línea] 2005. [Citado el: 15 de noviembre de 2010.] http://www.webestilo.com/mysql/intro.phtml.

GSInnova. [En línea] [Citado el: 15 de noviembre de 2010.] http://www.rational.com.ar/herramientas/rosedatamodeler.html.

Fuente, Albert Philippe de la. *Medusa: un cliente avanzado de Oracle.* Málaga : s.n., 2003.

Glosario de términos

BI: Inteligencia de Negocio.

CASE: Ayuda a la Ingeniería de Software por computadora.

DATEC: Centro de Tecnologías de Almacenamiento de Datos.

ETL: Extracción, Transformación y Carga.

GPL: Licencia Pública General.

MDX: Lenguaje de consulta a estructuras multidimensionales.

ONE: Oficina Nacional de Estadísticas.

UCI: Universidad de las Ciencias Informáticas.

UML: Lenguaje Unificado de Modelado.

Anexos

Analizar inversiones ejecutadas

Caso de Uso:	Analizar inversiones ejecutadas.
Tipo:	Información.
Actores:	Analista.
Resumen:	El CU comienza cuando el actor entra al sistema y decide analizar la información de las inversiones y finaliza cuando se muestran los resultados.
Precondiciones:	Carga de los datos.
Referencias	RI5
Prioridad	Crítico

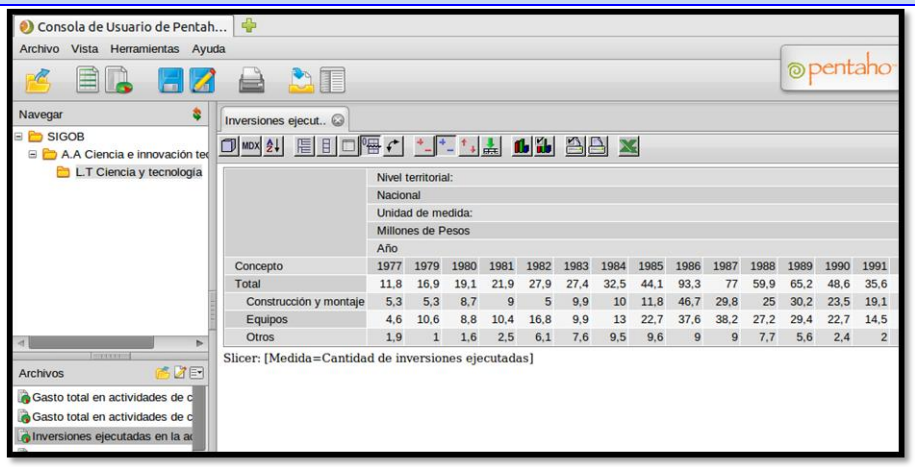
Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**
1. El actor se autentica y entrar al sistema.	2. El sistema muestra la interfaz principal con todas sus A.A.G.
3. El actor selecciona el A.A.G que desea analizar.	4. El sistema muestra las A.A que se encuentran en el A.A.G seleccionada por el actor.
5. El actor selecciona el A.A Ciencia e innovación tecnológica que desea analizar.	6. El sistema muestra los L.T correspondientes a dicha A.A seleccionada.
7. El actor selecciona el L.T Ciencia y tecnología.	8. El sistema muestra las opciones del reporte.
9. Selecciona el reporte analizar inversiones ejecutadas.	10. Muestra la información comprendida en el reporte analizar inversiones ejecutadas seleccionado, dando la posibilidad de realizarle cambios, donde se puede

	ver y analizar desde otra vista de análisis. Finaliza el CU.

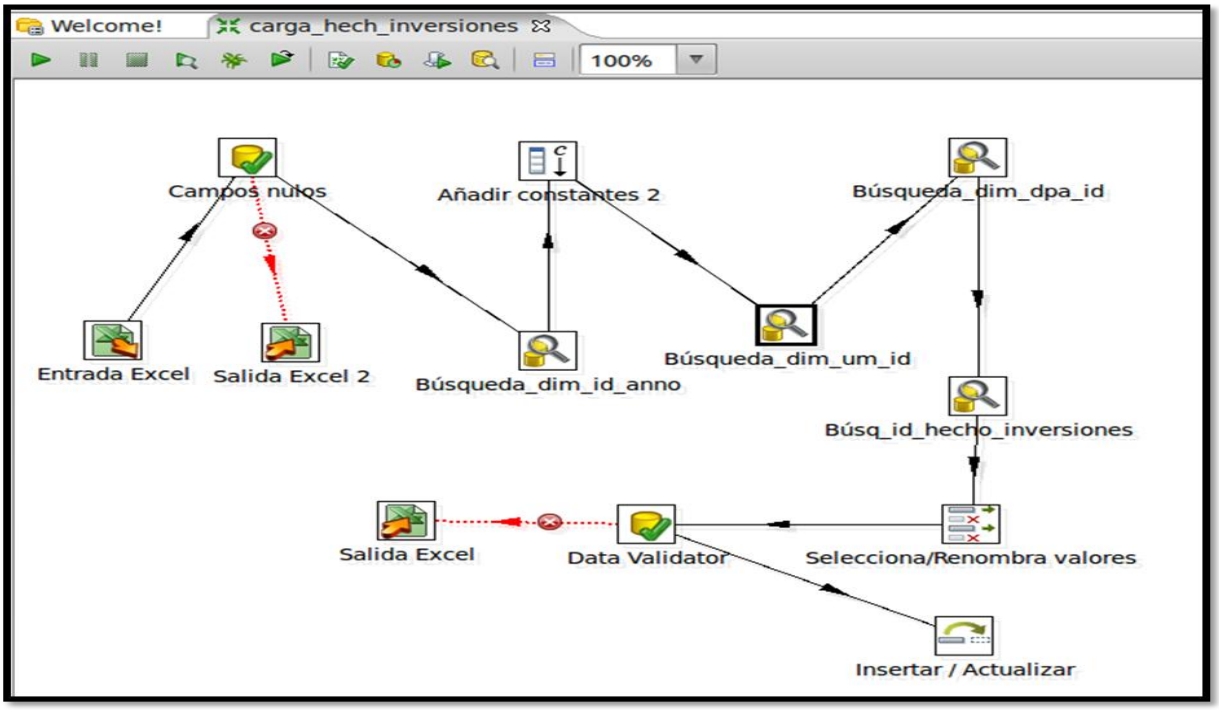
Opciones del reporte de inversiones

Entradas	Posibles resultados	
	Salidas	**Periodicidad**
Variables de entrada disponibles relacionadas con el caso de uso Analizar inversiones ejecutadas: ✓ Componentes ✓ Temporal ✓ Unidad de medida ✓ DPA	Variables de salida disponibles relacionadas con el caso de uso Analizar inversiones ejecutadas: ✓ Cantidad de inversiones ejecutadas.	Rango de tiempo en que se solicitan las variables de salida: ✓ Anual.

Prototipo de interfaz



Carga del hecho inversiones



Diseño del cubo inversiones