Universidad de las Ciencias Informáticas

Facultad 6

Título: Sistema de Información de Gobierno.
Mercado de datos para las áreas de Cultura y Deporte

TRABAJO DE DIPLOMA PARA OPTAR POR EL TÍTULO DE INGENIERO EN CIENCIAS INFORMÁTICAS

Autor: Fabian López García

Tutor: Ing. Loismarx Peña González

Ciudad de la Habana, junio de 2011

"Año 53 de la Revolución"

"El hombre debe transformarse al mismo tiempo que la producción progresa; no realizaríamos una tarea adecuada si fuéramos tan sólo productores de artículos, de materias primas y no fuéramos al mismo tiempo productores de hombres.

Declaro ser autor del presente trabajo “Sistema de Información de Gobierno. Mercado de datos para las áreas de Cultura y Deporte” 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 _____.

_____	_____
Autor	**Tutor**
Fabian López García	Loismarx Peña González

DATOS DE CONTACTO

Tutor: Ing. Loismarx Peña González
Graduado en el 2008 en la especialidad de Ingeniería en Ciencias Informáticas de la Universidad de las Ciencias Informáticas (UCI), Habana, Cuba. Profesor instructor de la Facultad 6 de la UCI. Email: lpgonzalez@uci.cu

Quiero agradecerle a mis padres, Marbelis y Eugenio, por el apoyo y la confianza depositada en mi cada minuto.

A mi madre que es la persona más importante de mi vida, gracias por el enorme esfuerzo realizado para que yo cumpliera mis sueños, por el apoyo en cada momento y gracias por estar siempre para mí.

A mi familia por los consejos que me dieron y por el apoyo brindado para poder seguir hacia adelante.

A todos los profesores que contribuyeron en mi formación profesional y personal durante estos cinco años.

A mis compañeros por haber estado siempre en los momentos buenos y malos.

A mis tutores por el tiempo y esfuerzo dedicado durante el desarrollo de la investigación.

A mi novia Martha Maricela Veliz Jaime y toda su familia por su cariño, dedicación y preocupación constantemente.

En fin, a todas las personas que de una forma u otra contribuyeron con el éxito del presente trabajo, les agradezco de todo corazón.

A Fidel Castro Ruz y a la revolución por la increíble oportunidad de estudiar en esta universidad de excelencia y por tanta confianza depositada en nosotros.

A todos los que de una forma u otra han estado ahí para lo que haga falta y de cierta manera han hecho que este sueño se haga realidad, especialmente a mis padres, a mi familia, a mis amigos, a mi novia y a todos los que se han encargado de formar parte de mi vida.

RESUMEN

La presente investigación surge como necesidad de construir un mercado de datos para las áreas de Cultura y Deporte de la Oficina Nacional de Estadística (ONE), con el propósito de apoyar el proceso de toma de decisiones que realizan los especialistas de la ONE a través del Sistema de Información de Gobierno (SIGOB). El proceso de almacenamiento y análisis de la información por parte de la ONE presenta muchas dificultades, debido a que no poseen una herramienta que permita la integración, centralización y disponibilidad de los datos, para que de esta forma se facilite el trabajo estadístico que se realiza en la institución. Como solución a esta problemática se plantea el desarrollo de un mercado de datos para las áreas de Cultura y Deporte que servirá de apoyo al proceso de toma de decisiones que se desarrolla en la ONE. Para el desarrollo de este mercado de datos se utilizó el ciclo de vida de la metodología de Ralph Kimball, complementada con lo planteado por Leopoldo Zenaido en su tesis de doctorado de guiar todo el proceso de desarrollo a través de casos de uso, y las herramientas utilizadas fueron: Visual Paradigm, PostgreSQL, PgAdmin III, Pentaho Data Integration, Schema Workbench y Pentaho BI Server. El mercado de datos fue aceptado por el cliente al cumplir con los requerimientos especificados y para validar la calidad del producto se realizaron un conjunto de pruebas, como las pruebas de integración y las pruebas de aceptación.

Palabras Claves: mercado de datos, toma de decisiones, ONE, SIGOB**.**

ÍNDICE

ÍNDICE DE FIGURA

ÍNDICE DE TABLA

INTRODUCCIÓN

La elección, desarrollo y uso de las tecnologías puede tener impactos variados en el quehacer humano. Hoy, cuando se habla de avances tecnológicos, no se puede dejar atrás el surgimiento de la computadora y el desarrollo que se ha ido adquiriendo en el campo de la Informática, este avance se evidencia en los diferentes campos de la sociedad, como el sector de la salud, educación y las áreas de cultura y deporte. Estas últimas áreas son las de mayor seguimiento por los usuarios por la versatilidad y dinamismo que poseen, por lo que traen gran cúmulo de información que necesita ser procesada, almacenada y transformada para un profundo análisis.

A nivel mundial la mayoría de las empresas han automatizado sus procesos o se encuentran involucradas en ello, dándole una mayor importancia al almacenamiento, manejo y uso de la información. Con esto se podría realizar un mejor análisis estadístico de los datos almacenados y una comparación de los resultados obtenidos por la empresa a través de su comportamiento histórico. De ahí el proceso de digitalización que se está llevando a cabo en cada una de las instituciones y órganos de trabajo con el objetivo de mejorar el proceso de análisis y manejo de la información.

Ante este desarrollo científico-técnico Cuba no se ha quedado al margen y ha ido madurando en estos temas con gran aceptación. En particular en el campo estadístico, sector encargado de recopilar, describir y analizar conjunto de datos para elevar o mantener la integridad de un país. Trabajo que nunca ha sido fácil y ha resultado ser un proceso muy tedioso sino se tienen los medios necesarios para manejar los grandes volúmenes de información que hay para procesar.

En Cuba se lleva a cabo la recopilación de información estadística de los diferentes sectores económicos a lo largo y ancho del país, proceso dirigido por la Oficina Nacional de Estadística (ONE), órgano rector que vela por el funcionamiento adecuado del sistema de información estadístico a nivel nacional dándole un tratamiento y difusión especial a los datos recogidos. Para ello se apoya en las entidades existentes en cada uno de los municipios y provincias del país encargadas de recopilar la información referente a su radio de acción, y de una serie de modelos estadísticos que guardan información referente a los campos de índole nacional. Además, la Universidad de Ciencias Informáticas (UCI), específicamente el Centro de Tecnologías de Gestión de Datos (DATEC) en conjunto con la ONE, están trabajando juntos para apoyar el proceso de toma de decisiones por parte del Sistema de Información de Gobierno (SIGOB).

Las tecnologías y herramientas que se usan actualmente en la ONE digitalizan la información recogida en formatos de difícil acceso para su consulta, información que solo puede ser consultada por especialistas del tema y con conocimiento del negocio, pues muchas veces los datos recogidos se encuentran codificados y se debe tener conocimientos relacionados con las temáticas en cuestión para

lograr entender el contenido de las fuentes. El proceso de obtención de la información por parte de la ONE varia en dependencia de las áreas de análisis, en algunas áreas la información se recoge mensualmente y en otras anualmente, esto provoca que existan grandes volúmenes de datos de diferentes zonas del país en disímiles formatos, trayendo como consecuencia dificultad a la hora de analizar los datos y lentitud en el proceso de toma de decisiones por parte de órganos del estado, además de dificultar los procesos de integración, centralización y disponibilidad de la información específicamente para las áreas de Cultura y Deporte, áreas que cuentan con grandes volúmenes de información. Con el objetivo de resolver este problema se desarrolló un trabajo previo "Análisis, Diseño e Implementación del mercado de datos para la Cultura y Deporte en la Oficina Nacional de Estadística", pero este solo fue enfocado al análisis y diseño de un mercado de datos para estas áreas. Por lo planteado anteriormente se define como **problema científico**: ¿Cómo apoyar a la toma de decisiones en las áreas de Cultura y Deporte del Sistema de Información de Gobierno?

Teniendo la presente investigación como **objeto de estudio** los almacenes de datos e inteligencia de negocio, enmarcado en el **campo de acción** mercado de datos y capa de visualización para las áreas de Cultura y Deporte del Sistema de Información de Gobierno.

Definiéndose como **objetivo general**: desarrollar el mercado de datos para las áreas de Cultura y Deporte del Sistema de Información de Gobierno que apoye a la toma de decisiones.

Para dar solución al objetivo general se plantearon los siguientes **objetivos específicos**:

- Refinar el análisis y diseño del mercado de datos de las áreas de Cultura y Deporte.
- Implementar el mercado de datos de las áreas de Cultura y Deporte.
- Validar el mercado de datos de las áreas de Cultura y Deporte.

Para darle solución a los objetivos específicos anteriormente mencionados se realizó la siguiente planificación de las tareas de investigación:

- Estudio y análisis de los almacenes de datos.
- Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.
- Refinamiento de los requisitos del almacén de datos.
- Refinamiento de las descripciones de los casos de uso del mercado de datos.
- Refinamiento de los hechos, las medidas y las dimensiones del mercado de datos.

- Refinamiento de la matriz dimensional
- Refinamiento del diseño del modelo de datos.
- Definición de la arquitectura del mercado de datos
- Diseño del subsistema de integración.
- Diseño del subsistema de visualización.
- Diseño de los casos de pruebas.
- Implementación del subsistema de integración.
- Implementación del subsistema de visualización.
- Aplicación de las listas de chequeo.
- Aplicación de los casos de pruebas.

El presente trabajo está estructurado en 4 capítulos:

Capítulo 1 Fundamentación teórica de los almacenes de datos

En este capítulo se abordará toda la fundamentación teórica del objeto de estudio, los principales conceptos, metodologías y herramientas para el desarrollo de un mercado de datos que permitirá a la ONE apoyar el proceso de toma de decisiones.

Capítulo 2 Análisis y diseño del mercado de datos Cultura y Deporte

En este capítulo se realizará el refinamiento al análisis y diseño realizado en el trabajo precedente.

Capítulo 3 Implementación del mercado de datos Cultura y Deporte

En este capítulo se realizarán todas las transformaciones necesarias como Extraer, Transformar y Cargar (ETL), proceso que deja todos los datos listos para la siguiente fase, Inteligencia del Negocio (BI); donde se realizará la implementación del modelo de datos, los cubos OLAP, las vistas de análisis y la política de seguridad de los usuarios.

Capítulo 4 Validación del mercado de datos Cultura y Deporte

En este capítulo se aplicarán los casos de prueba y las listas de chequeo para validar el mercado de datos.

CAPÍTULO 1: FUNDAMENTOS TEÓRICOS DE LOS ALMACENES DE DATOS

Introducción

En este capítulo se abordarán los principales conceptos, metodologías y herramientas para el desarrollo de un mercado de datos que permitirá a la ONE mejorar el proceso de toma de decisiones.

1.1 Almacén de datos

En una empresa no tener conocimiento sobre la información que en ella se maneja, resulta algo complicado, pero también resulta algo complejo poseer demasiada información y no saber cómo utilizarla, una solución para este problema son los almacenes de datos y la inteligencia de negocio.

Existen diversos autores que se han encargado de definir el término almacén de datos (Data Warehouse). Muchas de estas definiciones se concentran en los datos. La definición del Data Warehouse es más que datos, son también los procesos involucrados en obtener los datos desde las fuentes a las tablas y desde las tablas al analista. En otras palabras un Data Warehouse es el dato (metadato/ hecho/ dimensión/ agregación) y los administradores de proceso (carga/ almacén/ consulta) que hacen disponible la información, permitiendo a las personas tomar decisiones [1].

Un Almacén de Datos 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 [2].

La definición más difundida y aceptada de un almacén de datos pertenece a William H. Inmon, licenciado en Ciencias Matemáticas y máster en Ciencias de la Computación, además de ser un autor productivo en la construcción, uso y mantenimiento del almacén de datos y la fábrica de Información Corporativa fue el primero en acuñar el término. “Un Data Warehouse es un conjunto de datos orientados hacia una materia, integrados, no transitorios y que varían con el tiempo, los cuales apoyan el proceso de toma de decisiones de una administración [3].

Luego del estudio sobre las diferentes fuentes que hablan del tema, se puede decir que los almacenes de datos son un conjunto de datos integrados de diferentes fuentes, permitiendo el acceso a datos históricos y actuales, y permiten además el uso de distintas técnicas y herramientas de inteligencia de negocio para extraer el conocimiento de los datos almacenados y apoyar el proceso de toma de decisiones.

A continuación se mencionan las características, estructura, ventajas y desventajas de los almacenes de datos.

Características de los almacenes de datos

- Integrado: como los datos almacenados provienen de fuentes diferentes deben integrarse en una estructura consistente que elimine las inconsistencias existentes en los datos.

- Temático: los datos se organizan por los principales temas de la organización, esto es lo que lo diferencia de los sistemas operacionales que se basan en los procesos cotidianos de la organización.

- Histórico: el almacén contiene grandes volúmenes de información histórica sobre las actividades de la organización y va variando en el tiempo, recibiendo periódicamente nuevos datos.

- No volátil: la información que existe en un almacén de datos es permanente, los datos no cambian, esto garantiza que las consultas que se realicen sobre el almacén de datos siempre producirá el mismo resultado [3].

Estructura del almacén de datos

Los almacenes de datos están compuestos por una serie de elementos que definen, en su conjunto, el ambiente que estos poseen. Aunque cada desarrollo de un almacén de datos es diferente, debido a las especificaciones de cada organización, generalmente, cumplen con la realización de los componentes que a continuación se proponen [4].

- Sistemas de fuentes operacionales: estos son los sistemas que poseen las compañías o empresas para la gestión de sus transacciones diarias. Estas transacciones son almacenadas en los más diversos formatos, desde una base de datos relacional hasta cualquier tipo de ficheros, ya sea Excel, XML, DBF, texto plano, entre otros. Las prioridades principales de este componente es 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 manualmente debido a que no posean un sistema que las procese.

- Área de procesamiento (staging): es el área que almacena los datos temporalmente y realiza un conjunto de procesos comúnmente llamados de Extracción, Transformación y Carga (ETL). En este componente 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 toda esa información en bases de datos relacionales, generalmente, para realizar el análisis y procesamiento de los datos. Una vez los datos almacenados en bases de datos temporales se procede a su limpieza donde se detectan inconsistencias, duplicaciones, errores de formato e

inexistencias, estandarizándose la información almacenada en diferentes fuentes. Estas transformaciones son las que sirven de apoyo para realizar la carga de los datos, hacia el DW, en el área de presentación.

- 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 Mercados de Datos 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 Almacén de Datos. 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.

Ventajas del almacén de datos

- Integrar datos históricos sobre la actividad de la organización (o negocio) en un único repositorio.

- Analizar los datos del negocio desde la perspectiva de su evolución en el tiempo.

- Prever tendencias de evolución del negocio.

- Identificar nuevas oportunidades de negocio y tomar decisiones estratégicas.

- Reducir los costes materiales y humanos en la toma de decisiones [5].

Desventajas del almacén de datos

- Riesgo de fracaso en la construcción del sistema, al subestimar los costes de captura y preparación de los datos.
- Riesgo de fracaso en la construcción del sistema por cambios continuos en los requisitos de los usuarios.

- Problemas con la privacidad de los datos [5].

1.1.1 Mercado de datos

Un mercado de datos (Data Mart) es un subconjunto de datos de un almacén de datos relativos a los requisitos de un departamento o área de negocio concretos. Este subconjunto de datos puede funcionar de forma autónoma, o bien enlazado al almacén de datos. El motivo por el cual se crean mercados de datos es el crecimiento que tiene el Almacén y así facilitar su construcción y utilización [2].

Los mercados de datos presentan las mismas características que los almacenes de datos y su función es apoyar la toma de decisiones. Algunas características que lo diferencian de los almacenes de datos son:

- Se centran en los requisitos de los usuarios asociados a un departamento o área de negocio concretos.
- Son más sencillos a la hora de utilizarlos y comprender sus datos, debido a que la cantidad de información que contienen es mucho menor que en los Almacenes de Datos [2].

Ventajas de los mercados de datos

Dentro de las ventajas de aplicar un mercado de datos a un negocio, se han seleccionado las siguientes:

- 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 [6].

1.2 Tendencias actuales del uso de los mercados de datos

Mercados de datos a nivel mundial

Con la informatización de la sociedad y dentro de estas las empresas, ha crecido a nivel mundial la capacidad de generación y almacenamiento de la información. Muchas empresas han desarrollados sus propios mercados de datos, ya que las características que presentaban los sistemas relacionales no eran las más adecuadas para agilizar el proceso de toma de decisiones en comparación con los almacenes de datos. Ejemplos de ellos tenemos el Sistema de Información para Gestión Ambiental y Social en Colombia, que se encarga del seguimiento de las operaciones de crédito multinacional de la

CAF desde el punto de vista de la gestión ambiental y social. También cuentan en Colombia con un mercado de datos de Prepago encargado de analizar el uso y la utilidad generada por la plataforma de tarjetas prepago de Orbitel, a través de variables geográficas, alguna información técnica y los diferentes tipos de tarjetas de prepago [7].

Mercados de datos en Cuba

Las empresas cubanas en la época actual necesitan obtener de una forma cada vez más rápida la información necesaria para tomar decisiones. Con la utilización de las tecnologías de la información y las comunicaciones trata de alcanzar la mayor productividad posible y que los productos tengan una mejor calidad para obtener ventajas competitivas. El uso de mercados de datos es una de las estrategias para lograrlo. Ejemplos de tal desempeño:

La Empresa de Proyectos de Arquitectura e Ingeniería (EMPAI) de Matanzas posee un mercado de datos para el control de la información relevante del proceso de negocio de Gestión del Capital Humano para elevar la efectividad de la organización [8].

1.3 Etapas de desarrollo de un Almacén de datos

Para el proceso de desarrollo del almacén de datos se cuenta con tres etapas fundamentales:

- Análisis y diseño.
- Extracción, transformación y carga (ETL).
- Inteligencia de negocio.

1.3.1 Análisis y diseño

Toda fase inicial de un proyecto de gran alcance como son los almacenes de datos, debe contar con bases sólidas que le permitan avanzar hacia las otras fases de desarrollo con un mayor nivel de seguridad y organización. Se deben identificar los posibles riesgos. Debe contar además con un plan de acción efectivo que minimice los riesgos de fracaso y tener conocimiento de un grupo de elementos y herramientas que intervendrán en el desarrollo de dicho proyecto. Todo esto, sumado al suficiente dominio sobre el funcionamiento del proyecto y tener bien claro cuáles son los objetivos que se pretenden alcanzar, son las bases para comenzar el desarrollo de un almacén de datos.

Para un almacén de datos es prioritario reconocer las necesidades analíticas que tiene una organización y establecer los objetivos que se persiguen al desarrollar un almacén de datos en la organización en la que se desarrolle. Este es el primer paso a seguir, el análisis y la comprensión del nuevo entorno, el entorno analítico [3].

En cuanto al diseño del almacén, una arquitectura Data Warehouse establece el marco de trabajo, estándares y procedimientos a seguir para la construcción de un almacén de datos a nivel empresarial. El objetivo de las actividades de la arquitectura es simple, integrar al almacén de datos las necesidades de información empresarial [3]. Es Ralph Kimball quien propone la arquitectura mejor estructurada y de mejor comprensión para el diseño de un almacén.

De los resultados más importantes que se pueden obtener de la aplicación de esta arquitectura son:

- El modelo de datos fuente.
- El modelo de datos conceptual del almacén de datos.
- Arquitectura tecnológica Data Warehouse [3].

En cuanto al modelo de datos del almacén de datos se pueden encontrar dos formas a seguir:

Modelado relacional: se basa en el uso de Diagramas Entidad Relación (ERD).

Modelado multidimensional

Para el diseño de un almacén de datos, varios autores han seguido diferentes conceptos para abordar el modelado de un almacén, entre los que se encuentran, esquema en estrella, esquema en copo de nieve, esquema en constelación entre otros. Independientemente de la metodología seguida, todos ellos utilizan un modelo multidimensional de datos.

El modelado multidimensional es: modelos de datos como conjuntos de medidas descritas por dimensiones.

- Adecuado para resumir y organizar datos (p.ej. hojas de cálculo).
- Enfocado para trabajar sobre datos de tipo numérico.
- Más fácil de visualizar y entender que el modelado E/R [9].

Características del modelo multidimensional

La estructura básica de un almacén de datos para el modelo multidimensional está definida por dos elementos: esquemas y tablas.

Como cualquier base de datos relacional, el almacén de datos se compone de tablas, de dos tipos de tablas específicamente en el modelo multidimensional.

- Tablas fact o de hechos: tabla central en un esquema dimensional y en ella se guardan las medidas numéricas del negocio, por ejemplo: número de libros vendidos.

- Tablas lock- up o dimensionales: son las tablas que se alimentan de las tablas de hechos, por lo que contienen el detalle de los valores que se encuentran asociados a la tabla de hechos. [10]

La colección de tablas dentro del almacén se conoce como esquema y por lo general caen dentro de dos categorías básicas: esquema en estrellas y esquemas en copo de nieve.

Componentes del modelo multidimensional

El modelo multidimensional está compuesto por cubos, medidas, dimensiones, jerarquías, niveles, y atributos.

- Cubos: son los principales objetos del proceso analítico en línea (OLAP, Online Analytic Processing), una tecnología que proporciona rápido acceso a los datos de un almacén de datos, conjunto de datos que normalmente se construyen a partir de un subconjunto de un almacén de datos y se organiza y resume en una estructura multidimensional definida por un conjunto de dimensiones y medidas.

- Dimensiones: son un atributo estructural de los cubos y sirven como un mecanismo de selección de datos, organizadas en jerarquías de categorías y niveles que describen los datos de las tablas de hechos.

- Medidas: dentro del modelo multidimensional, las medidas o atributos numéricos describen un cierto proceso del mundo real, el cual va a ser proceso de un análisis.

- Jerarquías: es un conjunto de miembros de una dimensión y sus posiciones en relación con los demás, una forma de organizar los datos en diferentes niveles de agregación.

- Niveles: es un elemento de la jerarquía de dimensiones. Describen la jerarquía desde el nivel de datos más altos (más resumido) hasta el más bajo (más detallado).
- Atributos: provee información adicional acerca de los dato [9].

1.3.2 Procesos de extracción, transformación y carga de datos

En aquellos escenarios donde exista un almacén de datos, este necesita ser alimentado de datos ya existentes en otras fuentes, y para ello aparecen los procesos de ETL.

- **Extracción**: este es el primer paso para la obtención de la información de las diferentes fuentes hacia el almacén de datos. Como los datos provienen de fuentes distintas, generalmente se encuentran en formatos distintos y organizados a la forma de cada organización, la extracción deja los datos en un formato listo para la transformación.

- **Transformación**: una vez que la información es extraída, el proceso de transformación se encarga de preparar los datos de la manera adecuada para integrarlos en el almacén de datos. Para ello este proceso se compone de algunas actividades como: limpieza de datos, integración de formato, integración de datos.

- **Carga**: una vez que los datos han sido extraídos de las diferentes fuentes y transformados, se procede a cargar el almacén de datos, después de la carga inicial se procede a mantener el almacén actualizándolo periódicamente [11].

1.3.3 Inteligencia de negocio (Business Intelligence)

La gran mayoría de las organizaciones poseen abundancia de información pero carecen de métodos y herramientas que permitan el uso eficiente de la información. La inteligencia de negocios es la clave para la solución de dicho problema, ya que posee vías para mejorar el proceso de toma de decisiones y convertir esto en una ventaja competitiva.

La inteligencia de negocios o business intelligence (BI) se puede definir como el proceso de analizar los bienes o datos acumulados en la empresa y extraer una cierta inteligencia o conocimiento de ellos [12].

El proceso de inteligencia de negocio apoya a los usuarios que toman decisiones con la información correcta o que ellos necesitan, en el momento correcto y lugar correcto, lo que le facilita tomar mejores decisiones de negocios.

Una de las herramientas que permite una mejor visualización de los datos a los usuarios son las herramientas OLAP a través de reportes y gráficos.

Proceso Analítico en Línea

El término OLAP (Proceso Analítico en Línea), fue introducido en el año 1993 por E. F. Codd, para referirse a nuevas herramientas para el análisis de datos. Las herramientas OLAP se basan en el modelo multidimensional de datos, es decir presentan a los usuarios una visión multidimensional de los datos, independientemente del servidor que soporte el almacén de datos [5].

Estas herramientas poseen un grupo de ventajas frente a las herramientas que se basan en el modelo OLTP, dentro del grupo de herramientas que se basan en este modelo podemos encontrar: archivos de texto, hojas de cálculo y las bases de datos transaccionales. A continuación se muestra una tabla que refleja las diferencias entre ambas tecnologías.

Tabla 1. Diferencias entre OLAP y OLTP [10]

OLTP	OLAP	
Objetivo	Control de procesos operacionales	Toma de decisiones
Datos	Atómicos, actualizados y dinámicos	Consolidados, históricos y estables
Estructura	Normalizada	Dimensional
Tiempo de Respuesta	Segundos	De segundos a minutos
Acceso	Alto	Moderado a bajo

Modos de almacenamiento

Existen tres formas para almacenar la información de los cubos:

Herramientas ROLAP (Relational On-line Analytical Process)

Toda la información de los cubos es almacenada en una base de datos relacional, no realiza una copia de la base de datos, cuando necesita información accede directamente a las tablas originales, generalmente es mucho más lenta que las otras dos vías de almacenamiento [13].

Herramientas MOLAP (Multidimensional On-line Analytical Process)

Los datos fuentes del cubo son almacenados con sus agregaciones en una estructura multidimensional de alto rendimiento. Provee excelente rendimiento y compresión de los datos, tiene el mejor tiempo de respuesta, muy adecuado para cubos por su rápida respuesta [13].

Herramientas HOLAP (Hybrid On-line Analytical Process)

Combina elementos de MOLAP y ROLAP, las agregaciones de los datos son almacenadas en una estructura multidimensional usada por MOLAP y la base de datos fuentes en una base de datos relacional. Los cubos almacenados como HOLAP, son más pequeños que los MOLAP y responden más rápido que los ROLAP. Es generalmente usado para cubos que requieran rápida respuesta, para sumarizaciones basadas en grandes cantidades de datos [13].

Para el desarrollo del mercado de datos se escogió como modo de almacenamiento de datos al Procesamiento Analítico Relacional (ROLAP). Porque los usuarios finales pueden realizar sus análisis multidimensionales a través del motor ROLAP, que transforma dinámicamente sus consultas a consultas SQL, luego estas consultas se ejecutan en las bases de datos relacionales y sus resultados

se relacionan mediante tablas cruzadas y conjuntos multidimensionales para devolver los resultados a los usuarios. Además esta arquitectura accede directamente a los datos almacenados en el almacén de datos y soporta técnicas de optimización de accesos para mejorar el tiempo de respuesta de las consultas. También el tipo de gestor de BD a utilizar exige que sea este el modelo de almacenamiento a utilizar.

1.4 Metodologías para el desarrollo de un almacén de datos

Una metodología es aquella guía que se sigue a fin 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, disciplinada y con cierta disciplina [14].

En este mundo de los almacenes de datos existen dos tendencias que han marcado bien sus pautas frente a las demás, sirviendo de guía a la comunidad mundial en cuanto a este tema. Estas tendencias son llamadas como Metodología de Kimball y Metodología de Inmon, en honor a sus creadores Ralph Kimball y William H. Inmon. Dos de las personalidades referentes y más influyentes en el área de los almacenes de datos, Kimball especialista en el diseño de almacenes de datos y creador del enfoque multidimensional y Inmon creador del término almacenes de datos y considerado el padre de la disciplina.

La diferencia entre ambas tendencias viene dada por el enfoque que le da cada uno al problema.

El enfoque que propone Inmon se basa en un enfoque Top-Down (Descendentemente), este enfoque propone construir primero el almacén de datos y a partir de este construir los mercados de datos. Propone la construcción de un repositorio de datos corporativo como fuente de información consistente, consolidada, histórica y de calidad. Como el almacén de datos se construye descendentemente los mercados de datos se nutren del almacén corporativo, convirtiéndose en un complejo empresarial de bases de datos relacionales [15].

1.4.1 Ciclo de vida de Kimball

La metodología de Kimball se enfoca principalmente en el diseño de la base de datos que almacenará la información para la toma de decisiones. El diseño se basa en la creación de tablas de hechos (Facts) que son tablas que contienen la información numérica de los indicadores a analizar, es decir la parte cuantitativa de la información [16].

En comparación con el enfoque que propone Inmon, la metodología de Kimball propone una arquitectura Bottom-up (Ascendentemente), plantea que se deben construir primeramente los mercados de datos orientados al tema de su departamento y luego el almacén de datos sería la unión de estos. Kimball propone dividir el mundo de la inteligencia de negocio entre los hechos y las

dimensiones, esta es eficaz y conduce a una solución completa en un corto período de tiempo. Además, tiene abundante documentación y se puede encontrar una respuesta a casi todas las preguntas que se puedan tener [15]. Lo que es factible para el desarrollo de proyectos donde se quiera alcanzar éxito, ya que el otro modelo puede tardar mucho más y no llegar a finalizar el proyecto.

1.4.2 Metodología a utilizar para el desarrollo del mercado de datos en el centro DATEC

Existen diversas metodologías pera la construcción de un almacén de datos y un mercado de datos, pero el centro definió trabajar con el ciclo de vida de la metodología de Kimball por los siguientes elementos:

- Identifica la tabla de hechos y la tabla de dimensiones, lo cual agiliza el proceso de desarrollo y con ello la toma de decisiones.

- Propone la construcción de mercados de datos departamentales y después el almacén de datos, esto trae como ventaja, que la creación y la puesta en marcha de los mercados de datos se produce en un lapso de tiempo corto y después se valora si se construye o no el almacén de datos.

- Existe amplia documentación de la misma, así cualquier duda que exista puede ser atendida rápidamente.

- Es una metodología madura y reconocida por los usuarios dedicados al tema, además de tener bien definidas sus etapas, actividades, roles y artefactos [15].

Para complemento de la misma y por las características de trabajo de la universidad se tomó lo planteado por Leopoldo Zenaido Zepeda en su tesis de doctorado, incluir los casos de uso para guiar el proceso de desarrollo, y así lograr estar más alineados a las tendencias y normas de la universidad.
Flujos de trabajo que presenta esta metodología [15]:

- **Estudio preliminar o planeación**: se realiza el estudio de la entidad cliente para determinar lo que se desea construir y las condiciones que existen para el desarrollo de la misma, la planeación del proyecto, se definen los objetivos, el alcance preliminar, los costos estimados y otras series de actividades.

- **Requerimientos**: se realiza en dos direcciones, una, identificando las necesidades de información y reglas del negocio; y la otra con un levantamiento detallado de las fuentes de datos a integrar. Es aquí donde se definen los requerimientos a través de la comparación de las necesidades y las reglas del negocio.

- **Arquitectura y diseño**: aquí se definen las estructuras de almacenamiento, se diseñan las reglas de extracción, transformación y carga, así como la arquitectura de información que regirá el desarrollo de la solución.

- **Implementación**: se lleva a cabo el diseño físico del repositorio de datos, se crean las estructuras de almacenamiento, el área temporal de almacenamiento, se ejecutan las reglas de ETL y se configuran e implementan las herramientas de BI para la obtención de los requerimientos acordados con el cliente.

- **Prueba**: se realizan las pruebas de unidad, luego las pruebas de integración y sistema, hasta las pruebas de aceptación con el cliente final.

- **Despliegue**: consta de dos etapas, despliegue piloto en el cual se configuran los servidores y se instalan las herramientas según la arquitectura definida y se carga una muestra de los datos para demostrar al cliente que el sistema funciona. Posterior a la aceptación del cliente se realiza la carga histórica de los datos y la capacitación y transferencia tecnológica.

- **Soporte y mantenimiento**: después de haber implantado la solución, se brindan los servicios de soporte en línea, vía telefónica, web u otros, hasta el acompañamiento junto al cliente según el contrato firmado y las condiciones de soporte establecidas.

- **Gestión y administración del proyecto**: se lleva durante todo el ciclo de vida, es aquí donde se controla, gestiona y chequea todo el desarrollo, los gastos, las utilidades, los recursos, las adquisiciones, y demás actividades relacionadas con la gestión del proyecto.

1.5 Herramientas de modelado

Visual Paradigm 6.4

Se puede definir a las Herramientas CASE como un conjunto de programas y ayudas que dan asistencia a los analistas, ingenieros de software y desarrolladores, durante todos los pasos del Ciclo de Vida de desarrollo de un Software [17].

Ejemplo de esta herramienta se tiene el Visual Paradigm que una herramienta UML profesional que soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue. El software de modelado UML ayuda a una más rápida construcción de aplicaciones de calidad, mejores y a un menor coste. Permite dibujar todos los tipos de diagramas de clases, código inverso, generar código desde diagramas y generar documentación [18].

A continuación se enuncian algunas de sus características [19]:

- Soporte de UML 2.1.
- Ingeniería inversa.
- Generación de código.
- Importación y exportación de ficheros XMI.
- Integración con Visio.
- Disponibilidad de integrarse en los principales IDEs.
- Disponibilidad en múltiples plataformas.
- Distribución automática de diagramas.

Se seleccionó Visual Paradigm porque es una herramienta UML que soporta el ciclo de vida completo de desarrollo de un software. Permite dibujar todos los tipos de diagramas de clases, generación de código y la integración con los principales IDEs. También presenta licencia gratuita y comercial y la universidad cuenta con la licencia para su uso, es fácil de instalar y actualizar y permite la compatibilidad entre ediciones.

1.6 Herramientas informáticas para el proceso de extracción, transformación y carga

Kettle 4.1.0

El Pentaho Data Integration (PDI), antes llamado Kettle, es una de las soluciones más extendidas y mejor valoradas del mercado. Permite realizar transformaciones y trabajos de una forma muy sencilla e intuitiva. Igualmente los proyectos realizados con Data Integration son muy fáciles de mantener [20].
Se compone de 4 herramientas [21]:

- SPOON: permite diseñar de forma gráfica la transformación ETL.
- PAN: ejecuta las transformaciones diseñadas con SPOON.
- CHEF: permite, mediante una interfaz gráfica, diseñar la carga de datos incluyendo un control de estado de los trabajos.
- KITCHEN: permite ejecutar los trabajos batch diseñados con Chef.

A parte de ser open source y sin costes de licencia, las características básicas de esta herramienta son [21]:

- Entorno gráfico de desarrollo.
- Uso de tecnologías estándar: Java, XML, JavaScript.
- Fácil de instalar y configurar.
- Multiplataforma: Windows, Macintosh, Linux.
- Basado en dos tipos de objetos: Transformaciones (colección de pasos en un proceso ETL) y trabajos (colección de transformaciones).

DataCleaner 1.5

DataCleaner es una aplicación Open Source (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. Es la alternativa gratuita al software de gestión de datos maestros (MDM), almacenamiento de datos (DW), proyectos de investigación estadística, la preparación para la extracción, transformación y carga (ETL) de actividades y mucho más.

Está licenciado bajo los términos de la Licencia Pública General Menor (LGPL), que permite que cualquiera pueda utilizar el software para todos los efectos, pero ninguna de las modificaciones realizadas en el código debe ser aportado a la comunidad. Permite sin tener que esforzarte mucho obtener una visión perspicaz de los datos [22].

También puede acceder y analizar prácticamente cualquier almacén de datos, incluyendo [21]:

- Bases de datos como Oracle, Microsoft SQL Server, PostgreSQL, MySQL, Open Office (ODB) y más.
- Los archivos separados por comas y separados por tabuladores (.csv / .tsv).
- Hojas de cálculo Excel (. xls).
- Archivos XML.

1.7 Herramientas informáticas para la inteligencia de negocios, aplicando técnicas OLAP

El líder actual en cuanto a soluciones de inteligencia de negocio es el Pentaho, creado en el 2004, el mismo ofrece sus soluciones propias y cuenta con una amplia variedad de recursos para desarrollar,

mantener y explotar un proyecto de inteligencia de negocio. Para ello ha integrado diferentes proyectos ya existentes y de solvencia propia, como el Kettle (Data Integration) y el Mondrian [20].

Aplicación Web: Pentaho BI Server 3.6.0

El B.I. Server de Pentaho es una aplicación 100% Java2EE que permite gestionar todos los recursos de BI. Cuenta con una interfaz de usuario donde se encuentran disponibles todos los informes, vistas OLAP y cuadros de mando. Cuenta con una consola de administración que permite gestionar y supervisar la aplicación y los usuarios [20].

OLAP: Mondrian 3.0.4

Online Analitical Processing es la tecnología que permite organizar la información en una estructura dimensional que proporcionará la posibilidad de moverse por la información desplazándonos por sus dimensiones.

Mondrian es el motor OLAP de Pentaho, aunque puede ser integrado independientemente en cualquier otra plataforma, y de hecho es el componente, junto con Data Integration que más se utiliza independientemente. Es un motor Hybrid OLAP que combina la flexibilidad de los motores ROLAP con una caché que le proporciona velocidad.

Es una de las aplicaciones más importantes de la plataforma Pentaho BI. Es un servidor OLAP open source que gestiona comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuente, es decir, Mondrian actúa como JDBC para OLAP [20].

Servidor Apache Tomcat 5.5.5

Tomcat (también llamado Jakarta Tomcat o Apache Tomcat) funciona como un contenedor de servlets desarrollado bajo el proyecto Jakarta en la Apache Software Foundation. Implementa las especificaciones de los servlets y de JavaServer Pages (JSP) de Sun Microsystems. Es un servidor web con soporte de servlets y JSPs. Incluye el compilador Jasper, que compila JSPs convirtiéndolas en servlets. El motor de servlets de Tomcat a menudo se presenta en combinación con el servidor web Apache [23].

Puede funcionar como servidor web por sí mismo. En sus inicios existió la percepción de que el uso de Tomcat de forma autónoma era sólo recomendable para entornos de desarrollo y entornos con requisitos mínimos de velocidad y gestión de transacciones. Hoy en día ya no existe esa percepción y Tomcat es usado como servidor web autónomo en entornos con alto nivel de tráfico y alta disponibilidad. Dado que Tomcat fue escrito en Java, funciona en cualquier sistema operativo que disponga de la máquina virtual Java [23]. Es el servidor Web más utilizado a la hora de trabajar con Java en entornos web; también puede especificarse como el manejador de las peticiones de JSP y

servlets recibidas por servidores Web populares, como el servidor Apache HTTP de la Fundación de software de Apache o el servidor Microsoft Internet Information Server (IIS) [24].

Características

- Implementado de Servlet 2.5 y JSP 2.1
- Soporte para Unified Expression Language 2.1
- Diseñado para funcionar en Java SE 5.0 y posteriores

Schema Workbench 3.2.1

El esquema Mondrian Workbench es un entorno visual para el desarrollo y prueba de cubos OLAP Mondrian. El motor de Mondrian procesa las solicitudes de MDX. Estos archivos son los modelos de esquemas XML de metadatos que se crean en una estructura específica utilizada por el motor de Mondrian. Estos modelos XML se pueden considerar las estructuras de forma de cubo que utilizan HECHO existentes y tablas de dimensiones que se encuentran en su RDBMS.

Ofrece las siguientes funcionalidades [25]:

- Editor de esquema integrado con el origen de datos subyacente para su validación
- Permite la ejecución de consultas MDX contra el esquema y la base de datos y la navegación por la base de datos subyacente

1.8 Sistema Gestor de Base de Datos

PostgreSQL 8.4

PostgreSQL se caracteriza por ser un sistema estable, de alto rendimiento, gran flexibilidad ya que funciona en la mayoría de los sistemas Unix, además tiene características que permiten extender fácilmente el sistema. Puede ser integrado al ambiente Windows permitiendo de esta manera a los desarrolladores, generar nuevas aplicaciones o mantener las ya existentes. Permite desarrollar o migrar aplicaciones desde Access, Visual Basic, FoxPro, Visual FoxPro, C/C++ Visual C/C++, Delphi, etc., para que utilicen a PostgreSQL como servidor de BD [26].

Existen otras ventajas que hacen que PostgreSQL sea una alternativa a tomar seriamente [27]:

- Instalación ilimitada: con PostgreSQL, nadie puede demandarlo por violar acuerdos de licencia, puesto que no hay costo asociado a la licencia del software.
- Soporte: además de ofertas de soporte, existe una importante comunidad de profesionales y entusiastas de PostgreSQL de los que su compañía puede obtener beneficios y contribuir.

- Ahorros considerables en costos de operación: PostgreSQL ha sido diseñado y creado para tener un mantenimiento y ajuste mucho menor que otros productos, conservando todas las características, estabilidad y rendimiento.

- Extensible: El código fuente está disponible para todos sin costo. Si su equipo necesita extender o personalizar PostgreSQL de alguna manera, pueden hacerlo con un mínimo esfuerzo, sin costos adicionales. Esto es complementado por la comunidad de profesionales y entusiastas de PostgreSQL alrededor del mundo que también extienden PostgreSQL todos los días.

PgAdmin III

PgAdmin3 es una aplicación gráfica para gestionar el gestor de bases de datos PostgreSQL, siendo la más completa y popular con licencia open source. Está escrita en C++. PgAdmin3 está diseñado para responder a las necesidades de todos los usuarios, desde escribir consultas SQL simples hasta desarrollar bases de datos complejas. El interfaz gráfico soporta todas las características de PostgreSQL y facilita enormemente la administración. La aplicación también incluye un editor SQL con resaltado de sintaxis, un editor de código de la parte del servidor, un agente para lanzar scripts programados, soporte para el motor de replicación Slony-I y mucho más [28].

Se seleccionó PostgreSQL teniendo en cuenta sus características, además de ofrecer estabilidad, rendimiento y la eficiencia de un sistema operativo. Por otra parte posee mejor rendimiento en el trabajo con grandes volúmenes de datos como son los almacenes de datos y existe amplio respaldo por la comunidad de PostgreSQL para la obtención de información y uso de la herramienta.

1.9 Conclusiones

Lo planteado en este capítulo servirá de base para desarrollar un mercado de datos que apoye a la toma de decisiones en las áreas de Cultura y Deporte del Sistema de Información de Gobierno, para ello se realizó una caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de almacenes de datos.

Se utilizará como modelo de desarrollo el ciclo de vida de la metodología de Kimball complementada con la tesis de doctorado de Leopoldo Zepeda Zenaido para agilizar el proceso de desarrollo y adecuarnos a las tendencias de trabajo de la universidad. Se seleccionó como modelo de almacenamiento de datos ROLAP ya que el sistema gestor de base de datos que se utilizará será PostgreSQL, por ser una Base de datos Objeto - Relacional. Como herramienta para el proceso de ETL se seleccionó el Pentaho Data Integration (PDI) o Kettle por la facilidad de uso, mantenimiento y flexibilidad a la hora de realizar las transformaciones. Para la realización del diseño se utilizará Visual

Paradigm y para el proceso de inteligencia de negocio se utilizarán las herramientas Schema Workbench y el Pentaho BI Server.

CAPÍTULO 2: ANÁLISIS Y DISEÑO DEL MERCADO DE DATOS CULTURA Y DEPORTE

2.1 Introducción

En este capítulo se abordarán un grupo de elementos que se deben tener en cuenta en el desarrollo de la solución y para un mejor entendimiento del negocio. Se definirán los temas de análisis, roles, permisos y necesidades del usuario. También se definirán los requisitos de información, requisitos funcionales y no funcionales, requisitos multidimensionales, así como los casos de uso del sistema. Se confeccionará además el modelo de datos y la matriz bus, se identificarán los hechos y dimensiones del mercado de datos.

2.2 Descripción del negocio

La Oficina Nacional de Estadística órgano estadístico encargado de velar por el funcionamiento adecuado del sistema estadístico nacional, recibe grandes cantidades de información de diferentes sectores y áreas del país, entre ellos se encuentran las áreas de Cultura y Deporte. La recogida de información referente a estas áreas se realiza mediante una serie de modelos estadísticos definidos, y provienen de organismos y empresas distintas como son la UNESCO y el Instituto Nacional de Deporte, Educación Física y Recreación (INDER). Esta información puede ser recogida anualmente y por lo general son almacenadas en hojas de cálculos o como se conocen excel, esto provoca tardanza en la elaboración de informes, por lo tanto es costoso en tiempo y esfuerzo, además de retrasar el proceso de recuperación de los datos almacenados ya que son múltiples versiones. El análisis de esta información por los especialistas de la ONE constituye un proceso tedioso, ya que son grandes volúmenes de datos y varios indicadores a analizar, como son el análisis del comportamiento de las instalaciones culturales en el país, el comportamiento de la Feria Internacional del Libro, entre otros, haciendo tardía la respuesta al ministerio de Cultura. Por otro lado, en el área de Deporte, dentro de los indicadores que se analizan se encontraron los siguientes, el análisis del comportamiento de las instalaciones deportivas en el país, los títulos ganados por Cuba en eventos competitivos, las olimpiadas celebrados por deportes en el país, entre otros temas de análisis que son de interés para muchas instituciones, pero en especial para el INDER ya que es el encargado de esta área en el país.

2.3 Necesidades del usuario

Las necesidades de información son las necesidades que presentan los usuarios de la ONE en las diferentes áreas de la organización, específicamente se muestran las relacionadas con las áreas de Cultura y Deporte, partiendo de los temas de análisis de cada área, los cuales son:

Cultura

- Análisis del comportamiento de los títulos de libros publicados.

- Análisis del comportamiento de los libros y folletos publicados.
- Análisis del comportamiento de las instalaciones culturales en el país.
- Análisis del comportamiento de las ofertas culturales en el país.
- Análisis del comportamiento de los grupos profesionales.
- Análisis del comportamiento de las emisoras radiales y canales de televisión.
- Análisis del comportamiento de las horas de emisión por televisión.
- Análisis del comportamiento de los videos club juvenil.
- Análisis del comportamiento de los joven club de computación y palacios de computación.
- Análisis del comportamiento de las producciones cinematográficas.
- Análisis del comportamiento de la Feria Internacional del Libro en el país.

Deporte

- Análisis del comportamiento de las instalaciones deportivas.
- Análisis del comportamiento de los títulos ganados por Cuba en eventos competitivos.
- Análisis del comportamiento de los mundiales celebrados en Cuba.
- Análisis del comportamiento de las olimpiadas del deporte cubano.
- Análisis del comportamiento de los practicantes sistemáticos del deporte participativo y de alto rendimiento.
- Análisis del comportamiento de los participantes en competencias deportivas.
- Análisis del comportamiento de los festivales promovidos por el INDER.
- Análisis del comportamiento del personal deportivo pedagógico en el país.
- Análisis del comportamiento de los participantes en Juegos Paralímpicos Nacionales.

2.4 Especificación de requerimientos

Después de realizar un análisis de los principales indicadores y temas de análisis para las áreas de Cultura y Deporte se identificaron los requerimientos necesarios para satisfacer las necesidades del cliente.

2.4.1 Requisitos de información

Los requisitos de información coinciden con las necesidades de los usuarios por lo tanto es información que debe estar disponible para su consulta. A continuación se muestran algunos de los requisitos de información identificados durante el proceso de análisis, clasificados según los temas de análisis definidos, los restantes se encuentran en el expediente de proyecto del mercado de datos de Cultura y Deporte en el artefacto “Especificación de requerimientos”.

Cultura

- Obtener la cantidad de títulos de libros editados por tipo de materia, división política administrativa (DPA), indicador, unidad de medida (UM) y año.
- Obtener la cantidad de libros y folletos editados por tipo de publicación, DPA, indicador, UM y año.
- Obtener la cantidad de producciones cinematográficas por DPA, indicador, UM y año.
- Obtener la cantidad de grupos profesionales por tipo de manifestación artística, DPA, indicador, UM y año.
- Obtener la cantidad de ofertas artísticas por tipo de instalación cultural, DPA, indicador y año.
- Obtener la cantidad de emisoras por provincia, nivel, indicador y año.
- Obtener la cantidad horas de emisión de radio por provincia, nivel, indicador y año.
- Obtener la cantidad horas de emisión de radio de las emisoras provinciales por provincia, nivel, indicador y año.
- Obtener la cantidad de canales de televisión por provincias, nivel, indicador y año.
- Obtener la cantidad de canales municipales por provincia, nivel, indicador y año.
- Obtener la cantidad de salas de videos club juvenil por año, DPA, indicador, UM.
- Obtener la cantidad de expositores por año, evento cultural, DPA, indicador y UM.

Deporte

- Obtener la cantidad de instalaciones deportivas por provincia, indicador, UM y año.
- Obtener la cantidad de personal deportivo pedagógico por tipo de personal, año, DPA, indicador y UM.
- Obtener la cantidad de practicantes sistemáticos del deporte por tipo de practicante, DPA, indicador, UM y año.
- Obtener la cantidad de participantes de mayores en competencias deportivas por deportes, años, sexo, indicador, UM y nivel.
- Obtener la cantidad participantes en olimpiadas del deporte cubano por eventos deportivos, deporte, DPA, indicador y UM.
- Obtener la cantidad de títulos ganados por Cuba por tipo de evento deportivo, deporte, indicador y UM.
- Obtener la cantidad de campeonatos mundiales celebrados en Cuba por deporte, DPA, indicador, UM y año.
- Obtener la cantidad de participantes en los juegos paralímpicos nacionales por evento deportivo, DPA, indicador, UM y deporte.

- Obtener la cantidad de Carreras MaraCuba efectuadas en festivales promovidos por el INDER por año, DPA, indicador y UM.

2.4.2 Requisitos funcionales

Los requisitos funcionales son funcionalidades que debe cumplir la herramienta a desarrollar, de acuerdo con las necesidades y especificaciones del cliente. Entre los identificados se encuentran:

RF 1. Extraer datos de los sistemas fuentes.

RF 2. Transformar y cargar datos de los sistemas fuentes.

RF 3. Autenticar usuario.

RF 4. Adicionar usuario.

RF 5. Eliminar usuario.

RF 6. Adicionar rol.

RF 7. Eliminar rol.

RF 8. Adicionar reporte.

RF 9. Eliminar reporte.

RF 10. Modificar reporte.

RF 11. Configurar elementos del cubo.

RF 12. Ordenar resultados.

RF 13. Ocultar repeticiones.

RF 14. Mostrar padres.

RF 15. Mostrar propiedades.

RF 16. Suprimir filas vacías.

RF 17. Invertir eje.

RF 18. Detallar miembros.

RF 19. Entrar en detalles.

RF 20. Mostrar datos de origen.

RF 21. Mostrar gráfico.

RF 22. Configurar gráfico.

RF 23. Modificar las consultas MDX.

RF 24. Exportar reporte a pdf.

RF 25. Exportar reporte a Excel.

RF 26. Imprimir reporte.

2.4.3 Requisitos no funcionales

Los requisitos no funcionales son condiciones que el sistema debe cumplir para darle cumplimiento a los requisitos de información en el mercado de datos, por lo que se identificaron los siguientes:

Requerimientos de usabilidad

RNF 1. Cumplir con las pautas de diseño de las interfaces.

El sistema debe tener una interfaz gráfica uniforme que incluya pantallas, menús y opciones. Las pautas de diseño se realizarán siguiendo la arquitectura de información definida.

RNF 2. Mostrar los mensajes, títulos y demás textos que aparezcan en la interfaz del sistema en idioma español.

Los títulos de los componentes de la interfaz, los mensajes para interactuar con los usuarios y los mensajes de error, deben ser en idioma español y tener una apariencia uniforme en todo el sistema. Los mensajes de error deberán ser lo suficientemente informativos para dar a conocer la severidad del error.

RNF 3. Agilizar el acceso a los reportes del almacén de datos mediante la distribución de la información por áreas de análisis.

El usuario podrá acceder de manera rápida a la información que solicita en el área correspondiente de acuerdo al objetivo de su solicitud.

Requerimientos de fiabilidad

RNF 4. Asegurar la disponibilidad del sistema.

El sistema debe estar disponible durante el horario de trabajo. En caso de fallo, la recuperación del servicio no deberá ser de un período de tiempo muy prolongado.

RNF 5. Asegurar la recuperación ante un fallo.

El sistema debe ser capaz de recuperarse ante un fallo, teniendo en cuenta la complejidad y naturaleza de éste. El tiempo para su correcta recuperación fluctúa entre 10 minutos y 72 horas. Este tiempo comprende la solución al problema, así como su validación y prueba.

RNF 6. Garantizar la persistencia de la información.

Se debe realizar un respaldo total de los datos del almacén de datos con una frecuencia anual. Esta información se almacenará en el edificio correspondiente a la oficina de estadísticas de La Habana y será responsabilidad del grupo de administración de redes de la ONE.

Requerimientos de eficiencia

RNF 7. Garantizar la conexión de múltiples usuario al mismo tiempo.

El sistema debe permitir que existan varios usuarios conectados de forma simultánea.

Requerimientos de soporte

RNF 8. Lograr la homogeneidad de la estructura de los elementos definidos en el almacén.

Las estructuras del almacén de datos deben tener un nombre estándar teniendo en cuenta el tipo de estructura que sea. En la siguiente tabla se definen convenciones de nombrado con el objetivo de manejar un vocabulario común en todo el almacén de datos, permitiendo un entendimiento claro y conciso por parte de los desarrolladores.

Tabla 2. Convenciones de nombrado

Estructura	Descripción	Ejemplo
Tablas de hechos	Todas las tablas de hechos tendrán una cadena que demuestra que son hechos y el concepto que describen.	hech_<concepto>
Tablas de dimensiones	Todas las tablas de dimensiones tendrán una cadena que demuestra que son dimensiones y el concepto que describen.	dim_<concepto>
Llaves primarias	Todas las llaves primarias tendrán una cadena que demuestra que son llaves	<nombre_tabla>_id

	primarias y el nombre de la tabla a la que pertenecen.	
Atributos compuestos	En los atributos donde el nombre es compuesto se debe especificar el primer componente del atributo separado del segundo por un carácter de _.	<Primer nombre>_<Segundo nombre>

Requerimientos de restricciones de diseño

RNF 9. Utilizar los lenguajes de programación definidos durante la investigación.

Como lenguaje dentro del sistema gestor de base de datos para la programación en el almacén de datos se utilizará PL/pgSQL. En la implementación de los procesos de integración de datos se utilizará el lenguaje JavaScript. También se hará uso del lenguaje MDX para realizar las consultas.

RNF 10. Utilizar el Sistema Gestor de Base de Datos definido durante la investigación.

El gestor de base de datos que se utilizará es PostgreSQL y como interfaz de administración de dicho gestor PgAdmin.

RNF 11. Utilizar la herramienta de integración de datos definida durante la investigación.

Para el proceso de integración de datos se usará la herramienta Pentaho Data Integrator.

RNF 12. Utilizar las herramientas para la implementación de la capa de inteligencia de negocios definidas durante la investigación.

De la suite Pentaho, se usarán los siguientes componentes:

- Schema Workbench: herramienta gráfica que se utiliza para construir el esquema multidimensional que soportará la creación de los reportes multidimensionales.
- Pentaho BI Server: servidor que se encarga de visualizar los reportes, tableros de control digital, controlar el acceso a la información y unificar en una solución de inteligencia de negocios el uso de las demás herramientas que componen la suite.
- Pentaho Administrator Console: herramienta para administrar el Pentaho BI Server, que permite la administración de las conexiones a las bases de datos, tareas programadas así como los roles y usuarios.

Para el uso de las herramientas anteriores se requiere la instalación de la máquina virtual de java (Java Virtual Machine 6.0).

Requerimientos para la documentación de usuarios y ayuda del sistema

RNF 13. Confección de un manual de usuario.

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

Interfaz

RNF 14. Acceso al sistema.

El usuario deberá acceder a la aplicación a mediante el protocolo HTTP, usando preferiblemente el navegador web Firefox 2.0 en adelante.

Interfaces de usuario

RNF 15. Garantizar una interfaz amigable al usuario.

El sistema debe tener una interfaz amigable y sencilla de utilizar, teniendo en cuenta que los usuarios finales no son personas adiestradas en el campo de la informática.

Interfaces de hardware

RNF 16. Definir las interfaces de hardware que soportará el sistema.

El sistema podrá interactuar solamente con una interfaz de hardware: la impresora. Esta interacción se ocasionará cuando se necesite imprimir un reporte en formato físico. El acceso a la impresora será mediante el protocolo TCP/IP a través de la interfaz que ofrece el hardware.

RNF 17. Proporcionar características mínimas de hardware a las estaciones de trabajo.

Características de un cliente ligero.

RNF 18. Proporcionar características mínimas de hardware a los servidores.

Para lograr una explotación aceptable del sistema los servidores deben contar con los siguientes requerimientos de hardware:

- Windows XP o cualquier otro.
- 1 GB RAM.

- 1 Microprocesador Core2Duo.

Interfaces de software

RNF 19. Instalar en las estaciones de trabajo el software necesario para el correcto funcionamiento del sistema.

Las configuraciones de software de las máquinas clientes deben contar al menos con:

- Firefox 2.0 o superior.
- Java Virtual Machine 6.0 y Schema Workbench 3.2.1 en caso de que un usuario capacitado requiera la construcción de esquemas multidimensionales para el diseño de nuevos reportes.

Interfaces de comunicación

No se posee ningún requisito de interfaz de comunicación.

Requerimientos de licencia

No se posee ningún requisito de licencia o restricción en el uso del software, ya que se hará uso de herramientas libres.

Requerimientos legales de derechos de autor y otros

RNF 20. Entregar el sistema a la ONE.

El sistema debe ser transferido a la ONE mediante un proceso de transferencia una vez que esté en explotación, incluyendo el código fuente y la documentación correspondiente.

RNF 21. Requerimientos legales, de derecho de autor y otros.

- No se hace solicitud de derecho de autor, patentes, marca comercial o complacencia con logotipo para el software, debido a que se usan soluciones con Licencia Pública General (GNU GPL por sus siglas en inglés), bajo el principio de software libre.

2.5 Modelo de casos de uso del sistema

El diagrama de casos de uso del sistema refleja el comportamiento del sistema, mediante la interacción con los usuarios, o la interacción de los actores con los casos de uso del sistema y se utilizan para mostrar los requerimientos del sistema.

2.5.1 Actores del sistema

Los actores del sistema son todas aquellas entidades externas al sistema y que inicializan una funcionalidad en el, incluyendo al hombre y otras entidades abstractas como el tiempo.

Tabla 3. Actores del sistema

Actor	Descripción
Analista	Obtiene y consulta los reportes
Administrador	El administrador gestiona todas las operaciones de administración de roles y usuarios.
Administrador ETL	El administrador de ETL se encarga de realizar los procesos de extracción, transformación y carga.

2.5.2 Diagrama de casos de uso del sistema

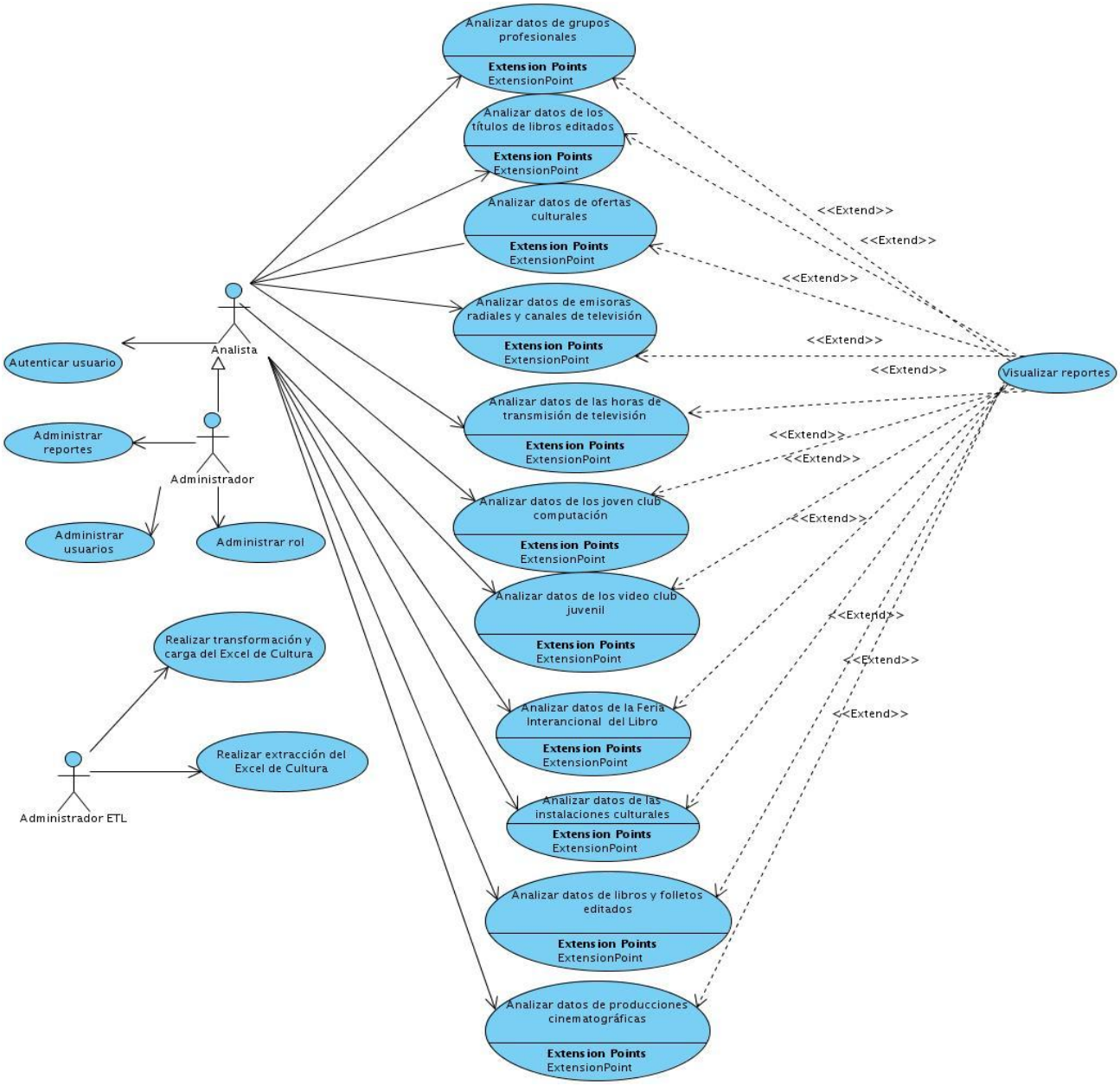

Figura 1. Diagrama de Casos de Uso del Sistema de Cultura

Figura 2. Diagrama de Casos de Uso del Sistema de Deporte

2.5.3 Especificaciones de casos de uso del sistema

Los casos de uso son la interacción del usuario con el sistema para obtener un objetivo específico, estos se utilizan para representar los requisitos del sistema a desarrollar.

A continuación se describen algunos de los casos de uso de información y funcionales de ambos temas de análisis. Las restantes descripciones se encuentran en el expediente de proyecto del mercado de datos de Cultura y Deporte en el artefacto “Modelo de Casos de Uso”.

Descripción de Casos de Uso críticos del sistema

Tabla 4. Descripción del caso de uso Extraer datos de los sistemas fuentes

Caso de Uso:	**Extraer datos de los sistemas fuentes**

Tipo:	**Funcional.**
Actores:	**Administrador ETL**
Resumen:	**El CU inicia cuando el actor vaya a insertar nuevos datos al mercado de datos. Este caso de uso finaliza cuando estos datos se encuentren en el área de intercambio o área temporal.**
Precondiciones:	**Disponibilidad de las fuentes.**
Referencias	**RF1**
Prioridad	**Crítico**
Complejidad	**Crítico**

Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Excel Cultura.**
1. El usuario interactúa con la herramienta Pentaho Data Integration (PDI) para realizar la extracción de los datos.	**1.1 El sistema muestra el área de trabajo y le da al usuario la posibilidad de cargar la transformación.**
2. Selecciona la transformación a cargar o realiza los pasos para una nueva transformación.	
3. Configura los parámetros de entrada de la transformación.	**3.1 El sistema almacena los datos en el área temporal.**
4. El usuario previsualiza los datos.	**4.1 El sistema muestra los datos existentes en esta área.**
5. El usuario realiza la acción de aceptar y finaliza.	

Flujos Alternos	
Acción del Actor	**Respuesta del Sistema**
	4.1 El sistema muestra un mensaje de error
2. El usuario vuelve al paso 3	

Poscondiciones	**Datos disponibles para transformar.**

Tabla 5. Descripción del caso de uso Realizar transformación y carga de los datos de los sistemas fuentes

Caso de Uso:	**Realizar transformación y carga de los datos de los sistemas fuentes**
Tipo:	**Funcional.**
Actores:	**Administrador ETL**
Resumen:	**El CU inicia cuando el actor selecciona las estructuras a transformar, carga**

	los datos seleccionados en el área temporal hacia el mercado de datos y finaliza con el éxito de la carga de los datos en la base de datos.
Precondiciones:	**Extracción completada en área temporal. Estructuras del mercado de datos disponibles.**
Referencias	**RF2**
Prioridad	**Crítico**
Complejidad	**Crítico**

Flujo Normal de Eventos	
Acción del Actor.	**Respuesta del mercado de datos.**
1. Selecciona estructuras del área temporal a transformar.	
2. Carga datos seleccionados en memoria.	
3. Aplica transformaciones pertinentes.	**3.1 Carga datos en el mercado de datos.**
4. El usuario realiza la acción de aceptar y finaliza.	**4.1 El sistema muestra un mensaje de afirmación.**

Flujos Alternos	
Acción del Actor	**Respuesta del Sistema**
	4.1 El sistema muestra un mensaje de error.

2.6 Modelo de datos dimensional

Dentro de la modelación de datos se encuentra el modelado dimensional, el cual está compuesto por hechos y dimensiones.

Los hechos son los datos que brindan información cuantitativa sobre las características del negocio que se quieren analizar, en este caso sobre los temas de análisis identificados para cada área. Por otra parte, las dimensiones serían los grupos de elementos que permitirían analizar esos valores contenidos en los hechos.

Este tipo de modelado posee la técnica de diseño para estructurar los datos según el tema de clasificación, se basa en la simplicidad, es decir, es más intuitivo para el usuario y está optimizado para lograr un mejor rendimiento de las consultas.

A continuación se listan las dimensiones y hechos identificados para ambas áreas de análisis.

Lista de dimensiones

1. dim_dpa: dimensión división política administrativa
2. dim_provincia: dimensión provincia

3. dim_sexo: dimensión sexo
4. dim_temporal_anno: dimensión temporal año
5. dim_um: dimensión unidad de medida
6. dim_deporte: dimensión deporte
7. dim_eventos_culturales: dimensión eventos culturales
8. dim_eventos_deportivos: dimensión eventos deportivos
9. dim_indicadores: dimensión indicadores
10. dim_instalaciones: dimensión instalaciones culturales
11. dim_manifestaciones_artisticas: dimensión manifestaciones artísticas
12. dim_materias_libros: dimensión materia libros
13. dim_nivel: dimensión nivel
14. dim_personal: dimensión personal
15. dim_practicantes: dimensión practicantes
16. dim_publicacion: dimensión publicación
17. dim_transmision: dimensión transmisión

Lista de hechos

1. hech_canales_television: hecho canales de televisión
2. hech_emisoras_radiales_canales_television: hecho emisoras radiales y canales de televisión
3. hech_feria_inter_libro: hecho Feria Internacional del Libro
4. hech_festivales_inder: hecho festivales promovidos por el INDER
5. hech_grupos_profesionales: hecho grupos profesionales
6. hech_instalaciones_culturales: hecho instalaciones culturales
7. hech_instalaciones_deportivas: hecho instalaciones deportivas
8. hech_jovenclub_computacion: hecho joven club de computación
9. hech_juegos_paralimpicos: hecho Juegos Paralímpicos Nacionales
10. hech_libros_folletos: hecho libros y folletos
11. hech_mundiales_en_cuba: hecho mundiales celebrados en Cuba
12. hech_ofertas_culturales: hecho ofertas culturales
13. hech_olimpiadas_deporte_cubano: hecho olimpiadas del deporte cubano
14. hech_participantes: hecho participantes en eventos competitivos
15. hech_personal_pedagogico_deportivo: hecho personal deportivo pedagógico

16. hech_practicantes_sistematicos_deporte: hecho practicantes sistemáticos del deporte
17. hech_producciones_cinematograficas: hecho producciones cinematográficas
18. hech_titulos_ganados_cuba: hecho títulos ganados por Cuba
19. hech_titulos_libros_editados: hecho títulos de libros editados
20. hech_videoclub_juvenil: hecho videos club juvenil

2.6.1 Matriz Bus

La Matriz Bus refleja gráficamente la relación que existe entre las dimensiones y hechos del mercado de datos, a continuación se presenta la tabla de dicha relación, clasificada por los temas de análisis definidos.

Matriz Bus																
Hechos	Dimensiones															
	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10	D11	D12	D13	D14	D15	D16	D17
H1	x			x	x				x								x
H2		x		x					x				x				
H3	x			x	x		x		x								
H4	x			x	x				x								
H5	x			x	x				x		x						
H6		x		x	x				x	x							
H7		x		x	x				x								
H8		x		x	x				x								
H9	x				x	x		x	x								
H10	x			x	x				x							x	
H11	x			x	x	x			x								
H12	x			x					x	x							
H13	x				x	x		x	x								
H14			x	x	x	x			x				x				
H15	x			x	x				x					x			
H16	x			x	x				x						x		
H17	x			x	x				x								
H18					x	x		x	x								
H19	x			x	x				x			x					
H20	x			x	x				x								

Figura 3. Matriz Bus

2.6.2 Tabla de dimensiones

A continuación se describen algunas de las dimensiones asociadas a las áreas de Cultura y Deporte identificadas en cada tema de análisis. Las tablas descriptivas de las dimensiones se encuentran en el

expediente de proyecto del mercado de datos de Cultura y Deporte en el artefacto especificado “Especificaciones del Modelo de Datos Dimensional”.

dim_indicadores

Todas las tablas de dimensiones tendrán un identificador que estará compuesto por el nombre de la tabla, guión bajo y seguido la palabra id.

Esta tabla contiene además del identificador tres atributos, el atributo temática recoge el área de análisis al que pertenece el indicador a medir, la columna indicador recoge todos los indicadores de ambas áreas de análisis y el atributo descripcion recoge la descripción de los indicadores.

dim_eventos_deportivos

Esta tabla contiene además del identificador cinco atributos, el atributo evento_deportivo abarca distintos eventos deportivos nacionales e internacionales como los Juegos Panamericanos, Juegos Olímpicos y los Juegos Paralímpicos Nacionales entre otros, la columna nombre_evento recoge el nombre de los juegos y la edición, el atributo anno recoge el año en que se desarrolló el evento, la columna lugar_evento abarca el lugar en que se desarrolla el evento y el atributo descripcion recoge la descripción de los juegos.

dim_eventos_culturales

Esta tabla contiene además del identificador cuatro atributos, el atributo evento_cultural abarca distintos eventos culturales como la Feria Internacional del Libro, la columna codigo recoge el código de los indicadores del evento, el atributo dim_eventos_culturales_padre recoge el identificador de los indicadores que son temáticas de otros eventos y el atributo descripcion recoge la descripción de los eventos culturales.

dim_deporte

Esta dimensión está compuesta por el identificador de la tabla, el atributo deporte recoge todos los deportes participativos, el atributo codigo recoge el código de cada deporte, el atributo dim_deporte_padre recoge el identificador de los deportes que engloban otros deportes y el atributo descripcion recoge la descripción de los deportes.

dim_publicacion

Esta dimensión contiene aparte del identificador, el atributo publicacion que abarca las publicaciones impresas como libros y folletos, el atributo codigo_publicacion recoge el código de las publicaciones, el atributo dim_publicacion_padre recoge el identificador de las publicaciones que agrupan otras publicaciones y el atributo descripcion recoge la descripción de las publicaciones.

dim_practicantes

Esta dimensión además del identificador está conformada por el atributo tipo_practicante que recoge las especialidades en las que las personas practican de manera sistemática, el atributo codigo recoge el código de los practicantes por especialidades, el atributo dim_practicantes_padre recoge el identificador de las especialidades que engloban otras actividades y el atributo descripcion recoge la descripción de las especialidades.

dim_temporal_anno

Esta dimensión recoge el identificador de la dimensión, el atributo anno_codigo que recoge el código del año, el atributo anno_nombre que recoge el nombre del año y el atributo anno_numero que agrupa el número del año.

dim_provincia

Esta dimensión recoge los atributos provincia_descripcion, provincia_codigo y provincia_nombre además del identificador. El atributo provincia_descripcion recoge las descripciones de las provincias, provincia_codigo agrupa los códigos de las provincias y el atributo provincia_nombre recoge el nombre de las provincias.

dim_um

Esta dimensión recoge los atributos codigo, descripcion y nombre además del identificador. El atributo codigo recoge los códigos de las unidades de medida, descripcion agrupa las descripciones de las unidades de medida y el atributo nombre recoge el nombre de las unidades de medida.

dim_dpa

Esta dimensión agrupa los atributos provincia_codigo que recoge el código de las provincias, provincia_nombre que recoge el nombre de las provincias, municipio_codigo que recoge el código de los municipios, municipio_nombre que recoge el nombre de los municipios, municipio_descripcion que recoge la descripción de los municipios, provincia_descripcion que recoge la descripción de las provincias, municipio_extension que recoge el tamaño de los municipios, municipio_ext_cayosady, municipio_ext_tierrafirme, dpa_anno_nombre que recoge el año de la división política administrativa.

2.6.3 Tablas de hechos

A continuación se describen algunas de las tablas de hechos identificadas en las áreas de Cultura y Deporte correspondientes a cada tema de análisis. Las tablas descriptivas se encuentran en el expediente de proyecto del mercado de datos de Cultura y Deporte en el artefacto especificado “Especificaciones del Modelo de Datos Dimensional”.

hech_producciones_cinematograficas

Esta tabla de hecho se corresponde con los indicadores referentes a las producciones cinematográficas, su identificador lo componen las llaves de las dimensiones con las que se asocia,

además se compone de 3 medidas, cantidad de largometrajes, cantidad de cortometrajes y cantidad de dibujos animados.

hech_titulos_ganados_cuba

Esta tabla de hecho se corresponde con los indicadores referentes a los títulos ganadas por Cuba en los distintos eventos deportivos, su identificador lo componen las llaves de las dimensiones con las que se asocia, además se compone de 3 medidas, cantidad de medallas de oro, cantidad de medallas de plata y cantidad de medallas de bronce.

hech_practicantes_sistematicos_deporte

Esta tabla de hecho se corresponde con los indicadores referentes a los practicantes sistemáticos del deporte, su identificador lo componen las llaves de las dimensiones con las que se asocia, además se compone de 1 medida, cantidad de practicantes.

hech_titulos_libros_editados

Esta tabla de hecho se corresponde con los indicadores referentes a los títulos de libros editados por materias, su identificador lo componen las llaves de las dimensiones con las que se asocia, además se compone de 1 medida, cantidad de títulos de libros editados.

hech_canales_television

Esta tabla de hecho se corresponde con los indicadores referentes a las horas de transmisión según el canal televisivo y los programas transmitidos, su identificador lo componen las llaves de las dimensiones con las que se asocia, además se compone de 1 medida, cantidad de horas de emisión.

hech_jovenclub_computacion

Esta tabla de hecho se corresponde con los indicadores referentes al total de instituciones por provincias, su identificador lo componen las llaves de las dimensiones con las que se asocia, además se compone de 4 medidas, cantidad de palacios de computación, cantidad de joven club de computación, cantidad de egresados de los cursos regulares y cantidad de municipios con varios joven club de computación.

hech_instalaciones_deportivas

Esta tabla de hecho se corresponde con los indicadores referentes a las instalaciones deportivas, su identificador lo componen las llaves de las dimensiones con las que se asocia, además se compone de 4 medidas, cantidad de terrenos al aire libre, cantidad de piscinas, cantidad de salas deportivas y cantidad de complejos deportivos.

hech_ofertas_culturales

Esta tabla de hecho se corresponde con los indicadores referentes a las ofertas culturales en el país, su identificador lo componen las llaves de las dimensiones con las que se asocia, además se compone de 2 medidas, cantidad de ofertas culturales y cantidad de asistentes a las ofertas.

hech_instalaciones_culturales

Esta tabla de hecho se corresponde con los indicadores referentes a las instalaciones culturales en el país, su identificador lo componen las llaves de las dimensiones con las que se asocia, además se compone de 1 medida, cantidad de instalaciones culturales.

2.8 Política de respaldo y recuperación

Las políticas de seguridad y respaldo que se utilizarán en el mercado de datos de Cultura y Deporte están divididas en tres puntos fundamentalmente.

- Periodicidad de las salvas: las salvas de toda la información contenida en la BD se realizan anualmente, así lo define la organización, certificando en todo momento que exista una copia escrita de la información presente en el servidor.
- Tablas involucradas: las tablas que se involucran en la realización son las 20 tablas de hechos identificadas en el proceso de análisis con sus 17 dimensiones asociadas, además de las 4 tablas closure que apoyan el proceso de ETL.
- Backups existentes: en cada año se realizan dos salvas, una en el mes de diciembre con los datos de cierre del año, y otra en el mes de julio, además se tiene un servidor de respaldo con todos los datos para en caso de que ocurra un incidente que atente contra la seguridad de la entidad. También se tiene copia de los datos en otros medios de almacenamiento como son los DVD.

2.9 Conclusiones del capítulo

En este capítulo se definieron los hechos y dimensiones que conforman la Matriz Bus, así como el diseño físico del modelo de datos de las áreas de Cultura y Deporte, como también se definieron un grupo de elementos expuestos en los epígrafes anteriores que sentaron las bases para el comienzo de la implementación del Mercado de datos Cultura y Deporte.

CAPÍTULO 3: IMPLEMENTACIÓN DEL MERCADO DE DATOS CULTURA Y DEPORTE

3.1 Introducción

En este capítulo se realiza la implementación de los subsistemas de integración y visualización de los datos para el mercado de datos Cultura y Deporte. Se realiza el proceso de ETL y almacenamiento de los datos en las tablas correspondientes de acuerdo al modelo de datos definido en el capítulo anterior, posteriormente se define la capa de Inteligencia de Negocios y con ella la implementación de los reportes candidatos que responden a las necesidades del cliente a través de consultas MDX, permitiendo visualizar los datos a través de textos, tablas y gráficos que permiten al usuario un mejor entendimiento y comprensión de los mismos.

3.2 Implementación del modelo de datos físico

El esquema de una base de datos define todas sus tablas, cada campo en cada tabla y las relaciones entre ellos y cada tabla.

Para el desarrollo del sistema propuesto en la presente investigación se definieron dos esquemas, el esquema dimensiones que recoge todas las tablas de dimensiones que son comunes para el almacén de datos central y el esquema mart_cultura_deporte que recoge las tablas de dimensiones y hechos propias del mercado de datos. La solución cuenta con 41 tablas, 17 tablas de dimensiones, 20 tablas de hechos y 4 tablas closure que apoyan el proceso de ETL.

Tabla 6. Esquemas y tablas de la base de datos

Esquemas	Tablas
dimensiones	dim_sexo
dimensiones	dim_provincia
dimensiones	dim_temporal_anno
dimensiones	dim_dpa
dimensiones	dim_um
mart_cultura_deporte	dim_nivel
mart_cultura_deporte	dim_deporte
mart_cultura_deporte	dim_eventos_culturales
mart_cultura_deporte	dim_eventos_deportivos
mart_cultura_deporte	dim_indicadores
mart_cultura_deporte	dim_instalaciones
mart_cultura_deporte	dim_manifestaciones_artisticas
mart_cultura_deporte	dim_materias_libros

mart_cultura_deporte	dim_personal
mart_cultura_deporte	dim_practicantes
mart_cultura_deporte	dim_publicacion
mart_cultura_deporte	dim_transmision
mart_cultura_deporte	hech_canales_television
mart_cultura_deporte	hech_emisoras_radiales_canales_television
mart_cultura_deporte	hech_feria_inter_libro
mart_cultura_deporte	hech_festivales_inder
mart_cultura_deporte	hech_grupos_profesionales
mart_cultura_deporte	hech_instalaciones_culturales
mart_cultura_deporte	hech_instalaciones_deportivas
mart_cultura_deporte	hech_joven_club_computacion
mart_cultura_deporte	hech_juegos_paralimpicos
mart_cultura_deporte	hech_libros_folletos
mart_cultura_deporte	hech_mundiales_en_cuba
mart_cultura_deporte	hech_ofertas_culturales
mart_cultura_deporte	hech_olimpiadas_deporte_cubano
mart_cultura_deporte	hech_participantes
mart_cultura_deporte	hech_personal_pedagogico_deportivo
mart_cultura_deporte	hech_practicantes_sistematicos
mart_cultura_deporte	hech_producciones_cinematograficas
mart_cultura_deporte	hech_titulos_ganados_cuba
mart_cultura_deporte	hech_titulos_libros_editados
mart_cultura_deporte	hech_videoclub_juvenil
mart_cultura_deporte	closure_deporte
mart_cultura_deporte	closure_eventos_culturales
mart_cultura_deporte	closure_practicantes_sistematicos_deporte
mart_cultura_deporte	closure_publicacion

3.3 Implementación de los subsistemas de integración

El proceso de Extracción, Transformación y Carga consiste en extraer los datos de las fuentes y se seleccionan los campos necesarios conforme al modelo de datos, luego estos datos se transforman,

limpian y estandarizan para eliminar inconsistencias y posibles errores que pudieran llegar a existir. Luego se realiza la carga de las dimensiones y hechos que componen el mercado de datos a través de un grupo de componentes que se encuentran en la herramienta definida en el capítulo uno, teniendo como salida la tabla correspondiente en la base de datos.

A continuación se ilustran algunos ejemplos de las transformaciones realizadas para poblar la base de datos correspondiente al mercado de datos Cultura y Deporte.

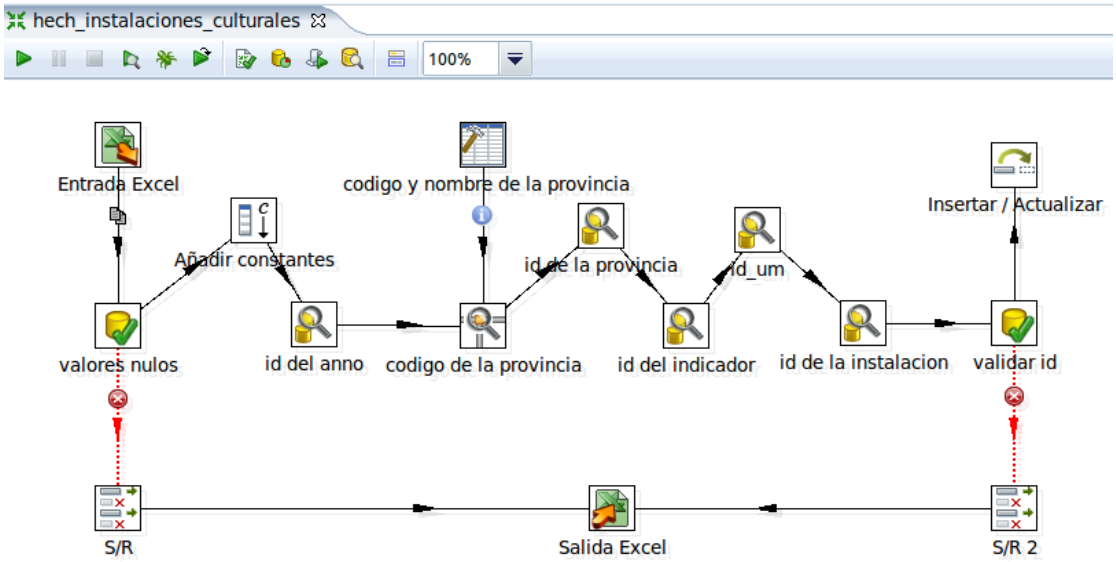

Figura 4. Carga del hecho instalaciones culturales

Se carga el hecho correspondiente a los indicadores pertenecientes a las instalaciones culturales: en la transformación se realiza la extracción de los datos fuentes, en este caso son entrada Excel, se validan que los campos no contengan valores nulos y que los id buscados sean válidos. Se almacenan las provincias viejas con sus respectivos códigos para luego realizar una búsqueda en flujo y obtener el código de la provincia, se utiliza el componente añadir constantes para añadir las constantes que no vienen directamente de la fuente, se busca en la base de datos el id de las dimensiones relacionados con el hecho para finalmente cargar el hecho.

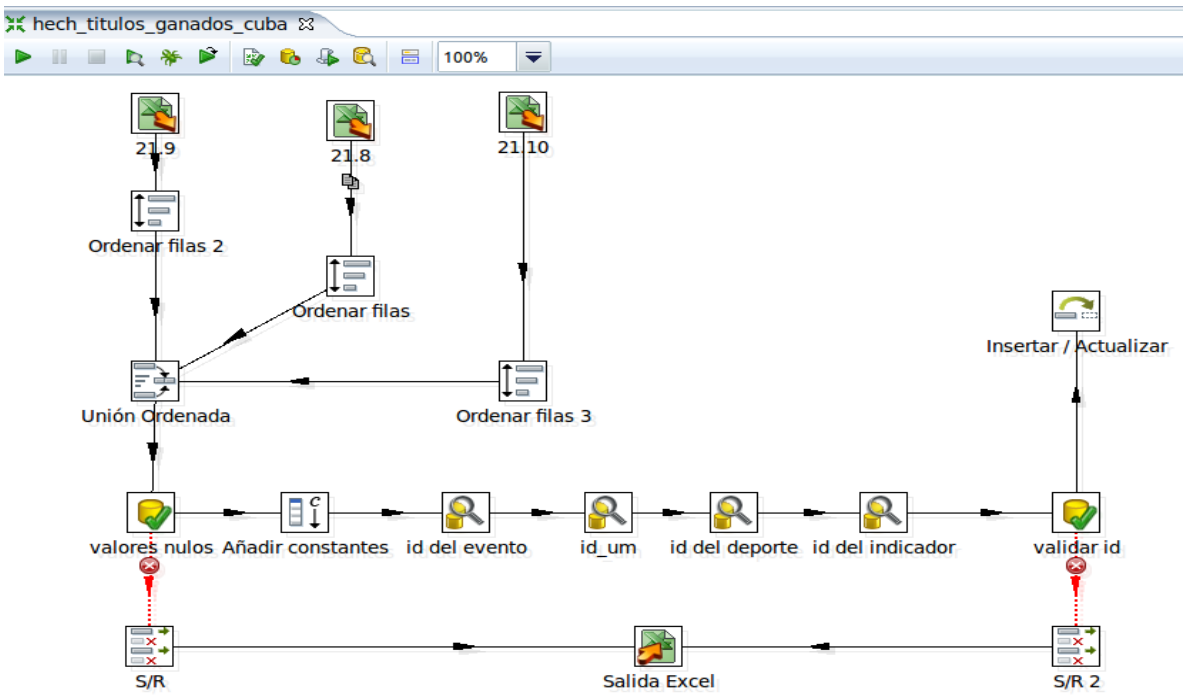

Figura 5. Carga del hecho títulos ganados por Cuba en eventos competitivos

Se carga el hecho correspondiente a los indicadores perteneciente a los títulos ganados por Cuba en los diferentes eventos deportivos: en la transformación se realiza la extracción de los datos fuentes, en este caso son entrada Excel, como son varias entradas Excel con los mismos campos se realiza una unión y obteniéndose como salida una sola tabla que contiene todos los campos con todos los valores, se validan que los campos no contengan valores nulos y que los id buscados sean válidos. Se utiliza el componente añadir constantes para añadir las constantes que no vienen directamente de la fuente y luego se busca en la base de datos el id de las dimensiones relacionados con el hecho para finalmente cargar el hecho.

3.4 Implementación de los trabajos

Un trabajo o job es similar al concepto de proceso. Un proceso es un conjunto sencillo o complejo de tareas con el objetivo de realizar una acción determinada. En los trabajos podemos utilizar pasos específicos que son diferentes a los disponibles en las transformaciones [29]. Además, podemos ejecutar una o varias transformaciones de las que hayamos diseñado y realizar una secuencia de ejecución de ellas, una transformación no se empieza a ejecutar si la anterior no ha terminado, en este caso se cargaron primero las dimensiones para que no exista referencia de llaves nulas en las tablas de hechos.

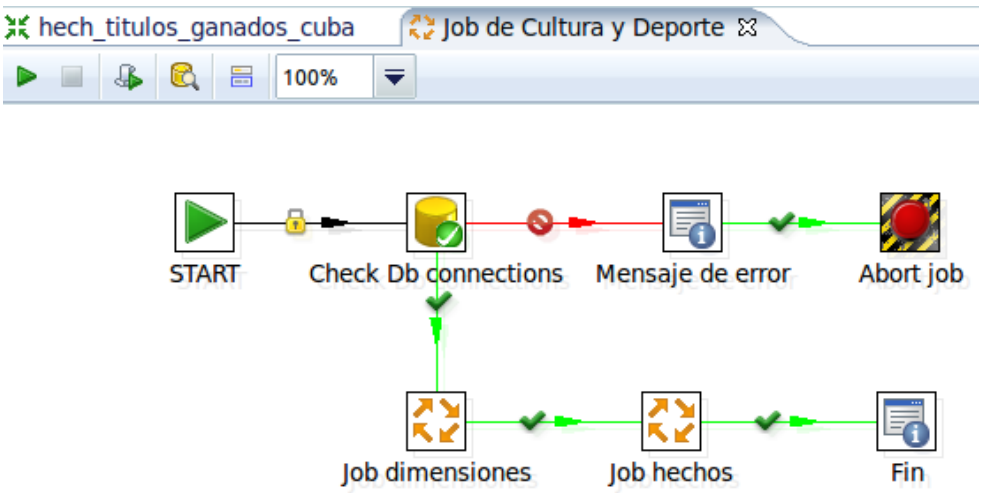

Figura 6. Implementación del Job

3.5 Implementación de los subsistemas de visualización de datos

La Inteligencia de Negocio consiste en transformar los datos en información, y la información en conocimiento, de forma que se pueda optimizar el proceso de toma de decisiones en los negocios. Para ello se realizarán un conjunto de tareas encaminadas a satisfacer las necesidades del cliente.

3.5.1 Implementación de los cubos OLAP

Uno de los factores claves en el procesamiento analítico en línea son los cubos OLAP, estos proveen rápido acceso a los datos almacenados independientemente de la cantidad de datos en el cubo, al mismo tiempo son un subconjunto de datos del almacén de datos. Luego de haber realizado los procesos de ETL se realizó la creación de los cubos correspondientes al mercado de datos. En la presente investigación se diseñaron 20 cubos multidimensionales, uno para cada tabla de hecho, especificando en los cubos las dimensiones y características que se corresponden con estas tablas, además el esquema cuenta también con 17 dimensiones y 63 medidas, de ellas 6 calculables, todas ellas enfocadas a las necesidades del cliente para darle solución al problema planteado.

Figura 7. Diseño de los cubos OLAP

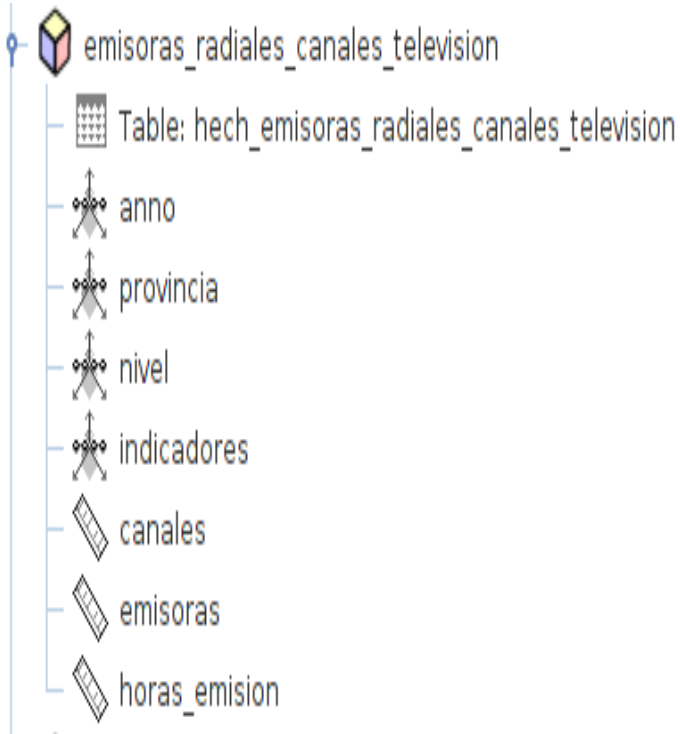

Figura 8. Elementos contenidos dentro del cubo emisoras radiales y canales de televisión

El cubo contiene las dimensiones relacionadas con el hecho, año, provincia, indicador y nivel, que constituyen las perspectivas de análisis de las tres medidas físicas que se analizan en el hecho correspondiente al cubo.

3.6 Arquitectura de información

A continuación se muestran en detalles los elementos que conforman el mapa de navegación correspondiente al mercado de datos Cultura y Deporte para la visualización de los datos. El mismo estará compuesto por un área de análisis general, un área de análisis que recogerá las áreas de análisis de Cultura y Deporte, el área de análisis de Cultura conformada por 11 libros de trabajos y 14 reportes asociados a cada libro de trabajo; y el área de análisis de Deporte estructurado en 9 libros de trabajos y 9 reportes asociados a estos.

A continuación se muestran figuras que ilustran el mapa de navegación.

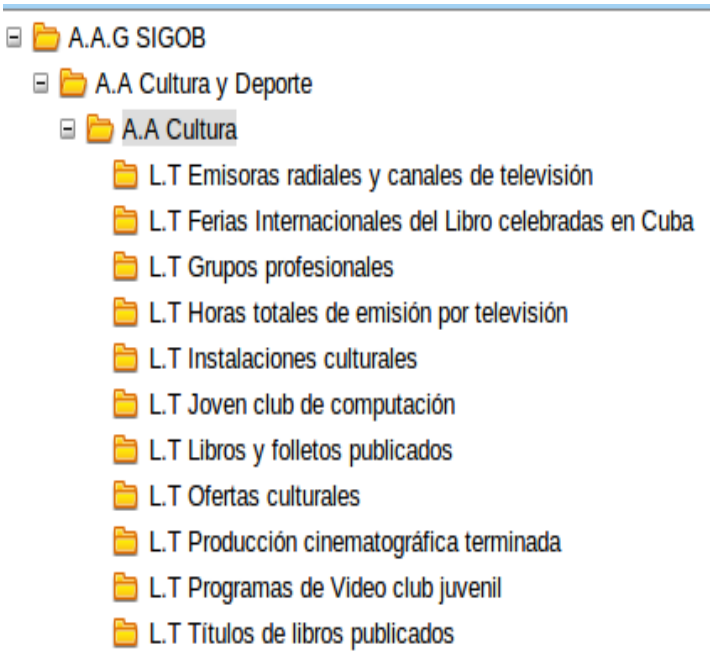

Figura 9. Área de análisis de Cultura

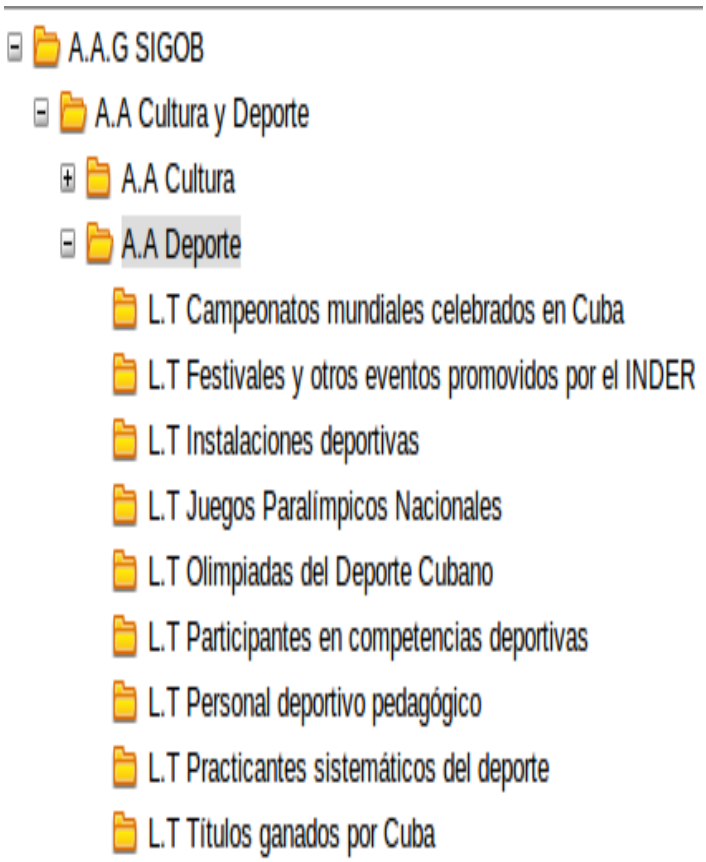

Figura 10. Área de análisis de Deporte

3.7 Implementación de los reportes candidatos

Los reportes candidatos o tablas de salidas como también se les conoce van a contener los valores asociados a los indicadores que son de interés para el cliente, estos son implementados a través de consultas MDX que son en los sistemas OLAP el equivalente a las consultas SQL en las bases de datos relacionales. A continuación se muestran algunas vistas de análisis implementadas a través de consultas MDX.

Producciones cinem..

	Indicador				
	Producción cinematográfica terminada				
	Unidad de medida				
	Unidad				
	Medidas				
Año	Total	Cantidad de largometrajes	Cantidad de cortometrajes	Cantidad de dibujos animados	Cantidad de noticieros
Total	4.074	303	1.384	877	1.510
1960	60		28	2	30
1961	90	3	32	4	51
1962	112	4	54	1	53
1963	107	4	46	5	52
1964	88	7	27	2	52
1965	102	3	36	12	51
1966	87	3	27	5	52
1967	95	4	28	13	50
1968	97	4	44	5	44
1969	66	1	19	3	43
1970	73	1	29	5	38

Figura 11. Vista de análisis del indicador general producción cinematográfica terminada

En la vista de análisis que se presentó anteriormente se miden cuatro medidas físicas pertenecientes al hecho producciones cinematográficas y que se definieron en el diseño del cubo OLAP que se corresponde con la tabla de hecho, estas medidas son la cantidad de largometrajes, cantidad de cortometrajes, cantidad de dibujos animados y la cantidad de noticieros; y la otra variable que se analiza es el total que es una variable calculable que se definió en el diseño del cubo, todas ellas pueden ser analizadas desde diferentes perspectivas de análisis, en este caso se pueden filtrar por años que se encuentra en la fila y por columna se tiene la unidad de medida y el indicador general a medir.

Practicantes siste..

	Año			
	1985	1986	1987	1988
	Unidad de medida	Unidad de medida	Unidad de medida	Unidad de medida
Practicantes sistemáticos del deporte	Unidad	Unidad	Unidad	Unidad
Practicantes sistemáticos	943.244	1.029.988	1.077.265	1.166.534
En la masividad	808.082	901.742	943.563	955.449
Total de practicantes sistemáticos en deporte participativo	396.480	449.302	487.942	555.537
Educación física para adultos	233.557	265.016	286.621	341.035
En círculos de abuelos		1.986	49.809	86.332
Otros	233.557	263.030	236.812	254.703
Preparación física				
Gimnasio de cultura física				
Grupos de salud				
Hipertensos				
Otros grupos de salud				
Gimnasia básica para la mujer	158.283	179.048	195.433	207.356

Figura 12. Vista de análisis del indicador practicantes sistemáticos del deporte

En la vista de análisis que se presentó anteriormente se miden los indicadores que se encuentran dentro de la dimensión practicantes sistemáticos del deporte por año y por unidad de medida.

3.8 Conclusión

En este capítulo se abordaron los elementos que permitieron el desarrollo de la implementación del mercado de datos. Se realizó la implementación del modelo de datos físico definiéndose dos esquemas: dimensiones y mart_cultura_deporte para lograr una mejor estructuración de la información. Se realizó el proceso de ETL, extrayendo los datos de las series históricas de Cultura y Deporte para luego realizar la carga de los datos a la base de datos. Luego de tener los datos cargados se realizó el proceso de inteligencia de negocio, donde se diseñaron los cubos OLAP, se implementaron las vistas de análisis correspondientes y se realizó la visualización de los datos convirtiéndolos en información valiosa para los usuarios. .

CAPÍTULO 4: VALIDACIÓN DEL MERCADO DE DATOS CULTURA Y DEPORTE

4.1 Introducción

En el presente capítulo se aplican diferentes tipos de pruebas manuales, tanto a la aplicación como a la documentación. Se validan las entradas de datos, verificando que los resultados obtenidos se correspondan con los resultados esperados, mediante casos de pruebas y las listas de chequeo a los que fue sometido el mercado de datos.

4.2 Validación y prueba

Para los ingenieros del software desarrollar un producto o servicio de buena calidad y aceptación por el cliente, constituye un requisito indispensable para que el propio ingeniero pueda ganar determinada reputación dentro de la entidad donde se desempeña como profesional y también la empresa u organización a la que pertenecen.

A nivel mundial se trata de lograr que las empresas que producen software, lo hagan de acuerdo a criterios comunes, para lograr uniformidad, muchos son los esfuerzos que se realizan para lograr que los productos de software tengan una alta calidad al salir al mercado y por lo tanto que la satisfacción del cliente sea elevada.

Muchos son los autores que han sido capaces de hablar de diferentes conceptos relacionados con la calidad de los productos de software. Uno de los conceptos más íntegros y fácil de comprender por el lector es el que a continuación se muestra.

“La calidad del software es el grado con el que un sistema, componente o proceso cumple los requerimientos específicos y las necesidades o expectativas del cliente o usuario [30].”

Durante el desarrollo de un producto de software la forma más eficaz de evitar que los errores se produzcan y, sobre todo que se propaguen es disponer de procedimientos de calidad y pruebas que acompañen al producto a lo largo de su ciclo de vida. Resulta importante llevar a cabo pruebas de productos y sistemas de manera independiente y buscar asesoría y colaboración en los procesos de certificación.

Para lograr obtener un producto con calidad existen diferentes tipos de pruebas que pueden ser aplicadas:

- **Prueba unitaria:** se enfocan en un programa o un componente que desempeña una función específica que puede ser probada y que se asegura que funcione tal y como lo define la especificación del programa. Los programadores siempre prueban el código durante el desarrollo, por lo que las pruebas unitarias son realizadas solamente después de que el programador considera que el componente está libre de errores [31].

- **Prueba de integración:** su objetivo es identificar errores introducidos por la combinación de programas o componentes probados unitariamente, además, verificar que las especificaciones de diseño sean alcanzadas. Componentes individuales son combinados con otros componentes para asegurar que la comunicación, enlaces y los datos compartidos ocurran apropiadamente. No son verdaderamente pruebas de sistema porque los componentes no están implementados en el ambiente operativo [31].
- **Prueba de sistema:** son usualmente conducidas para asegurar que todos los módulos trabajan como sistema, sin error. Es similar a la prueba de integración pero con un alcance mucho más amplio. Las pruebas del sistema examinan qué tan bien el sistema cumple con los requerimientos de la organización y su utilidad, seguridad y desempeño. También se realizan estas pruebas a la documentación del sistema [31].
- **Prueba de aceptación:** son realizadas principalmente por los usuarios con el apoyo del equipo del proyecto. El propósito es confirmar que el sistema está terminado, que desarrolla puntualmente las necesidades de la organización y que es aceptado por los usuarios finales [31].

A parte de las pruebas existentes y que se pueden aplicar, el centro realiza un grupo de pruebas encaminadas a la culminación del producto, estas pruebas realizadas son:

- **Pruebas internas:** son realizadas por el equipo de desarrollo y especialistas de los departamentos que participan en la solución.
- **Pruebas de liberación:** son realizadas por los especialistas de calidad, donde en un primer momento es desarrollada por los especialistas de calidad de DATEC y luego por los especialistas del Centro de Calidad para Aplicaciones Tecnológicas (CALISOFT).

Dentro de las herramientas utilizadas para que se apliquen los distintos tipos de pruebas se tienen los casos de prueba y las listas de chequeo.

4.2.1 Diseño de los casos de prueba

Para lograr la calidad del producto de software es necesario realizar un conjunto de evaluaciones durante todo el proceso de desarrollo que implique al cliente y desarrollador. Para ello se diseñaron casos de pruebas basados en los casos de uso y a su vez en los requerimientos funcionales, comparando cada funcionalidad implementada con la descrita, para verificar hasta qué punto cumplía con las necesidades del cliente [32]. Para el mercado de datos Cultura y Deporte se diseñaron 20 casos de prueba, a continuación se muestra un escenario del diseño de caso de prueba correspondiente al caso de uso “Analizar datos de emisoras radiales y canales de televisión” (Ver expediente de proyecto correspondiente al mercado de datos Cultura y Deporte).

Tabla 7. Diseño del caso de prueba "Analizar datos de emisoras radiales y canales de televisión"

Escenario	Descripción	Variables		Respuesta del sistema	Flujo central
		Perfiles de análisis	Indicadores a medir		
EC 1.1: Cantidad de canales	Permite visualizar el reporte con las variables presentes en el mismo.	Año Provincia Nivel	Cantidad de canales (U)	Se muestra la tabla con los valores correspondientes a cada escenario.	Se abre la aplicación. Se autentifica. Se entra al sistema. Se despliega hacia la derecha el componente ubicado en el lateral izquierdo que contiene el navegador. Se seleccionan las áreas de análisis **A.A Cultura y Deporte** y luego **A.A Cultura.** Se selecciona el libro de trabajo **L.T Emisoras radiales y canales de televisión.** En la parte inferior izquierda se selecciona el reporte deseado. En el área de trabajo se visualiza la tabla correspondiente al reporte.

4.2.2 Aplicación de listas de chequeo

La lista de chequeo es un documento que tiene un conjunto de parámetros a medir sobre un aspecto determinado, dígase documentación o aplicación. Es un instrumento de medición y evaluación que consiste básicamente en un formulario de preguntas referentes al atributo de calidad que se está probando y de las características del documento en el caso de la documentación. Cada pregunta tiene asociada una evaluación en una escala que da una medida del grado de cumplimiento y disponibilidad de la propiedad evaluada, de esta manera se determina la evaluación del elemento probado [32]. En el caso del mercado de datos Cultura y Deporte la lista de chequeo que se aplicó se divide en tres secciones de la siguiente forma (Ver expediente de proyecto correspondiente al mercado de datos Cultura y Deporte).

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

Tabla 8. Aplicación de la lista de chequeo al mercado de datos Cultura y Deporte

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1. ¿Están los documentos acorde a las planillas estándar definidas por el proyecto?	0		0	
crítico	2. ¿Contiene las secciones obligatorias definidas en el expediente? (Ver Expediente de Proyecto del Departamento)	0		0	
Elementos definidos por el modelo de desarrollo					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos**	**Comentarios**

				afectados	
	1. ¿Se realizaron estudios preliminares de la entidad cliente?	0		0	
crítico	2. ¿Se identificaron las necesidades de información y las reglas del negocio?	0		0	
crítico	3. ¿Se realizó el diseño del modelo de datos correspondiente al mercado de datos Cultura y Deporte en conjunto con el cliente y los especialistas del centro?	0		0	
	4. ¿Se realizaron los diseños de los procesos Extracción, Transformación y Carga para el mercado de datos Cultura y Deporte?	0		0	
crítico	5. ¿Se realizaron los procesos de Extracción, Transformación y Carga correspondiente al mercado de datos Cultura y Deporte?	0		0	
	6. ¿Las transformaciones se pueden ejecutar desde cualquier computadora?	0		0	
crítico	7. ¿Se realizaron las implementaciones de los trabajos para el mercado de datos Cultura y Deporte?	0		0	
crítico	8. ¿Se le dan tratamiento a los errores que ocurren en el proceso de Extracción, Transformación y Carga?	0		0	

crítico	9. ¿Se realizó el proceso de Inteligencia de Negocio correspondiente al mercado de datos Cultura y Deporte?	0		0	
crítico	10. ¿Los reportes que se muestran en la capa de visualización se corresponden con las necesidades del negocio identificadas?	0		0	
crítico	11. ¿La aplicación realizada apoya el proceso de toma de decisiones para las áreas de Cultura y Deporte?	0		0	
crítico	12. ¿Se realizaron los diseños de los casos de prueba?	0		0	
crítico	13. ¿Se realizaron las Pruebas de unidad?	0		0	
crítico	14. ¿Se realizaron las Pruebas de aceptación?	0		0	
crítico	15. ¿Se realizaron las Pruebas de integración?				
crítico	16. ¿Se realizaron las Pruebas de sistema?				
crítico	17. ¿Se realizó el despliegue de la aplicación?	0		0	
	18. ¿Se realizaron las Pruebas pilotos?	0		0	
crítico	19. ¿Se le da soporte y mantenimiento a la aplicación?	0		0	
Semántica del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**

crítico	1. ¿Se han identificado errores ortográficos en los entregables?	0		0	
crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?	0		0	
	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?	0		0	

4.3 Evaluación de los resultados

Después de que se realizó la evaluación del mercado de datos, es necesario realizar una valoración de los resultados, está valoración se basa en los parámetros que se midieron en la lista de chequeo antes mencionada aplicada a los artefactos rectores y en la aplicación de los diseños de casos de prueba aplicados a la funcionalidades de la aplicación. A continuación se muestra un gráfico que refleja el resultado de las pruebas internas y las pruebas de liberación:

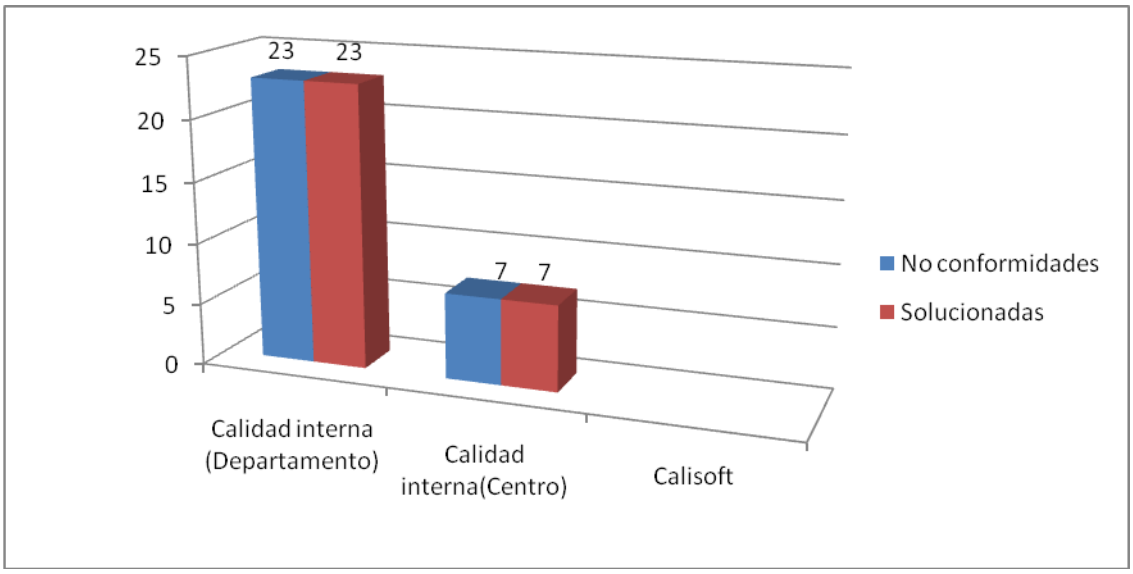

Figura 13. Resultados de la pruebas de liberación

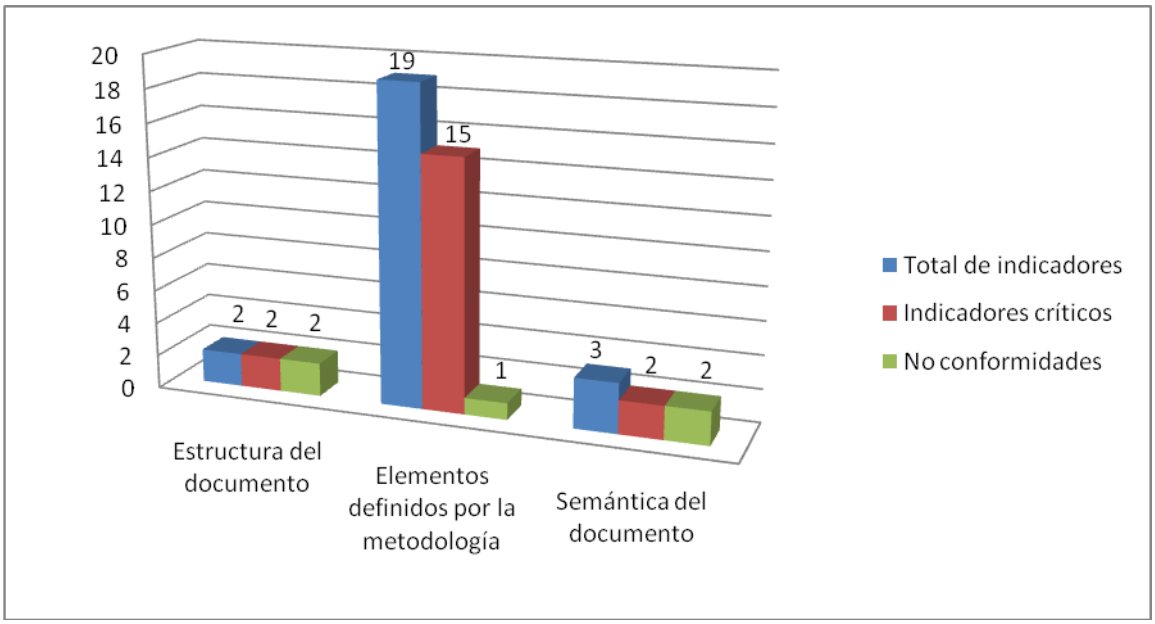

Figura 14. Aplicación de la lista de chequeo

4.4 Conclusiones

Con el objetivo de mejorar la organización y la ejecución de las pruebas se diseñaron y aplicaron 20 casos de pruebas para validar la calidad del mercado de datos, donde se obtuvieron 23 no conformidades en la primera revisión y 7 no conformidades en la segunda revisión, todas han sido resueltas. Una vez que se aplicaron las listas de chequeo al mercado de datos Cultura y Deporte se encontraron 2 no conformidades en la estructura del documento,1 no conformidad en elementos definidos por la metodología y 2 no conformidades en la semántica del documento, todas detectadas de forma general en la revisión de los artefactos y el documento de tesis, también han sido resueltas. Además se realizó la prueba de aceptación, donde el cliente hizo una revisión de las funcionalidades del mercado de datos verificando que cumpliera con los requerimientos especificados, donde se obtuvo finalmente la carta de aceptación por parte del cliente.

CONCLUSIONES

Luego de haber concluido el desarrollo del mercado de datos correspondiente a las áreas de Cultura y Deporte, como resultado se obtuvieron las siguientes conclusiones:

- El análisis y diseño del mercado de datos cumplió con todos los requerimientos especificados en el negocio.
- Se realizó el diseño del modelo de datos, identificando las estructuras que componen el mismo obteniéndose como resultado el modelo de datos para las áreas de Cultura y Deporte.
- Se realizaron los procesos de ETL para las estructuras dimensionales identificadas en el modelo de datos quedando poblada la base de datos.
- Se realizaron las implementaciones de las vistas de análisis para ambas áreas apoyando el proceso de toma de decisiones.
- Las pruebas utilizadas permitieron validar la calidad del producto y obtener resultados satisfactorios.

RECOMENDACIONES

- Se recomienda implementar una estrategia para realizar la carga de los ficheros de error obtenidos como resultado de la ejecución de las transformaciones y los trabajos para evitar la pérdida de datos.
- Se recomienda poner en explotación el mercado de datos para las áreas de Cultura y Deporte para apoyar el proceso de toma de decisiones.

Trabajos citados

1. **Jorge Eduardo Guevara Lenis, Janeth del Carmen Valencia Arcos.** [Online] Junio 2007. http://bibdigital.epn.edu.ec/bitstream/15000/445/1/CD-0827.pdf.

2. **Cabrera, María Evelia Casales.** Facultad de Ciencias. *Facultad de Ciencias.* [Online] Enero 2009. [Cited: Octubre 10, 2010.] http://hp.fciencias.unam.mx/~alg/bd/dwh.pdf.

3. **Ericka Graciela Sevilla Berríos.** [Online] Marzo 2003. [Cited: Octubre 12, 2010.] http://www.scribd.com/doc/39978465/GUIA-METODOLOGICA-PARA-LA-DEFINICION-Y-DESARROLLO-DE-UN-DATAWAREHOUSE.

4. **Sierra, Julio Ernesto Ortiz.** *Diseño e Implementación de un Mercado de Datos para la Oficina Nacional de Estadísticas.* Ciudad de la Habana : s.n., Junio de 2009.

5. **García, Gerardo Clemente.** Riunet. *Riunet.* [Online] Junio 2008 . [Cited: Noviembre 10, 2010.] http://riunet.upv.es/manakin/bitstream/handle/10251/2505/tesisUPV2842.pdf.

6. Dataprix. *Dataprix.* [Online] Mayo 12, 2009. http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/i-data-warehousing-investigacion-y-sistematizacion-concepto-13.

7. Colombia InterGrupo. *Colombia InterGrupo.* [Online] [Cited: Octubre 15, 2010.] http://www.intergrupo.com/Col_CasosExito_todos01.aspx.

8. **Rodríguez, Vivian Lorenzo.** Empai. *Empai.* [Online] 2008. [Cited: Octubre 15, 2010.] http://www.empai-matanzas.co.cu/revista/Vol.%202%20%20%20No.%202%20%20%20%202008.pdf.

9. **Berzal, Fernando.** [Online] [Cited: Noviembre 05, 2010.] http://elvex.ugr.es/idbis/db/docs/intro/F%20Modelo%20multidimensional.pdf.

10. **Wolff, Carmen Gloria.** [Online] Agosto 28, 2002. [Cited: Noviembre 05, 2010.] http://www.inf.udec.cl/~revista/ediciones/edicion4/modmulti.PDF.

11. **Sánchez, Leopoldo Zenaido Zepeda.** UNIVERSIDAD POLITÉCNICA DE VALENCIA. *Departamento de Sistemas Informáticos y Computación.* [Online] Junio 2008. http://riunet.upv.es/bitstream/handle/10251/2506/tesisUPV2841.pdf.

12. **Santos, Romina Elizabeth Dos.** Facultad de Ciencias Exactas y Naturales y Agrimensura. *Facultad de Ciencias Exactas y Naturales y Agrimensura.* [Online] 2005. [Cited: Noviembre 12, 2010.] http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/mromy.pdf.

13. MSDN. *Procesamiento y modos de almacenamiento de particiones.* [Online] 2011. [Cited: Junio 12, 2011.] http://msdn.microsoft.com/es-es/library/ms174915.aspx.

14. MIS RESPUESTAS.COM. *MIS RESPUESTAS.COM.* [Online] [Cited: Noviembre 14, 2010.] http://www.misrespuestas.com/que-es-una-metodologia.html.

15. **Alberto Límia Navarro, Anisley Delfino, Asnioby Hernández López, Doris Medina Mustelier, Iván M. Cárdenas Tandrón, Julio Ernesto Ortiz, Marisleidys Socas, Osniel Calvo, Yanet Peña Vazquez, Yanisbel Gonzales Hernández y otros compañeros del Centro UCI-Vill.** *Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocios en DATEC.* Ciudad de La Habana : s.n., 2010.

16. **Huamantumba, Rayner.** [Online] Agosto 31, 2007. [Cited: Noviembre 07, 2010.] WWW.RUEDATECNOLOGICA.COM.

17. Scribd. *Scribd.* [Online] Abril 14, 2010. http://www.scribd.com/doc/3062020/Capitulo-I-HERRAMIENTAS-CASE.

18. Free Download Manager. *Free Download Manager.* [Online] Marzo 5, 2007. [Cited: Noviembre 12, 2010.] http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML_%28M%C3%8D%29_14720_p/.

19. **Sierra, María.** [Online] http://personales.unican.es/ruizfr/is1/doc/lab/01/is1-p01-trans.pdf.

20. Stratebi. *Stratebi.* [Online] Junio 08, 2010. [Cited: Noviembre 16, 2010.] http://www.stratebi.es/todobi/jun10/Comparativa_OSBI.pdf.

21. Gravitar. *Gravitar.* [Online] http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/.

22. DataCleaner. *DataCleaner.* [Online] http://datacleaner.eobjects.org/.

23. Scribd. *Scribd.* [Online] http://www.scribd.com/doc/27519905/Servidores-Web.

24. Apache Tomcat. *Apache Tomcat.* [Online] http://tomcat.apache.org/.

25. Pentaho. *Pentaho.* [Online] http://mondrian.pentaho.com/documentation/workbench.php.

26. **Denzer, Patricio.** [Online] Octubre 23, 2002. http://profesores.elo.utfsm.cl/~agv/elo330/2s02/projects/denzer/informe.pdf.

27. **Espinoza, Humberto.** Links Global Services. *Links Global Services.* [Online] 2005. http://www.lgs.com.ve/pres/PresentacionES_PSQL.pdf.

28. pgAdmin. *PostgreSQL Tools.* [Online] [Cited: Noviembre 14, 2010.] http://www.pgadmin.org/.

29. **Espinosa, Roberto.** El Rincon del BI. [Online] Mayo 10, 2010. http://churriwifi.wordpress.com/2010/05/10/16-3-construccion-procesos-etl-utilizando-kettle-pentaho-data-integration/.

30. **Quispe-Otazu, Rodolfo.** ¿Que es la Calidad de Software? *Blog de Rodolfo Quispe-Otazu.* [Online] Diciembre 2008. [Cited: Junio 02, 2011.] http://www.rodolfoquispe.org/blog/que-es-la-calidad-de-software.php.

31. **Lamancha, Beatriz Pérez.** Proceso de Testing funcional independientemente. *Tesis de Maestría en Informática.* [Online] 2006. [Cited: Junio 04, 2011.] http://www.fing.edu.uy/~bperez/.../Tesis%20-%20Beatriz%20Perez-%202006.pdf.

32. **Lianet Lores Sánchez, Diana Monné Roque.** *Aplicación de las pruebas de liberación al Sistema Informático de Genética Médica. Trabajo de Diploma para optar por el título de Ingeniero Informático. Universidad de las Ciencias Informáticas.* Ciudad de La Habana : s.n., Junio 2009.

Bibliografía

1. **Alberto Límia Navarro Anisley Delfino, Asnioby Hernández López, Doris Medina Mustelier, Iván M. Cárdenas Tandrón, Julio Ernesto Ortiz, Marisleidys Socas, Osniel Calvo, Yanet Peña Vazquez, Yanisbel Gonzales Hernández y otros compañeros del Centro UCI-Vill** Metodología para el Desarrollo de Soluciones de Almacenes de Datos e Inteligencia de Negocios en DATEC [Report]. - Ciudad de La Habana : [s.n.], 2010.
2. **Antonio Aliaga Ibarra Marcos Agustin Miani Flores** I.E.S. San Vicente [Online] // I.E.S. San Vicente. - Enero 21, 2008. - Noviembre 2010. - http://www.iessanvicente.com/colaboraciones/postgreSQL.pdf.
3. Apache Tomcat [Online] // Apache Tomcat. - Noviembre 2010. - http://tomcat.apache.org/.
4. **B. Ing. Alexander Oré** [Online]. - 2009. - 05 10, 2011. - http://www.calidadysoftware.com/testing/pruebas_funcionales.php.
5. **Berzal Fernando** [Online]. - Noviembre 05, 2010. - http://elvex.ugr.es/idbis/db/docs/intro/F%20Modelo%20multidimensional.pdf.
6. **Cabrera María Evelia Casales** Facultad de Ciencias [Online] // Facultad de Ciencias. - Enero 2009. - Octubre 10, 2010. - http://hp.fciencias.unam.mx/~alg/bd/dwh.pdf.
7. Colombia InterGrupo [Online] // Colombia InterGrupo. - Octubre 15, 2010. - http://www.intergrupo.com/Col_CasosExito_todos01.aspx.
8. DataCleaner [Online] // DataCleaner. - Noviembre 2010. - http://datacleaner.eobjects.org/.
9. Dataprix [Online] // Dataprix. - Mayo 12, 2009. - Noviembre 2010. - http://www.dataprix.com/data-warehousing-y-metodologia-hefesto/i-data-warehousing-investigacion-y-sistematizacion-concepto-13.
10. **Denzer Patricio** [Online]. - Octubre 23, 2002. - Noviembre 2010. - http://profesores.elo.utfsm.cl/~agv/elo330/2s02/projects/denzer/informe.pdf.
11. Ericka Graciela Sevilla Berríos [Online]. - Marzo 2003. - Octubre 12, 2010. - http://www.scribd.com/doc/39978465/GUIA-METODOLOGICA-PARA-LA-DEFINICION-Y-DESARROLLO-DE-UN-DATAWAREHOUSE.
12. **Espinosa Roberto** El Rincon del BI [Online]. - Mayo 10, 2010. - Marzo 2011. - http://churriwifi.wordpress.com/2010/05/10/16-3-construccion-procesos-etl-utilizando-kettle-pentaho-data-integration/.

13. **Espinoza Humberto** Links Global Services [Online] // Links Global Services. - 2005. - Noviembre 2010. - http://www.lgs.com.ve/pres/PresentacionES_PSQL.pdf.
14. Free Download Manager [Online] // Free Download Manager. - Marzo 5, 2007. - Noviembre 12, 2010.http://www.freedownloadmanager.org/es/downloads/Paradigma_Visual_para_UML_%28M%C3%8D%29_14720_p/.
15. **García Gerardo Clemente** Riunet [Online] // Riunet. - Junio 2008 . Noviembre 10, 2010. - http://riunet.upv.es/manakin/bitstream/handle/10251/2505/tesisUPV2842.pdf.
16. **Geiger Claudia Imhoff. Nicholas Galemmo. Jonathan G.** Mastering Data Warehouse Design, Relational and Dimentional Techniques [Book]. - 2003. - 0-471-32421-3.
17. Gravitar [Online] // Gravitar. - Noviembre 2010. - http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/.
18. **Herrera Cristhian** Características del almacén de datos [Online]. - 10 30, 2007. - http://www.adictosaltrabajo.com/tutoriales/tutoriales.php?pagina=datawarehouse.
19. **Huamantumba Rayner** [Online]. - Agosto 31, 2007. - Noviembre 07, 2010. - WWW.RUEDATECNOLOGICA.COM.
20. **Jorge Eduardo Guevara Lenis Janeth del Carmen Valencia Arcos** [Online]. - Junio 2007. - Octubre 2010. - http://bibdigital.epn.edu.ec/bitstream/15000/445/1/CD-0827.pdf.
21. **Lamancha Beatriz Pérez** Proceso de Testing funcional independientemente [Online] // Tesis de Maestría en Informática. - 2006. - Junio 04, 2011. - http://www.fing.edu.uy/~bperez/.../Tesis%20-%20Beatriz%20Perez-%202006.pdf.
22. **Lianet Lores Sánchez Diana Monné Roque** Aplicación de las pruebas de liberación al Sistema Informático de Genética Médica. Trabajo de Diploma para optar por el título de Ingeniero Informático. Universidad de las Ciencias Informáticas [Report]. - Ciudad de La Habana: [s.n.], Junio 2009.
23. MIS RESPUESTAS.COM [Online] // MIS RESPUESTAS.COM. - Noviembre 14, 2010. - http://www.misrespuestas.com/que-es-una-metodologia.html.
24. Modelamiento multidimensional [Online]. - 05 30, 2011. - http://www.google.com/url?sa=t&source=web&cd=1&ved=0CBsQFjAA&url=http%3A%2F%2Fwww.inf.udec.cl%2F~revista%2Fediciones%2Fedicion4%2Fmodmulti.PDF&rct=j&q=modelo%20multidimensional%20&ei=PnDuTaTdB6L00gGztszfAw&usg=AFQjCNGUM1nUN0UrK8yOYBuxFI3dXBnnnQ&sig2=iGxzY.
25. MSDN [Online] // Procesamiento y modos de almacenamiento de particiones. - 2011. - Junio 12, 2011. - http://msdn.microsoft.com/es-es/library/ms174915.aspx.

26. Pentaho [Online] // Pentaho. - Noviembre 2010. - http://mondrian.pentaho.com/documentation/workbench.php.
27. **Peñaloza Lucía Victoria Hernández** Tesis para lograr el título de Magíster: Diseño y Construcción de un Data Mart para la mantención de Indicadores de Sostenibilidad de la Industria del Salmón. Chile [Book Section]. - Chile : [s.n.], 2008.
28. PgAdmin [Online] // PostgreSQL Tools. - Noviembre 14, 2010. - http://www.pgadmin.org/.
29. **Quispe-Otazu Rodolfo** ¿Que es la Calidad de Software? [Online] // Blog de Rodolfo Quispe-Otazu. - Diciembre 2008. - Junio 02, 2011. - http://www.rodolfoquispe.org/blog/que-es-la-calidad-de-software.php.
30. **Rodríguez Vivian Lorenzo** Empai [Online] // Empai. - 2008. - Octubre 15, 2010. - http://www.empai-matanzas.co.cu/revista/Vol.%202%20%20%20No.%202%20%20%20%202008.pdf.
31. **Ross Ralph Kimball y Margy** The Data Warehouse Toolkit. [Book]. - 2002.
32. **Sánchez Leopoldo Zenaido Zepeda** UNIVERSIDAD POLITÉCNICA DE VALENCIA [Online] // Departamento de Sistemas Informáticos y Computación. - Junio 2008. - Noviembre 2010. - http://riunet.upv.es/bitstream/handle/10251/2506/tesisUPV2841.pdf.
33. **Santos Romina Elizabeth Dos** Facultad de Ciencias Exactas y Naturales y Agrimensura [Online] // Facultad de Ciencias Exactas y Naturales y Agrimensura. - 2005. - Noviembre 12, 2010. - http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/mromy.pdf.
34. **Santos Romina Elizabeth Dos** Facultad de Ciencias Exactas y Naturales y Agrimensura [Online] // Facultad de Ciencias Exactas y Naturales y Agrimensura. - 2005. - Noviembre 12, 2010. - http://exa.unne.edu.ar/depar/areas/informatica/SistemasOperativos/mromy.pdf.
35. Scribd [Online] // Scribd. - Noviembre 2010. - http://www.scribd.com/doc/27519905/Servidores-Web.
36. Scribd [Online] // Scribd. - Abril 14, 2010. - Noviembre12 2010. - http://www.scribd.com/doc/3062020/Capitulo-I-HERRAMIENTAS-CASE.
37. **Sierra Julio Ernesto Ortiz** Diseño e Implementación de un Mercado de Datos para la Oficina Nacional de Estadísticas [Report]. - Ciudad de la Habana : [s.n.], Junio de 2009.
38. **Sierra María** [Online]. - Noviembre 2010. - http://personales.unican.es/ruizfr/is1/doc/lab/01/is1-p01-trans.pdf.
39. Sinnexus [Online] // Persistencia MOLAP, ROLAP, HOLAP. - Junio 2011. - http://www.dataprix.com/category/tags/rolap.

40. Stratebi [Online] // Stratebi. - Junio 08, 2010. - Noviembre 16, 2010. - http://www.stratebi.es/todobi/jun10/Comparativa_OSBI.pdf.
41. **Wolff Carmen Gloria** [Online]. - Agosto 28, 2002. - Noviembre 05, 2010. - http://www.inf.udec.cl/~revista/ediciones/edicion4/modmulti.PDF.

ANEXOS

UCI Universidad de las Ciencias Informáticas

Carta de aceptación del cliente.

Fecha: 1 / 6 /2011

Yo:

Elena Leonila Fernández García, representante de la Oficina Nacional de Estadísticas en la Universidad de las Ciencias Informáticas para el desarrollo del Sistema de Información de Gobierno. Apruebo:

1- El mercado de datos para el área de CULTURA Y DEPORTE cumple con los requisitos especificados por el cliente.

Elena Leonila Fernández García
Firma del cliente.

Figura 1. Carta de aceptación del cliente

GLOSARIO DE TÉRMINOS

ONE: Oficina Nacional de Estadísticas.

SIGOB: Sistema de Información de Gobierno.

DATEC: Centro de Tecnologías de Gestión de Datos.

Data Warehouse: almacén de datos.

Data Mart: mercado de datos.

ETL: proceso de extracción, transformación y carga.

BI: Inteligencia del negocio.

Indicador: son los valores que toman determinadas variables cuando se analizan

MOLAP: Multidimensional On Line Analytical Processing

OLAP: On Line Analytical Processing

OLTP: On Line Transactional Processing

ROLAP: Relational On Line Analytical Processing

SGBD: Sistemas de gestión de bases de datos

MDX: Multidimensional Query eXpression

UML: Unified Modeling Language

XML: Extensible Markup Language

SQL: Structured Query Language