Universidad de las Ciencias Informáticas

Facultad 6

Título: Sistema de Información de Gobierno. Mercado de datos Ocupación.

Autores:

Amaia Yoldi Iglesias.

Ransel Fidel Ferrer Lara.

Tutores:

Ing. Yuneimy Tellez Pérez.

Ing. Rayko Emilio Torres Cruz.

22 de junio del 2011

Declaramos ser autores de la presente tesis y reconocemos a la Universidad de las Ciencias Informáticas los derechos patrimoniales de la misma, con carácter exclusivo.

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

Autores: _____ _____

Amaia Yoldi Iglesias Ransel Fidel Ferrer Lara

Tutores: _____ _____

Ing. Yuneimy Tellez Pérez Ing. Rayko Emilio Torres Cruz

Tutores:

Tutor: Ing. Yuneimy Téllez Pérez

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: Instructor en Adiestramiento

Categoría Científica: Ingeniero

Años de experiencia en el tema: 2

Años de graduado: 2

Correo Electrónico: ytellez@uci.cu

Tutor: Ing. Rayko Emilio Torres Cruz

Especialidad de graduación: Ingeniería en Ciencias Informáticas

Categoría docente: Instructor en Adiestramiento

Categoría Científica: Ingeniero

Años de experiencia en el tema: 0

Años de graduado: 1

Correo Electrónico: retorres@uci.cu

Agradecimientos

Amaia:

A todos mis compañeros del proyecto, especialmente a Susy, Laydi y David por la ayuda brindada en el transcurso del curso y los buenos momentos compartidos en el laboratorio. A Roberto Tellez, jefe del departamento Almacenes de Datos, por toda la ayuda brindada durante esta etapa difícil.

A mis tutores y tribunal, por su preocupación y sus consejos, su ayuda, sus críticas constructivas, su paciencia y el apoyo brindado durante el transcurso del curso.

A todos mis amigos, que se convirtieron en mi familia de la UCI. Principalmente al antiguo grupo 6303.

A Jose, Felo, Leandro y Ernesto por apoyarme, soportarme y ayudarme en los momentos más difíciles de la carrera y sobre todo por su amistad.

Especialmente a Viviana y a Maylin, por permitirme ser su amiga y brindarme su confianza, lealtad y apoyo incondicional. Por estar junto a mí en los buenos momentos y sobre todo en los difíciles, brindándome su cariño, comprensión y consejos, convirtiéndose fácilmente en mis mejores amigas.

A mi familia, mis tías, mis primos, mis abuelos, mi papa y sobre todo a mi mami por el apoyo que me ha brindado en los momentos que más lo necesite y la confianza que ha tenido en mi.

A Leonardo por su apoyo, comprensión, amistad y cariño.

En fin, a todas aquellas personas que me ayudaron de una forma u otra y se preocuparon por que este sueño se haga realidad.

Muchas gracias.

Ransel:

Le agradezco a mi mamá por siempre confiar en mí y apoyarme en todo momento. A mi padre por ser tan positivo y por sus grandes consejos. A mi novia Darmalina por estar siempre a mi lado y hacerme ver que no hay nada imposible. A Lucía Cuza por estar pendiente constantemente de mis problemas. A Fina, Anita y Mercedes por ser como madres para mí. A mis tutores que se entregaron en esta tesis junto conmigo. A Amaia por compartir cada derrota, cada victoria y tantas noches de desvelo.

Le agradezco a mis amigos Annia, Luis, Playa, Dachelys, Yudith, Tenorio, Lachy, Alain, Jesús, Alexis y Montano, por apoyarme en todo momento, estar pendiente del progreso de este trabajo y ser amigos incondicionales.

Agradezco a mis compañeros de grupo de primer año especialmente a Manresa, Orelmis, Adrian, Leonel, Yoendi, Fabian, Yuniel, Liniuska, Taimi, Susel, Delmis, Amauri, Yelena, Arleydis, Jose, Liset y Yenisleydi que me han ayudado tanto durante el transcurso de la carrera y han estado a mi lado cuando más los he necesitado.

Agradecer también a mis compañeros del proyecto que han puesto a mi disposición su tiempo y todo lo que he necesitado, en especial a Laydi, Leo, Yulio, Patricia, Laritza, Aylenis y Dainelis.

Agradecer a todos los profesores que me apoyaron y confiaron en mí en especial a Yoendri Lacoste y Doris Medina.

Agradecer a todas las chicas del apartamento 105103, que siempre se preocuparon por mis resultados y me dieron ánimo a seguir siempre que lo necesite.

Dar las gracias a todas las personas que hicieron posible la realización de este trabajo, en especial a Elena.

Dedicatoria

Amaia:

Quiero dedicar este trabajo a mi madre, por apoyarme tanto en las buenas como en las malas y estar siempre presente cuando la necesito.

Ransel:

Dedicado a la memoria de Angeles del Carmen Orbeal y Suarez del Villar, mi abuela, por ser la persona que consagró su vida a cuidarme, enseñarme todo lo que se, y con paciencia y gran sabiduría saber guiarme.

Resumen

La Oficina Nacional de Estadística (ONE), es la entidad encargada de almacenar, analizar y gestionar toda la Información Estadística del país. La ONE trabaja con datos históricos que provienen de todas las provincias del país, por lo que la información que almacena está definida de diferentes formas y no se encuentra debidamente integrada. Para dar solución a estos problemas, comienza a trabajar conjuntamente con la Universidad de las Ciencias Informáticas (UCI), en la construcción de un Almacén de Datos (AD), denominado, Sistema de Información de Gobierno (SIGOB). El mismo está compuesto por varios Mercados de Datos (MD), que permitirán integrar y homogeneizar la información con la que trabaja el AD, brindando, el soporte con la información necesaria y oportuna, que sustente el desarrollo y adecuado funcionamiento, de un eficiente proceso de toma de decisiones, en los diferentes Sectores Socioeconómicos.

El presente Trabajo de Diploma, da continuidad a la Tesis titulada “Análisis, Diseño e Implementación, del Mercado de Datos para los indicadores específicos de la Ocupación de la Oficina Nacional de Estadística”. Partiendo de la misma base teórica e integrando las nuevas exigencias del cliente y la investigación de las metodologías y herramientas más utilizadas en el desarrollo de los MD. Se realiza el análisis, diseño e implementación, de los subsistemas de integración e Inteligencia de Negocio, que permita el análisis estadístico, partiendo de los reportes contenidos en el libro de trabajo del área Ocupación. Además, de realizar la aplicación de los casos de pruebas y la validación de las listas de chequeo, con el objetivo, de obtener un Mercado Datos que recoge toda la información del indicador de Ocupación del país, que forma parte del SIGOB y satisfaga las necesidades del cliente.

Palabras claves: Almacén de Datos, Mercado de Datos, Oficina Nacional de Estadísticas, Sistema de Información de Gobierno.

Tabla de Contenido

Índice de Figuras

Índice de Tablas

Introducción

En la actualidad, el mundo se encuentra inmerso en un acelerado desarrollo tecnológico. En las últimas décadas, no solo el desarrollo alcanzado en la informática, sino también en la microelectrónica y las telecomunicaciones, han dado al traste, a lo que de modo general se ha denominado, las nuevas tecnologías de la información y las comunicaciones, las cuales en un proceso acelerado de convergencia, penetran en diversos ámbitos de la vida humana: el trabajo, la enseñanza, el hogar y la distracción entre muchos otros, siendo el desarrollo alcanzado, cada vez más dinámico e impredecible.
El avance alcanzado en la informática a nivel mundial, se conoce como la nueva Revolución Tecnológica, su carácter sistémico, estratégico y de penetración generalizada, hace que cada día desempeñe un mayor rol, por ello se aprecia su inserción, en todos los procesos cotidianos, encontrándose, fuertemente vinculada a los cambios económicos, políticos y sociales de la época.
Cuba no ha estado exenta del progreso creciente y constante de la informática, la que constituye además, en su aplicación a los diversos procesos laborales, un instrumento vital para la obtención de información confiable y actualizada en los procesos de dirección y toma de decisiones. Debido a esto, muchas de las Empresas, se han visto obligadas a informatizarse y digitalizar la información que poseen, para lograr mayor eficiencia en su desempeño y a su vez, establecer procesos de retroalimentación y comunicación con otras entidades e instituciones, facilitando los convenios y relaciones entre las mismas.
En Cuba se pueden encontrar diversas Universidades, que cuentan con la preparación suficiente, para alcanzar un gran desarrollo en la informática. Una de las principales es la Universidad de las Ciencias Informáticas (UCI), que fue creada como proyecto de la Revolución y que ofrece múltiples servicios, que contribuyen a la informatización del país, por lo que se ha convertido, en la principal propulsora del desarrollo de producción de software en Cuba. En ella se desarrollan distintos proyectos productivos, que contribuyen con resultados significativos a la adquisición de nuevos conocimientos.
La UCI actualmente, se encuentra trabajando, conjuntamente con la Oficina Nacional de Estadísticas (ONE), en el proyecto Sistema de Información de Gobierno (SIGOB), constituyendo una de las prioridades, la centralización e integración de las bases de datos de diferentes fuentes, lo que servirá para proveer la información actualizada a los Órganos de Administración del Gobierno.
La ONE se encarga de gestionar la información estadística del país y es la responsable de resguardar, disponer y decidir, cuales datos son confidenciales y establecer el acceso a la misma. Obtiene la información a través de 15 Delegaciones Territoriales y 168 Oficinas Municipales, además de los Organismos de la Administración Central del Estado y otras Instituciones.

Entre las diferentes áreas en las que trabaja la ONE, se encuentra la de Ocupación, la cual se encarga de recopilar, toda la información referente a los indicadores relacionados con el empleo y los salarios. Además, recopila información sobre la población económicamente activa (ocupados y desocupados), la fuerza de trabajo ocupada según la situación del empleo, la estructura por clase de la actividad económica, el sexo, la edad, el nivel educacional, los tipos de contratos de trabajo y la categoría ocupacional. Se incluyen datos sobre el salario medio por territorios y clase de actividad económica, así como indicadores sobre accidentes de trabajo, seguridad social y asistencia social.

Una de las dificultades detectadas en la ONE, es que la información recogida, se encuentra contenida en archivos con formatos, que solo pueden ser consultados por un informático, o un especialista de la información que tenga conocimiento previo del negocio. También la existencia de múltiples versiones de los datos, intervienen negativamente en la posibilidad, de que estos se integren y estandaricen correctamente, logrando una buena calidad. Estos problemas, traen consigo, que existan limitaciones para recuperar indicadores desde diferentes perspectivas de análisis, además, de que provocan que el proceso de recuperación y elaboración de informes, resulte costoso en tiempo y esfuerzo, lo que dificulta la disponibilidad de información confiable para los órganos que la requieran. Anteriormente la ONE en conjunto con DATEC realizó el análisis y diseño del Mercado de Datos Ocupación, sin embargo, en la actualidad han surgido nuevas necesidades, por lo que se hizo necesario realizar el refinamiento de esta área debido a los problemas existentes.

A partir de esta situación surge el siguiente **problema de la investigación**: ¿Cómo contribuir a la toma de decisiones en el área Ocupación para el Sistema de Información del Gobierno?

En correspondencia con el problema definido se plantea como **objeto de estudio**: los Almacenes de Datos, enmarcado en el **campo de acción** el proceso de desarrollo del Mercado de Datos para el área Ocupación del Sistema de Información del Gobierno.

Para dar solución al problema de la investigación, se propone como **objetivo general**: desarrollar el Mercado de Datos, Ocupación del Sistema de Información del Gobierno, que contribuya a la toma de decisiones para la ONE.

Del mismo se desglosan los siguientes **objetivos específicos**:

1. Refinar el análisis y diseño del Mercado de Datos del área Ocupación.
2. Implementar el Mercado de Datos del área Ocupación.
3. Validar el Mercado de Datos del área Ocupación.

Para darle cumplimiento a los objetivos específicos propuestos, se plantearon las siguientes **tareas de investigación:**

1. Caracterización de las metodologías, herramientas y tecnologías a utilizar en el desarrollo de Almacenes de Datos.
2. Refinamiento de los requisitos.
3. Refinamiento de los hechos, las medidas y las dimensiones del Mercado de Datos.
4. Refinamiento del Modelo de Datos.
5. Refinamiento de la arquitectura del Mercado de Datos.
6. Diseño del subsistema de integración.
7. Diseño del subsistema de visualización.
8. Diseño de los casos de pruebas.
9. Implementación del subsistema de integración.
10. Implementación del subsistema de visualización.
11. Aplicación de las listas de chequeo.
12. Aplicación de los casos de pruebas.

Para dar cumplimiento a los objetivos trazados, el presente documento está estructurado en cuatro capítulos, en los que se describen los métodos y procedimientos a seguir. A continuación se expone una breve descripción de cada uno de ellos.

Capítulo 1 - Fundamento teórico

En este capítulo se abordan temas referentes a la descripción del objeto de estudio y el campo de acción; así como la selección de la metodología de desarrollo, tecnologías y herramientas que se utilizan para desarrollar el presente trabajo de investigación.

Capítulo 2 - Análisis y diseño del Mercado de Datos:

En este capítulo se definen los principales artefactos de la etapa de análisis y diseño, tales como: temas de análisis, las necesidades de información, requisitos funcionales y no funcionales. También se refinan las dimensiones del Mercado de Datos, los hechos asociados, la estructura del modelo dimensional y las transformaciones al diseño físico.

Capítulo 3 - Implementación del Mercado de Datos:
En este capítulo se describe como se implementa la base de datos y se desarrollan los mecanismos de integración y visualización de la información.

Capítulo 4 - Validación del Mercado de Datos de la Ocupación:
En este capítulo se realizan las validaciones a la solución Mercado de Datos del área de Ocupación aplicando casos de prueba y listas de chequeo.

Capítulo 1: Fundamento Teórico

En el presente capítulo se abordan los elementos teóricos necesarios para la realización del trabajo de investigación. Se realiza un estudio de las principales características, ventajas y desventajas de los Almacenes de Datos y los Mercados de Datos. También se detallan los elementos fundamentales de algunas de las metodologías, tecnologías y herramientas más utilizadas en el mundo para la implementación de los Almacenes de Datos.

1.1 Soluciones existentes en el ámbito internacional y nacional

En la actualidad, son numerosas las empresas del mundo interesadas en integrar y centralizar los datos con los que trabaja, con el objetivo de reducir costos y obtener una mejor información del negocio. Muchas de ellas, manejan un gran cúmulo de información, lo que provoca que el uso de los Almacenes de Datos sea imprescindible para su correcto funcionamiento. Para el desarrollo de este trabajo, se efectuó el estudio de algunas de estas empresas, las cuales tratan varios indicadores incluyendo el indicador de Ocupación.

1.1.1 Soluciones internacionales

Instituto Nacional de Estadísticas en Chile.

El Instituto Nacional de Estadísticas (INE) está encargado de producir, recopilar, analizar y publicar las estadísticas oficiales de Chile. La INE mediante la información recopilada ha permitido conocer cómo ha crecido la población chilena, el desarrollo cultural y la evolución de la economía, entre otros temas. Esta información es utilizada en el desarrollo de investigaciones y para la toma de decisiones. Algunos indicadores en los que trabaja son: Población, Vivienda, Empleo, Economía, Índice de precios al consumidor, Encuesta de presupuestos familiares, Censos agropecuarios, Cultura, Seguridad ciudadana, Social, Demográfica, Medioambiental, entre otras, con la finalidad que los agentes públicos, privados, investigadores y ciudadanos tomen decisiones informadas y así fortalecer una sociedad abierta y democrática. Su finalidad es coordinar y mejorar la pertinencia de la producción estadística, intercambiar puntos de vista y normar estándares comunes. El INE en Chile presenta la información a sus usuarios en un sistema que contiene datos relevantes. Su sección Micro Datos presenta un apéndice que aborda directamente el tema del indicador de Ocupación, permitiendo acceder a la base de datos y realizar cruces de variables (Ineatacama).

Instituto Nacional de Estadísticas en España.

La organización estadística en España se crea para la obtención de estadísticas oficiales de este país, pues su misión principal es la elaboración y perfección de los datos recogidos de distintos indicadores. Presentan un apartado de oferta pública, en el cual se recoge cualquier convocatoria de empleo, ya sea por contratación temporal o fija. La mayor complejidad de esta organización es la descentralización de la producción estadística dentro de la Administración General del Estado. La información de este INE se publica en su página web, donde presenta información de diversos indicadores, como es el caso del indicador sociedad, donde almacena datos sobre el mercado laboral. Recoge operaciones estadísticas elaboradas por el INE y por otros organismos. Se estructura por temas y apartados, para que la localización de la misma sea de forma sencilla. La información estadística está permanentemente actualizada (INE, 2010).

1.1.2 Soluciones nacionales

En Cuba existen diferentes entidades que han desarrollado Almacenes de Datos para la gestión de su información, entre ellos se encuentran: la Corporación COPEXTEL, la compañía de Unión Cuba Petróleo (CUPET) y la Corporación CIMEX.

La Oficina Nacional de Estadísticas es la entidad creada para proponer, organizar y ejecutar, la materia estadística del país. Se dedica a recolectar, organizar, resumir, presentar y analizar datos relativos a un conjunto de objetos, personas, procesos, entre otros y resulta una herramienta de suma utilidad para la toma de decisiones. Garantiza la producción de estadísticas de calidad a través del Sistema Estadístico Nacional, ejerciendo una adecuada dirección, ejecución y control de la captación de las cifras económicas y sociales, así como su adecuada difusión de acuerdo con las necesidades del país en información estadística.

La ONE tiene como objetivo integrar toda la información resultante de diferentes fuentes para lograr un mejor monitoreo y control de los datos, expresando la evolución de los diferentes indicadores con los que trabaja, entre los que se encuentra Ocupación. Este indicador[1] se relaciona directamente con el empleo y los salarios, los cuales tienen como propósito ofrecer los datos de forma que presenten una visión más integral sobre el tema.

[1] Magnitud utilizada para medir o comparar los resultados efectivamente obtenidos en la ejecución de un proyecto, programa o actividad. Los indicadores pueden ser cualitativos o cuantitativos.

La ONE ha decidido realizar un Almacén de Datos para cumplir su objetivo de forma satisfactoria. Luego de analizar los sistemas de la INE en España y Chile se decide construir un Mercado de Datos para los indicadores de Ocupación, dado que realizar una adaptación de estos sistemas no resolvería las necesidades que presenta la ONE en el área de Ocupación.

1.2 Tecnologías de almacenamiento de datos

1.2.1 Almacenes de datos

La aparición de los ordenadores para la automatización de la información introdujo un cambio considerable para la gestión de la misma, desplazando el centro de informática hacia la estructuración de los datos, convirtiéndose las bases de datos en una herramienta esencial para el control y el manejo de las operaciones del comercio, debido a la necesidad de las empresas de disponer de gran cúmulo de información almacenada en diferentes fuentes de datos. A su vez, esta acumulación de información evidencia que podría ser de gran ayuda tener almacenada en un mismo lugar la mayoría de las operaciones.

Debido a la idea de unir las distintas fuentes de información en un lugar único, para la futura introducción de la documentación relevante y como respuesta a la misma, es que surgen los almacenes de datos.

De forma general se puede decir que un Almacén de Datos es una base de datos que se caracteriza por la purificación e integración de la información recogida de diferentes fuentes para después procesarla.

Dos de los conceptos de Almacén de Datos más importantes, son los planteados por Ralph Kimball y Willian H. Inmon.

Ralph Kimball propone que: “Un Almacén de Datos es una copia de los datos transaccionales específicamente estructurada para la consulta y el análisis” (Kimball, 2006).

Willian H. Inmon propone que: “Un Almacén de Datos es una colección de datos orientados a temas, integrados, no volátiles y variante de tiempo, organizados para soportar las necesidades del cliente” (Inmon, 2006).

Entre las características de los almacenes de datos se pueden mencionar las siguientes:

Orientado al Tema: Los datos están almacenados por materias o temas. Estos se organizan desde la perspectiva del usuario final, mientras que en las bases de datos operacionales se organizan desde la perspectiva de la aplicación, con vistas a lograr una mayor eficiencia en el acceso a los datos.

Integrado: Todos los datos almacenados están integrados. Las bases de datos operacionales orientadas hacia las aplicaciones fueron creadas sin pensar en su integración, por lo que un mismo tipo de datos puede ser expresado de distinta manera en dos bases de datos operacionales (Por ejemplo, para representar el sexo: 'Femenino' y 'Masculino', 'F' y 'M' o '0' y '1').

No volátil: Únicamente hay dos tipos de operaciones en el Almacén de Datos: la carga de los datos procedentes de los entornos operacionales (carga inicial y carga periódica) y la consulta de los mismos. La actualización de datos no forma parte de la operativa normal.

Histórico: El tiempo debe estar presente en todos los registros. Las bases de datos operacionales contienen los valores actuales de los datos. Un Almacén de Datos no es más que una serie de instantáneas en el tiempo, tomadas periódicamente (Torres, 2006).

El uso de los sistemas de almacenes de datos provee las siguientes ventajas:

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

Además de las siguientes desventajas:

- ✓ Infravaloración del esfuerzo necesario para su diseño y creación.
- ✓ Infravaloración de los recursos necesarios para la captura, carga y almacenamiento 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.

Se puede concluir que un Almacén de Datos proporciona una visión global e integrada de la información de la entidad. Este se encargará de entregar la información correcta, en el momento óptimo al personal indicado, optimizando los procesos de la empresa. Para que el Almacén de Datos presente una correcta construcción, los datos contenidos de las diferentes fuentes deben estar bien integrados. Además, su información debe ser histórica, no volátil y orientada a la organización. Permitiendo el análisis de la información contenida en ella y brindando soporte al proceso de toma de decisiones.

1.2.2 Mercado de Datos

Un Mercado de Datos es un subconjunto de datos de un almacén con el objetivo de responder a los requisitos de un departamento o área específica del negocio. Es el almacén natural para los datos departamentales. En cambio, el ámbito del Almacén de Datos es la organización en su conjunto. El Mercado 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, facilitando su construcción y utilización.

Características de los Mercados de Datos.

- ✓ Se centra en los requisitos de los usuarios asociados a un departamento o área de negocio concreto.
- ✓ Como diferencia con los almacenes de datos, los Mercados no contienen datos operacionales detallados.
- ✓ 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 (Francisco José Lucas-Torres Torrillas, 2008-2009).

1.3 Metodologías de desarrollo

El vocablo metodología, se refiere a la ciencia que estudia los métodos o procedimientos de investigación, que se siguen, para alcanzar una gama de objetivos en una ciencia. En múltiples disciplinas, existen diferentes enfoques para abordar un mismo concepto o problema. La existencia de dichos enfoques, enriquece en gran modo la propia disciplina.

El diseño de un Almacén de Datos, como disciplina, ha alcanzado un grado de madurez considerable a lo largo de estos años, por lo que presenta diferentes enfoques. Existen dos criterios bien identificados y que han marcado claramente su tendencia, sirviéndole de guía a la comunidad mundial; estas tendencias son conocida como: Metodología Kimball y Metodología Inmon, en honor de sus creadores Ralph Kimball y William H. Inmon.

La visión de Inmon se basa principalmente en un enfoque descendente (Top-down). Los datos son extraídos de los sistemas operacionales por los procesos de Extracción, Transformación y Carga (ETL) y cargados en las áreas de almacenamiento, donde son validados y consolidados en el Almacén de Datos corporativo. Una vez realizado este proceso, los procesos de refresco de los Mercados de Datos departamentales obtienen la información de él, y con las consiguientes transformaciones, organizan los datos en las estructuras particulares requeridas por cada uno de ellos, refrescando su contenido.

La visión de Kimball se basa principalmente en un enfoque ascendente (Bottom-up), pues al final el Almacén de Datos corporativo no es más que la unión de los diferentes Mercados de Datos. Esta característica le hace más flexible y sencillo de implementar, pues podemos construir un Mercado de Datos como primer elemento del sistema de análisis, y luego ir añadiendo otros que comparten las dimensiones ya definidas o incluyen otras nuevas. En este sistema, los procesos de Extracción, Transformación y Carga (ETL) extraen la información de los sistemas operacionales y los procesan igualmente en el área de almacenamiento, realizando posteriormente el llenado de cada uno de los Mercados de Datos de una forma individual, aunque siempre respetando la estandarización de las dimensiones.

Para la implementación del Almacén de Datos Sistema de Información de Gobierno se propone como metodología una adaptación de la metodología planteada por Ralph Kimball y lo planteado en la tesis de doctorado de Leopoldo Zenaido Zepeda Sánchez.

Ralph Kimball conduce a una solución completa en un período de tiempo relativamente pequeño. Tiene bien definidas las fases de desarrollo, proporciona una mayor agilidad en el proceso de desarrollo y creación de las tablas de dimensiones y hechos de la solución. La relación entre las tablas concede a cualquier usuario la posibilidad de construir consultas muy sencillas. El enfoque ascendente (bottom-up) permite que, partiendo de cero, se pueda obtener información útil en cuestión de días. Los Mercados de Datos resultantes son fácilmente consultables tanto para los desarrolladores como para los usuarios finales. Cuenta con una abundante documentación, que permite encontrar una respuesta a casi todas las preguntas que puedan surgir. Asegura que los usuarios interactúen con un sistema fácil de entender (Sanz, 2010).

La tesis de doctorado de Leopoldo Zenaido Zepeda Sánchez fortalece la etapa de levantamiento de requisitos, orientando el trabajo a los casos de uso, lográndose estar más alineado con las tendencias y normas de la universidad. En el ciclo de vida tienen lugar las siguientes disciplinas de trabajo:

- ✓ Estudio preliminar o planeación.
- ✓ Requerimientos.
- ✓ Arquitectura y diseño.
- ✓ Implementación.
- ✓ Prueba.
- ✓ Despliegue.
- ✓ Soporte y mantenimiento.
- ✓ Gestión y administración del proyecto.

1.4 Modelos de datos

Los modelos de datos se utilizan para describir las características y las relaciones que existen entre los datos para su posterior manipulación. Con el desarrollo de las bases de datos se han empleado modelos para realizar el diseño de las mismas, entre los cuales se pueden mencionar el modelo relacional y el modelo dimensional.

Modelo relacional: "El diagrama entidad-relación es un lenguaje para realizar el modelado de los datos de un sistema de información, se basa en la separación de los datos en entidades para formar parte del diseño físico**"** (Castillo, 2008).

Modelo dimensional: "Es uno de los más reconocidos en el mundo de los almacenes de datos, contiene la misma información que un modelo de entidad-relación, pero la forma de empaquetar los datos en un formato simétrico tiene como objetivo garantizar una ejecución eficiente y rápida de las consultas, así como lograr una mayor comprensión del usuario. Dicho modelo separa sus datos en dos grandes tipos: las medidas, que generalmente son valores numéricos que se almacenan en las tablas de hechos, las cuales son las tablas primarias de este modelo; y las descripciones de los entornos, que son textuales y se almacenan en las tablas de dimensiones**"** (Verástegui, 2007).

Los datos contenidos en las tablas de hechos y dimensiones se pueden modelar de diferentes maneras. Por lo que se define como esquema, a la colección de tablas que existen en un Almacén de Datos. Los esquemas están divididos en varias categorías básicas: esquemas estrellas, copo de nieve y constelación de hechos.

El esquema en estrella es el más sencillo de los esquemas de almacenamiento de datos. Se llama así porque el diagrama se asemeja a una estrella, con los puntos que irradian desde un centro. El centro de la estrella consta de una o más tablas de hechos y los puntos de la estrella son las tablas de dimensiones. En concreto este esquema en estrella es ideal por su simplicidad y velocidad para ser usado en análisis multidimensionales como los Mercados de Datos, ya que permite acceder tanto a datos agregados como de detalle. Además, ofrece la posibilidad de implementar la funcionalidad de una base de datos multidimensional utilizando una clásica base de datos relacional.

En el esquema en estrella la tabla de hechos es la única tabla que tiene múltiples *joins* que la conectan con otras tablas. El resto de tablas del esquema (tablas de dimensión) únicamente hacen *join* con esta tabla de hechos. Las tablas de dimensión se encuentran además totalmente desnormalizadas, es decir, toda la información referente a una dimensión se almacena en la misma tabla.

El esquema en copo de nieve es un esquema de representación derivado del esquema en estrella, en el que las tablas de dimensión se normalizan en múltiples tablas. Por esta razón, la tabla de hechos deja de ser la única tabla del esquema que se relaciona con otras tablas, y aparecen nuevas *join* o uniones entre tablas gracias a que las dimensiones de análisis se representan ahora en tablas de dimensión normalizadas. En la estructura dimensional normalizada, la tabla que representa el nivel base de la dimensión es la que hace *join* directamente con la tabla de hechos. La diferencia entre ambos esquemas (estrella y copo de nieve) reside entonces en la estructura de las tablas de dimensión. Para conseguir un esquema en copo de nieve se ha de tomar un esquema en estrella y conservar la tabla de hechos, centrándose únicamente en el modelado de las tablas de dimensión, que si bien en el esquema en estrella se encontraban totalmente desnormalizadas, ahora se dividen en subtablas tras un proceso de normalización.
Es posible distinguir dos tipos de esquemas en copo de nieve, un *copo de nieve* completo (en el que todas las tablas de dimensión en el esquema en estrella aparecen normalizadas) o un copo de nieve parcial (sólo se lleva a cabo la normalización de algunas de ellas).

El esquema de constelación de hechos (fact constellation schema) se puede construir para cada esquema estrella o copo de nieve de un Almacén de Datos. Este esquema es más complejo que las otras arquitecturas debido al facto[2] de que contiene múltiples tablas de hechos. Con esta solución las tablas de dimensiones pueden estar compartidas entre más de una tabla de hechos. El esquema de constelación de hechos tiene mucha flexibilidad y esta es su gran virtud. Sin embargo, el problema es que cuando el número de las tablas vinculadas aumenta, la arquitectura puede llegar a ser muy compleja y difícil para mantener (Esquema de constelación de hechos, 2006).

El modelado del Mercado de Datos Ocupación se realiza con el esquema de estrella dado que se optimiza el rendimiento al mantener las consultas lo más simple posible y proporcionar servicios rápidos con un tiempo de respuesta lo más corto posible, al almacenar toda la información acerca de cada nivel en pocas tablas. Además, este Mercado de Datos cuenta solamente con el hecho hech_num_trab_ocupados y nueve dimensiones que se relacionan con él, por lo que el esquema estrella se ajusta perfectamente para ser utilizado.

[2] El hecho, en contraste con el dicho o con lo pensado.

1.5 Herramientas de modelado

Visual Paradigm es una herramienta CASE que utiliza Lenguaje Unificado de Modelado (UML por sus siglas en inglés Unified Modified Language) como lenguaje de modelado, con el uso del acercamiento orientado al objeto. Contiene varios productos o módulos que facilitan el trabajo durante la confección de un software, procurando garantizar la calidad en el producto final.

Se caracteriza por:

- ✓ Brinda facilidades para redactar Especificaciones de Casos de Uso del Sistema.
- ✓ Permite un diseño centrado en casos de uso y enfocado al negocio que genera un software de mayor calidad.
- ✓ Permite la sincronización entre Diagramas de Entidad Relación y Diagramas de Clases.
- ✓ Se integra a varias herramientas de Java como son: Eclipse, Oracle jdeveloper y Netbeans IDE.
- ✓ Proporciona soporte a varios lenguajes en generación de código e ingeniería inversa a través de plataformas Java.
- ✓ Es un poderoso generador de documentación pdf/html/ms Word.
- ✓ Es una herramienta CASE que soporta las últimas versiones del Lenguaje Unificado de Modelado y el modelado de procesos de negocio, desde un grupo administrador de objetos.
- ✓ En adicción al soporte UML, esta herramienta provee el modelado de procesos de negocio, además de un generador de mapeo de objetos relacionados para los lenguajes de programación Java, .net y php.

Se selecciona como herramienta CASE Visual Paradigm 6.4 por ser una herramienta UML profesional que soporta el ciclo de vida completo del desarrollo de software: análisis y diseño orientados a objetos, construcción, pruebas y despliegue; permitiendo la elaboración de aplicaciones con calidad. A pesar de ser una herramienta propietaria la universidad cuenta con la licencia. Además, es una herramienta multiplataforma y amigable en su entorno, lo que facilita su uso e interoperabilidad con otras aplicaciones y la definen como la mejor candidata.

1.6 Gestores de Bases de datos

Un sistema gestor de base de datos (SGBD) es la integración de un grupo de programas que administran y gestionan la información que se almacena en una base de datos. Es capaz de proporcionar una interfaz entre los datos, los programas que los manejan y los usuarios finales. A continuación se plantean los tres grupos de sistemas gestores de base de datos que existen con algunos de los sistemas gestores que los conforman. Se describen las principales características de

los gestores de datos PostgreSQL y Oracle por ser los más utilizados en el mundo para el desarrollo de los Almacenes de Datos.

SGBD libres: PostgreSQL, Firebird, SQLite y MySQL.
SGBD no libres: MySQL, Fox Pro, IBM Informix, IBM DB2: Universal Database (DB2 UDB), Interbase de CodeGear, Microsoft Access, Microsoft SQL Server, Oracle, Sybase ASA, Sybase IQ y Sybase ASE.
SGBD no libres y gratuitos: Microsoft SQL Server Compact Edition Basica, Sybase ASE Express Edition para Linux y Oracle Express Edition 10.

Oracle

Oracle Data base (Oracle DB) ha sido diseñado para que las organizaciones puedan controlar y gestionar grandes volúmenes de contenidos no estructurados en un único repositorio con el objetivo de reducir los costes y los riesgos asociados a la pérdida de información.
Oracle es un sistema gestor de datos relacionales de última generación, lo cual quiere decir que está orientado al acceso remoto y redes (internet). Hoy por hoy Oracle se puede implementar en diferentes plataformas: Familia de Microsoft, Unix, Linux, Vms, entre otros.
Las arquitecturas en las que se asienta pueden ser: Intel, Alpha, Sparc, Risc a nivel de procesadores. Oracle es perfectamente configurable en entornos "OLTP", paralelos, Cluster, e incluso resulta una genial solución a nivel de Datawarehouse.
Algunas de las ventajas que presenta Oracle son:

- ✓ Oracle es el motor de base de datos relacional más usado a nivel mundial.
- ✓ Puede ejecutarse en todas las plataformas, desde una PC hasta un supercomputador.
- ✓ Soporta todas las funciones que se esperan de un servidor serio: un lenguaje de diseño de bases de datos muy completo (PL/SQL) que permite implementar diseños activos, con triggers y procedimientos almacenados, con una integridad referencial declarativa bastante potente.
- ✓ Permite el uso de particiones para la mejora de la eficiencia, de replicación e incluso ciertas versiones admiten la administración de bases de datos distribuidas.
- ✓ El software del servidor puede ejecutarse en multitud de sistemas operativos. Existe incluso una versión personal para Windows 9x, lo cual es un punto a favor para los desarrolladores que se llevan trabajo a casa.
- ✓ Oracle es la base de datos con más orientación hacia internet.
- ✓ Un aceptable soporte

Entre las desventajas que presenta Oracle se encuentran:

- ✓ Han salido varias versiones con correcciones, hasta alcanzar la estabilidad en la 8.0.3. El motivo de tantos fallos fue, la remodelación del sistema de almacenamiento por causa de la introducción de extensiones orientadas a objetos.
- ✓ El mayor inconveniente de Oracle es quizás su precio, incluso las licencias de Personal Oracle son excesivamente caras.
- ✓ Otro problema es la necesidad de ajustes. Un error frecuente consiste en pensar que basta instalar el Oracle en un servidor y conectar directamente las aplicaciones clientes. Un Oracle mal configurado puede ser excesivamente lento.
- ✓ También es elevado el coste de la formación, y solo últimamente han comenzado a aparecer buenos libros sobre asuntos técnicos distintos de la simple instalación y administración (Daniel Muñoz, 2011).

PostgreSQL

El PostgreSQL es un poderoso sistema manejador de bases de datos, es decir, diseñado para administrar grandes cantidades de datos y es reconocido principalmente, por ser el gestor de base de datos de código abierto más avanzado del mundo.

PostgreSQL se ha preocupado por ser una solución real a los complejos problemas del mundo empresarial y a la vez mantener la eficiencia al consultar los datos. Con ese fin, se han desarrollado y añadido al PostgreSQL las más interesantes y útiles características que antes sólo podían hallarse en sistemas manejadores de bases de datos comerciales como Oracle, DB2 o Sybase; lo cual lo coloca, como su lema indica, como "El manejador de bases de datos de código abierto más avanzado del mundo". Debido a sus características PostgreSQL se ha ganado la admiración y el respeto de sus usuarios, así como el reconocimiento de la industria (ganador del Linux New Media Award for Best Database System y 3 veces ganador del The Linux Journal Editors' Choice Award for best DBMS) (Sistema Gestor de Base de Datos PostgreSQL, 2007).

Ventajas de PostgreSQL:

- ✓ 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 las ofertas de soporte, se tiene una importante comunidad de profesionales y entusiastas de PostgreSQL de los que su compañía puede brindar beneficios.

- ✓ 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.
- ✓ Estabilidad y Confiabilidad Legendarias: Es extremadamente común que compañías reporten que PostgreSQL nunca ha presentado caídas en varios años de operación de alta actividad. Ni una sola vez.
- ✓ 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.
- ✓ Multiplataforma: PostgreSQL está disponible en casi cualquier plataformas Unix (3.4 en la última versión estable), y ahora en versión nativa para Windows.
- ✓ Diseñado para ambientes de alto volumen: PostgreSQL usa una estrategia de almacenamiento de filas llamada MVCC para conseguir una mejor respuesta en ambientes de grandes volúmenes. Los principales proveedores de sistemas de bases de datos comerciales usan también esta tecnología, por las mismas razones.
- ✓ Herramientas gráficas de diseño y administración de bases de datos: Existen varias herramientas gráficas de alta calidad para administrar las bases de datos (pgAdmin, pgAccess) y para hacer diseños de bases de datos (Tora, Data Architect).
- ✓ Licencia: El código fuente de PostgreSQL está disponible bajo la licencia BSD. Esta licencia te da la libertad de usar, modificar y distribuir PostgreSQL de la forma que quieras, tanto de forma abierta como cerrada. Cualquier modificación, mejora o cambios que hagas son tuyas y puedes usarlas como quieras (Ventajas de PostgreSQL, 2003).

Se seleccionó PostgreSQL 8.4 como gestor de base de datos por brindar un servidor de base de datos altamente sofisticado, con alto rendimiento, estable y capacitado para lidiar con grandes volúmenes de datos. Además, por ser un producto de código abierto y sin costos de licencia, lo que ofrece un ahorro significativo de costos en activos. Esta versión incorpora la restauración de base de datos en forma paralela para acelerar la recuperación de las copias de seguridad (hasta 8 veces más rápido) y mejora el control en datos sensibles e incluso en la base de datos. A pesar de que PostgreSQL no admite el almacenamiento dimensional, soporta que el diseño de la base de datos sea dimensional. Esto permite que al utilizarlo conjuntamente con otras herramientas se pueda analizar la información de forma multidimensional.

1.7 Modos de Almacenamiento de Datos

Procesamiento Transaccional en Línea.

Los sistemas de Procesamiento Transaccional en Línea (OLTP por sus siglas en inglés On-Line Transactional Processing) son bases de datos orientadas al procesamiento de transacciones. Una transacción genera un proceso atómico (que debe ser validado o invalidado), y que puede involucrar operaciones de inserción, modificación y borrado de datos. El proceso transaccional es típico de las bases de datos operacionales.

Procesamiento Analítico en Línea.

Los sistemas de Procesamiento Analítico en Línea (OLAP por sus siglas en inglés On-Line Analytical Processing) son bases de datos orientadas al procesamiento analítico. Este análisis suele implicar, generalmente, la lectura de grandes cantidades de datos para llegar a extraer algún tipo de información útil: tendencias de ventas, patrones de comportamiento de los consumidores, elaboración de informes complejos, entre otros. Este sistema es típico de los mercados de datos.

- ✓ El acceso a los datos suele ser de solo lectura. La acción más común es la consulta, con muy pocas inserciones, actualizaciones o eliminaciones.
- ✓ Los datos se estructuran según las áreas de negocio y los formatos de los datos están integrados de manera uniforme en toda la organización.
- ✓ El historial de datos es a largo plazo, normalmente, de dos a cinco años.
- ✓ Las bases de datos OLAP se suelen alimentar de información procedente de los sistemas operacionales existentes, mediante un proceso de Extracción, Transformación y Carga (ETL) (Bases de datos OLTP Y OLAP, 2007).

Procesamiento Analítico Relacional en Línea.

Los sistemas de Procesamiento Analítico Relacional en línea (ROLAP por sus siglas en inglés Relational On-Line Analytical Processing), accede a los datos almacenados en un almacén para proporcionar los análisis OLAP. La premisa de los sistemas ROLAP es que las capacidades OLAP se soportan mejor contra las bases de datos relacionales.

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

El motor ROLAP se integra con niveles de presentación, a través de los cuales los usuarios realizan los análisis OLAP. Después de que el modelo de datos para el almacén se ha definido, los datos se cargan desde el sistema operacional. Se ejecutan rutinas de bases de datos para agregar el dato, si así es requerido por el modelo de datos entonces se crean los índices para optimizar los tiempos de acceso a las consultas.

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

La arquitectura ROLAP es capaz de usar datos pre-calculados si estos están disponibles, o de generar dinámicamente los resultados desde los datos elementales si es preciso. Esta arquitectura accede directamente a los datos del almacén y soporta técnicas de optimización de accesos para acelerar las consultas.

Procesamiento Analítico Multidimensional en Línea.

La arquitectura Procesamiento Analítico Multidimensional en Línea (MOLAP por sus siglas en inglés Multidimensional On-Line Analytical Processing) utiliza bases de datos multidimensionales para proporcionar el análisis, su principal premisa es que el OLAP está mejor implantado almacenando los datos multidimensionalmente. Por el contrario, la arquitectura ROLAP cree que las capacidades OLAP están perfectamente implantadas sobre bases de datos relacionales. Un sistema MOLAP usa una base de datos propietaria multidimensional, en la que la información se almacena multidimensionalmente, para ser visualizada en varias dimensiones de análisis.

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

La arquitectura MOLAP requiere unos cálculos intensivos de compilación. Lee de datos precompilados y tiene capacidades limitadas de crear agregaciones dinámicamente o de hallar ratios que no se hayan precalculados y almacenados previamente.

Procesamiento Analítico Híbrido en línea.

Un desarrollo un poco más reciente ha sido la solución OLAP híbrida (HOLAP por sus siglas en inglés Hybrid On-Line Analytical Processing), la cual combina las arquitecturas ROLAP y MOLAP para brindar una solución con las mejores características de ambas: desempeño superior y gran escalabilidad. Un tipo de ROLAP mantiene los registros de detalles (los volúmenes más grandes) en la base de datos relacional, mientras que mantiene las agregaciones en un almacén MOLAP separado (MOLAP, ROLAP, HOLAP, 2007).

ROLAP VS MOLAP.

ROLAP es una alternativa a MOLAP tecnología. Mientras que las herramientas analíticas de ROLAP y de MOLAP se diseñan para permitir el análisis de datos con el uso de un modelo multidimensional de los datos, ROLAP diferencia perceptiblemente en que no requiere el pre-cómputo y el almacenaje de la información. En lugar, ROLAP filetea el acceso a los datos en la base de datos emparentada y genera preguntas SQL para calcular la información en el nivel apropiado cuando un usuario final la solicita. Con ROLAP, es posible crear las tablas adicionales de la base de datos (tablas sumarias o agregaciones) las cuales resumen los datos en cualquier combinación deseada en dimensiones.

Ventajas de ROLAP

- ✓ ROLAP se considera ser más escalable en la manipulación de los grandes volúmenes de los datos.
- ✓ Con una variedad de herramientas del cargamento de los datos disponibles y la capacidad para la consonancia ETL final al código del modelo particular de los datos, se tiene que los tiempos de carga son generalmente más cortos que con el automatizado de cargas MOLAP.
- ✓ Los datos se almacenan en un estándar de bases de datos emparentadas y pueden ser alcanzados por cualquier herramienta SQL (la herramienta no tiene que ser OLAP).
- ✓ Por no estar parejo el almacenaje de datos del modelo multidimensional, es posible modelar con éxito los datos que no entrarían de otra manera en un modelo dimensional terminante (Rolap, 2011).

La selección de uno u otro modelo depende de cuán importante sea el rendimiento de las consultas para los usuarios y de la tecnología disponible a utilizar. En el modelo ROLAP la respuesta a las consultas y el tiempo de procesamiento suelen ser más lentos que con los modos de almacenamiento MOLAP o HOLAP. No obstante, ROLAP permite a los usuarios ver los datos en tiempo real y ahorrar espacio de almacenamiento al trabajar con conjuntos de datos grandes a los que no se suele consultar con frecuencia, como datos puramente históricos.

Se seleccionó ROLAP como modelo a utilizar dado que capacidades OLAP se soportan mejor contra las bases de datos relacionales. Mantiene los registros de detalles. Permite a los usuarios ver la información en tiempo real. Ahorrar espacio de almacenamiento al trabajar con conjuntos de datos grandes a los que no se suele consultar con frecuencia. Es capaz de usar datos pre-calculados y de generar dinámicamente los resultados desde los datos elementales.

1.8 Herramientas para el proceso Extracción, Transformación y Carga

Las herramientas para el proceso de Extracción, Transformación y Carga (ETL por sus siglas en inglés Extraction, Transformation, and Loading) seleccionadas para la construcción del Mercado de Datos fueron:

El Pentaho Data Integration por ser multiplataforma, fácil de instalar y de configurar. Además, de permitir cargar archivos en formato Excel, dado que la información se encuentra guardada en dicho formato. Ofrece soporte para metadatos e incluye operaciones de transformación, así como funciones que posibilitan operar con los campos en el flujo de datos, renombrando, calculando campos en función de otros, correlacionando valores y realizando búsquedas auxiliares en bases de datos.

El Data Cleaner por ser una herramienta gratuita y de código abierto, que permite la evaluación de la calidad que poseen los datos contenidos. Preparando a los mismos para las transformaciones que se les realizará en el proceso ETL. Es una aplicación muy fácil de usar, genera sofisticados informes y gráficos que permiten a los usuarios determinar el nivel de calidad de los datos, identificar y analizar la estructura de los datos y combinar los resultados, creando vistas fáciles de interpretar para evaluar la calidad de los datos.

A continuación se mencionan otras características que poseen las herramientas escogidas:

1.8.1 Pentaho Data Integration

Muchas organizaciones tienen información disponible en aplicaciones y base de datos separadas. Pentaho Data Integration, limpia e integra esta valiosa información y la pone en manos del usuario. Provee una consistencia, una sola versión de todos los recursos de información, que es uno de los más grandes desafíos para las organizaciones de las tecnologías de la información hoy en día. Pentaho Data Integration también conocido como Kettle permite una poderosa ETL. El uso del Kettle permite evitar grandes cargas de trabajo manual frecuentemente difícil de mantener y de desplegar. La arquitectura del Pentaho Data Integration viene representada por el siguiente esquema:

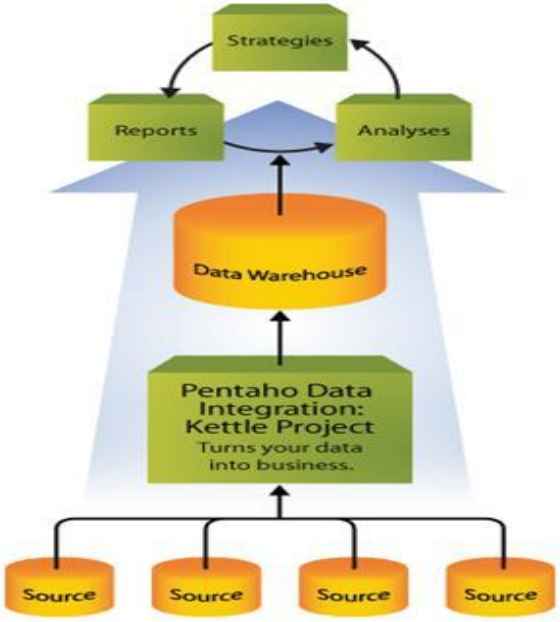


Figura 1. Arquitectura del Pentaho Data Integration.

Propiedades básicas:

A parte de ser de código abierto y sin costes de licencia, las características básicas de esta herramienta son:

- ✓ 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) (Pentaho: Características, 2011).

Con Kettle (Pentaho Data Integration) se pueden realizar diversas tareas, entre ellas cabe resaltar:

- ✓ Soporte para cambiar, enlazar dimensiones y otras operaciones en el Almacén de Datos.
- ✓ Exportar de bases de datos a ficheros u otras bases de datos.
- ✓ Importar en bases de datos ficheros en formato Excel o texto.
- ✓ Migración de datos entre diferentes bases de datos.
- ✓ Explotación de los datos existentes en bases de datos (tablas, vistas).
- ✓ Enriquecer la información mediante la búsqueda de datos en diferentes almacenes de información (bases de datos, ficheros de texto, hojas Excel).

- ✓ Limpieza de datos aplicando transformaciones de datos con condiciones complejas.
- ✓ Integración de aplicaciones.
- ✓ Transformación: (sirve para mover, copiar, transformar datos, filas entre una fuente y un destino).
- ✓ Trabajo: Coordinación de Transformaciones, secuencialidad y paralelismo (Control de flujo, ejecutar transformaciones, enviar correos en caso de error).

Se compone de 4 herramientas:

- ✓ 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 (Curto, ETL: Kettle: Pentaho Data Integration, 2009).

1.8.2 DataCleaner

DataCleaner es una aplicación de código abierto para el perfil, la validación y comparación de datos. Estas actividades ayudan a administrar y supervisar la calidad de los datos con el fin de garantizar que la información sea útil y aplicable a su situación de negocio. Esta es la alternativa gratuita al software de gestión de datos maestros, metodologías, almacenamiento de datos, proyectos de investigación estadística y la preparación para los procesos de extracción, transformación y carga.

Los procesos de ETL son los encargados de copiar y transformar datos de una o más fuentes de información a otra. Normalmente, se utiliza una herramienta como DataCleaner antes, durante y después de cualquier actividad de ETL.

DataCleaner:

- ✓ Está licenciado bajo los términos de la Licencia Pública General Menor (LGPL).
- ✓ Es posiblemente el más fácil para aplicar a los datos de calidad disponibles.
- ✓ Brinda la opción de personalizar los datos de una forma sencilla.
- ✓ Es un software de código abierto.
- ✓ Puede acceder y analizar prácticamente cualquier Almacén de Datos, incluyendo: Archivos XML, Hojas de cálculo Excel (. Xls) y bases de datos como Oracle, Microsoft SQL Server, PostgreSQL, MySQL, OpenOffice (ODB). Archivos separados por comas y separados por tabuladores (.csv / .tsv).

1.9 Herramientas para el proceso de Inteligencia de Negocio

Las herramientas para el proceso de Inteligencia de Negocio (BI por sus siglas en inglés Business Intelligence) seleccionadas para el desarrollo de la solución fueron:

El Pentaho Schema Worbench para el desarrollo, prueba y publicación de los cubos OLAP. El Mondrian OLAP Server para realizar las consultas al almacén de datos, presentando los resultados mediante un navegador. El Pentaho BI Server que suministra soporte e infraestructura para crear soluciones de inteligencia de negocio e incorpora un motor de solución que integra reportes, análisis y tableros de comandos. El mismo funciona como un sistema basado en administración web de informes. Por lo que se selecciona el Apache Tomcat 6.0 que posibilita la gestión de solicitudes y respuestas mediante el conector http en la solución. Permitiendo al usuario realizar las actividades mediante la navegación web.

A continuación se mencionan otras características de las herramientas seleccionadas:

1.9.1 Pentaho Schema Workbench

Pentaho proporciona en su plataforma BI una solución ROLAP a través de lo que llaman Pentaho Analysis Services. (PAS) está basado en Mondrian, que es el corazón de este y en Jpivot, que es la herramienta de análisis de usuario, con él se realiza la navegación dimensional sobre los cubos desde la plataforma BI y se visualizan los resultados de las consultas. Estas son ejecutadas por Mondrian, que traduce los resultados relacionales a resultados dimensionales, que a su vez son mostrados al usuario en formato Html por Jpivot.

El elemento principal del sistema son los ficheros xml donde se representan los esquemas dimensionales. Para construir estos ficheros xml, se podría utilizar cualquier editor de texto o xml, o bien la herramienta que ofrece Pentaho, que se llama Schema Workbench.

Pentaho Schema Workbench es la herramienta gráfica que permite la construcción de los esquemas de Mondrian y además permite publicarlos al servidor BI para que puedan ser utilizados en los análisis por los usuarios de la plataforma (Business Intelligence, OpenSource, Pentaho, 2010).

Mondrian Schema Workbench es un entorno visual para el desarrollo y prueba de cubos OLAP Mondrian. Con esta aplicación, se puede configurar una conexión JDBC como el modelo físico, para luego elaborar el esquema lógico de manera simple y efectiva. Para ello el entorno ofrece un editor de

esquemas con la fuente de datos subyacente para su validación. Permite la ejecución de consultas MDX contra el esquema y la navegación por la base de datos.

Algunas de las características de Workbench son:

- ✓ Libre, distribuida bajo la licencia GPL.
- ✓ Multiplataforma. disponible para Windows, GNU/Linux. Mac.
- ✓ Permite crear diagramas Entidad-Relación.
- ✓ Brinda una interfaz gráfica para el usuario que hace muy intuitivo, cómodo de usar y rápido de trabajar, lo que permite ahorrar tiempo (Epsilon, 2009).

1.9.2 Pentaho BI Server

Con esta herramienta se suministra soporte e infraestructura para crear soluciones de Inteligencia de Negocio. Proporciona servicios básicos además de incluir autenticación, registro, auditoría, servicios web y motor de reglas. Incorpora un motor de solución que integra reportes, análisis, tableros de comandos y componentes de minería de datos. Funciona como un sistema basado en administración web de informes, el servidor de integración de aplicaciones y un motor de flujo de trabajo ligero (secuencias de acción.). Además, está diseñada para integrarse fácilmente en cualquier proceso de negocio. Permite que puedan ejecutarse los informes y aplicaciones, se puede usar como base para construir un sistema propio de BI.

Tres de sus principales ventajas son:

- ✓ Administra y programa reportes.
- ✓ Administra seguridad de usuarios.
- ✓ Brinda la posibilidad de guardar la consulta que se ejecute.

1.9.3 Mondrian OLAP Server

OLAP es el acrónimo en inglés de Procesamiento Analítico en Línea (On-Line Analytical Processing). Es una solución utilizada en el campo de la llamada Inteligencia de Negocio cuyo objetivo es agilizar la consulta de grandes cantidades de datos. Para ello utiliza estructuras multidimensionales (o Cubos OLAP) que contienen datos resumidos de grandes bases de datos o Sistemas Transaccionales (OLTP). Se usa en informes de negocios de ventas, marketing, informes de dirección, minería de datos y áreas similares.

La razón de usar OLAP para las consultas es la velocidad de respuesta. Una base de datos relacional almacena entidades en tablas discretas si han sido normalizadas. Para obtener la funcionalidad de

OLAP se utilizan otras dos aplicaciones: el servidor OLAP Mondrian, que combinado con Jpivot, permiten realizar consultas a los Mercados de Datos, que los resultados sean presentados mediante un browser y que el usuario pueda realizar drill down y el resto de las navegaciones típicas.
Mondrian, ahora rebautizado como Pentaho Analysis Services, es el motor OLAP integrado en la suite de Business Intelligence de Pentaho. Mondrian se encarga de recibir consultas dimensionales (lenguaje MDX) y devolver los datos de un cubo, solo que este cubo no es algo físico sino un conjunto de metadatos que definen como se han de "mapear" estas consultas, que tratan conceptos dimensionales a sentencias SQL, ya tratando con conceptos relacionales que obtengan de la base de datos la información necesaria para satisfacer la consulta dimensional.

Algunas de las ventajas de este modelo son:

- ✓ El no tener que generar cubos estáticos ahorrando que cuesta generarlos y la memoria que ocupan.
- ✓ La posibilidad de utilizar siempre los datos residentes en la base de datos, de forma que se trabaja con datos actualizados. Muy útil en entorno de BI operacional.

Pentaho Análisis suministra a los usuarios un sistema avanzado de análisis de información. Con uso de las tablas dinámicas (pivot tables, crosstabs), generadas por Mondrian y JPivot, el usuario puede navegar por los datos, ajustando la visión de los datos, los filtros de visualización, añadiendo o quitando los campos de agregación. Los datos pueden ser representados en una forma de SVG o Flash, los dashboards widgets, o también integrados con los sistemas de minería de datos y los portales web (portlets). Además, con el Microsoft Excel Analysis Services, se puede analizar los datos dinámicos en Microsoft Excel (usando la conexión a OLAP server Mondrian) (Business Intelligence, OpenSource, Pentaho, 2010).

1.9.4 Apache Tomcat

A través del navegador instalado en su ordenador los usuarios solicitan y reciben los contenidos utilizando el protocolo HTTP. Por este motivo, es necesario que el sistema incorpore un servidor Web. Para relacionar el protocolo HTTP y el sistema de agentes este servidor tiene que poder ejecutar código Java.
Entre los múltiples servidores que cumplen estos requisitos el más conocido es Apache Tomcat. Entre sus características se puede mencionar que es gratuito, de código abierto, que está bien documentado

y que en él es fácil implementar aplicaciones Web, ponerlas en funcionamiento y publicarlas (Hall, 2003).

Conclusiones del capítulo.

En este capítulo se realizó un estudio sobre los conceptos, características, ventajas y desventajas de los Almacenes de Datos y los Mercados de Datos; se investigó sobre los tipos de tecnologías, modelos, esquemas, metodología y herramientas existentes para el desarrollo de los Mercados de Datos, con el objetivo de desarrollar el Mercado de Datos para los indicadores de la Ocupación. Se propone utilizar una adaptación de la metodología planteada por Ralph Kimball y lo planteado en la tesis de doctorado de Leopoldo Zenaido Zepeda Sánchez para el desarrollo del Mercado de Datos Ocupación. La herramienta de modelado a utilizar es el Visual Paradigm en su versión 6.4. Como gestor de bases de datos PostgreSQL en su versión 8.4. Se seleccionaron para el desarrollo de ETL, el DataCleaner en su versión 1.5.3 para la estandarización de los datos y el Pentaho Data Integration en su versión 4.0.1 para la implementación de los procesos ETL. Se seleccionaron para la implementación del BI el Schema Workbench en su versión 3.2.0 para el desarrollo de los cubos OLAP, el Pentaho BI Server para administrar los reportes y usuarios, el Servidor Mondrian en su versión 3.2.0 como servidor OLAP y el Apache Tomcat en su versión 5.5 como servidor web.

Capítulo 2: Análisis y diseño

En este capítulo se realiza una descripción del negocio para un mejor entendimiento de los indicadores de Ocupación, se refinan las necesidades de los usuarios donde se encuentran los requisitos de información, requisitos funcionales, requisitos no funcionales y las reglas del negocio. Se diseñan el diagrama de Casos de Uso del Sistema, la Matriz Bus y el Modelo de Datos partiendo del refinamiento de las tablas de hechos y dimensiones que los componen. Por último, se diseñan los procesos de integración y los subsistemas de visualización.

2.1 Análisis del Mercado de Datos Ocupación

2.1.1 Estudio preliminar del negocio

La Oficina Nacional de Estadística (ONE) es la entidad creada para proponer, organizar y ejecutar, según corresponda, la aplicación de la política estatal en materia de estadística del país. Esta se encarga de recolectar, organizar, resumir, presentar y analizar la información relativa a un conjunto de indicadores, entre los que se encuentra Ocupación.

Ocupación es el área que gestiona toda la información correspondiente al empleo y los salarios, este incluye datos sobre la población económicamente activa (ocupados y desocupados), fuerza de trabajo ocupada según la situación del empleo, su estructura por clase de actividad económica, sexo, edades, nivel educacional y categoría ocupacional. Las tasas de desocupación se muestran por sexo, se incluyen datos sobre el salario medio por territorios y clase de actividad económica, así como indicadores sobre accidentes del trabajo, seguridad social y asistencia social.

Esta información permite conocer la ocupación total y por sexo del país (Sector Estatal Civil y no Estatal), por territorio puro, así como su desglose por tipos de contratos de trabajo, mayores y menores de la edad laboral que trabajan, categoría ocupacional y sexo.

2.1.2 Temas de análisis

Para un correcto desarrollo del Mercado de Datos es preciso identificar los temas de análisis, con los que se adquieren diferentes perspectivas del análisis. Estos muestran el progreso de las tareas planteadas y garantizarán el éxito del diseño de la estructura que se desarrolla. La presente investigación se centra en dar cumplimiento a los requisitos identificados en el tema de análisis del Empleo del indicador Ocupación.

2.1.3 Necesidades de los usuarios

Es importante conocer las necesidades del cliente, dado que su implicación garantiza un buen análisis y diseño del proceso de negocio. La interacción con el usuario durante el ciclo de vida del producto permitirá que exprese su opinión con respecto a los resultados del producto, si son satisfactorios o insatisfactorios según sus necesidades.

Durante el análisis de Ocupación, los especialistas de la ONE se enfocan en la información contenida en el área de empleo, dado que es donde se almacenan los datos que serán estudiados. A continuación se especifican los requisitos de información, funcionales, no funcionales y las reglas del negocio para dar solución a las necesidades de los usuarios.

Requisitos de Información.

Los requisitos de información son las funcionalidades más importantes que el sistema debe incluir y mantenerlas disponibles cuando se realiza el análisis sobre los datos. Representan una entrada de datos fundamental para el proceso de Inteligencia de Negocio y los reportes futuros. Surgen a partir de la comparación entre las necesidades de la información y las reglas del negocio. A continuación se exponen los requisitos identificados:

RI 1. Obtener la cantidad total de trabajadores.

RI 2. Obtener la cantidad de trabajadores del sexo femenino.

RI 3. Obtener la cantidad de trabajadores del sexo masculino.

RI 4. Obtener la cantidad total de trabajadores contratados por tiempo determinado.

RI 5. Obtener la cantidad de trabajadores contratados por tiempo determinado del sexo femenino.

RI 6. Obtener la cantidad de trabajadores contratados por tiempo determinado del sexo masculino.

RI 7. Obtener la cantidad total de trabajadores mayores de la edad laboral.

RI 8. Obtener la cantidad de trabajadores mayores de la edad laboral del sexo femenino.

RI 9. Obtener la cantidad de trabajadores mayores de la edad laboral del sexo masculino.

RI 10. Obtener la cantidad total de trabajadores menores de la edad laboral.

RI 11. Obtener la cantidad de trabajadores menores de la edad laboral del sexo femenino.

RI 12. Obtener la cantidad de trabajadores menores de la edad laboral del sexo masculino.

RI 13. Obtener la cantidad total de trabajadores operarios.

RI 14. Obtener la cantidad de trabajadores operarios del sexo femenino.

RI 15. Obtener la cantidad de trabajadores operarios del sexo masculino.
RI 16. Obtener la cantidad total de trabajadores técnicos.
RI 17. Obtener la cantidad de trabajadores técnicos del sexo femenino.
RI 18. Obtener la cantidad de trabajadores técnicos del sexo masculino.
RI 19. Obtener la cantidad total de trabajadores administrativos.
RI 20. Obtener la cantidad de trabajadores administrativos del sexo femenino.
RI 21. Obtener la cantidad de trabajadores administrativos del sexo masculino.
RI 22. Obtener la cantidad total de trabajadores de servicios.
RI 23. Obtener la cantidad de trabajadores de servicios del sexo femenino.
RI 24. Obtener la cantidad de trabajadores de servicios del sexo masculino.
RI 25. Obtener la cantidad total de trabajadores dirigentes.
RI 26. Obtener la cantidad de trabajadores dirigentes del sexo femenino.
RI 27. Obtener la cantidad de trabajadores dirigentes del sexo masculino.
RI 28. Población económicamente activa.
RI 29. Ocupados por clase actividad económica.
RI 30. Distribución por edades de los trabajadores por categoría ocupacional.
RI 31. Distribución de la fuerza de trabajo por categoría ocupacional y sexos.

Requisitos funcionales.

Los requisitos funcionales son aquellos que el sistema va a incluir. Las funcionalidades deben estar dirigidas a las necesidades del cliente. Los requisitos que fueron identificados para el desarrollo del Mercado de Datos Ocupación quedan expuestos a continuación:

RF 1. Extraer datos ocupación.
RF 2. Realizar transformaciones y carga de los datos ocupación.
RF 3. Autenticar usuario.
RF 4. Insertar roles y permisos.
RF 5. Eliminar roles y permisos.
RF 6. Insertar usuario.
RF 7. Eliminar usuario.
RF 8. Insertar reporte OLAP.
RF 9. Eliminar reporte OLAP.
RF 10. Modificar reporte OLAP.
RF 11. Abrir navegador OLAP.

RF 12. Mostrar editor MDX.
RF 13. Configurar tabla OLAP.
RF 14. Mostrar padres.
RF 15. Ocultar repeticiones.
RF 16. Mostrar propiedades.
RF 17. Suprimir filas/columnas vacías.
RF 18. Intercambiar ejes.
RF 19. Detallar miembros.
RF 20. Abrir detalle.
RF 21. Entrar en detalle.
RF 22. Mostrar datos origen.
RF 23. Mostrar gráfico.
RF 24. Configurar gráficos.
RF 25. Configurar impresión.
RF 26. Exportar a PDF.
RF 27. Exportar a Excel.

Requisitos no funcionales.

Los requisitos no funcionales son características y condiciones que el sistema debe cumplir. Describen las propiedades no funcionales que los clientes desean que el producto posea, lo que permite desarrollar un producto usable, fiable y eficiente.

Usabilidad: Esfuerzo necesario para aprender, operar, preparar entradas e interpretar la salida de un programa.

Requisitos de Usabilidad

RNF 1. Cumplir con los patrones del diseño de la aplicación.

La aplicación debe tener una interfaz gráfica uniforme, donde se incluyan pantallas, menús y opciones.

RNF 2. La interfaz de la aplicación debe mostrar los mensajes, títulos y textos 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 toda la aplicación.

RNF 3. Diseñar un reporte del Mercado de Datos de manera sencilla y ágil.

Un usuario podrá diseñar un reporte del Mercado de Datos de manera rápida y sencilla sin necesidad de poseer un gran conocimiento de la aplicación.

RNF 4. Navegar en los reportes del Almacén de Datos de manera ágil.

El usuario podrá realizar las agrupaciones y cruces de variables en los reportes de forma fácil y dinámica en la misma área de trabajo. Esto permite agilizar la navegabilidad de los usuarios en un reporte.

Confiabilidad: Grado en el que un programa se espera que realice su función con una precisión requerida.

Requisitos de Fiabilidad

RNF 5. Garantizar la persistencia de la información.

Para garantizar la persistencia de la información se realizará un respaldo total de los datos del mercado, con una frecuencia anual. Toda esta información se almacenará en el área de la dirección de informática en un banco de datos especial. 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.

RNF 6. Garantizar el cumplimiento de actualización de los datos en el almacén.

La información contenida en el mercado tendrá una precisión y exactitud anual, en correspondencia con la periodicidad con que se recogen los datos.

Eficiencia: Cantidad de recursos y código requeridos por un programa para realizar una función.

Requisitos de Eficiencia

RNF 7. Tiempo promedio de respuesta para la obtención de un reporte

El tiempo promedio para la obtención de un reporte es de aproximadamente 5 segundos.

RNF 8. Cantidad de usuarios conectados de forma simultánea

El sistema debe permitir que existan al menos 1000 usuarios conectados de forma simultánea.

RNF 9. Cumplir con los recursos de hardware para un funcionamiento óptimo.

Restricciones de Hardware

RNF 10. Para los procesos de integración de datos la utilización de los recursos como mínimo es de:

Memoria RAM: 1 GB.

Disco duro: 40 GB

Comunicaciones: 5MB/seg.

Para el repositorio de datos la utilización de recursos como mínimo es de:

Memoria RAM: 1 GB.

Disco duro: 40 GB libres es disco

Comunicaciones: 10MB/seg.

Restricciones de diseño

Lenguaje de Programación

RNF 11. El lenguaje para la programación del proceso de integración será SQL para realizar consultas a la base de datos y Java script para implementar algunas reglas de transformaciones

Base de datos

RNF 12. El Sistema Gestor de Base de datos (SGBD) para implementar el Almacén de Datos es el PostgreSql en su versión 8.4. Como Interfaz de Administración del SGBD se usará el PgAdmin en su versión 1.10.

Modelado de los datos

RNF 13. Se empleará para el modelado de los datos Visual Paradigm 6.4

Integración de los datos

RNF 14. Para la integración de los datos se utilizará el Pentaho Data Integration 4.1.0

Capa de visualización

RNF 15. Para el desarrollo de la capa de visualización se utilizarán las herramientas Workbench en su versión 3.2.1 y el Pentaho BI Server en su versión 3.6.

Soporte

RNF 16. Definir el soporte o mantenimiento del sistema

El sistema tendrá acceso a la Plataforma de Soporte Online de ALBET.

RNF 17. Las estructuras del Almacén de Datos se nombrarán de una manera estándar teniendo en cuenta el tipo de estructura que se maneje.

Se definen convenciones de nombrado con el objetivo de manejar un vocabulario común en el Almacén de Datos que permita un entendimiento claro y conciso de las estructuras por parte de los desarrolladores que interactúen con el Almacén de Datos.

RNF 18. Establecer tiempo de entrenamiento requerido para que usuarios normales sean productivos operando el sistema.

El tiempo de entrenamiento requerido para que usuarios normales sean productivos operando el sistema deberá ser entre 7 y 14 días. Para aquellos usuarios con un nivel avanzado se define como

tiempo máximo 7 días. Para lograr el cumplimiento de los tiempos establecidos por parte de los usuarios es necesario un dominio del funcionamiento del negocio en correspondencia con el rol que ocupen.

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

RNF 19. Se dispone de un manual de usuario.

Se elaborará una guía de ayuda para la navegación y componentes del sistema, adjuntada al expediente del proyecto.

2.1.4 Reglas del Negocio

Las reglas del negocio son las encargadas de estipular las políticas, normas, procedimientos, definiciones, condiciones y restricciones que la entidad presenta. De estas depende el cumplimiento del objetivo de la organización.

Durante el procedimiento de almacenamiento de la información de los indicadores de Ocupación, se guardan los datos referentes al empleo, estos toman forma de clasificadores para un mejor control de las medidas y parámetros. Los clasificadores tienen una definición propia, que direcciona hacia las reglas del negocio.

Para el desarrollo del Mercado de Datos Ocupación se definieron las siguientes reglas:

RN 1. Todos los indicadores del Mercado de Datos deben poseer un valor mayor o igual a cero (0).

RN 2. Los indicadores poseen periodicidad anual.

RN 3. El proceso ETL verificará que los campos que representan los diferentes indicadores del Mercado de Datos Ocupación no sean nulos.

RN 4. El valor de los indicadores relacionados a la dim_sector_ocupacion debe estar en el rango del 01 al 05.

RN 5. La variante 07 (de trabajadores por cuenta propia) de la dim_sector_ocupacion se trabaja en el CAE, las otras variantes se trabajan en el DPA.

RN 6. El DPA tendrá un código de cuatro dígitos que identifica al municipio. Si dim_sector_ocupacion es el 07 (de trabajadores por cuenta propia) se reflejará el código de cuatro dígitos que identifica el sector y la rama donde clasifica cada actividad.

RN 7. La columna 01 es igual a la suma de las columnas 09, 11, 13, 15 y 17. Es también mayor o a lo sumo igual a la columna 02, para todas las filas del modelo.

RN 8. La columna 02 es un desglose de la columna 01, por lo que tiene que ser menor o a lo sumo igual a la misma, para todas las filas del modelo. Es también igual a la suma de las columnas 10, 12, 14, 16 y 18, para todas las filas del modelo.

RN 9. La columna 03 es un desglose de la columna 01, por lo cual debe ser menor o a lo sumo igual a la misma.

RN 10. La columna 04 es un desglose de la columna 02 y 03 separadamente, por lo cual debe ser menor o a lo sumo igual a cada una de ellas.

RN 11. La columna 05 es un desglose de la columna 01, por lo cual debe ser menor o a lo sumo igual a la misma.

RN 12. La columna 06 es un desglose de la columna 02 y 05 separadamente, por lo cual debe ser menor o a lo sumo igual a cada una de ellas.

RN 13. La columna 07 es un desglose de la columna 01, por lo cual debe ser menor o a lo sumo igual a la misma.

RN 14. La columna 08 es un desglose de la columna 02 y 07 separadamente, por lo cual debe ser menor o a lo sumo igual a cada una de ellas.

RN 15. La columna 09 es un desglose de la columna 01, por lo cual debe ser menor o a lo sumo igual a la misma.

RN 16. La columna 10 es un desglose de la columna 02 y 09 separadamente, por lo cual debe ser menor o a lo sumo igual a cada una de ellas.

RN 17. La columna 11 es un desglose de la columna 01, por lo cual debe ser menor o a lo sumo igual a la misma.

RN 18. La columna 12 es un desglose de la columna 02 y 11 separadamente, por lo cual debe ser menor o a lo sumo igual a cada una de ellas.

RN 19. La columna 13 es un desglose de la columna 01, por lo cual debe ser menor o a lo sumo igual a la misma.

RN 20. La columna 14 es un desglose de la columna 02 y 13 separadamente, por lo cual debe ser menor o a lo sumo igual a cada una de ellas.

RN 21. La columna 15 es un desglose de la columna 01, por lo cual debe ser menor o a lo sumo igual a la misma.

RN 22. La columna 16 es un desglose de la columna 02 y 15 separadamente, por lo cual debe ser menor o a lo sumo igual a cada una de ellas.

RN 23. La columna 17 es un desglose de la columna 01, por lo cual debe ser menor o a lo sumo igual a la misma.

RN 24. La columna 18 es un desglose de la columna 02 y 17 separadamente, por lo cual debe ser menor o a lo sumo igual a cada una de ellas.

Para un mejor entendimiento de las reglas del negocio remitirse al **Anexo 1** donde se podrá analizar el modelo M5200 de los indicadores de Ocupación.

2.1.5 Casos de Uso del Sistema

Los casos de uso del sistema son los encargados de representar la información de forma visual, representando los requisitos funcionales. Las relaciones del usuario con los casos de uso del sistema son representadas en un diagrama de casos de uso.

Para el desarrollo del Mercado de Datos se identificaron los casos de uso de información y funcionales, de los cuales se plantean sus descripciones textuales (para la especificación detallada revisar: Modelo de Casos de Uso en el expediente de proyecto).

Los casos de uso de información se dividen según los temas de análisis contenidos en el indicador Ocupación. En la tesis precedente se especifican 44 requisitos de información abordados en las áreas de empleo y salario, durante el refinamiento se concluyó que solo 31 referentes al área de empleo serían implementados, ya que los restantes no se relacionaban con las necesidades planteadas por el cliente en el presente trabajo. Estos requisitos se agrupan en el caso de uso Analizar indicadores de empleo.

Los casos de uso funcionales se basan en las diferentes operaciones que se realicen en la aplicación. Entre dichas operaciones se encuentran los procesos de integración a la fuente de Ocupación. También se encuentra presente la administración de los usuarios, roles, permisos y reportes de la aplicación, se autenticarán los usuarios y se visualizarán las modificaciones que el usuario realice sobre las consultas mostradas por el sistema.

Los casos de uso funcionales se definen a continuación:

- ✓ Administrar roles y permisos.
- ✓ Administrar usuarios.
- ✓ Administrar reportes OLAP.
- ✓ Autenticar usuario.
- ✓ Visualizar cambios en el reporte.
- ✓ Extraer datos ocupación.
- ✓ Realizar transformaciones y carga de los datos ocupación.

El diagrama de casos de uso que se muestra a continuación representa la relación entre los actores y los casos de uso que se identificaron en el área de Ocupación. Para un mejor entendimiento de este, remitirse a la especificación del Caso de uso Analizar indicadores de empleo en el **Anexo 2**.

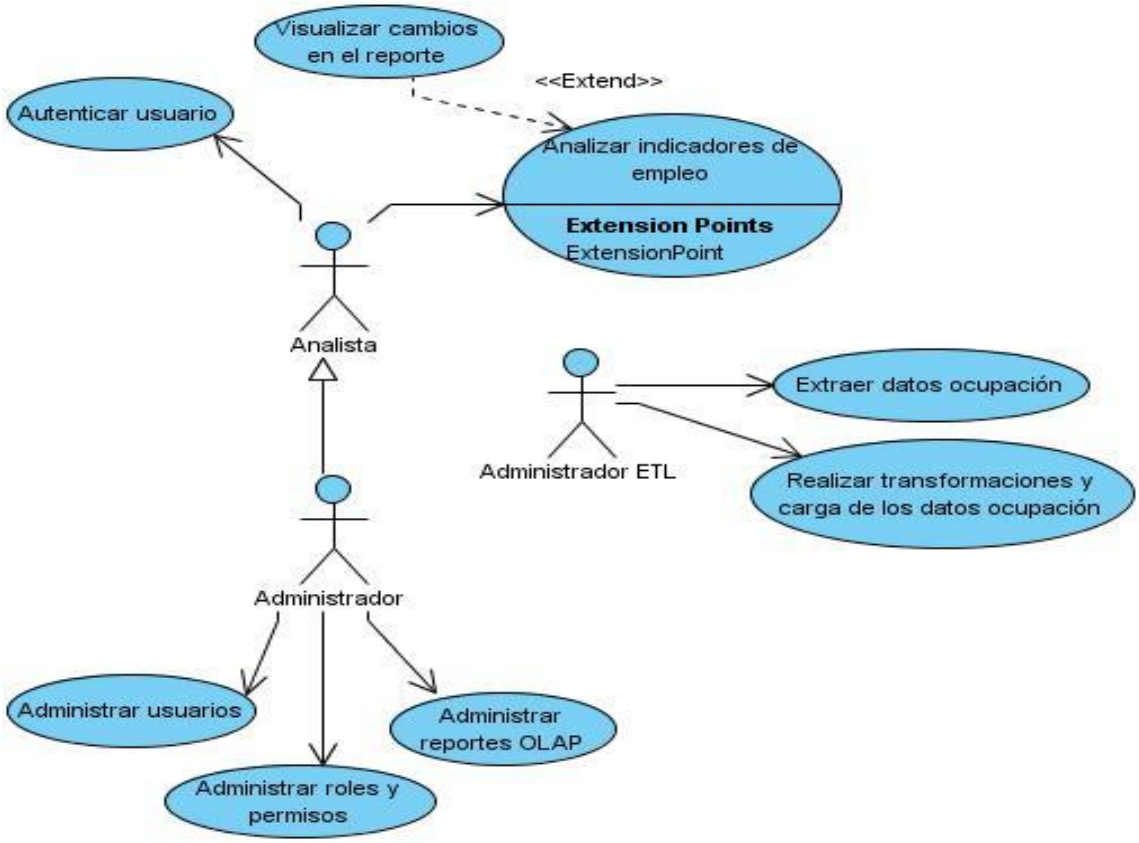


Figura 2. Diagrama de Casos de Uso del Sistema.

Descripción de los usuarios que interactúan en el diagrama de caso de usos.

Actor	Descripción
Administrador	El administrador interactúa con el sistema para realizar todas las operaciones de administración de los roles, permisos, usuarios y reportes.
Administrador ETL	El administrador ETL interactúa con el sistema para monitorizar los procesos de Extracción, Transformación y Carga de los datos.
Analista	El analista interactúa con el sistema para analizar y consultar la información.

Tabla 1. Descripción de los actores.

2.2 Diseño del Mercado de Datos Ocupación

2.2.1 Dimensiones, hechos y medidas

En el presente epígrafe se plantean las 9 dimensiones, 3 medidas, el hecho y los atributos que fueron identificados en el refinamiento del diseño del Mercado de Datos Ocupación. Además, se realiza una breve descripción de los componentes del Mercado de Datos.

Dimensiones	Descripción
dim_cae	Contiene el identificador de la dimensión, el código, el nombre y la descripción de los niveles sector, rama y subrama del que puede pertenecer un trabajador. Jerarquía: Sector → Rama → Subrama
dim_nae	Contiene el identificador de la dimensión, un código, nombre y descripción de los nomencladores de actividad económica en los niveles de sección, división y clase. Jerarquía: Sección → División → Clase
dim_dpa	Contiene el identificador de la dimensión, el código, el nombre y la descripción de cada municipio y provincia establecida nacionalmente, además de la extensión de cada municipio, la extensión de los cayos y la extensión de tierra firme. Jerarquía: Provincia → Municipio
dim_organismo	Contiene el identificador de la dimensión, el código, nombre y la descripción del organismo.
dim_sector_ocupacion	Contiene el identificador de la dimensión, el código, el nombre y la descripción de las variantes ocupacionales establecidas.
dim_entidad	Contiene el identificador de la dimensión, el código, el nombre y la descripción de la entidad que emite la información.
dim_inf_trabajador	Contiene el identificador de la dimensión, un código, el tiempo que es contratado un trabajador y si es mayor o menor de la

	edad laboral.
dim_temporal	Contiene el identificador de la dimensión y el número, nombre y código del año con que se interactúa.
dim_ffinanciamiento	Contiene el identificador de la dimensión, el código, nombre y descripción general y especifica de la forma de financiamiento.

Tabla 2. Descripción de las dimensiones.

Medidas	Descripción	Calculable	Hecho al que pertenece
cant_trabajadores_total	Almacena la cantidad total de trabajadores.	Sí	Número de trabajadores ocupados
cant_trabajadores_fem	Almacena la cantidad de trabajadores del sexo femenino.	Sí	Número de trabajadores ocupados
cant_trabajadores_masc	Almacena la cantidad de trabajadores del sexo masculino.	Sí	Número de trabajadores ocupados

Tabla 3. Descripción de las medidas.

Hecho	Descripción
hech_num_trab_ocupados	El hecho contiene todas las dimensiones y medidas definidas en el modelo.

Tabla 4. Descripción del hecho.

Atributos del hecho	Descripción
dim_temporal_id	Almacena la llave primaria de la dimensión temporal.
dim_nae_id	Almacena la llave primaria de la dimensión NAE.
dim_cae_id	Almacena la llave primaria de la dimensión CAE.
dim_dpa_id	Almacena la llave primaria de la dimensión DPA.
dpa_centro_id	Almacena la llave primaria del lugar donde radica el centro.

dpa_terr_id	Almacena la llave primaria del territorio donde radica el centro.
dim_organismo_id	Almacena la llave primaria de la dimensión del organismo.
organismo_centro_id	Almacena la llave primaria del centro al cual pertenece el organismo.
organismo_est_id	Almacena la llave primaria del centro al cual pertenece el organismo, después de la reestructuración.
dim_ff_id	Almacena la llave primaria de la dimensión.
rama_actividad_id	Almacena la llave primaria del CAE donde se realiza la actividad.
dim_entidad_id	Almacena la llave primaria de la dimensión de la entidad.
dim_sector_ocupacion_id	Almacena la llave primaria de la dimensión sector ocupación.
dim_inf_trabajador_id	Almacena la llave primaria de la dimensión información del trabajador.

Tabla 5. Descripción de los atributos que componen el hecho.

2.2.2 Matriz BUS o Dimensional

La Matriz Bus o Matriz Dimensional tiene como objetivo representar en una tabla las relaciones existentes entre las dimensiones y el hecho identificado, para conformar el Modelo de Datos. Esta permite determinar el impacto que provocaría un cambio en alguna tabla del modelo durante el desarrollo del sistema. A continuación se presentan dicha matriz:

dimensión/hecho	hech_num_trab_ocupados
dim_nae	X
dim_cae	X
dim_dpa	X
dim_organismo	X
dim_entidad	X
dim_ffinanciamiento	X
dim_temporal	X

dim_sector_ocupacion	X
dim_inf_trabajador	X

Tabla 6. Matriz BUS o Dimensional.

Una vez realizada la Matriz Buz se pudo observar las relaciones existentes entre el hecho y las dimensiones. Evidenciando que al relacionar el hecho hech_num_trab_ocupados con las dimensiones del Mercado de Datos se obtiene el esquema estrella.

2.2.3 Modelo de Datos

Un modelo de datos es un conjunto de conceptos, reglas y convenciones que permiten describir y en ocasiones manipular los datos de un cierto mundo real que se desea almacenar en la base de datos. El modelo de datos está formado por dos componentes: componente estática, relacionada con el lenguaje de definición de datos (LDD) y dinámica, relacionada con el lenguaje de manipulación de datos (LMD). La parte estática se refiere a la estructura y la dinámica a que operaciones se pueden realizar sobre cada objeto.

Estáticas: Entidades (u objetos), propiedades (o atributos) de las entidades, y relaciones entre las entidades.
Dinámicas: Operaciones sobre entidades, sobre propiedades o relaciones entre operaciones.

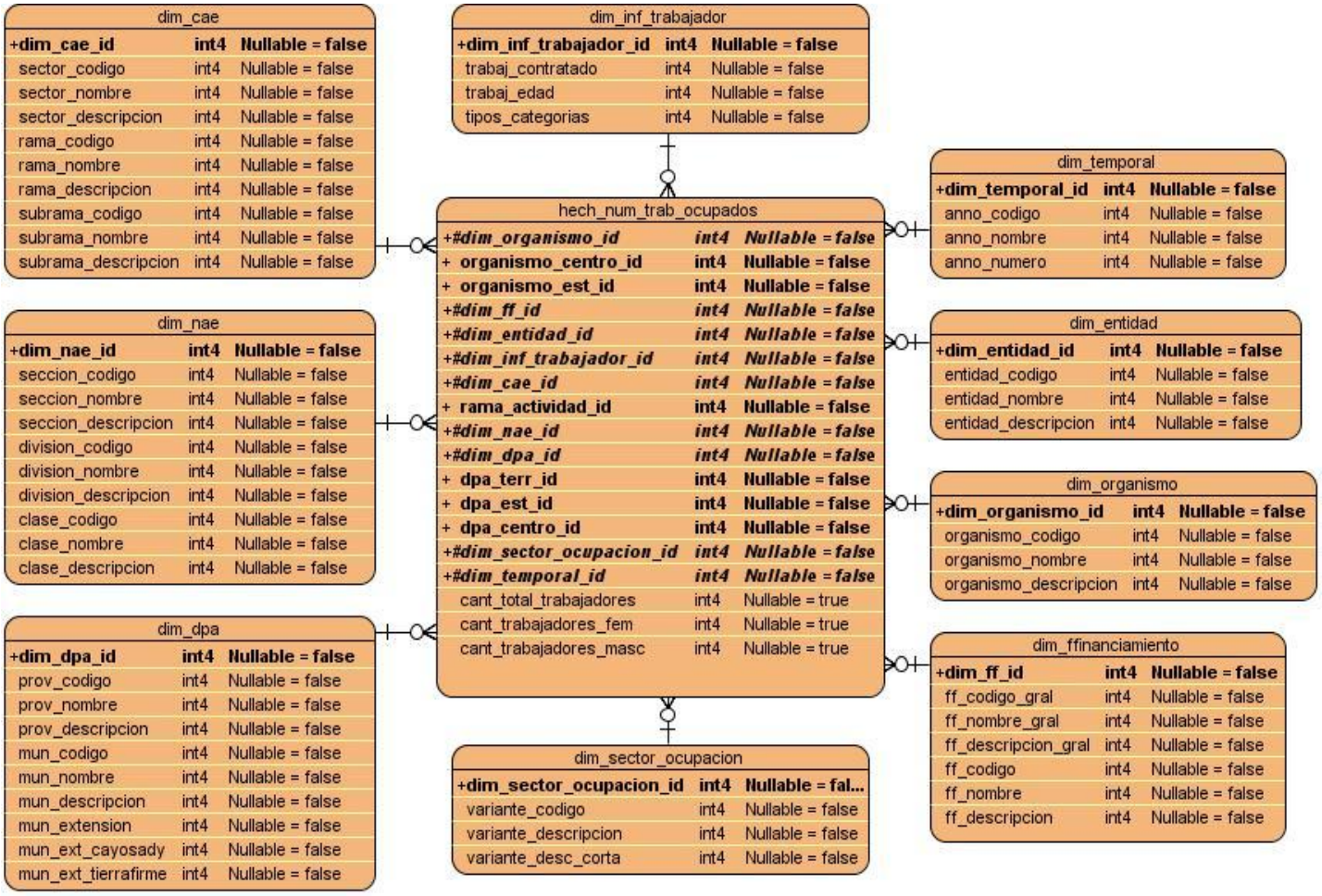


Figura 3. Modelo de Datos.

Al concluir el Modelo de Datos, se observan las tablas de dimensiones y hechos, con los atributos que las componen y las relaciones entre ellas.

2.2.4 Procesos de integración

Diseño de las transformaciones.

El diseño de transformaciones se realiza para definir cómo se comporta la transformación de los datos luego de ejecutado el proceso de ETL.

En la siguiente figura se puede observar cómo se desarrolla la transformación para cargar los datos a la base datos. La cual se inicializa al extraer la fuente de información contenida en las series, después se le aplica la transformación seleccionar/renombrar para ajustar los valores, luego se pasa a la conexión de la base de datos y se finaliza el proceso con la carga de los datos.

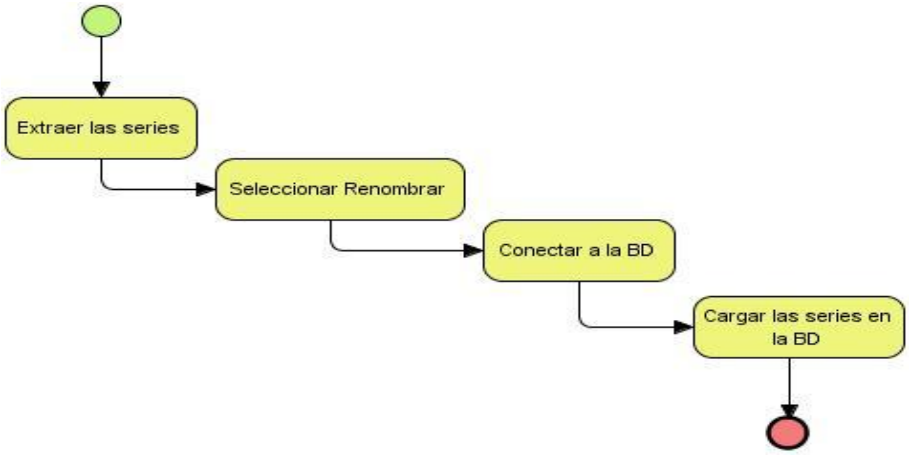


Figura 4. Transformación Carga de series.

2.2.5 Inteligencia de Negocio

Reportes Candidatos.

Los reportes candidatos responden a las necesidades de los usuarios, estos se implementan en la capa de Inteligencia de Negocio. Los reportes candidatos contienen los elementos que componen la estructura de los reportes. En los reportes se definen las medidas y dimensiones (categorías descriptivas) por las que se va a detallar la necesidad del usuario. Además, contienen la frecuencia con la que se actualizan los datos y el tipo de gráfico que se observa en la vista de análisis. A continuación se muestra un ejemplo del reporte candidato: “Obtener la cantidad total de trabajadores.”

Área de análisis (AA)	Ocupación
Libro de Trabajo (LT)	LT1 – Empleo
Reporte (Tabla de Salida – TS)	TS1 – Obtener la cantidad total de trabajadores.
Descripción	Reporte que muestra el total de trabajadores activos en un año determinado por el organismo y el DPA.
Elementos del reporte	dim_dpa: Dimensión División Política Administrativa. dim_organismo: Dimensión Organismo. dim_temporal: Dimensión Temporal. Medida: cant_total_trabajadores. Medida: cant_trabajadores_fem. Medida: cant_trabajadores_masc.
Frecuencia de emisión	Anual

Funciones	
Gráfico	Barras verticales 3D

Tabla 7. Obtener la cantidad total de trabajadores.

2.2.6 Política de respaldo y recuperación

La política de respaldo y recuperación que se utiliza en el Mercado de Datos Ocupación es sencilla y está establecida por tres parámetros:

Periodicidad de salvas del sistema: Se realizarán salvas de toda la información que conforma el Mercado mensualmente. En el servidor de la organización existirá una copia de la salva.

Tablas involucradas: Las tablas involucradas son: el hecho hech_num_trab_ocupados y las dimensiones dim_sector_ocupacion y dim_inf_trabajador.

Salvas existentes: En la actualidad no existe ninguna salva existente, aunque se prevé que la realización sea anual y el estado se chequee mensualmente realizando pruebas de rendimiento y fiabilidad.

2.2.7 Esquema de seguridad

La aplicación de una buena política de seguridad es uno de los aspectos que puede garantizar el buen funcionamiento de la aplicación, se necesita tener mucho cuidado con ello dado que una mala práctica podría afectar al desarrollo de la misma. Para el Mercado de Datos Ocupación la seguridad está determinada por los niveles de permisos y accesos al sistema, apoyándose en los roles que fueron definidos para la interacción de los usuarios con la base de datos de la aplicación.

Seguridad de la base de datos.

Para la interacción con la base de datos se definió el rol de Administrador ETL, el cual se encargará de realizar los procesos de Extracción, Transformación y Carga de los datos.

Seguridad de la aplicación.

Dada la gran interacción que existe entre los usuarios y las aplicaciones brindadas por el servidor Pentaho de Inteligencia de Negocio, es importante que se defina un esquema de seguridad que garantice que las aplicaciones del servidor sean sostenibles.

Los roles que se definieron para la interacción con la aplicación fueron: el administrador y el analista. El administrador es el que tiene acceso pleno a todas las áreas de análisis y es el encargado de administrar los usuarios, roles y permisos en el Sistema de Información de Gobierno. El analista tiene

acceso de solo lectura a las áreas de análisis del Mercado de Datos Ocupación y visualiza los reportes.

Para un mejor entendimiento sobre los permisos que posee cada usuario se presenta la siguiente tabla:

Roles/Permisos	Base de datos		Aplicación	
	Lectura	Escritura	Lectura	Escritura
Administrador de ETL	X	X		
Administrador	X		X	X
Analista	X		X	

Tabla 8. Roles y permisos que poseen los usuarios.

Conclusiones del capítulo.

Luego de desarrollar el capítulo se obtuvieron 24 reglas del negocio, se refinaron 31 requisitos de información, 27 requisitos funcionales y 19 requisitos no funcionales. Se identificaron 3 actores y 8 casos de uso del sistema, a los que se les realizó una descripción textual para su mejor entendimiento. Se realizó la Matriz Bus, con el objetivo de poder observar la trazabilidad existente entre las 9 dimensiones identificadas y el hecho. Se confeccionó el Modelo de Datos, con las 10 tablas identificadas, los atributos y medidas refinadas, para dar solución a las necesidades del cliente. Se diseñaron las principales transformaciones a realizar en la implementación de los procesos de integración y los reportes candidatos, para las vistas de análisis; además de diseñar el esquema de seguridad para los procesos de integración y visualización.

Capítulo 3: Implementación del Mercado de Datos

Durante el desarrollo del capítulo se implementa el modelo físico donde se crea la estructura de los datos, esquemas y tablas de la base de datos, dejando bien definido el modelo de datos. Se realiza la implementación de los procesos de Extracción, Transformación y Carga, efectuándose el perfilado de datos, los flujos de transformaciones y los trabajos a las fuentes de información. Por último, se realiza la implementación de los subsistemas de visualización.

3.1 Implementación del modelo de datos

En un Almacén de Datos es preciso poseer una adecuada distribución de la información, para ello se implementan las tablas del Modelo de Datos. Estas se dividen en dos esquemas: dimensiones y mart_ocupacion. El esquema dimensiones contiene aquellas dimensiones que son comunes en el almacén central, mientras que el mart_ocupacion contiene el hecho y dimensiones propios del Mercado de Datos Ocupación. En la siguiente tabla se distribuyen las dimensiones y hechos según el esquema que le corresponde:

Esquema	Tablas
dimensiones	dim_temporal_id
dimensiones	dim_nae_id
dimensiones	dim_cae_id
dimensiones	dim_dpa_id
dimensiones	dim_organismo_id
dimensiones	dim_ff_id
dimensiones	dim_entidad
mart_ocupacion	dim_sector_ocupacion_id
mart_ ocupacion	dim_inf_trabajador_id
mart_ ocupacion	hech_num_trab_ocupados

Tabla 9. Esquemas y tablas del Mercado de Datos.

3.2 Perfilado de datos

El perfilado de datos ayuda a comprender el estado en que se encuentran los datos, ya que muestra de forma gráfica las estadísticas de los mismos. A través de este proceso se definen los posibles errores que pueda contener la fuente de donde se extrae la información.

Entre los perfiles que se utilizaron se encuentra distribución de valores (value distribution), el cual permite saber cuántos valores están repetidos y qué cantidad de veces se repiten. También se utilizó el análisis de cadenas (string analysis) que muestra el valor máximo, el mínimo, el promedio, entre otros. Además, se utiliza el perfilado de estándares de medida (standard measures), para saber la cantidad de valores que contiene el documento donde se recogió la información y conocer cuáles de estos valores son nulos y vacíos.

Value distribution

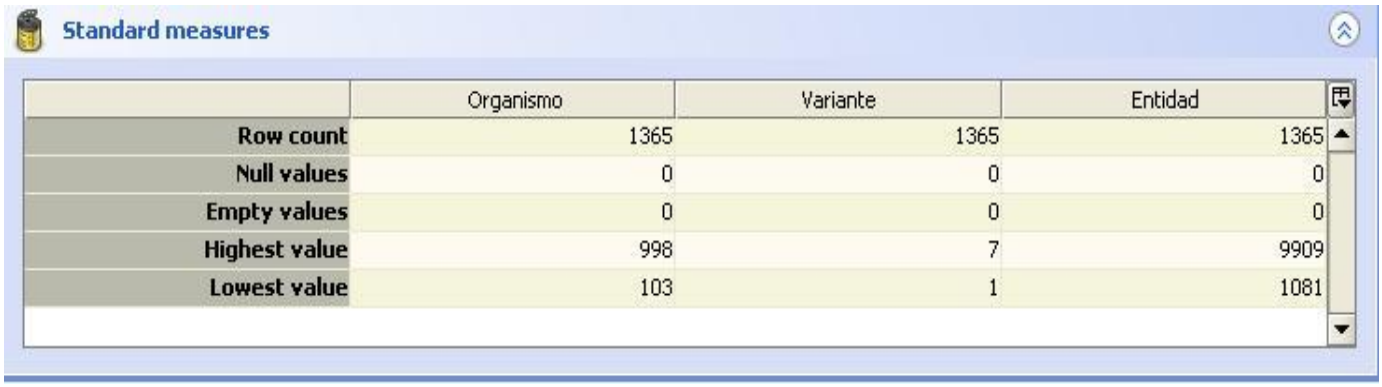
	Organismo	Variante	Entidad
top 1	318 (552)	1 (1040)	9289 (24)
top 2	131 (240)	12 (82)	89701 (24)
top 3	126 (108)	7 (72)	3658 (14)
top 4	108 (84)	3 (48)	3525 (13)
top 5	151 (38)	5 (36)	13690 (12)
bottom 5	233 (4)	<null>	1081 (3)
bottom 4	241 (4)	4 (15)	11098 (3)
bottom 3	174 (3)	14 (12)	1111 (3)
bottom 2	223 (3)	2 (12)	11263 (3)
bottom 1	304 (3)	6 (12)	11540 (3)

Figura 5. Perfilado de distribución de valores del modelo M5200CA.

String analysis

	Organismo	Variante	Entidad
Char count	4095	1495	5778
Max chars	3	2	5
Min chars	3	1	3
Avg chars	3	1,1	4,23
Max white spaces	0	0	0
Min white spaces	0	0	0
Avg white spaces	0	0	0
Uppercase chars	0%	0%	0%
Lowercase chars	0%	0%	0%
Non-letter chars	100%	100%	100%
Word count	1365	1365	1365
Max words	1	1	1
Min words	1	1	1

Figura 6. Perfilado de análisis de cadenas del modelo M5200CA.



	Organismo	Variante	Entidad
Row count	1365	1365	1365
Null values	0	0	0
Empty values	0	0	0
Highest value	998	7	9909
Lowest value	103	1	1081

Figura 7. Perfilado de estándares de medidas del modelo M5200CA.

3.3 Implementación de la base de datos

La implementación de la base de datos se apoyará en tres procesos principales; el primero es la extracción de los datos que se encuentran almacenados en fuentes de distintos formatos, en el caso particular, los datos se almacenan en documentos de formato Excel. El segundo proceso es realizar un grupo de transformaciones y una de las principales acciones que permite es conocer el total de trabajadores, desglosando este resultado en masculino y femenino. Por último, se carga cada una de las dimensiones que se requieren según los códigos establecidos.

La figura 8, muestra la transformación realizada para la carga del hecho hech_num_trab_ocupados. Se comienza por la extracción de los datos de la fuente, luego se utiliza el componente Seleccionar/Renombrar, para obtener y renombrar las columnas necesarias.
Con el objetivo de agrupar todos los datos que responden a los campos del total de trabajadores y total de trabajadores femeninos, se hace uso nuevamente del componente Seleccionar/Renombrar y Unión Ordenada.

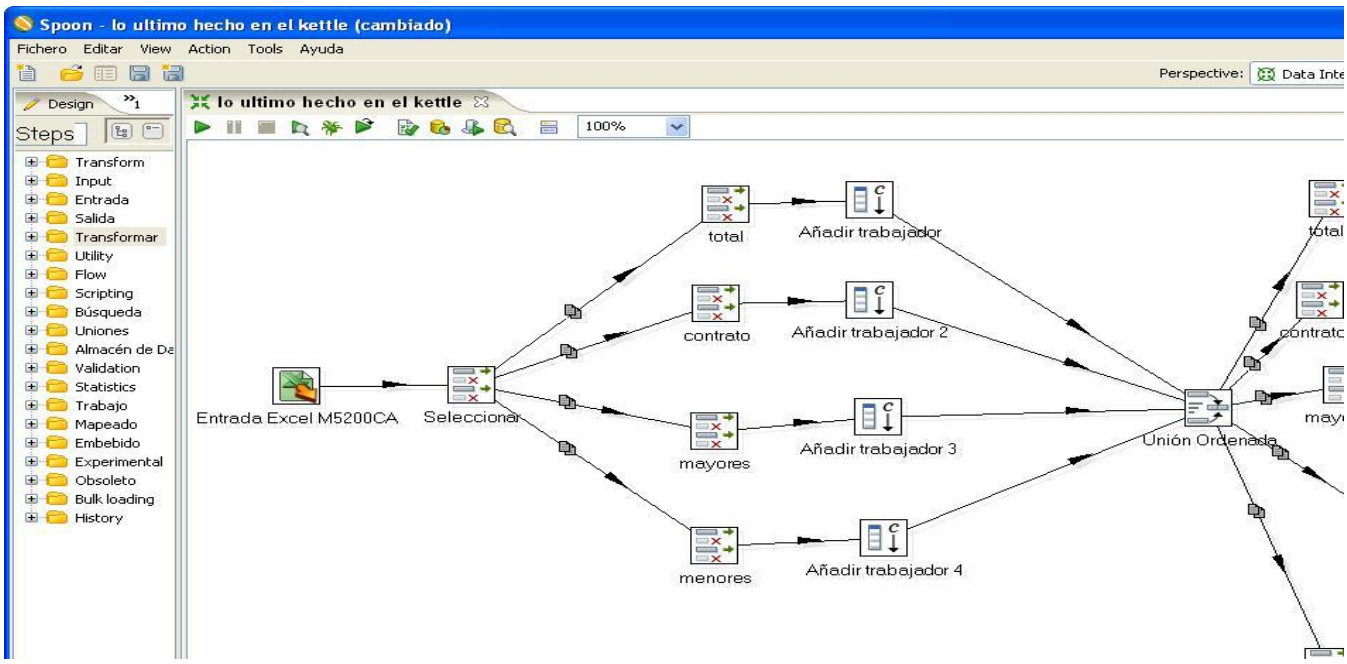


Figura 8. Transformación hech_num_trab_ocupados (parte 1)

En la figura 9 se repite el proceso descrito en la figura 8 para las categorías ocupacionales. Se utiliza, una calculadora para obtener el total de trabajadores masculinos, luego un script para validar que los datos cargados se correspondan con los de la ONE. Se hace uso del componente búsqueda en base de datos, para obtener los identificadores de las dimensiones partiendo de sus códigos. Se trabaja con el componente Añadir constantes, en el cual se definen valores constantes para la dimensión temporal_anno. Finalmente, los valores correctos se almacenan en la tabla hech_num_trabajadores que se encuentra en el Mercado de Datos Ocupación y los valores nulos o posibles errores se registran en un documento Excel para un análisis posterior.

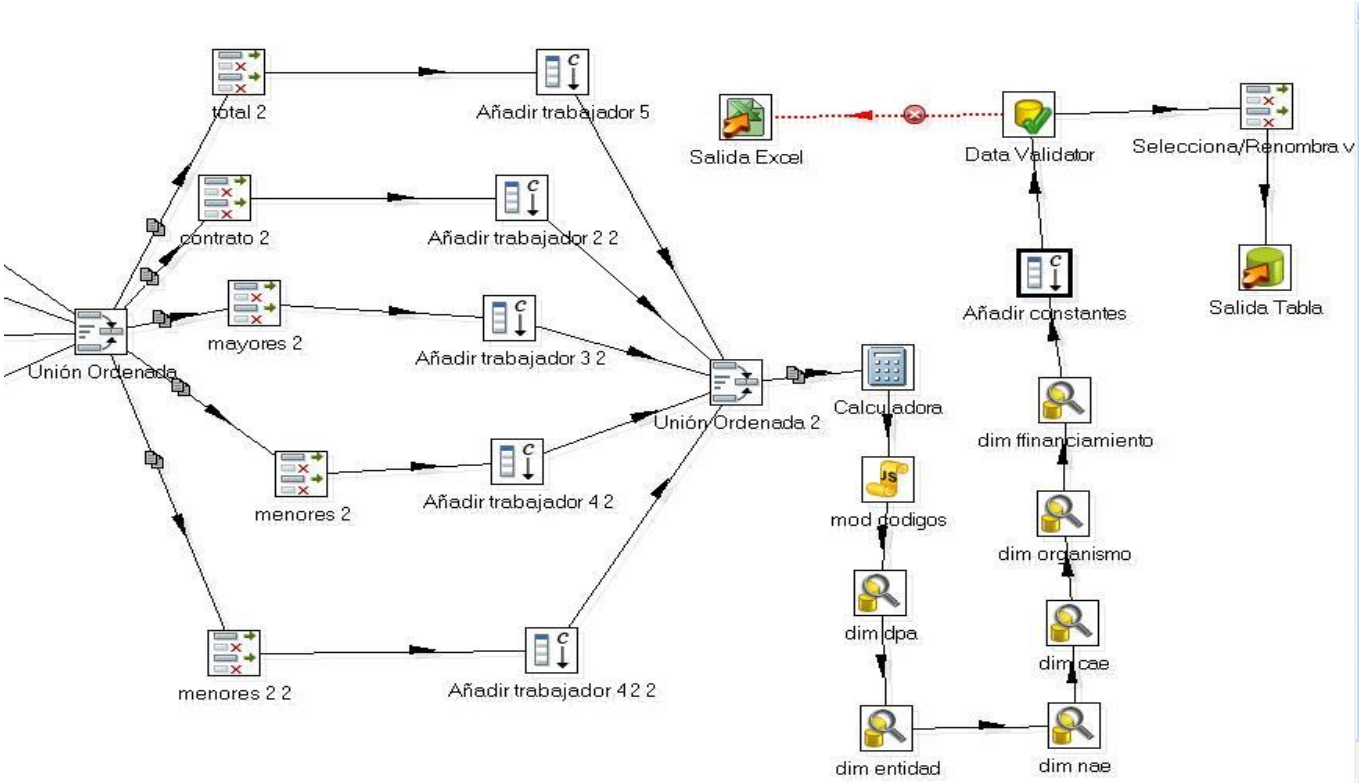


Figura 9. Transformación hech_num_trab_ocupados (parte 2)

3.3.1 Implementación de los Trabajos

Los Jobs o Trabajos son un grupo de tareas que permiten realizar acciones determinadas entre las que se encuentra la de realizar una o más transformaciones a los datos. Las transformaciones se encuentran en un nivel inferior a los Trabajos. La carga de las tablas se debe realizar luego de las transformaciones a los datos, de esta forma no se cargan llaves nulas de otras tablas. Los Trabajos permiten establecer la frecuencia de la carga y en el orden en que se pueden ejecutar las transformaciones.

Para la carga de los datos se realizaron dos trabajos, uno para las dimensiones y otro para el hecho. A continuación se muestran los ejemplos correspondientes de la carga de los mismos.

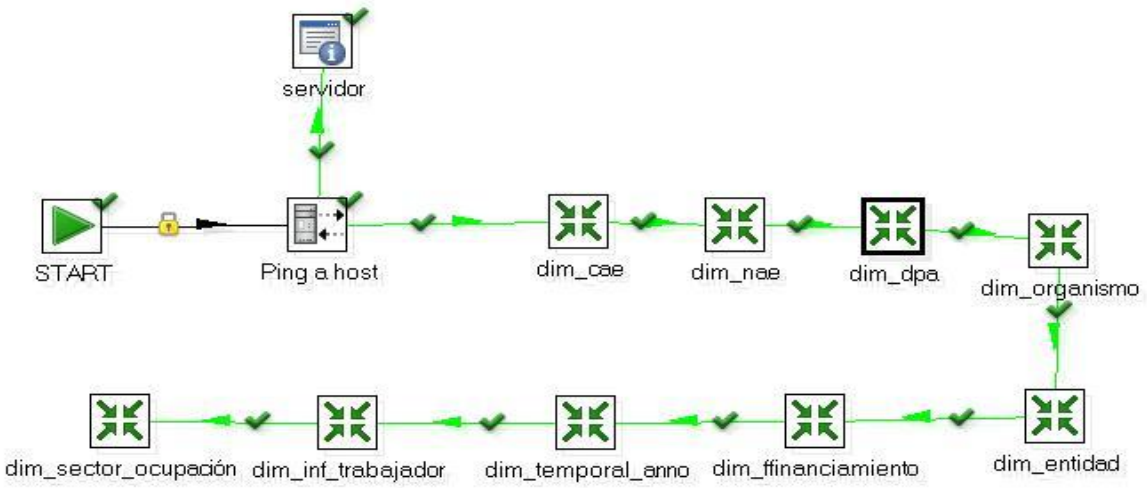


Figura 10. Carga de las dimensiones.

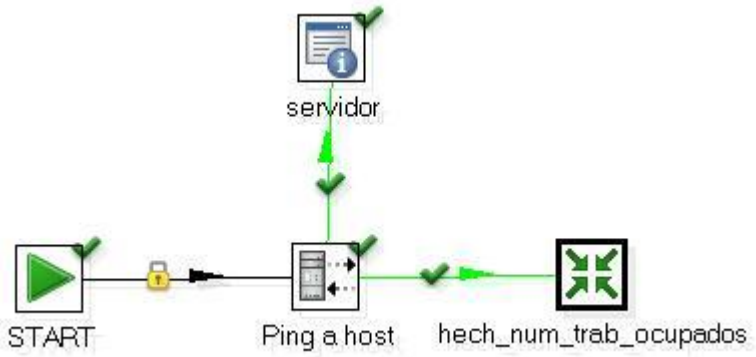


Figura 11. Carga del hecho.

3.4 Implementación del subsistema de visualización de datos

3.4.1 Cubos OLAP

Un cubo OLAP es una base de datos multidimensional, en la cual el almacenamiento físico de los datos se realiza en un vector multidimensional. Este está formado por dimensiones, las cuales son categorías descriptivas por los cuales los datos numéricos, son separados para el análisis. El cubo permite analizar su información de forma multidimensional, lo que permite que el usuario vea la información que desea desde diferentes perspectivas, además de cruzar todas las dimensiones que sean necesarias mostrando nuevas informaciones. El Mercado de Datos Ocupación se compone por un cubo multidimensional, 9 dimensiones, 2 medidas y un miembro calculable.

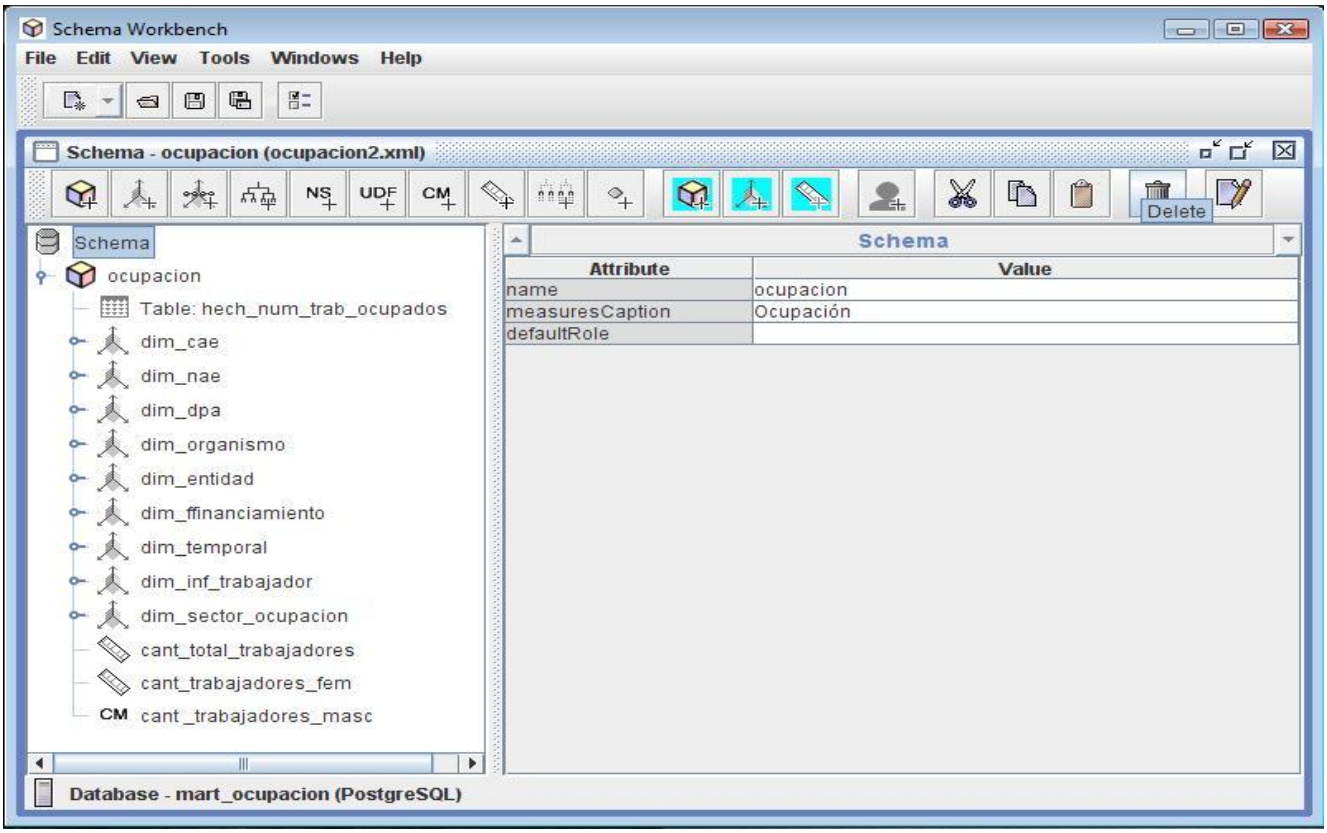


Figura 12. Diseño del cubo y sus componentes utilizando Pentaho Schema Workbench.

3.4.2 Arquitectura de la información

La estructura de navegación se compone por un área de análisis, un libro de trabajo y 13 reportes.

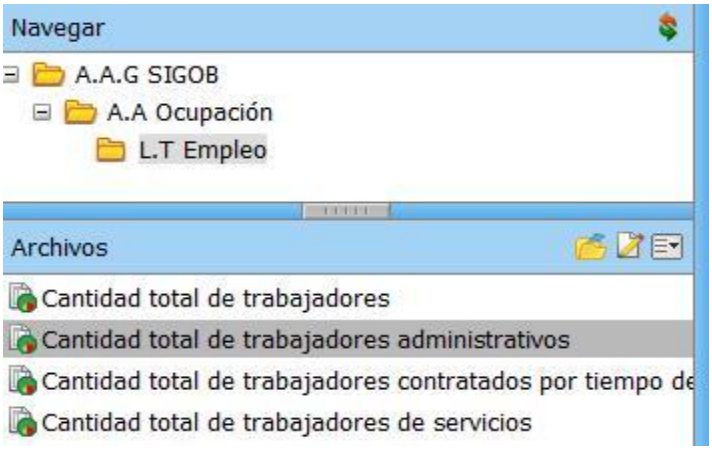


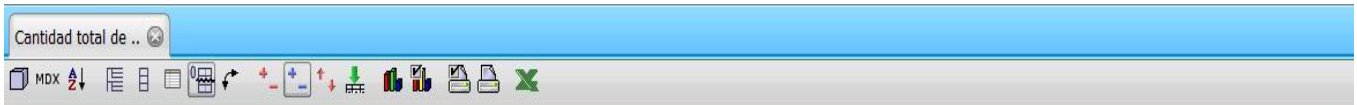
Figura 13. Estructura de navegación.

3.4.3 Reportes

Los reportes se visualizan a través de consultas de formato mdx. El siguiente reporte muestra la cantidad total de trabajadores, el total de trabajadores masculinos y femeninos administrativos según el año, la división política administrativa y el organismo al que pertenecen.

Consulta mdx:

```
select NON EMPTY Crossjoin({[dim_sector_ocupacion].[Administrativos]},
Crossjoin({[dim_temporal_anno].[Todos]}, {[Measures].[cant_trabajadores_fem],
[Measures].[cant_trabajadores_masc], [Measures].[cant_total_trabajadores]})) ON COLUMNS,
 NON EMPTY Crossjoin(Hierarchize(Union({[dim_dpa].[Todos]}, [dim_dpa].[Todos].Children)),
{[dim_organismo].[Todos]}) ON ROWS
from [ocupacion]
```



		Sector Ocupación		
		Administrativos		
		Temporal		
		Todos		
		Ocupación		
DPA	Organismo	Cantidad de trabajadores mujeres (u)	Cantidad de trabajadores hombres (u)	Cantidad total de trabajadores (u)
Todos	Todos	715.460	906.492	1.621.952
Pinar del Río	Todos	52.304	89.989	142.293
La Habana	Todos	1.164	7.662	8.826
Ciudad de La Habana	Todos	1.246	1.731	2.977
Matanzas	Todos	109.856	130.389	240.245
Villa Clara	Todos	8.016	29.226	37.242
Cienfuegos	Todos	67.659	82.465	150.124
Sancti Spíritus	Todos	85.726	102.564	188.290
Ciego de Ávila	Todos	115.718	115.694	231.412
Camagüey	Todos	125.311	132.848	258.159
Las Tunas	Todos	3.474	17.290	20.764
Granma	Todos	127.221	166.647	293.868
Santiago de Cuba	Todos	2.418	2.915	5.333

Figura 14. Reporte cantidad total de trabajadores administrativos.

3.5 Seguridad de los usuarios

El Pentaho BI Server permite agrupar a los usuarios por los roles que desempeñan y las necesidades de acceso de información que precisa cada rol, permitiendo que cada usuario interactúe solo con la información que le concierne del sistema.

Usuarios y Roles.

La interacción con la aplicación mediante la definición de roles y usuarios contribuye a garantizar la seguridad de la misma. En la aplicación se establecieron los siguientes roles.

- ✓ **Authenticated:** Este rol permite al usuario autenticarse en el sistema, por lo que todos los usuarios creados lo contendrán, es un rol propio de la aplicación.
- ✓ **Admin:** Es el rol del administrador del sistema, el cual tendrá acceso a toda la aplicación y será el encargado de administrar y configurar los reportes, roles y usuarios que componen la aplicación.
- ✓ **Analista:** Este rol permite al usuario consultar la información y analizarla desde las diferentes perspectivas.

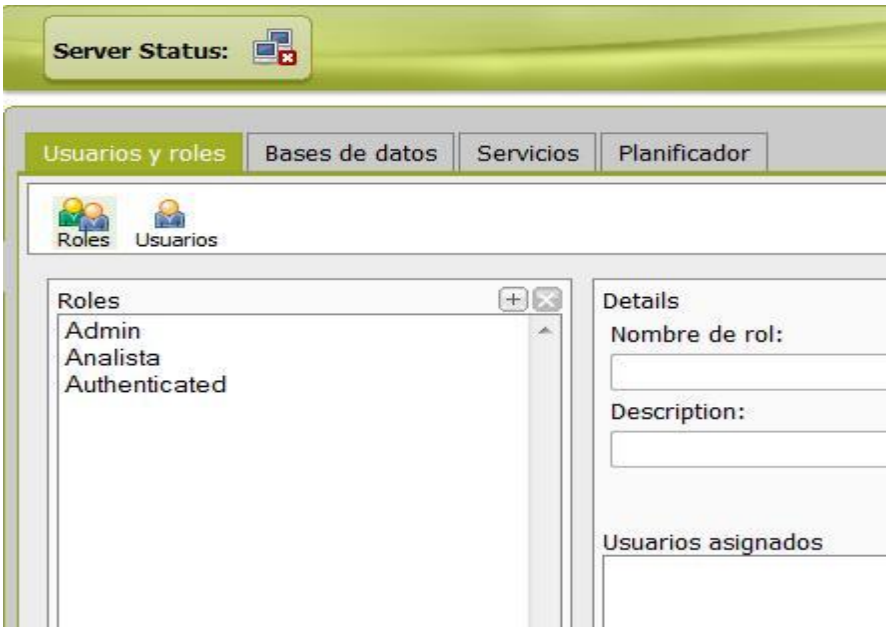


Figura 15. Consola de administración del Pentaho BI Server.

Privilegios:

Los privilegios de cada usuario serán asignados según el rol que desempeñe. Los usuarios obtendrán los siguientes permisos y roles:

- ✓ **Administrador:** Se le asigna el permiso de control absoluto de la aplicación y desarrolla el rol de Admin.

- ✓ **Analista:** Se le asigna el permiso de ejecutar los reportes y analizarlo desde las diferentes perspectivas, y desarrolla el rol de Especialista.

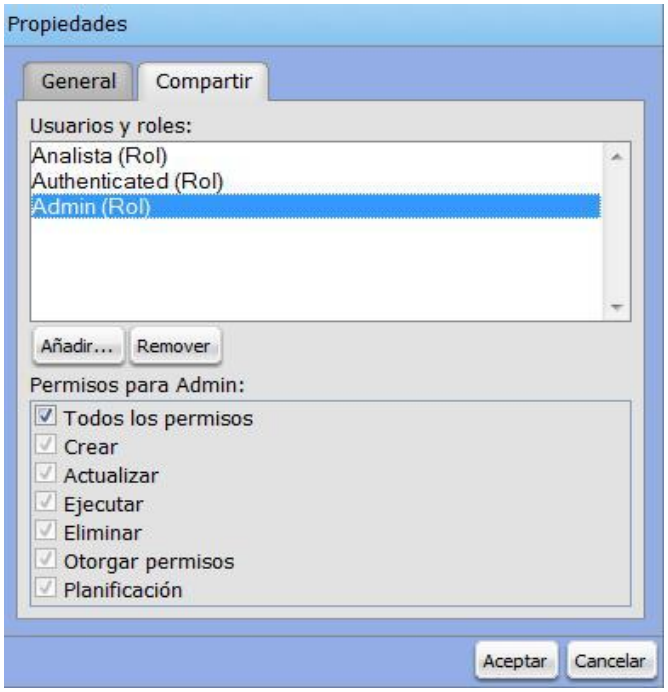


Figura16. Propiedades de los usuarios y roles.

Conclusiones del capítulo.

Durante el desarrollo del capítulo se implementó el modelo físico del Mercado de Datos Ocupación, quedando definidos 2 esquemas: dimensiones y mart_ocupacion. Los esquemas contienen las 10 tablas definidas en el modelo de datos, que se componen por un hecho y 9 dimensiones. Se realizó el perfilado de datos a las fuentes, permitiendo conocer el estado de la información con la que se trabaja. Se implementan los procesos de ETL, en los cuales, se realiza la extracción de los datos contenidos en los documentos Excel del área de Ocupación. Se aplican las transformaciones necesarias a los datos y se cargan en la Base de datos. Por último, se desarrollaron los subsistemas de visualización donde se diseñan e implementan los cubos OLAP, quedando definido un cubo, 9 dimensiones, 2 medidas y un miembro calculable. Se define el mapa de navegación, con un área de análisis, un libro de trabajo y 13 vistas de análisis y los usuarios, roles y permisos que existen en la solución.

Capítulo 4: Validación y pruebas del Mercado de Datos

En el presente capítulo se realizan diferentes pruebas al Mercado de Datos Ocupación, con el objetivo de verificar su correcto funcionamiento y que cuente con la calidad requerida por el cliente. La validación del Mercado de Datos se realiza mediante la aplicación de los casos de pruebas y las listas de chequeo.

4.1 Validación y prueba

Para obtener un producto con calidad, es necesario realizar varios procesos de validación y pruebas que garanticen que el software funciona correctamente. La etapa de validación y pruebas comienza al finalizar la etapa de desarrollo del producto. Las pruebas de software permiten verificar la calidad de un producto, con ellas se identifican los errores cometidos en la implementación del Mercado de Datos Ocupación. Para determinar el nivel de calidad se deben efectuar pruebas que permitan comprobar el grado de cumplimiento de las especificaciones iniciales del sistema.

Entre las diferentes pruebas de software que pueden ser aplicadas, se encuentran las:

Prueba de caja negra: Se refiere a las pruebas que se llevan a cabo sobre la interfaz del software, por lo que los casos de prueba pretenden demostrar que las funciones del software son operativas, que la entrada se acepta de forma adecuada y que se produce una salida correcta, así como que la integridad de la información externa se mantiene. Esta prueba examina algunos aspectos del modelo fundamentalmente del sistema sin tener mucho en cuenta la estructura interna del software.

Prueba de caja blanca: Se basa en el minucioso examen de los detalles procedimentales. Se comprueban los caminos lógicos del software proponiendo casos de prueba que examinen que están correctas todas las condiciones y/o bucles para determinar si el estado real coincide con el esperado o afirmado. Esto genera gran cantidad de caminos posibles por lo que hay que dedicar esfuerzos a la determinación de las condiciones de prueba que se van a verificar.

La Prueba es aplicada para diferentes tipos de objetivos, en diferentes escenarios o niveles de trabajo. Se distinguen los siguientes niveles de pruebas:

- Prueba de desarrollador
- Prueba independiente
- Prueba de Unidad
- Prueba de Integración
- Prueba de sistema

• Prueba de aceptación

Prueba de Desarrollador. Es la prueba diseñada e implementada por el equipo de desarrollo. Tradicionalmente estas pruebas han sido consideradas solo para la prueba de unidad, aunque en la actualidad en algunos casos pueden ejecutar pruebas de integración. Se recomienda que estas pruebas cubran más que las pruebas de unidad.

Prueba independiente: Es la prueba que es diseñada e implementada por alguien independiente del grupo de desarrolladores. El objetivo de estas pruebas es proporcionar una perspectiva diferente y en un ambiente más rico que los desarrolladores.

Prueba de unidad: Es la prueba enfocada a los elementos testeables más pequeño del software. Es aplicable a componentes representados en el modelo de implementación para verificar que los flujos de control y de datos están cubiertos, y que ellos funcionen como se espera. La prueba de unidad siempre está orientada a caja blanca.

Prueba de integración: Es ejecutada para asegurar que los componentes en el modelo de implementación operen correctamente cuando son combinados para ejecutar un caso de uso. Se prueba un paquete o un conjunto de paquetes del modelo de implementación. Estas pruebas descubren errores o incompletitud en las especificaciones de las interfaces de los paquetes.

Prueba de Sistema: Son las pruebas que se hacen cuando el software está funcionando como un todo. Es la actividad de prueba dirigida a verificar el programa final, después que todos los componentes de software y hardware han sido integrados.

Prueba de aceptación: Prueba de aceptación del usuario es la prueba final antes del despliegue del sistema. Su objetivo es verificar que el software está listo y que puede ser usado por usuarios finales para ejecutar aquellas funciones y tareas para las cuales el software fue construido.

Otros procesos para validar los Mercados de Datos son las pruebas de interfaz, casos de prueba y listas de chequeo. Estas se realizan para comprobar que lo planteado en la documentación del producto corresponde con lo desarrollado en la aplicación, y que no existan errores a la hora de entregar el producto final al cliente.

4.2 Pruebas aplicadas al Mercado de Datos Ocupación

Diseño de los casos de prueba.

Los casos de pruebas se diseñan según los casos de usos de información diseñados en el modelo de casos de uso del sistema. Estos se realizan con el propósito de comprobar que las vistas de análisis muestran la información requerida en los requisitos de información planteados por el usuario.

Se aplicaron por parte de los especialistas del departamento los casos de prueba diseñados. En el proceso se identificó 1 no conformidades, la cual se corrigió posteriormente para lograr una correcta disponibilidad de la información. El caso de prueba “Analizar indicadores de empleo” permitió realizar pruebas de caja negra donde se verificó que las funcionalidades establecidas en el levantamiento de requisitos por el cliente se cumplieron satisfactoriamente. Además, se aplicó la prueba de integridad del sistema a los procesos de ETL y BI, la cual mostró que la solución visualizaba los datos correctos en la capa de Inteligencia de Negocio, los datos arrojados en el BI coincidieron con los cargados en la base de datos mediante los procesos de ETL, concluyendo que los procesos de ETL y BI se integraron correctamente.

En el **Anexo 3** se encuentra el diseño del caso de prueba correspondiente al caso de uso de información “Analizar indicadores de empleo” del Mercado de Datos Ocupación.

Aplicación de las listas de chequeo.

Las listas de chequeo constituyen un mecanismo para el control de los riesgos y su función básica es la de detectar condiciones peligrosas que puedan generar incidentes al producto de software. Es un documento que tiene un conjunto de parámetros a medir sobre un aspecto determinado, sea 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 a 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.

Para elaborar la lista de chequeo del Mercado de Datos Ocupación, se tuvieron en cuenta elementos de evaluación que son importantes una vez realizado el proceso de ETL y BI, permitiendo recoger los puntos eficientes e ineficientes que posean dichos procesos. La lista de chequeo contiene diferentes indicadores a evaluar, los cuales se encuentran distribuidos en tres secciones fundamentales:

- ✓ **Estructura del documento**: Abarca todos los aspectos definidos por el expediente de proyecto o el formato establecido por el proyecto.
- ✓ **Indicadores definidos**: Abarca todos los indicadores a evaluar durante la etapa de desarrollo del Mercado de Datos según el modelo de desarrollo.
- ✓ **Semántica del documento**: Contempla todos los indicadores a evaluar respecto a la ortografía y redacción.

Para observar la lista de chequeo aplicada remitirse al **Anexo 4.**

Evaluación de las listas de chequeo.

Una vez aplicadas las listas de chequeo se realiza la evaluación del Mercado de Datos. Para ello es necesario realizar una apreciación de los resultados obtenidos con las listas de chequeo, en las cuales se identificaron 14 indicadores, 8 de ellos críticos y se generaron 6 no conformidades dándose solución a las mismas.

A continuación se muestra una gráfica que refleja el resultado de la evaluación según los parámetros de las listas:

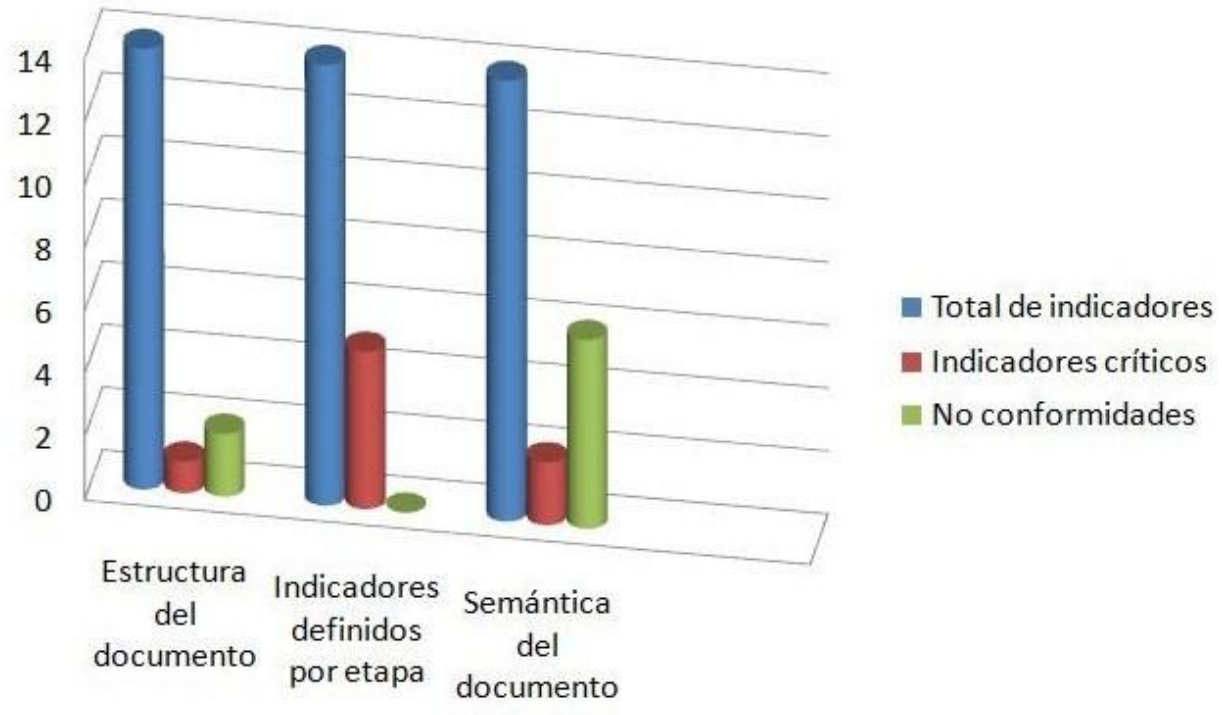


Figura 17. Comportamiento de los indicadores por secciones.

Prueba interna del departamento.

Para la validación del Mercado de Datos se aplicaron por parte de los especialistas del departamento los casos de prueba diseñados. Durante el proceso se detectaron 18 no conformidades, que no afectan el buen funcionamiento del producto pero si la calidad de la información presentada. Las no conformidades detectadas fueron resueltas de forma satisfactoria.

Pruebas DATEC.

Los especialistas de calidad del centro realizaron la evaluación de los casos de prueba diseñados, durante el procedimiento se encontraron 2 no conformidades afectando la calidad de la documentación. Posteriormente se les dio una solución correcta a las no conformidades.

Prueba de aceptación.

La validación es la comprobación de que la solución está acorde a las necesidades y exigencias de los clientes. Para aprobar la propuesta de solución se realizó un encuentro con la representante de la ONE en la universidad: Elena Leonila Fernández García, quien validó que los datos mostrados eran los requeridos.

Conclusiones del capítulo.

Durante el desarrollo del presente capítulo se describe el proceso de prueba y evaluación del Mercado de Datos Ocupación, donde se aplicaron casos de prueba y listas de chequeo. La evaluación del caso de prueba comprobó que lo planteado en la documentación corresponde con lo desarrollado en el sistema y las listas de chequeo arrojaron un total de 28 no conformidades, se identificaron también 14 indicadores de los cuales 8 fueron críticos. Se aplicaron las pruebas de caja negra y de integración, arrojando resultados positivos sobre la integración de la solución. Se realizaron las pruebas internas y de aceptación. Los resultados obtenidos fueron evaluados de bien, demostrando la calidad del proceso realizado.

Conclusiones generales.

Al concluir la solución, se puede plantear que las tareas de la investigación y los objetivos específicos definidos se cumplieron, arrojando las siguientes conclusiones:

- ✓ Con el refinamiento del análisis y el diseño del Mercado de Datos Ocupación se identificaron y refinaron las tablas de dimensiones y hecho.
- ✓ Con la implementación de los procesos de ETL se obtuvo un Mercado de Datos con información organizada, centralizada, consistente y homogénea.
- ✓ Con las tablas de hechos y dimensiones identificadas se implementaron los cubos multidimensionales OLAP.
- ✓ La implementación del subsistema de visualización queda compuesta por un área de análisis, un libro de trabajo y 13 reportes.
- ✓ La aplicación de las listas de chequeo y casos de pruebas permitieron evaluar la solución obteniendo resultados satisfactorios.

Recomendaciones.

✓ Integrar el Mercado de Datos Ocupación con el Sistema Informático de Gestión Estadística (SIGE).

✓ Incluir el año en el nombre de la fuente de datos para el Mercado de Datos Ocupación de la ONE.

✓ Extender el uso de la tecnología de los Almacenes de Datos a otras instituciones del país que almacenan grandes volúmenes de información.

Referencias Bibliográficas.

1. Business Intelligence, OpenSource, Pentaho. (2010). Recuperado el 2010, de http://churriwifi.wordpress.com/2010/07/04/17-3-preparando-el-analisis-dimensional-definicion-de-cubos-utilizando-schema-workbench/
2. Castillo, C. (2008). Sistemas de información 2.
3. Curto, J. (2009). ETL: Kettle: Pentaho Data Integration. Recuperado el 2010, de http://bi-businessintelligence.blogspot.com/2009/04/pentaho.html
4. Daniel Muñoz, F. F. (2011). Principales Bases de Datos que usa el mercado.
5. Epsilon. (2009). Mysql workbench. Una buena herramienta para las bases de datos . Recuperado el 2010, de http://www.rinconinformatico.net/mysql-workbench-una-buena-herramienta-para-las-bases-de-datos
6. Francisco José Lucas-Torres Torrillas, S. N. (2008-2009). Modelos Avanzados de bases de datos. ALMACENES DE DATOS y BASES DE DATOS XML. Recuperado el 2010, de http://alarcos.inf-cr.uclm.es/doc/bbddavanzadas/08-09/FUNCIONALIDAD%204.pdf
7. Hall, M. B. (2003). Core Servlets and JavaServer Pages: Volume 1: Core Technologies. Prentice Hall PTR.
8. http://www.gravitar.biz. (s.f.). Obtenido de http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/
9. http://www.http-peru.com/. (s.f.). Obtenido de http://www.http-peru.com/postgresql.php
10. http://www.sinnexus.com. (s.f.). Obtenido de http://www.sinnexus.com/business_intelligence/olap_vs_oltp.aspx
11. http://www.sinnexus.com. (s.f.). Obtenido de http://www.sinnexus.com/business_intelligence/olap_avanzado.aspx
12. http://www.worldlingo.com. (s.f.). Obtenido de http://www.worldlingo.com/ma/enwiki/es/ROLAP
13. INE. (2010). INE Instituto Nacional de Estadística. Recuperado el 2010, de http://www.ine.es/
14. Ineatacama. (s.f.). www.ineatacama.cl. Recuperado el 2010, de http://www.ineatacama.cl
15. Inmon, W. H. (2006). Definiciones de Almacén de Datos.
16. Kimball, R. (2006). Definiciones de Almacén de Datos.

17. Sanz, Miguel Rodriguez. 2010. Análisis y diseño de un DataMart. Madrid : s.n., 2010.
18. Sistema Gestor de Base de Datos PostgreSQL. (2007). Obtenido de http://www.http-peru.com/postgresql.php
19. Torres, L. P. (2006). Curso Almacenes de Datos. Importancia estándar. Recuperado el 2010, de http://www.mailxmail.com/curso-almacenes-datos-importancia-estandar/caracteristicas-almacen-datos
20. Ventajas de PostgreSQL. (2003). Obtenido de http://soporte.tiendalinux.com/portal/Portfolio/postgresql_ventajas_html
21. Verástegui, H. C. (2007). Modelo Dimensional de Datos.

Bibliografía.

1. Business Intelligence, OpenSource, Pentaho. (2010). Recuperado el 2010, de http://churriwifi.wordpress.com/2010/07/04/17-3-preparando-el-analisis-dimensional-definicion-de-cubos-utilizando-schema-workbench/
2. Castillo, C. (2008). Sistemas de información 2.
3. Curto, J. (2009). Gestion del Rendimiento.
4. Curto, J. (2009). ETL: Kettle: Pentaho Data Integration. Recuperado el 2010, de http://bi-businessintelligence.blogspot.com/2009/04/pentaho.html
5. Daniel Muñoz, F. F. (2011). Principales Bases de Datos que usa el mercado.
6. Epsilon. (2009). Mysql workbench. Una buena herramienta para las Bases de datos . Recuperado el 2010, de http://www.rinconinformatico.net/mysql-workbench-una-buena-herramienta-para-las-bases-de-datos
7. Francisco José Lucas-Torres Torrillas, S. N. (2008-2009). Modelos Avanzados de Bases de datos. ALMACENES DE DATOS y BASES DE DATOS XML. Recuperado el 2010, de http://alarcos.inf-cr.uclm.es/doc/bbddavanzadas/08-09/FUNCIONALIDAD%204.pdf
8. García, G. C. (2008). tesisUPV2842. Recuperado el 2010, de http://dspace.upv.es/xmlui/bitstream/handle/10251/2505/tesisUPV2842.pdf?sequence=1
9. Hall, M. B. (2003). Core Servlets and JavaServer Pages: Volume 1: Core Technologies. Prentice Hall PTR.
10. Hernández, J. O. (2003). Análisis y Extracción de Conocimiento en Sistemas de Información:Datawarehouse y Datamining.
11. http://datacleaner.eobjects.org/. (s.f.). Recuperado el 2010, de http://datacleaner.eobjects.org/
12. http://soporte.tiendalinux.com/. (s.f.). Obtenido de http://soporte.tiendalinux.com/portal/Portfolio/postgresql_ventaja
13. http://www.buenastareas.com/. (s.f.). Recuperado el 2010, de http://www.buenastareas.com/ensayos/Almacenes-De-Datos-Y-Miner%C3%ADa-De/614076.html
14. http://www.gravitar.biz. (s.f.). Obtenido de http://www.gravitar.biz/index.php/herramientas-bi/pentaho/caracteristicas-pentaho/
15. http://www.http-peru.com/. (s.f.). Obtenido de http://www.http-peru.com/postgresql.php
16. http://www.sinnexus.com. (s.f.). Obtenido de

http://www.sinnexus.com/business_intelligence/olap_vs_oltp.aspx

17. http://www.sinnexus.com. (s.f.). Obtenido de http://www.sinnexus.com/business_intelligence/olap_avanzado.aspx
18. http://www.slideshare.net/. (s.f.). Obtenido de http://www.slideshare.net/vanquishdarkenigma/visual- paradigm-for-uml
19. http://www.summan.com. (s.f.). Recuperado el 2010, de http://www.summan.com/index.php/productos/software/pentaho-.html
20. http://www.worldlingo.com. (s.f.). Obtenido de http://www.worldlingo.com/ma/enwiki/es/ROLAP
21. INE. (2010). INE Instituto Nacional de Estadística. Recuperado el 2010, de http://www.ine.es/
22. Ineatacama. (s.f.). www.ineatacama.cl. Recuperado el 2010, de http://www.ineatacama.cl
23. Inmon, W. H. (2006). Definiciones de Almacén de Datos.
24. Kimball, R. (2006). Definiciones de Almacén de Datos.
25. Kurniawan, B. (2002). JAVA for the Web with Servlets, JSP, and EJB: A Developer's Guide to J2EE Solutions. New Riders Publishing.
26. Rubia, J. M. (2007). Introducción a los almacenes de datos.
27. Sanz, Miguel Rodriguez. 2010. Análisis y diseño de un DataMart. Madrid : s.n., 2010.
28. Sistema Gestor de Base de Datos PostgreSQL. (2007). Obtenido de http://www.http-peru.com/postgresql.php
29. Torres, L. P. (2006). Curso Almacenes de Datos. Importancia estándar. Recuperado el 2010, de http://www.mailxmail.com/curso-almacenes-datos-importancia-estandar/caracteristicas-almacen-datos
30. Ventajas de PostgreSQL. (2003). Obtenido de http://soporte.tiendalinux.com/portal/Portfolio/postgresql_ventajas_html
31. Verástegui, H. C. (2007). Modelo Dimensional de Datos.
32. http://cantabria.fspugt.es/uploads/documentos/documentos_Comunicado_INE_15-feb-2008__1_809cbef7.pdf
33. http://ftp.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-227/paper05.pdf
34. http://kuainasi.ciens.ucv.ve/ideas07/documentos/articulos_ideas/Articulo82.pdf
35. http://dspace.ucbscz.edu.bo/dspace/bitstream/123456789/606/1/1633.pdf
36. http://dialnet.unirioja.es/servlet/dcfichero_articulo?codigo=1300079&orden=0

37. http://ficcte.unimoron.edu.ar/wicc/Trabajos/III%20-%20isbd/677-wicc2006bdmultimedia.pdf
38. http://nti.uji.es/docs/nti/Jordi_Adell_EDUTEC.html
39. http://recyt.fecyt.es/index.php/RIAII/article/view/10504/7303
40. http://scholar.google.es/scholar?q=related:fm4FItU-rC0J:scholar.google.com/&hl=es&as_sdt=2000
41. http://churriwifi.wordpress.com/2010/04/19/15-2-ampliacion-conceptos-del-modelado-dimensional/
42. http://egkafati.bligoo.com/content/view/302166/Datawarehouse-y-sus-principales-caracteristicas.html
43. http://fccea.unicauca.edu.co/old/datawarehouse.htm
44. http://etl-tools.info/es/bi/almacendedatos_arquitectura.htm
45. http://es.kioskea.net/contents/entreprise/datamining-olap.php3
46. http://technet.microsoft.com/es-es/library/ms175367%28SQL.90%29.aspx
47. http://en.scientificcommons.org/8811423
48. http://tortoisesvn.net/
49. http://tortoisesvn.tigris.org/
50. http://es.calameo.com/read/0001827671c89418c6890
51. http://profesores.elo.utfsm.cl/~agv/elo330/2s02/projects/denzer/informe.pdf
52. http://profesores.elo.utfsm.cl/~agv/elo330/2s02/projects/denzer/postgresql.ppt
53. http://postgresql.uci.cu/node/109
54. http://danielpecos.com/docs/mysql_postgres/b164.html
55. http://www.sinnexus.com/business_intelligence
56. http://www.ine.gub.uy/biblioteca/raza/MODULO_RAZA.pdf
57. http://www.univo.edu.sv:8081/tesis/018166/018166_Cap1.pdf
58. http://www.mitecnologico.com/Main/DefinicionesConceptosMercadosDatos
59. http://www.csae.map.es/csi/silice/DW21.html
60. http://www.elprofesionaldelainformacion.com/contenidos/1996/noviembre/data_warehouses_nuevas_perspectivas_en_la_gestin_de_los_sistemas_de_informacin_parte_ii.html
61. http://www.sqlmax.com/dataw1.asp
62. http://www.navactiva.com/es/asesoria/data-ware-house-ventajas-y-desventajas_19520
63. http://www.scribd.com/doc/27007744/Que-Es-Un-Data-Warehouse
64. http://www.rhernando.net/modules/tutorials/doc/bd/dw.html
65. http://www.dataprix.com/rolap-vs-molap
66. http://www.gnulamp.com/rolap.html
67. http://www.gestiopolis.com/delta/term/TER321.html

68. http://www.dataprix.com/empresa/recursos/molap
69. www.one.cu
70. http://www.stratebi.es/evento_pilaos/tools.php
71. http://www.lgs.com.ve/pres/PresentacionES_PSQL.pdf
72. http://www.postgresql.org/about/featurematrix
73. http://www.postgresql.org/community/contributors/
74. http://www.mysql.com/documentation/index.html .
75. http://www.fedora-es.com/node/189
76. http://www.aplicacionesempresariales.com/postgresql-84.html
77. http://www.ecm-spain.com/interior.asp?IdItem=7308
78. http://www.desarrolloweb.com/articulos/840.php
79. http://www.postgresql-es.org/sobre_postgresql

Anexo 1 - Modelo 5200 de la ONE

ONE OFICINA NACIONAL DE ESTADÍSTICAS	Sistema de Información de Estadística Nacional (SIEN)	NÚMERO DE TRABAJADORES POR CATEGORÍA OCUPACIONAL Y SEXO	INFORME AL CIERRE DE: 30 de septiembre del año: ____	MODELO No. 5200-03 Página ____ de ____ ANUAL
Centro informante:		Código del centro informante:	Variante:	UNIDAD DE MEDIDA: Uno

MUNICIPIOS	CÓDIGO D.P.A.	NÚMERO DE TRABAJADORES SEGÚN REGISTRO																	
		TOTAL		CONTRATADOS POR TIEMPO DETERMINADO		MAYORES DE LA EDAD LABORAL QUE TRABAJAN		MENORES DE LA EDAD LABORAL QUE TRABAJAN		CATEGORÍA OCUPACIONAL									
										Operarios		Técnicos		Administrativos		De servicios		Dirigentes	
		Total	Mujeres	Total	Mujeres	Total	Mujeres	Total	Mujeres	Total	Mujeres	Total	Mujeres	Total	Mujeres	Total	Mujeres	Total	Mujeres
A	B	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18
Total																			
Suma de control																			

Anexo 2 - Caso de uso Analizar indicadores de empleo.

Caso de Uso:	Analizar indicadores de empleo.
Tipo:	Información.
Actores:	Analista.
Resumen:	El analista inicia el caso de uso cuando necesita obtener información sobre los indicadores de empleo, luego selecciona el reporte que responde a su necesidad. El sistema muestra la información contenida en la consulta, finalizando el caso de uso.
Precondiciones:	Completitud del almacén. Carga de los Datos. Usuario autenticado.
Referencias:	R1 – R31.
Prioridad:	Crítico.
Flujo Normal de Eventos	
Acción del Actor	**Respuesta del Sistema**

	1. El sistema muestra las áreas de análisis existentes.
2. El analista selecciona el área de análisis relacionada a los indicadores de empleo.	3. El sistema muestra los libros de trabajos existentes en el área escogida.
4. El analista selecciona el libro de empleo.	5. El sistema muestra los reportes contenidos en el libro de empleo.
6. El analista selecciona el reporte deseado.	7. El sistema muestra la información del reporte seleccionado.

Opciones de reportes de Empleo		
Entradas	**Posibles resultados**	
	Salidas	**Periodicidad**
Variables de entrada relacionadas con el caso de uso Analizar indicadores de empleo: organismo, DPA[3], NAE[4], CAE[5], temporal, sector ocupación, información del trabajador, entidad y forma de financiamiento.	Variables de salida disponibles en el caso de uso Analizar indicador de empleo: ✓ *Cantidad total de trabajadores.* ✓ *Cantidad de trabajadores del sexo femenino.* ✓ *Cantidad de trabajadores del sexo masculino.*	El rango de tiempo en que se solicitan las variables de salida es anual.
Poscondiciones:	Disponibilidad de reportes relacionados con el caso de uso Analizar indicadores de empleo.	

Anexo 3 - Caso de prueba Analizar indicadores de empleo.

Escenario	Descripción	Perfiles de análisis	Indicadores a medir	Respuesta del sistema	Flujo central
EC 1.1: Cantidad total de trabajadores	Permite visualizar los reportes con las variables presentes en	Organismo DPA Temporal	Cantidad total de trabajadores Cantidad de trabajadores mujeres	Se muestra la tabla con los valores correspondientes a cada	Se ejecuta la aplicación. Se autentifica el usuario. Se entra al

[3] DPA: Distribución Política Administrativa
[4] NAE: Nomencladores de Actividad Económica
[5] CAE: Clasificador de Actividad Económica

	el mismo.		Cantidad de trabajadores hombres	escenario y su gráfico.	sistema. Se despliega hacia la derecha el componente ubicado en el lateral izquierdo que contiene el navegador. Se selecciona el área de análisis de **AA.Ocupación.** Se selecciona el libro de trabajo **LT.Empleo.** En la parte inferior izquierda se selecciona el reporte deseado. En el área de trabajo se visualiza la tabla correspondiente al reporte. Se visualiza el gráfico correspondiente a la información de la tabla. Se visualiza el gráfico en correspondencia con el tipo de gráfico que haya determinado el cliente.
EC 1.2: Cantidad total de trabajadores administrativos		Organismo DPA Temporal Sector Ocupación	Cantidad total de trabajadores Cantidad de trabajadores mujeres Cantidad de trabajadores hombres		
EC 1.3: Cantidad total de trabajadores contratados por tiempo determinado		Organismo DPA Temporal Información del trabajador	Cantidad total de trabajadores Cantidad de trabajadores mujeres Cantidad de trabajadores hombres		
EC 1.4: Cantidad total de trabajadores de servicios		Organismo DPA Temporal Sector Ocupación	Cantidad total de trabajadores Cantidad de trabajadores mujeres Cantidad de trabajadores hombres		
EC 1.5: Cantidad total de trabajadores dirigentes		Organismo DPA Temporal Sector Ocupación	Cantidad total de trabajadores Cantidad de trabajadores mujeres Cantidad de trabajadores hombres		
EC 1.6: Cantidad total de trabajadores mayores de la edad laboral		Organismo DPA Temporal Información del trabajador	Cantidad total de trabajadores Cantidad de trabajadores mujeres Cantidad de trabajadores hombres		
EC 1.7: Cantidad total de		Organismo DPA	Cantidad total de trabajadores		

trabajadores menores de la edad laboral		Temporal Información del trabajador	Cantidad de trabajadores mujeres Cantidad de trabajadores hombres		
EC 1.8: Cantidad total de trabajadores operarios		Organismo DPA Temporal Sector Ocupación	Cantidad total de trabajadores Cantidad de trabajadores mujeres Cantidad de trabajadores hombres		
EC 1.9: Cantidad total de trabajadores técnicos		Organismo DPA Temporal Sector Ocupación	Cantidad total de trabajadores Cantidad de trabajadores mujeres Cantidad de trabajadores hombres		
EC 1.10: Distribución de la fuerza de trabajo por categoría ocupacional y sexos		Organismo DPA Temporal Sector Ocupación	Cantidad total de trabajadores Cantidad de trabajadores mujeres Cantidad de trabajadores hombres		
EC 1.11: Distribución por edades de los trabajadores por categoría ocupacional		Organismo DPA Temporal Información del trabajador Sector Ocupación	Cantidad total de trabajadores Cantidad de trabajadores mujeres Cantidad de trabajadores hombres		
EC 1.12: Ocupados por clase actividad económica		Temporal NAE	Cantidad total de trabajadores Cantidad de trabajadores mujeres Cantidad de trabajadores hombres		

EC 1.13: Población económicamente activa		DPA Temporal NAE	Cantidad total de trabajadores Cantidad de trabajadores mujeres Cantidad de trabajadores hombres		

Anexo 4 - Lista de chequeo

Estructura del documento					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
crítico	1.¿Los entregables contienen las secciones obligatorias de la plantilla estándar definidas para un expediente de proyecto? (portada, control de versiones, reglas de confidencialidad, tabla de contenidos y contenido) (ver expediente de proyecto)	3		2	
Indicadores definidos en el desarrollo					
Peso	**Indicadores a evaluar**	**Eval**	**(NP)**	**Cantidad de elementos afectados**	**Comentarios**
	1. ¿Se utilizó un lenguaje cuyas sentencias son expresables mediante una sintaxis bien definida?	5		0	
crítico	2. ¿Los reportes son configurables a través de la interfaz del sistema?	5		0	
	3. ¿La interfaz está orientada a facilitar el uso de las funciones del sistema por parte de los usuarios?	5		0	
crítico	4. ¿No existen restricciones para construir cubos OLAP con dimensiones y niveles de agregación ilimitados?	5		0	

crítico	5. ¿Los usuarios son capaces de manipular los resultados de manera que se ajusten a sus necesidades, conformando nuevos reportes?	5		0	
	6. ¿El sistema responde de una forma rápida a la información que le sea solicitada por el usuario?	5		0	
	7. ¿El sistema refleja cualquier lógica del negocio para poder responder a preguntas específicas?	5		0	
crítico	8. ¿El sistema garantiza la confidencialidad y seguridad de acceso a los datos por rol de los usuarios?	5		0	
	9. ¿Los datos e información derivados del proceso de análisis realizado mediante la aplicación, apoyan la toma de decisiones en la Institución?	5		0	
crítico	10. ¿Los cambios en los datos se reflejan automáticamente en los reportes de forma instantánea?	5		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?	3		6	
crítico	2. ¿Se entiende claramente lo que se ha especificado en el documento?	5		0	
	3. ¿El número de página que aparece en el índice coincide con el contenido que se refleja realmente en dicha página?	5		0	

Glosario de términos.

ONE: Oficina Nacional de Estadística.

UCI: Universidad de las Ciencias Informáticas.

SIGOB: Sistema de Información de Gobierno.

INE: Instituto Nacional de Estadísticas.

MD: Mercado de Datos.

AD: Almacén de Datos.

UML: Lenguaje Unificado de Modelado.

OLTP: Procesamiento Transaccional en Línea.

OLAP: Procesamiento Analítico en Línea.

ROLAP: Procesamiento Analítico en Línea Relacional.

MOLAP: Procesamiento Analítico en Línea Multidimensional.

HOLAP: Procesamiento Analítico en Línea Híbrido.

Código abierto: término con el que se conoce al software distribuido y desarrollado libremente.

LGPL: Licencia Pública General Menor.

ETL: Extracción, Transformación y Carga.

PDI: Pentaho Data Integration.

BI: Inteligencia de Negocio.

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