Universidad de las Ciencias Informáticas

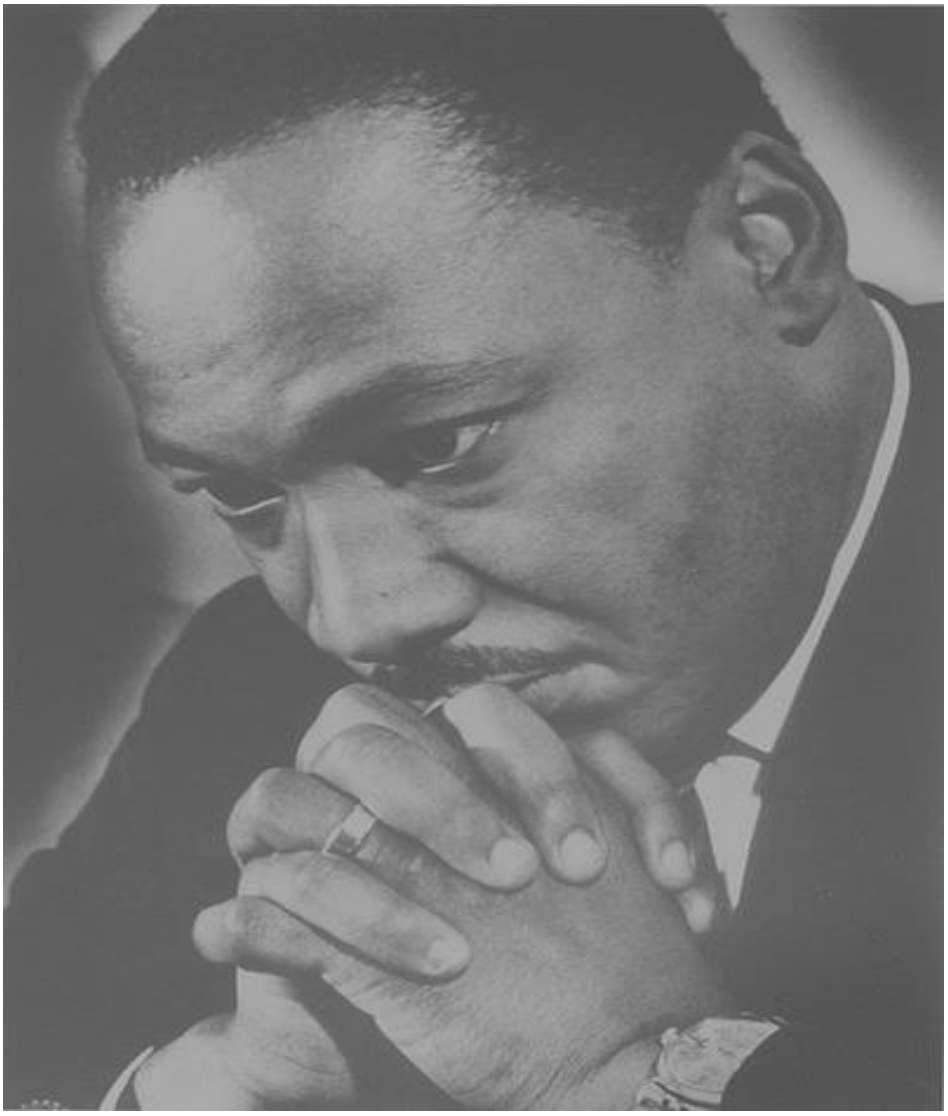
Facultad # 6

Almacén de datos para la Unión de Jóvenes Comunistas.

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

Autores: Angélica Vázquez Reinoso
Ronald Fernández Flores
Tutores: Ing. Yanet Cardoso García
Ing. Aldis Joan Abreu Medina

La Habana, julio 2016
"Año 58 de la Revolución"



"Un hombre no se mide por la posición que adopta en momentos de comodidad y conveniencia, sino por su posición en tiempos de desafío y controversia."

Martin Luther King, Jr.

DECLARACIÓN DE AUTORÍA

Declaramos ser autores de la presente tesis “Almacén de datos para la Unión de Jóvenes Comunistas” 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 _____.

_____	_____
Angélica Vázquez Reinoso	**Ronald Fernández Flores**
Firma del Autor	**Firma del Autor**
_____	_____
Ing. Yanet Cardoso García	**Ing. Aldis Joan Abreu Medina**
Firma del Tutor	**Firma del tutor**

DATOS DEL TUTOR

Nombre: Yanet Cardoso García

Profesión: Ingeniero en Ciencias Informáticas

Categoría docente: -

Año de graduado: 2013

Correo: ycardosog@uci.cu

DATOS DEL TUTOR

Nombre: Aldis Joan Abreu Medina

Profesión: Ingeniero en Ciencias Informáticas

Categoría docente: Instructor

Año de graduado: 2009

Correo: ajabreu@uci.cu

AGRADECIMIENTOS DE ANGELICA

A mis padres. A mi mamá porque es ese ser especial que me ha apoyado a lo largo de mi vida. Por complacerme en todo en cuanto pudo y por ser mí ejemplo a seguir. A mi padre por enseñarme de la vida y porque gracias a él se ha forjado mi carácter. A mi hermanita querida por ser especial y por quererme a su manera, espero que mis pasos te guíen a lo largo de tu carrera. A mi hermana Akemy que aunque la sangre no nos une, eres de las personas con las que siempre he contado. A mi amiga Nina por quererme tal y como soy, por ser esa niña pequeña con la cual convivo. A Mayi por esas fiestas que compartimos y esas deliciosas comidas. A Yoan, porque en estos meses te has convertido en mi cómplice, por hacerme sentir tan especial. A mi compañero de tesis Ronald por ser mi amigo, por los dolores de cabeza que me causaste y por los apuros de los cuales me sacaste. A mi tutora: Yanet te doy las gracias por el apoyo que me has brindado, por ser guía y compañera. Aldis aunque no compartimos desde el inicio gracias por el apoyo y la disposición. A mis compañeros de aula, esos con los que compartí cada curso: Yanitza, Lionel, Dayana, Elián. A mis nuevos amigos que me han acogido con cariño Danny, Mary, Dennis y Yunior. A las personas con las cuales compartí en estos años, a los que en su momento ocuparon un lugar especial y hoy son grandes amigos. Gracias a todos.

AGRADECIMIENTOS DE RONALD

Le agradezco a mi mamá, por ser esa madre ejemplar que me ha dado todo su apoyo y cariño en todos estos años de mi vida, mami te quiero. A mi papá, que a pesar de no poder estar presente hoy en día por trabajo …

DEDICATORIA DE ANGÉLICA

A mis padres por ser mi apoyo.

A mi hermanita Anélica, que mis pasos le sirvan de guía a lo largo de su vida.

A mi familia… A los que no están entre nosotros…tía, te lo prometí.

DEDICATORIA DE RONALD

Esta tesis va dedicada a mis padres, ya que son mi inspiración y los que me dan las fuerzas para seguir luchando por lo que quiero.

A mi hermano Rodney, para que esto le sirva de ejemplo a seguir y se haga otro profesional en la familia.

En fin a toda mi familia en su totalidad, puesto que cada uno puso su granito de arena, a lo largo de mi carrera…

RESUMEN

La Unión de Jóvenes Comunistas es la organización política de la juventud cubana, cantera y reserva combativa del Partido Comunista de Cuba. Esta organización necesita recopilar, describir y analizar un conjunto de datos estadísticos para mantener la integridad del país. Actualmente toda la información que se maneja es gestionada por el Sistema de Control de la Militancia para la Unión de Jóvenes Comunistas, desarrollado por el Centro de Tecnologías de Gestión de Datos de la Universidad de las Ciencias Informáticas. La organización cuenta con una base de datos relacional que almacena sus datos históricos, imposibilitando el análisis de tendencias que permita tomar las decisiones. La metodología utilizada fue la titulada Metodología de desarrollo para proyectos de Almacenes de datos. Se utilizaron para el desarrollo del sistema las herramientas PostgreSQL, Data Cleaner y la Suite Pentaho. La solución propuesta ofrece la posibilidad de consultar reportes operacionales, vistas de análisis y d*ashboard* para la visualización de la información relacionada con los militantes, niveles estructurales y procesos políticos que se manejan en la organización a través de un Almacén de datos, garantizando una mejor comprensión y análisis para la toma de decisiones.

Palabras clave: Almacén de datos, reportes operacionales, toma de decisiones, vistas de análisis, Unión de Jóvenes Comunistas.

Abstract

The Young Communist Union is the political organization of Cuban youth, pool and combative source of the Cuban Communist Party. This organization needs to analyze, collect and describe a set of statistic data to keep the integrity of the country. Today, all the information handled is managed by the Militancy Control System for the Young Communist Union developed by the Data Management Technology Center from the University of Informatics Sciences. Nowadays, the organization has a relational database that stores historical data, trend analysis impossible that allows making decisions. The methodology used was entitled Development Methodology for Data Warehouses projects. Free tools were also used for the development of the system. The solution proposed offers the possibility of consulting operational reports, analysis views, and dashboard to visualize the information related to the militants, structural levels, and political processes managed in the organization through a Data Warehouse, making possible better understanding and analysis for the decision making.

Keyswords: analysis views, Data Warehouse, decision making, operational reports, Young Communist Union.

ÍNDICE

ÍNDICE DE TABLAS

ÍNDICE DE FIGURAS

INTRODUCCIÓN

Algunos expertos comparan la era de la Revolución Industrial con la época que actualmente está viviendo la tecnología. Cada día se ven nuevos *software*, nuevos equipos, nuevas maneras de hacer las cosas y la organización que no esté preparada para estos cambios, que no tenga capacidad de información o que la misma sea muy débil simplemente no puede competir contra el resto de las organizaciones.

La toma de decisiones es una parte esencial en la existencia y supervivencia de las organizaciones, y por desgracia en muchas ocasiones no es una actividad sencilla y rápida. Según el paso de los años este proceso se vuelve más y más complejo, debido a la dinámica de las empresas y el ambiente en el que están inmersos, y conforme pase el tiempo se irá incrementando. De ahí surge la necesidad de estar siempre actualizados en cuanto a información relevante para la empresa.

Los sistemas con este fin se han convertido en un factor distinguido para las organizaciones a nivel mundial, mejorando la forma en que operan. Debido a la posibilidad que brindan los mismos de ejecutar una correcta gestión de la información.

Cuba no queda ajena al empleo de estas soluciones informáticas y ha puesto empeño en su uso, sobre todo en organizaciones que necesitan recopilar, describir y analizar un conjunto de datos estadísticos para mantener la integridad del país. En el desarrollo de estas aplicaciones para la gestión de la información, la Universidad de las Ciencias Informáticas (UCI) ha representado un papel primordial desde su creación en 2002 por el Comandante en Jefe Fidel Castro Ruz. El Centro de Tecnologías de Gestión de Datos (DATEC) es unos de los tantos centros productivos de la universidad que acoge a una parte de los profesores, estudiantes y especialistas de la Facultad 6, cuya especialidad es desarrollar tecnologías y proveer servicios relacionados con la gestión de datos y el análisis de información, los sistemas de información y los sistemas de inteligencia de negocios para apoyar los procesos de dirección de las organizaciones (Gespro, 2015).

La universidad ha desarrollado a lo largo de los años numerosos proyectos con empresas e instituciones cubanas y del extranjero, entre los cuales se encuentra la Unión de Jóvenes Comunistas (UJC) una organización de avanzada de la juventud, cantera y reserva combativa del Partido Comunista de Cuba (PCC) que tiene la responsabilidad de formar en sus filas a los jóvenes; para contribuir a la educación de las nuevas generaciones como constructores conscientes del socialismo.

Actualmente toda la información de la UJC es gestionada por el Sistema de Control de la Militancia para la UJC (SICOM-UJC), una aplicación desarrollada por DATEC. SICOM-UJC permite la recolección de la información de los militantes que integran sus filas, gestionar los procesos operativos en los distintos niveles de dirección de la organización y la consolidación de la información siguiendo el flujo establecido por la UJC.

A partir de la información generada por el sistema se evidencia que los datos muestran problemas de estandarización, ya que no figuran reglas que garanticen su homogeneidad. Además, se hace necesario por parte de los directivos la existencia de reportes dinámicos, es decir, reportes que puedan ser configurados y adecuados para obtener tendencias, que le sirvan de ayuda a tomar decisiones. Es necesario la utilización de gráficos, que permitan realizar el análisis de los datos en un lenguaje claro y sencillo de los principales resultados de la organización. Por otra parte, en la UJC Nacional no es posible realizar una comparación de la información existente en cada una de las provincias del país, imposibilitando el análisis estadístico, además necesitan de una fuente capaz de almacenar los datos históricos que en ella se manejan. A partir de la problemática planteada surge el siguiente **problema de investigación:** ¿Cómo apoyar a la toma de decisiones en la Unión de Jóvenes Comunistas?

Una vez analizado el problema se identifica como **objeto de estudio:** Almacén de datos, enmarcado en el **campo de acción:** Almacén de datos para la Unión de Jóvenes Comunistas. Se ha propuesto a su vez para dar respuesta al problema planteado como **objetivo general:** Desarrollar un Almacén de Datos para la Unión de Jóvenes Comunistas que apoye la toma de decisiones.

Para guiar la investigación se plantean las siguientes **preguntas científicas**:

- ✓ ¿Cuáles son los referentes teóricos relacionados con los almacenes de datos?
- ✓ ¿Cómo realizar el análisis y diseño del Almacén de datos para la UJC?
- ✓ ¿Cómo desarrollar el Almacén de datos para la UJC?
- ✓ ¿Cómo verificar que la implementación dio solución al problema planteado?

En el desarrollo del Almacén de datos para la UJC fueron identificadas las siguientes **tareas de investigación:**

- ✓ Definición del marco teórico-metodológico de la investigación y de las herramientas a utilizar, que permitan centrar las bases de la investigación.
- ✓ Levantamiento de requisitos, que permitan determinar las necesidades de información.
- ✓ Descripción de los casos de uso para especificar cada una de las funcionalidades del sistema.
- ✓ Definición de la arquitectura del Almacén de datos, que permita identificar los principales subsistemas que la componen.

- ✓ Definición de los hechos, las medidas y las dimensiones para determinar los elementos que forman parte del modelo lógico de datos.
- ✓ Diseño del modelo lógico de datos para así determinar los elementos que componen su modelo físico.
- ✓ Diseño del subsistema de integración como guía para la implementación de dicho subsistema.
- ✓ Diseño del subsistema de visualización, que permita realizar la capa de visualización.
- ✓ Implementación del modelo físico del Almacén de datos.
- ✓ Implementación del subsistema de integración para el poblado del Almacén de datos, cargando los hechos y las dimensiones correspondientes.
- ✓ Implementación del subsistema de visualización con el fin de obtener los reportes para los usuarios finales.
- ✓ Aplicación de las listas de chequeo para determinar que la estructura de los artefactos que corresponden a los procesos de Extracción, Transformación y Carga, tengan la calidad requerida.
- ✓ Aplicación de los casos de prueba para validar los reportes realizados.

En la investigación se utilizan métodos teóricos y empíricos, los cuales se definen a continuación:

Métodos Teóricos:

1. **Análisis Histórico – lógico:** este método está vinculado a la realización de un análisis de las distintas etapas lógicas sucesivas por las que han transcurrido los almacenes de datos para su desarrollo, así como de las tecnologías que existen actualmente para decidir cuáles emplear en el desarrollo de la aplicación.

2. **Modelación:** se emplea en la concepción de los diagramas establecidos por la metodología para el desarrollo del Almacén de datos.

3. **Analítico-sintético:** se emplea en el estudio de las fuentes bibliográficas que permiten la elaboración de la fundamentación teórica y en la síntesis de los conceptos asociados al dominio del problema.

Métodos Empíricos:

1. **Entrevista:** se emplea en la realización de entrevistas a los creadores del SICOM-UJC que estén directamente relacionados con el manejo de la información en el archivo técnico. De esta forma se

obtienen los datos necesarios de cómo se realizan estos procesos para su posterior informatización (Ver Anexo 2).

La presente investigación está dividida en tres capítulos. A continuación se expone un resumen de cada uno de ellos:

Capítulo 1: "Fundamentación teórica". En este capítulo se definen los conceptos fundamentales asociados al dominio del problema, se realiza un análisis del objeto de estudio, así como una caracterización de las soluciones existentes vinculadas al campo de acción, así como las tecnologías, metodología y herramientas a utilizar.

Capítulo 2: "Análisis y diseño del Almacén de Datos para la UJC". En este capítulo se realizará el estudio preliminar del negocio, se realizará el levantamiento de requisitos, se obtendrán las reglas de negocio, así como los hechos, dimensiones y medidas del almacén. Se hará el diseño de los subsistemas de almacenamiento, diseño general de las transformaciones y visualización.

Capítulo 3: "Implementación y pruebas del Almacén de Datos para la UJC". En este capítulo se implementarán los subsistemas de almacenamiento, transformación y visualización, y serán realizadas las pruebas al mismo a partir del diseño obtenido en el capítulo de Análisis y diseño del Almacén de Datos para la UJC.

CAPÍTULO I: Fundamentación teórica.

Introducción

En el presente capítulo se expone el marco teórico que sustenta la investigación realizada. Se relacionan los conceptos necesarios para el entendimiento de lo planteado en la situación problemática. Se analizan las herramientas, metodología y tecnologías necesarias que dan cumplimiento al objetivo general, justificando a su vez la selección y utilización de cada una de ellas.

1.1 Conceptos asociados a los almacenes de datos

Un concepto muy conocido relacionado con almacenes de datos es el planteado por William Harvey Inmon quien define que *"un almacén de datos es una colección de datos orientada al negocio, integrada, variante en el tiempo y no volátil para el soporte del proceso de toma de decisiones de la gerencia"* (Inmon, 2005). Mientras que Ralph Kimball por otra parte lo define como *"(…) una copia de las transacciones de datos específicamente estructurada para la consulta y el análisis; es la unión de todos los mercados de datos de una entidad"* (Kimball, y otros, 2002).

También es considerado una colección de datos orientada a un determinado ámbito (empresa, organización), integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza. Se trata, sobre todo, de un expediente completo de una organización, más allá de la información transaccional y operacional, almacenándolos en una base de datos diseñada para favorecer el análisis y la divulgación eficiente de datos (Juarez, 2011).

Una vez analizadas las definiciones antes expuestas se puede definir un Almacén de datos como una colección de datos que ayuda en la toma de decisiones; caracterizándose además por ser orientada a tema, integrada, variante en el tiempo y no volátil.

1.1.1 Características de los almacenes de datos.

Un Almacén de datos por lo general maneja un gran cúmulo de información. Entre sus principales características se destacan las que a continuación se presentan:

- ✓ **Orientado a tema:** Se basa en los aspectos que son de interés para la organización. Solo los datos necesarios para el proceso de generación del conocimiento del negocio se integran desde el entorno operacional (Fig. 1).

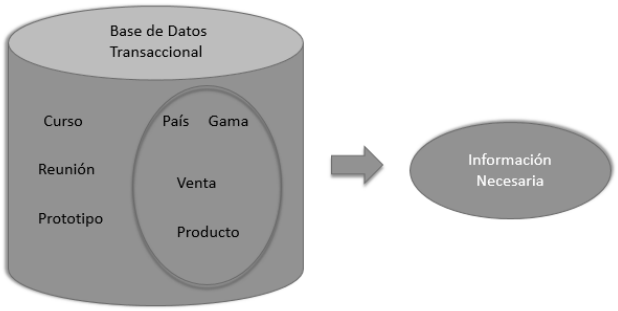


Fig. 1 Orientado a tema.

- ✓ **Integrado:** Integra la información que necesita ser almacenada en una única fuente de salida, aun cuando los sistemas operacionales de entrada almacenen los datos de manera diferente (Fig. 2).

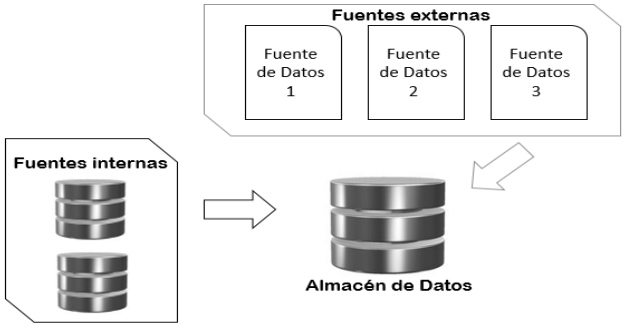


Fig. 2 Datos integrados.

- ✓ **Variante en el tiempo:** Los cambios producidos en los datos a lo largo del tiempo quedan registrados para ser consultados, dando solución a la problemática de los sistemas operacionales, que solo reflejan el estado actual del negocio (Fig. 3).

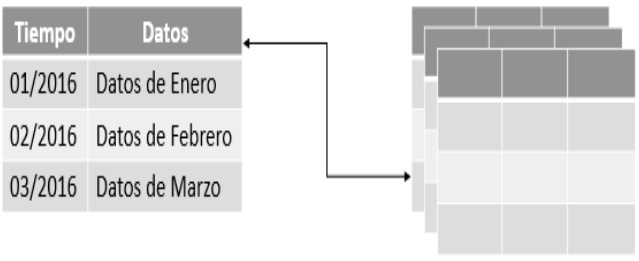


Fig. 3 Variante en el tiempo.

- ✓ **No volátil:** La información es útil para el análisis y la toma de decisiones solo cuando es estable (Bernabeu, 2010). La manipulación de la información se realiza mediante las operaciones de carga y acceso a los datos, no permitiendo la actualización de los mismos (Fig. 4).

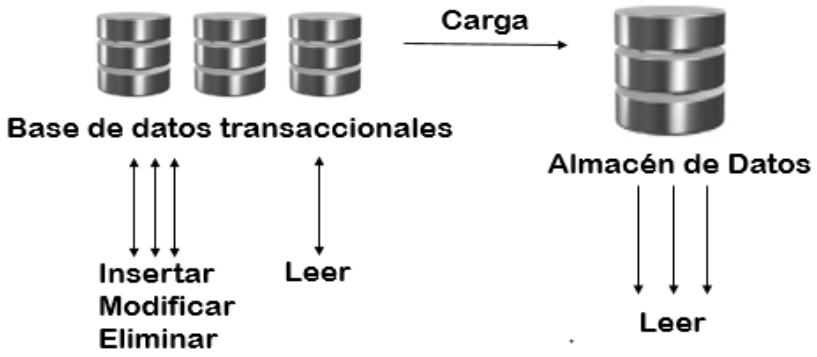


Fig. 4 No volátil.

1.1.2 Ventajas de los almacenes de datos.

Los almacenes de datos se han convertido para la mayoría de las organizaciones en una herramienta que ayuda al análisis de tendencias para una buena toma de decisiones, mostrando como principales ventajas:

- ✓ La integración y consolidación de diferentes fuentes de datos en una única plataforma sólida y centralizada.
- ✓ Los usuarios pueden acceder directamente a la información en línea, lo que contribuye a su capacidad para operar con mayor efectividad. Además, pueden tener a su disposición una gran cantidad de valiosa información multidimensional, presentada coherentemente como fuente única, confiable y disponible en sus estaciones de trabajo (HEFESTO, 2010).
- ✓ Permite la toma de decisiones estratégicas y tácticas mediante reportes, vistas de análisis, gráficos y cuadros de mando.
- ✓ La entrega de información a los usuarios va a ser completa, correcta, consistente, oportuna y accesible, en el momento adecuado.

1.2 Modo de almacenamiento de datos OLAP

La tecnología OLAP (por sus siglas en inglés de *On-Line Analytical Processing*, Procesamiento Analítico en Línea) permite a los usuarios analizar diferentes dimensiones de datos multidimensionales. Su objetivo

es agilizar la consulta de grandes cantidades de datos, lo que proporciona respuestas rápidas a consultas analíticas complejas e iterativas, utilizadas generalmente para sistemas que contribuyan a la toma de decisiones.

Existen tres modelos derivados de OLAP: ROLAP (por sus siglas en inglés de *Relational On-Line Analytical Processing*, Procesamiento Analítico Relacional en Línea), MOLAP (por sus siglas en inglés de *Multidimensional On-Line Analytical Processing*, Procesamiento Analítico Multidimensional en Línea) y HOLAP (por sus siglas en inglés de *Hybrid On-Line Analytical Processing*, Procesamiento Analítico Híbrido en Línea); su principal diferencia está dada por la forma de organizar y almacenar los datos.

ROLAP: los datos son almacenados en filas y columnas de forma relacional. Es soportado por Sistemas de Gestión de Bases de Datos Relacionales, mediante el uso de metadatos, evitando así la necesidad de crear una estructura de datos multidimensional estática. Este modelo presenta la información a los usuarios en forma de tablas de negocio. La principal ventaja de esta arquitectura es que permite el análisis de una gran cantidad de datos.

MOLAP: es una herramienta OLAP que accede a datos que no están almacenados en registros de tablas, sino que requiere un pre-procesamiento y almacenamiento de la información contenida en el cubo OLAP; el mismo almacena estos datos en una matriz multidimensional. Su principal premisa es que el MOLAP está mejor establecido almacenando los datos multidimensionalmente.

HOLAP: permite un análisis híbrido de la información, combinando atributos de los modos MOLAP y ROLAP. El análisis HOLAP ayuda a reducir costes de *hardware* ya que se necesita menos espacio en disco que en las bases de datos relacionales. Además, la respuesta de las consultas sobre las bases de datos multidimensionales son más rápidas que sobre las relacionales. Como aspecto negativo, los datos multidimensionales deben ser cargados antes de ser consultados y actualizados cuando se actualizan los datos de la organización.

En la investigación se decide utilizar como modo de almacenamiento ROLAP, pues se utiliza como Sistema de Gestión de Base de Datos PostgreSQL, que además de ser libre y poseer una serie de características, las cuales se detallan posteriormente, permite modelar bases de datos relacionales, no así multidimensionales. ROLAP accede directamente a los datos del almacén, soporta técnicas de optimización de accesos, tales como soporte a la desnormalización y uniones múltiples, para acelerar las consultas.

1.3 Metodologías para el desarrollo de almacenes de datos

Según la Real Academia metodología se le denomina al "*Conjunto de métodos que se siguen en una investigación científica o en una exposición doctrinal"* (Real Academia Española, 2016). Para el diseño de almacenes de datos Inmon propone una metodología *Top-Down* (descendente), mientras que Kimball por su parte propone el desarrollo de los *Data Marts* (mercados de datos departamentales) y el Almacén de datos sería la unión de ellos, denominada metodología *Bottom-Up* (ascendente).

Inmon propone la construcción de un repositorio de datos corporativo como fuente de información consistente, consolidada, histórica y de calidad. Como el Almacén de datos se construye descendentemente los mercados de datos se nutren del almacén corporativo, convirtiéndose en un complejo empresarial de base de datos relacionales (Hernández, 2010).

Metodología de desarrollo para proyectos de Almacenes de datos.

En el desarrollo del Almacén de datos para la UJC se utilizará la Metodología de desarrollo para proyectos de Almacenes de datos, la cual toma como base para su desarrollo el enfoque de Kimball. La misma define claramente los procesos y actividades que deben realizarse para el adecuado desarrollo de un Almacén de datos, además permite implementar las buenas prácticas definidas por *CMMI* (por sus siglas en inglés de *Capability Maturity Model Integration,* Integración de Modelos de Madurez de Capacidades) en su nivel 2. El ciclo de vida propuesto por la metodología es flexible y puede ser adaptado al ambiente de cualquier organización que desarrolle almacenes de datos (Hernández, 2010).

Dicha metodología propone siete fases de desarrollo y un flujo de trabajo, de los cuales solo se desarrollarán las cinco primeras fases. A continuación son expuestas las mismas:

- ✓ **Estudio preliminar y planeación:** Se realiza un estudio del ambiente que rodea al cliente, definiéndose además los objetivos del proyecto y otras actividades que ayudan a la planificación del proyecto.
- ✓ **Levantamiento de requisitos:** Se identifica la necesidad de información del cliente, se definen las reglas de negocio y se realiza un levantamiento de los datos a integrar para validar la disponibilidad de la información.
- ✓ **Arquitectura:** Se define la arquitectura de la solución, aspectos como la comunicación entre los subsistemas.

- ✓ **Diseño e Implementación:** Se diseñan e implementan los subsistemas que conforman la solución del problema (repositorio de datos, integración de datos, presentación de datos).
- ✓ **Prueba:** Se realizan las pruebas al sistema, necesarias para validar lo implementado y pactado con el cliente.

1.4 Herramienta de modelado a utilizar en el Almacén de datos para la UJC

UML (por sus siglas del inglés *Unified Modeling Lenguaje,* Lenguaje de Modelado Unificado) es un lenguaje que permite modelar, construir y documentar los elementos que forman un producto de *software* que responde a un enfoque orientado a objetos. Se ha convertido en el estándar internacional para definir, organizar y visualizar los elementos que configuran la arquitectura de una aplicación. Con este lenguaje, se pretende unificar las experiencias acumuladas sobre técnicas de modelado e incorporar las mejores prácticas actuales en un acercamiento estándar (Object Management Group, 1997).

Visual Paradigm 8.0 como herramienta de modelado

El nombre CASE proviene por sus siglas en inglés de *Computer-Aided Software Engineering* (Ingeniería de *Software* Asistida por Computadora), está diseñada para automatizar o apoyar una o más fases del ciclo de vida de desarrollo de un *software*. Visual Paradigm es una herramienta CASE para UML que propicia un conjunto de ayudas para el desarrollo de programas informáticos, desde la planificación, pasando por el análisis y el diseño, hasta la generación del código fuente de los programas y la documentación.

Características principales:

- ✓ *Software* libre.
- ✓ Disponibilidad en múltiples plataformas (*Windows*, *Linux*).
- ✓ Diseño centrado en casos de uso y enfocado al negocio.
- ✓ Uso de un lenguaje estándar común a todo el equipo de desarrollo que facilita la comunicación.
- ✓ Capacidades de ingeniería directa e inversa.
- ✓ Soporta aplicaciones *web*.
- ✓ Las imágenes y reportes generados no son de muy buena calidad.
- ✓ Soporta varios idiomas.
- ✓ Fácil de instalar y actualizar.
- ✓ Compatibilidad entre ediciones.
- ✓ Editor de Detalles de Casos de Uso.

- ✓ Generación de bases de datos.
- ✓ Generador de informes.

Ventajas de su utilización:

- ✓ Utiliza UML como lenguaje de modelado ofreciendo soluciones de *software* que permiten a las organizaciones desarrollar las aplicaciones con más calidad y más rápido.
- ✓ Es muy fácil de usar y presenta un ambiente gráfico agradable para el usuario.
- ✓ Permite aumentar la calidad del *software*, a través de la mejora de la productividad en el desarrollo y mantenimiento del mismo.
- ✓ Permite la reutilización del *software*, portabilidad y estandarización de la documentación, además del uso de las distintas metodologías propias de la Ingeniería.

Se decide utilizar esta herramienta porque permite modelar lógicamente la estructura del Almacén de datos para la UJC, generar el modelo físico a partir del modelo lógico y generar finalmente el script de base de datos para cargar el diseño en el sistema de gestión de bases de datos. Además de apoyar el ciclo de vida de desarrollo de la solución.

1.5 Sistema de Gestión de Bases de Datos

Un Sistema de Gestión de Bases de Datos (SGBD) es un programa de computador que permite la definición de bases de datos actuando de interfaz entre el usuario y las aplicaciones. Permite la elección de las estructuras de la información necesaria para su almacenamiento y búsqueda, ya sea de forma interactiva o a través de un lenguaje de programación. Se compone de un lenguaje de definición DDL (por sus siglas en inglés de *Data Definition Lenguage*, Lenguaje de Definición), de un DML (por sus siglas en inglés de *Data Manipulation Lenguaje*, Lenguaje de Manipulación) y de un SQL (por sus siglas en inglés de *Structured Query Lenguage*, Lenguaje de Consulta). Un SGBD permite definir los datos a distintos niveles de abstracción y manipularlos, garantizando la seguridad e integridad de los mismos (Bertino, et al., 1995).

Un SGBD facilita a los usuarios describir la información que será almacenada en la base de datos junto con un grupo de operaciones para manejarla, además de que permite a varios usuarios acceder a los datos de forma concurrente. Brindan un grupo de funciones con el objetivo de garantizar la confidencialidad, la calidad, la seguridad y la integridad de los datos que contienen, así como un acceso fácil y eficiente a los mismos.

Características:

- ✓ Abstracción de la información: ahorran a los usuarios detalles acerca del almacenamiento físico de los datos.
- ✓ Independencia: capacidad de modificar el esquema (físico o lógico) de una base de datos sin tener que realizar cambios en las aplicaciones que se sirven de ella.
- ✓ Redundancia mínima: un buen diseño de una base de datos logrará evitar la aparición de información repetida o redundante.
- ✓ Consistencia: en aquellos casos en los que no se ha logrado esta redundancia nula, será necesario vigilar que la información que aparece repetida se actualice de forma coherente, es decir, que todos los datos repetidos se actualicen de forma simultánea.
- ✓ Respaldo y recuperación: deben proporcionar una forma eficiente de realizar copias de respaldo de la información almacenada en ellos y de restaurar a partir de estas copias los datos que se hayan podido perder.
- ✓ Control de la concurrencia: En la mayoría de entornos lo más habitual es que sean muchas las personas que acceden a una base de datos, bien para recuperar información, bien para almacenarla. Es también frecuente que dichos accesos se realicen de forma simultánea. Un SGBD debe controlar este acceso concurrente a la información, que podría derivar en inconsistencias (Bertino, y otros, 1995).

PostgreSQL 9.4

PostgreSQL es un SGBD relacional, distribuido bajo licencia BSD (por sus siglas en inglés de *Berkeley Software Distribution*) y con su código fuente disponible libremente. Es el sistema de código abierto más potente del mercado y en sus últimas versiones no tiene nada que envidiarle a otras bases de datos comerciales.

Posee características significativas del motor de datos, entre las que se pueden incluir:

- ✓ Las subconsultas.
- ✓ Los valores por defecto.
- ✓ Las restricciones a valores en los campos (*constraints*).
- ✓ Los disparadores (*triggers*).

Ofrece funcionalidades en línea con el estándar SQL92, incluyendo claves primarias, identificadores entrecomillados, conversión de tipos y entrada de enteros binarios y hexadecimales. Debido a la liberación de la licencia, PostgreSQL se puede usar, modificar y distribuir de forma gratuita para cualquier fin, ya sea privado, comercial o académico (Postgres, 2010).

PgAdmin III 1.20.0

PgAdmin III es una aplicación gráfica para administrar el SGBD PostgreSQL, siendo la más completa y popular con licencia *Open Source*. Es capaz de gestionar versiones a partir de PostgreSQL 7.3 ejecutándose en cualquier plataforma. Está diseñado para responder a las necesidades de todos los usuarios, desde escribir consultas SQL simples, hasta desarrollar bases de datos complejas (PgAdmin, 2015). Entre sus principales características se tienen:

- ✓ Multiplataforma.
- ✓ Amplia documentación.
- ✓ Acceso a los datos.
- ✓ Acceso a todos los objetos de PostgreSQL (PostgreSQL, 2012).
- ✓ Diseñado para múltiples versiones de PostgreSQL.

Para la gestión de la base de datos se utilizará PostgreSQL en su versión 9.4 y para su administración se decide utilizar PgAdmin III 1.20.0 porque es un motor de base de datos de código abierto, multiplataforma y también funciona con otros motores comerciales basados en PostgreSQL. Se diseña para responder a las necesidades de la mayoría de los usuarios. La interfaz gráfica soporta todas las características de PostgreSQL y facilita la administración. Está disponible en más de una docena de lenguajes y para varios sistemas operativos, incluyendo Microsoft Windows, Linux, FreeBSD, Mac y OSX.

1.6 Herramienta para el perfilado de datos

El perfilado de datos es una de las primeras tareas que se suelen abordar en procesos calidad de datos, y consiste en realizar un primer análisis sobre los datos de origen, recopilar estadísticas e información sobre los mismos normalmente sobre tablas, con el objetivo de empezar a conocer su estructura, formato y nivel de calidad.

DataCleaner 3.1

DataCleaner es un motor de perfilado de datos, para descubrir y analizar la calidad de los datos. Encuentra los patrones, valores que faltan y otras características de los datos. El monitoreo es un aspecto central del DataCleaner para establecer el punto de partida, los objetivos, y para asegurar un proceso de

seguimiento de las cuestiones de calidad de datos. Además, brinda la posibilidad de tener un conocimiento general del estado de las fuentes (DataCleaner, 2012).

En el desarrollo del Almacén de Datos para la UJC se decide utilizar esta herramienta de perfilado de datos, la cual presenta las siguientes características:

- ✓ Consigue acceder a las bases de datos más utilizadas en el mercado, incluyendo Oracle, Microsoft SQL Server, MySQL, PostgreSQL, OpenOffice. Además, consigue interactuar con archivos en formato XML (por sus siglas en inglés de *Multidimensional Expressions*, Expresiones Multidimensionales) y planillas de Microsoft Excel.
- ✓ Código abierto y licencia de uso gratuita.
- ✓ Realiza la comparación de tablas, columnas y celdas con el fin de verificar la consistencia y veracidad de los datos.
- ✓ Garantiza la calidad de datos esenciales para el funcionamiento de la organización (DataCleaner, 2012).
- ✓ Averigua qué valores se presentan en mayor parte con el perfil de distribución de valores, así mismo como calcular la cantidad de valores nulos presentes en la fuente.
- ✓ Es una herramienta gráfica amigable y fácil de utilizar.

1.7 Suite Pentaho

La *Suite Pentaho* es una herramienta de BI (por sus siglas en inglés de *Business Intelligence*, Inteligencia de Negocio) desarrollada bajo la filosofía del *software* libre para la gestión y toma de decisiones empresariales. Cuenta con potentes capacidades para la gestión de procesos ETL (por sus siglas en inglés de *Extract-Transform-Load*, Extracción, Transformación y Carga de datos), creación de cuadros de mando para el usuario, informes interactivos, análisis multidimensionales de información (OLAP) o minería de datos. Todos estos servicios están integrados en una plataforma web, en la que el usuario puede consultar la información de una manera fácil e intuitiva (Pentaho Corporation, 2012).

1.7.1 Herramienta para el proceso de extracción, transformación y carga de datos

ETL es el proceso que organiza el flujo de los datos entre diferentes sistemas en una organización y aporta los métodos y herramientas necesarias para mover datos desde múltiples fuentes a un Almacén de datos, reformatearlos, limpiarlos y cargarlos en otra base de datos*, data marts* o bodega de datos. ETL

forma parte de la Inteligencia Empresarial también llamado “Gestión de los Datos” (*Data Management*) (ETL-Tools.Info, 2014).

Pentaho Data Integration 5.4

Desarrollado íntegramente en Java, posee LGPL (por sus siglas en inglés de *Lesser General Public Lincence,* Licencia Pública General Menor). Se utiliza para la integración de datos, carga de almacenes de datos y mercados de datos, limpieza de datos, análisis y perfilado de datos, migración de datos entre bases de datos y exportar datos de bases de datos a archivos planos y viceversa. Transforma e integra datos entre sistemas de información existentes y los mercados de datos que compondrán el sistema de inteligencia de negocio. Posee como principales características (Pierri, 2013):

- ✓ Entorno gráfico de desarrollo.
- ✓ Uso de tecnología 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 Data Integration está compuesto principalmente de las siguientes aplicaciones:

- ✓ SPOON: para diseñar transformaciones ETL usando un entorno gráfico.
- ✓ PAN: para ejecutar transformaciones diseñadas con SPOON.
- ✓ CHEF: permite diseñar la carga de datos incluyendo un control de estado de los trabajos.
- ✓ KITCHEN: permite ejecutar los trabajos diseñados con CHEF.

Para el desarrollo del Almacén de datos para la UJC fue seleccionada la versión 5.4 de Pentaho Data Integration, la cual incluye las siguientes ventajas:

- ✓ Reduce riesgos y costos de implementación.
- ✓ Permite probar de forma empírica y temprana la arquitectura de la aplicación de inteligencia de negocios.
- ✓ Soporta diferentes fuentes de información como son: Excel, PostgreSQL, MySql.

1.7.2 Herramienta para el proceso de inteligencia de negocio

Las herramientas de BI son usadas para acceder a los datos de los negocios y proporcionar reportes, análisis y visualizaciones a los usuarios. La gran mayoría son usadas por usuarios finales para acceder, analizar y reportar contra los datos que más frecuentemente residen en mercados de datos y almacenes de datos operacionales.

Entre sus características fundamentales se destacan:

- ✓ Poner a disposición de los usuarios la información necesaria para el análisis y la toma de decisiones.
- ✓ La información se obtiene sin dependencias de otros departamentos, con posibilidad de navegación *OLAP* por los propios usuarios, que permite profundizar en el análisis de forma interactiva en base a cualquiera de las dimensiones disponibles.
- ✓ Homogeneidad en la utilización de la información (interna y externa): utilización de la misma información al medir las cosas.
- ✓ Sistema soportado sobre plataformas tecnológicas sólidas y escalables (WorkMeter, 2010).

Pentaho Schema Workbench 3.12.0

Pentaho Schema Workbench es la herramienta gráfica que permite la construcción de los esquemas de *Mondrian*, y además permite publicarlos al *BI Server* para que puedan ser utilizados en los análisis por los usuarios de la plataforma (Softwarex, 2015).

Esta herramienta de la *suite Pentaho* tiene como objetivo facilitar el diseño de cubos *OLAP*. Su sencilla interfaz permite modelar un XML a través de opciones lógicas e intuitivas que no requieren de un manejo avanzado de este formato de archivo. Dentro de sus características se destacan:

- ✓ Diseñador intuitivo de esquemas *OLAP*.
- ✓ Permite crear, editar, actualizar y publicar esquemas *OLAP* para ser desplegados por aplicaciones de visualización de Pentaho.

Pentaho Report Designer 5.4

Pentaho Report Designer es una solución basada en el proyecto JFreeReports[1], permite generar informes de manera ágil es una herramienta para la creación de reportes operacionales y analíticos. Se puede utilizar para transformar los datos en información significativa, permitiendo la generación de informes en HTML, Excel, PDF, TXT, CSV y XML de manera dinámica.

Acepta datos de diferentes fuentes, incluyendo bases de datos SQL, fuentes de datos *OLAP*, e incluso la herramienta Pentaho Data Integration. El diseñador de reportes ofrece un entorno gráfico familiar, con herramientas intuitivas y fáciles de utilizar, y una estructura de reportes flexibles para darle libertad al diseñador de generar reportes que se adapten totalmente a su gusto y necesidad (Pentaho Community, 2015).

Pentaho BI Server 5.4

Esta herramienta suministra el soporte e infraestructura para crear soluciones de inteligencia de negocio. Proporciona servicios básicos además de incluir autenticación, motor de reglas, registro, auditoría y servicios *web*. Incorpora un motor de solución que integra reportes, análisis, tableros de comandos *(dashboard)* 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 mediante las publicaciones generadas por *Schema Workbench* y *Report Designer*, se puede usar como base para construir un sistema propio de inteligencia de negocios (SUMMAN, 2012).

Entre sus ventajas están:

- ✓ Aplicación Java2EE 100% extensible, adaptable y configurable.
- ✓ Administra y programa vistas de análisis.
- ✓ Administra seguridad de usuarios.
- ✓ Se integra con la mayoría de entornos y se puede comunicar con otras aplicaciones vía *webservices*.
- ✓ Permite la publicación de reportes operacionales y creación de *dashboard y* vistas de análisis.

[1] JFreeReports: es una librería de código abierto para la generación de reportes. Está desarrollada en el lenguaje Java.

Conclusiones del capítulo

Una vez analizado el estado del arte de los almacenes de datos con las principales características, se caracterizó la metodología, la tecnología y herramientas a utilizar en la solución. Para el proceso de la investigación se decidió utilizar la Metodología para el desarrollo de proyectos de Almacenes de datos ya que cubre las etapas por las que transita el desarrollo de un Almacén de datos y brinda diversas ventajas que facilitan su construcción; también se adapta a las tendencias de la universidad teniendo como base la metodología propuesta por Kimball. Se utiliza además como modo de almacenamiento ROLAP. Para la realización del modelado se decide utilizar el Visual Paradigm en su versión 8.0. El SGBD seleccionado es PostgreSQL en su versión 9.4 y como herramienta para la administración de los datos PgAdmin III en su versión 1.20.0. Para desarrollar el Almacén de datos para la UJC se utilizará DataCleaner en su versión 3.1 para el perfilado de datos, para el proceso de ETL se seleccionó la herramienta Pentaho Data Integration en su versión 5.4. El uso de Pentaho Schema Workbench en su versión 3.12.0. para la creación de los cubos OLAP, Pentaho Report Designer 5.4 para la implementación de reportes operacionales y Pentaho BI Server 5.4 para el desarrollo de las vistas de análisis y *dashboard*, herramientas que permitirán la implementación de la capa de visualización del Almacén de datos para la UJC.

CAPÍTULO II: Análisis y diseño del Almacén de datos para la UJC

Introducción

En este capítulo se realizará el análisis del negocio del Almacén de datos para la UJC. Se mostrará el diseño del mismo, especificando los requisitos de información, funcionales y no funcionales. Se confeccionará además el modelo de datos y la matriz bus, se identificarán los hechos y dimensiones del Almacén de datos para la UJC, así como el diseño de los subsistemas de almacenamiento, transformación y visualización.

2.1 Descripción del negocio

SICOM-UJC, es una herramienta implementada por DATEC que responde a las necesidades de la UJC, permitiendo la recolección de la información de los militantes que integran sus filas, automatizar los procesos operativos en los distintos niveles de dirección de la organización y la consolidación de la información siguiendo el flujo establecido por la UJC, utilizando mecanismos seguros para la transferencia de la misma. Además, permite la generación de resúmenes y reportes que se nutren de la información previamente gestionada en el sistema, generándose los datos estadísticos que son interpretados por la dirección de la organización como elementos esenciales para apoyarse en sus decisiones.

Teniendo en cuenta los requerimientos de la UJC, se hace necesario la creación de reportes dinámicos que puedan ser configurados y adecuados para obtener tendencias, así como la utilización de gráficos, que permitan realizar el análisis de los datos en un lenguaje claro y sencillo de los principales resultados de la organización. Por otra parte, es fundamental realizar comparaciones de la información existente en cada una de las provincias del país. Es por ello que se lleva a cabo el proceso de desarrollo del Almacén de datos para la UJC.

2.2 Especificación de requisitos

El análisis detallado de requisitos contribuye a una correcta elaboración del Almacén de datos, son definidos teniendo en cuenta las necesidades de la UJC y de su área de trabajo, de las cuales depende la implementación de la solución. Fueron identificados requisitos de información que responden a las necesidades del cliente para la toma de decisiones, funcionales que responden a las funciones del sistema y no funcionales, los cuales brindan las especificaciones de uso, finalidad que el sistema debe cumplir en su desarrollo y despliegue.

2.2.1 Requisitos de información

Los requisitos de información (RI) describen la información y los datos que un Almacén de datos debe proveer o debe acceder. Estos se definen a partir de las necesidades de información identificadas en el negocio, que permitan el análisis del comportamiento de los indicadores a medir según los objetivos y metas de la organización (Schiefer, 2002).

Los RI identificados durante el proceso de análisis del Almacén de datos para la UJC fueron clasificados por el tipo de información que brindan, ya sea militantes, niveles de organización y procesos políticos que se llevan a cabo. A continuación son expuestos los mismos:

RI1. Obtener la cantidad de altas por concepto, comité de base, comité provincial y comité municipal.
RI2. Obtener la cantidad de bajas por concepto y fecha de baja.
RI3. Obtener la cantidad de núcleos mixtos por fecha de activación, causa de activación y sector o rama de la economía.
RI4. Obtener la cantidad de sanciones por fecha de aprobación, provincia, municipio, comité de base, edad, tipo de sanción y causa de la sanción.
RI5. Obtener la cantidad de desactivaciones por fecha de aprobación, provincia, municipio, edad, tipo y causa de desactivación.
RI6. Obtener la cantidad de procesos de crecimiento por fecha de inicio, fecha de aprobación, comité de base, tipo comité, vía de ingreso, edad, sector o rama de la economía, clasificador ocupacional y ocupación laboral.
RI7. Obtener la cantidad de pendientes solucionados por fecha de solución, comité de base, fecha de recepción y causa.
RI8. Obtener la cantidad de pendientes solucionados como baja por fecha de recepción, fecha de baja, comité de base, comité provincial, comité municipal, causa y país.
RI9. Obtener la cantidad de pendientes por fecha de recepción, comité de base, comité provincial, comité municipal y causa.
RI10. Obtener la cantidad de organizaciones de base activas por fecha de activación, provincia, municipio, comité de base, tipo, causa de activación y sector o rama de la economía.
RI11. Obtener la cantidad de organizaciones de base desactivadas por fecha de desactivación, causa de desactivación, provincia, municipio, comité de base, tipo y sector o rama de la economía.
RI12. Obtener la cantidad de organizaciones de base por fecha de activación, sector o rama de la economía, provincia, municipio y tipo.

RI13. Obtener la cantidad de doble militantes por fecha de inicio en la UJC, fecha de inicio en el PCC, provincia, municipio, organizaciones de base, raza, sexo, edad, responsabilidad en la UJC, clasificador ocupacional y ocupación laboral.

RI14. Obtener la cantidad de militantes con cuotas atrasadas por fecha de inicio en la UJC, provincia, municipio, comité de base, nivel de enseñanza, nivel cultural, fecha de inicio laboral, números de meses que debe, sexo, raza, edad y ocupación laboral.

RI15. Obtener la cantidad de militantes en el exterior por fecha de inicio en la UJC, país al que viajó, motivo de salida, número de meses que debe, fecha de salida, edad, sexo, raza, provincia, municipio, nivel cultural y ocupación laboral.

RI16. Obtener la cantidad de militantes en grados terminales por fecha de inicio en la UJC, edad, sexo, raza, nivel de enseñanza, nivel cultural, provincia y municipio.

RI17. Obtener la cantidad de militantes por provincia, municipio, zona de trabajo, comité de base, clasificador ocupacional, ocupación laboral, nivel cultural, fecha de inicio en la UJC, fecha de alta, raza, sexo, edad y nivel de enseñanza.

2.2.2 Requisitos funcionales

Los requisitos funcionales (RF) describen lo que el sistema debe hacer. Estos dependen del tipo de *software* que se desarrolle, de los posibles usuarios del *software* y del enfoque general tomado por la organización (Sommerville, 2005). A continuación son enunciados los RF identificados:

RF1. Extraer datos de los sistemas fuentes.

RF2. Transformar y cargar datos de los sistemas fuentes.

RF3. Autenticar usuario.

RF4. Adicionar usuario.

RF5. Eliminar usuario.

RF6. Modificar usuario.

RF7. Mostrar usuario

RF8. Adicionar rol.

RF9. Eliminar rol.

RF10. Modificar rol.

RF11. Mostrar rol

RF12. Adicionar reporte.

RF13. Eliminar reporte.

RF14. Modificar reporte.

RF15. Mostrar reportes.

RF16. Visualizar reportes.

2.2.3 Requisitos no funcionales

Los requisitos no funcionales (RnF) son aquellos que no se refieren directamente a las funciones específicas que proporciona el sistema, sino a las propiedades emergentes de éste como la fiabilidad, el tiempo de respuesta y la capacidad de almacenamiento. De forma alternativa definen las restricciones del sistema como la capacidad de los dispositivos de entrada/salida y las representaciones de datos que se utilizan en las interfaces del sistema (Sommerville, 2005). A continuación son enunciados los RnF identificados:

Finalidad

RnF 1. Cumplir con las pautas de diseño de las interfaces: el sistema debe tener una interfaz gráfica uniforme que incluya pantallas, menús y opciones. Las pautas de diseño se realizarán siguiendo la arquitectura de información definida.

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

RnF 3. El acceso a los reportes del almacén será mediante la distribución de la información por áreas de análisis y libros de trabajos.

Confiabilidad

RnF 4. Asegurar la disponibilidad del sistema: el sistema debe estar disponible durante el horario de trabajo. En caso de fallo, la recuperación del servicio no deberá ser de un período de tiempo mayor de 30 minutos.

Restricciones de diseño

RnF 5. Utilizar la herramienta de modelado definida: Visual Paradigm 8.0.

RnF 6. Utilizar el Sistema de Gestión de Bases de Datos definido: PostgreSQL 9.4 y como interfaz de administración de dicho gestor PgAdmin III 1.20.0.

RnF 7. Utilizar la herramienta de integración de datos definida: Pentaho Data Integration 5.4.

RnF 8. Utilizar las herramientas de inteligencia de negocios definidas: *Pentaho Schema Workbench* 3.12.1, *Pentaho Report Designer* 5.4 y *Pentaho BI Server* 5.4.

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

Interfaz

RnF 9. El usuario deberá acceder a la aplicación mediante el protocolo HTTPS, usando preferiblemente el navegador *web Firefox* 32.0 o superior.

Ambiente

RnF 10. Las características de *hardware* y *software* requeridas para desplegar y utilizar la aplicación consisten en:

Servidor

- ✓ Microprocesador Core2Duo.
- ✓ Memoria RAM mínimo 2 GB.
- ✓ Disco duro mínimo 80 GB.
- ✓ Servidor de aplicaciones y bases de datos.

Cliente

- ✓ Memoria RAM mínimo 1 GB.

2.3 Reglas de negocio

Una vez definidos los requerimientos del Almacén de datos para la UJC son definidas las reglas de negocio (RN) que ayudarán a definir y controlar la estructura, el funcionamiento y la estrategia del mismo. Estas RN son clasificadas según la metodología como reglas de: almacenamiento, transformación y visualización, a continuación son expuestas:

Reglas de almacenamiento

RN1. Las cadenas de las dimensiones tendrán valor máximo de 100 caracteres.

RN2. Las cantidades tendrán como tipo de datos enteros.

Reglas de transformación

RN3. Si el militante no ha viajado el valor que tomará el país es CUBA.

RN4. Si el militante no debe meses de pago de cotización tomará valor 0.

RN5. La ocupación laboral será corregida ortográficamente y se mostrarán las cadenas en mayúsculas.

RN6. Si el militante no ha viajado el valor que tomará el motivo de salida es NO HA VIAJADO.

RN7. Si no es doble militante el valor de la fecha de inicio del PCC será el último valor de la dimensión tiempo (31 de diciembre de 2050).

RN8. Si el militante no ha viajado los valores de las fechas de salida y regreso serán el último valor de la dimensión tiempo (31 de diciembre de 2050).

Reglas de visualización

RN9. Las carpetas y reportes serán mostrados en mayúsculas.

RN10. Las cantidades estarán dadas por valores enteros positivos.

2.4 Caso de Uso del Sistema

Un Caso de Uso (CU) es una secuencia de interacciones que se desarrollarán entre un sistema y sus actores en respuesta a un evento que inicia un actor principal sobre el propio sistema. Los diagramas de casos de uso sirven para especificar la comunicación y el comportamiento de un sistema mediante su interacción con los usuarios y/u otros sistemas. O lo que es igual, un diagrama que muestra la relación entre los actores y los casos de uso del sistema (Sommerville, 2005).

2.4.1 Actores del sistema

Tabla 1. Descripción de los actores del sistema.

Actores	Descripción
Administrador	Usuario responsable de administrar los temas de seguridad del sistema gestionando los usuarios que interactuarán con la aplicación, así como los privilegios que puedan tener sobre los recursos del sistema.
Especialista	Responsable de inicializar los casos de uso relacionados con la visualización de los reportes del Almacén de datos para la UJC.
Administrador ETL	Responsable de llevar a cabo los procesos de extracción, transformación y carga de los datos.

2.4.2 Casos de uso de información

Los casos de uso de información describen los requisitos de información agrupados según el tema de análisis. Estos se encuentran de manera íntegra en el expediente de Proyecto de Almacén de datos para la UJC, en el artefacto de proyecto *"DATEC_Especificacion_de_casos_de_uso_AD_SICOM-UJC"*.

CUI1. Obtener reportes de las altas: muestra los reportes de los indicadores de las altas en la UJC.

CUI2. Obtener reportes de las bajas: muestra los reportes de los indicadores de las bajas en la UJC.

CUI3. Obtener reportes de los núcleos mixtos: muestra los reportes de los indicadores de los núcleos mixtos existentes en la UJC.

CUI4. Obtener reportes de las organizaciones de base: muestra los reportes de los indicadores de las organizaciones de base existentes en la UJC.

CUI5. Obtener reportes de los militantes: muestra los reportes de los indicadores de los militantes de la UJC.

CUI6. Obtener reportes de los pendientes: muestra los reportes de los indicadores de procesos y/o militantes pendientes de la UJC.

CUI7. Obtener reportes de los procesos: muestra los reportes de los indicadores de los procesos de crecimiento que se llevan a cabo en la UJC.

CUI8. Obtener reportes de las sanciones y desactivaciones: muestra los reportes de los indicadores de las sanciones y desactivaciones que se realizan en la UJC.

A continuación se muestra la descripción del caso de uso perteneciente al hecho núcleos mixtos (Tabla 2):

Tabla 2. Descripción del caso de uso Obtener reportes de los núcleos mixtos.

Objetivo	Mostrar información sobre los núcleos mixtos.	
Actores	Especialista.	
Resumen	El Caso de Uso inicia cuando el especialista desea consultar la información referente a los núcleos mixtos.	
Complejidad	Alta	
Prioridad	Alta	
Precondiciones	El especialista tiene que estar autenticado. El almacén tiene que estar poblado.	
Postcondiciones	Los reportes correspondientes fueron consultados por el especialista.	
Flujo de eventos		
Flujo básico Mostrar información sobre los Núcleos mixtos		
	Actor	**Sistema**
1	Selecciona Núcleos mixtos	
2.		Muestra los reportes asociados al tema de análisis Núcleos mixtos.
3.	Selecciona el Reporte	
4.		Visualiza el Reporte. Se brindan las opciones del usuario.
Opciones de Reportes		
	Perspectivas de análisis	**Posibles resultados (Medidas)**
Variables de entrada relacionadas con el CU: 1. fecha de activación, 2. causa de activación, 3. sector o rama de la economía		Variables de salida disponibles: 1. cantidad de núcleos mixtos
Prototipo de interfaz		
Relaciones	**CU incluidos**	No aplica.
	CU extendidos	No aplica.
Requisitos no funcionales		
Asuntos pendientes		

2.4.3 Casos de uso funcionales

Los casos de usos funcionales se basan en la gestión de los roles, reportes, usuarios y la autenticación de los usuarios, además de las consultas en la base de datos.

CU1. Autenticar usuario: permite al usuario entrar en el sistema según los permisos que posea en el mismo.
CU2. **Gestionar usuario:** agrupa los requisitos funcionales de Insertar usuario, Modificar usuario, Mostrar usuario y Eliminar usuario.

CU3. Gestionar reporte: agrupa los requisitos funcionales de Insertar reporte, Modificar reporte, Mostrar reporte y Eliminar reporte.
CU4. Gestionar rol: agrupa los requisitos funcionales de Insertar rol, Modificar rol, Mostrar rol y Eliminar rol.
CU5. Extraer datos de la fuente: responde el requisito donde se extraen los datos de la fuente.
CU6. Realizar transformación y carga datos: agrupa los requisitos de transformar los datos extraídos y cargar los datos en el almacén.
CU7. Visualizar reporte: permite al usuario visualizar los reportes.

2.4.4 Diagrama de casos de uso

El diagrama de casos de uso representa la forma en como un Cliente (Actor) interactúa con el sistema en desarrollo, así como la forma en que los elementos se relacionan (Fig. 5). La descripción de cada caso de uso se encuentran de manera íntegra en el expediente de Proyecto de Almacén de datos para la UJC, en el artefacto de proyecto *"DATEC_Especificacion_de_casos_de_uso_AD_SICOM-UJC"*. En la Tabla 2 se ejemplifica la descripción del caso de uso correspondiente a obtener los reportes de los Núcleos mixtos.

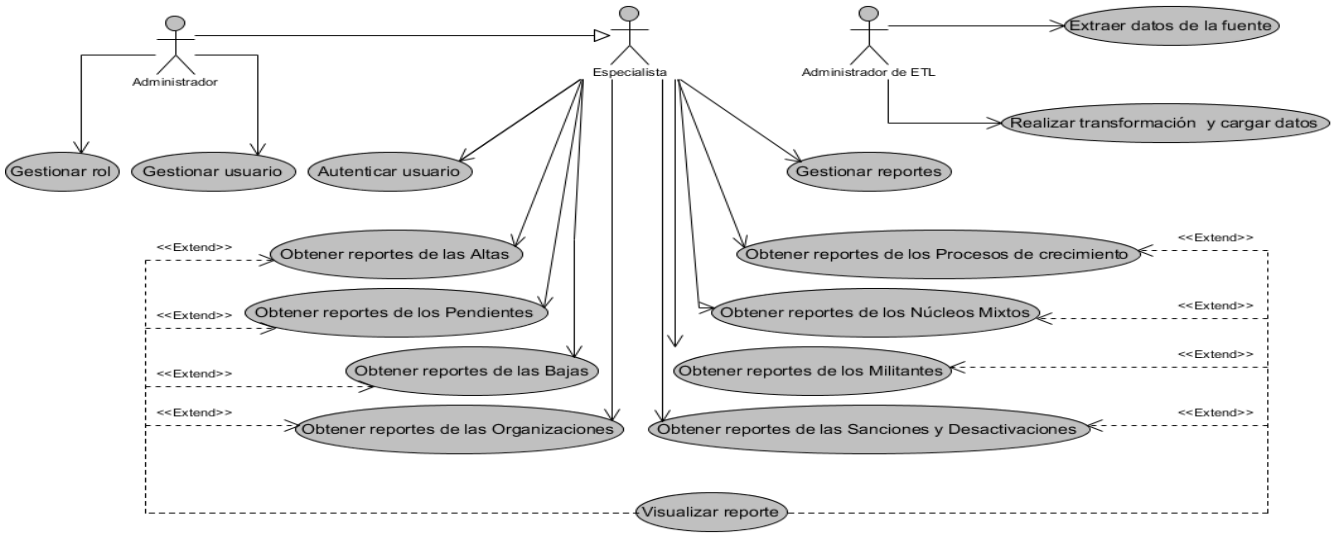


Fig. 5 Diagrama de casos de uso del sistema.

2.5 Definición de la arquitectura base del Almacén de datos

El Almacén de datos para la UJC está ordenado mediante una arquitectura compuesta por la fuente de datos, el Subsistema de integración, Subsistema de almacenamiento y Subsistema de visualización. A continuación una descripción de cada uno de estos subsistemas.

- ✓ La fuente de datos estará compuesta por la Base de datos generada por SICOM-UJC.
- ✓ En el subsistema de integración es donde se realizan todos los procesos ETL en los cuales se extraen, se limpian e integran los datos almacenados en los sistemas fuentes a través de transformaciones y trabajos.
- ✓ En el subsistema de almacenamiento es donde se guarda toda la información que ha sido transformada en el subsistema de integración.
- ✓ En el subsistema de visualización es donde se muestra toda la información almacenada al cliente, a través de vistas de análisis, reportes operacionales y *dashboard*. Los mismos permiten al cliente realizar un análisis de toda la información procesada.

La arquitectura del Almacén de datos para la UJC queda diseñada de la siguiente manera (Fig. 6).

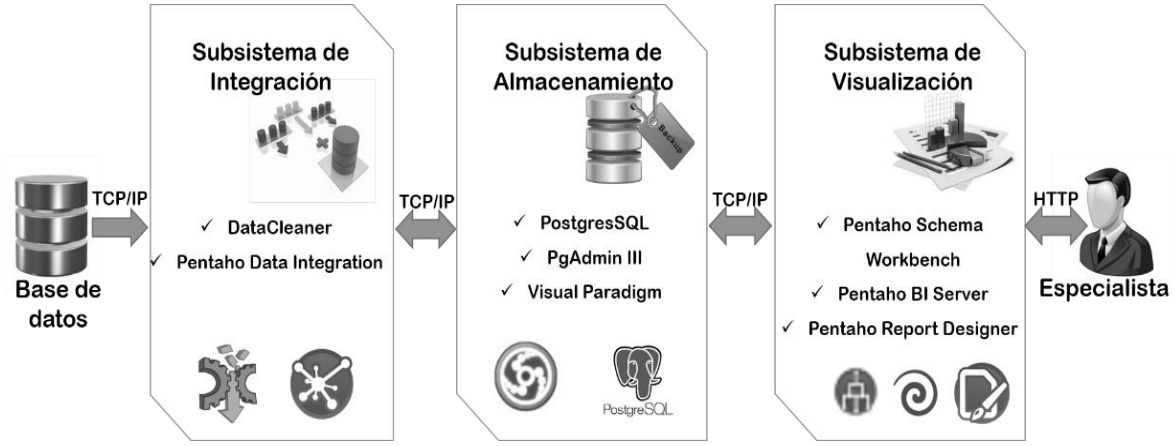


Fig. 6 Arquitectura del Almacén de datos para la UJC.

Primeramente se encuentra el Subsistema de integración que obtiene la información de la base de datos sicom_ujc; luego el subsistema de almacenamiento recibe los datos integrados y los almacena en la base de datos destino, la cual es soportada por el SGBD PostgreSQL y es administrada por los usuarios que tienen acceso a la información mediante pgAdmin III. Finalmente en el Subsistema de visualización se muestra la información a través de vistas de análisis mediante la herramienta Pentaho BI Server con el *framework* CDE y a través de reportes operacionales mediante la herramienta Pentaho Report Designer, la cual es accedida por los usuarios autorizados a la visualización de las vistas de análisis, reportes operacionales y gráficas, quienes se comunican mediante el protocolo HTTP[2] (por sus siglas en inglés de *Hypertext Transfer Protocol*, Protocolo de Transferencia de Hipertexto).

2.6 Diseño del subsistema de almacenamiento

El diseño del subsistema de almacenamiento comprende la identificación de dimensiones, hechos y medidas del Almacén de datos, así como la elaboración de la matriz dimensional y la selección de la topología a utilizar en el diseño de la solución.

2.6.1 Dimensiones

Las tablas de dimensiones definen como están los datos organizados lógicamente y proveen el medio para analizar el contexto del negocio. Contienen datos cualitativos. Representan los aspectos de interés, mediante los cuales los usuarios podrán filtrar y manipular la información almacenada en la tabla de hechos.

[2] HTTP: es el protocolo de comunicación que permite las transferencias de información en la web.

1. **Dimensión causa** (dim_causa): en esta dimensión se almacenan las causas por las que un militante es considerado pendiente.

2. **Dimensión causa de activación** (dim_causa_activacion): en esta dimensión se almacenan las causas por las cuales pueden ser activadas las organizaciones en la UJC.

3. **Dimensión causa de desactivación** (dim_causa_desactivacion): en esta dimensión se almacenan las causas por las cuales pueden ser desactivadas las organizaciones en la UJC **.**

4. **Dimensión causa de desactivación del militante** (dim_causa_desactivacion_militante): en esta dimensión se almacenan las causas por las cuales pueden ser desactivados los militantes en la UJC.

5. **Dimensión causa de sanciones** (dim_causa_sanciones): en esta dimensión se almacenan las causas por las cuales pueden ser sancionados los militantes en la UJC.

6. **Dimensión clasificador ocupacional** (dim_clasificador_ocupacional): en esta dimensión se almacena la clasificación ocupacional de los militantes por niveles.

7. **Dimensión edad** (dim_edad): en esta dimensión se almacenan las posibles edades de los militantes divididas además en rangos.

8. **Dimensión nivel cultural** (dim_nivel_cultural): en esta dimensión se almacenan los posibles niveles culturales que pueden tener los militantes.

9. **Dimensión nivel de enseñanza** (dim_nivel_enseñanza): en esta dimensión se almacenan los posibles niveles de enseñanza que pueden tener los militantes.

10. **Dimensión raza** (dim_raza): en esta dimensión se almacenan los valores de las razas.

11. **Dimensión sector o rama de la economía** (dim_sector_rama): en esta dimensión se almacenan los sectores de la economía.

12. **Dimensión tiempo** (dim_tiempo): en esta dimensión se almacenan los años, meses y días.

13. **Dimensión concepto** (dim_concepto): en esta dimensión se almacenan los conceptos asociados a las altas y bajas.

14. **Dimensión sexo** (dim_sexo): esta dimensión almacena el sexo.

15. **Dimensión responsabilidad en la UJC** (dim_responsabilidad): en esta dimensión se guardan todas las responsabilidades que puede ocupar un militante en la UJC.

16. **Dimensión DPA** (dim_dpa): en esta dimensión se guardan los niveles de la división política administrativa del país.

17. **Dimensión ocupación laboral** (dim_ocupacion_laboral): en esta dimensión se almacenan las ocupaciones laborales de los militantes.

18. **Dimensión interés personal** (dim_interes_personal): en esta dimensión se almacenan los posibles motivos del militante que viaja al exterior.

19. **Dimensión meses que debe** (dim_meses_debe): en esta dimensión se almacenan los meses que debe un militante sin pagar la cotización.

20. **Dimensión niveles de la UJC** (dim_nivel_ujc): en esta dimensión se almacenan los niveles de la UJC.

21. **Dimensión vía de ingreso** (dim_via_ingreso): en esta dimensión se almacenan los datos de las posibles vías de ingreso.

22. **Dimensión tipo de desactivación del militante** (dim_tipo_desactivacion_militante): en esta dimensión se almacena el tipo de desactivación que se le hizo al militante.

23. **Dimensión tipo de sanciones** (dim_tipo_sanciones): en esta dimensión se almacenan los tipos de sanciones.

24. **Dimensión país** (dim_pais): en esta dimensión se almacenan los posibles países a los que viaja un militante.

2.6.2 Hechos y medidas

Las tablas de hechos son las tablas primarias en el modelo dimensional. Generalmente, almacenan medidas numéricas, las que representan valores de las dimensiones. La llave de la tabla de hecho, es una llave compuesta, debido a que se forma de la composición de las llaves primarias de las tablas dimensionales a las que están unidas. Las tablas de hechos contienen, precisamente, los hechos que serán utilizados por los analistas de negocio para apoyar el proceso de toma de decisiones. Contienen datos cuantitativos.

1. **Hecho altas** (hech_altas): en este hecho se recogen las personas que se le dan de alta.

- Cantidad_altas

2. **Hecho bajas** (hech_bajas): en este hecho se recogen las personas que se le dan de baja.

- Cantidad_bajas

3. **Hecho militantes** (hech_militantes): en este hecho se recogen los datos de los militantes.

- Cantidad_militantes

4. **Hecho núcleos mixtos** (hech_núcleos_mixtos): en este hecho se recogen los datos referentes a los núcleos mixtos existentes en la UJC.

- Cantidad_nucleos_mixtos

5. **Hecho organizaciones de base** (hech_organizaciones_base): en este hecho se recogen los datos referentes a las organizaciones de base existente en la UJC.

- Cantidad_organizaciones_base

6. **Hecho pendientes** (hech_pendientes): en este hecho se recogen los militantes pendientes.

- Cantidad_pendientes

7. **Hecho procesos de crecimiento** (hech_procesos): en este hecho se recogen todos los procesos de crecimiento aplicados en la UJC.

- Cantidad_procesos

8. **Hecho sanciones y desactivaciones** (hech_sanciones_descativaciones):en este hecho se recogen todas las sanciones y desactivaciones que se le hacen a los militantes.

- Cantidad_sanciones
- Cantidad_desactivaciones

2.6.3 Matriz bus o matriz dimensional

La matriz bus es la herramienta esencial para el diseño y la comunicación de la arquitectura de un Almacén de datos. Las filas de la matriz son dimensiones y las columnas son hechos. Las celdas marcadas con X de la matriz indican si una dimensión se asocia con un hecho dado. El equipo de diseño escanea cada fila para probar si una dimensión candidata es bien definida para el hecho y también analiza cada columna para ver donde una dimensión debe ser conformada a través de múltiples hechos.

Tabla 3 Matriz dimensional

Dimensiones	Hechos							
	altas	bajas	militante	núcleos mixtos	organizaciones base	pendientes	procesos	sanciones desactivaciones
causa						X		
causa de activación				X	X			
causa de desactivación					X			
causa de desactivación militante								X
causas sanciones								X
clasificador ocupacional			X				X	
concepto	X	X						
dpa			X		X			X
edad			X				X	X
interés personal			X					
meses debe			X					
nivel cultural			X					
nivel enseñanza			X					
nivel UJC	X		X		X	X	X	
ocupación laboral			X				X	
país			X			X		
raza			X					
responsabilidad			X					
sector o rama				X	X		X	
sexo			X					
tiempo	X	X	X	X	X	X	X	X
tipo desactivación								X
tipo sanciones								X
vía de ingreso							X	

2.6.4 Topología

Esquema de estrella*:* es un tipo de esquema de base de datos relacional que consta de una sola tabla de hechos central rodeada de tablas de dimensiones (Fig. 7).

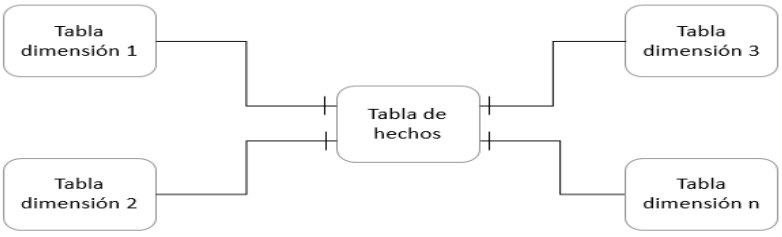


Fig. 7 Esquema de estrella.

Esquema copo de nieve: consta de una tabla de hechos que está conectada a muchas tablas de dimensiones, que pueden estar conectadas a otras tablas de dimensiones a través de una relación de muchos a uno. Las tablas de un esquema de copo de nieve generalmente se normalizan en tercera forma normal. Cada tabla de dimensiones representa exactamente un nivel en una jerarquía (Fig. 8).

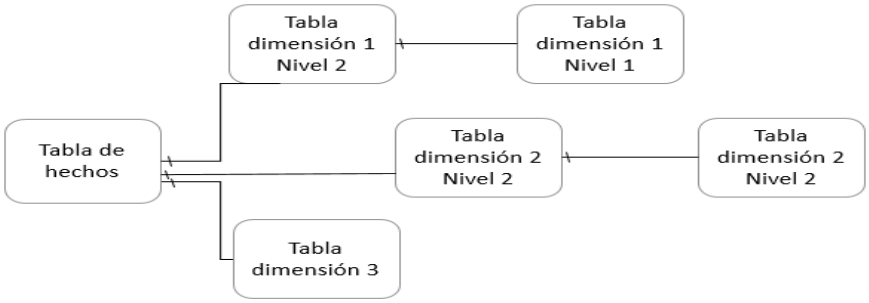


Fig. 8 Esquema copo de nieve.

Esquema constelación de hechos: son aquellos en los que solo algunas de las tablas de dimensiones se han desnormalizado. El objetivo es aprovechar las ventajas de los esquemas de estrella y copo de nieve. Las jerarquías de los esquemas de estrella están desnormalizadas, mientras que las jerarquías de los esquemas de copo de nieve están normalizadas. Estos están normalizados para eliminar las redundancias de las dimensiones. Para normalizar el esquema, las jerarquías dimensionales compartidas se colocan en estabilizadores (outrigger).

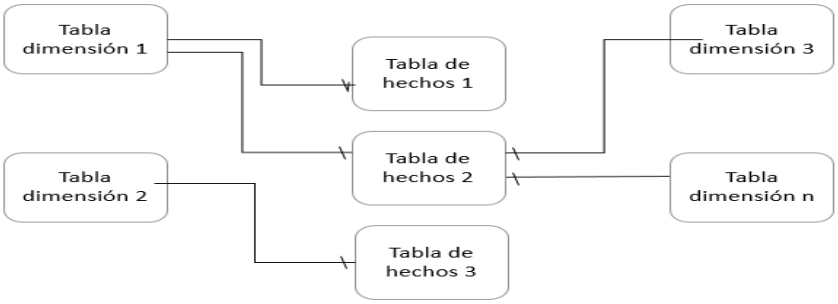


Fig. 9 Esquema constelación de hechos.

2.7 Modelo de datos

Según *Dittrich "los modelos de datos son un conjunto de herramientas conceptuales para describir la representación de la información en términos de datos comprenden aspectos relacionados con: estructuras y tipos de datos, operaciones y restricciones"* (Universidad de Sevilla, 2005).

Un modelo de datos se puede utilizar para describir un conjunto de datos y las operaciones para manipularlos. En el desarrollo del Almacén de datos para la UJC se realizó el modelo datos aplicando el esquema constelación de hechos para el mejor manejo de la información, obteniéndose 24 tablas de

dimensiones y 8 tablas de hechos a las que están asociados 9 medidas. A continuación se presenta un fragmento del mismo (Fig. 10), se encuentra de forma completa en el Anexo 1.

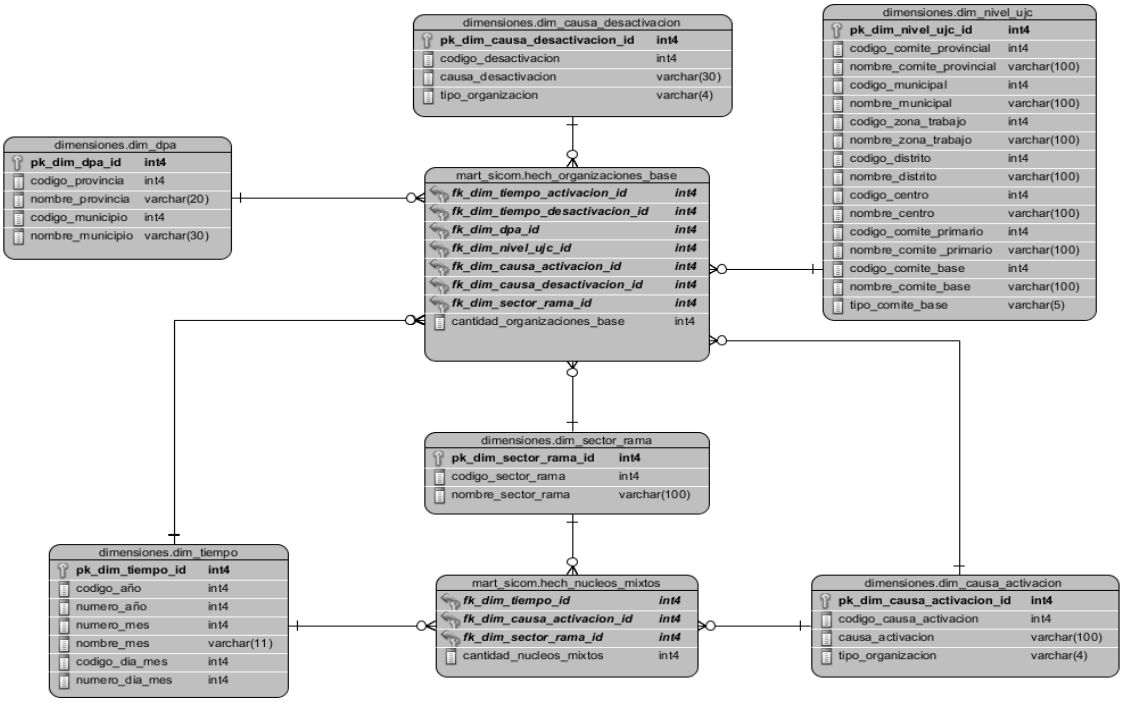


Fig. 10 Fragmento del Modelo de Datos.

2.7.1 Diseño del subsistema de integración

El subsistema de integración comprende el perfilado de los datos, que permite realizar un análisis profundo de los datos provenientes de la fuente para conocer el estado en que se encuentran, así como su calidad y estructura; y la extracción de los mismos desde los sistemas fuentes, los cuales sufren un conjunto de transformaciones. Conforma estos de manera que fuentes separadas puedan ser aprovechadas conjuntamente, y finalmente hace su entrega en un formato listo para el almacenamiento. El perfilado, el diccionario de datos y el diseño de las transformaciones constituyen elementos esenciales para lograr el diseño del subsistema de integración.

Perfilado de datos

El perfilado de datos consiste en realizar un primer análisis sobre los datos de origen, recopilar estadísticas e información sobre los mismos, normalmente sobre tablas, con el objetivo de empezar a conocer su estructura, formato y nivel de calidad.

Una vez realizado el perfilado de datos con la herramienta DataCleaner a la fuente de datos de SICOM-UJC se identificó que la mayoría de los tipos de datos son cadenas y fechas, estas últimas entre 2006 y 2014 con el formato aaaa-mm-dd. Se analizaron todos los campos de la fuente, detectándose errores descritos en el artefacto *"DATEC_Perfilado_de_datos_AD_SICOM-UJC.doc*", entre los cuales se encuentran valores nulos. También se muestran de los diferentes campos: la longitud de la cadena, la cantidad de valores nulos o negativos y los valores mínimos y máximos.

La Figura (Fig. 11) muestra la distribución de los datos en la base de datos de SICOM-UJC donde el 8% de los datos son de tipo fecha y el restante 92% de tipo cadena de un total de 109 columnas analizadas, además se encontraron un total de 4502 valores nulos correspondientes en su totalidad a la tabla de datos generales de los militantes (m_datos_generales).

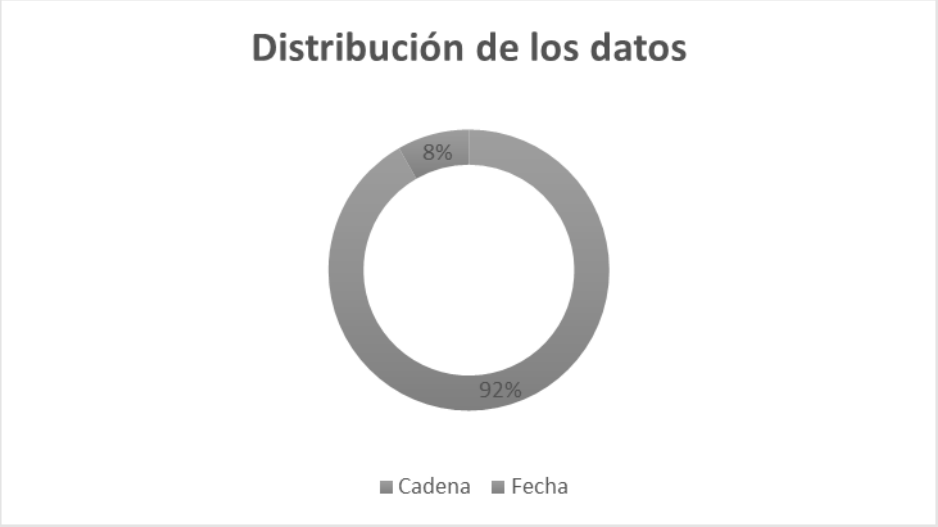


Fig. 11 Distribución de los tipos de datos.

Diseño general de las transformaciones

Las transformaciones son el elemento fundamental del proceso de ETL, en el diseño de las mismas se detalla cada uno de los pasos a seguir para efectuar la carga de las dimensiones y los hechos en el Almacén de datos para la UJC.

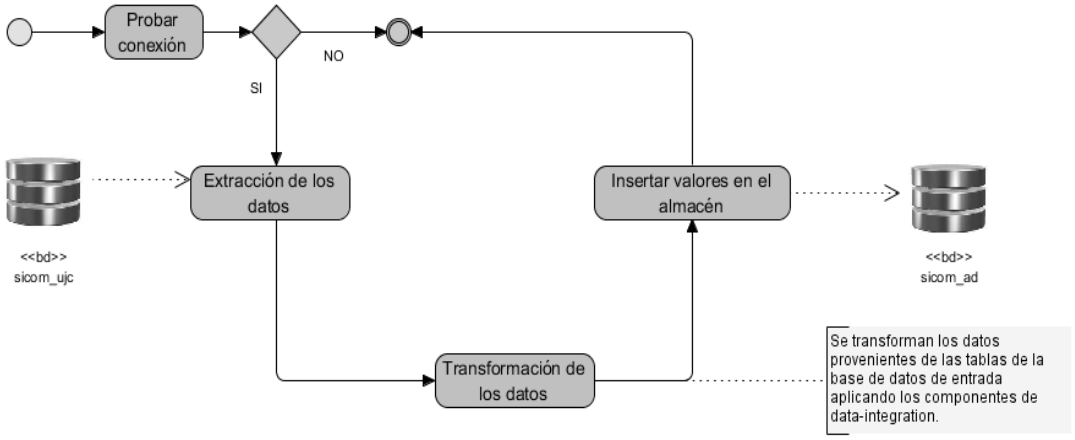


Fig. 12 Diseño general de las transformaciones de las dimensiones.

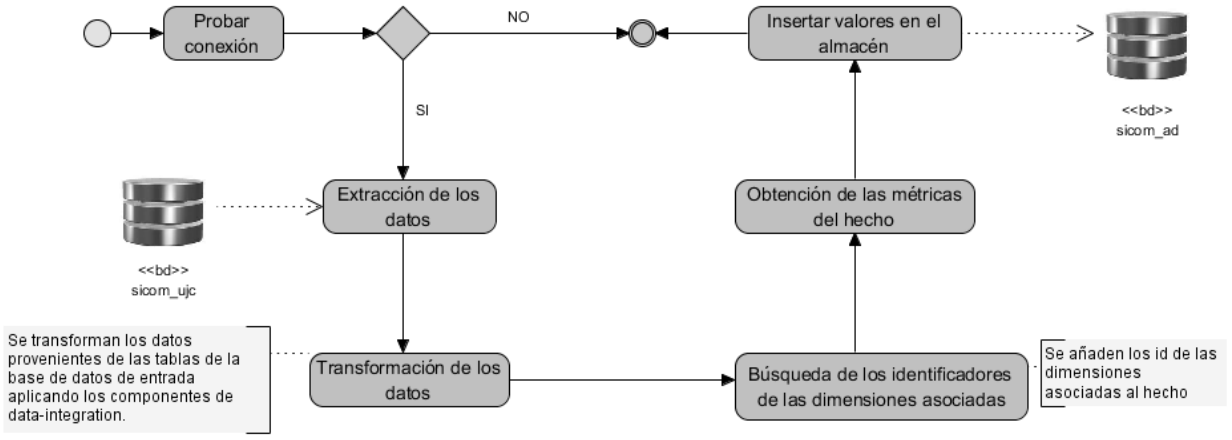


Fig. 13 Diseño general de las transformaciones de los hechos.

2.8 Diseño del subsistema de visualización

El diseño del Subsistema de visualización se realiza con el objetivo de organizar las vistas de análisis, reportes y *dashboard* por áreas de análisis, facilitando al usuario una búsqueda rápida de la información. Comprende la realización de los cubos *OLAP*, además de los distintos reportes que contribuyan a la toma de decisiones de los usuarios finales.

2.8.1 Arquitectura de la Información

La arquitectura de la información o mapa de navegación, se compone según las necesidades de los usuarios finales, por el Área de Análisis General (A.A.G) SICOM-UJC, las Áreas de Análisis (A.A) Gestión de la Información y Procesos Políticos donde están comprendidos cinco Libros de Trabajo (L.T) que incluyen los 30 reportes que conforman el Almacén de datos para la UJC(Fig. 14).

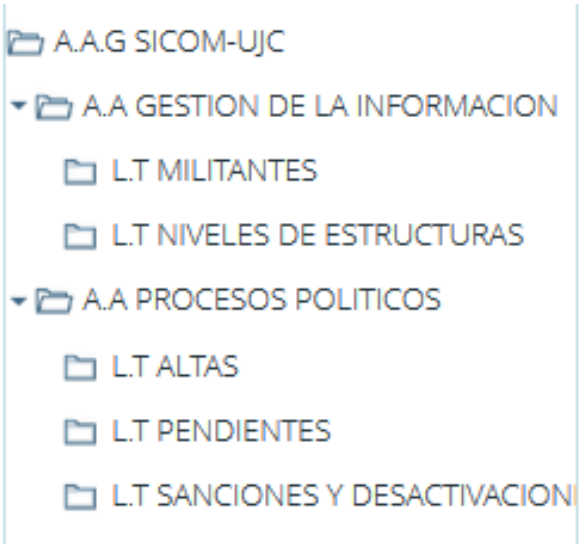


Fig. 14 Arquitectura de la información.

2.8.2 Diseño de las vistas de Análisis

Tienen como objetivo agrupar determinada información estadística que servirá de apoyo al proceso de toma de decisiones, permitiendo visualizar las distintas formas de análisis e interpretación de los datos desglosando la información hasta un nivel mínimo de detalle. Para el desarrollo del Almacén de datos para la UJC se identificaron 8 vistas de análisis para hacer más fácil el empleo de la aplicación por parte del usuario final, haciendo correspondencia con cada uno de los hechos del propio Almacén de datos.

2.8.3 Diseño de los cubos OLAP

En el Almacén de datos para la UJC se definieron 8 cubos multidimensionales desarrollados en la herramienta *Schema Workbench* correspondientes a cada uno de los hechos identificados, además se definieron 18 dimensiones compartidas y 6 que guardan relación con un solo hecho. La figura muestra el diseño correspondiente al cubo del hecho núcleos mixtos (Fig. 15).

Fig. 15 Diseño de los cubos multidimensionales.

2.9 Política de respaldo y recuperación

Con el objetivo de garantizar la persistencia de la información, se establece una política de respaldo y recuperación basada fundamentalmente en las copias de seguridad. Debido a que el sistema posee una carga histórica, que solo se cargará una vez y no se tiene definido por parte del cliente carga incremental, se realizarán copias de seguridad a la base de datos para salvaguardar los mismos. De igual forma en caso de perder la información se pueden volver a ejecutar las trasformaciones y se obtiene nuevamente la base de datos poblada. También se propone salvar la información en algún dispositivo de almacenamiento, ya que es de gran importancia mantener su disponibilidad y seguridad.

2.9.1 Esquema de seguridad

En el Almacén de datos para la UJC es necesaria la seguridad de la información, pues los datos que maneja son de vital importancia para el país. Con este objetivo se definieron roles para darle permisos a cada uno de los usuarios que interactúan directamente con el sistema.

Seguridad en el Subsistema de Almacenamiento

Para la seguridad de la base de datos se creó el rol Administrador de base de datos (BD) el que posee acceso total a la BD y el Administrador de ETL que se encarga de los procesos de extracción transformación y carga. En la Tabla 4 se muestra el esquema de seguridad definido para el Subsistema de Almacenamiento.

Tabla 4. Esquema de Seguridad definido para el Subsistema de Almacenamiento

Rol	Permisos
Administrador de BD	Total acceso a la BD. Realiza la administración de la BD, que contiene todas las tablas de hechos y dimensiones del almacén para la UJC. Autoriza permiso a cada uno de los usuarios.
Administrador de ETL	Realiza los procesos de extracción, transformación y carga de los datos. Tiene todos los permisos sobre las tablas de hechos y dimensiones del almacén.
Administrador de BI	Realiza la consulta a la base de datos para la obtención de los cubos multidimensionales.

Seguridad del subsistema de Integración

La seguridad de los datos durante el proceso de ETL constituye una tarea de gran importancia, se hace necesario garantizar la misma pues de esta depende la confidencialidad e integridad de los datos. El esquema de seguridad representa el respaldado por los niveles de acceso, específicamente por el rol definido:

Tabla 5 Esquema de seguridad definido para el Subsistema de Integración

Rol	Permisos
Administrador ETL	Tiene permiso sobre las transformaciones y trabajos implementados.

Seguridad del subsistema de Visualización

Para la seguridad de la aplicación se definió el rol Administrador, que posee total acceso al A.A.G SICOM-UJC, además de ser el encargado de la creación de nuevos usuarios, así como asignarles los roles y permisos a los mismos y el rol especialista que tiene acceso al A.A.G para de este modo consultar las vistas de análisis, reportes operacionales y gráficas correspondientes a cada uno de los libros de trabajo de esta área. En la Tabla 6 se muestra el esquema de seguridad definido para el Subsistema de Visualización.

Tabla 6. Esquema de seguridad definido para el Subsistema de Visualización

Rol	Permisos
Administrador	Acceso total al A.A.G SICOM-UJC, creación de grupos, usuarios y gestión de permisos.
Especialista	Acceso para visualizar y realizar el trabajo sobre las vistas de análisis del A.A.G SICOM-UJC.

Conclusiones del capítulo

Se realizó un estudio de las necesidades de información de la UJC, permitiendo la identificación de 17 requisitos de información agrupados en ocho casos de uso de información, 16 funcionales agrupados en siete casos de uso, 10 no funcionales y 10 reglas del negocio, las cuales han sido aplicadas durante el diseño de los subsistemas. Mediante el diseño del modelo de datos fueron identificadas 24 tablas dimensionales y ocho tablas de hechos, que garantizan el correcto funcionamiento del sistema. El perfilado de datos realizado a la base de datos de SICOM-UJC permitió obtener una noción del estado de la misma, así como el establecimiento de nuevas reglas del negocio aplicables durante el proceso de transformación. El diseño de las transformaciones para la carga de las dimensiones y los hechos constituye una aproximación a los pasos que se deben realizar para lograr la estandarización de la información y su almacenamiento. Las políticas de recuperación y respaldo establecidas contribuyen a mantener la integridad de los datos almacenados. Lo antes planteado centra las bases para la implementación del Almacén de datos para la UJC.

CAPÍTULO III: Implementación y pruebas del Almacén de datos para la UJC

Introducción

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

3.1 Implementación del Subsistema de Almacenamiento

Una vez planteado el modelo dimensional siguiendo una estandarización de los nombres, se dio lugar al modelo físico, permitiendo describir el almacenamiento de los datos y la relación entre las tablas. Además fueron creados los esquemas, así como las tablas correspondientes a cada uno de ellos.

3.1.1 Estándares de codificación

Los estándares de codificación persiguen el objetivo de organizar la forma en que se nombran las estructuras con el fin de lograr un patrón que contribuya a la correcta normalización de los términos utilizados. Esta codificación está más bien dirigida a los desarrolladores, para que exista un vocabulario común en todo el Almacén de datos para la UJC, que permita un entendimiento claro.

En la solución propuesta se mantiene la misma nomenclatura atendiendo a la clasificación de las estructuras, teniendo en cuenta si la misma es una dimensión, un hecho, una transformación, un metadatos o un trabajo. Si la tabla es una dimensión, el nombre estaría compuesto por las letras “dim” separadas del nombre de la misma por el caracter “_”, ejemplo dim_clasificador_ocupacional. En caso de ser una tabla de hecho, se le antepone las letras “hech” e igualmente se separa del nombre del hecho por el caracter “_”, ejemplo hech_sanciones_desactivaciones.

En el caso de los atributos de las dimensiones se siguió la misma estrategia para cada una de ellas. Las llaves primarias de las dimensiones fueron denominadas de la forma “pk_dim_dimension_id”, ejemplo pk_dim_sexo_id. Si el atributo fuera un código del negocio se le especificó “codigo_dimension “, ejemplo codigo_sexo. Se procedió de igual forma para el nombre y la descripción: “nombre_dimension” y “descripcion_dimension”, ejemplos nombre_sexo y descripcion_sexo respectivamente. Las medidas

contienen las letras “cantidad”, el caracter “_” y luego se especifica lo que se va a contar, ejemplo cantidad_militantes.

El nombre de las transformaciones comienzan con las letras “trans”, luego el caracter especial “_” y finalmente el nombre de la misma, ejemplos trans_dim_sexo y trans_hech_sanciones_desactivaciones; mientras que los trabajos comienzan con las letras “trab”, luego el carácter especial “_” y finalmente el nombre de la misma, ejemplo trab_hechos.

Luego de finalizar el proceso de estandarización de los nombres, queda organizada la nomenclatura utilizada para la denominación de las tablas, atributos y medidas dentro de la base de datos, así como de las transformaciones y trabajos. Se procede entonces a la implementación del modelo de las estructuras físicas.

3.1.2 Implementación del modelo físico de datos

El modelo de datos físico es una colección de entidades que describen la estructura de los datos, las restricciones de su integración y las operaciones de manipulación de los mismos. Este modelo se genera partiendo del modelo dimensional mostrado en el capítulo anterior.

En la base de datos se encuentran los datos organizados de manera estructurada facilitando la correcta manipulación de los mismos. Estas estructuras se denominan esquemas y tablas. En la investigación realizada se definieron dos esquemas: mart_sicom y dimensiones. El esquema dimensiones contendrá las dimensiones del Almacén de datos para la UJC, mientras que mart_sicom contendrá los hechos del mismo.

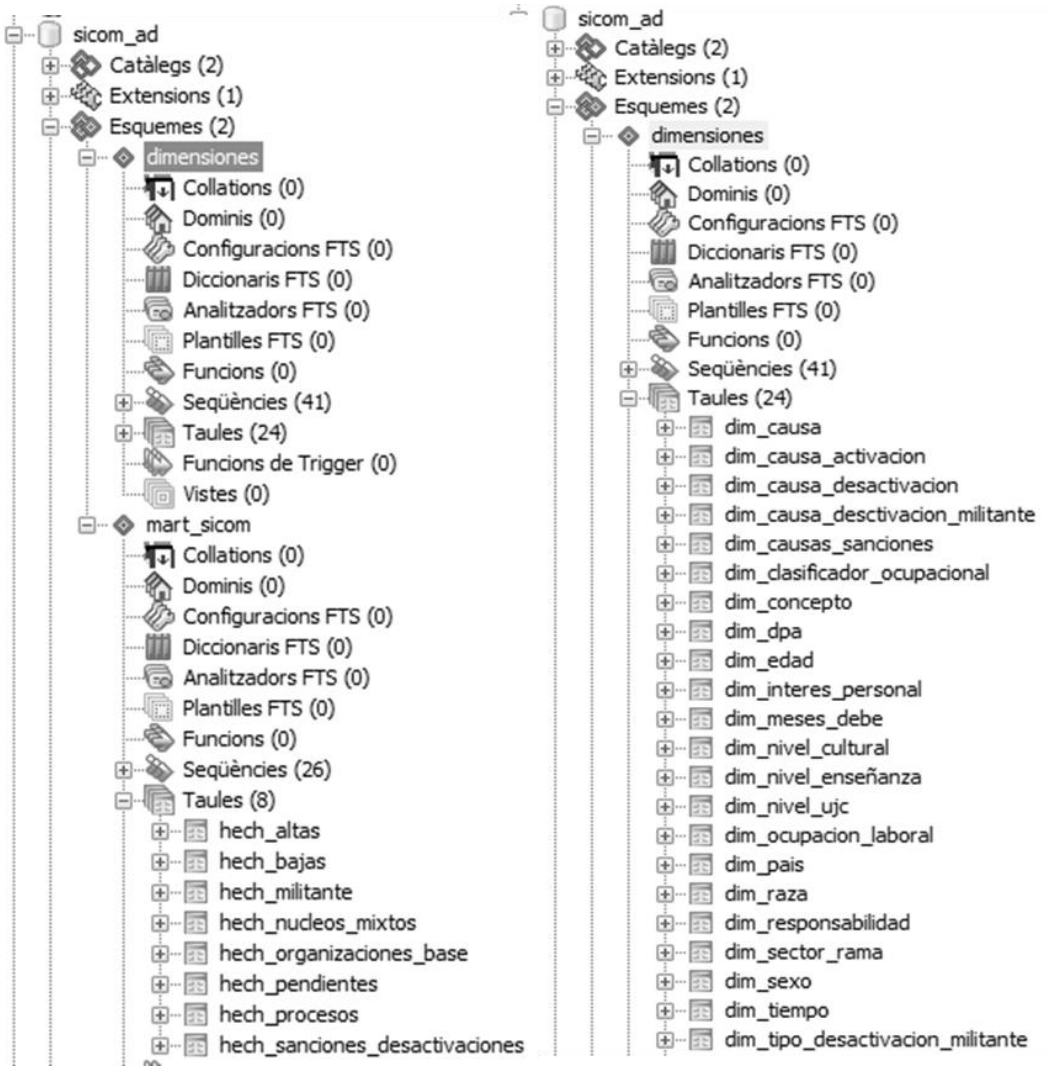


Fig. 16 Esquemas del Almacén de datos para la UJC.

3.2 Implementación del Subsistema de Integración

El proceso de ETL consiste en extraer los datos de las fuentes y se seleccionan los campos necesarios conforme al modelo de datos, luego estos datos se transforman, limpian y estandarizan para eliminar inconsistencias y posibles errores que pudieran llegar a existir. Luego se realiza la carga de las dimensiones y hechos que componen el Almacén de datos para la UJC a través de un grupo de componentes que se encuentran en la herramienta definida en el capítulo uno, teniendo como salida la tabla correspondiente en la base de datos.

A continuación se muestran algunos ejemplos de las transformaciones realizadas para poblar la base de datos correspondiente al Almacén de datos para la UJC.

Se carga la dimensión correspondiente a los datos de los sectores o ramas de la economía (dim_sector_rama): primeramente se extraen los datos de las diferentes tablas de la fuente, luego son corregidos ortográficamente los valores de entrada, se añade el código a la dimensión, se seleccionan los valores a mostrar ajustando los datos y se insertan en el almacén.

Fig. 17 Transformación de la dimensión sector o rama de la economía

Se carga el hecho correspondiente a los indicadores pertenecientes a los núcleos mixtos: primeramente se extraen los datos de la base datos, se filtran las filas para eliminar las inconsistencias existentes, se realiza el tratamiento de los valores nulos de entrada, se busca en la base de datos los identificadores de las dimensiones relacionados con el hecho (llaves subrogadas), se validan los valores a cargar, se seleccionan los valores a mostrar, se agrupan y calculan las medidas para finalmente cargar el hecho.

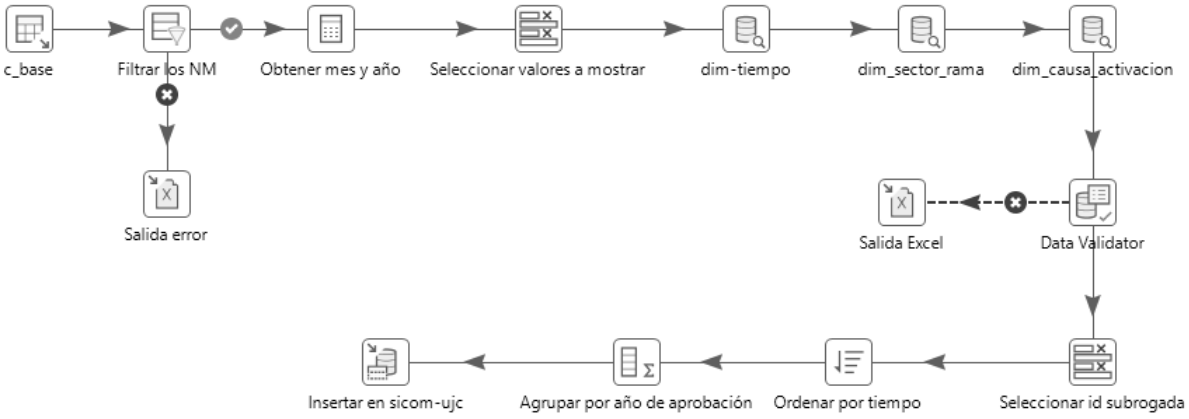


Fig. 18 Transformación del hecho núcleos mixtos.

3.2.1 Implementación de los trabajos

Un trabajo o *Job* es la realización de un conjunto de tareas para realizar determinadas acciones. También se pueden realizar un grupo de transformaciones dependiendo de la secuencia a seguir, una transformación no se empieza a ejecutar si la anterior no ha terminado, en este caso se cargaron primero las dimensiones para que no haya referencia de llaves nulas en las tablas de hechos (Fig.19).

Para dar comienzo al trabajo se comprueba la conexión si no existe se envía un mensaje de error y se aborta el trabajo, mientras que si la conexión es correcta se ejecuta el trabajo correspondiente a las dimensiones y luego el correspondiente a los hechos, para luego enviar un mensaje que confirma el fin del trabajo; de no ejecutarse alguna transformación el trabajo es abortado.

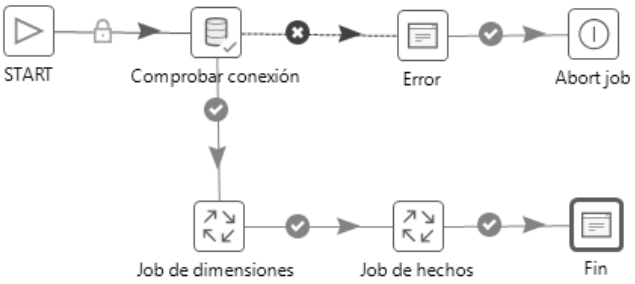


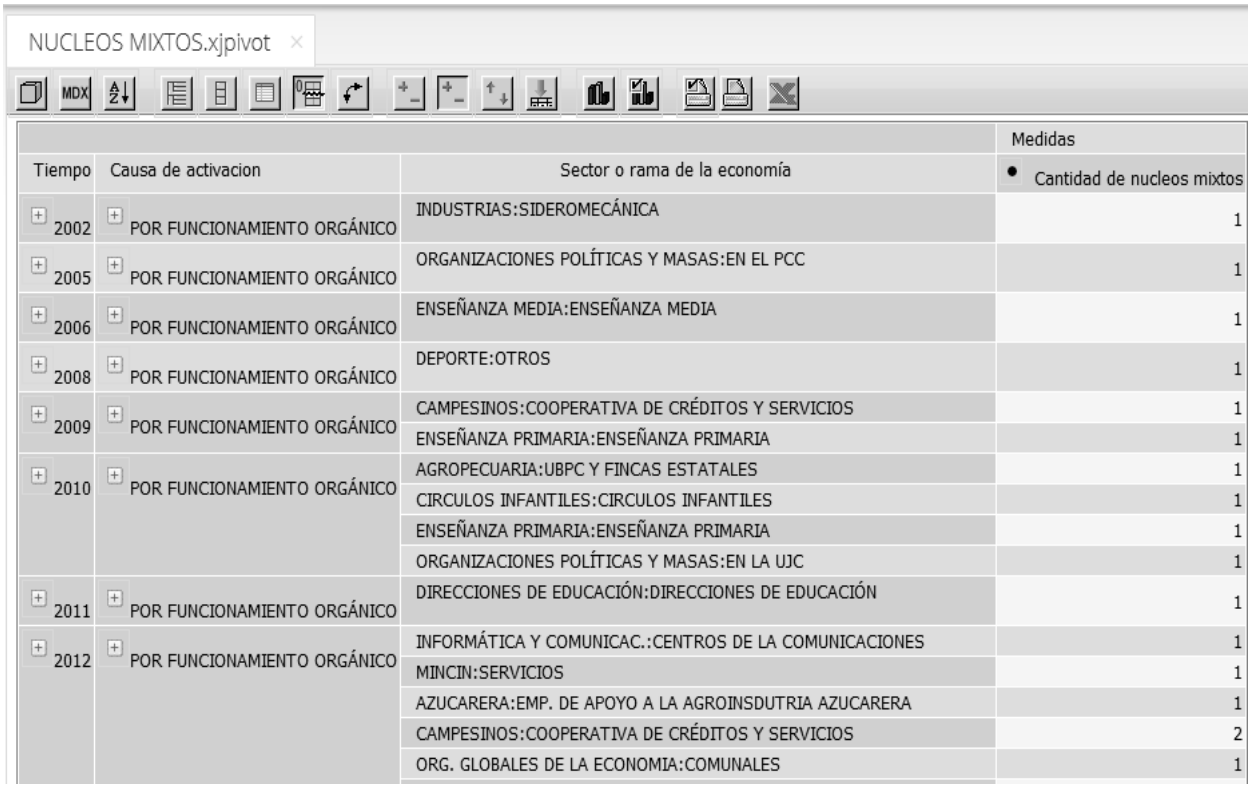
Fig. 19 Transformación del trabajo.

3.3 Implementación del Subsistema de Visualización

Una vez implementados los subsistemas de almacenamiento e integración, se procede a la implementación del subsistema de visualización. Se utilizó como base la arquitectura de la información que contiene las Áreas de Análisis y los Libros de Trabajo. La implementación del subsistema está comprendida por las vistas de análisis, *dashboard* y reportes operacionales.

3.3.1 Implementación de las vistas de análisis

Las vistas de análisis pueden ser creadas o consultadas por el usuario en la herramienta de *BI Server* después de publicados los cubos *OLAP*. Para la realización de cada reporte se define una consulta MDX, por lo que se tienen 17 consultas. A continuación se muestra la vista de análisis correspondiente al RI Obtener la cantidad de núcleos mixtos por fecha de activación, causa de activación y sector o rama de la economía (Fig. 20).



			Medidas
Tiempo	Causa de activacion	Sector o rama de la economía	Cantidad de nucleos mixtos
2002	POR FUNCIONAMIENTO ORGÁNICO	INDUSTRIAS:SIDEROMECÁNICA	1
2005	POR FUNCIONAMIENTO ORGÁNICO	ORGANIZACIONES POLÍTICAS Y MASAS:EN EL PCC	1
2006	POR FUNCIONAMIENTO ORGÁNICO	ENSEÑANZA MEDIA:ENSEÑANZA MEDIA	1
2008	POR FUNCIONAMIENTO ORGÁNICO	DEPORTE:OTROS	1
2009	POR FUNCIONAMIENTO ORGÁNICO	CAMPESINOS:COOPERATIVA DE CRÉDITOS Y SERVICIOS	1
		ENSEÑANZA PRIMARIA:ENSEÑANZA PRIMARIA	1
2010	POR FUNCIONAMIENTO ORGÁNICO	AGROPECUARIA:UBPC Y FINCAS ESTATALES	1
		CIRCULOS INFANTILES:CIRCULOS INFANTILES	1
		ENSEÑANZA PRIMARIA:ENSEÑANZA PRIMARIA	1
		ORGANIZACIONES POLÍTICAS Y MASAS:EN LA UJC	1
2011	POR FUNCIONAMIENTO ORGÁNICO	DIRECCIONES DE EDUCACIÓN:DIRECCIONES DE EDUCACIÓN	1
2012	POR FUNCIONAMIENTO ORGÁNICO	INFORMÁTICA Y COMUNICAC.:CENTROS DE LA COMUNICACIONES	1
		MINCIN:SERVICIOS	1
		AZUCARERA:EMP. DE APOYO A LA AGROINSDUTRIA AZUCARERA	1
		CAMPESINOS:COOPERATIVA DE CRÉDITOS Y SERVICIOS	2
		ORG. GLOBALES DE LA ECONOMIA:COMUNALES	1

Fig. 20 Vista de análisis perteneciente al hecho núcleos mixtos.

3.3.2 Implementación de los dashboard

Los *dashboard* (tableros de mando) permiten ver la información más importante de una empresa. Es uno de los recursos más potentes y utilizados en inteligencia de negocio. Un *dashboard* es un informe que incluye gráficos, tablas e indicadores. Su objetivo es mostrar mucha información y hacerla visible y comprensible a primera vista (Serrano, 2014). *Pentaho dashboard* es una plataforma integrada a la herramienta Pentaho BI Server, la cual será utilizada para su implementación y visualización. A continuación se muestra un ejemplo de un *dashboard* perteneciente al Resumen de los núcleos mixtos del municipio de Florida en Camagüey.

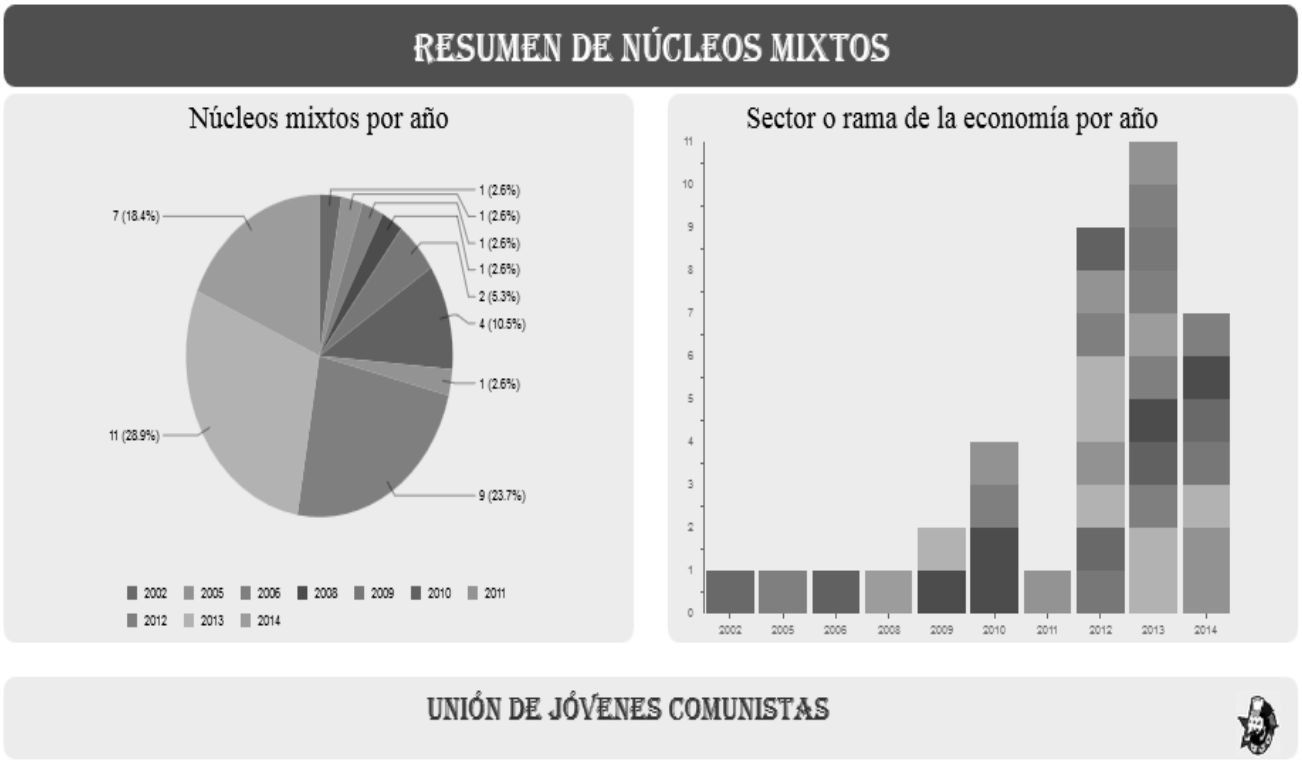


Fig. 21 *Dashboard* Resumen de los núcleos mixtos

3.3.3 Implementación de los reportes operacionales

Los reportes que son creados para el análisis de la información se realizan a través de las herramientas Report Designer y BI *server*. Estos brindan la oportunidad de que el usuario pueda filtrar el reporte de acuerdo a la información que desee analizar. Los reportes operacionales en el Almacén de datos para la UJC fueron implementados en su totalidad a través de consultas SQL. A continuación se muestra información seleccionada por un usuario, referente a la cantidad de organizaciones de base en el año 2005, provincia Camagüey, mostrando el mes, la causa de activación y sector o rama de la economía a la que pertenece un núcleo mixto (Fig. 22).

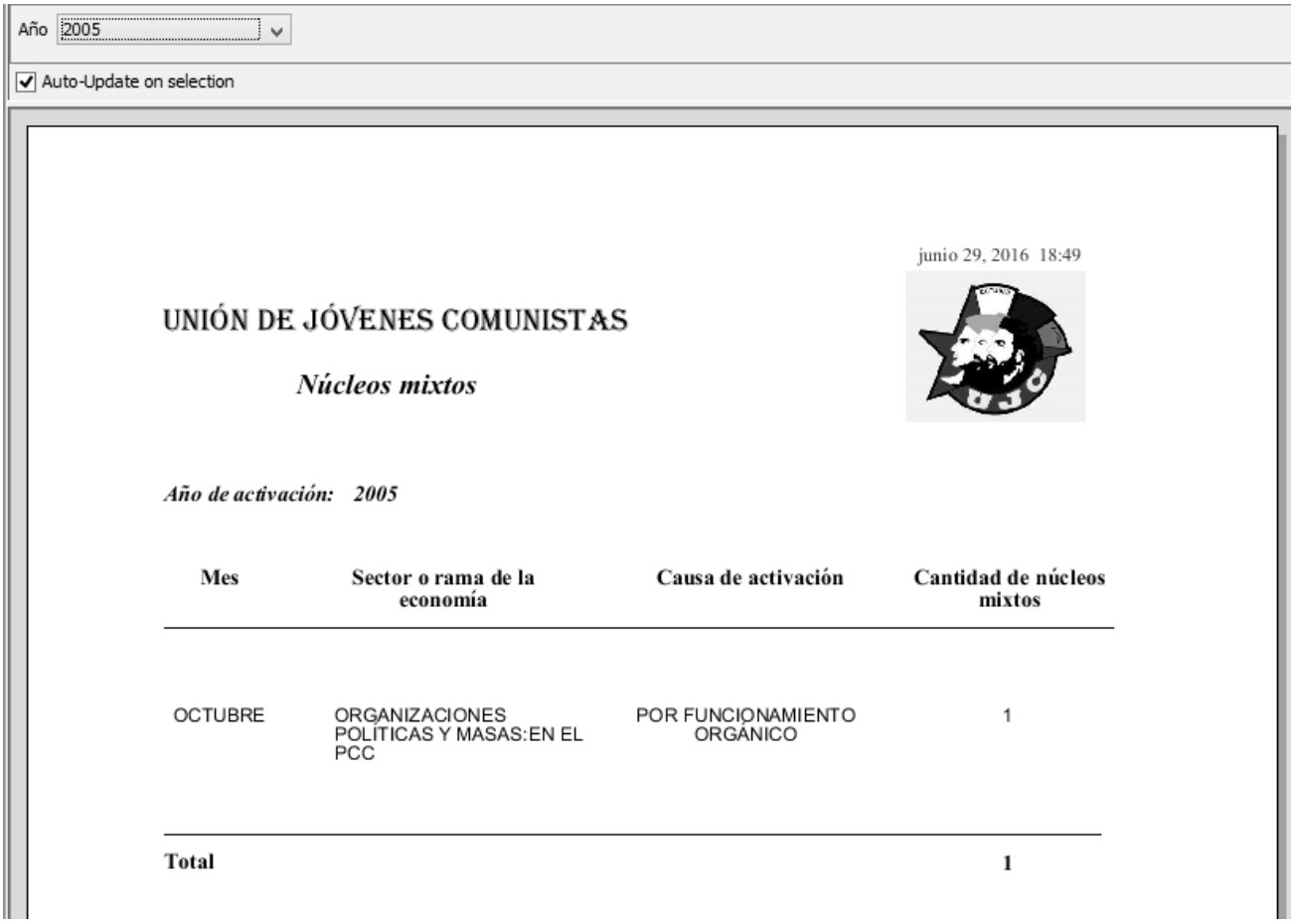


Fig. 22 Reporte operacional núcleos mixtos.

3.4 Pruebas

El único instrumento adecuado para determinar el estado de la calidad de un producto de *software* es el proceso de pruebas. En este proceso se ejecutan pruebas dirigidas a componentes del *software* o al sistema en su totalidad, con el objetivo de medir el grado en que cumple con los requerimientos. En las pruebas se usan casos de prueba, especificados de forma estructurada mediante técnicas de prueba (Pruebas de Software, 2005).

La metodología utilizada propone la realización de pruebas internas ya sean operacionales, unitarias y de integración y pruebas de aceptación por el cliente final. A continuación se muestran las pruebas aplicadas al Almacén de datos para la UJC propuestas por la metodología utilizada.

- ✓ **Pruebas unitarias:** Es el proceso de probar los subsistema individuales de un programa. El propósito es descubrir discrepancias entre la especificación de la interfaz de los módulos y su

comportamiento real. Estas pruebas son diseñadas y ejecutadas por el desarrollador una vez terminado el desarrollo de cada componente.

- ✓ **Pruebas de integración:** Son las pruebas que se realizan para determinar la integración de los componentes dentro de un sistema y evaluar su correcta interfaz, funcionalidad y desempeño. Estas pruebas son diseñadas y ejecutadas por el desarrollador cuando la solución está completa junto a los especialistas del centro.

- ✓ **Pruebas del sistema:** Están basadas en los requerimientos generales y abarca todas las partes combinadas del sistema. Permiten validar el cumplimiento de los requisitos de información y funcionales definidos por los clientes. Son las pruebas más cercanas a la realidad del cliente, debido a que los probadores utilizan el sistema de la misma manera que será usado por los clientes. Estas pruebas constituyen las actividades fundamentales de la fase de prueba (Hernández, 2013).

3.4.1 Herramientas de pruebas

Las herramientas de pruebas son el medio para comprobar si lo implementado cumple con los objetivos trazados. En el Almacén de datos para la UJC fueron realizados los casos de prueba y las listas de chaqueo a los artefactos del proceso de ETL: perfilado de datos, diccionario de datos, mapa lógico y registro de sistemas fuente.

Casos de Prueba

Los casos de prueba son utilizados para identificar posibles fallos de implementación y comprobar el grado de cumplimiento de los requisitos especificados para el sistema. En el Subsistema de visualización del Almacén de datos para la UJC fueron diseñados ocho casos de prueba asociados a cada CUI identificados en la etapa de análisis, con el fin de comprobar que estén almacenadas las variables correspondientes. A continuación se muestra el caso de prueba perteneciente al CU Mostrar información de los núcleos mixtos (Tabla 7), el resto podrá ser consultado en el expediente de proyecto.

Tabla 7 Caso de prueba del CU Mostrar información de los núcleos mixtos

Escenario	Descripción	Variables de Entrada	Variables de Salida	Respuesta del sistema	Flujo central
EC 1: Obtener la cantidad de núcleos mixtos por fecha de activación, causa de activación y sector o rama de la economía.	Permite mantener disponible la información de la cantidad de núcleos mixtos en la unión de jóvenes comunistas por fecha de activación, causa de activación y sector o rama de la economía.	Fecha Causa de activación Sector o rama de la economía	cantidad de núcleos mixtos	El sistema muestra todas las variables disponibles para el análisis, ubicados en las filas y las columnas que pueden ser visualizadas en el reporte.	Se abre la aplicación. Se autentica. Se entra al sistema. Se selecciona el área de análisis general de A.A.G SICOM-UJC. Se selecciona el área de análisis de A.A. GESTION DE LA INFORMACION. Se selecciona el libro de trabajo L.T NIVELES DE LA ESTRUCTURA. Se selecciona el reporte NUCLEOS MIXTOS.

Lista de Chequeo

La lista de chequeo consta de una serie de preguntas, en forma de cuestionario, mediante el cual se verifica el grado de cumplimiento de determinadas reglas establecidas para los procesos de desarrollo del sistema, además de medir la calidad de los artefactos de los procesos de ETL generados durante la realización del producto. Esta evaluación se desarrolla a través del análisis de un grupo de indicadores, 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.
- ✓ **Semántica del documento:** contempla todos los indicadores a evaluar respecto a la ortografía y redacción.

La estructura de la lista de chequeo está formada por los siguientes elementos:

- ✓ **Peso:** define si el indicador a evaluar es crítico o no. El mismo se describe con una C si es crítico.
- ✓ **Indicadores a evaluar:** constituyen los indicadores a evaluar en las secciones Estructura del documento, Semántica del documento e Indicadores definidos para el artefacto a evaluar.

- ✓ **Evaluación:** es la forma de evaluar el indicador en cuestión. El mismo se evalúa de uno en caso de que exista alguna dificultad sobre el indicador y de cero, en caso de que el indicador revisado no presente problemas.
- ✓ **No Procede (N.P):** se usa para especificar que no es necesario evaluar el indicador en ese caso.
- ✓ **Cantidad de elementos afectados (CEA):** especifica la cantidad de errores encontrados sobre el mismo indicador.
- ✓ **Comentario:** especifica los señalamientos o sugerencias que quiera incluir la persona que aplica la lista de chequeo. Pueden o no existir señalamientos o sugerencias.

Las listas de chequeo se le aplicaron a los artefactos de ETL “Registro del sistema fuente (RSF)”, “Perfilado de datos (PD)”, “Diccionario de datos (DD)” y “Mapa lógico de datos (MLD)”. Seguidamente se muestra una tabla en la que se encuentran los principales aspectos que fueron evaluados y los resultados que arrojó la aplicación de los mismos.

Tabla 8 Aplicación de la Lista se chequeo a los artefactos de ETL.

Secciones	RSF	PD	DD	MLD
Estructura	2	2	2	2
Indicadores	1	4	3	1
Semántica	3	3	3	3
Total de indicadores	6	9	8	6
Indicadores críticos	4	5	5	4
No Conformidades	3	3	3	2

En la Fig. 23 se encuentra un resumen de los resultados obtenidos después de la aplicación de las Listas de Chequeo a los cuatro artefactos mencionados anteriormente.

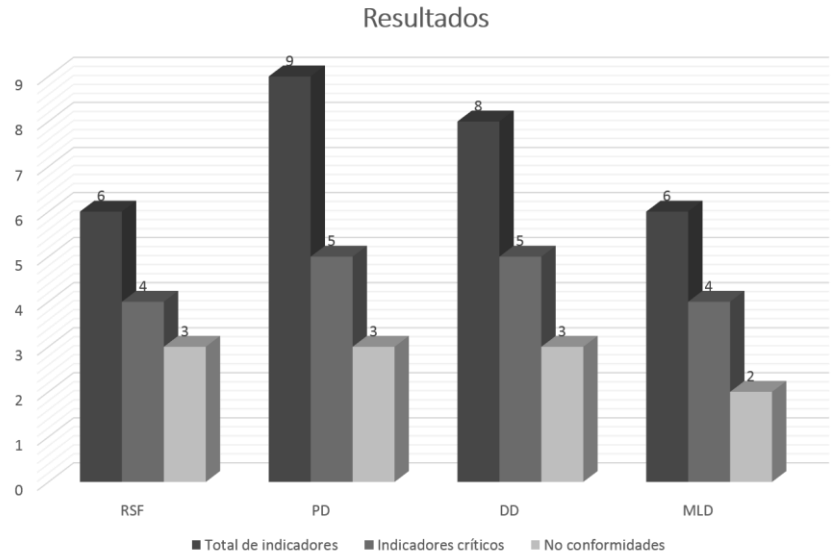


Fig. 23 Resultado de la aplicación de las Listas de chequeo a los artefactos

3.4.2 Resultados de las pruebas

Pruebas unitarias

Estas pruebas fueron realizadas por los especialistas de almacenes de datos en DATEC en almacenes de datos. Una vez concluida la etapa de análisis y diseño se comprobó el subsistema de almacenamiento, donde se detectaron dos No Conformidades (NC), solucionadas en su totalidad.

NC1: No están bien definidas las llaves primarias en la dimensiones.

NC2: Existían tablas fantasmas en el Almacén de datos generados a través de la herramienta Visual Paradigm.

Durante la realización de la etapa de implementación se le realizaron pruebas unitarias al subsistema de integración de datos, detectando cuatro NC, solucionadas en su totalidad.

NC1: Los componentes utilizados en la implementación de las transformaciones no son los ideales.

NC2: El orden de los componentes no permite la correcta agrupación de los valores.

NC3: Los componentes no sugieren la función que realizan, imposibilitando el entendimiento común.

Una vez concluida la etapa de implementación se le realizaron pruebas unitarias al subsistema de visualización de datos, detectando una NC, solucionada en su totalidad.

NC1: Los miembros mostrados en las vistas de análisis no permiten el entendimiento de la misma.

A continuación se muestran los resultados de las pruebas unitarias (Fig. 23):

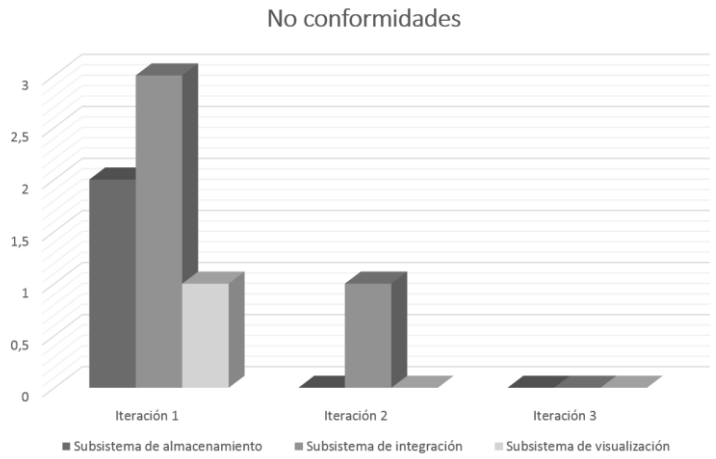


Fig. 24 Resultado de las pruebas unitarias

Conclusiones del capítulo

Se realizó la implementación del subsistema de integración en el que se ejecutaron 24 transformaciones de las dimensiones, 8 de hechos y 3 trabajos. Fue poblado el Almacén de datos para la UJC en su totalidad, permitiendo la creación de un total de 15 vistas de análisis, 12 reportes operacionales y 3 *dashboard* que conforman el subsistema de visualización. La obtención del mismo permitió comprender mejor la información, así como el análisis de tendencias en apoyo a la toma de decisiones. Se realizaron las pruebas para validar el correcto funcionamiento del Almacén de datos para la UJC, dando respuesta a las necesidades del cliente definidas al inicio de la investigación.

CONCLUSIONES

El estudio de los distintos temas relacionados con el desarrollo de los almacenes de datos, proporcionó la elaboración del presente trabajo, del cual se obtuvo como resultado el “Almacén de para la Unión de Jóvenes Comunistas”. Los siguientes resultados demuestran el cumplimiento de los objetivos propuestos en la investigación:

- ✓ Se logró mediante una investigación detallada la selección de la metodología, herramientas y tecnologías a utilizar en el desarrollo de la solución, permitiendo centrar las bases en el proceso de construcción del Almacén de datos para la UJC.
- ✓ Se realizó el análisis y diseño del Almacén de datos para la UJC. Fueron identificados 17 requisitos de información, 12 funcionales, 10 no funcionales y 10 reglas del negocio. Además, fueron diseñados los subsistemas de integración, almacenamiento y visualización, que permitirían la posterior implementación.
- ✓ La implementación de los subsistemas de almacenamiento, integración y de visualización posibilitaron la obtención del Almacén de datos para la UJC correctamente poblado, con información disponible para ser consultada por parte de los usuarios, brindando apoyo al proceso de toma de decisiones.
- ✓ Se demostró mediante las pruebas unitarias y de sistema, aplicando casos de prueba y lista de chequeo que el sistema cumple con las necesidades del cliente.

RECOMENDACIONES

Con el propósito de mejorar la propuesta realizada en este trabajo de diploma, se recomienda:

- ✓ Realizar la integración de los datos de la Unión de Jóvenes Comunistas de la nación.

BIBLIOGRAFÍA REFERENCIADA

WorkMeter . 2010. El blog de WorkMeter . [En línea] 27 de Julio de 2010. [Citado el: 20 de Enero de 2016.] http://es.workmeter.com/blog/bid/192978/Principales-herramientas-de-Business-Intelligence.

Ayala, Alejandro Peña. 2006. *Sistemas basados en Conocimiento:Una Base para su Concepción y Desarrollo.* Mexico : INSTITUTO POLITÉCNICO NACIONAL, 2006. 970-94797-4-1 .

Bernabeu, Ing. Ricardo Dario Córdoba. Julio de 2010.. *DATA WAREHOUSING: Investigación y Sistematización de Conceptos. HEFESTO: Metodología para la Construcción de un Data Warehouse.* Córdoba, Argentina, : s.n., Julio de 2010.

Bertino, E.A y Martino, L.A. 1995. *Sistemas de bases de datos orientadas a objetos.* s.l. : Ediciones Díaz de Santos, 1995.

Blogger. 2010. [En línea] 1 de Mayo de 2010. http://marycasasola.blogspot.com/2010/05/sistemas-de-soporte-decisiones-en-grupo.html.

Caralt, Jordi Conesa. 2011. *Introducción al Business Intelligencie.* Barcelona : OUC, 2011. ISBN: 978-84-9788-886-8.

Carrillo, A. 2009. *Herramienta Multimedia de apoyo a la Enseñanza de la Metodología RUP de Ingeniería del Software.* 2009.

Comunidad Postgres. 2010. PosgreSQL. [En línea] 10 de Octubre de 2010. [Citado el: 10 de Noviembre de 2015.] http://www.postgresql.org/about/.

Crosse, Favier. 2014. SlideShare. [En línea] Linkendl, 11 de Junio de 2014. [Citado el: 10 de Diciembre de 2015.] http://es.slideshare.net/ssmendez07/sistema-de-apoyo-a-la-toma-de-decisiones.

DataCleaner. 2012. DataCleaner. [En línea] 2012. [Citado el: 17 de enero de 2016.] http://datacleaner.org/.

DATEC. 2012. *Manual de Usuario SICOM-UJC MVR.* La Habana : s.n., 2012.

Díaz, Josep Curto y Conesa Caralt , Jordi. 2011. *Introducción al Business Intelligence.* Barcelona : El Ciervo 96, 2011. 978-84-9788-886-8.

Espinosa, Roberto. 2010. El Rincón del BI. [En línea] 10 de Julio de 2010. [Citado el: 11 de Novimbre de 2015.] https://churriwifi.wordpress.com/2010/07/04/17-3-preparando-el-analisis-dimensional-definicion-de-cubos-utilizando-schema-workbench/.

ETL-Tools.Info. 2014. ETL-Tools.Info. [En línea] 2014. [Citado el: 20 de Enero de 2016.] http://etl-tools.info/es/bi/proceso_etl.htm.

Gespro. 2015. Suite de Gestion de Proyectos. [En línea] 2015. [Citado el: 30 de Octubre de 2015.] http://gespro.datec.prod.uci.cu/.

Grupo Aitec. 2013. Business Intelligence. [En línea] 2013. [Citado el: 10 de Noviembre de 2015.] http://www.linkconsulting.com/BI/detalhe_artigo.aspx?idsc=5380&idl=3.

Guzmán, Eric. 2013. Prezi. [En línea] 3 de Septiembre de 2013. [Citado el: 10 de Marzo de 2016.] https://prezi.com/jekyx5ssiryg/sistemas-de-informacion-para-ejecutivos/.

HEFESTO. 2010. *DATA WAREHOUSING: Investigación y Sistematización.* Argentina : s.n., 2010.

Hernández, Ing. Yanisbel González. 2013. *METODOLOGÍA DE DESARROLLO PARA PROYECTOS DE ALMACENES DE DATOS.* La Habana : s.n., 2013.

Inmon, William Harvey. 2005. *Building the Data Warehouse, Fourth Edition .* Indianapolis : Wiley Publishing, Inc, 2005. ISBN-10: 0-7645-9944-5.

Kimball, Ralph y Ross, Margy. 2002. *The data warehouse toolkit : the complete guide to dimensional modeling.* Toronto : Wiley Computer Publishing, 2002. ISBN 0-471-20024-7 .

Juarez, María España D. **2011.** *Almacen de datos.* [En línea] 2011. [Citado el: 7 de Octubre de 2015.] http://es.slideshare.net/MaritaEspaaDJurez/almacn-de-datos-20869086..

Microsoft. 2015. Devaloped Network. [En línea] Microsoft, 2015. [Citado el: 1 de Diciembre de 2015.] https://msdn.microsoft.com/es-es/library/aa995548.aspx.

Object Management Group. 1997. Unified Modeling Lenguage. [En línea] 1997. [Citado el: 21 de Octubre de 2015.] http://www.uml-diagrams.org/.

Pedro Alves. 2015. Webdetails. [En línea] 22 de Julio de 2015. [Citado el: 17 de Noviembre de 2015.] http://www.webdetails.pt/ctools/cde/.

Pentaho Community. 2015. Pentaho. [En línea] Pentaho community, 6 de Febrero de 2015. [Citado el: 20 de Enero de 2016.] http://wiki.pentaho.com/display/Reporting/Pentaho+Reporting+-+User+Guide+for+Report+Designer.

Pentaho Corporation. 2012. pentaho. [En línea] Pentaho Corporation, 2012. http://community.pentaho.com/.

PgAdmin.PostgreSQL Tools. 2015. PgAdmin.PostgreSQL Tools. [En línea] 2015. https://www.pgadmin.org/.

Pierri, M. 2013. slideshare. [En línea] 2013. [Citado el: 9 de Noviembre de 2015.] http://www.slideshare.net/mpierri/manipulacion-de-datos-con-kettle..

Pirone, Ángel Luis Peréz y Roldán Bonilla, Kelly D´yana. 2004. *Sistema de apoyo para la toma de decisiones en el control de riesgos de procesos de facturación de una compañía de telefonía movil.* Caracas : s.n., 2004.

Point, Tutorials. 2014. Tutorials Point. [En línea] 2014. [Citado el: 29 de Abril de 2016.] www.tutorialspoint.com/pentaho/pentaho_tutorial.pdf.

PostgreSQL. 2012. pgAdmin PostgreSQL. [En línea] 2012. [Citado el: 9 de Noviembre de 2015.] http://www.pgadmin.org/;http://rpm.pbone.net/index.php3/stat/4/idpl/17347092/dir/fedora_16/com/pgadmin3-1.14.0-1.fc16.x86_64.rpm.html.

Pruebas de Software. 2005. Gestión de Calidad y Pruebas de Software. [En línea] 2005. [Citado el: 11 de Mayo de 2016.] http://www.pruebasdesoftware.com/laspruebasdesoftware.htm.

Real Academia Española. 2016. Real Academia Española. [En línea] 2016. http://dle.rae.es/?id=P7eTCPD.

Schiefer, Josef. 2002. *A Holistic Approach for Managing Requirements of Data Warehouse Systems.* Vienna University oh Technology : s.n., 2002.

Serrano, Erica María del Carmen Palma. 2014. Gestiopolis. [En línea] 14 de Noviembre de 2014. [Citado el: 12 de Abril de 2016.] http://www.gestiopolis.com/inteligencia-de-negocios-business-intelligence/.

Sommerville, I. 2005. *Ingeniería de Software.* Madrid : Pearson Educación, 2005.

Sosa, Ángel Gabriel Olivera. 2010. Scribd. [En línea] 10 de Septiembre de 2010. [Citado el: 12 de Enero de 2016.] https://es.scribd.com/doc/37187866/Requerimientos-funcionales-y-no-funcionales.

SUMMAN. 2012. *Pentaho BI Platform Server.* 2012.

TIBCO Softwarex. 2015. Jaspersoft Community. [En línea] TIBCO Software, 2015. [Citado el: 12 de Diciembre de 2015.] http://community.jaspersoft.com/wiki/jaspersoft-olap-schema-workbench.

UBUNTU. 2013. GUÍA DOCUMENTADA PARA UBUNTU. [En línea] 2013. [Citado el: 9 de Noviembre de 2015.] http://www.guia-ubuntu.com/index.php?title=PgAdmin_III..

Universidad de Sevilla. 2005. Base de datos. Tema 3 Modelo de datos. [En línea] marzo de 2005. [Citado el: 15 de Febrero de 2016.] http://www.lsi.us.es/docencia/get.php?id=4525.

BIBLIOGRFÍA CONSULTADA

WorkMeter . 2010. El blog de WorkMeter . [En línea] 27 de Julio de 2010. [Citado el: 20 de Enero de 2016.] http://es.workmeter.com/blog/bid/192978/Principales-herramientas-de-Business-Intelligence.

Ayala, Alejandro Peña. 2006. *Sistemas basados en Conocimiento:Una Base para su Concepción y Desarrollo.* Mexico : INSTITUTO POLITÉCNICO NACIONAL, 2006. 970-94797-4-1 .

Bernabeu, Ing. Ricardo Dario Córdoba. Julio de 2010.. *DATA WAREHOUSING: Investigación y Sistematización de Conceptos. HEFESTO: Metodología para la Construcción de un Data Warehouse.* Córdoba, Argentina, : s.n., Julio de 2010.

Bertino, E.A y Martino, L.A. 1995. *Sistemas de bases de datos orientadas a objetos.* s.l. : Ediciones Díaz de Santos, 1995.

2010. Blogger. [En línea] 1 de Mayo de 2010. http://marycasasola.blogspot.com/2010/05/sistemas-de-soporte-decisiones-en-grupo.html.

Caralt, Jordi Conesa. 2011. *Introducción al Business Intelligencie.* Barcelona : OUC, 2011. ISBN: 978-84-9788-886-8.

Carrillo, A. 2009. *Herramienta Multimedia de apoyo a la Enseñanza de la Metodología RUP de Ingeniería del Software.* 2009.

Comunidad Postgres. 2010. PosgreSQL. [En línea] 10 de Octubre de 2010. [Citado el: 10 de Noviembre de 2015.] http://www.postgresql.org/about/.

Crosse, Favier. 2014. SlideShare. [En línea] Linkendl, 11 de Junio de 2014. [Citado el: 10 de Diciembre de 2015.] http://es.slideshare.net/ssmendez07/sistema-de-apoyo-a-la-toma-de-decisiones.

DataCleaner. 2012. DataCleaner. [En línea] 2012. [Citado el: 17 de enero de 2016.] http://datacleaner.org/.

DATEC. 2012. *Manual de Usuario SICOM-UJC MVR.* La Habana : s.n., 2012.

Díaz, Josep Curto y Conesa Caralt , Jordi. 2011. *Introducción al Business Intelligence.* Barcelona : El Ciervo 96, 2011. 978-84-9788-886-8.

Espinosa, Roberto. 2010. El Rincón del BI. [En línea] 10 de Julio de 2010. [Citado el: 11 de Novimbre de 2015.] https://churriwifi.wordpress.com/2010/07/04/17-3-preparando-el-analisis-dimensional-definicion-de-cubos-utilizando-schema-workbench/.

ETL-Tools.Info. 2014. ETL-Tools.Info. [En línea] 2014. [Citado el: 20 de Enero de 2016.] http://etl-tools.info/es/bi/proceso_etl.htm.

Gespro. 2015. Suite de Gestion de Proyectos. [En línea] 2015. [Citado el: 30 de Octubre de 2015.] http://gespro.datec.prod.uci.cu/.

Gorman, Will. 2009. *Pentaho Reporting 3.5 for Java Developers.* Birmingham : Packt, 2009. ISBN 978-1-847193-19-.

Grupo Aitec. 2013. Business Intelligence. [En línea] 2013. [Citado el: 10 de Noviembre de 2015.] http://www.linkconsulting.com/BI/detalhe_artigo.aspx?idsc=5380&idl=3.

Guzmán, Eric. 2013. Prezi. [En línea] 3 de Septiembre de 2013. [Citado el: 10 de Marzo de 2016.] https://prezi.com/jekyx5ssiryg/sistemas-de-informacion-para-ejecutivos/.

HEFESTO. 2010. *DATA WAREHOUSING: Investigación y Sistematización.* Argentina : s.n., 2010.

Hernández, Ing. Yanisbel González. 2013. *METODOLOGÍA DE DESARROLLO PARA PROYECTOS DE ALMACENES DE DATOS.* La Habana : s.n., 2013.

Inmon, William Harvey. 2005. *Building the Data Warehouse, Fourth Edition .* Indianapolis : Wiley Publishing, Inc, 2005. ISBN-10: 0-7645-9944-5.

Kimball, Ralph y Ross, Margy. 2002. *The data warehouse toolkit : the complete guide to dimensional modeling.* Toronto : Wiley Computer Publishing, 2002. ISBN 0-471-20024-7 .

Juarez, María España D. 2011. *Almacen de datos.* [En línea] 2011. [Citado el: 7 de Octubre de 2015.] http://es.slideshare.net/MaritaEspaaDJurez/almacn-de-datos-20869086..

Microsoft. 2015. Devaloped Network. [En línea] Microsoft, 2015. [Citado el: 1 de Diciembre de 2015.] https://msdn.microsoft.com/es-es/library/aa995548.aspx.

Object Management Group. 1997. Unified Modeling Lenguage. [En línea] 1997. [Citado el: 21 de Octubre de 2015.] http://www.uml-diagrams.org/.

Pedro Alves. 2015. Webdetails. [En línea] 22 de Julio de 2015. [Citado el: 17 de Noviembre de 2015.] http://www.webdetails.pt/ctools/cde/.

Pentaho Community. 2015. Pentaho. [En línea] Pentaho community, 6 de Febrero de 2015. [Citado el: 20 de Enero de 2016.] http://wiki.pentaho.com/display/Reporting/Pentaho+Reporting+-+User+Guide+for+Report+Designer.

Pentaho Corporation. 2012. pentaho. [En línea] Pentaho Corporation, 2012. http://community.pentaho.com/.

PgAdmin.PostgreSQL Tools. 2015. PgAdmin.PostgreSQL Tools. [En línea] 2015. https://www.pgadmin.org/.

Pierri, M. 2013. slideshare. [En línea] 2013. [Citado el: 9 de Noviembre de 2015.] http://www.slideshare.net/mpierri/manipulacion-de-datos-con-kettle..

Pirone, Ángel Luis Peréz y Roldán Bonilla, Kelly D´yana. 2004. *Sistema de apoyo para la toma de decisiones en el control de riesgos de procesos de facturación de una compañía de telefonía movil.* Caracas : s.n., 2004.

Point, Tutorials. 2014. Tutorials Point. [En línea] 2014. [Citado el: 29 de Abril de 2016.] www.tutorialspoint.com/pentaho/pentaho_tutorial.pdf.

PostgreSQL. 2012. pgAdmin PostgreSQL. [En línea] 2012. [Citado el: 9 de Noviembre de 2015.] http://www.pgadmin.org/;http://rpm.pbone.net/index.php3/stat/4/idpl/17347092/dir/fedora_16/com/pgadmin3-1.14.0-1.fc16.x86_64.rpm.html.

Pruebas de Software. 2005. Gestión de Calidad y Pruebas de Software. [En línea] 2005. [Citado el: 11 de Mayo de 2016.] http://www.pruebasdesoftware.com/laspruebasdesoftware.htm.

Real Academia Española. 2016. Real Academia Española. [En línea] 2016. http://dle.rae.es/?id=P7eTCPD.

Rivadera, Gustavo R. 2010. *La metodología de Kimball para el diseño de almacenes de datos.* 5, Buenos Aires : Cuadernos de la Facultad, 2010, Vol. I.

Schiefer, Josef. 2002. *A Holistic Approach for Managing Requirements of Data Warehouse Systems.* Vienna University oh Technology : s.n., 2002.

Serrano, Erica María del Carmen Palma. 2014. Gestiopolis. [En línea] 14 de Noviembre de 2014. [Citado el: 12 de Abril de 2016.] http://www.gestiopolis.com/inteligencia-de-negocios-business-intelligence/.

Sommerville, I. 2005. *Ingeniería de Software.* Madrid : Pearson Educación, 2005.

Sosa, Ángel Gabriel Olivera. 2010. Scribd. [En línea] 10 de Septiembre de 2010. [Citado el: 12 de Enero de 2016.] https://es.scribd.com/doc/37187866/Requerimientos-funcionales-y-no-funcionales.

SUMMAN. 2012. *Pentaho BI Platform Server.* 2012.

TIBCO Softwarex. 2015. Jaspersoft Community. [En línea] TIBCO Software, 2015. [Citado el: 12 de Diciembre de 2015.] http://community.jaspersoft.com/wiki/jaspersoft-olap-schema-workbench.

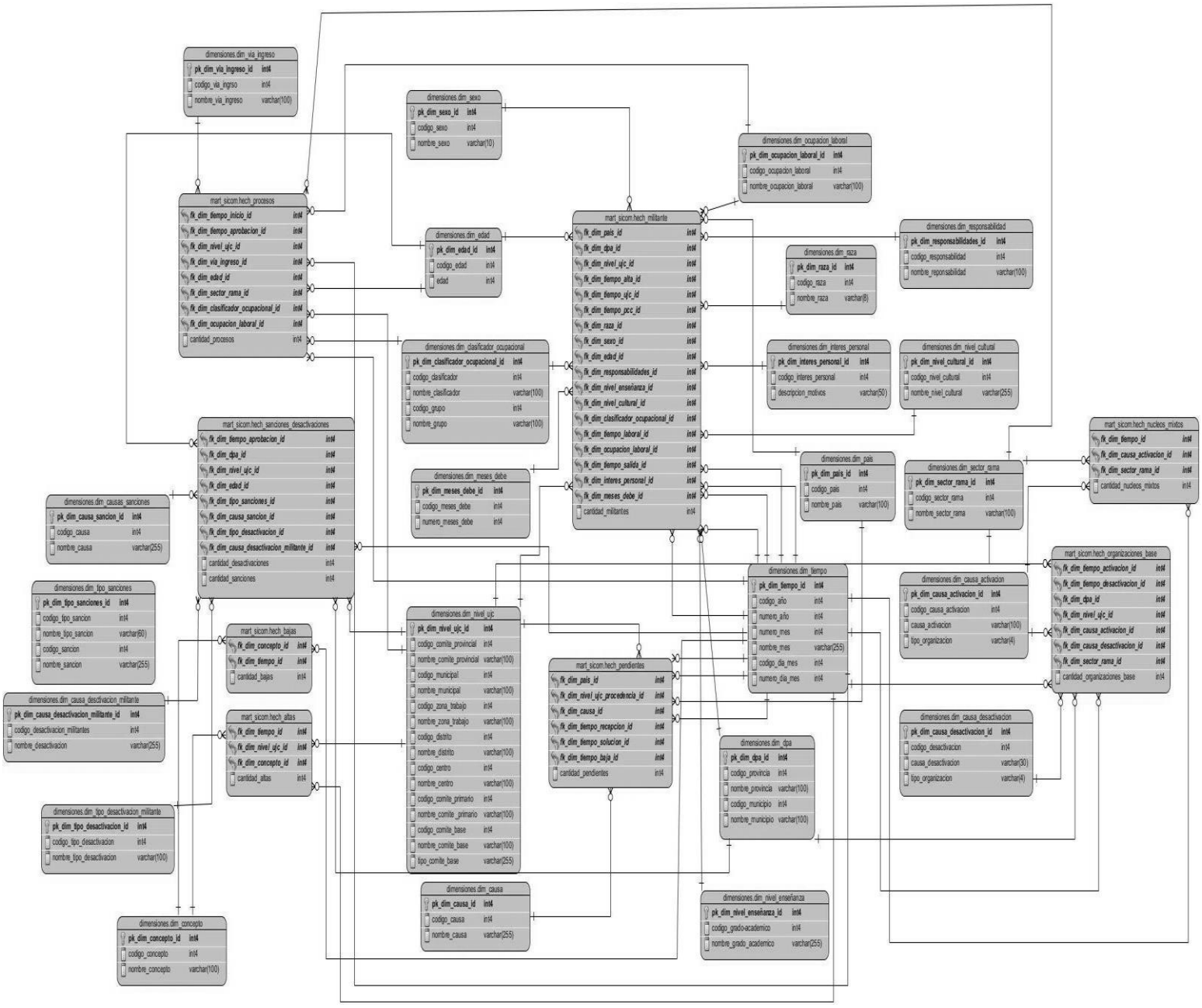
UBUNTU. 2013. GUÍA DOCUMENTADA PARA UBUNTU. [En línea] 2013. [Citado el: 9 de Noviembre de 2015.] http://www.guia-ubuntu.com/index.php?title=PgAdmin_III..

Underdahl, Brian. 2014. *Data Integration for Dummles.* New Jesrsey : John Wiley & Sons, Inc, 2014. ISBN 978-1-118-89658-7.

Universidad de Sevilla. 2005. Base de datos. Tema 3 Modelo de datos. [En línea] marzo de 2005. [Citado el: 15 de Febrero de 2016.] http://www.lsi.us.es/docencia/get.php?id=4525.

ANEXOS

Anexo 1: Modelo de datos del Almacén de datos para SICOM-UJC.



Anexo 2: Entrevista realizada a los creadores de SICOM-UJC

1-) ¿Cuáles son los objetivos de la organización?

2-) ¿Qué situación existe en la actualidad para el análisis de la información?

3-) ¿Cómo se organiza la información en la organización?

4-) ¿Qué áreas de análisis son de prioridad para la UJC? ¿Cuáles son los procesos que se manejan en estas áreas de acuerdo al Sistema de Control de la militancia para la UJC?

5-) ¿Qué tipos de reportes se obtienen del sistema?

6-) ¿Con que frecuencia se obtiene la información?

7-) ¿Cantidad de información que se maneja en el sistema?

8-) ¿Que indicadores y nomencladores se tienen en cuenta para el análisis de la información?

9-) ¿Cuáles son los reportes más solicitados por los especialistas?

10-) ¿Qué tipos de análisis les gustaría realizar sobre los indicadores mencionados anteriormente?

11-) ¿Cómo les sería más fácil la presentación de la información en pantalla?