# Universidad de las Ciencias Informáticas

# Facultad 10


**Título:** Almacén de datos para los subsistemas de Reclutamiento y Potencial Humano.

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

**Autor:** Cadete Orestes Rodríguez Lorenzo.

**Tutor**: Tte. Ing. Jacinto Torres Fernández.

**Co Tutor:** Tte. Ing. Yadir Martinez Vergara.

Ciudad de La Habana, Junio de 2010.

Año del 50 aniversario de la Revolución

## Declaración de autoría

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

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

| | |
|---|---|
| _____ | _____ |
| Cadete Orestes Rodríguez Lorenzo | Tte. Ing. Jacinto Torres Fernández |
| Autor | Tutor |

## Datos de contacto

**Tutor:** Tte. Ing. Jacinto Torres Fernández.

**Profesión:** Ingeniero en Ciencias Informáticas.

**Clasificación del área de desarrollo:** UCID – Software de gestión.

**Síntesis:** Ingeniero en Ciencias Informáticas, graduado en julio de 2007. Líder del proyecto Comités Militares de la Unidad de Compatibilización, Integración y Desarrollo de Software para la Defensa (UCID).

**Años de graduado:** 3

**Correo electrónico:** jtfernandez@uci.cu

**Co-Tutor:** Tte. Ing. Yadir Martinez Vergara.

**Profesión:** Ingeniero en Ciencias Informáticas.

**Clasificación del área de desarrollo:** UCID – Software de gestión.

**Síntesis:** Ingeniero en Ciencias Informáticas, graduado en julio de 2008. Líder de la línea de Personal Vinculado a la Defensa de la Unidad de Compatibilización, Integración y Desarrollo de Software para la Defensa (UCID).

**Años de graduado:** 2

**Correo electrónico:** yvergara@uci.cu

*"... no hay nada que eduque más a un hombre honrado que vivir dentro de una revolución."*

## Agradecimientos

A toda mi familia, especialmente a mis padres, mi hermana y mi abuela, por darme su apoyo incondicional en todo momento.

A mi novia por apoyarme y estar siempre a mi lado todos estos años.

A mis amigos Osmany, Machín, Yadier, Mailén, Yamel, Adonys, Anaivys y Yordi.

A mis tutores Jacinto y Yadir, por apoyarme siempre y guiarme en el desarrollo de este trabajo.

A Dailén, Yordan, Abiague, Liesky, Walter, Percy y Rogelio Silverio por su ayuda incondicional.

A Palma, Dayana, Germán, Carlos, Maikel y su equipo, por ayudarme durante todo el tiempo en Matanzas.

A Yenis, Asdrubal, Lisdanay, Lisandra y mis compañeros del proyecto.

A mis profesores, especialmente a Dunia, Dainel y Liset.

A los compañeros con los que compartido todos los años de la universidad.

## Dedicatoria

A mis padres y mi hermana, por ser parte de mi corazón.

A la Revolución y a nuestro Comandante en Jefe, por la maravillosa idea de crear esta universidad.

## Resumen

En distintos niveles de las Fuerzas Armadas Revolucionarias los jefes deben tomar decisiones como parte de su trabajo cotidiano. Con el desarrollo tecnológico han surgido un conjunto de aplicaciones informáticas que gestionan las disímiles actividades que se realizan a diario. El presente trabajo propone la construcción de un almacén de datos que servirá de apoyo al proceso de toma de decisiones para los subsistemas de Reclutamiento y Potencial Humano. Los datos almacenados por la institución no tienen utilidad si no son convertidos en información útil para la toma de decisiones.

Este trabajo se encuentra inmerso en todo un proceso de inteligencia de negocios; para el desarrollo del mismo fue necesario realizar un análisis de las diversas necesidades de información para dar paso al diseño del almacén de datos, y finalmente se crearon y programaron un conjunto de procesos para la extracción, transformación y carga de los datos. También se realizaron un conjunto de pruebas de calidad a lo largo del desarrollo de la solución para garantizar su correcto funcionamiento.

PALABRAS CLAVES

Almacén de datos, Data Warehouse, Inteligencia de negocios, Hefesto, Pentaho, Procesos ELT

## Índice de contenidos

**INTRODUCCIÓN ..... 1**

**CAPÍTULO 1: FUNDAMENTACIÓN TEÓRICA. ..... 4**

INTRODUCCIÓN. ..... 4
1.1. INTELIGENCIA DE NEGOCIOS. ..... 4
1.1.1. HERRAMIENTAS DE INTELIGENCIA DE NEGOCIOS. ..... 5
1.1.2. APORTES DE LA INTELIGENCIA DE NEGOCIOS. ..... 6
1.2. ALMACÉN DE DATOS. ..... 6
1.2.1. DEFINICIÓN DE ALMACÉN DE DATOS. ..... 7
1.2.2. CARACTERÍSTICAS DE LOS ALMACENES DE DATOS. ..... 7
1.2.3. FUNCIONES DE LOS ALMACENES DE DATOS. ..... 9
1.2.4. ESTRUCTURA DE UN ALMACÉN DE DATOS. ..... 9
1.2.5. VENTAJAS DEL USO DE ALMACENES DE DATOS. ..... 10
1.2.6. DESVENTAJAS DEL USO DE ALMACENES DE DATOS. ..... 10
1.2.7. ALMACENES DE DATOS EN EL MUNDO. ..... 11
1.3. BASES DE DATOS MULTIDIMENSIONALES. ..... 12
1.3.1. ESQUEMA ESTRELLA. ..... 12
1.3.2. ESQUEMA COPO DE NIEVE. ..... 12
1.3.3. ESQUEMA CONSTELACIÓN. ..... 13
1.4. CUBOS MULTIDIMENSIONALES. ..... 14
1.5. PROCESAMIENTO ANALÍTICO EN LÍNEA (OLAP). ..... 15
1.6. INTEGRACIÓN DE DATOS. ..... 16
1.6.1. PROCESO DE EXTRACCIÓN, TRANSFORMACIÓN Y CARGA (ETL). ..... 16
1.6.2. HERRAMIENTAS ETL. ..... 17
1.7. METODOLOGÍAS PARA EL DESARROLLO DE UN ALMACÉN DE DATOS. ..... 19
1.8. OTROS CONCEPTOS DE INTERÉS. ..... 20
1.8.1. PROCESAMIENTO DE TRANSACCIONES EN LÍNEA (OLTP). ..... 20
1.8.2. DATA MART. ..... 20
1.8.3. MINERÍA DE DATOS. ..... 21
1.8.4. SISTEMAS DE SOPORTE A DECISIONES. ..... 21
1.8.5. SISTEMAS GESTORES DE BASES DE DATOS. ..... 21
1.9. METODOLOGÍA, TÉCNICAS Y HERRAMIENTAS PROPUESTAS PARA LA SOLUCIÓN. ..... 22
1.9.1. METODOLOGÍA. ..... 22
1.9.2. HERRAMIENTA DE MODELADO. ..... 24
1.9.3. GESTOR DE BASES DE DATOS. ..... 25
1.9.4. HERRAMIENTAS PARA LA ADMINISTRACIÓN DE BASES DE DATOS. ..... 25
1.9.5. PLATAFORMA DE INTELIGENCIA DE NEGOCIOS. ..... 25

1.9.6. HERRAMIENTAS ETL. ..... 25
1.9.7. SERVIDOR WEB. ..... 26
CONCLUSIONES. ..... 26

**CAPÍTULO 2: ANÁLISIS. ..... 27**

INTRODUCCIÓN. ..... 27
2.1. DESCRIPCIÓN DEL NEGOCIO. ..... 27
2.2. ANÁLISIS DE LOS REQUERIMIENTOS. ..... 28
2.2.1. PERSPECTIVAS E INDICADORES. ..... 29
2.2.2. MODELO CONCEPTUAL. ..... 30
2.3. ANÁLISIS DEL ESTADO DE LAS FUENTES DE DATOS. ..... 30
2.4. DETERMINACIÓN DE INDICADORES. ..... 31
2.5. CORRESPONDENCIAS. ..... 32
2.6. NIVEL DE GRANULARIDAD. ..... 33
2.7. PERFILADO DE DATOS. ..... 37
2.8. REGLAS PARA TRANSFORMACIONES DE DATOS. ..... 38
CONCLUSIONES. ..... 38

**CAPÍTULO 3: DISEÑO. ..... 40**

INTRODUCCIÓN. ..... 40
3.1. ARQUITECTURA DEL ALMACÉN DE DATOS. ..... 40
3.2. TIPO DE MODELO PARA EL DISEÑO DEL ALMACÉN DE DATOS. ..... 41
3.3. DEFINICIÓN DE ESTÁNDARES PARA OBJETOS FÍSICOS. ..... 42
3.4. IDENTIFICACIÓN DE DIMENSIONES. ..... 42
3.5. IDENTIFICACIÓN DE HECHOS. ..... 46
3.6. UNIONES ENTRE DIMENSIONES Y HECHOS. ..... 48
CONCLUSIONES. ..... 49

**CAPÍTULO 4: PROCESOS ETL Y PRUEBAS. ..... 50**

INTRODUCCIÓN. ..... 50
4.1. MAPEO DE DATOS. ..... 50
4.2. CONDICIONES ADICIONALES. ..... 51
4.3. ESTÁNDARES Y ACCIONES PREVIAS AL DISEÑO DEL PROCESO ETL. ..... 52
4.4. EXTRACCIÓN, TRANSFORMACIÓN Y CARGA DE LOS DATOS. ..... 52
4.5. CARGAS INCREMENTALES DE LOS DATOS. ..... 53
4.6. DISEÑO Y CONSTRUCCIÓN DE LA AUTOMATIZACIÓN DE LOS PROCESOS ETL. ..... 53
4.7. PRUEBAS DE CALIDAD. ..... 54
CONCLUSIONES. ..... 55

**CONCLUSIONES. ..... 56**

**RECOMENDACIONES. ..... 57**

# ÍNDICE DE CONTENIDOS

**BIBLIOGRAFÍA. ..... 58**
**REFERENCIAS BIBLIOGRÁFICAS. ..... 60**
**GLOSARIO DE TÉRMINOS. ..... 61**

## Índice de figuras

FIGURA 2.1. MODELO CONCEPTUAL. ..... 30
FIGURA 3.1. MODELO DEL ALMACÉN DE DATOS. ..... 49

## Índice de tablas

TABLA 1.1. ALMACENES DE DATOS VS. SISTEMAS TRADICIONALES..... 8
TABLA 2.1. DESCRIPCIÓN DE LOS DATOS DISPONIBLES PARA LA PERSPECTIVA “PERSONA”. ..... 34
TABLA 2.2. DESCRIPCIÓN DE LOS DATOS DISPONIBLES PARA LA PERSPECTIVA “ESTRUCTURA”. ..... 34
TABLA 2.3. DESCRIPCIÓN DE LOS DATOS DISPONIBLES PARA LA PERSPECTIVA “GENERACIÓN”..... 34
TABLA 2.4. DESCRIPCIÓN DE LOS DATOS DISPONIBLES PARA LA PERSPECTIVA "FUENTE DE INGRESO"..... 35
TABLA 2.5. DESCRIPCIÓN DE LOS DATOS DISPONIBLES PARA LA PERSPECTIVA "CAUSA NO INSCRITO". ..... 35
TABLA 2.6. DESCRIPCIÓN DE LOS DATOS DISPONIBLES PARA LA PERSPECTIVA "PRERRECLUTA"..... 35
TABLA 2.7. DESCRIPCIÓN DE LOS DATOS DISPONIBLES PARA LA PERSPECTIVA "SITUACIÓN EN EL REGISTRO MILITAR". ..... 36
TABLA 2.8. ANÁLISIS DE LOS DATOS. ..... 37
TABLA 3.1. DESCRIPCIÓN DE LA ENTIDAD "*DIM_PERSONA*"..... 43
TABLA 3.2. DESCRIPCIÓN DE LA ENTIDAD "*DIM_GENERACION*". ..... 43
TABLA 3.3. DESCRIPCIÓN DE LA ENTIDAD "*DIM_ESTRUCTURA*". ..... 44
TABLA 3.4. DESCRIPCIÓN DE LA ENTIDAD "*DIM_FECHA*". ..... 44
TABLA 3.5. DESCRIPCIÓN DE LA ENTIDAD "*DIM_FUENTEINGRESO*". ..... 44
TABLA 3.6. DESCRIPCIÓN DE LA ENTIDAD "*DIM_CAUSANOINSC*"..... 45
TABLA 3.7. DESCRIPCIÓN DE LA ENTIDAD "*DIM_PRERRECLUTA*". ..... 45
TABLA 3.8. DESCRIPCIÓN DE LA ENTIDAD "*DIM_SITUACIONRM*"..... 45
TABLA 3.9. DESCRIPCIÓN DE LA ENTIDAD "*CUB_LISTADOUNICO*". ..... 46
TABLA 3.10. DESCRIPCIÓN DE LA ENTIDAD "*CUB_INSCRITOS*". ..... 47
TABLA 3.11. DESCRIPCIÓN DE LA ENTIDAD "*CUB_NOINSCRITOS*". ..... 47
TABLA 3.12. DESCRIPCIÓN DE LA ENTIDAD "*CUB_SGC*"..... 48
TABLA 4.1. MAPEO DE DATOS. ..... 51

## Introducción

La información es un fenómeno que proporciona significado a las cosas, o puede verse también como un conjunto organizado de datos procesados que a su vez constituyen un mensaje sobre determinada situación. En la edad media el almacenamiento, acceso y uso limitado de la información se realizaba en las bibliotecas de los monasterios. Actualmente, en un corto período de tiempo el mundo desarrollado se ha propuesto lograr la globalización del acceso a los enormes volúmenes de información existentes en medios cada vez más complejos, con capacidades ascendentes de almacenamiento y en soportes más reducidos. Los datos se perciben, se integran y generan la información necesaria para producir el conocimiento que finalmente permite tomar decisiones. Hoy en día las empresas buscan obtener la mayor cantidad de información posible para el beneficio de sus negocios. El análisis de la información con que se cuenta es fundamental para que la misma logre un alto nivel de competitividad y posibilidades de desarrollo, por lo que se hace imprescindible el uso de sistemas informáticos que sean capaces de procesar los datos y la información de la manera más adecuada posible en función de sus objetivos.

La toma de decisiones en cualquier empresa es un proceso complejo debido a la repercusión que puede tener, más aún si se trata de una institución que influye directamente en la economía de una nación. El análisis de la información es una de las bases para el proceso de toma de decisiones, por lo que, cuanto más información útil tenga una empresa en su poder y mejor calidad de análisis, mayor serán las probabilidades de alcanzar el éxito.

En los últimos años, se está llevando a cabo en las Fuerzas Armadas Revolucionarias un proceso de automatización en todos los niveles, contando hoy con un grupo de sistemas informáticos que automatizan diversos procesos en la institución. Debido al volumen de datos generados por estos, se hace necesario el uso de herramientas que generen la información necesaria y permitan su análisis posterior para apoyar la toma de decisiones de los dirigentes. Uno de los sistemas con los que se cuenta actualmente es el Sistema Automatizado de los Comités Militares **DATAFAR** que gestiona los procesos de Reclutamiento y Potencial Humano. Debido al tiempo que el mismo lleva en explotación contiene gran cantidad de datos almacenados, los cuales pueden ser analizados de forma limitada, por lo que se requiere el uso de herramientas de análisis de información para el apoyo a la toma de decisiones.

# INTRODUCCIÓN

Luego de un estudio de la situación antes reflejada se plantea el siguiente **problema a resolver**: ¿Cómo mejorar el proceso de toma de decisiones a partir de los datos de los subsistemas de Reclutamiento y Potencial Humano?

El problema planteado anteriormente tiene como **objeto de estudio** el proceso de análisis de información, y el **campo de acción** abarca el proceso de toma de decisiones a mediano y largo plazo para los subsistemas Reclutamiento y Potencial Humano.

Para guiar el desarrollo del trabajo se plantea la siguiente **idea a defender**: Con el desarrollo de un almacén de datos para los subsistemas Reclutamiento y Potencial Humano, se obtendrá una herramienta funcional para el análisis de la información, logrando mayor eficiencia en los procesos para la toma de decisiones.

El **objetivo general** de este trabajo es desarrollar un almacén de datos para los subsistemas de Reclutamiento y Potencial Humano, que mejore el análisis de la información durante el proceso de toma de decisiones.

Para un mejor desarrollo de la investigación se plantean los siguientes **objetivos específicos**:

- Obtener el diseño teórico de la investigación.
- Obtener el estado y tendencias actuales de las metodologías y herramientas para el desarrollo de un almacén de datos.
- Comprender las necesidades de información a partir del análisis de los requerimientos.
- Conocer mediante un análisis las características y el estado que presentan las fuentes de datos.
- Lograr el diseño del almacén de datos.
- Lograr la correcta ejecución de los procesos de extracción, transformación y carga.
- Garantizar la calidad requerida del producto mediante un conjunto de pruebas de calidad.

La implantación de este almacén de datos proporcionará los siguientes **aportes prácticos**:

- Integrará y consolidará diferentes fuentes de datos y departamentos en la institución.
- Mejorará la entrega de la información.
- Permitirá la toma de decisiones estratégicas y tácticas.

El presente documento se encuentra estructurado en cuatro capítulos:

# INTRODUCCIÓN

Capítulo 1: En este capítulo se realiza una breve explicación del estado del arte de las metodologías, técnicas y herramientas que se tuvieron en cuenta para dar solución al problema; así como los principales conceptos y características de las mismas.

Capítulo 2: En este capítulo se plasma el análisis realizado para llevar a cabo la solución propuesta, donde se identifican las principales necesidades de información del usuario, así como las características y el estado actual de las fuentes de datos.

Capítulo 3: El capítulo presenta los aspectos relacionados con el modelado y diseño de la solución propuesta, además se explican algunas características y conceptos relevantes que se tuvieron en cuenta.

Capítulo 4: Este capítulo plasma todo lo relacionado al diseño y construcción de los procesos de extracción, transformación y carga de los datos, y un conjunto de las pruebas de calidad efectuadas con el fin de garantizar la calidad de la solución propuesta.

# Capítulo 1: Fundamentación teórica.

**Introducción.**

El creciente interés por parte de las grandes empresas del uso de sistemas que permitan el análisis de la información y apoyen la toma de decisiones ha dado paso al surgimiento de nuevas tendencias, concepciones, métodos y herramientas. El Ministerio de las Fuerzas Armadas Revolucionarias cuenta ya con gran cantidad de información producto de las aplicaciones informáticas que posee y necesita de estos sistemas de análisis.

En este capítulo se refleja el resultado de un estudio a escala nacional e internacional del estado y las tendencias de las metodologías y herramientas existentes que se tuvieron en cuenta para dar solución al problema.

## 1.1. Inteligencia de negocios.

La inteligencia de negocios o también llamada inteligencia empresarial no es más que el conjunto de estrategias y herramientas dirigidas a la administración y creación de conocimiento mediante el análisis de los datos de una organización, con el fin de obtener una ventaja competitiva. Este término abarca la comprensión del funcionamiento actual de la institución, así como la anticipación a eventos futuros, con el objetivo de apoyar la toma de decisiones.

El complejo proceso de creación de inteligencia de negocios se encuentra dividido en cinco fases o etapas, para una mayor comprensión se explica a continuación cada una de estas.

En la primera fase, denominada “Dirigir y Planear”, se debe capturar los requerimientos de información de los diferentes usuarios, también se debe entender claramente las diversas necesidades del mismo para luego formular las preguntas que ayudarán a alcanzar sus objetivos.

Es en la segunda fase “Identificación de la información” donde se realiza el proceso de extracción de los datos de las distintas fuentes de información de la empresa que son necesarios para dar respuestas a las preguntas generadas en la etapa anterior.

En la tercera fase, llamada “Procesamiento de datos”, se integran y cargan los datos en crudo en un formato utilizable para el análisis. Esta actividad se puede realizar agregando los datos a una base de datos existente, también se puede crear una base de datos nueva o consolidando la información.

En la cuarta fase, denominada “Análisis y Producción”, se procede al trabajo sobre los datos ya extraídos e integrados, utilizando un conjunto de herramientas y técnicas propias de inteligencia de negocios para crear el conocimiento. El resultado de esta fase es la respuesta a las preguntas formuladas en la primera etapa mediante la creación de reportes, indicadores, etc.

En la quinta y última fase llamada “Difusión” se entrega a los usuarios que lo requieran las herramientas necesarias para explotar los datos de manera sencilla y eficaz. [1]

### 1.1.1. Herramientas de inteligencia de negocios.

Las herramientas de inteligencia de negocios son aplicaciones diseñadas con el fin de colaborar con la inteligencia de negocios en los procesos de las empresas. Estas herramientas se basan fundamentalmente en la asistencia a los análisis y la presentación de los datos al usuario.

Existen varios tipos de herramientas de inteligencia de negocios, de acuerdo con el nivel de complejidad, estas soluciones se pueden clasificar en aplicaciones de reporte con consultas e informes simples, cubos OLAP (Procesamiento Analítico en Línea), minería de datos y sistemas de previsión empresarial.

Entre las características principales de este conjunto de herramientas se destacan la accesibilidad de la información, el apoyo a la toma de decisiones y la orientación al usuario final.

En la actualidad existen innumerables herramientas para la inteligencia de negocios, entre los productos de fuente abierta se destacan Freereporting.com, Eclipse BIRT Project y Pentaho; por otra parte, entre los productos comerciales, también se destacan IBM Cognos, Crystal Reports, Oracle Business Intelligence Suite y Microsoft Business Intelligence.

Crystal Reports es una aplicación de inteligencia de negocios, la cual permite crear con facilidad informes interactivos. Esta aplicación posee una gran facilidad de integración, principalmente con componentes diseñados en Flash y Adobe Flex. [2]

Oracle Business Intelligence Suite es una plataforma de las más completas en la actualidad para inteligencia de negocios, cubre un amplio espectro de necesidades de inteligencia de negocios, incluidos los tableros interactivos, alertas e inteligencia proactiva, publicación e informes avanzados, análisis predictivos en tiempo real y de tecnología móvil, entre otras más. [3]

Microsoft Business Intelligence es una suite completa de productos integrados, que proporciona acceso continuo a aplicaciones de amplia difusión e informes, de esta forma, brinda cobertura a los aspectos del proceso de toma de decisiones. Está basado en la plataforma de inteligencia de negocio de Microsoft SQL Server y tiene una gran difusión entre el público a través de Microsoft Office. [4]

La plataforma Pentaho Open Source Business Intelligence cubre las necesidades de análisis de los datos y de los informes empresariales. Las soluciones de Pentaho están escritas en Java y su ambiente de implementación también está basado en Java. Esto hace de Pentaho una solución flexible que cubre una amplia gama de necesidades empresariales. Pentaho cuenta con diversos módulos para reportes, análisis, minería de datos e integración de datos. [5]

### 1.1.2. Aportes de la inteligencia de negocios.

Entre los aportes más significativos que la inteligencia de negocios proporciona a las organizaciones, se destacan:

- Reduce el tiempo mínimo que se requiere para recoger toda la información relevante del negocio.
- Automatiza la asimilación de la información.
- Proporciona herramientas de análisis para establecer comparaciones y tomar decisiones.
- Permite al usuario no depender de reportes programados.
- Posibilita la formulación de preguntas y respuestas claves en la institución.

## 1.2. Almacén de datos.

En el proceso de inteligencia de negocios es necesario gestionar los datos almacenados en diversos formatos y fuentes para luego ser integrados, también es fundamental almacenar los mismos en un mismo destino que permita el posterior análisis de estos; por tanto, es muy importante el uso de una herramienta que cubra estas necesidades. Un almacén de datos o *data warehouse* en inglés, es la

herramienta que responde a estos intereses, básicamente se encarga de consolidar, integrar y centralizar los datos de la institución.

### 1.2.1. Definición de almacén de datos.

Uno de los primeros autores en escribir sobre el tema de almacenes de datos fue Bill Inmon, quien lo define en términos de las características de un repositorio de datos orientado a temas, variante en el tiempo, no volátil e integrado. Otro reconocido autor es Ralph Kimball define un almacén de datos como una copia de las transacciones de datos específicamente para la consulta y el análisis.

Se puede definir a un almacén de datos como una colección de datos orientados a una determinada empresa u organización, integrado, no volátil y variante en el tiempo que facilita el análisis de la información para la toma de decisiones.

### 1.2.2. Características de los almacenes de datos.

Para una mayor comprensión de la definición anterior podemos definir los siguientes conceptos:

- Orientado a temas: En la base de datos, los datos están organizados de manera que los elementos de datos relativos al mismo evento u objeto del mundo real queden unidos entre sí.
- Variante en el tiempo: Los cambios producidos en los datos a lo largo del tiempo quedan registrados para que los informes que se puedan generar reflejen esas variaciones.
- No volátil: La información no se modifica ni se elimina, una vez almacenado el dato, este se convierte en información de solo lectura, y se mantiene para futuras consultas.
- Integrado: La base de datos contiene los datos de todos los sistemas operacionales de la organización, estos datos deben de ser consistentes.

Una de las principales características de los almacenes de datos, es que estos son capaces de manejar un gran volumen de datos debido a que consolida en su estructura la información recolectada durante años, proveniente de diversas fuentes en un solo lugar centralizado; por esta razón es que el depósito puede ser soportado y mantenido sobre diversos medios de almacenamiento. También, como ya se había mencionado anteriormente los almacenes de datos manejan información histórica. Tienen la capacidad de organizar y almacenar los datos que se necesitan para el procesamiento analítico e informático, con el propósito de responder a preguntas de negocios y brindarles a los usuarios finales una interfaz amigable, comprensible y fácil de utilizar para

que los mismos puedan tomar decisiones sobre los datos sin tener conocimientos avanzados de informática.

El almacén de datos permite un acceso más directo a la información, ya que esta gira en torno al negocio, por lo que facilita al usuario realizar una exploración de los datos y encontrar relaciones complejas entre ellos.

Debido a que el almacén de datos recibe información histórica de diferentes fuentes, se podría suponer que existe una repetición masiva de datos entre el ambiente del almacén de datos y el operacional; este razonamiento es superficial y erróneo aunque puede existir una mínima redundancia entre estos ambientes. Se debe considerar que los datos en el ambiente operacional se filtran antes de pertenecer al almacén de datos, por lo que existen muchos datos que no ingresarán debido a que no son parte de la información necesaria para la toma de decisiones. Los datos experimentan una considerable transformación antes de ser cargados al almacén de datos. Luego de analizar estos factores, se puede afirmar que la redundancia encontrada es mínima ya que generalmente tiene un porcentaje menor del 1%. [1]

Las diferencias entre un almacén de datos y un sistema tradicional se pueden resumir en la siguiente tabla:

| Sistema tradicional | Almacén de datos |
|---|---|
| Predomina la actualización. | Predomina la consulta. |
| La actividad fundamental es de tipo operativo. | La actividad fundamental es el análisis y la decisión estratégica. |
| Mayor importancia a la estabilidad. | Mayor importancia al dinamismo. |
| Datos en general desagregados. | Datos en distintos niveles de detalle y agregación. |
| Importancia del dato actual. | Importancia del dato histórico. |
| Estructura relacional. | Vista multidimensional. |
| Usuarios de perfiles medios y bajos. | Usuarios de perfiles altos. |
| Explotación de la información relacionada con la operativa de cada aplicación. | Explotación de la información interna y externa relacionada con el negocio. |

**Tabla 1.1. Almacenes de datos vs. sistemas tradicionales.**

### 1.2.3. Funciones de los almacenes de datos.

El objetivo en un almacén de datos es contener datos que son necesarios o útiles para una organización, o sea, se utiliza como un repositorio de datos para posteriormente transformarlos en información útil para el usuario. Un almacén de datos debe entregar la información correcta a la persona indicada en el momento oportuno y en el formato adecuado. Los almacenes de datos dan respuesta a las necesidades de los usuarios expertos, utilizando Sistemas de Soporte a Decisiones, Sistemas de Información Ejecutiva o herramientas para hacer consultas o informes. Los usuarios finales pueden crear consultas sobre el almacén de datos sin afectar la operación del sistema.

Durante el funcionamiento de un almacén de datos es muy importante la integración de los datos provenientes de bases de datos distribuidas por la organización y que con frecuencia tendrán diferentes estructuras; también la separación de los datos usados en operaciones diarias de los usados en el almacén de datos para los propósitos de divulgación, de ayuda en la toma de decisiones, para el análisis y para operaciones de control, estos tipos de datos no deben coincidir en la misma base de datos, ya que cumplen objetivos distintos.

Frecuentemente se importan datos de distintas fuentes relacionadas con el negocio para luego transformarlos. Es usual normalizar los datos antes de combinarlos en el almacén de datos mediante el uso de herramientas de extracción, transformación y carga. [1]

### 1.2.4. Estructura de un almacén de datos.

Los almacenes de datos estructuran los datos de una particular manera y existen diferentes niveles de esquematización y detalles que los delimitan. Los mismos están compuestos por diversos tipos de datos, que se organizan y dividen de acuerdo con el nivel de detalle o granularidad que posean. A continuación se explican cada uno de estos tipos de datos:

- **Detalle de datos actuales.** Son aquellos que reflejan las ocurrencias más recientes. Generalmente se almacenan en disco, aunque su administración sea compleja y costosa, con el objetivo de lograr que el acceso a la información sea rápido y sencillo debido a su volumen. El gran tamaño está dado a que los datos residentes poseen el más bajo nivel de granularidad.
- **Detalle de datos históricos.** Representan aquellos datos antiguos que no son frecuentemente consultados, también se almacena a nivel de detalle, generalmente sobre alguna forma de almacenamiento externa.

- **Datos ligeramente resumidos.** Son los datos que provienen desde un bajo nivel de detalle y se sumarizan o agrupan bajo algún criterio o condición de análisis.
- **Datos altamente resumidos.** Son aquellos que compactan aún más a los datos ligeramente resumidos.

### 1.2.5. Ventajas del uso de almacenes de datos.

A continuación, aparecen un conjunto de ventajas que son facilitadas por el uso de un almacén de datos en una institución:

- Transforma los datos orientados a las aplicaciones en información orientada a la toma de decisiones.
- Integra y consolida diferentes fuentes de datos y departamentos empresariales en una única plataforma sólida y centralizada.
- Provee la capacidad de analizar y explotar diferentes áreas de trabajo y de realizar un análisis inmediato de las mismas.
- Permite reaccionar rápidamente ante los cambios.
- Aumenta la competitividad de los responsables de la toma de decisiones.
- Elimina la producción y el procesamiento de datos que no son utilizados ni necesarios, producto a aplicaciones mal diseñadas o no utilizadas.
- Mejora la entrega de la información que los usuarios necesitan, una información completa, correcta, consistente y oportuna.
- Logra un impacto positivo sobre los procesos empresariales. Cuando los usuarios tienen acceso a una información de mayor calidad, la empresa puede lograr aprovechar el valor potencial de sus recursos de información y transformarlo en valor verdadero, así como eliminar retardos en los procesos empresariales e integrar y optimizar procesos.
- Permite la toma de decisiones estratégicas y tácticas.
- Los almacenes de datos pueden trabajar en conjunto y aumentar el valor operacional de las aplicaciones empresariales, en especial la gestión de relaciones con clientes.

### 1.2.6. Desventajas del uso de almacenes de datos.

Es importante conocer las desventajas que trae consigo el uso de los almacenes de datos, de las cuales se mencionan algunas a continuación:

- Requiere una gran inversión, debido a que su construcción no es tarea sencilla y consume muchos recursos, además su implementación implica la adquisición de herramientas de consulta y análisis y la capacitación de los usuarios.
- Generalmente existe resistencia al cambio por parte de los usuarios.
- Generalmente los beneficios de un almacén de datos son realmente apreciados en el mediano y largo plazo, derivado del punto anterior ya que algunos usuarios no confían en el almacén de datos desde una primera instancia.
- Si se incluyen datos personales o confidenciales de los clientes atentará con la privacidad de los mismos, ya que cualquier usuario podrá acceder a los datos.
- Infravaloración de los recursos necesarios para la captura, carga y almacenamiento de los datos.
- Infravaloración del esfuerzo necesario para su diseño y creación.
- Incremento continuo de los requerimientos de los usuarios.
- Los almacenes de datos se pueden quedar obsoletos relativamente pronto.

### 1.2.7. Almacenes de datos en el mundo.

Muchas empresas en el mundo utilizan los almacenes de datos como vía para mejorar la toma de decisiones, en la mayoría de estas se han experimentado resultados positivos. Ejemplo de las mismas son Jazztel y France Telecom, ambas son empresas de telecomunicaciones, también se destacan por su uso prestigiosas aerolíneas como British Airways y Air France, y otras prestigiosas empresas como son Coca Cola, Adidas, y Nike. La característica común de todas estas empresas es el gran volumen de información que se genera, principalmente en lo que respecta a los clientes finales.

En nuestro país, algunas instituciones han comenzado a hacer uso de las ventajas de un almacén de datos como es el caso de la corporación CIMEX y la Oficina Nacional de Estadísticas (ONE), ambos almacenes de datos desarrollados en la Universidad de las Ciencias Informáticas.

Las empresas que utilizan almacenes de datos generalmente son aquellas que manejan grandes volúmenes de datos relativos a clientes, marketing, transacciones y operaciones, por lo que es muy difícil encontrar pequeñas y medianas empresas que utilicen estos sistemas.

## 1.3. Bases de datos multidimensionales.

Las bases de datos multidimensionales, proveen una estructura que permite, mediante la creación y consulta a una estructura de datos determinada tener acceso flexible a los datos para explorar y analizar sus relaciones y resultados.

Las bases de datos multidimensionales implican tres variantes posibles de modelación, las cuales se mencionan a continuación:

- Esquema estrella.
- Esquema copo de nieve.
- Esquema constelación.

### 1.3.1. Esquema estrella.

El esquema estrella o *Star Scheme* en inglés, es la arquitectura más simple de un almacén de datos. El esquema estrella consta de una tabla de variables o tabla de hechos central y de varias tablas de dimensiones relacionadas a esta a través de sus claves respectivas.

Este modelo debe estar totalmente desnormalizado, es decir, no puede presentarse en tercera forma normal (3ra. FN). Cuando se normaliza, se pretende eliminar la redundancia, la repetición de datos y que las claves sean independientes de las columnas. La ventaja que proporciona la desnormalización es obviar las uniones entre las tablas cuando se realizan consultas, logrando de esta manera un mejor tiempo de respuesta y una mayor sencillez con respecto a su utilización y como desventaja genera un cierto grado de redundancia. [1]

A continuación se mencionan algunas de las características de este modelo:

- Posee mejores tiempos de respuesta.
- Su diseño es fácilmente modificable.
- Existe paralelismo entre su diseño y la forma en que los usuarios visualizan y manipulan los datos.
- Facilita la interacción con herramientas de consulta y análisis.

### 1.3.2. Esquema copo de nieve.

El esquema copo de nieve o *Snowflake Scheme* en inglés representa una extensión del modelo en estrella cuando las tablas se organizan en jerarquías de dimensiones. Este esquema presenta una

tabla de hechos central que está relacionada con una o más tablas de dimensiones, las cuales a su vez pueden estar relacionadas o no con una o más tablas de dimensiones. [1]

Este modelo es más cercano a un modelo entidad relación, que al modelo en estrella, debido a que sus tablas de dimensiones están normalizadas.

Un motivo para utilizar este tipo de modelo es la posibilidad de separar los datos de las tablas de dimensiones y proveer un esquema que sustente los requerimientos del diseño. Otra razón es que es muy flexible y puede implementarse después de haber desarrollado un esquema estrella.

Para este tipo de modelo se puede destacar las siguientes características:

- Posee mayor complejidad en su estructura.
- Hace una mejor utilización del espacio.
- Puede desarrollar clases de jerarquías fuera de las tablas de dimensiones, que permiten realizar un análisis de lo general a lo detallado y viceversa.
- Las tablas de dimensiones están normalizadas.
- Es muy útil en tablas de dimensiones de muchas tuplas.

### 1.3.3. Esquema constelación.

El esquema constelación o *Starflake Scheme* está compuesto por una serie de esquemas en estrellas, y el mismo está formado por una tabla de hechos principal y por una o más tablas de hechos auxiliares, las cuales pueden ser sumarizaciones de la principal. Estas tablas están relacionadas con sus respectivas tablas de dimensiones. [1]

No es necesario que las tablas de hechos compartan las mismas tablas de dimensiones, ya que las tablas de hechos auxiliares pueden vincularse con solo algunas tablas de dimensiones asignadas de la tabla de hechos principal, y también pueden hacerlo con nuevas tablas de dimensiones.

Su diseño y cualidades son muy similares a las del esquema estrella, pero tiene un conjunto de características que lo destacan y caracterizan, de las cuales se puede mencionar:

- Permite tener más de una tabla de hechos, por lo que se podrán analizar más aspectos claves del negocio como un mínimo esfuerzo adicional del diseño.
- Contribuye a la reutilización de las tablas de dimensiones, ya que la misma tabla de dimensión puede utilizarse para varias tablas de hechos.

- No es soportado por todas las herramientas de consulta y análisis.

### 1.4. Cubos multidimensionales.

Un cubo multidimensional o hipercubo, representa o convierte los datos planos que se encuentran en las filas y columnas en una matriz de n dimensiones. Los objetos más importantes que se pueden incluir en un cubo multidimensional, son los siguientes:

- Indicadores. Son las sumarizaciones que se efectúan sobre algún hecho, perteneciente a una tabla de hechos.
- Atributos. Son campos o criterios de análisis, que pertenecen a las tablas de dimensiones.
- Jerarquías. Representa una relación lógica entre dos o más atributos.

De esta forma, en un cubo multidimensional los atributos existen a lo largo de varios ejes o dimensiones, y la intersección de las mismas representa el valor que tomará el indicador que se está evaluando.

Por ejemplo, dada una relación de orden N, se considera la posibilidad de proyección que dispone de los campos X, Y, Z, como clave de la relación y W como atributo residual. Categorizando lo anterior como una función obtenemos:

$W: (X, Y, Z) \rightarrow W$

Los atributos X, Y, Z, se corresponden con los ejes de cubo, mientras que el valor de W devuelto por cada tripleta (X, Y, Z) se corresponde con el elemento que se ocupa en cada celda del cubo.

Cada una de las dimensiones de un cubo multidimensional puede resumirse mediante una jerarquía. Vincular o enlazar los cubos es un mecanismo para superar la dispersión, la cual se produce cuando no todas las celdas del cubo no están llenas con datos. El tiempo de procesamiento es muy valioso por lo que se debe adoptar la manera más efectiva de procesar los valores nulos. En ocasiones es mejor crear un cubo distinto en lugar de uno disperso, pero vinculado donde un subconjunto de datos se puede analizar con gran detalle. Esta vinculación asegura que los datos de los dos cubos mantengan coherencia. [1]

## 1.5. Procesamiento analítico en línea (OLAP).

OLAP es el acrónimo en inglés de procesamiento analítico en línea (*On-Line Analytical Processing*), es una solución empleada en el campo de la inteligencia empresarial, con el objetivo de agilizar la consulta de grandes cantidades de datos. Para esto se utilizan estructuras multidimensionales o cubos OLAP que contienen datos resumidos de grandes bases de datos o sistemas transaccionales (OLTP). Su uso se enfoca en informes de negocios, marketing, informes de dirección, minería de datos y otras áreas similares; entonces en la base de cualquier sistema OLAP se encuentra el concepto de cubo OLAP.

Los sistemas OLAP se clasifican en las siguientes categorías:

- ROLAP: Procesamiento Analítico Relacional en Línea (*Relational OnLine Analytical Processing*), trata sobre sistemas y herramientas OLAP que son construidos sobre una base de datos relacional. La arquitectura está compuesta por un servidor de base de datos relacional y el motor OLAP que se encuentra en un servidor dedicado. La ventaja de esta arquitectura es que permite el análisis de una gran cantidad de datos.
- MOLAP: Procesamiento Analítico Multidimensional en Línea (*Multidimensional OnLine Analytical Processing*), es una alternativa a la tecnología ROLAP, aunque estas están diseñadas para realizar análisis de datos a través de un modelo de datos dimensional. La diferencia significativa está en que MOLAP requiere un procesamiento y almacenamiento de la información contenida en el cubo OLAP (cubo multidimensional). MOLAP almacena los datos en una matriz de almacenamiento multidimensional optimizada.
- HOLAP: Procesamiento Analítico Híbrido en Línea (*Hybrid OnLine Analytical Processing*), es una combinación de ROLAP y MOLAP. Este permite almacenar una parte de los datos como en un sistema MOLAP y el resto como en uno ROLAP. El grado de control que el operador de la aplicación tiene sobre este particionamiento varía de unos productos a otros. Los particionamientos se pueden clasificar en vertical y horizontal.

También existen otros tipos de sistemas OLAP, pero no son generalizados como los mencionados anteriormente, estos son:

- Web OLAP (WOLAP) está basado u orientado para la web.
- Desktop OLAP (DOLAP) basado en escritorio.
- Real Time OLAP (RTOLAP): OLAP en tiempo real.

- Spatial OLAP (SOLAP): OLAP espacial.

### 1.6. Integración de datos.

La integración de datos consiste en la combinación de los datos que residen en diferentes fuentes y brindar a los usuarios una vista unificada de los mismos. Esta integración se puede lograr mediante los procesos de extracción, transformación y carga.

#### 1.6.1. Proceso de extracción, transformación y carga (ETL).

El término ETL son las siglas en inglés de extraer, transformar y cargar (*Extract-Transform-Load*), y se refiere a los datos de una empresa. El proceso ETL organiza el flujo de datos entre diferentes sistemas y aporta los métodos y herramientas necesarias para mover los datos de múltiples fuentes, reformatearlos, limpiarlos y cargarlos en una base de datos, data mart o almacén de datos. Este proceso forma parte de la inteligencia empresarial, y está enfocado en la integración de datos. El proceso ETL fortalece los datos para la construcción de bases de datos permanentes dedicadas al análisis o la generación de informes, también es utilizado para conversión de bases de datos de un tipo o formato a otro, entre otras funciones que forman parte de este proceso. [6]

Las funciones específicas del proceso ETL son: extracción, transformación y carga.

**Extracción**

Esta función del proceso ETL consiste en explorar las fuentes OLTP que la empresa tenga a disposición y extraer los datos necesarios de las fuentes antes mencionadas, basándose en las necesidades y requisitos de los usuarios. Las fuentes pueden encontrarse sobre arquitecturas o estructuras heterogéneas, cada sistema puede usar una organización diferente de los datos o diferentes formatos. Por otro lado, los formatos de las fuentes normalmente se encuentran en bases de datos relacionales o ficheros planos, aunque se pueden incluir bases de datos no relacionales u otras estructuras diferentes. Una vez que los datos son seleccionados y extraídos, se guardan en un almacenamiento intermedio, lo que permite manipular los datos sin interrumpir el funcionamiento de los OLTP, ni el del almacén de datos. También permite almacenar y gestionar los metadatos que se generan en los procesos ETL y facilita la integración de las diversas fuentes de datos. Las herramientas utilizadas en la extracción de datos deben ser adaptables, extensibles y capaces de

filtrar los datos relevantes de las fuentes, permitiendo la compresión, descompresión y encriptación de datos. [1]

**Limpieza y Transformación**

La función de transformación es la encargada de convertir aquellos datos extraídos en un conjunto de datos compatibles y congruentes con el repositorio destino, sin perder su veracidad con respecto a las fuentes. Estas acciones se llevan a cabo debido a que pueden existir diferentes fuentes de información, y es importante conciliar un formato y una forma única, definiendo estándares para que todos los datos que ingresan al almacén de datos estén integrados. En esta etapa del proceso ETL los datos deben ser limpiados, debido a que con frecuencia se muestran incompletos, contienen errores o valores fuera de límites, manteniendo discrepancias en nombres o códigos. Para eliminar estos problemas se realiza un proceso de limpieza de datos cuyo objetivo principal es realizar distintos tipos de acciones contra la mayor cantidad de datos erróneos, inconsistentes e irrelevantes. Es importante identificar la anomalía a la hora de elegir una acción, para lograr actuar en consecuencia, con el fin de agregar más valor a los datos de la organización. Una vez que los datos son limpiados se realizan las transformaciones. [1]

**Carga**

Esta función es la responsable de cargar la estructura del almacén de datos mediante la transferencia de aquellos datos que han sido transformados y residen en el almacenamiento intermedio, y los datos de los OLTP que tienen correspondencia directa con el depósito de datos. Es preciso tener en cuenta que los datos antes de ser movidos deben de ser analizados con el objetivo de asegurar la calidad. [1]

### 1.6.2. Herramientas ETL.

Es una herramienta o un conjunto de herramientas que permiten desarrollar el proceso ETL. A la hora de seleccionar las mismas, se debe tener en cuenta un conjunto de características:

- Multiplataforma: la herramienta debe de funcionar en cualquier plataforma, aunque basta con que sea compatible con la plataforma seleccionada.
- Independencia del tipo de fuente o destino: la herramienta debe ser capaz de leer y escribir directamente desde y hacia las fuentes o destinos, independientemente de su tipo.

- Soporte para metadatos: la herramienta debe tener disponible la información sobre todos los datos durante el desarrollo y ejecución de los procesos.
- Soporte funcional: debe ser posible la realización eficiente de operaciones para la limpieza de los datos, transformaciones, agregaciones, reorganización y carga.
- Soporte al modelo dimensional: la herramienta debe tener incorporado soporte para la creación de tareas de dimensiones lentamente cambiantes, generación de llaves sustitutas y construcción de dimensiones agregadas.
- Paralelismo: debe posibilitar la ejecución de actividades en paralelo.
- Facilidad de uso: la herramienta debe ser de propósito general y amigable, de forma tal que el usuario pueda identificarse con la misma.
- Corrección de errores y registro de eventos: debe ser posible rastrear los errores en las transformaciones en tiempo de ejecución, así como ver los datos antes y después de efectuar las mismas También debe controlar el proceso registrando los eventos durante la ejecución.
- Planificación de la ejecución: la herramienta debe ofrecer una manera de planificar la ejecución de los trabajos de forma automática.
- Reusabilidad: la herramienta debe aportar la necesidad de aprovechar parte de la lógica de tareas anteriormente solucionadas.
- Extensibilidad: debe permitir al usuario definir nuevas funciones y utilizarlas al igual que las herramientas.
- Perfil de datos: la herramienta debe permitir realizar el perfilado de datos a las fuentes.

También es necesario tener en cuenta otras características como son los requerimientos de hardware y de software, documentación, preparación y soporte técnico, gestión de la calidad de datos, gestión de las dimensiones lentamente cambiantes y gestión de la sustitución de llaves, así como las utilidades del sistema operativo y servicios de transporte. En la actualidad los desarrolladores de estas herramientas centran su objetivo en estas características, que sin dudas continuará creciendo en el futuro. [6]

A la hora de realizar la selección de alguna de estas herramientas se podría tener en cuenta varios productos. Entre las herramientas y aplicaciones ETL más importantes del mercado se encuentra IBM InfoSphere DataStage es el componente base de IBM WebSphere Data Integration Suite, esta herramienta ofrece tres funciones claves y necesarias para conseguir una correcta integración de

datos empresariales: la conectividad global para acceder con rapidez y facilidad a cualquier origen o destino; herramientas avanzadas de desarrollo y mantenimiento, que agilizan la implementación y simplifican la administración; y una plataforma escalable que permite gestionar los actuales volúmenes masivos de datos empresariales. [7]

Otra herramienta es Microsoft Integration Services, que es una plataforma para la creación de soluciones empresariales de transformación e integración de datos. Entre las principales funciones de Integration Services se destacan la limpieza y minería de datos, actualización de almacenes de datos y la administración de objetos y datos de SQL Server. [8]

Oracle Warehouse Builder es una herramienta completa para el proceso de extracción transformación y carga, así como el modelado relacional y dimensional, la calidad de los datos y la gestión del ciclo de vida de los datos y los metadatos. [3]

Kettle Pentaho Data Integration es una potente herramienta de extracción transformación y carga que permite la integración de ambientes y datos para soportar las áreas de negocio. Esta herramienta cuenta con una interfaz gráfica e intuitiva, probada y escalable con muchas facilidades para los usuarios. Una de las principales características es que permite realizar transformaciones complejas sin tener que generar algún código personalizado. [5]

### 1.7. Metodologías para el desarrollo de un almacén de datos.

En la actualidad existen numerosas metodologías para el desarrollo de aplicaciones informáticas. Una de las peculiaridades de los almacenes de datos es que hace que modelos y metodologías tradicionales no resulten en ocasiones apropiadas para su diseño, por lo que han surgido metodologías propias para este tipo de sistemas. Entre las metodologías propias para la construcción de un almacén de datos se destacan la metodología descendente, propuesta por Bill Inmon, la metodología ascendente propuesta por Ralph Kimball, Hefesto, CRISP-DM para la minería de datos y Rapid Warehousing Methodology propuesta por SAS Institute.

La metodología de Ralph Kimball se enfoca principalmente en el diseño de datos que almacenará la información para la toma de decisiones. Rapid Warehousing Methodology es una metodología iterativa y está basada en el desarrollo incremental del proyecto dividido en cinco fases: definición de objetivos, definición de requerimientos, diseño y modelación, implementación y división. Por otro lado,

está la metodología Hefesto, que es una metodología propia que puede ser embebida en cualquier ciclo de vida que no requieran de fases extensas de requerimientos y análisis, esta metodología se encuentra dividida en 4 fases: análisis de los requerimientos, análisis de las fuentes de datos, modelado lógico y procesos ETL.

## 1.8. Otros conceptos de interés.

### 1.8.1. Procesamiento de Transacciones en Línea (OLTP).

El Procesamiento de Transacciones en Línea, en inglés On-Line Transaction Processing (OLTP) es un tipo de sistemas que facilitan y administran sistemas transaccionales, generalmente para entradas de datos y recuperación y procesamiento de transacciones. Los paquetes de software para OLTP se basan en la arquitectura cliente servidor que suelen ser utilizados por empresas con una red distribuida. Los OLTP representan toda aquella información transaccional que se genera en la organización en su accionar diario, además de las fuentes externas que puede llegar a disponer. Estas fuentes tienen diversas características, varían en formato, procedencia, función, etc. Entre los OLTP más comunes en cualquier organización se encuentran los archivos de textos, hojas de cálculos, informes con determinada frecuencia y las bases de datos transaccionales. [1]

### 1.8.2. Data mart.

Otro concepto de gran importancia en lo que respecta a la inteligencia de negocios es el Data mart. Un Data mart es una versión especial de un almacén de datos; son subconjuntos de datos con el objetivo de ayudar a que una determinada área en el negocio pueda tomar decisiones mejores, por lo que su alcance de contenido es limitado. Los datos existentes en este contexto pueden ser agrupados, explorados y propagados de diversas formas para que varios grupos de usuarios realicen la exploración de los mismos de la forma más conveniente.

Es válido aclarar que un Data mart por sí solo no es un almacén de datos, porque este último tiene mayor cantidad de usuarios, aborda más temas y provee una vista completa de las áreas funcionales de la organización. [1]

Los Data mart son sistemas orientados a la consulta, que son consultados mediante herramientas OLAP que ofrecen una visión multidimensional de la información. Sobre estas bases de datos se

pueden construir sistemas de información para directivos (EIS) y sistemas de ayuda a la toma de decisiones (DSS).

### 1.8.3. Minería de datos.

La minería de datos o Data Mining en inglés consiste en la extracción no trivial de información que reside de manera implícita en los datos. Esta información era previamente desconocida y podía resultar muy útil. Este término engloba un conjunto de técnicas encaminadas a la extracción de conocimiento procesable implícito en las bases de datos.

### 1.8.4. Sistemas de soporte a decisiones.

Un sistema de soporte a decisiones DSS del inglés *Decision Support System*, es un sistema de información basado en un computador interactivo, flexible y adaptable, especialmente desarrollado para apoyar la solución de un problema de gestión no estructurado para mejorar la toma de decisiones. También utiliza los datos de la organización, proporciona una interfaz amigable y permite la toma de decisiones en el propio análisis de la situación.

### 1.8.5. Sistemas gestores de bases de datos.

Los sistemas de gestión de bases de datos o SGBD, también conocidos en inglés como *database management system* DBMS, son un tipo de software dedicado a servir de interfaz entre la base de datos, el usuario y las aplicaciones que la utilizan. La principal función de estos sistemas es manejar de manera sencilla y ordenada un conjunto de datos que posteriormente se convertirán en información relevante para una organización.

Entre los objetivos que deben cumplir los SGDB se encuentran:

- Abstracción de la información. Los SGDB ahorran a los usuarios detalles acerca del almacenamiento físico de los datos.
- Independencia. La independencia de los datos consiste en la capacidad de modificar un esquema de una base de datos sin tener que realizar cambios en las aplicaciones que se sirven de ella.
- Seguridad. Los SGBD deben garantizar que la información se encuentre segura, permitiendo otorgar diversas categorías de permisos.

- Tiempo de respuesta. Los SGBD debe minimizar el tiempo en que demora en devolver la operación realizada sobre la base de datos.

Los SGBD proveen facilidades para manipular los grandes volúmenes de datos de las cuales se pueden mencionar las siguientes:

- Simplifican la programación de equipos de consistencia.
- Organizan los datos con un impacto mínimo en el código de los programas.
- Disminuyen los tiempos de desarrollo.
- Usualmente proveen interfaces y lenguajes de consulta que simplifican la recuperación de los datos.

En la actualidad existen en el mercado internacional productos de este tipo disponibles, entre los productos libres los más destacados son MySQL y PostgreSQL y entre los no libres podemos mencionar Oracle, Microsoft SQL Server, Microsoft Acces, IBM DB2 y PervasiveSQL.

## 1.9. Metodología, técnicas y herramientas propuestas para la solución.

### 1.9.1. Metodología.

Para guiar el desarrollo del almacén de datos, se utiliza un proceso que se obtiene como resultado de un análisis de las metodologías antes expuestas y un estudio realizado al proceso de desarrollo y gestión de proyecto de software de la Unidad de Compatibilización, Integración y Desarrollo de software para la defensa UCID. Este proceso tiene como base la metodología Hefesto, por las facilidades que la misma presenta y se agregan un conjunto de actividades que describe la metodología ascendente de Kimball.

A continuación se muestran de manera general los pasos con los que cuenta dicha guía:

Fase I: Análisis de los requerimientos.

1.1. Planificar entrevistas.
1.2. Identificar preguntas.
1.3. Identificar perspectivas e indicadores.
1.4. Construcción del modelo conceptual.
1.5. Aprobación del cliente.

Fase II: Análisis de las fuentes.

2.1. Definir estado general de los sistemas fuentes.
2.2. Determinación de indicadores.
2.3. Establecer correspondencias.
2.4. Nivel de granularidad.
2.5. Ampliación del modelo conceptual.
2.6. Definir reglas del negocio.

Fase III: Modelo del almacén de datos.

3.1. Definir tipo de modelo lógico del almacén de datos.
3.2. Definir estándares para objetos físicos.
3.3. Identificar dimensiones.
3.4. Identificar hechos.
3.5. Realizar uniones entre dimensiones y hechos.
3.6. Diseñar tablas y columnas físicas.

Fase IV: Procesos ETL.

4.1. Mapeo de datos.
4.2. Establecer condiciones adicionales y restricciones.
4.3. Cargas incrementales de datos.
4.4. Diseño y construcción de la automatización del sistema ETL.

Fase V: Representación de la información.

Esta guía hereda las siguientes características de la metodología base:

- Los objetivos y resultados esperados en cada fase se distinguen fácilmente y son sencillos de comprender.
- Se basa en los requerimientos del usuario, lo que le permite a su estructura adaptarse con facilidad y rapidez a cualquier cambio en el negocio.
- Reduce la resistencia al cambio, debido a que el usuario final está involucrado en cada etapa para que tome decisiones respecto al comportamiento y funcionamiento del almacén de datos.
- Utiliza modelos conceptuales y lógicos.
- Es independiente de las herramientas que se utilicen para su implementación.
- Es independiente de las estructuras físicas que contengan el almacén de datos y de su respectiva distribución.

- Los resultados obtenidos al concluir una fase se convierten en el punto de partida para llevar a cabo la siguiente.

### 1.9.2. Herramienta de modelado.

En la actualidad existen potentes herramientas que facilitan el modelado de bases de datos, entre las que se destacan el Embarcadero ER/Studio, Rational Rose Data Modeler, DBDesigner y Visual Paradigm.

El Embarcadero ER/Studio es una potente herramienta para el modelado de bases de datos que presenta como una de las principales características el modelo impulsado por el entorno de diseño. También contiene características básicas de los modeladores de bases de datos, posee un generador de SQL y autogenera reportes. [9] Rational Rose Data Modeler acelera el diseño de las bases de datos mediante un entorno de modelado sofisticado y una transformación flexible entre los modelos lógicos y físicos, además de la capacidad de un modelado visual para bases de datos. [7] La principal desventaja de estas dos herramientas de modelado es que no son software libres.

Por otra parte DBDesigner es una herramienta multiplataforma, que posee control de versiones, manejador de SQL y otras características. Una de las ventajas de esta herramienta es que es libre y la principal desventaja es que es una herramienta a la que le faltan funcionalidades por desarrollar y otras necesitan ser mejoradas. [10]

Visual Paradigm es una herramienta libre que ayuda a construir aplicaciones rápidamente, soporta a varios usuarios trabajando en el mismo proyecto. Entre las características que se deben de destacar de esta poderosa herramienta es su robustez, usabilidad y portabilidad. Facilita el modelado colaborativo con Subversion, proporciona la transformación de diagramas Entidad-Relación en tablas de bases de datos, ingeniería inversa de bases de datos, además posee un potente generador de informes.

A partir del estudio realizado de las principales herramientas de modelado, se decidió realizar el modelo de datos en Visual Paradigm, por las ventajas que la misma posee.

### 1.9.3. Gestor de bases de datos.

En cuanto al gestor de bases de datos se ha optado por utilizar PostgreSQL, debido a que es un gestor libre, publicado bajo la licencia BSD. Este gestor soporta integridad referencial, posee una facilidad de configuración e instalación y una gran escalabilidad pudiéndose aplicar a grandes bases de datos. La presencia de integridad referencial es importante para el desarrollo, ya que facilita la depuración de errores en el proceso ETL. Vale destacar que este gestor es multiplataforma y existe una gran comunidad de desarrollo en el mundo que permite la rápida y eficaz evolución del mismo. [11]

### 1.9.4. Herramientas para la administración de bases de datos.

La herramienta de administración de bases de datos seleccionada fue pgAdmin III, la cual es multiplataforma, y está diseñada para responder las necesidades de los usuarios. Esta poderosa herramienta está desarrollada por una comunidad de expertos en PostgresSQL dispersos en todo el mundo. En la selección de esta herramienta fue necesario tener en cuenta el gestor de bases de datos seleccionado.

### 1.9.5. Plataforma de inteligencia de negocios.

Luego de analizar las herramientas de inteligencia de negocios que se tuvieron en cuenta, se optó por utilizar la plataforma Pentaho Open Source Business Intelligence. Unas de las principales características es que es multiplataforma, soporta la ejecución de dashboard y reportes en tiempo real, permite gestionar usuarios, colgar documentación y monitorear la ejecución de trabajos. Esta herramienta posee abundante documentación disponible y tiene una interfaz amigable e incluso, es posible editar dicha interfaz para adecuarse a las necesidades de cada usuario.

### 1.9.6. Herramientas ETL.

Para el proceso ETL, se utilizará Kettle Pentaho Data Integration de la plataforma Pentaho Open Source Business Intelligence. Esta herramienta cuenta con cuatro módulos:

- SPOON, que permite diseñar de forma gráfica las transformaciones del proceso ETL.
- PAN, que ejecuta las transformaciones diseñadas con SPOON.
- CHEF permite mediante una interfaz gráfica diseñar la carga de datos incluyendo un control de estado de los trabajos.

- KITCHEN permite ejecutar los trabajos diseñados con CHEF.

En general, Kettle es una herramienta gráfica de las más antiguas de código abierto, por lo que cuenta con una amplia comunidad de usuarios. Permite trabajar con diferentes fuentes de datos y conectarse a diversos motores de bases de datos, ya sean fuentes o destinos. Otra de las ventajas significativas de esta potente herramienta es que posee rutinas que facilitan el proceso de limpieza de datos.

### 1.9.7. Servidor web.

Apache Tomcat, también llamado Jakarta Tomcat funciona como un contenedor de servlets desarrollado en el proyecto Jakarta en la Apache Software Fundation. Tomcat implementa las especificaciones de los servlets y de Java Server Pages (JSP) de Sun. También incluye el compilador Jasper, que compila JSP convirtiéndola en servlet. El motor de servlets de Tomcat a menudo se presenta en combinación con el servidor web Apache. Debido a que Tomcat fue escrito en Java, funciona en cualquier sistema operativo que disponga de la maquina virtual Java.

### Conclusiones.

En el presente capítulo se hizo referencia a los principales conceptos y características de las herramientas y tecnologías analizadas para dar solución al problema planteado. En la selección de las mismas se tuvo en cuanta las políticas de desarrollo de software en las Fuerzas Armadas Revolucionarias.

Las herramientas seleccionadas para la solución son:

- Pentaho Open Source Business Intelligence como plataforma de inteligencia de negocios.
- Kettle Pentaho Data Integration como herramienta para el proceso ETL.
- Cube Designer, Schema Workbench y Mondrian, pertenecientes a la suite de Pentaho Open Source Business Intelligence para el trabajo con los cubos.
- Apache Tomcat como servidor web.
- Visual Paradigm para el modelado de las bases de datos.
- PostgreSQL como gestor de bases de datos.
- pgAdmin III como cliente para administrar las bases de datos.

Para guiar el desarrollo del almacén de datos se propone el guía que se obtuvo como resultado.

# Capítulo 2: Análisis.

**Introducción.**

El proceso de construcción de un almacén de datos comienza por un análisis de los requerimientos donde se identificarán las necesidades de información de los usuarios, luego se realiza un análisis de las fuentes de datos existentes donde se determina el estado general de las mismas. En el presente capítulo se brindará los diferentes análisis que se realizaron para dar cumplimiento a las primeras fases de la construcción de esta herramienta.

## 2.1. Descripción del negocio.

El departamento de recursos humanos de cualquier entidad es el órgano encargado de controlar la aplicación de las políticas y los lineamientos establecidos como: la estructura, composición y las cifras de efectivos.

En la institución analizada la dirección de recursos humanos cuenta con varios departamentos que atienden diferentes sectores de la institución, entre los que se pueden mencionar el de personal, movilización y defensa civil. La dirección básicamente comprende dos frentes principales de trabajo: reclutamiento o llamado al Servicio Militar Activo (SMA) y el registro, actualización y control del personal que cubre las plantillas de las unidades en tiempo de paz. Ambas tareas se encuentran encaminadas a dos objetivos principales: preparar las reservas y mantener un nivel de fuerzas activas que garanticen el cumplimiento de las misiones para el incremento progresivo hasta alcanzar la completa disposición combativa en tiempo de guerra. Otros de los objetivos que se persiguen en los departamentos antes mencionados son:

- Dirigir todo el proceso de reclutamiento.
- Dirigir los trabajos relacionados con el registro, control, movimiento, preparación y atención del personal.
- Obtener, consolidar e informar a los niveles superiores toda la información necesaria sobre el personal.

Las tareas de los departamentos analizados se ejecutan mediante los niveles inferiores. En estas tareas existe una activa participación de órganos, organismos y organizaciones del estado y la

sociedad cubana. Todas las unidades de la dirección de recursos humanos interactúan mutuamente, nutriéndose y brindando información a otros frentes de trabajo.

De manera general los indicadores en el departamento se obtienen a través de informaciones regulares emitidas por los órganos de personal de los niveles inferiores. Al nivel que se encuentran los departamentos, las informaciones que se manejan son meramente estadísticas y en algunos casos muy puntuales se requiere información nominal relacionada con los datos asociados a las personas, sus características y su empleo en la defensa. Una vez que se cuenta con toda la información es necesario consolidarla e integrarla para lograr realizar un eficiente análisis. Con cierta frecuencia se realizan análisis de comportamiento y tendencias con datos históricos, por lo que los mismos son de gran valor para la institución.

Generalmente el tiempo con que se cuenta para realizar algún análisis en el departamento es limitado debido a la premura de la necesidad de dicha información en niveles superiores.

Actualmente se encuentra en explotación el sistema DATAFAR que automatiza algunos procesos de la dirección de recursos humanos.

## 2.2. Análisis de los requerimientos.

La primera etapa del desarrollo de la solución comienza identificando las necesidades de información de los usuarios, dicha actividad se puede llevar a cabo mediante diferentes técnicas como son las encuestas, cuestionarios, entrevistas y observaciones. Para dar cumplimiento a la misma se elaboraron un conjunto de cuestionarios (ver anexos 1, 2 y 3) con objetivos bien definidos, además fue necesaria la planificación de una serie de entrevistas con los usuarios.

Luego de un profundo análisis de los resultados obtenidos a partir de los cuestionarios realizados y las entrevistas se identificaron las siguientes necesidades de información:

- Se requiere conocer la cantidad de personas con presencia en el listado único en una fecha dada, que pertenezcan a una determinada fuente de ingreso, generación, y lugar.
- Se precisa conocer la cantidad de personas inscritas en una fecha determinada, de cada generación y lugar, además saber si las mismas tienen presencia en el listado único.
- Se necesita conocer la cantidad de personas no inscritas teniendo en cuenta un tiempo determinado, por cada causa, generación y lugar.

- Se exige conocer la cantidad de prerreclutas de cada generación y lugar que conforman la preselección, ya sean disponibles o estudiantes, y también posean una determinada situación en el registro militar.
- Se pide conocer la cantidad de prerreclutas de cada generación y lugar que no fueron preseleccionados, ya sean disponibles o estudiantes, y además, posean una determinada situación en el registro militar.
- Se necesita conocer la cantidad de prerreclutas controlados de cada generación y lugar en una determinada fecha, presenten una determinada situación en el registro militar.

### 2.2.1. Perspectivas e indicadores.

En esta sección del documento se detallan las perspectivas de observación que intervienen en el análisis de la información, así como los indicadores que se utilizarán. Los indicadores generalmente son valores numéricos y representan lo que se desea analizar, por ejemplo: promedios, cantidades, sumatorias, etc.; en cambio, las perspectivas se refieren a los objetos mediante los cuales se quiere examinar los indicadores, con el fin de responder las preguntas planteadas. [1]

Luego de un análisis de las necesidades de información mencionadas anteriormente se determinaron las siguientes perspectivas e indicadores.

**Perspectivas:**

- Persona.
- Generación.
- Lugar.
- Fecha.
- Fuente de ingreso.
- Causa de no inscripción.
- Prerrecluta.
- Situación en el registro militar.

**Indicadores:**

- Cantidad de personas en el listado único.
- Cantidad de personas inscritas.
- Cantidad de personas no inscritas.

- Cantidad de prerreclutas preseleccionados.
- Cantidad de prerreclutas no preseleccionados.
- Cantidad de prerreclutas controlados.

### 2.2.2. Modelo conceptual.

A continuación, en la figura 2.2 se muestra un modelo conceptual construido a partir de los indicadores y las perspectivas identificadas, además de las relaciones que existen entre ellos, en dicho modelo se puede observar con claridad cuáles son los alcances del almacén de datos.

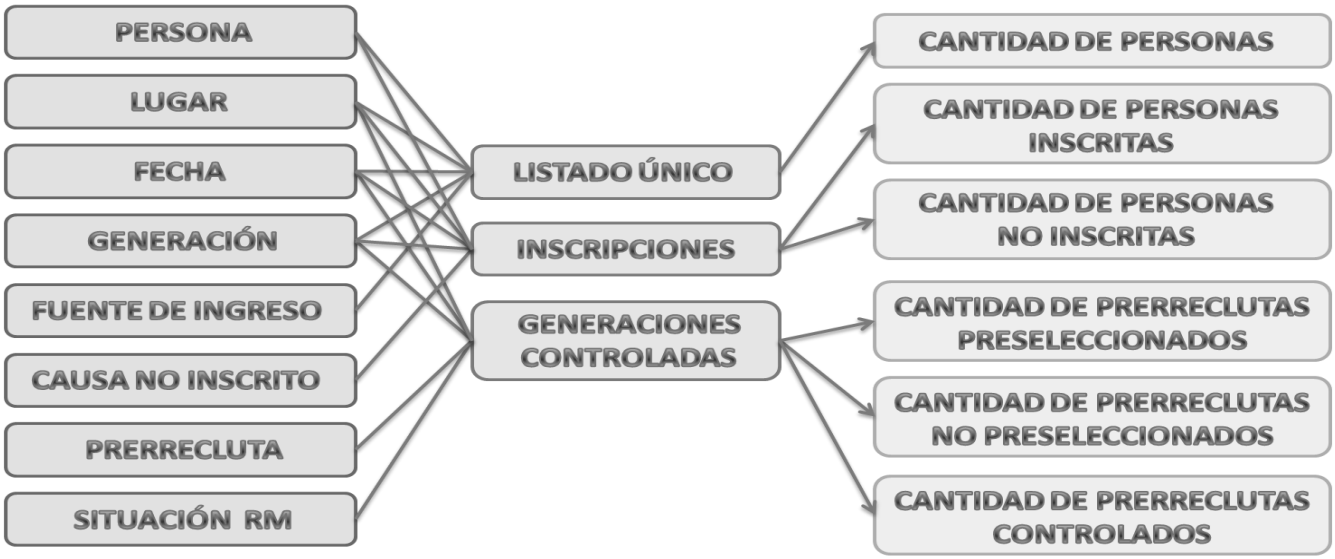


**Figura 2.1. Modelo conceptual.**

Al paso que se desarrolla el análisis este modelo conceptual se irá enriqueciendo ampliando los conceptos que se muestran, agregando valor a los mismos.

## 2.3. Análisis del estado de las fuentes de datos.

Para gestionar los datos correspondientes al sistema DATAFAR se cuenta con una base de datos relacional y como gestor de base de datos se utiliza PostgreSQL en la versión 8.0.

Como medida organizativa y técnica tomada para asegurar el funcionamiento del sistema de los Comités Militares, los servidores deben estar disponibles las 24 horas del día, logrando de esta manera que la base de datos se encuentre en plena disponibilidad. El sistema DATAFAR cuenta con

una funcionalidad para la réplica de datos, por lo que adquiere una completa actualización de los servidores que se encuentran en diferentes niveles.

### 2.4. Determinación de indicadores.

En este paso se explica cómo se calculan los indicadores, definiendo conceptos para ellos como son los hechos que lo componen, fórmula o función de sumarización que se utilizará para su agregación en el almacén de datos.

El indicador "**cantidad de personas en el listado único**" representa la cantidad de las personas que conforman el listado único por una o más fuentes de ingreso.

Hecho: total de personas en el listado único.

Función de sumarización: COUNT.

El indicador "**cantidad de personas inscritas**" representa el total de las personas que ya se encuentran inscritas.

Hecho: total de personas inscritas.

Función de sumarización: COUNT.

El indicador "**cantidad de personas no inscritas**" constituye la cantidad de personas que no han sido inscritas por un determinado motivo.

Hecho: total de personas no inscritas.

Función de sumarización: COUNT.

El indicador "**cantidad de prerreclutas preseleccionados**" constituye el total de prerreclutas preseleccionados.

Hecho: total de prerreclutas preseleccionados.

Función de sumarización: COUNT.

El indicador “**cantidad de prerreclutas no preseleccionados**” constituye el total de prerreclutas que no han sido preseleccionados por alguna causa.

Hecho: total de prerreclutas no preseleccionados.

Función de sumarización: COUNT.

El indicador “**cantidad de prerreclutas controlados**” constituye el total de personas inscritas sin importar la generación y que tengan la categoría de prerreclutas.

Hecho: total de prerreclutas controlados.

Función de sumarización: COUNT.

### 2.5. Correspondencias.

Luego de un análisis de las fuentes de datos disponibles en busca de que las mismas contengan los datos requeridos se establece una correspondencia entre el modelo conceptual y las fuentes de datos.

A continuación se muestran las relaciones identificadas:

- La perspectiva “Persona” se relaciona con la tabla “*dat_persona*”.
- La perspectiva “Generación” se relaciona con la tabla “*aux_generación*”.
- La perspectiva “Estructura” se relaciona con la tabla “*nom_nodosapp*”.
- La perspectiva “Fuente de ingreso” se relaciona con la tabla “*nom_fuenteingreso*”.
- La perspectiva “Causa no inscrito” se relaciona con la tabla “*nom_causaninsc*”.
- La perspectiva “Prerrecluta” se relaciona con la tabla “*dat_reclutamiento*”.
- La perspectiva “Situación en el registro militar” se relaciona con la tabla “*nom_causasituacrm*”.
- El indicador “cantidad de personas en el listado único” se relaciona con la tabla “*dat_listadounico*”.
- El indicador “cantidad de personas inscritas” se relaciona con la tabla “*dat_reclutamiento*”.
- El indicador “cantidad de personas no inscritas” se relaciona con la tabla “*dat_noinscritos*”.
- El indicador “cantidad de prerreclutas preseleccionados” se relaciona con la tabla “*dat_proceso*”.
- El indicador “cantidad de prerreclutas no preseleccionados” se relaciona con la tabla “*dat_proceso*”.

- El indicador “cantidad de prerreclutas controlados” se relaciona con la tabla “*dat_situacrm*”.

La perspectiva fecha tiene relación con todos los campos denominados “*rmodific*” de las tablas involucradas.

### 2.6. Nivel de granularidad.

Teniendo en cuenta las correspondencias establecidas, se analizaron los campos de cada tabla de la base de datos a la que se hace referencia, examinándolos para distinguir los significados de cada campo.

Respecto a la perspectiva persona los datos disponibles son los siguientes:

| Atributo | Tipo | Nulo | Descripción |
|---|---|---|---|
| idorig | BIGINT | No | Es el identificador del origen de los datos de la persona. |
| idpers | BIGINT | No | Es el identificador de la persona y representa unívocamente una sola persona en un determinado origen de datos. |
| numid | VARCHAR(11) | No | Representa el número de identidad de la persona. |
| nombre | VARCHAR(30) | No | Nombre de la persona. |
| papell | VARCHAR(30) | No | Primer apellido de la persona. |
| sapell | VARCHAR(30) | No | Segundo apellido de la persona. |
| nompa | VARCHAR(30) | No | Nombre del padre. |
| nomma | VARCHAR(30) | No | Nombre de la madre. |
| idprovincia | SMALLINT | No | Representa mediante una clave foránea la provincia a la que pertenece la persona. |
| idmunicipio | SMALLINT | No | A través de una clave foránea representa el municipio al que pertenece la persona. |
| direccion | VARCHAR(100) | No | Dirección particular de la persona. |
| iddireccion | BIGINT | No | Representa mediante una clave foránea la dirección particular actual de la persona. |
| idocupacion | SMALLINT | No | Representa la ocupación actual de la persona. |
| idextsocial | SMALLINT | No | Representa a través de una clave foránea la extracción social de la persona. |
| idcolorpiel | SMALLINT | No | Representa mediante una clave foránea el color de la piel de la persona. |
| idcolorpelo | SMALLINT | No | Representa a través de una clave foránea el color de la piel de la persona. |
| idcolorojos | SMALLINT | No | Representa mediante una clave foránea el color de la piel de la persona. |
| sexo | BIT(1) | No | Es el sexo de la persona. |

| | | | |
|---|---|---|---|
| idnivesc | SMALLINT | No | Mediante una clave foránea representa el nivel escolar de la persona. |
| fnac | DATE | | Representa la fuente de ingreso de la persona. |
| idmilita | SMALLINT | No | Representa a través de una clave foránea la militancia de la persona. |

**Tabla 2.1. Descripción de los datos disponibles para la perspectiva "Persona".**

Respecto a la perspectiva "Estructura" los datos disponibles son los siguientes:

| Atributo | Tipo | Nulo | Descripción |
|---|---|---|---|
| idorig | BIGINT | No | Es el identificador del origen de los datos. |
| idnivelacc | SMALLINT | No | Representa el nivel de acceso dentro de una determinada estructura. |
| idmando | SMALLINT | No | Indica a que nodo del nivel 1 pertenece cada provincia. |
| idprovincia | SMALLINT | No | Mediante una clave foránea representa la provincia a la que pertenece una determinada estructura. |
| idmunicipio | SMALLINT | No | Representa a través de una clave foránea el municipio al que pertenece una determinada estructura. |
| nodo | VARCHAR(40) | No | Constituye la base de una estructura determinada. |
| padre | BIGINT | No | Representa el identificador del nodo padre de la estructura. |
| codminint | VARCHAR(4) | No | Representa el código de la provincia y el municipio para el MININT. |

**Tabla 2.2. Descripción de los datos disponibles para la perspectiva "Estructura".**

Respecto a la perspectiva "Generación" los datos disponibles son los siguientes:

| Atributo | Tipo | Nulo | Descripción |
|---|---|---|---|
| idorig | BIGINT | No | Es el identificador del origen de los datos. |
| generacion | SMALLINT | No | Representa el número del año de la generación. |
| ainscribir | BIT(1) | No | Representa cual es la generación que se debe de inscribir en el año actual. |

**Tabla 2.3. Descripción de los datos disponibles para la perspectiva "Generación".**

Respecto a la perspectiva "Fuente de ingreso" los datos disponibles son los siguientes:

| Atributo | Tipo | Nulo | Descripción |
|---|---|---|---|
| idfingreso | SMALLINT | No | Representa el identificador de cada fuente de ingreso. |
| fingreso | VARCHAR(30) | No | Representa la fuente de ingreso de una persona al listado único. |

| | | | |
|---|---|---|---|
| abrev | CHAR(5) | No | Representa la abreviatura que se emplea para cada fuente de ingreso. |

**Tabla 2.4. Descripción de los datos disponibles para la perspectiva "Fuente de ingreso".**

Respecto a la perspectiva “Causa no inscrito” los datos disponibles son los siguientes:

| Atributo | Tipo | Nulo | Descripción |
|---|---|---|---|
| idcausaninsc | SMALLINT | No | Representa el identificador de la causa por la que la persona no fue inscrita. |
| causaninsc | VARCHAR(50) | No | Representa la causa por la que una persona no fue inscrita. |
| abrev | VARCHAR(25) | No | Representa la abreviatura que se emplea para cada causa por la que una persona no fue inscrita. |

**Tabla 2.5. Descripción de los datos disponibles para la perspectiva "Causa no inscrito".**

Respecto a la perspectiva “Prerrecluta” los datos disponibles son los siguientes:

| Atributo | Tipo | Nulo | Descripción |
|---|---|---|---|
| idorig | BIGINT | No | Constituye el identificador del origen de los datos. |
| idpers | BIGINT | No | Representa el identificador de la persona que posee el prerrecluta. |
| padec | VARCHAR(95) | No | Constituye el padecimiento que reporta el prerrecluta a la hora que se realizó la inscripción |
| idpadec | SMALLINT | No | Constituye el identificador del padecimiento reportado por el prerrecluta. |
| aatenc | SMALLINT | No | Número del área de atención a la que pertenece el prerrecluta. |
| faltarm | DATE | No | Constituye la fecha de alta en el registro de prerreclutas. |
| finscripc | DATE | | Representa la fecha en que el prerrecluta fue inscrito en el registro. |
| faplaz | DATE | | Constituye la fecha de aplazamiento en el registro del prerrecluta. |

**Tabla 2.6. Descripción de los datos disponibles para la perspectiva "Prerrecluta".**

Respecto a la perspectiva “Situación en el registro militar” los datos disponibles son los siguientes:

| Atributo | Tipo | Nulo | Descripción |
|---|---|---|---|
| idsituacrm | INTEGER | No | Representa la situación genérica en el registro militar a través de una clave foránea. |
| idcausasituacrm | INTEGER | No | Constituye el identificador de una situación específica en el registro. |
| causa | VARCHAR(50) | No | Representa la situación específica en el registro |

| | | | |
|---|---|---|---|
| | | | militar. |
| grupo | SMALLINT | No | Identifica al grupo al que pertenecen, los mismos pueden ser prerreclutas o bajas |

**Tabla 2.7. Descripción de los datos disponibles para la perspectiva "Situación en el Registro Militar".**

Respecto a la perspectiva “Fecha” los datos que se pueden emplear son los siguientes:

- Día.
- Mes.
- Año.
- Semestre.
- Trimestre.
- Hora.
- Minutos.
- Segundos.

Los campos que se describieron anteriormente, son aquellos que se consideraron con mayor importancia.

Una vez recolectada toda la información apropiada y consultado con los usuarios cuáles eran los datos que se consideraban de interés para analizar los indicadores expuestos anteriormente, se obtuvieron los siguientes resultados:

- Para conformar la perspectiva “Persona” solo se tendrá en cuenta el número de identidad, el nombre y los apellidos, el sexo, el color de la piel, la dirección particular, la extracción social, el nivel escolar y la ocupación.
- En la perspectiva “Generación” se tendrá en cuenta el año de la generación.
- Para conformar la perspectiva “Estructura” se usará el identificador del origen, el mando superior, el ejército al que pertenece cada provincia, así como la provincia, el municipio y el área de atención.
- En la perspectiva “Fuente de ingreso” solo se utilizará la fuente de ingreso.
- En la perspectiva “Causa no inscrito” solo se usará la causa por la que no fue inscrita alguna persona.
- Para la perspectiva “Prerrecluta” solo se utilizará la fecha de alta en el registro de prerrecluta, la fecha de inscripción y el área de atención.

- Para la perspectiva “Situación en el registro militar” se tendrá en cuenta la situación genérica en el registro militar y la situación específica en el registro militar.
- En la perspectiva “Fecha” se tendrá en cuenta el día, mes y año.

### 2.7. Perfilado de datos.

Se entiende por perfilado de datos el análisis de datos de los sistemas para comprender su contenido, estructura, calidad y dependencias. Esta tarea es vital, ya que, a la hora de plantear un análisis de los datos de origen encontramos en muchas ocasiones que se desconoce qué se ha de preguntar, ni donde pueden residir algunos problemas.

El perfilado de datos dará un análisis exacto de la estructura y el modelo de datos a tratar, dejando de lado la forma más común de análisis.

Los datos se analizan mediante perfiles de heurística donde se establecen reglas, que se utilizan para medir los datos y corregirlos en caso de errores.

| PERFILADO DE DATOS | | |
|---|---|---|
| **PROBLEMAS** | **METADATOS** | **OBSERVACIÓN** |
| Valores llegales | Max, Min | Rango de valores permitidos. |
| Valores escritos incorrectamente | Valores de los atributos | El ordenamiento de los valores ayuda a detectar valores mal escritos.<br>Contar los valores puede ayudar a detectar errores. |
| Valores ausentes | Valores nulos | Por ciento/Cantidad de valores nulos. |
| | Valores de atributos + valores por defecto | La presencia de valores por defecto puede indicar que valores reales estén ausentes. |
| Diferentes representaciones del mismo valor | Valores de atributos | Comparando el conjunto de valores de una columna de una tabla con los valores de otra columna de otra tabla. |
| Duplicados | Cardinalidad + Unicidad | La cardinalidad de un atributo identificador debe corresponderse con la cantidad de filas de la tabla. |
| | Valores de atributos | Ordenado los valores por el número de ocurrencias, más de una ocurrencia puede indicar valores duplicados. |

**Tabla 2.8. Análisis de los datos.**

Durante la ejecución de este paso del análisis de los sistemas fuentes, se logró detectar un conjunto de errores que pueden tener cierta repercusión en el resultado final. Entre los posibles problemas detectados se encuentran:

- La existencia de campos que contienen valores nulos.
- La existencia de campos que toman valores por defecto.

Cada uno de estos posibles errores fue analizado cuidadosamente con los especialistas, lo que permitió llegar a la siguiente conclusión: Los campos que contienen valores nulos representan un 0.26 % aproximadamente, y los que residen en tablas significativas para la construcción del almacén de datos no presentan riesgos una vez que sean tratados correctamente. Los valores por defecto indican datos temporales, por lo que de la misma manera no constituyen una amenaza.

### 2.8. Reglas para transformaciones de datos.

Las reglas que se definen en este paso contienen detalles necesarios para llevar a cabo las transformaciones que se realizarán en los procesos ETL.

Teniendo en cuenta el análisis realizado anteriormente se establecieron las siguientes reglas para la transformación y carga de los datos:

- No cargar los datos temporales debido a que los usuarios no lo tienen en cuenta a la hora de realizar un determinado análisis.
- Tener en cuenta el nivel 5 para la conformación de las estructuras.
- Asignar un nuevo identificador para las personas y las estructuras.
- En el caso particular del campo que contiene los valores asociados a la preselección de los prerreclutas que contiene valores nulos, los mismos indican que los prerreclutas no están preseleccionados.
- Los prerreclutas son todas las personas que tienen como identificador de categoría los valores 1 y 4.

**Conclusiones.**

En el desarrollo del presente capítulo se brinda una breve descripción del negocio para lograr una eficiente comprensión a la hora de analizar y diseñar el almacén de datos. También se realizó un detallado análisis de las diferentes necesidades de información que existen determinando las

principales perspectivas e indicadores que intervienen comúnmente durante el análisis de información. También se analizó el estado actual de las fuentes de datos con que se cuenta actualmente. Uno de los logros más significativos fue el de establecer las correspondencias entre el resultado del análisis de las necesidades de información con las fuentes de datos, dando paso al posterior diseño del almacén de datos.

# Capítulo 3: Diseño.

**Introducción.**

En el desarrollo del presente capítulo se pretende tratar los temas relacionados con el diseño del almacén de datos, así como la arquitectura definida para el mismo. Se identificarán las tablas de hechos y dimensiones y también se realizarán las uniones entre las mismas para conformar el modelado del depósito de datos.

## 3.1. Arquitectura del almacén de datos.

Teniendo en cuenta las características generales de un almacén de datos se definen los componentes que intervienen en su arquitectura o ambiente.

El ambiente está formado por diversos elementos que interactúan entre sí y que cumplen una función específica dentro del sistema. Básicamente la forma de operar del esquema superior se resume de la siguiente manera:

- Los datos son extraídos de las aplicaciones, bases de datos, archivos, etc. Esta información reside generalmente en diferentes tipos de sistemas, orígenes y arquitecturas, también tienen formatos muy variados.
- Los datos son transformados, limpiados e integrados para luego ser cargados en el almacén de datos.
- La información del almacén de datos se estructura en cubos multidimensionales, aunque también se pueden utilizar otros tipos de estructuras para representar la información del almacén de datos.

Como se había visto anteriormente los OLTP representan toda la información transaccional que genera la institución en su accionar diario, además de todas las fuentes de datos externas que puedan existir. En este caso la información de la que se habla se encuentra en las bases de datos que dispone el sistema DATAFAR.

En el gestor de carga se utilizará un sistema que es capaz de extraer los datos, manipularlos, transformarlos e integrarlos, este sistema es comúnmente conocido como sistemas ETL. A continuación se explica en síntesis el accionar del proceso de extracción, transformación y carga.

Los pasos que se siguen en este proceso son los siguientes:

- Se extraen los datos más relevantes desde los OLTP y se colocan en un almacenamiento intermedio.
- Se integran y se transforman los datos para evitar inconsistencias.
- Se cargan los datos desde el almacenamiento intermedio hasta el almacén de datos, si existen correspondencias directas entre los datos de los OLTP y el almacén de datos se procede a su respectiva carga. [1]

Los sistemas ETL son los encargados de realizar una carga inicial y actualizar periódicamente.

El gestor del almacén de datos presenta las siguientes características y funciones:

- Transforma e integra los datos fuentes y del almacenamiento intermedio en un modelo adecuado para la toma de decisiones.
- Gestiona el depósito de datos a través de tablas de hechos y tablas de dimensiones y lo organiza en torno a una base de datos multidimensional. Esto permite que se puedan crear cubos multidimensionales, modelos de negocio u otras estructuras de datos.
- Permite realizar todas las funciones de definición y manipulación del depósito de datos, para poder soportar todos los procesos de gestión del mismo.
- Es el encargado de ejecutar y definir las políticas de particionamiento.
- Realiza copias de resguardo incremental o total de los datos del almacén de datos.
- Se constituye comúnmente al combinar un SGBD con aplicaciones especializadas.
- Posee un depósito de datos propio.
- Gestiona y mantiene los metadatos.

## 3.2. Tipo de modelo para el diseño del almacén de datos.

Para seleccionar el modelo que se empleará para el diseño del almacén de datos se debe tener en cuenta los requerimientos y necesidades de los usuarios. Es de suma importancia definir cuál esquema se empleará para el diseño de la estructura del depósito de datos debido a que esta selección afectará considerablemente la elaboración del modelo lógico. [1]

El modelo que se empleará es un híbrido a partir de los esquemas constelación y copo de nieve, debido a las necesidades de los usuarios, así como las perspectivas e indicadores identificados. Se selecciona el esquema constelación debido a que se crearán tablas de dimensiones que se

relacionarán con más de una tabla de hechos. También se basa en el modelo copo de nieve porque se diseñarán tablas de dimensiones que se asociarán a otras dimensiones.

### 3.3. Definición de estándares para objetos físicos.

Para realizar el diseño del almacén de datos se utilizó los estándares definidos por el centro en el documento que recoge las políticas y estándares del rol arquitecto de datos.

Teniendo en cuenta los estándares de nomenclatura, los nombres de las bases de datos comienzan con la primera letra en mayúscula y el resto en minúscula, y en caso de que sea un nombre compuesto se empleará la notación PascalCasing. Para el caso particular de los almacenes de datos se utilizará el prefijo “Dwh”.

Los nombres de las tablas deben escribirse en minúsculas para evitar problemas con el Case Sensitive del gestor, también se utilizará el prefijo “dim_” en las tablas de dimensiones y “cub_” en las tablas de hechos.

El nombre a emplear para los campos se escribirá en minúsculas y lo más legible posible, de forma tal que se logre entender el propósito del mismo. En caso de que el campo sea un identificador se empleará el prefijo “id” y si forma parte de una llave foránea recibirá el mismo nombre de donde proviene, seguidamente un símbolo “_” y finalmente el nombre de la tabla que la recibe.

El nombre de las restricciones se escribirá en minúsculas, para el caso de las claves primarias se utilizará el prefijo “pk_” y luego el nombre de la tabla sin especificar el prefijo de la misma. Los nombres de las llaves foráneas se escribirán con minúsculas y comenzarán con el prefijo “fk_”.

Todas las tablas y atributos de la misma serán comentadas con el fin de definir el propósito de cada uno de estos elementos.

### 3.4. Identificación de dimensiones.

Generalmente cada perspectiva definida en el modelo conceptual constituirá una dimensión. Para llevar a cabo este proceso primeramente es necesario determinar el nombre que identificará la tabla de dimensión. Seguidamente se añadirá un campo que constituya la clave principal de dicha tabla y finalmente se redefinirán los nombres de los campos que conformarán la dimensión en caso que no sean lo necesariamente intuitivos.

La dimensión asociada a la perspectiva “Persona” tendrá como nombre “*dim_persona*”. El atributo denominado “*idpers*” que representa el identificador de la persona en la base de datos conformará la clave primaria, la misma será denominada “*pk_persona*”. A continuación se muestra la descripción de dicha tabla de dimensión:

| **Nombre: dim_persona** | | |
|---|---|---|
| **Descripción:** Tabla de dimensión del almacén de datos que almacena los datos asociados a la perspectiva persona. | | |
| **Atributo** | **Tipo** | **Descripción** |
| idpers | NUMERIC (18,0) | Identificador de la persona. |
| numid | VARCHAR (11) | Número de identidad de la persona. |
| nombre | VARCHAR (50) | Nombre(s) de la persona. |
| papellido | VARCHAR (50) | Primer apellido de la persona. |
| sapellido | VARCHAR (50) | Segundo apellido de la persona. |
| nombapells | VARCHAR (150) | Nombre(s) y apellidos de la persona. |
| sexo | VARCHAR (9) | Sexo de la persona. |
| colorpiel | VARCHAR (20) | Color de piel de la persona. |
| dirpart | VARCHAR (300) | Dirección particular de la persona. |
| extracsocial | VARCHAR (15) | Extracción social de la persona. |
| nivelescolar | VARCHAR (17) | Nivel escolar de la persona. |
| ocupacion | VARCHAR (28) | Ocupación de la persona. |

**Tabla 3.1. Descripción de la entidad "*dim_persona*".**

La dimensión asociada a la perspectiva “Generación” recibirá el nombre *“dim_generacion*”. El atributo llamado “*idgeneracion*” conformará la clave primaria de la tabla que recibirá la denominación “*pk_generacion*”. En la siguiente tabla se resume la descripción de la entidad “dim_generación”:

| **Nombre: dim_generacion** | | |
|---|---|---|
| **Descripción:** Tabla de dimensión del almacén de datos que almacena los datos asociados a la perspectiva generación. | | |
| **Atributo** | **Tipo** | **Descripción** |
| idgeneracion | NUMERIC (3,0) | Identificador de la generación. |
| generacion | NUMERIC (4,0) | Número del año de la generación. |

**Tabla 3.2. Descripción de la entidad "*dim_generacion*".**

La tabla dimensión asociada a la perspectiva “Estructura” se denominará *“dim_estructura*”. El atributo denominado “*idestructura*” que representará el identificador de la estructura o el lugar en la base de datos conformará la llave primaria denominada “*pk_estructura*”. La siguiente tabla resume la descripción de la entidad “*dim_estructura*”:

| **Nombre: dim_estructura** | | |
|---|---|---|
| **Descripción:** Tabla de dimensión que almacena los datos relacionados a la estructura. | | |
| **Atributo** | **Tipo** | **Descripción** |
| idestructura | NUMERIC (5,0) | Identificador de la estructura. |
| pnivel | VARCHAR (6) | Primer nivel de la jerarquía estructura. |
| snivel | VARCHAR (19) | Segundo nivel de la jerarquía estructura. |
| tnivel | VARCHAR (55) | Tercer nivel de la jerarquía estructura. |
| cnivel | VARCHAR (55) | Cuarto nivel de la jerarquía estructura. |
| qnivel | NUMERIC (3,0) | Quinto nivel de la jerarquía estructura. |

**Tabla 3.3. Descripción de la entidad "*dim_estructura*".**

La dimensión asociada a la perspectiva “Fecha” se denominará *“dim_fecha*”. El atributo “*idfecha*” que representa el identificador de la fecha en la base de datos constituirá una clave primaria llamada “*pk_fecha*”. La siguiente tabla muestra la descripción de la entidad “*dim_fecha*”:

| **Nombre: dim_fecha** | | |
|---|---|---|
| **Descripción:** Tabla de dimensión que almacena los datos relacionados con el tiempo. | | |
| **Atributo** | **Tipo** | **Descripción** |
| idfecha | NUMERIC (15,0) | Identificador de la fecha. |
| anno | NUMERIC (4,0) | Número del año. |
| mes | NUMERIC (2,0) | Número del mes. |
| dia | NUMERIC (2,0) | Número del día. |
| denominacion | VARCHAR (10) | Nombre que recibe el día de la semana. |

**Tabla 3.4. Descripción de la entidad "*dim_fecha*".**

La tabla de dimensión relacionada a la perspectiva “Fuente de ingreso” será denominada “*dim_fuenteingreso*”, la misma tendrá el atributo “*idfuenteingreso*” que representa el identificador de la fuente de ingreso, el mismo conformará la clave primaria, la cual recibirá el nombre “pk_fuenteingreso”. A continuación se muestra una tabla que recoge la descripción de esta entidad:

| **Nombre: dim_fuenteingreso** | | |
|---|---|---|
| **Descripción:** Tabla de dimensión que almacena los datos relacionados con la fuente de ingreso de la persona al listado único. | | |
| **Atributo** | **Tipo** | **Descripción** |
| idfuenteingreso | NUMERIC (2,0) | Identificador de la fuente de ingreso. |
| fuenteingreso | VARCHAR (30) | Fuente de ingreso al listado único. |

**Tabla 3.5. Descripción de la entidad "*dim_fuenteingreso*".**

La dimensión asociada a la perspectiva “Causa no inscrito” recibirá el nombre “*dim_causanoinsc*”, esta tabla tendrá al atributo “*idcausa*” como clave primaria, la cual será denominada “*pk_causanoinsc*”. La siguiente tabla muestra la descripción de dicha entidad:

| **Nombre: dim_causanoinsc** | | |
|---|---|---|
| **Descripción:** Tabla de dimensión que almacena los datos relacionados a las causas por las que una persona no fue inscrita. | | |
| **Atributo** | **Tipo** | **Descripción** |
| idcausa | NUMERIC (3,0) | Identificador de la causa de no inscripción. |
| causanoinsc | VARCHAR (50) | Causa de no inscripción de la persona. |

**Tabla 3.6. Descripción de la entidad "*dim_causanoinsc*".**

La tabla dimensión relacionada a la perspectiva “Prerrecluta” se denominará “*dim_prerrecluta*”. En esta tabla se representará el identificador de la persona “*idpers*” mediante una llave foránea denominada “*fk_persona_prerrecluta*”, dicho atributo conformará la clave primaria y se denominará “*pk_prerrecluta*”. En la tabla que se muestra a continuación se resume la descripción de la entidad “*dim_prerrecluta*”:

| **Nombre: dim_prerrecluta** | | |
|---|---|---|
| **Descripción:** Tabla de dimensión que almacena las personas que son prerreclutas. | | |
| **Atributo** | **Tipo** | **Descripción** |
| idpers | NUMERIC (18,0) | Identificador de la persona. |
| preseleccionado | NUMERIC (1,0) | Determina si un prerrecluta está preseleccionado o no. |

**Tabla 3.7. Descripción de la entidad "*dim_prerrecluta*".**

La dimensión relacionada con la perspectiva “Situación en el Registro Militar” tendrá como nombre “*dim_situacionrm*”, esta tabla tendrá un atributo denominado “*idsituacion*” que formará la clave primaria “*pk_situacionrm*”. En la siguiente tabla se resume la descripción de dicha entidad:

| **Nombre: dim_situacionrm** | | |
|---|---|---|
| **Descripción:** Tabla de dimensión que almacena la situación en el registro militar de una persona. | | |
| **Atributo** | **Tipo** | **Descripción** |
| idsituacion | NUMERIC (3,0) | Identificador de la situación en el registro militar. |
| situacion | VARCHAR (50) | Situación genérica que presenta la persona en el registro militar. |
| causasituacion | VARCHAR (50) | Causa de la situación en el registro militar. |

**Tabla 3.8. Descripción de la entidad "*dim_situacionrm*".**

### 3.5. Identificación de hechos.

Las tablas de hechos contienen los hechos que serán analizados para apoyar el proceso de toma de decisiones. Los hechos son datos instantáneos en el tiempo, que son filtrados, agrupados y explorados a través de condiciones definidas en las tablas de dimensiones. Los registros del hecho poseen una clave primaria que está compuesta por las claves primarias de las tablas de dimensiones relacionadas a este. [1]

Para los esquemas estrella y copo de nieve se debe asignar un nombre a la tabla de hechos que represente la información analizada, área de investigación, negocio, etc. La clave primaria estará compuesta por la combinación de las claves primarias de cada tabla de dimensión relacionada y se crearán tantos campos de hechos como indicadores se hayan definido en el modelo conceptual.

Para los esquemas constelación las tablas de hechos se deben confeccionar teniendo en cuenta el análisis realizado en las etapas anteriores

A continuación se definirán las tablas de hechos, que son aquellas que contendrán los hechos a través de los cuales se construirán los indicadores de análisis.

La primera tabla de hechos que surge de la relación identificada como “listado único” será denominada “*cub_listadounico*”, esta tabla contendrá los campos “*idestructura*”, “*idpers*”, “*idgeneracion*”, “*idfecha*” y “*idfuenteingreso*” que constituyen las claves foráneas “*fk_estructura_listadounico*”, “*fk_persona_listadounico*”, “*fk_generacion_listadounico*”, “*fk_fecha_listadounico*” y “*fk_fuenteingreso_listadounico*” respectivamente. Todos estos atributos constituyen la clave primaria denominada “*pk_listadounico*”. En la siguiente tabla se muestra la descripción de la entidad “*cub_listadounico*”:

| **Nombre: cub_listadounico** | | |
|---|---|---|
| **Descripción:** Tabla de hechos que almacena los datos asociados a las personas que se encuentran en el listado único. | | |
| **Atributo** | **Tipo** | **Descripción** |
| idestructura | NUMERIC (5,0) | Identificador de la estructura. |
| idpers | NUMERIC (18,0) | Identificador de la persona. |
| idgeneracion | NUMERIC (3,0) | Identificador de la generación. |
| idfecha | NUMERIC (15,0) | Identificador de la fecha. |
| idfuenteingreso | NUMERIC (2,0) | Identificador de la fuente de ingreso. |

**Tabla 3.9. Descripción de la entidad "*cub_listadounico*".**

Para la relación identificada como inscripción es necesario realizar dos tablas de hechos, una para los inscritos llamada "*cub_inscritos*" y otra para los no inscritos denominada "*cub_noinscritos*" porque los indicadores poseen diferentes perspectivas de análisis. La entidad "*cub_inscritos*" tendrá una clave primaria denominada "*pk_inscritos*" que estará compuesta por los atributos "*idestructura*", "*idpers*", "*idgeneracion*" y "*idfecha*", estos campos a su vez constituirán las llaves foráneas "*fk_estructura_inscritos*", "*fk_persona_inscritos*", "*fk_generación_inscritos*" y "*fk_fecha_inscritos*" respectivamente. En la tabla 3.10 se muestra la descripción de la entidad "*cub_inscritos*". Los campos de la tabla de hechos "*cub_noinscritos*": "*idestructura*", "*idpers*", "*idgeneracion*", "*idfecha*" y "*idcausa*" formarán las claves foráneas llamadas "*fk_estructura_noinscritos*", "*fk_persona_noinscritos*", "*fk_generacion_noinscritos*", "*fk_fecha_noinscritos*" y "*fk_causanoinsc_noinscritos*" respectivamente, además todos estos atributos formarán la clave primaria denominada "*pk_noinscritos*". La descripción de la entidad aparece en la tabla 3.11.

| **Nombre: cub_inscritos** | | |
|---|---|---|
| **Descripción:** Tabla de hechos que almacena los datos relacionados a las personas inscritas. | | |
| **Atributo** | **Tipo** | **Descripción** |
| idestructura | NUMERIC (5,0) | Identificador de la estructura. |
| idpers | NUMERIC (18,0) | Identificador de la persona. |
| idgeneracion | NUMERIC (3,0) | Identificador de la generación. |
| idfecha | NUMERIC (15,0) | Identificador de la fecha. |
| listunico | NUMERIC (1,0) | Determina si la persona tiene presencia en el listado único o no. |

**Tabla 3.10. Descripción de la entidad "*cub_inscritos*".**

| **Nombre: cub_noinscritos** | | |
|---|---|---|
| **Descripción:** Tabla de hechos que almacena los datos relacionados a las personas no inscritas. | | |
| **Atributo** | **Tipo** | **Descripción** |
| idestructura | NUMERIC (5,0) | Identificador de la estructura. |
| idpers | NUMERIC (18,0) | Identificador de la persona. |
| idgeneracion | NUMERIC (3,0) | Identificador de la generación. |
| idfecha | NUMERIC (15,0) | Identificador de la fecha. |
| idcausa | NUMERIC (3,0) | Identificador de la causa de no inscripción. |

**Tabla 3.11. Descripción de la entidad "*cub_noinscritos*".**

Por último, la tabla de hechos que se relaciona con la relación de las generaciones controladas se denominará "*cub_sgc*". Los campos "*idestructura*", "*idpers*", "*idgeneracion*", "*idfecha*" y "*idsituacion*" forman las claves foráneas "*fk_estructura_preseleccion*", "*fk_prerrecluta_preseleccion*",

"*fk_generacion_preseleccion*", "*fk_fecha_preseleccion*" y "*fk_situacionrm_preseleccion*" respectivamente y a su vez conforman la clave primaria denominada "*pk_sgc*". En la siguiente tabla se resume la descripción de esta entidad:

| **Nombre: cub_sgc** | | |
|---|---|---|
| **Descripción:** Tabla de hechos que almacena los datos relacionados a la situación de las prerreclutas controlados por generaciones. | | |
| **Atributo** | **Tipo** | **Descripción** |
| idestructura | NUMERIC (5,0) | Identificador de la estructura. |
| idpers | NUMERIC (18,0) | Identificador de la persona. |
| idgeneracion | NUMERIC (3,0) | Identificador de la generación. |
| idfecha | NUMERIC (15,0) | Identificador de la fecha. |
| idsituacion | NUMERIC (3,0) | Identificador de la situación que presenta el prerrecluta en el registro militar. |
| preselec | NUMERIC (18,0) | Hecho para la cantidad de personas controladas. |

**Tabla 3.12. Descripción de la entidad "*cub_sgc*".**

### 3.6. Uniones entre dimensiones y hechos.

Las uniones entre las tablas dimensiones y las tablas de hechos se realizarán independientemente del modelo de bases de datos multidimensionales que se utilice. Estas uniones se realizan teniendo en cuenta las relaciones que existen entre las tablas de hechos y las dimensiones. La tabla de hechos "cub_listadounico" se relaciona con las dimensiones "dim_persona", "dim_estructura", "dim_generacion", "dim_fecha" y "dim_fuenteingreso". La tabla de hechos "cub_inscritos" se relaciona con las dimensiones: "dim_persona", "dim_estructura", "dim_generacion" y "dim_fecha"; del mismo modo la tabla de hechos "cub_noinscritos" se relaciona también con la dimensión "dim_causanoinsc". Finalmente, la tabla de hecho "cub_sgc" se relaciona con las dimensiones "dim_prerrecluta", "dim_estructura", "dim_generacion", "dim_fecha", "dim_situacionrm". En la siguiente figura se puede observar con claridad estas relaciones:

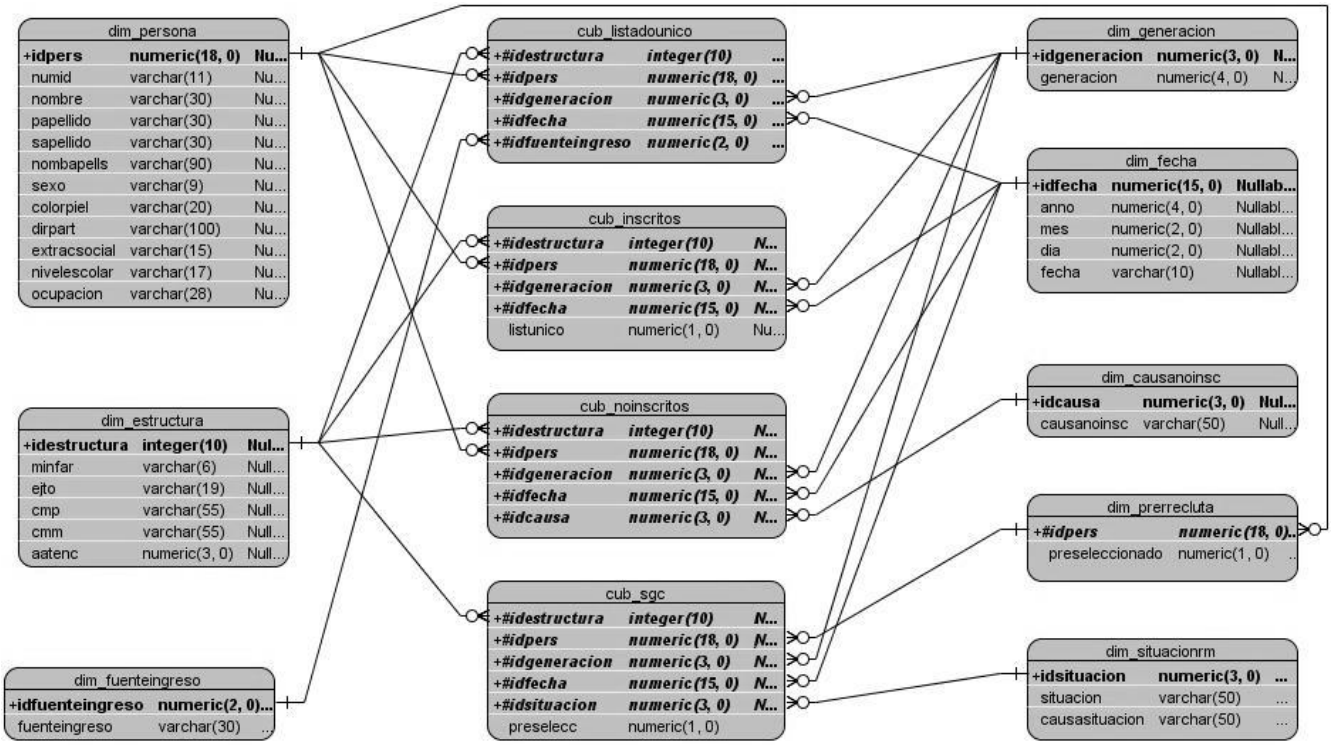


**Figura 3.1. Modelo del almacén de datos.**

Esta imagen corresponde al modelo realizado del almacén de datos para los subsistemas de Reclutamiento y Potencial Humano.

**Conclusiones.**

En este capítulo se detallan todos los aspectos relacionados al diseño del almacén de datos. Primeramente se abarca sobre la arquitectura que tendrá la solución propuesta. También se muestran un conjunto de aspectos que conforman los estándares para el diseño de los almacenes de datos. El principal resultado en el desarrollo del capítulo es la identificación de las dimensiones y hechos, además de las uniones establecidas entre las mismas, que forman el diseño del almacén de datos.

# Capítulo 4: Procesos ETL y Pruebas.

### Introducción.

Una vez diseñado el almacén de datos, se procederá a la carga de los datos mediante los procesos de extracción, transformación y carga. Para comenzar esta compleja fase es necesario realizar primeramente un mapeo de datos detallando el origen y destino de los mismos, luego se establecen un conjunto de restricciones y condiciones. Una vez diseñadas todas las transformaciones, se procede a la carga incremental de los datos hacia su destino según corresponda y finalmente se diseña la automatización de todo el proceso.

## 4.1. Mapeo de datos.

Con el objetivo de evitar las pérdidas de datos y posibilitar una ejecución amena de los próximos pasos, se realiza un mapeo de datos, indicando la fuente y el destino de los mismos, así como el tipo de dato que presenta en el origen y el que poseerá en el destino. De esta manera, se podrá apreciar con claridad algunas de las transformaciones a que se someterán los datos extraídos. En la siguiente tabla se muestran detalles del mapeo de datos realizado:

| Origen | | Destino | |
|---|---|---|---|
| **Campo** | **Tipo de dato** | **Campo** | **Tipo de dato** |
| persona.dat_persona.numid | VARCHAR(11) | public.dim_persona.numid | VARCHAR(11) |
| persona.dat_persona.nombre | VARCHAR(30) | public.dim_persona.nombre | VARCHAR(30) |
| persona.dat_persona.papel | VARCHAR(30) | public.dim_persona.papellido | VARCHAR(30) |
| persona.dat_persona.sapell | VARCHAR(30) | public.dim_persona.sapellido | VARCHAR(30) |
| persona.dat_persona.sexo | BIT(1) | public.dim_persona.sexo | VARCHAR(9) |
| persona.dat_persona.idcolorpiel | SMALLINT | public.dim_persona.colorpiel | VARCHAR(20) |
| persona.dat_persona.direccion | VARCHAR(100) | public.dim_persona.dirpart | VARCHAR(100) |
| persona.dat_persona.idextsocial | SMALLINT | public.dim_persona.extracsocial | VARCHAR(15) |
| persona.dat_persona.idnivesc | SMALLINT | public.dim_persona.nivelescolar | VARCHAR(17) |
| persona.dat_persona.idocupacion | SMALLINT | public.dim_persona.ocupacion | VARCHAR(28) |
| reclutamiento.nom_nodosapp. idmando | SMALLINT | public.dim_estructura.nivel1 | VARCHAR(19) |
| reclutamiento.nom_nodosapp. idprovincia | SMALLINT | public.dim_estructura.nivel2 | VARCHAR(55) |
| reclutamiento.nom_nodosapp. idmunicipio | SMALLINT | public.dim_estructura.nivel3 | VARCHAR(55) |
| persona.dat_direcciones.aatenc | SMALLINT | public.dim_estructura.nivel4 | NUMERIC(3,0) |

| | | | |
|---|---|---|---|
| reclutamiento.aux_generacion. generacion | SMALLINT | public.dim_generacion. generacion | NUMERIC(4,0) |
| reclutamiento.nom_fuenteingreso. idfingreso | SMALLINT | public.dim_fuenteingreso. idfuenteingreso | NUMERIC(2,0) |
| reclutamiento.nom_fuenteingreso. fingreso | VARCHAR(30) | public.dim_fuenteingreso. fuenteingreso | VARCHAR(30) |
| reclutamiento.nom_causaninsc. idcausaninsc | SMALLINT | public.dim_causanoinsc.idcausa | NUMERIC(2,0) |
| reclutamiento.nom_causaninsc. causaninsc | VARCHAR(50) | public.dim_causanoinsc. causanoinsc | VARCHAR(50) |
| reclutamiento.dat_reclutamiento. idpers | BIGINT | public.dim_prerrecluta. idpers | NUMERIC(18,0) |
| reclutamiento.dat_proceso. preselec | BIT(1) | public.dim_prerrecluta. preseleccionado | NUMERIC(1,0) |
| reclutamiento. nom_causasituacrm. idcausasituacrm | INTEGER | public.dim_situacionrm. idsituacion | NUMERIC(2,0) |
| reclutamiento. nom_causasituacrm.idsituacrm | INTEGER | public.dim_situacionrm.situacion | VARCHAR(50) |
| reclutamiento. nom_causasituacrm.causa | VARCHAR(50) | public.dim_situacionrm. causasituacion | VARCHAR(50) |

**Tabla 4.1. Mapeo de datos.**

En la tabla anterior, para agregar más información acerca de los campos que se muestran, se utiliza el siguiente formato: [esquema de la base de datos].[nombre de la tabla].[nombre del campo].

### 4.2. Condiciones adicionales.

Con el fin de lograr la carga de la información deseada en el almacén de datos, se establecerá un conjunto de condiciones adicionales teniendo en cuenta el análisis realizado en los capítulos anteriores. Estas condiciones deben analizarse cuidadosamente para evitar pérdidas significativas de datos.

Teniendo en cuenta los problemas detectados en el análisis realizado a los datos se establecen las siguientes condiciones:

- A los campos que contengan valores nulos se les asignará el valor: “DESCONOCIDO”, siempre que el tipo de dato en el destino lo permita. En el caso de que sean valores numéricos se le establecerá un valor que se encuentre en el rango admitido pero que permita detectar la anomalía.

- Los datos que tengan como identificador de origen (*idorig*) el valor 101010 no serán cargados, debido a que los mismos no son de importancia para los análisis realizados por los usuarios.
- Teniendo en cuenta que el capo "*preselec*", cuando tiene valor *null* presenta un determinado significado, se agregará el valor 0.
- Añadir un nuevo identificador de persona y estructura, pero se almacenarán los identificadores originales y los nuevos que se generaron en una tabla auxiliar para realizar las transformaciones.

### 4.3. Estándares y acciones previas al diseño del proceso ETL.

Antes de comenzar a diseñar los procesos de extracción, transformación y carga que se realizarán a los datos, fue necesario establecer un estándar para organizar el trabajo. Para dar cumplimiento a esto, se determinó nombrar a todas las transformaciones con el nombre que tiene la tabla que constituye el destino de los datos transformados, además se agregará al nombre el prefijo "Transf_". De igual forma, el nombre de cada paso en la transformación debe sugerir la acción que se realiza.

A modo de seguridad se creó un repositorio o catálogo donde se almacenarán todas las transformaciones y trabajos diseñados, además serán almacenados en formato XML y en el que provee la herramienta Kettle Pentaho Data Integration (ktr y kjb).

### 4.4. Extracción, transformación y carga de los datos.

Antes de comenzar el diseño de las transformaciones de los datos fue necesario realizar una configuración de las conexiones a las bases de datos fuente y destino (ver anexo 19). También fue necesario crear dos tablas auxiliares donde se almacenarán temporalmente un conjunto de datos que serán útiles para la ejecución de las transformaciones. Estas tablas recopilarán datos relacionados a las personas y las estructuras, las mismas recibieron el nombre "*aux_persona*" y "*aux_estructura*" respectivamente.

Para realizar la extracción de los datos que serán transformados y cargados en el almacén de datos se crearon un conjunto de consultas SQL que dan cumplimiento a este objetivo. Una vez obtenidos los datos, se procede a la transformación de los mismos en caso que lo requiera, las mismas se realizarán mediante una secuencia de pasos, utilizando los componentes que brinda la herramienta utilizada para la integración (ver anexos 4-18). Entre los pasos más comunes que se realizaron se

encuentran: la agregación de nuevos identificadores, búsquedas, cambios de valores y renombramiento y selección de los campos. Luego del diseño y construcción de estos pasos se procede a agregar otros que facilitan la carga de los mismos en el almacén de datos. En los anexos del 20 al 30 se muestran las transformaciones diseñadas mediante imágenes.

### 4.5. Cargas incrementales de los datos.

Las cargas incrementales en el almacén de datos se realizaron a partir de las transformaciones previamente diseñadas. Primeramente se diseñó un trabajo para cada tabla de hechos, donde se carga inicialmente las dimensiones asociadas y luego la tabla de hecho. También fue necesario crear un trabajo denominado “carga_inicial” donde se realiza una carga de datos de manera general. Esta carga inicial se refiere precisamente a la primera carga de datos que se realiza al almacén de datos, por lo general esta tarea consume bastante tiempo, ya que se deben de insertar todos los registros que han sido generados. Por otra parte, los restantes trabajos tienen la función de realizar los mantenimientos periódicos que mueven pequeños volúmenes de datos, y su frecuencia esta dada por el gránulo y los requerimientos de los usuarios.

Estos trabajos se diseñaron haciendo uso de los componentes que brinda la herramienta utilizada, permitiendo generar en todos los casos los registros detallados de todas las actividades ejecutadas con el fin de detectar posibles errores durante la ejecución de los mismos; también se muestra un mensaje un mensaje indicando el fin del mismo. En los anexos del 31 al 35 se muestran en detalle los trabajos diseñados a través de imágenes.

### 4.6. Diseño y construcción de la automatización de los procesos ETL.

Teniendo en cuenta la frecuencia de actualización de los datos en el sistema DATAFAR, se propone la automatización de los procesos de extracción, transformación y carga. En acuerdo con los especialistas del sistema se acordó realizar la carga de los datos relacionados al listado único mensualmente, con la misma frecuencia se cargarán los asociados a las generaciones controladas; por otra parte los datos relacionados a la inscripción se cargaran con una periodicidad de 15 días.

El objetivo principal de esta automatización es garantizar el mantenimiento o actualización de los datos, o sea, añadir al depósito aquellos datos nuevos que fueron generados después de la última

carga. La misma se llevó a cabo mediante la programación de los trabajos diseñados, haciendo uso de lo componente de inicio.

### 4.7. Pruebas de calidad.

El propósito de esta sección es describir el plan para realizar un conjunto de pruebas de calidad al almacén de datos. A continuación se explican un conjunto de características que deben cumplir los datos que poseen una adecuada calidad:

- Exactitud: mide el grado en que la información refleja lo que está pasando en el negocio.
- Totalidad: medición que refleja que las bases de datos contienen toda la información relevante para el negocio.
- Oportunidad: medición con la que se conoce si la información esta disponible cuando se requiere.
- Relevancia: permite determinar que la información sea útil a los usuarios que la están usando.
- La información debe tener el nivel de detalle requerido, dependiendo del nivel organizacional y el tipo de decisión al cual esté destinada la información.
- Consistencia: permite determinar que la información sea la misma en todas las áreas.

Estas características describen como se deben de encontrar los datos luego de haber realizado los procesos de extracción, transformación y carga. Para lograr la calidad requerida en el almacén de datos, se utiliza una estrategia propuesta por Ralph Kimball, que asegura la calidad de los datos mediante los siguientes pasos:

- Definir las reglas de calidad de los datos. Este paso consiste en definir un conjunto de reglas con el fin de realizar un análisis de los datos.
- Documentar los defectos de los datos.
- Pruebas. Validación del control de la calidad de los datos.

En el presente trabajo las reglas de calidad de los datos se crearon durante un análisis realizado a través del perfilado de datos, todos los defectos identificados quedaron documentados con todos los detalles.

También se diseñaron un conjunto de pruebas a realizar, las cuales tienen como objetivo probar las siguientes características:

- Acceso al almacén de datos.
- Proceso de extracción, transformación y carga de datos.
- Programación del proceso de extracción, transformación y carga.
- Veracidad de los datos cargados a partir de muestras tomadas.

Para lograr la ejecución correcta de las pruebas se utilizaron las siguientes herramientas:

- Kettle Pentaho Data Integration.
- PgAdmin III.

**Conclusiones.**

En este capítulo se explicaron los pasos realizados para diseñar los procesos de extracción, transformación y carga, así como el diseño de las cargas incrementales de datos y la automatización de dicho proceso. También se explicaron las características que deben ser probadas para garantizar la calidad del almacén de datos. Además, se puso en práctica un plan de pruebas que responde a las necesidades requeridas.

## Conclusiones.

En el presente trabajo se establece una solución que facilita el análisis de la información y tiene como objetivo el apoyo al proceso de toma de decisiones. Durante el desarrollo del mismo se obtuvo una guía para el desarrollo de almacenes de datos; además se realizó una selección de herramientas para el desarrollo del almacén de datos teniendo en cuentas las necesidades existentes y las políticas establecidas en el centro para el desarrollo de software. También se logró exitosamente la construcción y ejecución de los procesos de extracción, transformación y carga de los datos.

Es importante destacar que todos los objetivos del trabajo se cumplieron en el tiempo estimado, teniendo en cuenta la planificación realizada para cada una de las tareas.

## Recomendaciones.

Continuar el desarrollo de la solución, implementando las herramientas y técnicas para el análisis de la información.

Extender la solución a los restantes módulos del sistema DATAFAR.

Lograr la integración con el sistema de recuperaciones dinámicas y la representación geoespacial con el fin de enriquecer las posibilidades de un mejor análisis de información.

Tener en cuenta las nuevas necesidades de información que presenten los usuarios.

## Bibliografía.

1. **Bernabeu, Ricardo Dario.** *DATA WAREHOUSING: Investigación y Sistematización de Conceptos - HEFESTO: Metodología propia para la Construcción de un Data Warehouse.* [PDF] Córdoba, Argentina : s.n., 21 de Abril de 2009. Vol. 1.1.

2. **Systeme, Anwendungen and Produkte.** SAP. [En línea] 2010. http://www.sap.com.

3. **Oracle.** Oracle. [En línea] 2010. http://www.oracle.com/global/es/index.html.

4. **Microsoft.** Microsoft Business Intelligence. [En línea] Microsoft, 2009. http://www.microsoft.com/bi/.

5. **Pentaho Corporation.** Pentaho Open Source Business Intelligence. [En línea] 2010. http://www.pentaho.com/.

6. **Hartman Díaz, Yohanlena y Ramón Zequeira, Dailen.** *Implementación del proceso de extracción, transformación y carga en un almacén de datos operacional para CIMEX.* Facultad 3, Universidad de las Ciencias Informáticas. Ciudad de la Habana : s.n., 2009. págs. 24-35, Trabajo de Diploma.

7. **International Business Machines.** IBM. [En línea] 2010. http://www.ibm.com/es/es/.

8. **Microsoft.** Microsoft Developer Network. *MSDN.* [En línea] 2010. http://msdn.microsoft.com/es-es/default.aspx.

9. **Embarcadero Technologies, Inc.** Embarcadero. [En línea] 2010. http://www.embarcadero.com.

10. **fabFORCE.** fabFORCE.net. *Fabulous Force Database Tools.* [En línea] http://www.fabforce.net.

11. **PostgreSQL Global Development Group .** PostgreSQL. [En línea] 2010. http://www.postgresql.org.

12. **Villanueva Ojeda, Álvaro.** *Análisis, Diseño e Implementación de un Data Warehouse de Soporte de Decisiones para un Hospital del Sistema de Salud Público.* Facultad de Ciencias e Ingeniería, Pontificia Universidad Católica del Perú. Lima, Perú : s.n., 2008. Trabajo de Diploma.

13. **InformationWeek.** *Analytics Alerts. Next-Gen BI Is Here.* 2009.

14. **Ibermática.** *Business Intelligence.* 2008.

15. **Ec. Hochsztain, Esther y Ing. Tasistro, Andrómaca.** *DATA MINING y DATA WAREHOUSE. METODOLOGÍAS PARA ANALISIS Y EXPLORACIÓN DE DATOS. APLICACIONES EMPRESARIALES.* Facultad de Ciencias Económicas y de Administración, Universidad de la República. Montevideo, Uruguay : s.n., 2009. 301/09.

16. **Chapman, Pete, y otros.** *Metodología CRISP-DM para minería de datos.* Dataprix. 2007.

# BIBLIOGRAFÍA

17. **Wolff, Carmen.** *Implementando un DataWarehouse.* Ingeniería Informática y Ciencias de la Computación, Universidad de Concepción. Concepción, Chile : s.n., 2002.

18. **Kimball, Ralph, y otros.** *The Data Warehouse Lifecycle Toolkit. Practical Techniques for Building Datawarehouse and Business Intelligence Systems.* [ed.] Robert Elliott, y otros. Segunda Edicion. s.l. : John Wiley & Sons, 2008. ISBN: 978-0-470-14977-5.

19. **Kimball, Ralph y Caserta, Joe.** *The Data Warehouse ETL Toolkit. Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* [ed.] Mary Bednarek, y otros. s.l. : John Wiley & Sons, 2004. ISBN: 0-764-57923-1.

20. **Kimball, Ralph y Ross, Margy.** *The Data Warehouse Toolkit. The Complete Guide to Dimensional Modeling.* [ed.] Robert Elliott, y otros. Segunda Edición. s.l. : John Wiley & Sons, 2002. ISBN 0-471-20024-7.

21. **Kimball, Ralph, y otros.** *The Kimball Group Reader. Relentlessly Practical Tools for Data Warehousing and Business Intelligence.* [ed.] Robert Elliott, y otros. s.l. : John Wiley & Sons, 2010. ISBN 978-0-470-56310-6.

22. **Ruiz Perez, Rafael y Fernandez de Castro Espinosa, Francisco.** *Normativas para la implantación de los módulos del sistema informático para el trabajo en los Comités Militares y Órganos de Organización y Personal en las unidades de las FAR.* Ministerio de las Fuerzas Armadas Revolucionarias. La Habana, Cuba : s.n., 2008. Normativa.

23. **López Miera, Álvaro.** *Indicaciones a las Dirección de Organización y Personal.* Ministerio de las Fuerzas Armadas Revolucionarias. La Habana, Cuba : s.n., 2009. Indicación. 14282-0.

# Referencias Bibliográficas.

1. **Bernabeu, Ricardo Dario.** *DATA WAREHOUSING: Investigación y Sistematización de Conceptos - HEFESTO: Metodología propia para la Construcción de un Data Warehouse.* [PDF] Córdoba, Argentina : s.n., 21 de Abril de 2009. Vol. 1.1.

2. **Systeme, Anwendungen and Produkte.** SAP. [En línea] 2010. http://www.sap.com/solutions/sapbusinessobjects/sap-crystal-solutions/index.epx.

3. **Oracle.** Oracle. [En línea] 2010. http://www.oracle.com/lang/es/appserver/business-intelligence/index.html.

4. **Microsoft.** Microsoft Business Intelligence. [En línea] Microsoft, 2009. http://www.microsoft.com/bi/productsbi/.

5. **Pentaho Corporation.** Pentaho Open Source Business Intelligence. [En línea] 2010. http://www.pentaho.com/products/.

6. **Hartman Díaz, Yohanlena y Ramón Zequeira, Dailen.** *Implementación del proceso de extracción, transformación y carga en un almacén de datos operacional para CIMEX.* Facultad 3, Universidad de las Ciencias Informáticas. Ciudad de la Habana : s.n., 2009. págs. 24-35, Trabajo de Diploma.

7. **International Business Machines.** IBM. [En línea] 2010. http://www-01.ibm.com/software/es/websphere/index.html?&S_TACT=none&S_CMP=none?.

8. **Microsoft.** Microsoft Developer Network. *MSDN.* [En línea] 2010. http://msdn.microsoft.com/en-us/library/ms141026.aspx.

9. **Embarcadero Technologies, Inc.** Embarcadero. [En línea] 2010. http://www.embarcadero.com/products/er-studio-business-architect.

10. **fabFORCE.** fabFORCE.net. *Fabulous Force Database Tools.* [En línea] http://www.fabforce.net/dbdesigner4/.

11. **PostgreSQL Global Development Group.** PostgreSQL. [En línea] 2010. http://www.postgresql.org/docs/.

## Glosario de términos.

**[A]**

**Apache Software Fundation:** Es una organización no lucrativa creada para dar soporte a los proyectos de software bajo la denominación *Apache*.

**Arquitectura cliente – servidor:** Consiste básicamente en un cliente que realiza peticiones a un servidor, y el mismo le da una respuesta. Esta idea también puede ser aplicada a programas que se ejecutan sobre una sola computadora.

**Automatización:** Sistema de producción en el que se usan máquinas en lugar de mano de obra; también puede verse como un proceso para lograr operaciones automáticas.

**[B]**

**Base de datos:** También llamado banco de datos, es un conjunto de datos pertenecientes a un mismo contexto y almacenados sistemáticamente para su posterior uso.

**[C]**

**Cardinalidad:** Indica el número o cantidad de elementos de un conjunto, sea esta cantidad finita o infinita.

**Case Sensitive:** Literalmente “sensible a las mayúsculas/minúsculas”. Es una expresión usada en jerga informática que se aplica a los textos en los que tiene alguna relevancia escribir un caracter en mayúsculas o minúsculas.

**Copia de resguardo o seguridad (*backup*):** En informática es un archivo digital, un conjunto de archivos o la totalidad de los datos considerados lo suficientemente importantes para ser conservados.

**COUNT**: Función definida en el lenguaje SQL para contar, también se puede encontrar definida en otros lenguajes.

**CRISP-DM:** Proceso estándar para realizar minería de datos.

**[D]**

**Dashboard:** Este término engloba a varias herramientas que muestran información relevante para la empresa a través de una serie de indicadores de rendimiento.

**[F]**

**Fuente abierta (Código abierto – *Open Source*):** Es el término con el que se conoce al software distribuido y desarrollado libremente. El código abierto tiene un punto de vista más orientado a los beneficios prácticos de compartir el código que a las cuestiones morales y/o filosóficas.

**[H]**

**Hardware:** Corresponde a todas las partes físicas y tangibles de una computadora.

**Heurística:** Se denomina con este término a la capacidad de un sistema para realizar de forma inmediata innovaciones positivas para sus fines.

**[I]**

**IBM (*International Business Machines*):** Es una empresa multinacional que fabrica y comercializa herramientas, programas y servicios relacionados con la informática.

**[J]**

**Java:** Es un lenguaje de programación orientado a objetos desarrollado por *Sun Microsystems*.

**Java Server Page (JSP):** Es una tecnología Java que permite generar contenido dinámico para la web, en forma de documentos HTML, XML o de otro tipo.

**[L]**

**Listado único:** Es el listado que se conforma durante un año con las personas que se deben inscribir en el Registro Militar en el siguiente año.

**Llave foránea:** Es un término de bases de datos, empleado en las relaciones entre tablas.

**Llave primaria:** Es un término de bases de datos que constituye una identificación unívoca, esta puede ser un atributo o una combinación de estos.

**[M]**

**Marketing:** Es el conjunto de técnicas, que a través de estudios de mercado intenta lograr el máximo beneficio.

**Microsoft:** Es una empresa multinacional estadounidense, fundada en 1975 por Bill Gates y Paul Allen, dedicada al sector de la informática.

**Ministerio de las Fuerzas Armadas Revolucionarias (MINFAR):** Es el organismo encargado de dirigir, ejecutar y controlar la aplicación política del Estado y el Gobierno en cuanto a preparación del país para la defensa. La defensa de la soberanía del Estado sobre todo el territorio nacional, la preparación y realización de la lucha armada; así como, la contratación, adquisición, producción y uso material de guerra que satisfaga las necesidades de la defensa; todas estas obligaciones son cumplidas con la participación de los demás órganos y organismos estatales, las entidades económicas, instituciones sociales y los ciudadanos.

**Ministerio del Interior (MININT):** Es el encargado mediante los órganos y estructuras que lo conforman de cumplir funciones de seguridad ciudadana, y de establecimiento del orden interior.

**[N]**

**Nivel de granularidad:** La granularidad representa el nivel de detalle al que se desea almacenar la información sobre el negocio que se esté analizando.

**[O]**

**Oracle:** Es una de las mayores compañías de software del mundo, sus productos van desde bases de datos hasta sistemas de gestión.

**[R]**

**Reclutamiento:** Es la actividad mediante la cual se incorpora a los ciudadanos en el Servicio Militar Activo, según los requisitos establecidos por la ley No. 75 de la Defensa Nacional. Es ejecutada por

las comisiones de reclutamiento, que son las que determinan individualmente la situación de los jóvenes prerreclutas con relación al Servicio Militar Activo.

**Recursos Humanos:** En la administración de empresas, se emplea este término para el trabajo que aporta el conjunto de los empleados o colaboradores de una organización.

**Red distribuida:** Es una topología de red caracterizada por la ausencia de un centro individual o colectivo.

**Registro Militar:** Constituye un sistema único, que incluye los procedimientos y documentos de control individual de los prerreclutas, reservistas y milicianos, así como los medios y equipos de la reserva militar.

**Repositorio:** Es un sitio centralizado donde se almacena y mantiene información digital.

**Requerimiento:** Es una necesidad documentada sobre el contenido, forma o funcionalidad de un producto o servicio.

**Reservista:** Son todas las personas que mayores de 28 años de edad que no tienen un vínculo permanente y activo con las Fuerzas Armadas Revolucionarias, incluyendo los milicianos.

**[S]**

**Servicio Militar Activo (SMA):** Consiste en el cumplimiento de las obligaciones militares por los ciudadanos en las unidades o dependencias de las Fuerzas Armadas Revolucionarias.

**Servlet:** Programas de aplicación que es ejecutado por un servidor.

**Sistemas de información ejecutiva (Executive Information System - EIS):** Es una herramienta de inteligencia empresarial, orientada a usuarios de nivel gerencial, que permite monitorizar el estado de las variables de un área o unidad de una empresa a partir de información interna y externa a la misma.

**Software:** Es el conjunto de los programas de cómputo, procedimientos, regla, documentación y datos asociados que forman parte de las operaciones de un sistema de computación.

**Software de intermediación (*middleware*):** Este término abarca a todo el software distribuido necesario para el soporte de interacciones entre clientes y servidores.

**Solo lectura (Read-only):** Describe un registro o área de memoria que puede leerse, pero no se puede escribir.

**Subversion**: Software de sistema de control de versiones.

**SQL:** Es un lenguaje formal declarativo para manipular información en una base de datos.

**SUM:** Función del lenguaje SQL que devuelve el valor acumulado de una expresión.

**Sun Microsystems:** Es una empresa informática. Las siglas *SUN* se derivan de "*Stanford University Network*".

**[X]**

**XML:** Son las siglas en inglés de *Extensible Markup Language*, es un metalenguaje extensible de etiquetas desarrollado por *World Wide Web Consortium* (W3C).