DEPARTAMENTO DE CIENCIAS DE LA COMPUTACIÓN

CARRERA DE INGENIERÍA EN SISTEMAS

TESIS PREVIO A LA OBTENCIÓN DEL TÍTULO DE INGENIERO EN SISTEMAS E INFORMÁTICA

TEMA: ANÁLISIS Y ESTRUCTURACIÓN DE LA INFORMACIÓN HIDROCARBURÍFERA NACIONAL Y GEOESPACIAL PARA EL DISEÑO Y CONSTRUCCIÓN DE UN DATA WAREHOUSE PARA LA TOMA DE DECISIONES SOCIO-AMBIENTALES DEL PROGRAMA DE REPARACIÓN AMBIENTAL Y SOCIAL - PRAS

AUTOR: DÍAZ RAZO, RICARDO MIGUEL

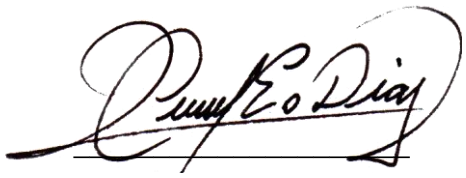
DIRECTOR: DÍAZ RODRIGUEZ, OSWALDO EFRAÍN

SANGOLQUÍ

2015

CERTIFICADO

Certifico que el presente proyecto titulado "Análisis y estructuración de la información hidrocarburífera nacional y geoespacial para el diseño y construcción de un data warehouse para la toma de decisiones socio-ambientales del Programa de Reparación Ambiental y Social - PRAS", fue desarrollado en su totalidad por el señor Ricardo Miguel Díaz Razo, bajo mi dirección.

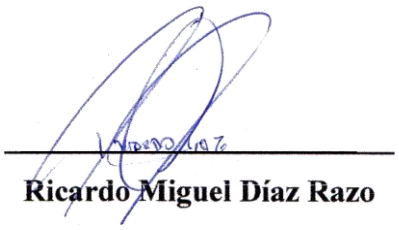


Ing. Oswaldo Díaz

AUTORÍA DE RESPONSABILIDAD

El presente proyecto titulado "Análisis y estructuración de la información hidrocarburífera nacional y geoespacial para el diseño y construcción de un data warehouse para la toma de decisiones socio-ambientales del Programa de Reparación Ambiental y Social - PRAS", ha sido desarrollado considerando los métodos de investigación existentes, así como también se ha respetado el derecho intelectual de terceros considerándolos en citas a pie de página y como fuentes en el registro bibliográfico.

Consecuentemente declaro que este trabajo es de mi autoría, en virtud de ello me declaro responsable del contenido, veracidad y alcance del proyecto en mención.

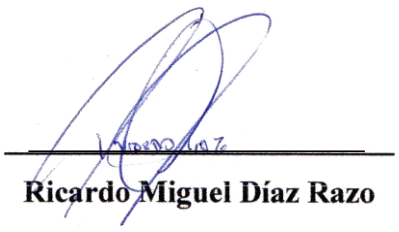


Ricardo Miguel Díaz Razo

AUTORIZACIÓN

Yo, Ricardo Miguel Díaz Razo, autorizo a la Universidad de las Fuerzas Armadas "ESPE" a publicar en la biblioteca virtual de la institución el presente trabajo "Análisis y estructuración de la información hidrocarburífera nacional y geoespacial para el diseño y construcción de un data warehouse para la toma de decisiones socio-ambientales del Programa de Reparación Ambiental y Social - PRAS", cuyo contenido, ideas y criterios son de mi autoría y responsabilidad.

Sangolquí, julio del 2015



Ricardo Miguel Díaz Razo

DEDICATORIA

El presente proyecto de tesis va dedicado a mis hijos Ricardo y Stefanía por su amor y su paciencia durante esta etapa de mi vida. También a mis padres Miguel Díaz y Eugenia Razo y por su apoyo, su confianza y su amor, y a todas las personas que han caminado en este largo camino junto a mi.

AGRADECIMIENTO

Agradezco a Dios por permitirme aprender cada día de mis errores y levantarme con más fuerza cada caida, al Ingeniero Oswaldo Díaz, por ponerme un gran reto para terminar con la frente en alto esta etapa de mi formación educativa, a mis padres por ser el mejor ejemplo que he podido tener. A la mujer de mi vida, Gabriela Jiménez Puente, por ser mi apoyo, mis fuerzas, mi guía y darme la vida que siempre imaginé tener. Gracias por tanto amor, por tanto sacrificio, por tanta paciencia y por estar a mi lado en todo momento, gracias a ti soy la persona, padre e hijo que ahora soy.

ÍNDICE DE GENERAL

ÍNDICE TABLAS

ÍNDICE DE FIGURAS

RESÚMEN

El Programa de Reparación Ambiental y Social- PRAS almacena información Hidrocarburífera y de Gestión Social del País en archivos semiestructurados y georeferenciados, lo cuál dificulta el cruce de información para el análisis y elaboración de reportes para su personal Directivo y Estadístico. El proyecto ha sido orientado a estructurar la información dentro de un Data Warehouse creado a partir del punto de equilibrio entre nivel de granularidad y grado de cohesión, además de utilizar las mejores prácticas de las metodologías más conocidas de la industria de Data Warehousing (Kimball e Inmon), y como herramienta integradora de datos y visualización se utilizó la plataforma de inteligencia de negocios Pentaho CE. Los resultados mostraron que, mediante el diseño de un modelo de punto de equilibro, la creación de cada uno de los Data Marts en esquema estrella al utilizar medidas compartidas partiendo de un modelo EDW proporciona una estructura de información eficiente y escalable además de rápida a nivel de procesamiento y consulta de la información lo cuál facilita el diseño y construcción del Data Warehouse. En conclusión, este proyecto propone una nueva metodología de creación de Data Warehouses, disminuyendo la complejidad del modelo y compartiendo información de dimensiones entre las diferentes tablas de hechos.

PALABRAS CLAVE:

- **PUNTO DE EQUILIBRIO**
- **NIVEL DE GRANULARIDAD**
- **GRADO DE COHESIÓN**
- **INMON**
- **PENTAHO COMMUNITY**

ABSTRACT

El Programa de Reparación Ambiental y Social- PRAS stores information about Hydrocarbons and Social Management of the Country in semistructured and georeferenced files, thus makes difficult the crossing of information for analysis and reporting to Directors and Statistical staff. The project was aimed to structure the information in a data warehouse created from the point of balance between granularity and cohesion, in addition to using the best practices of the best known methodologies in the industry of Data Warehousing (Kimball and Inmon) and as inte-grating data visualization tool was used Pentaho BI CE platform. The results showed that, by designing a model of balance point, the creation of each data mart star schema using shared dimensions that comes from an EDW model, provides an efficient and scalable structure information in addition to the design and construction of Data Warehouse. In conclusion, this project proposes a methodology for creating data warehouses, reducing the complexity of the model and sharing dimensions information among different fact tables.

KEY WORDS:

- **BALANCE POINT**
- **GRANULARITY LEVEL**
- **COHESION DEGREE**
- **INMON**
- **PENTAHO COMMUNITY**

CAPÍTULO 1
INTRODUCCIÓN

1.1 TEMA

Análisis y estructuración de la información hidrocarburífera nacional y geoespacial para el diseño y construcción de un data warehouse para la toma de decisiones socio-ambientales del Programa de Reparación Ambiental y Social – PRAS.

1.2 INTRODUCCIÓN

El Programa de Reparación Ambiental y Social – PRAS, tiene como finalidad visualizar y hacer comparativas mediante reportes entre los diferentes indicadores con los que cuenta actualmente la institución para poder analizar, investigar y tomar decisiones sobre la situación socio ambiental actual y las unidades territoriales con mayor afectación en el desarrollo de la actividad hidrocarburífera.

1.3 PLANTEAMIENTO DEL PROBLEMA

La Dirección de Investigación, orientada al procesamiento y sistematización de la información existente sobre daños históricos, demandas de actores afectados y notificaciones de los responsables, desarrolló en una primera etapa el Sistema de Indicadores de Pasivos Ambientales y Sociales dirigida a la actividad Hidrocarburífera Nacional (SIPAS-HN), para en un segundo momento enfocarse a otras actividades económicas.

El Programa de Reparación Ambiental y Social actualmente posee gran cantidad de información sobre la actividad hidrocarburífera nacional en archivos planos, la cual se encuentra sin estructurar ni depurar para ser analizada de una manera inmediata; esto se convierte en un gran inconveniente al momento de realizar reportes o informes para las autoridades y directivos

del Ministerio del Ambiente, lo que demanda un gran esfuerzo y trabajo para los técnicos del PRAS para poder entregar esta información en el momento requerido.

Es así que la demanda de un proceso de sistematización de datos, transformación y carga de una base de datos analítica es de suma importancia, por lo cual se ha decidido implementar una plataforma de Business Intelligence Open Source para el apoyo a la gestión del PRAS y las correspondientes áreas de la institución para permitir el uso racional de la información que facilite el proceso de control, evaluación y de toma decisiones.

1.4 JUSTIFICACIÓN E IMPORTANCIA

La implantación de una herramienta de inteligencia de negocios y la elaboración automatizada de reportes, van a reducir el esfuerzo de las coordinaciones para recopilar información y van a permitir al PRAS contar con información en el momento requerido y de manera fiable.

El Plan Nacional del Buen Vivir tiene como objetivo 1: "El Consolidar el Estado democrático y la construcción del poder popular" promueve la generación de proyectos que permitan:

- Mejorar continuamente los procesos, la gestión estratégica y la aplicación de tecnologías de la información y comunicación, para optimizar los servicios prestados por el Estado.
- Estandarizar procedimientos de la administración pública con criterios de calidad y excelencia con la aplicación de buenas prácticas.
- Maximizar el acceso a la información pública, oportuna, de calidad, comprensible y diversa.

La ley del Comercio electrónico establece que: para el cumplimiento de los objetivos del plan de gobierno electrónico el proyecto cumpla con las siguientes estrategias:

1. Acceso centralizado, entendiendo que la solución está disponible e integrada en un portal único de acceso, cumpliendo para ello con los estándares definidos en las normativas para el efecto.

2. Contenidos de capacitación, entendiendo que la solución concibe el acceso a contenidos actualizados para desarrollar capacidades para el buen uso de la mismas.

3. Documentos Electrónicos, entendiendo que la solución tiene un enfoque de "cero papeles", es decir, que genere documentos electrónicos; esto implica el uso de firma electrónica cuando sea necesario.

4. Autenticación Única, entendiendo que para el acceso a los servicios proporcionados por la solución se requiera un usuario y clave único.

5. Interoperable, entendiendo que la solución, en su concepto y arquitectura, facilita el intercambio de información pertinente con otras soluciones, para brindar un servicio más eficiente.

6. Mecanismos de evaluación de la percepción ciudadana, entendiendo que para los servicios desarrollados en la solución existen mecanismos claros y eficientes para receptar la percepción de los usuarios e incorporar la misma en el proceso de mejora contínua.

7. Esquema de datos abiertos, entendiendo que la solución en su diseño y arquitectura define esquemas para la apertura y reutilización de datos.

8. Accesibilidad y Usabilidad, entendiendo que la solución contempla que los servicios, por ella generados, son accesibles y de fácil uso indistintamente de la condición del usuario y del medio de acceso, para

lo cual deberá cumplir con los estándares definidos en la normatividad para el efecto.

1.5 ALCANCE

1. Levantamiento y análisis de la información.

   Se organizará una serie de entrevistas con las personas involucradas para la revisión de las fuentes de información necesarias para el cumplimiento exitoso de los requerimientos, utilizando la metodología Hefesto v2[1].

2. Elaboración de las siguientes soluciones:

   - Elaboración del Data Warehouse que contenga información estadística existente sobre pozos, plataformas y estaciones pertenecientes a la situación hidrocarburífera nacional, e información sobre conflictos, convenios y reclamos pertenecientes a gestión social.

   - Creación de procesos de Extracción, Transformación y Carga (ETLs), los que servirán para poblar cada uno de los Data Marts.

   - Creación del Cubo de Información para el análisis y visualización de Datos.

   - Creación de tres tableros de control Georeferenciados.

3. Instalación y Configuración de la Herramienta de Inteligencia de Negocios.

---

[1] **HEFESTO v2:** Se describe la metodología en el capítulo 2 “MARCO TEÓRICO”.

- Se realizará una instalación que permita un posterior crecimiento, escalabilidad y modularidad del sistema BI.
- Afinamiento de la Herramienta a implementar.
- Configuración de los siguientes ítems:
  - Ambiente Java, Sun JDK 1.7.
  - Variables Globales para el Sistema Operativo y la Herramienta a implementarse.
  - Conexiones a Bases de Datos.
  - Look & Feel.
  - Esquema de Seguridades (autenticación/autorización) de Pentaho (Framework Spring versión 2.0) en BDD de Pentaho.

1.6 OBJETIVOS

1.6.1. Objetivo General

Diseñar una solución tecnológica para la toma de decisiones socio-ambientales del Programa de Reparación Ambiental y Social – PRAS.

1.6.2. Objetivos Específicos

- Organizar, estandarizar y estructurar la información del PRAS registrada en archivos semiestructurados.
- Implementar el Data Warehouse para la integración de la información Hidrocarburífera y de Gestión Social existente.
- Generar procesos de extracción, transformación y carga para poblar el Data Warehouse.
- Visualizar información georeferenciada en tableros de control.

1.7 HERRAMIENTAS

1.7.1 MOTOR DE BASE DE DATOS

POSTGRESQL

Es un sistema de base de datos relacional "orientada a objetos" (diseñada para trabajar en conjunción con lenguajes de programación orientados a objetos como Java, C#, Visual Basic.NET y C++) que soporta distintos tipos de datos. Además del soporte para los tipos de datos base, también soporta datos de tipo fecha, monetarios, elementos gráficos, datos sobre redes, cadenas de bits, etc..

Entre sus características más importantes se pueden mencionar:

- Gran sistema de seguridad mediante la gestión de usuarios, grupos de usuarios , permisos y contraseñas.
- Gran capacidad de almacenamiento.
- Licencia de tipo BDS[2], lo cual permite manejar libremente su código fuente.

1.7.2 HERRAMIENTAS DE INTELIGENCIA DE NEGOCIOS

PENTAHO

Pentaho se define a si mismo como una plataforma de BI "orientada a la solución" y "centrada en procesos" que incluye todos los principales componentes requeridos para implementar soluciones basadas en procesos y ha sido concebido desde el principio para estar basada en procesos. Las soluciones que Pentaho pretende ofrecer se componen fundamentalmente de una infraestructura de herramientas de análisis e informes, integrado con un motor de workflow de procesos de negocio. La plataforma será capaz de

[2] **BSD:** Es una licencia de software libre permisiva como la licencia de OpenSSL o la MIT License. La licencia BSD al contrario que la GPL permite el uso del código fuente en software no libre.

ejecutar las reglas de negocio necesarias, expresadas en forma de un conjunto de actividades, que entregan la información adecuada en el momento adecuado.

Pentaho maneja dos modelos de ingresos, el primero esta orientado a los servicios (soporte, formación, consultoría y soporte a vendedores independiente a través de su portal web http://support.pentaho.com y distribuidores OEM[3]) y el segundo, otorgando funcionalidades "Premium" dentro de sus suscripciones Enterprise y Profesional.

En su web presenta una organización por productos: Reporting, Analysis, Dashboards y Data Mining, acompañado por dos introducciones: a la plataforma y a los productos. En dichas introducciones se hace mención específica al workflow como una de las capacidades BI claves de la plataforma.

1.7.3 FACTIBILIDAD

1.7.3.1 Factibilidad técnica

El desarrollador tiene amplia experiencia en la herramienta de inteligencia de negocios, además de ser **certificado** por "Pentaho Corporation" como consultor y **certificado** por "2nd Quadrant" como administrador de PostgreSQL , lo cual garantiza el cumplimiento del proyecto; por su parte la institución ha evaluado al responsable y ha concluido que está apto para la ejecución del mismo.

---

[3] **OEM:** Originalmente abreviatura de "Original Equipment Manufacturer" (fabricante de equipos originales). Actualmente son empresas revendedoras de un producto que tienen una relación directa con la fabrica con su propio nombre y marca.

1.7.3.2 Factibilidad tecnológica

A continuación se detallará los recursos requeridos para el desarrollo exitoso del proyecto:

- Hardware
  - Procesador Intel i7
  - Memoria RAM de 8 gigas
  - Disco duro de 750 gigas
- Software
  - Sistema Operativo Windows o Linux.
  - Java Development Kit (JDK) 1.7.
  - Pentaho BA 5.4 Community Edition

1.7.3.3 Factibilidad operativa

El Programa de Reparación Ambiental y Social PRAS requiere que la información, vistas de análisis y tableros de control se encuentren disponibles para el público en general 8x5, es decir, ocho horas diarias durante los cinco días laborables de la semana sobre una IP pública disponible para cualquier persona que disponga de una conexión a internet desde cualquier parte del país y del mundo.

1.7.3.4 Factibilidad operacional

La facilidad de uso de las herramientas a utilizar se han catalogado como: Media-Alta.

Para la ejecución exitosa del proyecto, los usuarios técnicos deben poseer conocimientos en:

- Diseño e Implementación de Base de Datos.

- Data Warehousing.
- ETL.

1.7.3.5 Factibilidad económica

Los detalles para la estimación económica del proyecto se detallan en la **Tabla 1**:

Tabla 1 Factibilidad Económica

CANTIDAD	DETALLE	VALOR UNTARIO	SUBTOTAL
HARDWARE			
1	Laptop Intel core i7	$1.200	$1.200
SOFTWARE			
1	Pentaho BA Community Edition 5.2	$0	$0
1	Java Development Kit	$0	$0
OTROS			
1	Desarrollador BI (456 horas)	$40	$18.240
		TOTAL	$19.440

El valor final del proyecto está valorado en $19.440, dicho costo será asumido en su totalidad por el estudiante con el objetivo de obtener su titulo de INGENIERO EN SISTEMAS E INFORMÁTICA.

1.7.3.6 Factibilidad legal

- Plan nacional del buen vivir, Objetivo Nº 1.
- Estrategias de la ley de comercio electrónico:

1. Acceso centralizado, entendiendo que la solución está disponible e integrada en un portal único de acceso, cumpliendo para ello con los estándares definidos en las normativas para el efecto.

2. Contenidos de capacitación, entendiendo que la solución concibe el acceso a contenidos actualizados para desarrollar capacidades para el buen uso de la mismas.

3. Documentos Electrónicos, entendiendo que la solución tienen un enfoque de "cero papeles", es decir, que genere documentos electrónicos. Esto implica el uso de firma electrónica cuando sea necesario.

4. Autenticación Única, entendiendo que para el acceso a los servicios proporcionados por la solución se requiera un usuario y clave único.

5. Interoperable, entendiendo que la solución, en su concepto y arquitectura, facilita el intercambio de información pertinente con otras soluciones, para brindar un servicio más eficiente.

6. Mecanismos de evaluación de la percepción ciudadana, entendiendo que para los servicios desarrollados en la solución existe mecanismos claros y eficientes para receptar la percepción de los usuarios e incorporar la misma en el proceso de mejora contínua.

7. Esquema de datos abiertos, entendiendo que la solución en su diseño y arquitectura define esquemas para la apertura y reutilización de datos.

8. Accesibilidad y Usabilidad, entendiendo que la solución contempla que los servicios, por ella generados, son

accesibles y de fácil uso indistintamente de la condición del usuario y del medio de acceso, para lo cual deberá cumplir con los estándares definidos en la normatividad para el efecto.

1.7.4 RECURSOS DEL PERSONAL TÉCNICO

Ya que es un proyecto de Inteligencia de Negocios fundamentado en la parte técnica y gerencial del área de ingeniería en sistemas se ve necesario la intervención de:

- **Director de tesis:** Ingeniero de Sistemas.
- **Informantes (para elaboración de plan de tesis)**: Ingenieros de Sistemas.
- **Tesista:** Egresado de la carrera de Ingeniería de Sistemas.

CAPÍTULO 2
MARCO TEÓRICO

2.1 INTELIGENCIA DE NEGOCIOS

2.1.1. ¿Qué es Inteligencia de Negocios?

Suele definirse como la transformación de los datos de la compañía en conocimiento para obtener una ventaja competitiva (Gartner Group).

Es una estrategia empresarial que persigue incrementar el rendimiento de la empresa o la competitividad del negocio, a través de la organización inteligente de sus datos históricos (transacciones u operaciones diarias), usualmente residiendo en Data Warehouses corporativos o Data Marts departamentales. Desde un punto de vista más pragmático, y asociándolo directamente a las tecnologías de la información, se puede definir como el conjunto de metodologías, aplicaciones y tecnologías que permiten reunir, depurar y transformar datos de los sistemas transaccionales e información desestructurada (interna y externa a la compañía) en información estructurada, para su explotación directa (reporting, análisis OLAP.) o para su análisis y conversión en conocimiento soporte a la toma de decisiones sobre el negocio. (Indensa, n.d.)

Esta definición pretende abarcar y describir el ámbito integral del entorno BI, reflejado resumidamente en la **Figura 1** :

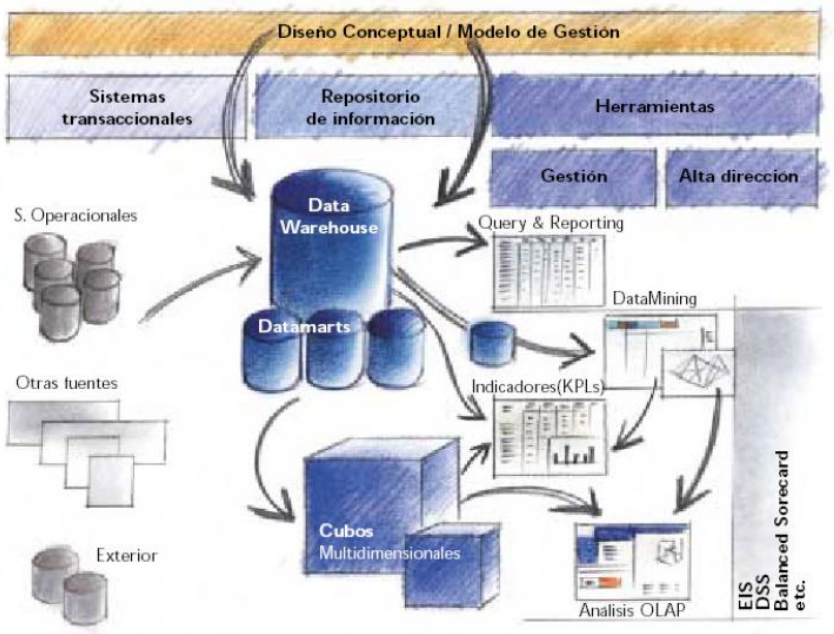


Figura 1: Modelo Integral de una solución BI
Fuente: (Espinoza, 2010)

Es importante considerar cualquier proyecto BI como un modelo objetivo integral. Algunas organizaciones han desarrollado proyectos parciales BI, sin tener en cuenta esta visión global, comprometiendo la calidad y efectividad de los resultados obtenidos.

2.1.2. Proceso BI

El proceso de BI que permite tener un acceso a la información de manera oportuna y acertada se describe en base a la **Figura 2**:

Figura 2: Proceso BI

2.1.3. Componentes de una solución de Inteligencia de Negocios

Una solución integral BI se compone de los siguientes elementos:

2.1.3.1 Diseño conceptual de los sistemas.

Para resolver el diseño de un modelo BI, se deben contestar a tres preguntas básicas: ¿Cuál es la información requerida para gestionar y tomar decisiones?; ¿Cuál debe ser el formato y composición de los datos a utilizar?; y ¿De dónde proceden esos datos y cuál es la disponibilidad y periodicidad requerida?. En otras palabras, el diseño conceptual tiene diferentes momentos en el desarrollo de una plataforma BI: En la fase de construcción del data warehouse y Data Marts, primarán los aspectos de estructuración de la información según potenciales criterios de explotación. En la fase de implantación de herramientas de soporte a la alta dirección, se desarrolla el análisis de criterios directivos: misión, objetivos estratégicos, factores de seguimiento, indicadores clave de gestión o KPIs, modelos de gestión.. en definitiva, información para el qué, cómo, cuándo, dónde y para qué de sus necesidades de información. Estos momentos no son necesariamente correlativos, sino que cada una de las etapas del diseño condiciona y es condicionada por el resto.

2.1.3.2 Construcción y alimentación del Data Warehouse.

Un data warehouse es una base de datos corporativa que replica los datos transaccionales una vez seleccionados, depurados y especialmente estructurados para actividades de query y reporting. Un Data Mart (o mercado de datos) es una base de datos especializada, departamental, orientada a satisfacer las necesidades específicas de un grupo particular de usuarios (en otras palabras, un data warehouse departamental, normalmente es un subconjunto del corporativo con transformaciones específicas para el área al que va dirigido).

La vocación del data warehouse es aislar los sistemas operacionales de las necesidades de información para la gestión, de forma que, posibles cambios realizados en aquéllos no afecten a éstas, y viceversa (únicamente cambiarán los mecanismos de alimentación, no la estructura, contenidos, etc.). No diseñar y estructurar convenientemente desde un punto de vista corporativo el data warehouse y los Data Marts generará problemas que pueden condenar al fracaso cualquier esfuerzo posterior: información para la gestión obtenida directamente a los sistemas operacionales, florecimiento de Data Marts descoordinados en diferentes departamentos, etc.

En definitiva, según la estructuración y organización de cada compañía, pueden originarse situaciones no deseadas y caracterizadas generalmente por la ineficiencia y la falta de calidad en la información resultante.

2.1.3.3 Herramientas de explotación de la información

Es el área donde más avances se han producido en los últimos años. Sin embargo, la proliferación de soluciones mágicas y su aplicación coyuntural para solucionar aspectos puntuales ha llevado, en ocasiones, a una situación de desánimo en la organización respecto a los beneficios de una solución BI.

2.1.3.4 Query & reporting:

Herramientas para la elaboración de informes y listados, tanto en detalle como sobre información agregada, a partir de la información de los data warehouses y Data Marts. Desarrollo a medida y/o herramientas para una explotación libre.

2.1.3.5 Cuadro de mando analítico (EIS tradicionales):

Elaboración, a partir de Data Marts, de informes resumen e indicadores clave para la gestión (KPI), que permitan a los gestores de

la empresa analizar los resultados de la misma forma rápida y eficaz. En la práctica, es una herramienta de query orientada a la obtención y presentación de indicadores para la dirección (frente a la obtención de informes y listados).

2.1.3.6 Cuadro de mando integral o estratégico:

Este modelo parte de que la estrategia de la empresa, es el punto de referencia para todo proceso de gestión interno.

Con él, los diferentes niveles de dirección y gestión de la organización disponen de una visión de la estrategia de la empresa traducida en un conjunto de objetivos, iniciativas de actuación e indicadores de evolución.

Los objetivos estratégicos se asocian mediante relaciones causa-efecto y se organizan en cuatro áreas o perspectivas: financiera, cliente, procesos y formación o desarrollo. El cuadro de mando integral es una herramienta que permite alinear los objetivos de las diferentes áreas o unidades con la estrategia de la empresa y seguir su evolución.

2.1.4. Niveles de uso de Datos

La Inteligencia de Negocios permite a los usuarios finales acceder y analizar de una forma eficaz y sencilla la información para la toma de decisiones dentro de la organización, a nivel operativo, táctico y estratégico, representado en la **Figura 3**.

Figura 3: Niveles de manejo de Datos

En el ámbito empresarial se requiere más que intuición para tomar decisiones correctas basadas en información exacta y actualizada. Los usuarios a través de las herramientas de análisis, consulta y reporte de datos son capaces de navegar en un mar de información y obtener en un espacio corto de tiempo informes confiables.

2.1.4.1 Nivel Operativo

Se compone básicamente de herramientas de reportes u hojas de cálculo con un formato fijo y cuya información se actualiza frecuentemente para que los usuarios operativos (empleados) tengan acceso a la misma de una manera oportuna y exacta.

2.1.4.2 Nivel Táctico

Los analistas de datos y gerencia media de la empresa tendrá el apoyo de herramientas de análisis y consulta con la finalidad de analizar la información sin la necesidad de intervención de terceros que les provean de información específica en un formato definido.

Este tipo de herramientas permitirán identificar por períodos, comportamientos que se estén produciendo dentro del negocio y de esta manera tomar ventaja con decisiones que tengan un impacto positivo para la empresa.

2.1.4.3 Nivel Estratégico

La Inteligencia de Negocios permite a la alta dirección de las empresas analizar y monitorear tendencias, patrones, metas y objetivos estratégicos de la organización.

Algunas ventajas de su implementación se describen a continuación:

- Promueve la alineación estratégica de toda la organización a partir de la transformación de la estrategia en planes concretos de acción.
- Con la finalidad de una estrategia bien definida, fomenta el trabajo en equipo y la colaboración de toda la organización.
- De una manera sencilla, hace eficaz la comunicación de los planes estratégicos a toda la empresa.
- Facilita la integración de un gran volumen de datos con sus indicadores que provienen de la gestión diaria de la organización.
- Permite a la organización cumplir con sus objetivos estratégicos en base al conocimiento.(Indensa, s.f.)

2.1.5. Aplicación de la Inteligencia de Negocios

2.1.5.1 Medición de Datos

La Cultura de Medición en una organización, permite proyectar habilidades de medición y evaluación de la gestión operativa, mediante los cuales se llegue a conocer los puntos críticos de desempeño, fomentando la generación de planes de acción que permitan trabajar sobre las debilidades, al igual que permite identificar las oportunidades de mejoramiento a nivel empresarial y personal.

Una organización puede diseñar un sistema de indicadores de gestión según sus necesidades. Su objetivo es entregar herramientas confiables que faciliten la visualización de la información actualizada y de manera oportuna que permitan acciones específicas para corregir situaciones encontradas.

Medir es un requisito para la mejora continua. Establecer el desempeño de la organización actual es el punto de partida con el cual se pueden plantear objetivos claros que generen verdaderos cambios en la madurez de la compañía. (Buenas Tareas, 2012).

2.2 GESTIÓN DE ALMACENAMIENTO

2.2.1 BASE DE DATOS

Es una colección integrada de datos organizada para satisfacer los requerimientos de información de los usuarios de una empresa, por medio de procesos de captura, validación, almacenamiento, actualización, integridad, cálculo, presentación, respaldo y restauración de datos; además de incluir los recursos, políticas y métodos de diseminación de la información. Los tipos de base de datos son:

2.2.1.1 Base de datos Transaccionales:

Son bases de datos cuyo único fin es el envío y recepción de datos a grandes velocidades, estas bases son muy poco comunes y están dirigidas por lo general al entorno de análisis de calidad, datos de producción e industrial, es importante entender que su fin único es recolectar y recuperar los datos a la mayor velocidad posible, por lo tanto la redundancia y duplicación de información no es un problema como con las demás bases de datos, por lo general para poderlas

aprovechar al máximo, permiten algún tipo de conectividad a bases de datos relacionales.

2.2.1.2 Base de datos Relacionales:

Éste es el modelo utilizado en la actualidad para representar problemas reales y administrar datos dinámicamente. El lugar y la forma en que se almacenen los datos no tienen relevancia (a diferencia de otros modelos como el jerárquico y el de red). Esto tiene la considerable ventaja de que es más fácil de entender y de utilizar para un usuario esporádico de la base de datos. La información puede ser recuperada o almacenada mediante "consultas" que ofrecen una amplia flexibilidad y poder para administrar la información.

2.2.1.3 Base de datos Multidimensional:

Son bases de datos ideadas para desarrollar aplicaciones muy concretas, como creación de Cubos OLAP. Básicamente no se diferencian demasiado de las bases de datos relacionales (una tabla en una base de datos relacional podría serlo también en una base de datos multidimensional), la diferencia está más bien a nivel conceptual; en las bases de datos multidimensionales los campos o atributos de una tabla pueden ser de dos tipos, o bien representan dimensiones de la tabla, o bien representan métricas que se desean aprender.

2.2.2 COMPONENTES DE LA BASE DE DATOS

- **Base de datos:** Es el depósito físico donde se almacenan los datos por medio de tablas, índices, ventanas, procedimientos y otras facilidades, cuya administración, respaldo y restauración requiere una estrecha relación con los recursos físicos y lógicos del computador.

- **Sistema manejador de base de datos (DBMS):** Es el programa que permite la definición y construcción de los elementos (tablas, reglas y procedimientos) de la base de datos, con la finalidad de controlar el ingreso, almacenamiento, actualización, integridad y recuperación de la información.

- **Repositorio:** Son las definiciones de base de datos, tablas, tipos de datos, consultas, ventanas, reglas, valores por omisión "default", procedimientos, reportes y otras definiciones que establecen la naturaleza del sistema y base de datos del usuario.

- **Lenguaje estructurado de consulta (SQL):** Es un programa orientado a crear, administrar y explotar la base de datos, por medio de un lenguaje estándar equivalente al inglés que se puede usar en cualquier manejador de base de datos.

- **Programas para desarrollo de aplicaciones:** Son programas que facilitan la creación, prueba y mantenimiento de procesos de consulta, cálculo y explotación de la base de datos.

- **Programas de aplicación:** Son los procedimientos creados para servir de interface entre el usuario y la base de datos para introducir, validar, actualizar y explotar la información, ejecutar procesos de cálculo, conversión, exportación, replicación y administración de datos, los cuales emplean instrucciones de SQL y programas para desarrollo de aplicaciones.

- **Administrador:** Crea, mantiene y administra la base de datos, supervisa su operación y empleo de recursos, establece y aplica las políticas de acceso, seguridad e integridad en el uso de datos a cargo de los usuarios. También vigilia el rendimiento y tiempo de respuesta del sistema.

- **Desarrollador:** Es el personal técnico encargado de crear los programas para operar la base de datos.

- **Usuario:** Son los interesados en introducir, actualizar y consultar los datos, conforme a las políticas establecidas por el administrador, utilizando los programas de aplicación.

2.2.3 NORMALIZACIÓN DE UNA BASE DE DATOS

- **Primera forma normal (1FF):** Remueve grupos repetidos de datos, dedicando un solo valor en cada atributo, es decir, el dato almacenado en el espacio representado en la intersección entre cada renglón y columna de una tabla .

- **Segunda forma normal (2FF):** Elimina dependencias parciales, generando agrupaciones de datos uniformes que pueden ser referencias entre sí para evitar redundancias.

- **Tercera forma normal (3FF):** Quita las dependencias transitivas, atributos que no son “llaves” llamados “determinantes”, al no contener valores únicos o no ser usados como referencias para acceso directo, son dependientes de uno o más atributos que tampoco son no llaves.

- **Boyce – Codd forma normal (BCFF):** Remueve las anomalías resultantes de dependencias funcionales cuando hay más de un atributo candidato a ser llave o bien cuando se usan varios atributos para formar una llave (compuesta), generalmente se deben separar los atributos determinantes que solo dependan de una parte de esos atributos candidatos a llave.

- **Cuarta forma normal (4FF):** Elimina dependencias multivaluadas, cuando por ejemplo existe una relación con tres atributos y por cada valor del primero hay un conjunto bien definido de valores del segundo y otro tanto para el tercero, sin

embargo el conjunto de valores de los dos últimos atributos son independientes uno del otro, se generan dos relaciones, una para el primer y segundo atributo y la otra para el primero con el tercero.

- **Quinta forma normal (5FF):** Remueve anomalías restantes, particularmente se enfoca a la situación que se genera cuando una relación tiene una dependencia tipo "join". Es aquella liga entre dos o más tablas para virtualmente formar una, al no poder dividirse en dos o más relaciones, por lo que las tablas resultantes llegan a convertirse en la tabla original, dando lugar a la necesidad de normalizar de nuevo la tabla. Sin embargo, este tipo de casos ocasionalmente son tolerados en aras de reducir el tiempo de respuesta a los accesos a varias tablas para satisfacer varias consultas del mismo tipo.(Ayala Peña, 2006)

2.2.4 CLAVES SUBROGADAS

Son aquellas que se definen artificialmente, son de tipo numérico secuencial, no tienen relación directa con ningún dato y no poseen ningún significado en especial.

- Ocupan menos espacio y son más eficientes que las tradicionales claves naturales, y más aún si estas últimas son de tipo texto.
- Son de tipo numérico entero (autonumérico o secuencial).
- Permiten que la construcción y mantenimiento de índices sea una tarea sencilla.
- El Data Warehouse no dependerá de la codificación interna del OLTP.
- Si se modifica el valor de una clave en el OLTP, el Data Warehouse lo tomará como un nuevo elemento, permitiendo de esta manera, almacenar diferentes versiones del mismo dato.

- Permiten la correcta aplicación de técnicas SCD (Dimensiones lentamente cambiantes). (Dario, 2010)

2.2.5 DATA WAREHOUSE

Las empresas de hoy en día se identifican por su gestión dinámica, donde las personas que las conforman deben tomar decisiones en forma rápida y efectiva basados en la última información disponible, para poder así conservar la ventaja competitiva.

Por otro lado, las compañías almacenan grandes volúmenes de datos en sus bases de datos operativas que se incrementan, en promedio, el doble cada año. Aun así, un mínimo porcentaje de estos datos es aprovechado para obtener una ventaja en las decisiones de sus negocios.

Actualmente las organizaciones están comprendiendo la importancia de la extracción de la información que se encuentra en sus bases de datos, necesaria para soportar las decisiones que deben ser tomadas por sus directivos, llegando así al concepto de data warehousing.

Un **Data warehouse** es una base de datos corporativa que se caracteriza por integrar y depurar información de una o más fuentes de procedencia variada, para luego procesarla permitiendo su análisis desde infinidad de perspectivas y con grandes velocidades de respuesta. La creación de un data warehouse representa en la mayoría de las ocasiones el primer paso, desde el punto de vista técnico, para implantar una solución completa y fiable de Business Intelligence.

El tipo de almacenamiento que permite este tipo de bases de datos: modelos de tablas en estrella, en copo de nieve, cubos relacionales, etc,

se considera una ventaja cuando se trata de analizar la información. (Sinnexus, © Copyright 2007 - 2012)

Las principales características de un Data Warehouse son:

- **Fuente exclusiva:** La procedencia de los datos de una organización es de distintas fuentes, tanto internas como externas, y en una gran diversidad de formatos. Sin importar cómo o de dónde vienen, para poder presentar estos datos al usuario final tienen que ser depurados, para asegurar su calidad e integridad;

- **Disponibilidad de la información:** Sin importar la procedencia de la información, se asegura un alto rendimiento satisfaciendo necesidades de movilidad y confidencialidad.

- **La información enfocada en la organización:** Los usuarios entienden mejor los datos si son presentados dentro del giro de negocio al que se dedican. Diccionarios de datos y catálogos de información creados por expertos en las áreas respectivas se convierten en una importante fuente para el desarrollo de sus actividades.
- **Automatización de la información:** A medida que los datos se convierten en información, van atravesando un camino cada vez más complejo. La automatización de estos mecanismos junto con los de distribución es un paso fundamental.

- **Calidad de la información y seguridad:** La información es el activo principal de toda compañía, y como cualquier otro activo debe ser administrado y protegido. Su calidad debe estar asegurada.

2.2.5.1 Principales aportaciones de un Data Warehouse

- Proporciona una herramienta para la toma de decisiones en cualquier área funcional, basándose en información integrada y global de la compañía.
- Facilita la aplicación de técnicas estadísticas de análisis y modelización para encontrar relaciones ocultas entre los datos del almacén; obteniendo un valor añadido para el negocio de dicha información.
- Proporciona la capacidad de aprender de los datos del pasado y de predecir situaciones futuras en diversos escenarios.
- Simplifica dentro de la empresa la implantación de sistemas de gestión integral.
- Supone una optimización tecnológica y económica en entornos de Centro de Información, estadística o de generación de informes con retornos de la inversión significativos.

2.2.5.2 Tipos de modelamiento de un Data Warehouse

- **Esquema en Estrella (Star)**

Una tabla de hechos y una tabla adicional por dimensión, su modelamiento se encuentra representado en la **Figura 4**.

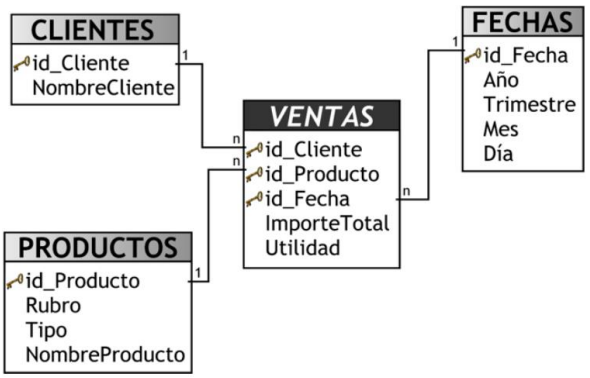


Figura 4 Esquema en Estrella

Fuente: (Leporati, n.d.)

- **Esquema Copo de Nieve (Snowflake)**

Una tabla de hechos central que está relacionada con una o más tablas de dimensiones, quienes a su vez pueden estar relacionadas o no con una o más tablas de dimensiones, su modelamiento se encuentra representado en la **Figura 5**.

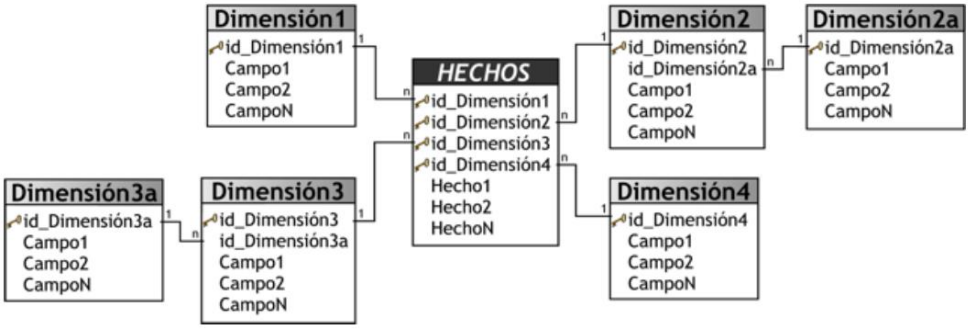


Figura 5 Esquema Copo de Nieve

Fuente: (Leporati, n.d.)

- **Esquema Constelación**

  Para cada esquema estrella o esquema del copo de nieve en almacén de datos es posible construir un esquema de constelación de hechos.

  Este esquema es más complejo que las otras arquitecturas debido al facto de que contiene multiples tablas de hechos. Con este solución las tablas de dimensiones pueden estar compartidas entre mas que una tabla de los factos, su modelamiento se encuentra representado en la **Figura 6**. (Leporati, s.f.)

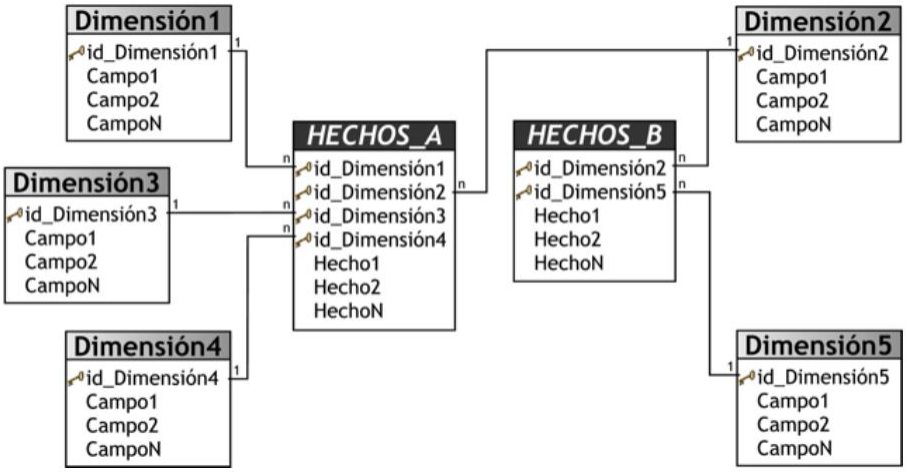


Figura 6 Esquema Constelación

Fuente: (Leporati, n.d.)

2.2.6 DATA MART

Un **Data Mart** es una base de datos departamental, especializada en el almacenamiento de los datos de un área de negocio específica. Se caracteriza por disponer la **estructura óptima de datos** para analizar la información al detalle desde todas las perspectivas que afecten a los procesos de dicho departamento.

Por tanto, para crear el Data Mart de un área funcional de la empresa es preciso encontrar la estructura óptima para el análisis de su información, estructura que puede estar montada sobre una base de datos OLTP, como el propio data warehouse, o sobre una base de datos OLAP. La designación de una u otra dependerá de los datos, los requerimientos y las características específicas de cada departamento. De esta forma se pueden plantear dos tipos de Data Marts:

2.2.6.1 Data Mart OLAP

Se basan en los populares cubos OLAP, que se construyen agregando, según los requerimientos de cada área o departamento, las dimensiones y los indicadores necesarios de cada cubo relacional. El modo de creación, explotación y mantenimiento de los cubos OLAP es muy heterogéneo, en función de la herramienta final que se utilice.

2.2.6.2 Data Mart OLTP

Pueden basarse en un simple extracto del data warehouse, no obstante, lo común es introducir mejoras en su rendimiento (las agregaciones y los filtrados suelen ser las operaciones más usuales) aprovechando las características particulares de cada área de la empresa. Las estructuras más comunes en este sentido son las *fact-tables* reducidas (que agregan las dimensiones oportunas), y las vistas materializadas, que se construyen con la misma estructura que las anteriores, pero con el objetivo de explotar la reescritura de queries (aunque sólo es posible en algunos SGBD avanzados, como Oracle). (Sinnexus, © Copyright 2007 - 2012)

Los Data Marts que están dotados con estas estructuras óptimas de análisis presentan las siguientes ventajas:

- Poco volumen de datos
- Mayor rapidez de consulta
- Validación directa de la información
- Facilidad para la historización de los datos

2.3 OLAP

Es un conjunto de tecnologías y aplicaciones de software que permiten recoger los datos de la organización, almacenarlos e indagar sobre ellos de forma rápida e intuitiva.

2.3.1 Cubos OLAP

Los cubos OLAP se generan mediante los esquemas sobre el Data Warehouse. La **Figura 7** muestra la generación de un Cubo OLAP.

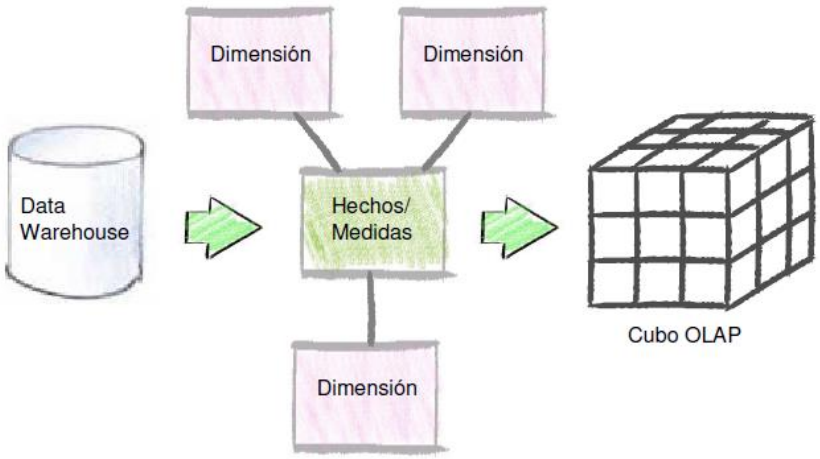


Figura 7 Cubos OLAP

Fuente: (Kimbal, 1996)

2.3.1.1 Modelado Multidimensional

Los datos en un Data Warehouse se modelan en "Cubos de Información", estructuras multidimensionales (Hipercubos) cuyas operaciones más comunes son:

- **Roll up:** Incremento en el nivel de agregación de los datos.
- **Drill down:** Incremento en el nivel de detalle, opuesto a roll up.
- **Slice:** Reducción de la dimensionalidad de los datos mediante selección.
- **Dice:** Reducción de la dimensionalidad de los datos mediante proyección.
- **Pivotaje o rotación:** Reorientación de la visión multidimensional de los datos. (Kimbal, 1996)

2.3.1.2 Tabla de Hechos.

Tabla central de la estructura. Contiene datos numéricos y proporciona información histórica.

2.3.1.3 Tabla de Dimensiones

Tablas adicionales relacionadas con la tabla de hechos.

2.3.1.4 Medida

Son valores de la tabla de hechos que se analizan y agregan.

2.3.1.5 Dimensión

Son valores de las tablas de dimensiones y describen un conjunto de miembros para ser analizados.

2.3.1.6 Jerarquía

Los miembros de las dimensiones se suelen organizar en forma de jerarquías, tal como lo muestra la **Figura 8**.

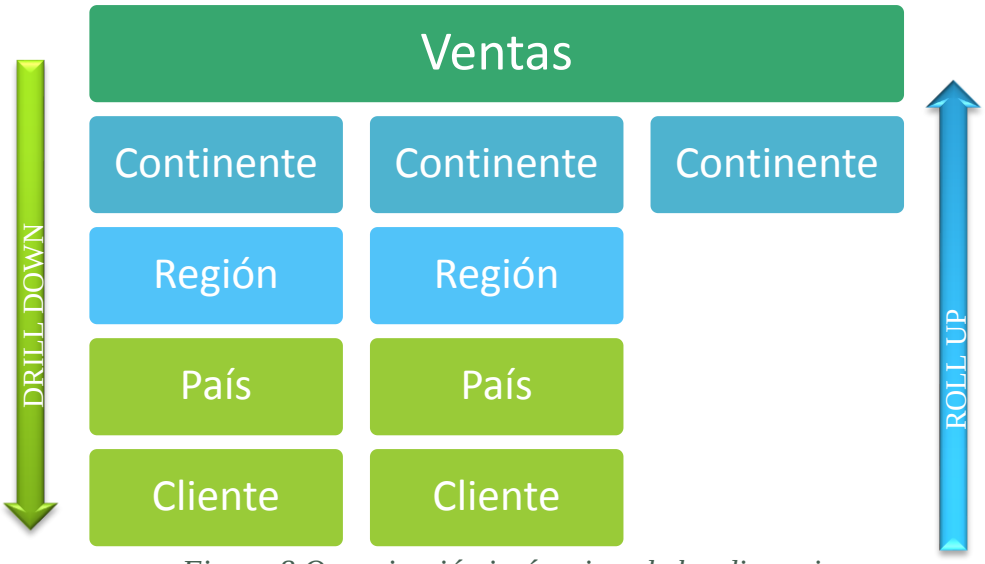


Figura 8 Organización jerárquica de las dimensiones

2.4 NIVEL DE GRANULARIDAD

La granularidad representa el nivel de detalle al que se desea almacenar la información sobre el negocio que se esté analizando. Por ejemplo, los datos referentes a ventas o compras realizadas por una empresa, pueden registrarse día a día, en cambio, los datos pertinentes a pagos de sueldos o cuotas de socios, podrán almacenarse a nivel de mes.

Mientras mayor sea el nivel de detalle de los datos, se tendrán mayores posibilidades analíticas, ya que los mismos podrán ser resumidos o sumarizados.(Bernabeu, 2010)

2.5 GRADO DE COHESIÓN

Grado en el cual los componentes de un módulo (típicamente las instrucciones individuales que lo conforman) son necesarios y suficientes para llevar a cabo una sola función bien definida. En la práctica, esto significa que el diseñador debe asegurarse de no fragmentar los procesos esenciales en módulos, y también debe asegurarse de no juntar procesos no relacionados en módulos sin sentido. Los mejores módulos son aquellos que son funcionalmente cohesivos (es decir, módulos en los cuales cada instrucción es necesaria para poder llevar a cabo una tarea bien definida). Los peores módulos

son los que son coincidentalmente cohesivos (es decir, donde sus instrucciones no tienen una relación significativa entre uno y otro).

Los grados de cohesión, de menor a mayor son:

- **Cohesión Coincidental:** No existe una relación significativa entre los elementos del módulo.
- **Cohesión Lógica:** La relación entre los elementos del módulo está basada en obtener ventajas en el procesamiento, por ejemplo, todos manipulan el mismo dato. Normalmente esto implica tener un código truculento o compartido, que degrada los propósitos de un buen diseño.
- **Cohesión Temporal:** Los elementos del módulo constituyen un conjunto que se ejecuta secuencialmente en un punto fijo en el tiempo. Aunque tiende, a veces, a confundirse con la cohesión lógica, la diferencia está en que este tipo de módulo es más simple y se ejecuta sin la intervención de otras aplicaciones.
- **Cohesión Comunicacional:** Se dice que un módulo tiene cohesión comunicacional si realiza actividades paralelas que usan los mismos datos de entrada y/o los mismos datos de salida.
- **Cohesión Secuencial**: Implica que la salida de un elemento es la entrada para el próximo.
- **Cohesión Funcional:** Un módulo tiene cohesión funcional si contiene elementos que contribuyen todos a la implementación de una sola función relacionada con el entorno del problema. (Bernal, 2012).

2.5.1 Determinación del grado de Cohesión

Se utiliza para determinar la cohesión:

2.5.1.1 Una sentencia descriptiva

Consiste en inspeccionar el módulo y tratar de escribir una sentencia que describe que hace el módulo (hay que mirar dentro de un módulo) de esta forma se puede deducir que cohesión tiene el módulo.

- Los módulos con cohesión funcional están descritos por una sentencia formada por un verbo imperativo y un nombre.
- Los módulos con cohesión secuencial quedan descritos mediante sentencias que contienen nombres de varias funciones.
- Los módulos con cohesión comunicacional quedan descritos mediante sentencias que contienen varios nombres de funciones estando relacionadas estas funciones por el hecho de que trabajan con los mismos datos de entrada y salida.
- Los módulos con cohesión temporal suelen contener referencias al tiempo.
- Los módulos con cohesión lógica suelen contener nombres de propósito general.
- Los módulos con cohesión casual suelen contener nombres poco significativos.

2.5.1.2 Un árbol de decisión

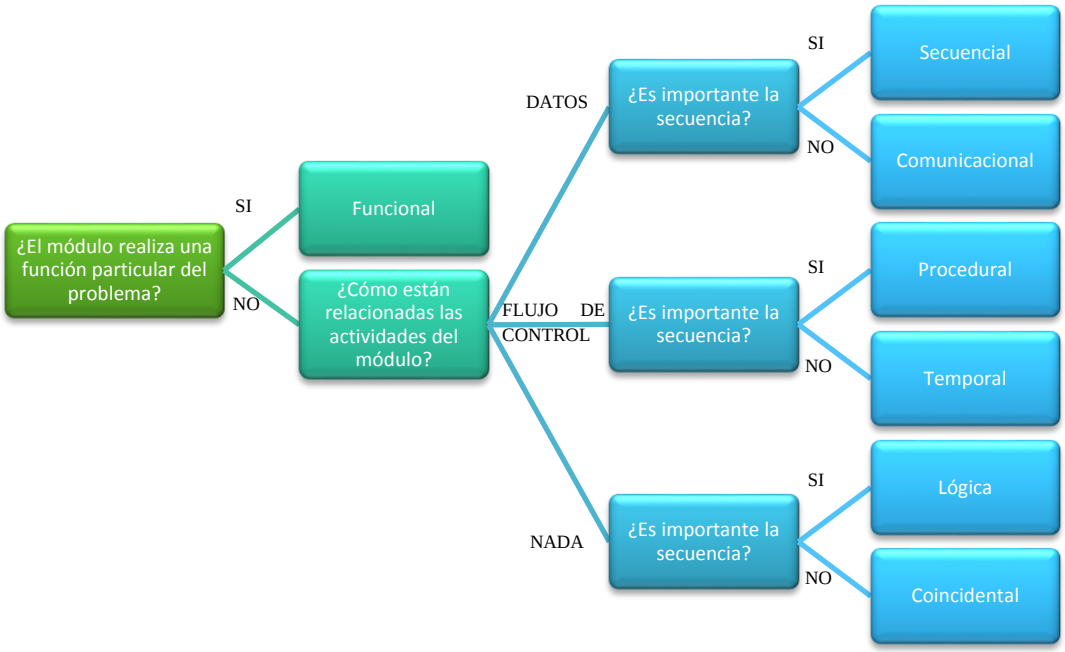


Figura 9 Árbol de Decisión Cohesión

2.6 ETL (extract, transform, load)

- **Extract:** Permite capturar la información desde los sistemas operacionales, para integrar:
  - Bases de datos
  - Ficheros planos
- **Transformar:** Es la adaptación de los datos fuente al formato destino definido en el Data warehouse, en el cual:
  - Se agregan datos numéricos
  - Se transpone información
  - Componer textos a partir de otros
- **Load (Cargar):** Es el proceso en que los nuevos datos son finalmente almacenados en el Data warehouse en su formato definitivo.(Bernabeu, 2010).

2.7 PENTAHO

Pentaho se define a sí mismo como una plataforma de BI “orientada a la solución” y “centrada en procesos” que incluye todos los principales componentes requeridos para implementar soluciones cuyo fundamento sean los procesos y ha sido concebido desde el principio para estar basada en procesos.

Las soluciones que Pentaho pretende ofrecer se componen fundamentalmente de una infraestructura de herramientas de análisis e informes, integrado con un motor de workflow de procesos de negocio. La plataforma será capaz de ejecutar las reglas de negocio necesarias, expresadas en forma de un conjunto de actividades, que entregan la información adecuada en el momento adecuado.

Su modelo de ingresos parece estar orientado a los servicios (soporte, formación, consultoría y soporte a ISVs y distribuciones OEM) aunque en alguno de los documentos y páginas que se ha examinado aparecen algunas funcionalidades “Premium” que hacen pensar en ingresos por futuras versiones o funcionalidades de pago.

En su web presenta una organización por productos: Reporting, Analysis, Dashboards y Data Mining, acompañado por dos introducciones: a la plataforma y a los productos. En dichas introducciones se hace mención específica al workflow como una de las capacidades BI claves de la plataforma.

2.7.1. CARACTERÍSTICAS

- Extracción, trasformación y carga de datos.
- Soporte para BI ágil.
- Presentación de informes.
- Análisis OLAP.
- Funcionalidades para Hadoop.
- Manejo de meta-datos.
- Cuadros de mando integral.
- Indicadores, métricas, KPI's.
- Integración y despliegue.
- Alertas configurables.
- Completamente estandarizado (XML, LOG4J, HTML, SQL, MDX, JOLAP, XML/A, CWM, etc.).
- Multiplataforma (Windows, Linux, Unix, MacOS).
- Servidor BI liviano.
- Desarrollado en lenguaje Java J2EE.
- Código abierto flexible.
- Conexión a cualquier BDD o fuente de información.
- Trabaja con múltiples servidores de aplicaciones.
- Soporta distribución y calendarización.
- Manejo de servicios Web (SOA).
- Integración fácil con portales Web.
- Manejo de Hadoop.
- Integración con bases de datos No SQL.
- Integración con R Studio.
- Integración con Alfresco.
- Auditoría en pantalla.
- Calendarización de reportes.

2.7.2. MÓDULOS DE PENTAHO

En la **Tabla 2** se muestra una breve descripción de los módulos de Pentaho:

Tabla 2 Módulos de Pentaho

MÓDULO	DESCRIPCIÓN
 PENTAHO DATA INTEGRATION	Consultar, limpiar e integrar los datos donde quiera que se encuentren (ETL).
 PENTAHO REPORTING	Acceder a los datos y entregar información a la organización en todos sus niveles.
 PENTAHO ANALYSIS	Explorar y analizar variables de manera interactiva con una respuesta rápida.
 PENTAHO DASHBOARD	Obtener una visibilidad inmediata del estado del negocio a través de las métricas y KPI's.
 PENTAHO DATA MINING	Descubrir patrones ocultos y los indicadores de desempeño futuro

2.7.3. PENTAHO DATA INTEGRATION (PDI)

Pentaho Data Integration es una herramienta de extracción, transformación y carga (ETL) robusta, que se puede utilizar para integrar, manipular y visualizar sus datos.

Se puede usar PDI para importar, transformar y exportar datos de múltiples fuentes de datos, incluyendo archivos planos, bases de datos relacionales, Hadoop, bases de datos NoSQL, bases de datos analíticas, flujos de medios sociales, y tiendas operativas en línea. También se puede utilizar PDI para limpiar y enriquecer los datos, mover datos entre bases de datos, y visualizarlos.

Pentaho Data Integration se compone de los siguientes módulos principales:

- **Spoon:** Es una aplicación de escritorio que utiliza una interfaz gráfica para la edición y creación de transformaciones y trabajos (jobs). Proporciona una forma visual para crear procesos ETLs complejos sin tener que leer o escribir código.

  Cuando se piensa en Pentaho Data Integration como un producto, Spoon es lo que viene a la mente porque, como desarrollador de base de datos, esta es la aplicación en la que pasará la mayor parte de su tiempo. Cada vez que el autor edite, ejecute o depure una transformación o trabajo, va a utilizar Spoon.

- **Pan:** Es un proceso independiente de línea de comandos que se puede utilizar para ejecutar transformaciones y trabajos creados en Spoon. El motor de transformación de datos "Pan" lee la información y escribe datos en varias fuentes de datos. Pan también permite manipular datos.

- **Kitchen:** Es un proceso independiente de línea de comandos que se puede utilizar para ejecutar los trabajos (jobs). Es el programa que ejecuta los "jobs" diseñados en la interfaz gráfica de Spoon, ya sea en XML o en un repositorio de la base de datos.

  Los "Jobs" se programan para ejecutarse en modo batch en intervalos regulares.

- **Carte:** Es un contenedor web ligero que le permite configurar un servidor ETL dedicado, a distancia. Esto proporciona capacidades de ejecución remota similares a las del servidor de integración de datos, pero no proporciona la programación, integración de seguridad ni un sistema de gestión de contenidos.

2.7.4. PENTAHO REPORTING: PENTAHO REPORT DESIGNER

Pentaho Report Designer es una herramienta de creación de informes sofisticada que se puede utilizar independiente o como parte de la gran distribución "Pentaho Business Analytics". Permite a los profesionales crear informes altamente detallados basados en datos preparados adecuadamente desde, prácticamente, cualquier fuente de datos.

El motor de Pentaho Reporting ofrece una funcionalidad única que no se encuentra en la competencia de soluciones integrables:

- **No requiere un JDK.** Mientras que usted no necesite un kit de desarrollo de Java instalado en su equipo de desarrollo, no es necesario un JDK para ejecutar un programa que incorpora el motor de Pentaho Reporting - sólo un Sun Java Runtime Environment estándar.

- **Todo el procesamiento se realiza en memoria.** No hay archivos temporales creados por el motor de informes. Un programa que se basa en el motor de Pentaho Reporting para la generación de informes se puede ejecutar en un sistema sin disco.

- **Componentes Dinámicos y ajustables.** El motor de Pentaho Reporting detecta los JARs para añadir funcionalidad en tiempo de ejecución, por lo que se puede añadir nuevos archivos JARs para ampliar las capacidades del motor, o eliminar JARs innecesarios para reducir la memoria y espacio en disco de su aplicación.

- **Bajo consumo de memoria.** Una aplicación basada en informes de Pentaho puede funcionar con un mínimo de 64 MB de memoria (128MB aunque aumentaría drásticamente la velocidad de procesamiento de informes).

- **Totalmente configurable a través de parámetros en tiempo de ejecución.** Cada estilo, función, consulta y elemento del informe es totalmente personalizable por el paso de parámetros para el motor de informes.

- **Integración OpenFormula.** OpenFormula es un estándar abierto para las fórmulas matemáticas. Se puede crear fácilmente sus propias fórmulas personalizadas, o puede personalizar las construidas en el motor de Pentaho Reporting.

- **Simple gestión de recursos.** Como utilizar el formato OpenDocument (ODF), el motor de Pentaho Reporting reúne todos los recursos de informe, incluyendo la información de la fuente de datos de conexión, consulta y recursos incluso binarios como imágenes en un archivo canónico. Esto simplifica la gestión de los recursos físicos y elimina los problemas de ruta relativa.

2.7.5. PENTAHO ANALYSIS: SAIKU

Las aplicaciones OLAP son uno de los pilares de cualquier solución de Inteligencia de Negocios, ya que proveen a los usuarios acceso a información sumarizada mediante convenientes métodos de navegación que le permitan analizar y mantener una conversación fluida con los datos de la organización, con óptimos tiempos de respuesta. **Saiku Analytics** es un proyecto para Pentaho Community que permite desarrollar rápidamente reportes AD HOC.

Entre las principales características de Saiku se tiene:

- Diseño de reportes Drag & drop.

- Permite exportar a: PDF, CSV, XLS, CDA, PRPT.

2.7.5.1 Importancia de la Visualización

La visualización de la información es crucial cuando se trata del análisis de datos de una organización (Ver **Figura 10, 11 y 12**). El cerebro humano puede tardar en reconocer patrones o marcas dentro de la información en reportes muy largos; esto se puede considerar como algo del pasado.

Poder representar los datos de manera visual mediante patrones reduce la complejidad de su análisis. En la actualidad, el usuario será capaz de visualizar gran cantidad de información en un espacio reducido facilitando la búsqueda de datos específicos, lo que incrementa la capacidad de toma de decisiones dentro de la organización.

La capa de presentación de datos contiene diferentes tipos de gráficos y además contiene una nueva forma de interactuar con los resultados - zoom.

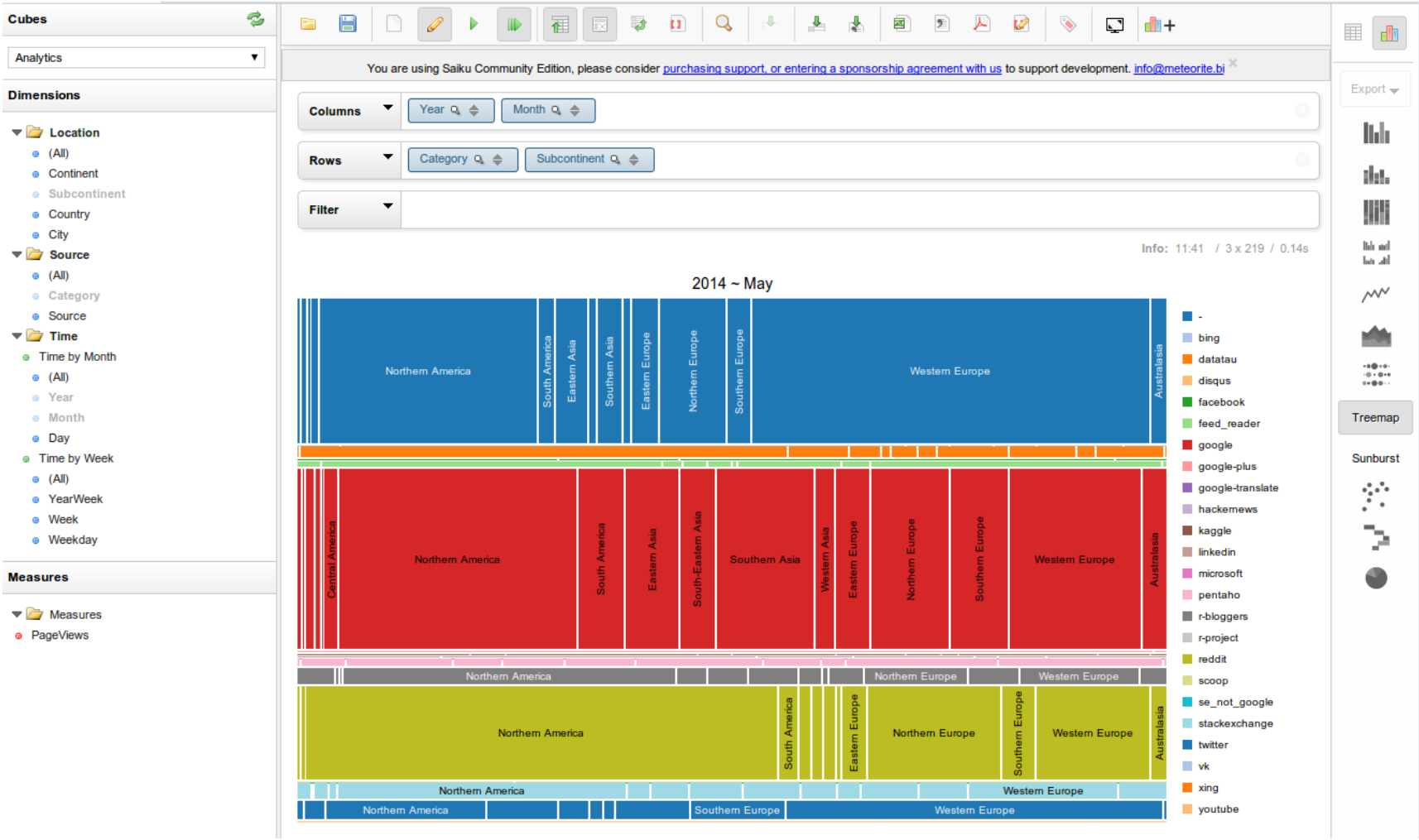
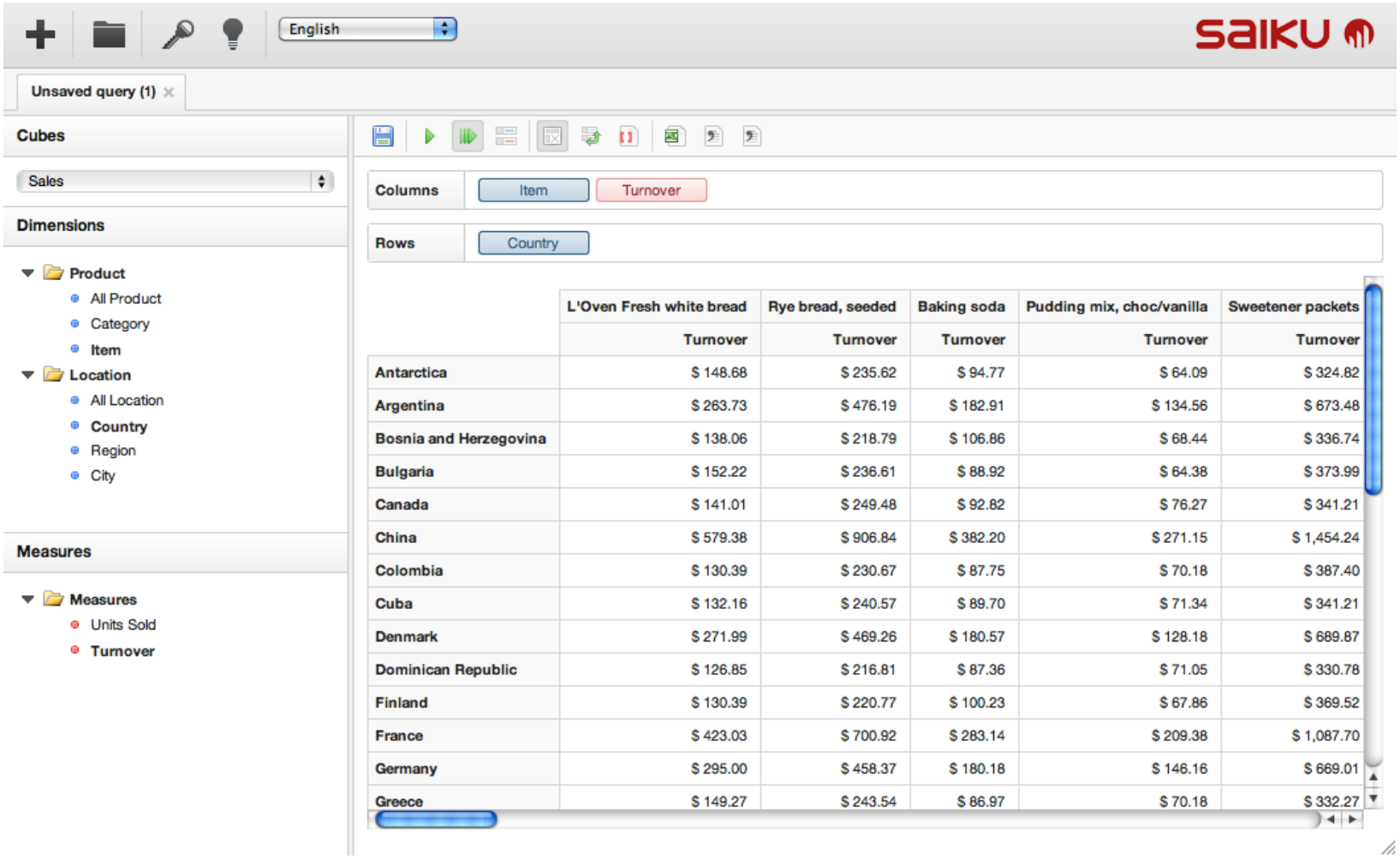


Figura 10 Saiku Gráficos

Figura 11 Saiku Gráfico Heat Map



	L'Oven Fresh white bread	Rye bread, seeded	Baking soda	Pudding mix, choc/vanilla	Sweetener packets
	Turnover	Turnover	Turnover	Turnover	Turnover
Antarctica	$ 148.68	$ 235.62	$ 94.77	$ 64.09	$ 324.82
Argentina	$ 263.73	$ 476.19	$ 182.91	$ 134.56	$ 673.48
Bosnia and Herzegovina	$ 138.06	$ 218.79	$ 106.86	$ 68.44	$ 336.74
Bulgaria	$ 152.22	$ 236.61	$ 88.92	$ 64.38	$ 373.99
Canada	$ 141.01	$ 249.48	$ 92.82	$ 76.27	$ 341.21
China	$ 579.38	$ 906.84	$ 382.20	$ 271.15	$ 1,454.24
Colombia	$ 130.39	$ 230.67	$ 87.75	$ 70.18	$ 387.40
Cuba	$ 132.16	$ 240.57	$ 89.70	$ 71.34	$ 341.21
Denmark	$ 271.99	$ 469.26	$ 180.57	$ 128.18	$ 689.87
Dominican Republic	$ 126.85	$ 216.81	$ 87.36	$ 71.05	$ 330.78
Finland	$ 130.39	$ 220.77	$ 100.23	$ 67.86	$ 369.52
France	$ 423.03	$ 700.92	$ 283.14	$ 209.38	$ 1,087.70
Germany	$ 295.00	$ 458.37	$ 180.18	$ 146.16	$ 669.01
Greece	$ 149.27	$ 243.54	$ 86.97	$ 70.18	$ 332.27

Figura 12 Saiku Análisis

Utilizando el Drill down la herramienta fácilmente permite analizar un área específica que es de interés, utilizando el Drill up permite regresar a los datos de una manera general, esto es posible manejarlo en los gráficos y tablas que son parte de un reporte.

2.7.6. PENTAHO COMMUNITY DASHBOARD

Para elaborar tableros de control, Pentaho utiliza cuatro componentes desarrollados por la empresa portuguesa Webdetails, los cuales se detallan a continuación:

- **Community Dashboard Framework (CDF):** Es un framework de código abierto que permite la creación de cuadros de mando altamente personalizables. CDF se basa en estándares de desarrollo web como CSS, HTML5 y JavaScript (aprovechando algunos esquemas de uso frecuente como jQuery o Bootstrap).

  CDF está dirigido a desarrolladores y consiste en una solución eficaz para combinar datos con una capa de visualización atractiva.

- **Community Data Access (CDA):** Es un plugin diseñado para acceder a datos de cualquier fuente de información con una gran flexibilidad. Permite además conectarse a diferentes fuentes de datos simplemente editando un archivo XML y entregar los datos en diferentes formatos (csv, xls, etc.) a través de la consola de usuario Pentaho.

  CDA puede ser utilizado como un complemento independiente en el servidor Pentaho BI o en combinación con CDE / CDF.

- **Community Chart Components (CCC**): Proporciona a los desarrolladores la ruta para incluir en sus cuadros de mando gráficos básicos sin perder el principio fundamental:

extensibilidad. Los gráficos de la CCC son visualmente atractivos, son flexibles y permiten la interacción además de ser altamente personalizables.

- **Community Graphic Generator (CGG):** Permite al usuario exportar informes CCC / CDE como imágenes. Este plugin es capaz de hacer, en el servidor, exactamente el mismo gráfico que se representa en el navegador por CDE / CDF.

Todos los componentes mencionados permiten el desarrollo y despliegue de Pentaho Dashboards de una manera rápida y eficaz.

2.7.7. PENTAHO DATA MINING (WEKA)

Es una suite de herramientas que usa estrategias de aprendizaje automático y minería de datos. Cuenta con series de clasificación, de regresión, reglas de asociación y de algoritmos de clustering para apoyar las tareas de análisis predictivo.

Weka está diseñado como una herramienta orientada a la extensibilidad por lo que añadir nuevas funcionalidades es una tarea sencilla

La licencia de Weka es GPL, lo que permite una libre distribución y difusión. Además, ya que Weka está programado en Java, es independiente de la arquitectura, funciona en cualquier plataforma sobre la que haya una máquina virtual Java disponible.

Sus características en base a su arquitectura y funcionalidad son:

- Interface gráfica de usuario (visualización de datos)
- Ambiente para comparar los algoritmos de aprendizaje.
- Arquitectura modular orientada a objetos.

En la **Figura 13** se observa la aplicación Weka en funcionamiento:

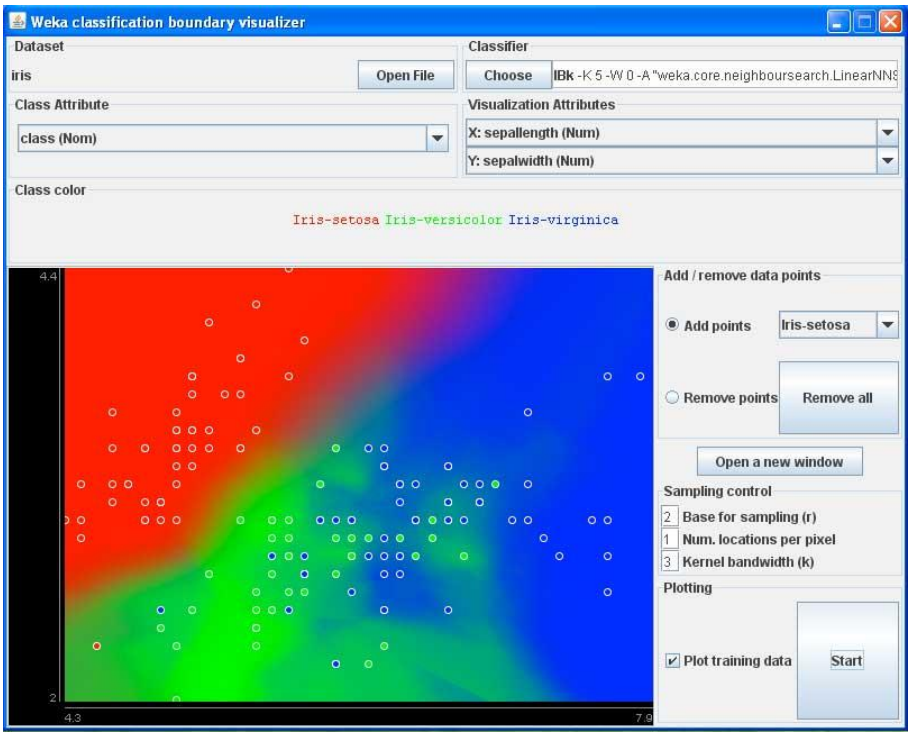


Figura 13 Pentaho Data Mining

2.8 METODOLOGÍA HEFESTO v2

HEFESTO es una metodología propia, cuya propuesta está fundamentada en una muy amplia investigación, comparación de metodologías existentes, experiencias propias en procesos de confección de almacenes de datos.

La metodología HEFESTO puede resumirse como lo muestra la **Figura 14**:

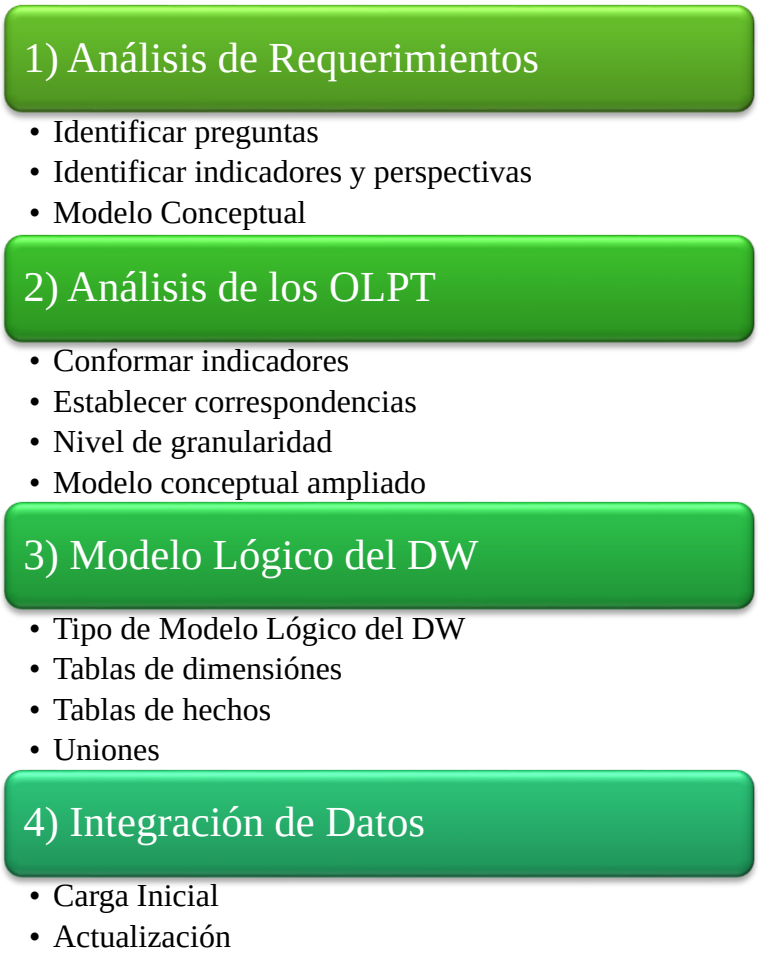


Figura 14 Pasos Metodología Hefesto v2

Fuente: (Bernabeu, 2010)

Como se puede apreciar, se comienza recolectando las necesidades de información de los usuarios y se obtienen las preguntas claves del negocio. Luego, se deben identificar los indicadores resultantes de los interrogativos y sus respectivas perspectivas de análisis, mediante las cuales se construirá el modelo conceptual de datos del DW.

Después, se analizarán los OLTP para determinar cómo se construirán los indicadores, señalar las correspondencias con los datos fuentes y para seleccionar los campos de estudio de cada perspectiva.

Una vez hecho esto, se pasará a la construcción del modelo lógico del depósito, en donde se definirá cuál será el tipo de esquema que se implementará. Seguidamente, se confeccionarán las tablas de dimensiones y las tablas de hechos, para luego efectuar sus respectivas uniones.

Por último, utilizando técnicas de limpieza y calidad de datos, procesos ETL, etc, se definirán políticas y estrategias para la Carga Inicial del DW y su respectiva actualización. (Bernabeu, 2010)

CAPÍTULO 3
METODOLOGÍA

Se han analizado las principales metodologías existentes (ver **Tabla 3**), con lo cual se pretende encontrar un modelo híbrido utilizando las mejores prácticas de cada una para la correcta ejecución del proyecto:

Tabla 3 Metodologías de Diseño de DW

	TOP-DOWN	BOTTOM-UP	HÍBRIDO
Profesional	**Inmon**	**Kimball**	**Ricardo Díaz**
Énfasis	General	Específico	Punto de Equilibrio entre nivel de granularidad y grado de cohesión
Diseño	Modelo normalizado basado en la empresa	El modelo dimensional de datamarts, usa esquema de estrella	Diseño basado en el punto de equilibrio soportado por las reglas del negocio
Arquitectura	Compuesto de varios niveles de áreas de interés y data marts dependientes	Área de interés y data marts	Área del negocio
Data Set	DWH datos a nivel atómico; data marts datos sumarizados	Contiene datos atómicos y sumarizados	Carga data marts con datos atómicos y sumarizados vía ETLs y aporte técnico científico

Para comprender sus diferencias y semejanzas, se explica a continuación sus principales fundamentos y características:

3.1 METODOLOGÍA DE RALPH KIMBALL

La metodología de Kimball (ver **Figura 15**), llamada Modelo Dimensional (Dimensional Modeling), se basa en lo que se denomina Ciclo de Vida Dimensional del Negocio (Business Dimensional Lifecycle). Esta metodología es considerada una de las técnicas más utilizadas cuando se construye un Data Warehouse.

Un almacén de datos (Data Warehouse) es una estructura de datos donde la información contenida esta diseñada para favorecer el análisis y la divulgación eficiente de datos.

En el Modelo Dimensional se constituyen modelos de tablas y relaciones con el propósito de optimizar la toma de decisiones, mediante las consultas hechas en una base de datos relacional que están ligadas con la medición o un conjunto de mediciones de los resultados de los procesos de negocio.

El Modelo Dimensional es una técnica de diseño lógico que tiene como objetivo presentar los datos dentro de un marco de trabajo estándar e intuitivo para permitir su acceso con un alto rendimiento.

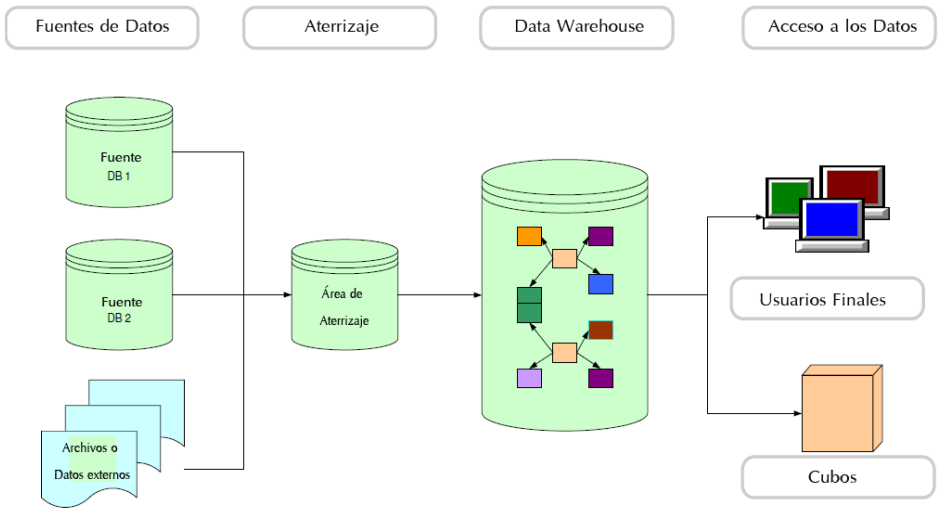


Figura 15 Ciclo de Vida Kimball

Fuente: (Kimbal, 1996)

3.1.1 Modelado dimensional.

Básicamente se comienza con una matriz donde se determina la dimensionalidad de cada indicador y luego se especifican los diferentes grados de detalle (atributos), dentro de cada concepto del negocio (dimensión), así como la granularidad de cada indicador (variable o métrica) y las jerarquías que dan forma al modelo dimensional del negocio.

3.1.2 Diseño físico.

El diseño físico se focaliza sobre la selección de estructuras necesarias para soportar el diseño lógico. Los elementos principales de este proceso son la definición de estándares de nombres específicos del ambiente de la base de datos.

3.1.3 Diseño y desarrollo de presentación de datos.

Las principales sub-etapas de esta zona del ciclo de vida son: la extracción, la transformación y la carga (ETL process). Los procesos de carga de datos sirven para poblar el Data Warehouse.

3.1.4 Diseño de la Arquitectura Técnica

Los ambientes de data warehousing requieren la integración de numerosas tecnologías. Se debe tener en cuenta tres factores: los requerimientos del negocio, los actuales ambientes técnicos y las directrices técnicas estratégicas futuras planificadas para de esta forma poder establecer el diseño de la arquitectura técnica del ambiente de data warehousing.

Esta metodología también se referencia como **Bottom-up** , pues al final el Data warehouse Corporativo no es más que la unión de los diferentes datamarts, que están estructurados de una forma común a través de la estructura de bus. Esta característica le hace más flexible y sencilla de implementar, pues se puede construir un Datamart como

primer elemento del sistema de análisis, y luego ir añadiendo otros que comparten las dimensiones ya definidas o incluyen otras nuevas. En este sistema, los procesos ETL extraen la información de los sistemas operacionales y los procesan igualmente en las áreas de aterrizaje, realizando posteriormente el llenado de cada uno de los Datamart de una forma individual, aunque siempre respetando la estandarización de las dimensiones (dimensiones conformadas).

La metodología para la construcción del Data warehouse incluye las 4 fases que son:

- Selección del proceso de negocio.
- Definición de la granularidad de la información.
- Elección de las dimensiones de análisis.
- Identificación de los hechos o métricas. Tratamiento de los cambios, Dimensiones Lentamente Cambiantes (SCD). (Kimball, 2000)

3.2 METODOLOGÍA DE W. H. INMON

William Inmon ve la necesidad de transferir la información de los diferentes OLTP (Sistemas Transaccionales) de las organizaciones a un lugar centralizado donde los datos puedan ser utilizados para el análisis a la Fábrica de Información Corporativa (CIF o Corporate Information Factory). Insiste además en que ha de tener las siguientes características:

- **Orientado a temas:** Los datos en la base de datos están organizados de manera que todos los elementos de datos relativos al mismo evento u objeto del mundo real queden unidos entre sí.
- **Integrado:** La base de datos contiene los datos de todos los sistemas operacionales de la organización, y dichos datos deben ser consistentes.

- **No volátil**: La información no se modifica ni se elimina, una vez almacenado un dato, éste se convierte en información de sólo lectura, y se mantiene para futuras consultas.

- **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.

La información debe estar a los máximos niveles de detalle. Los Data warehouses departamentales o datamarts son tratados como subconjuntos de este Data warehouse corporativo, son construidos para cubrir las necesidades individuales de análisis de cada departamento, y siempre a partir del Data warehouse central (de este también se pueden construir los ODS (Operational Data Stores) o similares), el enfoque Inmon se muestra en la **Figura 16**.

Figura 16 Enfoque Inmon

Fuente: (Inmon, 2002)

La metodología Inmon también se referencia normalmente como **Top-down** . Los datos son extraídos de los sistemas operacionales por los procesos ETL y cargados en las area stage, donde son validados y consolidados en el DW corporativo, donde además existen los llamados metadatos que documentan de una forma clara precisa el contenido del DW. Una vez realizado este proceso, los procesos de refresco de los Datamart departamentales obtienen la información de este, y con las consiguientes transformaciones,

organizan los datos en las estructuras particulares requeridas por cada uno de ellos, refrescando su contenido.

La metodología para la construcción de un sistema de este tipo es la habitual para construir un sistema de información, utilizando las herramientas habituales (esquema Entidad Relación, DIS (Data Item Sets, etc.). Para el tratamiento de los cambios en los datos, usa la Gestión de las dimensiones continuas y discretas (inserta fechas en los datos para determinar su validez para la dimensión continua o bien mediante el concepto de snapshot o foto para la dimensión discreta).

Al tener este enfoque global, es más difícil de desarrollar en un proyecto sencillo (pues se intentará abordar el "todo", a partir del cual se irá al "detalle").(Inmon, 2002)

3.2.1 Pre-requerimientos

3.2.1.1 Modelo de datos

El punto de partida para el diseño y desarrollo del data warehouse es el modelo de datos.

Sin el modelo de datos, es difícil contemplar la construcción del data warehouse. El modelo de datos actúa como hoja de ruta (roadmap) para el desarrollo.

El modelo de datos del Data Warehouse se compone de al menos dos componentes principales: Un modelo de alto nivel y un modelo de nivel medio. La **Figura 17** representa gráficamente el modelo de datos del Data Warehouse:

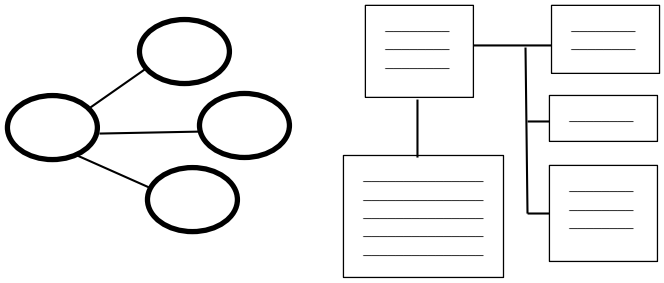


Figura 17 El modelo de datos está realizado
Fuente: (Inmon, 2002)

3.2.1.2 Selección de tecnología

La selección de la tecnología de almacenamiento de datos, tanto de hardware como de software depende de muchos factores, tales como:

- El volumen de datos que pueden alojarse.
- La velocidad con que se necesitan datos.
- La historia de la organización.
- Que nivel de datos se construirá
- Cuántos usuarios habrá
- Qué tipo de análisis se va a realizar,
- El coste de la tecnología

3.2.1.3 Dimensionando el data warehouse

Otro requisito previo para la primera iteración del diseño y población del Data Warehouse es el dimensionamiento del almacén de datos, como se representa en la **Figura 18**:

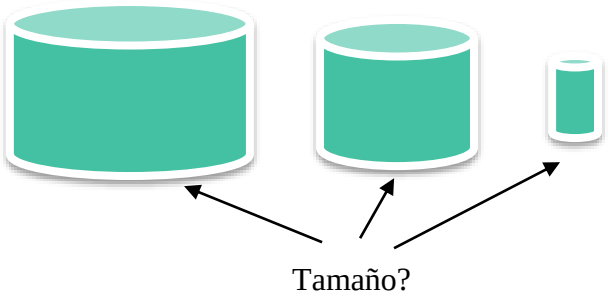


Figura 18 La dimensión aproximada del Data Warehouse es calculada
Fuente: (Inmon, 2002)

Aquí se muestra un dimensionamiento aproximado de los datos para determinar la aptitud de las plataformas de hardware y software. Si las plataformas de hardware y software son demasiado grandes o demasiado pequeñas para la cantidad de datos que residen en el data warehouse, entonces no debe producirse el desarrollo iterativo hasta que el ajuste este realizado correctamente.

3.2.1.4 Recolección de requerimientos de información

Los requerimientos de información de la organización deben ser recogidos por medio de un "time box". Un "time box"es una limitación de tiempo - desde una semana hasta seis meses – en la que todos los requerimientos de información conocidos y obvios se reunen. Una vez que ocurre la fecha límite, los requerimientos de información se introducen en el modelo de datos del data warehouse.

La **Figura 19** muestra los medios típicos por el cual dichos requerimientos informativos se identifican y se reúnen.

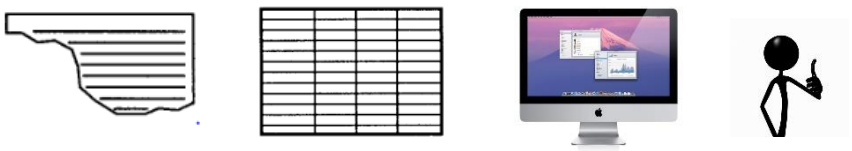


Figura 19 Análisis existentes de SSD son reunidos
Fuente: (Inmon, 2002)

3.2.2 Primera iteración

3.2.2.1 ¿Cuántos datos se van a cargar?

El primer problema del diseño y planificación para la primera iteración del data warehouse a ser desarrollado es exactamente que cantidad y variedad de datos se van a cargar. Es muy poco probable el caso en el que grandes cantidades de datos se carguen como un resultado de la primera iteración. Una pauta general para la carga de la primera iteración es:

La primera iteración DEBE contener datos que sean lo suficientemente grandes para que tenga sentido y lo suficientemente pequeños para ser rápidos y factibles.

Existen varias maneras diferentes de reducir el tamaño de los datos sin perder su eficacia. La **Figura 20** ilustra algunas de esas formas:

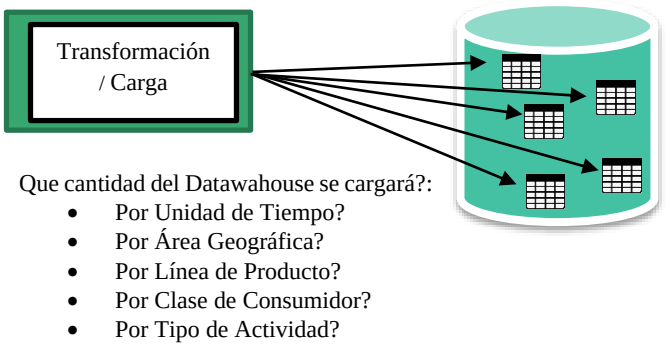


Figura 20 Tipos de Reducción de Tamaño de Datos
Fuente: (Inmon, 2002)

El arquitecto de datos debe tener cuidado de no poner demasiada información en la primera iteración de desarrollo. Por otro lado, el arquitecto de datos debe tener cuidado de no incluir tan poca información para que la espontaneidad de descubrimiento por el analista SSD no esté limitada.

3.2.2.2 Pescando en el estanque correcto

Existen muchas áreas funcionales en las que el procesamiento de información puede dar fruto. Pero hay algunas áreas funcionales clásicas que con los años han dado más frutos que otras. La **Figura 21** representa las arenas funcionales:

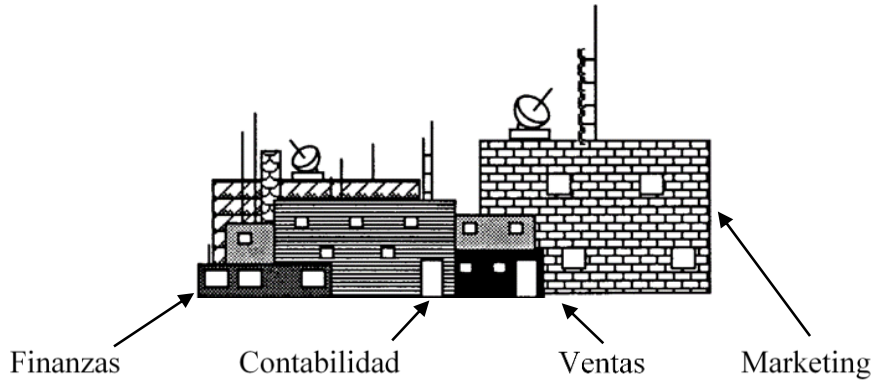


Figura 21 Elección de área funcional "Primera Iteración" del Data Warehouse
Fuente: (Inmon, 2002)

3.2.2.3 Seleccionar un área temática

Otra decisión importante que se necesita tomar para la construcción de la primera iteración del entorno del data warehouse es la selección de la primera área a ser implementada. La **Figura 22** muestra esta selección.

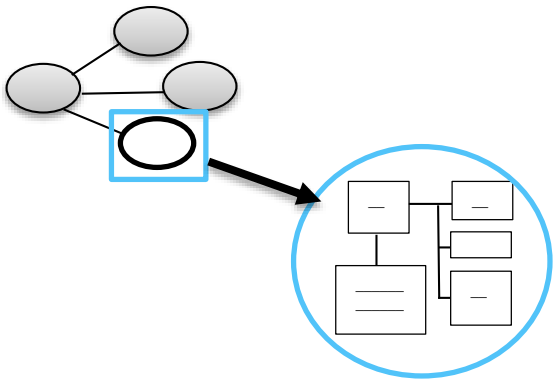


Figura 22 Selección del Área Temática
Fuente: (Inmon, 2002)

Las diferentes opciones que se aplican a la construcción de la primera iteración del data warehouse son:

- Aplicar una sola área temática,
- Implementar un subconjunto de una sola área temática, o
- Implementar un subconjunto de múltiples áreas temáticas.

Después de este punto se describirán los pasos de diseño para el almacenamiento de datos. Hay que aclarar que estos pasos se realizan para muy pequeñas cantidades de datos - no hay nada de barrido en el diseño que está implícito aquí. Normalmente hay de cinco a quince tablas físicas que resultarán del trabajo de diseño descrito. Si hay más, entonces todo el objetivo de hacer el diseño y el desarrollo iterativo se ha perdido.

3.2.2.4 Añadiendo Atributos Físicos

Una vez que se ha seleccionado el área temática, la siguiente tarea es ir desde el modelo de datos del data warehouse al diseño de base de datos físico. Los pasos a seguir son:

Primer paso: Especificación de los atributos físicos del modelo de datos. Hay que tomar en cosideración que:

- Los usuarios finales comprendan los datos.
- El formato de los datos, ya que reside en el entorno de los sistemas heredados.

- Los diferentes formatos de sistemas heredados que deben consolidarse.
- El volumen de datos que se va a crear, y así sucesivamente.

Segundo paso: Identificación de la unidad de tiempo que reside en cada unidad de datos del data warehouse. En la mayoría de los casos, la unidad de tiempo año, trimestre, mes, día, etc. se adjunta a la clave de la tabla. En algunos casos, la unidad de tiempo que identifica a cada ocurrencia de datos del data warehouse se identifica implícitamente con el data warehouse.

Tercer paso: Definir el sistema de registro. El sistema de registro es la fuente de datos que se encuentra en el entorno operativo, los sistemas heredados. El sistema de registro son los datos: más completos, más precisos, más oportunos, tiene la mejor conformidad estructural al modelo de datos, y es más cercano al punto de entrada operacional. Una vez que se han especificado las transformaciones:

- Se transfieren a la infraestructura de metadatos que se encuentra "encima" del data warehouse.
- Las transformaciones se converten en código. Tener un sistema automático generador de código como el Administrador de PRISM convierte esta tarea en un ejercicio automático y eficiente.
- Simultáneamente con la generación de código se encuentra la tarea de asignación de espacio para el almacenamiento de datos donde van a residir.
- Una vez que se ha realizado la asignación de espacio, los programas que se han creado se pueden poner en ejecución y los datos automáticamente en el data warehouse.
- Una vez que ocurre la población del data warehouse, el usuario final o el analista DSS es expuesto a los datos.

En este punto la primera iteración de desarrollo del data warehouse finaliza. Si bien se han realizado muchas actividades, ha existido relativamente pocos

datos y algunos tipos de datos para realizar las actividades del negocio. Se entiende que los errores en el diseño ocurrirán y que esos errores se corregirán en la segunda iteración (y más) en el diseño y el desarrollo del data warehouse.

3.2.3 Segunda iteración

Las semillas de la segunda iteración de desarrollo del data warehouse se sembraron en la finalización de la primera fase de desarrollo. Un aspecto importante y esencial del ciclo de vida de desarrollo es la supervisión del usuario final para la retroalimentación.

El arquitecto de datos escucha muy atentamente al analista DSS para determinar dónde se pueden hacer mejoras en el data warehouse. Mientras que la conversación entre el arquitecto de datos y el analista DSS es un proceso continuo, la conversación que se produce mientras el data warehouse es poblado la primera vez es, probablemente, la más importante. Es en los primeros usos del data warehouse que las recomendaciones y sugerencias más valiosas ocurren normalmente. (Inmon W., 2000)

3.3 Kimball vs Inmon

Desde un punto de vista “Arquitectónico”, la mayor diferencia entre ambas metodologías es el sentido de la construcción del Data Warehouse:

Ralph Kimball - Arquitectura Bottom-Up: Enfocado en que los Data Marts se crean primero para proporcionar información y capacidad de análisis para los procesos de negocio (ver **Figura 23**).

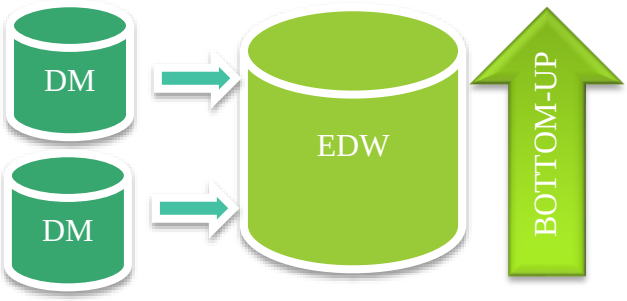


Figura 23Bottom Up
Fuente: (Drazda, 2013)

Bill Inmon – Arquitectura Top-Down: Conceptualiza el Data Warehouse como un repositorio centralizado para toda la empresa. Diseño del Datawarehouse usando el modelo de datos de la empresa normalizado para, posteriormente, crear los Data Marts. (ver **Figura 24**). (Drazda, 2013)

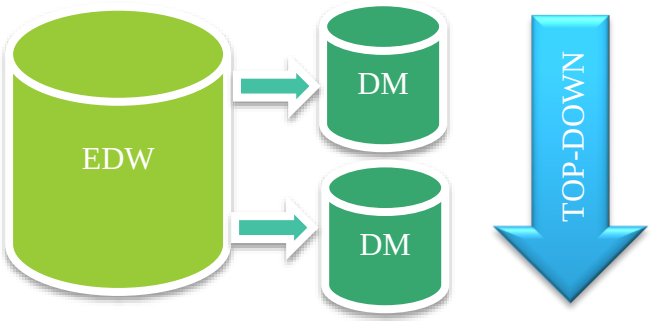


Figura 24 Top Down
Fuente: (Drazda, 2013)

Desde un punto de vista conceptual, las principales diferencias entre ambas metodologías estan representadas en la **Tabla 4**:

Tabla 4 Diferencias entre Kimball e Inmon

KIMBALL	INMON
Bodega de Datos Lógico (BUS) creado para aplicarse a escenarios empresariales 'Data Mart' (Ej.: Ventas, Finanzas, etc)	Modelo de Datos Empresariales (CIF) que es un Almacén (Bodega) de Datos Empresarial (EDW – Enterprise Data Warehouse)
Aplicado empresarialmente con el fin de tener una máxima participación de todos los usuarios	Aplicado a usuarios IT con el fin de tener una mínima participación de dichos usuarios
Data Mart descentralizado (No requiere ser separado físicamente del Almacén de Datos)	Centralización Atómica de Tablas Normalizadas (Fuera del límite de uso de usuarios finales)
Optimizado para escenarios de Reportes Analíticos y Análisis de Datos OLAP independientemente de la dimensionalidad del Data Mart	Posteriormente crea dependencia de los Data Marts que se encuentran separados físicamente de los subconjuntos de datos y pueden ser usados por múltiples propósitos
Integrado vía 'Dimensiones Conformadas' (Provee consistencia de orígenes de datos cruzados)	Integrado vía 'Modelo de Datos Empresariales'
2 capas (Data Mart, Cubos OLAP), Menos ETLs, Duplicación de Datos nula	3 capas (Data Warehouse, Data Mart, Cubos OLAP), Duplicación de Datos

Es importante reconocer que la **arquitectura** del data warehouse identifica los componentes, sus características, y las relaciones entre las partes, mientras que la **metodología** identifica las actividades que tienen que realizarse y su secuencia. Con demasiada frecuencia, los términos de **arquitectura** y **metodología** se utilizan indistintamente, lo que crea confusión.

La arquitectura es el producto final, mientras que una metodología es el proceso para la elaboración de un producto final. Pero, aunque que la arquitectura y la metodología son diferentes, deben ser compatibles. Es importante utilizar una metodología que sea consistente con la arquitectura que se está aplicando.

A veces la arquitectura "**hub and spoke**" (CIF o "Corporate Information Factory") se conoce como un enfoque de **top down** y la arquitectura de "**bus**" como **bottom up**.

La razón de esto es que la aquitectura **hub and spoke** hace gran énfasis en poner inicialmente la infraestructura y procesos en su lugar para crear un data warehouse empresarial y la arquitectura de **bus** se centra en la entrega de una solución que responde a una necesidad de negocio actual. Estas son metodologías en lugar de arquitecturas porque describen procesos de desarrollo.

Con el tiempo, tanto los enfoques **top down** y **bottom up** se han vuelto cada vez más similares. Los defensores del enfoque **top down** están de acuerdo en la importancia de desarrollar de forma incremental y entrega temprana. Los defensores del enfoque **bottom up** reconocen la importancia de contar con un plan empresarial para la integración incremental de los data marts desarrollados. Como resultado, **las dos metodologías no son tan diferentes como mucha gente cree**. (Watson, 2005)

3.4 ¿Qué metología utilizar?

Los Modelos no son diferentes, ya que llegan a ser similares con el pasar de los años en un ecosistema que terminan complementándose mutuamente.

Inmon reduce la creación de DW normalizados antes de crear un Data Mart dimensional, y Kimball pasa por alto la normalización de un DW.

Se puede optimizar cada modelo, mostrándose cada uno similar al otro (Por supuesto, agregando un EDW normalizado bajo el modelo Kimball o estructurándo dimensionadamente los Data Mart como Inmon).

Vale la pena recalcar que, comprendiendo ambos enfoques y seleccionando partes de ambas metodologías para cubrir las necesidades que nuestro escenario requiere, se logrará exitosamente la decisión a tomar, no se necesita establecer un solo enfoque. (Redondo, 2010)

La metodología a utilizar será híbrida, tomando en cuenta un diseño basado en el punto de equilibrio entre el nivel de granularidad y grado de cohesión, soportado por las reglas del negocio, como se encuentra representado en **Figura 25**:

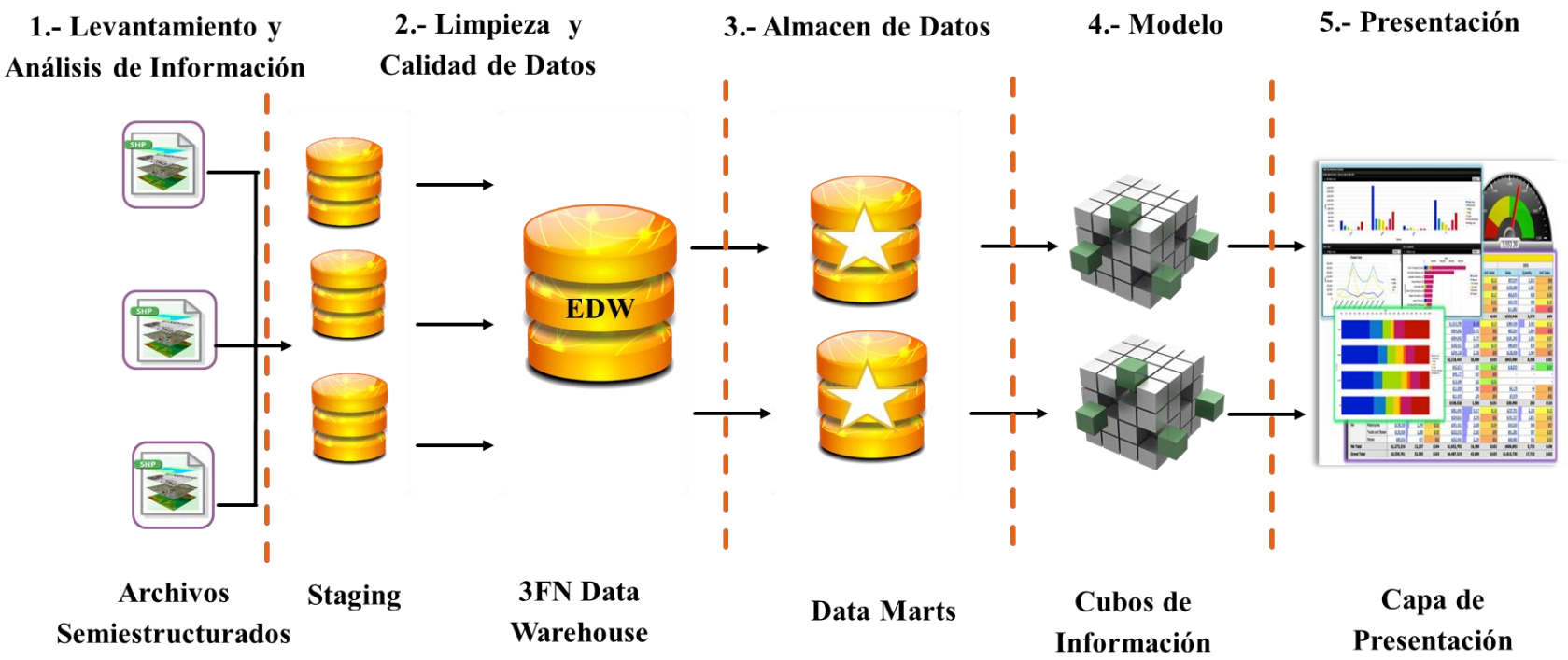


Figura 25 Metodología a Utilizar

CAPÍTULO 4
DESARROLLO DEL PROYECTO

4.1 PLANEACIÓN Y ADMINISTRACIÓN DEL PROYECTO

4.1.1 Antecedentes y Justificación del Proyecto

El SIPAS es un elemento constitutivo del SINARI desarrollado por el MAE a través del PRAS, dirección SINPAS, que tiene la finalidad de mostrar indicadores ambientales los cuales sirven de insumo para la toma de decisiones respecto a las afectaciones que comprometen al ambiente y personas por causa de los pasivos ambientales. En este sentido, el SIPAS HN tiene, entre sus principales objetivos, generar información para la toma de decisiones y contribuir a la reducción de pasivos socio-ambientales, por esta razón se propone la implantación de la suite de Inteligencia de negocios Pentaho, para la elaboración rápida de reportes y tableros de control con el fin de reducir el esfuerzo de las coordinaciones en el análisis de la información que permitan al PRAS disponer de información fiable en el momento requerido.

4.1.2 Planificación del Proyecto

Es necesario contar con personal de la institución involucrado para la correcta ejecución del proyecto, es por esta razón que se ha asignado a los siguientes responsables (ver **Tabla 5**):

Tabla 5 Planificación del Proyecto

PERSONAL	RESPONSABLE
Patrocinador del proyecto	Programa de Reparación Ambiental y Social – PRAS
Administrador de base de datos	Ricardo Díaz (tesista)
Director del proyecto	Ing. Juan Carlos Dueñas
Desarrollador, diseñador, arquitecto de datos	Ricardo Díaz (tesista)
Personal involucrado en el proyecto	Empleados de la dirección del SINPAS

4.1.3 Administración del Proyecto

El proyecto se ha realizado en las instalaciones del PRAS, por esta razón, se contó con el total apoyo, monitoreo y gestión del personal de la institución.

4.1.4 Situación Actual

Para el análisis de la información hidrocarburífera, el personal encargado realiza todo el proceso manualmente desde la generación de los archivos semiestructurados, depuración y cruce de información hasta la creación de tablas dinámicas para, posteriormente, generar los reportes requeridos por las autoridades de la institución, clientes y público en general.

Un informe puede tardar desde una a varias semanas en ser entregado o publicado debido a la asignación del personal no solamente a una tarea específica, sino a un conjunto de tareas de acuerdo al cargo que posea.

4.1.5 Fuentes de Datos

Los datos se encuentran en archivos semiestructurados (Excel y Shapefiles) correctamente georeferenciados por el departamento GIS de la institución.

La fuentes de datos a utilizar se encuentran detalladas en la **Tabla 6**:

Tabla 6 Fuentes de Datos

FUENTE DE DATOS	ELEMENTOS	Nº REGISTROS
Infraestructura de la actividad hidrocarburífera	Estaciones	180
	Pozos	5.223
	Plataformas	1.079
Conflictos	Conflictos	81
Convenios	Convenios	357
Reclamos	Reclamos	1.013

4.2 FASE 1: LEVANTAMIENTO Y ANÁLISIS DE INFORMACIÓN

Durante esta fase se identificó los procesos de negocio que se desea modelar:

- Seguimiento, monitoreo y evaluación de la Infraestructura Hidrocarburífera.
- Ejecución de acciones previas y/o complementarias para la Gestión Social de la Reparación Integral.

4.2.1 Requerimientos del proceso de negocio

Los requerimientos identificados para los procesos de negocio se detallan a continuación:

- Ubicar geográficamente las infraestructuras hidrocarburíferas para conocer en qué localidad del país se encuentran.
- Ver en qué área protegida se encuentra cada infraestructura hidrocarburífera.
- Ver a que bloque petrolero (Catastro) y campo petrolero pertenece cada infraestructura hidrocarburífera.
- Determinar las infraestructuras que existen dentro de los diferentes territorios indígenas que hay en el país.
- Identificar en qué localidad se encuentra la mayor parte de infraestructuras estatales y no estatales.
- Identificar los tipos de estaciones que existen actualmente.
- Identificar los estados de los pozos que existen actualmente.
- Determinar la ubicación geográfica de cada uno de los eventos sociales (Conflictos, Convenios y Reclamos).
- Identificar cada uno de los eventos sociales relacionados con las actividades hidrocarburíferas por año, mes y día.
- Identificar cada uno de los eventos sociales relacionados con las actividades hidrocarburíferas por área protegida, campo petrolero y bloque petrolero.

- Identificar los actores involucrados, agravantes y tipos de acción generados por los conflictos.
- Identificar los ámbitos de retribución de los convenios y sus tipos de documentos firmados así como los tipos de causas y beneficiarios.
- Identificar el tipo de afectación de los reclamos así como sus actores y las figuras de reclamo.

4.2.2 Definir Indicadores

Se han definido los indicadores (ver **Tabla 7**) en base a dos módulos requeridos:

Gestión Social: Comprende los Conflictos y Reclamos ejecutados por el pueblo ecuatoriano ante la contaminación o desgaste del patrimonio natural por la ejecución de actividades hidrocarburíferas y los Convenios realizados con el estado.

Infraestructura: Comprende el detalle de Pozos, Estaciones y Plataformas que se encuentran en el país.

Tabla 7 Tabla de Indicadores

MÓDULO	REQUERIMIENTO	INDICADOR
Gestión Social	Conflictos	Número de conflictos por fecha
		Número de conflictos según el tipo de acciones
		Número de conflictos según los actores involucrados en la acción
		Número de conflictos según agravantes de la acción
		Número de conflictos por localidad
	Convenios	Número de convenios por fecha
		Número de convenios por tipo de documento
		Número de convenios por tipo de causa
		Número de convenios por beneficiarios
		Número de convenios por ámbito de la retribución
		Número de convenios por localidad
	Reclamos	Número de reclamos por fecha

CONTINÚA ➔

		Número de reclamos por tipo de figura
		Número de actores que promueven el reclamo
		Número de reclamos por tipo de afectación denunciada
		Número de reclamos por localidad
Hidrocarburífero Nacional	Infraestructura	Número de Estaciones
		Número de Pozos
		Número de Plataformas

4.2.3 Dimensiones de alto nivel que son comunes en diversos procesos

El depósito de datos proporcionó información coherente para las consultas que eran requeridas por la institución. Para mantener la coherencia, un método fue crear tablas de dimensiones compartidas que utilice la aplicación.

Generalmente, cada proceso de negocio tiene su propio esquema que contiene una tabla de hechos, varias tablas de dimensiones compartidas y tablas de dimensiones exclusivas de la función de negocio específica como se establecerá en la siguiente fase.

4.2.4 Identificar el Grano (Nivel de Granularidad)

Identificando el grano se pudo especificar que contiene un registro de una tabla de hechos exactamente. El grano muestra el nivel de detalle asociado a las medidas de las tablas de hechos. Fue necesario establecer el nivel de detalle que se desea estará disponible en el modelo dimensional, se determinó la granularidad para cada módulo en la **Tabla 8**:

Tabla 8 Identificación de Granularidad

MÓDULO	GRANULARIDAD
Infraestructura	Número de infraestructuras por nombre, sector censal, micro cuenca, área protegida, bloque petrolero, campo petrolero, territorio indígena, estado pozo, tipo estación, tipo estatal y fecha.
Gestión Social	Número de eventos por detalle, sector censal, micro cuenca, área protegida, bloque petrolero, campo petrolero, territorio indígena, tipo, actor beneficiario, tipo documento, ámbito agravante figura y fecha.

4.2.5 Identificar Grado de Cohesión

Para la identificación del grado de cohesión del Data Warehouse se utilizará el método "Árbol de decisión".

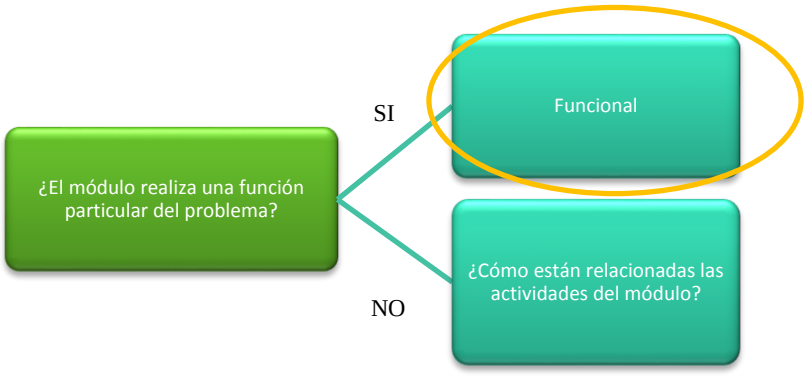


Figura 26 Árbol de Decisión Cohesión Data Warehouse

Las dimensiones compartidas : área protegida, bloque petrolero, campo petrolero, cuenca hidrográfica, localidad, territorio indígena y tiempo proporcionan funciones particulares para ambos Data Marts (Infraestructura y Gestión Social), y las demás dimensiones de ambos módulos cumplen también funciones particulares para cada uno de los Data Marts, lo cual permite una cohesión **funcional** del problema.

4.2.6 Identificar el Punto de Equilibrio entre Nivel de Granularidad y Grado de Cohesión

Una vez identificado en nivel de granularidad y el grado de cohesión, se procede a establecer un punto de equilibrio entre ambos; para lo cual se analizará dimensión por dimensión (ver **Tabla 9**):

Tabla 9 Punto de Equilibrio

INFRAESTRUCTURA	GESTIÓN SOCIAL	PUNTO DE EQUILIBRIO
Nombre	Detalle	
Sector Censal	Sector Censal	X
Micro Cuenca	Micro Cuenca	X
Área Protegida	Área Protegida	X
Bloque Petrolero	Bloque Petrolero	X
Campo Petrolero	Campo Petrolero	X
Territorio Indígena	Territorio Indígena	X
Estado Pozo	Tipo	
Tipo Estación	Actor Beneficiario	
Tipo Estatal	Tipo Documento	
Fecha	Fecha	X
	Ámbito Agravante Figura	

4.2.6.1 Modelo de punto de equilibrio

Partiendo de la **Tabla 9** se realizará un modelo de alto nivel, en el cual se coloca en el centro de la figura las dimensiones compartidas, en la parte superior las dimensiones únicas para el módulo de Infraestructura y en la parte inferior las dimensiones únicas para el módulo de Gestión Social, tal como se muestra en la **Figura 27** .

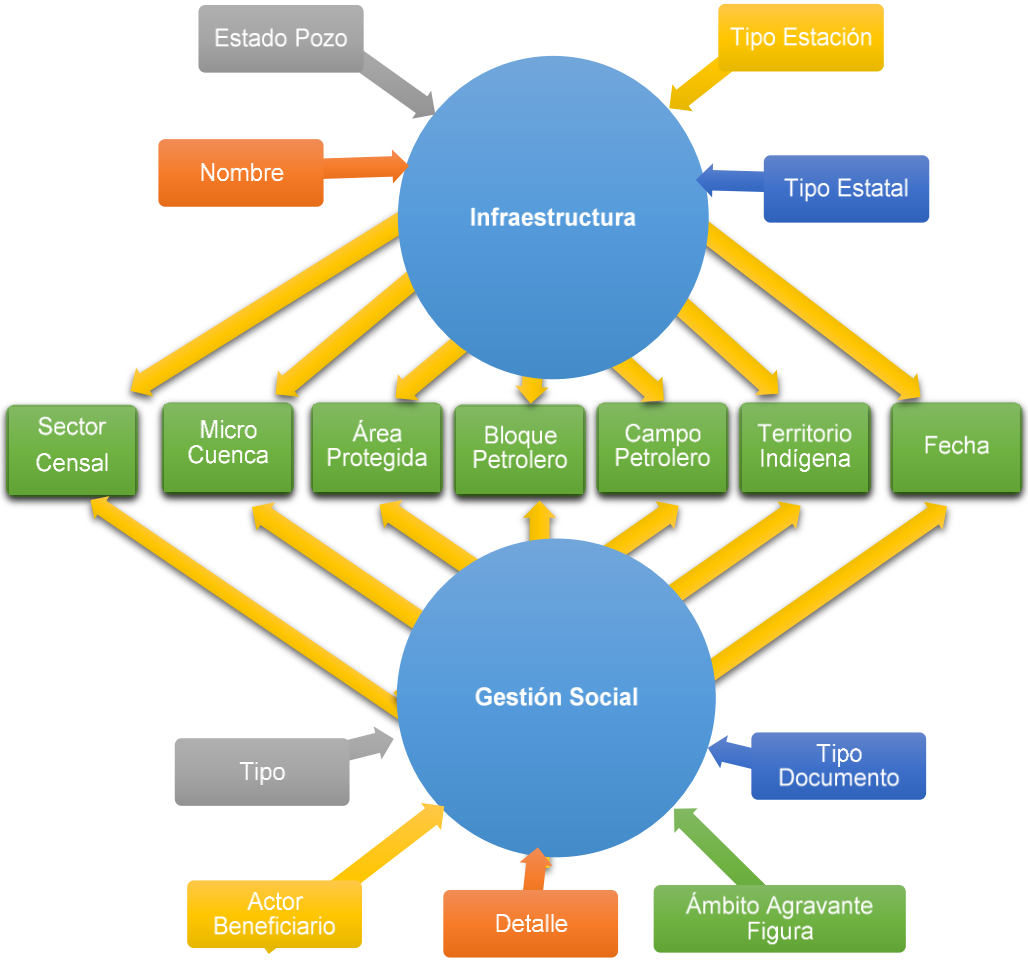


Figura 27 Modelo Punto de Equilibrio

4.2.7 Identificar las dimensiones y medidas

Después de identificar el punto de equilibrio entre nivel de granularidad y grado de cohesión, se estableció las tablas de dimensiones y medidas para el módulo de Infraestructura (ver **Tabla 10**) y para el módulo de Gestion Social (ver **Tabla 11**).

Tabla 10 Tabla de Hechos Infraestructura

TABLA DE HECHOS: INFRAESTRUCTURA		
	Nombre Dimensión	**Jerarquías**
Dimensiones	Área Protegida	Área Protegida
	Bloque Petrolero	Bloque Petrolero
	Campo Petrolero	Campo Petrolero
	Cuenca Hidrográfica	Cuenca SubCuenca Microcuenca
	Estado Pozo	Estado Pozo
	Estatal	Tipo Estatal
	Infraestructura	Tipo Nombre
	Localidad	Provincia Cantón Parroquia
	Territorio Indígena	Territorio Indígena
	Tiempo	Año Mes Día
	Tipo Estación	Tipo Estación
Medidas	Número de Infraestructuras	

Tabla 11 Tabla de Hechos Gestión Social

TABLA DE HECHOS: GESTIÓN SOCIAL		
	Nombre Dimensión	**Jerarquías**
Dimensiones	Área Protegida	Área Protegida
	Bloque Petrolero	Bloque Petrolero
	Campo Petrolero	Campo Petrolero
	Cuenca Hidrográfica	Cuenca SubCuenca Microcuenca
	Gestión Social	Tipo Evento

CONTINÚA ➔

		Detalle
	Localidad	Provincia Cantón Parroquia
	Territorio Indígena	Territorio Indígena
	Tiempo	Año Mes Día
	Actor Beneficiario	Actor Beneficiario
	Ámbito Agravante Figura	Ámbito Agravante Figura
	Tipo	Tipo
	Tipo Documento	Tipo Documento
Medidas	Número de Eventos	

Las dimensiones que comparten las tablas de hechos son:

- Área protegida
- Bloque petrolero
- Campo petrolero
- Cuenca Hidrográfica
- Localidad
- Territorio indígena
- Tiempo

4.2.8 Identificar tableros de control

Se han identificado tres tableros de control, para los cuales se ha establecido los siguientes diseños (ver **Figura 28, 29, 30**):

4.2.8.1 TABLERO DE CONTROL ESTACIONES

*1.1.2 Número de estaciones
*BDD estaciones
*Gráfico de pastel: % de operación estatal y no estatal [const_estatal_noestatal (BDD estaciones; sumatoria de estatales y no estatales]

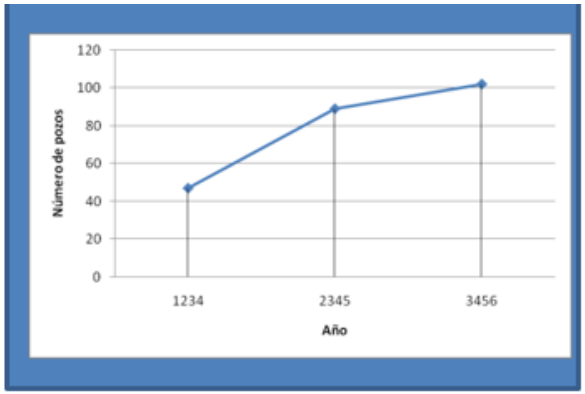
Campo petrolero	Tipo de estación
A	explotación
B	productora
C	reductora
D	etc...

*1.1.2 Número de estaciones
*BDD estaciones
*Tabulado: Tipo de estación por campo petrolero [tipo_estacion (BDD estaciones; campo (BDD estaciones)]

Figura 28 Diseño Tablero de Control Estaciones

4.2.8.2 TABLERO DE CONTROL POZOS

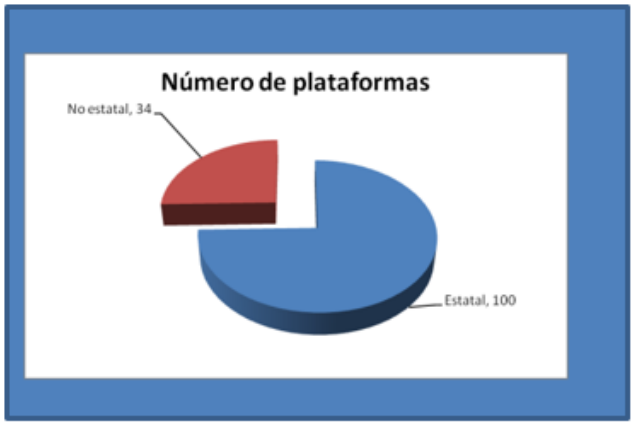
*1.1.3 Número de pozos
*BDD pozos
*Gráfico de pastel: Número de pozos por operadora estatal y no estatal [perf_estatal_noestatal (BDD pozos)]



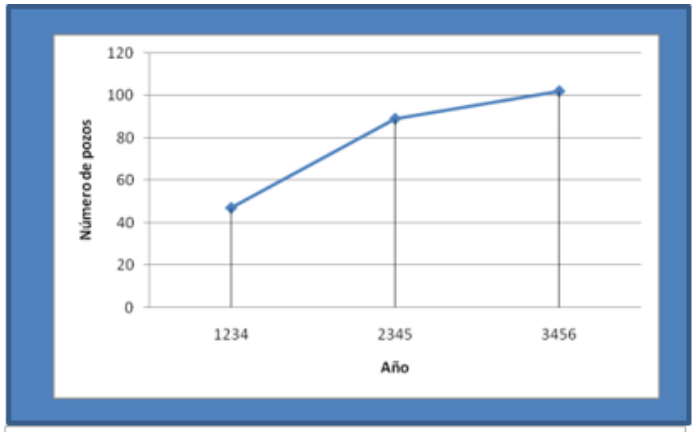
*1.1.3 Número de pozos
*BDD pozos
*Gráfico de tendencia: Número de pozos perforados por año [anio_perforacion (BDD pozos)]

Figura 29 Diseño Tablero de Control Pozos

4.2.8.3 TABLERO DE CONTROL PLATAFORMAS



*1.1.5 Número de plataformas
*BDD plataformas
*Gráfico de pastel: Número de plataformas por operadora estatal y no estatal [const_estatal_noestatal (BDD plataformas)]



*1.1.5 Número de plataformas
*BDD plataformas
*Gráfico de tendencia: Número de construidas por año [anio_construccion (BDD plataformas)]

Figura 30 Diseño de Tablero de Control Plataformas

4.2.9 Resúmen de levantamiento de información

Tabla 12 Resúmen de Levantamiento de Información

SOLUCIONES	TOTAL
Total indicadores	19
Total tableros de control	3

4.2.10 Análisis de Drill Down y Drill Up

4.2.10.1 Infraestructura

De acuerdo al análisis de jerarquías de cada dimensión, se ha establecido el nivel máximo de detalle (Drill Down) del esquema "Insfraestructura"en la **Figura 31**, así como el máximo nivel de agrupación (Drill up).

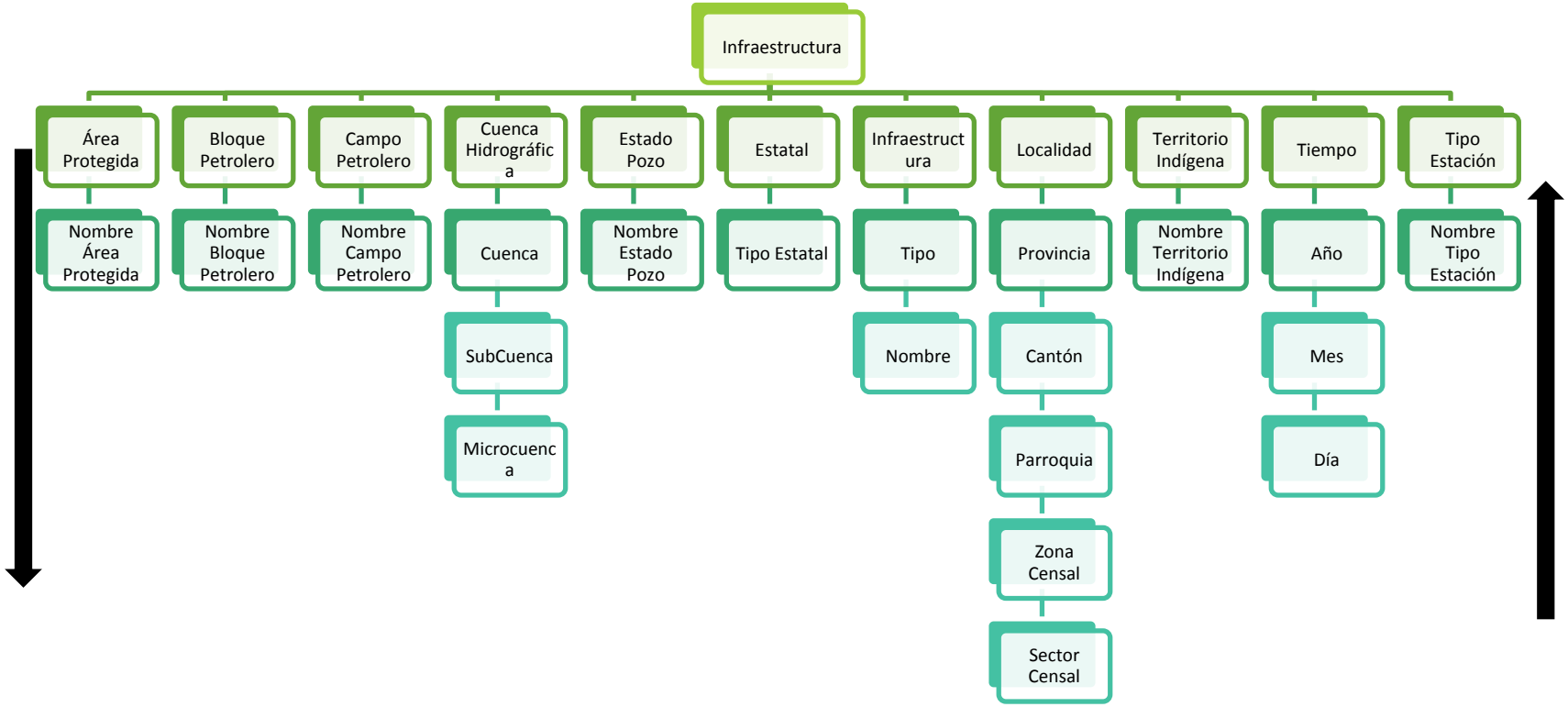


Figura 31 Drill Down y Drill Up Infraestructura

4.2.10.2 Gestión Social

De acuerdo al análisis de jerarquías de cada dimensión, se ha establecido el nivel máximo de detalle (Drill Down) del esquema "Gestión Social"en la **Figura 32**, así como el máximo nivel de agrupación (Drill up).

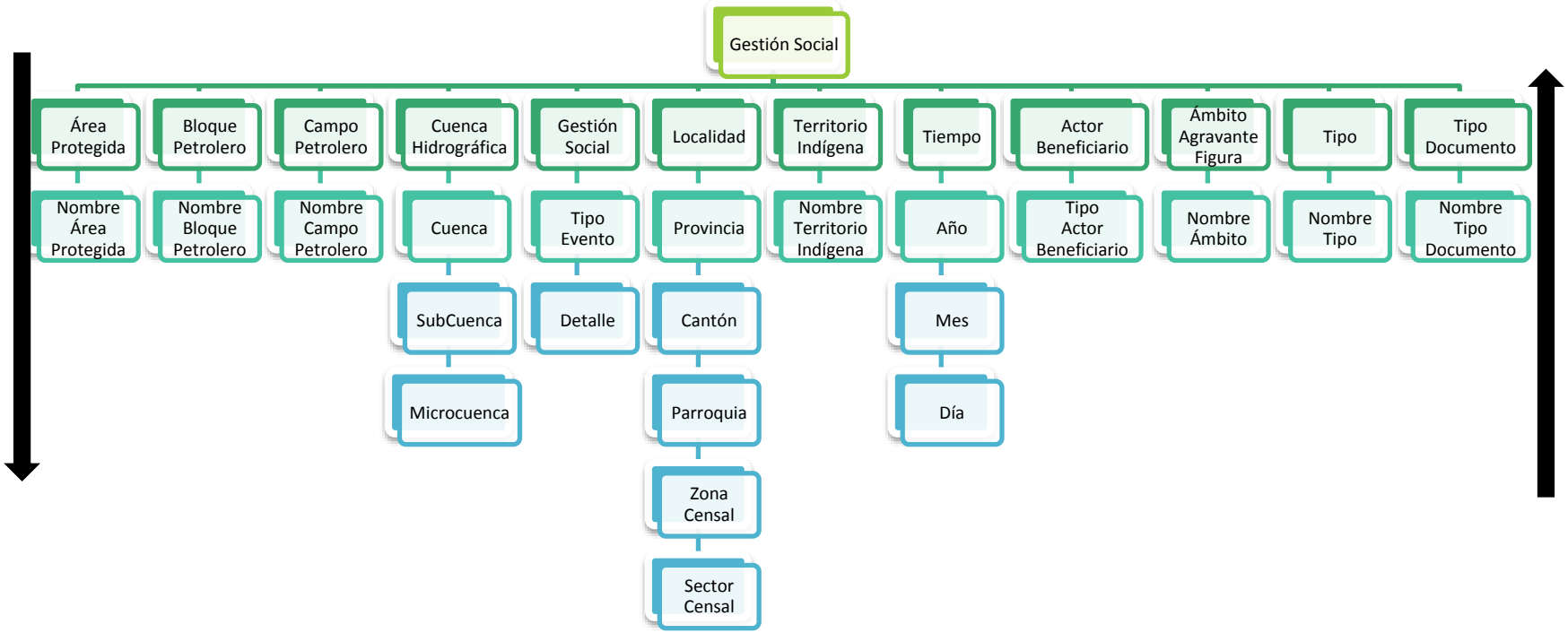


Figura 32 Drill Down y Drill Up Gestión Social

4.2.11 Identificación de Drill Across

De acuerdo al análisis previo, se ha determinado las dimensiones en común entre ambas fact_tables (tablas de hechos), lo cual ha permitido generar el drill across entre ambos esquemas considerando sus medidas (ver **Figura 33**):

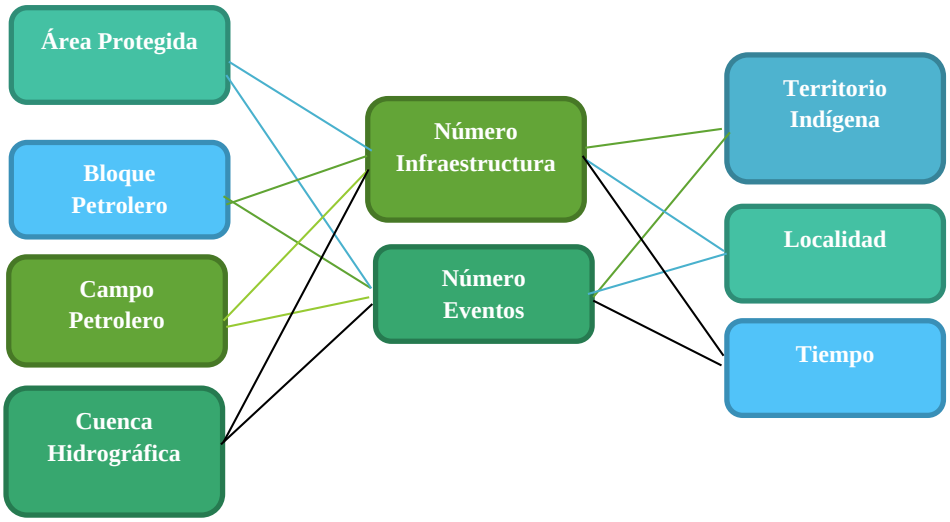
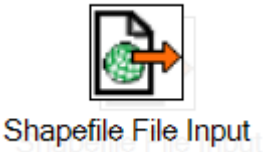


Figura 33 Drill Across

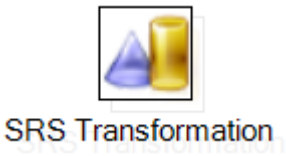
4.3 FASE 2: LIMPIEZA Y CALIDAD DE DATOS

4.3.1 Tablas de Staging

Para extraer la información desde Shapefiles (Archivo Georeferenciados), se ha utilizado el step "Shapefile File Input", el cual permite leer todos los datos desde el archivo ".shp" y poder transformarlos en información útil para nuestro Data Warehouse.



Los archivos Georeferenciados proporcionados por el PRAS tienen coordenadas geográficas en el sistema espacial “WSG 84: EPSG:3260”, pero Pentaho permite georeferenciar en OpenStreet Maps o Google Maps en el sistema espacial “WGS 84 EPSG:4326”, para lo cual se utilizó el step “SRS Transformation”.



4.3.2 Módulo Gestión Social

4.3.2.1 Conflictos

ETL Población Tabla stg_conflictos

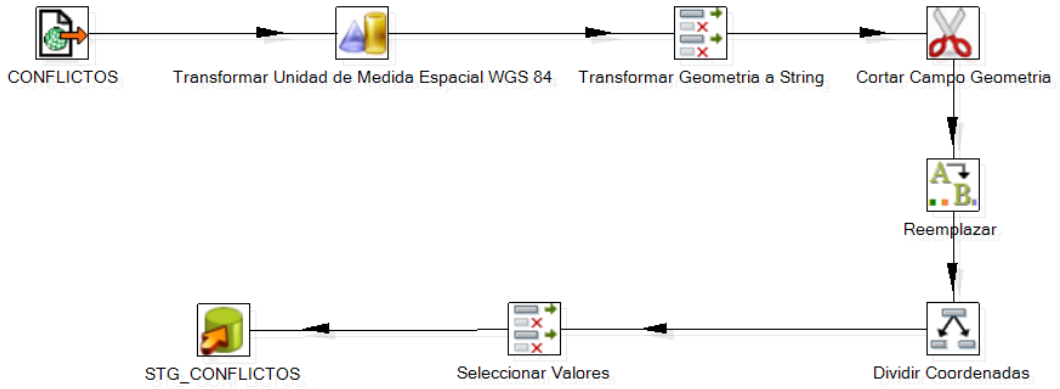


Figura 34 ETL stg_conflictos

4.3.2.2 Convenios

ETL Población Tabla stg_convenios

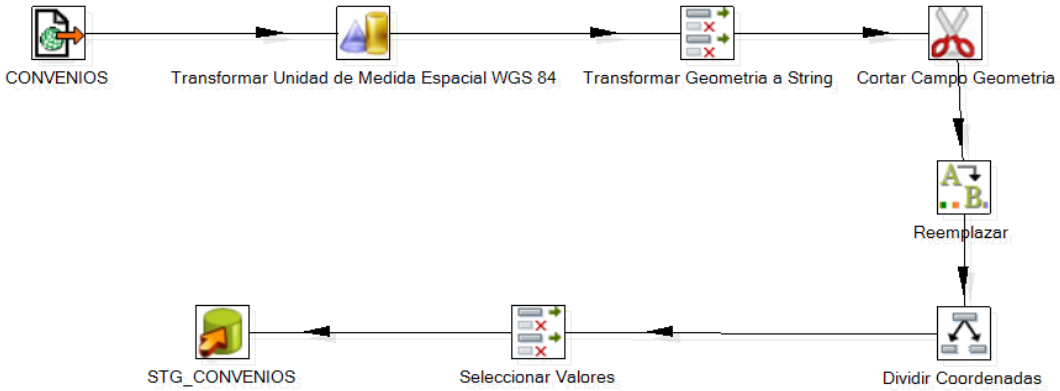


Figura 35 ETL stg_convenios

4.3.2.3 Reclamos

ETL Población Tabla stg_reclamos

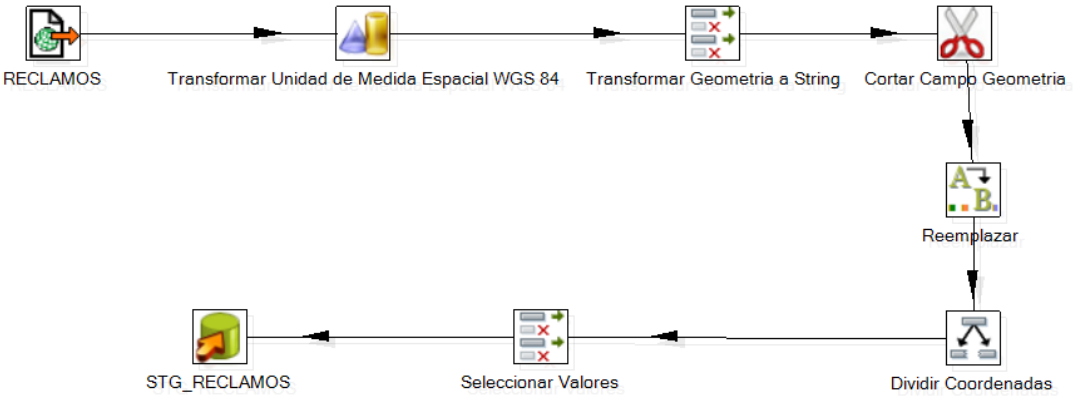


Figura 36 ETL stg_reclamos

4.3.2.1 Gestión Social

ETLs Población Tabla stg_gestion_social

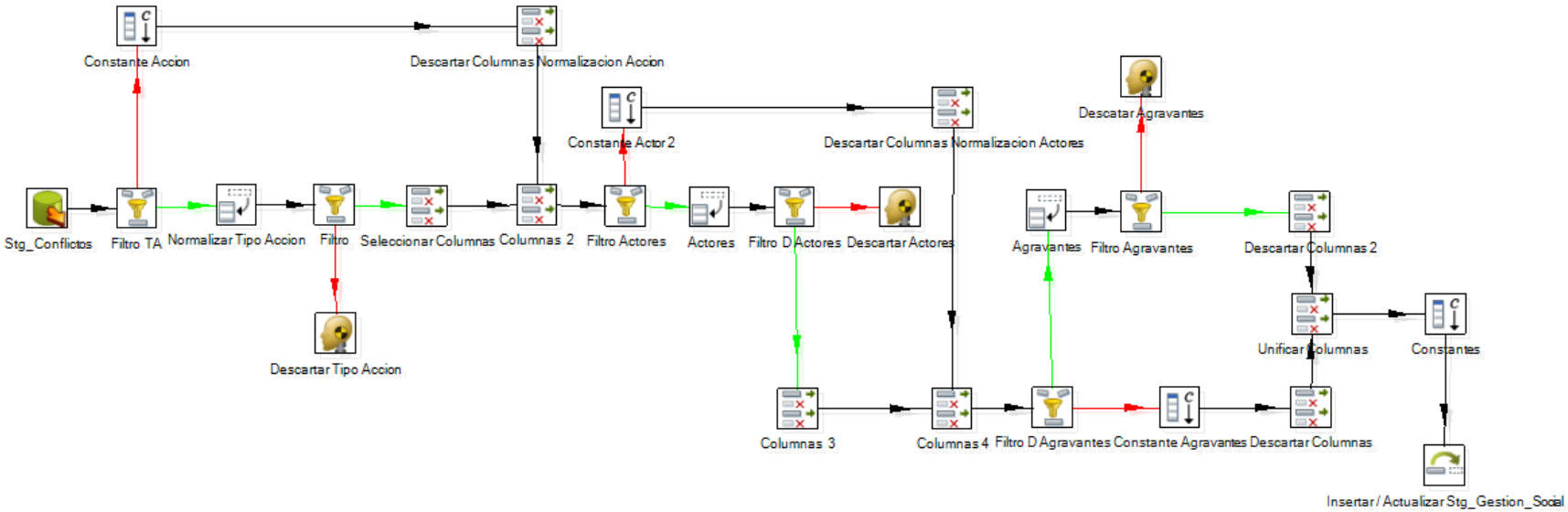


Figura 37 ETL stg_gestion_social – Conflictos

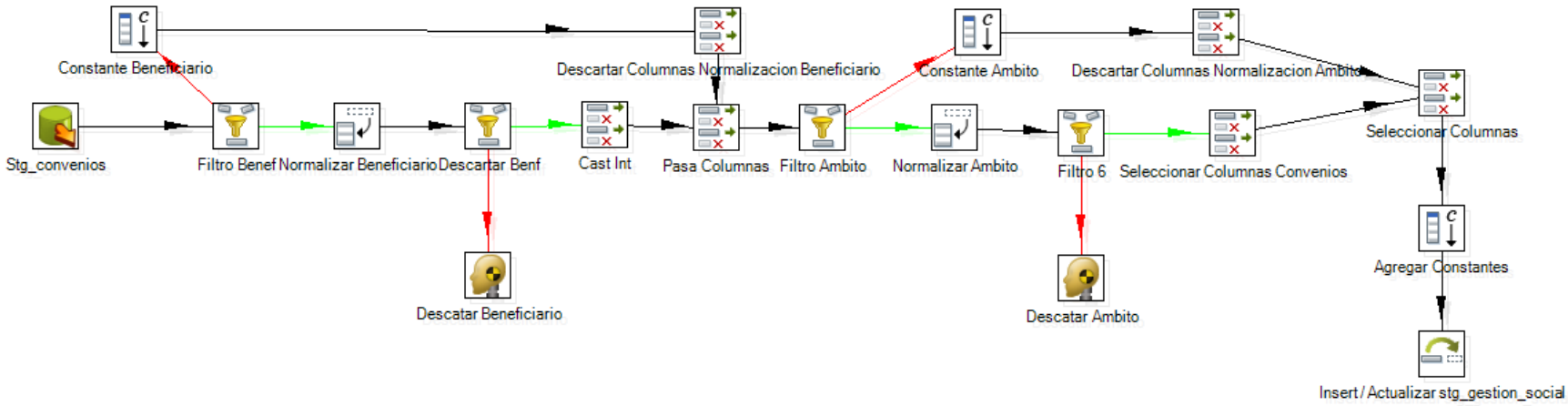


Figura 38 ETL stg_gestion_social – Convenios

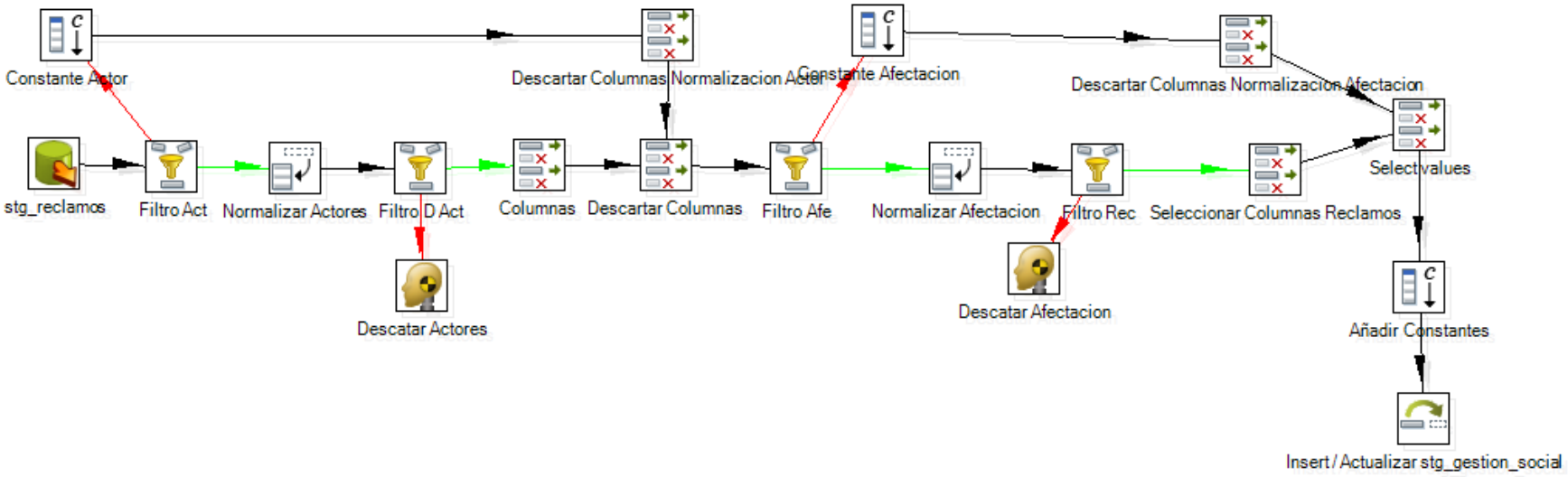


Figura 39 ETL stg_gestion_social – Reclamos

4.3.3 Módulo Hidrocarburífero Nacional

4.3.3.1 Estaciones

ETL Población Tabla stg_estaciones

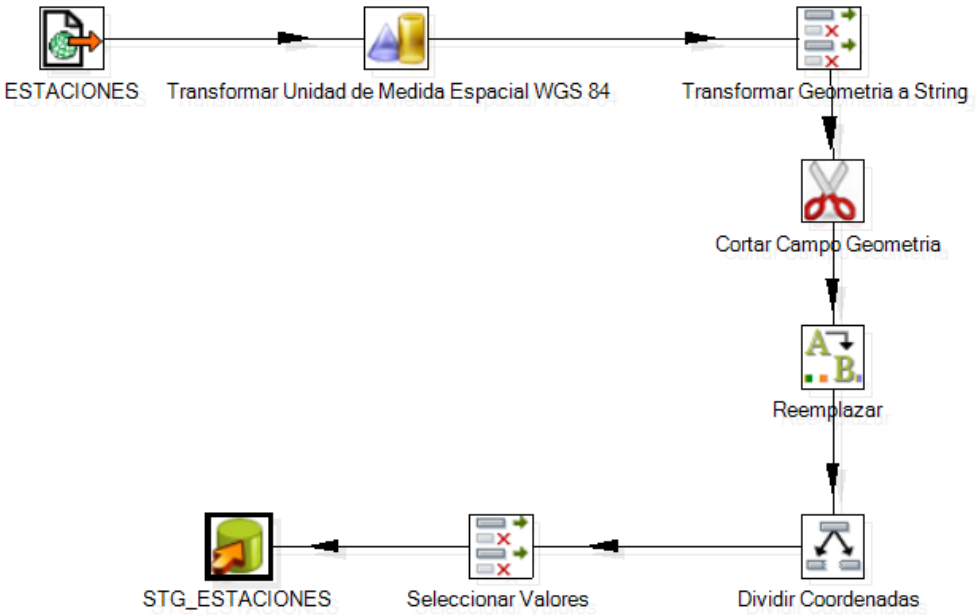


Figura 40 ETL stg_estaciones

4.3.3.2 Plataformas

ETL Población Tabla stg_plataformas

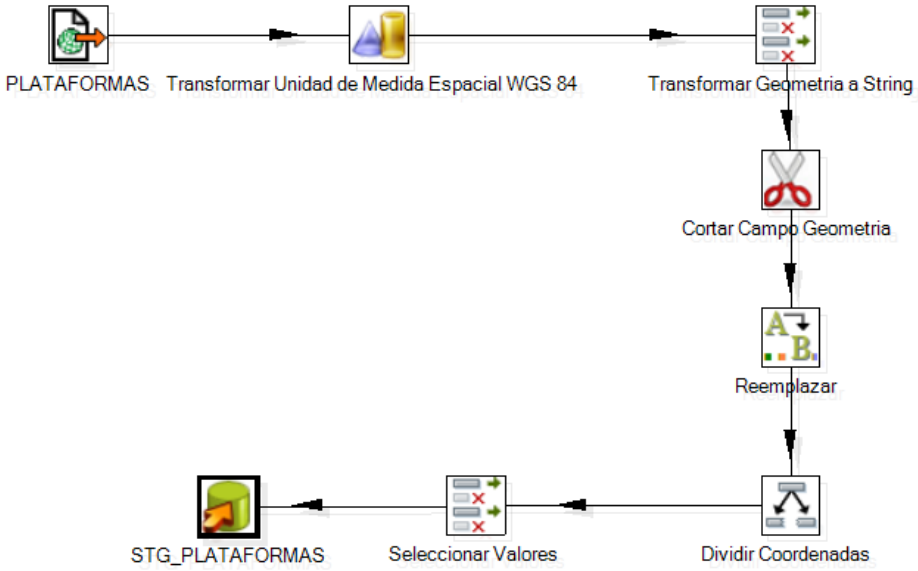


Figura 41 ETL stg_plataformas

4.3.3.3 Pozos

ETL Población Tabla stg_pozos

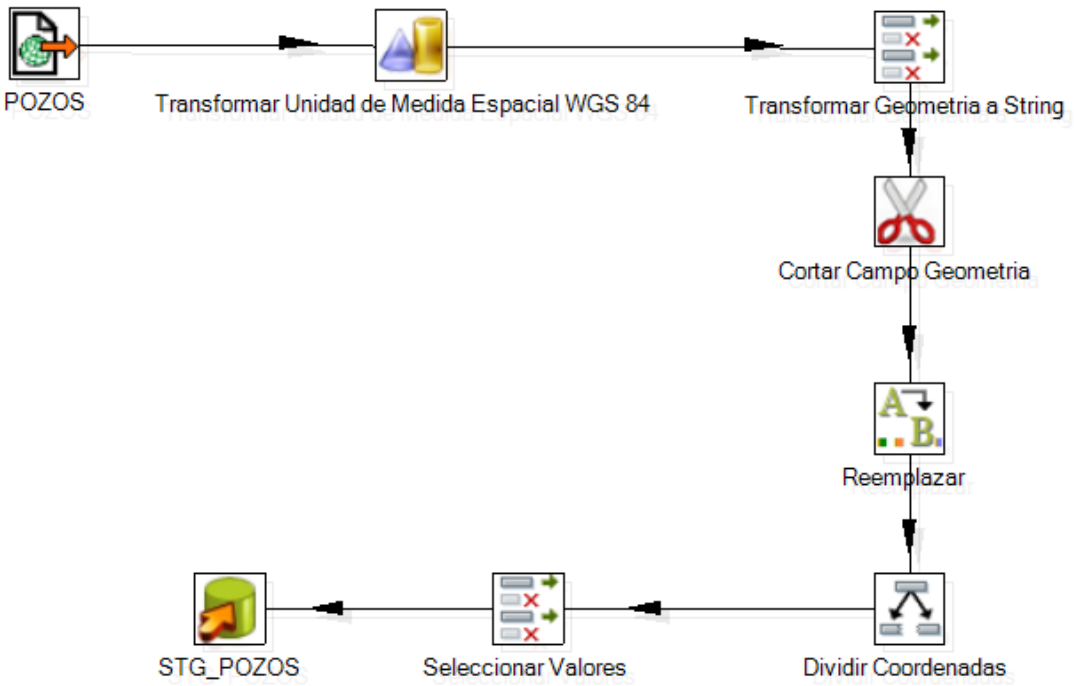


Figura 42 ETL stg_pozos

4.4 FASE 3: ALMACEN DE DATOS

4.4.1 EDW

4.4.1.1 Área Protegida

ETL Población Tabla area_protegida

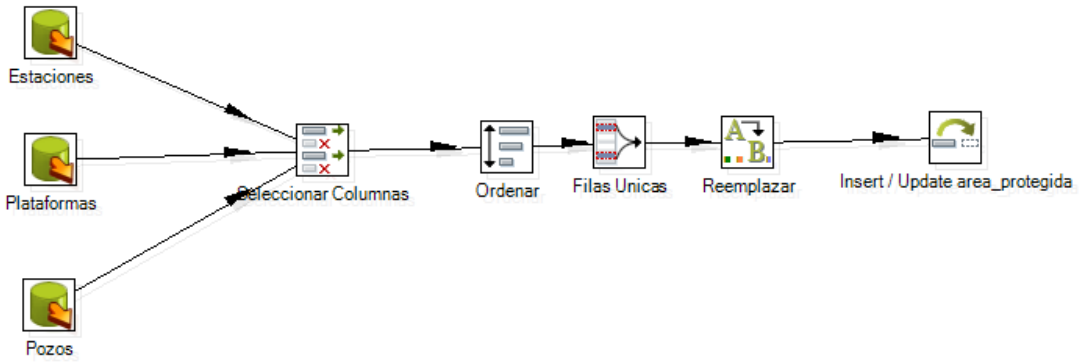


Figura 43 ETL area_protegida

4.4.1.2 Campo

ETL Población Tabla campo

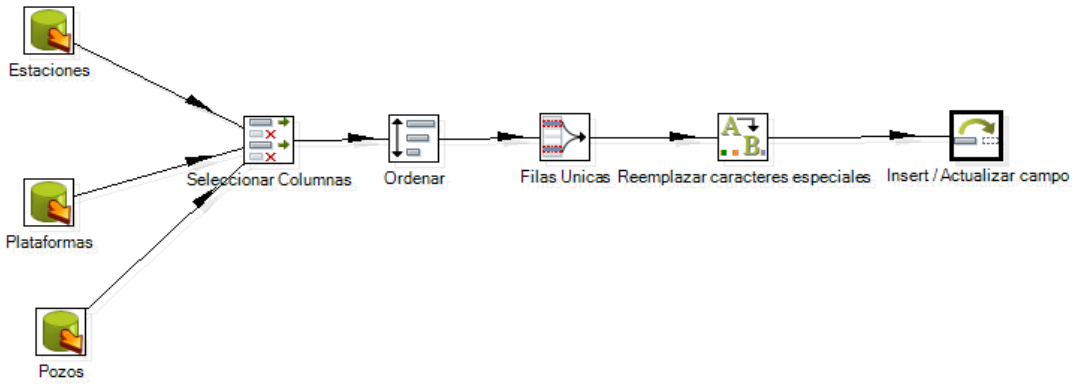


Figura 44 ETL campo

4.4.1.3 Cuenca

ETL Población Tabla cuenca

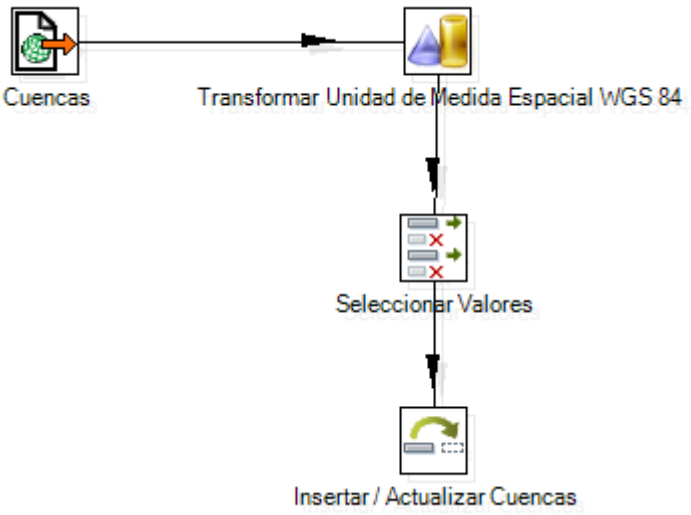


Figura 45 ETL Cuenca

4.4.1.4 Bloque Petrolero

ETL Población Tabla bloque_petrolero

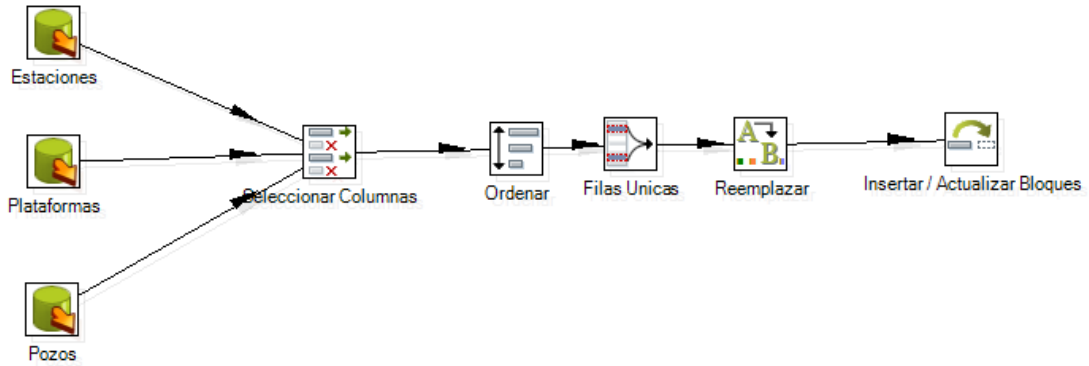


Figura 46 ETL bloque_petrolero

4.4.1.5 Territorio Indígena

ETL Población Tabla territorio_indigena

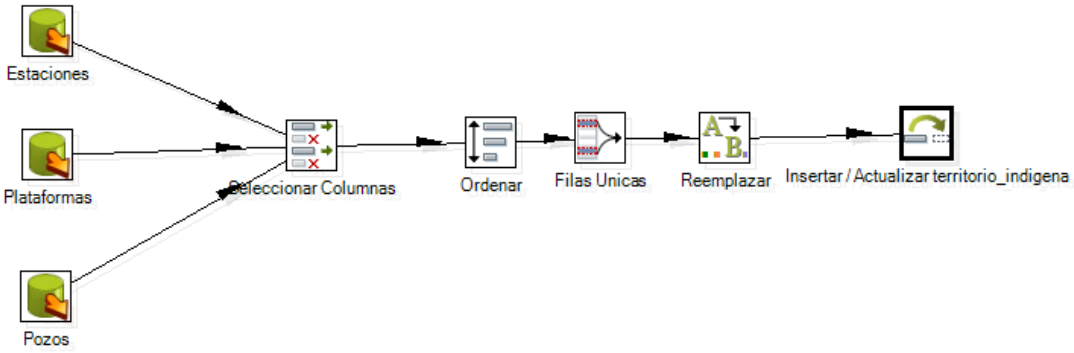


Figura 47 ETL territorio_indigena

4.4.1.6 Localidad

ETL Población Tabla localidad_i

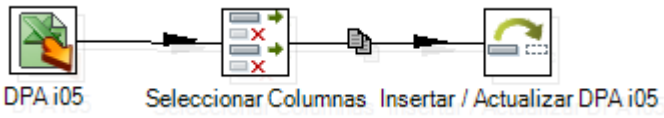


Figura 48 ETL localidad

4.4.1.7 Gestion Social

ETL Población Tabla gestion_social

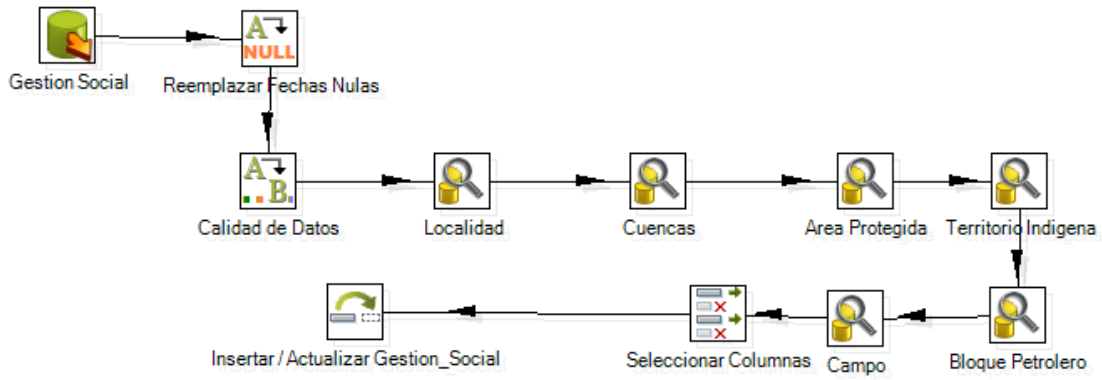


Figura 49 ETL gestion_social

4.4.1.8 Estado_Pozo

ETL Población Tabla estado_pozo

Figura 50 ETL estado_pozo

4.4.1.9 Estatal

ETL Población Tabla estatal

Figura 51 ETL estatal

4.4.1.10 Tipo_Estacion

ETL Población Tabla tipo_estacion

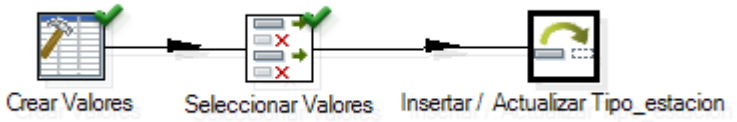


Figura 52 ETL tipo_estacion

4.4.1.11 Infraestructura

ETL Población Tabla infraestructura

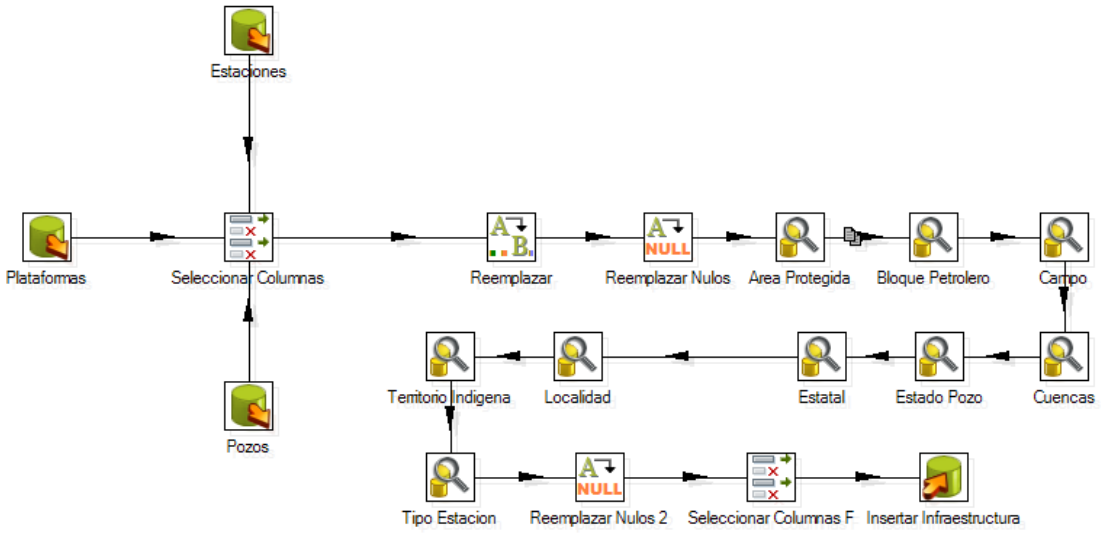


Figura 53 ETL infraestructura

4.4.2 Modelo Final– EDW

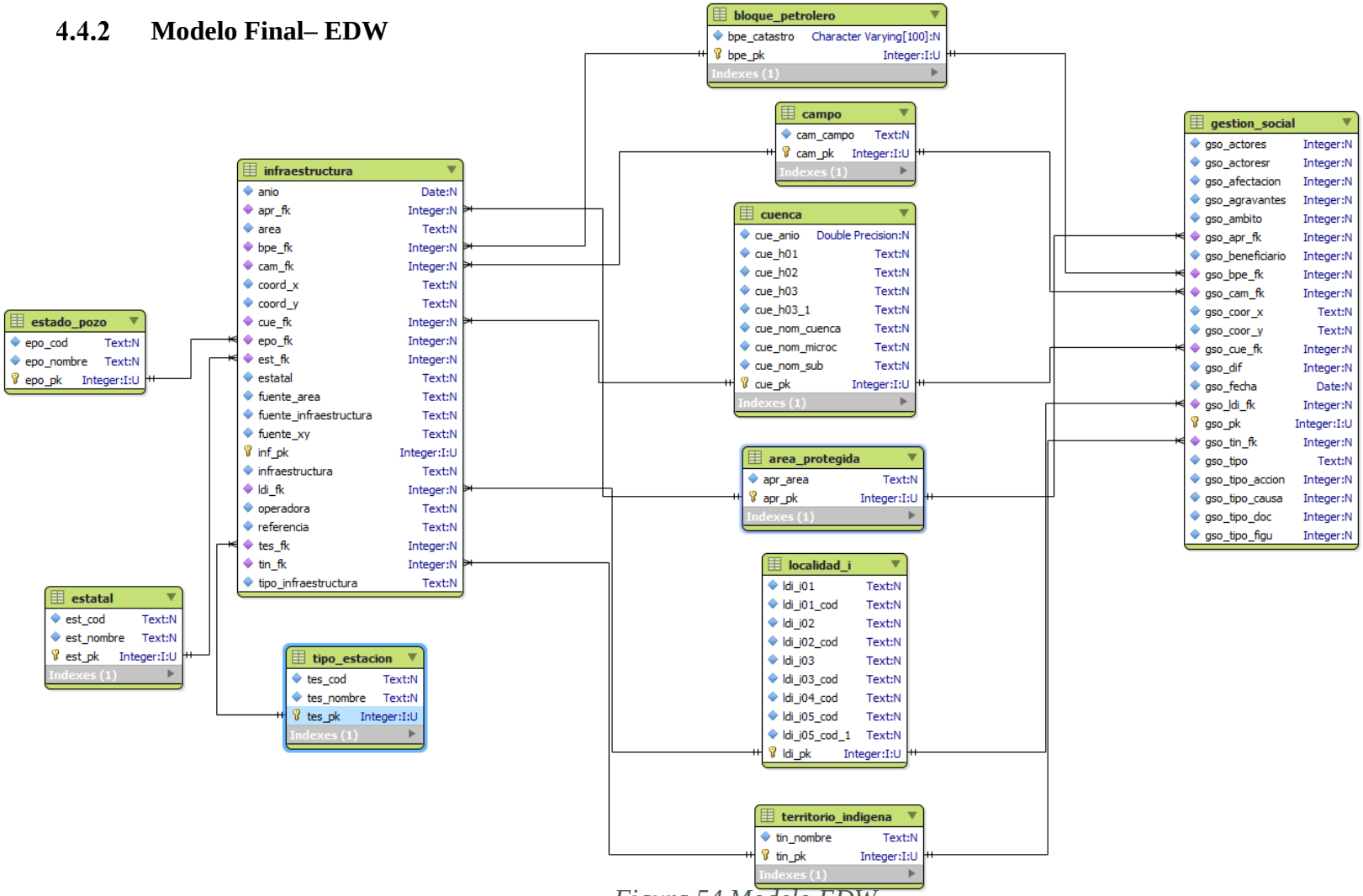


Figura 54 Modelo EDW

4.4.3 DATA MARTS

4.4.4 Dimensiones Compartidas

4.4.4.1 Dim_tiempo

ETL Población Tabla dim_tiempo

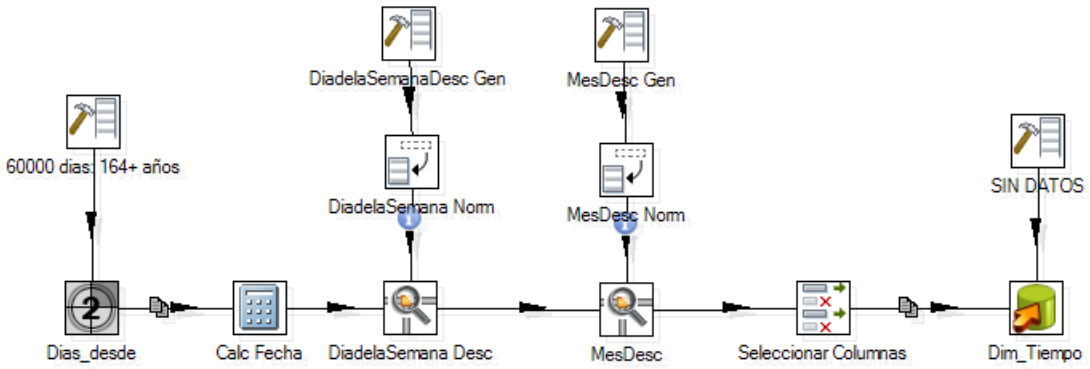


Figura 55 ETL dim_tiempo

4.4.4.2 Dim_area_protegida

ETL Población Tabla dim_area_protegida

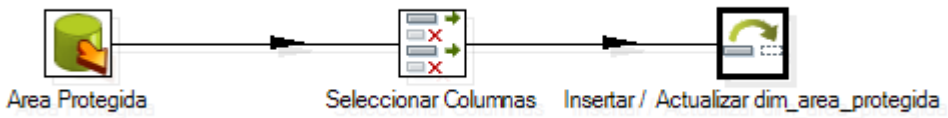


Figura 56 ETL dim_area_protegida

4.4.4.3 Dim_bloque_petrolero

ETL Población Tabla dim_bloque_petrolero

Figura 57 ETL dim_bloque_petrolero

4.4.4.4 Dim_campo

ETL Población Tabla dim_campo

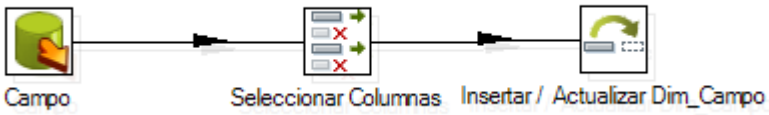


Figura 58 ETL dim_campo

4.4.4.5 Dim_cuenca

ETL Población Tabla dim_cuenca

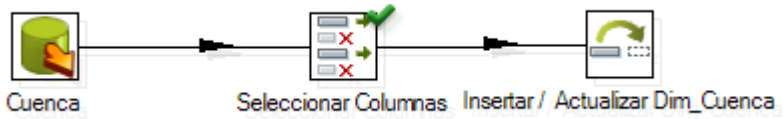


Figura 59 ETL dim_cuenca

4.4.4.6 Dim_territorio_indigena

ETL Población Tabla dim_territorio_indigena

Figura 60 ETL dim_territorio_indigena

4.4.4.7 Dim_localidad

ETL Población Tabla dim_localidad

Figura 61 ETL dim_localidad

4.4.5 Hidrocarburífero Nacional

4.4.5.1 Dim_infraestructura

ETL Población Tabla dim_infraestructura

Figura 62 ETL dim_infraestructura

4.4.5.2 Dim_estado_pozo

ETL Población Tabla dim_estado_pozo

Figura 63 ETL dim_estado_pozo

4.4.5.3 Dim_estatal

ETL Población Tabla dim_estatal

Figura 64 ETL dim_estatal

4.4.5.4 Dim_tipo_estacion

ETL Población Tabla dim_tipo_estacion

Figura 65 ETL dim_tipo_estacion

4.4.5.5 Fact_infraestructra

ETL Población Tabla fact_infraestructura

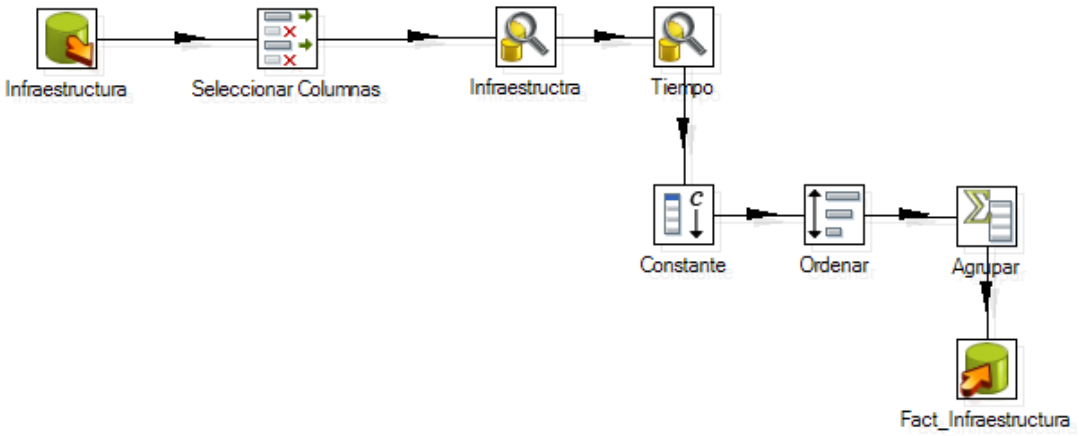


Figura 66 ETL fact_infraestructura

4.4.6 Modelo Estrella– Hidrocarburífero Nacional

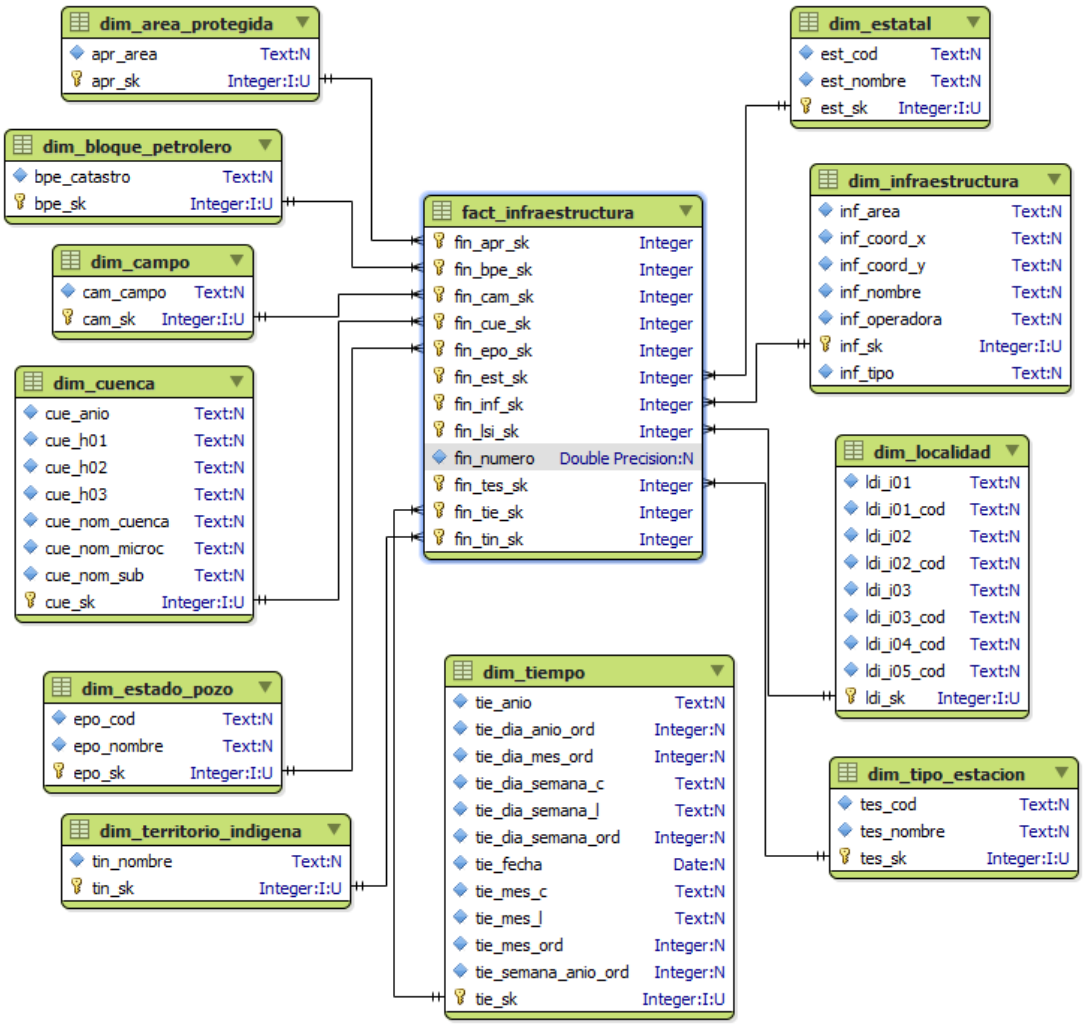


Figura 67 Modelo Estrella HN

4.4.7 Gestión Social

4.4.7.1 Dim_gestion_social

ETL Población Tabla dim_gestion_social

Figura 68 ETL dim_gestion_social

4.4.7.2 Dim_actor_beneficiario

ETL Población Tabla dim_actor_beneficiario

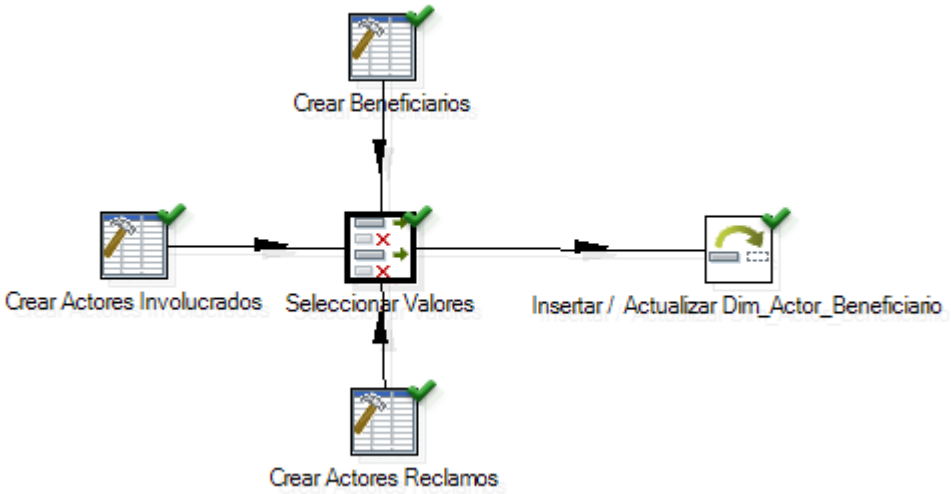


Figura 69 ETL dim_actor_ beneficiario

4.4.7.3 Dim_ambito_agravante_figura

ETL Población Tabla dim_ambito_agravante_figura

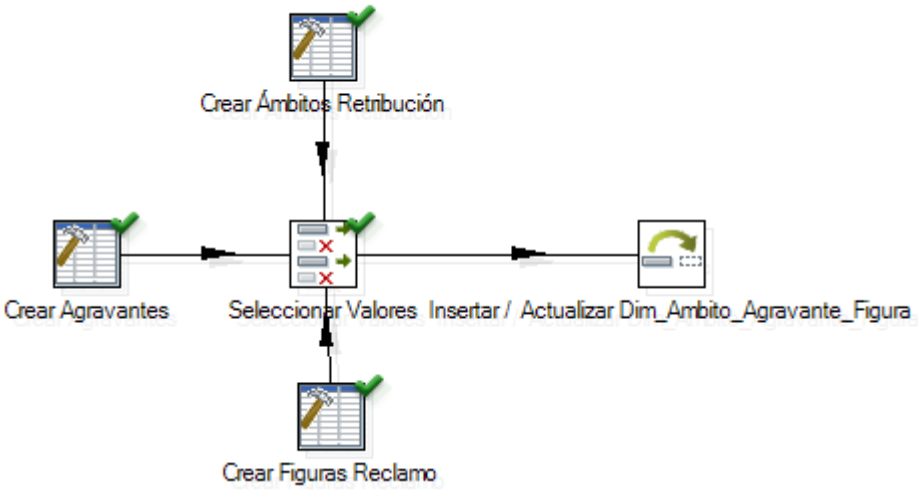


Figura 70 ETL dim_ambito_agravante_figura

4.4.7.4 Dim_tipo

ETL Población Tabla dim_tipo

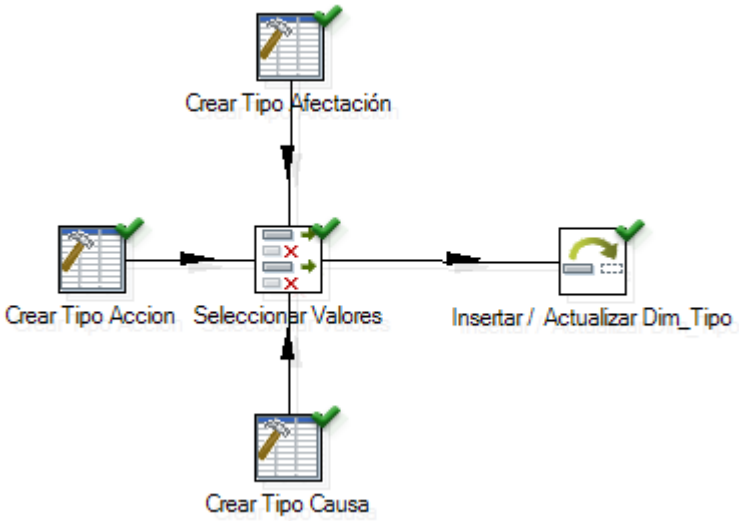


Figura 71 ETL dim_tipo_accion

4.4.7.5 Dim_tipo_documento

ETL Población Tabla dim_tipo_documento

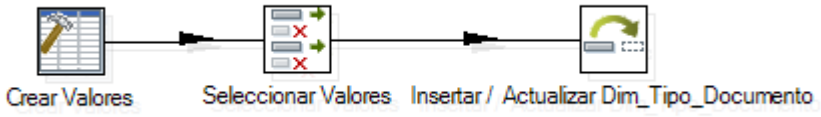


Figura 72 ETL dim_tipo_documento

4.4.7.6 Fact_Gestion_Social

ETL Población Tabla fact_gestion_social

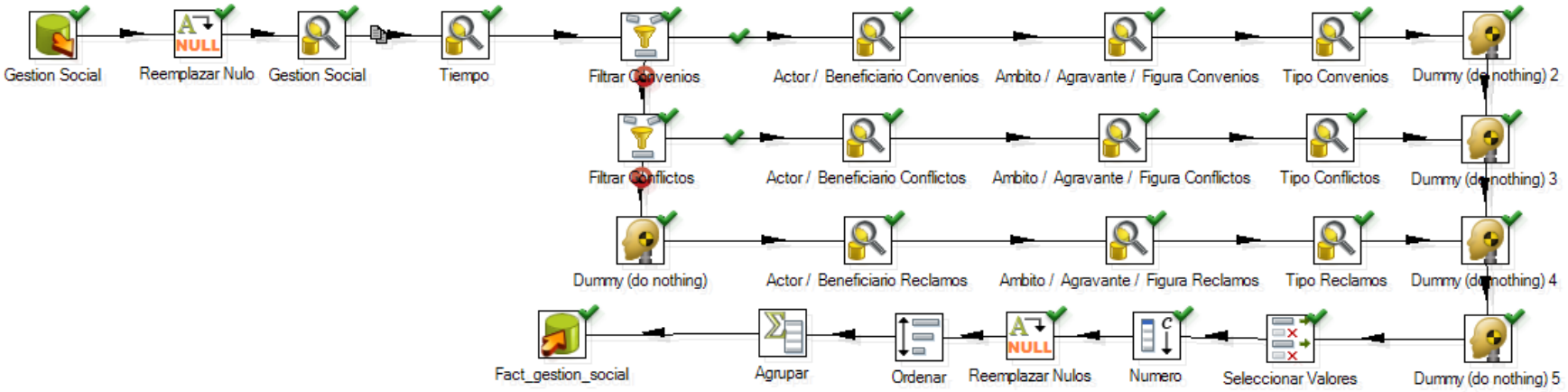


Figura 73 ETL fact_gestion_social

4.4.8 Modelo Estrella– Gestión Social

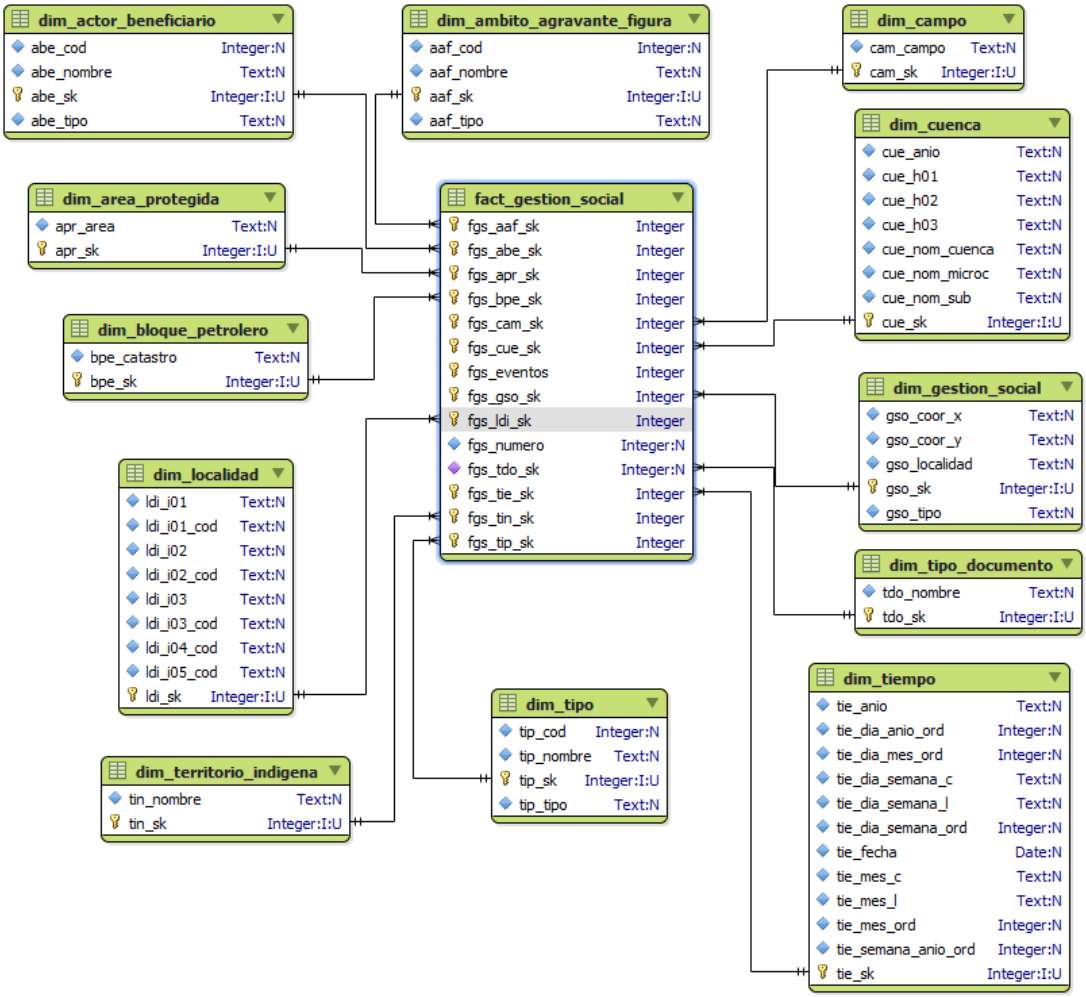


Figura 74 Modelo Estrella Gestión Social

4.5 FASE 4: MODELO

Los cubos de información han sido elaborados considerando el **"Drill-Across"** analizado en la FASE 1:

4.5.1 Dimensiones Compartidas (Dimension Usage)

Las dimensiones compartidas son aquellas determinadas en el "Drill-Across", las cuales serán creadas una única vez para todo el esquema que contrendrá a todos los cubos de información desde los cuales serán instanciadas como lo muestra la **Figura 75**.

Figura 75 Dimensiones Compartidas

4.5.2 Cubo Infraestructura

El cubo Infraestructura (ver **Figura 76**) muestra la tabla de hechos, las dimensiones únicas, las dimensiones compartidas y la medida Número Infraestructura.

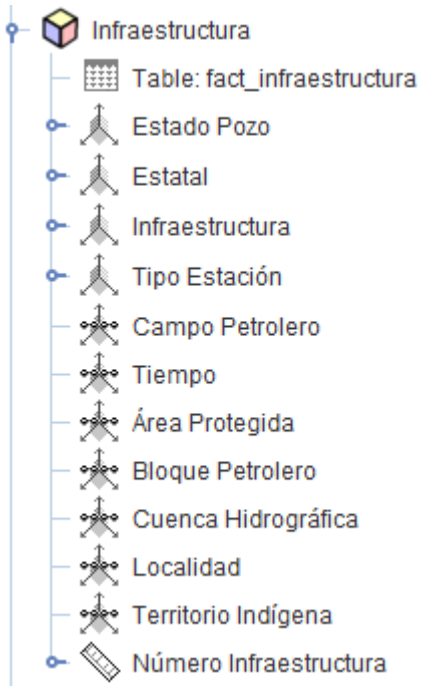


Figura 76 Cubo Infraestructura

4.5.3 Cubo Gestión Social

El cubo Gestión Social (ver **Figura 77**) muestra la tabla de hechos, las dimensiones únicas, las dimensiones compartidas y la medida Número Eventos.

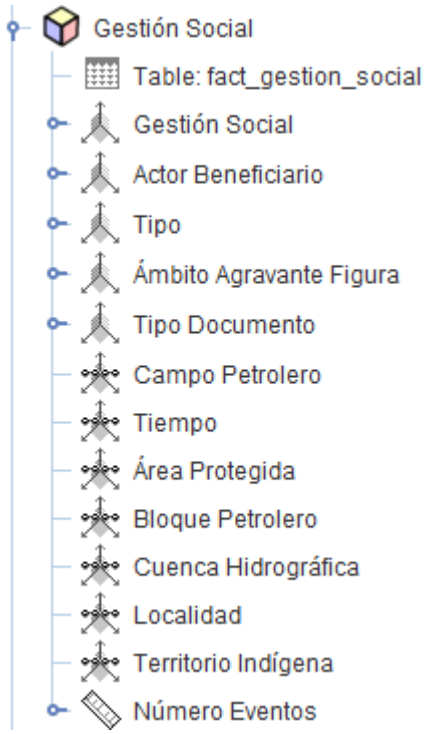


Figura 77 Cubo Gestión Social

4.5.4 Georeferenciación Cubos de información

Para la georeferenciación en Schema Workbench se debe agregar las siguientes propiedades a las jerarquias requeridas:

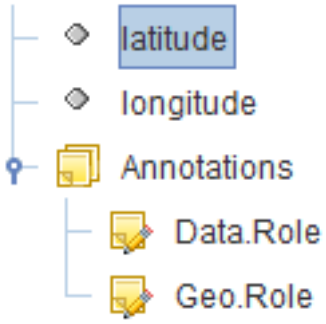


Figura 78 Propiedades Georeferenciación

Donde:

Tabla 13 Configuración Georeferenciación Mondrian

Campo	Configuración
Latitude	Coordenada X
Longitude	Coordenada Y
Data.Role	Geography
Geo.Role	location

Configurando a nivel de XML el reultado es:

```
<Annotations>
  <Annotation name="Data.Role">
    <![CDATA[Geography]]>
  </Annotation>
  <Annotation name="Geo.Role">
    <![CDATA[location]]>
  </Annotation>
</Annotations>
<Property name="latitude" column="gso_coor_x" type="String">
</Property>
<Property name="longitude" column="gso_coor_y" type="String">
</Property>
</Level>
```

4.6 FASE 5: PRESENTACIÓN

En el proceso de elaboración de tableros de control se ha generado las siguientes soluciones (ver **Figura 79, 80, 81, 82**):

4.6.1 Página principal de tableros de control

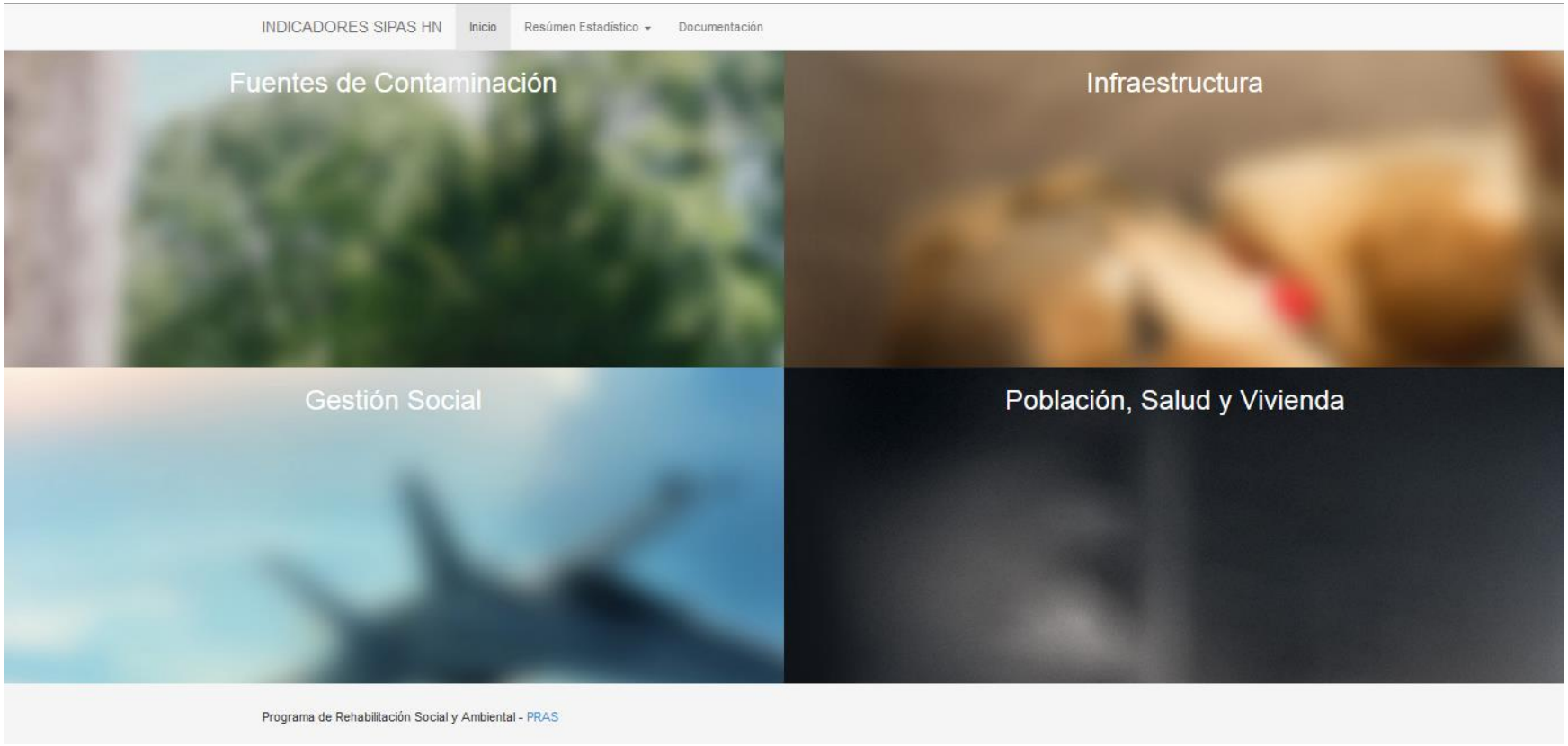


Figura 79 Página principal de tableros de control

4.6.2 Tablero de control Estaciones

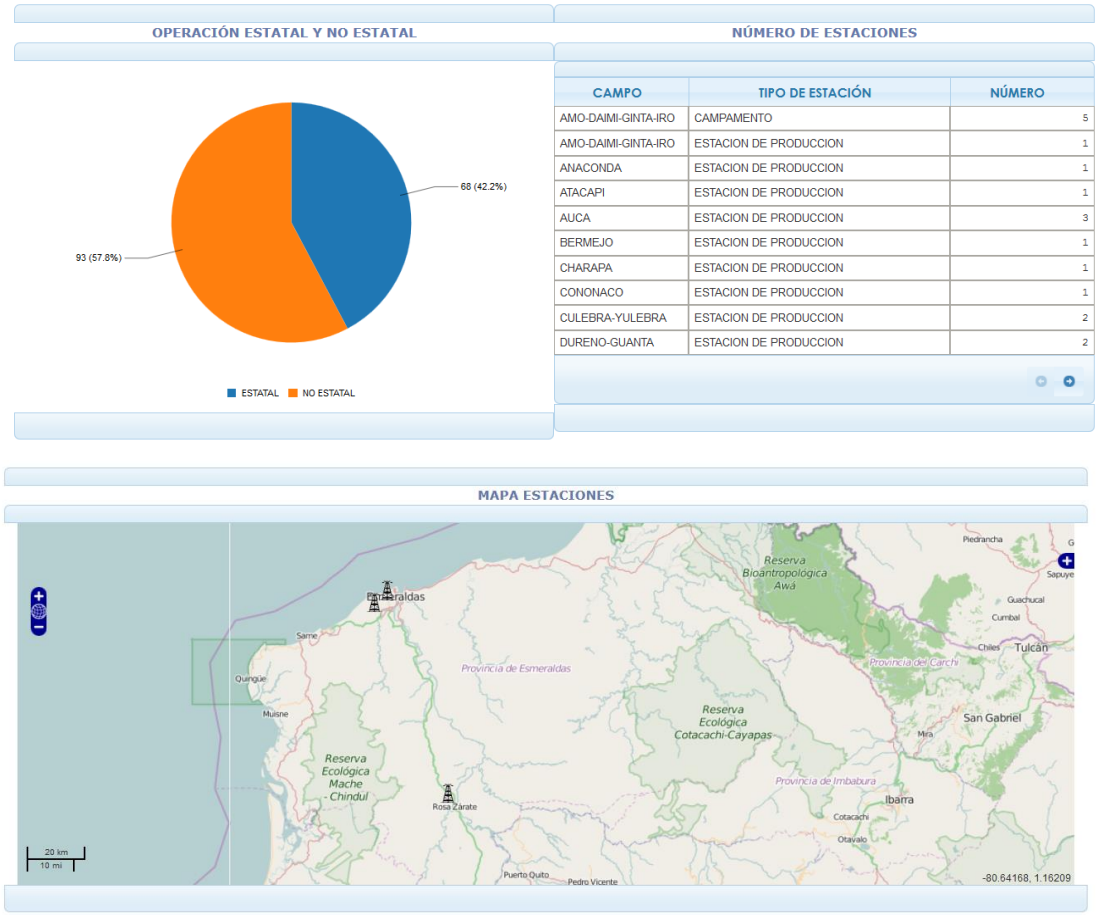


Figura 80 Tablero de Control Estaciones

4.6.3 Tablero de control Pozos

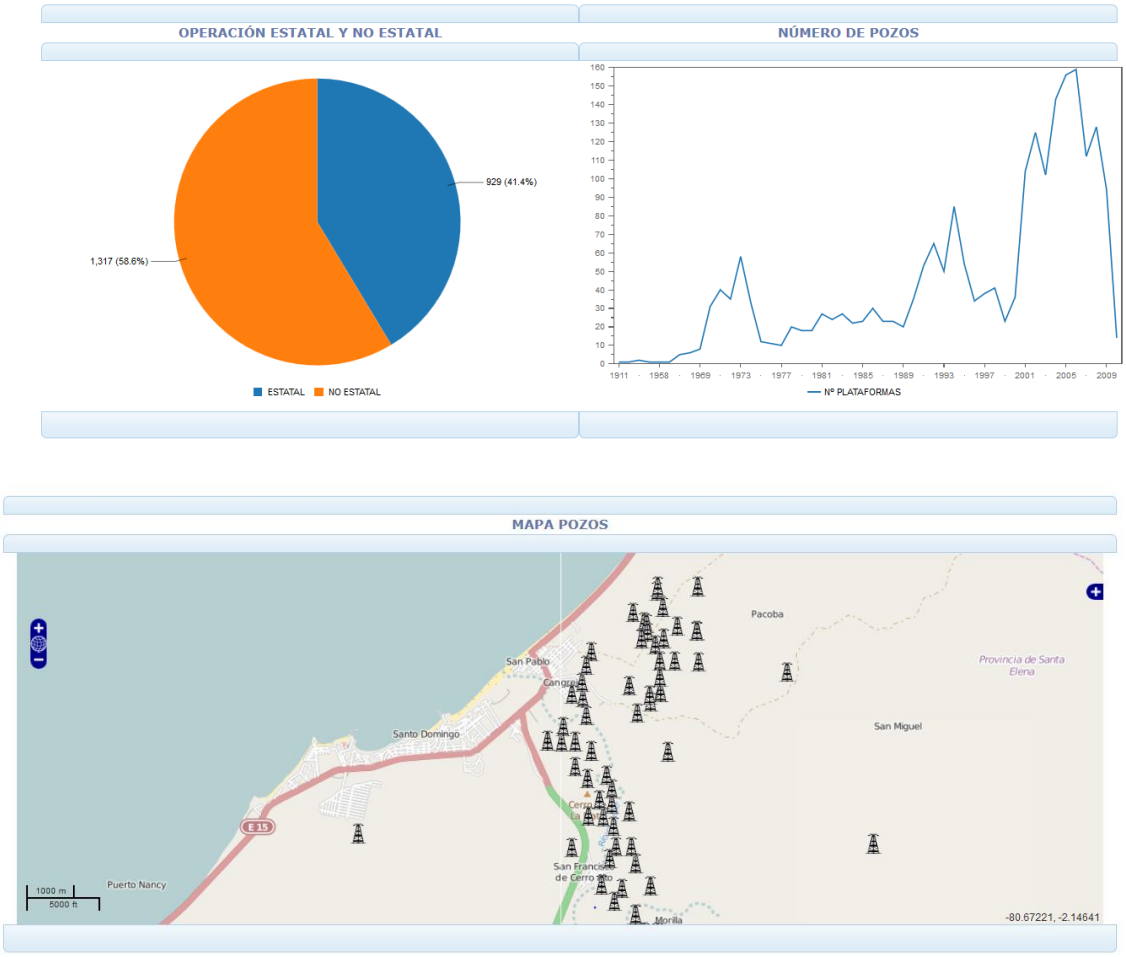


Figura 81 Tablero de control Pozos

4.6.4 Tablero de control Plataformas

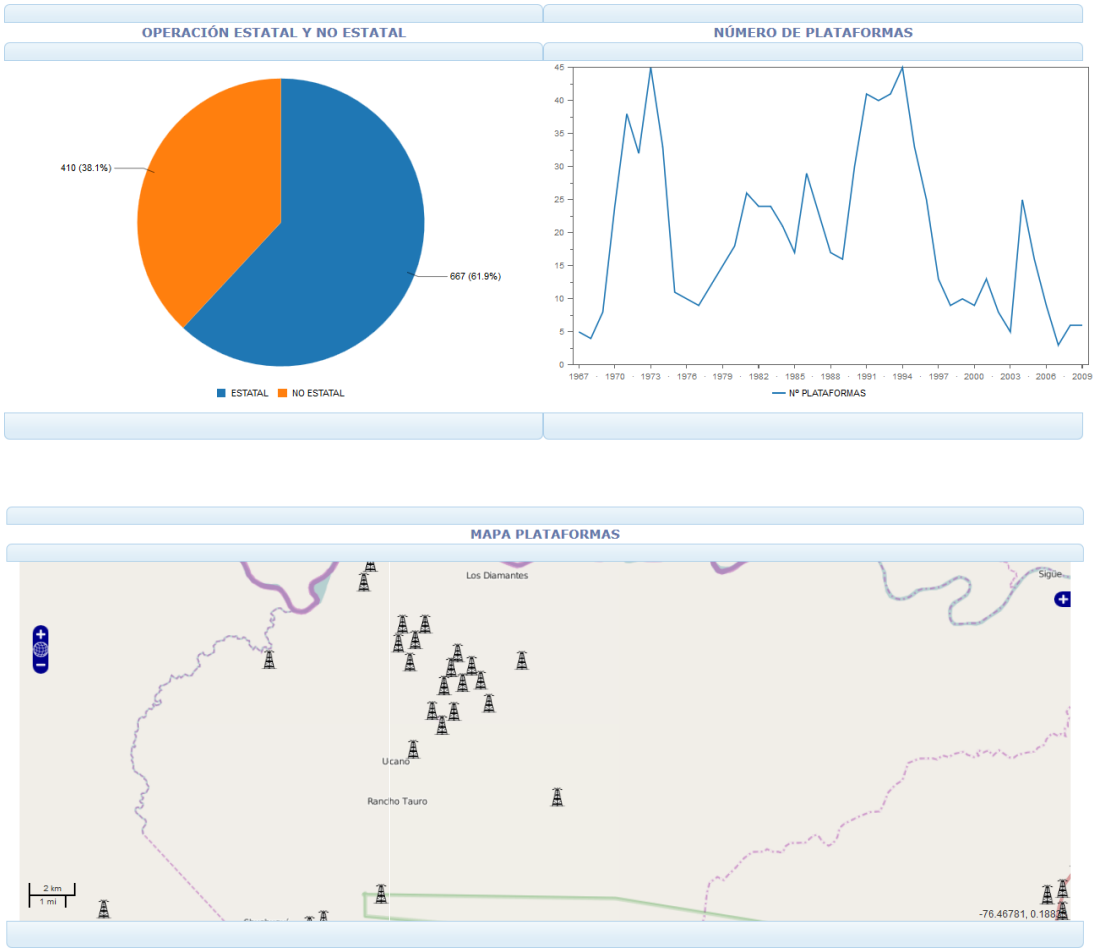


Figura 82 Tablero de Control Plataformas

CAPÍTULO 5
CONCLUSIONES Y RECOMENDACIONES

5.1 CONCLUSIONES

- El data warehouse construido para el desarrollo del proyecto "Análisis y estructuración de la información hidrocarburífera nacional y geoespacial para el diseño y construcción de un data warehouse para la toma de decisiones socio-ambientales del Programa de Reparación Ambiental y Social - PRAS", basado en una metodología híbrida (Inmon y Kimball), además de la correcta definición del nivel de detalle (nivel granularidad) y grado de cohesión, ha permitido solventar las necesidades de información para la gestión, de esta manera el rendimiento de la herramienta de inteligencia de negocios y los resultados satisfacen los requerimientos inicialmente planteados, permitiendo la simplificación del acceso a la información y la presentación de informes avanzados.

- La implantación del sistema de inteligencia de negocios "Pentaho" ha sido útil para observar, de primera mano, el gran aporte dentro de la organización, por ser un sistema fácil, potente, asequible convirtiendose en la base para futuras implementaciones.

- El desarrollo de la solución sobre la herramienta de código abierto Pentaho, ha permitido demostrar sus capacidades, aprovechando el acceso a su código fuente para potenciar: la visualización de la consola de usuario y el diseñador de tableros de control, las posibilidades de integración vía portal web a toda la información, destacando enormemente los informes a nivel geográfico.

5.2 RECOMENDACIONES

- Para asegurar una gestión eficaz de la información dentro de la institución se debe construir un EDW en tercera forma normal para asegurar un posterior crecimiento, además de tener en consideración las metodologías y establecimiento del nivel de detalle que requiere el desarrollo.

- Para una mejor comprensión y mantenimiento del sistema de inteligencia de negocios, se recomienda que las personas parte de alta gerencia y usuarios finales sean capacitadas en el uso funcional y técnico en la herramienta para certificar su uso eficiente.

- Es recomendable la utilización de herramientas open source para llevar a cabo el desarrollo de este tipo de proyectos, destinando un tiempo adecuado para el levantamiento de requerimientos para el total cumplimiento de los productos desarrollados, reutilizando lo ya existente y desarrollando o parametrizando lo que no exista en la aplicación.

BIBLIOGRAFÍA

Ayala Peña, A. (2006). *Inteligencia de Negocios: Una Propuesta para su Desarrollo en las organizaciones.* México.

Bernabeu, I. R. (2010). *Hefesto V2.* Córdoba, Argentina.

Bernal, I. J. (2012). Fundamentos de Desarrollo de Sistemas.

Buenas Tareas. (2012, Octubre). From Aplicación de Inteligencia de Negocios: http://www.buenastareas.com/ensayos/Ambitos-De-Aplicacion-De-Inteligencia-De/5691982.html

Dario, I. B. (2010, Enero 27). *Hefesto Blogspot*. From http://tgx-hefesto.blogspot.com/2010/01/claves-subrogadas.html

Drazda. (2013, Abril 7). From http://drazda.blogspot.com/2013/04/data-warehousing-concepts-kimball-vs.html

Indensa. (n.d.). From Inteligencia de Negocios: http://www.idensa.com/?start=3

Inmon. (2002). *Building the Data Warehouse, (Third Edition).* John Wiley & Sons.

Inmon, W. H. (2000). *Buiildiing the Data Warehouse: Gettiing Started.*

Kimbal, R. (1996). *The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data.* John WIley & Sons.

Kimball, R. a. (2000). *The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling (Second Edition).* New York: John Wiley & Sons.

Leporati, C. L. (n.d.). *Gestión de Almacenamiento, Universidad ULACIT*. From http://www.gestiopolis.com/

Redondo, J. (2010). Construyendo La Arquitectura de una eficiente Bodega de Datos.

Sinnexus. (© Copyright 2007 - 2012). *Sinnexus*. From http://www.sinnexus.com/business_intelligence/index.aspx

Watson, H. J. (2005). *Data Warehouse Architectures: Factors in the Selection Decision and the Success of the Architectures.* Athens, Georgia.

GLOSARIO

- **BI:** Business Intelligence.
- **DW:** Data Warehouse.
- **ETL:** Extract Transformation Load (Extracción, Transformación y Carga)
- **OLAP:** On Line Analytical Processing.
- **OLTP:** On Line Transaction Processing.
- **MDX:** Multidimensional Expression.
- **CDE:** Community Dashboard editor.
- **XML:** eXtensible Markup Language.

BIOGRAFÍA

Nombres y Apellidos:

Ricardo Miguel Díaz Razo.

Lugar y Fecha de Nacimiento:

Quito, 09 de Septiembre de 1987.

Educación Primaria:

Colegio “Paulo Sexto” – Quito.

Educación Secundaria:

Colegio “Paulo Sexto” – Quito.

Educación Superior:

Universidad de las Fuerzas Armadas ESPE – Sangolquí
Ingeniería de Sistemas e Informática.

Títulos Obtenidos:

Suficiencia en el idioma Inglés.

HOJA DE LEGALIZACIÓN DE FIRMAS

ELABORADO POR:

Ricardo Miguel Díaz Razo

DIRECTOR DE LA CARRERA:

Ing. Mauricio Campaña MsC.

Sangolquí, Julio del 2015