ISMMM
INSTITUTO SUPERIOR MINERO
METALURGICO DE MOA
DR. ANTONIO NUÑEZ JIMENEZ

Ingeniería Informática

Facultad: Geología y Minas

Trabajo de Diploma

Para Optar por el Título de

Ingeniero Informático

Título: Mercado de Datos para el análisis de los datos hidrológicos en la provincia Holguín

Autor: Alianne Sánchez Rómulo

Tutor(es): Ing. Edgar Núñez Torres

Ing. Oscar Reyes Pérez

Moa, 2013

"Año 55 de la Revolución

Pensamiento

Si supiera lo que estoy haciendo no lo llamaría Investigación.

Albert Einstein.

Dedicatoria

A mis padres Bárbara y Leonardo sin ustedes no estaría hoy aquí a ustedes va este trabajo.

Agradecimientos

A mis padres por apoyarme incondicionalmente

A mi familia por el amor que siempre me ha rodeado

A mis amigas y amigos por estar en las buenas y en las malas (soportándome)

A mis tutores por la paciencia

A mi novio por el cariño

A todos los que de una manera u otra se interesaron por el desarrollo de esta investigación:

GRACIAS.

DECLARACIÓN DE AUTORÍA.

Declaro que yo soy la única autora de este trabajo y autorizo al Instituto Superior Minero Metalúrgico de Moa "Dr. Antonio Núñez Jiménez" para que hagan el uso que estimen pertinente del mismo.

Para que así conste firmamos la presente a los 14 días del mes de Junio del 2013.

Alianne Sánchez Rómulo

_____

Firma del autor

Ing. Edgar Núñez Torres

_____

Firma del tutor

Ing. Oscar Reyes Pérez

_____

Firma del tutor

Resumen

Los almacenes de datos (Data Warehouse DWH) han progresado paulatinamente y son repositorios diseñados para facilitar la confección de informes y la realización de análisis para la toma de decisiones, los cuales se subdividen en unidades lógicas más pequeñas llamadas mercado de datos (Data Mart DM). Los mercados de datos son una versión del almacén que resuelven estudios a nivel de departamento, en específico para una necesidad de datos seleccionados.

Esta investigación presenta el diseño lógico de un DM para el análisis de datos hidrológicos en la Empresa de Aprovechamiento de Recursos Hidráulicos Holguín (EAHHLG). La misma estuvo basada en las siguientes metodologías: Ralph Kimball para el diseño de la arquitectura del DM y HEFESTO para el desarrollo del DWH. Además fueron empleadas las siguientes herramientas: Pentaho Bussiness Intelligence, Pentaho Data Integration 4.2.1, Pentaho Schema Workbench, PostgreSQL 9.0 y Embarcadero ERStudio 8.0.

Palabras claves: Mercado de datos o Data Mart, datos hidrológicos, Pentaho.

Abstract

Data warehouses (DWH Data Warehouse) have progressed slowly and are repositories designed to facilitate the preparation of reports and analyzes for decision-making, which are subdivided into smaller logical units called data mart (DM Data Mart). The market data is a store version that meets departmental level studies, specifically for a selected data needs.

This research presents the logical design of a DM for the analysis of hydrological data in Hydrologic Resource Development Corporation Holguin (EAHHLG). It was based the following methodology on Ralph Kimball for designing DM architecture and for the development HEFESTO. They were also employed the following tools: Pentaho Business Intelligence, Pentaho Data Integration 4.2.1, Pentaho Workbench Schema, PostgreSQL 9.0, and 8.0 ERStudio Embarcadero.

Keywords: Market Data or Data Mart, hydrological data, Pentaho

Índice

Introducción

Actualmente en la provincia de Holguín existe la necesidad de tener una visión analítica y universal de la evolución de situaciones ambientales, sociales, administrativas a través del acceso a bases de información que se alimentan de datos de diferente naturaleza. Sin embargo, muy pocos trabajos hoy en día atacan de manera frontal los problemas de la integración de datos del medio ambiente, en particular los datos de las lluvias precipitadas y la situación de los embalses.

La Empresa Aprovechamiento de los Recursos Hidráulicos Holguín (EAHHLG) maneja un gran número de información, la cual se obtiene de un amplio proceso que ocurre de manera conjunta con las Delegaciones Territoriales de la provincia. Representantes de cada delegación trabajan coordinadamente con la provincia para suministrar y actualizar la información concerniente al ámbito de competencia. La empresa EAHHLG se encarga de procesar las informaciones referentes a la lluvia precipitada y la situación de los embalses en cada municipio y en la provincia en general. La Delegación Territorial Moa atiende además de este municipio a los de Sagua de Tánamo y Frank País. En cuanto a los embalses hay que decir que el municipio de Moa atiende la Presa Nuevo Mundo y la Derivadora Moa.

Toda la información que poseen estas entidades es entregada a la EAHHLG donde se va registrando, a medida que va llegando, ya sea en soporte digital o en documento impreso, también se pueden realizar los partes a través del teléfono o planta de radio. Toda esta información es registrada en ficheros Excel que contienen información resumida, basada en datos recolectados diariamente de las principales fuentes hidrológicas de los municipios y la provincia. Cada uno de estos ficheros almacena información referente a un tema dado, el cual permite conocer medidas específicas pero no realizar un análisis de los datos a través de un período de tiempo dado, ni asegurar que todos sean almacenados en un mismo lugar para una posterior consulta, constituyendo esto la situación problémica que da lugar a esta investigación.

De ahí que el presente trabajo de diploma para brindar solución a la situación descrita con anterioridad, posee el siguiente **problema a resolver**: ¿Cómo lograr el

almacenamiento integrado de los datos hidrológicos históricos obtenidos en la Delegación Territorial Moa para su posterior análisis?

En busca de una solución al problema antes planteado, se propone como **objeto de estudio:** Proceso de desarrollo de Data Warehouse.

El **campo de acción:** viene enmarcado en el proceso de desarrollo de un Mercado de Datos para el análisis de los datos hidrológicos obtenidos en la Provincia Holguín

Como guía en esta investigación se plantea la siguiente **idea a defender:** Con el desarrollo de un Mercado de Datos que contenga los datos hidrológicos en la provincia se le facilitará el trabajo a los especialistas de la Empresa Aprovechamiento de los Recursos Hidráulicos Holguín (EAHHLG).

Teniendo en la misma como **objetivo general:** Desarrollar un Mercado de Datos para el análisis de los datos hidrológicos en la provincia Holguín.

Para el cumplimiento del objetivo general se implementa el siguiente **sistema de tareas:**

- ✓ Estudio del proceso para el manejo de información en la EAHHLG
- ✓ Investigación de las metodologías y herramientas de construcción de Mercados de Datos.
- ✓ Realización del diseño del modelo lógico del Mercado de Datos.
- ✓ Realización del diseño del modelo físico del Mercado de Datos.
- ✓ Realización de proceso de Extracción Transformación y Carga (ETL por sus siglas en inglés) para poblar el mercado.

Para dar respuesta a estas tareas propuestas se emplearon los métodos científicos de la investigación: teóricos y empíricos.

Los métodos empíricos estudian las características y relaciones esenciales del objeto que son accesibles directamente desde la percepción sensorial (conocimiento sensorial).

Los métodos teóricos se aplican durante el proceso de explicación, predicción, interpretación y comprensión de la esencia del objeto. Posibilitan la interpretación conceptual de los datos empíricos, revelan las relaciones esenciales del objeto de investigación que no son observados a simple vista, participan en la construcción del modelo y la hipótesis de la investigación

Entre los métodos empíricos se encuentran:

- Entrevista: Necesaria en la recopilación de la información para el conocimiento del problema en general. En esta investigación, se realizaron varias entrevistas con el experto, que radica en la Delegación Territorial Moa, con el fin de obtener los requisitos necesarios para llevar a cabo el proyecto.
- Comparación: Esta se utilizó en la búsqueda y solución de problemas, donde pudimos comparar todas las herramientas estudiadas y así definir cuál utilizar.
- Revisión de documentos utilizados para la recopilación de información: en el estudio de diferentes bibliografías para la selección de metodologías y herramientas, lo cual aporta elementos para la fundamentación de la solución.
- La observación: se empleó para percibir cómo se gestiona la información en la Empresa Aprovechamiento de los Recursos Hidráulicos Holguín.

Entre los métodos teóricos se encuentran:

- Análisis y síntesis: empleado en la recopilación y el procesamiento de la información obtenida en los métodos empíricos y de esta forma arribar a las conclusiones.
- Método de Modelación: empleado en la construcción de modelos como el modelo físico de la Base de Datos y el modelo lógico de la misma.
- Inducción–deducción: empleado para la aplicación de la metodología de desarrollo y la interpretación de los resultados.

El documento fue elaborado siguiendo la siguiente estructura:

Capítulo 1 FUNDAMENTACIÓN DEL MARCO TEÓRICO: Se expone el estado del arte, donde se realiza la fundamentación teórica del tema. Al mismo tiempo se describe el objeto de estudio, se explica la metodología para la construcción de un Data Mart, se realiza el estudio y selección de las herramientas, así como de los artefactos para su elaboración.

Capítulo 2 ANÁLISIS Y DISEÑO DEL DATA MART: Se realiza el diseño e implementación del Data Mart, se definen las dimensiones y las medidas, y se elaboran los artefactos seleccionados en el capítulo anterior para su construcción a partir de la metodología escogida.

Capítulo 3 INTEGRACIÓN DE DATOS: En este capítulo es donde se desarrolla la última fase de la metodología en la cual se realiza el proceso ETL para poblar el mercado.

Capítulo 1 FUNDAMENTACIÓN DEL MARCO TEÓRICO

Introducción

En este capítulo se hace referencia a una serie de conceptos, de los cuales sería de gran importancia tener dominio para la futura comprensión de algunos términos que serán tratados más adelante. Se realiza un estudio de las herramientas con las que se obtendría el modelado lógico y físico de la base de datos, además de las metodologías existentes para determinar cuál es la más conveniente para el desarrollo de este trabajo.

1.1 Empresa Aprovechamiento de los Recursos Hidráulicos Holguín: manejo de información.

La Empresa Aprovechamiento de los Recursos Hidráulicos Holguín (EAHHLG) maneja un gran número de información, la cual se obtiene de un amplio proceso que ocurre de manera conjunta con las Delegaciones Territoriales de la provincia. Representantes de cada delegación trabajan coordinadamente con la EAHHLG para suministrar y actualizar la información concerniente al ámbito de competencia. Toda la información que poseen estas entidades es entregada a la EAHHLG, donde se va registrando, a medida que va llegando, en soporte digital o en documento impreso; para esto se destina un responsable para la administración del flujo de información, ya que recibe, recopila, procesa y elabora la información resultante del monitoreo y la comunica a la EAHHLG para la toma de decisiones.

La Empresa Aprovechamiento de los Recursos Hidráulicos Holguín se encarga de procesar las informaciones referentes a la lluvia precipitada y la situación de los embalses en cada municipio y en la provincia en general.

Existe una Red Pluviométrica en estos municipios que cuenta con 11 equipos (pluviómetros) en Moa, con 23 en Sagua y con 9 en Frank País. Esta red está compuesta por la Red Informativa del día 3 y 15, la Red Informativa Especial, la Red Informativa Básica (esta red está compuesta por los equipos que tienen más de 50 años informando) y la Red Informativa Diaria (esta red está integrada por los equipos que informan diariamente). Los pluviómetros que brindan la información diaria son: en

Moa el 1547 ubicado en la Delegación Territorial, el 1695 en la Presa Moa y el 1696 en la Derivadora; en Sagua el 58 en el Tele correo Sagua, el 1585 en el Tele correo Naranjo y el 982 en el Tele correo Calabaza y en Frank País el 601 ubicado en el Acueducto municipal.

El funcionamiento de la Red Informativa Diaria funciona cuando cada operador o especialista ubicados en estos equipos toman la lectura a la 7:00 am y la informan al especialista de la Delegación Territorial Moa y al de la provincia, esta lectura corresponde al día anterior. La información se va registrando en un fichero Excel para los estudios pertinentes. La provincia con cada información que recibe de cada municipio obtiene un parte donde refleja la información más relevante. Este proceso es prácticamente el mismo para los datos referidos a lluvias precipitadas como a los de la situación de los embalses.

1.2 Definiciones fundamentales de Data Warehouse

A partir de mediados de los ochenta, en el entorno empresarial, ha cobrado importancia el concepto de Data Warehouse (DWH por sus siglas en inglés) o almacén-factoría de datos, entendido como la plataforma que concentra toda la información de interés para la organización. Sus fuentes de información son tanto las bases de datos corporativas, como otras fuentes externas dígase hojas de cálculo Excel o algún otro fichero que contenga información.

En el mundo existen numerosas definiciones para el DWH, la más conocida fue propuesta por Inmon (considerado el padre de los Data Warehouse) en 1992, el planteaba que: “Un DWH es una colección de datos orientados a temas, integrados, no-volátiles y variante en el tiempo, organizados para soportar necesidades empresariales”. (Inmon , 1996)

Otra definición acertada fue dada en 1993 por Susan Osterfeldt enfocándolo como “algo que provee dos beneficios empresariales reales, en primer lugar la Integración y acceso de datos, eliminación de una gran cantidad de datos inútiles y no deseados; así como también el procesamiento desde el ambiente operacional clásico”. (Second International Symposium on Information Technologies in Environmental Engineering, 2005) Por otro

lado Ralph Kimball, considerado uno de los grandes exponentes en este campo planteó que: “El Data Warehouse es la unión de todos los Data Mart de una entidad”. (Kimball, y otros) Asumiendo para esta investigación la definición dada por Inmon y la de Kimball que si bien no se contradicen manifiestan las características fundamentales que presentan estos sistemas.

1.2.1 Objetivos y Características de un Data Warehouse

Los principales objetivos de un Data Warehouse son:

- Comprender las necesidades de los usuarios por áreas dentro del negocio.
- Determinar qué decisiones se pueden tomar con la ayuda del DWH.
- Seleccionar un subconjunto del sistema de fuentes de datos que sea el más efectivo y procesable para presentar el DWH.
- Asegurar que los datos sean precisos, correctos y confiables y que mantengan la consistencia.
- Monitorear continuamente la precisión y exactitud de los datos y el contenido de los reportes generados.

Como ya se mencionó, un DWH debe tener cuatro características primarias. Es una colección de datos orientada a un tema, integrada, variable en el tiempo y no volátil, que sea útil para la toma de decisiones.

- Orientada a un tema porque tiene en cuenta los procesos de negocio de la empresa que se deseen priorizar.
- Integrado porque agrupa a todos los sistemas operacionales en un sistema de información con formatos y códigos consistentes.
- Histórico o variante en el tiempo porque los datos se organizan y almacenan en jerarquías en el tiempo, lo que permiten realizar análisis comparativos de estados actuales y de períodos anteriores.
- No volátil ya que se usa principalmente para operaciones de recuperación de información y no para actualizaciones.

1.3 Data Mart

Un Data Mart es una base de datos departamental, la cual se especializa en el almacenamiento de datos de un área específica del negocio. Los DM poseen una estructura de datos óptima para analizar detalladamente la información que en ellos se almacenan desde todas las perspectivas que pueden afectar los procesos del departamento.

Un Data Mart puede obtener información desde los datos almacenados en un Data Warehouse o integrar por sí mismo un compendio de distintas fuentes de información. Es considerado un Data Warehouse con función departamental o regional contando con sus mismas características y brindando sus mismas facilidades, pero está orientado a una sola actividad y no a satisfacer las necesidades de toda la empresa. Por tanto, no se puede pensar en un Data Mart en los términos de un DWH más pequeño, porque no es su tamaño lo que lo define sino su objetivo en la organización. (Kimball, y otros)

1.3.1 Procesos que intervienen en la obtención de un Data Mart

La construcción de un Data Mart es el resultado de un proceso consciente y ordenado por parte de cada organización, en gran número de ocasiones puede implicar cambios en el entorno del negocio tanto por las mejoras que introduce como por la reorganización de los procesos que durante su construcción se llevan a cabo. La aplicación de un Data Mart está orientada a la toma de decisiones, por lo que un buen diseño de su base de datos (BD) favorece el análisis y la recuperación de datos para obtener una ventaja estratégica; todo esto unido al crecimiento de la complejidad de las BD de los sistemas operacionales y el aumento de los requisitos por parte de los clientes que utilizarán los reportes generados a partir del mismo, torna complejo el procedimiento para su implementación, de ahí que sea necesaria la realización de procesos que contribuyan a la creación del mismo, dichos procesos son:

- Proceso de extracción, transformación y carga de datos.
- Proceso de explotación.
- Metadatos.

Proceso de extracción, transformación y carga de datos

Los datos para almacenarlos en un Data Mart necesitan agregarse, analizarse, computarse, procesarse matemáticamente, etc., y en muchos casos también se hace necesario realizar transformaciones específicas, los procesos de extracción, transformación y carga de datos conocidos como ETL (por sus siglas en inglés) recuperan los datos de todos los sistemas operacionales, permitiendo que la información de los mismos pueda transformarse y moverse desde el sistema operacional u otros sistemas al Data Mart. Este constituye uno de los procesos más importantes en el desarrollo del Data Mart, se puede decir que la exactitud de la plataforma de Inteligencia del Negocio (BI por sus siglas en inglés) entera va a depender en gran medida de la calidad de los procesos ETL.

Proceso de extracción y limpieza de datos:

En este proceso se procede a extraer la información de las diferentes fuentes de origen de datos para su posterior limpieza posibilitando de esta manera eliminar redundancias, gestionar la corrección de errores, entre otros problemas que pueden presentar los mismos.

Estas fuentes de origen son las encargadas de alimentar el Data Mart, estando en muchos casos diseñadas para registrar grandes cantidades de transacciones. Una de las fuentes más comunes es la base de datos operacional (OLTP por sus siglas en inglés), estas son bases de datos que como requisito fundamental posee soportar procesos transaccionales presentando algunas de las siguientes características:

- Son pobladas por usuarios finales.
- Se optimizan en función de los procesos transaccionales.
- Se actualizan constantemente.
- Contienen mucha información de detalle.
- Generalmente se encuentran normalizadas hasta segunda o tercera forma normal.

Proceso de transformación:

Luego de ser identificadas las diferentes fuentes de alimentación de datos, es necesario realizar la unificación de los mismos. En este marco se define un único tipo de datos para cada uno de los campos, al que se tendría que llevar la información de las distintas fuentes de alimentación externas.

Proceso de carga:

Una vez realizado el proceso de extracción, limpieza y transformación de datos se procede a cargar los mismos en el almacén de datos. Este proceso es el responsable de cargar la estructura de datos del DWH con:

- Aquellos datos que han sido transformados y que residen en el almacenamiento intermedio.
- Aquellos datos de los OLTP que tienen correspondencia directa con el depósito de datos.

Dependiendo de los requerimientos de la organización, este proceso puede abarcar una amplia variedad de acciones diferentes. En algunos casos las bases de datos sobrescriben la información antigua con nuevos datos.

En la siguiente figura se puede apreciar mejor lo antes descrito (Síntesis del accionar del proceso ETL, y cuál es la relación existente entre sus diversas funciones). Los pasos que se siguen son:

- Se extraen los datos relevantes desde los OLTP.
- Estos datos se depositan en un almacenamiento intermedio.
- Se integran y transforman los datos, para evitar inconsistencias.
- Finalmente los datos depurados son cargados desde el almacenamiento intermedio hasta el DWH. Si existiesen correspondencias directas entre datos de los OLTP y el DWH, se procede también a su respectiva carga.

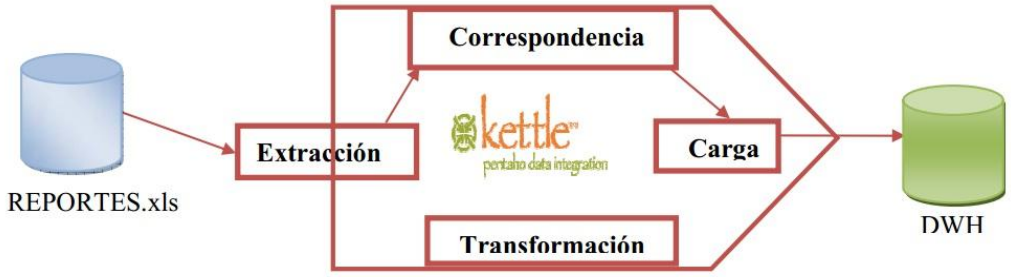


Figura 1: Proceso de extracción transformación y carga.

Proceso de explotación

Estando la información en el almacén de datos se puede pasar al proceso de explotación del Data Mart, que mediante el empleo de diferentes herramientas de consulta realiza la extracción y análisis de la información en los distintos niveles de agrupación que se han definido o que el usuario final desee.

Metadatos

Estos describen los tipos de datos, las definiciones físicas y lógicas de los mismos, las consultas e informes predefinidos, las reglas de validación y negocio, las definiciones de las fuentes de datos, las rutinas de transformación y de proceso. En definitiva, se refieren a cualquier estructura que define un objeto del Data Mart.

1.4 Data Warehouse vs Data Mart

Como ya se ha abordado previamente, el concepto Data Mart es una extensión natural del Data Warehouse estando enfocado a un departamento o área específica del negocio, permitiendo de esta manera un mejor control de la información que se está abarcando. Esta es la principal característica que diferencia a los mismos, aunque existen otros aspectos por los que se pueden realizar comparaciones entre ellos, por ejemplo, costo de diseño e implementación, alcance, conexiones de usuarios, entre otros por solo citar algunos ejemplos.

Debido al alcance de los Data Mart en ocasiones se hace difícil coordinar el flujo de los datos a través de los múltiples departamentos que ellos representan, por otro lado cada departamento tendrá su propia visión de cómo un Data Mart debe lucir, siendo cada Data Mart específico para cada uno de ellos. Por el contrario, un Data Warehouse se

encuentra diseñado en torno a toda la organización en su conjunto, en lugar de ser propiedad de un departamento, será propiedad de la compañía entera. También hay que considerar que la implementación de varios Data Mart para una organización puede acarrear conflictos si presentan diferentes fuentes o diferentes periodos de actualización. Algunas de las diferencias más relevantes entre ambos se presentan en la Tabla 1.

Tabla 1: Comparación entre un Data Warehouse y un Data Mart

	Data Warehouse	**Data Mart**
Costos en diseño e implementación.	Es más costoso de diseñar e implementar.	Es menos costoso de diseñar e implementar.
Distribución de información útil para la toma de decisiones.	No optimiza la distribución de información útil para la toma de decisiones.	Optimiza la distribución de información útil para la toma de decisiones.
Radio de acción.	Se ajusta a varias áreas de procesos.	Se ajusta mucho a las necesidades que tienen un área específica.
Conexiones de usuarios.	Soporta mayor cantidad de usuarios.	Soporta menos usuarios.
Alcance.	Posee un alcance menos limitado.	Posee un alcance más limitado.
Consistencia.	Consistente.	La proliferación de estos puede llevar a la inconsistencia.

Integración a un DWH.	No se integra.	Puede integrarse eventualmente a un DWH.
Obtención de resultados.	Los resultados no se muestran hasta el final del proceso de implementación.	Los resultados pueden observarse en la medida que se termine cada uno de los mismos.

No obstante, por lo costoso que resulta la implementación de un Data Warehouse en cuanto a tiempo y recursos, además de no poder contar con los resultados hasta el final del proceso; muchas organizaciones deciden la construcción paulatina de Data Mart departamentales garantizándole la recuperación de información de forma gradual resultando ventajoso para la organización.

1.5 Modelo

Un modelo es una representación de la realidad que contiene las características generales de algo que se va a realizar. Es la representación en pequeña escala de alguna cosa, aunque también se le conoce como un esquema teórico, generalmente en forma matemática, de un sistema o de una realidad compleja, como la evolución económica de un país, que se elabora para facilitar su comprensión y el estudio de su comportamiento. (Mod13) Puede involucrar tanto planos detallados como planos más generales que ofrecen una visión global del sistema en consideración.

El modelado es común en los proyectos de software exitosos, constituye una técnica de ingeniería probada y bien aceptada, que entre otros factores ayuda a:

- Visualizar a los usuarios el producto final.
- Comprender mejor el sistema.
- Comunicar las ideas a otros.

En el desarrollo de software existen varias formas de enfocar un modelo, ocupando primordial importancia el modelado de la Base de Datos, esta representación se realiza

de forma gráfica, pero un único modelo o vista no es suficiente. Cualquier sistema no trivial se aborda mejor a través de un pequeño conjunto de modelos casi independientes con múltiples puntos de vista, significa tener modelos que se puedan construir y estudiar separadamente, pero que aun así estén interrelacionados, de ahí que a la hora de diseñar una Base de Datos sean generados varios de estos modelos.

1.5.1 Modelo Multidimensional

La modelación dimensional es un nuevo nombre para una técnica antigua que permite hacer simples y comprensibles bases de datos, la cual puede ser visualizada como un "*cubo*" de tres, cuatro, cinco o más dimensiones, donde cualquier punto interior es una intersección de las coordenadas definidas por los ejes del cubo. (Mod13)

Para poder entender la definición presentada así como el modelo multidimensional, se deben comprender algunos conceptos fundamentales:

Un *cubo* es la unidad de representación de la información, equivalente a las tablas de las bases de datos relacionales, las que representan la información como matrices.

A los ejes de la matriz se les llama *dimensiones* representando los criterios de análisis, y a los datos almacenados en la matriz se denominan *medidas* y representan los indicadores o valores a analizar. Las dimensiones caracterizan a una actividad o hecho, permitiendo su análisis posterior en el proceso de toma de decisiones, brindando una perspectiva adicional al hecho dado. Son agrupaciones lógicas de atributos con un significado común y atómico. (Wolf, 2012)

Se llama *hecho* a una operación que se realiza en el negocio, la cual está estrechamente relacionada con el tiempo y es objeto de análisis para la toma de decisiones. También puede verse como un valor numérico que representa una actividad específica casi siempre con cifras que se suman entre sí. La estructura que forman los hechos y las dimensiones puede verse como el plano o la visión desplegada de un cubo. (Wolf, 2012) Ver Figura 2.

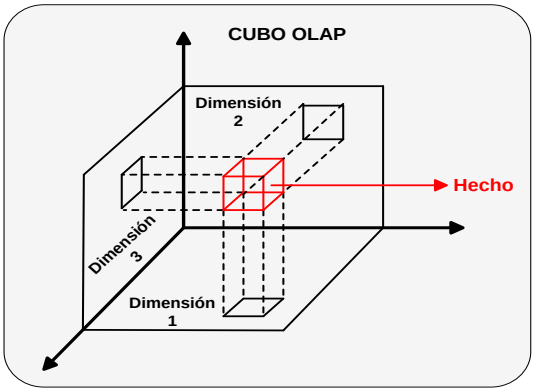


Figura 2: Cubo OLAP

La principal característica del modelo dimensional es su sencillez, permitiéndoles a los usuarios una fácil comprensión de las bases de datos, además de posibilitar al software un recorrido eficiente de sus estructuras.

1.5.2 Componentes del Modelo Multidimensional

Para una correcta compresión y construcción de un modelo multidimensional es necesario conocer cuáles son los elementos que lo componen y lo que significa cada uno de ellos.

Tablas de dimensiones

Las dimensiones organizan los datos en función de un área de interés para los usuarios. Cada dimensión describe un aspecto del negocio y proporciona el acceso intuitivo y simple a datos. Una dimensión provee al usuario de un gran número de combinaciones e intersecciones para analizar datos.

Cada nivel de una dimensión debe corresponderse con una columna en la tabla de la dimensión. Los niveles se ordenan por grado de detalle y se organizan en una estructura jerárquica. Cada nivel contiene miembros, los miembros son los valores de la columna que define el nivel.

Todos los elementos que componen una dimensión están enmarcados en una determinada estructura jerárquica excepto los que conformaran las propiedades de un determinado elemento.

Tablas de hechos

Las tablas de hechos son las tablas primarias del modelo dimensional conteniendo los valores del negocio que se desea analizar así como cada una de las llaves de las dimensiones involucradas en el mismo. Todas estas columnas son valores numéricos calculables durante el proceso extracción, transformación y carga de datos.

Medidas o métricas

El modelo dimensional divide el mundo de los datos en dos grandes tipos: las medidas y las dimensiones de estas medidas. Las medidas, siempre son numéricas, se almacenan en las tablas de hechos y las dimensiones que son textuales se almacenan en las tablas de dimensiones. Las medidas son los valores de datos que se analizan.

Una medida es una columna cuantitativa, numérica, en la tabla de hechos. Las medidas representan los valores que son analizados constituyendo:

- Valores que permiten analizar los hechos.
- Bases a partir de las cuales el usuario puede realizar cálculos.

Las medidas pueden clasificarse en:

- Naturales
- Calculadas

Medidas naturales: Son el resultado de la aplicación de operaciones matemáticas sencillas a un solo campo existente en la tabla de hechos. Cuando se define una medida se debe tener en cuenta cuál será la forma de agregación (agrupación de la misma) al subir por la estructura dimensional.

Medidas calculadas: son el resultado de las diferentes operaciones que se pueden realizar con los valores de las medidas naturales. Debe tenerse en cuenta que estas medidas calculadas se pueden obtener durante el proceso ETL y después del mismo.

La decisión de cuando usar cada cual está en dependencia de los requisitos definidos por el cliente.

En sentido general la expresión medidas calculadas es muy amplia y engloba a cualquier manipulación de las medidas naturales que faciliten el análisis de los hechos, pero por lo general en una medida calculada puede haber:

- Cálculos Matemáticos
- Expresiones condicionales
- Alertas

Estos tres tipos (cálculos, condiciones y alertas) usualmente pueden existir juntos dentro de la misma medida calculada. (Alonso, y otros, 2005)

1.5.3 Ventajas del Modelo Multidimensional

- Por presentar una visión casi desnormalizada en su totalidad, agiliza el proceso de recuperación de datos disminuyendo en gran medida el tiempo de respuesta del sistema.
- Agrupa los datos en pocas dimensiones, cada una de las cuales se convierte en una tabla en la BD.
- Se torna en un modelo simple consistente de pocas tablas.
- No es simétrico, pudiéndose determinar cuál tabla es más importante.
- Se puede apreciar con facilidad qué/cuáles tablas del modelo contienen medidas numéricas.
- Son muy sencillas para las personas (usuarios finales o diseñadores) visualizar y conservar en sus mentes todas las tablas que componen el modelo.
- Para llegar a la información que el usuario desea obtener de una consulta donde se encuentren involucradas dos tablas solo existe un único camino de posibles conexiones entre ellas.

1.6 Antecedentes del uso de la tecnología Data Warehouse para el trabajo con datos hidrológicos

Al ir en ascenso en el mundo la tecnología del Warehouse va incrementándose su uso en múltiples sectores para diversos fines. En materia de recursos hidráulicos son varios

ya los sistemas desarrollados dentro y fuera del país. Para el desarrollo de esta investigación se tuvo en cuenta los siguientes:

- ✓ *Cuba*, "Sistema automatizado de alerta temprana ante el peligro de inundaciones", esta propuesta presentada en el 2012, es una herramienta para la asistencia a la toma de decisiones que tiene como objetivos fundamentales: preservar el recurso agua, combatir la sequía y la prevención y control de inundaciones de una cuenca hidrográfica. El sistema fue aplicado a la cuenca San Pedro, Camagüey, para un evento extremo particular obteniendo buenos resultados. Posee cuatro componentes principales: Modelo de simulación (HEC-HMS y HEC-RAS), Sistema de Información Geográfica empleando la herramienta ArcGis 9.3, Adquisición y Supervisión de datos en Tiempo Real y Diferido a través de un sistema SCADA empleando la herramienta Wizcon 8.3, el cual permitirá obtener información de variables como: lámina de lluvia, nivel (cota de agua), escurrimiento e intensidad de la lluvia y por último, la Base de datos histórica actualizada cuyo gestor de base de datos empleado es el sistema PostgreSQL. (Garrido, 2012)
- ✓ *Ecuador*, "Sistema de información del Instituto Nacional de Meteorología e Hidrología (INAMHI)" conformado por 300 equipos de medición. Los datos medidos por 32 equipos son enviados 5 veces al día por radio mediante comunicación verbal. El resto son recibidos en diferido por correo convencional. Se dispone de unos 20 años de datos correspondientes a los recibidos por la red de radiofonía. Mediante otros métodos se tienen datos desde 1965. La información se almacena con distintas agregaciones temporales (diaria, mensual, por décadas). Cuenta con una página web de acceso público. La base de datos fue desarrollada con Oracle y se emplea el SIG ArcView de ESRI para la visualización y análisis de la información. (Molina, 2006)
- ✓ *España*, desde el año 1983 a través de la Dirección General del Agua (DGA) del Ministerio de Medio Ambiente y Medio Rural y Marino ha desarrollado el "Sistema automático de información hidrológica (SAIH)", el cual es un sistema de información en tiempo real para la gestión de recursos hídricos y entre otros,

para el control de inundaciones, lo que permite el seguimiento en tiempo real de variables hidrometeorológicas, que se almacenan en una base de datos en MySQL, así como visualización de series de datos temporales, mapas de la red de estaciones de medición meteorológica, presentación de mapas temáticos en tiempo real, así como acceso al libro digital del agua donde se presentan gráficos y mapas exportables. (Mosteiro, 2009)

- ✓ *Estados Unidos de América,* para el año 2010, "The Consortium of Universities for the Advancement of Hydrologic Sciences (CUAHSI)", desarrolló el "Hydrologic information systems (CUASHI-HIS)" el cual ofrece varias funciones incluyendo entre ellas la consulta de series de datos, visualización de mapas basado en descarga de datos, construcción de gráficos, edición, impresión y modelado con series de datos y exportación a determinados formatos de modelos específicos y la vinculación con los sistemas integrados de modelación. La base de datos emplea el gestor ODM SQL Server e integra el SIG ArcGIS. ("Introducing the Open Source CUAHSI Hydrologic Information System Desktop Application (HIS Desktop)", 2009).

Después de haberse realizado un estudio de estos, se concluye que los mismos no presentan las funcionalidades requeridas para el caso que ocupa esta investigación y en muchos la selección de herramientas privativas dificultan su uso en la Empresa Aprovechamiento de los Recursos Hidráulicos Holguín (EAHHLG), no obstante estas investigaciones sirven de apoyo para el diseño del mercado de datos que se propone en esta investigación.

1.7 Metodologías a emplear

En la búsqueda de la metodología para realizar este trabajo se propone como idea principal, comprender cada paso que se realizará, para mejorar el tener que seguir un método al pie de la letra sin saber exactamente qué se está haciendo, ni por qué.

Existen varias metodologías empleadas en la construcción de un almacén de datos, no es posible decir cuál es la mejor o si alguna es mejor que otra, su selección está en dependencia de las características del negocio y de la organización para la cual se va a

construir, entre todas, las seleccionadas para el estudio son las siguientes: Metodología HEFESTO y DATEC.

1.7.1 Metodología para el diseño de la arquitectura

Metodología de Ralph Kimball

Según Kimball: "El Data Warehouse es la unión de todos los Data Mart de una entidad", siendo además una copia de los datos transaccionales, estructurados de una forma especial para realizar su análisis, de acuerdo al modelo dimensional no normalizado. Este enfoque también es conocido como Bottom-up. Esta característica le hace más flexible y sencillo de implementar, pues se puede construir un Data Mart como primer elemento del sistema de análisis, y luego ir añadiendo otros que compartan las dimensiones ya definidas o incluyan otras nuevas. En este sistema, los procesos ETL extraen la información de los sistemas operacionales y los procesan realizando posteriormente el llenado de cada uno de los Data Mart de una forma individual, aunque siempre respetando la estandarización de las dimensiones. Este enfoque es eficaz y conduce a una solución completa en un corto periodo de tiempo. (Kimball, y otros) Ver Figura 6.

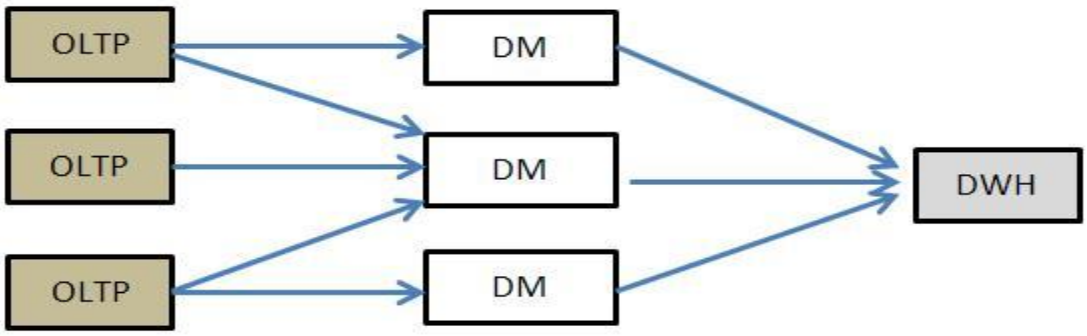


Figura 3: Enfoque Kimball

Metodología de Bill Inmon

Inmon ve la necesidad de transferir la información de los diferentes OLTP de las organizaciones a un DWH centralizado, donde los datos puedan ser utilizados para su posterior análisis. La información ha de estar a los máximos niveles de detalle. Los

DWH departamentales o Data Marts son tratados como subconjuntos de este DWH corporativo, y son construidos para cubrir las necesidades individuales de análisis de cada departamento, siempre a partir de este DWH central.

El enfoque Inmon también se referencia normalmente como **Top-Down**. Los datos son extraídos de los sistemas operacionales por los procesos ETL donde son validados y consolidados para su posterior almacenamiento en el DWH, donde además existen los llamados metadatos que documentan de una forma clara y precisa el contenido del DWH. Una vez realizado este proceso, los procesos de actualización de los Data Mart departamentales obtienen la información de él, y con las consiguientes Transformaciones, organizan los datos en las estructuras particulares requeridas por cada uno de ellos. Al tener este enfoque global, es más difícil de aplicar en un proyecto sencillo (pues se intenta abordar el "todo", a partir del cual luego se irá al "detalle"). (Inmon , 1996) Ver figura 7.

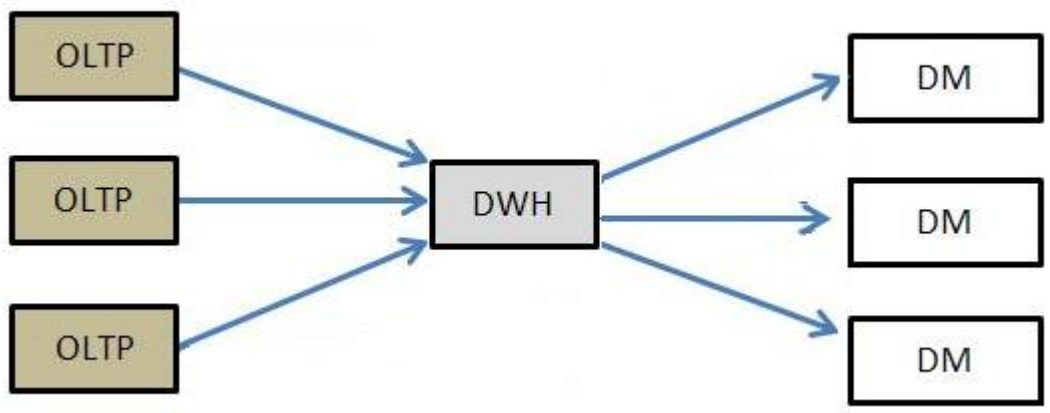


Figura 4: Enfoque Inmon

1.7.2 Metodología para el desarrollo del DWH

Metodología HEFESTO

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

utilizadas por su fácil implementación y aporte práctico, aunque no propone de forma explícita los artefactos y entregables a generar en cada fase.

Esta metodología cuenta con las siguientes características:

- Se basa en los requerimientos del usuario, por lo cual su estructura es capaz de adaptarse con facilidad y rapidez ante los cambios en el negocio.
- Reduce la resistencia al cambio, ya que involucra al usuario final en cada etapa para que tome decisiones respecto al comportamiento y funciones del DWH.
- Se aplica tanto para el desarrollo de Data Mart como para DWH. (Bernau Ricardo, 2010)

Pasos de la Metodología HEFESTO

Paso # 1: Análisis de Requisitos.

a) Identificar Preguntas.
b) Identificar Indicadores y Perspectivas.
c) Modelo Conceptual.

Paso # 2: Análisis de OLTP.

a) Establecer correspondencias con los requerimientos.
b) Seleccionar los campos que integrarán cada perspectiva. Nivel de granularidad.

Paso # 3: Elaboración del Modelo Lógico de la estructura del DWH.

a) Diseñar tablas de dimensiones.
b) Diseñar tablas de hechos.
c) Realizar Uniones.
d) Determinar Jerarquías.

Paso # 4: Procesos ETL, Limpieza de datos y sentencias SQL

Metodología DATEC

Esta metodología está basada fundamentalmente en el enfoque Kimball, consta de 5 fases y un total de 43 artefactos. En esta se definen además los hitos de desarrollo así

como los roles y sus responsabilidades en el mismo definiendo también las herramientas a emplear en cada una de sus fases. Su aporte radica en la integración de algunas prácticas de RUP con este enfoque. Si bien esto ayuda a obtener una documentación amplia para el proyecto, implica el empleo de mayores recursos como tiempo, esfuerzo y personas en la generación de estos artefactos para cada una de sus fases.

1.7.3 Justificación de las metodologías a utilizar

En cuanto a estas metodologías se puede decir que el enfoque Inmon es más apropiado para sistemas complejos, en los que se quiera además asegurar su perdurabilidad y consistencia aunque cambien los procesos de negocio en la organización. Aunque, para pequeños proyectos, en los que además se haga necesario asegurar la usabilidad de los usuarios con un sistema fácil de entender y garantizar un rápido desarrollo en la solución, el enfoque Kimball es más apropiado, siendo el utilizado para el caso que ocupa esta investigación.

En relación a las metodologías de desarrollo, si bien la metodología de DATEC posibilita organizar de forma eficaz el proceso de creación de los DWH; su empleo implica la creación de gran cantidad de artefactos para la documentación haciendo que el proceso de implementación del Data Mart se torne engorroso y tedioso. Por esta entre otras razones se selecciona la metodología HEFESTO, por su aporte práctico es fácil de entender e implementar.

1.8 Herramientas para la construcción de un Data Warehouse

Las herramientas para la construcción del Data Warehouse se dividieron en 3 grupos principales, los relacionados con los sistemas gestores de bases de datos, herramientas de integración de datos, y herramientas para diseñar cubos OLAP, realizando un estudio basados en sus principales características, su usabilidad, así como la documentación sobre el uso de las mismas, resultando escogidas las siguientes:

1.8.1 Sistemas gestores de Bases de Datos: **PostgreSQL**

PostgreSQL es un sistema de base de datos objeto-relacional de código abierto bajo licencia GPL. Cuenta con más de 15 años de desarrollo activo y una arquitectura probada que se ha ganado una sólida reputación de fiabilidad, integridad y corrección de datos. Se ejecuta en todos los principales sistemas operativos, como son Linux, UNIX, Mac OS X y Windows. Tiene soporte completo para claves foráneas, uniones, vistas, disparadores y procedimientos almacenados. Incluye la mayoría de las sentencias SQL. También soporta almacenamiento de objetos binarios grandes, como imágenes, sonidos o videos. Cuenta con interfaces nativas de programación para C/C++, Java, .Net, Python, Ruby, entre otros y la documentación excepcional.

Algunas de sus características principales son:

- Instalación ilimitada.
- Mejor soporte que los proveedores comerciales.
- Ahorros considerables en costos de operación.
- Estabilidad y confiabilidad.
- Extensible.
- Multiplataforma.
- Herramientas gráficas de diseño y administración de bases de datos.(Sit 2011)

1.8.2 Herramienta de integración de datos: KETTLE

Utiliza como entorno gráfico la herramienta de diseño (Spoon) basada en SWT(*Conjunto de herramientas flash de código fuente abierto, para Java, diseñado para proporcionar un acceso eficiente y portátil para la interfaz de usuario de los sistemas operativos en los que se aplica*) el entorno para su ejecución es desde la herramienta de diseño, o desde línea de comandos con las utilidades Pan y Kitchen.

- SPOON: permite diseñar de forma gráfica la transformación ETL.
- PAN: ejecuta un conjunto de Transformaciones diseñadas con SPOON, conocidas como trabajos (jobs), creando dependencias entre dichas

Transformaciones.

- CHEF: permite, mediante una interfaz gráfica, diseñar la carga de datos incluyendo un control de estado de los trabajos.
- KITCHEN: permite ejecutar los trabajos diseñados con Chef. (ETL13)

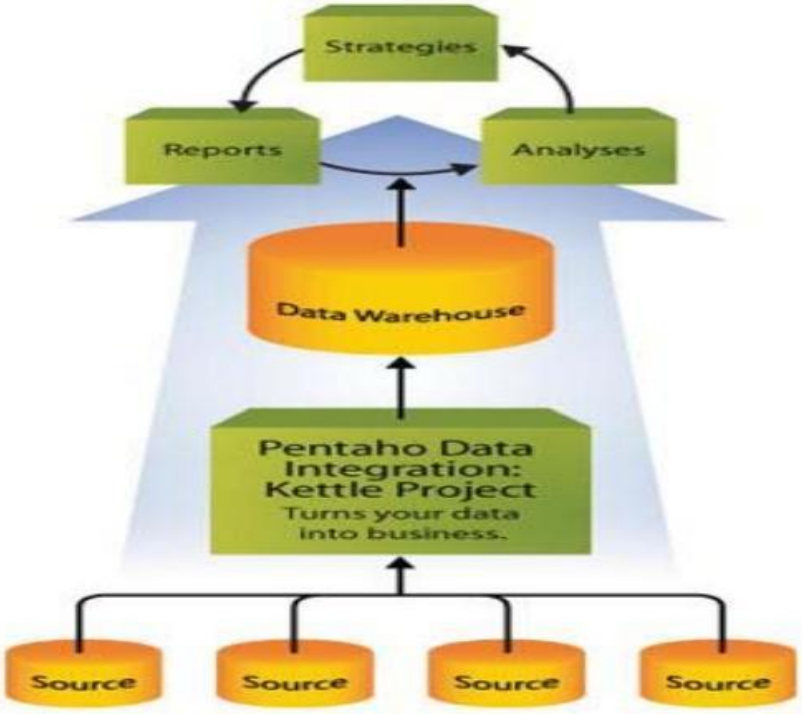


Figura 5: Representación de la arquitectura de Kettle.

1.8.3 Herramientas para diseñar cubos OLAP: Mondrian

Mondrian es el motor OLAP integrado en la suite de Business Intelligence Open Source Pentaho. Es un proyecto Open Source y actúa bajo la licencia Mozilla Public License (MPL).

Schema workbench es un entorno visual para el desarrollo y prueba de cubos OLAP. Esta herramienta se utiliza para la creación de los archivos XML que se usan para la construcción de los cubos. Además permite la ejecución de consultas MDX(*Las expresiones multidimensionales (MDX es el acrónimo de MultiDimensional eXpressions) son un lenguaje de consulta para bases de datos multidimensionales sobre cubos*

OLAP, se utiliza en Business Intelligence para generar reportes para la toma de decisiones basados en datos históricos, con la posibilidad de cambiar la estructura, o permitiendo rotar el cubo) contra el esquema y la base de datos. (Bravo Martínez)

Conclusiones del capítulo

Una vez finalizado el presente capítulo llegamos a las siguientes conclusiones:

- La metodología de arquitectura seleccionada es la correspondiente con el enfoque de Ralph Kimball por adecuarse a las características del negocio y las ventajas que presenta la misma.
- La metodología de desarrollo seleccionada es HEFESTO por su aplicación sencilla para la implementación del DM.
- Después del estudio de las diferentes herramientas utilizadas para la construcción del Data Mart se seleccionaron por sus características: Como sistema gestor de base de datos al PostgreSQL, para el proceso de extracción, transformación y carga de los datos a la herramienta Kettle perteneciente a la suite del Pentaho, como herramienta para el diseño del cubo OLAP al Schema Workbench (Mondrian).

Capítulo 2. ANÁLISIS y DISEÑO DEL DATA MART

Introducción

En el presente capítulo se realizará el análisis, diseño del Data Mart, en correspondencia con las etapas del proceso de desarrollo planteado en la metodología seleccionada en el capítulo anterior. Se describe paso a paso cada una de las fases de dicha metodología que incluye desde el análisis de los requerimientos, pasando por la identificación de indicadores y perspectivas, dimensiones y medidas, así como la definición de procesos de extracción, transformación y carga de datos.

2.1 Análisis de los requerimientos

El análisis de los requerimientos de los diferentes usuarios, es el punto de partida de esta metodología, ya que ellos son los que deben, en cierto modo, guiar la investigación hacia un desarrollo que refleje claramente lo que se espera del depósito de datos, en relación a sus funciones y cualidades.

2.1.2. Identificar preguntas.

El objetivo principal de este paso, es obtener e identificar las necesidades de información clave de alto nivel, que es esencial para llevar a cabo las metas y estrategias de la empresa

- ✓ Lluvia promedio en mm y % en un municipio en una fecha dada.
- ✓ Acumulado de lluvias en mm y % en un municipio en una fecha dada.
- ✓ Promedio histórico en mm en un municipio en una fecha dada.
- ✓ Lluvias más significativas en un municipio en una fecha dada.
- ✓ Volumen total de un embalse en un municipio en una fecha dada.
- ✓ Volumen muerto de un embalse en un municipio en una fecha dada.
- ✓ Volumen disponible de un embalse de un municipio en una fecha dada.
- ✓ Por ciento de llenado de un embalse en una fecha dada.
- ✓ Consumo de energía en la zona de explotación Moa para una fecha dada.
- ✓ Capacidad de agua embalsada en un momento dado.

- ✓ Capacidad de agua embalsada anteriormente en un embalse específico en una fecha dada
- ✓ Agua diaria disponible de un embalse
- ✓ Incremento de

2.1.3. Identificar perspectivas e indicadores

Luego de haber establecido las preguntas, se debe proceder a su descomposición para descubrir los indicadores que se utilizarán y las perspectivas de análisis que intervendrán. Para cada uno se puede utilizar como indicador el representativo numérico cantidad, total.

Para una correcta implementación del DM es necesario realizar un detallado estudio de las perspectivas a indicadores necesarios para el estudio. Para ello, se debe tener en cuenta que los indicadores, para que sean realmente efectivos son, en general, valores numéricos y representan lo que se desea analizar concretamente, por ejemplo: saldos, promedios, cantidades, sumatorias, fórmulas, etc. En cambio, las perspectivas se refieren a los objetos mediante los cuales se quiere examinar los indicadores, con el fin de responder a las preguntas planteadas, por ejemplo: clientes, proveedores, sucursales, países, productos, rubros, etc. Cabe destacar, que el Tiempo es muy comúnmente una perspectiva.

Perspectivas Identificadas:

- ✓ Provincia
- ✓ Municipio
- ✓ Embalse
- ✓ Estación
- ✓ Plan energético
- ✓ Tiempo

Indicadores Identificados:

1. Lluvia promedio en mm y %
2. Acumulado de lluvias en mm y %
3. Promedio histórico en mm
4. Lluvias más significativas
5. Volumen total
6. Volumen muerto

7. Volumen disponible
8. Por ciento de llenado
9. Consumo de energía
10. Capacidad de agua embalsada
11. Capacidad de agua embalsada anterior
12. Agua diaria disponible
13. Incremento

2.1.4 Modelo conceptual

Luego que identificamos las perspectivas y los indicadores, se realizó el modelo conceptual para observar con claridad cuáles son los alcances del proyecto, para luego poder trabajar sobre ellos, además al poseer un alto nivel de definición de los datos, permite que pueda ser presentado ante los usuarios y explicado con facilidad.

Figura 6. Modelo conceptual de las lluvias precipitadas

Figura 7. Modelo conceptual de la situación de los embalses

2.2 Análisis de los OLTP.

El objetivo de este análisis, es el de examinar los OLTP disponibles que contengan la información requerida, como así también sus características, para poder identificar las correspondencias entre el modelo conceptual y las fuentes de datos. La idea es, que todos los elementos del modelo conceptual estén correspondidos en los OLTP.

2.2.1 Determinación de indicadores.

En este paso se explican cómo se calcularán los indicadores, definiendo los siguientes conceptos para cada uno de ellos:

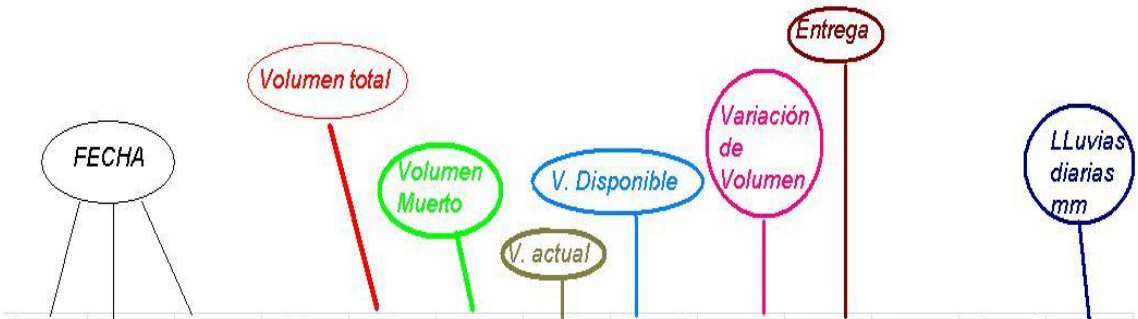
- ✓ *"Lluvia promedio":*
  - *Hechos: Lluvia promedio.*
  - *Función de sumarización: AVG.*
- ✓ *" Acumulado de lluvias del mes":*
  - *Hechos: Acumulado de lluvias del mes.*

- *Función de sumarización: SUM.*

- ✓ *" Promedio Histórico del mes":*
  - *Hechos: Promedio Histórico del mes.*
  - *Función de sumarización: AVG.*
- ✓ *" Acumulado de lluvias en un municipio":*
  - *Hechos: Acumulado de lluvias en un municipio.*
  - *Función de sumarización: SUM.*
- ✓ *"Promedio histórico del municipio":*
  - *Hechos: Promedio histórico del municipio.*
  - *Función de sumarización: AVG.*
- ✓ *"Lluvias más significativas":*
  - *Hechos: Lluvias más significativas.*
  - *Función de sumarización: SUM.*
- ✓ *"Volumen total":*
  - *Hechos: Volumen total.*
  - *Función de sumarización: SUM.*
- ✓ *"Volumen muerto":*
  - *Hechos: Volumen muerto.*
  - *Función de sumarización: SUM.*
- ✓ *"Volumen disponible":*
  - *Hechos: Volumen disponible.*
  - *Función de sumarización: SUM.*
- ✓ *"Por ciento de llenado":*
  - *Hechos: Por ciento de llenado.*
  - *Función de sumarización: AVG.*
- ✓ *"Total de lluvias":*
  - *Hechos: Total de lluvias.*
  - *Función de sumarización: SUM.*
- ✓ *"Consumo de energía":*
  - *Hechos: (Consumo Derivadora)+ (Consumo UEB Moa).*
  - *Función de sumarización: SUM.*

2.2.2 Establecer correspondencias.

El objetivo de este paso, es el de examinar los OLTP disponibles que contengan la información requerida, como así también sus características, para poder identificar las correspondencias entre el modelo conceptual y las fuentes de datos. Un ejemplo se puede observar en la siguiente figura:



día	mes	año	Embalse	Vtotal	Vmuerto	Vactual	Vdisponible	V.Variacion	Entrega gasti	EntregaV	Cobertura di	lluviasmm
4	2	2013	Cacoyugüin	5.620	0.250	5.237	4.987	-0.018	93.2	0.220	19008	203
4	2	2013	Güirabo	15.200	0.800	7.000	6.200	-0.025	46.1	0.446	38549	195
4	2	2013	Gibara	65.600	0.600	43.884	43.284	-0.048	66.9	0.477	41213	322
4	2	2013	San Andrés	6.700	1.080	4.224	3.144	-0.043	63.0	0.546	47200	187
4	2	2013	Tres Palmas	6.630	0.105	6.630	6.525	0.000	100.0	0.000	0	-
4	2	2013	Santa Clara	21.500	1.000	21.458	20.458	0.000	99.8	0.046	3974	-
4	2	2013	Magueyal	12.781	0.500	6.900	6.400	0.000	54.0	0.020	1750	409
4	2	2013	Colorado	38.000	1.000	38.000	37.000	0.000	100.0	0.288	24855	289
4	2	2013	Der Colorado	0.310	0.168	0.310	0.142	0.000	100.0	-	-	-
4	2	2013	Las Lajas	4.847	0.080	2.645	2.565	-0.023	54.6	0.290	25058	-
4	2	2013	Santa Inés	3.080	0.130	1.590	1.460	-0.012	51.6	0.004	346	-
4	2	2013	Tacajó	12.000	1.000	11.769	10.769	-0.013	98.1	0.140	12107	859
4	2	2013	Birán	30.600	3.750	25.554	21.804	-0.055	83.5	0.400	34560	487
4	2	2013	Nipe	112.200	46.400	110.396	63.996	0.164	98.4	0.783	67622	642

Figura 8. Representación en los sistemas operacionales de los Embalses.

Nota: Para el caso de las restantes correspondencias referirse a los anexos.

2.2.3 Nivel de granularidad.

Una vez que se han establecido las relaciones con los OLTP, se deben seleccionar los campos que contendrá cada perspectiva, ya que será a través de estos por los que se examinarán y filtrarán los indicadores.

- ✓ Perspectiva: Fecha
  - Id_ tiempo
  - Año
  - Mes
  - Día
- ✓ Perspectiva: Estación
  - Id_estación
  - Descripción
  - Estación
  - Altura
  - Norte

- Este

- ✓ Perspectiva: Municipio
  - Id_ municipio
  - Id_provincia
  - Nombre
- ✓ Perspectiva: Embalse
  - Id_embalse
  - Nombre
  - lectura
  - consumo_kw
  - nivel
  - entrega
  - gasto
  - entrada de agua
  - salida de agua
- ✓ Perspectiva: Plan energético
  - Id_plan energetico
  - Plan año
  - Plan mes

2.3 Modelo lógico del DWH

A continuación, se confeccionará el modelo lógico de la estructura del DWH, teniendo como base el modelo conceptual que ya ha sido creado. Para ello, primero se definirá el tipo de modelo que se utilizará y luego se llevarán a cabo las acciones propias al caso, para diseñar las tablas de dimensiones y de hechos. Finalmente, se realizarán las uniones pertinentes entre estas tablas.

2.3.1 Tipo de modelo lógico del DWH: Estilo constelación de hechos.

El estilo constelación de hechos, referencia las situaciones en que un único modelo multidimensional posee múltiples hechos, y por lo tanto, crea múltiples estilos estrellas, básicamente esta estructura es requerida cuando los hechos no comparten todas las dimensiones.

Para cada estilo estrella o 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 a que contiene múltiples tablas de hechos. Con esta

solución las tablas de dimensiones pueden estar compartidas para más de una tabla de hechos.

El estilo de constelación de hechos posee mucha flexibilidad y esta es su gran virtud; sin embargo, el problema es que cuando el número de las tablas vinculadas aumenta, la arquitectura puede llegar a ser muy compleja y difícil de mantener.

En un estilo de constelación de hechos, las distintas tablas de hechos están asignadas a las dimensiones relevantes para cada uno de ellos. Esto puede ser útil cuando los hechos están asignados a un nivel de una dimensión y los otros hechos a otro nivel de detalle de otra dimensión

2.3.2 Tablas de dimensiones.

En este paso se deben diseñar las tablas de dimensiones que formaran parte del DWH. Cada perspectiva definida en el modelo conceptual constituirá una tabla de dimensión.

Figura 9. Representación de las tablas Dimensión.

2.3.3 Tablas hechos.

En este paso, se definirán las tablas de hechos, que son las que contendrán los hechos a través de los cuales se construirán los indicadores de estudio.

Figura 10. Representación de las tablas Hechos.

2.3.4 Uniones.

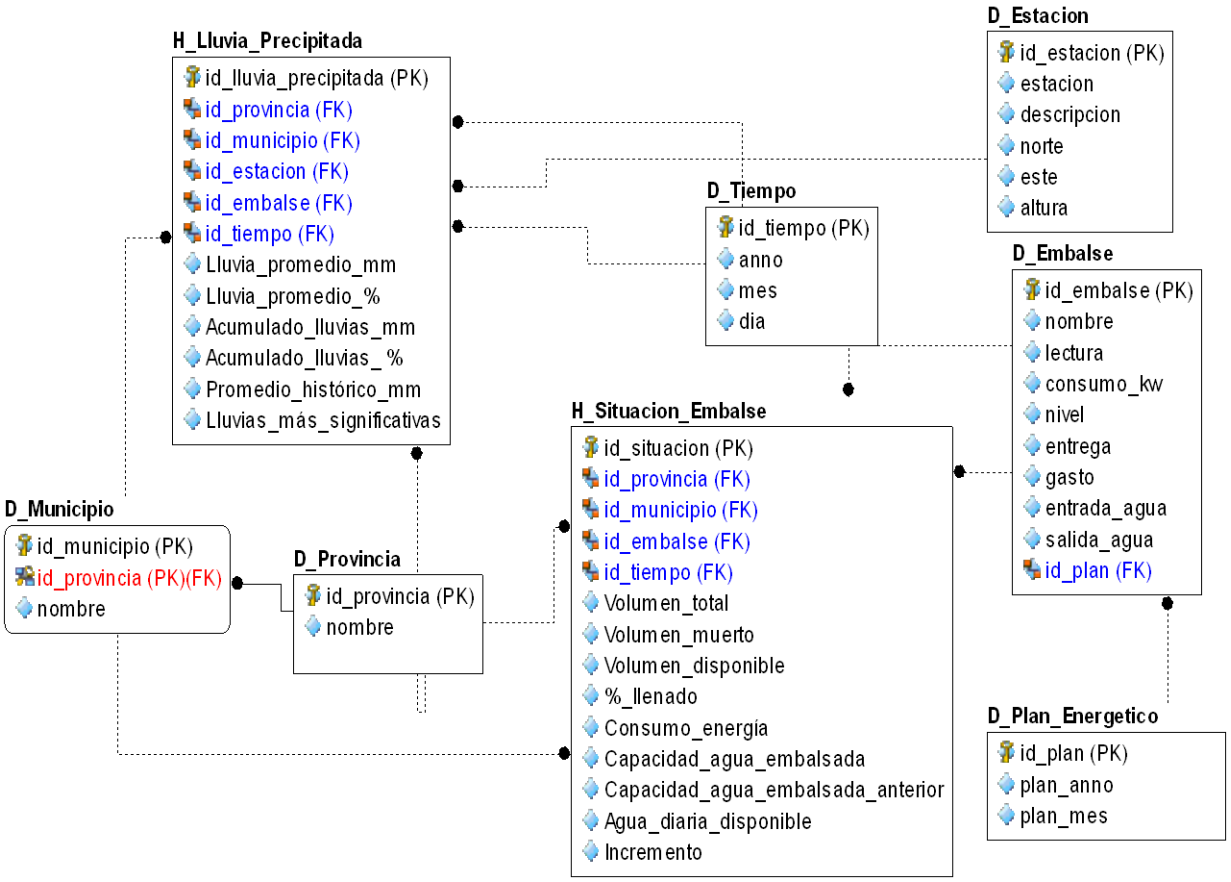


Figura 11: Modelo lógico del Data Mart.

Conclusiones del capítulo

La identificación y definición de los indicadores, perspectivas y hechos constituyeron la base a partir de la cual fue posible realizar el diseño del Data Mart. De igual forma, el proceso de análisis de los OLTP posibilitó organizar y limpiar los datos que van desde las fuentes de origen hacia la destino, logrando así una alta confiabilidad en los mismos, empleando además el modelo constelación de hechos para realizar varios tipos de análisis en el Data Mart con el fin de responder a las necesidades requeridas por el cliente.

Capítulo 3 INTEGRACIÓN DE DATOS.

Introducción

En este capítulo se desarrollara el último paso de la metodología seleccionada. Una vez construido el modelo lógico, se deberá proceder a poblarlo con datos, utilizando técnicas de limpieza y calidad de datos, procesos ETL. Existen varios software que facilitan estas tareas, en esta investigación se realizó el proceso de extracción, transformación y carga de los datos con la herramienta Pentaho Data Integration versión 4.2.1.

3.1 Extracción y Transformación

Al comenzar a trabajar con la herramienta se definió una conexión (Ver figura12) con el Sistema Gestor de Base de Datos PostgreSQL utilizada en todas las transformaciones, esta conexión es recomendable probarla antes de comenzarla a utilizar (Ver figura 12).

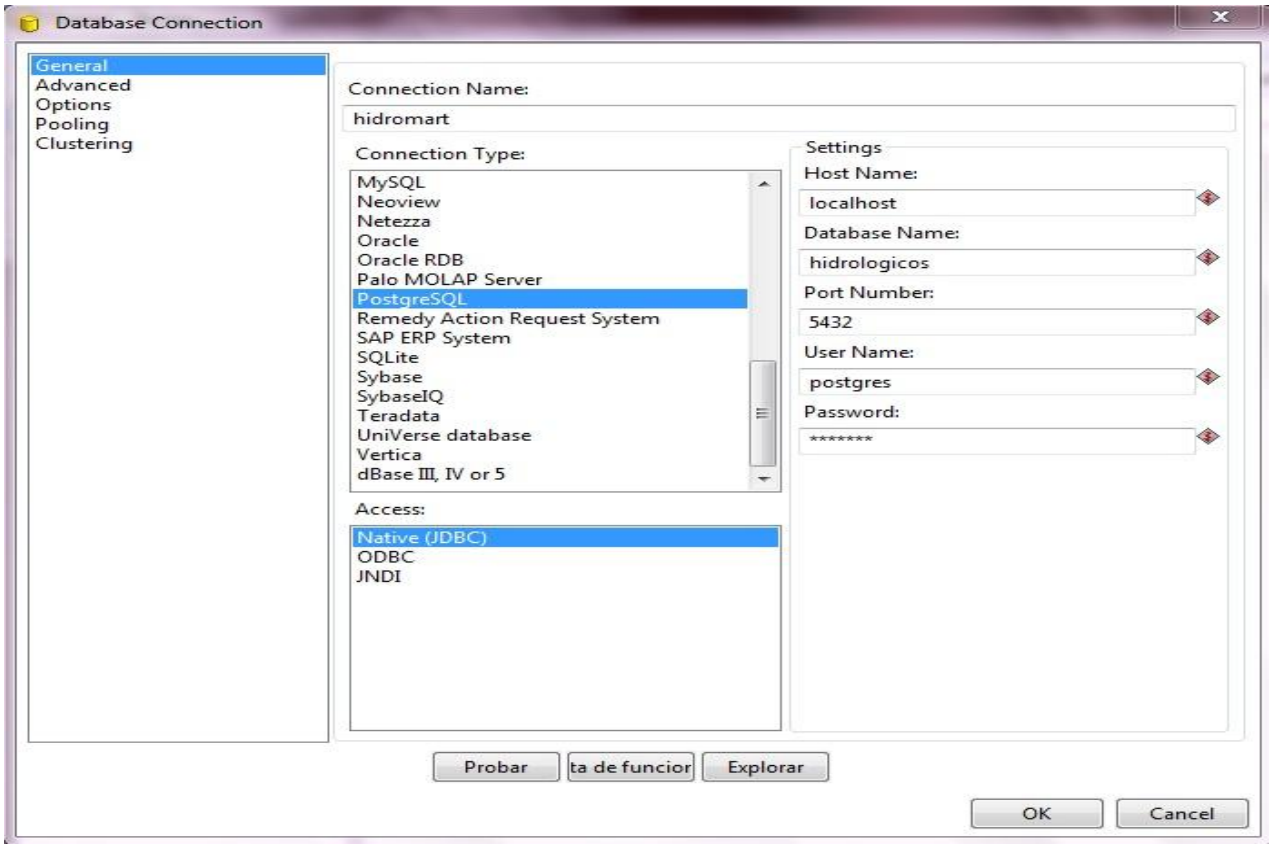


Figura 12: Conexión Spoon

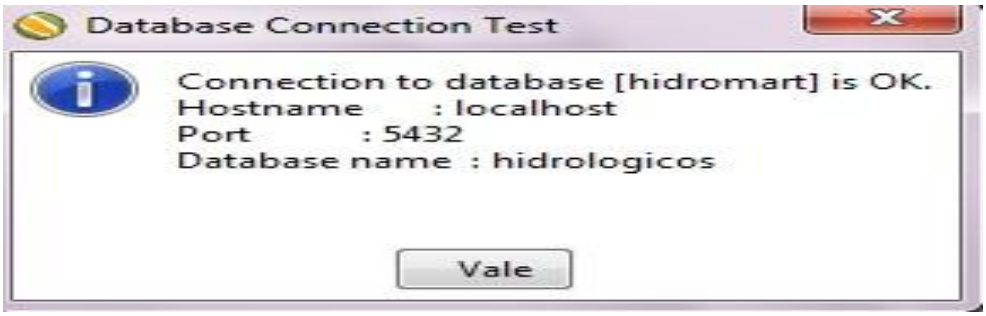


Figura 13: Prueba de Conexión Spoon

Lo primero que se realiza son las transformaciones donde el primer paso es la extracción de los atributos necesarios de los sistemas operacionales para llenar las tablas dimensiones (Ver figura 14)

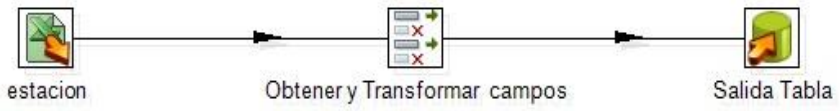


Figura 14: transformación Estación.

Con el primer componente se extraen del fichero en Microsoft Office Excel los datos a cargar, este paso se edita (Ver figura15, 16, y17) de forma tal que se defina qué hojas de cálculo serán extraídas para el caso en particular que se va a realizar.

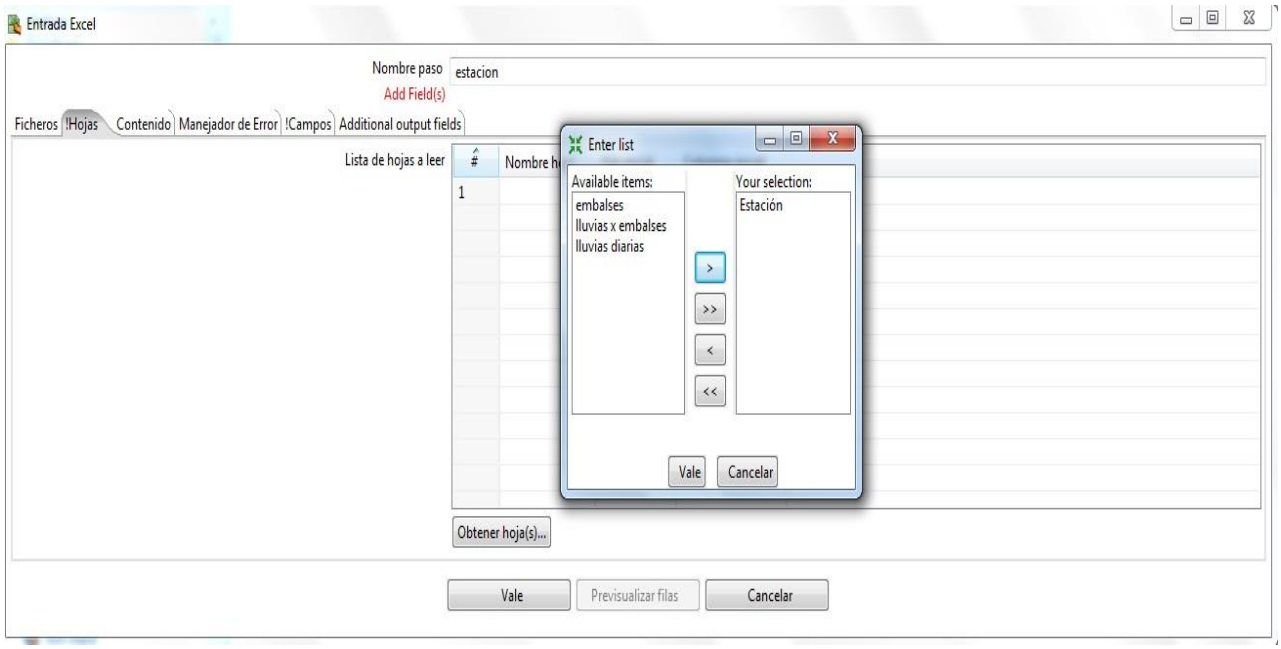


Figura: 15 Componente entrada Excel: obtener hojas de cálculo.

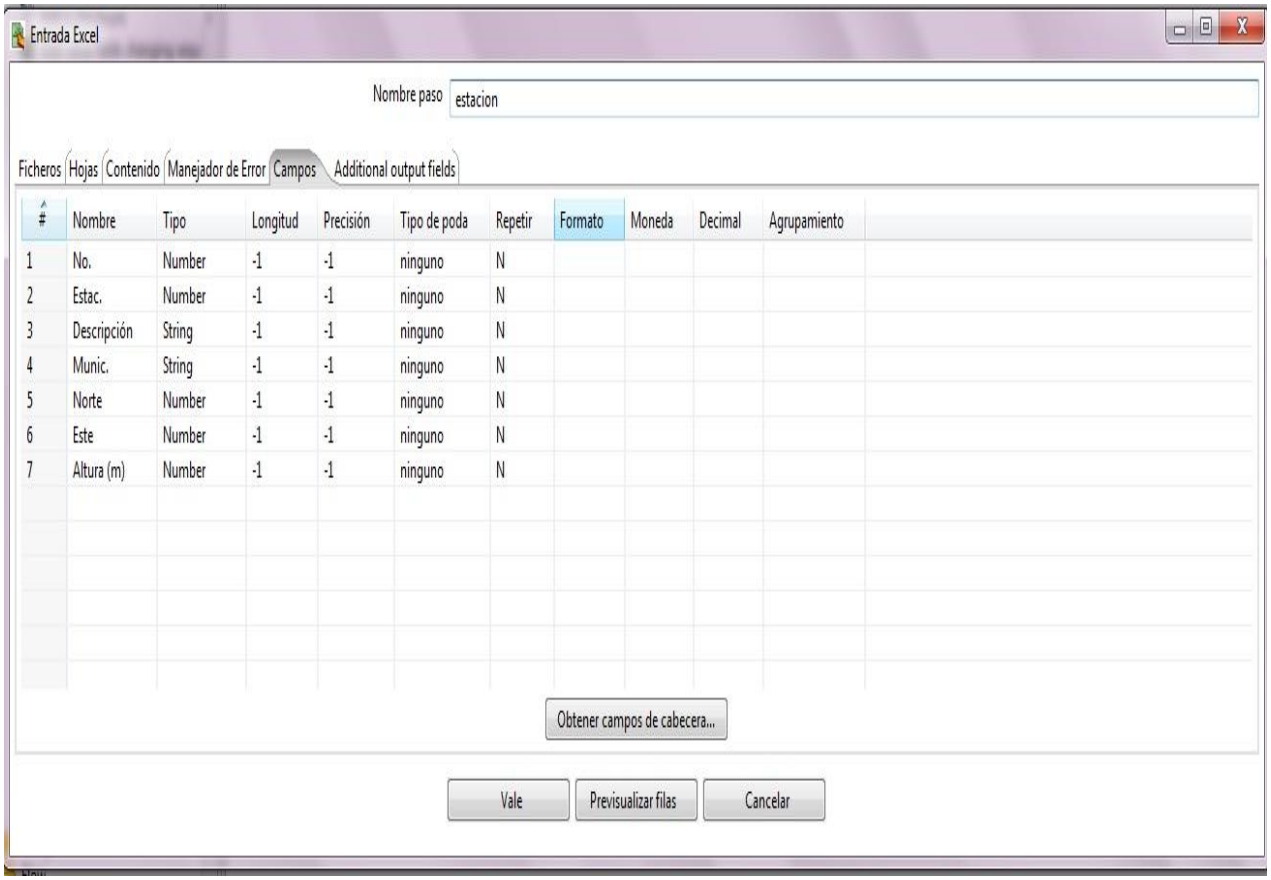
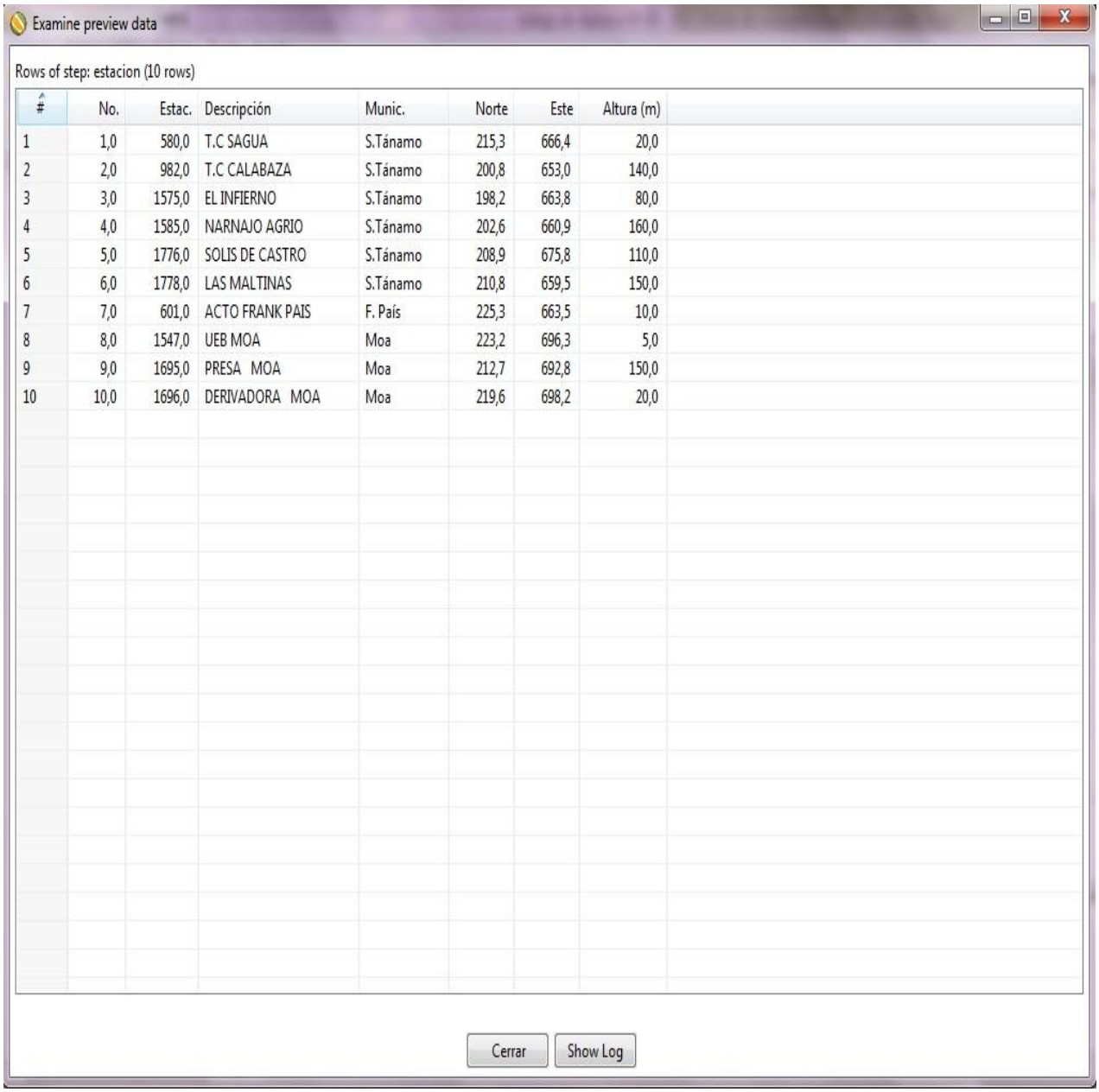


Figura: 16 Componente entrada Excel: obtener campos cabecera.



#	No.	Estac.	Descripción	Munic.	Norte	Este	Altura (m)
1	1,0	580,0	T.C SAGUA	S.Tánamo	215,3	666,4	20,0
2	2,0	982,0	T.C CALABAZA	S.Tánamo	200,8	653,0	140,0
3	3,0	1575,0	EL INFIERNO	S.Tánamo	198,2	663,8	80,0
4	4,0	1585,0	NARNAJO AGRIO	S.Tánamo	202,6	660,9	160,0
5	5,0	1776,0	SOLIS DE CASTRO	S.Tánamo	208,9	675,8	110,0
6	6,0	1778,0	LAS MALTINAS	S.Tánamo	210,8	659,5	150,0
7	7,0	601,0	ACTO FRANK PAIS	F. País	225,3	663,5	10,0
8	8,0	1547,0	UEB MOA	Moa	223,2	696,3	5,0
9	9,0	1695,0	PRESA MOA	Moa	212,7	692,8	150,0
10	10,0	1696,0	DERIVADORA MOA	Moa	219,6	698,2	20,0

Figura: 17 Componente entrada Excel: visualizar filas.

A través de un componente *obtener y transformar campos* se especifica qué valores tomarán las variables de salida definida por el componente (TD_Estación) y además se eliminan los campos que no se van a incluir en el Data Mart (ver Fig.18).

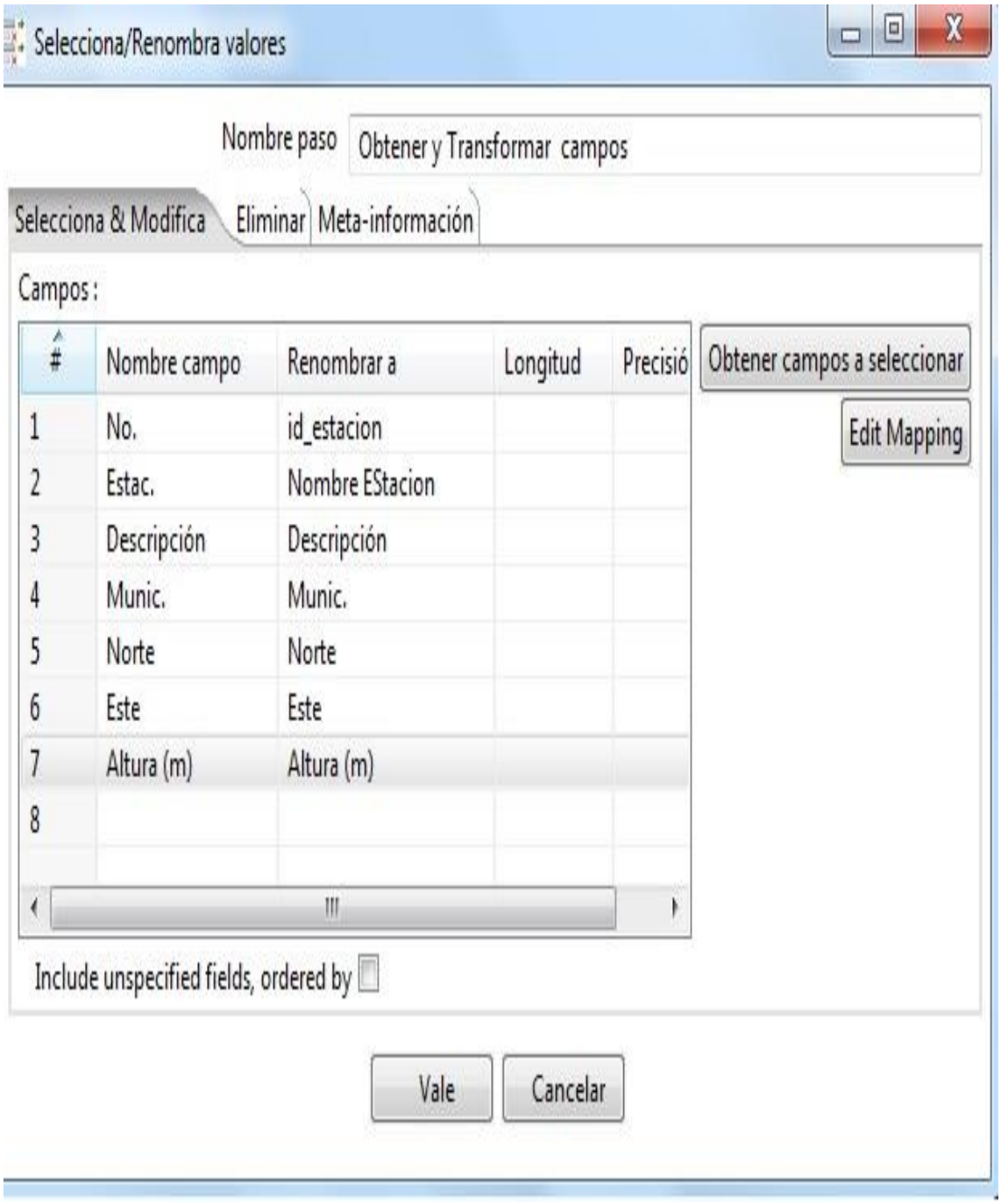


Figura 18: Valores del componente obtener campos.

Una vez seleccionados los mismos y renombrados como han de quedar en la tabla, se añade el componente *Salida Tabla*, este se edita para lograr la conexión con el gestor de bases de datos (Ver figura 19) y se selecciona la tabla en la cual se han de guardar los datos (Ver figura 20).

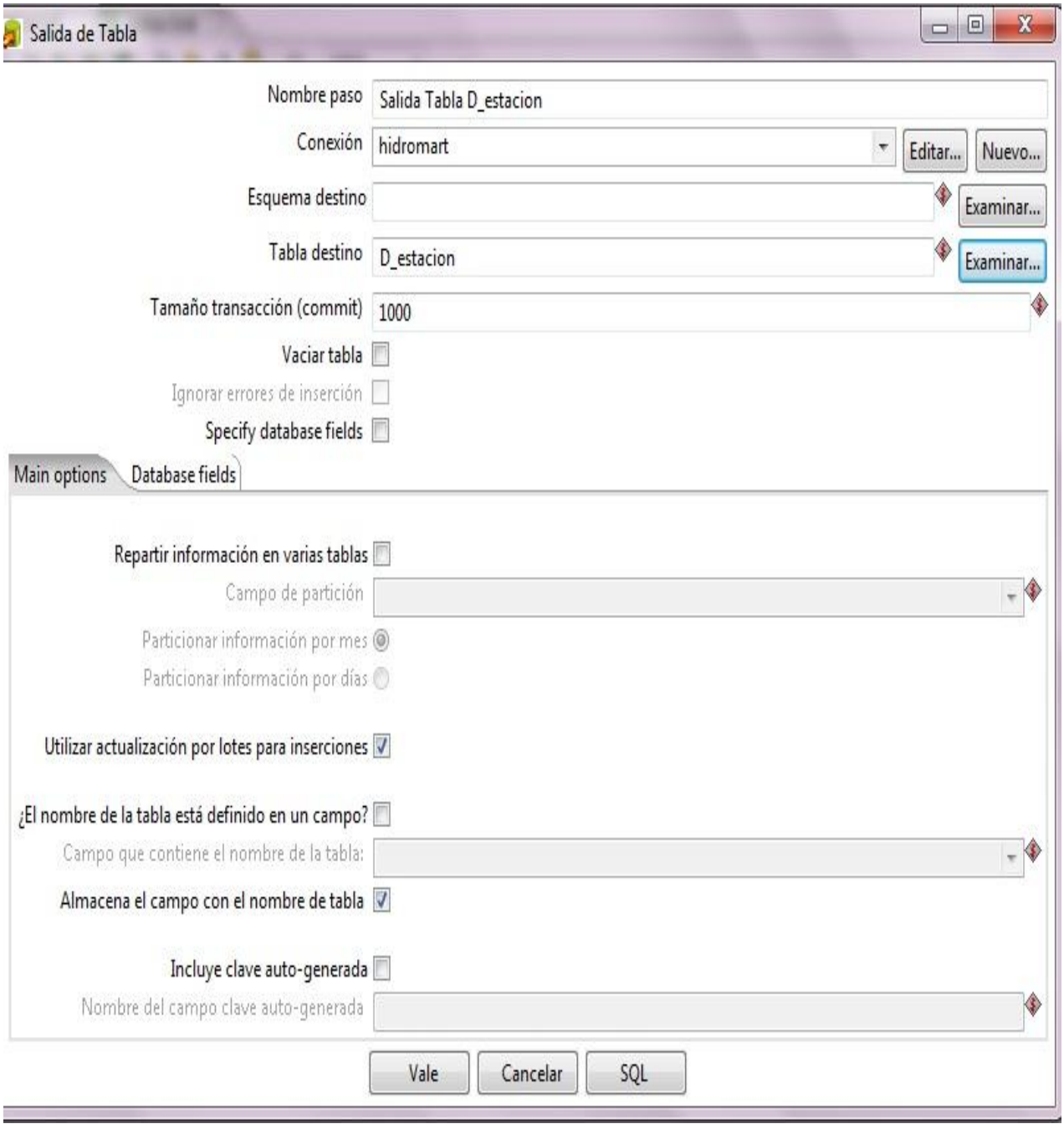


Figura 19: Componente salida tabla edición.

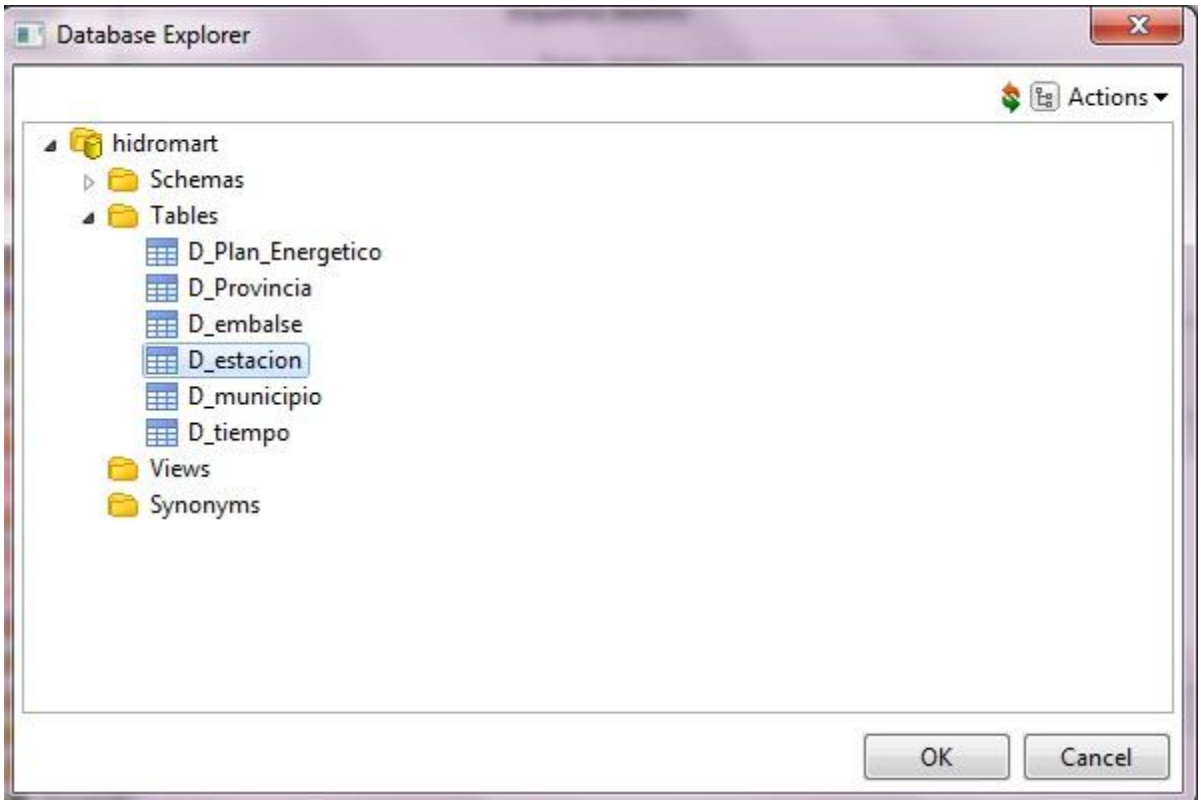


Figura 20: Componente salida tabla edición selección.

Una vez dentro de este se selecciona la opción SQL para entrar al editor de código SQL (Ver figura 21). De esta forma se puede visualizar de forma previa la parte de las sentencias SQL que corresponden a la ejecución de esta transformación.

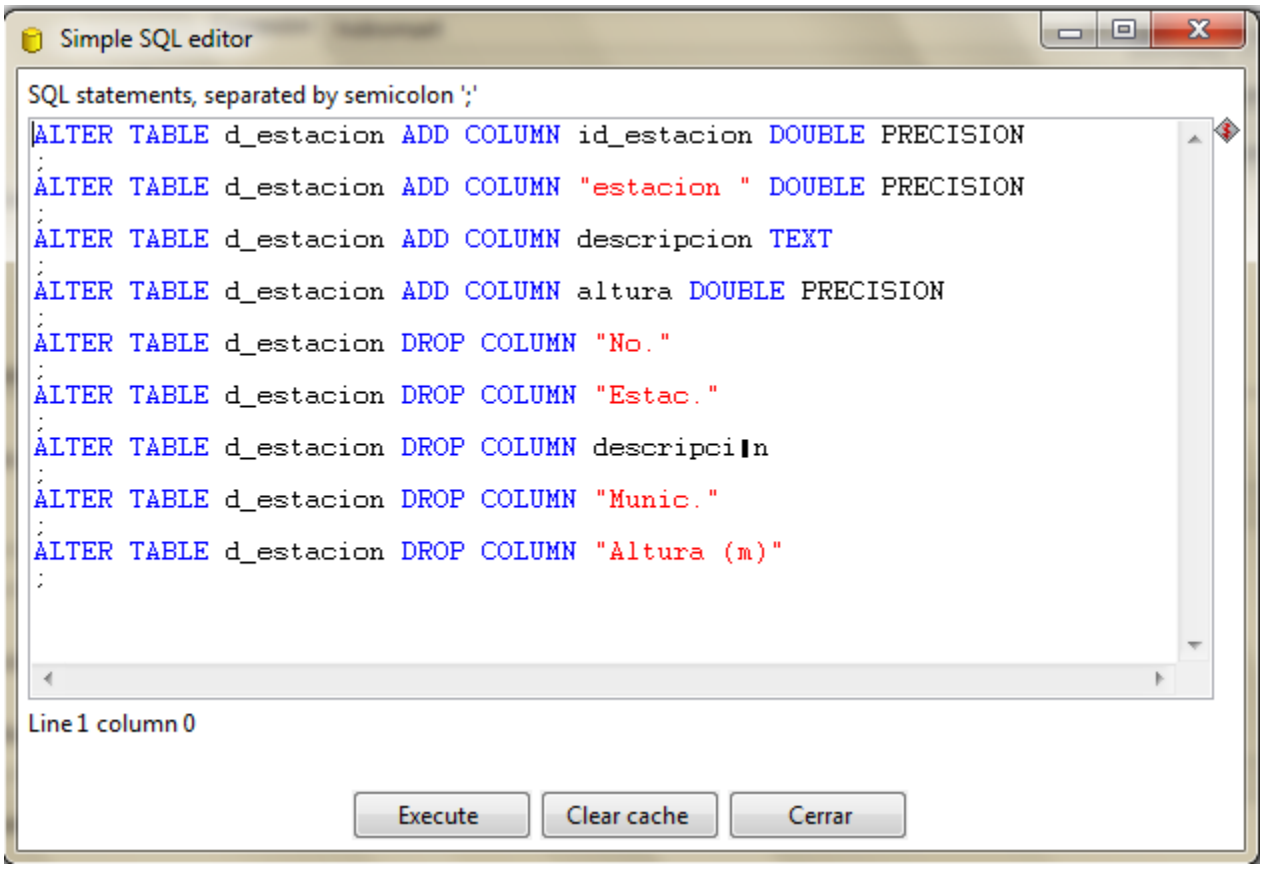


Figura 21: Editor SQL.

En el caso de la dimensión Tiempo ocurre que los datos no son obtenidos de una sola hoja de cálculo y los mismos han de coincidir en formato para evitar incoherencias (Ver figura 22).

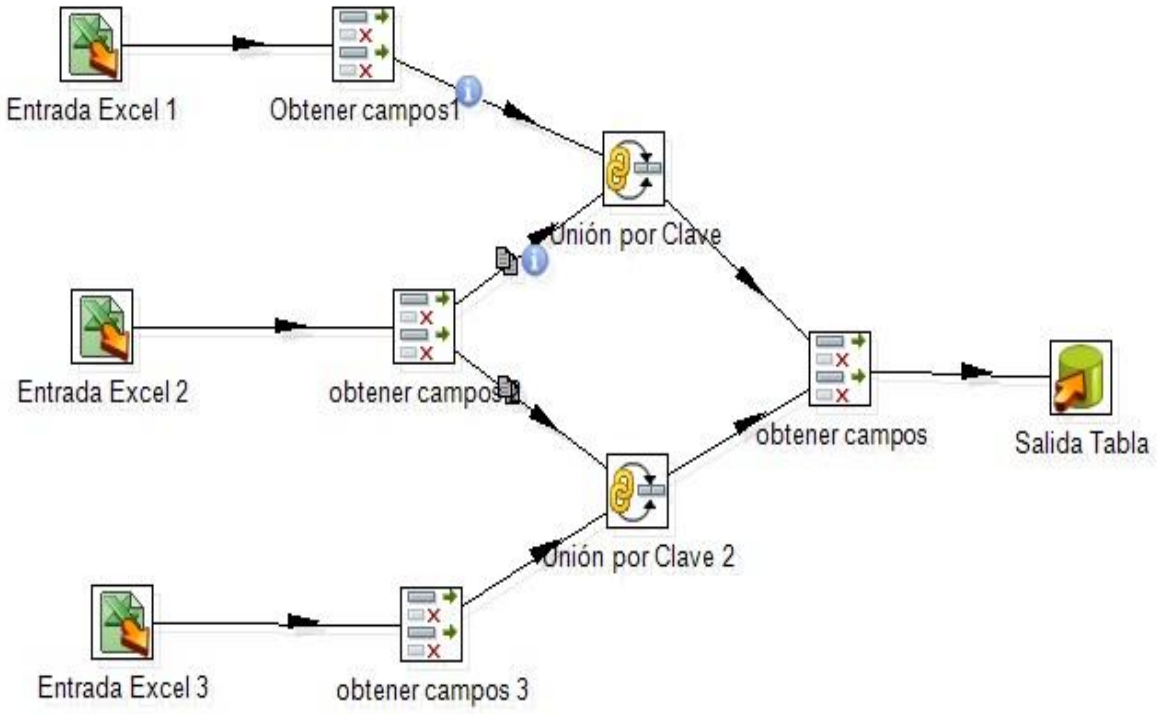


Figura 22: Transformación fecha.

Uno de los componentes a utilizar en esta transformación es el componente *unión por clave*, el cual tiene por objetivo realizar comparaciones entre campos iguales (Ver figura 23)

Figura 23: Componente unión por clave.

El proceso realizado es similar para la carga de las 6 Transformaciones restantes (Ver anexos).

3.2 Carga

Para la realización de la carga como último elemento del proceso ETL, se diseñará mediante la misma herramienta Pentaho Data Integration: un trabajo (Job) (Ver Figura 24)

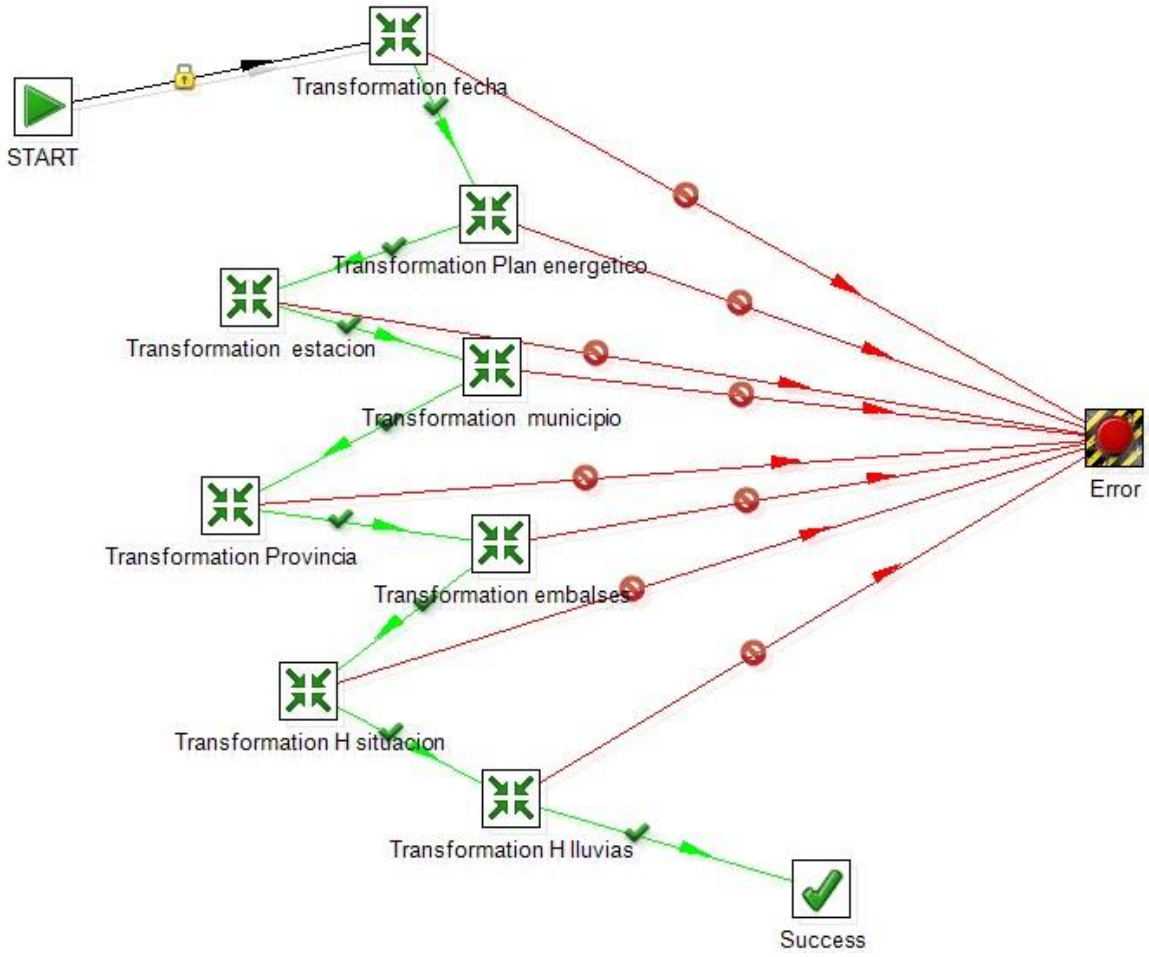


Figura 24: Trabajo Hidromart.

El primer componente que se puede observar es el Start, este se edita para determinar la frecuencia con que se realiza el mismo. Observar en la figura 25. Luego se añaden las transformaciones y se ejecutan de forma secuencial si el resultado de la misma es verdadero (Ver figura 26 y 27). En caso de ser falso el salto sería al componente de *error* donde se configura el mensaje a mostrar (Ver Figura 28)

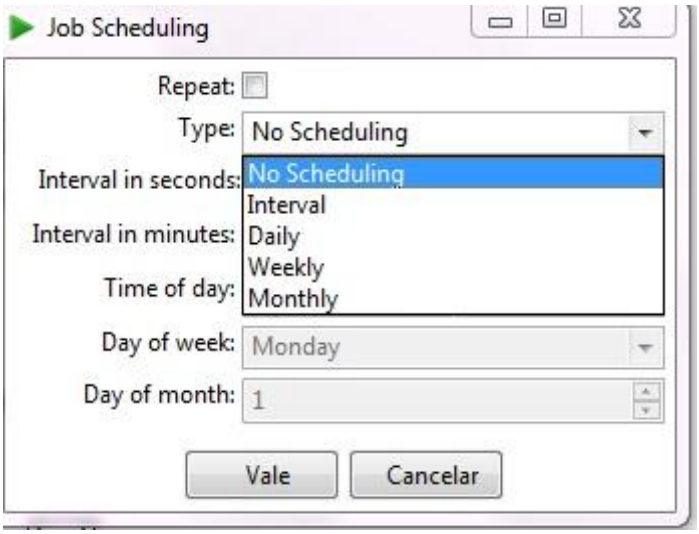


Figura 25: Componente Star.

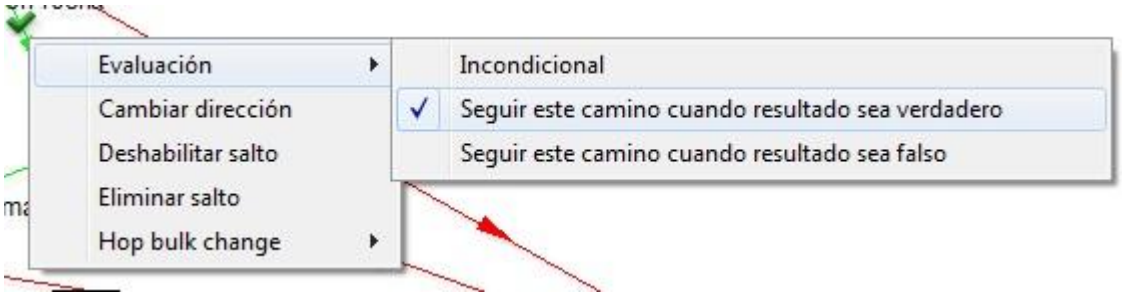


Figura 26: Saltos a siguiente transformación.

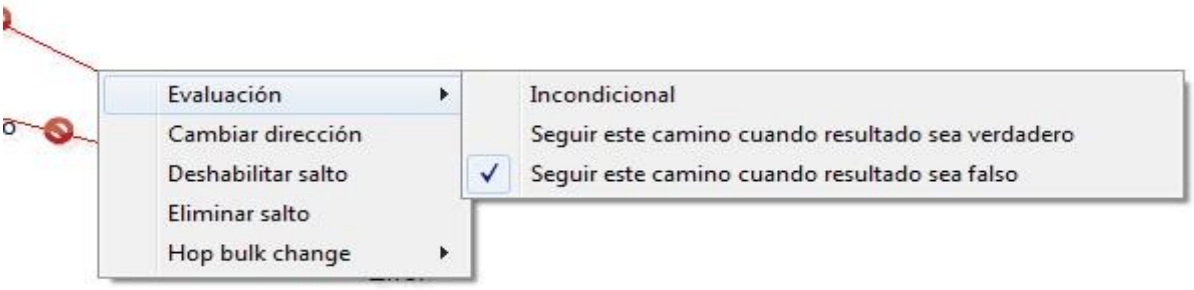


Figura 27: Saltos a Componente Error.

Figura 28: Componente Error.

3.3 Validación.

La validación de todo sistema constituye el mecanismo mediante el cual se garantiza el cumplimiento de las funcionalidades definidas por los clientes al inicio de su desarrollo y su correcto funcionamiento. En este caso como el alcance de la investigación se ha limitado al diseño del Mercado de Datos, es a este diseño al que se le realiza una prueba con el objetivo de comprobar su correcto desempeño así como el cumplimiento de las necesidades del cliente. Para ello se hizo necesario la creación de dos cubos OLAP.

3.3.1. Validación funcional

Para efectuar esta validación fue necesario el uso de una herramienta para la representación de los cubos, el Schema Workbench, definiendo dos cubos con las dimensiones y medidas correspondientes. Estos cubos se denominaron Lluvias_Precipitadas y Situación_Embalse. Cada uno de estos cubos tiene como base fundamental las tablas de hechos "H_lluvias_precipitadas" y "H_situacion_embalses" respectivamente.

El primer paso es realizar la conexión entre las herramientas antes mencionadas (Ver Figura 29).

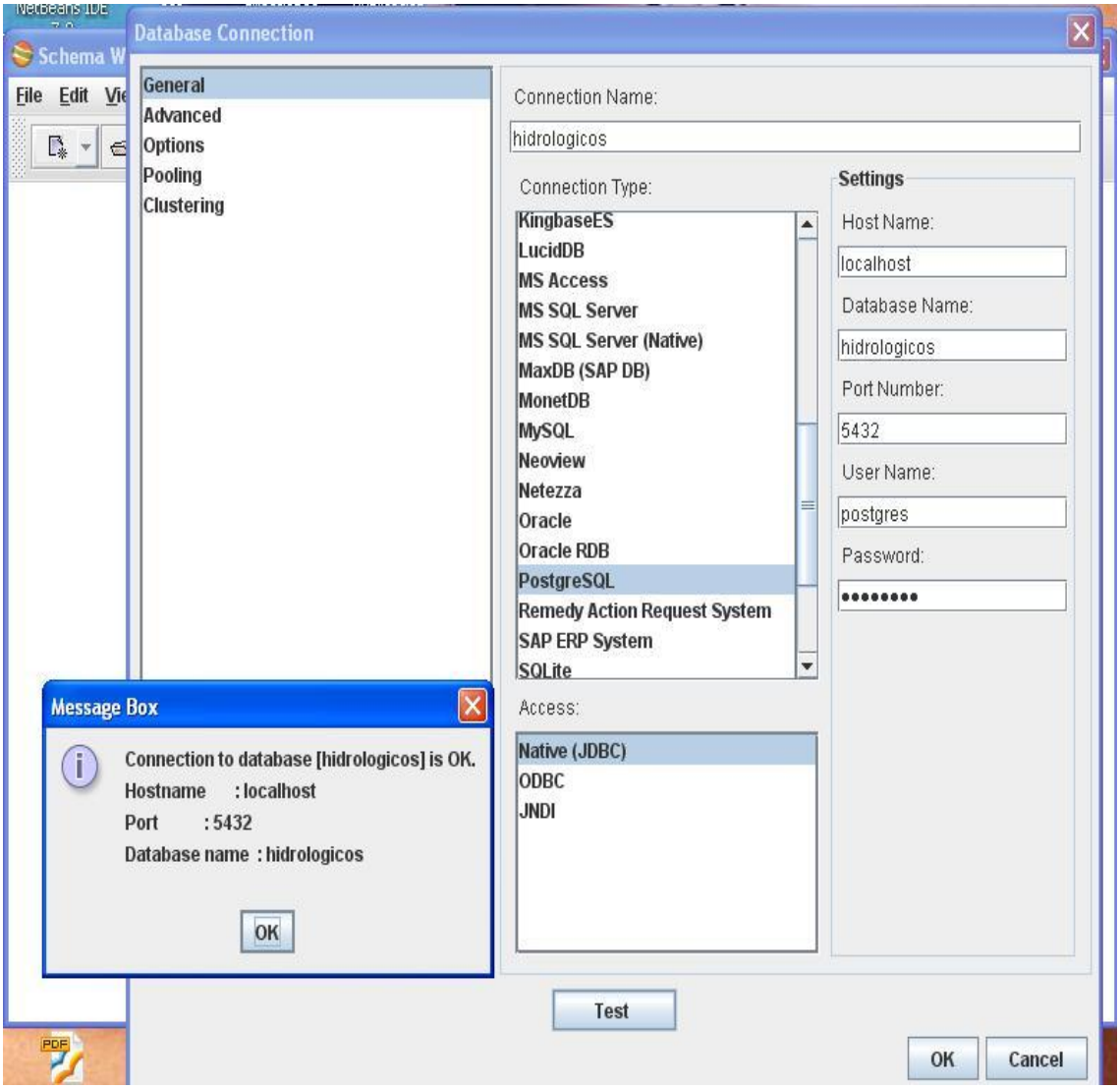


Figura 29. Conexión entre Schema y PostgreSQL.

Una vez conectado se procede a diseñar el modelo, el cual consta de dos cubos Lluvias_Precipitadas y Situación_Embalse, con 4 dimensiones compartidas por ambos cubos y dos individuales (Ver Figura 30)

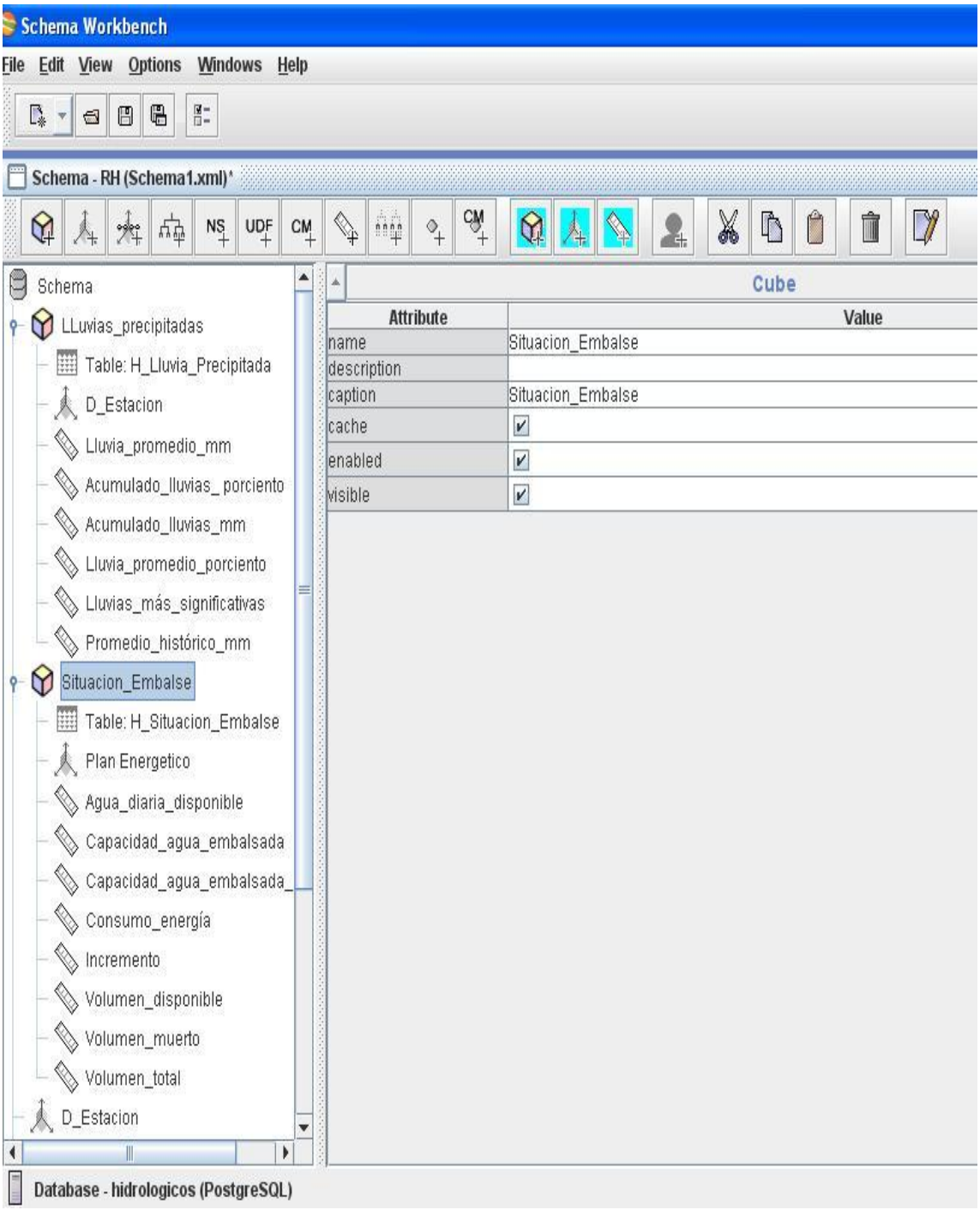


Figura 30. Diseño del modelo multidimensional en el Schema Workbench.

Uno de los componentes generados por esta herramienta es un archivo XML (Ver figura 31) el cual permite comprobar la funcionalidad del mercado.

Cube

```
<Cube name="LLuvias_precipitadas" caption="LLuvias_precipitadas" visible="true" cache="true" enabled="true">
  <Table name="H_Lluvia_Precipitada" schema="public">
  </Table>
  <Dimension type="StandardDimension" visible="true" highCardinality="false" name="D_Estacion" caption="D_Estacion">
  </Dimension>
  <Measure name="Lluvia_promedio_mm" column="Lluvia_promedio_mm" datatype="Integer" aggregator="avg" caption="Lluvia_promedio_mm" visible="true">
  </Measure>
  <Measure name="Acumulado_lluvias_porciento" column="Acumulado_lluvias_porciento" datatype="Integer" aggregator="sum" caption="Acumulado_lluvias_porciento" visible="true">
  </Measure>
  <Measure name="Acumulado_lluvias_mm" column="Acumulado_lluvias_mm" datatype="Integer" aggregator="sum" caption="Acumulado_lluvias_mm" visible="true">
  </Measure>
  <Measure name="Lluvia_promedio_porciento" column="Lluvia_promedio_porciento" datatype="Integer" aggregator="avg" caption="Lluvia_promedio_porciento" visible="true">
  </Measure>
  <Measure name="Lluvias_m&#225;s_significativas" column="Lluvias_m&#225;s_significativas" datatype="Integer" aggregator="max" caption="Lluvias_m&#225;s_significativas" visible="true">
  </Measure>
  <Measure name="Promedio_hist&#243;rico_mm" column="Promedio_hist&#243;rico_mm" datatype="Integer" aggregator="avg" caption="Promedio_hist&#243;rico_mm" visible="true">
  </Measure>
</Cube>
```

Figura 31. Schema Workbench: XML.

Conclusiones de capítulo

En este capítulo se realizó el proceso ETL, de vital importancia para el éxito del DM, en el cual se transformó y cargó todos los datos seleccionados desde los sistemas operacionales al sistema gestor de bases de datos PostgreSQL. Este proceso concluyó satisfactoriamente con la exitosa ejecución del trabajo nombrado Hidromart como tarea final del desarrollo del diseño del Data Mart para el análisis de datos hidrológicos en la provincia Holguín.

Conclusiones Generales

Una vez culminado el trabajo de diploma se puede concluir que se le dio cumplimiento al objetivo general trazado para el proceso de la investigación, puesto que:

- Se desarrolló el diseño de un Data Mart para el análisis de los datos hidrológicos en la Empresa Aprovechamiento de los Recursos Hidráulicos Holguín (EAHHLG), el cual tiene el objetivo aportar una ayuda al proceso de toma de decisiones a partir de datos que se almacenan en el mismo de forma integrada y segura y posibilitan la creación de modelos para la anticipación de situaciones de los embalses de la provincia y la situación de lluvias precipitadas, mejorando los tiempos de análisis y trabajo con dichos datos; validado a través de pruebas realizadas a las funcionalidades, las que arrojaron resultados favorables, dando cumplimiento a las necesidades y exigencias previstas para su desarrollo hasta la etapa propuesta para esta investigación.

Recomendaciones

Para un posterior desarrollo y perfeccionamiento del DM desarrollado se recomienda:

- Desarrollar el sistema para la generación de reportes.
- Realizar vistas de análisis mediante la visualización de información a través de pizarras de gráficos.
- Gestionar la seguridad del sistema para el acceso a la información a nivel de entidades.
- Realizar la integración del sistema con un marco de trabajo (como el del CEDRUX).

Bibliografía

"Introducing the Open Source CUAHSI Hydrologic Information System Desktop Application (HIS Desktop)". **Ames, D.P. 2009.** Australia : 18th World IMACS / MODSIM Congress, 2009. pp:4353 - 4359.

2012. A practical guide to getting stared with Dta Warehousing. [En línea] Marzo de 2012. http://www.techguide.com/.

Alonso y Infante, Eddy Manuel. 2005. *Diseño e implemetación de Akademo Mart.* La Habana : s.n., 2005.

Bernau Ricardo, Dario. 2010. *HEFESTO Metodología para la construcción de un Data Warehouse.* Córdova : s.n., 2010.

Bravo Martínez, Edgar Caheri. Manejo de datos fuente para la integración en sistemas de informacion geográficos. *Colección de Tesis Digitales Universidad de las Américas Puebla. .* [En línea] [Citado el: 03 de Febrero de 2013.] http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/bravo_m_ec/..

Conferencia 2 Modelo de datos. **Universidad de las Ciencias Informáticas, UCI. 2011.** La Habana : s.n., 2011.

ETL-Tools.Info. *Business Intelligence - Almacenes de Datos - ETL.* [En línea] [Citado el: 20 de Enero de 2013.] http://etl-tools.info/es/bi/almacendedatos_esquema-constelacion.htm. .

Galindo Aparicio, Luis Alberto. SINAI: Toolkit de indexacion de datos multidimensionales. *Colección de Tesis Digitales Universidad de las Américas Puebla.* [En línea] [Citado el: 03 de Febrero de 2013.] http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/galindo_a_la/..

Garrido, Maray. 2012. *"Integración de herramientas informáticas para la alerta temprana ante el peligro de inundaciones".* La Habana : Instituto Superior Politécnico “José Antonio Echeverría”, 2012.

Guanche Cañizares, Mauricio. 2011. *Desarrollo de un DataMart para la obtención de las razones financieras de los subsistemas de Cobros y Pagos, Caja y Banco del proyecto ERP-Cuba.* La habana : UCI, 2011.

Inmon , Bill. 1996. *Building the data Warehouse.* New York : Wiley Computer, 1996.

Kimball, Ralph y Ross, Margy. *The Data Warehouse Toolkid.* Segunda edicion. Canada : s.n.

Michel Diaz, Llerena . 2007. *Data Mart para la gestión del conocimiento.* 2007.

Modelo Relacional. *Modelo Relacional.* [En línea] [Citado el: 12 de Enero de 2013.] http://www.virtual.unal.edu.co/cursos/sedes/manizales/4060029/lecciones/cap4-1.html..

Molina, Martín. 2006. *Proyecto PREDECAN apoyo a la prevención de desastres en la Comunidad Andina. Análisis de Sistemas de Información de Prevención y Atención de Desastres en la Comunidad Andina.* Lima, Perú : Comunidad Andina, 2006.

Mosteiro, Luis. 2009. *Sistema Integrado de Información del Agua.* 2009.

ONJUÁN, G. 2002. *Gestión de Información en las organizaciones. Principios, conceptos y aplicaciones.* Habana : Felix Varela, 2002.

Pérez Pedraza, Alejandro. Implementación y explotación de un datawarehouse empresarial para la toma de decisiones: aplicación a la empresa Textiles Carmelita. *Colección de Tesis Digitales Universidad de las Américas Puebla.* [En línea] [Citado el: 03 de Febrero de 2013.] http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/perez_p_a/.

Quintero Orea, Moises. Interfaz en español para recuperación de información de datos geográficos. *Colección de Tesis Digitales Universidad de las Américas Puebla.* [En línea] [Citado el: 03 de Febrero de 2013.] http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/quintero_o_m/.

Rodríguez Fernández, Osvaldo. 2010. SAEP: Data Warehouse para apoyar el análisis de evaluación de profesores. *Colección de Tesis Digitales Universidad de las Américas Puebla.* [En línea] 2010. [Citado el: 03 de Febrero de 2013.] .http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/ydirin_p_mm/portada.html..

Second International Symposium on Information Technologies in Environmental Engineering. **Shaker-Verlag. 2005.** Alemania : s.n., 2005. Modelling of a data warehouse system for environmental information – A case study.

2011. Sitio oficial de postgre. *Sitio oficial de postgre.* [En línea] 2011. http://www.postgresql.org/..

Valdes Yero, Livan. 2011. *Desarrollo de un Data Mart para la obtención de los indicadores financieros de los subsistemas de Contabilidad y Costos y Procesos del proyecto ERP.* La Habana : UCI, 2011.

Wiley. 2002. *Modeling (Second Edition).* New york : s.n., 2002.

Wolf, Carmen Gloria. 2012. La tecnología data warehousing. *La tecnología data warehousing.* [En línea] Noviembre de 2012. [Citado el: 15 de Diciembre de 2012.] http://www.inf.udec.cl/revistaedicion3/cwolf.html.

Ydirín Préstamo, María Magdalena. 2004. Construcción de un Data Warehouse de datos del medio ambiente para la toma de decisiones: aplicación a los datos hidrológicos. *Colección de Tesis Digitales Universidad de las Américas Puebla.* [En línea] 2004. [Citado el: 03 de Febrero de 2013.] .http://catarina.udlap.mx/u_dl_a/tales/documentos/lis/ydirin_p_mm/portada.html..

Glosario de términos

BD: Una Base de Datos es un conjunto de datos relacionados entre sí, entendiéndose por dato los hechos conocidos, que pueden registrarse y que tienen significado implícito.

Data Warehouse (DWH): El almacenamiento de información homogénea y fiable, en una estructura basada en la consulta y el tratamiento jerarquizado de la misma, y en un entorno diferenciado de los sistemas operacionales.

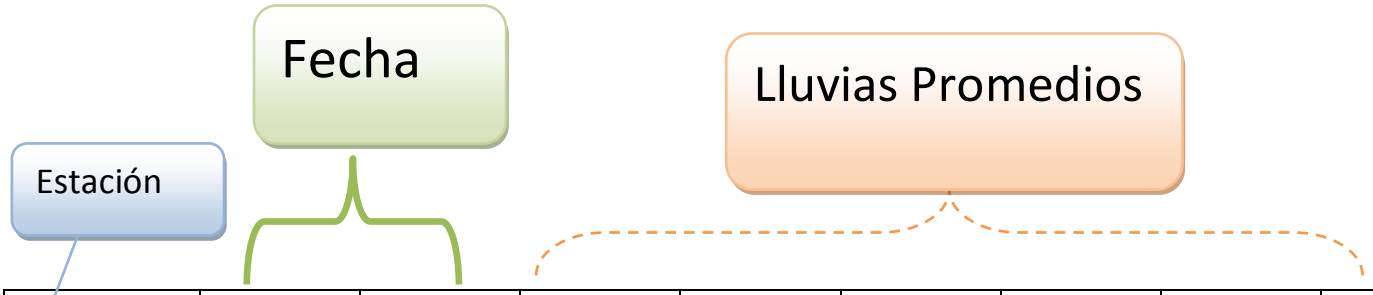
Data Mart: Son subconjuntos de datos con el propósito de ayudar a que un área específica dentro del negocio pueda tomar mejores decisiones. Los datos existentes en este contexto pueden ser agrupados, explorados y propagados de múltiples formas para que diversos grupos de usuarios realicen la explotación de los mismos de la forma más conveniente según sus necesidades.

Inteligencia de Negocio (BI): El proceso de analizar los bienes o datos acumulados en la empresa y extraer una cierta inteligencia o conocimiento de ellos.

SQL: Lenguaje de consulta estructurado o SQL (por sus siglas en inglés structured query language) es un lenguaje declarativo de acceso a bases de datos relacionales que permite especificar diversos tipos de operaciones en éstas. Una de sus características es el manejo del álgebra y el cálculo relacional permitiendo efectuar consultas con el fin de recuperar -de una forma sencilla- información de interés de una base de datos, así como también hacer cambios sobre ella.

Anexos

Anexo 1 Correspondencias entre los OLTP y los indicadores a conformar el mercado de datos:



Estac.	Ano	IdMes	Dia01	Dia02	Dia03	Dia04	Dia05	Dia
580	1990	1	0	30	3	0	0	0
601	1990	1	1	12	15,5	1	0	0
982	1990	1	0	3	5	0	0	0
1547	1990	1	0	5	7	1	3	1,
1575	1990	1	0	0	0	0	0,1	0,
1585	1990	1	0	0	16	0	0	0
1695	1990	1	2,3	31,4	0	7,9	0	4,
1696	1990	1	0	13	8	6	0,3	3
1776	1990	1						
1778	1990	1						
580	1990	2	0	0	0	0	0	0
601	1990	2	0	0	0	0	0	0
982	1990	2	0	0	0	0	0	1
1547	1990	2	0	0	0	1	1,5	8
1575	1990	2	0	1,7	0	0	0	2
1585	1990	2	0	0	0	0	0	0
1695	1990	2	0	0	3,4	2,8	3,6	8,
1696	1990	2	0	1,5	0	1,5	5,5	0
1776	1990	2						
1778	1990	2						
580	1990	3	10	0	18	0	0	0
601	1990	3	48	6	1	0	0	0

Figura 1 Caso refente al promedio de lluvias diarias.

DATOS DE ESTACIÓN

No.	Estac.	Descripción	Munic.	Norte	Este	Altura (m)
1	580	T.C SAGUA	S.Tánamo	215,30	666,40	20,0
2	982	T.C CALABAZA	S.Tánamo	200,80	653,00	140,0
3	1575	EL INFIERNO	S.Tánamo	198,20	663,80	80,0
4	1585	NARNAJO AGRIO	S.Tánamo	202,60	660,90	160,0
5	1776	SOLIS DE CASTRO	S.Tánamo	208,9	675,8	110,0
6	1778	LAS MALTINAS	S.Tánamo	210,8	659,5	150,0
7	601	ACTO FRANK PAIS	F. País	225,30	663,50	10,0
8	1547	UEB MOA	Moa	223,20	696,30	5,0
9	1695	PRESA MOA	Moa	212,70	692,80	150,0
10	1696	DERIVADORA MOA	Moa	219,60	698,20	20,0

Figura 2 Caso datos estación

dia	mes	año	municipio	Lluvia promedio		Acumes mm	Acumes%	PromHistmm	AcumFechm	AcumFech%
27	2	2013	Calixto Garci	0	0,0	2,0	7,1	28,2	35,2	57,3
27	2	2013	Cacocum	0	0,0	5,9	31,8	18,5	25,5	70,5
27	2	2013	Holguin	0	0,0	8,0	22,7	35,1	90,2	113,1
27	2	2013	Gibara	0	0,0	14,5	43,3	33,4	161,2	207,2
27	2	2013	Rafael Freyre	0	0,0	14,8	40,2	36,9	103,5	114,1
27	2	2013	Baguanos	0	0,0	15,0	41,0	36,6	132,4	163,5
27	2	2013	Urbano Noris	0	0,0	1,8	8,6	20,9	8,0	18,1
27	2	2013	Cueto	0	0,0	16,5	42,3	39,0	41,3	48,4
27	2	2013	Banes	0	0,0	21,3	28,6	74,4	131,0	82,4
27	2	2013	Antilla	0	0,0	50,2	66,9	75,0	199,4	127,1
27	2	2013	Mayari	0	0,0	49,7	59,7	83,2	121,7	70,1
27	2	2013	Sagua de T.	0	0,0	61,7	80,4	76,7	90,2	57,2
27	2	2013	Frank País	0	0,0	82,9	93,3	88,9	184,6	100,5
27	2	2013	Moa	0	0,0	187,6	133,7	140,3	351,3	118,6
27	2	2013	Provincia	0	0,0	31,2	54,5	57,3	106,1	86,7
20	3	2013	Calixto Garci	0	0,0	13,6	30,3	45,0	48,8	53,4
20	3	2013	Cacocum	0	3,7	7,9	17,4	45,3	33,4	50,5

Figura 3 Caso lluvias por embalses

Anexo 2 Transformaciones:

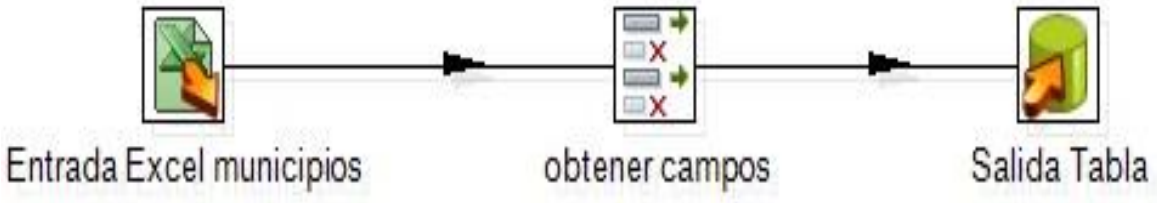


Figura 4 Transformación realizada para el llenado de la tabla dimension

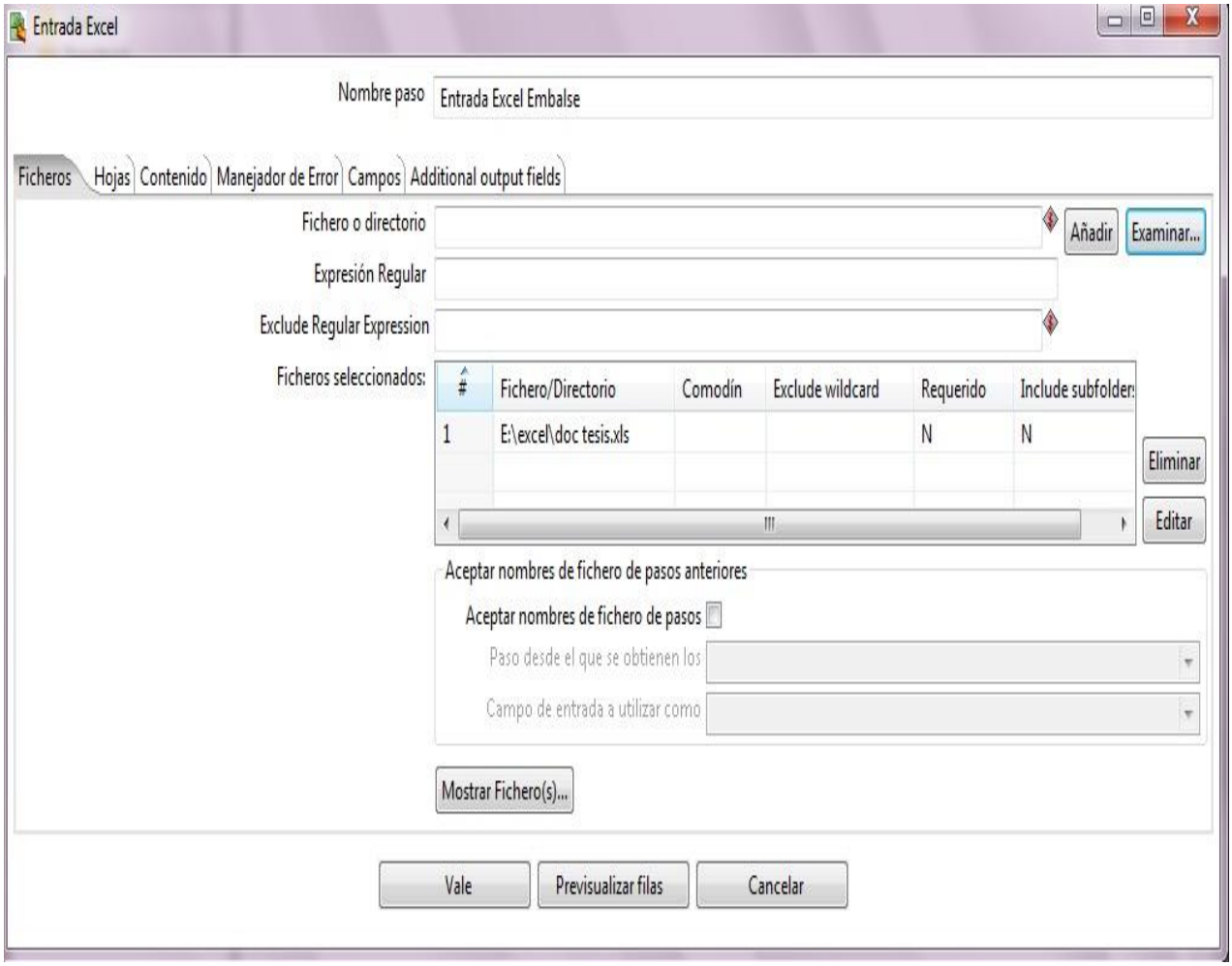


Figura 5 Entrada Excel municipios selección de fichero.

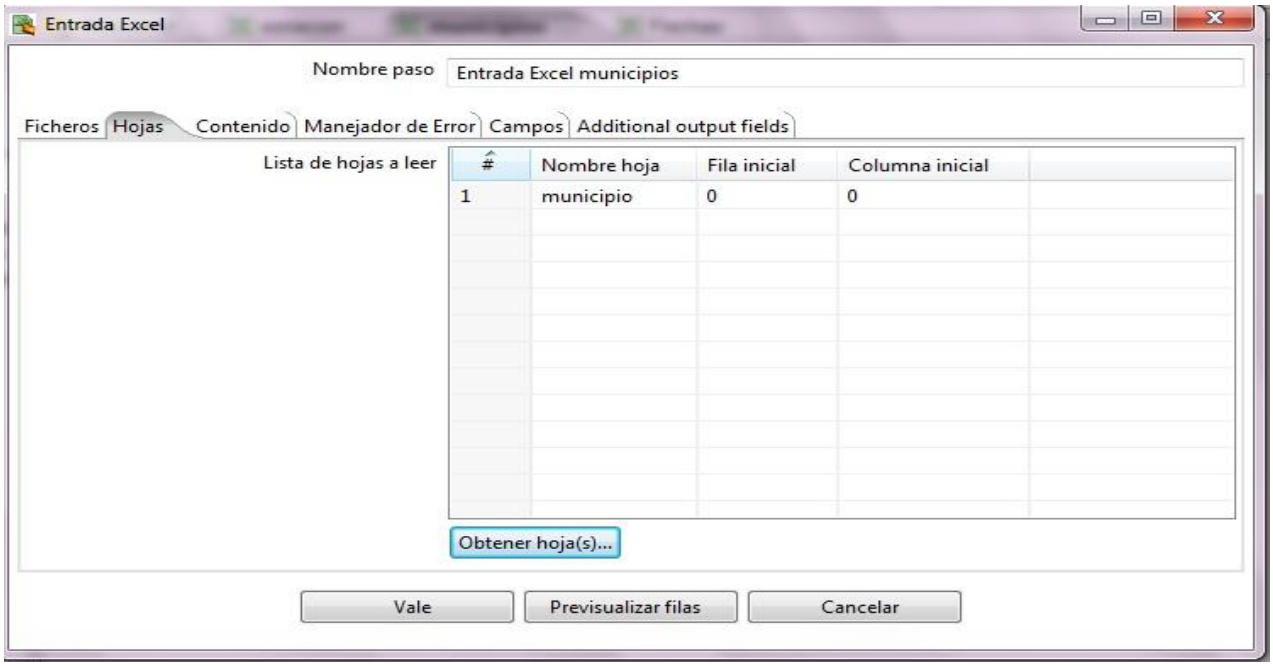


Figura 6 Entrada Excel municipios obtener hoja de cálculo

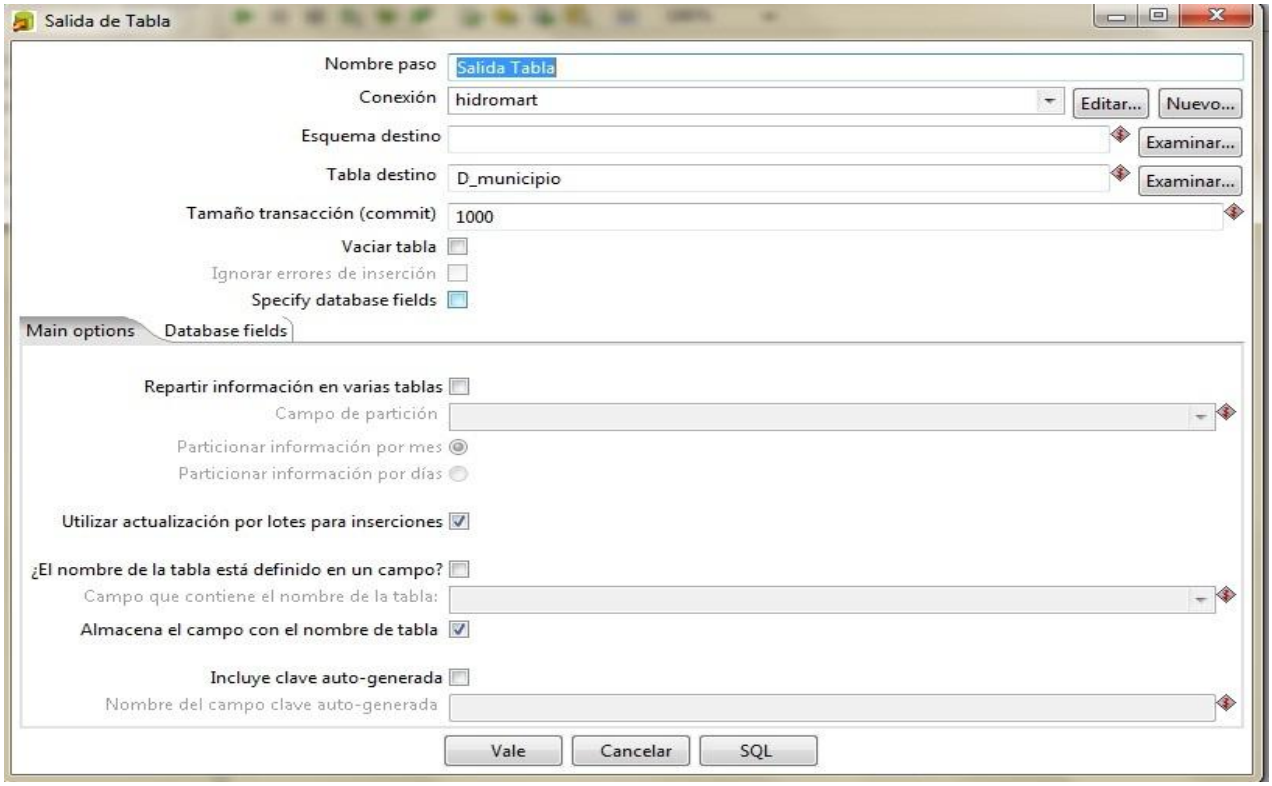


Figura 7 Salida tabla (municipios)

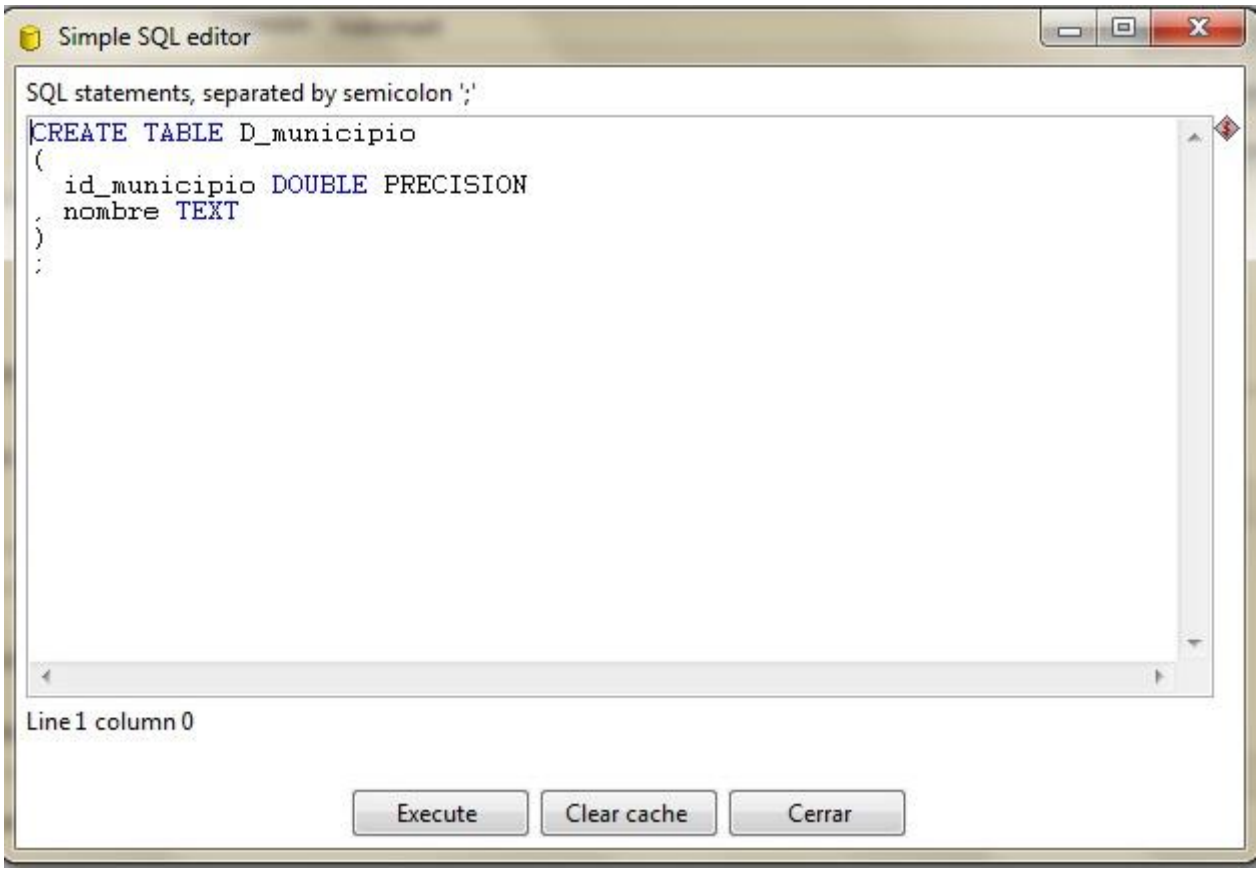


Figura 8 editor SQL transformación municipios

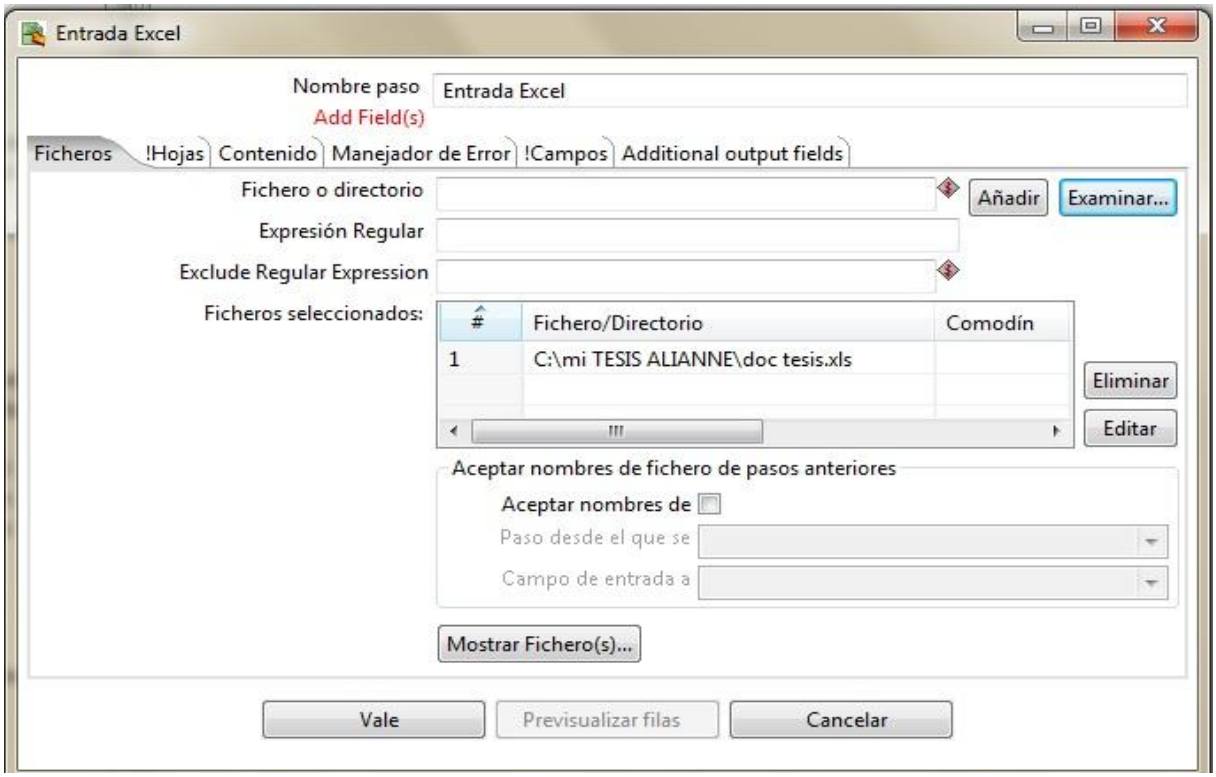
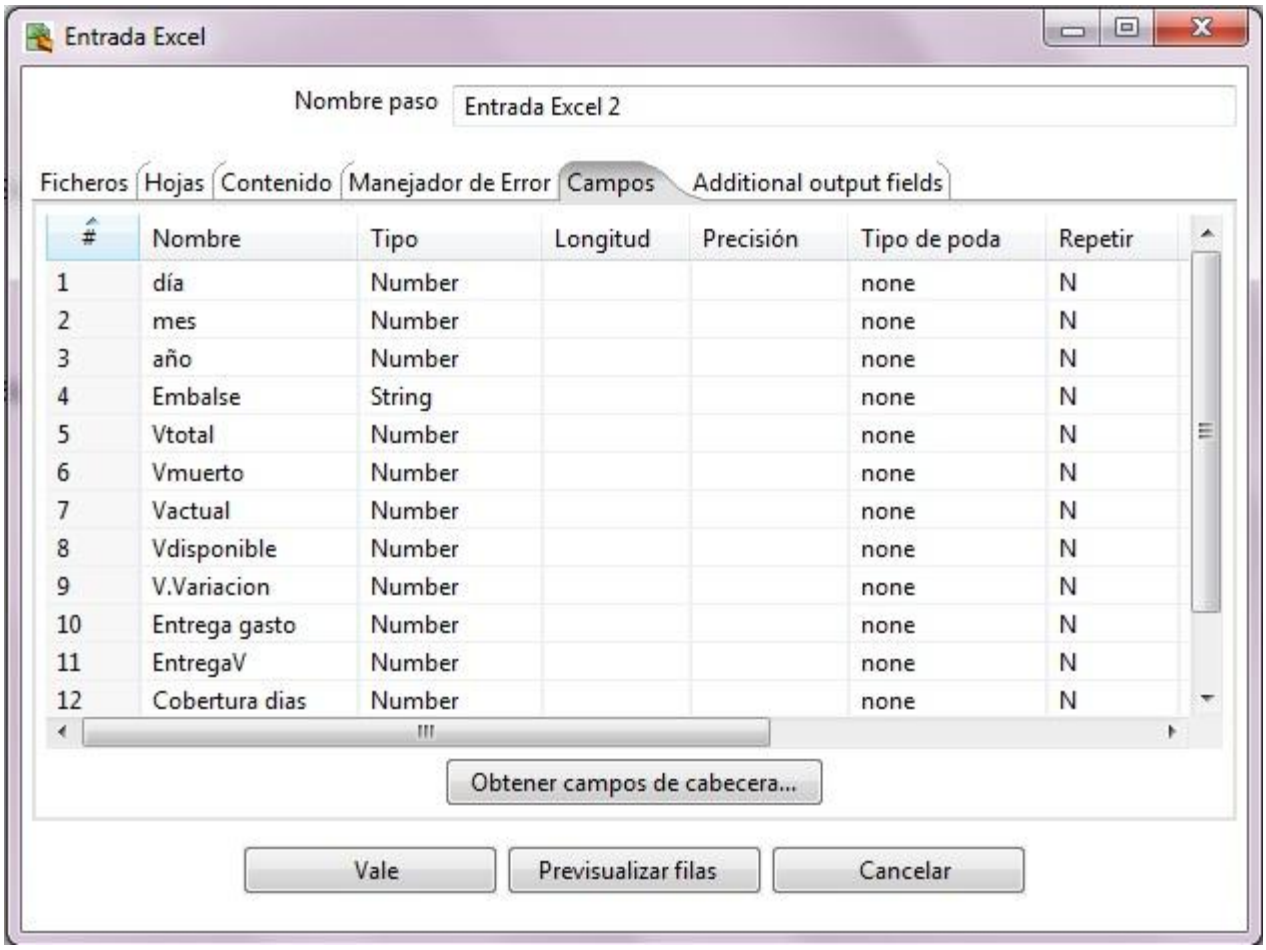
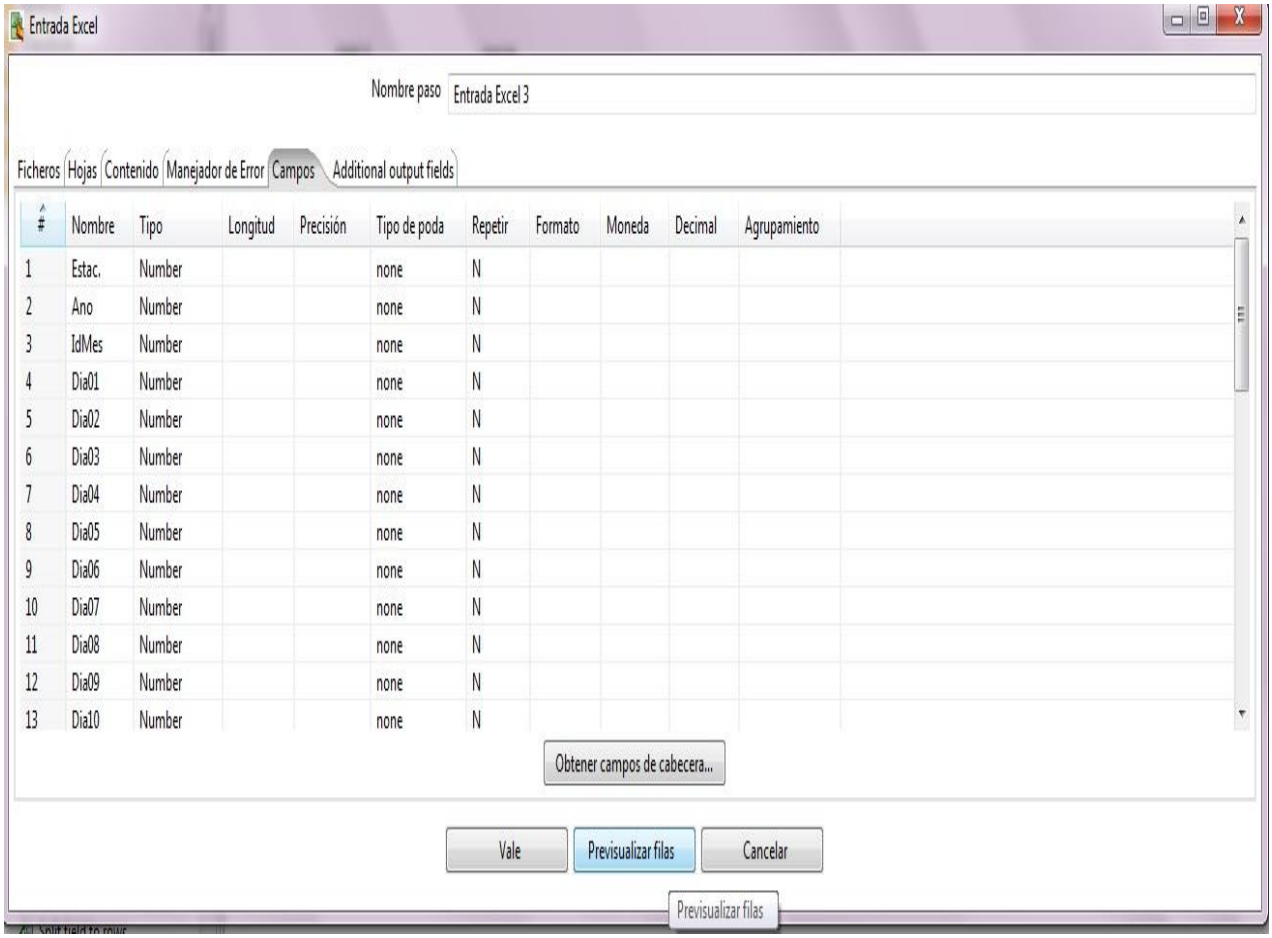


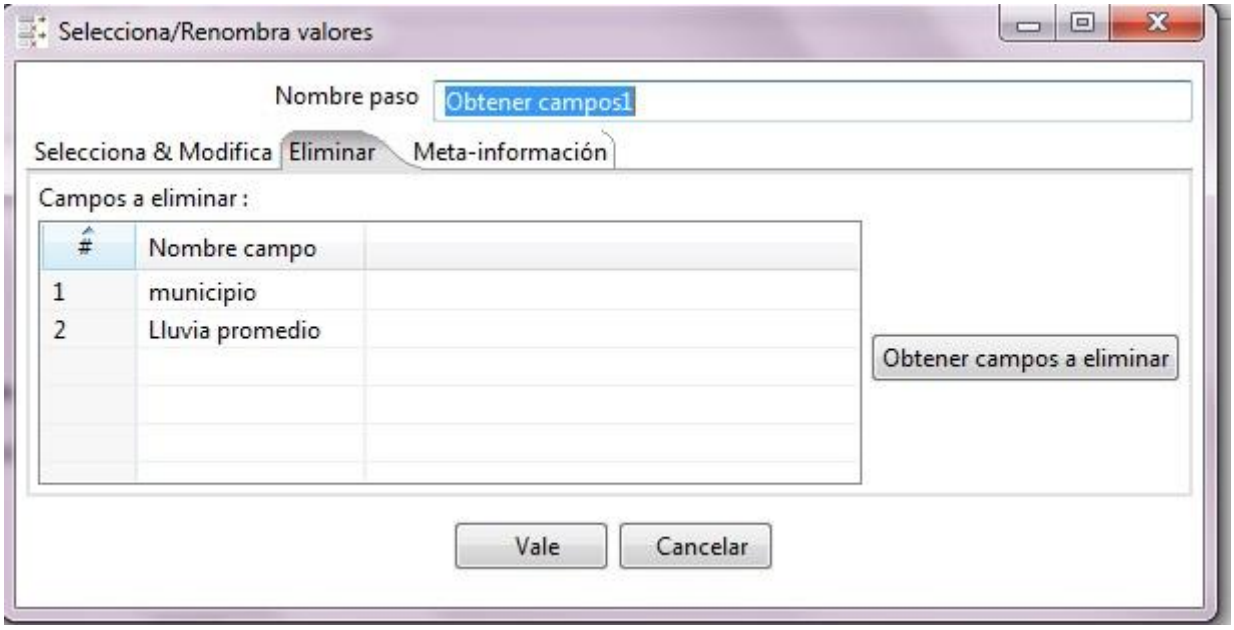
Figura 9 Entrada Excel



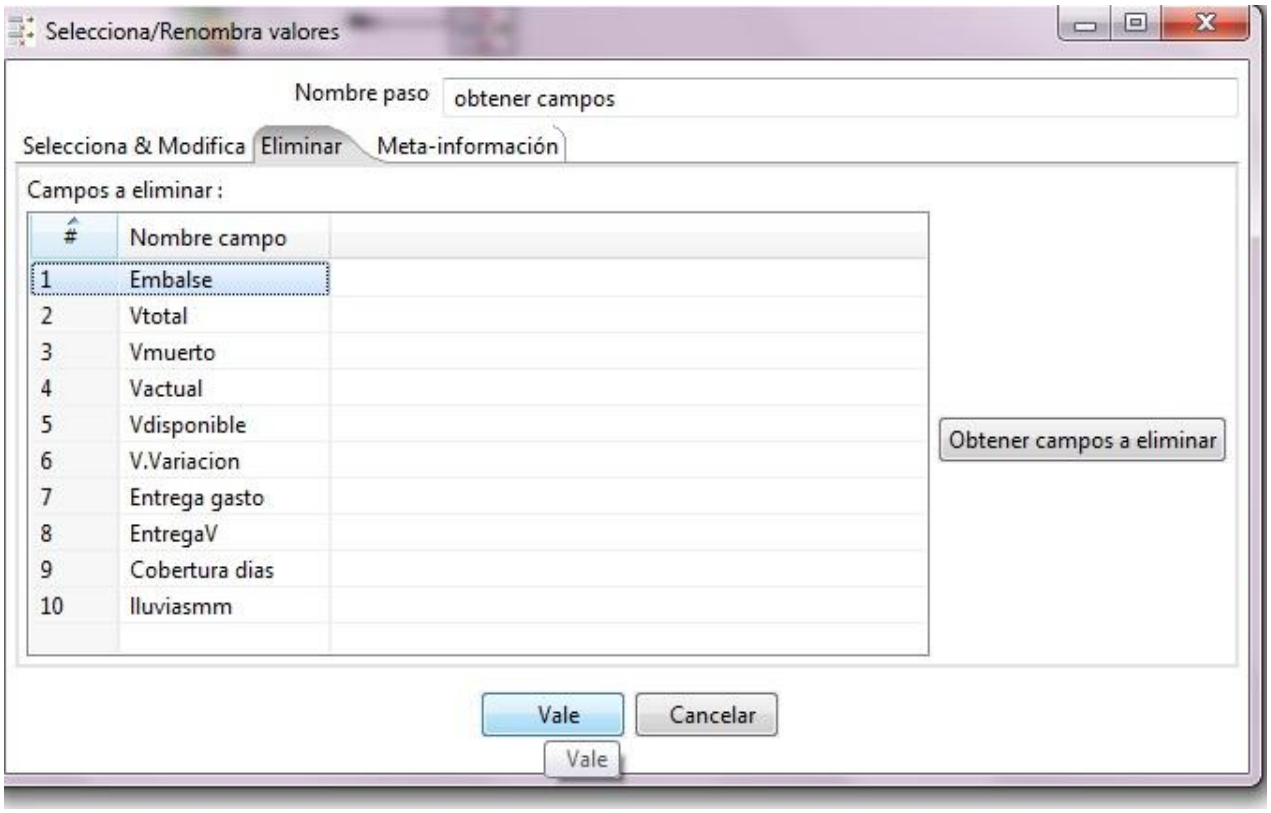
Anexo 8 Entrada Excel 2 Transformación Fecha



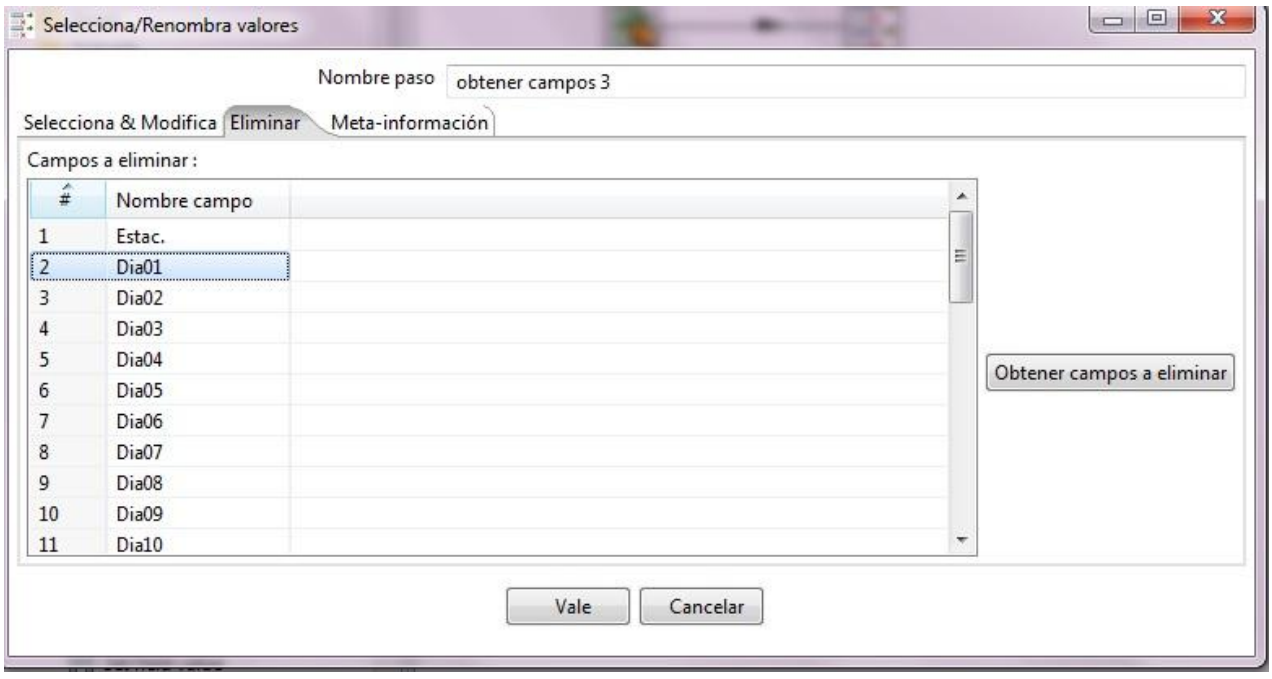
Anexo 9 Entrada Excel 3 Transformación Fecha



Anexo 10 Obtener Campos 1 Transformación Fecha



Anexo 11 Obtener Campos 2 Transformación Fecha



Anexo 12 Obtener Campos 3 Transformación Fecha