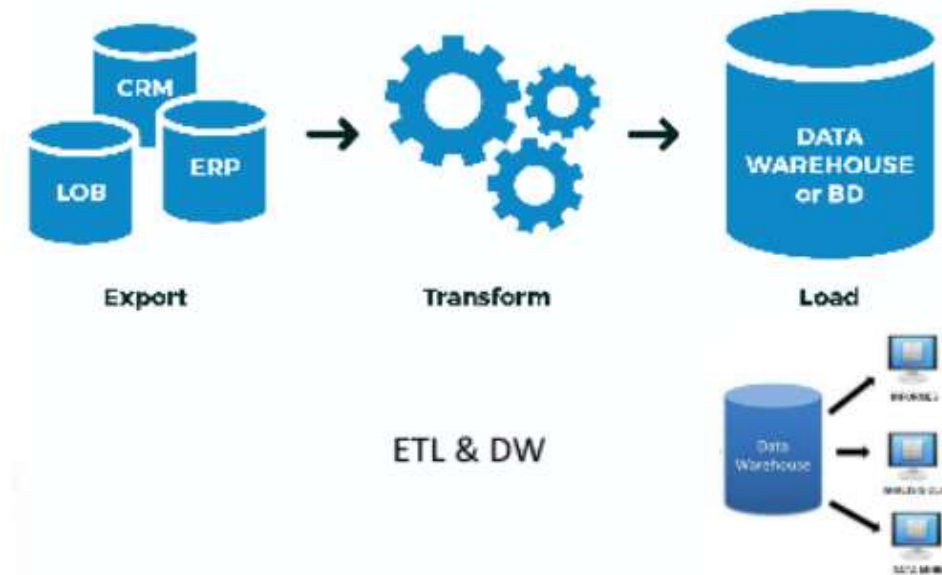


Fuente de datos



Procesos ETL

Introducción a Data Warehousing: Revisar los conceptos del proceso ETL para introducir data warehousing

Mg. Pablo Recalde

S3
uisrael.edu.ec



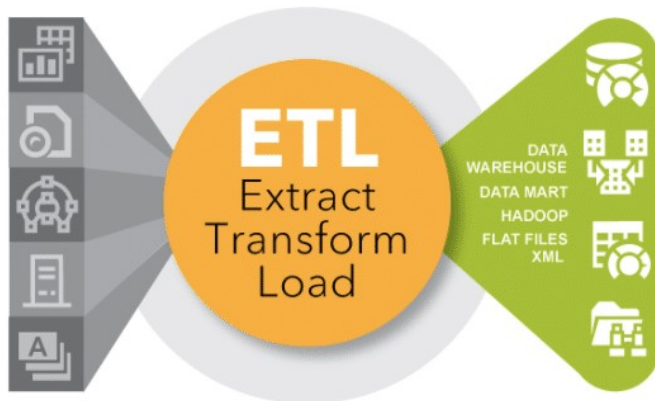
¿Qué es un Data Warehouse?

Un Data Warehouse (DW): En el contexto de la informática, un almacén de datos es una colección de datos orientada a un determinado ámbito, integrado, no volátil y variable en el tiempo, que ayuda a la toma de decisiones en la entidad en la que se utiliza





ETL: Extract, Transform and Load es el proceso que permite a las organizaciones mover datos desde múltiples fuentes, reformatearlos y limpiarlos, y cargarlos en otra base de datos, data mart, o data warehouse para analizar, o en otro sistema operacional para apoyar un proceso de negocio.



ELT: Extract, Load and Transform es un método diferente de acercarse al flujo de datos, en el que los datos extraídos se cargan primero en el sistema de destino. Las transformaciones se realizan después de que carguemos los datos en el almacén de datos. En lugar de transformar los datos antes de que se escriban, ELT permite que el sistema de destino realice la transformación. Los datos primero se copian en el data lake y luego se transforman in situ. El ELT generalmente funciona bien cuando el sistema objetivo es lo suficientemente potente como para manejar transformaciones a gran escala. ELT generalmente se usa con bases de datos NOSQL como el clúster de Hadoop, un dispositivo de datos o una instalación en la nube. Las bases de datos analíticas como Amazon Redshift o Google BigQuery se usan a menudo en pipelines ELT porque son altamente eficientes para realizar transformaciones.

Diferencias ETL y ELT

Parámetros	ETL	ELT
Procesamiento	Los datos se transforman en el servidor de almacenamiento intermedio y luego se transfieren al Datawarehouse.	Los datos permanecen en el data lake.
Código de uso	Usado para: – Transformaciones de computación intensiva – Pequeña cantidad de datos	Cantidades grandes de datos
Transformación	Las transformaciones se realizan en el servidor ETL / área de ensayo.	Las transformaciones se realizan en el sistema de destino.
Tiempos de carga	Los datos se cargan primero en el almacenamiento intermedio y luego se mueven al sistema objetivo. Tiempo intensivo.	Los datos cargados en el sistema de destino solo una vez. Más rápido.
Tiempos de Transformación	El proceso ETL necesita esperar a que se complete la transformación. A medida que crece el tamaño de los datos, aumenta el tiempo de transformación.	En el proceso ELT, la velocidad nunca depende del tamaño de los datos.
Tiempos de mantenimiento	Necesita altos niveles de mantenimiento ya que necesita seleccionar datos para cargar y transformar.	Bajo mantenimiento ya que los datos están siempre disponibles.
Complejidad de implementación	En una etapa temprana, es más fácil de implementar.	Para implementar el proceso de ELT, la organización debe tener un conocimiento profundo de las herramientas y los skills necesarios.
Soporte para Data warehouse	Modelo de ETL utilizado para datos locales, relacionales y estructurados.	Se utiliza en una infraestructura de cloud escalable que admite orígenes de datos estructurados y no estructurados.
Soporte Data lake	No soportado.	Permite usar un Data lake con datos no estructurados.

Diferencias ETL y ELT

Complejidad	El proceso ETL carga solo los datos importantes, como se identificaron en el momento del diseño.	Este proceso implica el desarrollo desde la salida hacia atrás y la carga de solo datos relevantes.
Costes	Costes elevados para pequeñas y medianas empresas.	Bajos costes de entrada utilizando plataformas de Software as a Service.
Búsquedas	En el proceso de ETL, tanto los hechos como las dimensiones deben estar disponibles en el área de preparación.	Todos los datos estarán disponibles porque la extracción y la carga se producen en una sola acción.
Agregaciones	Aumento de la complejidad con la cantidad adicional de datos en el conjunto de datos.	El poder de la plataforma de destino puede procesar una cantidad significativa de datos rápidamente.
Cálculos	Sobrescribe la columna existente o necesidad de adjuntar el conjunto de datos y empujar a la plataforma de destino.	Permite agregar fácilmente la columna calculada a la tabla existente.
Madurez	El proceso se utiliza desde hace más de dos décadas. Está bien documentado y las mejores prácticas fácilmente disponibles.	Concepto relativamente nuevo y complejo de implementar.
Hardware	La mayoría de las herramientas tienen requisitos de hardware únicos que son caros.	En formato SaaS el coste de hardware no es un factor crucial.
Soporte para datos no estructurados	En su mayoría soporta datos relacionales	Soporte para datos no estructurados fácilmente disponibles.

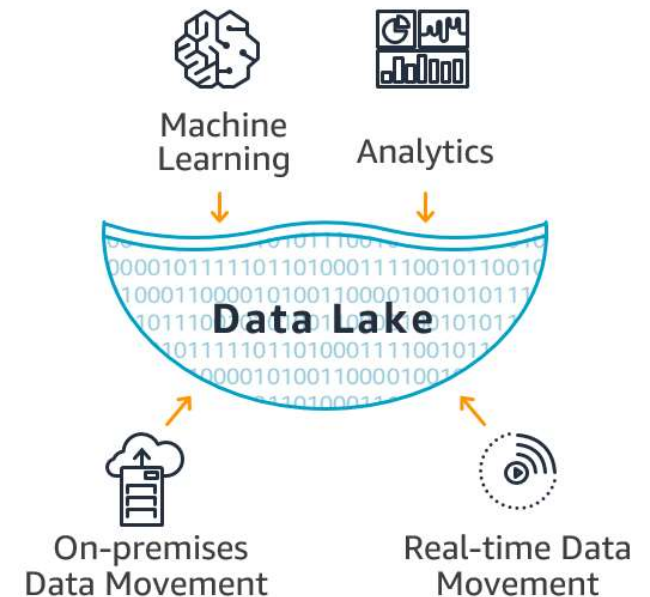


¿Qué es Data Lake?

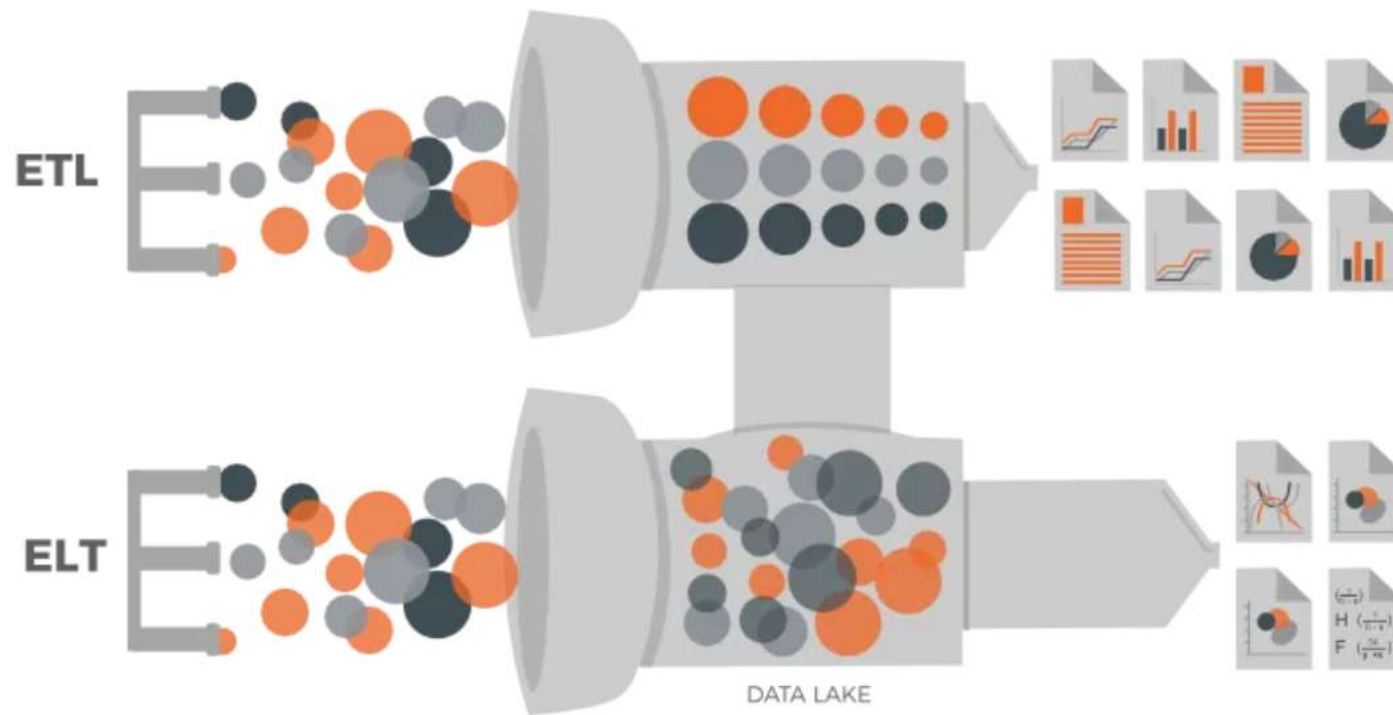
Un data lake es un repositorio de almacenamiento que contienen una gran cantidad de datos en bruto y que se mantienen allí hasta que sea necesario. *A diferencia de un data warehouse jerárquico que almacena datos en ficheros o carpetas, un data lake utiliza una arquitectura plana para almacenar los datos.*

A cada elemento de un data lake se le asigna un identificador único y se etiqueta con un conjunto de etiquetas de metadatos extendidas. Cuando se presenta una cuestión de negocios que debe ser resuelta, podemos solicitarle al data lake los datos que estén relacionados con esa cuestión. Una vez obtenidos podemos analizar ese conjunto de datos más pequeño para ayudar a obtener una respuesta.

El data lake se asocia a menudo con el almacenamiento de objetos orientado a **Hadoop**. En este escenario, los datos de una organización se cargan primero en la plataforma Hadoop y, a continuación, se aplican las herramientas de análisis y de minería de datos a los datos que residen en los nodos clúster de Hadoop.



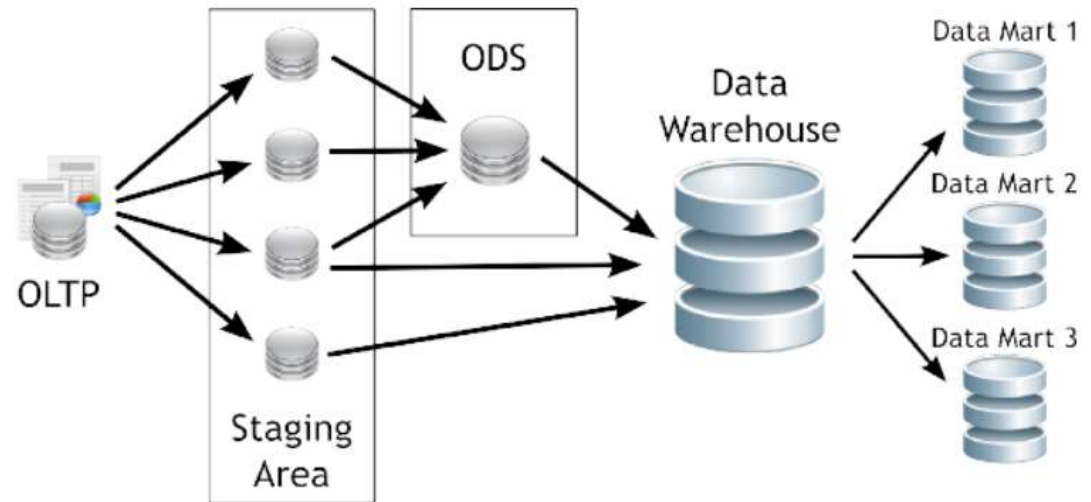
Resumiendo los dos tipos de transformación:



Áreas de datos del sistema

Staging Area

Es un área temporal donde se recogen los datos que se necesitan de los sistemas origen. Se recogen los datos estrictamente necesarios para las cargas, y se aplica el mínimo de transformaciones a los mismos. No se aplican restricciones de integridad ni se utilizan claves, los datos se tratan como si las tablas fueran ficheros planos. De esta manera se minimiza la afectación a los sistemas origen, la carga es lo más rápida posible para acotar la ventana horaria necesaria, y se reduce también al mínimo la posibilidad de error.

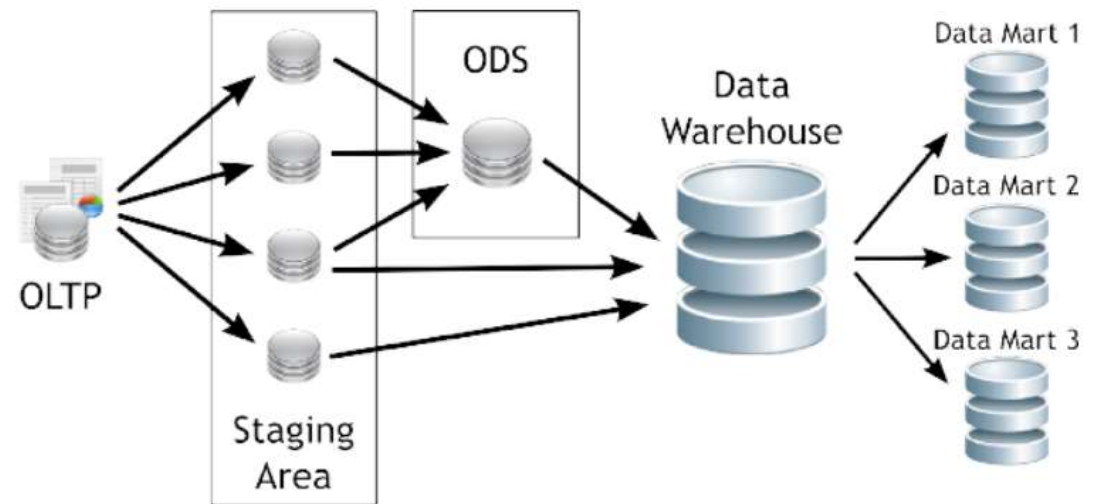


Una vez que los datos han sido traspasados, el DW se independiza de los sistemas origen hasta la siguiente carga. Lo único que se suele añadir es algún campo que almacene la fecha de la carga.

Operational Data Store (ODS)

Es la que da soporte a los sistemas operacionales. El modelo de datos del Almacén de Datos Operacional sigue una estructura relacional y normalizada, para que cualquier herramienta de reporting o sistema operacional pueda consultar sus datos. Está dentro del Data Warehouse porque se aprovecha el esfuerzo de integración que supone la creación del Almacén de Datos Corporativo para poder atender también a necesidades operacionales, pero no es obligatorio. Ni siquiera es algo específico del BI, los ODS ya existían antes de que surgieran los conceptos de Data Warehousing y Business Intelligence.

No almacena datos históricos, muestra la imagen del momento actual, aunque eso no significa que no se puedan registrar los cambios.



Los datos del ODS se recogen de la Staging Area, y en este proceso sí que se realizan transformaciones, limpieza de datos y controles de integridad referencial para que los datos estén perfectamente integrados en el modelo relacional normalizado.



Almacén de Datos Corporativo (DW)

El Almacén de Datos Corporativo sí que contiene datos históricos, y está orientado a la explotación analítica de la información que recoge. Las herramientas DSS o de reporting analítico consultan tanto los Data marts como el Almacén de Datos Corporativo. El DW puede servir consultas en las que se precisa mostrar a la vez información que se encuentre en diferentes Datamarts.

En él se almacenan datos que pueden provenir tanto de la Staging Area como del ODS. Si ya se realizan procesos de transformación e integración en el ODS no se repiten para pasar los mismos datos al Almacén Corporativo. Lo que no se pueda recoger desde el ODS sí que hay que ir a buscarlo a la Staging Area.

Data Warehouse



El esquema se parece al de un modelo relacional normalizado, pero en él ya se aplican técnicas de desnormalización. No debería contener un número excesivo de tablas ni de relaciones ya que, por ejemplo, muchas relaciones jerárquicas que en un modelo normalizado se implementarían con tablas separadas aquí ya deberían crearse en una misma tabla, que después representará una dimensión.

Otra particularidad es que la mayoría de las tablas han de incorporar campos de fecha para controlar la fecha de carga, la fecha en que se produce un hecho, o el periodo de validez del registro.

Si el Data Warehouse no es demasiado grande, o el nivel de exigencia no es muy elevado en cuanto a los requerimientos 'operacionales', para simplificar la estructura se puede optar por prescindir del ODS, y si es necesario adecuar el Almacén de Datos Corporativo para servir tanto al reporting operacional como al analítico. En este caso, el área resultante sería el DW Corporativo, pero en ocasiones también se denomina como ODS.



Data Mart (DM)

Otro área de datos es el lugar donde se crean los Data marts. Éstos acostumbran a obtenerse a partir de la información recopilada en el área del Almacén Corporativo, aunque también puede ser a la inversa. Cada Data Mart es como un subconjunto de este almacén, pero orientado a un tema de análisis, normalmente asociado a un departamento de la empresa.

El Data Mart se diseña con estructura multidimensional, cada objeto de análisis es una tabla de hechos enlazada con diversas tablas de dimensiones. Si se diseña siguiendo el Modelo en Estrella habrá prácticamente una tabla para cada dimensión, es la versión más desnormalizada.

Si se sigue un modelo de Copo de Nieve las tablas de dimensiones estarán menos desnormalizadas y para cada dimensión se podrán utilizar varias tablas enlazadas jerárquicamente.

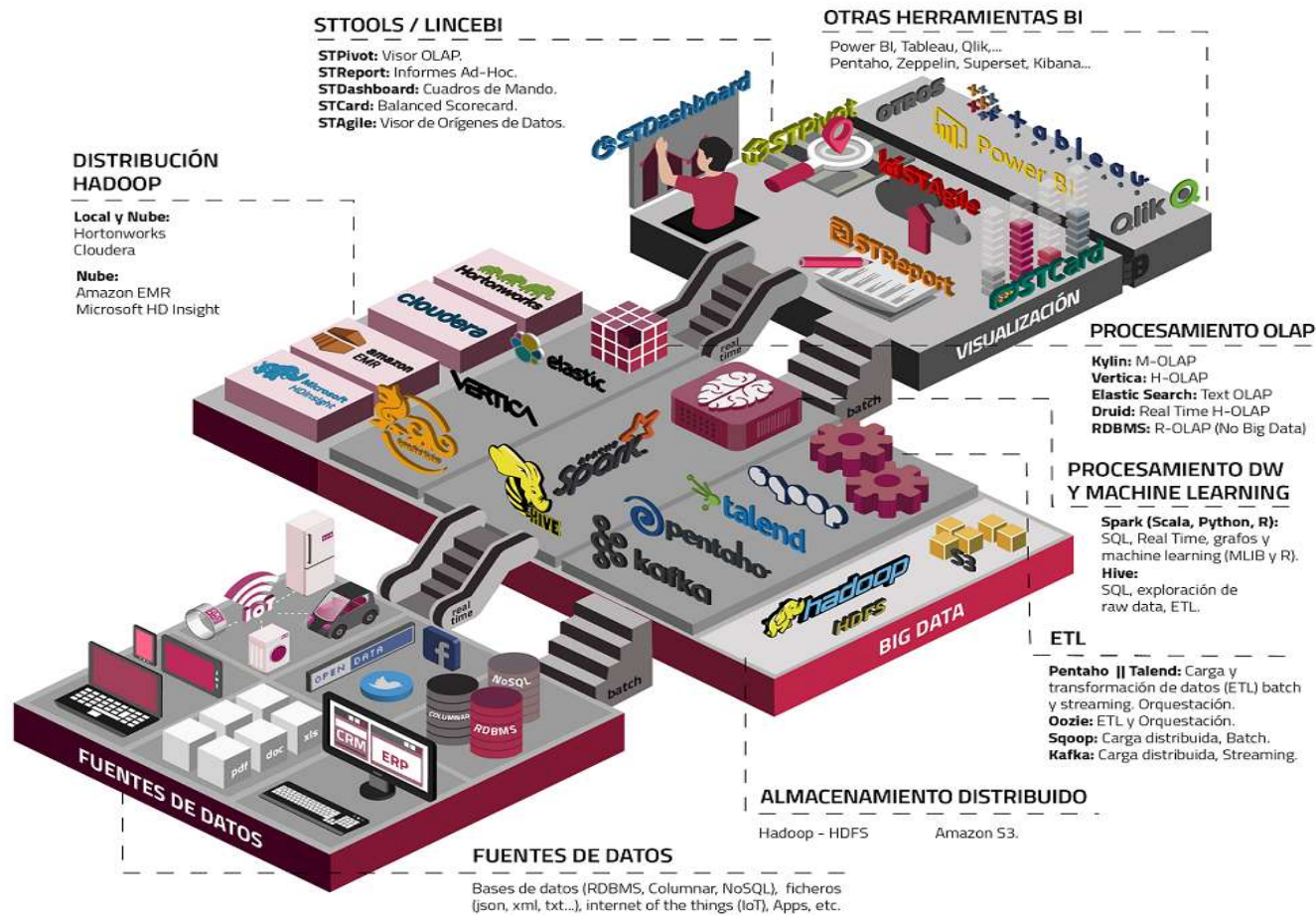
Este área puede residir en la misma base de datos que las demás si la herramienta de explotación es de tipo ROLAP, o también puede crearse ya fuera de la BD, en la estructura de datos propia que generan las aplicaciones de tipo MOLAP, más conocida como los cubos multidimensionales.



Si se sigue una aproximación Top-down para la creación de los Data mart, el paso del área de DW a esta ha de ser bastante simple, cosa que además proporciona una cierta independencia sobre el software que se utiliza para el reporting analítico. Si por cualquier razón es necesario cambiar la herramienta de OLAP hay que hacer poco más que redefinir los metadatos y regenerar los cubos, y si el cambio es entre dos de tipo ROLAP ni siquiera esto último sería necesario. En cualquier caso, las áreas anteriores no tienen porqué ser modificadas.



Herramientas para ETL y DW



Talend

Descargas

Buscar en Descargas


Nombre	Fecha de modificación	Tipo
Talend-DataPreparation-Free-Desktop-2.5.1.exe		

Setup

talend

Setup - Talend Data Preparation 2.5.1

Welcome to the Talend Data Preparation 2.5.1 Setup Wizard.



**TALEND DATA
PREPARATION
FREE DESKTOP**

< Back Next > Cancel



Talend

127.0.0.1:9090/#/home/preparations/

DATA PREPARATION

Q | ? Help | i

Home

PREPARATIONS

DATASETS

ADD PREPARATION

ADD FOLDER

Display

NAME ^	AUTHOR	CREATED	MODIFIED	DATASET	STEPS
Clean HRMS Data	precalde	a few seconds ago	a few seconds ago	HRMS Export	2
Create Email Address	precalde	a few seconds ago	a few seconds ago	Marketing Leads	21
CRM Phone Numbers	precalde	a few seconds ago	a few seconds ago	CRM Export	7
Marketing Upload	precalde	a few seconds ago	a few seconds ago	Customer Contact Data	9

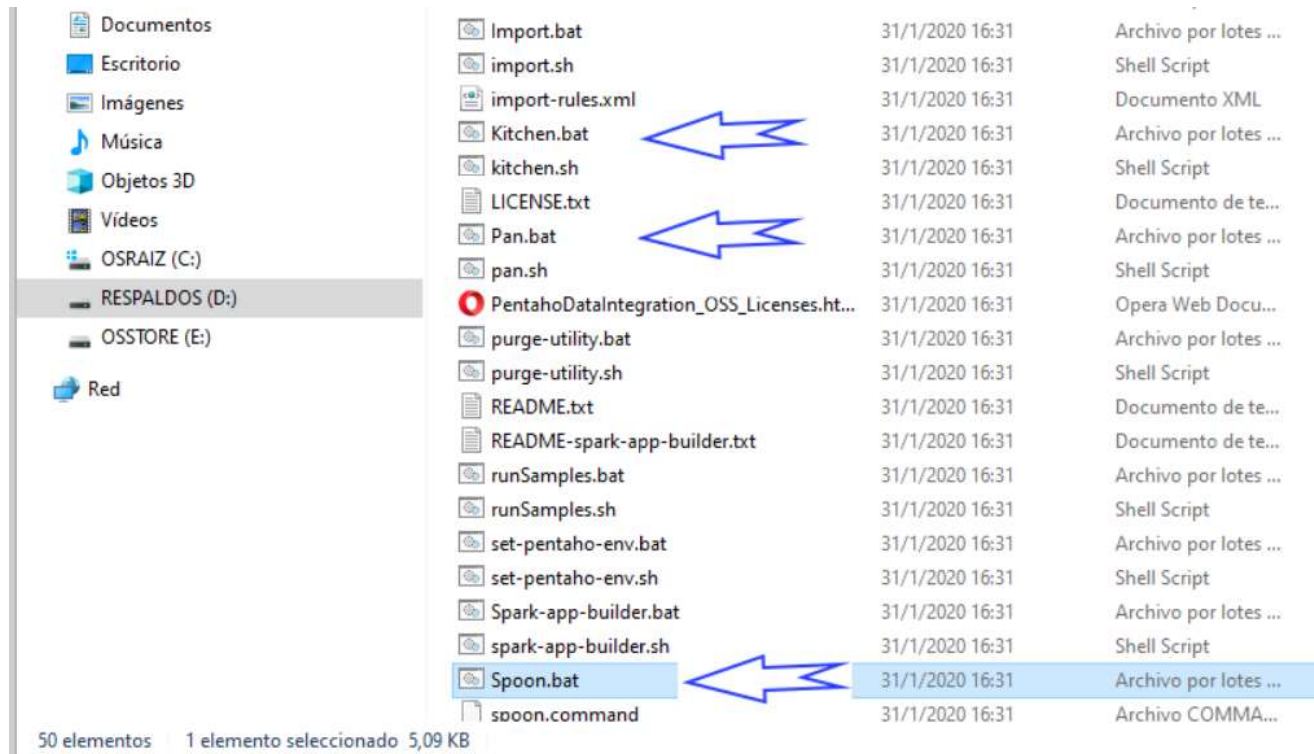
Pentaho Data Integration: PDI

← → ↕ ↗ > Este equipo > RESPALDOS (D:) > pdi-ce-9.0.0.0-423 > data-integration

Nombre	Fecha de modificación	Tipo
.telemetry	17/11/2020 23:30	Carpeta de archivos
ADDITIONAL-FILES	31/1/2020 18:38	Carpeta de archivos
classes	31/1/2020 16:31	Carpeta de archivos
Data Integration.app	31/1/2020 16:31	Carpeta de archivos
Data Service JDBC Driver	31/1/2020 18:38	Carpeta de archivos
docs	31/1/2020 16:31	Carpeta de archivos
drivers	26/11/2020 10:26	Carpeta de archivos
launcher	31/1/2020 16:31	Carpeta de archivos
lib	26/11/2020 10:18	Carpeta de archivos
libswt	31/1/2020 18:38	Carpeta de archivos
logs	30/11/2020 11:44	Carpeta de archivos
plugins	31/1/2020 16:32	Carpeta de archivos
pwd	31/1/2020 16:31	Carpeta de archivos
samples	31/1/2020 16:31	Carpeta de archivos
simple-jndi	31/1/2020 16:31	Carpeta de archivos
static	31/1/2020 16:31	Carpeta de archivos
system	31/1/2020 18:38	Carpeta de archivos
ui	31/1/2020 16:31	Carpeta de archivos
Carte.bat	31/1/2020 16:31	Archivo por lotes ...
carte.sh	31/1/2020 16:31	Shell Script
Encr.bat	31/1/2020 16:31	Archivo por lotes ...
encr.sh	31/1/2020 16:31	Shell Script

←

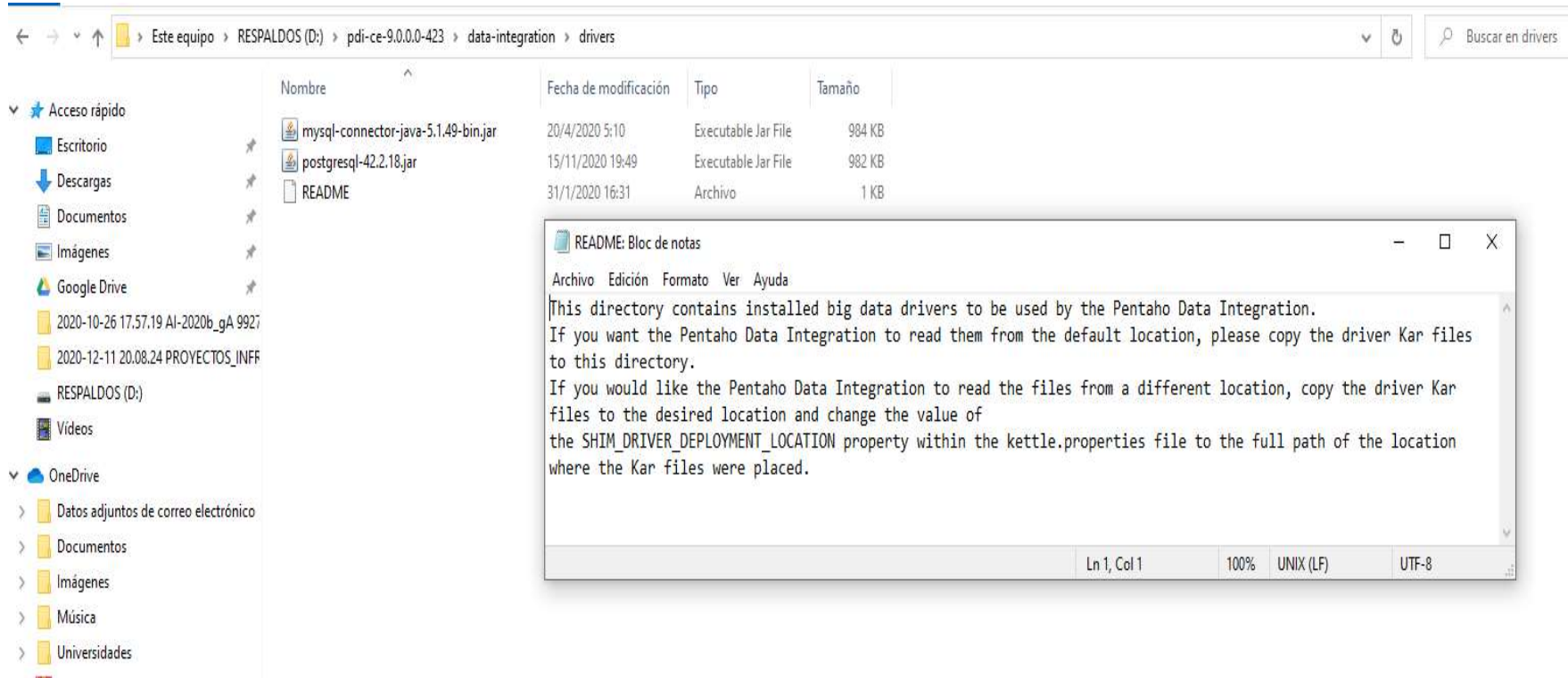
Pentaho Data Integration: PDI



Nombre	Fecha y hora	Tipo
Import.bat	31/1/2020 16:31	Archivo por lotes ...
import.sh	31/1/2020 16:31	Shell Script
import-rules.xml	31/1/2020 16:31	Documento XML
Kitchen.bat	31/1/2020 16:31	Archivo por lotes ...
kitchen.sh	31/1/2020 16:31	Shell Script
LICENSE.txt	31/1/2020 16:31	Documento de te...
Pan.bat	31/1/2020 16:31	Archivo por lotes ...
pan.sh	31/1/2020 16:31	Shell Script
PentahoDataIntegration_OSS_Licenses.ht...	31/1/2020 16:31	Opera Web Docu...
purge-utility.bat	31/1/2020 16:31	Archivo por lotes ...
purge-utility.sh	31/1/2020 16:31	Shell Script
README.txt	31/1/2020 16:31	Documento de te...
README-spark-app-builder.txt	31/1/2020 16:31	Documento de te...
runSamples.bat	31/1/2020 16:31	Archivo por lotes ...
runSamples.sh	31/1/2020 16:31	Shell Script
set-pentaho-env.bat	31/1/2020 16:31	Archivo por lotes ...
set-pentaho-env.sh	31/1/2020 16:31	Shell Script
Spark-app-builder.bat	31/1/2020 16:31	Archivo por lotes ...
spark-app-builder.sh	31/1/2020 16:31	Shell Script
Spoon.bat	31/1/2020 16:31	Archivo por lotes ...
spoon.command	31/1/2020 16:31	Archivo COMMA...

50 elementos 1 elemento seleccionado 5,09 KB

Pentaho Data Integration: PDI



Este equipo > RESPALDOS (D:) > pdi-ce-9.0.0.0-423 > data-integration > drivers

Nombre	Fecha de modificación	Tipo	Tamaño
mysql-connector-java-5.1.49-bin.jar	20/4/2020 5:10	Executable Jar File	984 KB
postgresql-42.2.18.jar	15/11/2020 19:49	Executable Jar File	982 KB
README	31/1/2020 16:31	Archivo	1 KB

README: Bloc de notas

Archivo Edición Formato Ver Ayuda

This directory contains installed big data drivers to be used by the Pentaho Data Integration.

If you want the Pentaho Data Integration to read them from the default location, please copy the driver Kar files to this directory.

If you would like the Pentaho Data Integration to read the files from a different location, copy the driver Kar files to the desired location and change the value of the SHIM_DRIVER_DEPLOYMENT_LOCATION property within the kettle.properties file to the full path of the location where the Kar files were placed.

Ln 1, Col 1 100% UNIX (LF) UTF-8



Spoon es el diseñador gráfico de transformaciones y trabajos del sistema de ETTLS de Pentaho Data Integration (PDI), también conocido como Kettle (acrónimo de "Kettle E.T.T.L. Environment"). Está diseñado para ayudar en los procesos ETTLS, que incluyen la Extracción, Transformación, Transporte y Carga de datos.

Pan es un motor de transformación de datos que realiza muchas funciones tales como lectura, manipulación, y escritura de datos hacia y desde varias fuentes de datos.

Kitchen es un programa que ejecuta los Trabajos diseñados por Spoon en XML o en un catálogo de base de datos.

Los Trabajos normalmente se planifican en **modo batch** (por lotes) para ejecutarlos automáticamente en intervalos regulares.

Las Transformaciones y Trabajos se pueden describir usando un archivo XML o se pueden colocar en un catálogo de base de datos de Kettle.

Luego Pan o Kitchen pueden leer los datos para ejecutar los pasos que se describen en la Transformación o ejecutar el Trabajo.

En resumen, PDI facilita la construcción, actualización, y mantenimiento de Data Warehouses.

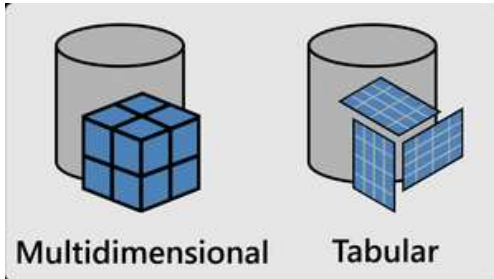
Modelos OLAP tienen como principal característica que cumplir **FASMI (Fast Analysis of Shared Multidimensional Information)** y en este orden:

- 1. Fast:** el sistema tiene que entregar la mayor cantidad de datos en el menor tiempo posible en aproximadamente 5 segundos. Para el análisis básico y elemental, no debe tardar más de 1 segundo y pocas veces puede llegar a ascender a 20 segundos.
- 2. Analysis:** lógica de negocio relevante y análisis de información que se simple para los analistas de negocio no expertos.
- 3. Shared:** poder gestionar múltiples actualizaciones de forma segura y rápida.
- 4. Multidimensional:** requisito básico, el sistema resultado debe proporcionar una vista conceptual multidimensional de los datos, es decir, podemos combinar dimensiones para obtener el resultado (valor) buscado.
- 5. Information:** el sistema debe contener todos los datos necesarios para las aplicaciones.

Modelo TABULAR.

Este modelo se basa en bases de datos in-memory (en memoria) que utilizan el motor analítico de proceso y almacenamiento xVelocity (inicialmente denominado VertiPaq). Este motor utiliza un almacenamiento basado en columnas y sofisticados algoritmos de compresión para ofrecernos tiempos de consulta muy rápidos, incluso con grandes volúmenes de datos. El modelo tabular, muestra los datos en formato relacional, es decir, interactuamos con tablas y relaciones en lugar de cubos. Además, podemos decir que el rendimiento de xVelocity puede superar al rendimiento de índices ColumnStore (creados en la estructura de las tablas de la base de datos) lógicamente al tratarse de un modelo en memoria. La arquitectura ideal para un rendimiento ultra rápido, sería poder disponer de ColumnStore en la BDs.

Los modelos tabulares son adecuados tanto para aplicaciones personales como para soluciones departamentales con grandes volúmenes de datos.

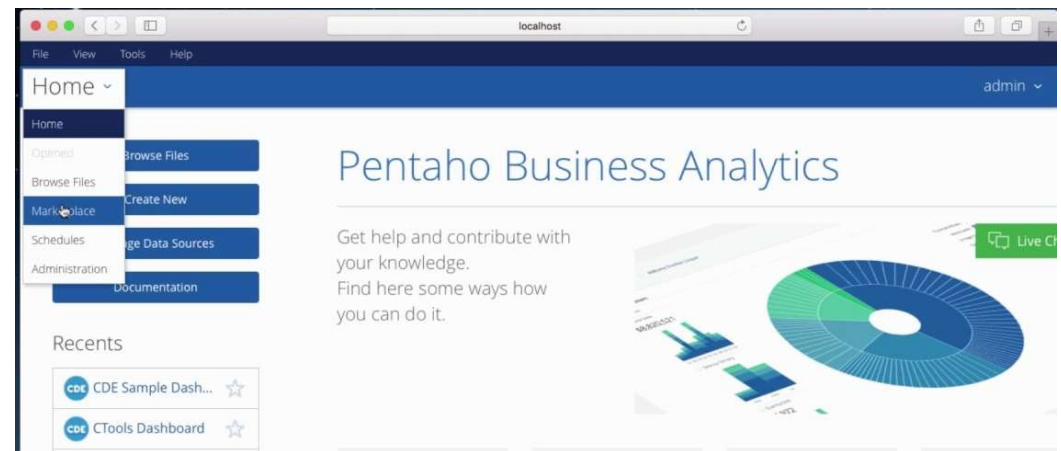


Como conclusión

El modelo óptimo para montar un cubo multidimensional sería **siempre MOLAP**. En cualquier otro caso, ir a un modelo tabular en lugar de ROLAP o HOLAP, sobre todo si luego las aplicaciones finales estarán desarrolladas con Power BI.

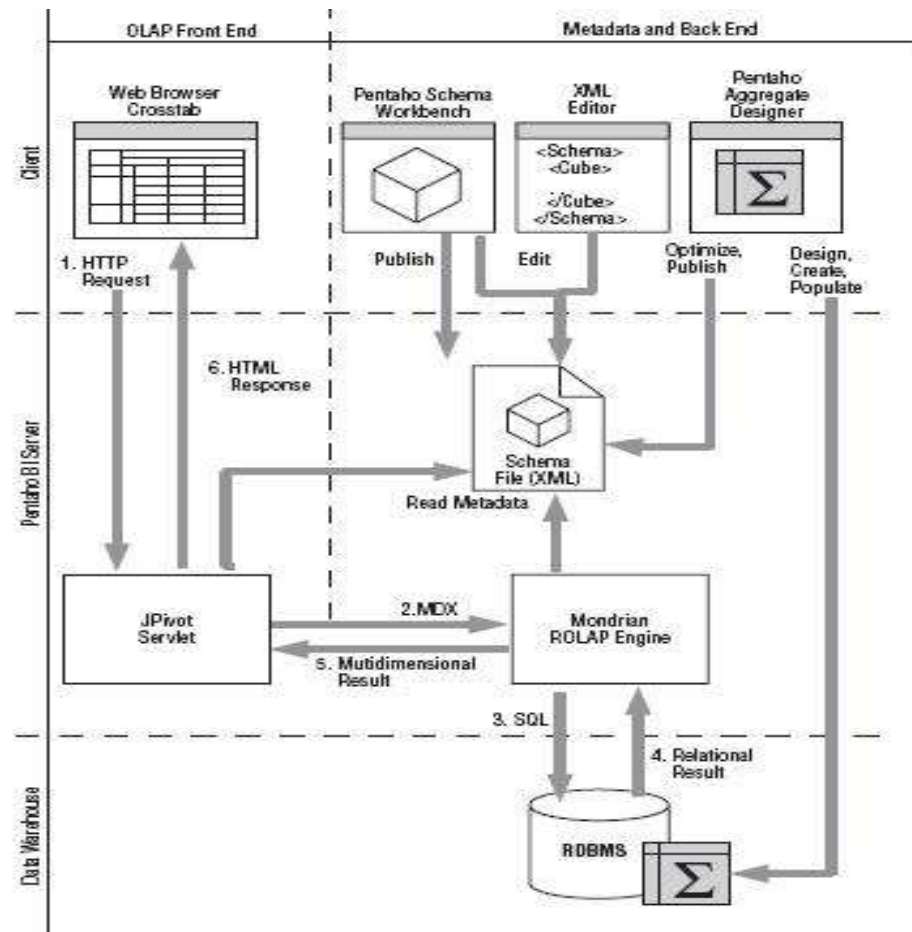
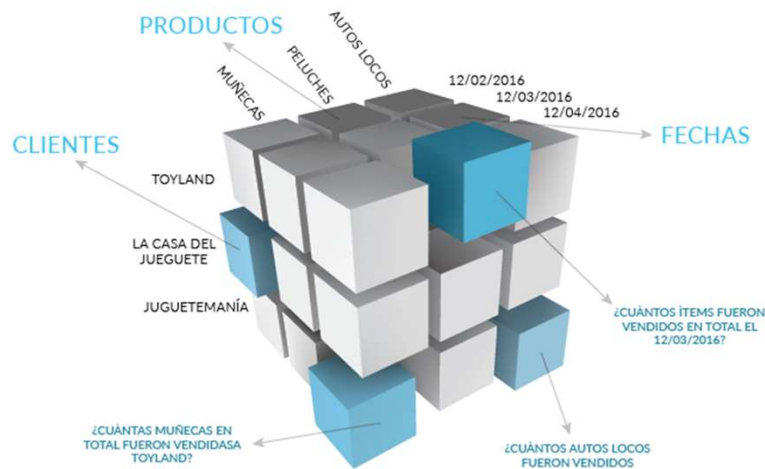
SERVIDOR OLAP Mondrian

Mondrian es una de las aplicaciones más importantes de la plataforma Pentaho BI. Mondrian es un servidor OLAP open source que gestiona comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuente. Es decir, Mondrian actúa como “JDBC para OLAP”.



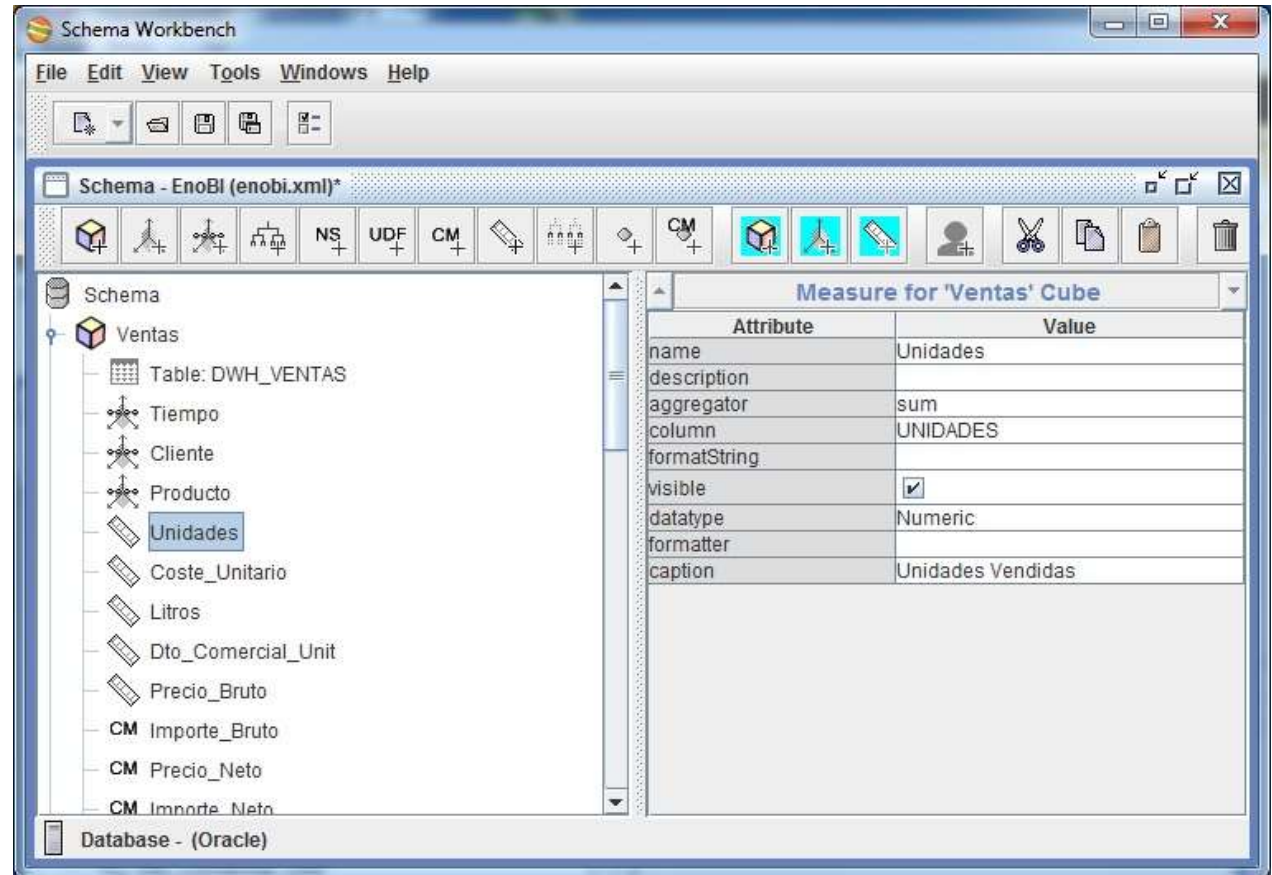
Arquitectura Pentaho Analysis Services

Pentaho Schema Workbench



SERVIDOR OLAP Mondrian

Mondrian es una de las aplicaciones más importantes de la plataforma Pentaho BI. Mondrian es un servidor OLAP open source que gestiona comunicación entre una aplicación OLAP (escrita en Java) y la base de datos con los datos fuente. Es decir, Mondrian actúa como “JDBC para OLAP”.





Mini-Taller

GRACIAS

