

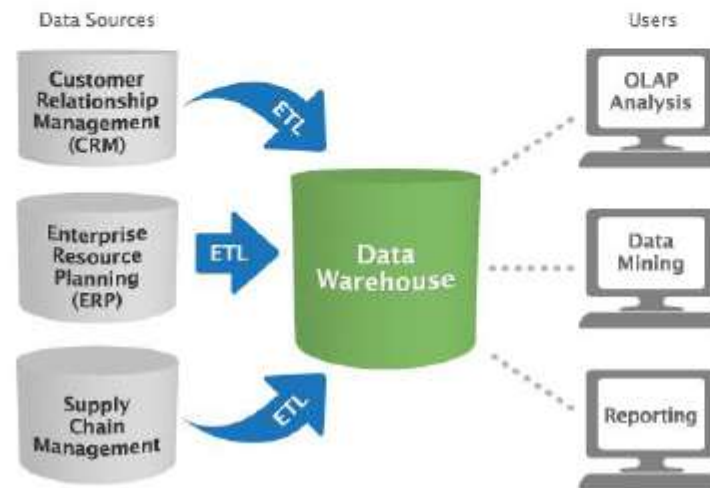


Introducción a Big Data

Fundación Telefónica Movistar
CURSO DE INTRODUCCIÓN A BIG DATA.

Data Warehouse (DW)

*“Es un repositorio de datos **integrado, no volátil, variable en el tiempo, orientado al negocio**, organizado de forma tal que facilita el análisis de grandes volúmenes de datos para la toma de decisiones”.*



Objetivos de un Data Warehouse

- Proporciona una herramienta para la **toma de decisiones** en cualquier área funcional, basándose en información integrada y global del negocio.
- Facilita la aplicación de **técnicas estadísticas de análisis y modelización** para encontrar relaciones ocultas entre los datos del almacén; obteniendo un valor añadido para el negocio de dicha información.
- Proporciona la capacidad de **aprender de los datos del pasado y de predecir situaciones futuras** en diversos escenarios.

Características de un Data Warehouse

Integrado

- Información proveniente de sistemas heterogéneos. (BD, excels, archivos planos, etc.)
- Procesos de integración de datos y limpieza de información. (unificación de formatos, códigos, etc.)

No volátil

- Los datos perduran en el tiempo (Sólo inserts y updates).

Variable en el tiempo

- Es un repositorio de datos históricos.
- El tiempo de conservación de los datos es mayor que en sistemas transaccionales.
- La fecha es un dato fundamental, para poder analizar en el tiempo.

Orientado al negocio

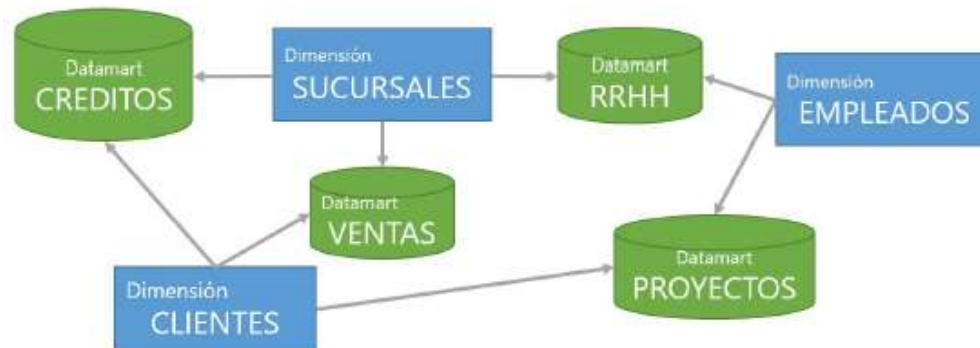
- Los datos están organizados y presentados como se manejan en el negocio.
- Los datos tienen el nivel de detalle y estructura necesarios para los que toman decisiones.



/ Datamarts & Data Warehouse

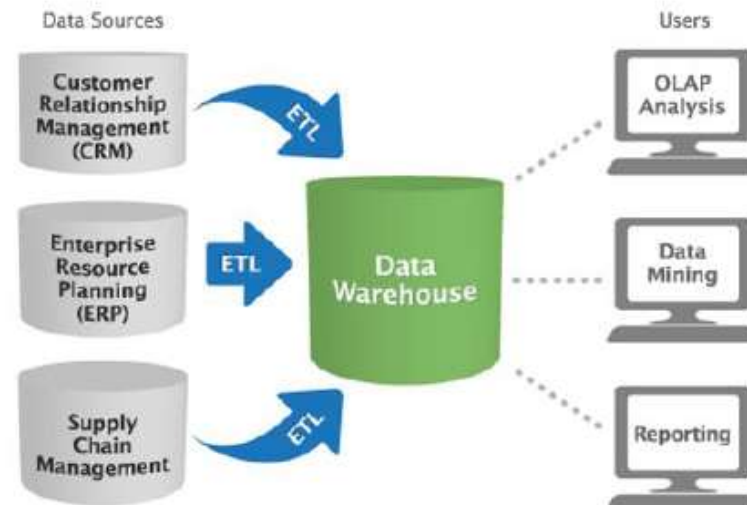
Datamart es una **partición o subconjunto del Data Warehouse**:


- Acotados a un proceso de negocio.
- De interés para un grupo de usuarios o área del negocio.
- Suelen tener alcance parecidos a los de un proyecto.



Componentes de la Arquitectura

- Source/data
- ETLs
- Data Warehouse
- Presentation
(Explotación & Visualización)





Data Warehouse - Ingesta (ETL)

- El **Datawarehouse** se alimenta mediante procesos **ETL** (extracción, transformación y carga), desde diferentes fuentes heterogéneas,
- El proceso ETL se comunica con los sistemas Transaccionales, archivos propios de los usuarios, bases de datos internas/externas, aplicaciones para extraer los datos y cargarlos.
- **Extracción:** obtención de información de las distintas fuentes tanto internas como externas.
- **Transformación:** filtrado, limpieza, depuración, homogeneización y agrupación de la información.
- **Carga:** organización y carga de los datos y los metadatos en la base de datos.

Antes de realizar la carga final, se pueden reformatear, limpiar, filtrar y posteriormente cargarlos en otra base de datos, en archivos de texto, o en diferentes tecnologías.

Data Lake – Definición

Un **Data Lake** es un repositorio centralizado que nos permite almacenar todos nuestros datos estructurados y no estructurados en cualquier escala, independientemente de su fuente o formato.

Normalmente se implementa, de manera On Premise utilizando la Plataforma Hadoop. Aunque existen distintas alternativas en la nube para armar nuestro Data Lake (por ej. Amazon S3 + Redshift / Amazon EMR / Azure HD insight)



Data Lake

Recomendaciones para su creación

Claves

- Contar con Datos crudos y datos modelados y refinados.
- Autenticación Centralizada con un Single Sign On.
- Autorización al acceso a datos clasificados.
- Data Governance.
- Auditoría y control.
- Integración con otras Plataformas.
- Contar con un Plan a Mediano Plazo.
- Tener en cuenta una estrategia para Recuperación Ante Desastres.



Data Warehouse vs Data Lake

Ingestar y enriquecer más datos para responder con más velocidad a las nuevas necesidades comerciales

Enterprise DW lives on

Con foco en información del negocio

Las plataformas tradicionales relacionales aún son las preferidas por los usuarios.

Data Lake complementa el DW

Con foco en datos externos estructurados y/o no estructurados

Se enlaza con el DW en múltiples maneras.

Se realiza el enriquecimiento de datos.

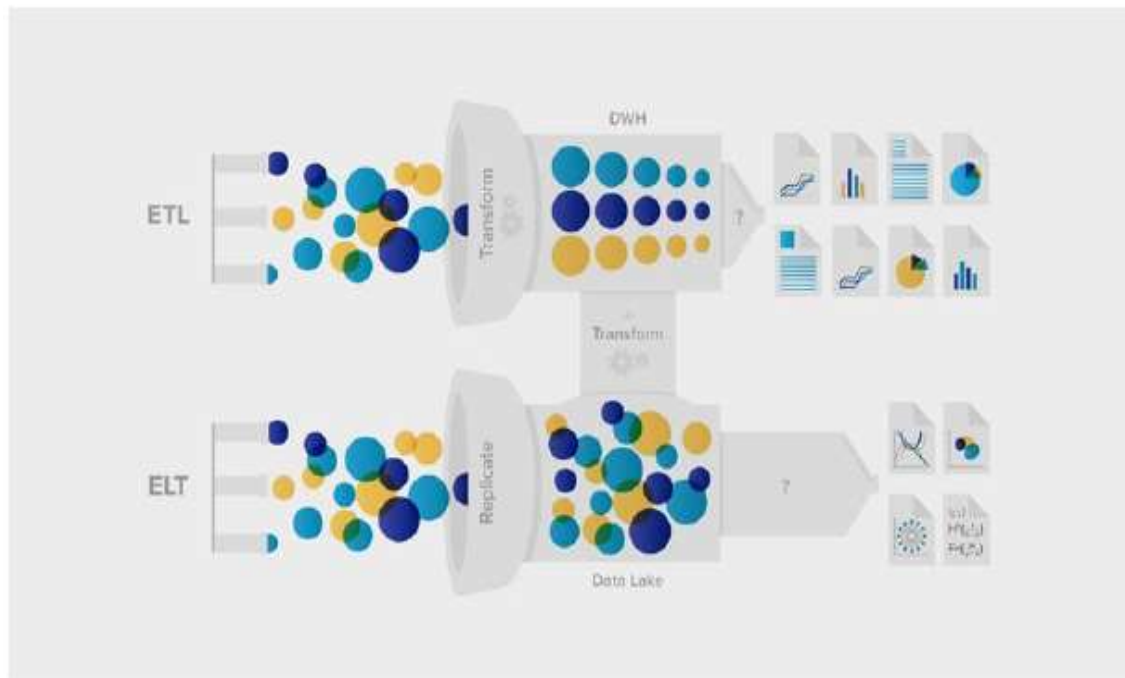
El Data Lake puede asistir/descargar al DW

Usar el almacenamiento y potencialidad de procesamiento reduce costos y mejora la performance.

DATA WAREHOUSE
LAKE



Data Warehouse



ETL

Extract Transform Load

vs.

ELT

Extract Load Transform



Sandboxes

El **Sandbox** es parte integral del Data Lake porque **permite a los data scientists y data engineers crear casos de uso exploratorios** ad hoc sin la necesidad de involucrar a la TI Departamento o dedicar recursos para crear entornos adecuados dentro de los cuales probar los datos.

Los datos se pueden **importar al sandbox desde** cualquiera de las **zonas**, así como directamente **de la fuente**. Dependiendo de los **permisos** con que cuente **el usuario** que realiza la acción.

Esto permite a las empresas **explorar, cómo, ciertas variables podrían afectar los resultados comerciales** y, por lo tanto, obtener más información para ayudar a hacer negocios decisiones de gestión.

Podrían enviarse un set de datos resultante directamente a la zona sin procesar, permitiendo que los datos derivados actúen como datos de origen.



Sandboxes

Utilización de carpetas de su Data Lake como SANDBOX Individual o Grupal.

El usuario podría:

- Realizar **Upload de archivos** externos.
- Crear **Tablas externas apuntando a las carpetas** en el sandbox.
- Realizar **Bajada de datos desde Tablas con** resultados puntuales.
- Utilizar toda la **potencia del Data Lake** con los datos distribuidos y replicados.



Se pueden definir **cuotas por cantidad de archivos creados** para cada carpeta del Sandbox o **por cantidad de bytes grabados**.

Nuevos Perfiles



- Data Analyst/Business Analyst

Se especializa en el negocio y en realizar análisis de datos, extrayendo conclusiones. En general no tienen mucho conocimiento técnico, pero utilizan herramientas de explotación y análisis gráficos para realizar su trabajo.

Habilidades: los analistas de datos deben tener una comprensión básica de algunas habilidades: recopilación de datos, visualización de datos, análisis exploratorio de datos.

Herramientas: Microstrategy, Qlik, Tableau, entre otras.





- Data Architect - orientado a modelado de datos (Data Modeler)

Se especializa en el **modelado de datos**, tanto **relacional o no relacional** y fundamentalmente **dimensional**.

Definen cómo los datos **serán ingestados y almacenados, conformados e integrados, administrados y explotados** por diferentes aplicaciones.

Responsable de la **capa semántica** para presentar la información en términos de negocio.

Herramientas: **MicroStrategy, Governance Catalog, Bus Matrix** (para documentar hechos y dimensiones), Diagrama ER en **Erwin**.





- Data Architect - orientado a tecnologías y plataformas

Se especializa en el conocimiento de tecnologías desarrolladas para Big Data, encargándose de diseñar la topología, definir las soluciones, configurar, mantener y soportar las plataformas tecnológicas.

Definen cómo los datos serán ingestados, almacenados, consumidos, integrados, administrados y explotados por diferentes aplicaciones, asegurando la integración de todas las plataformas con las bases de datos y tecnologías tradicionales existentes en la organización.

Habilidades: Tiene conocimiento avanzado en plataformas de almacenamiento y procesamiento de datos distribuidos tanto on premise como cloud.

Exhaustivo conocimiento de la arquitectura de la base de datos NoSQL.
Conocimiento de Herramientas de Extracción, transformación y carga (ETL/ELT) Batch, microbatch, near real time y real time.

Herramientas: Plataforma Hadoop, Nifi, Kafka, Spark, MapReduce, Hive, MongoDB, otras Bases de datos relacionales y nosql, entre otros





- Data Engineer - Funcional

Se encargan de analizar requerimientos y necesidades, investigar potenciales fuentes de datos, diseñar modelos y procesos de carga/ingesta.


Entre sus principales tareas se incluye investigar fuentes de datos, descubrir patrones y relaciones, modelar dimensionalmente procesos de negocio, establecer los mapeos desde las fuentes de datos hasta el modelo propuesto (matriz SourceToTarget), diseñar procesos de preparación de datos contemplando la extracción, limpieza, conformación y delivery para su posterior uso analítico u operacional.



Data Engineer - Técnico


Se encargan del **desarrollo** e implementación de **modelos y algoritmos**, además de encargarse de **extraer, transformar, ingestar y gestionar los datos** requeridos en los distintos proyectos.

Entre sus principales tareas se incluye la **preparación de datos** para su **posterior uso analítico** u operacional. Esta preparación implica la de **construir y automatizar pipelines** para buscar datos de diferentes fuentes, ingestarlos, integrarlos, consolidarlos, limpiarlos y enriquecerlos para su uso posterior.



Habilidades: Conocimiento de sistemas de base de Datos Relacionales y noSQL. Modelado de datos y herramientas ETL/ELT Batch y Near real time. Desarrollo de APIs para acceso a datos. Desarrollar y utilizar plataformas de almacenamiento y procesamiento distribuido.

Herramientas: BD Relacionales, BD NoSQL, Hadoop, Spark, MapReduce, SQL, Hive, Sqoop, Kafka, Nifi, entre otros.



Lenguajes: Python, JavaScript, Java, Scala entre otros.



Data Scientist

Se encarga de realizar análisis y extracción de conocimiento y conclusiones a partir de grandes volúmenes de datos.

Para realizar dicho análisis se basa en utilizar y combinar la informática, con matemáticas y estadística, debiendo comunicar el conocimiento extraído de los datos y la visión de negocio.

Habilidades: Conocimiento de Procesamiento distribuido, algoritmos predictivos, matemáticas, estadísticas, algoritmos orientados a machine learning, buena comunicación y conocimiento del negocio, entre otros.

Herramientas: BD Relacionales, BD NoSQL, Hadoop, Spark, MapReduce, Hive, Notebook Jupyter, Notebook Zeppelin, entre otros.

Lenguajes: R, Python, Scala.



