

INFRAESTRUCTURA Y ARQUITECTURA DE TI

Ángela Villota Gómez
apvillota@icesi.edu.co



1

INTRODUCCIÓN

Contexto

2

DATA WAREHOUSES

Concepts

3

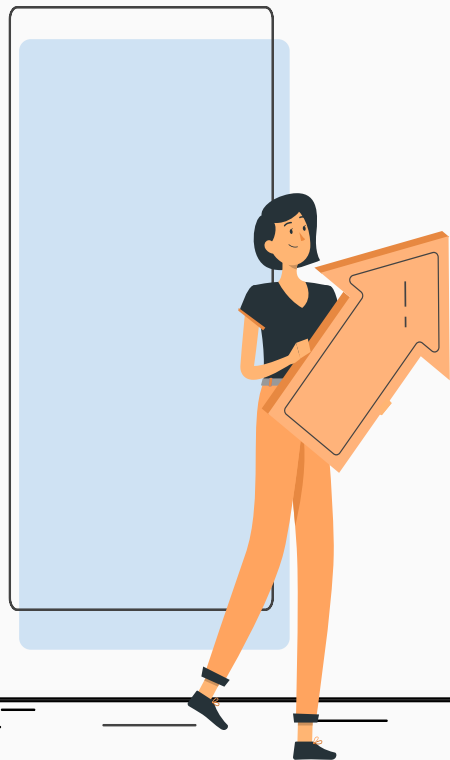
DATA LAKE

Concepts

4

PRÁCTICA

DW



INTRODUCCIÓN



LA SEMANA PASADA

Temas

1. Gestión de datos:
Conceptos generales de los patrones de arquitectura
 - Arquitectura de procesamiento por lotes
2. Almacenamiento y procesamiento de datos (Conceptos):
 - Tipos de procesamiento de datos (OLTP y OLAP)
 - Bases de datos relacionales (práctica)

Tareas:

- Quiz de la unidad 1 (para el inicio de la Sesión 3)
- Terminar el ejercicio del modelo relacional
- Mapa mental de la lectura

RECURSOS DE ESTA CLASE

En la carpeta de la Semana 3 Lecturas:

- Data warehousing quick guide
- CH 32, 33, 34 Connolly-Begg (lectura complementaria)
- Slides

QUIZ UI EN INTU



DATA WAREHOUSE



DATA WAREHOUSE/ING

A **Data Warehouse** is a centralized repository of integrated, structured, and historical data that supports business intelligence. It is designed to facilitate efficient data retrieval and analysis for decision-making purposes.

Data Warehousing is the process to operational extract, cleanse, transform, control, and load processes that maintain the data in a data warehouse.

DATA WAREHOUSING [VIDEO]

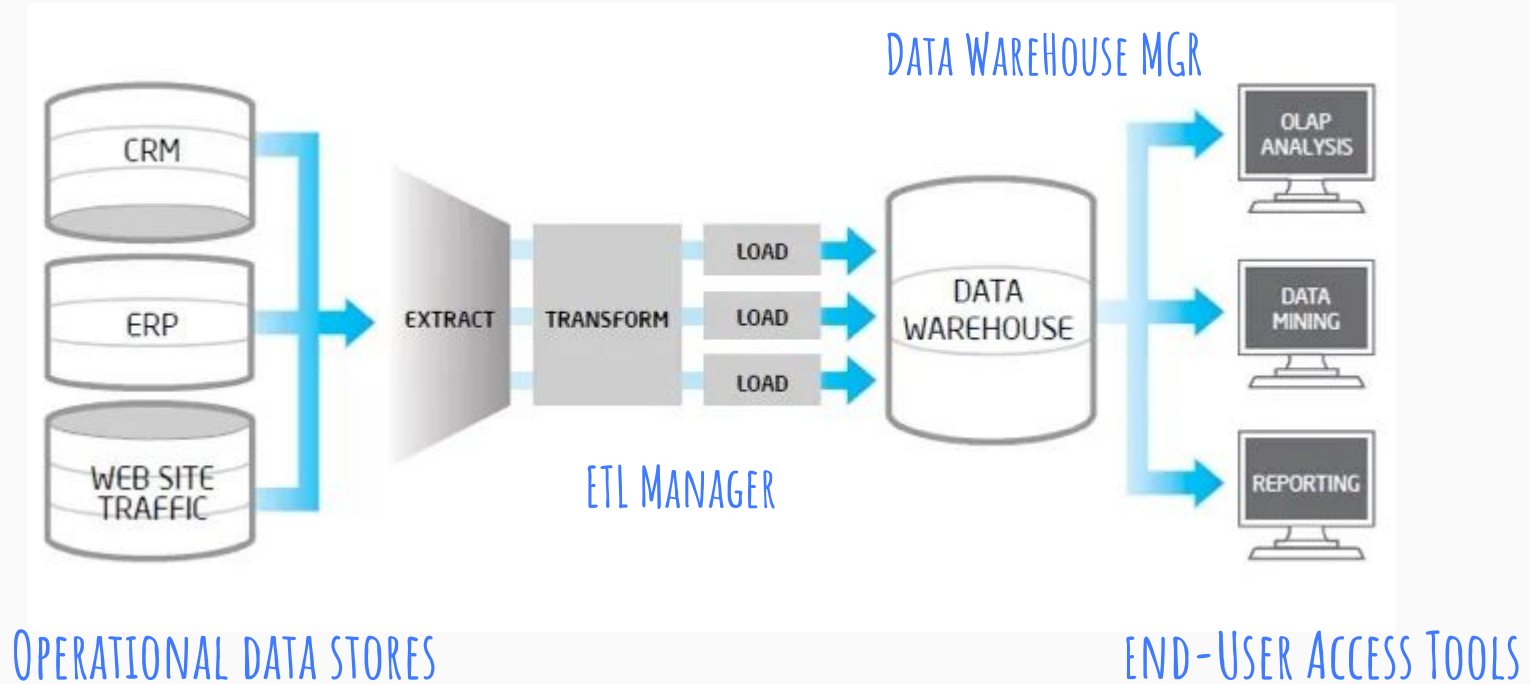


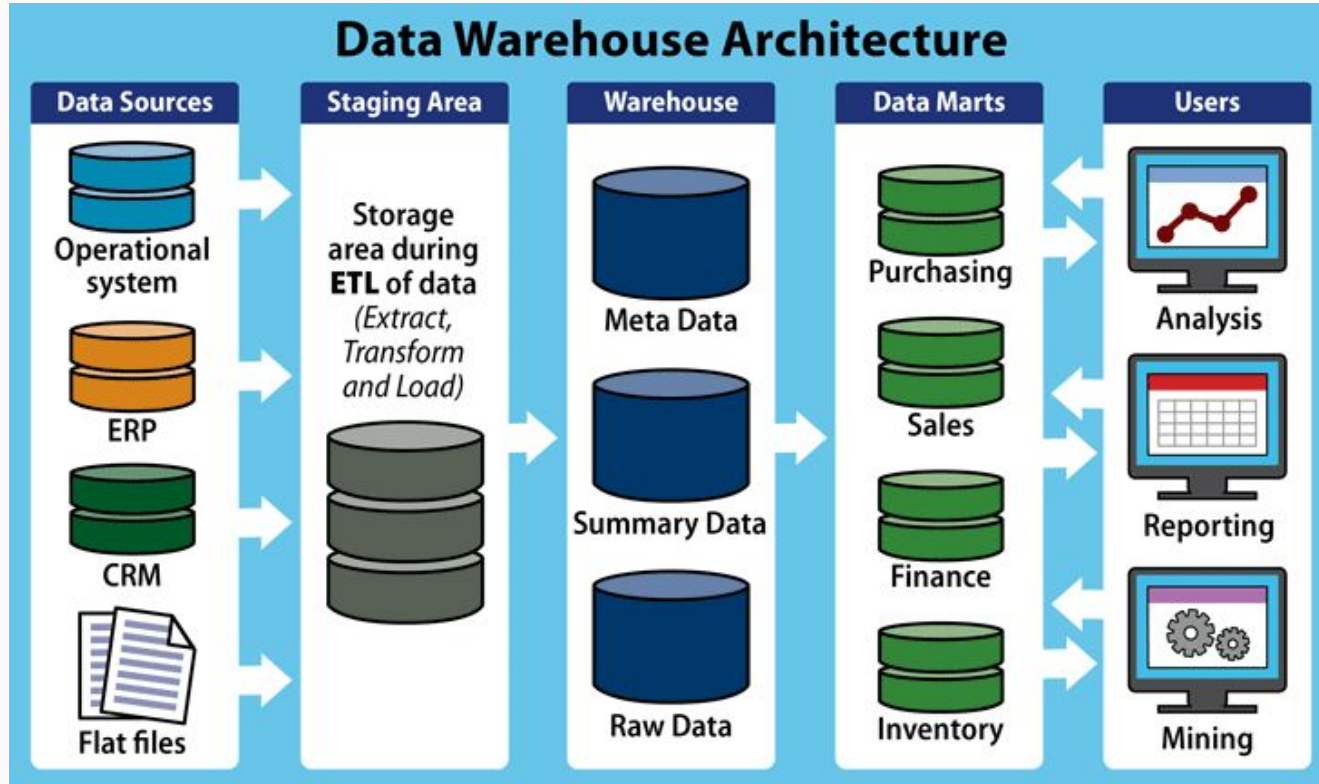
Image taken from [\[here\]](#)

DATA IN A DATA WAREHOUSE

Data in data warehouses is:

- Subject-oriented
- Integrated
- Time-variant
- Non-volatile

DATA WAREHOUSE ARCHITECTURE



DATA WAREHOUSE SCHEMA

A schema refers to the logical structure or blueprint that defines how data is organized and stored within a data warehouse.

The scheme determines the relationships between tables, the attributes of each table, and the rules for data integrity and consistency. There are two commonly used types of schemas in data warehousing: **star** schema, **snowflake** schema

Both types of schemas are composed by **Fact** tables and **Dimensions**

FACTS VS DIMENSIONS

Fact Table

- Central table in a data warehouse schema.
- Contains quantitative measures or metrics.
- Represents the primary focus of analysis.
- Associated with dimension tables through foreign keys.
- Contains aggregated values for efficient querying.
- Granularity determines the level of detail stored.
- Measures include sales revenue, quantity sold, etc.

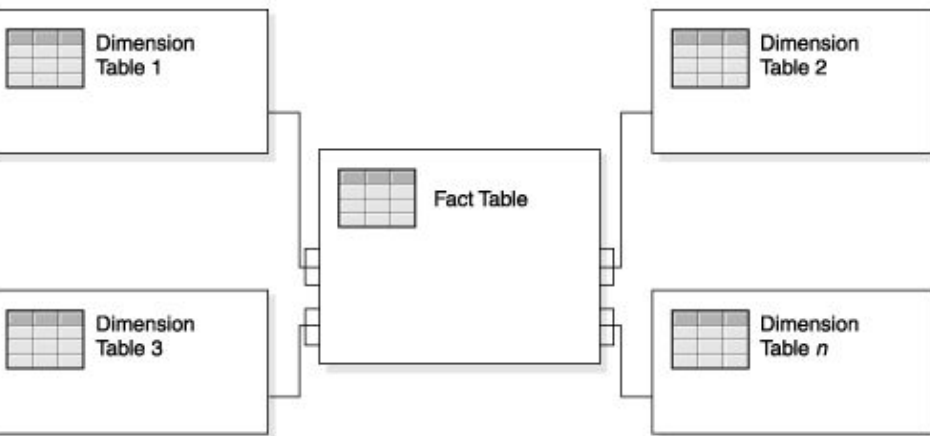
Dimensions:

- Supporting table in a data warehouse schema.
- Provides descriptive attributes to add context.
- Defines characteristics of the data.
- Often has hierarchies for drilling down or rolling up.
- Linked to the fact table through foreign keys.
- Examples: time, geography, product, customer.

THE RELATIONAL MODEL - CONCEPTS

- **Relational database.** A collection of *normalized* relations with different names.
- **Normalization:** Table normalization is a process in database design that aims to eliminate redundancy and improve data integrity by organizing data into separate tables.
- **Primary Key:** identifier, there are no duplicate tuples in a table.
- **Foreign Key:** It is a column or a set of columns in one table that refers to the primary key of another table. The foreign key creates a link between the two tables, enabling data integrity and enforcing referential integrity constraints.

Star Schema



Snowflake Schema

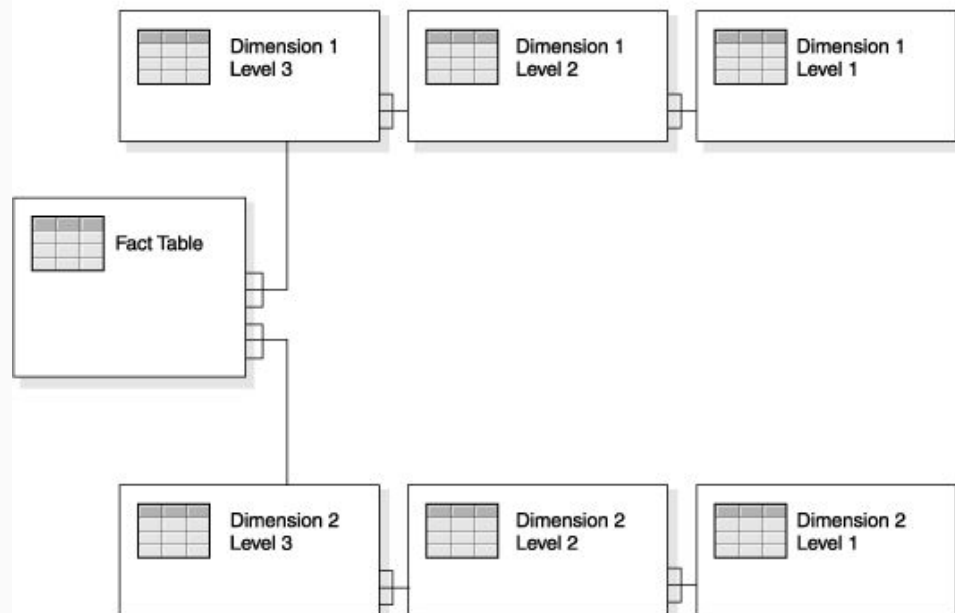


Image taken from [\[here\]](#)

MODELING A DATA WAREHOUSES

Star

Structure: data is organized around a **central fact** table, which contains the measures or key performance indicators (KPIs) of interest. The fact table is surrounded by **dimension tables**, each representing a specific attribute or perspective of the data, such as time, product, customer, or location.

Relationships: The fact table is connected to dimension tables through **foreign keys**, establishing one-to-many relationships. The dimension tables provide descriptive attributes that provide context to the measures in the fact table.

Snowflake

Structure: **extends** the star schema by further normalizing the dimension tables. This means that dimension tables are split into multiple tables, reducing redundancy by storing attribute hierarchies and relationships in separate tables.

Relationships and Normalization: The relationships between tables in the snowflake schema are more granular, with additional join paths. Dimension tables are typically normalized, resulting in a more complex structure compared to the star schema.

SCENARIO

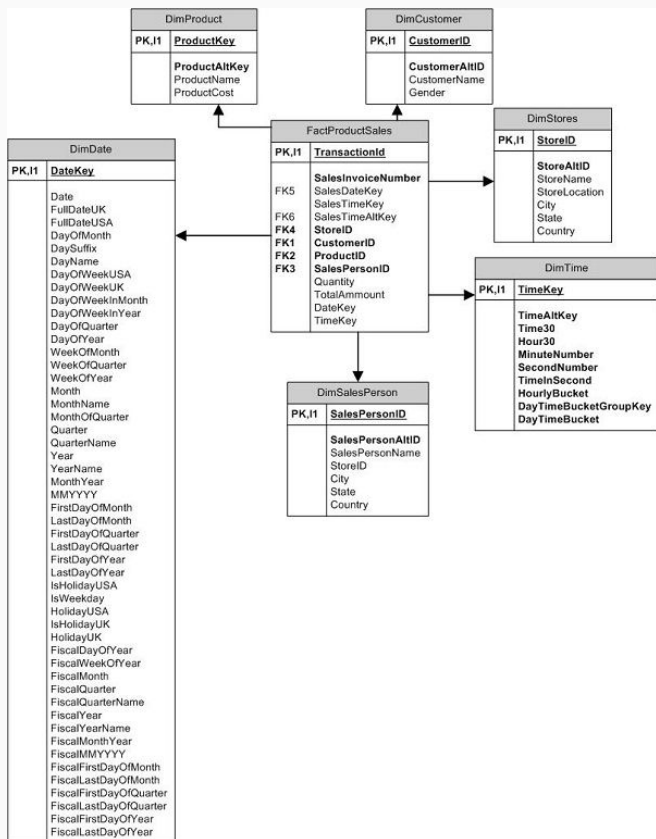
X-Mart es una compañía que tiene diferentes centros comerciales en nuestra ciudad, donde se realizan ventas diarias de varios productos. La alta dirección enfrenta un problema al tomar decisiones debido a la falta de disponibilidad de datos integrados; no pueden estudiar sus datos según sus necesidades. Por lo tanto, nos pidieron que diseñemos un sistema que pueda ayudarles rápidamente en la toma de decisiones y proporcionar Retorno de la Inversión (ROI).



SCENARIO - REQUIREMENTS

1. Necesidad de visualizar las ganancias diarias, semanales, mensuales y trimestrales de cada tienda.
2. Comparación de las ventas y ganancias en diversos períodos de tiempo.
3. Comparación de las ventas en diferentes franjas horarias del día.
4. Identificar cuál producto tiene más demanda en cada ubicación.
5. Estudiar la tendencia de las ventas según el período del día a lo largo de la semana, el mes y el año.
6. Identificar en qué día de la semana las ventas son más altas.
7. Obtener las ventas y ganancias de cada domingo de este mes.
8. Analizar la tendencia de las ventas en días laborables y fines de semana.
9. Comparar las ventas semanales, mensuales y anuales para conocer el crecimiento y los KPI (Indicadores Clave de Desempeño).

DATA WAREHOUSE SCHEMA



EJERCICIO

Para el siguiente ejercicio trabajaremos con una DW de ejemplo que nos permite aterrizar los conceptos de la clase.

Herramientas

- Oracle Live [[link](#)]
- Script e inserts [en la carpeta code]

ELEMENTOS

Dimensiones (Dimension) tablas: Product, Customer, Store, Date, Time, SalesPerson

Medidas (Measure) son columnas con datos cuantificables , que pueden ser agregados, hacen parte de la tabla de hechos. Las medidas en este ejemplo son: Actual Cost, Total Sales, Quantity, Total de filas de la tabla de hechos

Tabla de Hechos (Fact Table)

→ Llaves foráneas: Sales Date key, Sales Time key, Invoice Number, Sales Person ID, Store ID, Customer ID

EXPLORACIÓN - PARTE I

1. Inicie sesión en Oracle live para trabajar con el código del caso de estudio.
2. Cargue los scripts que se encuentran en la carpeta Code (archivos dw.sql e inserts.sql)
3. Haga un inventario de los objetos en la base de datos
 - a. ¿Están todas las dimensiones?
 - b. ¿Están todas las medidas?
 - c. ¿Está completa la tabla de hechos?
4. ¿Se puede responder a las preguntas con los datos actuales?

COMPLETANDO - PARTE 2

1. Cargue la tabla con las fechas (archivo DATE_DIM)
2. Explore el contenido de la nueva tabla (usando la sentencia select)
3. Completando la tabla de hechos

La tabla de hechos almacena todas las entradas transaccionales de ventas del día anterior con las llaves foráneas apropiadas que hacen referencia a las llaves primarias de las dimensiones.

Por ejemplo:

El cliente Henry Ford ha comprado 2 artículos (1 kg de aceite de girasol y 2 jabones Nirma) en una sola factura el 1 de enero de 2013 en D-mart en Sivranjani y el vendedor fue Jacob. El tiempo de facturación registrado es a las 13:00.

Antes de llenar la tabla de hechos, debe identificar y buscar los valores de la columna de clave primaria en las dimensiones según el ejemplo dado y completar las columnas de clave externa de la tabla de hechos con los valores de clave apropiados.

RESOLVIENDO LAS CONSULTAS

1. Necesidad de visualizar las ganancias diarias, semanales, mensuales y trimestrales de cada tienda.
2. Comparación de las ventas y ganancias en diversos períodos de tiempo.
3. Comparación de las ventas en diferentes franjas horarias del día.
4. Identificar cuál producto tiene más demanda en cada ubicación.
5. Estudiar la tendencia de las ventas según el período del día a lo largo de la semana, el mes y el año.
6. Identificar en qué día de la semana las ventas son más altas.
7. Obtener las ventas y ganancias de cada domingo de este mes.
8. Analizar la tendencia de las ventas en días laborables y fines de semana.
9. Comparar las ventas semanales, mensuales y anuales para conocer el crecimiento y los KPI (Indicadores Clave de Desempeño).

ONLINE ANALYTICAL PROCESSING (OLAP)



OLAP CUBE

An OLAP cube, is a data structure to facilitate fast and efficient multidimensional analysis of data. It provides a multidimensional view of data that allows users to analyze and explore data from different dimensions and levels of granularity.

OLAP cubes organize data based on **Dimensions and Measures**,

Dimensions represent the different attributes or characteristics of the data, e.g., time, geography, product, or customer.

Measures represent the metrics that are being analyzed. Measures are associated with the intersection points of dimensions in the cube.

MULTIDIMENSIONAL DATA, EXAMPLE

Go to [Miro board]

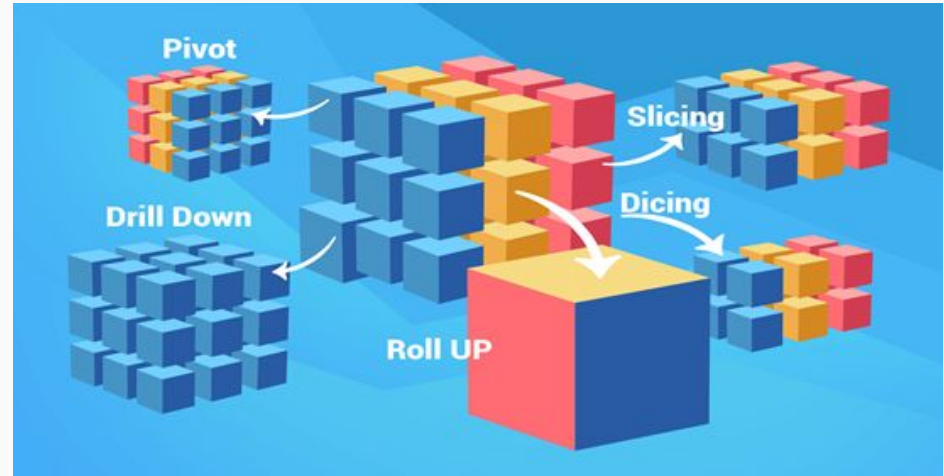
MULTIDIMENSIONAL OPERATIONS, EXAMPLE

Roll-up. Performs aggregations by moving up the dimensional hierarchy

Drill-down. Reverse to roll-up, reveals the detailed data confirming the aggregated data.

Slice and dice. Refers to the ability to look at data from different viewpoints, performs a selection on one dimension on the data

Pivot. Rotates the data to provide an alternative view of the same data.



CASE STUDY - DREAMHOUSE

DreamHouse= {Branch, Staff, PropertyForRent, Client, PrivateOwner, Viewing}

Branch (branchNo, street, city, postcode)

Staff (staffNo, fName, lName, position, sex, DOB, salary,
branchNo)

PropertyForRent (propertyNo, street, city, postcode, type, rooms, rent,
ownerNo, staffNo, branchNo)

Client (clientNo, fName, lName, telNo, prefType, maxRent, eMail)

PrivateOwner (ownerNo, fName, lName, address, telNo, eMail, password)

Viewing (clientNo, propertyNo, viewDate, comment)

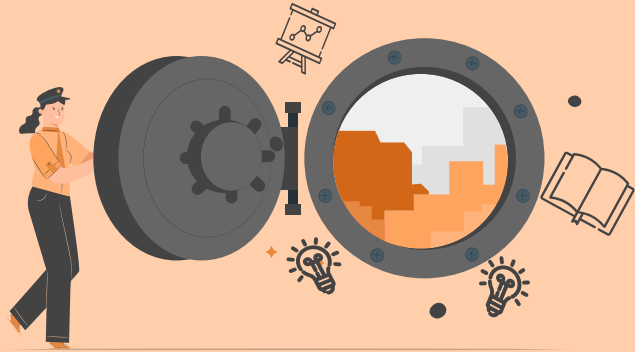
Registration (clientNo, branchNo, staffNo, dateJoined)

Link to the spreadsheet of the example [[open](#)]

DATA WAREHOUSING QUERIES

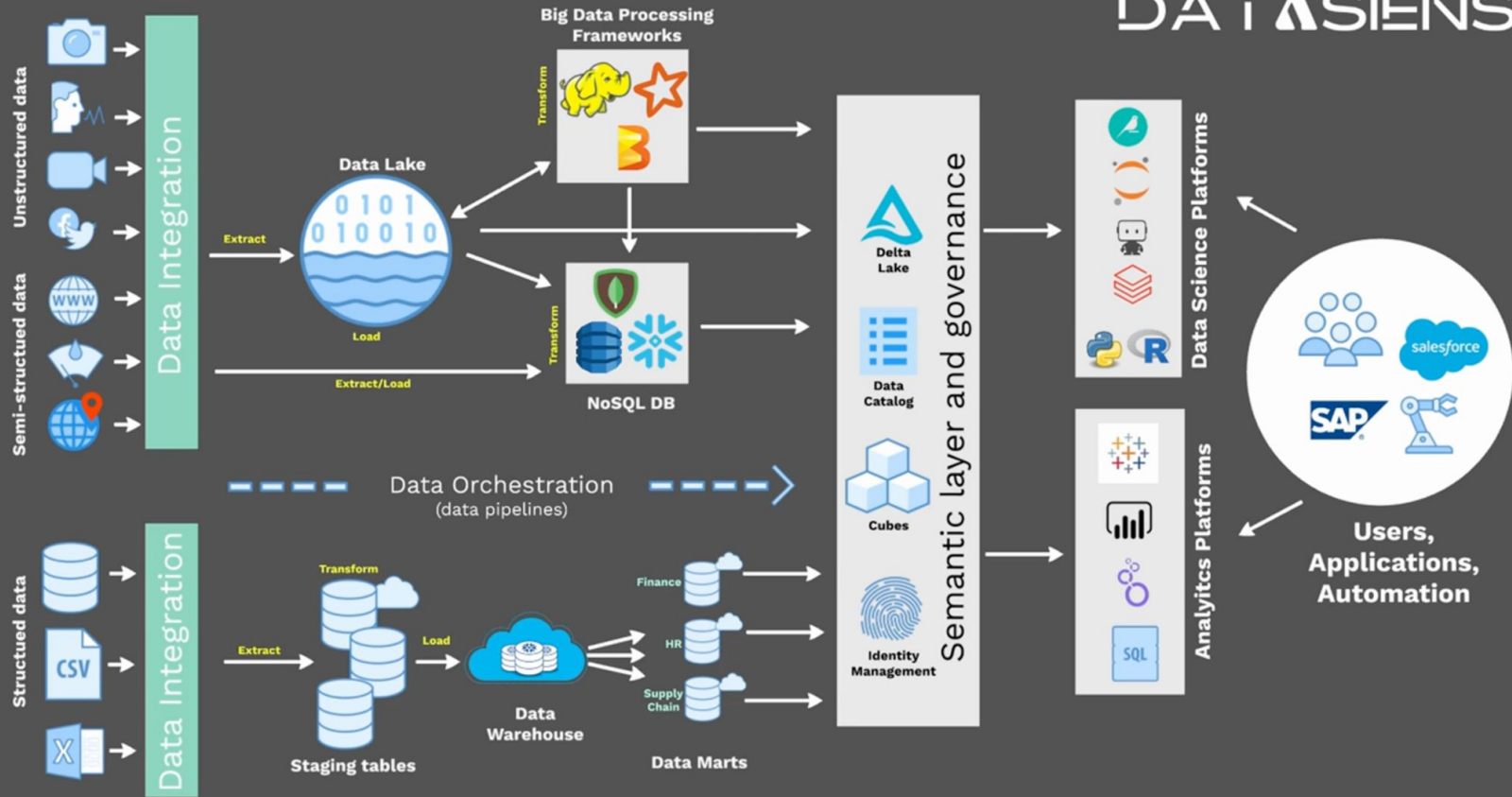
DWH systems can answer ad-hoc queries and store historical data, which is necessary for analyzing trends.

- What is the total revenue for Scotland in the third quarter of 2008?
- What are the three most popular areas in each city for the renting of property in 2008 and how do these results compare with the results for the previous two years?
- What is the relationship between the total annual revenue generated by each branch office and the total number of sales staff assigned to each branch office?



CIERRE DE LA CLASE





LA PRÓXIMA SEMANA

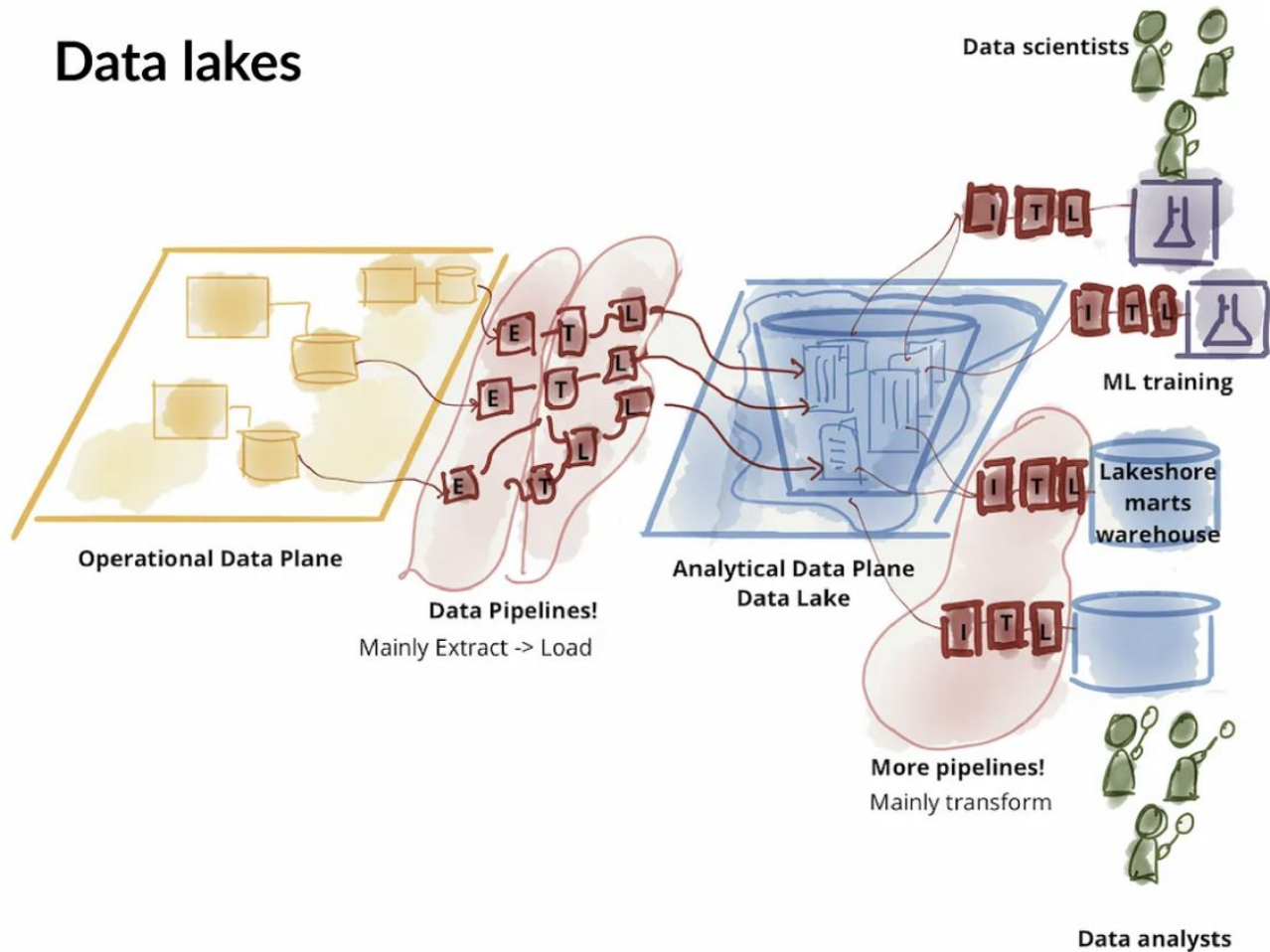
Fecha: Martes, 7 de mayo de 2024

Tema: Tipos de repositorios y flujos de ingesta de datos. (Datalake, práctica)

Asignaciones:

- Terminar el ejercicio del DW
- Lectura y videos sobre Data bases vs Data Warehouses vs Data Mart

Data lakes



THANKS

apvillota@icesi.edu.co

mmrojas@icesi.edu.co

CREDITS: This presentation template was created by **Slidesgo**, including icons by **Flaticon**, and infographics & images by **Freepik** and illustrations by **Storyset**

Does anyone have any questions?



CATEGORIES OF OLAP SERVERS










OLAP Servers are categorized regarding the architecture used to store and process multidimensional data.



- Multidimensional OLAP
- Relational OLAP
- Hybrid OLAP
- Desktop OLAP

CATEGORIES OF OLAP SERVERS

OLAP Servers are categorized regarding the architecture used to store and process multidimensional data.

- **Multidimensional OLAP** works directly with a multidimensional OLAP cube
- **Relational OLAP** is multidimensional data analysis that operates directly on data on relational tables, without first reorganizing the data into a cube.
- **Hybrid OLAP** attempts to create the optimal division of labor between relational and multidimensional databases within a single OLAP architecture.

	MOLAP	HOLAP	ROLAP
Cube Structure			
Preprocessed Aggregates			
Detail-Level Values			

 Multidimensional Storage  Relational Storage