DATA WAREHOUSE FUNDAMENTALS

What’s a data warehouse:

As the name indicates, it’s a warehouse but filled with data. Data warehouse isn’t the same as database, but it’s often built on top of relational databases: such as, MS SQL Server, or Oracle. Data comes from different sources in a data warehouse.

Data Sources 🡺 ETL 🡺 Data Warehouse (Staging Layer 🡺 User Access Layer)🡺 ETL 🡺 Data Marts

Cloud-based data warehouse examples:

1. Amazon Redshift
2. Google BigQuery
3. Snowflake
4. Microsoft Azure Synapse

Tools to implement ETL processes:

1. Apache Nifi
2. AWS Glue
3. Matillion
4. Databricks

DATA MARTS BY CAMPUSES

Campus 1 Data Mart

Data Warehouse

Campus 1 🡺 Faculty Management

Student Management

Facilities Management

Campus 2 🡺 Faculty Management

Campus 2 Data Mart

Student Management

Facilities Management 🡺 🡺 🡺

Campus 3 🡺 Faculty Management

Campus 3 Data Mart

Student Management

Facilities Management

Campus 4 🡺 Faculty Management

Campus 4 Data Mart

Student Management

Facilities Management

**ETL:**

Is a process of Extracting, Transforming and Loading data from various sources into a data warehouse. ETL jobs are typically run on a scheduled basis. In some cases, particularly with cloud data warehouses, the order might be changed to ELT (Extract, Load and Transform), where data is loaded before transformation.

ELT:

Process where Loading comes before transformation. ELT “blast” data into big environment. Raw form in Hadoop HDFS, AWS S3, etc.

Two Types of ELT 🡺 Initial and Incremental ETL:

Initial ETL:

Normally one time only, right before the data warehouse goes live and redo if data warehouse blows up.

Incremental ETL:

Incrementally refreshes the data warehouse. Brings in new data, modifies data.

Four major incremental ETL patterns:

1. Append
2. In-place update
3. Complete replacement
4. Rolling append