DATA WAREHOUSE

Introduction:

A data warehouse centralizes and consolidates large amounts of data from multiple sources. Its analytical capabilities allow organizations to derive valuable business insights from their data to improve decision-making

Data warehouse:

Data Warehouses are large, centralized repositories of data that help organizations store, manage, and analyze large amounts of data.

Schema:

A schema is a collection of database objects, including tables, views, indexes, and synonyms. There is a variety of ways of arranging schema objects in the schema models designed for data warehousing. One data warehouse schema model is a star schema.

Major types:

Following are the three major types of schemas:

- Schema.
- Snowflake Schema.
- Galaxy Schema

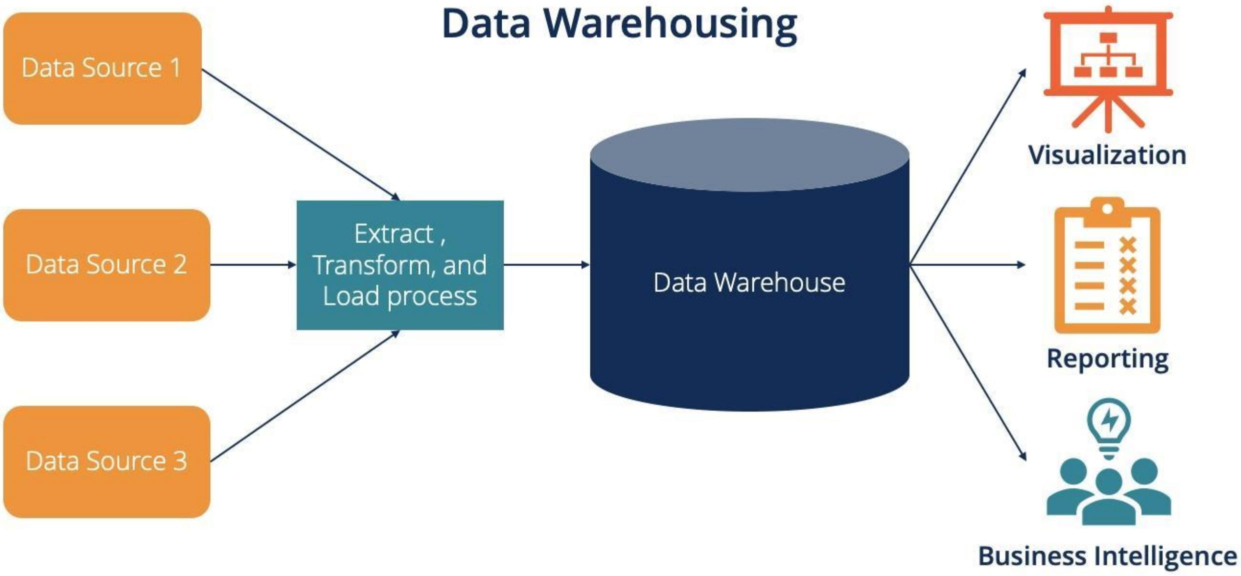
*The star schema is suitable for data warehouses that have stable and well-defined dimensions and facts, and that require high query efficiency and scalability. The main advantage of the snowflake schema is its flexibility and normalization.

Structure of Data Warehouse:

A typical data warehouse has four main components: a central database, ETL (extract, transform, load) tools, metadata, and access tools. All of these components are engineered for speed so that you can get results quickly and analyze data on the fly. Diagram showing the components of a data warehouse.

Data source:

A data source is a place where information is obtained. The source can be a database, a flat file, an XML file, or any other format that a system can read. The input is recorded as a collection of records that contain information used in the business process. That information can include customer details, accounting figures, sales, logistics, and more.



Why they are important?

Knowledge can help businesses respond to changing market conditions, deal with logistics challenges or identify new ways to improve the customer experience. These details can provide you with a unique perspective of your business operations

Scope of data warehouse:

The Data Warehouse (DW) or the Enterprise Data Warehouse (EDW) is the essential component for Business Intelligence (BI) systems, in which the process of assembling, administering, and manipulating the data from multiple varieties of data sources is performed in order to turn up with the significant business decision

Future scope:

Data warehouses are at an exciting point of expansion and evolution. With the global data warehousing market size estimated to grow at a CAGR of 10% until 2028, you'll see a greater reliance on them and the tools that make them easier to use than ever

Types of data warehouse:

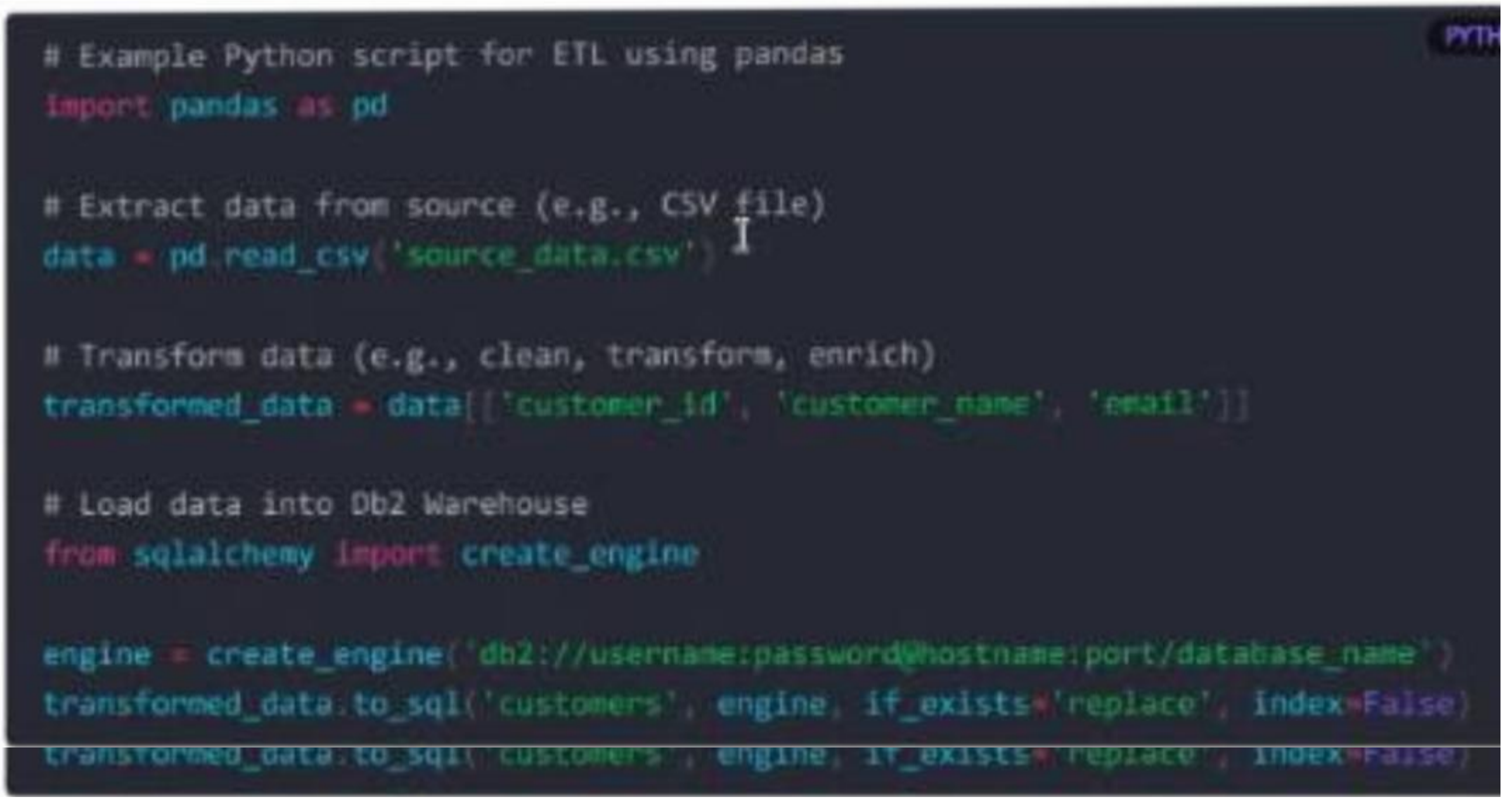
Once in the data warehouse, the data is ingested, transformed, processed, and made accessible for use in decision-making. The three main types of data warehouses are enterprise data warehouse (EDW), operational data store (ODS), and data mart.

Characteristics of data warehouse:

The four characteristics of a data warehouse, also called features of a data warehouse are: subject-oriented, time-variant, integrated, and non-volatile**.**

CSV file:

The Full Form Of CSV is Comma Separated Value. CSV (Comma Separated Values) stores tabular data. Each line of files is a record and CSV uses a comma to separate the values. CSV stores tabular data in such a way where each line



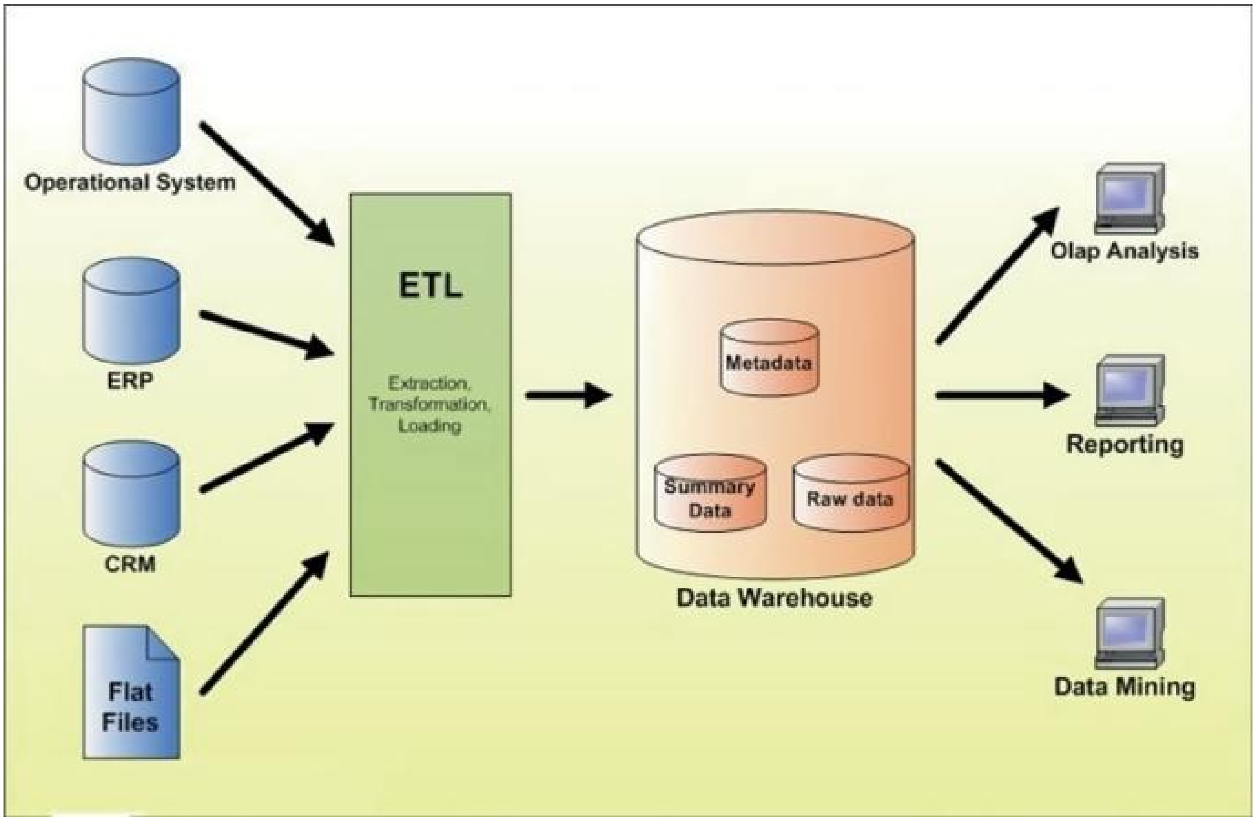
```
# Example Python script for ETL using pandas
import pandas as pd

# Extract data from source (e.g., CSV file)
data = pd.read_csv('source_data.csv')

# Transform data (e.g., clean, transform, enrich)
transformed_data = data[['customer_id', 'customer_name', 'email']]

# Load data into Db2 Warehouse
from sqlalchemy import create_engine

engine = create_engine('db2://username:password@hostname:port/database_name')
transformed_data.to_sql('customers', engine, if_exists='replace', index=False)
```

transformed_data.to_sql('customers', engine, if_exists='replace', index=False)

Architecture of data warehouse:

Data warehouse architecture is an intentional design of data services and subsystems that consolidates disparate data sources into a single repository for business intelligence (BI), AI/ML, and analysis

There are three main data warehouse architecture types: single-tier, two-tier and three-tier data warehouses. Every data warehouse has the same vital components within its architecture, namely: ETL tools, databases, metadata, bus & data marts and access tools



Design strategy:

Modeling

Visualize the relationships between data.

Set standardized naming conventions.

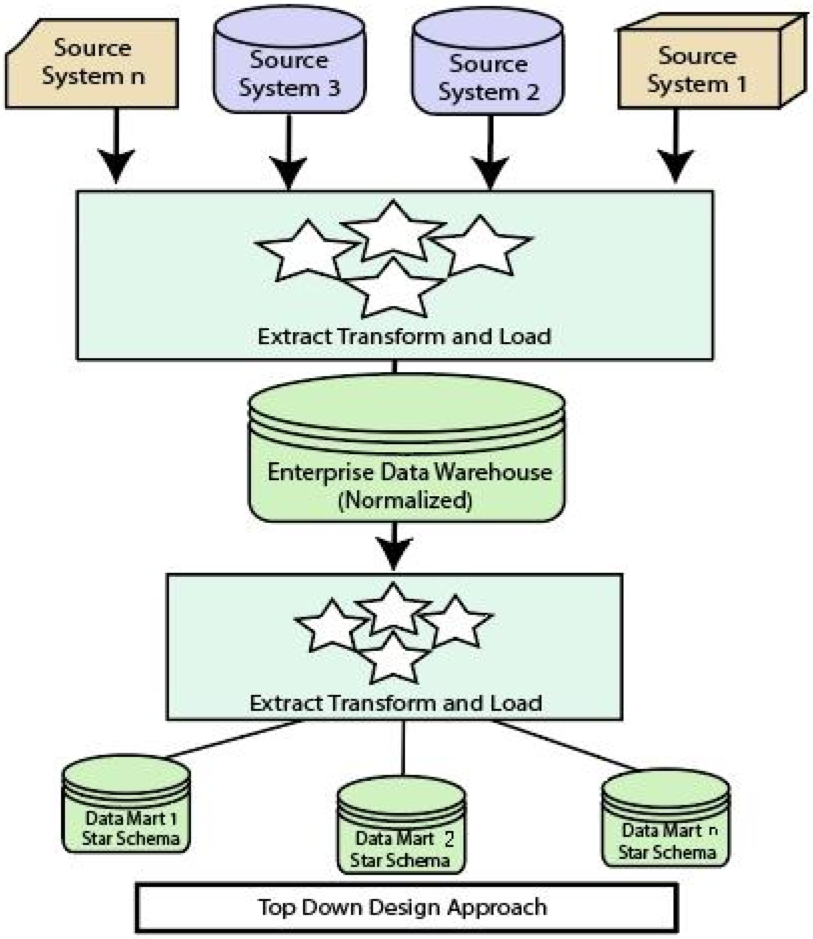
Create relationships between data sets.

Establish compliance and security processes.

Align your processes with your overarching IT goals.

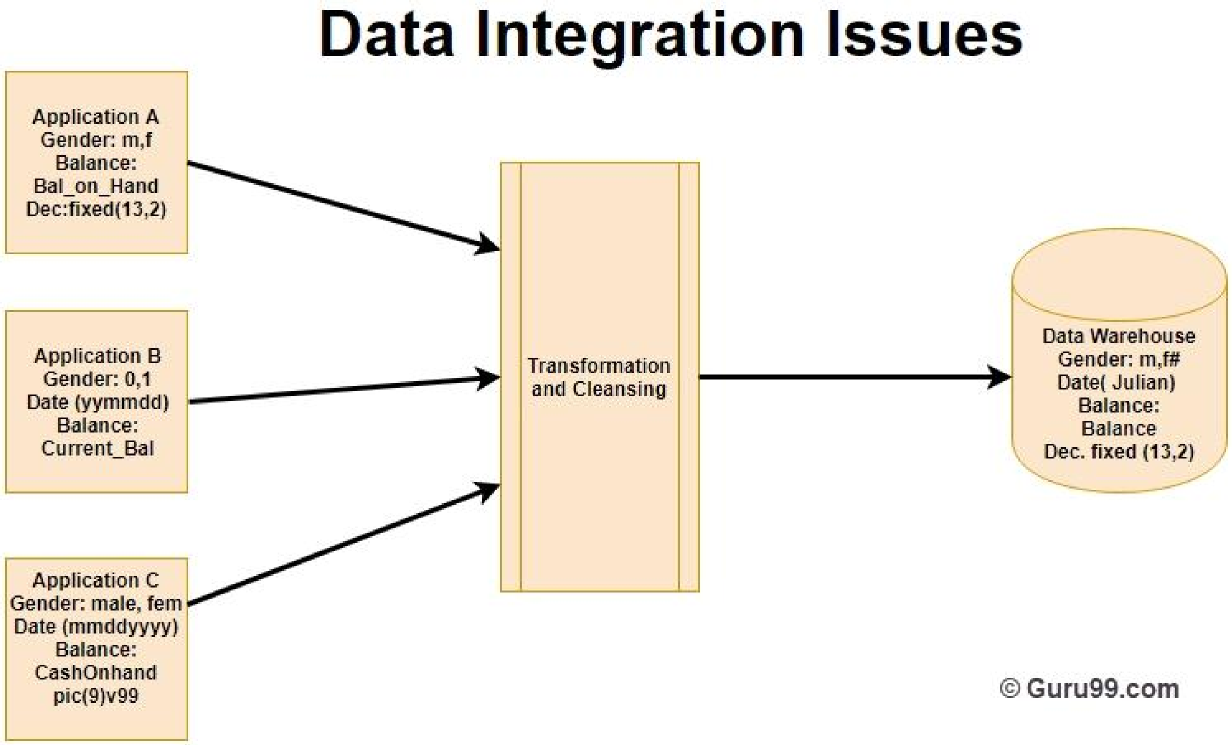
There are 2 approaches for constructing data-warehouse: Top-down approach and Bottom-up approach are explained as below. External source is a source from where data is collected irrespective of the type of data. Data can be structured, semi structured and unstructured as well.

Designing data models for the data warehouse and data marts. Identifying data objects as entities or attributes; identifying relationships between entities. Mapping data objects into the data warehouse. Designing ETL/ELT processes for data integration and data flow control.



Top Down Design Approach

Data Integration issues:



Advantage

- Data warehouse benefits
- Provide a stable, centralized repository for large amounts of historical data. Improve business processes and decision-making with actionable insights. Increase a business's overall return on investment (ROI) Improve data quality.
- The primary purpose of a data warehouse is to enable companies to access and analyze all of their data to derive the most accurate business insights and forecasting models

Disadvantage:

Costly setup and maintenance – Building a data warehouse can be a time-consuming and expensive process. It requires specialized knowledge and resources to design and maintain, which can be a significant investment for companies.

Limited flexibility – Once a data warehouse is set up, it can be challenging to make changes to the data structure. This can make it difficult to adapt to changes in the business environment, or to incorporate new types of data.

Data silos – Because data warehouses are designed to store specific types of data, it can be challenging to integrate data from different sources. This can lead to data silos, where different teams or departments have their own sets of data that are not shared with others.

Data latency – Depending on the size and complexity of a data warehouse, it can take some time for data to be processed and analyzed. This can create delays in decision-making, which can be a disadvantage in fast-paced business environments.

Data security – Data warehouses contain large amounts of sensitive information, which can make them a target for cyberattacks or data breaches. Ensuring the security of a data warehouse requires ongoing vigilance and investment in cybersecurity measures.

That's it.

Create table:

```
CREATE TABLE MyTable (col1 int, col2 int );
```

Create a table

- For Warehouse, you can create a table as a new empty table. You can also create and populate a table with the results of a select statement. The following are the T-SQL commands for creating a table.

T-SQL Statement   Description

- CREATE TABLE   Creates an empty table by defining all the table columns and options.

CREATE TABLE AS SELECT Populates a new table with the results of a select statement. The table columns and data types are based on the select statement results. To import data, this statement can select from an external table.

This example creates a table with two columns:

- Primary key, foreign key, and unique key

For Warehouse, PRIMARY KEY and UNIQUE constraint are only supported when NONCLUSTERED and NOT ENFORCED are both used.

FOREIGN KEY is only supported when NOT ENFORCED is used.

For syntax, check ALTER TABLE.

For more information, see Primary keys, foreign keys, and unique keys in Warehouse in Microsoft Fabric.

Conclusion:

A Data Warehouse is a collection of software tools that facilitates analysis of a large set of business data used to help an organization make decisions

Data summaries usually present the dataset's average (mean, median, and/or mode); standard deviation from mean or interquartile range; how the data is distributed across the range of data (for example is it skewed to one side of the range); and statistical dependence (if more than one variable was captured in the …