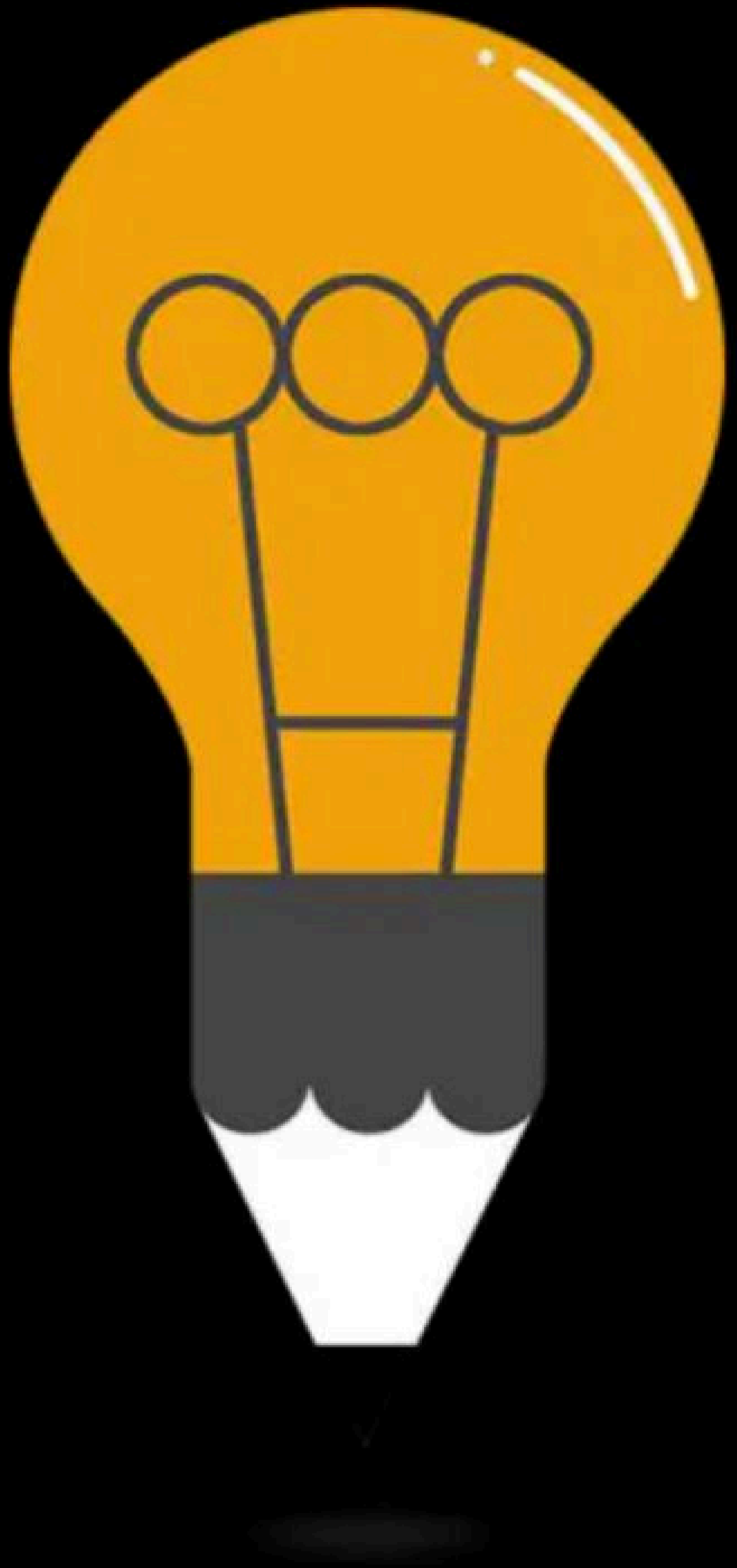
Data Warehousing Basics - Part I

Complete Course on Data Warehousing

Vishvadeep Gothi • Lesson 1 • Nov 17, 2023



Datawarehousing
Introduction

By: Vishvadeep Gothi

Vishvadeep Gothi

- **GATE Ranks:**
  - 682 (2009) – 3rd year
  - 19 (2010) – 4th year
  - 119, 440 etc.
- **Education:**
  - ME from IISc Bangalore
  - Mtech from BITS-pilani in Data Science
- **Work:**
  - 16+ Year Teaching Experience
  - 13+ in GATE/IES (GateForum, Gate Academy, ACE)
  - Worked in Cisco, Audience Communication

- **Professions:**
  - Freelance S/W developer
  - Educator
  - CrossFit Trainer

Course Structure

Topics
Basics
Data Transformation
Datawarehouse Schema
Data Discretization
Data Sampling
Data Compression
Datawarehouse Measures

Data

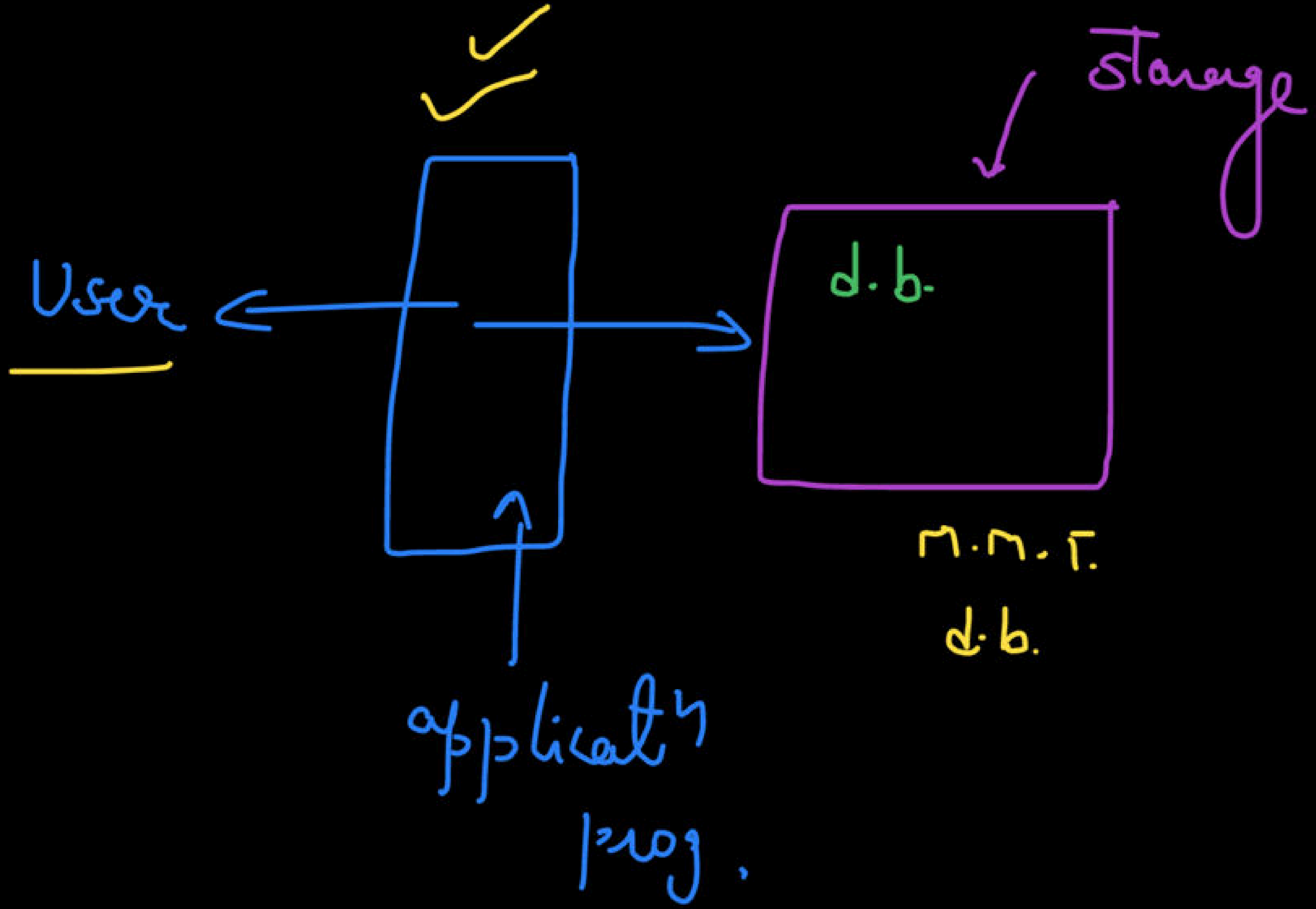
collect^{n} of raw facts

Database

The collection of data, usually referred to as the database, contains information relevant to an enterprise.

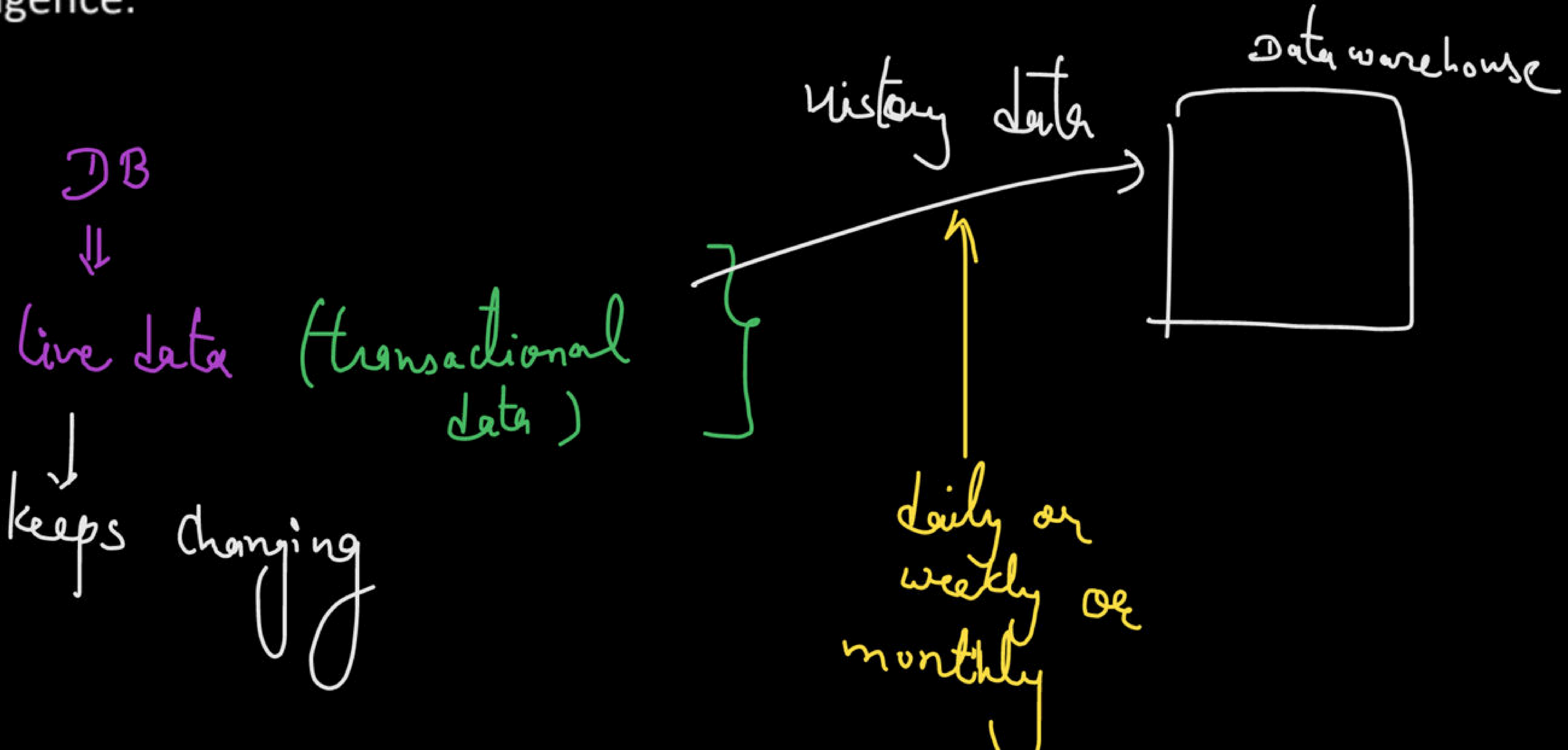
DBMS

A database-management system (DBMS) is a collection of interrelated data and a set of programs to access those data.



Data Warehousing

In computing, a data warehouse, also known as an enterprise data warehouse, is a system used for reporting and data analysis and is considered a core component of business intelligence.

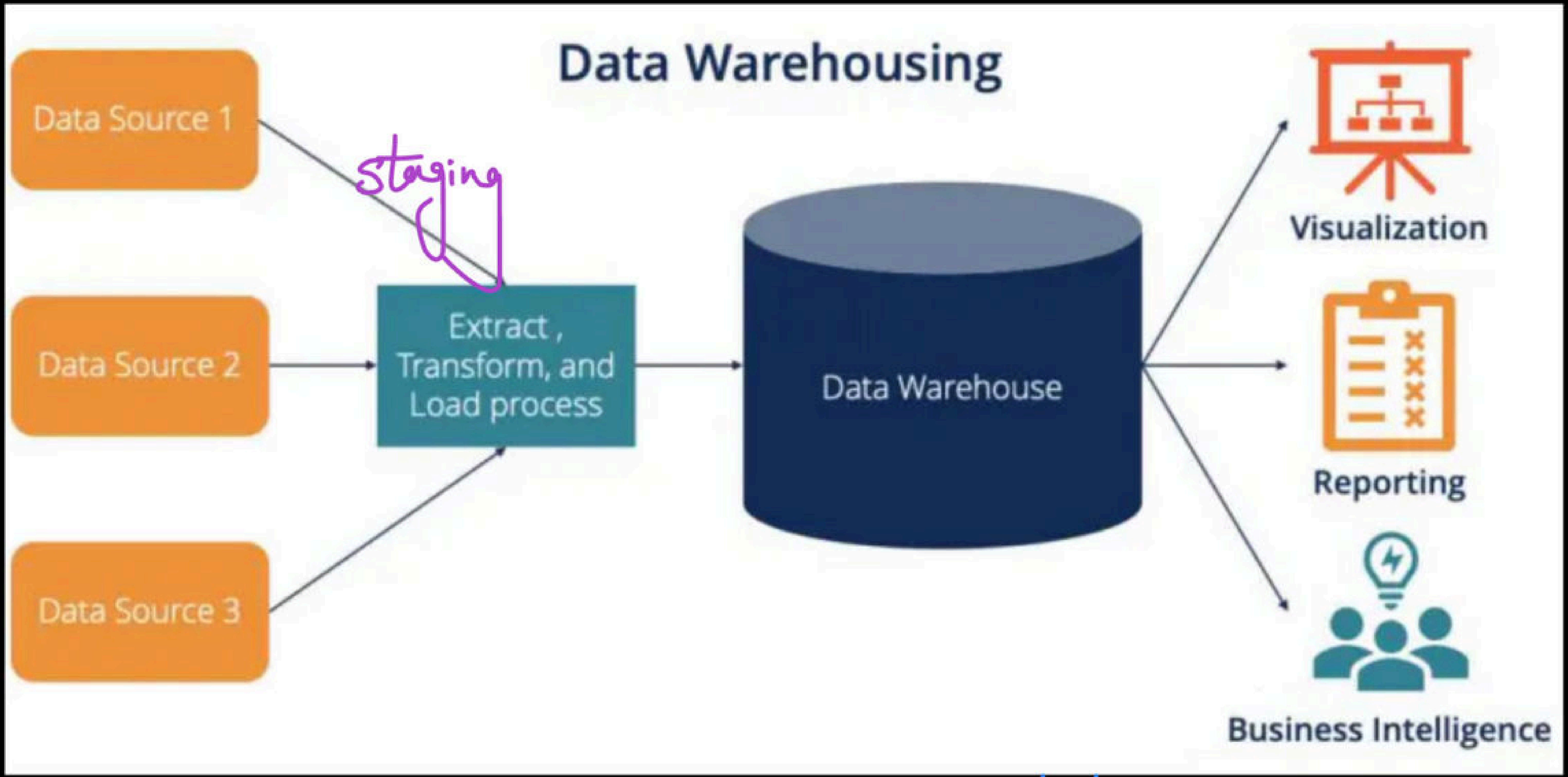


Data Warehousing

In computing, a data warehouse, also known as an enterprise data warehouse, is a system used for reporting and data analysis and is considered a core component of business intelligence.

Data warehouses are central repositories of integrated data from one or more ~~disparate~~ different sources. They store current and historical data in one single place

Data Warehousing



data mining

Steps in Data Warehousing

1. Extraction of data
2. Cleaning of data
3. Conversion of data
4. Storing in a warehouse

ETL (Extract, Transform, Load)

Three-phase process where data is extracted, transformed and loaded into an output data container.

ETL (Extract, Transform, Load)

Three-phase process where data is extracted, transformed and loaded into an output data container.

Done by application software but can be done manually also

✓ETL Tools

- Integrate.io
- IBM DataStage
- Oracle Data Integrator
- Fivetran
- Coupler.io
- SAS Data Management
- Talend Open Studio
- Whatagraph
- Pentaho Data Integration
- Singer
- Hadoop
- Dataddo
- AWS Glue
- Azure Data Factory
- Google Cloud Dataflow
- Stitch
- Informatica PowerCenter
- Skyvia
- Portable

ETL (Extract, Transform, Load)

Extract:

same type diff. type

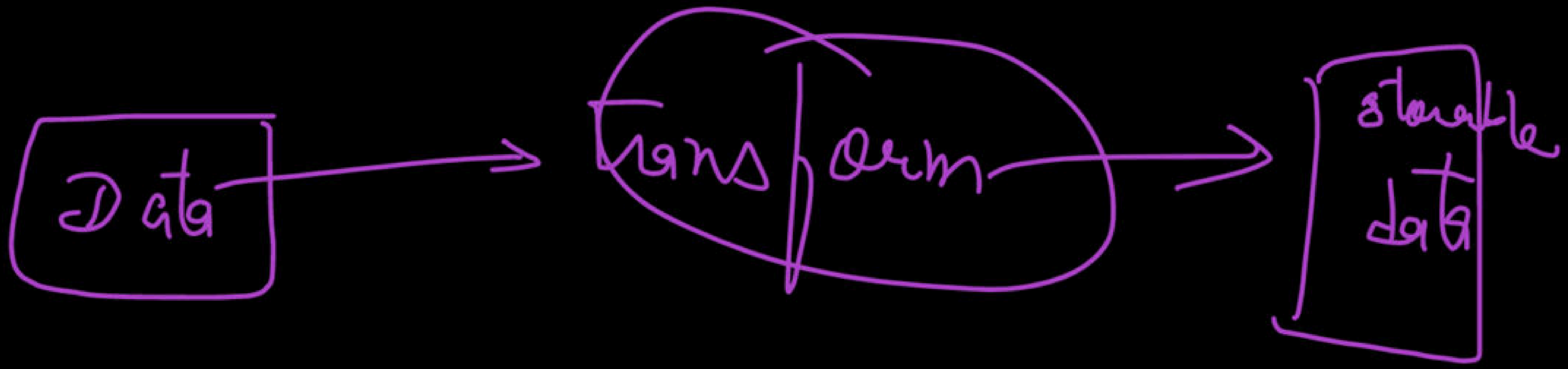
involves extracting data from homogeneous or heterogeneous sources

An intrinsic part of the extraction involves data validation

ETL (Extract, Transform, Load)

Transform:

A series of rules or functions are applied to the extracted data in order to prepare it for loading



ETL (Extract, Transform, Load)

Transform:

A series of rules or functions are applied to the extracted data in order to prepare it for loading

Cleansing: aims to pass only "proper" data to the target

Transform

- Selecting only certain columns to load
- Translating coded values (e.g., if the source system codes male as "1" and female as "2", but the warehouse codes male as "M" and female as "F")
- Encoding free-form values (e.g., mapping "Male" to "M")
- Deriving a new calculated value (e.g., sale_amount = qty * unit_price)
- Sorting or ordering the data based on a list of columns to improve search performance
- Joining data from multiple sources (e.g., lookup, merge) and deduplicating the data

Transform

- Aggregating (for example, summarizing multiple rows of data — total sales for each store, and for each region, etc.)
- Generating surrogate-key values
- Transposing or pivoting (turning multiple columns into multiple rows or vice versa)
- Splitting a column into multiple columns (e.g., converting a comma-separated list, specified as a string in one column, into individual values in different columns)
- Applying any form of data validation; failed validation may result in a full rejection of the data, partial rejection, or no rejection at all.

ETL (Extract, Transform, Load)

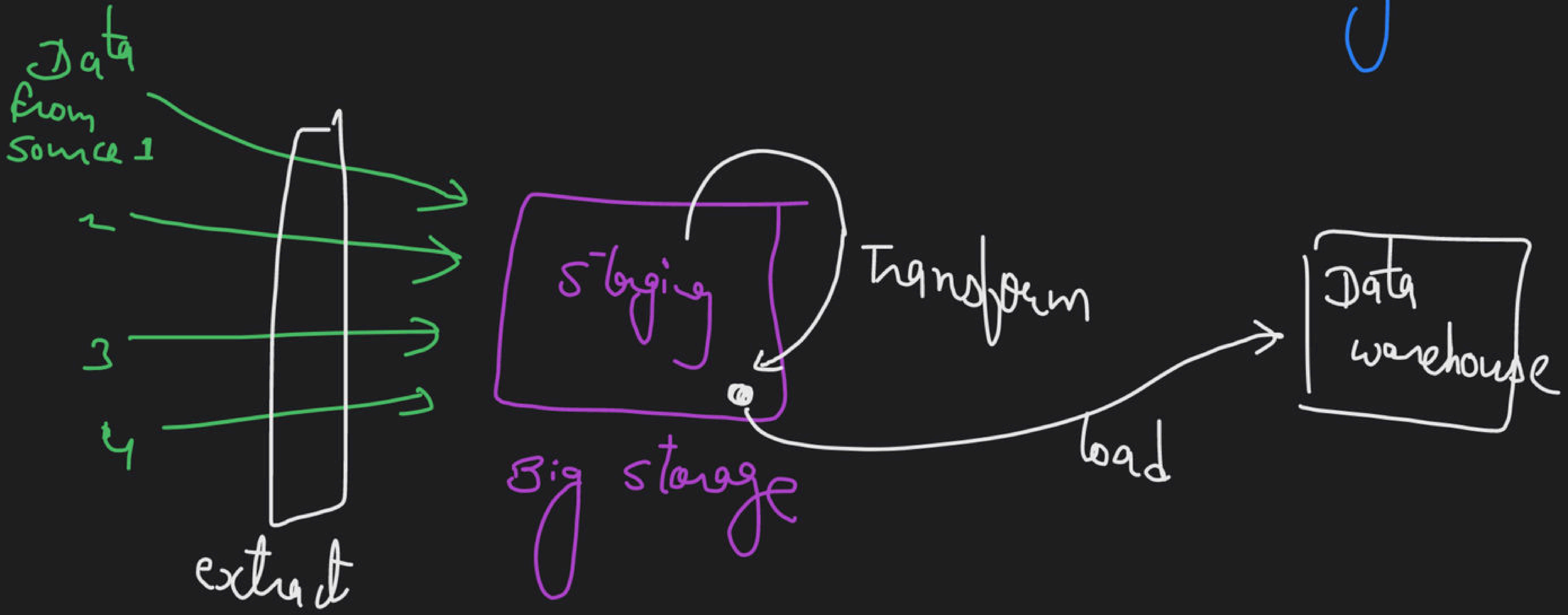
Load:

This phase loads the data into the end target

Data Warehousing

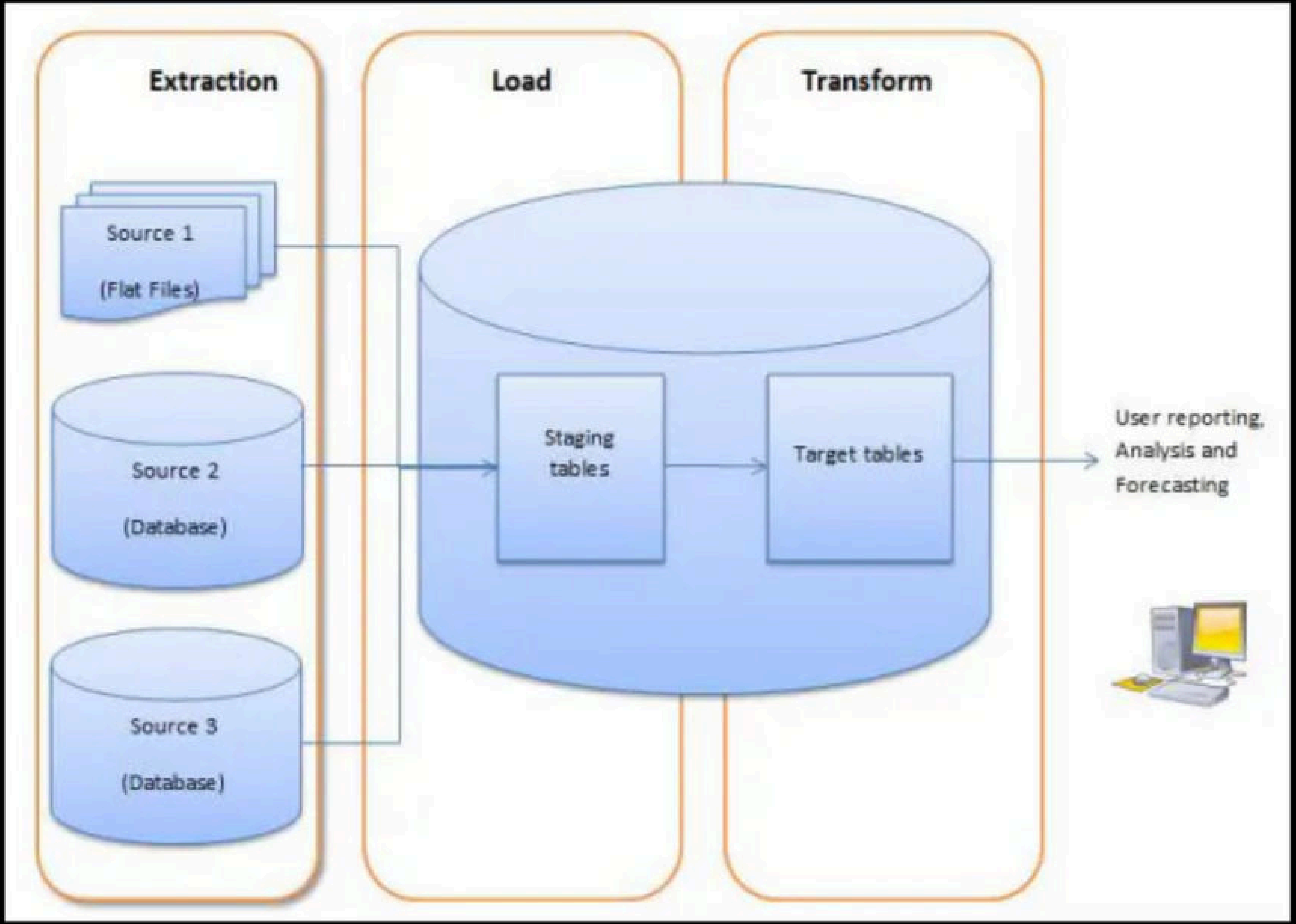
- **ETL Based:** Uses Staging
- **ELT Based:** gets rid of a separate ETL tool for data transformation. Instead, it maintains a staging area inside the data warehouse itself.

ETL Based Data warehousing



unacademy

ELT Based Data Warehousing



Data Warehousing Architecture

Happy Learning.!