ASSESSMENT-1

DATA ENGINEERING

|  |
| --- |
|  |

*Presented by:*

SreeVidya Vempalli

* **AN INTRODUCTION TO DATA :-**
  + - * It is information or particular quantity that is in the form of text, observations, images, pie charts, numbers, graphs used to analyze and make decisions.
      * In simple words, data is said to be information.
* ***Example:*** Data can be anything like ages, population, distances, temperatures, prices, names, favorite sports etc.
* We can also define data in 3 types:

1. Raw data
2. Processed data
3. Cooked data.

* ***Example:*** Rice is the perfect example for it.

1. Raw data means it is about to process, which means no format is applied on it, i.e., so many other un wanted sources are mixed with it

* To remove them, we will move to processing of data.

1. Processed data means during this process unwanted things are removed from this
2. Cooked data is defined as summarized data

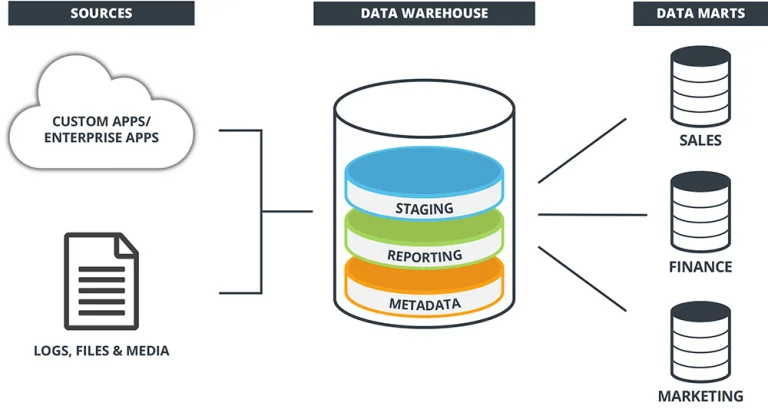
Which means ready to use for multiple purposes.

* This replicates that cooked data is more important for us but that doesn’t mean the other two can be ignored.
* We can say that the two are needed but cooked data is important.
* **INTRODUCTION TO DATA ENGINEERING:-**
* To provide solutions data engineering comes into picture.
* Data engineering comes under useful and multiple purposes like we can design it and practice building systems for collecting, storing and most of it analysing data on scale
* We apply this mechanism almost in every industry
* Simply, data engineering helps to make data more useful and accessible for consumers.
* A data engineer is a type of software engineer who creates big data ETL pipelines to manage the flow of data through the organisation.

***Example:*** Data stored in relational database is managed as table, like a Microsoft excel spreadsheet.

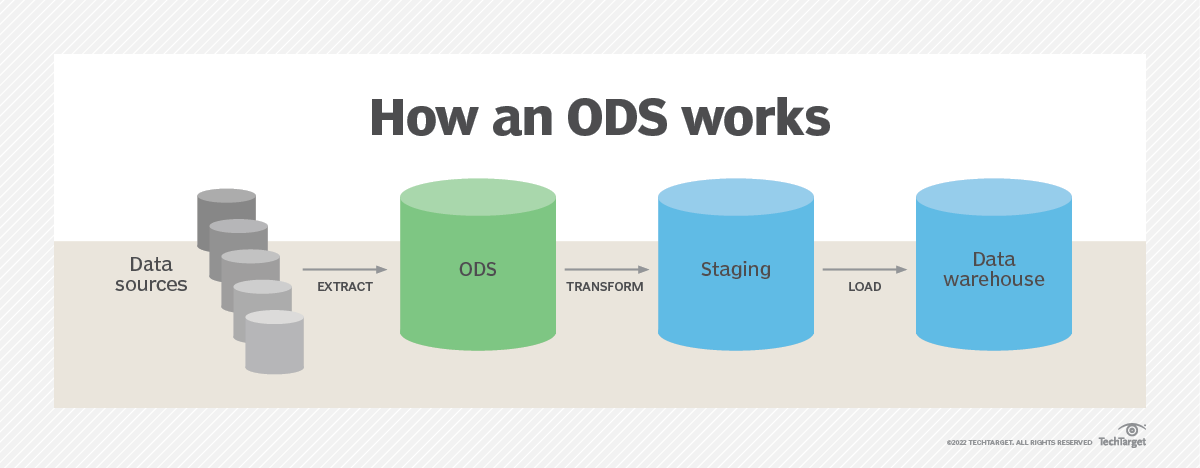
* **WARE HOUSING**: -
* Generally warehousing is a process of storing goods in a warehouse i.e., building for the purpose of distribution, sale or manufacturing
* For example : Godowns for products
* Like that here warehouse is also a type of collection of all databases that it can distribute from here itself.
* As it can hold all databases in one management system it is called data warehousing.
* **PURPOSE OF DATA WAREHOUSING** :-
* It is a type of Data management system that is to designed, to enable and support business intelligence activities, especially analytics.
* So its purpose is to provide a central repository of information that can be quickly analyzed and request to generate relevant insight.

ARCHITECTURE OF DATA WAREHOUSING



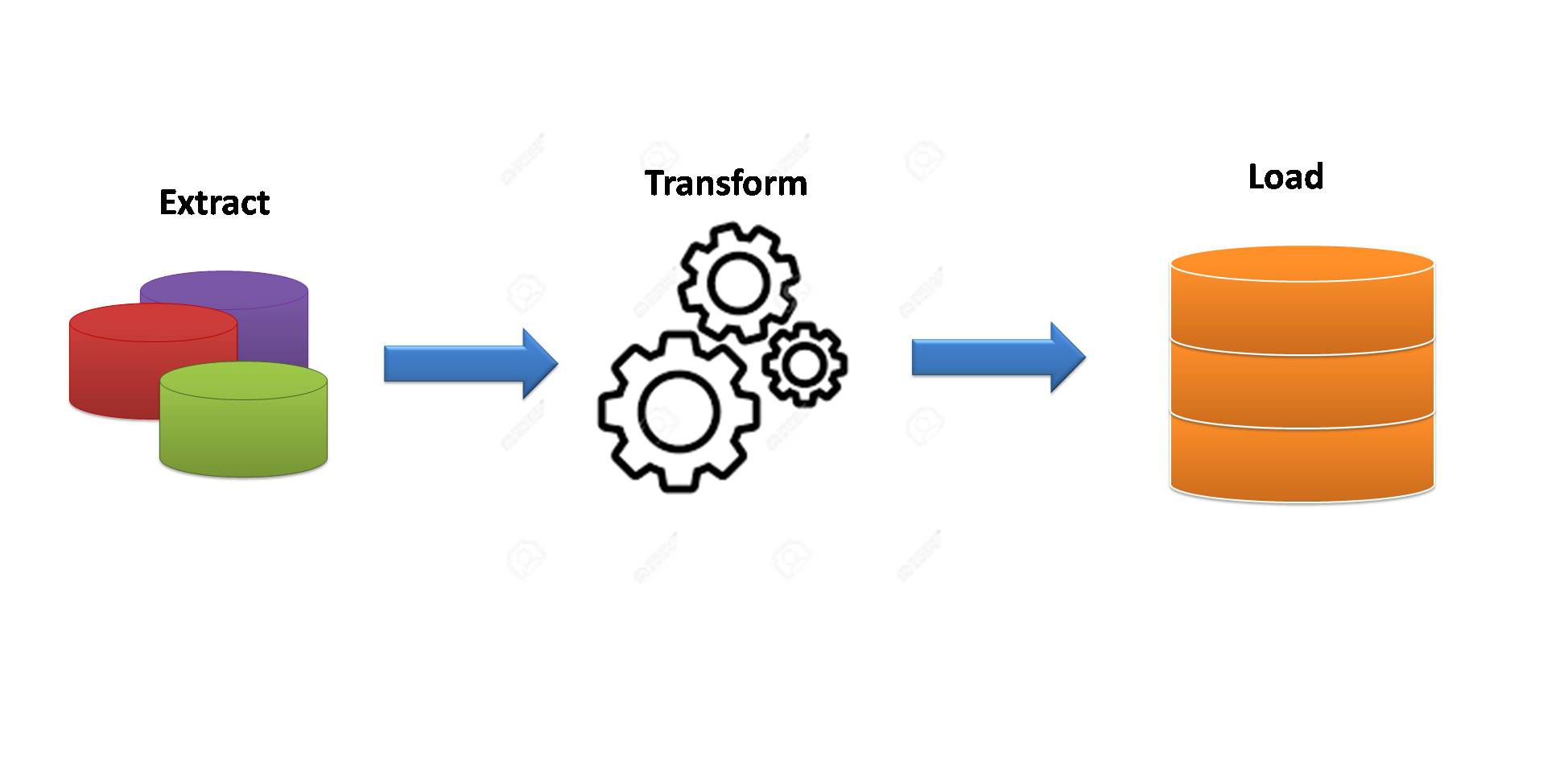
* **OPERATIONAL DATA STORE: -**

1. Data storing can be in two tyoes
2. Operational data
3. External data
4. Business data can be extracted from both the data which are related to business like stock markets etc.



* ETL means extract transform load

**ETL :**



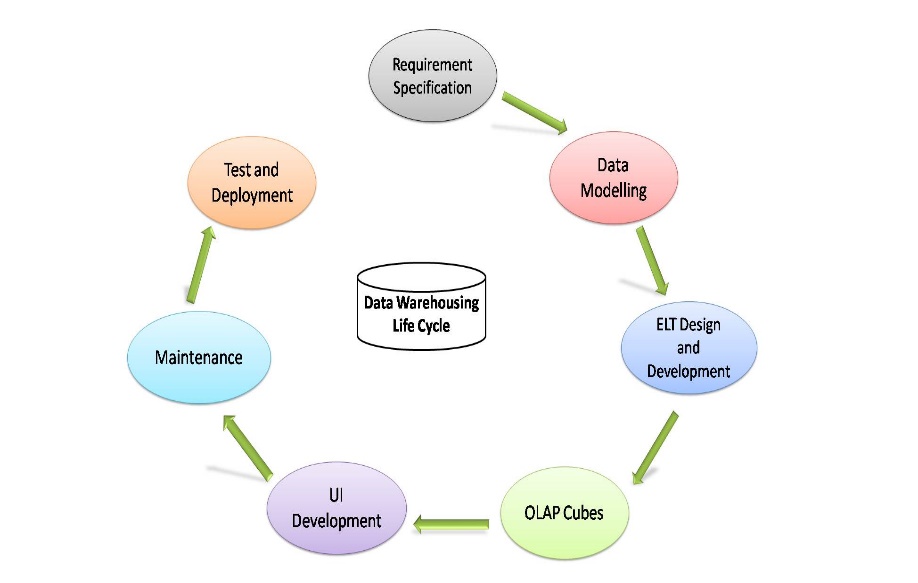
* **OLTP VS WAREHOUSE : -**

|  |  |
| --- | --- |
| OLTP | WAREHOUSE |
| 1. This data can be latest and most recent one because updates are carried out continously | 1. This data is acollection of data from a period that is updated periodically and removes unnecessary data |
| 1. OLTP works and only supports pre-defined operations | 1. It is designed to optimize to perform well for variety of possible query operations |
| 1. It allows us to use it to read, update and delete the data | 1. This allows us only to read without making it to update and delete |
| 1. It has more detailed in terms of explanation because it is obtained at present time | 1. It has a slightly less detailed explanation than OLTP because it is obtained from past |

* **APPLICATIONS : -**
* Business intelligence
* Decision making
* Finance
* Distribution
* Data mining
* Retail data management
* Banking
* **DATA MARTS : -**
* It is data storage system that contans information specific to an organizations business unit
* It is also called the subset of datawarehouse
* For example marketing team may use data marts to analyze consumer behaviors.
* **DATA MARTS VS DATAWAREHOUSE :-**

|  |  |
| --- | --- |
| DATA MART | DATA WAREHOUSE |
| 1. It is just a part of database | 1.It is a collection of databases |
| 1. It is decentralised system | 2.It is a centralised system |
| 1. It is bottom- up model | 3.It is top-down model |
| 1. It is easy to built | 4.Difficult to built |
| 1. This is not flexible | 5.this is flexible |
| 1. It is project-oriented in nature | 6.Object oriented in nature |
| 1. Data is in summarized form | 7.Data is in detailed form |
| 1. Size is less than 100gb | 8.It is in between 100gb to 1tb |
| 1. Generally, stores data from data warehouse | 9.It collects data from various data sources |
| 1. Easy to design | 10.Complicated to design |

* **DATA WAREHOUSE LIFE CYCLE : -**



Life cycle includes following phases: -

Requirement analysis

Design

Construction

Testing

Deployment

Operation

Maintenance

Retirement

