Data Engineering

Day 1 Assignment

Submitted By Mihir Kumar

An introduction to Data Engineering

Data engineering is about collecting, organizing, and fixing data so that it's easy to use and understand. It's like arranging and fixing Lego blocks so that you can build cool things with them later.

Data Enginner

A data engineer is someone who works with data. They focus on designing, building, and managing systems to collect, store, and process data.

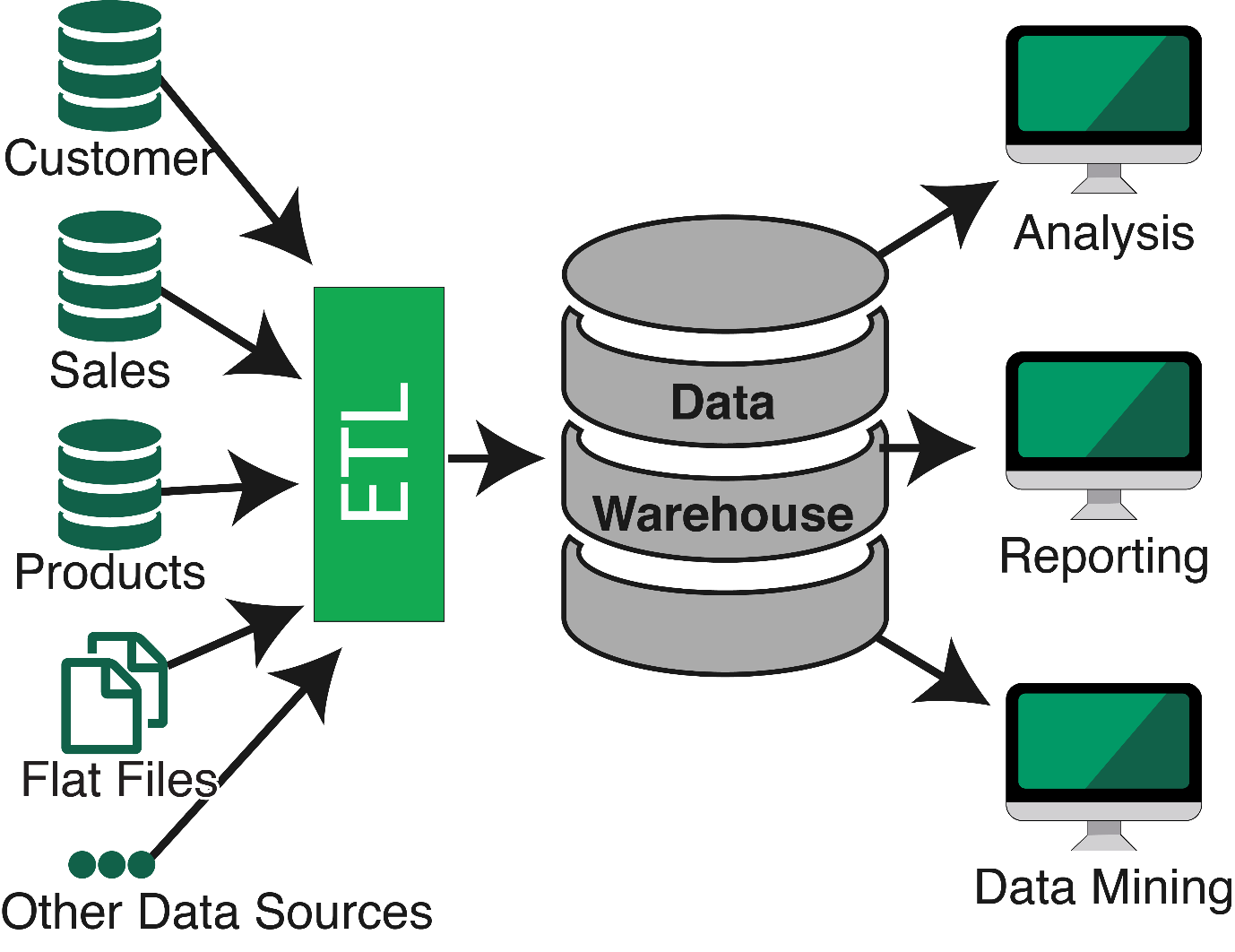
ETL

Stands for Extract , Transform and Load.

Extract- This step involves getting data from various sources.

Transform- Once the data is extracted, it often needs to be cleaned, organized, and transformed into a usable format.

Load- Once the data is transformed ,it is loaded in targeted database or data warehouse . So that it can used for various purposes like analytics , machine learning etc.



Data-It is a collection of raw facts , symbols , images etc.

Types of Data

1. Raw Data:- It is the Data that we generally get from different sources that contains duplicate , unstructured data which can’t be used for learning purposes.
2. Processed Data:- It is the Data that we get after processed of raw data. It removed all the unwanted data.
3. Cooked Data:- It is the one which we use to for learning purposes because it is the most summerized form of data

Vs of BigData

* Volume:- It specify the amount of data
* Velocity:- It specify how fast data is travelled
* Variety:- It specify the different kind of data
* Veracity:-It specify reliablity of data

Batch Data Processing

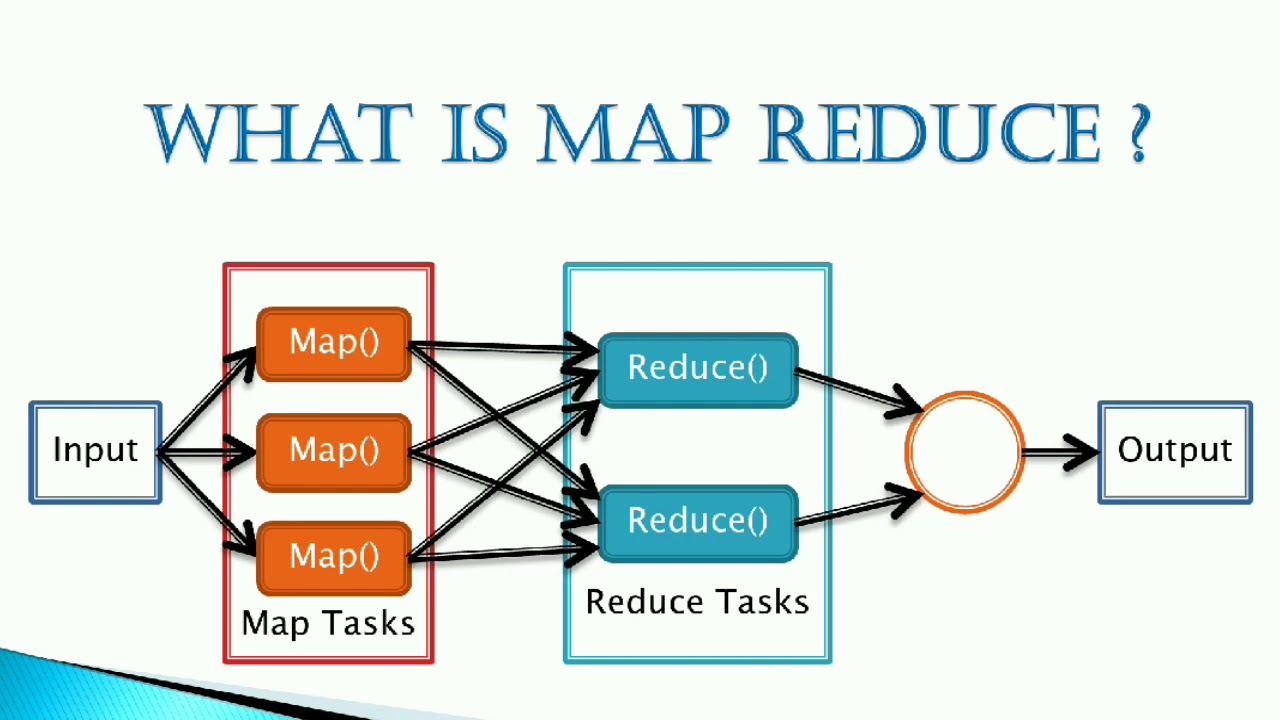
Batch data processing is a way of processing data in batches or chunks.In this processing , data is collected from different data sources , store and then use that data for analytics so that we can get some insight about the data.

Stream Data Processing

It is a another type of Data Processing which process the data continuously.

MapReduce

It is a type of model which is used for data processing by doing key value pairing .It splits data in key value pair and then after rearranging the key it reduces to single file.



Data Warehouse

It is a system that is used to store different kinds of data from different sources.

Purpose of Data Warehouse

* It is subject oriented means it provides topic wise information like sales
* It is integrated means it is a collection of other databases
* It is time varient means it provides the information about the past
* It is non-volatile means the data that is store in the data warehouse stays in as well as we keep the data

Data Warehouse architecture

Data warehouse architecture is the structure and layout of a system designed to store, manage, and analyze large volumes of data.

It contains many components:-

* Data sources
* ETL
* Data storage
* Analysis

Operational Data Store

It is like a traditional Data store but it focuses on real-time data it does not provide historical data

OLTP

It stands for online transaction processing

It is a system which is used for day to day transaction like in ATM

Data warehouse Applications

Data warehouse has many applications in different areas:-

* Businesss analysis
* Market analysis
* Healthcare analytics etc.

Data Marts

It is subset of data warehouse which includes only particular type of data like sales etc.

Data Marts vs Data warehouse

Data Marts is a subset of data warehouse which contain only particular bussiness unit of data. But

Data warehouse is a collection of all kinds of data .

Data warehouse lifecycle

* Planning
* Design
* ETL
* Implementation
* Testing
* Deployment
* Maintenance