**Name: K. Praveen Kumar**

**What is Data Engineering?**

Designing, building and scaling systems that organize the data for analytics.

The analytics refers to the amount of data that is to be measured.

**ETL :**

**ETL** stans for Extract, Transform and Load

**Architecture of ETL**:

IOT Devices

Analytics

Front end

SQL dB

API

**My Data Science Platform Blueprint :**

Connect Visualize

Web UI’s

Store

Bigdata sql DB/Warehouse

Processing Framework

Stream Batch

Buffer

Caches Message Queues

API

Data flow apps

BI Tools

External

Enterprise data warehouse

SQL DB’s

Mobile Apps

The Data is classified into different types and they are :

1. **Raw data** : It comes from different data sources.

It is an unprocessed data where no schema is applied.

1. **Processed data** : In this the raw data is processed with applied schema.

This data will be stored in tables/events in schema pipeline.

1. **Cooked data**: This data comes as an output to the end users.

All of this data will be stored in Big Data. There are number of applications are there to store the big data like My SQL , Mongodb , etc…

To Provide solutions to this data the **Data Engineering** Comes into picture.

**Properties of Big Data:**

There are four properties called as 4v’s and they are:

1. **Volume** – How much amount of data is transferred.
2. **Velocity** – How faster the data is transferred.
3. **Variety** – Different types of data.
4. **Veracity** – How reliable the data is.

**Batch Processing:**

Data Storage Analytics Insight

In order to store and perform the analytics of the data there are some required tools .

Insights refers to the Tables, Pictures and Dashboard.

**Stream Processing:**

It processes the data on the flow as it comes in.

For example, if we take Hotstar and Youtube the videos will be continuously streamed and the data will be continuously processed.

**Map Reduce**:

Big data works on the concept of map reduce.

Map reduce uses key-value pairing and it organize the data into keys and values.

It sorts and combines the data by using the key.

It will repeat until the final key-value outcome was found.

There are several tools for the map reduce to execute like Apache Spark, Hadoop, Beam and Samza.

**What is Data WareHousing?**

To understand what is Data warehousing, firstly we need to understand what is meant by Data Warehouse?

**Data Warehouse:**

A data warehouse is a Subject oriented, Integrated, Time variant and Non-volatile collection of data in support of managements system.

Or

Simply we can say that the data warehouse is a relational database designed for analytical needs.

* It organizes the collection of data which is collected from different data sources.

**Data Warehousing**: Data warehousing is the act of organizing and storing the data in a way so as to make its retrieval efficient

**Features of Data Warehousing:**

* **Subject-Oriented :** It organizes the data according to subject instead of application.

It mainly focuses on modeling and analysis of data.

* **Integrated:** Constructed by integrating multiple data sources like Relation dB, Flat files and Online Transaction records.

Here Integrating means combining of data sources.

* **Time Variant:** Time horizon for data warehouse is significantly longer i.e., it can provide past data from 5-10 years.
* **Non – Volatile**: Once the data is entered it will be never removed and No updates are allowed for the data.

**Decision Support Staff:**

The need for DSS in business is that it helps us to assess and resolve everyday business questions.

The DSS will have both the structured and Unstructured components.

DSS Architectural Styles:

1. **OLTP** (Online Transaction Processing) : used by RDBMS.
2. **OLAP** (Online Analytical Processing) : used by warehouses.

**Operational database :** It is the one which is accessed by the operational system to carry out at regular operations.

It uses OLTP Architecture.

**OLTP**: To Provide end users with access to large amount of data.

**OLTP Architecture**:

To understand the architecture an ATM is considered and the architecture is shown below

Auditors

Analysis Harinya Trainer (Guest)

An introduction to Data Warehousing  
Purpose of Data Warehouse  
Data Warehouse Architecture  
Operational Data Store  
OLTP Vs Warehouse Applications  
Data Marts  
Data marts Vs Data Warehouses  
Data Warehouse Life cycle"

Warehouse

Raw Audit data Table

Reporting

Mining

Advantages of OLTP:

1. Simple and Efficient
2. Data Integrity
3. Faster Query Processing

Pitfalls of OLTP:

1. It requires instant update.
2. The data that is acquired from OLTP is not suitable for data analysis.

**Operational Data Store:**

The data which is at the warehouse is first passed to the Operational Data store(ODS).

The data which is integrated is passed back to the operational systems for decision making.

The Data Store contains two main types of data :

1. **Business data :** This is extracted from operational database and from external sources.

Operational data

Business data

External data

1. **Business data model**

**DSS Data vs Operational data** :

The DSS data which is extracted from multiple sources differ from operational data in mainly 3 areas:

1. Time Span
2. Granularity
3. Dimensionality

A Typical ETL based warehouses uses:

1. Staging
2. Integration
3. Access Layers (Data marts) to house its key functions.

**Data Marts:**

The data which is stored in the warehouses are in the form of data marts.

Data Mart

Data Warehouse

Data Mart

Data Mart

**Data Marts vs Data Warehouse :**

A data warehouse stores structured data and it is a central repository of raw data for analytics.

A data mart is a part of a data warehouse that serves the needs of a specific unit.