DataEngineering

Data Engineering is the process of designing, building and scaling systems that organize the data for analytics

Data Engineering uses Technique called ETL for Analytics

**ETL** – **E**xtract, **T**ransform and **L**oad

**My Data Science Platform Blueprint**

**Visualize**

Web UI

BI Tools

Mobile Apps

**Processing Framework**

Stream Batch

**Buffer**

Caches Message Query Queries

**Store**

Big Data SQL DB

.,

External

Enterprise Data warehouse

SQL DB’s

**Connect**

API’s

Data Flow Apps

**Data Classification:**

1. **Raw Data:**

* It is unprocessed data
* Example: JSON
* No Scheme was applied in this data

1. **Processed Data:**

* In this processed data schema will be applied
* It is stored in Tables

1. **Cooked Data:**

* The processed id further reduced according to the problem statement or necessity

**Big Data:**

In order to store the larger volumes of data, Big Data is the solution for that.

**Big Data Properties:**

1. Volume - How much data?
2. Velocity – How fast data is getting to you?
3. Variety – How different your data is?
4. Veracity – How reliable your data is?

**Batch Processing:**

Processing of high volume of data in batch, where we can able to get insights

**Stream Processing:**

The processing takes place with the help of real time data

**Map Reduce:**

**Key – Value Pairing**

* Organizes the data into key and values
* Sort by the key
* Combine the data with matching keys
* Process will be repeated until getting the final key
* Big Data works in **Map Reduce**

**Tools in Big Data:**

* Hadoop
* Apache spark
* Beam
* Samza

**Two Types of Databases:**

1. Relational Database – SQL
2. Document Store – NoSQL

**Data Warehousing:**

It is a subject-oriented, integrated, time variant, non-volatile collection of data in support of management system.

**Features of Data Warehousing:**

1. Subject - oriented
2. Integrated
3. Time variant
4. Non – Volatile

**DSS Architectural styles:**

* OLTP – Online Transaction Processing (Used by RDBMS)
* OLAP – Online Analytical Processing (Used by Data warehouse)

**Operational Database:**

It is accessed by an operational system to carry out the regular operations of an organizations which uses OLTP architecture

**OLTP:**

Here it refers to class of systems that manages transactions – oriented applications which is typical for data entry and retrieval transaction processing

Example: ATM

**Benefits of OLTP:**

* It’s simple and efficient
* Supports data integrity
* Faster Query Processing

**Drawbacks of OLTP:**

* It requires instant update
* Data from OLTP not suitable for Data Analysis

**Types of Data Storage:**

1. Business Data
2. Business Data Model

**Business Data:**

The business data are extracted from operational databases and from external data sources. The business data represent the actual snapshot of the company’s situation