**G. Rupa Manasa**

**DAY 1 ASSESSMENT**

***INTRODUCTION TO DATA ENGINEERING***

***DEFINITIONS:***

**Data Engineering:** It is a process of designing, building and scaling systems that organize data for analytics. Analytics refers to how much amount of data is generated with respect to some time frame.

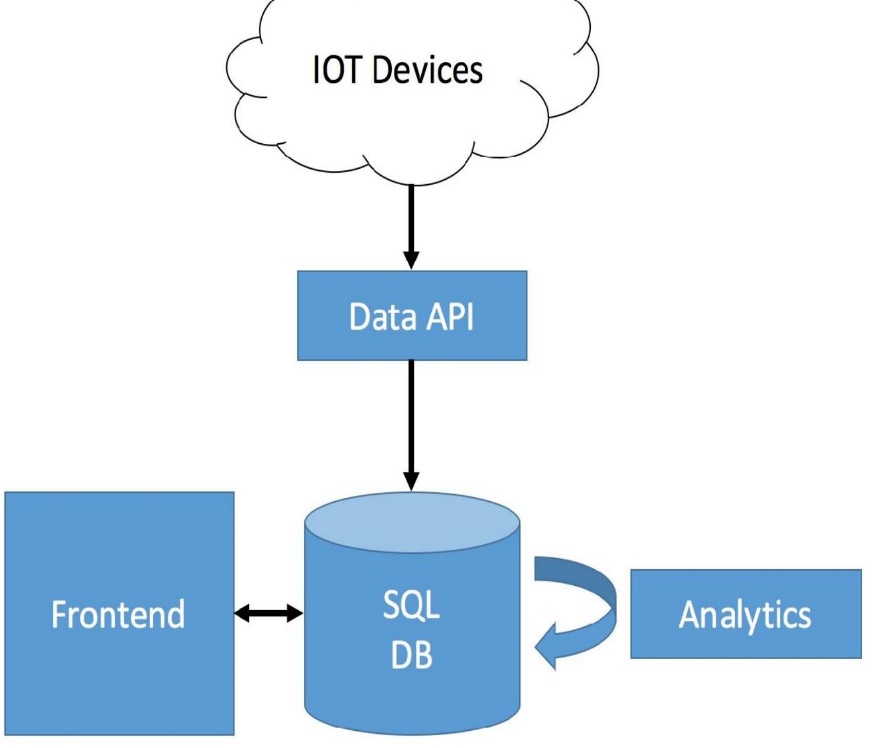
**What does Data Engineers Do?**They build and scale the platforms that enable data collection, processing and storage for data science/business analytics use.

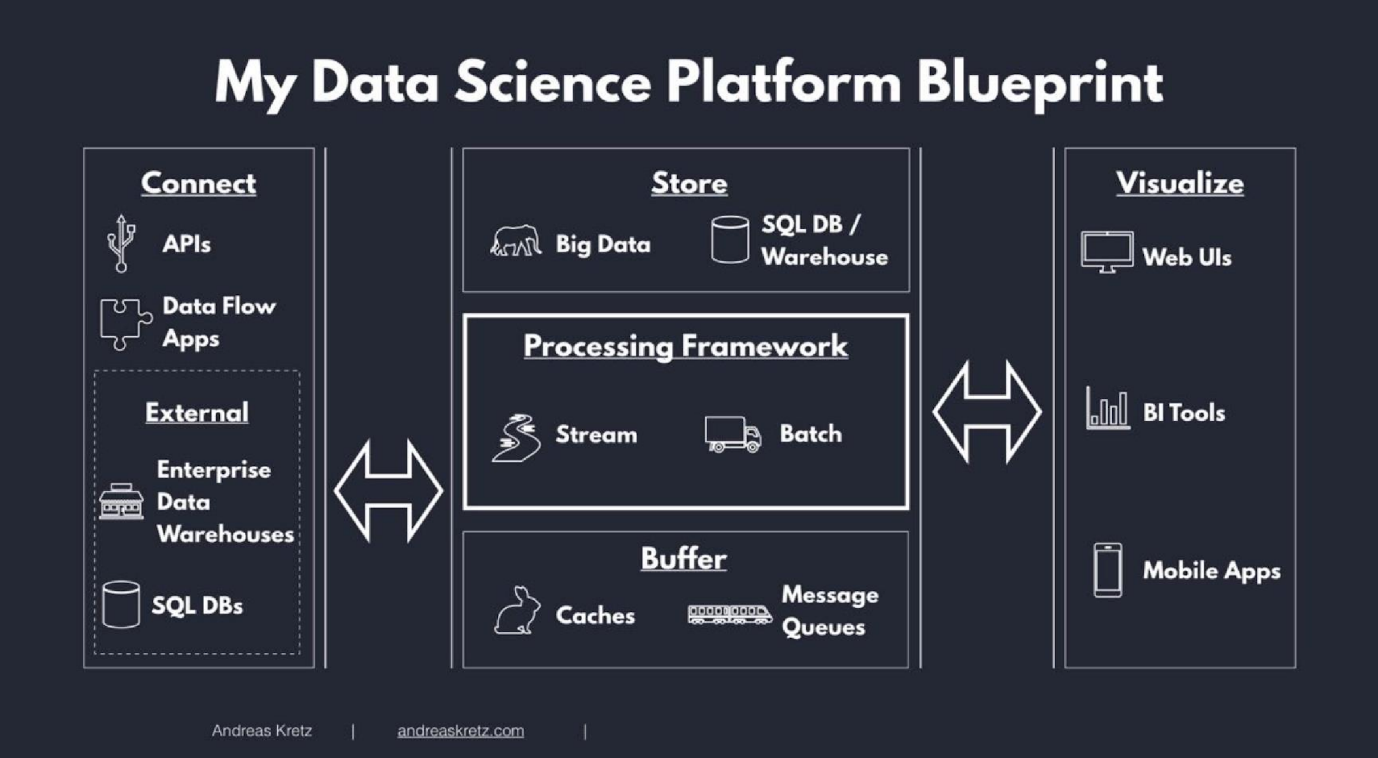
One of the data engineering tool is ETL.

**ETL:**

ETL stands for Extract Transform Load.

The basic architecture of ETL is shown below. It involves, the data is extracted from the IoT devices and then stored in the database in which it is then organized for frontend and analytics.



**Data Science Platform Blue print:** The blue print involves the connect, external, store, processing framework and visualize.

**Data Classification:**

It consists of Raw data, Processed data and cooked data.

*Raw data:* It is an unprocessed data just as it is come from data sources. No schema applied.

*Processed data:* It is raw data with schema applied.

*Cooked data:* It is the data which is finally summarized. It is used for analytics and it is the important data.

**Big Data**

In order to store all the data either it may be raw data or processed data or cooked data, we need big data.

**Properties of Big Data:** It has 4 properties which is popularly known as 4V’s.

1. Volume: Means how much data we have.
2. Velocity: Means how fast the data is getting to you.
3. Variety: Means how different the data is.
4. Veracity: Means how reliable our data is.

**Data Processing Methods:** 2 methods

1. Batch Processing: Here, the data is stored and then processed for analytics and finally the data will go through insight which means we are going to justify whether the data is appropriate or not.
2. Stream Processing: Process data on the fly, as it comes in. Atleast once, Atmost once and Exact once are the methods in stream processing.

**Map Reduce:**

Big data works on map reduce.

It involves Key – Value pairing.

Organize the data into keys and values, sort by the key.

**Some Data Engineering Tools:**

Azure databricks, Hadoop, apache Spark, Samza.

**Data Warehouse (DW)**: It is a Subject oriented, integrated, time variant, non-volatile collection of data in support of management’s system.

It is a kind of database.

It is a place where we are keeping entire data.

Consists of database machines and virtual machines.

It is a database used for data reporting and analysis.

**Features of Data Warehouse:**

1. *Subject Oriented*: Data is structured based on subjects rather than applications.
2. *Integrated:* Constructed by integrating multiple, heterogeneous data sources like relational databases, flat files, on-line transaction records.
3. *Time Variant:* The time horizon for the data warehouse is significantly longer than that of operational systems.
4. *Non – Volatile:* The data is never deleted and no updates are allowed.

**Purpose of Data Warehouse:** The purpose of a data warehouse is used for storing, managing, and analysing large volumes of data from various sources.

The data stored in warehouse are uploaded from the operational system.

**Decision Support System (DSS):**

DSS helps us to assess and resolve everyday business operations.

It works by compiling useful information from a combination of raw data, documents, personal knowledge, or business models.

There are 2 architectural styles in DSS.

1. OLTP – Online Transaction Processing.
2. OLAP – Online Analytical Processing.

**OLTP:**

OLTP is a methodology to provide end users with access to large amounts of data.

OLTP refers to a class of systems that facilitate and manage transaction-oriented applications.

It is known for its simplicity, efficiency, data integrity and fast query processing.

Best Example is ATM.

The drawback is that the data what we get from OLTP is not suitable for data analysis and it requires instant update.

**Data Store in Data Warehouse:**

The data store contains two main types of data.

Business Data

Business Data Model

The business data are extracted from operational database and from external sources.

**Operational Database:**

It is accessed by an operational system to carry out regular operations of an organization.

It uses an OLTP architecture for fast transaction processing.

**ETL based Data Warehouse:**

It uses staging, integration and access layer (data marts).

Staging is a phase which involves keeping of data into warehouse by any mediator like Relational database.

The data that arrived at data warehouse are first passed to Operational Data Store (ODS).

Data is integrated from multiple sources for additional operations on the data.

The data stored in data warehouse is stored in the form of Data Marts i.e., data mart is a subset of data warehouse.