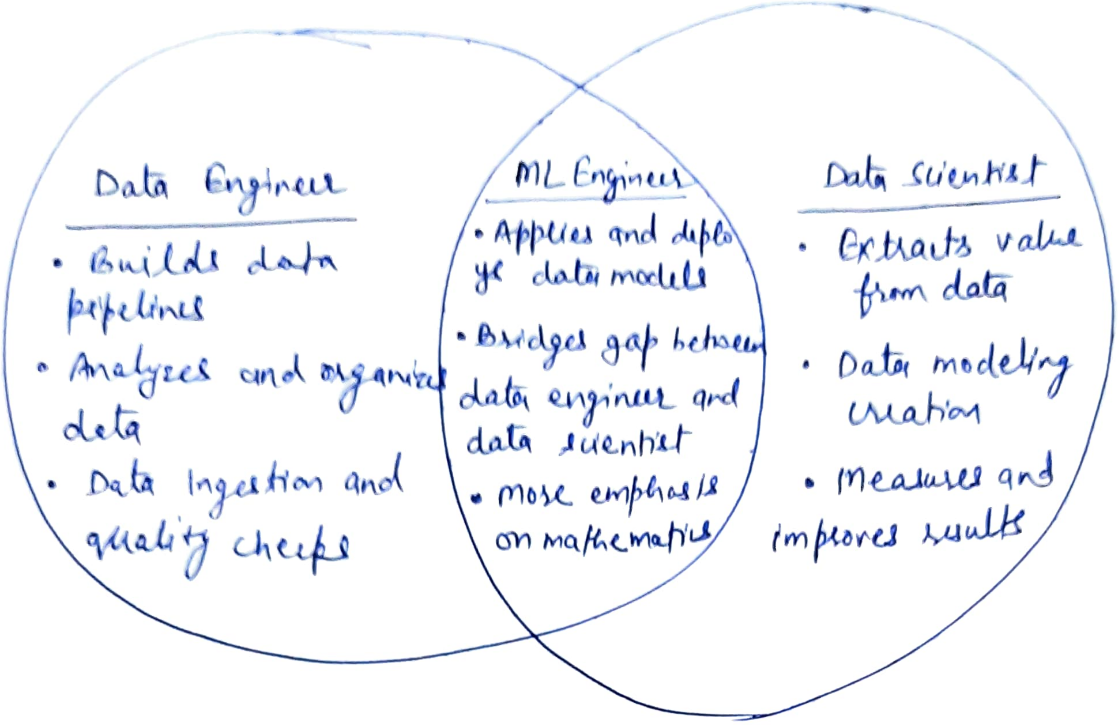
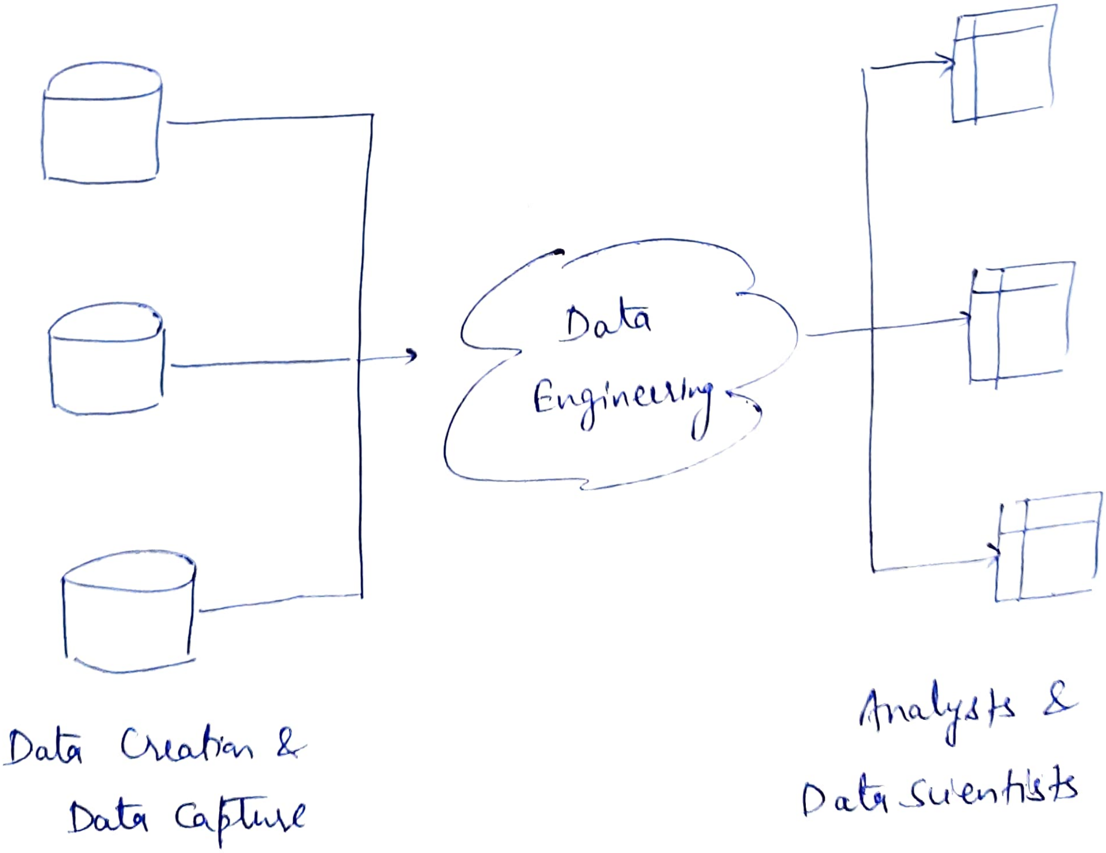
Data Team at a glance


Data Engineer
• Builds data pipelines
• Analyses and organizes data
• Data Ingestion and quality checks
ML Engineer
• Applies and deploys data models
• Bridges gap between data engineer and data scientist
• more emphasis on mathematics
Data Scientist
• Extracts value from data
• Data modeling creation
• Measures and improves results



Difference between Data Engineering and Data Mining

- **Data Engineering** is typically more focused on the back-end solution.
- They develop the architecture or schema on how all of the relationship between disparate data sources integrates together to tell one story.
- Their work will include data modeling, ETL structures and framework, and integration of multiple data sources into a single usable model

- **Data Mining** means to dig through something hoping to find something of value.
- It is the ability to develop models to get actionable information out of the data once it is in a usable format (from the Data Engineers)
- This includes, but not limited to, machine learning techniques, statistical model, regression, time series techniques, or even something as simple as grouping and average.

Key differences between Data science and Data Mining

(i) Data science includes the processes of capturing of data, analyzing and deriving insights from it. Data mining is mainly about finding useful information in a dataset and utilizing that Information to uncover hidden patterns.

(ii) In data science you are not only finding patterns and analyzing them which are key components of the data mining instead, with the help of data science tools and technologies, you are expected to be able to forecast future events by leveraging the present and historical data.

(iii) Data science deals with every type of data whether structured, semi-structured or unstructured on the other hand, data mining mostly deals with structured data.

(iv) The role of a data science professional can be considered as a combination of an AI researcher, a deep learning engineering, machine learning engineer or a data analyst, to some extent. and might be able to perform the role of a data engineer as well.

- on the contrary, a data mining professional does not necessarily have to be able to perform all these roles.

Big Data Ecosystem

- Big data ecosystem is the comprehension of massive functional components with various enabling tools.
- Capabilities of the big data ecosystem are not only about computing and storing big data, but also the advantages of its systematic platform and potentials of big data analytics.

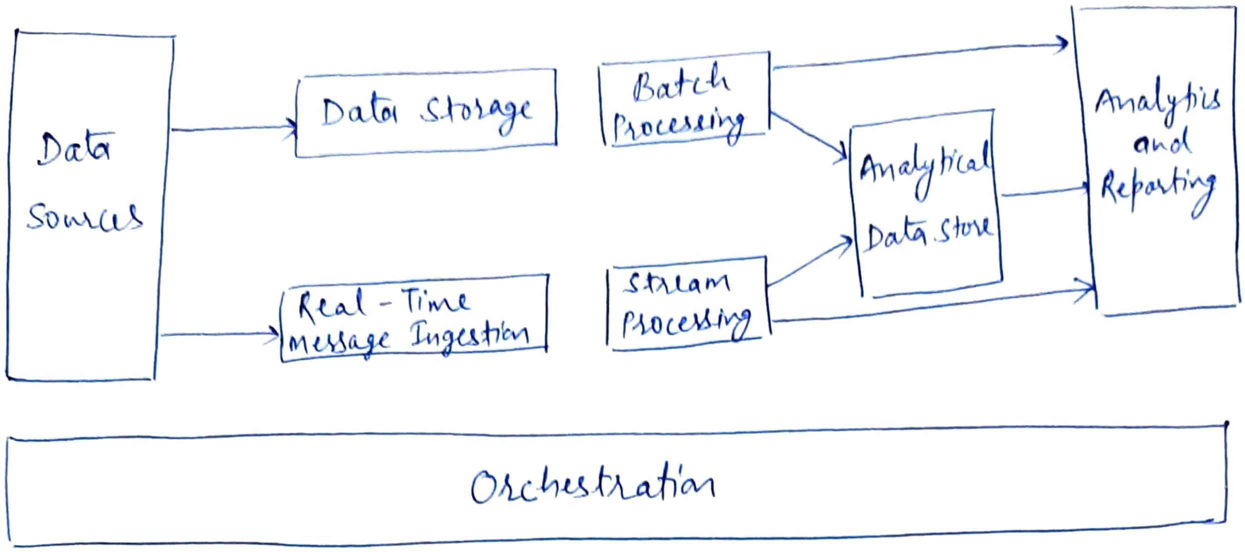
Componets and architecture of Big data :-

(i) Data sources

(ii) Data management (Integration, storage and processing)

(iii) Data analytics, Business intelligence (BI) and knowledge discovery (KD).

(i) Data sources :→ In a modern data ecosystem, the data sources layer is composed of both private and public data sources



- The corporate data originates from internal systems, cloud - based systems, as well as external data provided from partners and third parties.
- Any type of data can be acquired and stored but the most challenging task is to capture the heterogeneous datasets from various service providers.
- In order to allow developers to create new applications on top of open datasets, machine -readable formats are needed.
- As such, XML and JSON have quickly become the de facto format for the web and mobile applications due to their ease of integration into browser technologies and server technologies that support Java script.

- The public data sources (statistics, trends, conversations, Image, videos, audios, and podcasts for instance from google trends, Twitter, Instagram and others) provide real-time information and on-demand insights that enable businesses to analyse user interactions, draw patterns and conclusions.

Examples of big data ecosyetems

Facebook

Facebook (2018) has more than two billion users on millions of servers, running thousands of configuration changes every day involving trillions of configuration checks.

Alibaba

The 402,000 web-facing computers that Alibaba hosts (2017) from china-allocated IP addresses would alone be sufficient to make Alibaba the second largest hosting company in the world today.

Google

There's no official data on how many servers there are in Google's data centers, but Gartner estimated in a July 2016 report that Google at the time had 2.5 million servers.

LinkedIn

Amazon ..-----

Twitter

(ii) Data Management :→

- As data become increasingly available (from social media, web logs, IoT sensors etc.), the challenge of managing (selecting, combining, storing) and analyzing large and growing data sets is growing more urgent.
- From a data analytics point of view, that means that data processing has to be designed taking into consideration the diversity and scalability requirements of targeted data analytics applications.

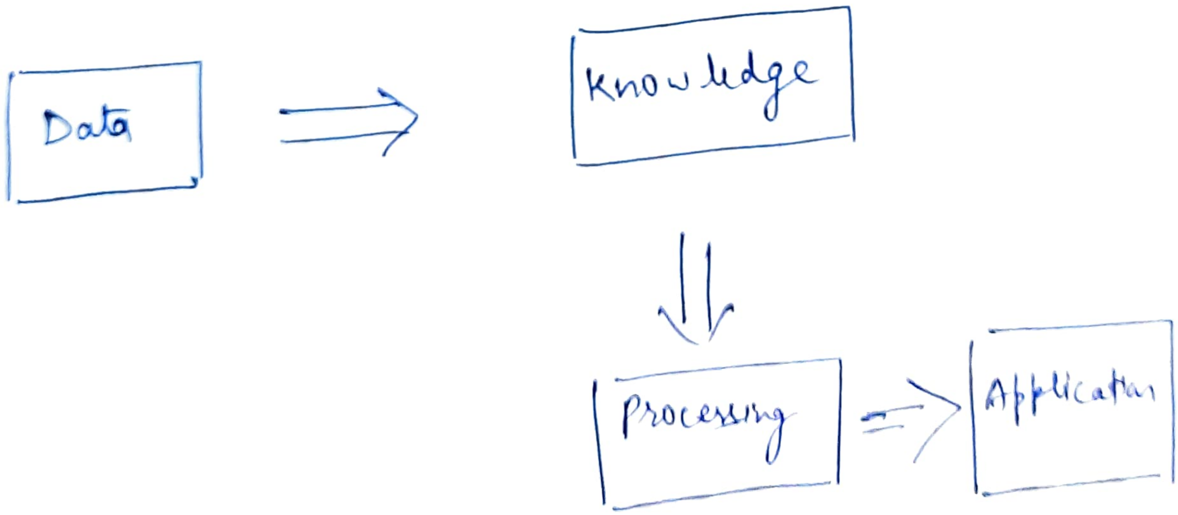
Over the last two decades, the emerging challenges in the design of end to end data processing pipelines were addressed by computer scientists and software provider in the following ways!

→ In addition to operational database management systems (present on the market since 1970s), different NoSQL stores appeared that lack aherence to the time-honored SQL principles of ACID (atomicity, consistency, isolation, durability).

→ Cloud computing emerged as a paradigm that focuses on sharing data and computations over a scalable network of nodes including end user computers, data centers, and web services.

→ The Data Lake concept as a new storage architecture was promoted where raw data can be stored regardless of source, structure and (usually) size.

The data warehousing approach is thus perceived as outdated as it creates certain issues with respect to data integration and the addition of new data sources.



Components of big data ecosystem