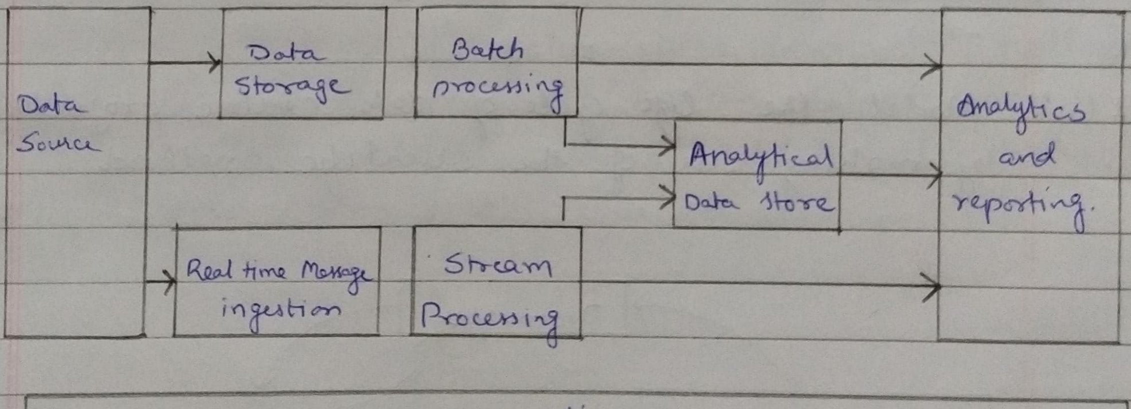
Name - Vasu Kalariya
Roll - PE29 (Batch 1)
Sub - BDA

Theory Assignment - 1

Q1 (a) Draw and explain the Big Data architecture of the System.

Big data architecture is the foundation for big data Analytics. It is the overarching system used to manage large amounts of data so that it can be analysed for business purposes, steer data analystics and provide an in which big data analystics tools can extract vital business information from otherwise ambigueous data

Big Data Solutions typically involve one or more of following types of workhood
→ Batch processing of Big Data source at rest.
→ Real time processing of Big Data in motion
→ Interaction explained of big data



The components are:

→ Data Sources: All big data solution start with one or more data source

Vasu Kalariya (DE29)

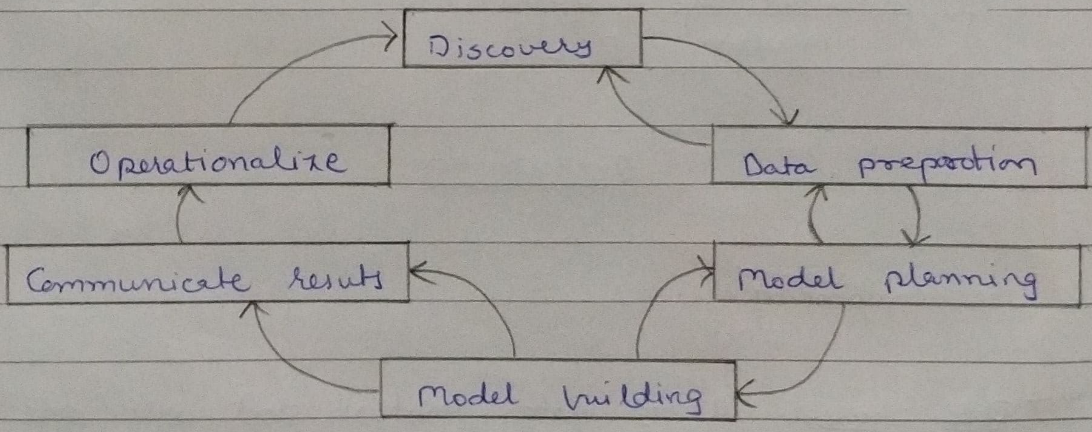
→ Data Storage: Data for batch processing operation is typically stored in distributed fill store that can hold high volume of large file in various formatted. This kind of store is often called Data lake.

→ Batch processing: Because data sets are so large often a big data solution must process data files using long running batch jobs to filters aggregate and otherwise prepare the data for analysis.

→ Real time message ingestion: If the solution indeed real time sources, the architechure must include a way to capture and store real time messages for stream processing.

→ Stream preprocessing: After capturing real time messages, the solution must process them by filtering, aggregating and otherwise ~~preprocess~~ preparing the data for analysis.

Q1 (b) Model the life cycle of data centric projects ~~&~~ by making use of the scientific method



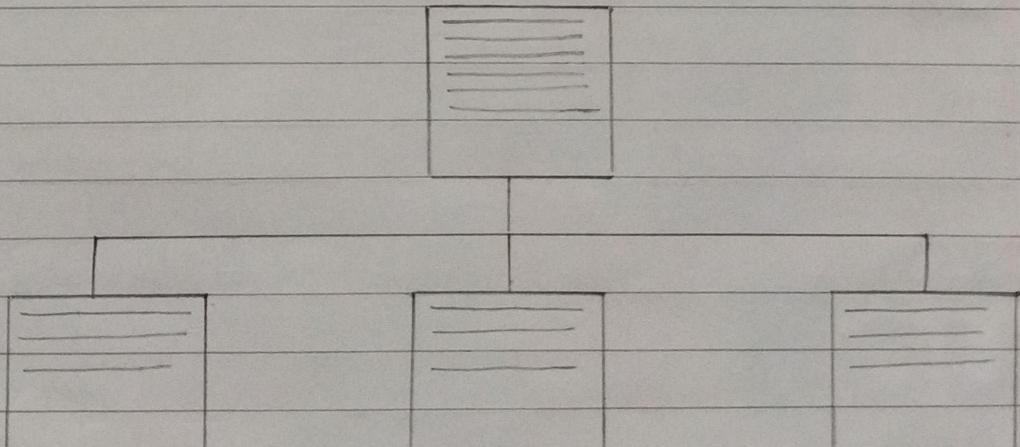
Vasu Kalariya (PE29)

Data analysis lifecycle define analytics process best practices. spanning discovery for project completion the life cycle draws from established methods in the real of data analystics and decision ~~sinoa~~ science

Phase: 1] Discovery
2] Data preparation
3] Model planning
4] Model Bulding
5] Communicatea results
6] Operationalize.

Q 2 (a) MongoDB use horizontal scaling for handling huge amounts of data. Defend it with the help of suitable diagram

Sharading is a method for distributing data across multiple machines. MongoDB uses horizontal scalling which involveds, dividing the system dataset and load multiple servers, adding additional servers to increase capacity as required.



Vasu Kalariya (PE29)

Q 2 (b) Consider Employee database having.
{ Eid, Department, Ename, Address : [street : streetname, city : cityname, state : statename, pincode : pincode-value }, phone : [home contact, mobile : contact], age, salary }

1) Display only 3 employees residing in the state : "Maharastra". Skip the first employee document

```
→ db. Employee. find ({ "Address. City" } : "Maharastra" }). limit(3).
  skip(1).
```

2 Eid should allow only unique value

```
→ db. Employee. create Index ({ "Eid" : 1 }, unique : true )
```

3 Display ~~dip~~ department-wise average salary of employee having the average salary > 5000.

```
→ db. Employee. aggregate ([
  { $group : { _id : " $ DeptName", Avg-Sal : { $ avg : "$salary"
  }}},
  { $ match : { Avg-sal : { $gt : 5000 }}}.
]);
```