UNIT-5

Ms. Shalaka Patil

Big data Analytics : Introduction to the Big Data Problem, challenges, trends, 2 applications, Technologies for big data management. Big data technology tools, Map-Reduce paradigm & the Hadoop]

* Introduction to Big data

- Big data is defined as large amount of data which requires new technologies & architectures to make possible to extract value from it by capturing & analysis process.
- New sources of big data include location specific data arising from traffic management and from the tracking of personal devices such as smartphones
- Big data has emerged because we are living in a society which makes increasing use of data intensive technologies.
- Due to such large size of data it becomes very difficult to perform effective analysis using the existing traditional techniques
- Big data concept means a datasets which continues to grow so much that it becomes difficult to manage it using existing database management concepts & tools.
- The difficulties can be related to data capture, storage, search, sharing, analytics & visualization etc
- Big data due to its various properties like volume, velocity, variability, value & complexity put forward many challenges:

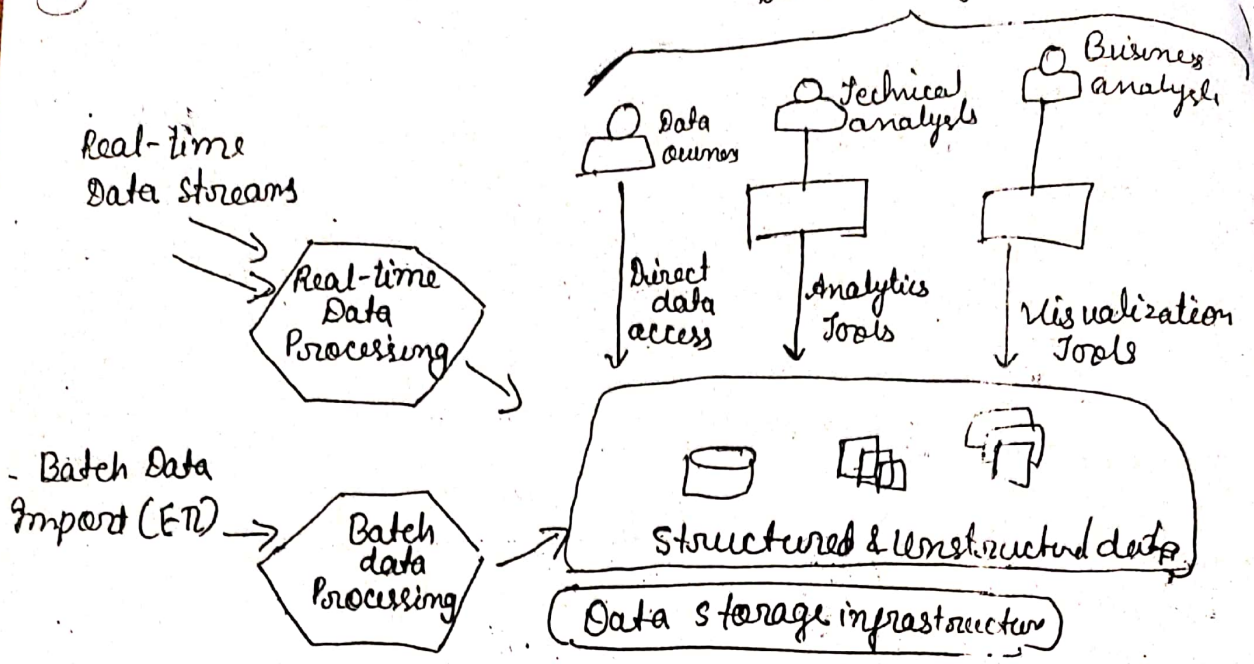

Fig Example of Big Data Architecture.

- The various challenges faced in large data management include - scalability, unstructured data, accessibility, real time analytics, fault tolerance & many more.

⇒ BIG DATA Characteristics

(i) Data Volume → The Big word in Big data itself defines the volume. At present the data existing is in petabytes (10^{15}) & is supposed to increase to zettabytes (10^{21}) in nearby feature.

(ii) Data Velocity :- Velocity in Big data is a concept which deals with the speed of the data coming from various sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows & aggregated.

(iii) Data Variety :- Data variety is a measure of the richness of the data representation - text, images video, audio etc

iv) Data value : Data Value measures the usefulness of data in making decisions

v) Complexity :- It measures the degree of interconnectedness & interdependence in big data structures such that a small change in one or a few elements can yield very large changes or small change that ripple across or cascade through the system.

~~Introduction to the Big Data Problem~~

Issues in Big data

- The issues in Big data are some of the conceptual points that should be understood by the organization to implement the technology effectively.

☆ Introduction to Big data Problems

1) More data than we are used to handling :-
- organisations storing or maintaining big data, is not always extremely high volumes
- As we are not dealing with lots of high volumes can produce a challenge

2) Data is unstructured
- Another problem with Big data is that they produce data which naturally doesn't fit into the conventional storage. (unstructured)
- As we have various kinds of data format such as images, videos, text and there are really poor match.
- so organisations often have problem how to to deal with this kind of unstructured data.

3) Data arrives quickly, may have very narrow usefulness window

Example - If you have fradulent section, then we have very less time to react to it.

4) So much data - but not sure what to analyze?

how you are going to get benifits from these data, what will be right mathematical models to be apply, what kind of mining approaches to be applied.

5) Integrate data with conventional systems

- Big data are not Isolated, we need to integrate these Big data systems with conventional systems
- There information flow need to be united and there should be provision for conversion or transform information from big data systems to smaller conventional systems, which can be handled by conventional systems

lack of Knowledgable people
process, manage & analyze data.

7) organization silos :- Many organizations are afraid of using Big data, as a new change, there is a competition between organizations who will be using Big data.

* Challenges in BIG DATA

The challenges in Big Data are usually the real implementation hurdles which require immediate attention

- Any implementation without handling these challenges may lead to the failure of the technology implementation & some unpleasant results

(i) Privacy & security

- It is the most important challenges with Big data which is sensitive & includes conceptual, technical as well as legal significance
  - → the personal information of a person when combined with external large data sets, leads to the inference of new facts about that person & its possible that these kinds of facts about the person are secretive & the person might want the data owner to know. or any person to know about them.
  - → Information regarding the people is collected & used in order to add value to the business of the organization
  - → Another important consequence arising would be social stratification where a literate person would be taking advantage of the Big data predictive analysis.

→ Big data used by law enforcement will increase the chances of certain tagged people to suffer from adverse consequences without the ability to fight back or even having knowledge that they are being discriminated.

ii) Data Access & Sharing of Information

- If the data in the companies information systems is to be used to make accurate decisions in time it becomes necessary that it should be available in accurate, complete & timely manner.
- This makes the data management & governance process bit complex adding the necessity to make data open & make it available to government agencies in standardized manner with standardized APIs, metadata & formats thus leading to better decision making.

ii) Analytical Challenges

- The main challenging questions are as:-
→ What if data volumes gets so large & varied & it is not known how to deal with it?
→ Does all data need to be stored?
→ Does all data need to to analyzed?
→ How to find out which data points are really important?
→ How can the data be used to best advantage?
- Big Data brings along with it some huge analytical challenges.
The type of analysis to be done on this huge amount of data which can be unstructured, semi structured or structured requires a large number of advance skills.
- Moreover the type of analysis which is needed to be done on the data depends highly on the results to be obtained i.e decision making

iv) Human Resources & Manpower

since Big data is at its youth & an emerging technology so it needs to attract organizations & youth with diverse new skill sets.

- These skills should not be limited to technical ones but also should extend to research, analytical, interpretive & creative ones
- These skills need to be developed in individuals hence requires training programs to be held by the organizations

(v) Technical Challenges

→ Fault Tolerance :- With the incoming of new technologies like Cloud computing & Big data it is always intended that ~~what~~ whenever the failure occurs the damage done should be within acceptable threshold rather than beginning the whole task from the scratch.

- Fault tolerance computing is extremely hard.
- Two methods which seem to increase the fault tolerance in Big data are as follows :-
  (i) First is to divide the whole computation being done into tasks & assign these tasks to different nodes for computation
  (ii) Second is, one node is assigned the work of observing that these nodes are working properly.

→ Scalability :- The scalability issue of Big data has lead towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters.

- ~~which~~ This requires a high level of sharing of

resources which is responsible for dealing with all various challenges like how to run & execute various jobs so that we can meet the goal of each workload cost effectively

→ Quality of Data :- Collection of huge amount of data & its storage comes to at a cost.
- More data is used for decision making or for predictive analysis in business will definitely lead to better results.
- Big data basically focuses on quality data storage rather than having very large irrelevant data so that better results & conclusions can be drawn.
- This further leads to various questions like how to it can be ensured that which data is relevant, how much data would be enough for decision making & whether the stored data is accurate or not to draw the conclusion from it etc

→ Heterogeneous Data :- Unstructured data represents almost every kind of data being produced like social media interactions, to recorded meeting, to handling of PDF documents, fax transfers, to emails & more
- Working with unstructured data is cumbersome & of course costly too.

* Top Big Data Technologies & Tools- Hadoop & NoSQL

Here we are going to discuss the tools that are used for solving big data from technology standpoint- Hadoop (HDFS, MapReduce) which is an open source computing framework & NoSQL which is non-relational database

⇒ Big Data Technologies & Tools

① HADOOP [High availability distributed object-oriented platform]

- It is a software framework which analyse structured & unstructured data & distribute applications on different servers.

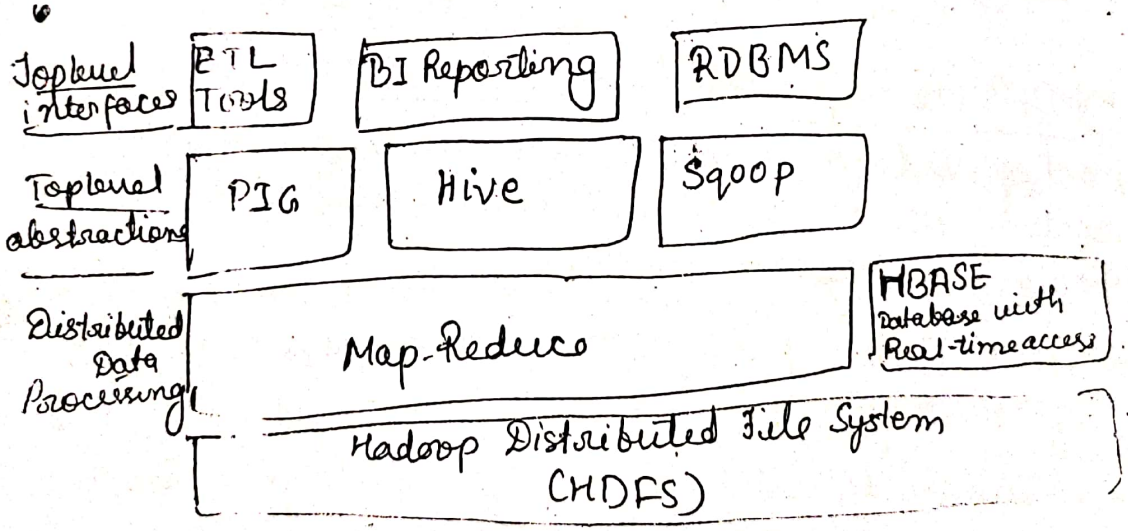

fig Hadoop architecture

- Basic Application Of Hadoop
  - Hadoop is used in maintaining, scaling, error ~~dealing~~ handling, self healing & securing large scale of data

- These data can be structured or unstructured
- i.e if data is large then traditional systems are unable to handle it.
- Thus, Hadoop comes in the picture.
- Below are some basic features of Hadoop
  - → Hadoop maintains & secures the data by storing & keeping its replica
  - → It is focused on scaling according to data usage
  - → It can detect & delete the failed task & as well as failed transaction of data.
  - → It not only recovers the data but also automatically restores the data at its place

⇒ Typical Hadoop Platform Stack - HDFS + Hive + HBase + Pig

(i) HDFS (Hadoop distributed file system) is a part of Hadoop & is known as a special file system which deals with distribution & storage of large set of data

- HDFS stores file as sequence of same size of block except the last block
- It also deals with hardware failure & smoothen the data handling

(ii) Hive :- Hive was initiated by facebook

- Hive is data warehouse tool which is based on Hadoop & converts query language into Mapreduce ~~task~~ jobs.
- It deals with the storage, analysis & queries of large set of data

Query language in hive used as HQL statement.
(Hive Query language)
- HQL is similar to standard SQL statement

(iii) Hbase → is a Hadoop application which runs on top of HDFS
- Hbase system represents set of table but Hbase is column oriented database management system
- Generally if we talk about database then we think of relational database system but unlikely Hbase is not relational database at all & also it doesn't support structured query language like SQL.
- Java is prefered language use for Hbase application
- One most important feature of Hbase is to real time read or write to large set of data

iv) Pig → initiated by Yahoo, became open source in 2007.
- It is named as Pig! because it can handle any type of data!! (strange but true)
- Pig is a high level procedural programming platform developed for simplifying large data sets query in Hadoop & MapReduce
- Pig has 2 components - (i) PigLatin → which is programming language & the other is (ii) run time environment → where PigLatin programs are executed

(2) NoSQL :- It means non relational or Non SQL database, refer to Hbase, Cassandra, MongoDb, Riak, CouchDB.

- It is not based on table formats & that's the reason we don't use SQL for data access
- A traditional database deals with structured data while a relational database deals with the vertical as well as horizontal storage system.
- NoSQL deals with the unstructured, unpredictable kind of data according to the system requirement
- NoSQL Technologies HBase, Cassandra, MongoDB, Riak, CouchDB.

→ Cassandra :- Database is used to handle the large set of data when we need to scale the database with high performance

- It deals with fault tolerance & replication of the data
- It is a partial relational database system, supports best query capability but don't have joins feature
- It follows the column family model map with 2 dimensional & 3 dimensional.
- 2D model includes column family with some column in it, while 3D model created by associating super column in column family

→ MongoDB is a agile NoSQL document Database unlike the traditional database which store the data in rows & column, MongoDB stores the document data in binary form of JSON document which is also known as BSON format
- It is used for high scalability, availability & performance.

ii) Human Resources & Manpower

Since Big Data is at its youth & an emerging technology so it needs to attract organizations & youth with diverse new skill sets.

- These skills should not be limited to technical ones but also should extend to research, analytical, interpretive & creative ones.
- These skills need to be developed in individuals hence requires training programs to be held by the organizations
- Moreover the Universities need to introduce curriculum on Big Data to produce skilled employees in this expertise