Module 1

PAGE NO :
DATE :

Topics covered

- Introduction to big data analytics
  - Big data
  - Scalability & Parallel Processing
  - Designing data architecture
  - Data sources
  - Quality
  - Pre processing & storing
  - Data storage & Analysis
  - Applications
  - Case studies.

What is Big-Data (Data in Peta bytes). 10^{15} bytes

: very complex and too large data sets are known as Big Data - which is too complex to handle by traditional data processing softwares.

Defn: A high volume, high velocity and/or high variety information asset that requires new form of processing for enhanced decision making, insight discovery and process optimization. (Gartner 2012)

3 V's are necessary conditions to consider a data set as big data., these 3 V's are also known as characteristics of Big Data

Volume: This refers to the amount of data generated by applications in terms of size. This is a measurable unit. The size also determines how to process the data as well.

Velocity.

This refers to how fast the data

is getting generated by the applications. This also plays a huge role in processing and analysing the big data.

Variety : In traditional databases the data is well structured and can be stored in rows & columns but now data is available in different formats and forms. The format of the data is to be known to decide the type of Big Data.

One more V is also added along with 3Vs, i.e Veracity - this refers to the quality of data, it is important to know the accuracy of the data so vegacity plays a crucial role.

Types of Data : In big data it is important to know the types of data.
There are different types of data generated by the softwares
1. structured, 2. unstructured 3. semi structured
4. multistructured.

1. Structured Data : The data which is to the point, factual and conform and associate with data schemas and data models. eg: RDBMS tables.
SQL is used to query structured data.
it is highly organized and understandable for machine languages.

2. Unstructured data : This is raw data. It lacks any predefined model or format. This does not conforms to any model - or unstructured data needs more storage & providing security is hard. Managing analysing & searching is also very hard.

3. Semi structured data :

This is a structured data but does not follow tabular form. It uses other forms of markers or heirarchy within the data.

eg: XML, email

4. Multi structured data :

This refers to data consisting of multiple formats of data. structured, unstructured, semi structured. They can have many formats. They are found in non-transactional systems.

Another way of classifying big data is as follows.

1. Social networks & web-data
2. Transactions data & basiness process
3. customer master data
4. Machine generated data.
5. Human generalid data.

Big data classification :

Data sources — eg - RDBMS

Data formats — structured vs Semi structured

PAGE NO :
DATE :

Big Data sources – eg. distributed file system.

Big data formats – unstructured, semi structured & multi structured.

Data stores structure – web, cloud servers.

Processing data rates – eg. Batch, real time.

Processing big data rates – High volume, velocity

Analysis types – Batch, scheduled

Big data processing methods – Batch processing, real time processing.

Data analysis methods – statistical analysis, predictive analysis

Data usage – Human, Business process.

Big Data Handling techniques :

PAGE NO:
DATE:

1.3: Scalability & Parallel processing.

- Hundreds of terabytes to peta bytes of data
- Needs hundreds of computing nodes.
- Requirement here is processing of data in short time at minimum cost.

Scalability enables increase or decrease in the capacity of data storage, processing and analytics. It is the capacity of the system to handle the magnitude of the work.

When the workload increases the capability needs to increase. When the workload & complexity exceed the system capacity scale it up & scale it out.

Vertical Scalability: it means Scaling up the given systems resources & increasing the systems analytics, reporting and visualization capabilities.

Scaling up: Designing the algorithm according to the architecture that uses resources efficiently.

eg:

Data - x terabytes
processing time - t
Code size - n (increased with complexity).

Scaling up means processing takes equal less or much less than (n x t).

DATE :

Horrizontal Scalability :

Increasing the numbers of systems working in coherence & scaling out the workload.

Scaling out - using more resources & distributing the processing & storage tasks in parallel

r resources
x terabytes of data in
t time
then (p x x) terabytes of data process on p parallel distributed nodes.

Easy way to scale up & scale out is buying faster CPU's bigger & faster RAM etc. but this is expesive.
Alternatively we can go for

MPP is (masively parallel processing platforms) cloud, grid, cluster & distributed computing s/w.

1. Massively parallel processing platforms (MPP):

Many programs are so big and complex that they can't be run on a single CPU. The solution is either scale up the computer system or. use MPP platforms parallelization can be done at 3 levels
i) thread level
ii) multiple CPU's

iii). separate computers

Distributed computing model:
it uses cloud, grid or clusters which processes & analyze big & large data sets on distributed computing nodes. connected by high speed networks.

2. Cloud computing: Internet based computing that provides shared processing resources & data to the computers and other devices on demand.

benefits: perform parallel & distributed computing in a cloud computing environment
- multiple nodes perform automatically & interchangeably.
- offers high data security.
- Failure of one node does not cause failure of entire system.

Eg: AWS - Amazon web services. Elastic compute cloud. (EC2)
Microsoft Azure or Apache cloudstack.

features of cloud computing -
a) on demand service
b) Resource pooling
c) scalability
d) accountability.
e) Broad n/w access.

PAGE NO :
DATE :

Types of cloud computing - H/d, n/w connection, database storage, data centre etc.

① Infrastructure as service (Iaas)
- eg: Tata communications
Amazon data centers & virtual servers
Apache open stack — open source.

② Platform as service (Paas) — Provides run time environments to allows developers to build applications & services.

eg: Hadoop cloud services.

③ Software as service (Saas) — Providing software applications as a service to end users.
eg: SQL google SQL, IBM Big SQL, microsoft polybase oracle big data SQL.

③ Grid and Cluster computing:

group of computers from several locations are connected with each other to achieve a common task.

features of grid:
- scalable
- sharing of resources
- distributed n/w of resource integration.

drawbacks - single point failure leads to failure of entire grid.

PAGE NO:
DATE:

cluster computing:

- Group of computers connected by a n/w to accomplish a single task.
- It is mainly used for load balancing.
- Among the group of computers the processes are shifted between the nodes for load balancing.

eg: Hadoop architecture.

Differences

Distributed computing	cluster computing	Grid computing
① Loosely coupled	Tightly coupled	Large scale.
② Heterogeneous	Homogeneous	cross organizational.
③ Single administration	cooperative working	Geographical distribution Distributed management

④ Volunteer Computing

- Distributed computing paradigm which uses resources of volunteers — they are organizations or members who own personal computers.

issues

- heterogeneity
- Drop outs from the n/w.
- sporadic availability
- Incorrect results are unaccountable due anonymity.

PAGE NO :
DATE :

Designing Data architectural

Layer 5 Data consumption	Export of data sets to cloud web etc	Data Sets usage BPs, BIs Knowledge discovery	Analytics (real time, near-real time).
Layer 4 Data processing	Processing Technology: map reduce, Hive, Pig, Spark	Processing in real time: scheduled batches or hybrid	Synchronous or asynchronous processing
Layer 3 Data storage	Considerations of type, formats compression frequency	Hadoop distributed file system	NO SQL databases ~~Ingestion of data~~ Hbase, Mongo db
Layer 2 Data ingestion & Acquisition	Ingestion using Extract load & Transform (ELT)	Data Semantics (such as replace append ...)	Ingestion of data sources in batches or real time
Layer 1 Identification of internal & external sources of data	sources for ingestion of data	Push or Pull of data from sources for ingestion	Data types for files, web ...	Data formats structured unstructured ...

Design of logical layers in data processing architectures & functions in the layers.

DATE :

Data sources, quality, preprocessing & storing :-

- Applications
- programs & tools use data.

Sources can be external
eg: sensors, trackers, weblogs, computer systems logs & feeds.
sources can be machines, which source data from data creating programs.
- Data sources can be structured, semistructured or unstructured.
- can be social media.
- can be internal
- can be data repositories such as database, rdbms, flat file, spread sheet, mail server, web server, csv
- sources may be data store for applications.

- structured.
- unstructured.
- sensors signals & GPS.

Data Quality :

High quality data - enables all the required operations, analysis, decisions planning & knowledge discovery correctly.
A high quality data has 5 R's.
- Relevancy - this is utmost importance.
- Recency
- Range
- Robustness
- Reliability.

PAGE NO:
DATE:

Data inligrity: refers to the maintenance of consistency & accuracy in data over its usual life.

Data Pre processing:
- Step before data ingestion.
- A must before running a machine ~~learning~~ Learning algorithm.
- Prior screening is of data quality is needed.

why is it needed:
- Dropping out of range, inconsistent & out-lier values.
- Filtering unreliable, irrelevant & redundant info.
- Data cleaning, editing, reduction or wrangling.
- Data validation, transformation or toarscoding.
- ELT processing.

Data cleaning:
- Process of removing or correcting incomplete, incorrect, inaccurate or irrevelant parts of the data after detecting them.

eg: correcting the grade outlier or mistakenly entered values means cleaning & correcting the data.

data cleaning tools:
It is done before mining of data.

Incomplete or irrelevant data may result

PAGE NO :
DATE :

into misleading decisions.
- It helps in refining & structuring data into usable data.
eg: OpenRefine. & DataCleaner.

Data enrichment:

Data enrichment refers to operations or processes which refine, enhance or improve the raw data.

Data Editing :- process of reviewing & adjusting the aquired datasets.
i) Interactive
ii) selective
iii) automatic
iv) aggregating

Data reduction :-

Enables the transformation into an ordered, correct and simplified form. Only meaningful data will be ingested into the data base.

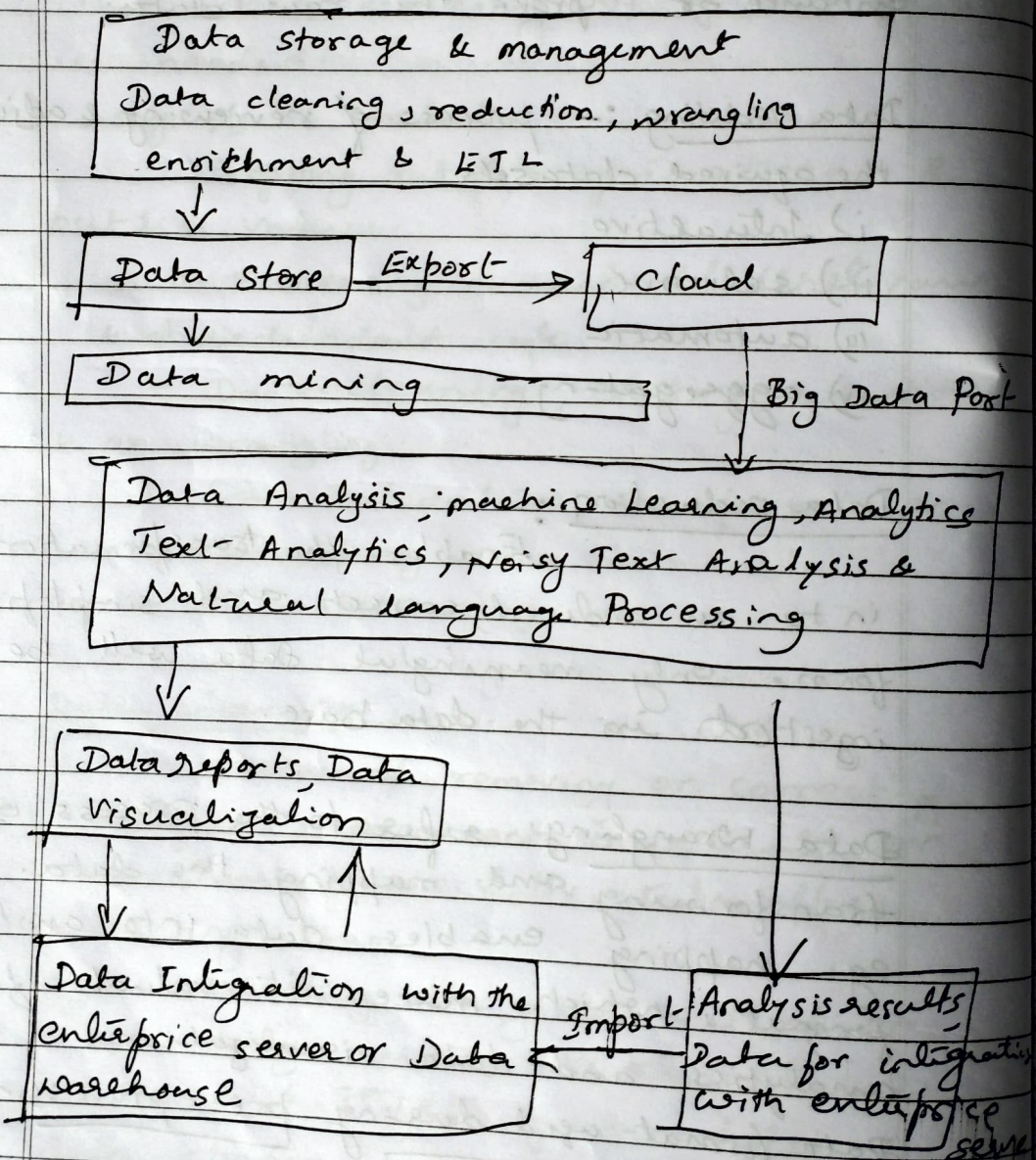
Data wrangling: refers to the process of transforming and mapping the data.
eg: mapping enables data into another format, which makes it valuable for analytics and data visualizations.
Data format used during pre processing

examples of formats for data transfer from a) Data storage b) Analytics c) service or d) cloud can be

i) Comma Separated values (CSV)
ii) JSON - Java script object notation
iii) Tag length values (TLV)

CSV format :

Data store export to cloud



Data Pre processing analysis, visualization data store export.

PAGE NO.:
DATE:

Cloud services : cloud offers various services. These services can be accessed through a cloud client, such as a web browser i.e web services.

Data Storage and Analysis :

The data store in traditional data base management systems and Big data are discussed here.

Big Data storage

Big data NoSQL or Not only SQL

- No SQL databases are considered as semi-structured data. Big data store uses NoSQL.

NoSQL stands for No SQL or Not only SQL. The store do not integrate with applications using SQL.

NoSQL is also used in cloud data store.

i) ~~class consist~~ It is a class of non relational data storage systems, and the flexible data models & multiple schema.

i) class consisting of uninterrupted key/value or big hash table.

ii) class consisting of unordered keys & using JSON.

iii) class consisting of Ordered keys & semi structured data storage systems.

iv) class consisting of JSON (MongoDB)

PAGE NO :
DATE :

v) Class consisting of name/value in the text (CouchDB)

vi) Man not use fixed table schema.

vii) Do not use JOINS.

viii) Data written at one node can replicate at multiple nodes, therefore data storage is fault tolerant.

ix) may relax ACID properties during the data store transactions.

x) Data store can be partitioned and follows CAP theorem

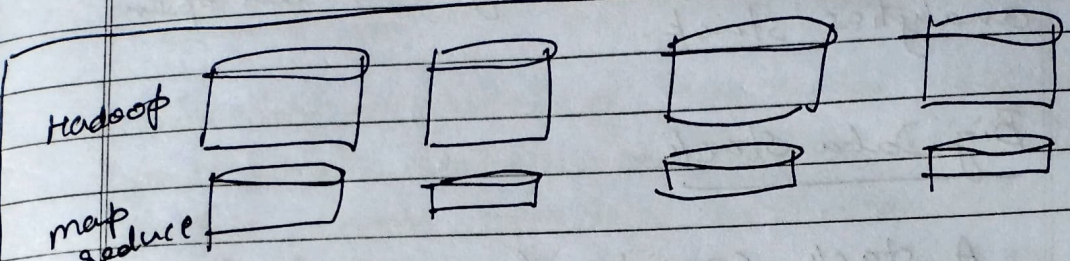
Coexistence of Big Data, NOSQL and Traditional data stores.

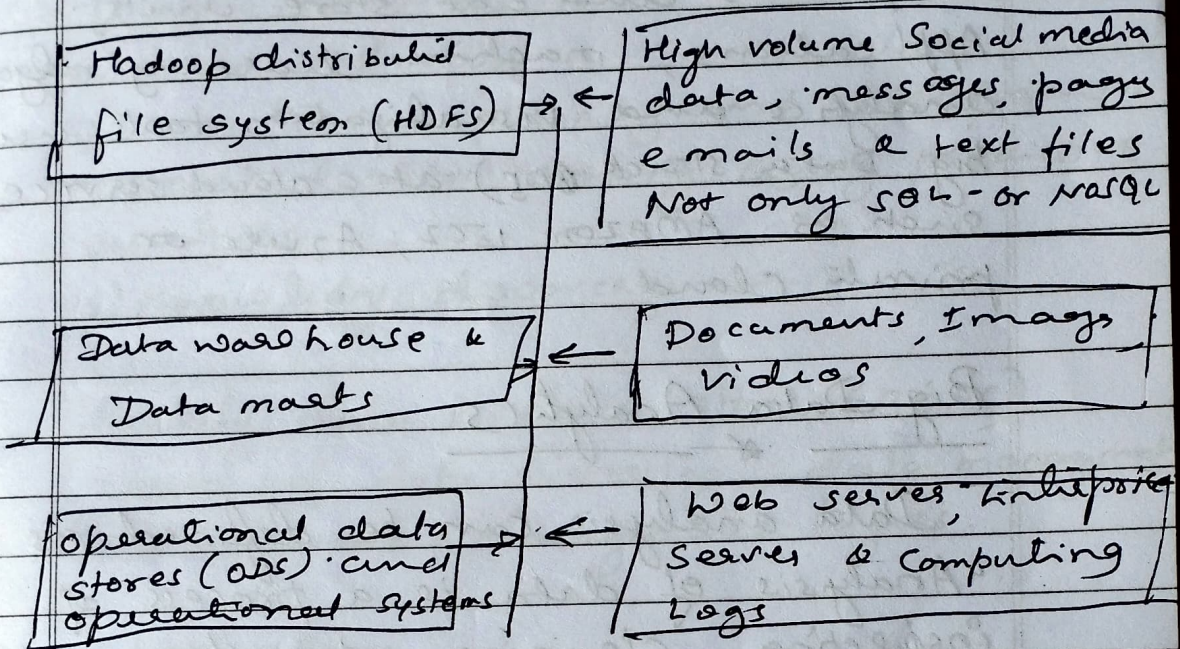
Big Data Platform

- Big data platform should provision tools and services for.

1. storage, processing and analytics.
2. Developing, deploying, operating & managing big data environment.
3. Reducing the complexity of multiple data sources & integration of applications into cohesive solution.
4. custom development, querying & integration with other systems &
5. The traditional as well as Big data techniques.

Hadoop





Hadoop based Big Data Environment.

Big data platform consists of Big data storage, servers & data management & buying intelligence s/w.

storage can deploy Hadoop distributed file system (HDFS), NOSQL data stores, such as HBase, MongoDB, cassandra. HDFS system is an open source storage system.

Mesos:

Mesos v0.9 is a resource management platform which enables sharing of cluster of nodes by multiple frameworks &

which has compatibility with an open analytics stack.

Big Data Stack.

A stack consists of a set of software components and data store units. Applications, machine learning algorithms analytics and visualization tools use big Data stack (BDS) at a cloud service such as Amazon EC2, Azure or private cloud.

Big Data Analytics

Data analysis can be defined as "Analysis of data is a process of inspecting, cleaning, transforming & modeling data with the goal of discovering useful information, suggesting conclusions and supporting decision making.

Data analytics can be defined as the statistical and mathematical data analysis that clusters, segments, ranks and predicts future possibilities.

Phases in Analytics

Descriptive analytics
Predictive analytics
Prescriptive analytics
Cognitive analytics.

Berkeley Data analytics stack (BDAS)

The importance of Big Data lies in the fact that what one does with it rather than how big or large it is.

1) cost reduction
2) time reduction.
3) new product planning and development.
4) smart decision making using predictive analysis.
5) knowledge discovery.

Berkely data analytics stack (BDAS) consists of data processing, data managment and resource management layers.

1. Applications.
2. Data Processing combines batch, steaming and interactive computations
3. Resource managment software. component provides for sharing the infracture across various frameworks.

PAGE NO :
DATE :

1.7 Big Data Analytics applications and case studies :-

- Big data in marketing & sales.
- Big data analytics in dilection of marketing frauds.
- Big data risks.
- Big data credit risk management.
- Big Data and algorithm trading.

Big data & healthcare.
Big data in medicine.
Big Data in advertising.