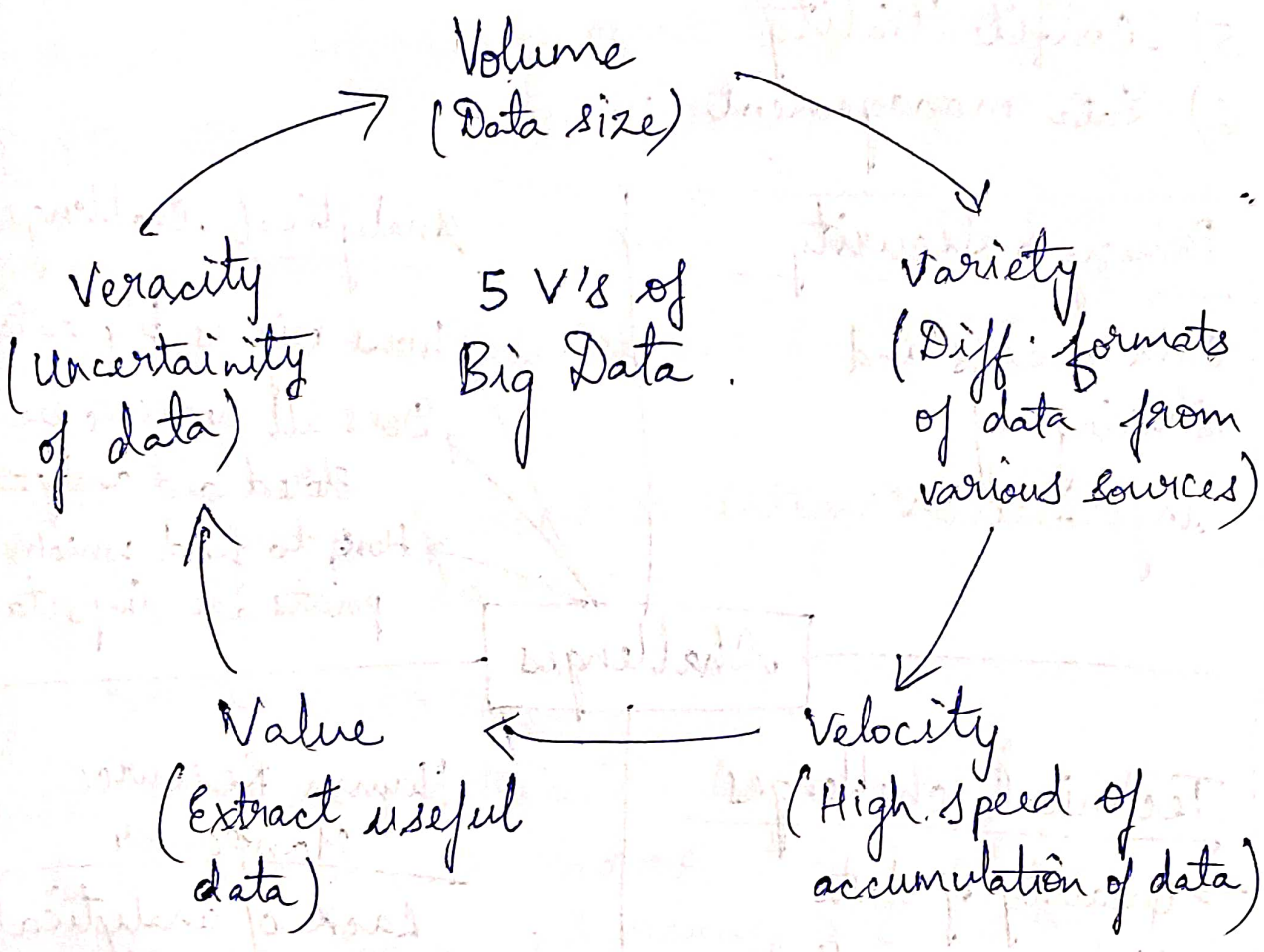
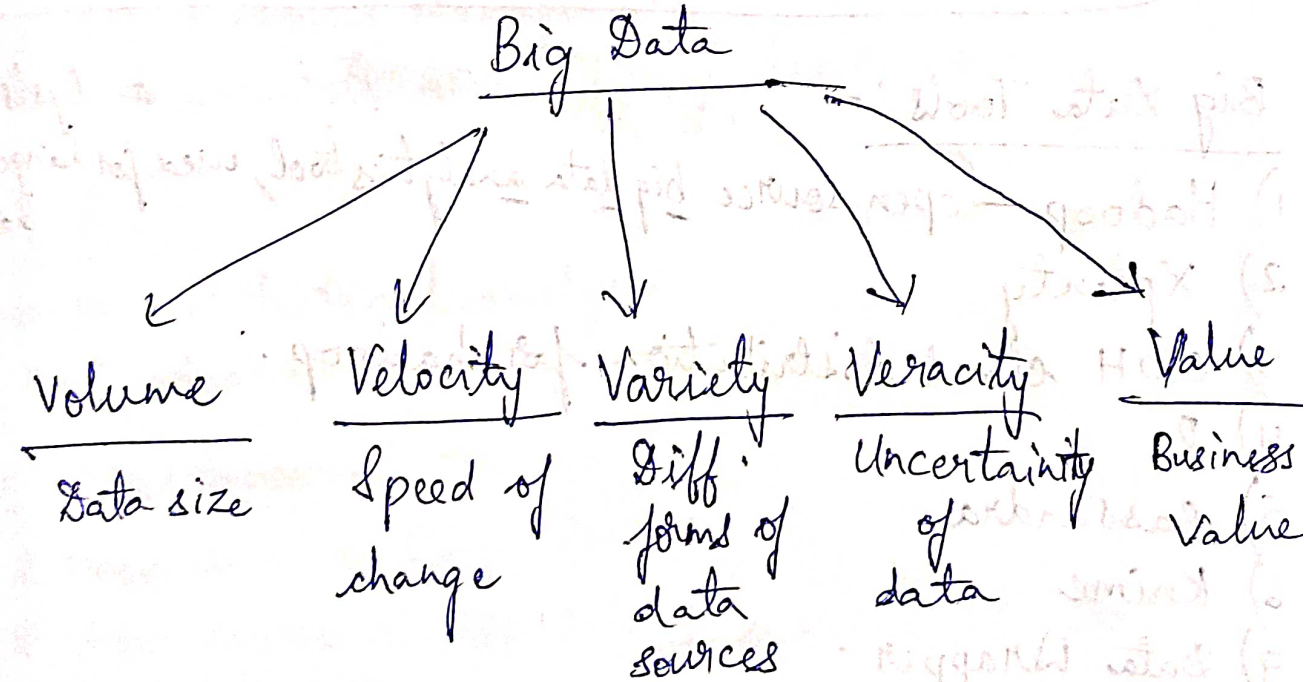
2/3/22.

Big Data Analytics.





Other Issues :-

1) Transporting issue.
2) Processing issue.

3) Data redundancy.
4) Data representation.
5) Confidentiality.
6) Data management.

Challenges

- **Privacy & Security**
  Data access and sharing of information.
- **Analytical Challenges**
  - how big is big data
  - Does all needs to be stored and analyzed
  - How to find which points are important
- **Technical challenges**
  - → Quality of data
  - → Heterogenous Data.
- **Human Resource Manpower**
  Lack of analytical Skilled Manpower.

Big Data Tools :- → common for all tools.

1) Hadoop — [open source big data analytics tool, used for large analysis of data]
2) Xplenty
3) CDH cloud distribution for hadoop.
4) R
5) Cassandra
6) Knime.
7) Data Wrapper.
8) Mongo DB.
9) Lumify
10) HPCC
11) Storm
12) Drill.

1) **Hadoop :-**
* Open-source BDA tool.
* Analysing, processing power ↑, strength
* ~~Strength~~ HDFS holds : all types of data (audio, video, XML, plain text)
* Easy to access.
* Usage : R & D (Research and Development).

2) **X plenty.**
* Open-source & used to integrate data from various sources.

3) **CDH :-**
* Cloud around Distribution Hadoop tool.
* It includes Apache Hadoop, Apache Spark.
* Acquiring, Analysing & managing the data.
* Hadoop cluster is very well managed by the cloud around Manager.
* High security and easily deployed.

4) **R :-**
* For statistical analysis.
* Open source.

5) **Cassandra :-**
* Open source BDA tool.
* Open source NoSQL DBMS uses CQL to interact with database.

6) **Knime :-**
* Konstanz information miner tool.

* Used for enterprise reporting, business intelligence analysing, management, processing.
* Supports Linux, Windows OS.

7) Data Wrapper:-
* Open source BDA tool.
* Exclusively for data visualisation (charts).

8) Mongo DB:-
* This is alternative to DB.
* To handle structured as well as unstructured data.
* Working on datasets which vary / change frequently.

9) Lumify:-
* Analyse as well as visualise data.
* 2D and 3D viewing of data, multimedia analysis, graphical viewings.
* Works very efficiently with Amazon AWS.

10) HPCC:-
* High Performance Computing Cluster.
* Also known as DAS (Data Analytics Super Computer) used to handle large data.
* Developed by LexisNexis Risk Solutions.
* Thor architecture — data parallelism, System parallelism.
* Highly scalable super computing platform.

Cloud Computing:-

Cloud is of 3 types:
1) Private cloud
2) Public cloud
3) Hybrid cloud.

Cloud Services.

- Apps
- Data
- Middleware
- O/S.
- Virtualisation
- Servers
- Storage
- N/W

IaaS: Virtualisation, Servers, Storage, N/W
Paas: Middleware, O/S, Virtualisation, Servers, Storage, N/W
SaaS: Apps, Data, Middleware, O/S, Virtualisation, Servers, Storage, N/W

Advantages / Benefits:-
1) Cost
2) Speed
3) Performance
4) Reliability
5) Security.

Big Data	Cloud
* Big data refers data size.	* Cloud refers on-demand availability of resource.
* Big data includes structured, unstructured data.	* Cloud includes IaaS, PaaS, SaaS services.
* Used to describe huge volume of data.	* Used to store data & information on remote

* sources where big data is generated — transport, black box data, social media, weather data, sensors data.

servers and process using remote servers.

* Cloud service providers are (Public) IBM cloud, Google Cloud, Amazon AWS, drop box, Microsoft Azure)
  Private — HPE, VMWare, Dell EMC.
  Hybrid — Google Cloud.

Architecture of IOT :-

sensors / device → collecting data through sensors
↓
Connectivity → Wifi/Bluetooth / Internet → transfer to cloud
↓
Data preprocessing → using BDA tools & algo.s
↓
user interface → output.

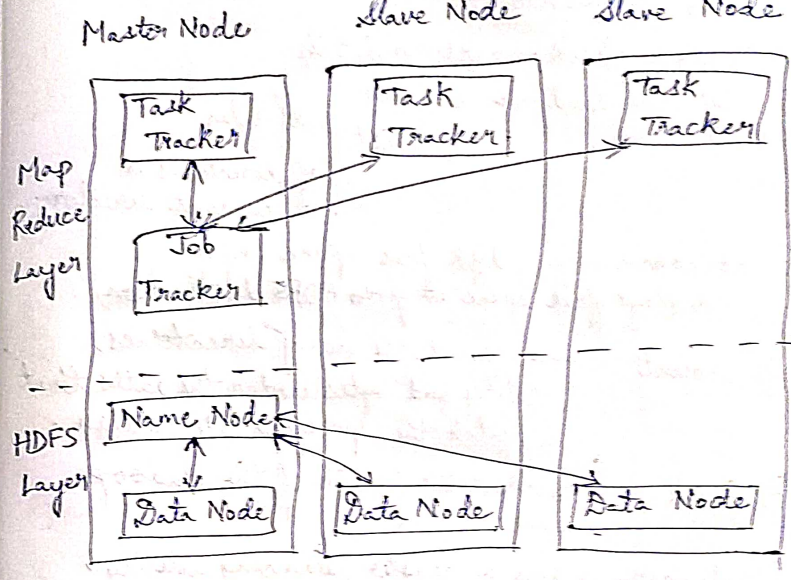
Hadoop :-

* Big data analytics uses Hadoop technology.
* Hadoop is a s/w framework written in java.
* Hadoop uses parallel processing.

4 modules of Hadoop ---

① HDFS
② YARN
③ Negotiator
④ MapReduce

Hadoop Architecture :-



Features of Hadoop :-

* Fault Tolerance solution.
* Low cost.
* Scalable.
* Parallel processing.

HDFS
Map Reduce Layer
Name node, datanode, task tracker, job tracker.
Metadata, master & slave node
Map & data splits.
3 acknowledgements.
Log report.

* Name node needs high reliable hardware.

HDFS Commands :-

1. Version check → version details
2. List command → ls / dir
   ↳ list — of files
   — of directories in current directory.
3. df command → disk free space info.
   (displays free space at given HDFS destination).
4. count → count the no. of directories, files and bytes under the paths that match the specified file pattern.
5. fsck → to check the health of the Hadoop file system.
6. balancer → run a cluster balancing utility.
7. mkdir → to create directory.
8. put → copy files from single/multiple sources from local system to destination file system.
9. du → disk usage.
   (displays size of the files & directories contained in the given directory).
10. rm → to remove file from HDFS (DOS/Linux)
11. rm -r → to remove entire directory and all its contents from HDFS.
12. chmod → change mode of operation

    r w x
    add + ↗ ↖ ← set =

    r w x r w x r w x
    Owner group Others
    u g o

    chmod g+x f1
    chmod g-x f
    chmod g=x f1

    ex: chmod 7 7 7 f1 → u+2+1
    u g o
    (7 implies r w x)
    (1 implies only x)
    ⋮

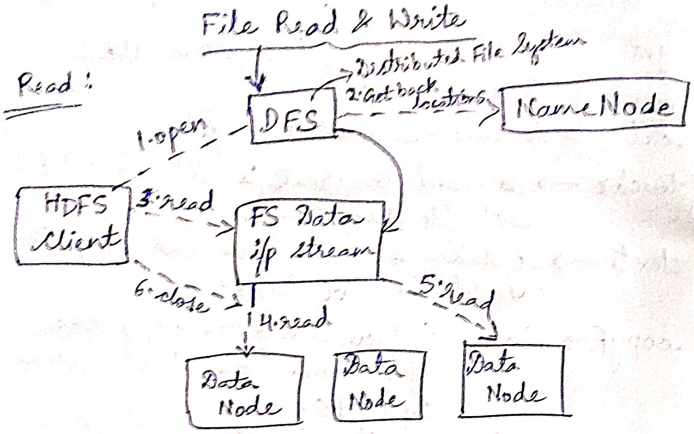
13. get → used to copy files from HDFS to
14. cat → to view the content of file.
15. touchz → is used to create a file in HDFS with file size 0 bytes.
16. text → it takes a source file and outputs the file in text format.
17. copyFromLocal → command to copy the file from local file system to HDFS.
18. copyToLocal → similar to get command, except that the destination is restricted to local file systems.
19. mv → moves file from source to destination.
    ex: mv f1 dir
    mv f1 f2 dir
20. cp → to copy files from source to destination
    ex: cp f1 f2
    source dest.
21. tail → to view last n lines of a file.
    ex: tail -2 → displays last 2 lines.
22. chown → change owner name.
    ex: chown ls cse4 f1 → prev own → new owner; f1 → is file

(22) stat, → it is to print the statistics about that file.

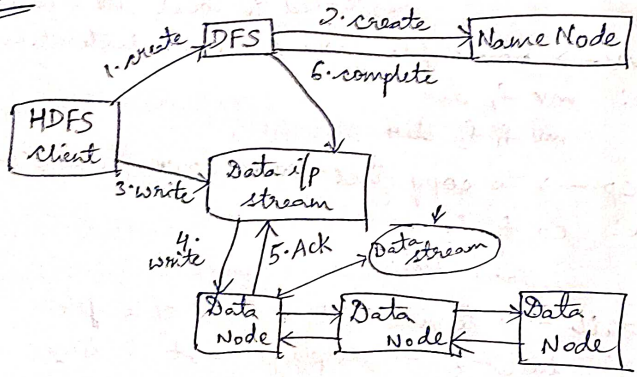
ex: file size in blocks, type, group name, owner name, block size, replication, modification date, file created date and time are the statistics.

File Read & Write

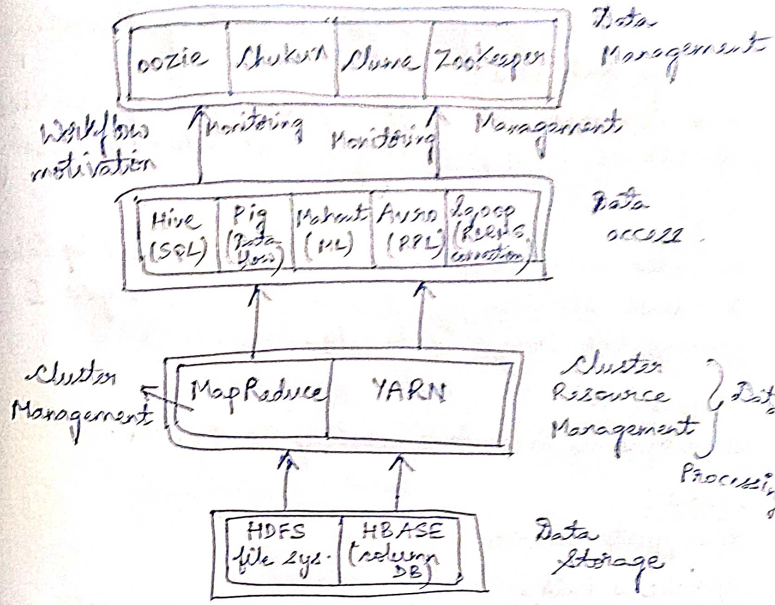
Read:



Write:



Hadoop Ecosystem:-



HBase:

It supports ~~any~~ all kinds of databases.
Helps to sort data, search and indexing is also fast

YARN:

Yet Another Resource Negotiator.
It helps to manage the resources across the cluster.
It helps in managing & scheduling the resource allocation.
Resource Manager, Application Manager and Node Manager — are its components.
Helps in allocation resources like CPU, memory...

Data access:

Hive:

Hive Query language performs reading & writing of large data sets.
Hive Command line helps in processing of queries.
It allows bash processing & real time processing

Pig:

It is also query based lang. that is developed by Yahoo
It allows structuring the data flow, analyse & process the large datasets.

Mahout (ML):

It allows machine learnability to system.

AVRO:

It supports Remote Procedure call.

Apache's Path:

It is used to handle all processing tasks like bash processing, realtime processing, iterative processing.

Zoo Keeper:

It is management of coordination, synchronization.

Oozie: (Also for chukux and Chume)

Job scheduler — reading, writing, closing

Lucene:

Searching and indexing.