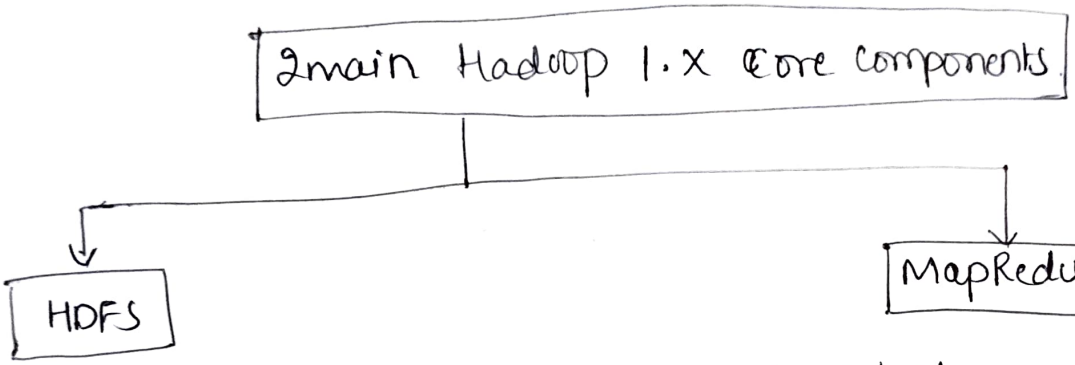
Big data Analytics

Assignment - II

Thanuja S.N
10X20MC108
B-section

1) Discuss critical Components of hadoop and there working along with a neat diagram

Ans: Two Critical Components of Hadoop.



→ Distributed across "nodes"
→ Natively redundant
→ NameNode tracks locations

→ splits a task across processors
→ "goes" the data & assemble results.
→ self-handling, high bandwidth.
→ Jobtracker manages the task trackers.
→ Clustered storage.

Hadoop Distributed file system

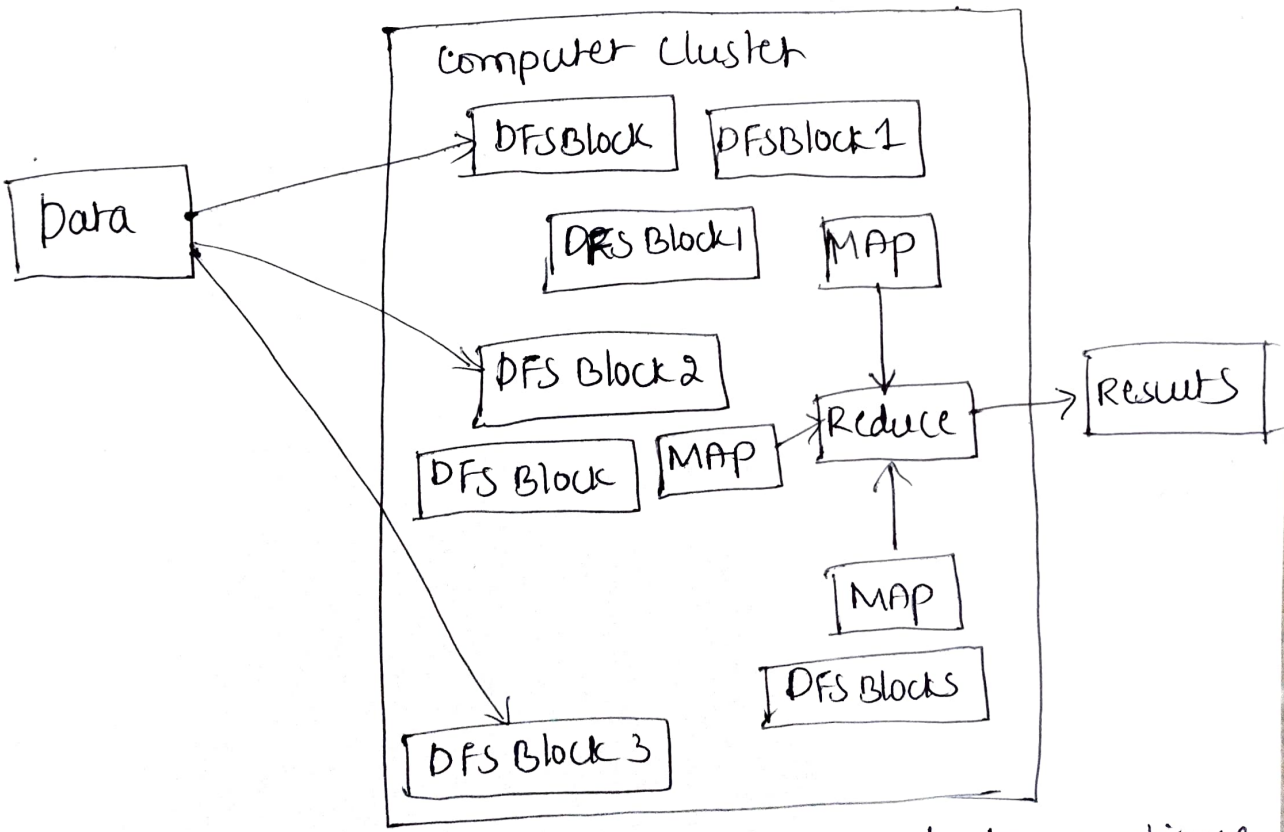
* HDFS is the storage system for a Hadoop cluster
* When data lands in the cluster, HDFS breaks it into pieces & distributes those pieces among the different servers participating in the cluster.

Map Reduce :

* Because Hadoop stores the entire dataset in small piece across a collection of servers, analytical job can be distributed in parallel to each of the servers storing part of the data.

* Each server evaluate the question against its local fragment simultaneously and reports its result back for collection into a compreh sive answer.

Working together HDFS and MR.



* Both HDFS & Map reduce are designed to continue to work in the face of system failure.

2) Explain crowd sourcing analytics and inter trans fire wall analytics

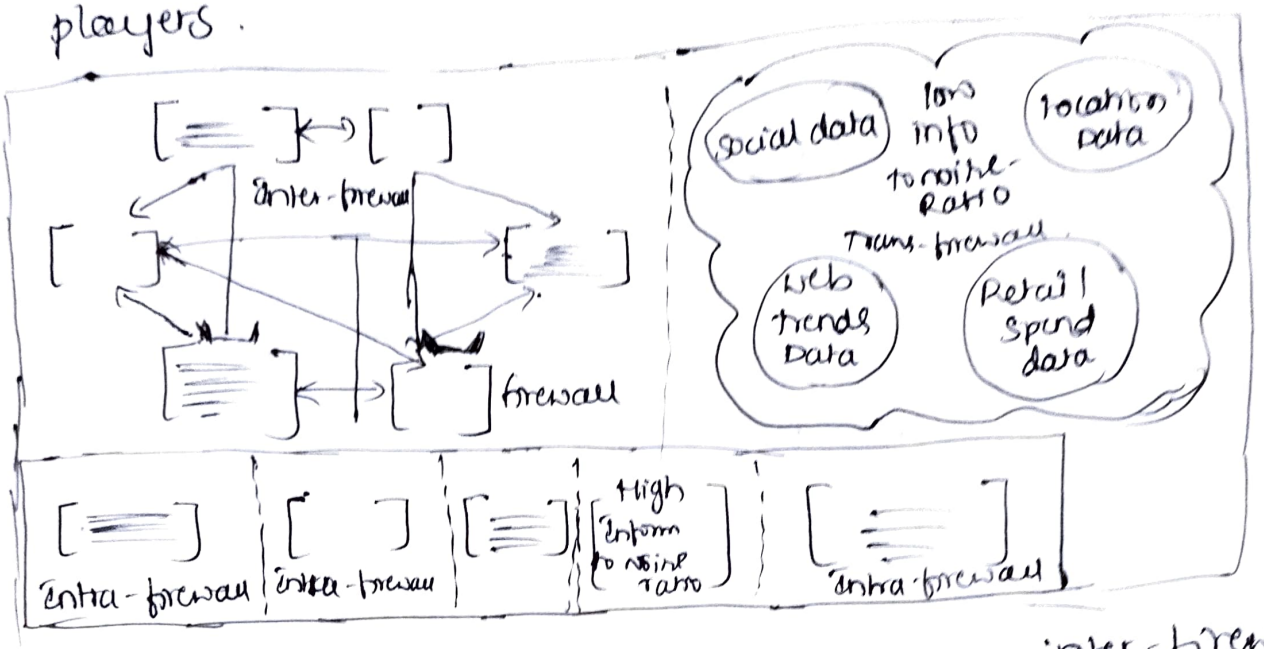
Ans: CrowSourcing analytics:

Crow Sourcing is the process of getting work (or finding, usually online from a crowd of people.

* Netflix was an innovator in a space now being termed crowd Sourcing.
* It is a recognition that you can't possibly always have the best & brighter internal people to Drive all your big problems.
* For example in kaggle, an australian company. describes it self as "an innovation solution for analytical outsourcing"
* kaggle converts problems into contests that are paired on website.
* The contest features with prize from $100 to $3millions
* The idea is that someone comes to us with a problem we put it up on our websites. & then people from all overe the world can complete to see who can provide best solution.

Inter and trans firewall analytics

* there are instances where retailer & social media company can come together to share insights on cust consumer behaviour that will benefit both players.



* In above figure depicts setting up inter-firewall & transfirewall analytics can add significant value.

* There are some challenges:

→ as one moves outside the firewall, the info-to-noise ratio increased, putting additional requirements on analytical methods & technology requirements.

→ organizations are often limited by a fear of collection & an overreliance on proprietary information

3) What is predictive analysis why all they required discuss the lead trends of predictive analysis.

Ans: Enterprises will move from business intelligence to forward learning position.

* using all the data available - tranditional internal data source combined with new rich external data source - will make the prediction more accurate and meaningful.

* Some leading trends that are making their way to the forefront of business today.

  * Recommendation engines.
  * Risk engines
  * Innovation engines
  * Customer insight engines
  * optimization engines.

→ Recommendation engines:-

Similar to those used in netflix and amazon that use past purchases and buying behaviour to recommand new purchases.

* Risk engines:

for a wide variety of business areas, including market and credit risk, catastrophic risk

* Innovation engines:

for new product innovation, drug discovery and consumer.

* Customer insight engines:

That integrate a wide variety of customer, related information including sentiment, behaviour and even emotions.

* Optimization engines

That optimize complex interrelated operations and decisions that are too overwhelming for people to systematically handle at scales.

* Using all the data available - traditional internal data sources combined with new rich external data sources will make the predictions more accurate & meaningfull.

4) List the differences b/w Mapreduce and RDBMS.

Ans:

Mapreduce	RDBMS
* Map reduce suits in an application where the data is written once and read many times.	* RDBMS is suits for an application where data size is limited like it's an GBs
* like Facebook.	
* Map Reduce suits for an appln where data size is in petabytes	* RDBMS good for data sets that are continuously update.
* Map Reduce access the data in batch mode.	* The RDBMS accessed data in interactive & batch mode.
* Map Reduce schema is dynamic	* RDBMS schema structure is static.
* MapReduce suits with unstructure data sets.	* The RDBMS suits with structure data sets
* Map Reduce is linear.	* RDBMS scaling is non-linear.

5) Write about volunteer Computing and Grid Computing.

Ans: Volunteer Computing

→ It is a type of distributed computing in which people donate their computer's unused resources to a research-oriented project.

→ A program running on a volunteer's computer periodically contacts a research application to request jobs and report results.

→ A middleware system usually serves as an intermediary

→ Since there are more than one billion PCs in the world, volunteer computing can supply more computing power to researches.

→ Supercomputers that have huge computing power are extremely expensive and are available only to some applications only if they can afford it.

* Grid Computing

→ Grid computing is the use of widely distributed computer resources to reach a common goal

→ A computing grid can be thought of as a distributed system with non-interactive workloads that involve many files.

→ Grid computing is distinguished from conventional high-performance computing systems.

→ Cluster computing in that gird computers have each nodes set to perform a different task/application.

→ Grid computers also tend to be more heterogeneous and geographically dispersed than cluster computers.

→ Grid sizes can be quite large.

6) Explain HDFS concepts with blocks, name nodes and data nodes, HDFS Federation & HDFS availability.

Ans:-

* Name nodes:

Name nodes is the centerpiece of the Hadoop Distributed file system.

→ It maintains and manages the file system name-space and provides the right access permission to the clients.

* Data nodes:

Data nodes are the slave nodes in Hadoop HDFS.

→ Data Nodes are inexpensive commodity hardware. They store blocks of a file.

→ DataNode is responsible for serving the client read / write requests.

* Blocks.

→ Block storage layer has two parts:

* Block Management: Namenode performs block management

→ Block management provides DataNodes cluster membership by handling registrations.

HDFS Federation:

→ HDFS Federation Architecture, we have horizontal scalability of name service.

→ We have multiple NameNodes with are federated i.e independent from each other

→ Each Datanode registers with all the NameNodes in the cluster.

→ The DataNodes transmit periodic heartbeats, blocks reports & handles commands from the NameNodes.

HDFS Availability:

→ Hadoop HDFS is a distributed file system.

→ HDFS distributes data among the nodes in the Hadoop cluster by creating a replica of the file.

→ Hadoop framework store these replicas of file on the other machines present in the cluster.

→ When an HDFS client wants to access this data he can easily access that data from a number of machines present in the cluster.

7) Explain with a neat diagram the sequence of events that takes place when writing a file to HDFS.

Ans: to write a file in HDFS a client needs to interact with master i.e namenode (master)

→ Now namenode provides the address of the datanodes (slaves) on which client will start writing the data

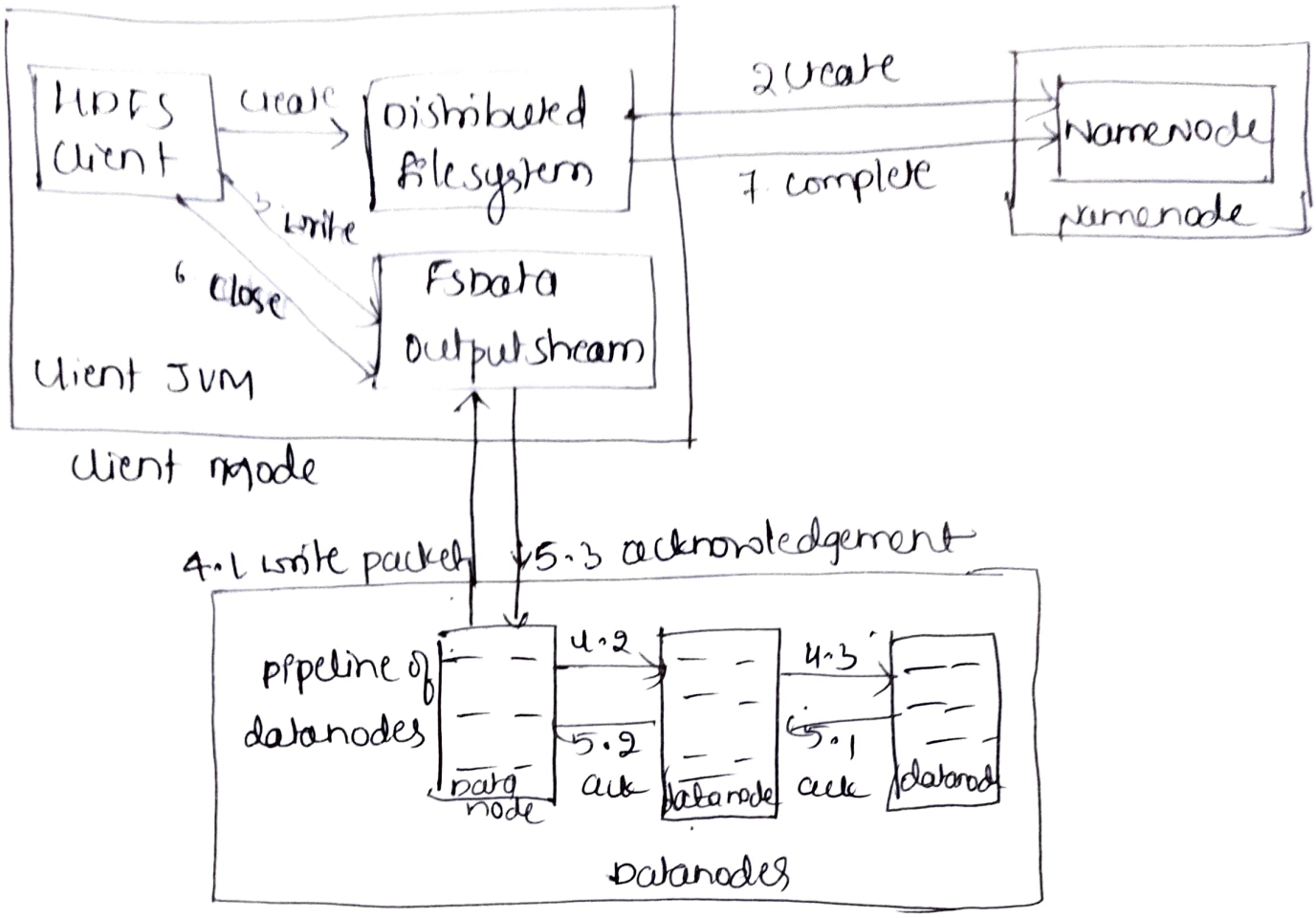
→ client directly writes data on the datanodes now datanode will create data write pipeline

→ The first datanode will copy the block to another datanode, which intern copy it to the third datanode.

→ HDFS data write pipeline works

→ understand complete end to end HDFS data write pipeline

i) The HDFS client sends a create request on distributed Filesystem APIs.



→ which I is responsible for asking the name node in the pipeline.

→ Datanode sends the acknowledgment once required replicas are created (3 by default)

→ When the client has finished writing data it calls close() on the stream.

→ This action flushes all the remaining packets to the datanode pipeline and waits for acknowledgments.