Roll No 22/25073    Big data Analytics    Course- BSc Prog C.S.

Ans1:- FSck (File System check) is the utility of HDFS which checks the overall health of HDFS.& provide a detailed Status report.

Ans 2-: Outlier.

Ans 3.: filtering.

Ans 4:- map phase in mapReduce programming tool,

Ans 5:- Reduce Phase.

Ans 6 commudity hardware

Ans 7: HDFS (Hadoop Distributed file system).

Ans 8: Edge nodes.

Ans 9: ./sbin/start-all.sh is used to start all Hadoop daemon, such as NameNode, DataNode & Resource manager and ~~/sbin~~ ./sbin/stop-all.sh commands are used for stopping all Hadoop daemons.

Ans 10:- jps

Ans 11. Speculative Execution.

Ans 12:- Data locality.

Ans 13 -: YARN (Yet Another Resource Negotiator)

Ans 14 Big data

Ans 15: Variety characteristics of Big data

Ans 16: Quintillions - A term representing extremely large quantities of data ( 1 quintillion = $10^{18}$)

Axons :: Likely refers to neuron - like connections in machine learning or neural networks, facilitating data processing & communications

Ans 17: the Hadoop Ecosystem consists of tools & frameworks that support big data storage & processing. Core components include:

HDFS :- for distributed Storage

MapReduce :- for processing data in parallel

YARN :- for Resource management

Hive :- Data quering & warehousing

Apache Pig, Apache HBase, Apache flume, Zookeeper, Apache oozie.

HDFS is the primary storage system. HDFS is a Java based file system that provide scalable. fault tolerance. reliable and cost efficient data storage of Big data. HDFS runs on commodity hardware. HDFS is already configured with default config. for many installation.

compo. of HDFS are (i) NameNode (ii) DataNode,

Map Reduce. • NameNode. known as master node.

- doesnot store actual data or dataset.
- stores metadata ie no of blocks. location, which DataNode, the data stored. and other details. consist of file & directories
- manage file system namespace
- Regulate client's access to files
- Execute file system execution such as naming, closing, opening file & Directories

(ii) DataNode

- known as slave
- perform operation like block replica. creation, deletion & replication according to instruction of NameNode
- DataNode manages data storage of system.

MapReduce: provide. data processing.

- process vast structured & unstructured data stored in HDFS.
- parallel in nature, hence very useful for large scale data analysis using multiple machines in cluster.

Two phase of map Reduce

(i) Map Phase (ii) Reduce Phase

YARN: provide resource - the resource management
YARN is also one ~~of~~ the most important component
of Hadoop ecosystem.
• called as operating system of Hadoop. responsible
• for managing & monitoring workload.

HIVE: opensource data warehouse system for
querying & analyzing large datasets stored in
Hadoop files.
does 3 main function: data summarization, query
and analysis.
uses language called HiveQL (HQL) similarto
SQL. HQL automatically translates 'SQL-like
queries into 'MapReduce' jobs.

HBase: distributed database, designed to store structured
data in tables that could have billions of row
and millions of columns. HBase is scalable,
distributed and NoSQL database that is built on
top of HDFS.
• provide access to read or write data in HDFS in
real time.

Apache Sqoop: import data from external source
into related Hadoop ecosystem component like
HDFS, HBase or Hive.
• export data from Hadoop to external sources
• works with relational database such as
teradata, Netazza, oracle, MySQL

Apache flume :- collects , aggregate & moves a large amount of data from its origin & sending it back to HDFS.

- It is fault tolerant & reliable mechanism.
- It allows dataflow from source into Hadoop ecosytem