Sample questions (SET 1)

1) What is Hadoop Map Reduce ?

For processing large data sets in parallel across a hadoop cluster, Hadoop MapReduce framework is used. Data analysis uses a two-step map and reduce process.

2) How Hadoop MapReduce works?

In MapReduce, during the map phase it counts the words in each document, while in the reduce phase it aggregates the data as per the document spanning the entire collection. During the map phase the input data is divided into splits for analysis by map tasks running in parallel across Hadoop framework.

3) Explain what is shuffling in MapReduce ?

The process by which the system performs the sort and transfers the map outputs to the reducer as inputs is known as the shuffle

4) Explain what is distributed Cache in MapReduce Framework ?

Distributed Cache is an important feature provided by map reduce framework. When you want to share some files across all nodes in Hadoop Cluster, DistributedCache is used. The files could be an executable jar files or simple properties file.

5) Explain what is NameNode in Hadoop?

NameNode in Hadoop is the node, where Hadoop stores all the file location information in HDFS (Hadoop Distributed File System). In other words, NameNode is the centrepiece of an HDFS file system. It keeps the record of all the files in the file system, and tracks the file data across the cluster or multiple machines

6) Explain what is JobTracker in Hadoop? What are the actions followed by Hadoop?

In Hadoop for submitting and tracking MapReduce jobs, JobTracker is used. Job tracker run on its own JVM process

Hadoop performs following actions in Hadoop

- Client application submit jobs to the job tracker
- JobTracker communicates to the Namemode to determine data location
- Near the data or with available slots JobTracker locates TaskTracker nodes
- On chosen TaskTracker Nodes, it submits the work
- When a task fails, Job tracker notify and decides what to do then.
- The TaskTracker nodes are monitored by JobTracker

7) Explain what is heartbeat in HDFS?

Heartbeat is referred to a signal used between a data node and Name node, and between task tracker and job tracker, if the Name node or job tracker does not respond to the signal, then it is considered there is some issues with data node or task tracker

8) Explain what combiners is and when you should use a combiner in a MapReduce Job?

To increase the efficiency of MapReduce Program, Combiners are used. The amount of data can be reduced with the help of combiner's that need to be transferred across to the reducers. If the operation performed is commutative and associative you can use your reducer code as a combiner. The execution of combiner is not guaranteed in Hadoop

9) What happens when a datanode fails ?

When a datanode fails

- Jobtracker and namenode detect the failure
- On the failed node all tasks are re-scheduled
- Namenode replicates the users data to another node

10) Explain what is Speculative Execution?

In Hadoop during Speculative Execution a certain number of duplicate tasks are launched. On different slave node, multiple copies of same map or reduce task can be executed using Speculative Execution. In simple words, if a particular drive is taking long time to complete a task, Hadoop will create a duplicate task on another disk. Disk that finish the task first are retained and disks that do not finish first are killed.

11) Explain what are the basic parameters of a Mapper?

The basic parameters of a Mapper are

- LongWritable and Text
- Text and IntWritable

12) Explain what is the function of MapReducer partitioner?

The function of MapReducer partitioner is to make sure that all the value of a single key goes to the same reducer, eventually which helps evenly distribution of the map output over the reducers

13) Explain what is difference between an Input Split and HDFS Block?

Logical division of data is known as Split while physical division of data is known as HDFS Block

14) Explain what happens in textinformat ?

In textinputformat, each line in the text file is a record. Value is the content of the line while Key is the byte offset of the line. For instance, Key: longWritable, Value: text

15) Mention what are the main configuration parameters that user need to specify to run Mapreduce Job ?

The user of Mapreduce framework needs to specify

- Job’s input locations in the distributed file system
- Job’s output location in the distributed file system

- Input format
- Output format
- Class containing the map function
- Class containing the reduce function
- JAR file containing the mapper, reducer and driver classes

16) Explain what is sqoop in Hadoop ?

- To transfer the data between Relational database management (RDBMS) and Hadoop HDFS a tool is used known as Sqoop. Using Sqoop data can be transferred from RDMS like MySQL or Oracle into HDFS as well as exporting data from HDFS file to RDBMS

17) Explain how JobTracker schedules a task ?

- The task tracker send out heartbeat messages to Jobtracker usually every few minutes to make sure that JobTracker is active and functioning. The message also informs JobTracker about the number of available slots, so the JobTracker can stay upto date with where in the cluster work can be delegated

18) Explain what is Sequencefileinputformat?

- Sequencefileinputformat is used for reading files in sequence. It is a specific compressed binary file format which is optimized for passing data between the output of one MapReduce job to the input of some other MapReduce job.

19) Explain what is Hadoop?

It is an open-source software framework for storing data and running applications on clusters of commodity hardware. It provides enormous processing power and massive storage for any type of data.

20) Mention what is the difference between an RDBMS and Hadoop?

RDBMS	Hadoop
RDBMS is relational database management system	Hadoop is node based flat structure
It used for OLTP processing whereas Hadoop	It is currently used for analytical and for BIG DATA processing
In RDBMS, the database cluster uses the same data files stored in shared storage	In Hadoop, the storage data can be stored independently in each processing node.
You need to preprocess data before storing it	you don’t need to preprocess data before storing it

21) Mention Hadoop core components?

Hadoop core components include,

- HDFS
- MapReduce

22) What is NameNode in Hadoop?

NameNode in Hadoop is where Hadoop stores all the file location information in HDFS. It is the master node on which job tracker runs and consists of metadata.

23) Mention what are the data components used by Hadoop?

Data components used by Hadoop are

- Pig

- Hive

24) Mention what is the data storage component used by Hadoop?

The data storage component used by Hadoop is HBase.

25) Mention what are the most common input formats defined in Hadoop?

The most common input formats defined in Hadoop are;

- TextInputFormat
- KeyValueInputFormat
- SequenceFileInputFormat

26) In Hadoop what is InputSplit?

It splits input files into chunks and assign each split to a mapper for processing.

27) For a job in Hadoop, is it possible to change the number of mappers to be created?

No, it is not possible to change the number of mappers to be created. The number of mappers is determined by the number of input splits.

28) Explain what is a sequence file in Hadoop?

To store binary key/value pairs, sequence file is used. Unlike regular compressed file, sequence file support splitting even when the data inside the file is compressed.

29) When Namenode is down what happens to job tracker?

Namenode is the single point of failure in HDFS so when Namenode is down your cluster will set off.

30) Explain how indexing in HDFS is done?

Hadoop has a unique way of indexing. Once the data is stored as per the block size, the HDFS will keep on storing the last part of the data which say where the next part of the data will be.

31) Explain is it possible to search for files using wildcards?

Yes, it is possible to search for files using wildcards.

32) List out Hadoop’s three configuration files?

The three configuration files are

- core-site.xml
- mapred-site.xml
- hdfs-site.xml

33) Explain what is “map” and what is "reducer" in Hadoop?

- In Hadoop, a map is a phase in HDFS query solving. A map reads data from an input location, and outputs a key value pair according to the input type.
- In Hadoop, a reducer collects the output generated by the mapper, processes it, and creates a final output of its own.

34) In Hadoop, which file controls reporting in Hadoop?

- In Hadoop, the hadoop-metrics.properties file controls reporting.
- **Explain what is a Task Tracker in Hadoop?**
- A Task Tracker in Hadoop is a slave node daemon in the cluster that accepts tasks from a JobTracker. It also sends out the heartbeat messages to the JobTracker, every few minutes, to confirm that the JobTracker is still alive.
- **Mention what is distributed cache in Hadoop?**
- Distributed cache in Hadoop is a facility provided by MapReduce framework. At the time of execution of the job, it is used to cache

file. The Framework copies the necessary files to the slave node before the execution of any task at that node.

35) How big data analysis helps businesses increase their revenue? Give example.

Big data analysis is helping businesses differentiate themselves – for example Walmart the world's largest retailer in 2014 in terms of revenue - is using big data analytics to increase its sales through better predictive analytics, providing customized recommendations and launching new products based on customer preferences and needs. Walmart observed a significant 10% to 15% increase in online sales for $1 billion in incremental revenue. There are many more companies like Facebook, Twitter, LinkedIn, Pandora, JPMorgan Chase, Bank of America, etc. using big data analytics to boost their revenue.

36) Differentiate between Structured and Unstructured data.

Data which can be stored in traditional database systems in the form of rows and columns, for example the online purchase transactions can be referred to as Structured Data. Data which can be stored only partially in traditional database systems, for example, data in XML records can be referred to as semi structured data. Unorganized and raw data that cannot be categorized as semi structured or structured data is referred to as unstructured data. Facebook updates, Tweets on Twitter, Reviews, web logs, etc. are all examples of unstructured data.

37) On what concept the Hadoop framework works?

Hadoop Framework works on the following two core components-

1) HDFS – Hadoop Distributed File System is the java based file system for scalable and reliable storage of large datasets. Data in HDFS is stored in the form of blocks and it operates on the Master Slave Architecture.
2) Hadoop MapReduce-This is a java based programming paradigm of Hadoop framework that provides scalability across various Hadoop clusters. MapReduce distributes the workload into various tasks that can

run in parallel. Hadoop jobs perform 2 separate tasks- job. The map job breaks down the data sets into key-value pairs or tuples. The reduce job then takes the output of the map job and combines the data tuples to into smaller set of tuples. The reduce job is always performed after the map job is executed.

38) What are the main components of a Hadoop Application?

Hadoop applications have wide range of technologies that provide great advantage in solving complex business problems.

Core components of a Hadoop application are-

1) Hadoop Common

2) HDFS

3) Hadoop MapReduce

4) YARN

Data Access Components are - Pig and Hive

Data Storage Component is - HBase

Data Integration Components are - Apache Flume, Sqoop, Chukwa

Data Management and Monitoring Components are - Ambari, Oozie and Zookeeper.

Data Serialization Components are - Thrift and Avro

Data Intelligence Components are - Apache Mahout and Drill.

39) What is Hadoop streaming?

Hadoop distribution has a generic application programming interface for writing Map and Reduce jobs in any desired programming language like Python, Perl, Ruby, etc. This is referred to as Hadoop Streaming. Users can create and run jobs with any kind of shell scripts or executable as the Mapper or Reducers.

40) What is the best hardware configuration to run Hadoop?

The best configuration for executing Hadoop jobs is dual core machines or dual processors with 4GB or 8GB RAM that use ECC memory. Hadoop highly benefits from using ECC memory though it is not low - end. ECC memory is recommended for running Hadoop because most of the Hadoop users have experienced various checksum errors by using non ECC memory. However, the hardware configuration also depends on the workflow requirements and can change accordingly.

41) What are the most commonly defined input formats in Hadoop?

The most common Input Formats defined in Hadoop are:

- Text Input Format- This is the default input format defined in Hadoop.
- Key Value Input Format- This input format is used for plain text files wherein the files are broken down into lines.
- Sequence File Input Format- This input format is used for reading files in sequence.

42) What is BigData?

Big data is defined as the voluminous amount of structured, unstructured or semi-structured data that has huge potential for mining but is so large that it cannot be processed using traditional database systems. Big data is characterized by its high velocity, volume and variety that requires cost effective and innovative methods for information processing to draw meaningful business insights. More than the volume of the data – it is the nature of the data that defines whether it is considered as Big Data or not.

43) Explain the difference between NameNode, Backup Node and Checkpoint NameNode.

NameNode: NameNode is at the heart of the HDFS file system which manages the metadata i.e. the data of the files is not stored on the

NameNode but rather it has the directory tree of all the files present in the HDFS file system on a hadoop cluster. NameNode uses two files for the namespace-
fsimage file- It keeps track of the latest checkpoint of the namespace.

edits file-It is a log of changes that have been made to the namespace since checkpoint.

Checkpoint Node-
Checkpoint Node keeps track of the latest checkpoint in a directory that has same structure as that of NameNode's directory. Checkpoint node creates checkpoints for the namespace at regular intervals by downloading the edits and fsimage file from the NameNode and merging it locally. The new image is then again updated back to the active NameNode.

BackupNode:
Backup Node also provides check pointing functionality like that of the checkpoint node but it also maintains its up-to-date in-memory copy of the file system namespace that is in sync with the active NameNode.

44) What is commodity hardware?

Commodity Hardware refers to inexpensive systems that do not have high availability or high quality. Commodity Hardware consists of RAM because there are specific services that need to be executed on RAM. Hadoop can be run on any commodity hardware and does not require any super computer s or high end hardware configuration to execute jobs.

45) What is a rack awareness and on what basis is data stored in a rack?

All the data nodes put together form a storage area i.e. the physical location of the data nodes is referred to as Rack in HDFS. The rack information i.e. the rack id of each data node is acquired by the NameNode. The process of selecting closer data nodes depending on the rack information is known as Rack Awareness.

The contents present in the file are divided into data block as soon as the client is ready to load the file into the hadoop cluster. After consulting with the NameNode, client allocates 3 data nodes for each data block. For each data block, there exists 2 copies in one rack and the third copy is present in another rack.

46) What happens to a NameNode that has no data?

There does not exist any NameNode without data. If it is a NameNode then it should have some sort of data in it.

Whenever a client submits a hadoop job, who receives it?

NameNode receives the Hadoop job which then looks for the data requested by the client and provides the block information. JobTracker takes care of resource allocation of the hadoop job to ensure timely completion.

Explain about the partitioning, shuffle and sort phase

Shuffle Phase-Once the first map tasks are completed, the nodes continue to perform several other map tasks and also exchange the intermediate outputs with the reducers as required. This process of moving the intermediate outputs of map tasks to the reducer is referred to as Shuffling.

Sort Phase- Hadoop MapReduce automatically sorts the set of intermediate keys on a single node before they are given as input to the reducer.

Partitioning Phase-The process that determines which intermediate keys and value will be received by each reducer instance is referred to as partitioning. The destination partition is same for any key irrespective of the mapper instance that generated it.

47) When should you use HBase and what are the key components of HBase?

HBase should be used when the big data application has –

1)A variable schema

2)When data is stored in the form of collections

3)If the application demands key based access to data while retrieving.

Key components of HBase are –

Region- This component contains memory data store and Hfile.

Region Server-This monitors the Region.

HBase Master-It is responsible for monitoring the region server.

Zookeeper- It takes care of the coordination between the HBase Master component and the client.

Catalog Tables-The two important catalog tables are ROOT and META.ROOT table tracks where the META table is and META table stores all the regions in the system.

48) What is Row Key?

Every row in an HBase table has a unique identifier known as RowKey. It is used for grouping cells logically and it ensures that all cells that have the same RowKeys are co-located on the same server. RowKey is internally regarded as a byte array.

49) Explain the difference between RDBMS data model and HBase data model.

RDBMS is a schema based database whereas HBase is schema less data model.

RDBMS does not have support for in-built partitioning whereas in HBase there is automated partitioning.

RDBMS stores normalized data whereas HBase stores de-normalized data.

50) Explain the difference between HBase and Hive.

HBase and Hive both are completely different hadoop based technologies-Hive is a data warehouse infrastructure on top of Hadoop whereas HBase is a NoSQL key value store that runs on top of Hadoop. Hive helps SQL savvy people to run MapReduce jobs whereas HBase supports 4 primary operations-put, get, scan and delete. HBase is ideal for real time querying of big data where Hive is an ideal choice for analytical querying of data collected over period of time.

51) What are the additional benefits YARN brings in to Hadoop?

- Effective utilization of the resources as multiple applications can be run in YARN all sharing a common resource.In Hadoop MapReduce there are seperate slots for Map and Reduce tasks whereas in YARN there is no fixed slot. The same container can be used for Map and Reduce tasks leading to better utilization.

- YARN is backward compatible so all the existing MapReduce jobs.

- Using YARN, one can even run applications that are not based on the MaReduce model

52) Compare RDBMS with Hadoop MapReduce.

53) What is MapReduce?

54) Illustrate a simple example of the working of MapReduce.