1. What is Big Data?

There are many definition of Big Data….Bigdata is term that describes the large amount of data…

2. How to classify Bigdata?

IBM classified Bigdata on four major points (4v’s)

i) **Volume**—Volume of data should be huge like tera byte, peta byte (not small data). That means single machince is incapable to handles that’s. Facebook generating 500+ terabytes data per day.

ii) **Velocity**—Speed of new data uploading…(either low or high both can be tackled in any situation). Like 900 million of photo uploading in fb and 3.5 billion search in google…

iii) **Variety**—Different forms of data (structured, semi-structured or unstructured).

iv) **Veracity**—Nature of data/abonormality in data. That means poor quality and uncleaned data. Like in a table lot of null data.

If any data characterized this four which is called Bigdata.

3. Why we need Bigdata?

Just one purpose: We will be able to process huge amount of data that we can not do in our traditional system.

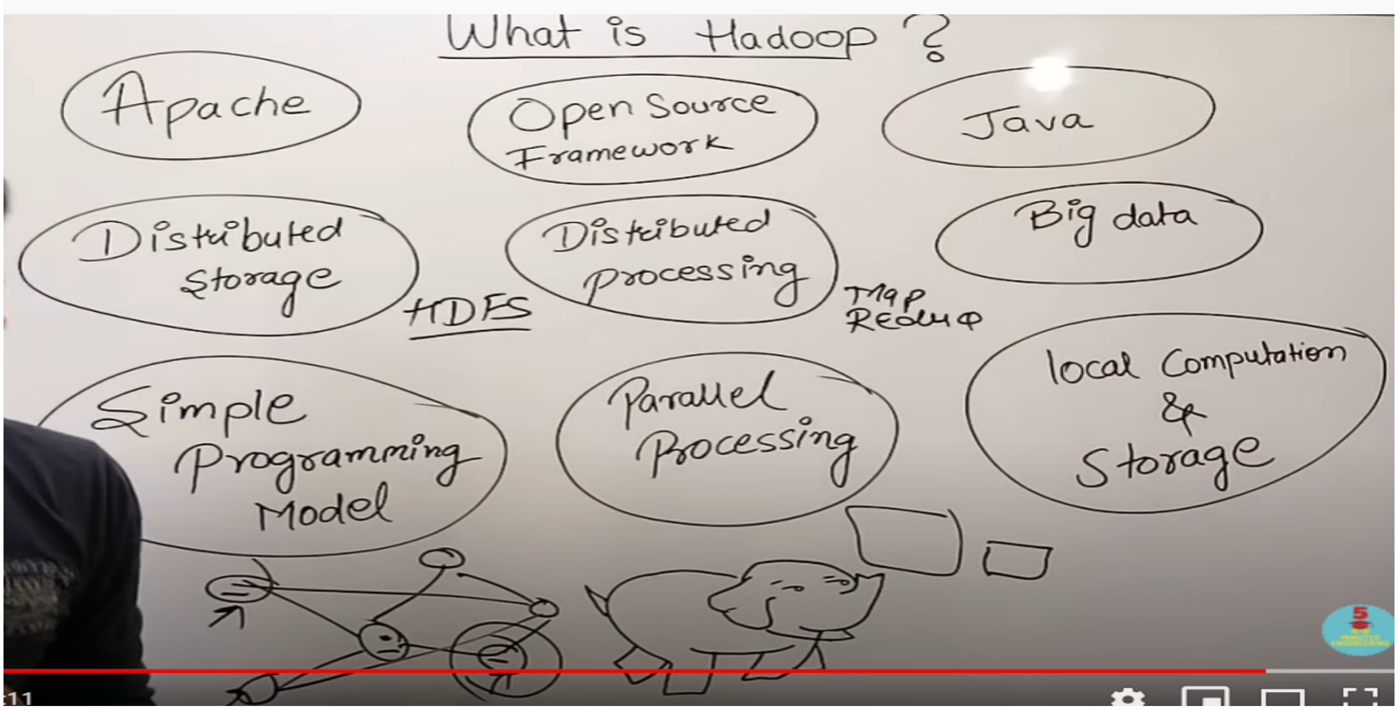
4. Differentiate between Structured and Unstructured data.

Data that 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 that 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, weblogs, etc. are all examples of unstructured data.

i) **Structured data:** Schema-based data, Datastore in SQL, Postgresql databases etc
ii) **Semi-structured data :** Json objects , json arrays, csv , txt ,xlsx files,web logs ,tweets etc
iii) **Unstructured data :** Audio, Video files, etc

5. What is Hadoop?

Hadoop is framework to solve bigdata problem.



6. Why do we need Hadoop?

Everyday a large amount of unstructured data is getting dumped into our machines. The major challenge is not to store large data sets in our systems but to retrieve and analyze the big data in the organizations, that too data present in different machines at different locations. In this situation a necessity for Hadoop arises. Hadoop has the ability to analyze the data present in different machines at different locations very quickly and in a very cost effective way. It uses the concept of MapReduce which enables it to divide the query into small parts and process them in parallel. This is also known as parallel computing.

7. Which are the three modes in which Hadoop can be run?

i) **Standalone (local) mode.**
ii) **Pseudo-distributed mode.**
iii) **File distributed mode.**

8. What are the most commonly defined input formats in Hadoop?

i) **Text Input Format:** Default input format defined in Hadoop.
ii) **Key Value Input Format:** Used for plain text file wherein the files are broken down into lines.
iii) **Sequensce File Input Format :** Used for reading files in sequence.

9. What are the steps involved in deploying a big data solution?

i) Data Ingestion – The foremost step in deploying big data solutions is to extract data from different sources which could be an Enterprise Resource Planning System like SAP, any CRM like Salesforce or Siebel , RDBMS like MySQL or Oracle, or could be the log files, flat files, documents, images, social media feeds. This data needs to be stored in HDFS. Data can either be ingested through batch jobs that run every 15 minutes, once every night and so on or through streaming in real-time from 100 ms to 120 seconds.

ii) Data Storage – The subsequent step after ingesting data is to store it either in HDFS or NoSQL database like HBase. HBase storage works well for random read/write access whereas HDFS is optimized for sequential access.

iii) Data Processing – The ultimate step is to process the data using one of the processing frameworks like mapreduce, spark, pig, hive, etc.

10. How Bigdata analysis helps businesses increase their revenue?

In every business, there target is expand the business by earning more revenue. For earning more revenue, management need to take proper decision time to time. Management will be able to do this when they have actual market/product information or data. By analysing this data, management will be able to understand the customer demand, whether they need to launching new product or need to increase the stock of existing product etc. There are many more companies like Facebook, Twitter, LinkedIn, Chase, Bank of America using big data analytics to boost their revenue.

11. On what concept Hadoop framework works?

1) **HDFS** (Hadoop Distributed File System, for distributed storage) We keep all information here. **Data in HDFS is stored in the form of block and it operates on Master-Slave Architecture.**
2) **MapReduce** (Distributed processing unit)
3) **YARN** (Mainly responsible for resource management)

Hive: Datawarehouse tool for providing data query and analysis.
PIG: A scripting language for data manipulation. Transfer unstructured data to structure format.
Sqoop: A command line interface application for transferring data between relational databases and Hadoop.

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.

❖ **Point for remember:**

Block: A block is a minimum amount of data that can be read and write. For Hadoop 2 version: Industrially recommended that every block size should be 128 Mb but it is flexible. we can up & down.

Block Scanner: Tracks the list of blocks present on a datanode and verifies them to find any kind of checksum errors.

Masterdata is called as a Namenode and Slavenode is called as a Datanode.

______The Namenode hold the information which is metadata (data about the data) in the form of table. It maintains and manages the blocks which present in the datanode i.e. controls the datanode. It is a high-availability machine and single point of failure in HDFS.

The datanode hold the actual data in the form of blocks. Also responsible for serving write and read request for the client.

Client node means you seat on a laptop and you made a request for reading file1 but know where is the file1. You always request to namenode because it cap all the information which one you want to read. Then it check the metadata which one you want to read. If he get the information of file1 then it will give you a information of file1(metadata). Then you read your file from datanode. Here namenode don't read the file.

It works internally.

Datanodes are made commodity hardware where namenode are made by high quality hardware only for less cost cluster. That's why datanode might fail frequently and namenode less chance to fail. So when a datanode be failed then a part of file would be miss. It is big problem. So, how can we recover this?

The solution is replication factor.

12. How could NameNode know whether a DataNode is failure?

This is the concept of heartbeat mechanism. Each datanode sends heartbeat to namenode in every 3 seconds in default. If a namenode doesn't get 10 consecutive heartbeats then it assumes the datanode is dead. The name node automatically delete the data node and create another copy of this information. Because they 3 copies by default. This critical condition is called fault tolerance.

If a datanode goes down, the replication factor also came down under 3. When a datanode failed down, internally it will create another replica.

13. What will be happen if Namenode is failure?

In Haddop version-1 when name node failure then whole system shutdown but HV-2 the whole system don't shutdown when failure of name node. What happen when name node will fail .that means we don't access metadata.without metadata we can not find block from datanode which one we need . Name node like a index pages of a book that’s mean first page. Say in a big books without index it is very difficult find a particular topics. Name like a heart component of HDFS. All communication will stop. SO Name node fails mean no access to metadata & NO matadata means no access to cluster. It is very painful situation that’s why name node is very high quality Hardware.

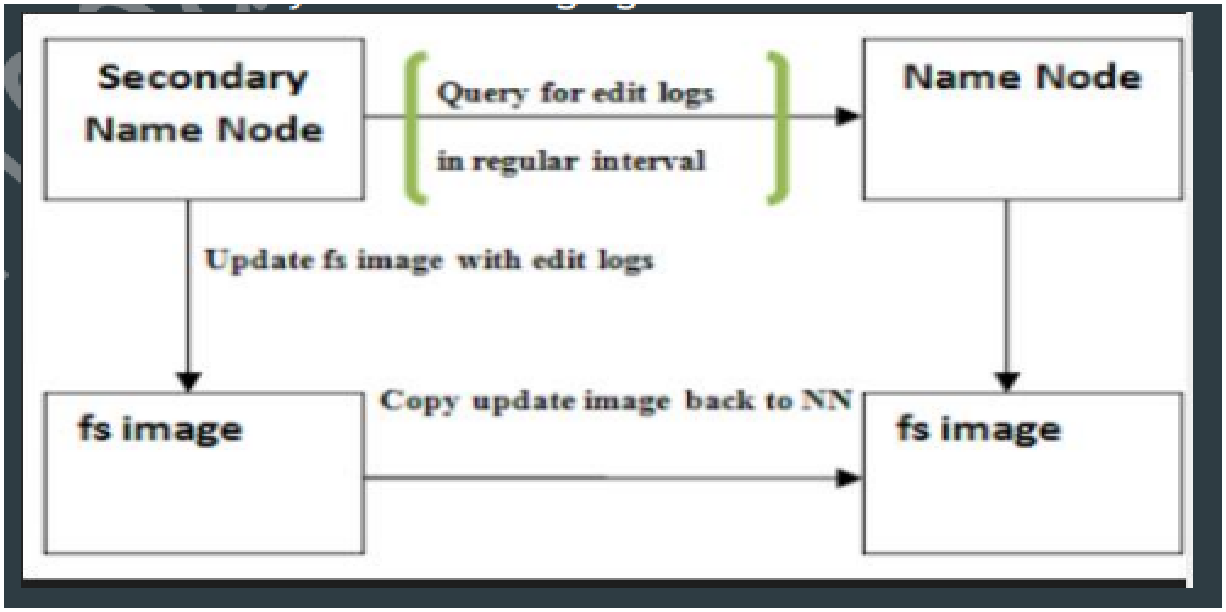
All communication will stop due to failure of the namenode.

Problem is that we will be loosing the mapping information, Now what the solution would be if we have the latest block mapping information(metadata) then we can make sure there is no downtime involve.

• There two extra things that help to restore the current metadata.

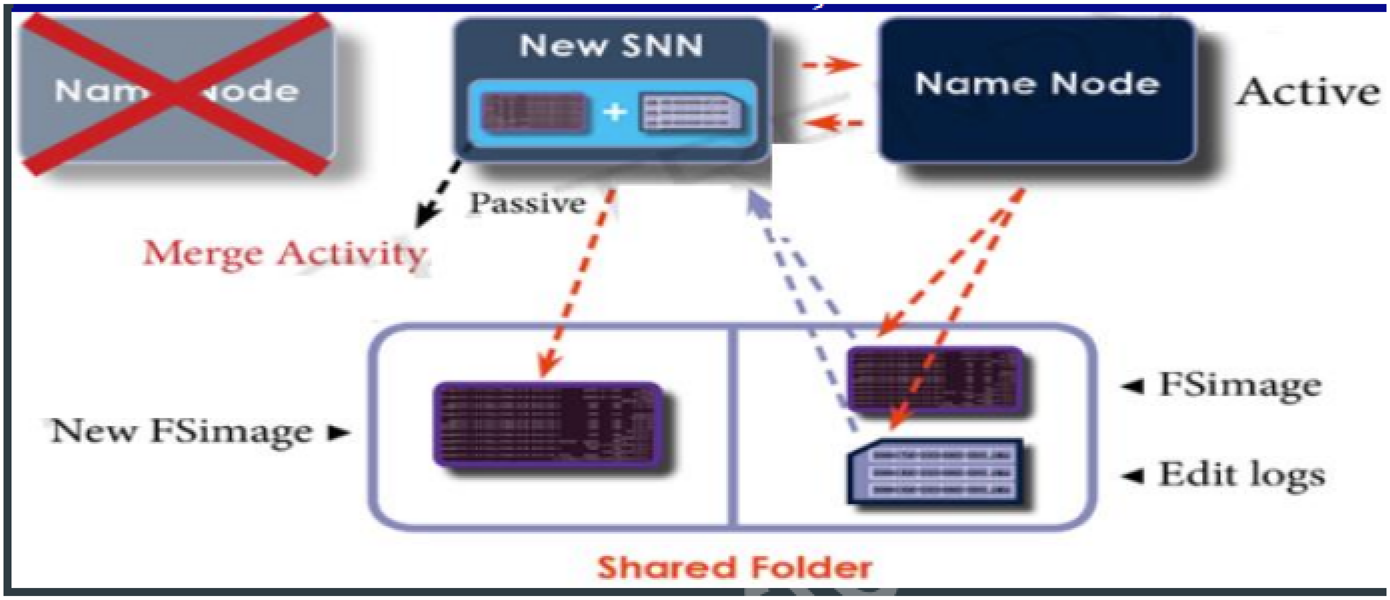
(1) Current information of the metadata @ Fsimage @ Edit logs

(i) Fsimage is a snapshot of in-memory filesystem at a given moment. That means the condition of blocks. Edit logs is all the new changes that happen after taken the latest snap shot.
Old fsimage + Edit logs = latest fsimage. That is a compute a heavy process. Name node is always so busy . so it should not take the activity of merging these 2 files. So, we need a secondary name node
(2) Secondary Name node (passive) This node only for the merging of the 2 files will take care. It is not active.


Secondary Name Node
Query for edit logs in regular interval
Name Node
Update fs image with edit logs
fs image
Copy update image back to NN
fs image

Here namenode always share the folder Fsimage & edit logs and access to the passive secondary namenode.Then new fs image is updated. These process repeats after every 30 seconds. When the merge activity completed the previous fsimage and edit logs reset to empty.

At the time name node failure the the secondary name node becomes active name node.



This time hadoop admin responsible to make sure a new secondary name node because the passive node now act as a active node now.

14. What happens if a namenode 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.

15. What happens when a user submits a Hadoop job when the NameNode is down- does the job get in to hold or does it fail.
The Hadoop job fails when the NameNode is down.

16. What happens when a user submits a Hadoop job when the Job Tracker is down- does the job get in to hold or does it fail.
The Hadoop job fails when the Job Tracker is down.

17. 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.

18. What do you understand by edge nodes in Hadoop?
Edges nodes are the interface between hadoop cluster and the external network. Edge nodes are used for running cluster adminstration tools and client applications.Edge nodes are also referred to as gateway nodes.

19. Explain the difference between NAS (Network Attached Storage) and HDFS.
NAS runs on a single machine and thus there is no probability of data redundancy whereas HDFS runs on a cluster of different machines thus there is data redundancy because of the replication protocol.

20. 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.

21. What is commodity hardware (System with average configuration and inexpensive)?

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 computers or high end hardware configuration to execute jobs.

22. What is the port number for NameNode, Task Tracker and Job Tracker?

NameNode 50070
Job Tracker 50030
Task Tracker 50060

23. Explain about the process of inter cluster data copying.

HDFS provides a distributed data copying facility through the DistCP from source to destination. If this data copying is within the hadoop cluster then it is referred to as inter cluster data copying. DistCP requires both source and destination to have a compatible or same version of hadoop.

24. What is Fault Tolerance?

Suppose you have a file stored in a system, and due to some technical problem that file gets destroyed. Then there is no chance of getting the data back present in that file. To avoid such situations, Hadoop has introduced the feature of fault tolerance in HDFS. In Hadoop, when we store a file, it automatically gets replicated at two other locations also. So even if one or two of the systems collapse, the file is still available on the third system.

25. Replication causes data redundancy, then why is it pursued in HDFS?

HDFS works with commodity hardware (systems with average configurations) that has high chances of getting crashed any time. Thus, to make the entire system highly fault-tolerant, HDFS replicates and stores data in different places. Any data on HDFS gets stored at least 3 different locations. So, even if one of them is corrupted and the other is unavailable for some time for any reason, then data can be accessed from the third one. Hence, there is no chance of losing the data. This replication factor helps us to attain the feature of Hadoop called Fault Tolerant.

26. Since the data is replicated thrice in HDFS, does it mean that any calculation done on one node will also be replicated on the other two?

No, calculations will be done only on the original data. The master node will know which node exactly has that particular data. In case, if one of the nodes is not responding, it is assumed to be failed. Only then, the required calculation will be done on the second replica.

27. Why do we use HDFS for applications having large data sets and not when there are lot of small files?

HDFS is more suitable for large amount of data sets in a single file as compared to small amount of data spread across multiple files. This is because Namenode is a very expensive high performance system, so it is not prudent to occupy the space in the Namenode by unnecessary amount of metadata that is generated for multiple small files. So, when there is a large amount of data in a single file, name node will occupy less space. Hence for getting optimized performance, HDFS supports large data sets instead of multiple small files.

28. How is indexing done in HDFS?

Hadoop has its own way of indexing. Depending upon the block size, once the data is stored, HDFS will keep on storing the last part of the data which will say where the next part of the data will be.

29. Are job tracker and task trackers present in separate machines?

Yes, job tracker and task tracker are present in different machines. The reason is job tracker is a single point of failure for the Hadoop MapReduce service. If it goes down, all running jobs are halted.

30. What is the communication channel between client and namenode/datanode?

The mode of communication is SSH.

31. What is Rack?

Rack is nothing but a group of servers. That placed on different geographical location.

One reck is nearest of our location and other should different location. Because when server is so far location then it needed very high bandwith network to get input and output result. When any local calamity occurs or damage one racks then it connect other location and continuously running the service.

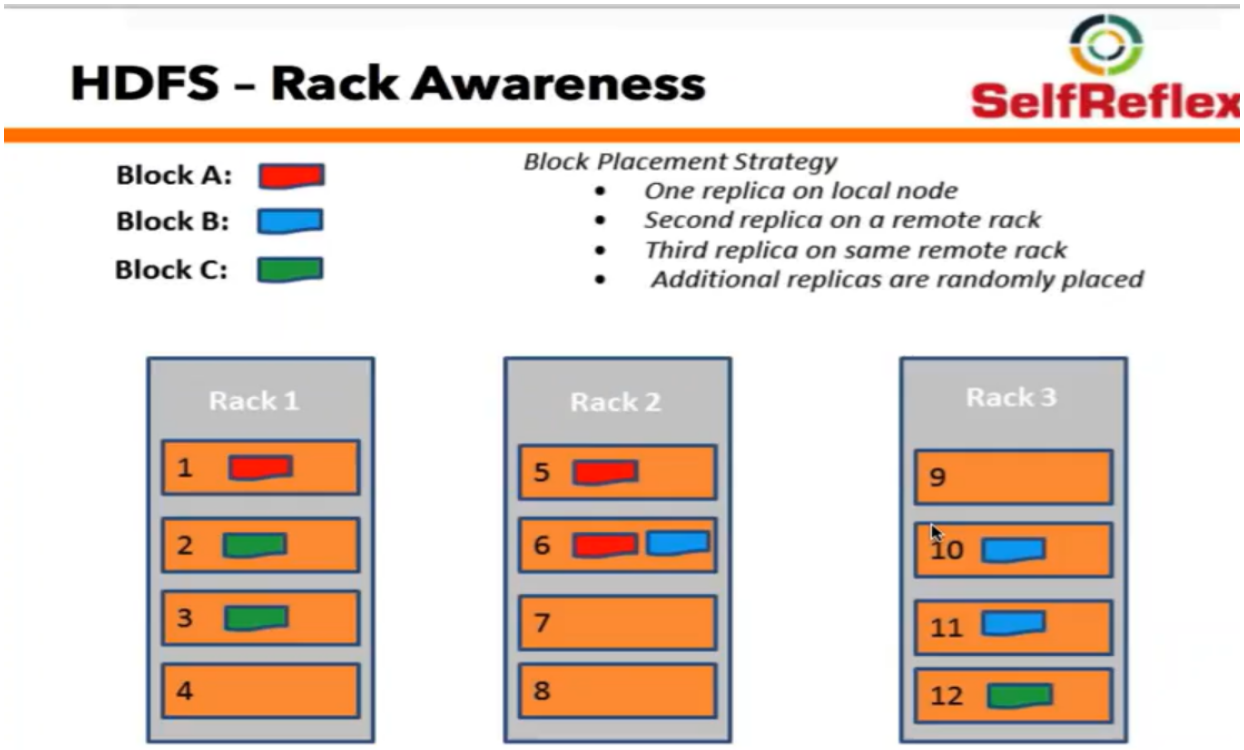
Hadoop default 3 replica of a block put 1 block a rack and others 2 block put another rack.(not other two rack because it needs high bandwith network.)but admin can change this as he likes.

Block report: Each datanode sends a block report to the name node at a fixed frequency if any blocks are corrupted then at timely name node delete the block & metadata ,finaly make new one same block.

32. What is Rack Awareness?

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. This is generally referred to as the Replica Placement Policy.



HDFS – Rack Awareness

SelfReflex

Block A: 🟥

Block B: 🟦

Block C: 🟩

Block Placement Strategy

- One replica on local node
- Second replica on a remote rack
- Third replica on same remote rack
- Additional replicas are randomly placed

33. What is Master Job Tracker and Slave Task Tracker?

Master Job Tracker---One (Like father in a family)

i) Managing Resources or Resource Management

ii) Scheduling Task

iii) Monitoring Task

Slave Task Trace---Many (Like Son & Daughters in family)

i) Executes the Task

ii) Provide Task status

34. What happens when two clients try to write into the same HDFS file?

HDFS supports exclusive writes only.

When the first client contacts the name-node to open the file for writing, the name-node grants a lease to the client to create this file. When the second client tries to open the same file for writing, the name-node will see that the lease for the file is already granted to another client, and will reject the open request for the second client

35. What does "file could only be replicated to 0 nodes, instead of 1" mean?

The namenode does not have any available DataNodes.

36. Consider case scenario: In M/R system, - HDFS block size is 64 MB

- Input format is FileInputFormat

– We have 3 files of size 64K, 65Mb and 127Mb

How many input splits will be made by Hadoop framework?

Hadoop will make 5 splits as follows –

- - 1 split for 64K files
- - 2 splits for 65MB files
- - 2 splits for 127MB files

37. Suppose Hadoop spawned 100 tasks for a job and one of the task failed. What will Hadoop do?

It will restart the task again on some other TaskTracker and only if the task fails more than four (the default setting and can be changed) times will it kill the job.

38. What are Problems with small files and HDFS?

HDFS is not good at handling large number of small files. Because every file, directory and block in HDFS is represented as an object in the namenode's memory, each of which occupies approx 150 bytes So 10 million files, each using a block, would use about 3 gigabytes of memory. when we go for a billion files the memory requirement in namenode cannot be met.

39. What is speculative execution in Hadoop?

If a node appears to be running slow, the master node can redundantly execute another instance of the same task and first output will be taken .this process is called as Speculative execution.

40. Can Hadoop handle streaming data?

Yes, through Technologies like Apache Kafka, Apache Flume, and Apache Spark it is possible to do large-scale streaming.

41. Why is Checkpointing Important in Hadoop?

As more and more files are added the namenode creates large edit logs. Which can substantially delay NameNode startup as the NameNode reapplies all the edits. Checkpointing is a process that takes an fsimage and edit log and compacts them into a new fsimage. This way, instead of replaying a potentially unbounded edit log, the NameNode can load the final in-memory state directly from the fsimage. This is a far more efficient operation and reduces NameNode startup time.

42. How can you overwrite the replication factors in HDFS?

The replication factor in HDFS can be modified or overwritten in 2 ways-

1)Using the Hadoop FS Shell, replication factor can be changed per file basis using the below command-

$hadoop fs –setrep –w 2 /my/test_file (test_file is the filename whose replication factor will be set to 2)

2)Using the Hadoop FS Shell, replication factor of all files under a given directory can be modified using the below command-

3)$hadoop fs –setrep –w 5 /my/test_dir (test_dir is the name of the directory and all the files in this directory will have a replication factor set to 5)

43. Explain the difference between NAS and HDFS.

NAS runs on a single machine and thus there is no probability of data redundancy whereas HDFS runs on a cluster of different machines thus there is data redundancy because of the replication protocol.

44. Explain about the process of inter cluster data copying.

HDFS provides a distributed data copying facility through the DistCP from source to destination. If this data copying is within the hadoop cluster then it is referred to as inter cluster data copying. DistCP requires both source and destination to have a compatible or same version of hadoop.

. Explain about the indexing process in HDFS.

Indexing process in HDFS depends on the block size. HDFS stores the last part of the data that further points to the address where the next part of data chunk is stored.

46. What do you understand by edge nodes in Hadoop?

Edges nodes are the interface between hadoop cluster and the external network. Edge nodes are used for running cluster adminstration tools and client applications.Edge nodes are also referred to as gateway nodes.

47. Copy a directory from one node in the cluster to another

Use '-distcp' command to copy,

48. Default replication factor to a file is 3.

Use '-setrep' command to change replication factor of a file to 2.

49. Which file does the Hadoop-core configuration?

core-default.xml

50. Is there a hdfs command to see available free space in hdfs

hadoop dfsadmin -report

51. The requirement is to add a new data node to a running Hadoop cluster; how do I start services on just one data node?

You do not need to shutdown and/or restart the entire cluster in this case.

First, add the new node's DNS name to the conf/slaves file on the master node.

Then log in to the new slave node and execute −

```
$ cd path/to/hadoop

$ bin/hadoop-daemon.sh start datanode

$ bin/hadoop-daemon.sh start tasktracker
```

then issuehadoop dfsadmin -refreshNodes and hadoop mradmin -refreshNodes so that
the NameNode and JobTracker know of the additional node that has been added.

52. How do you gracefully stop a running job?

Hadoop job –kill jobid

53. Does the name-node stay in safe mode till all under-replicated files are fully replicated?

No. During safe mode replication of blocks is prohibited. The name-node awaits when all or majority of data-nodes report their blocks.

54. What happens if one Hadoop client renames a file or a directory containing this file while another client is still writing into it?

A file will appear in the name space as soon as it is created. If a writer is writing to a file and another client renames either the file itself or any of its path components, then the original writer will get an IOException either when it finishes writing to the current block or when it closes the file.

How to make a large cluster smaller by taking out some of the nodes?

Hadoop offers the decommission feature to retire a set of existing data-nodes. The nodes to be retired should be included into the *exclude file*, and the exclude file name should be specified as a configuration parameter **dfs.hosts.exclude**.

The decommission process can be terminated at any time by editing the configuration or the exclude files and repeating the **-refreshNodes** command

Hadoop MapReduce Interview Questions and Answers:

- **Explain how do 'map' and 'reduce' works.**

Namenode takes the input and divide it into parts and assign them to data nodes. These datanodes process the tasks assigned to them and make a key-value pair and returns the intermediate output to the Reducer. The reducer collects this key value pairs of all the datanodes and combines them and generates the final output.

1. Explain the usage of Context Object.
Context Object is used to help the mapper interact with other Hadoop systems. Context Object can be used for updating counters, to report the progress and to provide any application level status updates. ContextObject has the configuration details for the job and also interfaces, that helps it to generating the output.

2. What are the core methods of a Reducer?
The 3 core methods of a reducer are –
1)setup () – This method of the reducer is used for configuring various parameters like the input data size, distributed cache, heap size, etc.
Function Definition- *public void setup (context)*
2)reduce () it is heart of the reducer which is called once per key with the associated reduce task.
Function Definition -public void reduce (Key,Value,context)
3)cleanup () - This method is called only once at the end of reduce task for clearing all the temporary files.
Function Definition -public void cleanup (context)

3. Explain about the Recorde Reader, partitioning, shuffle and sort phase

Recored Reader-Record reader is a tool that takes each line from datanode and convert each line as key value pair.
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.

4. How to write a custom partitioner for a Hadoop MapReduce job?

Steps to write a Custom Partitioner for a Hadoop MapReduce Job-

- A new class must be created that extends the pre-defined Partitioner Class.
- getPartition method of the Partitioner class must be overridden.
- The custom partitioner to the job can be added as a config file in the wrapper which runs Hadoop MapReduce or the custom partitioner can be added to the job by using the set method of the partitioner class.

5. What are side data distribution techniques in Hadoop?

The extra read only data required by a hadoop job to process the main dataset is referred to as side data. Hadoop has two side data distribution techniques -

i) Using the job configuration - This technique should not be used for transferring more than few kilobytes of data as it can pressurize the memory usage of hadoop daemons,particularly if your system is running several hadoop jobs.

ii) Distributed Cache - Rather than serializing side data using the job configuration, it is suggested to distribute data using hadoop's distributed cache mechanism.

Hadoop Interview FAQ’s – An Interviewee Should Ask an Interviewer

For many hadoop job seekers, the question from the interviewer – “Do you have any questions for me?” indicates the end of a Hadoop developer job interview. It is always enticing for a Hadoop job seeker to immediately say “No” to the question for the sake of keeping the first impression intact.However, to land a hadoop job or any other job, it is always preferable to fight that urge and ask relevant questions to the interviewer.

Asking questions related to the Hadoop technology implementation, shows your interest in the open hadoop job role and also conveys your interest in working with the company.Just like any other interview, even hadoop interviews are a two-way street- it helps the interviewer decide whether you have the desired hadoop skills they in are looking for in a hadoop developer, and helps an interviewee decide if that is the kind of big data infrastructure and hadoop technology implementation you want to devote your skills for foreseeable future growth in the big data domain.

Candidates should not be afraid to ask questions to the interviewer. To ease this for hadoop job seekers, ProjectPro has collated few hadoop interview FAQ’s that every candidate should ask an interviewer during their next hadoop job interview-

1) What is the size of the biggest hadoop cluster a company X operates?

Asking this question helps a hadoop job seeker understand the hadoop maturity curve at a company.Based on the answer of the interviewer, a candidate can judge how much an organization invests in Hadoop and their enthusiasm to buy big data products from various vendors. The candidate can also get an idea on the hiring needs of the company based on their hadoop infrastructure.

2) For what kind of big data problems, did the organization choose to use Hadoop?

Asking this question to the interviewer shows the candidates keen interest in understanding the reason for hadoop implementation from a business perspective. This question gives the impression to the interviewer that the candidate is not merely interested in the hadoop developer job role but is also interested in the growth of the company.

3) Based on the answer to question no 1, the candidate can ask the interviewer why the hadoop infrastructure is configured in that particular way, why the company chose to use the selected big data tools and how workloads are constructed in the hadoop environment.

Asking this question to the interviewer gives the impression that you are not just interested in maintaining the big data system and developing products around it but are also seriously thoughtful on how the infrastructure can be improved to help business growth and make cost savings.

4) What kind of data the organization works with or what are the HDFS file formats the company uses?

The question gives the candidate an idea on the kind of big data he or she will be handling if selected for the hadoop developer job role. Based on the data, it gives an idea on the kind of analysis they will be required to perform on the data.

5) What is the most complex problem the company is trying to solve using Apache Hadoop?

Asking this question helps the candidate know more about the upcoming projects he or she might have to work and what are the challenges around it. Knowing this beforehand helps the interviewee prepare on his or her areas of weakness.

6) Will I get an opportunity to attend big data conferences? Or will the organization incur any costs involved in taking advanced hadoop or big data certification?

This is a very important question that you should be asking these the interviewer. This helps a candidate understand whether the prospective hiring manager is interested and supportive when it comes to professional development of the employee.

Hadoop Interview FAQ’s – Interviewer Asks an Interviewee

So, you have cleared the technical interview after preparing thoroughly with the help of the Hadoop Interview Questions shared by ProjectPro. After an in-depth technical interview, the interviewer might still not be satisfied and would like to test your practical experience in navigating and analysing big data. The expectation of the interviewer is to judge whether you are really interested in the open position and ready to work with the company, regardless of the technical knowledge you have on hadoop technology.

There are quite a few on-going debates in the hadoop community, on the advantages of the various components in the hadoop ecosystem-- for example what is better MapReduce, Pig or Hive or Spark vs. Hadoop or when should a company use MapReduce over other alternative? etc. Interviewee and Interviewer should both be ready to answer such hadoop interview FAQs, as there is no right or wrong answer to these questions.The best possible way to answer these Hadoop interview FAQs is to explain why a particular interviewee favours an option. Answering these hadoop interview FAQs with practical examples as to why the candidate favours an option, demonstrates his or her understanding of the business needs and helps the interviewer judge the flexibility of the candidate to use various big data tools in the hadoop ecosystem.

Here are a few hadoop interview FAQs that are likely to be asked by the interviewer-

1) If you are an experienced hadoop professional then you are likely to be asked questions like –

- The number of nodes you have worked with in a cluster.
- Which hadoop distribution have you used in your recent project.
- Your experience on working with special configurations like High Availability.
- The data volume you have worked with in your most recent project.
- What are the various tools you used in the big data and hadoop projects you have worked on?

Your answer to these interview questions will help the interviewer understand your expertise in Hadoop based on the size of the hadoop cluster and number of nodes. Based on the highest volume of data you have handled in your previous projects, interviewer can assess your overall experience in debugging and troubleshooting issues involving huge hadoop clusters.

The number of tools you have worked with help an interviewer judge that you are aware of the overall hadoop ecosystem and not just MapReduce. To be selected, it all depends on how well you communicate the answers to all these questions.

2) What are the challenges that you faced when implementing hadoop projects?

Interviewers are interested to know more about the various issues you have encountered in the past when working with hadoop clusters and understand how you addressed them. The way you answer this question tells a lot about your expertise in troubleshooting and debugging hadoop clusters.The more issues you have encountered, the more probability there is, that you have become an expert in that area of Hadoop. Ensure that you list out all the issues that have trouble-shooted.

3) How were you involved in data modelling, data ingestion, data transformation and data aggregation?

You are likely to be involved in one or more phases when working with big data in a hadoop environment. The answer to this question helps the interviewer understand what kind of tools you are familiar with. If you answer that your focus was mainly on data ingestion then they can expect you to be well-versed with Sqoop and Flume, if you answer that you were involved in data analysis and data transformation then it gives the interviewer an impression that you have expertise in using Pig and Hive.

4) What is your favourite tool in the hadoop ecosystem?

The answer to this question will help the interviewer know more about the big data tools that you are well-versed with and are interested in working with. If you show affinity towards a particular tool then the probability that you will be deployed to work on that particular tool, is more.If you say that you have a good knowledge of all the popular big data tools like pig, hive, HBase, Sqoop, flume then it shows that you have knowledge about the hadoop ecosystem as a whole.

5) In you previous project, did you maintain the hadoop cluster in-house or used hadoop in the cloud?

Most of the organizations still do not have the budget to maintain Hadoop cluster in-house and they make use of Hadoop in the cloud from various vendors like Amazon, Microsoft, Google, etc. Interviewer gets to know about your familiarity with using hadoop in the cloud because if the company does not have an in-house implementation then hiring a candidate who has knowledge about using hadoop in the cloud is worth it.

BigData Interview Questions asked at Top Tech Companies

1) Write a MapReduce program to add all the elements in a file. (Hadoop Developer Interview Question asked at KPIT)

2) What is the difference between HashSet and hashmap? (Big Data Interview Question asked at Wipro)

3) Write a Hive program to find the number of employees department-wise in an organization. (Hadoop Developer Interview Question asked at Tripod Technologies)

4) How will you read a CSV file of 10GB and store it in the database as it is in just few seconds? (Hadoop Interview Question asked at Deutsche Bank)

5) How will a file of 100MB be stored in Hadoop? (Hadoop Interview Question asked at Deutsche Bank)

6) Given the HTTP URL, write a regular expression to extract different parts of the URL like server, path, and protocol.

7) If you store a file from HDFS to Apache Pig using PIGSTORAGE in grunt shell but if that file is not really available in that HDFS path will the command work?

8) Apart from partitioning and bucketing, what are the other methods you can use for improving performance?

9) Which file format is better Parquet or ORC?