Top 100 Hadoop Interview Questions and Answers 2016

21 Aug 2015

Latest Update made on May 20, 2016.

Big Data and Hadoop is a constantly changing field which required people to quickly upgrade their skills, to fit the requirements for Hadoop related jobs. If you are applying for a Hadoop job role, it is best to be prepared to answer any Hadoop interview question that might come your way. We will keep updating this list of Hadoop Interview questions, to suit the current industry standards.

If you would like more information about Big Data careers, please click the orange "Request Info" button on top of this page.

With more than 30,000 open Hadoop developer jobs, professionals must familiarize themselves with the each and every component of the Hadoop ecosystem to make sure that they have a deep understanding of what Hadoop is so that they can form an effective approach to a given big data problem. To help you get started, DeZyre presented a comprehensive list of Top 50 Hadoop Developer Interview Questions asked during recent Hadoop job interviews.

With the help of DeZyre's Hadoop Instructors, we have put together a detailed list of Hadoop latest interview questions based on the different components of the Hadoop Ecosystem such as MapReduce, Hive, HBase, Pig, YARN, Flume, Sqoop, HDFS, etc.

We had to spend lots of hours researching and deliberating on what are the best possible answers to these interview questions. We would love to invite people from the industry – hadoop developers, hadoop admins and architects to kindly help us and everyone else – with answering the unanswered questions.

Here are top Hadoop Developer Interview Questions and Answers based on different components of the Hadoop Ecosystem-

1) Hadoop Basic Interview Questions

2) Hadoop HDFS Interview Questions

3) MapReduce Interview Questions

4) Hadoop HBase Interview Questions

5) Hadoop Sqoop Interview Questions

6) Hadoop Flume Interview Questions

7) Hadoop Zookeeper Interview Questions

8) Pig Interview Questions

9) Hive Interview Questions

10) Hadoop YARN Interview Questions

Big Data Hadoop Interview Questions and Answers

These are Hadoop Basic Interview Questions and Answers for freshers and experienced.

1. What is Big Data?Click here to Tweet

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.

Here is an interesting and explanatory visual on "What is Big Data?"

1/11 Hadoop Training- What is Big Data by DeZyre.com

2. What do the four V's of Big Data denote? Click here to Tweet

IBM has a nice, simple explanation for the four critical features of big data:
a) Volume –Scale of data
b) Velocity –Analysis of streaming data
c) Variety – Different forms of data
d) Veracity –Uncertainty of data

Here is an explanatory video on the four V's of Big Data

Mobile App Development
NoSQL Database
Web Development

You might also like

Sqoop Interview Questions a for 2016
Working with Spark RDD for F Processing
Improve Your LinkedIn Profile right Hadoop Job!
Top 5 Apache Spark Use Cas
Life Cycle of a Data Science
Recap of Data Science News
Recap of Apache Spark News
Recap of Hadoop News for M
HDFS Interview Questions an for 2016
Data engineer vs. Data scient does your company need?

Tutorials

PySpark Tutorial-Learn to use Spark with Python
R Tutorial- Learn Data Visuali using GGVIS
Neural Network Training Tuto
Python List Tutorial
MatPlotLib Tutorial
Decision Tree Tutorial
Neural Network Tutorial
Performance Metrics for Mach Algorithms

Introduction to Apache Spark
R Tutorial: Importing Data fror
R Tutorial: Importing Data fror Database

2/11 Hadoop Training- Four V's of Big Data by DeZyre.com

3. How big data analysis helps businesses increase their revenue? Give example. Click here to Tweet

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.

Here is an interesting video that explains how various industries are leveraging big data analysis to increase their revenue

5/11 Hadoop Training- Top 10 industries using Big Data by...

4. Name some companies that use Hadoop. Click here to tweet this question

Yahoo (One of the biggest user & more than 80% code contributor to Hadoop)
Facebook
Netflix
Amazon
Adobe
eBay
Hulu
Spotify
Rubikloud
Twitter

What companies are you applying to for Hadoop job roles?

Enter your name here...

Write your answer here...

SUBMIT

Click on this link to view a detailed list of some of the top companies using Hadoop.

5. Differentiate between Structured and Unstructured data. Click here to Tweet

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.

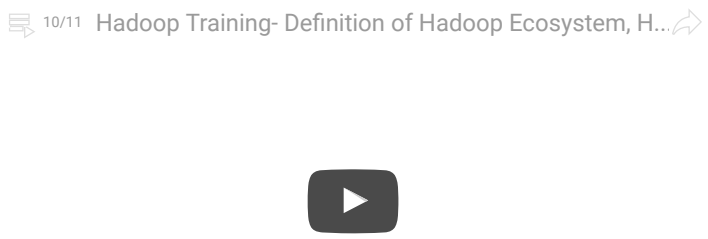
6. On what concept the Hadoop framework works? Click here to Tweet

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.

Here is a visual that clearly explain the HDFS and Hadoop MapReduce Concepts-



7) What are the main components of a Hadoop Application? Click here to Tweet

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

- Hadoop HDFS Tutorial
- Hadoop hBase Tutorial
- Hadoop Flume Tutorial
- Hadoop 2.0 YARN Tutorial
- Hadoop MapReduce Tutorial
- Big Data Hadoop Tutorial for I Hadoop Installation

Online Courses

- Hadoop Training
- Spark Certification Training
- Data Science in Python
- Data Science inR
- Data Science Training
- Hadoop Training in California
- Hadoop Training in New York
- Hadoop Training in Texas
- Hadoop Training in Virginia
- Hadoop Training in Washingt
- Hadoop Training in New Jers

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.

8. What is Hadoop streaming? Click here to Tweet

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.

9. What is the best hardware configuration to run Hadoop? Click here to Tweet

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.

10. What are the most commonly defined input formats in Hadoop? Click here to Tweet

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.

We have further categorized Big Data Interview Questions for Freshers and Experienced-

- Hadoop Interview Questions and Answers for Freshers - Q.Nos- 1,2,4,5,6,7,8,9
- Hadoop Interview Questions and Answers for Experienced - Q.Nos-3,8,9,10

For a detailed PDF report on Hadoop Salaries - CLICK HERE

Hadoop HDFS Interview Questions and Answers

1. What is a block and block scanner in HDFS? Click here to Tweet

Block - The minimum amount of data that can be read or written is generally referred to as a “block” in HDFS. The default size of a block in HDFS is 64MB.

Block Scanner - Block Scanner tracks the list of blocks present on a DataNode and verifies them to find any kind of checksum errors. Block Scanners use a throttling mechanism to reserve disk bandwidth on the datanode.

2. Explain the difference between NameNode, Backup Node and Checkpoint NameNode. Click here to Tweet

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.

3. What is commodity hardware? Click here to Tweet

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.

4. What is the port number for NameNode, Task Tracker and Job Tracker? Click here to Tweet

NameNode 50070

Job Tracker 50030

Task Tracker 50060

5. Explain about the process of inter cluster data copying. Click here to Tweet

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.

6. How can you overwrite the replication factors in HDFS? Click here to Tweet

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)

7. Explain the difference between NAS and HDFS. Click here to Tweet

- 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.
- NAS stores data on a dedicated hardware whereas in HDFS all the data blocks are distributed across local drives of the machines.
- In NAS data is stored independent of the computation and hence Hadoop MapReduce cannot be used for processing whereas HDFS works with Hadoop MapReduce as the computations in HDFS are moved to data.

What technologies are you working on currently? (Java, Datawarehouse, Business Intelligence, ETL, etc.)

Enter your name here...

Write your answer here...

SUBMIT

8. Explain what happens if during the PUT operation, HDFS block is assigned a replication factor 1 instead of the default value 3. Click here to Tweet

Replication factor is a property of HDFS that can be set accordingly for the entire cluster to adjust the number of times the blocks are to be replicated to ensure high data availability. For every block that is stored in HDFS, the cluster will have n-1 duplicated blocks. So, if the replication factor during the PUT operation is set to 1 instead of the default value 3, then it will have a single copy of data. Under these circumstances when the replication factor is set to 1 ,if the DataNode crashes under any circumstances, then only single copy of the data would be lost.

9. What is the process to change the files at arbitrary locations in HDFS? Click here to Tweet

HDFS does not support modifications at arbitrary offsets in the file or multiple writers but files are written by a single writer in append only format i.e. writes to a file in HDFS are always made at the end of the file.

10. Explain about the indexing process in HDFS. Click here to Tweet

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.

11. What is a rack awareness and on what basis is data stored in a rack? Click here to Tweet

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.

12. 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.

13. 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.

14. 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.

15. 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.

We have further categorized Hadoop HDFS Interview Questions for Freshers and Experienced-

- Hadoop Interview Questions and Answers for Freshers - Q.Nos- 2,3,7,9,10,11,13,14
- Hadoop Interview Questions and Answers for Experienced - Q.Nos- 1,2, 4,5,6,7,8,12,15

Hadoop MapReduce Interview Questions and Answers

1. Explain the usage of Context Object. Click here to Tweet

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? Click here to Tweet

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 partitioning, shuffle and sort phase Click here to Tweet

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? Click here to Tweet

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.

We have further categorized Hadoop MapReduce Interview Questions for Freshers and Experienced-

- Hadoop Interview Questions and Answers for Freshers - Q.Nos- 2
- Hadoop Interview Questions and Answers for Experienced - Q.Nos- 1,3,4,

Here are a few more frequently asked Hadoop MapReduce Interview Questions and Answers

Hadoop HBase Interview Questions and Answers

1. 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.

2. What are the different operational commands in HBase at record level and table level?

Record Level Operational Commands in HBase are –put, get, increment, scan and delete.

Table Level Operational Commands in HBase are-describe, list, drop, disable and scan.

3. 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.

4. 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.

5. Explain about the different catalog tables in HBase?

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

6. What is column families? What happens if you alter the block size of ColumnFamily on an already populated database?

The logical deviation of data is represented through a key known as column Family. Column families consist of the basic unit of physical storage on which compression features can be applied. In an already populated database, when the block size of column family is altered, the old data will remain within the old block size whereas the new data that comes in will take the new block size. When compaction takes place, the old data will take the new block size so that the existing data is read correctly.

7. 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.

8. Explain the process of row deletion in HBase.

On issuing a delete command in HBase through the HBase client, data is not actually deleted from the cells but rather the cells are made invisible by setting a tombstone marker. The deleted cells are removed at regular intervals during compaction.

9. What are the different types of tombstone markers in HBase for deletion?

There are 3 different types of tombstone markers in HBase for deletion-

1)Family Delete Marker- This markers marks all columns for a column family.

2)Version Delete Marker-This marker marks a single version of a column.

3)Column Delete Marker-This markers marks all the versions of a column.

10. Explain about HLog and WAL in HBase.

All edits in the HStore are stored in the HLog. Every region server has one HLog. HLog contains entries for edits of all regions performed by a particular Region Server.WAL abbreviates to Write Ahead Log (WAL) in which all the HLog edits are written immediately.WAL edits remain in the memory till the flush period in case of deferred log flush.

We have further categorized Hadoop HBase Interview Questions for Freshers and Experienced-

- Hadoop Interview Questions and Answers for Freshers - Q.Nos-1,2,4,5,7
- Hadoop Interview Questions and Answers for Experienced - Q.Nos-2,3,6,8,9,10

Hadoop Sqoop Interview Questions and Answers

1. Explain about some important Sqoop commands other than import and export.

Create Job (--create)

Here we are creating a job with the name my job, which can import the table data from RDBMS table to HDFS. The following command is used to create a job that is importing data from the employee table in the db database to the HDFS file.

```
$ Sqoop job --create myjob \

--import \

--connect jdbc:mysql://localhost/db \

--username root \

--table employee --m 1
```

Verify Job (--list)

'--list' argument is used to verify the saved jobs. The following command is used to verify the list of saved Sqoop jobs.

```
$ Sqoop job --list
```

Inspect Job (--show)

'--show' argument is used to inspect or verify particular jobs and their details. The following command and sample output is used to verify a job called myjob.

$ Sqoop job --show myjob

Execute Job (--exec)

'--exec' option is used to execute a saved job. The following command is used to execute a saved job called myjob.

$ Sqoop job --exec myjob

2. How Sqoop can be used in a Java program?

The Sqoop jar in classpath should be included in the java code. After this the method Sqoop.runTool () method must be invoked. The necessary parameters should be created to Sqoop programmatically just like for command line.

3. What is the process to perform an incremental data load in Sqoop?

The process to perform incremental data load in Sqoop is to synchronize the modified or updated data (often referred as delta data) from RDBMS to Hadoop. The delta data can be facilitated through the incremental load command in Sqoop.

Incremental load can be performed by using Sqoop import command or by loading the data into hive without overwriting it. The different attributes that need to be specified during incremental load in Sqoop are-

1)Mode (incremental) –The mode defines how Sqoop will determine what the new rows are. The mode can have value as Append or Last Modified.

2)Col (Check-column) –This attribute specifies the column that should be examined to find out the rows to be imported.

3)Value (last-value) –This denotes the maximum value of the check column from the previous import operation.

4. Is it possible to do an incremental import using Sqoop?

Yes, Sqoop supports two types of incremental imports-

1)Append

2)Last Modified

To insert only rows Append should be used in import command and for inserting the rows and also updating Last-Modified should be used in the import command.

5. What is the standard location or path for Hadoop Sqoop scripts?

/usr/bin/Hadoop Sqoop

6. How can you check all the tables present in a single database using Sqoop?

The command to check the list of all tables present in a single database using Sqoop is as follows-

Sqoop list-tables –connect jdbc: mysql: //localhost/user;

7. How are large objects handled in Sqoop?

Sqoop provides the capability to store large sized data into a single field based on the type of data. Sqoop supports the ability to store-

1)CLOB 's – Character Large Objects

2)BLOB's –Binary Large Objects

Large objects in Sqoop are handled by importing the large objects into a file referred as "LobFile" i.e. Large Object File. The LobFile has the ability to store records of huge size, thus each record in the LobFile is a large object.

8. Can free form SQL queries be used with Sqoop import command? If yes, then how can they be used?

Sqoop allows us to use free form SQL queries with the import command. The import command should be used with the –e and – query options to execute free form SQL queries. When using the –e and –query options with the import command the –target dir value must be specified.

9. Differentiate between Sqoop and distCP.

DistCP utility can be used to transfer data between clusters whereas Sqoop can be used to transfer data only between Hadoop and RDBMS.

10. What are the limitations of importing RDBMS tables into Hcatalog directly?

There is an option to import RDBMS tables into Hcatalog directly by making use of –hcatalog –database option with the –hcatalog –table but the limitation to it is that there are several arguments like –as-avrofile , -direct, -as-sequencefile, -target-dir , -export-dir are not supported.

We have further categorized Hadoop Sqoop Interview Questions for Freshers and Experienced-

- Hadoop Interview Questions and Answers for Freshers - Q.Nos- 4,5,6,9
- Hadoop Interview Questions and Answers for Experienced - Q.Nos- 1,2,3,6,7,8,10

Hadoop Flume Interview Questions and Answers

1) Explain about the core components of Flume.

The core components of Flume are –

Event- The single log entry or unit of data that is transported.

Source- This is the component through which data enters Flume workflows.

Sink-It is responsible for transporting data to the desired destination.

Channel- it is the duct between the Sink and Source.

Agent- Any JVM that runs Flume.

Client- The component that transmits event to the source that operates with the agent.

2) Does Flume provide 100% reliability to the data flow?

Yes, Apache Flume provides end to end reliability because of its transactional approach in data flow.

3) How can Flume be used with HBase?

Apache Flume can be used with HBase using one of the two HBase sinks –

- HBaseSink (org.apache.flume.sink.hbase.HBaseSink) supports secure HBase clusters and also the novel HBase IPC that was introduced in the version HBase 0.96.
- AsyncHBaseSink (org.apache.flume.sink.hbase.AsyncHBaseSink) has better performance than HBase sink as it can easily make non-blocking calls to HBase.

Working of the HBaseSink –

In HBaseSink, a Flume Event is converted into HBase Increments or Puts. Serializer implements the HBaseEventSerializer which is then instantiated when the sink starts. For every event, sink calls the initialize method in the serializer which then translates the Flume Event into HBase increments and puts to be sent to HBase cluster.

Working of the AsyncHBaseSink-

AsyncHBaseSink implements the AsyncHBaseEventSerializer. The initialize method is called only once by the sink when it starts. Sink invokes the setEvent method and then makes calls to the getIncrements and getActions methods just similar to HBase sink. When the sink stops, the cleanUp method is called by the serializer.

4) Explain about the different channel types in Flume. Which channel type is faster?

The 3 different built in channel types available in Flume are-

MEMORY Channel – Events are read from the source into memory and passed to the sink.

JDBC Channel – JDBC Channel stores the events in an embedded Derby database.

FILE Channel –File Channel writes the contents to a file on the file system after reading the event from a source. The file is deleted only after the contents are successfully delivered to the sink.

MEMORY Channel is the fastest channel among the three however has the risk of data loss. The channel that you choose completely depends on the nature of the big data application and the value of each event.

5) Which is the reliable channel in Flume to ensure that there is no data loss?

FILE Channel is the most reliable channel among the 3 channels JDBC, FILE and MEMORY.

6) Explain about the replication and multiplexing selectors in Flume.

Channel Selectors are used to handle multiple channels. Based on the Flume header value, an event can be written just to a single channel or to multiple channels. If a channel selector is not specified to the source then by default it is the Replicating selector. Using the replicating selector, the same event is written to all the channels in the source's channels list. Multiplexing channel selector is used when the application has to send different events to different channels.

7) How multi-hop agent can be setup in Flume?

Avro RPC Bridge mechanism is used to setup Multi-hop agent in Apache Flume.

8) Does Apache Flume provide support for third party plug-ins?

Most of the data analysts use Apache Flume has plug-in based architecture as it can load data from external sources and transfer it to external destinations.

9) Is it possible to leverage real time analysis on the big data collected by Flume directly? If yes, then explain how.

Data from Flume can be extracted, transformed and loaded in real-time into Apache Solr servers using

MorphlineSolrSink

10) Differentiate between FileSink and FileRollSink

The major difference between HDFS FileSink and FileRollSink is that HDFS File Sink writes the events into the Hadoop Distributed File System (HDFS) whereas File Roll Sink stores the events into the local file system.

Hadoop Flume Interview Questions and Answers for Freshers - Q.Nos- 1,2,4,5,6,10

Hadoop Flume Interview Questions and Answers for Experienced- Q.Nos- 3,7,8,9

Hadoop Zookeeper Interview Questions and Answers

1) Can Apache Kafka be used without Zookeeper?

It is not possible to use Apache Kafka without Zookeeper because if the Zookeeper is down Kafka cannot serve client request.

2) Name a few companies that use Zookeeper.

Yahoo, Solr, Helprace, Neo4j, Rackspace

3) What is the role of Zookeeper in HBase architecture?

In HBase architecture, ZooKeeper is the monitoring server that provides different services like –tracking server failure and network partitions, maintaining the configuration information, establishing communication between the clients and region servers, usability of ephemeral nodes to identify the available servers in the cluster.

4) Explain about ZooKeeper in Kafka

Apache Kafka uses ZooKeeper to be a highly distributed and scalable system. Zookeeper is used by Kafka to store various configurations and use them across the hadoop cluster in a distributed manner. To achieve distributed-ness, configurations are distributed and replicated throughout the leader and follower nodes in the ZooKeeper ensemble. We cannot directly connect to Kafka by bye-passing ZooKeeper because if the ZooKeeper is down it will not be able to serve the client request.

5) Explain how Zookeeper works

ZooKeeper is referred to as the King of Coordination and distributed applications use ZooKeeper to store and facilitate important configuration information updates. ZooKeeper works by coordinating the processes of distributed applications. ZooKeeper is a robust replicated synchronization service with eventual consistency. A set of nodes is known as an ensemble and persisted data is distributed between multiple nodes.

3 or more independent servers collectively form a ZooKeeper cluster and elect a master. One client connects to any of the specific server and migrates if a particular node fails. The ensemble of ZooKeeper nodes is alive till the majority of nods are working. The master node in ZooKeeper is dynamically selected by the consensus within the ensemble so if the master node fails then the role of master node will migrate to another node which is selected dynamically. Writes are linear and reads are concurrent in ZooKeeper.

6) List some examples of Zookeeper use cases.

- Found by Elastic uses Zookeeper comprehensively for resource allocation, leader election, high priority notifications and discovery. The entire service of Found built up of various systems that read and write to Zookeeper.
- Apache Kafka that depends on ZooKeeper is used by LinkedIn
- Storm that relies on ZooKeeper is used by popular companies like Groupon and Twitter.

7) How to use Apache Zookeeper command line interface?

ZooKeeper has a command line client support for interactive use. The command line interface of ZooKeeper is similar to the file and shell system of UNIX. Data in ZooKeeper is stored in a hierarchy of Znodes where each znode can contain data just similar to a file. Each znode can also have children just like directories in the UNIX file system.

Zookeeper-client command is used to launch the command line client. If the initial prompt is hidden by the log messages after entering the command, users can just hit ENTER to view the prompt.

8) What are the different types of Znodes?

There are 2 types of Znodes namely- Ephemeral and Sequential Znodes.

- The Znodes that get destroyed as soon as the client that created it disconnects are referred to as Ephemeral Znodes.
- Sequential Znode is the one in which sequential number is chosen by the ZooKeeper ensemble and is pre-fixed when the client assigns name to the znode.

9) What are watches?

Client disconnection might be troublesome problem especially when we need to keep a track on the state of Znodes at regular intervals. ZooKeeper has an event system referred to as watch which can be set on Znode to trigger an event whenever it is removed, altered or any new children are created below it.

10) What problems can be addressed by using Zookeeper?

In the development of distributed systems, creating own protocols for coordinating the hadoop cluster results in failure and frustration for the developers. The architecture of a distributed system can be prone to deadlocks, inconsistency and race conditions. This leads to various difficulties in making the hadoop cluster fast, reliable and scalable. To address all such problems, Apache ZooKeeper can be used as a coordination service to write correct distributed applications without having to reinvent the wheel from the beginning.

Hadoop ZooKeeper Interview Questions and Answers for Freshers - Q.Nos- 1,2,8,9

Hadoop ZooKeeper Interview Questions and Answers for Experienced- Q.Nos-3,4,5,6,7, 10

Hadoop Pig Interview Questions and Answers

1) What are different modes of execution in Apache Pig?

Apache Pig runs in 2 modes- one is the “Pig (Local Mode) Command Mode” and the other is the “Hadoop MapReduce (Java) Command Mode”. Local Mode requires access to only a single machine where all files are installed and executed on a local host whereas MapReduce requires accessing the Hadoop cluster.

2) Explain about co-group in Pig.

COGROUP operator in Pig is used to work with multiple tuples. COGROUP operator is applied on statements that contain or involve two or more relations. The COGROUP operator can be applied on up to 127 relations at a time. When using the COGROUP operator on two tables at once-Pig first groups both the tables and after that joins the two tables on the grouped columns.

We have further categorized Hadoop Pig Interview Questions for Freshers and Experienced-

- Hadoop Interview Questions and Answers for Freshers - Q.No-1
- Hadoop Interview Questions and Answers for Experienced - Q.No- 2

Here are a few more frequently asked Pig Hadoop Interview Questions and Answers for Freshers and Experienced

Hadoop Hive Interview Questions and Answers

1) Explain about the SMB Join in Hive.

In SMB join in Hive, each mapper reads a bucket from the first table and the corresponding bucket from the second table and then a merge sort join is performed. Sort Merge Bucket (SMB) join in hive is mainly used as there is no limit on file or partition or table join. SMB join can best be used when the tables are large. In SMB join the columns are bucketed and sorted using the join columns. All tables should have the same number of buckets in SMB join.

2) How can you connect an application, if you run Hive as a server?

When running Hive as a server, the application can be connected in one of the 3 ways-

ODBC Driver-This supports the ODBC protocol

JDBC Driver- This supports the JDBC protocol

Thrift Client- This client can be used to make calls to all hive commands using different programming language like PHP, Python, Java, C++ and Ruby.

3) What does the overwrite keyword denote in Hive load statement?

Overwrite keyword in Hive load statement deletes the contents of the target table and replaces them with the files referred by the file path i.e. the files that are referred by the file path will be added to the table when using the overwrite keyword.

4) What is SerDe in Hive? How can you write your own custom SerDe?

SerDe is a Serializer DeSerializer. Hive uses SerDe to read and write data from tables. Generally, users prefer to write a Deserializer instead of a SerDe as they want to read their own data format rather than writing to it. If the SerDe supports DDL i.e. basically SerDe with parameterized columns and different column types, the users can implement a Protocol based DynamicSerDe rather than writing the SerDe from scratch.

We have further categorized Hadoop Hive Interview Questions for Freshers and Experienced-

Hadoop Hive Interview Questions and Answers for Freshers- Q.Nos-3

Hadoop Hive Interview Questions and Answers for Experienced- Q.Nos-1,2,4

Here are a few more frequently asked Hadoop Hive Interview Questions and Answers for Freshers and Experienced

Hadoop YARN Interview Questions and Answers

1)What are the stable versions of Hadoop?

Release 2.7.1 (stable)

Release 2.4.1

Release 1.2.1 (stable)

2) What is Apache Hadoop YARN?

YARN is a powerful and efficient feature rolled out as a part of Hadoop 2.0.YARN is a large scale distributed system for running big data applications.

3) Is YARN a replacement of Hadoop MapReduce?

YARN is not a replacement of Hadoop but it is a more powerful and efficient technology that supports MapReduce and is also referred to as Hadoop 2.0 or MapReduce 2.

4) 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

5) How can native libraries be included in YARN jobs?

There are two ways to include native libraries in YARN jobs-

1) By setting the -Djava.library.path on the command line but in this case there are chances that the native libraries might not be loaded correctly and there is possibility of errors.

2) The better option to include native libraries is to the set the LD_LIBRARY_PATH in the .bashrc file.

6) Explain the differences between Hadoop 1.x and Hadoop 2.x

- In Hadoop 1.x, MapReduce is responsible for both processing and cluster management whereas in Hadoop 2.x processing is taken care of by other processing models and YARN is responsible for cluster management.
- Hadoop 2.x scales better when compared to Hadoop 1.x with close to 10000 nodes per cluster.
- Hadoop 1.x has single point of failure problem and whenever the NameNode fails it has to be recovered manually. However, in case of Hadoop 2.x StandBy NameNode overcomes the SPOF problem and whenever the NameNode fails it is configured for automatic recovery.
- Hadoop 1.x works on the concept of slots whereas Hadoop 2.x works on the concept of containers and can also run generic tasks.

7) What are the core changes in Hadoop 2.0?

Hadoop 2.x provides an upgrade to Hadoop 1.x in terms of resource management, scheduling and the manner in which execution occurs. In Hadoop 2.x the cluster resource management capabilities work in isolation from the MapReduce specific programming logic. This helps Hadoop to share resources dynamically between multiple parallel processing frameworks like Impala and the core MapReduce component. Hadoop 2.x Hadoop 2.x allows workable and fine grained resource configuration leading to efficient and better cluster utilization so that the application can scale to process larger number of jobs.

8) Differentiate between NFS, Hadoop NameNode and JournalNode.

HDFS is a write once file system so a user cannot update the files once they exist either they can read or write to it. However, under certain scenarios in the enterprise environment like file uploading, file downloading, file browsing or data streaming –it is not possible to achieve all this using the standard HDFS. This is where a distributed file system protocol Network File System (NFS) is used. NFS allows access to files on remote machines just similar to how local file system is accessed by applications.

Namenode is the heart of the HDFS file system that maintains the metadata and tracks where the file data is kept across the Hadoop cluster.

StandBy Nodes and Active Nodes communicate with a group of light weight nodes to keep their state synchronized. These are known as Journal Nodes.

9) What are the modules that constitute the Apache Hadoop 2.0 framework?

Hadoop 2.0 contains four important modules of which 3 are inherited from Hadoop 1.0 and a new module YARN is added to it.

1. Hadoop Common – This module consists of all the basic utilities and libraries that required by other modules.
2. HDFS- Hadoop Distributed file system that stores huge volumes of data on commodity machines across the cluster.
3. MapReduce- Java based programming model for data processing.
4. YARN- This is a new module introduced in Hadoop 2.0 for cluster resource management and job

scheduling.

CLICK HERE to read more about the YARN module in Hadoop 2.x.

10) How is the distance between two nodes defined in Hadoop?

Measuring bandwidth is difficult in Hadoop so network is denoted as a tree in Hadoop. The distance between two nodes in the tree plays a vital role in forming a Hadoop cluster and is defined by the network topology and java interface DNStoSwitchMapping. The distance is equal to the sum of the distance to the closest common ancestor of both the nodes. The method getDistance(Node node1, Node node2) is used to calculate the distance between two nodes with the assumption that the distance from a node to its parent node is always 1.

We have further categorized Hadoop YARN Interview Questions for Freshers and Experienced-

- Hadoop Interview Questions and Answers for Freshers - Q.Nos- 2,3,4,6,7,9
- Hadoop Interview Questions and Answers for Experienced - Q.Nos- 1,5,8,10

What other questions do you have regarding your Hadoop career?

Enter your name here...

Write your answer here...

SUBMIT

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, DeZyre 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.

Stay Tuned to the blog for more updates on Hadoop Interview FAQ's!!!

We hope that these Hadoop Interview Questions and Answers have pre-charged you for your next Hadoop Interview.Get the Ball Rolling and share your hadoop interview experiences in the comments below.Please do! It's all part of our shared mission to ease Hadoop Interviews for all prospective Hadoopers.We invite you to get involved.

Click here to know more about our IBM Certified Hadoop Developer course

Related Posts

How much Java is required to learn Hadoop?

Top 50 Hadoop Interview Questions for 2016

Top Hadoop Admin Interview Questions and Answers for 2016

Hadoop Developer Interview Questions at Top Tech Companies

PREVIOUS NEXT

Answers
Currently have 55 answers

Q: What companies are you applying to for Hadoop job roles?

Anonymous uhg
Jun 23 2016, 03:03 AM

CTS Big data support
Jun 22 2016, 12:21 AM

Nilesh facebook
Jun 09 2016, 03:18 PM

rajendra babu moka twitter
Jun 08 2016, 11:49 AM

Kandaswamy ebay
Jun 06 2016, 07:58 PM

View 20 more answers

Q: What technologies are you working on currently? (Java, Datawarehouse, Business Intelligence, ETL, etc.)

Alka JAVA
Jun 23 2016, 07:10 PM

Ankita sahoo ETL Abinitio
Jun 23 2016, 06:36 PM

C sdfsa
Jun 17 2016, 03:16 PM

gurjinder BI,ETL
Jun 15 2016, 05:04 PM

santosh jain ETL
Jun 14 2016, 04:12 PM

View 23 more answers

Q: What other questions do you have regarding your Hadoop career?

Anonymous Do I really need Java to sit for these interviews since I am not so proficient in this?
Apr 15 2016, 03:25 PM

Anonymous What hands-on projects should I be working on to get the right jobs.
Apr 15 2016, 03:25 PM

Follow