Here are frequently asked data engineer interview questions Part-2 for freshers as well as experienced candidates to get the right job.

(MADE BY RISHABH PANDEY)

SET OF 57 QUESTIONS

1) Explain Data Engineering.

Data engineering is a term used in big data. It focuses on the application of data collection and research. The data generated from various sources are just raw data. Data engineering helps to convert this raw data into useful information.

2) What is Data Modelling?

Data modeling is the method of documenting complex software design as a diagram so that anyone can easily understand. It is a conceptual representation of data objects that are associated between various data objects and the rules.

3) List various types of design schemas in Data Modelling

There are mainly two types of schemas in data modeling: 1) Star schema and 2) Snowflake schema.

4) Explain all components of a Hadoop application

Following are the components of Hadoop application:

Hadoop Common: It is a common set of utilities and libraries that are utilized by Hadoop.

HDFS: This Hadoop application relates to the file system in which the Hadoop data is stored. It is a distributed file system having high bandwidth.

Hadoop MapReduce: It is based according to the algorithm for the provision of large-scale data processing.

Hadoop YARN: It is used for resource management within the Hadoop cluster. It can also be used for task scheduling for users.

6) What is NameNode?

It is the centerpiece of HDFS. It stores data of HDFS and tracks various files across the clusters. Here, the actual data is not stored. The data is stored in DataNodes.

7) Define Hadoop streaming

It is a utility which allows for the creation of the map and Reduces jobs and submits them to a specific cluster.

8) What is the full form of HDFS?

HDFS stands for Hadoop Distributed File System.

9) Define Block and Block Scanner in HDFS

Blocks are the smallest unit of a data file. Hadoop automatically splits huge files into small pieces.

Block Scanner verifies the list of blocks that are presented on a DataNode.

10) What are the steps that occur when Block Scanner detects a corrupted data block?

Following are the steps that occur when Block Scanner find a corrupted data block:

1) First of all, when Block Scanner find a corrupted data block, DataNode report to NameNode

2) NameNode start the process of creating a new replica using a replica of the corrupted block.

3) Replication count of the correct replicas tries to match with the replication factor. If the match found corrupted data block will not be deleted.

11) Name two messages that NameNode gets from DataNode?

There are two messages which NameNode gets from DataNode. They are 1) Block report and 2) Heartbeat.

12) List out various XML configuration files in Hadoop?

There are five XML configuration files in Hadoop:

Mapred-site

Core-site

HDFS-site

Yarn-site

13) What are four V's of big data?

Four V's of big data are:

Velocity

Variety

Volume

Veracity

14) Explain the features of Hadoop

Important features of Hadoop are:

It is an open-source framework that is available freeware.

Hadoop is compatible with the many types of hardware and easy to access new hardware within a specific node.

Hadoop supports faster-distributed processing of data.

It stores the data in the cluster, which is independent of the rest of the operations.

Hadoop allows creating 3 replicas for each block with different nodes.

15) Explain the main methods of Reducer

setup (): It is used for configuring parameters like the size of input data and distributed cache.

cleanup(): This method is used to clean temporary files.

reduce(): It is a heart of the reducer which is called once per key with the associated reduced task

16) What is the abbreviation of COSHH?

The abbreviation of COSHH is Classification and Optimization based Schedule for Heterogeneous Hadoop systems.

17) Explain Star Schema

Star Schema or Star Join Schema is the simplest type of Data Warehouse schema. It is known as star schema because its structure is like a star. In the Star schema, the center of the star may have one fact table and multiple associated dimension table. This schema is used for querying large data sets.

18) How to deploy a big data solution?

Follow the following steps in order to deploy a big data solution.

1) Integrate data using data sources like RDBMS, SAP, MySQL, Salesforce

2) Store data extracted data in either NoSQL database or HDFS.

3) Deploy big data solution using processing frameworks like Pig, Spark, and MapReduce.

19) Explain FSCK

File System Check or FSCK is command used by HDFS. FSCK command is used to check inconsistencies and problem in file.

20) Explain Snowflake Schema

A Snowflake Schema is an extension of a Star Schema, and it adds additional dimensions. It is so-called as snowflake because its diagram looks like a Snowflake. The dimension tables are normalized, that splits data into additional tables.

21) Explain Hadoop distributed file system

Hadoop works with scalable distributed file systems like S3, HFTP FS, FS, and HDFS. Hadoop Distributed File System is made on the Google File System. This file system is designed in a way that it can easily run on a large cluster of the computer system.

22) Explain the main responsibilities of a data engineer

Data engineers have many responsibilities. They manage the source system of data. Data engineers simplify complex data structure and prevent the reduplication of data. Many times they also provide ELT and data transformation.

23) What is the full form of YARN?

The full form of YARN is Yet Another Resource Negotiator.

24) List various modes in Hadoop

Modes in Hadoop are 1) Standalone mode 2) Pseudo distributed mode 3) Fully distributed mode.

25) How to achieve security in Hadoop?

Perform the following steps to achieve security in Hadoop:

1) The first step is to secure the authentication channel of the client to the server. Provide time-stamped to the client.

2) In the second step, the client uses the received time-stamped to request TGS for a service ticket.

3) In the last step, the client use service ticket for self-authentication to a specific server.

26) What is Heartbeat in Hadoop?

In Hadoop, NameNode and DataNode communicate with each other. Heartbeat is the signal sent by DataNode to NameNode on a regular basis to show its presence.

27) What is Big Data?

It is a large amount of structured and unstructured data, that cannot be easily processed by traditional data storage methods. Data engineers are using Hadoop to manage big data.

28) What is FIFO scheduling?

It is a Hadoop Job scheduling algorithm. In this FIFO scheduling, a reporter selects jobs from a work queue, the oldest job first.

29) Mention default port numbers on which task tracker, NameNode, and job tracker run in Hadoop

Default port numbers on which task tracker, NameNode, and job tracker run in Hadoop are as follows:

Task tracker runs on 50060 port

NameNode runs on 50070 port

Job Tracker runs on 50030 port

33) How to disable Block Scanner on HDFS Data Node

In order to disable Block Scanner on HDFS Data Node, set dfs.datanode.scan.period.hours to 0.

30) How to define the distance between two nodes in Hadoop?

The distance is equal to the sum of the distance to the closest nodes. The method getDistance() is used to calculate the distance between two nodes.

31) Why use commodity hardware in Hadoop?

Commodity hardware is easy to obtain and affordable. It is a system that is compatible with Windows, MS-DOS, or Linux.

32) Define replication factor in HDFS

Replication factor is a total number of replicas of a file in the system.

33) What data is stored in NameNode?

Namenode stores the metadata for the HDFS like block information, and namespace information.

34) What do you mean by Rack Awareness?

In Haddop cluster, Namenode uses the Datanode to improve the network traffic while reading or writing any file that is closer to the nearby rack to Read or Write request. Namenode maintains the rack id of each DataNode to achieve rack information. This concept is called as Rack Awareness in Hadoop.

35) What are the functions of Secondary NameNode?

Following are the functions of Secondary NameNode:

FsImage which stores a copy of EditLog and FsImage file.

NameNode crash: If the NameNode crashes, then Secondary NameNode's FsImage can be used to recreate the NameNode.

Checkpoint: It is used by Secondary NameNode to confirm that data is not corrupted in HDFS.

Update: It automatically updates the EditLog and FsImage file. It helps to keep FsImage file on Secondary NameNode updated.

36) What happens when NameNode is down, and the user submits a new job?

NameNode is the single point of failure in Hadoop so the user can not submit a new job cannot execute. If the NameNode is down, then the job may fail, due to this user needs to wait for NameNode to restart before running any job.

37) What are the basic phases of reducer in Hadoop?

There are three basic phases of a reducer in Hadoop:

1. Shuffle: Here, Reducer copies the output from Mapper.

2. Sort: In sort, Hadoop sorts the input to Reducer using the same key.

3. Reduce: In this phase, output values associated with a key are reduced to consolidate the data into the final output.

38) Why Hadoop uses Context object?

Hadoop framework uses Context object with the Mapper class in order to interact with the remaining system. Context object gets the system configuration details and job in its constructor.

We use Context object in order to pass the information in setup(), cleanup() and map() methods. This object makes vital information available during the map operations.

39) Define Combiner in Hadoop

It is an optional step between Map and Reduce. Combiner takes the output from Map function, creates key value pairs, and submit to Hadoop Reducer. Combiner's task is to summarize the final result from Map into summary records with an identical key.

40) What is the default replication factor available in HDFS What it indicates?

Default replication factor in available in HDFS is three. Default replication factor indicates that there will be three replicas of each data.

41) What do you mean Data Locality in Hadoop?

In a Big Data system, the size of data is huge, and that is why it does not make sense to move data across the network. Now, Hadoop tries to move computation closer to data. This way, the data remains local to the stored location.

42) Define Balancer in HDFS

In HDFS, the balancer is an administrative used by admin staff to rebalance data across DataNodes and moves blocks from overutilized to underutilized nodes.

43) Explain Safe mode in HDFS

It is a read-only mode of NameNode in a cluster. Initially, NameNode is in Safemode. It prevents writing to file-system in Safemode. At this time, it collects data and statistics from all the DataNodes.

44) What is the importance of Distributed Cache in Apache Hadoop?

Hadoop has a useful utility feature so-called Distributed Cache which improves the performance of jobs by caching the files utilized by applications. An application can specify a file for the cache using JobConf configuration.

Hadoop framework makes replica of these files to the nodes one which a task has to be executed. This is done before the execution of task starts. Distributed Cache supports the distribution of read only files as well as zips, and jars files.

45) What is Metastore in Hive?

It stores schema as well as the Hive table location.

Hive table defines, mappings, and metadata that are stored in Metastore. This can be stored in RDBMS supported by JPOX.

46) What do mean by SerDe in Hive?

SerDe is a short name for Serializer or Deserializer. In Hive, SerDe allows to read data from table to and write to a specific field in any format you want.

47) List components available in Hive data model

There are the following components in the Hive data model:

Tables

Partitions

Buckets

48) Explain the use of Hive in Hadoop eco-system.

Hive provides an interface to manage data stored in Hadoop eco-system. Hive is used for mapping and working with HBase tables. Hive queries are converted into MapReduce jobs in order to hide the complexity associated with creating and running MapReduce jobs.

49) List various complex data types/collection are supported by Hive

Hive supports the following complex data types:

Map

Struct

Array

Union

50) Explain how .hiverc file in Hive is used?

In Hive, .hiverc is the initialization file. This file is initially loaded when we start Command Line Interface (CLI) for Hive. We can set the initial values of parameters in .hiverc file.

51) Is it possible to create more than one table in Hive for a single data file?

Yes, we can create more than one table schemas for a data file. Hive saves schema in Hive Metastore. Based on this schema, we can retrieve dissimilar results from same Data.

52) Explain different SerDe implementations available in Hive

There are many SerDe implementations available in Hive. You can also write your own custom SerDe implementation. Following are some famous SerDe implementations:

OpenCSVSerde

RegexSerDe

DelimitedJSONSerDe

ByteStreamTypedSerDe

53) List table generating functions available in Hive

Following is a list of table generating functions:

Explode(array)

JSON_tuple()

Stack()

Explode(map)

53) What is a Skewed table in Hive?

A Skewed table is a table that contains column values more often. In Hive, when we specify a table as SKEWED during creation, skewed values are written into separate files, and remaining values go to another file.

54) List out objects created by create statement in MySQL.

Objects created by create statement in MySQL are as follows:

Database

Index

Table

User

Procedure

Trigger

Event

View

Function

55) How to see the database structure in MySQL?

In order to see database structure in MySQL, you can use

DESCRIBE command. Syntax of this command is DESCRIBE Table name;.

56) How to search for a specific String in MySQL table column?

Use regex operator to search for a String in MySQL column. Here, we can also define various types of regular expression and search for using regex.

57) Explain how data analytics and big data can increase company revenue?

Following are the ways how data analytics and big data can increase company revenue:

Use data efficiently to make sure that business growth.

Increase customer value.

Turning analytical to improve staffing levels forecasts.

Cutting down the production cost of the organizations.