	codewitharrays.in freelance project available to buy contact on 8007592194	
SR.NO	Project NAME	Technology
1	Online E-Learning Platform Hub	React+Springboot+MySql
2	PG Mates / RoomSharing / Flat Mates	React+Springboot+MySql
3	Tour and Travel management System	React+Springboot+MySql
4	Election commition of India (online Voting System)	React+Springboot+MySql
5	HomeRental Booking System	React+Springboot+MySql
6	Event Management System	React+Springboot+MySql
7	Hotel Management System	React+Springboot+MySql
8	Agriculture web Project	React+Springboot+MySql
9	AirLine Reservation System / Flight booking System	React+Springboot+MySql
10	E-commerce web Project	React+Springboot+MySql
11	Hospital Management System	React+Springboot+MySql
12	E-RTO Driving licence portal	React+Springboot+MySql
13	Transpotation Services portal	React+Springboot+MySql
14	Courier Services Portal / Courier Management System	React+Springboot+MySql
15	Online Food Delivery Portal	React+Springboot+MySql
16	Muncipal Corporation Management	React+Springboot+MySql
17	Gym Management System	React+Springboot+MySql
18	Bike/Car ental System Portal	React+Springboot+MySql
19	CharityDonation web project	React+Springboot+MySql
20	Movie Booking System	React+Springboot+MySql

	freelance_Project available to buy contact on 8007592194	
21	Job Portal web project	React+Springboot+MySql
22	LIC Insurance Portal	React+Springboot+MySql
23	Employee Management System	React+Springboot+MySql
24	Payroll Management System	React+Springboot+MySql
25	RealEstate Property Project	React+Springboot+MySql
26	Marriage Hall Booking Project	React+Springboot+MySql
27	Online Student Management portal	React+Springboot+MySql
28	Resturant management System	React+Springboot+MySql
29	Solar Management Project	React+Springboot+MySql
30	OneStepService LinkLabourContractor	React+Springboot+MySql
31	Vehical Service Center Portal	React+Springboot+MySql
32	E-wallet Banking Project	React+Springboot+MySql
33	Blogg Application Project	React+Springboot+MySql
34	Car Parking booking Project	React+Springboot+MySql
35	OLA Cab Booking Portal	React+NextJs+Springboot+MySql
36	Society management Portal	React+Springboot+MySql
37	E-College Portal	React+Springboot+MySql
38	FoodWaste Management Donate System	React+Springboot+MySql
39	Sports Ground Booking	React+Springboot+MySql
40	BloodBank mangement System	React+Springboot+MySql

41	Bus Tickit Booking Project	React+Springboot+MySql
42	Fruite Delivery Project	React+Springboot+MySql
43	Woodworks Bed Shop	React+Springboot+MySql
44	Online Dairy Product sell Project	React+Springboot+MySql
45	Online E-Pharma medicine sell Project	React+Springboot+MySql
46	FarmerMarketplace Web Project	React+Springboot+MySql
47	Online Cloth Store Project	React+Springboot+MySql
48	Train Ticket Booking Project	React+Springboot+MySql
49	Quizz Application Project	JSP+Springboot+MySql
50	Hotel Room Booking Project	React+Springboot+MySql
51	Online Crime Reporting Portal Project	React+Springboot+MySql
52	Online Child Adoption Portal Project	React+Springboot+MySql
53	online Pizza Delivery System Project	React+Springboot+MySql
54	Online Social Complaint Portal Project	React+Springboot+MySql
55	Electric Vehical management system Project	React+Springboot+MySql
56	Online mess / Tiffin management System Project	React+Springboot+MySql
57		React+Springboot+MySql
58		React+Springboot+MySql
59		React+Springboot+MySql
60		React+Springboot+MySql

Spring Boot + React JS + MySQL Project List

Sr.No	Project Name	YouTube Link
1	Online E-Learning Hub Platform Project	https://youtu.be/KMjyBaWmgzg?si=YckHuNzs7eC84-IW
2	PG Mate / Room sharing/Flat sharing	https://youtu.be/4P9clHg3wvk?si=4uEsi0962CG6Xodp
3	Tour and Travel System Project Version 1.0	https://youtu.be/-UHOBywHaP8?si=KHHfE_A0uv725f12
4	Marriage Hall Booking	https://youtu.be/VXz0kZQi5to?si=llOS-QG3TpAFP5k7
5	Ecommerce Shopping project	https://youtu.be/vJ_C6LkhrZ0?si=YhcBylSErvdn7paq
6	Bike Rental System Project	https://youtu.be/FIzsAmIBCbk?si=7ujQTJqEgkQ8ju2H
7	Multi-Restaurant management system	https://youtu.be/pvV-pM2Jf3s?si=PgvnT-yFc8ktrDxB
8	Hospital management system Project	https://youtu.be/IynIouBZvY4?si=CXzQs3BsRkjKhZCw
9	Municipal Corporation system Project	https://youtu.be/cVMx9NVyI4I?si=qX0oQt-GT-LR_5jF
10	Tour and Travel System Project version 2.0	https://youtu.be/_4u0mB9mHXE?si=gDiAhKBowi2gNUKZ

Sr.No	Project Name	YouTube Link
11	Tour and Travel System Project version 3.0	https://youtu.be/Dm7nOdpasWg?si=P_Lh2gcOFhlyudug
12	Gym Management system Project	https://youtu.be/J8_7Zrkg7ag?si=LcxV51ynfUB7OptX
13	Online Driving License system Project	https://youtu.be/3yRzsMs8TLE?si=JRI_z4FDx4Gmt7fn
14	Online Flight Booking system Project	https://youtu.be/m755rOwdk8U?si=HURvAY2VnizIyJlh
15	Employee management system project	https://youtu.be/lD1iE3W_GRw?si=Y_jv1xV_BljhrD0H
16	Online student school or college portal	https://youtu.be/4A25aEKfei0?si=RoVgZtxMk9TPdQvD
17	Online movie booking system project	https://youtu.be/Lfjv_U74SC4?si=fiDvrhhrjb4KSlSm
18	Online Pizza Delivery system project	https://youtu.be/Tp3izreZ458?si=8eWAOzA8SVdNwlyM
19	Online Crime Reporting system Project	https://youtu.be/0UlzReSk9tQ?si=6vN0e70TVY1GOwPO
20	Online Children Adoption Project	https://youtu.be/3T5HC2HKyT4?si=bntP78niYH802I7N

Hadoop Admin Interview Questions and Answers

1. RDBMS vs Hadoop?

Name	RDBMS	Hadoop
Data volume	RDBMS cannot store and process a large amount of data	Hadoop works better for large amounts of data. It can easily store and process a large amount of data compared to RDBMS.
Throughput	RDBMS fails to achieve a high Throughput	Hadoop achieves high Throughput
Data variety	Schema of the data is known in RDBMS and it always depends on the structured data.	It stores any kind of data. Whether it could be structured, unstructured, or semi-structured.
Data processing	RDBMS supports OLTP(Online Transactional Processing)	Hadoop supports OLAP(Online Analytical Processing)
Read/Write Speed	Reads are fast in RDBMS because the schema of the data is already known.	Writes are fast in Hadoop because no schema validation happens during HDFS write.
Schema on reading Vs Write	RDBMS follows schema on write policy	Hadoop follows the schema on reading policy
Cost	RDBMS is a licensed software	Hadoop is a free and open-source framework

2. Explain Big data and its characteristics?

Big Data refers to a large amount of data that exceeds the processing capacity of conventional database systems and requires a special parallel processing mechanism. This data can be either structured or unstructured data.

Characteristics of Big Data:

- **Volume** - It represents the amount of data that is increasing at an exponential rate i.e. in gigabytes, Petabytes, Exabytes, etc.
- **Velocity** - Velocity refers to the rate at which data is generated, modified, and processed. At present, Social media is a major contributor to the velocity of growing data.
- **Variety** - It refers to data coming from a variety of sources like audios, videos, CSV, etc. It can be either structured, unstructured, or semi-structured.
- **Veracity** - Veracity refers to imprecise or uncertain data.
- **Value** - This is the most important element of big data. It includes data on how to access and deliver quality analytics to the organization. It provides a fair market value on the used technology.

3. What is Hadoop and list its components?

Hadoop is an open-source framework used for storing large data sets and runs applications across clusters of commodity hardware.

It offers extensive storage for any type of data and can handle endless parallel tasks.

Core components of Hadoop:

- Storage unit– HDFS (DataNode, NameNode)
- Processing framework– YARN (NodeManager, ResourceManager).

[Learn Complete Hadoop Tutorial]

4. What is YARN and explain its components?

Yet Another Resource Negotiator (YARN) is one of the core components of Hadoop and is responsible for managing resources for the various applications operating in a Hadoop cluster, and also schedules tasks on different cluster nodes.

YARN components:

- **Resource Manager -** It runs on a master daemon and controls the resource allocation in the cluster.
- **Node Manager -** It runs on a slave daemon and is responsible for the execution of tasks for each single Data Node.
- **Application Master -** It maintains the user job lifecycle and resource requirements of individual applications. It operates along with the Node Manager and controls the execution of tasks.
- **Container -** It is a combination of resources such as Network, HDD, RAM, CPU, etc., on a single node.

5. What is the difference between a regular file system and HDFS?

Regular File Systems	HDFS
A small block size of data (like 512 bytes)	Large block size (orders of 64MB)
Multiple disks seek large files	Reads data sequentially after single seek

6. What are the Hadoop daemons and explain their roles in a Hadoop cluster?

Generally, the daemon is nothing but a process that runs in the background. Hadoop has five such daemons. They are:

- **NameNode -** Is is the Master node responsible to store the meta-data for all the directories and files.
- **DataNode -** It is the Slave node responsible to store the actual data.

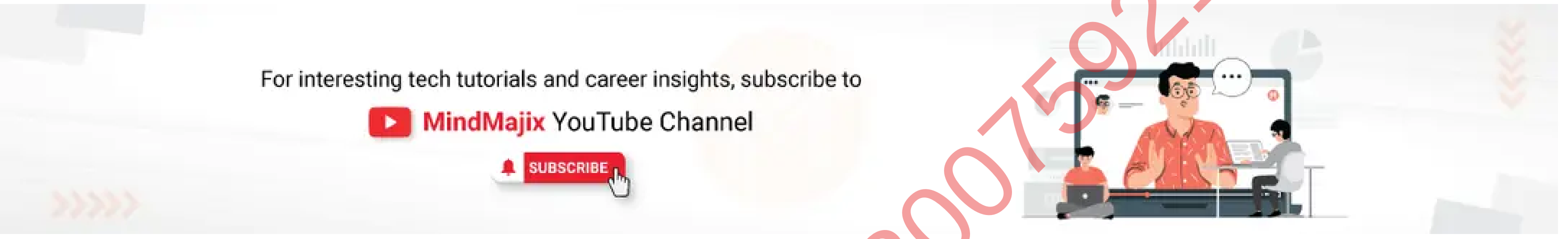
- **Secondary NameNode -** It is responsible for the backup of NameNode and stores the entire metadata of data nodes like data node properties, addresses, and block reports of each data node.
- **JobTracker -** It is used for creating and running jobs. It runs on data nodes and allocates the job to TaskTracker.
- **TaskTracker -** It operates on the data node. It runs the tasks and reports the tasks to JobTracker.

7. What is Avro Serialization in Hadoop?

- The process of translating objects or data structures state into binary or textual form is called Avro Serialization. It is defined as a language-independent schema (written in JSON).
- It provides AvroMapper and AvroReducer for running MapReduce programs.

8. How can you skip the bad records in Hadoop?

Hadoop provides a feature called SkipBadRecords class for skipping bad records while processing mapping inputs.



Hadoop HDFS Interview Questions and Answers

9. Explain HDFS and its components?

- HDFS (Hadoop Distributed File System) is the primary data storage unit of Hadoop.
- It stores various types of data as blocks in a distributed environment and follows master and slave topology.

HDFS components:

- **NameNode -** It is the master node and is responsible for maintaining the metadata information for the blocks of data stored in HDFS. It manages all the DataNodes.

  Ex: replication factors, block location, etc.

- **DataNode -** It is the slave node and responsible for storing data in the HDFS.

10. What are the features of HDFS?

- Supports storage of very large datasets
- Write once read many access model
- Streaming data access
- Replication using commodity hardware
- HDFS is highly Fault Tolerant
- Distributed Storage.

11. What is the HDFS block size?

By default, the HDFS block size is 128MB for Hadoop 2.x.

12. What is the default replication factor?

- Replication factor means the minimum number of times the file will replicate(copy) across the cluster.
- The default replication factor is 3.

13. List the various HDFS Commands?

The Various HDFS Commands are listed below:

- version
- mkdir
- ls
- put
- copy from local
- get
- copyToLocal
- cat
- mv
- cp.

14. Compare HDFS (Hadoop Distributed File System) and NAS (Network Attached Storage)?

HDFS	NAS
It is a distributed file system used for storing data by commodity hardware.	It is a file-level computer data storage server connected to a computer network, provides network access to a heterogeneous group of clients.
It includes commodity hardware which will be cost-effective	NAS is a high-end storage device that includes a high cost.
It is designed to work for the MapReduce paradigm.	It is not suitable for MapReduce.

15. What are the limitations of Hadoop 1.0?

- NameNode: No Horizontal Scalability and No High Availability
- Job Tracker: Overburdened.
- MRv1: It can only understand Map and Reduce tasks.

16. How to commission (adding) the nodes in the Hadoop cluster?

- Update the network addresses in the dfs.include and mapred.include
- Update the NameNode: Hadoop dfsadmin -refreshNodes
- Update the Jobtracker: Hadoop mradmin-refreshNodes
- Update the slave file.
- Start the DataNode and NodeManager on the added Node.

[Check out Hadoop Archive Files in HDFS]

17. How to decommission (removing) the nodes in the Hadoop cluster?

- Update the network addresses in the dfs.exclude and mapred.exclude
- Update the Namenode: $ Hadoop dfsadmin -refreshNodes
- Update the JobTracker: Hadoop mradmin -refreshNodes
- Cross-check the Web UI it will show “Decommissioning in Progress”
- Remove the Nodes from the include file and then run: Hadoop dfsadmin-refreshNodes, Hadoop mradmin -refreshNodes.
- Remove the Nodes from the slave file.

18) Compare Hadoop 1.x and Hadoop 2.x

Name	Hadoop 1.x	Hadoop 2.x
1. NameNode	In Hadoop 1.x, NameNode is the single point of failure	In Hadoop 2.x, we have both Active and passive NameNodes.
2. Processing	MRV1 (Job Tracker & Task Tracker)	MRV2/YARN (ResourceManager & NodeManager)

19. What is the difference between active and passive NameNodes?

- Active NameNode works and runs in the cluster.
- Passive NameNode has similar data as active NameNode and replaces it when it fails.

20. How will you resolve the NameNode failure issue?

The following steps need to be executed to resolve the NameNode issue and make the Hadoop cluster up and running:

- Use the FsImage (file system metadata replica) to start a new NameNode.
- Now, configure DataNodes and clients, so that they can acknowledge the new NameNode, that is started.
- The new NameNode will start serving the client once it has completed loading the last checkpoint FsImage and enough block reports from the DataNodes.

21. What is a Checkpoint Node in Hadoop?

Checkpoint Node is the new implementation of secondary NameNode in Hadoop. It periodically creates the checkpoints of filesystem metadata by merging the edits log file with FsImage file.

22. List the different types of Hadoop schedulers.

- Hadoop FIFO scheduler
- Hadoop Fair Scheduler
- Hadoop Capacity Scheduler.

23. How to keep an HDFS cluster balanced?

However, it is not possible to limit a cluster from becoming unbalanced. In order to give a balance to a certain threshold among data nodes, use the Balancer tool. This tool tries to subsequently even out the block data distribution across the cluster.

24. What is DistCp?

- **DistCp** is the tool used to copy large amounts of data to and from Hadoop file systems in parallel.
- It uses MapReduce to effect its distribution, reporting, recovery, and error handling.

25. What is HDFS Federation?

- HDFS Federation enhances the present HDFS architecture through a clear separation of namespace and storage by enabling a generic block storage layer.
- It provides multiple namespaces in the cluster to improve scalability and isolation.

26. What is HDFS High Availability?

HDFS High availability is introduced in Hadoop 2.0. It means providing support for multiple NameNodes to the Hadoop architecture.

27. What is a rack-aware replica placement policy?

- Rack Awareness is the algorithm used for improving the network traffic while reading/writing HDFS files to the Hadoop cluster by NameNode.
- NameNode chooses the Datanode which is closer to the same rack or nearby rack for reading/Write request. The concept of choosing closer data nodes based on racks information is called Rack Awareness.
- Consider the replication factor is 3 for data blocks on HDFS it means for every block of data two copies are stored on the same rack, while the third copy is stored on a different rack. This rule is called Replica Placement Policy.

28. What is the main purpose of the Hadoop fsck command?

Hadoop fsck command is used for checking the HDFS file system.

There are different arguments that can be passed with this command to emit different results.

- **Hadoop fsck / -files:** Displays all the files in HDFS while checking.
- **Hadoop fsck / -files -blocks:** Displays all the blocks of the files while checking.
- **Hadoop fsck / -files -blocks -locations:** Displays all the files block locations while checking.
- **Hadoop fsck / -files -blocks -locations -racks:** Displays the networking topology for data-node locations.
- **Hadoop fsck -delete:** Deletes the corrupted files in HDFS.
- **Hadoop fsck -move:** Moves the corrupted files to a particular directory.

29. What is the purpose of a DataNode block scanner?

- The purpose of the DataNode block scanner is to operate and periodically check all the blocks that are stored on the DataNode.
- If bad blocks are detected they will be fixed before any client reads.

30. What is the purpose of the admin tool?

- dfsadmin tool is used for examining the HDFS cluster status.

- dfsadmin – report command produces useful information about basic statistics of the cluster such as DataNodes and NameNode status, disk capacity configuration, etc.
- It performs all the administrative tasks on the HDFS.

31. What is the command used for printing the topology?

.hdfs dfsadmin -point topology is used for printing the topology. It displays the tree of racks and DataNodes attached to the tracks.

32. What is RAID?

RAID (redundant array of independent disks) is a data storage virtualization technology used for improving performance and data redundancy by combining multiple disk drives into a single entity.

33. Does Hadoop requires RAID?

- In DataNodes, RAID is not necessary as storage is achieved by replication between the Nodes.
- In NameNode's disk RAID is recommended.

34. List the various site-specific configuration files available in Hadoop?

- conf/Hadoop-env.sh
- conf/yarn-site.xml
- conf/yarn-env.sh
- conf/mapred-site.xml
- conf/hdfs-site.xml
- conf/core-site.xml.

[Check out** Installation and Configuration in Hadoop **]

35. What is the main functionality of NameNode?

It is mainly responsible for:

- Namespace - Manages metadata of HDFS.
- Block Management - Processes and manages the block reports and their location.

36. Which command is used to format the NameNode?

```
$ hdfs namenode -format
```

37. How a client application interacts with the NameNode?

- Client applications associate the Hadoop HDFS API with the NameNode when it has to copy/move/add/locate/delete a file.
- The NameNode returns to the successful requests by delivering a list of relevant DataNode servers where the data is residing.
- The client can talk directly to a DataNode after the NameNode has given the location of the data

Hadoop MapReduce Interview Questions

38. What is MapReduce and list its features?

MapReduce is a programming model used for processing and generating large datasets on the clusters with parallel and distributed algorithms.

The syntax for running the MapReduce program is

```
hadoop_jar_file.jar /input_path /output_path.
```

39. What are the features of MapReduce?

- Automatic parallelization and distribution.
- Built-in fault tolerance and redundancy are available.
- MapReduce Programming model is language independent
- Distributed programming complexity is hidden
- Enable data local processing
- Manages all the inter-process communication.

40. What does the MapReduce framework consist of?

MapReduce framework is used to write applications for processing large data in parallel on large clusters of commodity hardware.

It consists of:

ResourceManager (RM)

- Global resource scheduler
- One master RM.

NodeManager (NM)

- One slave NM per cluster node.

Container

- RM creates Containers upon request by AM
- The application runs in one or more containers.

ApplicationMaster (AM)

- One AM per application
- Runs in Container.

41. What are the two main components of ResourceManager?

- **Scheduler**

It allocates the resources (containers) to various running applications based on resource availability and configured shared policy.

- **ApplicationManager**

It is mainly responsible for managing a collection of submitted applications.

[Check out MapReduce Implementation in Hadoop]

42. What is a Hadoop counter?

Hadoop Counters measures the progress or tracks the number of operations that occur within a MapReduce job. Counters are useful for collecting statistics about MapReduce jobs for application-level or quality control.

43) What are the main configuration parameters for a MapReduce application?

The job configuration requires the following:

- Job’s input and output locations in the distributed file system
- The input format of data
- The output format of data
- Class containing the map function and reduced function
- JAR file containing the reducer, driver, and mapper classes.

44. What are the steps involved to submit a Hadoop job?

Steps involved in Hadoop job submission:

- Hadoop job client submits the job jar/executable and configuration to the ResourceManager.
- ResourceManager then distributes the software/configuration to the slaves.
- ResourceManager then scheduling tasks and monitoring them.
- Finally, job status and diagnostic information are provided to the client.

45. How does the MapReduce framework view its input internally?

It views the input data set as a set of pairs and processes the map tasks in a completely parallel manner.

46. What are the basic parameters of Mapper?

The basic parameters of Mapper are listed below:

1. LongWritable and Text
2. Text and IntWritable.

47. What are Writables and explain their importance in Hadoop?

- Writables are interfaces in Hadoop. They act as a wrapper class to almost all the primitive data types of Java.
- A serializable object which executes a simple and efficient serialization protocol, based on DataInput and DataOutput.
- Writables are used for creating serialized data types in Hadoop.

48. Why comparison of types is important for MapReduce?

- It is important for MapReduce as in the sorting phase the keys are compared with one another.
- For a Comparison of types, the WritableComparable interface is implemented.

49. What is "speculative execution" in Hadoop?

In Apache Hadoop, if nodes do not fix or diagnose the slow-running tasks, the master node can redundantly perform another instance of the same task on another node as a backup (the backup task is called a Speculative task). This process is called Speculative Execution in Hadoop.

50. What are the methods used for restarting the NameNode in Hadoop?

The methods used for restarting the NameNodes are the following:

- You can use **/sbin/hadoop-daemon.sh stop namenode** command for stopping the NameNode individually and then start the NameNode using **/sbin/hadoop-daemon.sh start namenode.**
- Use **/sbin/stop-all.sh** and then use **/sbin/start-all.sh** command for stopping all the demons first and then start all the daemons.

These script files are stored in the sbin directory inside the Hadoop directory store.

51. What is the difference between an "HDFS Block" and "MapReduce Input Split"?

- HDFS Block is the physical division of the disk which has the minimum amount of data that can be read/write, while MapReduce InputSplit is the logical division of data created by the InputFormat specified in the MapReduce job configuration.
- HDFS divides data into blocks, whereas MapReduce divides data into input split and empowers them to mapper function.

52. What are the different modes in which Hadoop can run?

- **Standalone Mode(local mode) -** This is the default mode where Hadoop is configured to run. In this mode, all the components of Hadoop such as DataNode, NameNode, etc., run as a single Java process and useful for debugging.
- **Pseudo Distributed Mode(Single-Node Cluster) -** Hadoop runs on a single node in a pseudo-distributed mode. Each Hadoop daemon works in a separate Java process in Pseudo-Distributed Mode, while in Local mode, each Hadoop daemon operates as a single Java process.
- **Fully distributed mode (or multiple node cluster) -** All the daemons are executed in separate nodes building into a multi-node cluster in the fully-distributed mode.

53. Why aggregation cannot be performed in Mapperside?

- We cannot perform Aggregation in mapping because it requires sorting of data, which occurs only at the Reducer side.
- For aggregation, we need the output from all the mapper functions, which is not possible during the map phase as map tasks will be running in different nodes, where data blocks are present.

54. What is the importance of “RecordReader” in Hadoop?

- RecordReader in Hadoop uses the data from the InputSplit as input and converts it into Key-value pairs for Mapper.
- The MapReduce framework represents the RecordReader instance through InputFormat.

55. What is the purpose of Distributed Cache in a MapReduce Framework?

- The Purpose of Distributed Cache in the MapReduce framework is to cache files when needed by the applications. It caches read-only text files, jar files, archives, etc.
- When you have cached a file for a job, the Hadoop framework will make it available to each and every data node where map/reduces tasks are operating.

56. How do reducers communicate with each other in Hadoop?

Reducers always run in isolation and the Hadoop Mapreduce programming paradigm never allows them to communicate with each other.

57. What is Identity Mapper?

- Identity Mapper is a default Mapper class that automatically works when no Mapper is specified in the MapReduce driver class.
- It implements mapping inputs directly into the output.
- IdentityMapper.class is used as a default value when JobConf.setMapperClass is not set.

58. What are the phases of MapReduce Reducer?

The MapReduce reducer has three phases:

- **Shuffle phase -** In this phase, the sorted output from a mapper is an input to the Reducer. This framework will fetch the relevant partition of the output of all the mappers by using HTTP.
- **Sort phase -** In this phase, the input from various mappers is sorted based on related keys. This framework groups reducer inputs by keys. Shuffle and sort phases occur concurrently.
- **Reduce phase -** In this phase, reduce task aggregates the key-value pairs after shuffling and sorting phases. The OutputCollector.collect() method, writes the output of the reduce task to the Filesystem.

59. What is the purpose of MapReduce Partitioner in Hadoop?

The MapReduce Partitioner manages the partitioning of the key of the intermediate mapper output. It makes sure that all the values of a single key pass to same reducers by allowing the even distribution over the reducers.

60. How will you write a custom partitioner for a Hadoop MapReduce job?

- Build a new class that extends Partitioner Class
- Override the get partition method in the wrapper.
- Add the custom partitioner to the job as a config file or by using the method set Partitioner.

[Related Article: Hadoop MapReduce Architecture]

61. What is a Combiner?

A Combiner is a semi-reducer that executes the local reduce task. It receives inputs from the Map class and passes the output key-value pairs to the reducer class.

62. What is the use of SequenceFileInputFormat in Hadoop?

SequenceFileInputFormat is the input format used for reading in sequence files. It is a compressed binary file format optimized for passing the data between outputs of one MapReduce job to the input of some other MapReduce job.

Apache Pig Interview Questions

63. What is Apache Pig?

- Apache Pig is a high-level scripting language used for creating programs to run on Apache Hadoop.
- The language used in this platform is called Pig Latin.
- It executes Hadoop jobs in Apache Spark, MapReduce, etc.

64. What are the benefits of Apache Pig over MapReduce?

- Pig Latin is a high-level scripting language while MapReduce is a low-level data processing paradigm.
- Without many complex Java implementations in MapReduce, programmers can perform the same implementations very easily using Pig Latin.
- Apache Pig decreases the length of the code by approx 20 times (according to Yahoo). Hence, this reduces development time by almost 16 times.
- Pig offers various built-in operators for data operations like filters, joins, sorting, ordering, etc., while to perform these same functions in MapReduce is an enormous task.

65. What are the Hadoop Pig data types?

Hadoop Pig runs both atomic data types and complex data types.

- **Atomic data types:** These are the basic data types that are used in all the languages like int, string, float, long, etc.
- **Complex Data Types:** These are Bag, Map, and Tuple.

[Related Article: Hadoop Apache Pig Data Types and Syntax]

66. List the various relational operators used in “Pig Latin”?

- SPLIT
- LIMIT
- CROSS

- COGROUP
- GROUP
- STORE
- DISTINCT
- ORDER BY
- JOIN
- FILTER
- FOREACH
- LOAD.

Apache Hive Interview Questions

67. What is Apache Hive?

Apache Hive offers a database query interface to Apache Hadoop. It reads, writes, and manages large datasets that are residing in distributed storage and queries through SQL syntax.

68. Where do Hive stores table data in HDFS?

/usr/hive/warehouse is the default location where Hive stores the table data in HDFS.

[Learn Complete** Overview on Hadoop Hive **]

69. Can the default "Hive Metastore" be used by multiple users (processes) at the same time?

By default, Hive Metastore uses the Derby database. So, it is not possible for multiple users or processes to access it at the same time.

70. What is a SerDe?

SerDe is a combination of Serializer and Deserializer. It interprets the results of how a record should be processed by allowing Hive to read and write from a table.

71. What are the differences between Hive and RDBMS?

Hive	RDBMS
Schema on Reading	Schema on write
Batch processing jobs	Real-time jobs
Data stored on HDFS	Data stored on the internal structure
Processed using MapReduce	Processed using database

Apache HBase Interview Questions

72. What is an Apache HBase?

Apache HBase is multidimensional and a column-oriented key datastore runs on top of HDFS (Hadoop Distributed File System). It is designed to provide high table-update rates and a fault-tolerant way to store a large collection of sparse data sets.

73. What are the various components of Apache HBase?

- **Region Server:** These are the worker nodes that handle read, write, update, and delete requests from clients. The region Server process runs on each and every node of the Hadoop cluster
- **HMaster:** It monitors and manages the Region Server in the Hadoop cluster for load balancing.
- **ZooKeeper:** ZHBase employs ZooKeeper for a distributed environment. It keeps track of each and every region server that is present in the HBase cluster.

74. What is WAL in HBase?

- Write-Ahead Log (WAL) is a file storage and it records all changes to data in HBase. It is used for recovering data sets.
- The WAL ensures all the changes to the data can be replayed when a RegionServer crashes or becomes unavailable.

Visit here to learn Hadoop Training Online in NewYork

75. What are the differences between the Relational database and HBase?

Relational Database	HBase
It is a row-oriented datastore	It is a column-oriented datastore
It's a schema-based database	Its schema is more flexible and less restrictive
Suitable for structured data	Suitable for both structured and unstructured data
Supports referential integrity	Doesn't supports referential integrity
It includes thin tables	It includes sparsely populated tables
Accesses records from tables using SQL queries.	Accesses data from HBase tables using APIs and MapReduce.

Apache Spark Interview Questions

76. What is Apache Spark?

Apache Spark is an open-source framework used for real-time data analytics in a distributed computing environment. It is a data processing engine that provides faster analytics than Hadoop MapReduce.

77. Can we build "Spark" with any particular Hadoop version?

Yes, we can build "Spark" for any specific Hadoop version.

78. What is RDD?

RDD(Resilient Distributed Datasets) is a fundamental data structure of Spark. It is a distributed collection of objects, and each dataset in RDD is further distributed into logical partitions and computed on several nodes of the cluster

Apache ZooKeeper Interview Questions

79. What is Apache ZooKeeper?

Apache ZooKeeper is a centralized service used for managing various operations in a distributed environment. It maintains configuration data, performs synchronization, naming, and grouping.

80. What is Apache Oozie?

Apache Oozie is a scheduler that controls the workflow of Hadoop jobs.

There are two kinds of Oozie jobs:

- **Oozie Workflow -** It is a collection of actions sequenced in a control dependency DAG(Direct Acyclic Graph) for execution.
- **Oozie Coordinator -** If you want to trigger workflows based on the availability of data or time then you can use Oozie Coordinator Engine.

81. How can you configure the “Oozie” job in Hadoop?

Integrate Oozie with the Hadoop stack, which supports several types of Hadoop jobs such as Streaming MapReduce, Java MapReduce, Sqoop, Hive, and Pig.

Apache Flume Interview Questions

82. What is an Apache Flume?

- Apache Flume is a service/tool/data ingestion mechanism used to collect, aggregate, and transfer massive amounts of streaming data such as events, log files, etc., from various web sources to a centralized data store where they can be processed together.
- It is a highly reliable, distributed, and configurable tool that is specially designed to transfer streaming data to HDFS.

83. List the Apache Flume features.

- It is fault-tolerant and robust
- Scales horizontally
- Selects high volume data streams in real-time
- Streaming data is gathered from multiple sources into Hadoop for analysis.
- Ensures guaranteed data delivery.

Apache Sqoop Interview Questions

84. What is the use of Apache Sqoop in Hadoop?

Apache Sqoop is a tool particularly used for transferring massive data between Apache Hadoop and external datastores such as relational database management, enterprise data warehouses, etc.

85. Where do Hadoop Sqoop scripts are stored?

```
/usr/bin/Hadoop Sqoop
```

https://www.youtube.com/@codewitharrays

https://www.instagram.com/codewitharrays/

https://t.me/codewitharrays Group Link: https://t.me/ccee2025notes

+91 8007592194 +91 9284926333

codewitharrays@gmail.com

https://codewitharrays.in/project