<https://www.edureka.co/blog/interview-questions/top-50-hadoop-interview-questions-2016/>

<https://www.dezyre.com/article/pig-interview-questions-and-answers-for-2018/244>

<https://www.gocit.vn/files/Oreilly.Programming.Pig-www.gocit.vn.pdf>

<https://intellipaat.com/blog/interview-question/big-data-hadoop-interview-questions/>

<https://www.edureka.co/blog/interview-questions/hive-interview-questions/>

To learn :

1. Developed ETL pipelines using Spark and Kafka with average load around 1 million+ records per minute.
2. Used Spark to parse and clean log files based on custom regular expressions and load into Impala where it can be further processed.
3. Managed and scheduled ingestion process for ETL pipelines using Airflow.
4. Deployed custom configured Cloudera clusters with Amazon EC2 on AWS.
5. Optimized clusters by optimizing YARN configurations

<https://www.dezyre.com/article/how-to-ensure-best-performance-for-your-hadoop-cluster/200>

<https://www.corejavaguru.com/bigdata/hadoop/yarn-interview-questions>

<https://docs.cloudera.com/documentation/enterprise/5-6-x/topics/cdh_ig_yarn_tuning.html>

1. Implemented scripts to backup/restore cluster to amazon S3 using DistCp

<https://docs.cloudera.com/documentation/enterprise/5-8-x/topics/cdh_admin_distcp_data_cluster_migrate.html#distcp_and_s3>

1. Devised ways to separate compute and storage for replacing HDFS for Hive/Impala with Amazon S3.
2. Configured ways to store Hive meta-store and Sentry security configuration on a remote database instance to support separation of compute and storage.
3. Created build pipelines using Jenkins scripts.
4. Migrated Microsoft T-SQL queries to Apache Hive queries.
5. Exported Hive databases as targeted exports and bulk exports to SQL databases using Sqoop scripts.
6. Parsed text files using map reduce jobs to populate Hive tables.
7. Scheduled jobs using oozie manager.
8. Bulk importing of data from various data sources into Spark and transform data in flexible ways by using Apache Spark / Apache Hadoop
9. Execute work with a high level of quality.
10. Writing python scripts for automating different processes on AWS.
11. Storing and Ingestion of data in HDFS
12. Creating a largescale application with Apache Spark, Scala, Pyspark.
13. Execute work with a high level of quality.
14. Performing data visualization using AWS Quicksight
15. Handson knowledge of diferent AWS services
16. Importing and Exporting Data from AWS S3 datalake
17. Performing ETL using AWS GLUE, AWS EMR
18. Creating clusters on AWS EMR
19. Triggering and Scheduling jobs using AWS Lambda
20. Collecting log details in AWS Cloudwatch
21. Data cleaning using Data frames on spark engine using pyspark
22. Troubleshooting data cleaning using python programming
23. Used Spark API to perform analytics on data in Hive.
24. Developed analytical components using Spark , Scala and Spark Stream
25. Developed Spark scripts by using Scala Shell commands as per the requirement.
26. Manipulating Data Frames using Pandas, Numpy, Matplotlib, Seaborn
27. Applying Machine Learning Algorithm using Scikit-Learn
28. ETL Applications from Data Lake using Spark, Dataframe, Scala, Java, Kudu, Scylla, Janusgraph
29. Design and develop near real time pipeline for data ingestion from salesforce to Kafka
30. Developed Graph QL user interface
31. Analysing data with SQL and python
32. Develop an automated web scraping tool using Selenium packages in Python, Cucumber and TestNG frameworks
33. Develop first generation ETL ingestion process and monitoring (bash, scoop, pig, python, psql)
34. Collect data and move from servers to Hive/HBase using Apache Spark
35. Implement a continuous streaming bus using Kafka
36. What is avro?

Avro is a row-based storage format for Hadoop which is widely used as a serialization platform. Avro stores the data definition (schema) in JSON format making it easy to read and interpret by any program. The data itself is stored in binary format making it compact and efficient.

1. What is parquet file format?

Parquet, an open source file format for Hadoop. Parquet stores nested data structures in a flat columnar format. Compared to a traditional approach where data is stored in row-oriented approach, parquet is more efficient in terms of storage and performance

1. What is orc file format?
2. Difference between managed and external tables?

When managed table is deleted the metadata information is lost, not the case with external tables.