UNIT 2 Question Bank

1. Explain what is Sqoop in Hadoop? Please explain the usage
2. What are the components used in Hive query processor?
3. What is Bucket in Hive?
4. For each Sqoop copying into HDFS, how many MapReduce jobs and tasks will be submitted? Please explain.
5. I am having around 500 tables in a database. I want to import all the tables from the database except the tables named Table 498, Table 323 and Table 199. How can we do this without having to import the tables one by one?
6. Explain the significance of using split-by clause in Apache Sqoop.
7. I want to see the present working directory in UNIX from Hive. Is it possible to run this command from Hive?
8. What is the use of explode in Hive?
9. Is it possible to change the default location of managed tables in Hive, if so how?
10. Why do we need Hive?
11. What is partitioning? When we may need to customize the default partition? Please explain the scenario with an example.
12. If you run a select * query in Hive, why does it not run MapReduce? Please explain it.
13. What is the difference between external table and managed table?
14. Why do we perform partitioning in Hive? Please explain the advantage of it.
15. Suppose, we create a table that contains details of all the transactions done by the customers of year 2018. CREATE TABLE customer_transaction_details (cust_id INT, amount FLOAT, month STRING, country STRING) ROW FORMAT DELIMITED FIELDS TERMINATED BY ‘,’; Now, after inserting 50,000 tuples in this table, we want to know the total revenue generated for each month. But the problem is, Hive is taking too much time in processing this query. How will you solve this problem and list the steps that we will be taking in order to do so?
16. Explain the data flow in Hive with a diagram. Please describe each and every step.
17. What is the usage of Metastore in Hive. If Metastore is not present in Hive, then what will be the problem?

18. How will you update the rows that are already exported? Write Sqoop command to show all the databases in MySQL server.
19. How to create a table in MySQL and how to insert the values into the table? Please import this table into Hive/HDFS using Apache Sqoop.
20. Please explain how apache Flume works. Also please describe a flow about how to extract a log file from source path and ingest into HDFS by Flume.
21. What is the usage of Oozie and what are the main components in it. Please explain.
22. Explain the core components of Apache Flume. What is Agent and Channel?

23. Why is the application program layer different from support layer in Big Data platform? Explain the Hadoop features. (LO 2.1)
24. List Hadoop two core components. Describe their usages. (LO 2.1)
25. Explain using a diagram the distributed storage, resource manager layer, processing framework and application APIs layers in Hadoop. (LO 2.1)
26. Give the meanings of Hadoop distributed file system, clusters, Racks, DataNodes, Data Blocks, MasterNode, NameNode and metadata of files. Explain these. (LO 2.2)
27. How do multiple NameNodes ensure high availability of Data in HDFS? (LO 2.2)
28. How does MapReduce function as a programming model for distributed computing? (LO 2.3)
29. How does MapReduce enables process huge amounts of data, in parallel, on large clusters of servers reliably. (LO 2.3)
30. List the resources required to run an application. How does the separation of resource management and processing components help the number of tasks and sub-tasks (threads) when running in parallel? (LO 2.4)
31. How does YARN resource manager do the following: (i) keep track of the active node managers and available resources and (ii) allocate the containers to the appropriate sub- tasks and monitors the Application Master?
32. List and explain the features of the MapReduce programming model? How does MapReduce program enable parallel processing? (LO 4.1)
33. How does a Map task implement using key-value pairs in an input file? What are the uses of Shuffle in processing the aggregates for all the Mapper output by grouping key values of the Mapper output and the value which gets appended in a list of values? (LO 4.1)

34. How does ‘Group By’ operate for creating Mapper output? What are the roles of partitioning and combining? (LO 4.1)
35. How does MapReduce program find the distinct values and count the unique values? (LO 4.2)
36. How does the MapReduce implement the relational algebraic functions, union, projection, difference, intersection, natural join, grouping and aggregation? Explain each with an example. (LO 4.2)
37. How do MapReduce tasks implement a matrix multiplication by a vector? (LO 4.2)
38. Describe the Hive architecture components. Why are HiveQL, SQL-like scripts used in place of RDBMS, such as MySQL for Big Data? (LO 4.3)
39. Describe the Hive architecture components. Why are HiveQL, SQL-like scripts used in place of RDBMS, such as MySQL for Big Data? (LO 4.3)
40. What are the types of built-in functions available in Hive? What are the uses of each of these? (LO 4.3)
41. Why should partitions be created in databases and tables in Hive data warehouse for very large datasets? (LO 4.4)
42. What are aggregation commands provisioned in HiveQL? What are the partitioning commands? (LO 4.4)
43. What are the differences between Pig programming model with MapReduce, relational database and Hive programming models? (LO 4.5)
44. Describe Pig data types and operators: Group, Join, Filter, Limit, Order by, parallel, sort and split. (LO 4.5)
45. Describe usages of Pig operations: parallel, split and defining a UDF. Give one example of each. (LO 4.5)