UNIT III
(HADOOP)

1. What are the two major parts of Hadoop Environment?

   HDFS-Hadoop Distributed File System for storing very large file in a distributed environment.

   MapReduce – A Programming Model used to implement Map and Reduce tasks

2. What is MapReduce Job?

   A MapReduce job is a unit of work that the client wants to be performed: it consists of the input data, the MapReduce program, and configuration information. Hadoop runs the job by dividing it into tasks, of which there are two types: map tasks and reduce tasks.

3. What is JobTracker and TaskTracker?

   The jobtracker coordinates all the jobs run on the system by scheduling tasks to run on tasktrackers.

   Tasktrackers run tasks and send progress reports to the jobtracker, which keeps a record of the overall progress of each job.

4. What do you mean by scaling out?

   It is the way of adding more data or very large dataset into processing. In Hadoop, HDFS, a distributed filesystem is used to store the data. It allows the Hadoop to move the MapReduce computation to each machine hosting a part of data.

5. Define data locality optimization.

   It is the process of running map task by Hadoop on a node where input data resides in HDFS. Optimal split is the largest size of input that can be guaranteed to be stored on a single node.

6. Define Hadoop streaming

   Hadoop uses UNIX standard streaming interface between Hadoop and User program. We can use any language to read standard input and write standard output to write MapReduce program. This interface will perform transformation to perform standard I/O operations.

7. State the purpose Hadoop pipe?

   Hadoop pipe is one of the facility uses sockets implemented in c++, as the channels over which tasktracker can communicate with other.

8. What is the role of combiner function in Hadoop.

   This is a function run on the output of Map task and the output of combiner function forms input to the reduce function. It is not an alternative to reduce function, but it can help to cut down amount of data shuffled between maps and reduces.

9. Define HDFS

   Hadoop Distributed File System is a file system that manages the storage across a network of machines. It is designed for storing very large file with streaming data acess patterns, running on clusters of commodity hardware.

10. List the advantages of HDFS

    Write once, read many times pattern

    Low-latency access

11. Name the components of HDFS

    - HDFS –Hadoop Distributed File System.
    - MapReduce – A Programming Model
    - Common – Library of components for other modules
    - Hive –Distributed dataware house
    - HBase-Distributed, column –oriented database
    - Pig-Data flow language
    - YARN- Resource management tool

12. Define the anatomy of MapReduce Job Run.

    The process of running job on MapReduce includes the following

    - Job submission
    - Job Initialization
    - Task Assignment
    - Task Execution
    - Progress and status updates
    - Job Completion

13. Define speculative execution of tasks.

    Instead of diagnosing and fixing slow-running tasks; Hadoop tries to detect when a task is running slower than expected and launches another, equivalent, task as a backup. This is called speculative execution.

14. When speculation tasks are launched?

    Speculative task is launched only after all the tasks for a job have been launched, and then only for tasks that have been running for some time (at least a minute) and have failed to make as much progress, on average, as the other tasks from the job.

15. Name the categories of counters used in MapReduce.

    In-built counters – They are used for keeping statistical information about map and reduce tasks

    User defined counters – It can be defined with Java enum types

    Dynamic counters- They can create with String class

    User defined streaming counters – counters used for standard I/O opeartions

16. Give the types and formats of MapReduce

    Input Formats –Input splits and records, FileInputStream

    Text Input –TextInputFormat, KeyValueTextInputFormat,

    Binary Input-SequentialFileInputFormat, SequentialFileAsTextInputFormat,

    Output Format -FileOutputStream

    Text Output –TextOutputFormat

    Binary Output-SequentialFileOutputFormat, MapFileOutputFormat

17. Define shuffle.

    The process by which the system performs the sort—and transfers the map outputs to the reducers as inputs is known as the shuffle

18. Give the name for the failures in MapReduce

    Task Failures

    Tasktracker failure – runtime exception thrown by map/reduce tasks, sudden exit of child JVM, streaming errors, hanging task.

    Jobtracker failure –Serious problem halts all the tasks running under this control.

19. Give some features of MapReduce.

    Counters – for keeping statistical information about the progress of jobs

    Sorting – sorting the data records either in in-memory or disks

    Joins – merges the large data sets.

    Side data distribution – extra read-only data needed by a job to process the main dataset.

    Distributed cache – used for copying files and archives to task nodes in time

    MapReduce library classes- library of mappers and reducers for common functions

20. What is the use of distributed cache in MapReduce?

    It is an extension of traditional cache that may span multiple servers

21. Define side data.

    It can be defined as extra read-only data needed by a job to process the main dataset

16 Marks

1. Explain the architecture and working principle of MapReduce.
2. What are the basic blocks of HDFS architecture? Explain
3. Discuss about the components of HDFS
4. How an application is developed in MapReduce Programme. Explain the steps involved into it.
5. Illustrate the data analysis with Hadoop
6. Write about the schedulers used in job scheduling in Hadoop
7. Discuss on different types of data and formats of MapReduce
8. Discuss about the features of MapReduce.
9. How shuffling and sorting are done in Hadoop. Explain
10. What are the possible failures occur in Hadoop Environment? Brief them