K. K. Wagh Institute of Engineering Education and Research, Nashik.

Department of Computer Engineering

Academic Year: 2020 – 2021 Semester: I
Class & Div: BE – B
Course Name & Code: Laboratory Practice - I (410246)
Teaching Scheme: Practical (04 Hrs / week)
Name of Faculty: Prof. N. G. Sharma

ASSIGNMENT 04 (DATA ANALYTICS) SAMPLE PROBLEM STATEMENTS

1. What is Hadoop Map Reduce?
   For processing large data sets in parallel across a Hadoop cluster, Hadoop MapReduce framework is used. Data analysis uses a two-step map and reduce process.

2. How Hadoop mapreduce works?
   In MapReduce, during the map phase, it counts the words in each document, while in the reduce phase it aggregates the data as per the document spanning the entire collection. During the map phase, the input data is divided into splits for analysis by map tasks running in parallel across Hadoop framework.

3. What is distributed Cache in mapreduce Framework?
   Distributed Cache is an important feature provided by the MapReduce framework. When you want to share some files across all nodes in Hadoop Cluster, Distributed Cache is used. The files could be an executable jar files or simple properties file.

4. What is namenode in Hadoop?
   NameNode in Hadoop is the node, where Hadoop stores all the file location information in HDFS (Hadoop Distributed File System). In other words, NameNode is the centerpiece of an HDFS file system. It keeps the record of all the files in the file system and tracks the file data across the cluster or multiple machines

5. What is jobtracker in Hadoop?
   In Hadoop for submitting and tracking MapReduce jobs, JobTracker is used. Job tracker run on its own JVM process

6. What are the actions followed by Hadoop?

Job Tracker performs following actions in Hadoop

- Client application submit jobs to the job tracker
- JobTracker communicates to the Name mode to determine data location
- Near the data or with available slots JobTracker locates TaskTracker nodes
- On chosen TaskTracker Nodes, it submits the work
- When a task fails, Job tracker notifies and decides what to do then.
- The TaskTracker nodes are monitored by JobTracker

7. What happens when a data node fails?

When a data node fails

- Jobtracker and namenode detect the failure
- On the failed node all tasks are re-scheduled
- Namenode replicates the user's data to another node

8. What are the basic parameters of a Mapper?
   The basic parameters of a Mapper are

   - LongWritable and Text
   - Text and IntWritable

9. What is the function of mapreduce partitioner?
   The function of MapReduce partitioner is to make sure that all the value of a single key goes to the same reducer, eventually which helps even distribution of the map output over the reducers

10. What is a difference between an Input Split and HDFS Block?
    The logical division of data is known as Split while a physical division of data is known as HDFS Block

11. What happens in text format?
    In text input format, each line in the text file is a record. Value is the content of the line while Key is the byte offset of the line. For instance, Key: longWritable, Value: text

12. What is the difference between an RDBMS and Hadoop?

RDBMS	Hadoop
RDBMS is a relational database management system	Hadoop is a node based flat structure
It used for OLTP processing whereas Hadoop	It is currently used for analytical and for BIG DATA processing
In RDBMS, the database cluster uses the same data files stored in a shared storage	In Hadoop, the storage data can be stored independently in each processing node.
You need to preprocess data before storing it	you don't need to preprocess data before storing it

13. Mention Hadoop core components?
    Hadoop core components include,

    - HDFS
    - MapReduce

14. What is "map" and what is "reducer" in Hadoop?

Hadoop Mapper is a function or task which is used to process all input records from a file and generate the output which works as input for Reducer. It produces the output by returning new key-value pairs.

15. Mention how Hadoop is different from other data processing tools?

Reducer in Hadoop MapReduce reduces a set of intermediate values which share a key to a smaller set of values.

Then, Reducer aggregate, filter and combine key-value pairs and this requires a wide range of processing.

16. What are the main components of MapReduce Job?

Mapper & Reducer

17. What is Shuffling and Sorting in MapReduce?

The process of transferring data from the mappers to reducers is shuffling. It is also the process by which the system performs the sort. Then it transfers the map output to the reducer as input. This is the reason shuffle phase is necessary for the reducers.

MapReduce Framework automatically sort the keys generated by the mapper. Thus, before starting of reducer, all intermediate key-value pairs get sorted by key and not by value. It does not sort values passed to each reducer. They can be in any order.Sorting in a MapReduce job helps reducer to easily distinguish when a new reduce task should start.This saves time for the reducer.

18. What are the basic parameters of a Reducer?

The four basic parameters of a reducer are **Text, IntWritable, Text, IntWritable**. The first two represent intermediate output parameters and the second two represent final output parameters.

19. What platform and Java version is required to run Hadoop?

Version 2.7 and later of Apache Hadoop requires Java 7. It is built and tested on both OpenJDK and Oracle (HotSpot)'s JDK/JRE.
Earlier versions (2.6 and earlier) support Java 6

20. Can MapReduce program be written in any language other than Java?

It also supports running non-Java applications in Ruby, Python, C++ and a few other programming languages, via two frameworks, namely the **Streaming** framework and the **Pipes** framework.

Prof. N. G. Sharma

Course Teacher

(BE-B)