**Session 12: Advanced MapReduce-I**

**Assignment 1**

**1. What is the purpose of RecordReader in Hadoop?**

**Answer :** Record Reader an interface between input splits and mappers. A RecordReader uses the data within the boundaries created by the input split to generate key/value pairs.

RecordReader, typically, converts the byte-oriented view of the input, provided by the InputSplit, and presents a record-oriented view for the mapper & reducer tasks for processing. It thus assumes the responsibility of processing record boundaries and presenting the tasks with keys and values.

**2. What happens if the number of reducers is 0?**

**Answer:** IF we set the no of reducer to 0 than it eill not perform sorting of map task and the job is called as **map only.**

**3. What is meant by Map-side and Reduce-side join in Hadoop?**

**Answer:**

**Map-Side Join:**

a. Joins performed by Mapper

b. It can be used only when one data set is small enough to accommodate into distributed cache.

c. The inputs for to each map must be partitioned and sorted in a specific way. Each input dataset must be divided into the same number of partitions, and stored into the hash table.

d. All the mappers can load this hash table back into the memory and do the join work in mapper stage.

e. Map-side join helps in minimizing the cost that is incurred for sorting and merging in the shuffle and reduce stages.

**Reduce-Side Join:**

a. Joins performed by Reducer

b. A mapper’s job during Map Stage is to “read” the data from join tables and to “return” the **‘join key’** and **‘join value’** pair into an intermediate file.

c. in the shuffle stage, this intermediate file is then sorted and merged. The reducer’s job during reduce stage is to take this sorted result as input and complete the task of join.

d. It is less efficient as both datasets have to go through the MapReduce shuffle phase

**4. What is the significance of conf.setMapper class?**

**Answer:** Conf.setMapper class sets the mapper class for the job. Do all the stuff related to map job such as reading a data and generating a key-value pair out of mapper.

**5. Give an example scenario on the usage of counters.**

**Answer:** Counters are used when we want to execute same map-reduce job again with slight alteration such as change in input and output file.

It is used to gather several statistics about job. Two types of counters:

a. Built-in Counters:-

**MapReduce Task Counters** - Collects task specific information (e.g., number of input records) during its execution time.

**FileSystem Counters** - Collects information like number of bytes read or written by a task

**FileInputFormat Counters** - Collects information of number of bytes read through FileInputFormat

**FileOutputFormat Counters** - Collects information of number of bytes written through FileOutputFormat

**Job Counters -** These counters are used by JobTracker. Statistics collected by them include e.g., number of task launched for a job.

b.User-Defined Counters:- The MapReduce Framework offers a provision of user-defined Counters, which can be effectively utilized to monitor the progress of data across nodes of distributed clusters.

**Example:-** The counters are used to solve the problem like Breadth First Search(as here traversing of nodes are required until each node get finished).

**6. Elaborate some problems which can only be solved by MapReduce and cannot be solved by PIG?**

**Answer:**

When we need to use custom partitioner than we need to use MapReduce over Pig as there is facility with MapReduce where we can create custom partitioner but not with Pig.

**Example:** Let sat we want to count population of 2 cities, let say one is Ahmendbad and other is Gandhinagar. So we need to consider key of  Ahmedabad city  similar to Gandhinagar through which we can bring the population data of these two cities to one reducer. So for this we have to instruct map reduce program – whenever you find city with the name ‘Ahmedabad‘ and city with the name ‘Gandhinagar’,  you create the alias name which will be the common name for these two cities so that  you create a common key for both the cities and it get passed to the same reducer. For this, we have to write  custom partitioner.

In mapreduce when you create a ‘key’ for city, you have to consider ‘city’ as the key. So, whenever the framework comes across a different city, it considers it as a different key. Hence, we need to use customized partitioner. There is a provision in mapreduce only, where you can write your custom partitioner and mention if city = Ahmedabad or Gandhinagar then pass similar hashcode.

**7. In what kind of scenarios, MR jobs will be more useful than PIG?**

**Answer:**

1. When developers need **definite driver program** control then they should make use of Hadoop MapReduce.
2. Whenever the job requires implementing a **custom partitioner,** developers can choose MapReduce over Pig and Hive.
3. If there already exists pre-defined library of Java Mappers or Reducers for a job then it is a wise decision to use Hadoop MapReduce.
4. If the developers require **good amount of testability** when combining lots of large data sets then they should use MapReduce.
5. If the job requires **optimization** at a particular stage of processing by making the best use of tricks like in-mapper combining then Hadoop MapReduce can prove to be a better coding approach.
6. If the job has some tricky **usage of distributed cache** (replicated join), cross products, groupings or joins then Hadoop MapReduce is a better programming approach .

**8. What are combiners and when are these used in a MapReduce job?**

**Answer:** When a MapReduce Job is run on a large dataset, Hadoop Mapper generates large chunks of intermediate data that is passed on to Hadoop Reducer for further processing, which leads to massive network congestion.

Combiner is used in between map and reduce task and act as a semi-reducer to reduce this network congestion. It is optional and used to reduce the volume of data transfer between map and reduce class. Hadoop Combiner is to process the output data from Hadoop Mapper, before passing it to a Hadoop Reducer.

The Combiner phase reads each key-value pair, combines the common words as key and values as collection.