Session12 Assignment 1- answers

1. What is the purpose of RecordReader in Hadoop?

RecordReader, typically, converts the byte-oriented view of the input, provided by the InputSplit, and presents a record-oriented view for the [Mapper](https://hadoop.apache.org/docs/r2.6.2/api/org/apache/hadoop/mapred/Mapper.html) & [Reducer](https://hadoop.apache.org/docs/r2.6.2/api/org/apache/hadoop/mapred/Reducer.html) 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?

when you use "No reducer" there is no partitioning&shuffling processes after MAP stage.

Which also means, reduce step will be skipped and mapper output will be the final out

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

Joins performed by Mapper are called as Map-side Joins.

A map-side join can be used to  
join the outputs of several jobs that had the same number of reducers, the same keys and output files that are no bigger than the HDFS block size.

Joins performed by Reducer can be treated as Reduce-side joins.

Reduce-Side joins are more simple than Map-Side joins since the input datasets need not to be structured. But it is less efficient as both datasets have  
to go through the MapReduce shuffle phase. the records with the same key are brought together in the reducer.

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

Conf.setMapperclass sets the mapper class and all the stuff related to map job such as reading a data and generating a key-value pair out of the mapper.

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

Hadoop MapReduce Counter provides a way to measure the progress or the number of operations that occur within MapReduce programs.

These counters are very useful especially when you evaluate some MapReduce programs.

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

Custom partitioning is available only in Mapreduce and not in PIG. Wherever we may need to use custom partitioner, we have to choose Mapreduce over PIG. For example:,

Lets consider the data set containing population details for different Indian cities. Suppose we need to calculate total population based on North Indian cities and South Indian cities, we can make use of a common key for say, Bangalore + Chennai + Hyderabad and another Common key for Noida + Delhi + Surat + etc. These common key can be passed to same reducer by using custom practitioner.

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

In the scenarios, where we need to make use of custom practitioner, MP Jobs are prefeeered.

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

The combiner function is used as an optimization for the MapReduce job. The combiner function runs on the output of the map phase and is used as a filtering or an aggregating step to lessen the number of intermediate keys that are being passed to the reducer. In most of the cases the reducer class is set to be the combiner class. The difference lies in the output from these classes. The output of the combiner class is the intermediate data that is passed to the reducer whereas the output of the reducer is passed to the output file on disk.

This you can primarily use for decreasing the amount of data needed to be processed by Reducers.