**Session 13: Advanced MR 2**

**Assignment 2**

**1. What is the difference between TextInputFormat and KeyValueInputFormat class?**

**Answer:**

**TextInputFormat:** It reads lines of text files and provides the offset of the line as key to the Mapper and actual line as Value to the mapper.  
**KeyValueInputFormat**: Reads text file and parses lines into key, Val pairs. Everything up to the first tab character is sent as key to the Mapper and the remainder of the line is sent as value to the mapper.

TextInputFormat is the default file format in Hadoop .

**2. How is the splitting of file invoked in Hadoop framework?**

**Answer:** It is invoked by the Hadoop framework by running **getInputSplit()** method of the **InputFormat** class (like FileInputFormat) defined by the user.

**3. Consider case scenario: In M/R system, - HDFS block size is 64 MB**

**- Input format is FileInputFormat – We have 3 files of size 64K, 65Mb and 127Mb**

**How many input splits will be made by Hadoop framework for each file?**

**Answer:** Hadoop will make 5 splits as follows:

1 split for 64K files

2 splits for 65MB files

2 splits for 127MB files

**4. After the Map phase finishes, the Hadoop framework performs “Partitioning, Shuffle and sort”. Explain each event in brief.**

**Answer:**

**Partitioning:** It is the process of determining which reducer instance will receive which intermediate keys and values. Each mapper must determine for all of its output (key, value) pairs which reducer will receive them. It is necessary that for any key, regardless of which mapper instance generated it, the destination partition is the same.

**Shuffle:** After the first map tasks have completed, the nodes may still be performing several more map tasks each. But they also begin exchanging the intermediate outputs from the map tasks to where they are required by the reducers. This process of moving map outputs to the reducers is known as shuffling.

**Sort:** Each reduce task is responsible for reducing the values associated with several intermediate keys. The set of intermediate keys on a single node is automatically sorted by Hadoop before they are presented to the Reducer.

**5. What is a Combiner?**

**Answer:** The Combiner is a ‘mini-reduce’ process which operates only on data generated by a mapper. Combiners are used to increase the efficiency of a MapReduce program. They are used to aggregate intermediate map output locally on individual mapper outputs. Combiners can help us to reduce the amount of data that needs to be transferred across to the reducers. We can use our reducer code as a combiner if the operation performed is commutative and associative. The execution of combiner is **not guaranteed**; Hadoop may or may not execute a combiner. Also, if required it may execute it more than 1 times. Therefore our MapReduce jobs should not depend on the combiners execution.

**6. What is Hadoop streaming?**

**Answe: Hadoop Streaming** is a generic API which allows writing Mappers and Reduces in any language. But the basic concept remains the same. Mappers and Reducers receive their input and output on stdin and stdout as (key, value) pairs.

**7. What are the most commonly defined input formats in Hadoop and explain each in brief.**

**Answer:**  The most common Input Formats defined in Hadoop are:

1. **TextInputFormat:-** It reads lines of text files and provides the offset of the line as key to the Mapper and actual line as Value to the mapper.
2. **KeyValueInputFormat:-** Reads text file and parses lines into key, Val pairs. Everything up to the first tab character is sent as key to the Mapper and the remainder of the line is sent as value to the mapper.
3. **SequenceFileInputFormat:-** Sequencefileinputformat is used for reading files in sequence.

 TextInputFormat is the Hadoop default.

**8. Explain what is distributed Cache in MapReduce Framework ?**

**Answer:** Distributed cache is an important feature provide by map reduce framework. Distributed cache can cache text, archive, jars which could be used by application to improve performance. Application provide details of file to jobconf object to cache.

DistributedCache tracks the caching with timestamp. Cached file should not be changed during the job execution.

**9. Explain what happens in textinputformat ?**

**Answer:** In textinputformat, 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

**10. Explain what is Sequencefileinputformat?**

**Answer:**

1. Sequencefileinputformat is used for reading files in sequence.
2. It is a specific compressed binary file format which is optimized for passing data between the output of one MapReduce job to the input of some other MapReduce job.
3. They are intrinsically more compact than text files.
4. It can be splitted and processed in parallel.