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

**Ans:** TextInputFormat:   
1) 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) TextInputFormat is the default file format in Hadoop .  
  
KeyValueInputFormat:   
1) 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.

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

**Ans:** It is invoked by the Hadoop framework by running getInputSplit()method of the Input format 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?**

**Ans:** 64- 1 split

65- 2 split

127- 2 split

**4. After the Map phase finishes, the Hadoop framework does “Partitioning, Shuffle and sort”. Explain what happens in this phase?**

**Ans: 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?**

**Ans:** The Combiner is a ‘mini-reduce’ process which operates only on data generated by a mapper. The Combiner will receive as input all data emitted by the Mapper instances on a given node. The output from the Combiner is then sent to the Reducers, instead of the output from the Mappers.

**6. What is Hadoop Streaming?**

**Ans:** Streaming is a generic API that allows programs written in virtually any language to be used as Hadoop Mapper and Reducer implementations.

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

**Ans:** The 3 most common Input Formats defined in Hadoop are:

a. TextInputFormat

b. KeyValueInputFormat

c. SequenceFileInputFormat

  TextInputFormat is the Hadoop default.

**8. What is Distributed Cache in Hadoop?**

**Ans:** Distributed Cache is a facility provided by the MapReduce framework to cache files (text, archives, jars and so on) needed by applications during execution of the job. The framework will copy the necessary files to the slave node before any tasks for the job are executed on that node.

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

**Ans:** textinputformat are key being simply the offset within the file it is common for each line in a  
file to be key value pair separate by a delimiter .

**10. Explain what is Sequencefileinputformat?**

**Ans:** the *SequenceFileInputFormat* reads special binary files that are specific to Hadoop. These files include many features designed to allow data to be rapidly read into Hadoop mappers. Sequence files are block-compressed and provide direct serialization and deserialization of several arbitrary data types (not just text). Sequence files can be generated as the output of other MapReduce tasks and are an efficient intermediate representation for data that is passing from one MapReduce job to another.