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

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)the\_default\_file\_format. .  
  
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.

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

Files are split into HDFS blocks and the blocks are replicated. Hadoop assigns a node for a split based on data locality principle. Hadoop will try to execute the mapper on the nodes where the block resides. Because of replication, there are multiple such nodes hosting the same block.

In case the nodes are not available, Hadoop will try to pick a node that is closest to the node that hosts the data block. It could pick another node in the same rack, for example. A node may not be available for various reasons; all the map slots may be under use or the node may simply be down.

**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?**

* 1 split for 64K files
* 2 splits for 65MB files
* 2 splits for 127MB files

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

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?**

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?**

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.**

These are the most common input formats defined in Hadoop:

TextInputFormat

KeyValueInputFormat

SequenceFileInputFormat

Default input format in Hadoop is TextInputFormat.

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

**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 ?**

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?**

Sequencefileinputformat is used for reading files in sequence. 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.