Session13 Assignment 2- answers

|  |  |
| --- | --- |
| **TextInputFormat** | **KeyValueInputFormat** |
| 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. | 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) TextInputFormat is the default file format in Hadoop . |  |

1. **What is the difference between TextInputFormat and KeyValueInputFormat class?**
2. **How is the splitting of file invoked in Hadoop framework?**

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?**

File with size = 64K, will have 1 input split

File with size = 65Mb will have 2 input split

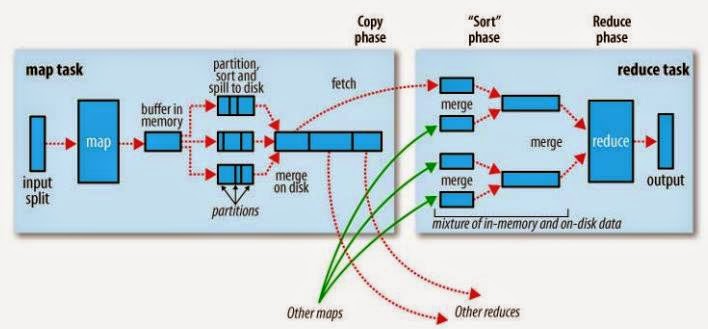
File with size = 127Mb will have 2 input split

**4. 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?**

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](http://en.wikipedia.org/wiki/Network_congestion). So how do go about reducing this network congestion? Is there any function in Hadoop to address this issue? The MapReduce framework offers a function known as ‘*Combiner*’ that can play a crucial role in reducing network congestion. As a matter of fact ‘*Combiner’* is also termed as ‘*Mini-reducer*’. It is important to note that the primary job of a Hadoop Combiner is to process the output data from Hadoop Mapper, before passing it to a Hadoop Reducer. Technically speaking, Combiner and Reducer use the same code.

**6. What is Hadoop streaming?**

Hadoop streaming is a utility that comes with the Hadoop distribution. The utility allows you to create and run Map/Reduce jobs with any executable or script as the mapper and/or the reducer.

The basic concept remains the same. Mappers and Reducers receive their input and output on stdin and stdout as (key, value) pairs. Apache Hadoop uses streams as per UNIX standard between your application and Hadoop system. Streaming is the best fit for text processing. The data view is line oriented and processed as a key/value pair separated by 'tab' character. The program reads each line and processes it as per the requirement.

In Streaming, input and output are always represented as text. The 'tab' character is used to separate key and value. The Streaming program uses the 'tab' character to split a line into key/value pair.

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

The most common input formats defined in Hadoop are;

* 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.
* SequenceFileInputFormat - 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

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

Distributed Cache is an important feature provided by map reduce framework. When you want to share some files across all nodes in Hadoop Cluster, DistributedCache  is used.  The files could be an executable jar files or simple properties file.

**9. Explain what happens in 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.

**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.