MapReduce 1.0 – Processing API of Hadoop

Apache Hadoop Tutorial – We shall learn about MapReduce 1.0, which is the Processing API of Hadoop.

MapReduce 1.0 is the initial version of MapReduce in Hadoop. Over time, to support distributed processing models as well, MapReduce has evolved to 2.0 with inclusion of YARN.

MapReduce 1.0

MapReduce is a programming paradigm that has caused Hadoop a big sensation in Big Data industry.

MapReduce is kind of approach to a problem. MapReduce originally was a result of research at Google for the problem of indexing all the content on Internet.

MapReduce has many levels analogous to layers of onion. We shall go through the understanding of each level and go deeper. As the name MapReduce suggests, visibly there are two parts in this approach: **Map** and **Reduce**.In addition to the Map and Reduce there is another part called **Shuffle** between Map and Reduce. We shall learn in detail about these three components.

Map

Mapping is kind of breaking down the input data into key-value pairs based on the given problem statement.

Input to Mapper : Raw DataOutput from Mapper : <***key***,***value***> pairs

Shuffle

Shuffling is sorting of the <***key***,***value***> pairs based on ***key***.

Input to Shuffler : <***key***,***value***> pairsOutput from Shuffler : Sorted <***key***,***value***> pairs

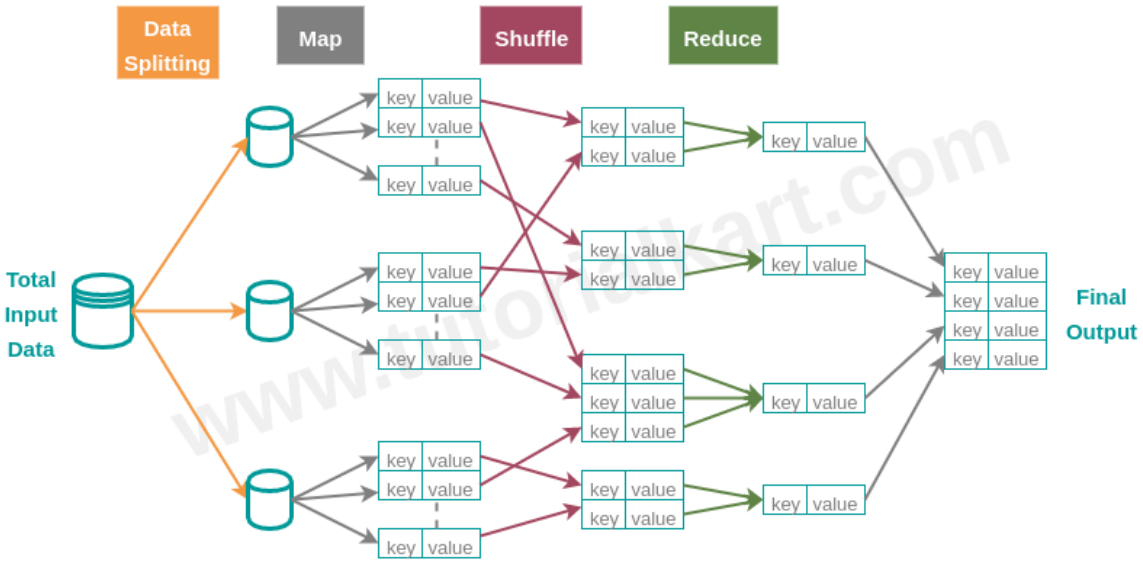
Reduce

Reducing is the task of aggregating those <***key***,***value***> pairs with same ***key*** to a single <***key***,***value***> pair with updated ***value***.

Input to Reducer : Sorted <key,value> pairsOutput from Reducer : <key,value> pairs which are reduced by key.

Stages in MapReduce

Following diagram shows typical stages in a Hadoop MapReduce along with the data flow from one stage to other :



Where does the efforts of a programmer go while implementing MapReduce ?

For a given problem and data, usually the programmer has to implement Mapper and Reducer. Shuffling is done automatically, unless you want to override it for some reasons. As a Hadoop developer, you are aware that MapReduce algorithm deals with <***key***,***value***> pairs. So, your first task is to realize the breaking of your input data into <***key***,***value***> pairs, which is the programming logic that goes into Mapper class. The Reducer class receives the list of ***value***s for each ***key***. You need to provide the logic of how you want those multiple ***value***s to be reduced to a single ***value***.

Conclusion

In this Apache Hadoop Tutorial, we have learnt about MapReduce 1.0 and the stages in the approach, the areas in which a programmer has to keep his/her efforts.

Learn Apache Hadoop

- Hadoop Tutorial
- Install Hadoop on Ubuntu