MapReduce

MapReduce is the processing layer of Hadoop.

MapReduce programming model is designed for processing large volumes of data in parallel by dividing the work into a set of independent tasks.

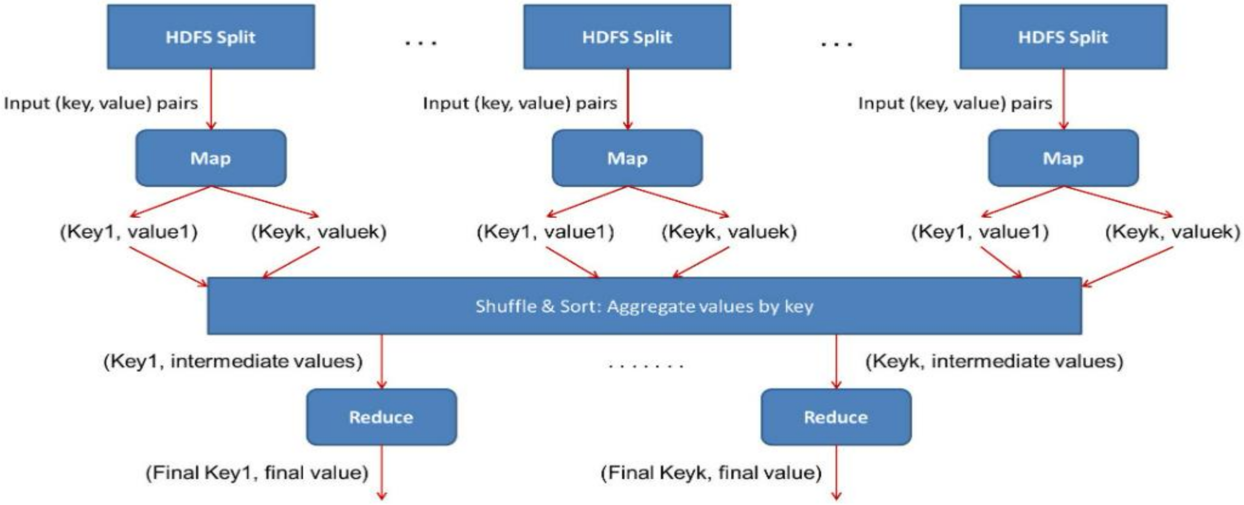
It is a processing layer of Hadoop.

There are two processes one is Mapper, and another is the Reducer.

1. Mapper
    - It is used to process the input data, Input data is in the form of file or directory which resides in HDFS.
    - Client needs to write the map reduce program and need to submit the input data.
    - The input file is passed to mapper line by line. It will process the data and produce the output which is called intermediate output.
    - The output of map is stored on the local disk from where it is shuffled to reduce nodes.
2. Reducer

    It takes an intermediate key/value pairs produced by Map. Reducer has 3 primary phases: shuffle, sort and reduce.

    - shuffle – Input to the reducer is sorted output of mappers. In this phase, framework fetches all output of mappers.
    - Sort – The framework groups reducer input by keys.
    - Reducer is the second phase of processing when the client needs to write the business logic.
    - The output of reducer is the final output that is written to HDFS.


HDFS Split
. . .
HDFS Split
. . .
HDFS Split
Input (key, value) pairs
Input (key, value) pairs
Input (key, value) pairs
Map
Map
Map
(Key1, value1)
(Keyk, valuek)
(Key1, value1)
(Keyk, valuek)
(Key1, value1)
(Keyk, valuek)
Shuffle & Sort: Aggregate values by key
(Key1, intermediate values)
. . . .
(Keyk, intermediate values)
Reduce
Reduce
(Final Key1, final value)
(Final Keyk, final value)