Map Reduce

MapReduce

- MapReduce is a python programming model.
- Map Reduce allows us to process large volume of data by dividing entire task into small pieces.
- It also allows parallel processing across clusters of machines.

Every MapReduce structure is composed of three major phases.

- **Map**
- **Shuffle and sort**
- **Reduce**

Map

This is first stage of MapReduce application.
A function called mapper routes a series of key-value pairs inside the map stage.

Shuffle and Sort

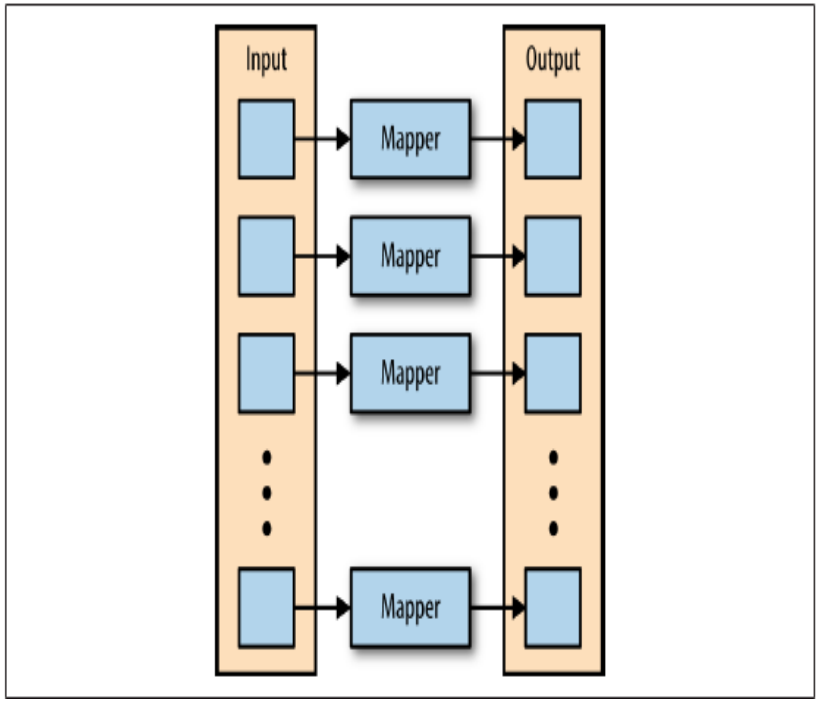
- This is second stage of MapReduce.
- The intermediate outputs from the map stage are moved to the reducers as the mappers bring into being completing.
- This process of moving output from the mappers to the reducers is recognized as shuffling.
- Shuffling is moved by a divider function, named the partitioner.
- The partitioner is used to handle the flow of key-value pairs from mappers to reducers.
- The last stage before the reducers begin processing data is the sorting process.

MapReduce

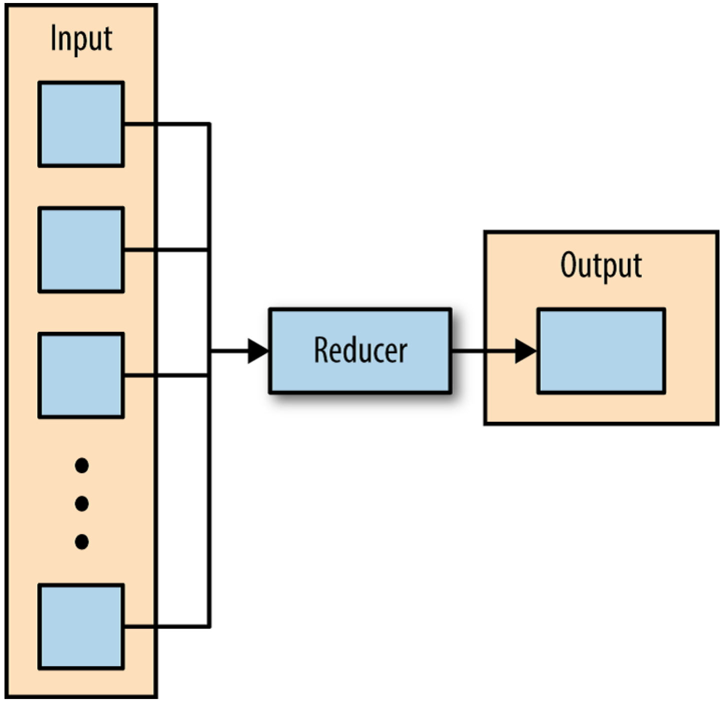
Reduce

This is third stage in MapReduce Stage.

An iterator of values is given to a function known as the reducer inside the reducer phase.



Map Stage



Reduce Stage

MapReduce

Example:

Suppose you have marks of 100 students in 6 subjects and you want to compute their average.

Input : 100 rows consisting 6 values each

Map : array of 6 marks for each students.

Reduce : Average function

Output : Average marks of all 100 students

Input	Map	Reduce	Output
43,58,78,85,99,66	[43,58,78,85,99,66]	**AVERAGE**([43,58,78,85,99,66])	71.5
85,91,93,87,95,99	[85,91,93,87,95,99]	**AVERAGE**([85,91,93,87,95,99])	91.66667
75,81,78,65,58,66	[75,81,78,65,58,66]	**AVERAGE**([75,81,78,65,58,66])	70.5
.	.	.	.
.	.	.	.
.	.	.	.