MapReduce

Why MapReduce:

- Distributes the processing of data on cluster
- Divide data into partition that are MAPPED (transform) and REDUCE (aggregated) by the mapper and reducer function.
- Resilient to failure- An application master monitors a mapper and reducer on each partition.

How MapReduce Works:

Mapping

The MAPPER converts raw source data into key/value pairs

Mapping Example:
Movie Data

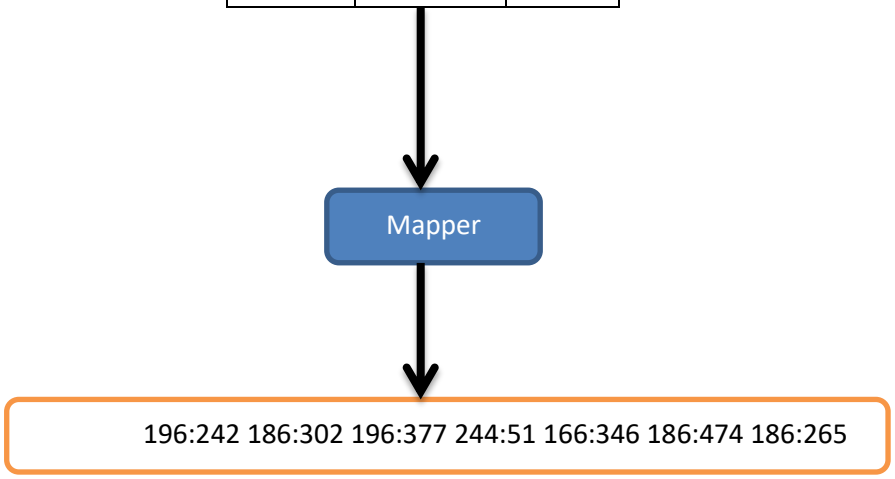
Input Data

User ID	Movie ID	Rating
196	242	3
186	302	3
196	377	1
244	51	2
166	346	1
186	474	4
186	265	2

Map user to movies ID they watched

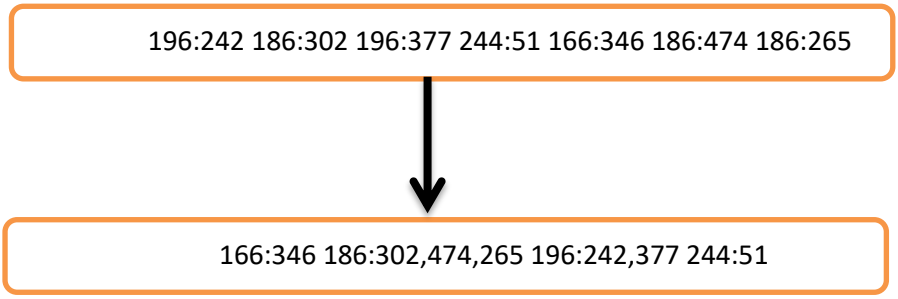
Input Data

User ID	Movie ID	Rating
196	242	3
186	302	3
196	377	1
244	51	2
166	346	1
186	474	4
186	265	2

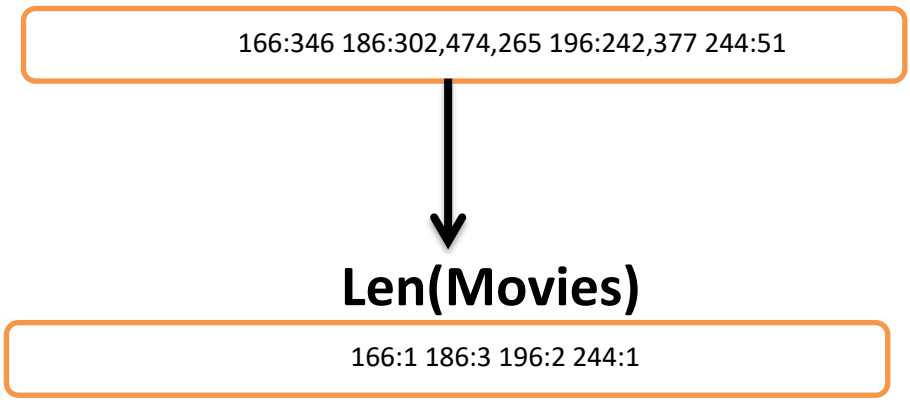


Shuffle and Sort

MapReduce sorts and groups the mapped data



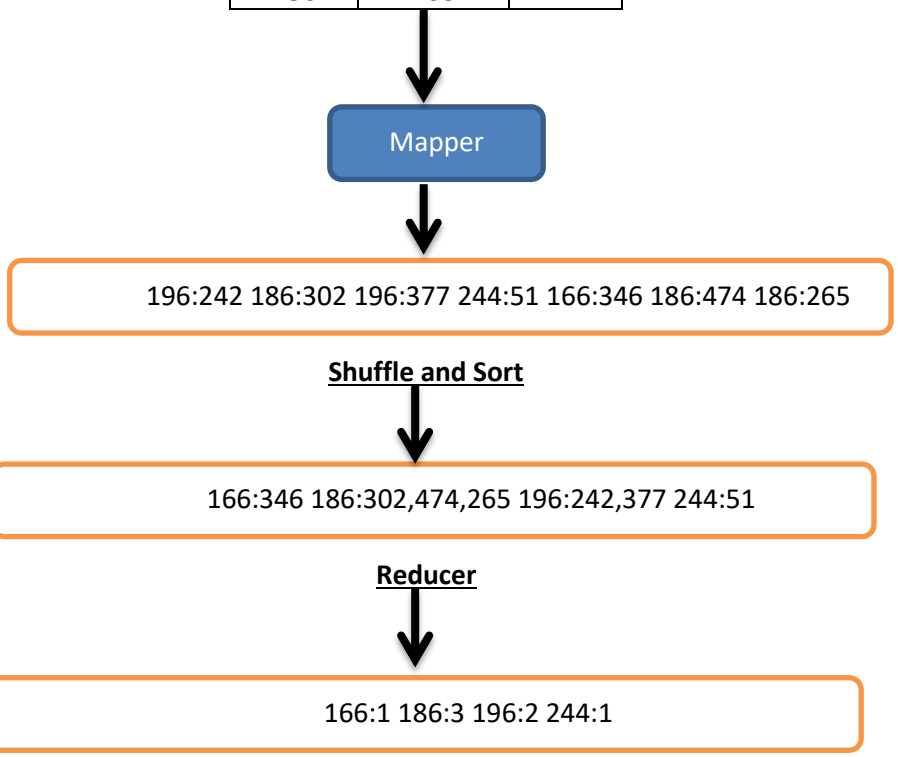
The Reducer processes each key's value



Entire MapReduce Process

Input Data

User ID	Movie ID	Rating
196	242	3
186	302	3
196	377	1
244	51	2
166	346	1
186	474	4
186	265	2



YARN- Yet Another Resource Negotiator

- YARN is introduced in Hadoop 2 system, which is used to separates the problem of managing resources on your cluster from MapReduce.
- It enables a development of MapReduce alternatives like Sparks, Tez built on top of YARN.

YARN Architecture