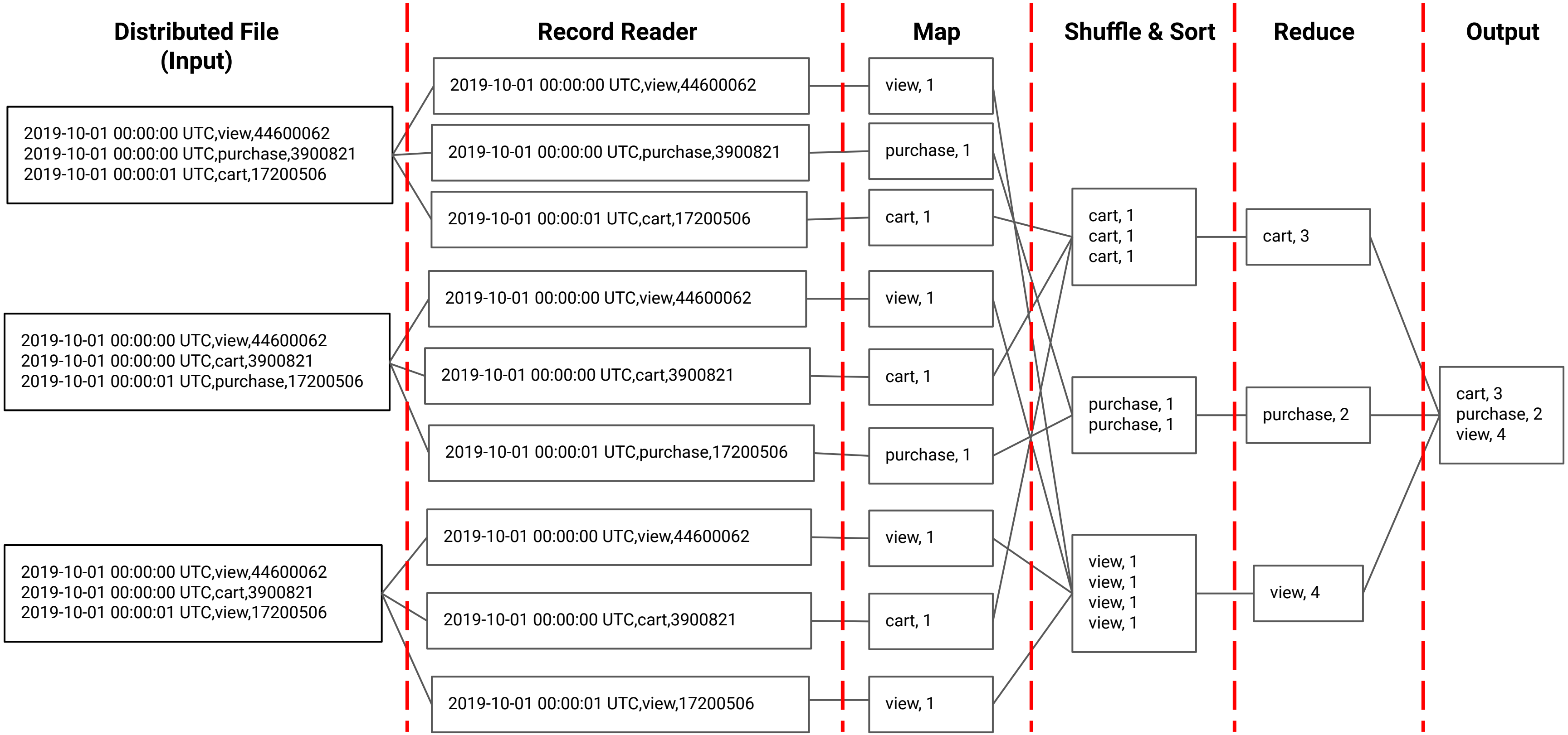
Doing Data Analytics II

31 Dec 2022

The MapReduce Framework

- MapReduce is a distributed execution framework within the Apache Hadoop ecosystem.
- The framework was developed by Google in 2004 to retrieve and process data stored in distributed file system (GFS).
- MapReduce once was the only available method to retrieve data from Hadoop Distributed File System (HDFS).
- However, now there are several alternatives to MapReduce such as Hive, Pig and Spark.

The MapReduce Framework



Distributed File (Input)

2019-10-01 00:00:00 UTC,view,44600062
2019-10-01 00:00:00 UTC,purchase,3900821
2019-10-01 00:00:01 UTC,cart,17200506

2019-10-01 00:00:00 UTC,view,44600062
2019-10-01 00:00:00 UTC,cart,3900821
2019-10-01 00:00:01 UTC,purchase,17200506

2019-10-01 00:00:00 UTC,view,44600062
2019-10-01 00:00:00 UTC,cart,3900821
2019-10-01 00:00:01 UTC,view,17200506

Record Reader

2019-10-01 00:00:00 UTC,view,44600062

2019-10-01 00:00:00 UTC,purchase,3900821

2019-10-01 00:00:01 UTC,cart,17200506

2019-10-01 00:00:00 UTC,view,44600062

2019-10-01 00:00:00 UTC,cart,3900821

2019-10-01 00:00:01 UTC,purchase,17200506

2019-10-01 00:00:00 UTC,view,44600062

2019-10-01 00:00:00 UTC,cart,3900821

2019-10-01 00:00:01 UTC,view,17200506

Map

view, 1

purchase, 1

cart, 1

view, 1

cart, 1

purchase, 1

view, 1

cart, 1

view, 1

Shuffle & Sort

cart, 1
cart, 1
cart, 1

purchase, 1
purchase, 1

view, 1
view, 1
view, 1
view, 1

Reduce

cart, 3

purchase, 2

view, 4

Output

cart, 3
purchase, 2
view, 4

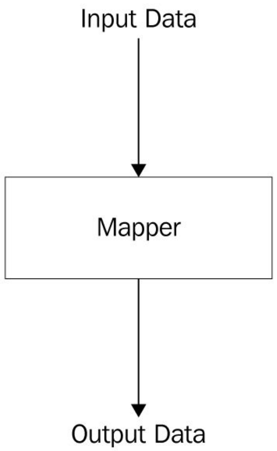
The MapReduce Framework

- **Record reader:** Divides the input data into appropriately sized splits, reads input data and generates key/value pairs.
- **Map:** Takes a series of key/value pairs, process each and generate zero or more key/value pairs.
- **Combiner:** Is an optional localised reducer which can group data in the map phase.
- **Partitioner:** Takes the key/value pairs from the maps and split them up into shards, one shard per reducer.
- **Shuffle and Sort:** Takes the output files written by the partitioners, downloads them to the reducer local machine and then sort by key.
- **Reduce:** The data can be aggregated, filtered and combined in a number of ways.
- **Output format:** Writes out the reducer output into HDFS.

MapReduce Job Types

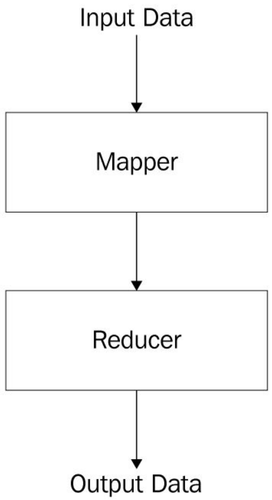
Single mapper job:

Single mapper jobs are used in transformations such as data format conversions.



Single mapper reducer job

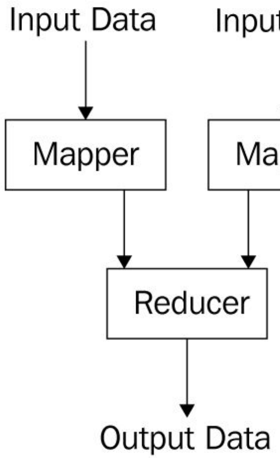
Single mapper reducer jobs are used in aggregation use cases such as count, sum or average.



MapReduce Job Types

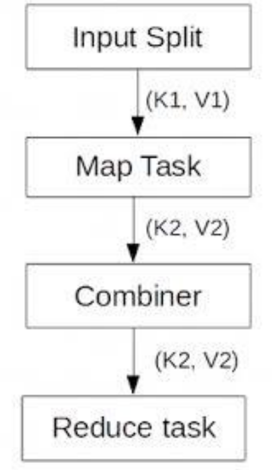
Multiple mappers reducer job

Multiple mappers reducer jobs are used in the join use cases.



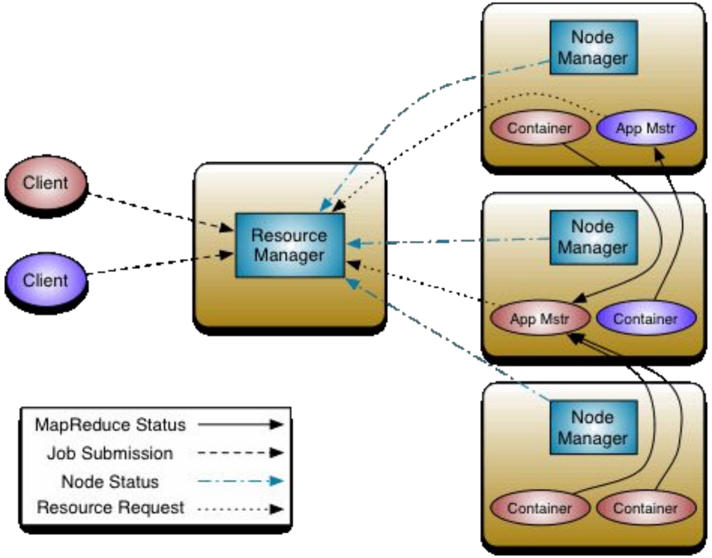
Single Mapper Combiner Reducer job

Combiner is used as a mini reducer to reduce the workload of the reducer. It is an optional class in between the mapper and reducer.



Yet Another Resource Negotiator (YARN)

Since Hadoop 2.0, YARN (Yet Another Resource Negotiator) is used to schedule the mappers and reducers tasks.



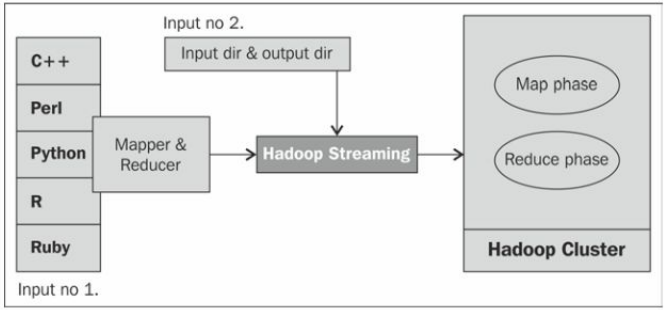
Hadoop Streaming

- Hadoop streaming is a Hadoop utility for running the Hadoop MapReduce job with executable scripts such as Mapper and Reducer.

- With this, the text input file is printed on stream (stdin), which is provided as an input to Mapper. And the output (stdout) of Mapper is provided as an input to Reducer. Finally, Reducer writes the output to the HDFS directory.

- The main advantage of the Hadoop streaming utility is that it allows Java as well as non-Java programmed MapReduce jobs to be executed over Hadoop clusters.

- Here, creating the driver file for running the MapReduce job is optional when we are implementing MapReduce with R and Hadoop.



Hadoop streaming components

Questions & Answers

Use your Google Classroom stream to post any questions or start discussions.
https://classroom.google.com/c/NDk3MzI3NzgxNDM5?cjc=ltzuypk