2.3.4 MapReduce

- MapReduce is a popular programming model for data intensive applications. It is a parallel processing framework.
- Computational processing can occur on data (even semi-structured and unstructured data) stored in a file system without loading it into any kind of database.
- It’s neither a database nor a direct competitor to databases.
- The run-time systems take care of tasks such partitioning the data, scheduling of jobs and communication between nodes in the cluster. This makes it easier for programmers to analyze massive scale data without worrying about tasks such as data partitioning and scheduling.
- This model has been implemented by the open source software Hadoop.

MapReduce phases

MapReduce model has two phases: Map and Reduce.

- In the Map phase, data is read from a distributed file system, partitioned among a set of computing nodes in the cluster, and sent to the nodes as a set of key-value pairs.
- When all the Map tasks are completed, the Reduce phase begins with the shuffle and sort step, in which the intermediate data is sorted by the key and the key-value pairs are grouped and shuffled to the reduce tasks.
- The reduce tasks then take the key-value pairs grouped by the key and run the reduce function for each group of key-value pairs.

An optional Combine task can be used to perform data aggregation on the intermediate data of the same key for the output of the mapper before transferring the output to the Reduce task.