The volume of digital data now is extremely large. One of the positive aspects of big data is reducing the advantages of “having more data”, which makes better or sophisticated algorithms more meaningful. However, storing and analyzing big data is a problem. The Hadoop Distributed Filesystem(HDFS) is introduced to solve the hardware failure of parallel writing and reading data from multiple disks. MapReduce provides a programming model that abstracts the problem from disk reads and writes, transforming it into a computation over sets of keys and values(1) to combine analysis tasks and input data. Hadoop is reliable, scalable, open source and affordable.

The premise of using MapReduce is that the dataset can be processed for each query. It is a powerful batch query processor that is able to run ad hoc queries against the whole dataset and get results in a reasonable time(1). It is not suitable for interactive analysis. However, Hadoop introduced a cluster resource management system to allow distributed programs: Yet Another Resource Negotiator(YARN), to enable new processing models and patterns such *Interactive SQL, Iterative processing, Stream processing* and *Search*.

1. Tom White. 2015. *Hadoop: The Definitive Guide (4th. ed.)*. O'Reilly Media, Inc.