Hadoop Frameworks

Bhavneet Soni

**Hadoop Frameworks**

Hadoop is an open ware framework based on the GFS. Hadoop was developed by Doug Cutting in Jan 2006 based on the white papers on Google File system and MapReduce: Simplified Data Processing on Large Clusters (https://en.wikipedia.org/, n.d.). Over the years many companies have developed their own versions of Hadoop framework adding and optimizing some of the features for specific uses.

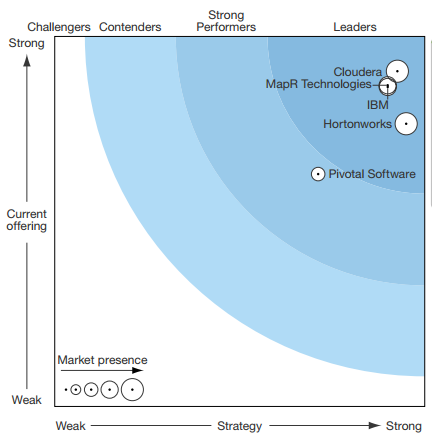


Figure 1 Popular Hadoop Distributions (Gualtieri & Yuhanna, 2016)

There are over 30 different distributions of Hadoop framework are available, however in this paper we will try to compare three widely used versions of Hadoop offered by:

1. Apachae
2. MapR
3. HartonWorks

Developers of Hadoop donated the project to Apache.org and maintains the main development of the Hadoop framework. Apachae Hadoop has two main components a storage part, known as Hadoop Distributed File System (HDFS), and a processing part which is a responsible for splitting files in to large blocks and distributing it over different node in a cluster, known as MapReduce programming model. Apart from thee Hadoop library also consists of Hadoop Common: The common utilities that support the other Hadoop modules, and Hadoop YARN: A framework for job scheduling and cluster resource management. The latest Apache Hadoop uses Java 8 for compiling.

Hortonworks, founded in 2011, is one of the major contributor to the Hadoop community, helping make it more robust and easier to use. Its similar to Apache Hadoop as in its also a open source platform for analyzing, storing and managing big data. It’s the only commercial vendor to distribute complete open source Apache Hadoop without additional proprietary software. Hortonworks has come up with innovative Yarn, which is better than MapReduce in the sense that it will enable inclusion of more data processing frameworks. (https://www.experfy.com/, 2014)

Major contributors to Apache Hadoop and dedicated to working with the community to make Apache Hadoop more robust and easier to install, manage, use, integrate and extend. Hortonwork had put in a lot of resources to make it a truly open source framework (Hadoop Wiki, 2016).

Another widely popular Hadoop distribution is offered by MapR, team at MapR has reengineered MapReduce components to provide significantly higher performance, higher level of fault tolerance, distributed incremental backups, read-write access to the cluster file store via NFS and other features (the HDFS team would dispute some of these assertions). (Hadoop Wiki, 2016) MapR Technologies distribution optimizes the Hadoop to improve its performance and scale potential with minimal effort. The MapR filesystem, which implements the HDFS API, is fully read/write, and can store trillions of files (versus the complex configuration for HDFS that requires separated namespaces). MapR has been at the forefront in providing a reliable and efficient distribution for large-cluster implementations so much so that most of the upcoming large, mission-critical Hadoop clusters and want to use MapR-DB and MapR Streams (which implement the HBase and Kafka APIs, respectively). (Gualtieri & Yuhanna, 2016)

References

(n.d.). Retrieved from https://en.wikipedia.org/: https://en.wikipedia.org/wiki/Apache\_Hadoop

(2014, Sept 5). Retrieved from https://www.experfy.com/: https://www.experfy.com/blog/cloudera-vs-hortonworks-comparing-hadoop-distributions/

Gualtieri, M., & Yuhanna, N. (2016). *The Forrester Wave™: Big Data Hadoop Distributions, Q1 2016.* Forrester Research, Inc.

*Hadoop Wiki*. (2016, 01 20). Retrieved from https://wiki.apache.org/hadoop/Distributions%20and%20Commercial%20Support