# Hadoop Challenges in Big Data and Cloud Era

Gavin Ni*

August 16, 2016

**Abstract**

Living in this fast changing cloud era, not only enterprise customers but also government customers prefer hassle-free cloud services e.g. Software as a Service(SaaS), Platform as a Service(PaaS) and Infrastructure as a Service(IaaS) than the traditional way of buying hardwares and softwares stacks before they can even deploying the service. This trend is particular true to Big Data area e.g. Hadoop system, which requires a fairly large number of hardware resources and deeper knowledge of system administrative skills before they can deploy and use the service since the first Apache Hadoop has released from 2011 (Manyika et al., 2011). This research reviews the technical solution of Hadoop and explores the challenges that Big Data engineers facing in order to run Hadoop as a Service(HDaaS).

## 1 Introduction

In early days, industries only require RDBMS e.g. Oracle, DB2, MSSQL, MySQL, Postgres etc. to process data up to terabytes and the costs of processing such data increase dramatically when the volume of data increases. Engineers desperately try to develop a system that can process data beyond terabytes in some cost effective ways. In 2008, Yahoo! Engineer Doug Cutting leaded a project which was originally designed to create web page search index on internet, was made as a top Apache project named Apache Hadoop™(White, 2012). Hadoop system implemented the Hadoop Distributed File System (HDFS) and the MapReduce algorithm published by Google (Dean & Ghemawat, 2008). Hadoop cluster can scale from under 10 nodes up to tens of thousands nodes. For instance, Yahoo! has been running the biggest Hadoop cluster in the world of 42,000 nodes to process petabyte data source from Hortonworks Hadoop summit 2011 keynote and Facebook has been running multiple clusters from group of 10 to group of thousands nodes (Borthakur et al., 2011). As Hadoop plays increasingly important role to Big Data, several commercial distribution also available for Enterprise to choose besides open source Apache Hadoop™, Pivotal HD™, HortonWorks Data Platform™and Cloudera Enterprise™. The following

*Student ID: 2566358 Email: jni001@ec.auckland.ac.nz

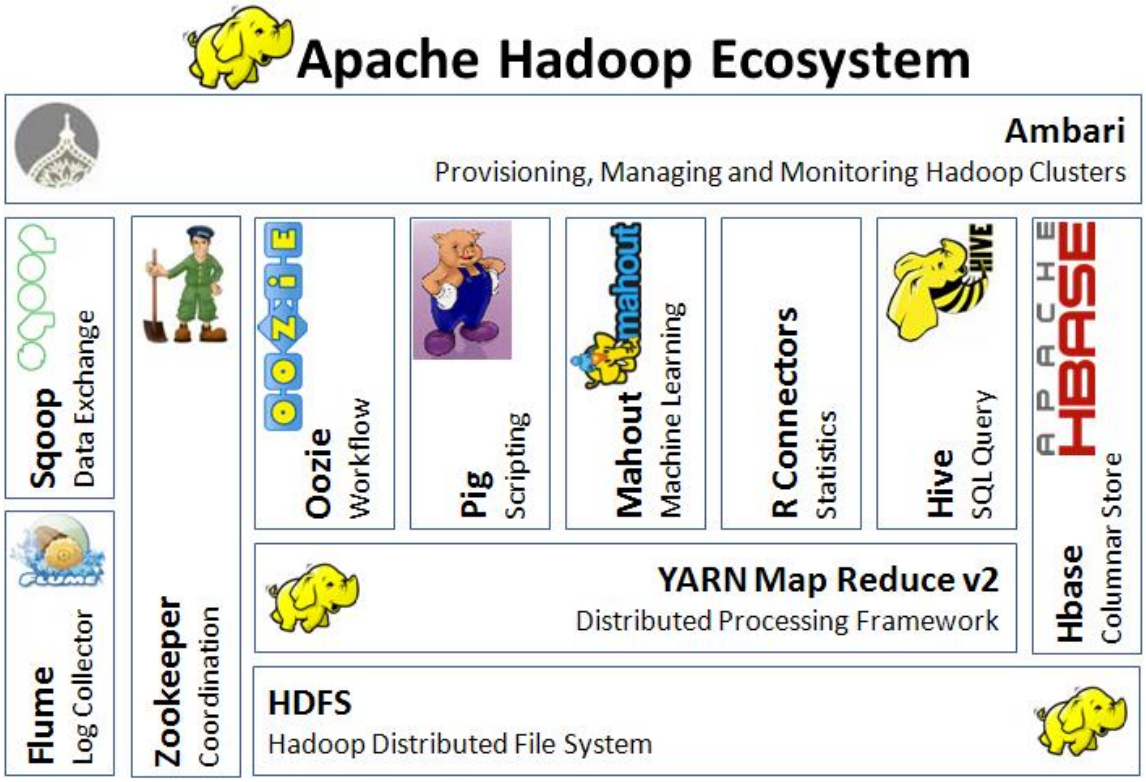


Figure 1: Hadoop Ecosystem

sections introduce HDFS and MapReduce principal then followed review the industrial Big Data solutions based on Hadoop case studies and their challenges, finally discuss the future Big Data technology trends.

## 2 Apache Hadoop Ecosystem

Apache Hadoop Ecosystem is a framework of various types of complex and evolving tools see Figure 1 and components from Apache Foundation which have proficient advantage in solving big Data and Business Intelligence(BI) problems. In (?, ?), the Hadoop tools set had been labeled as Data Management, Data Access, Data Process and Data Storage.

To understand Hadoop Ecosystem and what problem it can solve, in the follow subsections we will explain Hadoop Core system, HDFS and MapReduce, and Hadoop extension tools that make Hadoop core system even powerful and easy of use and maintain.

## 2.1 Hadoop Core System

The Apache Hadoop software library is a framework that allows for the distributed processing of large data sets across clusters of computers using simple programming models. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Rather than rely on hardware to deliver high-availability, the library itself is designed to detect and handle failures at the application layer, so delivering a highly-available service on top of a cluster of computers, each of which may be prone to failures.

### 2.1.1 HDFS™

The Hadoop Distributed File System (HDFS) is a distributed file system designed to run on commodity hardware. It has many similarities with existing distributed file systems. However, the differences from other distributed file systems are significant. HDFS is highly fault-tolerant and is designed to be deployed on low-cost hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. (Shvachko, Kuang, Radia, & Chansler, 2010)

### 2.1.2 YARN and MapReduce

YARN is Hadoop job scheduling system for parallel processing of large data sets using MapReduce. see Figure 3

## 2.2 Data Access

OOzie, Pig, Mahaut, R and Hive are developed running on Yarn while HBase access data directly off HDFS.

## 2.3 Operation Management

Tools like Zookeeper, Flume, Sqoop and Ambari are developed for deployment, configuration, diagnostics, and reporting.

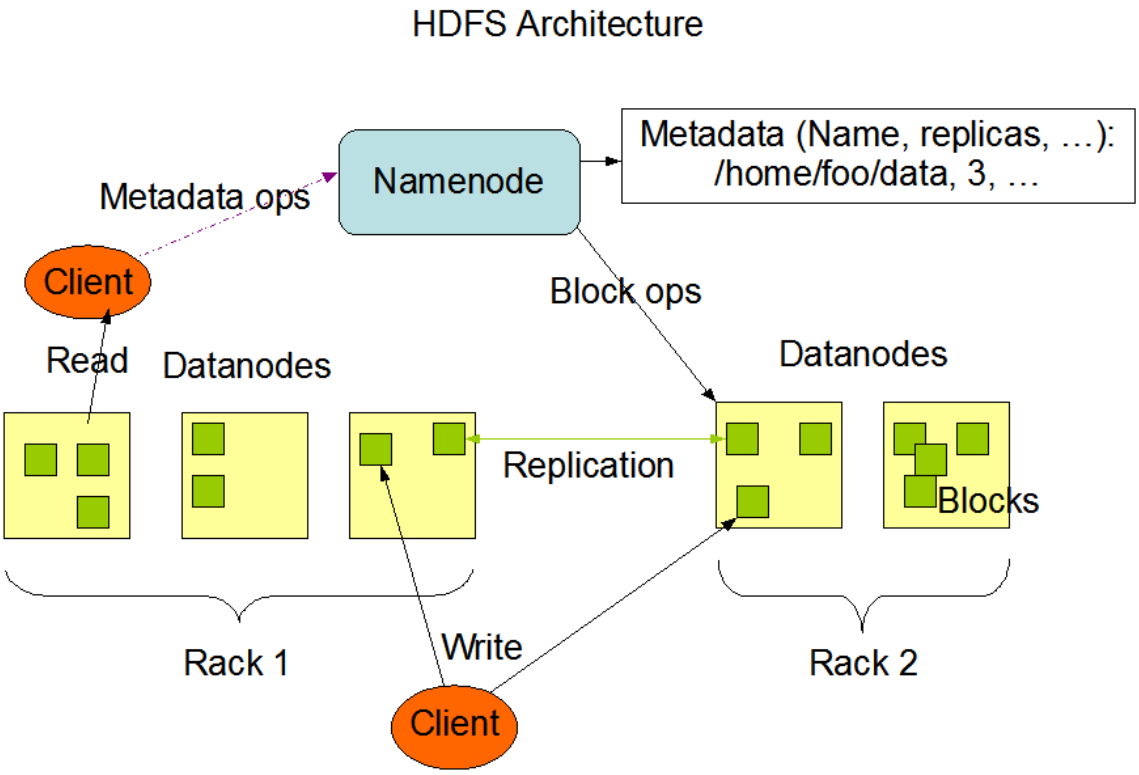


Figure 2: HDFS Architecture

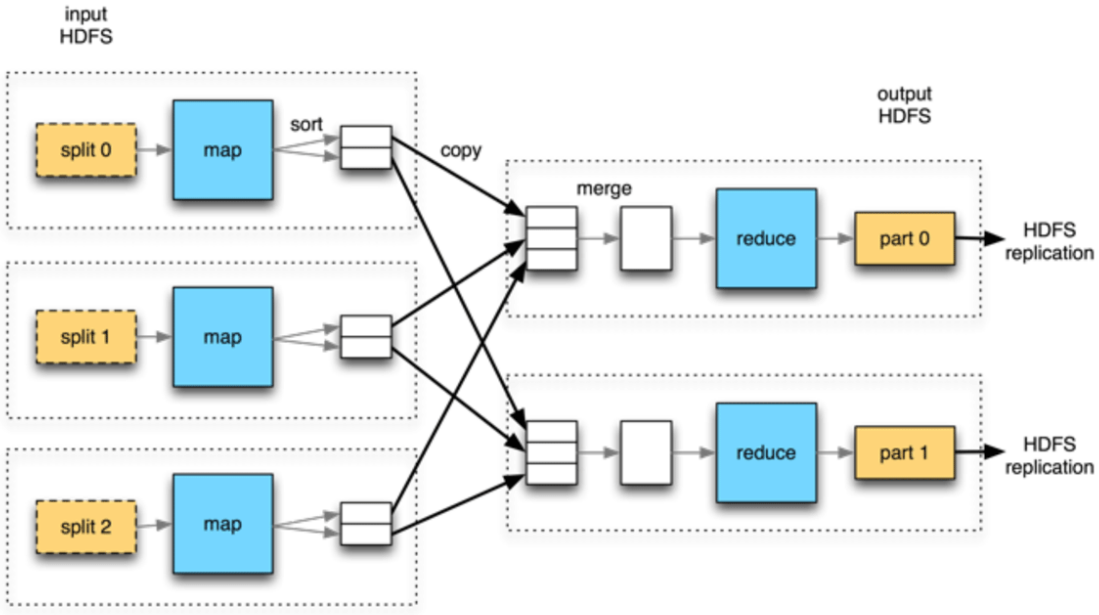


Figure 3: MapReduce

# 3 Big Data and Analytics Problem

## 3.1 Batch Processing

## 3.2 Near Realtime Processing

## 3.3 Realtime Processing

# 4 IT Operation Challenges

# 5 Conclusion

# 6 Future Work

# References

Borthakur, D., Gray, J., Sarma, J. S., Muthukkaruppan, K., Spiegelberg, N., Kuang, H., ... others (2011). Apache hadoop goes realtime at facebook. In *Proceedings of the 2011 acm sigmod international conference on management of data* (pp. 1071–1080).

Dean, J., & Ghemawat, S. (2008). Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, *51*(1), 107–113.

Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Byers, A. H. (2011). Big data: The next frontier for innovation, competition, and productivity.

Shvachko, K., Kuang, H., Radia, S., & Chansler, R. (2010). The hadoop distributed file system. In *2010 ieee 26th symposium on mass storage systems and technologies (msst)* (pp. 1–10).

White, T. (2012). *Hadoop: The definitive guide*. " O'Reilly Media, Inc.".