Volume 6, Issue 11, November 2016 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

# Analysis of Big Data with Hadoop Framework

**Swarupkumar Shivaji Suradkar**
Department Of Computer Engineering, Pune Institute of Computer Technology, Savitribai Phule Pune University, Pune, Maharashtra, India

***Abstract*— Big Data concern large volume complex data sets generated from an autonomous sources. With the fast growing of networking, Big Data are now rapidly expanding in all engineering and technology domains, including physical and medical sciences. A large volume datasets found in Big Data projects is difficult to analyze by using conventional databases, visualization tools and statistical software. Hadoop is an open source tool running on tens or thousands of nodes is more suitable for handling Big Data challenges. Web Data Commons comprises web pages crawled from the Internet. Parsing and Querying on large datasets has become easier with Big Data technologies such as Hadoop framework. Large volume public datasets are more available and that can be found on the Amazon Web Service (AWS).This paper explains the experimental work on Big Data challenges and its optimal solution using Hadoop framework.***

***Keywords*— Hadoop framework, HDFS, Map Reduce, Hive, Sqoop, AWS**

## I. INTRODUCTION

The size of the database has been growing at exponential rates day by day. Simultaneously, there is a need to store, process and analyze the large volumes of complex data for business decision making. In several scientific and business applications, there is a need to store and process terabytes of data in efficient manner on daily bases. The Big Data challenges faced by the industry due to the inability of statistical software tools and conventional database systems to store and process the Big Data sets within tolerable time limits. Analyze and processing of data can include various operations depending upon usage like parsing, querying, searching, indexing, highlighting, tagging, etc. It is not possible for single or few machines to store and process this large volume of data in a finite time period. Hadoop is an open source framework which stores large volume of data sets using Hadoop Distributed File System (HDFS) and process the data sets in parallel using Map Reduce programming framework. Hive is a data warehouse infrastructure built on top of Hadoop framework for providing data summarization, query, and analysis. Hive gives an SQL-like interface to query data stored in various databases and file systems that integrate with Hadoop framework. Sqoop is a Hadoop component designed for efficiently transferring bulk data between Hadoop framework and structural database such as relational databases.

## II. METHODOLOGY

For performing the Big Data experiments, setup of Hadoop single or multi node cluster comprising of Name node and Data nodes. Before moving to multi node cluster, single node cluster was first configured and tested. In a cluster one node was configured as Name node (Master Node) and other nodes were designated as Data nodes (Slave Node). The Master Node runs the "master" daemons: Name Node for the HDFS storage layer and Job Tracker for the Map Reduce processing layer. The slave nodes run the "slave" daemons: Data Node for the HDFS layer and Task Tracker for Map Reduce processing layer. The Master Node is also used as Slave Node to increase the processing nodes. Hadoop Distributed File System (HDFS) is used for storing large volumes of data while Map Reduce framework is used for processing large volumes of data (Big Data).

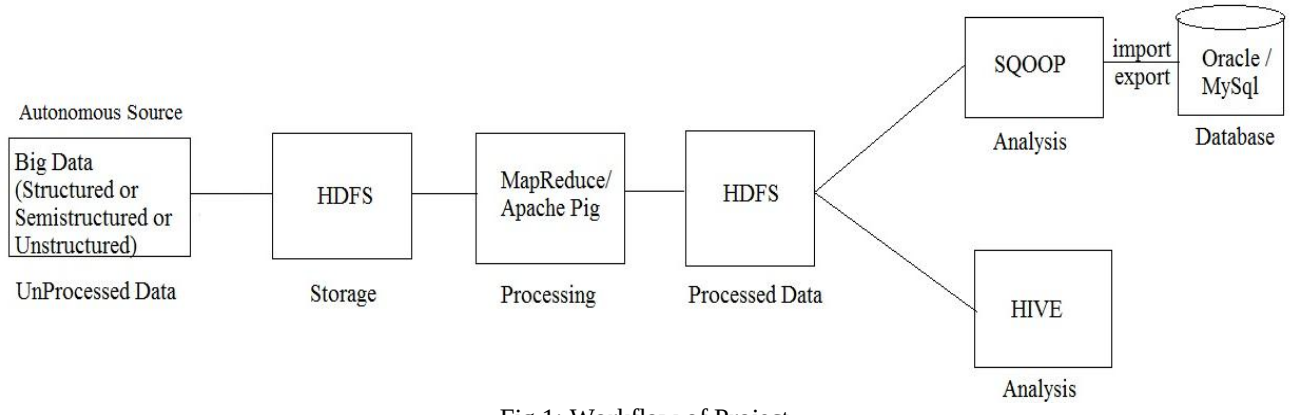


Fig 1: Workflow of Project

 

Hadoop has too many configuration parameters to describe here, but the most relevant for the purpose of this evaluation is the number of Map and Reduce tasks that are allowed to run on each node. Each Map Reduce program that is run is partitioned into M map tasks and R reduce tasks. Input and output data for the Map Reduce programs is stored in HDFS. The output of Map Reduce is used for analysis purpose and it is done by using Hadoop componants like HIVE, SQOOP.

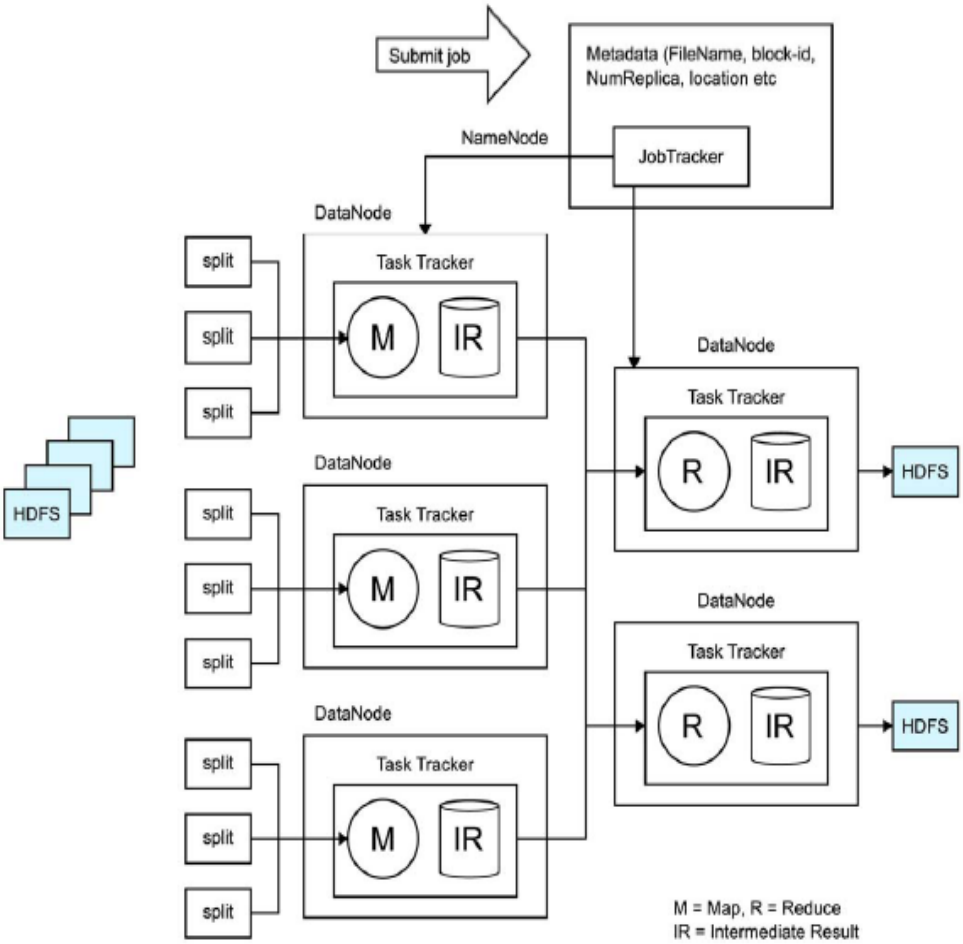


Fig 2: Hadoop Framework

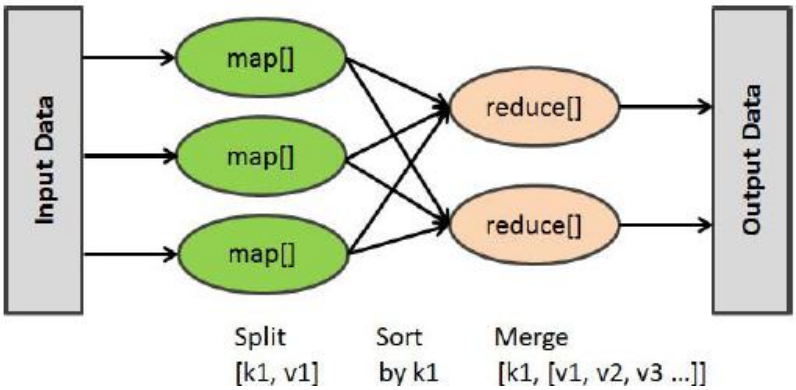


Fig 3: Map Reduce Framework

**An algorithm comprises the following three phases for processing Big Data by using Map Reduce framework.**

**1. Mapper Phase**

Map (K1,V1)→ List (K2,V2)

**2. Sort and Shuffle phase (framework driven phase)**

Output of Mapper Phase (List(K2,V2)) →(K3,List(V3))

**3. Reducer Phase**

Reduce (Output of Sort and Shuffle phase (K3,List(V3)))→List(K4,V4)

**Algorithm:**

map(input_record) {

...

 

```
Output (key 1, value 1)
...
Output (key 2, value 2)
...
}
reduce (key, value) {
aggregate = initialize()
while (values.has_next) {
aggregate = merge(values.next)
}
collect(key, aggregate_values)
}
```

## III. CONCLUSIONS

The most fundamental challenge for Big Data applications is to store, process and analyze the large volumes of data and extract useful information or knowledge for future actions. We have done implementation of Hadoop cluster, HDFS storage and Map Reduce framework for storing and processing large volumes of data sets by considering prototype of Big Data application scenarios. This paper explains the experimental work on big data problem and its optimal solution using Hadoop framework. The results obtained from various experiments indicate favorable results of above approach to address Big Data problem.

## REFERENCES

[1] Ted Garcia and Taehyung ("George") Wang,"Analysis of Big Data Technologies and Methods," 2013 IEEE Seventh International Conference on Semantic Computing.

[2] Xindong Wu,Fellow,Xingquan Zhu, Senior Member, Gong-Qing Wu, and Wei Ding, Senior Member, " Data Mining with Big Data," IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 26, NO. 1, JANUARY 2014.

[3] Weikuan Yu, Member, IEEE, Yandong Wang, and Xinyu Que,"Design and Evaluation of Network-Levitated Merge for Hadoop Acceleration" IEEE TRANSACTIONS ON PARALLEL AND DISTRIBUTED SYSTEMS, VOL. 25, NO. 3, MARCH 2014.

[4] Yanfeng Zhang, Shimin Chen, Qiang Wang, and Ge Yu, Member, "i2MapReduce: Incremental MapReduce for Mining Evolving Big Data ",IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 27, NO. 7, JULY 2015.

[5] Aditya B. Patel, Manashvi Birla and Ushma Nair, " Addressing Big Data Problem Using Hadoop and Map Reduce," NIRMA UNIVERSITY INTERNATIONAL CONFERENCE ON ENGINEERING, NUiCONE-2012, 06-08DECEMBER, 2012..

[6] A. Amer,D.D. E. Long, and R. C.Burns, "Group-based management of distributed file caches," in *Proc. 22nd Int. Conf. Distrib. Comput. Syst.(ICDCS'02)*,Washington,DC, USA, 2002, p. 525, IEEE Computer Socity.

[7] D. Yuan, Y. Yang, X. Liu, and J. Chen, "A data placement strategy in scientific cloud workflows," *Future Gener. Comput. Syst.*, vol. 26, pp. 1200–1214, Oct. 2010..

[8] S. Tripathi and R. S. Govindaraju, "Change detection in rainfall and temperature patterns over india," in *Proc. 3rd Int. Workshop on Knowledge Discov. Sens. Data, SensorKDD'09*, New York, NY, USA, 2009, pp. 133–141, ACM.

[9] X. Jiong, Y. Shu, R. Xiaojun, D. Zhiyang, T. Yun, J. Majors, A. Manzanares, and Q. Xiao, "Improving mapreduce performance through data placement in heterogeneous hadoop clusters," Apr. 2010.

[10] S. Sehrish, G. Mackey, J. Wang, and J. Bent, "Mrap: A novel mapreduce-based framework to support HPC analytics applications with access patterns," in *Proc. 19th ACM Int. Symp. High Perform. Distrib. Comput. HPDC'10*, New York, NY, USA, 2010, pp. 107–118, ACM.

[11] Qi Chen, Cheng Liu, and Zhen Xiao, Member," Improving MapReduce Performance Using Smart Speculative Execution Strategy", IEEE TRANSACTIONS ON COMPUTERS, VOL. 63, NO. 4, APRIL 2014.

 