International Journal of Innovative Research in Advanced Engineering (IJIRAE) ISSN: 2349-2163
Issue 4, Volume 2 (April 2015) www.ijirae.com

BIG DATA ANALYSIS USING RHADOOP

HARISH D* ANUSHA M.S Dr. DAYA SAGAR K.V
ECM & KLUNIVERSITY *ECM & KLUNIVERSITY* *ECM & KLUNIVERSITY*

Abstract— In this electronic age, increasing number of organizations are facing the problem of explosion of data and the size of the databases used in today's enterprises has been growing at exponential rates. Data is generated through many sources like business processes, transactions, social networking sites, web servers, etc. and remains in structured as well as unstructured form. Processing or analyzing the huge amount of data or extracting meaningful information is a challenging task. The way big data - heavy volume, highly volatile, vast variety and complex data - has entered our lives, it is becoming day by day difficult to manage and gain business advantages out of it. This paper describes as what big data is, how to process it by applying some tools and techniques so as to analyze, visualize and predict the future trend of the market.

Keywords— Big Data Problem, Hadoop cluster, Hadoop Distributed File System, MapReduce and R language.

I. INTRODUCTION

Big data is a buzzword, or a catch-phrase, used to describe the massive volume of both structured and unstructured data which is difficult to process using traditional relational database and software techniques as per organization's hardware and infrastructure.

Three major attributes as:
Variety – different type of data including text, audio, video, click streams, log files, and more which can be structured, semi-structure or unstructured.
Volume - hundreds of terabytes and petabytes of information.
Velocity – Speed of data to be analysed in real time to maximize the data's business value.

Figure 1: Attributes of Big Data

1.1 Apache Hadoop

The Apache Hadoop software is a framework that allows for the distributed processing of large data sets across clusters of computers using a thousands of computational independent computers and petabytes of data. Hadoop was derived from Google's Map Reduce and Google File System (GFS).

1.1.2 HDFS (Hadoop Distributed File System):

The Hadoop Distributed File System (HDFS) is a distributed file system providing fault tolerance and designed to run on commodity hardware. HDFS provides high throughput access to application data and is suitable for applications that have large data sets.

1.1.3 HBASE:

HBase is column-oriented database management system that runs on top of HDFS. It is well suited for sparse data sets, which are common in many big data use cases.

1.2 Some samples of data size of few leading companies

Data Generated by FACEBOOK in one day

- 2.5 billion Content items shared per day
- 2.7 billion Likes per day.
- 300 million photos uploaded per day.
- 100+ petabytes of disk space in one of FB's largest Hadoop (HDFS) clusters.

105 terabytes of data scanned via Hive, Facebook's Hadoop query language, every 30 minutes.
70,000 queries executed on these databases per day.
500+terabytes of new data ingested into the databases every day.

Data Generated by FACEBOOK in every 15 days
37.5 billion Content items shared per 15 days.
40.5 billion Likes per 15 days.
40% projected growth in global data generated per 15 days vs 5% growth in global IT spending.

1.3 Let us look out what is big data analysis, need of analysis and how we can do analysis in optimized way through different approaches.
We need to store, clean, brush up, apply some mathematical and algorithmic model to research so beautify our [data] to induce a visualized and delightful story out of it for the senior management team to take some. We need to store, clean, brush up, apply some mathematical and algorithmic model to research so beautify our [data] to induce a visualized and delightful story out of it for the senior management team to take some decisions.

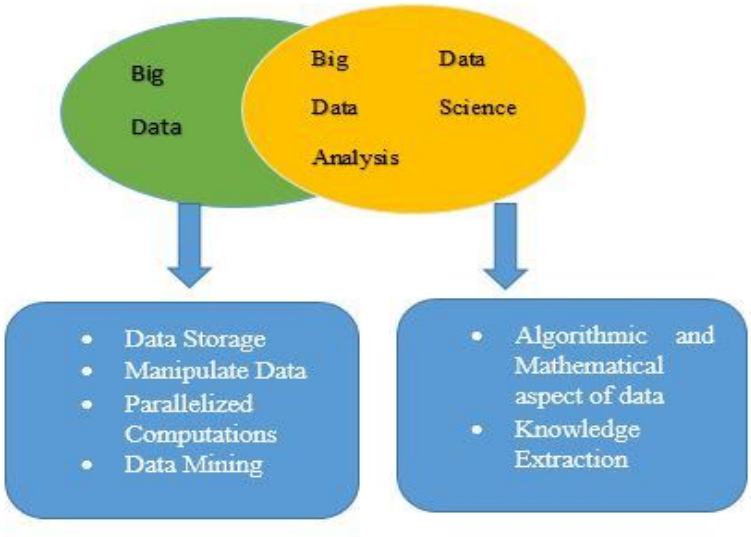


Fig 2: Big Data Analysis

1.3.1 **Need**

Can you imagine how your organization handles Big Data during daily operations? Just to give you an idea, consider the following scenarios:

"Can you quickly draw a graph of sentiment analysis for fundamentals news article revealed yesterday for a XYZ region?" What one ought to do therefore as you'll be able to answer your VP quickly by scanning your massive data?

1.3.2 **How**

Previously, it was the statisticians whom to play with data and come up with some models to reach out to some decision. Data person could be a mixed mix of a knowledge base knowledgeable, a statistician and a story author. In order to make their life easier we have R language wherein we can either store information or use existing information (from some info like SQL server or oracle) so we will perform our analysis victimization some predefined packages among R.
R will handle huge information victimization ff, fbase, RODBC, RHadoop packages.

II. Big Data Analysis using R

2.1 Introduction to R

R is an open source language which is used for data modelling, manipulation, statistics, forecasting, time series analysis and visualization of data. Gradually R got some libraries like ff, ffbase, Rodbc, rmr2 and Rhdfs to handle huge information. Rmr2 and rhdfs along use the facility of Hadoop so as to handle huge information effectively.
The latest version of R is going to be R 3.0.2

2.2 Big Data Analysis in R using ff, ffbase Packages

Native R stores everything into RAM. R objects will take memory up to 2-4 GB, depends on hardware configuration.

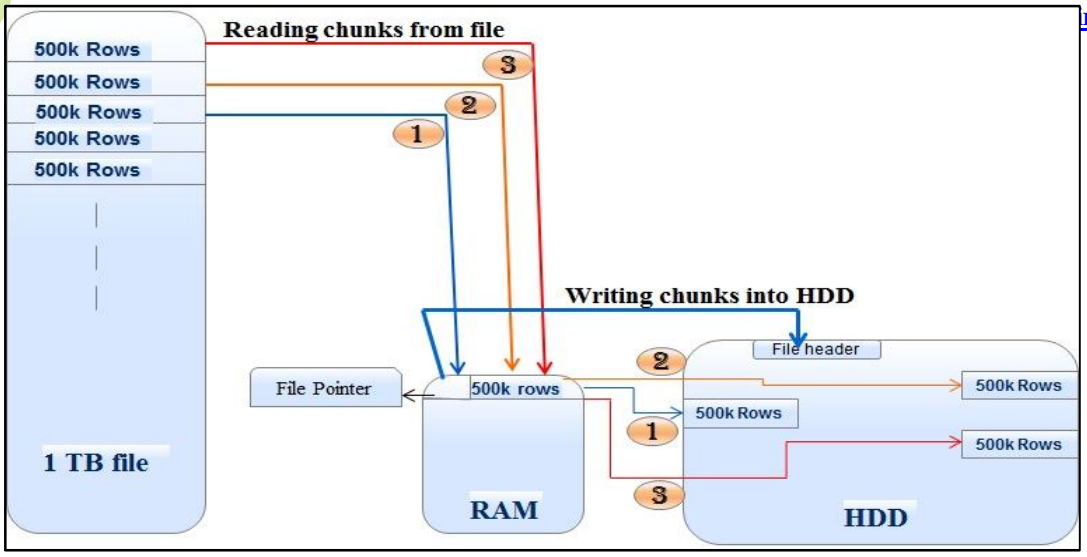


Fig 3: Functioning of ff package

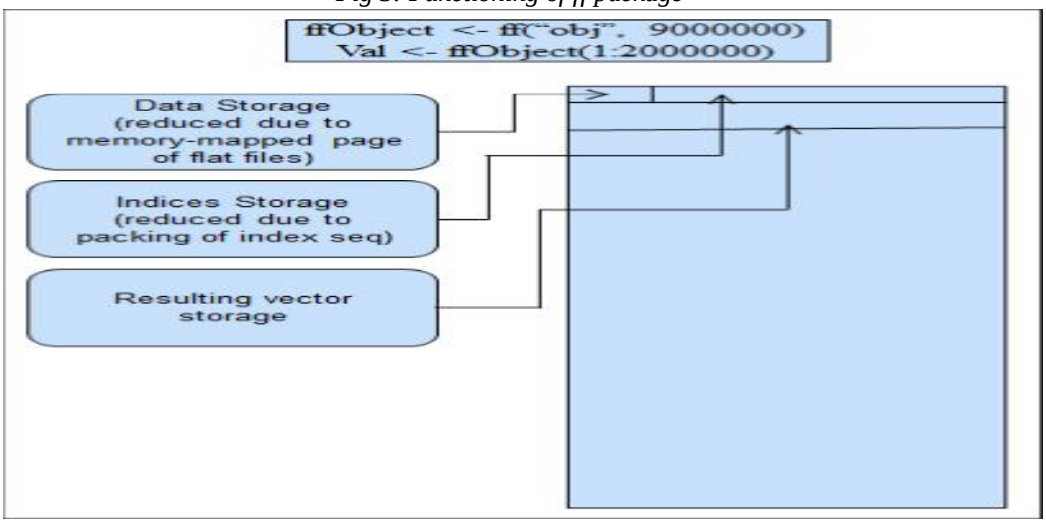


Fig 4: Data Storage with ff object

2.3 RODBC:

Hey don't worry if you've got keep your knowledge in any on-line database like SQL server. R has a privilege to connect to the SQL server (using RODBC package) and pick your data from there itself with quite an ease.

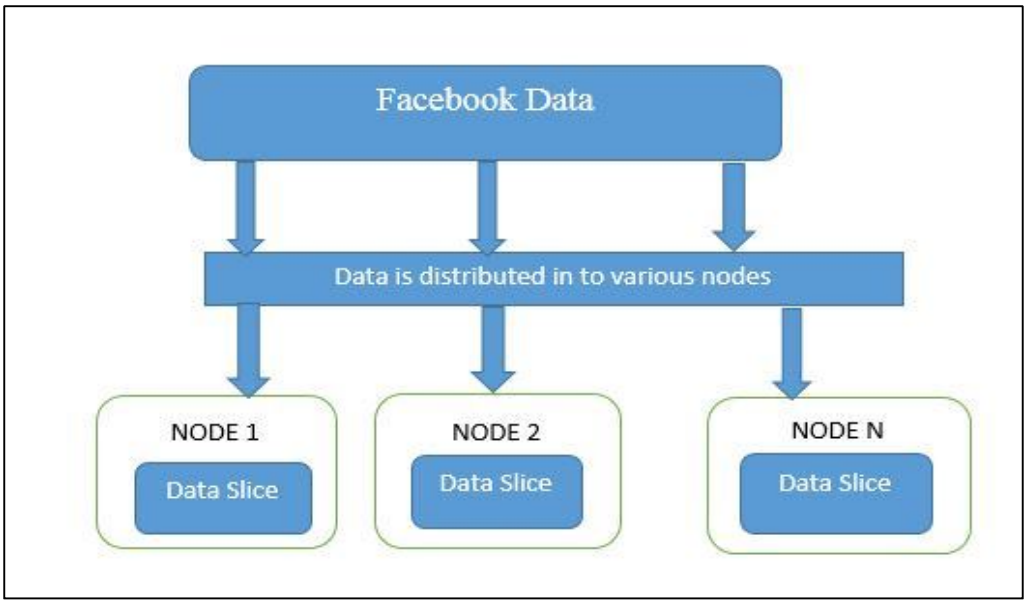


Fig 5: RODBC Connectivity

III. Introduction to Hadoop

Hadoop is an open source Apache software written in JAVA for running distributed applications on big data. It contains a distributed file system namely Hadoop distributed file system (HDFS) and a parallel processing batch framework.

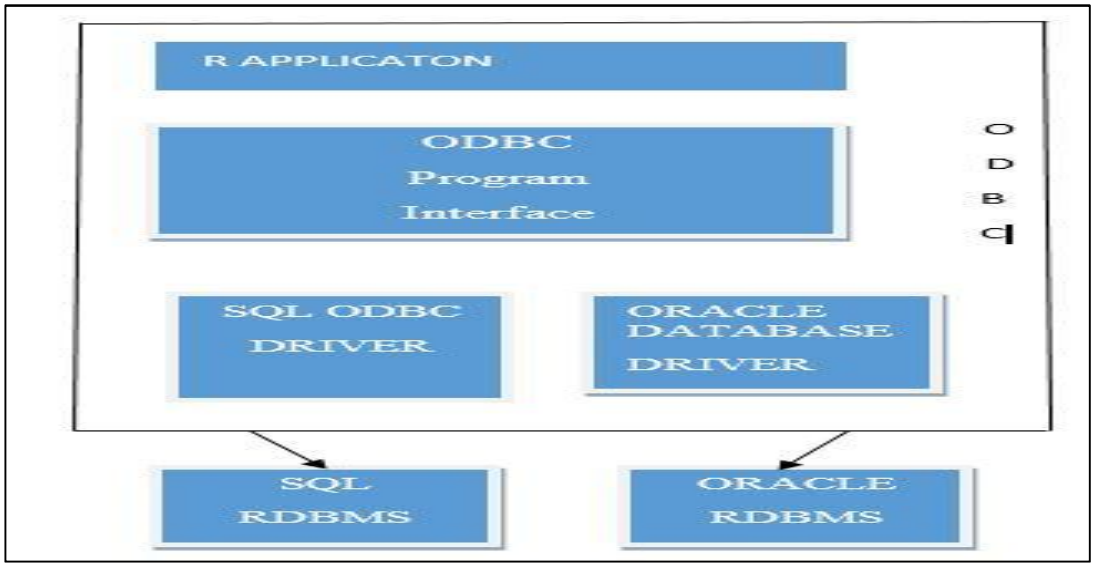
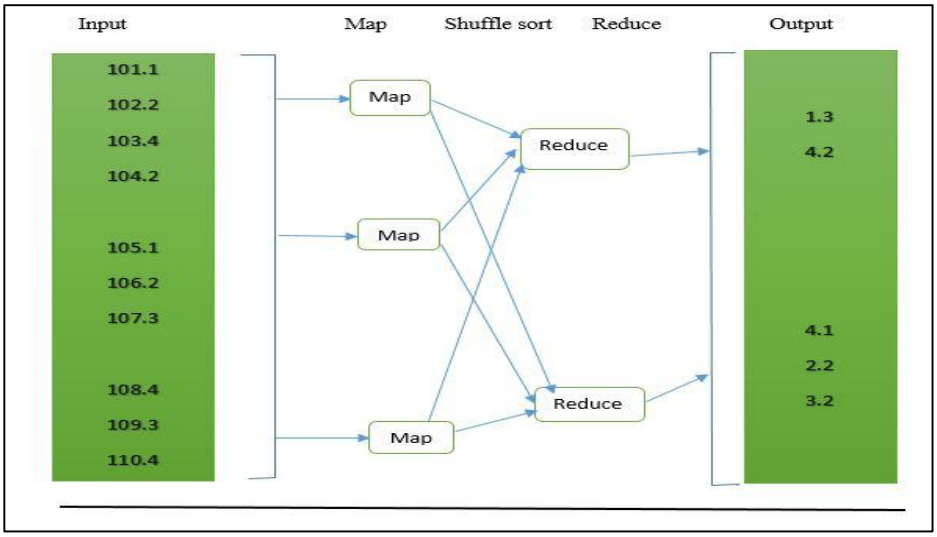


Fig 6: Example of distribute data system

Hadoop provides a great sense of data reliability and movement across the nodes in cluster. The core functionality lies with MapReduce which is a popular computational algorithm. MapReduce takes the approach of divide and conquer wherein the data is divided based on some mapping function and is then applied to various nodes within the cluster so as to execute parallel. Hadoop and MapReduce provides a high level of fault tolerance wherein by default data is replicated at three different data nodes (Slave nodes). Hadoop distribute its tasks and data into different nodes and each node is responsible for execution of tasks and processing of data and send result back to main node. According to The Apache Software Foundation, the primary objective of HDFS is to store data reliably even in the presence of failures including NameNode failures, DataNode failures and network partitions. The NameNode is a single point of failure for the HDFS cluster and a DataNode stores data in the Hadoop file management system. Facebook has more than 100 PB (PB= 1 M GB) of data in Hadoop clusters.

3.1 MapReduce – Data Reduction

MapReduce is a computational model working on divide and conquer approach using the key value pair wherein Map function divide the data set into subset and then the results are reduced to get the final output using the key. Hadoop has the power for processing the data efficiently but it does not have the powerful capability for analysis. To overcome this we require Hadoop to be integrated with some statistical language like R.



3.2 Big data analysis using RHadoop

RHadoop is a collection of three R packages: rmr, rhdfs and rhbase. . rmr require Rcpp, RJSONIO, bitops, digest, functional, stringr, plyr, reshape2. Rhdfs require rJava package. We need to put in these packages before install rmr and rhdfs severally. Rmr package provides Hadoop MapReduce practicality R, rhdfs HDFS provides file management and rhbase provides HBase management from among R. Below mentioned packages provides United States the practicality of Hadoop among R

Rmr2 - rmr2 offer United States Hadoop MapReduce practicality in R.

Rhdfs - rhdfs offer United States file management of the HDFS with R

Rhbase - rhbase offer United States management for the HBase distributed information with R.

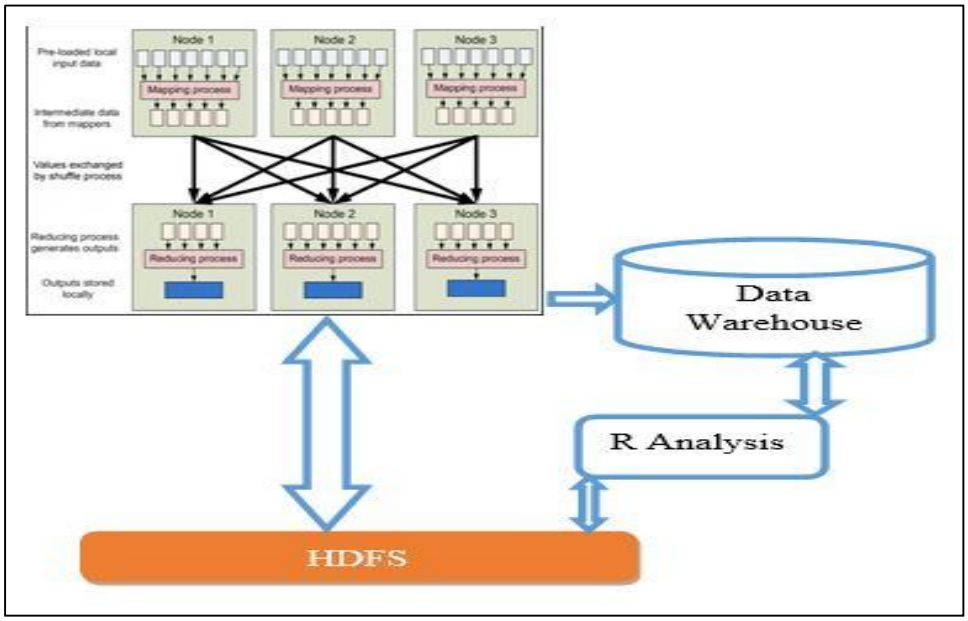


Fig 8: RHadoop Architecture

MapReduce takes the divide and conquer approach, which runs in parallel.

IV. **Conclusion**

RHadoop is complete set where we can process our data efficiently, perform some meaningful analysis. It removes the dependency of temp files that we used to do with ff package and any middle layer that has the limitation of memory. Finally, there is the approach of developing algorithms that have been explicitly parallelized to run within Hadoop. For example if you wanted to do a linear or logistic regression in R on a 1TB of data stored in HDFS, this requires that the algorithms themselves be implemented in way to use a distributed computing model. Revolution Analytics has a framework for developing these kinds of algorithms to be optimized within Hadoop.

We have examined the design and architecture of Hadoop's MapReduce framework in great detail. Particularly, our analysis has focused on data processing. We would conclude by saying that big data is the new buzz word and Hadoop MapReduce is the best tool available for processing data and its distributed, column-oriented database, HBase which uses HDFS for its underlying storage, and support provides more efficiency to the system. With Revolution Analytics' RHadoop packages and MapR's enterprise grade Hadoop distribution, data scientists can utilize the full potential of Hadoop from the familiar R environment. If someone needs to combine strong data analytics and visualization features with big data capabilities supported by Hadoop, it is certainly worth to have a closer look at RHadoop features. It has packages to integrate R with MapReduce, HDFS and HBase, the key components of the Hadoop ecosystem. For more details, please read the R and Hadoop Big Data Analytics whitepaper.

ACKNOWLEDGEMENT

This work is supported by KLUNIVERSITY. We express deepest gratitude to our project guide Assoc.Prof. Dr.K.V.Daya Sagar, Head of the Department Prof. M.Suman. We would like to thank our Principal Dr. A.ANAND KUMAR who provide a healthy environment for all of us to work in best possible way. We also express our deep gratitude towards all the people who have helped us to completion of this project successfully.

REFERENCES

1. Data Mining with BigData by Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, 1041-4347/13/$31.00 © 2013 IEEE.
2. Ahmed and Karypis 2012, Rezwan Ahmed, George Karypis, Algorithms for mining the evolution of conserved relational states in dynamic networks, *Knowledge and Information Systems*, December 2012, Volume 33, Issue 3, pp 603- 630.
3. Apache Hadoop Project, http://hadoop.apache.org/, 2013.
4. Jiang, B.C. Ooi, L. Shi, and S. Wu, "The Performance of MapReduce: An In-Depth Study," Proc. VLDB Endowment, vol. 3, no. 1, pp. 472-483, 2010.
5. Bughin et al. 2010, J Bughin, M Chui, J Manyika, Clouds, big data, and smart assets: Ten tech enabled business trends to watch, McKinSey Quarterly, 2010.
6. Gillick et al., 2006, Gillick D., Faria A., DeNero J., MapReduce: Distributed Computing for Machine Learning, Berkley, December 18, 2006.
7. IBM 2012, What is big data: Bring big data to the enterprise, http://www-01.ibm.com/software/data/bigdata/, IBM. [10] Michel F. 2012, how many photos are uploaded to Flickr?
8. http://www.flickr.com/photos/franckmichel/6855169886/
9. J. Dean and S. Ghemawat, "MapReduce: Simplified Data Processing on Large Clusters," Proc. Sixth Symp. Operating System Design and Implementation (OSDI '04), pp. 137-150, Dec. 2004.
10. D. Jiang, B.C. Ooi, L. Shi, and S. Wu, "The Performance of MapReduce: An In-Depth Study," Proc. VLDB Endowment, vol. 3, no. 1, pp. 472-483, 2010.
11. T. Condie, N. Conway, P. Alvaro, J.M. Heller stein, K. Elmeleegy, and R. Sears, "MapReduce Online," Proc. Seventh USENIX Symp. Networked Systems Design and Implementation (NSDI), pp. 312-328, Apr.
12. http://cran.rproject.org/web/packages/ff/ff.pdf
13. http://cran.rproject.org/web/packages/ffbase/ffbase.pdf
14. http://cran.rproject.org/web/packages/RODBC/ RODBC.pdf
15. http://rhandbook.wordpress.com/tag/rodbc/
16. http://cran.rproject.org/web/packages/HSAUR/vignettes /Ch_introduction_to_R.pdf
17. https://rhandbook.wordpress.com/tag/ffbase/
18. Wal TV. Folksonomy coinage and definition' http://www.vanderwal.net/folksonomy.html (2007, accessed 25 April 2011)
19. Bao S, Xue G, Wu X, Yu Y, Fei B, Su Z. Optimizing web search using social annotations. In: Proceedings of the 16th International World Wide Web Conference. Banff, Alberta, Canada, 8–12 May 2007, pp. 501–510.
20. Yanbe Y, Jatowt A, Nakamura S, Tanaka K. Can social bookmarking enhance search in the web? In: Proceedings of the 7th ACM/IEEE-CS Joint Conference on Digital Libraries. Vancouver, Canada, 17–22 June 2007, pp. 107–116.

