ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

A Study on Performance Enhancement of Cloud-based Bigdata Analytics

Shilpa B L[1], Basant Kumar[2], Abhisek Singh[3], Arjun Saxena[4,] Akash Kumar Keshri[5]

Assistant Professor, Dept. of CSE., SJBIT, Bengaluru, Karnataka, India[1]

8th Semester Student, Dept. of CSE., SJBIT, Bengaluru, Karnataka, India[2,3,4,5]

ABSTRACT: An open source framework called Hadoop, implementation of MapReduce provides efficient platform for BigData analytics. The performance of Hadoop MapReduce mainly depends on its configuration parameters. Tuning the job configuration parameters is an effective way to improve performance so that we can reduce the execution time and the disk utilization. The performance tuning mainly based on CPU usage, disk I/O rate, memory usage, network traffic components.
The requirement to perform complicated statistical analysis of big data by institutions of engineering, scientific research, health care, commerce, banking and computer research is immense. However, the limitations of the widely used current desktop software like excel, minitab and SPSS gives a researcher limitation to deal with big data. The big data analytic tools like IBM BigInsight, Revolution Analytics, and tableau software are commercial and heavily license. Still, to deal with big data, client has to invest in infrastructure, installation and maintenance of hadoop cluster to deploy these analytical tools. Apache Hadoop is an open source distributed computing framework that uses commodity hardware. In this paper, we intend to collaborate Apache Hadoop and R software over the on the Cloud. Objective is to build a SaaS (Software-as-a-Service) analytic platform that stores & analyzes big data using open source Apache Hadoop and open source R software.

KEYWORDS: Apache Hadoop; cloud; SaaS; Bigdata Analytics

I. INTRODUCTION

This CBA (Cloud Based Big Data Analytics) is a system to deal with big data to perform linear regression and similar predictive analysis with ease and prove to be very helpful for engineering research, business, health care, scientific research, banking & finance and machine learning where complicated statistical analysis need to be performed. Analysis of large data is very complicated for traditional analytic environment which can be done with ease in distributed environment without undermining the quality of the result. Statistics based on these customer feedback data will help expand businesses and a company that has such data to its disposal, surely has a far stronger feel on the pulse of the market. To perform such analysis for Small Medium.
Significant role on the performance, i.e. even a small change to one configuration parameter value makes a huge difference to performance, when running the same MapReduce job with the same size of data input. Job configuration of MapReduce model is black box kind feature, so it is difficult to find a straightforward mathematical model to the cluster configuration for a specific job. And same type of configuration is not applicable for all kinds of MapReduce jobs. This experiment started by creating a performance baseline for the system. We kept the Hadoop default configuration settings saying it as baseline of the system and compared this result with new methods job configuration. In our experiment we chosen most important parameters of Hadoop such as number of map slots, reduce slots, buffer size and compressing the mappers output. We proposed new methods to optimize these Hadoop parameters, so that overall performance can be improved.

II. METHODOLOGY

The methodologies described in this paper are based on experience of designing and configuration of Hadoop systems for different parameters. In this section, we provide a brief overview of Hadoop architecture i.e. HDFS and MapReduce and proposed new methods to tune the Hadoop parameters. The methodology presented here is universally applicable

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

to any Hadoop performance optimization operation. Hadoop Distributed File System (HDFS) is the distributed file system used by the Hadoop project, which is popular because of its scalability, reliability and storing capacity of very large

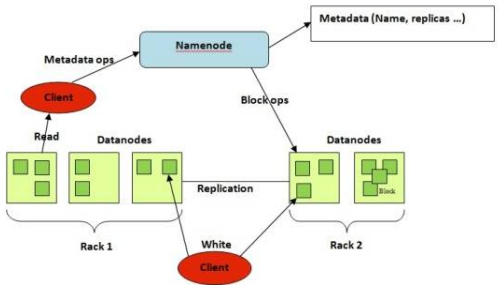

Fig.1 HDFS Architecture

files. HDFS is of Master/Slave architecture. Hadoop deployment has a single NameNode, which is the master and a set of DataNodes, which serve as slaves. The NameNode is unique in HDFS cluster which manages files meta information and stores in memory, thus limiting the number of files that can be stored by the system. HDFS makes one metadata file and several data blocks for a file. DataNodes are for actual data. Data blocks are distributed over DataNodes, by default, data blocks are replicated in HDFS, for higher chance of data locality and fault-tolerance. Compared to other existing distributed file systems, Hadoop Distributed File System (HDFS) provides high throughput access to data. It is designed to work on low-cost commodity hardware. Normally, a TaskTracker and a DataNode is a pair installed in a node. Fig 1 shows typical HDFS architecture.

MapReduce model:

MapReduce a programming model from Google, IS used to solve the problems across huge datasets in the multi node clusters environment. There are two main functions involved in a MapReduce framework, Map and Reduce function. The Map function takes in a key/value pair and outputs an intermediate key/value pairs. The Reduce function will then takes all values associated to the same key and produce the final output. Actual process, in Map Phase the master node called 10btracker divides the job and distributes this sub jobs to slave nodes called Tasktrackers. Tasktrackers will process the sub jobs and passes the answer back to its master node. In the Reduce step, master node combines the answer from the slave nodes to get a solution for main job. In Fig.2 we can seethe Map Reduce process Tuning Process There are more than 180 parameters available for us to manipulate in the Hadoop MapReduce framework to get better usage of resources. Depending on the way in which they make an impact on the performance of MapReduce job, configuration parameters are divided into three patterns: Core-related parameters, MapReduce relevant parameters and HDFS-relevant parameters. The Hadoop framework uses configuration files for setting the values of these parameters in each group. 1. Core related parameters: These are used for defining the most important features of a MapReduce cluster.The parameters in this group are associated only to the cluster itself such as where the temporary data is stored, how large the buffer size e.g. size of sort buffer, and what the threshold of shuffle group is, maximum merge of spill files etc. 2. MapReduce-relevant parameters: The parameters in this group are relevant to the MapReduce procedure:

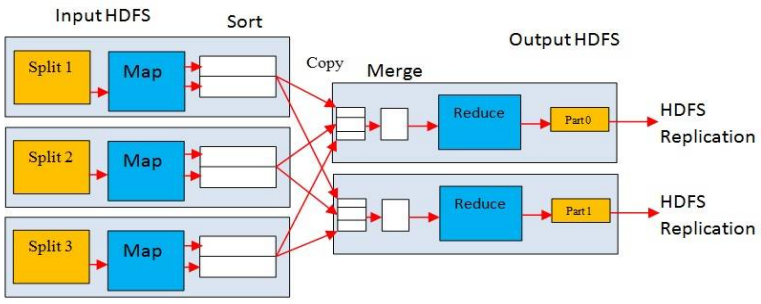

Fig.2 MapReduce Model

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

Some of them have a direct effect only on the Map phase or Reduce phase, while others may have an effect on both phases. E.g. number of map task, reduce task, maximum number of map slots, reduce slots etc
HDFS-relevant parameters: The parameters, such as specifying how many replicas should be stored in a cluster, HDFS block size etc. Hadoop performance tuning is an iterative process. Initially we launch a job then analyze resource usage, if it underutilized then adjust the parameter and re-run the job. Repeat this process until we reach the better performance. The following steps describe the process and Fig.3 shows the performance tuning cycle.
1. Run the job first time using the default configuration setting to get the overall system performance. This will be our baseline
2. Modify and tune the job configuration parameter and re-run the job, compare this result with baseline.
3. When the analysis is complete, we can review the results and accept or reject the inferences.
4. Repeat this step 2 till we get shortest execution time for our job.
For our experiment we considered main five parameters and had mentioned how to tune the parameter, we call this as new methods to tune the parameters.

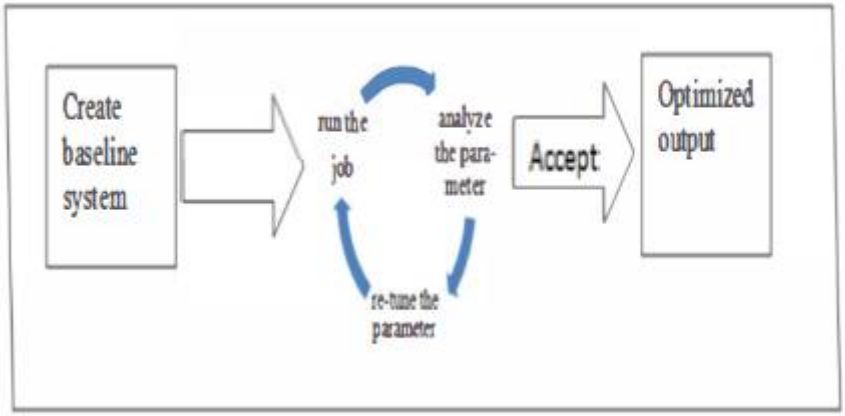

Fig.3.Performance tuning cycle.

III. SYSTEM MODEL OF CBA

Cloud Based BigData Analytics (CBA), constitutes four vital components:

- R statistical software to perform statistical analysis
- Hadoop framework to distribute data and compute tasks in the cluster computing
- Rhadoop to link R and Hadoop
- Amazon EMR to deploy system to provide SaaS.

Access to the CBA can be made by any data set that is a part of the cluster and a user can crawl through the clouds multiple files and folders collecting required data to analysis and store it back into the cloud as desired.

A. Web Application & Web Services for Accessing Analytical Services:

CBA is designed in such a way that anyone with basic computer knowledge can utilise it conveniently. User can upload data, select analysis which they want to perform and run the analysis in browser interface and view the result as well as store it.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

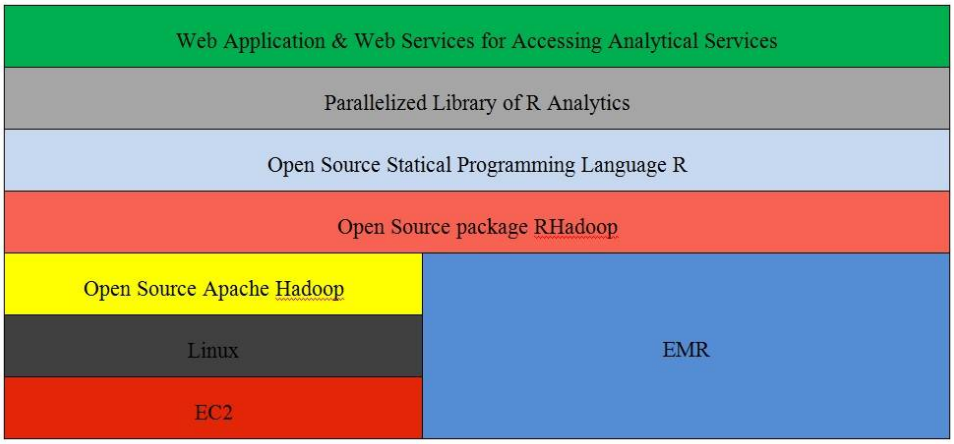

Figure 4. System Model

The features of web application are:
• Authenticity check
• Convenient
• Data upload and file management
• Select analytical feature to perform analysis
• View result etc.

B. Parallelized library of R Analytics:

Open source Apache Hadoop and Open source high level statistical programming language R are collaborated to create a parallelized library of R analytics. Data is read using HDFS, analysis performed using map reduce. The hadoop distributed file system facilitates the user to store, read, and write in hadoop. Since R is incompetent in the statistical analysis of large scale information, it is linked with hadoop to create a more efficient system. One that will prove useful for several industries including stock market, artificial intelligence designing, scientific research and many more where business analytics need to be performed.

C. Open Source Statistical Programming Language R:

R is a free software programming language. It is used by statisticians and data miners for statistical computing, graphics and several such applications. Ross Ihaka and Robert Gentleman created "R" in the year 1993 and it was an instant hit with the general public. This prompted them to make R available to the general public over the internet under the GPL. R soon became more accessible and before one realized, it took over much of computerized statistical analysis. With time, the language developed and R is still the most eligible and updated software programming language in the industry. R was rereleased on 29th February 2000, can be used freely, distributed and sold to any receivers who possess the same rights. R is based on formal computer language and is thus flexible. This makes it the most sought after software programming language available in the market. R is rich in various statistical analysis packages. There are 5922 packages available in CRAN packages repository and the number keeps rising. There are many R users and many forums as well as groups from where one can learn R. Online tutorial support available to those interested in learning the language.

D. Open Source package RHadoop

RHadoop is an open source project aimed at large scale data analysts to empower them to use
the horizontal scalability of Hadoop using the R language. ravro, plymr, rmr, rhdfs and rhbase, the 5 R packages enable its users to manage and analyze massive quantities of information using Hadoop.
• ravro – is the R package which permits the reading and writing of files in avro format, to R
• Plyrmr – is a more recent R package that makes R and Hadoop perform in near perfect if not perfect harmony, in the analysis of higher level plyr like data.
• Rmr - is a R package that came into being to allow users to write map reduce programs in R since it is
more productive and far more easier
• Rhdfs - is a R package that gives administration of HDFS files from within R. It uses Hadoop common to give access to map reduce file services

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

• Rhbase - is a R package that allows its users to connect with hbase and deal with hbase functions.

E. The Open Source Apache Hadoop:

Apache Hadoop is an easy to use, dependable and accurate distributed computing framework. It prompts it's users to shift from single server to multiple connected machines each functioning as a separate unit for data storage and computing. Since Apache Hadoop software library that has several thousand computers over a cluster, accuracy of analysis can be assured, as each unit is so programmed to spot and rectify errors. Hadoop stores massive quantities of data over several systems in the cluster. The solutions it leads us to, are reached through a highly scalable, distributed batch processing system. So, a large workload is distributed over the cluster of several inexpensive datasets and big data analysis is completed in no time.

IV. ARCHITECTURAL MODEL OF CBA

As it is evident from figure 5, the architectural model of CBA permits a user to interact with Amazon S3 as well as hadoop cluster, using the web browser. The environment to perform file management within cloud storage is provided by Amazon S3, cloud storage with the help of which a user can upload files in hdfs or can copy a file to hdfs from Amazon S3 in order to perform mapreduce task in hadoop cluster for high performance. After selecting the required files from hdfs, researchers can opt for analysis whose results are achieved after being performed over several datasets in a cluster since computation occurs in distributed environment thereby optimizing the performance of the system. After completion of analysis, the result of analysis is displayed to user in browser and users can store the result of analysis in cloud storage Amazon S3 for future reference. Amazon EMR allows businessmen, researchers, data analysts, and developers to process vast amounts of data with ease and at a far lesser cost. It uses a hosted Hadoop framework running on the web-scale infrastructure of EC2 and Amazon S3. Since CBA is cloud based, availability is one of the features of this system for users equipped with the internet anywhere around the world, all round the clock

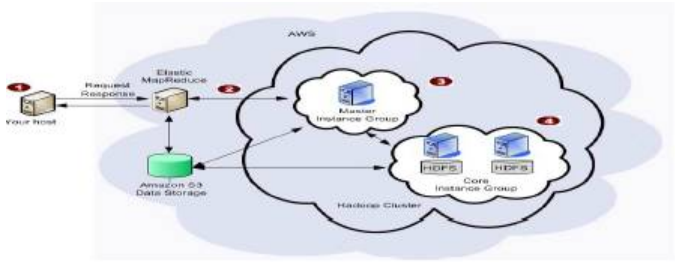

Figure 5. Architectural Model of CBA.

Amazon Elastic MapReduce (Amazon EMR) makes hadoop cluster easy to provision and manage in the AWS Cloud interface. Amazon EMR is available in two different distribution of hadoop, one is Amazon Distribution and next is the MapR Distribution of hadoop. MapR distribution of hadoop comes with additional hadoop application like spark, hive, and so on. MapR distribution also serves support for client and have enhanced many feature of hadoop to make ease-of-use as well as features. MapR distribution of hadoop has good presence in market and most of the leading enterprises like oracle, ibm and so on follow MapR distribution of hadoop. Inorder to perform predictive analytics in cloud platform on the top of MapR distribution of Hadoop Cluster CBA is put to use. Host or Client of CBA can access system after authorization and can perform predictive analysis of bigdata in cloud platform from anywhere with internet connectivity.

V. CONCLUSION AND FUTURE WORK

CBA is a SaaS to analyze bigdata in cloud and is nowadays focused on linear regression and time series analysis. Computing framework distribution, increasing number of nodes in cluster by memory scaling and parallel processing in distributed environment makes it easy to deal with big data and performance thereby decreasing the limitations of processing of large scale data. We now neither have to pick a random sample and end up with an inaccurate result nor do we have to choose vertical scaling and face a dead end. The above discussed model becomes even more powerful if one adds in kmean, correlation, market basket, time series and other such functionalities. The combination of

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 5, May 2016

opensource R and opensource Apache hadoop is from every angle an unmatchable marvel. It is safe, scalable, easily accessible and highly efficient. Additional programs like business intelligence or a link up with systems like oracle or SAP will without doubt make it the best analytic tool available in the industry. If in future the number of nodes could somehow be decreased the model can become far more economical. CBA can be deployed in private cloud using openstack. CBA has hbase installed in it and the integration of CBA with hbase in future will only make it all the more desirable. By integrating database in system, CBA becomes compatible with any SME using any database system. In business, time is money. And if cloud based big data analytics can perform crucial analysis in a short time, big wealth too can be accumulated in a short time.

REFERENCES

1. Roger S. Barga, Jaliya Ekanayake, Wei Lu, "Project Daytona: Data Analytics as a Cloud Service", IEEE 28th International Conference on Data Engineering, 2012.
2. Nikolay Laptev, Kai Zeng, Carlo Zaniolo., "Very Fast Estimation for result and Accuracy for Big Data Analytics: the EARL System", IEEE's, ICDE Conference 2013.
3. *(Accessed on 12th August 2014). Hadoop [Online] Available:http://hortonworks.com*
4. (Accessed on 14th August 2014) Hadoop [Online] Available: http://en.wikipedia.org/wiki/Apache_Hadoop
5. (Accessed on 14th August 2014) Hadoop [Online] Available: http://hadoop.apache.org/
6. (Accessed on 12th June 2014) R [Online] Available: http://www.revolutionanalytics.com
7. Vignesh Prajapati, "Big Data Analytics with R and Hadoop", Packt publication, 2013.
8. (Accessed on 18th January 2014) R [Online] Available: http://cran.rproject.org/
9. Q. Ethan McCallum & Stephen Weston, "Parallel R", O'Reilly, 2012.
10. Josep Adler, "R IN A NUTSHELL", second edition, O'Reilly, 2012
11. .Paul C. Zikopoulos, Chris Eaton, Dirk Deross, Thomas Deutsch.
12. Tom White, "Hadoop:The Definitive Guide, 3rd Edition", O'Reilly,2012.
13. Sudipto Das, Yannis Sismanis Kevin, S. Beyer, Rainer Gemulla,Peter J. Haas, and John McPherson, "Ricardo: Integrating R andHadoop", In Proc. SIGMOD'10, Indianapolis, Indiana, USA , June 6–11, 2010.
14. (Accessed on 16th September 2014) RHadoop [Online] Available:https://github.com/RevolutionAnalytics/RHadoop/wiki
15. Availabe:https://github.com/RevolutionAnalytics/rmr2/blob/master/docs/tutorial.md
16. (Accessed on 26th October 2014) Cran Packages Repository, [Online].Available: http://cran.r-project.org/web/packages/.
17. Antonio Piccolboni, "Rhadoop" [Online]. Available:https://github.com/RevolutionAnalytics/RHadoop/wiki. [Accessed:Oct. 11, 2014]. Available:http://info.mapr.com/rs/mapr/images/The_Forrester_Wave_Big_Data _Hadoop_Q12014.pdf
18. (Accessed on 5th December 2014) EMR [Online]. Available:http://aws.amazon.com/elasticmapreduce.
19. (Accessed on 5th December 2014) Amazon S3 [Online]. Available:http://aws.amazon.com/s3.
20. (Accessed on 5th December 2014) EC2 instances [Online]. Available:http://aws.amazon.com/ec2
21. (Accessed on 3rd October 2014) ff ffbase [Online] Availabel:https://cran.r-project.org/web/packages/biglm/index.html
22. (Accessed on 2nd October 2014) ff ffbase [Online] Available:http://cran.r-project.org/web/packages/ff/index.html

BIOGRAPHY

Shilpa B.L is currently working as Assistant Professor in the Department of Computer Science and Engg in SJB Institute of Technology, Bangalore. She obtained her B.E in Computer Science and Engg in 2009 and M.Tech in Computer Science and Engg from RVCE, Bangalore in 2011. Her research interests are in the area of Text Analytics and Predictive Analytics.