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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

Performance Analysis Model for Big Data Applications in Cloud Computing

Protim Ganguly

Dept. of ECE, Bengal Institute of Technology, Kolkata, West Bengal University of Technology, Kolkata , India.

ABSTRACT: Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources. With the fast development of networking, data storage, and the data collection capacity, Big Data is now rapidly expanding in all science and engineering domains, including physical, biological and bio-medical sciences. This article presents a HACE theorem that characterizes the features of the Big Data revolution, and proposes a Big Data processing model, from the data mining perspective. This data-driven model involves demand-driven aggregation of information sources, mining and analysis, user interest modeling, and security and privacy considerations. We analyze the challenging issues in the data-driven model and also in the Big Data revolution.

KEYWORDS: Information management, Data handling, Data storage systems, Data privacy, Data models, Distributed databases,complex and evolving associations, Big Data, data mining, heterogeneity, autonomous sources

I. INTRODUCTION

Data is the collection of values and variables related in some sense and differing in some othersense. In recent years the sizes of databases have increased rapidly. This has lead to a growinginterest in the development of tools capable in the automatic extraction of knowledge from data[1]. Data are collected and analyzed to create information suitable for making decisions. Hencedata provide a rich resource for knowledge discovery and decision support. A database is anorganized collection of data so that it can easily be accessed, managed, and updated. Data miningis the process discovering interesting knowledge such as associations, patterns, changes,anomalies and significant structures from large amounts of data stored in databases, datawarehouses or other information repositories. A widely accepted formal definition of data miningis given subsequently. According to this definition, data mining is the non-trivial extraction ofimplicit previously unknown and potentially useful information about data [2]. Data mininguncovers interesting patterns and relationships hidden in a large volume of raw data. Big Data is anew term used to identify the datasets that are of large size and have grater complexity [3]. So wecannot store, manage and analyze them with our current methodologies or data mining softwaretools. Big data is a heterogeneous collection of both structured and unstructured data. Businessesare mainly concerned with managing unstructured data. Big Data mining is the capability ofextracting useful information from these large datasets or streams of data which were not possiblebefore due to its volume, variety, and velocity.

The extracted knowledge is very useful and the mined knowledge is the representation ofdifferent types of patterns and each pattern corresponds to knowledge. Data Mining is analyzing the data from different perspectives and summarizing it into useful information that can be usedfor business solutions and predicting the future trends. Mining the information helpsorganizations to make knowledge driven decisions. Data mining (DM), also called KnowledgeDiscovery in Databases (KDD) or Knowledge Discovery and Data Mining, is the process ofsearching large volumes of data automatically for patterns such as association rules [4]. It appliesmany computational techniques from statistics, information retrieval, machine learning andpattern recognition. Data mining extract only required patterns from the database in a short timespan. Based on the type of patterns to be mined, data mining tasks can be classified intosummarization, classification, clustering, association and trends analysis.

DOI: 10.15680/IJIRCCE.2016. 0401153

II. LITERATURE SURVEY

1. **Puneet Singh Duggal, Sanchita Paul, "Big Data Analysis : Challenges and Solutions", international Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV**

This paper presents various methods for handling the problems of big data analysis through MapReduce framework over Hadoop Distributed File System (HDFS). Map Reduce techniques havebeen studied in this paper which is implemented for Big Data analysis using HDFS.

2. **ChanchalYadav, Shullang Wang, Manoj Kumar, "Algorithm and Approaches to handle large Data- A Survey", IJCSN, Vol 2, Issuue 3, 2013 ISSN:2277-5420**

This paper presents a review of various algorithms from 1994-2013 necessary for handling big data set. It gives an overview of architecture and algorithms used in large data sets. These algorithms define various structures and methods implemented to handle Big Data and this paperlists various tools that were developed for analyzing them. It also describes about the various security issues, application and trends followed by a large data set [9].

3. **Wei Fan, Albert Bifet, "Mining Big Data: Current Status, and Forecast to the Future", SIGKDDExplorations, Volume 14, Issue 2**

The paper presents a broad overview of the topic big data mining, its current status, controversy, and forecast to the future. This paper also covers various interesting and state-of-the-art topics onBig Data mining.

4. **Priya P. Sharma, Chandrakant P. Navdeti, "Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution", IJCSIT, Vol 5(2), 2014, 2126-2131**

This paper discusses about the big data security at the environment level along with the probing of built in protections. It also presents some security issues that we are dealing with today and 134 Computer Science & Information Technology (CS & IT) propose security solutions and commercially accessible techniques to address the same. The paper also covers all the security solutions to secure the Hadoop ecosystem.

5. **Richa Gupta, Sunny Gupta, AnuradhaSinghal, "Big Data : Overview", IJCTT, Vol 9, Number 5,March 2014**

This paper provides an overview on big data, its importance in our live and some technologies tohandle big data. This paper also states how Big Data can be applied to self-organizing websiteswhich can be extended to the field of advertising in companies.

III. ARCHITECTURE AND PROPOSED SYSTEM MECHANISM

Explanation-
Hadoop is a scalable, open source, fault tolerant Virtual Grid operating system architecture for data storage and processing. It runs on commodity hardware, it uses HDFS which is fault tolerant high bandwidth clustered storage architecture. It runs MapReduce for distributed data processing and is works with structured and unstructured data [11]. For handling the velocity heterogeneity of data, tools like Hive, Pig and Mahout are used which are parts of Hadoop and HDFS framework. Hadoop and HDFS (Hadoop Distributed File System) by Apache is widely used for storing and managing big data. Hadoop consists of distributed file system, data storage and analytics platforms and a layer that handles parallel computation, rate of flow (workflow) and configuration administration [6].

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

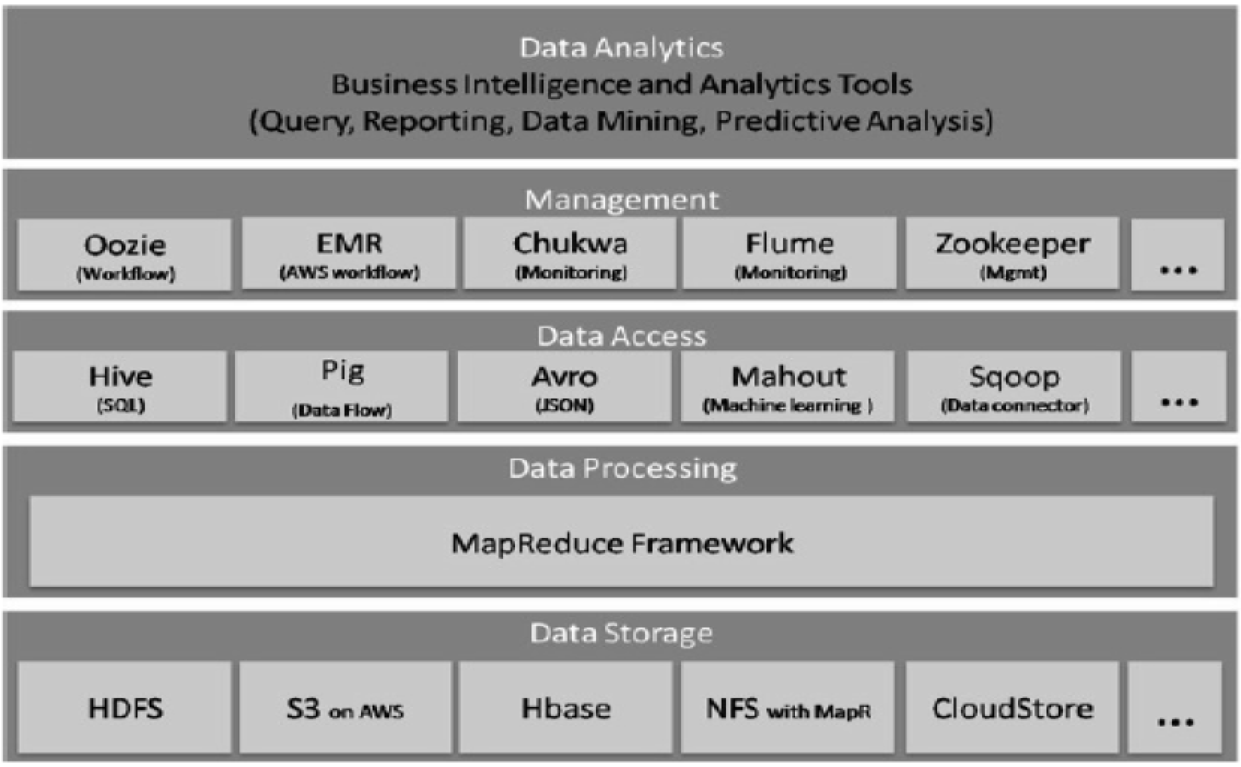

Fig 1. Big data Mining

HDFS runs across the nodes in a Hadoop cluster and together connects the file systems on many input and output data nodes to make them into one big file system. The present Hadoop ecosystem, as shown in Figure 1, consists of the Hadoop kernel, MapReduce, the Hadoop distributed file system (HDFS) and a number of related components such as Apache Hive, HBase, Oozie, Pig and Zookeeper and these components are explained as below [6]:

- HDFS: A highly faults tolerant distributed file system that is responsible for storing data on the clusters.
- MapReduce: A powerful parallel programming technique for distributed processing of vast amount of dataon clusters.
- HBase: A column oriented distributed NoSQL database for random read/write access.
- Pig: A high level data programming language for analyzing data of Hadoop computation.
- Hive: A data warehousing application that provides a SQL like access and relational model.
- Sqoop: A project for transferring/importing data between relational databases and Hadoop.
- Oozie: An orchestration and workflow management for dependent Hadoop jobs.

Figure gives an overview of the Big Data analysis tools which are used for efficient and precise data analysis and management jobs. The Big Data Analysis and management setup can be understood through the layered structured defined in the figure. The data storage part is dominated by the HDFS distributed file system architecture and other architectures available areAmazon Web Service, Hbase and Cloud Store etc. The data processing tasks for all the tools is Map Reduce and it is the Data processing tool which effectively used in the Big Data Analysis

IV. EXPERIMENTAL RESULT

In order to give more certainty to the results, the proposedtechnique was applied to the same data set. HACE theorem estimates the quality of features according to how well their values distinguish between samples (performance measures of MapReduce Job records) close to each other. Thus, after applying the mining algorithm to the data set, results are presented in Table 7 where the algorithm detects those features that are statistically relevant to the target classification which are measures with highest quality score.

Performance measures	Test values
MapReduceJob_ProcessingTime*	9.214837
MapReduceJob_TurnAround*	9.214828
SystemHDWriteBytes_Utilization*	8.176328
SystemUpTime	7.923577
SystemLoadMapCapacity	6.613519
SystemNetworkTxBytes	6.165150
SystemNetworkRxBytes	5.930647
SystemCPU_Utilization	5.200704
SystemLoadReduceCapacity	5.163010
MapReduceJob_ResponseTime	5.129339
SystemMemory_Utilization	3.965617
SystemHDReadBytes_Utilization	0.075003
NetworkRxDropped	0.00
NetworkTxCollisions	0.00
NetworkRxErrors	0.00
NetworkTxErrors	0.00

Table 1 Results of HACE Theorem

Performance measure	Quality score (W)
MapReduceJob_ProcessingTime*	0.74903
MapReduceJob_TurnAround*	0.74802
SystemHDWriteBytes_Utilization*	0.26229
SystemUpTime	0.25861
SystemCPU_Utilization	0.08189
SystemLoadMapCapacity	0.07878
SystemMemory_Utilization	0.06528
SystemNetworkTxBytes	0.05916
MapReduceJob_ResponseTime	0.03573
SystemLoadReduceCapacity	0.03051
SystemNetworkRxBytes	0.02674
SystemHDReadBytes_Utilization	0.00187
NetworkRxDropped	0.00
NetworkTxCollisions	0.00
NetworkRxErrors	0.00
NetworkTxErrors	0.00

Table 2 K-means clustering Algorithm results

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

International Journal of Innovative Research in Computer and Communication Engineering

(A High Impact Factor, Monthly, Peer Reviewed Journal)

Vol. 4, Issue 1, January 2016

The Algorithm results show that the performance measures job processing time and job turnaround, have the highest quality scores (W) and also have the potential to be distinguishing features between the two classes. In this case the performance measure ?hard disk bytes written? is also selected by means of the same approach as in the means and variance analysis: in other words, this has in terms of their use to stand out from the rest of the measures and give more certainty to the analysis of relationships. Thus, the measures job processing time, job turnaround and hard disk bytes written are also selected as candidates to represent the performance of BDA in the Hadoop system.

Trial	Expt	Time of up Time	Map tasks System	Reduce tasks Capacity	Network Rx Capacity bytes	Network Tx bytes	CPU utilization a	Job processing
1	1	−0.44091	−0.08601	−0.03342	−0.04170	−0.08030	−0.00762 ?a	−0.183902878
1	2	−0.34488	−0.07100	−0.03342	−0.02022	−0.18002	0.16864 ?a	−0.170883497
1	3	−0.49721	−0.08601	0.79990	0.01329	0.02184	−0.03221 ?a	−0.171468597
1	4	−0.39277	0.01307	−0.03342	0.02418	0.08115	−0.02227 ?a	−0.13252447

Table 3 Trials, experiments, and resulting values for job processing time output objective

a. Corresponding values for HD bytes read and Memory utilization.
b. Corresponding values for the set of experiments 5 to 12 of trial 1.

	Time of Response	Map tasks	Reduce tasks	Net. Rx	Net. Tx	CPU	HD bytes	Memory	
	System	Capacity	byte	byte	Utilization	read	Utilization	time	
Average SNR Level 1	3.18205	4.1784165	5.4175370	3.3712	3.8949	6.57901	5.11036	2.005514	4.011035
Average SNR Level 2	7.85630	5.8091173	4.8417803	7.5914	6.0116	3.58260	5.15667	8.253802	6.248281
Factor effect (Difference)	4.67424	1.6307007	0.5757566	4.2202	2.1166	2.99641	0.04630	6.248288	2.237245
Rank	2	7	8	3	6	4	9	1	5

Table 4 Factor effect rank on the job processing time output objective

DOI: 10.15680/IJIRCCE.2016. 0401153

According to HACE Theorem the factor effect is equal to the difference between the highest average SNR and the lowest average SNR for each factor (see Table 4). This means that the larger the factor effect for a parameter, the larger the effect the variable has on the process, or, in other words, the more significant the effect of the factor. Table 12 shows the factor effect for each variable studied in the experiment. Similar factor effect tables for job turnaround time and hard disk bytes written output values were also developed to obtain their results.Characteristic (job processing time used per a Map Reduce Job), it is always necessary to maximize the SNR parameter values. Consequently, the optimum level of a specific factor will be the highest value of its SNR. It can be seen that the optimum level for each factor is represented by the highest point in the graph (as presented that is, L2 for time of system up, L2 for map task capacity, L1 for reduce task capacity, etc.Using the findings presented in Table 4, it can be concluded that the optimum levels for the nine (9) factors for processing time output objective in this experiment based on our experimental configuration cluster are presented.

V. CONCLUSION

The amounts of data is growing exponentially worldwide due to the explosion of social networking sites, search and retrieval engines, media sharing sites, stock trading sites, news sources and so on. Big Data is becoming the new area for scientific data research and for business applications. Big data analysis is becoming indispensable for automatic discovering of intelligence that is involved in the frequently occurring patterns and hidden rules. Big data analysis helps companies to take better decisions, to predict and identify changes and to identify new opportunities. In this paper we discussed about the issues and challenges and result Analysis related to big data mining and also Big Data analysis tools like Map Reduce over Hadoop and HDFS which helps organizations to better understand their customers and the marketplace and to take better decisions and also helps researchers and scientists to extract useful knowledge out of Big data. In addition to that we introduce some big data mining tools and how to extract a significant knowledge from the Big Data.

REFERENCES

1) Julie M. David, KannanBalakrishnan, (2011), Prediction of Key Symptoms of LearningDisabilities in School-Age Children using Rough Sets, Int. J. of Computer and Electrical Engineering, Hong Kong, 3(1), pp163-169.
2) Julie M. David, KannanBalakrishnan, (2011), Prediction of Learning Disabilities in School-AgeChildren using SVM and Decision Tree, Int. J. of Computer Science and Information Technology, ISSN 0975-9646, 2(2), pp829-835.
3) Albert Bifet, (2013), “Mining Big data in Real time”, Informatica 37, pp15-20.
4) Richa Gupta, (2014), “Journey from data mining to Web Mining to Big Data”, IJCTT, 10(1),pp18-20.
5) http://www.domo.com/blog/2014/04/data-never-sleeps-2-0/
6) Priya P. Sharma, Chandrakant P. Navdeti, (2014), “ Securing Big Data Hadoop: A Review of Security Issues, Threats and Solution”, IJCSIT, 5(2), pp2126-2131.
7) Richa Gupta, Sunny Gupta, AnuradhaSinghal, (2014), “Big Data:Overview”, IJCTT, 9 (5).
8) Wei Fan, Albert Bifet, “Mining Big Data: Current Status and Forecast to the Future”, SIGKDD Explorations, 14 (2), pp1-5.
9) ChanchalYadav, Shullang Wang, Manoj Kumar, (2013) “Algorithm and Approaches to handle large Data- A Survey”, IJCSN, 2(3), ISSN:2277-5420(online), pp2277-5420.
10) Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, “Data Mining with Big Data”
11) Puneet Singh Duggal, Sanchita Paul, (2013), “Big Data Analysis:Challenges and Solutions”, Int. Conf. on Cloud, Big Data and Trust, RGPV