International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 19 (2017) pp. 8131-8136
© Research India Publications. http://www.ripublication.com

A Comparative Study of Big Data Computational Approaches

Sheikh Ikhlaq[1] and Bright Keswani [2]

[1]Research Scholar, Department of Computer Sciences, Suresh Gyan Vihar University, Jaipur, India.

[1]Orcid Id: 0000-0002-7175-6214
[2]Associate Professor, Department of Computer Sciences, Suresh Gyan Vihar University, Jaipur, India.

Abstract

In this paper we have done a comparative study of diverse methodologies (technologies/methods) based on certain parameters to chalk out which of them is best possible (optimal) or what else needs to be done, to have an optimal solution~~.~~ As the data is being generated and accumulated at a very high velocity with diversity in addition to it, processing has become a tedious task. Whereas the fact remains that, by processing of this data we will uncover gold from these huge mountains but if it's left untreated it will become Everest's of garbage .Since the foundation of term Big Data many methodologies or technologies are present which are being used to process these data mountains. These technologies have their own area of interest due to which there are obvious drawbacks and mode of operations between them. After analysis we found that common to them is low agility and the Data management. Once the data Management issue of Cloud technology is resolved it will prove a boon and cure to the other drawbacks of cloud implementation for processing of BigData. It makes as sure that Cloud technology with better Data Management will be implemented in every phase of life from governments to business.

INTRODUCTION

Almost 2.5 quintillion bytes of data is produced each day. 90% of this data is estimated to be created in just last two years only. Data at this point accumulates from everywhere, climate information gathering sensors, social media websites, videos and pictures, transaction records(e.g. banking records), and cell phone calls[1], GPS signals etc. Data gathered by all this is called as Big Data. Roger Mougalas at O'Reilly Media formed the term Big Data for the very first time, just after a term Web 2.0 was created. This refers to a outsized set of data which is nearly impractical to handle and process by traditional business intelligence methods and tools.

The expansion of data can by no means stop or it can't be restricted anyhow. IDC Digital Universe in its Study, published in year 2011 stated, nearly 130 Exabyte's of data was produced and stored in 2005. This quantity grew severely to 1,227 Exabyte's in 2010 and it was projected to produce at 45.2% to 7,910 Exabyte's in year 2015· This data can do wonder by extracting the buried assets of information that reside inside it[2]. In the year 2004 "Google" introduced MapReduce to the world, which later laid foundation to Hadoop and other related technologies (methods).2005 was the year developed and released by Yahoo!(built on top of Google's MapReduce).Goal of Hadoop was to index the entire WWW(World Wide Web) and today the open-source Hadoop technology is being used many organizations to munch through large volumes of data.

As there is start of increase in social networks and the Web 2.0 takes speed, additional and more data gets produced on a day today basis. Pioneering startups gradually set up to mine into this massive volume of data and also many governments have started to work on Big Data projects[3],[4].

Big data is explained as large volume of data that requires new set of technologies and architecture that will make it possible to mine value from it by the process of capturing and analysis[5].

IMPORTANCE OF BIG DATA PROCESSING

"Big data processing is the method of probing big data to uncover hidden patterns, correlations and other useful information that can be used to make improved decisions. With big data analytics, data scientists and others can analyze massive volumes of data that usual analytics and business intelligence solutions can't tap. Consider that your business could build up billions of rows of data with millions of data combinations in numerous data stores and plentiful formats. High-performance analytics is essential to process that much data in order to outline out what's significant and what isn't. For most organizations, big data analysis is defied. Consider the total volume of data and the unlike formats of the data that is composed transversely the entire business and the many unlike ways unlike types of data can be pooled, contrasted and analyzed to gather patterns and other valuable business information.

The first confront is in churn down data to right of entry all data in an organization stores in multiple places and often in multiple systems. The next big data challenge is in making platforms that can drag in unstructured data as effortlessly as structured data. This enormous [6]volume of data is in general so large that it's complex to process using traditional database and software techniques and methods.

There are generally four approaches to data analytics, and each them comes under either reactive or proactive category:

i. Reactive (business intelligence). In this category, business intelligence provides pattern business reports, ad hoc reports, OLAP and even gives alerts and notifications that are based on analytics. Here ad hoc analysis checks at the static past, which has its reason in a limited number of conditions.

ii. Reactive – big data BI. When reporting pulls from huge data sets, we can say this is performing big data BI. But decisions based on these two methods are still reactionary.

iii. Proactive (big analytics). Creation of forward-looking, proactive decisions needs proactive big analytics similar to optimization, predictive modelling, text mining, and forecasting. They allow us to recognize trends, mark weaknesses and determine circumstances for creation of decisions about the future. But though it is proactive, big analytics can’t still be performed on big data since traditional storage techniques and processing times cannot stay up.

iv. Proactive (big data analytics). Using big data analytics we can mine only the appropriate information from mountains of, and then analyze it to convert our business decisions for the better future. Flattering proactive with big data analytics is not a one-shot attempt; it is a culture change, a new method of acquisition by freeing our analysts and decision creators to gather the future with appropriate knowledge and insight[7].

It becomes very difficult for the traditional data analysis, processing and storing tools to deal with the five characteristic of data simultaneously. Since big data is a recent upcoming technology in the market which can bring huge benefits to the business organizations, it becomes necessary that various challenges and issues” [8], [9] related to it must be taken care of and resolved. If data is to be used to make accurate decisions in time it becomes necessary that it should be available in accurate, complete and timely manner. To deal with the Big Data analytics challenge of Fast Data Processing many techniques have been implemented in literature. Some of them are:

HADOOP

Hadoop is an open source software records that permits for distributed processing of large data sets across multiple clusters of computers with simple programming model implemented. Hadoop mainly consists of HDFS (Hadoop Distributed File System), MapReduce (distributed computing system).Hadoop is at present extensively used in the Internet, and has attracted universal attention from research society [10].

Hadoop includes following Components:

Core Hadoop: General utility which supports other Components.

Hadoop-Distributed File System: It is a distributed file system which is used for storing data files on clusters of computers.

HBase: It is distributed database for unsystematic Read/Write access.

Pig: It is high level data processing structure which analyzes data sets that come about in high level language(s).

Hive: Its data warehousing application which gives a SQL type interface and a relational model.

Sqoop: It is project used for transferring data between the relational databases and the Hadoop.

Avro: It is structure of data serialization.

Oozie: It is workflow for dependent Hadoop works.

Chukwa: This is a Hadoop sub- project used as data gathering system for monitoring the distributed systems.

Flume: This is a consistent and distributed streaming log collection component.

Zookeeper: This is a centralized server used to provide a distributed synchronization and group services [11].

HADOOP DISTRIBUTED FILE SYSTEM [12]

HDFS is a very huge distributed file system which provides us fault tolerance and also has high outputs. Hadoop Distributed File System stores and saves the files as a sequence of blocks and replicates these data blocks for fault tolerance. It stores gigantic data sets and gives the global access to files in clusters computers. Hadoop Distributed File saves metadata on a particular and dedicated server, termed “Name Node”. Application data is stretched on another network termed “Data Node” All servers are completely connected and they communicate with each other using TCP-IP based protocols. Hadoop Distributed File architecture is made up of four parts:

Name Node: Name Node is in charge of managing all the metadata and file system activities. It manages the file system namespace works like open, close and rename both on file and directory. It also, creates decisions concerning replication of the blocks. NameNode has to maintain the tree of namespace and has to maps the file blocks to DataNodes (physical location of file's data).A single NameNode is thought to be a bottleneck for managing requests in scientific environments

Data Node: DataNode saves the data in the HDFS. Each and every DataNode saves the data blocks on account of local or the remote patrons. Each block is stored as a separate file in the local file system of the node. On starting -up, DataNode is connected to the NameNode and then creates a handshake. The idea of the handshake is to authenticate the namespace ID and the version of software used by DataNode. Incase NameNode gets matched with DataNode, the DataNode shuts down automatically. When the handshake is victorious, the DataNode gets registered with NameNode. DataNodes continually stores their unique storage IDs. This storage ID is used as an internal identifier for the DataNode which makes it as identifiable even if it is started with a altered IP address or a port. This storage ID is attached to the DataNode, when it is registered with the NameNode at the very first time and then never changes later. This DataNode then provides response to the requirements that are approaching from the NameNode, for operations of file system. Then this DataNodes provide service of read, write and file replication requests which are based on the way coming from NameNode.

Job Tracker: Job Tracker speaks to the NameNode in order to establish the location of the data. JobTracker schedules every single map reduces or intermediate reconciling operations to the specific machines. It checks the success and losses of these individual (single) tasks. It works to finish the entire batch job. If the task fails, the JobTracker then automatically relaunchs the task,probably on a changed node, up to a certain amount of retries.

TaskTracker: JobTracker is considered the main master for supervision of overall implementation of a MapReduce job. The TaskTrackers administers the execution of individual (single) task on every slave node. Even though there is a single TaskTracker/slave node. Each and every TaskTracker can produce or have multiple Java Virtual Machines (JVMs) so as to handle multiple map or reduce tasks in parallel. This TaskTracker transmits major messages to the JobTracker, typically after every few minutes, to restore confidence that JobTracker is still a live and working [13] .

HADOOP DRAWBACKS

From the Architecture of Hadoop and its workflow (Dataflow) of data computation, many drawbacks of Hadoop arise. These include:

i. Hadoop requires very high memory and large storage to perform replication technique.

ii. Hadoop chains allotment of tasks only and does not have any plan to support the scheduling of tasks.

iii. Single master (Name Node) that requires care.

iv. Load time is very long.

CLOUD COMPUTING

Cloud computing is a currently trending in Information technology that takes computing and data away from personal computers into outsized data centers. It considers to applications which are delivered as the services which are over the Internet also to the real cloud infrastructure [14].

A cloud is composed of processing, network and the storage devices (elements).Cloud architecture comprises of three intangible layers. Infrastructure layer is the lowest layer responsible for delivering basic storage and computing capability as standardized services over the internet (network).Components like Servers, storage systems, switches, routers, and the other systems hold particular types of workloads, which consists from batch processing to server or storage amplification during the peak loads.

The middle platform layer gives the higher abstractions and services so as to develop, test, deploy, host, and maintain the applications within the same integrated development surroundings [15].The highest layer is the application layer and characterizes a complete application, to be offered as a service.

OTHER METHODS OF BIG DATA PROCESSING AND ANALYTICS

A. MOBILE AGENT BASED [16]

A latest framework which is termed as MRAM is created by using mobile agent and MapReduce archetype under the JADE. Here, mobile agents can send both the code and the data to whichever machine and react energetically for any of the changes in environment. Also, the mobile agents do have capability to move with code and data, to different machine or the environment if the pervious is down. Moreover, Hadoop has a problem that it is still single master which needs care. This setback is resolved in MRAM by using send metadata that holds map of network and all of the data about the tasks and dependences occurring between them. MRAM increases the performance by providing the server or control node, the chance to execute tasks as done by the others nodes. One more disadvantage of Hadoop is that it doesn't bear scheduling of tasks, also it does not work with tasks which are dependent, but MRAM provides this feature. A new approach is coded in JAVA programming language formed on JADE, Giving it platform independence means it can run on dissimilar and multiple machines and diverse operating system without any trouble[17]. Mobile Agent is software that can migrate during execution across a heterogeneous or homogenous network. An example of MRAM is shown below in Fig 1.

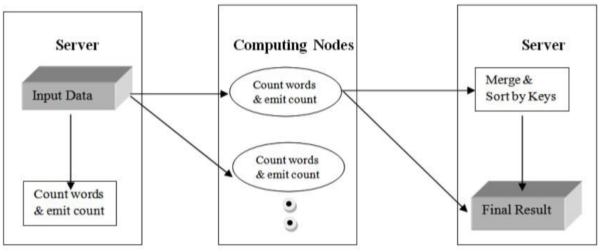

Figure 1: MRAM architecture

B.HIGH PERFORMANCE COMPUTING CLUSTER [18]

High Performance Computing Cluster, as shown in Fig 4. is open source; data concentrated computing system which holds a software architecture that is implemented on a product computing cluster to grant elevated performance parallel data processing for applications by using big data. The three core HPCC components are

High Performance Computing -Cluster Data Refinery (Thor) is a extremely parallel ETL (Extraction Transformation Loading) engine which enables us the data integration on a degree and gives the batch oriented data management.

High Performance Computing Cluster-Data Delivery Engine (Roxie) is a particularly parallel, elevated throughput, mega fast, low latency, which allows competent multi user retrieval of the data and the structured query reaction engine.

Enterprise Control Language (ECL)-that automatically distributes the workload involving nodes, it has regular (Automatic) synchronization of algorithms, to develop extensive machine learning library, it obtains simple usage of programming language that are optimized for Bigdata operations and query connections.

C. ANALYTIC TECHNIQUES USING DATA MINING [19]

Classification: Classification which is the procedure of arranging a particular object in set of category, based on each model attributes.

Prediction: Prediction is used to predict unremitting values whereas classification predicts the discrete values similar to class label.

Clustering: Classification is performed by learning and understanding the samples in advance, while clustering explains a class without a previous learning to categorize data. Clustering is the procedure of combining data which have several properties based on related attributes without any predefined criteria.

Association: This discovery rule is the procedure of identifying appealing relationship or correlation among the various databases [20] .

COMPARTIVE ANALYSIS BASED ON VARIOUS PARAMETERS

A variety of Parameters are used depending on how these above explained methods (Technologies) execute to satisify the 3 V`s of Big Data. outcome of which shows us comparitive analysis of these technologies(methods).Various comparitive paramteres with its particulars are listed below

Scalability: One of the major components of the Hadoop is redundancy built inside the environment. This redundancy makes Hadoop to level out workloads crosswise to huge clusters of reasonably priced machines to work on the problems associated with BigData. As the volume of data raises, so does the volume of the nodes increase in a Hadoop framework .This is not a problem in High Performance Computing Architecture (HPCC),Cloud technology and the mobile based agent.Swell in size of hadoop suggests that single master has lot of work to do and any loos or faluire in Master will mean that thewhole architecture has failed[21].

Storage: Hadoop requires a very large volume of storage and its mangement is very difficult. Storage problem is also present in mobile agent technology.While in cloud and High Performance Computing Architecture (HPCC) storage is not a issue but since there is no peferect data mangement a sloution is required to be found[22].

Fault Tolerance: Hadoop is excellent at fault tolerance when it is restricted to nodes and its failures but it fails when master gets faulty. In Hadoop, it is not only the data redundantly that is stored in multiple places crosswise the cluster, but also the programming model is so that failures are predictable and are solved automatically by running sections of the program on multiple servers present in the cluster. This is due to this redundancy, that it is possible to allocate the data and its associated programs across extremely large cluster of product components. It is very well known that commodity hardware components have to (will) fail, but it is this redundancy that provides the fault tolerance and ability for the Hadoop cluster to repair itself. Mobile Agent technology is modest to fault tolerance where as cloud is the best survivor when it comes to fault tolerance. HPCC is also good at fault tolerance.

Agility: In terms of agility no method is good as we have high load time in each technology due to the poor Data management. Agility is a need of an hour [23] [24].

Virtualization: A Browser-based visualization tool termed BigSheets, which is used to enable users to harness the supremacy of Hadoop using a recognizable spreadsheet interface. BigSheets don't requires any programming or special administration. BigSheets makes many visualization tools available: Tag Cloud, Pie Chart, Map, Heat Map and Bar Chart. There are no methods of virtualization when it comes to other technologies [25].

Cost: The cost of using and buying these technologies (methods) is very high. Governments and mid size organization in Large cannot bear that cost. Technology which has proven to be best in terms of cost is cloud. It is very cheap as compared to rest and can be obtained easily.

Ease of use: All of the above technologies require professionals that highly qualified and trained. This means that only qualified people can implement these technologies.

Data Management: No proper data management technique is available in these technologies as of now. Data is found scattered. A better data management technique is yet to be created or designed for these technologies. This has a direct effect on operations that are performed on data and its storage.

CONCLUSION

There are a variety of technologies for Bigdata computation: Hadoop, Cloud, HPCC, and Mobile Agent. These technologies are having their own procedure for computation of BigData. These technologies have their own area of interest due to which there are obvious drawbacks and mode of operations between them. Major drawback that is of common to them is low agility and the Data management. The job here is to improve the Data management technique with which all the technologies can provide better results. The data management technique will also cure the other short comings associated with studied Technologies.

REFERENCES

[1] "A Comparison of Big Data Analytics Approaches Based on Hadoop Map Reduce", 2016, http://www.academia.edu/3502325/A_Comparison_of_Big_Data_Analytics_Approaches_Based_on_Hadoop_MapReduce.

[2] "The 2011 Digital Universe Study: Extracting Value from Chaos", 2015, http://www.emc.com/collateral/demos/microsites/emc-digital-universe-2011/index.htm.

[3] Ikhlaq, S., & Keswani, B., 2016," Enhanced Approach for BigData Agility," International Journal of Advanced research in Computer Science and Software Engineering, 6(1), pp. 579-584.

[4] Ikhlaq, S., & Keswani, B., 2016,"Computation of BigData in Hadoop and Cloud Environment," IOSR Journal of Engineering, 6(1), pp.31-39.

[5] Sagiroglu, S., & Sinang, D., 2013,"Big Data: A Review. Collaboration Technologies and Systems (CTS)", Proceedings of International Conference on IEEE, pp. 42-47.

[6] Ikhlaq, S., & Keswani, B., 2016," Enhanced Approach for BigData Agility," International Journal of Advanced research in Computer Science and Software Engineering, 6(1), pp. 579-584.

[7] "Big Data Analytics", 2016, http://www.sas.com/it_it/insights/analytics/big-data-analytics html.

[8] Chandarana, P., & Vijayalakshmi, M., 2014, "Big Data Analytics Framework", Proceedings of International Conference on Circuits, System, Communication and Information Technology Applications IEEE.

[9] Saha, B., & Srivastava, D., 2014,"Data Quality: The other face of Big Data", Proceedings of 30th International Conference on Data Engineering IEEE".

[10] "Hadoop Online Training", 2015, http://es.slideshare.net/smconsultantdaniel/hadoop-online-training-at-s-m-consultant.

[11] Petrazickis, L., & Butuc, M., 2014, "Crunching Big Data with Hadoop and Big Insights in the Cloud", Information Management Technologies, 2(1), pp. 241-42.

[12] Chandarana, P., & Vijayalakshmi, M., 2014, "Big Data Analytics Framework", Proceedings of International Conference on Circuits, System, Communication and Information Technology Applications IEEE.

[13] "Hadoop Tutorials", 2014, http://www.authorstream.com/Presentation/smconsultantdaniel-2799775-hadoop-tutorial/

[14] Dikaiakos. M, D., & Pallis, G., 2009, " Cloud Computing: Distributed Internet computing for IT and scientific research", IEEE Internet Computing 13 (5), pp. 10-13.

[15] "Cloud Computing Distributed Internet Computing for IT and Scientific Research", 2016, http://www.academia.edu/8417516/Cloud_Computing_Distributed_Internet_Computing_for_IT_and_Scientific_Research.

[16] Essa Y.M., "Mobile Agent Based New Framework for improving Big Data Analysis", 2016 , Proceedings of International Conference on Cloud Computing and Big Data IEEE.

[17] "Mobile Agent based New Framework for Improving BigData Analysis", 2016, http://www.academia.edu/8953353/Mobile_Agent_based_New_Framework_for_Improving_Big_Data_Analysis.

[18] Katal A., et al., "Big Data : Issues, Challenges, Tools and good Practices", 2013, Proceeding of Contemporary Computing (IC3), Sixth International Conference on IEEE, pp. 404-09.

[19] Kim S. H., & et al., "Attribute Relationship Evaluation Methodology for Big Data Security", 2013, Proceedings of International Conference on Information Science and Applications (ICISA), IEEE.

[20] "Big Data: A review", 2016, .https://es.scribd.com/document/171804158/16.

[21] "A Comparison Of Big Data Analytics Approaches Based On Hadoop MapReduce", 2016, http:// www.academia.edu/3502325/A_Comparison_of_Big_ Data_Analytics_Approaches_Based_on_Hadoop_Map Reduce.

[22] Ikhlaq, S., & Keswani, B., 2016,"Computation of BigData in Hadoop and Cloud Environment," IOSR Journal of Engineering, 6(1), pp.31-39.

[23] Ikhlaq, S., & Keswani, B., 2016," Enhanced Approach for BigData Agility," International Journal of Advanced research in Computer Science and Software Engineering, 6(1), pp. 579-584.

[24] "A Comparison of Big-Data Analytics Approaches", 2016, https://www.scribd.com/document/290363390 /A-Comparison-of-Big-Data-Analytics-Appro

[25] A Comparison of Big-Data Analytics Approaches", 2016, https://www.scribd.com/document/290363390 /A-Comparison-of-Big-Data-Analytics-Appro