IJDCST @ Nov-Dec-2018, Issue- V-6, I-5, SW-01
ISSN-2320-7884 (Online)
ISSN-2321-0257 (Print)

Analysis on Processing Remotely Distributed Data Stores with Cutting Edge Mining Techniques

P.Siva Parvathi, Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh, India.

Sudhakar Putheti, Computer Science and Engineering, Vasireddy Venkatadri Institute of Technology, Nambur, Guntur Dt., Andhra Pradesh, India.

Abstract: Today business organizations are utilizing their raw data to obtain the useful information, which helps in decision making and forecasting of business needs. In order to process the high volumes of raw data, several types of big data analytics are using in this domain. As data has been distributed remotely in different locations of the world, gathering the data from different locations to a central location for processing became cumbersome as it needs more time, bandwidth and other resources. Present data mining analytics following procedure is "data should be transferred to processor for mining", which is trying to integrate the different sources of data from various remote locations to a center location for processing. As aforementioned this procedure consumes more time to transfer the data from remote locations to central store, expects more bandwidth, exceeded utilization of resources, loss of data confidentiality and also raises several privacy and security issues. In this paper, we discuss a new big data mining procedure "Processor should be transferred to data", which consumes very less bandwidth and completes the mining process at high speed compared to previous procedures. This paper explains the possible methodologies available to implement the proposed procedure. Simulations shown that the proposed "processor should be transferred to data" approach proven as the best choice than data should be transferred to processor.

Keywords: Big data Analytics, Information Mining, Distributed data stores and Cutting edge data mining techniques

I. Introduction

Information has been a backbone of any industry and will do continue in future as well. Storing, extracting and using information has been vital to many organization's tasks. In the past when there were no interconnected frameworks, information would remain and be devoured at one place. With the beginning of Internet innovation, capacity and necessity to share and change information has beena need.

The measure of information produced each day on the planet is detonating. The expanding volume of digital and social networking and IoT, is filling it considerably further.With the introduction of digital innovations and micro or mobile gadgets, a lot of advanced information is being generated each day. Advances in computerized sensors and correspondence innovations have tremendously added to this immense measure of information, collecting significant data for business organizations.This Big data[1, 2], which is generated from organizations in high volume, is difficult to process utilizing ordinary technologies and waiting for monstrous parallel processing. Today advanced technologies [3] that can store and process hexa bytes, terabytes, petabytes of information without immensely raising the data warehousing cost is a need of time.

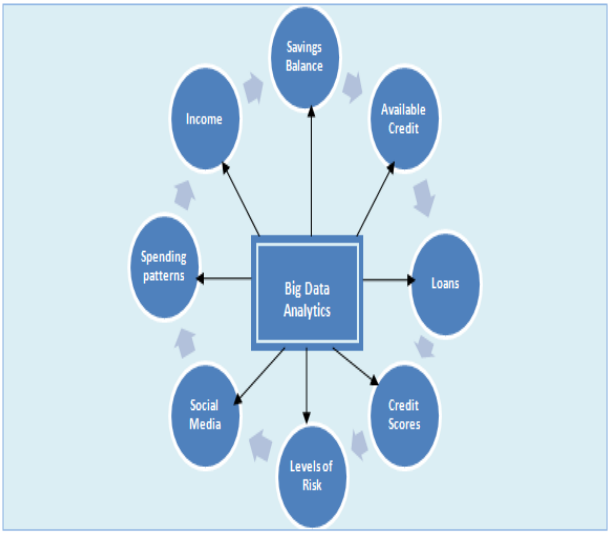

Figure 1. Big Data Analytics with their use in different operations

Advantages from Big data processing obtained by many organizations today are from simple marketing stores to big financial industries. Figure 1 is displaying the possible operations which are possible today by using Bigdata analytics[4, 5].Big data open source technologies have gained a lot of traction due to the proven ability to process massive amounts of information in parallel. These key features and the ability to process extensive data were a

great motivation to evaluate the industry architecture of Apache, Hadoop's leading Big Data Processing framework[6]. Lot of information means that heap of hidden insights. The flexibility to quickly analyze massive information means the chance to find out regarding customers, market trends, selling and advertising drives, instrumentality observation and performance analysis and far a lot of. And this is often an important reason that a lot of massive data contained enterprises are in a very need of sturdy big data analytics tools and technologies.

Big data analytics is currently broadly utilized in the fields of software engineering, for example, recommendation frameworks, pursuit and data recovery, computer vision and image processing, and is making its raid into this present reality as far as business intelligence, healthcare and supply chain analysis. It is utilized even inside the networks in areas such as network security. Traditionally, collections of information is stored and processed in an exceedingly single datacenter. Because the volume of information grows at an incredible rate, it's less efficient for less than one datacenter to handle such massive volumes of information from a performance point of view. Giant cloud service suppliers are deploying datacenters geographically around the world for higher performance and accessibility. A wide used approach for analytics of geo-distributed information [7] is that the centralized approach, that aggregates all the data from native datacenters to a central datacenter.

Nonetheless, it has been seen that this methodology devours a significant measure of data transfer capacity, prompting more terrible execution. Various instruments have been proposed to accomplish ideal execution when information examination is performed over geo-distributed datacenters. Present data mining analytics following procedure is “data should be transferred to processor for mining”, which is trying to integrate the different sources of data from various remote locations to a center location for processing. As aforementioned this procedure consumes more time to transfer the data from remote locations to central store, expects more bandwidth, exceeded utilization of resources, loss of data confidentiality and also raises several privacy and security issues. In this paper, we discuss a new big data mining procedure “Processor should be transferred to data”, which consumes very less bandwidth and completes the mining process at high speed compared to previous procedures. This paper explains the possible methodologies available to implement the proposed procedure.

II. Related Work

Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it’s not the amount of data that’s important. It’s what organizations do with the data that matters. Big data [1, 8] can be analyzed for insights that lead to better decisions and strategic business moves.

While the term “big data” is relatively new, the act of gathering and storing large amounts of information for eventual analysis is ages old. The concept gained momentum in the early 2000s when industry analyst Doug Laney articulated the now-mainstream definition of big data as the three Vs:

Volume: Organizations collect data from a variety of sources, including business transactions, social media and information from sensor or machine-to-machine data. In the past, storing it would’ve been a problem – but new technologies (such as Hadoop) have eased the burden.

Velocity: Data streams in at an unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time.

Variety: Data comes in all types of formats – from structured, numeric data in traditional databases to unstructured text documents, email, video, audio, stock ticker data and financial transactions. Recently we consider two additional dimensions when it comes to big data:

Variability: In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Is something trending in social media? Daily, seasonal and event-triggered peak data loads can be challenging to manage.

Complexity: Today's data comes from multiple sources, which makes it difficult to link, match, cleanse and transform data across systems. However, it’s necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control.

III. Distributed Big Data Processing Techniques

In this section we discuss about the big data processing techniques, which are implemented based on Processor should be transferred to data procedure, will be discussed. Fast-forward and a lot has changed in data processing for main scale industries. Over the last several years, the cost to purchase computing and storage resources has decreased dramatically. Aided by virtualization, commodity servers that could be clustered and blades that could be networked in a rack changed the economics of computing. This change coincided with innovation in software automation solutions that dramatically improved the manageability of these systems. The capability to leverage distributed computing and parallel processing

techniques dramatically transformed the landscape and dramatically reduce latency. There are special cases, such as High Frequency Trading (HFT) [9], in which low latency can only be achieved by physically locating servers in a single location. One of the perennial problems with managing data — especially large quantities of data — has been the impact of latency. Latency is the delay within a system based on delays in execution of a task. Latency is an issue in every aspect of computing, including communications, data management, system performance, and more. If you have ever used a wireless phone, you have experienced latency firsthand. It is the delay in the transmissions between you and your caller. At times, latency has little impact on customer satisfaction, such as if companies need to analyze results behind the scenes to plan for a new product release. This probably doesn't require instant response or access. However, the closer that response is to a customer at the time of decision, the more that latency matters.

By itself, stored data does not generate business value, and this is true of traditional databases, data warehouses, and the new technologies such as Hadoop [10, 11] for storing big data. Once the data is appropriately stored, however, it can be analyzed, which can create tremendous value. A variety of analysis technologies, approaches, and products have emerged that are especially applicable to big data, such as in-memory analytics, in-database analytics, and appliances.

It is useful to distinguish between the three kinds of analytics (Processor should be transferred to data) because the differences have implications for the technologies and architectures used for big data analytics. Some types of analytics are better performed on some platforms than on others.

Descriptive analytics, such as reporting/OLAP, dashboards/scorecards, and data visualization, have been widely used for some time, and are the core applications of traditional BI [11]. Descriptive analytics are backward looking (like a car's rear view mirror) and reveal what has occurred. One trend, however, is to include the findings from predictive analytics, such as forecasts of future sales, on dashboards/scorecards.

Predictive analytics suggest what will occur in the future (like looking through a car's windshield). The methods and algorithms for predictive analytics such as regression analysis, machine learning, and neural networks have existed for some time. Recently, however, software products such as SAS Enterprise Miner have made them much easier to understand and use. They have also been integrated into specific applications, such as for campaign management. Marketing is the target for many predictive analytics applications; here the goal is to better understand customers and their needs and preferences.

Some people also refer to exploratory or discovery analytics, although these are just other names for predictive analytics. When these terms are used, they normally refer to finding relationships in big data that were not previously known. The ability to analyze new data sources—that is, big data—creates additional opportunities for insights and is especially important for firms with massive amounts of customer data. Golden path analysis is a new and interesting predictive or discovery analytics technique. It involves the analysis of large quantities of behavioral data [12] (i.e., data associated with the activities or actions of people) to identify patterns of events or activities that foretell customer actions such as not renewing a cell phone contract, closing a checking account, or abandoning an electronic shopping cart. When a company can predict a behavior, it can intercede, perhaps with an offer, and possibly change the anticipated behavior. Whereas predictive analytics tells you what will happen, prescriptive analytics suggests what to do (like a car's GPS instructions). Prescriptive analytics can identify optimal solutions, often for the allocation of scarce resources. It, too, has been researched in academia for a long time but is now finding wider use in practice. For example, the use of mathematical programming for revenue management is increasingly common for organizations that have "perishable" goods such as rental cars, hotel rooms, and airline seats. For example, Harrah's Entertainment, a leader in the use of analytics, has been using revenue management for hotel room pricing for many years.

IV. Advantages of Bigdata Analytics

As has been discussed, collecting and storing big data does not create business value. Value is created only when the data is analyzed and acted on. The benefits from big data analytics can be varied, substantial, and the basis for competitive advantage. Because of its potential benefits, some people add a fourth to the characteristics of big data, high value. This value is realized, however, only when an organization has a carefully thought out and executed big data strategy. Research shows the benefits of using data and analytics in decision making. One study of 179 large publicly traded firms found that companies that have adopted data-driven decision making have output and productivity that is 5% to 6% higher than that of other firms. The relationship extends to other performance measures such as asset utilization, return on equity, and market value. In 2010, the MIT Sloan Management Review[12], in collaboration with the IBM Institute for Business Value, surveyed a global sample of nearly 3,000 executives. Among the findings were that top performing

organizations use analytics five times more than do lower performers, and that 37% of the respondents believe that analytics creates a competitive advantage. A follow-up study in 2011 found that the percentage of respondents who reported that the use of analytics was creating a competitive advantage rose to 58%, which is a 57% increase. Although these studies do not focus exclusively on big data, they do show the positive relationships between data-driven decision making, organizational performance, and competitive position.

V. Bigdata Hadoop Framework

Proposed System: Apache Hadoop is a software framework for processing large amounts of data across potentially massively parallel clusters of servers. To illustrate, Yahoo has over 42,000 servers in its Hadoop installation. Hadoop is open source and can be downloaded at www.apache.org. The key component of Hadoop is the Hadoop Distributed File System (HDFS) [11, 14], which manages the data spread across the various servers. It is because of HDFS that so many servers can be managed in parallel. HDFS is file based and does not need a data model to store and process data. It can store data of any structure, but is not a RDBMS. HDFS can manage the storage and access of any type of data (e.g., Web logs, XML files) as long as the data can be put in a file and copied into HDFS.

The Hadoop infrastructure typically runs MapReduce programs (using a programming or scripting language such as Java, Python, C, R, or Perl) in parallel. MapReduce takes large datasets, extracts and transforms useful data, distributes the data to the various servers where processing occurs, and assembles the results into a smaller, easier-to-analyze file. It does not perform analytics per se; rather, it provides the framework that controls the programs (often written in Java) that perform the analytics. Currently, jobs can be run only in batch, which limits the use of Hadoop/MapReduce for near real-time applications. Although Hadoop and MapReduce are discussed and typically used together, they can be used separately. That is, Hadoop can be used without MapReduce and vice versa.

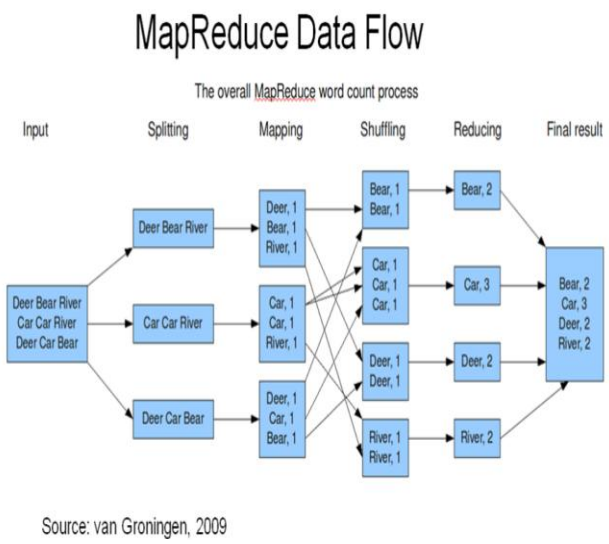

Figure 2 Hadoop/MapReduce internal process block diagram

Figure 2 illustrates how processing occurs with Hadoop/MapReduce [van Groningen, 2009] [2]. This is a simple processing task that could also be done with SQL and a RDBMS, but provides a good example of Hadoop/MapReduce processing. At the left is a data file with records containing Deer, Bear, River, and Car. The objective is to count the number of times each word occurs. The first step is to split the records and distribute them across the clusters of servers (there are only three in this simple example). These splits are then processed by multiple map programs such as Java and R running on the servers. The objective in this example is to group the data by a split based on the words. The MapReduce system then merges the shuffle/sort results for input to the reduce program, which then summarizes the number of times each word occurs. This output can then be input to a data warehouse where it may be combined with other data for analysis or accessed directly by various BI tools (e.g., Tableau, MicroStrategy).

Simulations: In order to mimic the process of geo-distributed data with big data analytics, we have designed a platform with windows-7 OS, Cygwin platform to install Linux, Hadoop Framework 1.08 version, My Sql database, Tomcat Server etc. After installation of Linux OS with the help of Cygwin, JDK 7 open source software has been installed on top of Linux. This JDK software executes the hadoop map reduce code, which is in compressed and executable jar file format. After executing the jar file with jdk for hadoop, it processed the predefined sample data from .csv file. This data is a structured and tabular data, contains the information in the form of records which is having the raw data belongs to different organizations. This data processed by hadoop

frame work analytics and results the values, which has been processed by the proposed "processor should be transferred to data" procedure. The below figure 4 and 5 are representing the results obtained after simulations we can see from them.

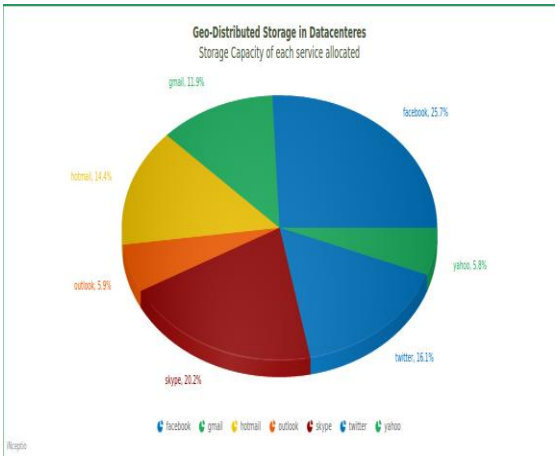

Figure 4. Pie chart of a few Geo distributed data centers storage capacities

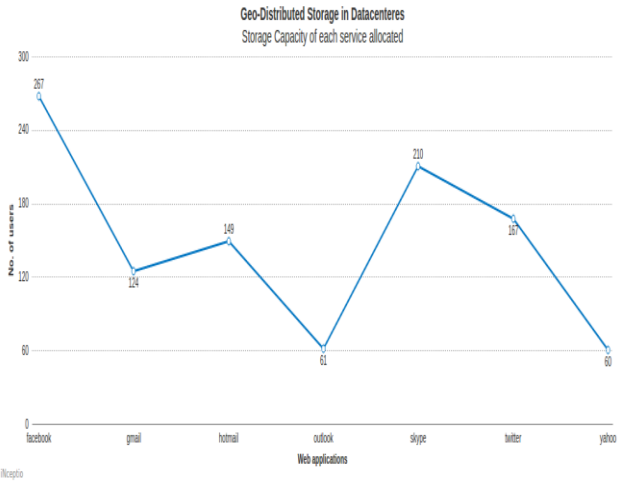

Figure 4. A graph representing geo distributed data processing analytics output

VI. Conclusion

In this paper we discussed about the distributed data processing with parallel technologies has been explained in detailed manner. Present data mining analytics following procedure is "data should be transferred to processor for mining", which is trying to integrate the different sources of data from various remote locations to a center location for processing. As aforementioned this procedure consumes more time to transfer the data from remote locations to central store, expects more bandwidth, exceeded utilization of resources, loss of data confidentiality and also raises several privacy and security issues. In this paper, we discuss a new big data mining procedure "Processor should be transferred to data", which consumes very less bandwidth and completes the mining process at high speed compared to previous procedures. This paper explains the possible methodologies available to implement the proposed procedure. Finally we discussed about the high level mining framework hadoop with map reduce technique.

VII. References

[1]. V. Mayer-Schoönberger and K. Cukier. Big data – a revolution that will transform how we live, work, and think. Eamon Dolan/Houghton Mifflin Harcourt, Chicago, Illinois 2013.

[2]. Wikipedia. Big data, 2014. http://en.wikipedia.org/wiki/Big_data, accessed April 2014.

[3]. VITRIA. The Operational Intelligence Company, 2014. http://blog.vitria.com, accessed April 2014.

[4]. E. Dumbill. What is Big Data? An Introduction to the Big Data Landscape, 2012. http://strata.oreilly.com/2012/01/what-is-big-data.html, accessed April 2014.

[5]. M. Stonebraker, P. Brown, and D. Moore. Object-relational DBMSs, tracking the next great wave. Morgan Kauffman Publishers, Inc., San Francisco, California, 2 edition, 1998.

[6]. Apache Hadoop. What Is Apache Hadoop?, 2014. http://hadoop.apache.org/, accessed April 2014.

[7]. Wikipedia. Apache Hadoop, 2014. http://en.wikipedia.org/wiki/Apache_Hadoop, accessed April 2014.

[8]. T. White. Hadoop – the definitive guide. O'Reilly Media, Inc., Sebastopol, California, 1 edition, 2009.

[9]. V. S. Patil and P. D. Soni. Hadoop Skeleton and Fault Tolerance in Hadoop Clusters, 2011. http://salsahpc.indiana.edu/b534projects/sites/default/files/public/0_Fault%20Tolerance%20in%20Hadoop%20for%20Work%20Migration_Evans,%20Jared%20Matthew.pdf, accessed April 2014.

[10]. Apache Hadoop. MapReduce Tutorial, 2013. https://hadoop.apache.org/docs/r1.2.1/mapred_tutorial.html, accessed April 2014.

[11]. Apache Hadoop. HDFS Architecture Guide, 2013. http://hadoop.apache.org/docs/r1.2.1/hdfs_design.html, accessed April 2014.

[12]. K. Kline, D. Kline, and B. Hunt. SQL in a nutshell, a desktop quick reference. O'Reilly Media, Sebastopol, California, 3 Edition, 2008.

[13]. P. J. Sadalage, and M. Fowler. NoSQL distilled, a brief guide to the emerging world of polygot persistence. Addison-Wesley, Reading, Massachusetts, 3 edition, 2013.

[14]. Johnston. Seminar on Collaboration as a Service – Cloud Computing, 2012. http://www.psirc.sg/events/seminar-on-collaboration-as-a-service-cloud-computing, accessed April 2014.

[15].