Volume: 08, Issue: 05 Pages: 10-13 (2017) Special Issue

A Review on Big Data Analytics with Hadoop Technology

S. Srimathy
Department of Information technology, Madurai Sivakasi Nadars Pioneer Meenakshi Women's College, Poovanthi.
Email:shreeharshini7@gmail.com

----------**ABSTRACT**----------

Big data is increasingly becoming factor in production, market competitiveness and the growth. Big data refers to massive, heterogeneous, and often unstructured digital content that is difficult to process using traditional data management tools and techniques. The term encompasses the complexity and variety of data and data types, real-time data collection and processing needs, and the value that can be obtained by smart analytics. Huge challenges must be overcome if the benefits are to be leveraged effectively. Matters of concern alongside increasing volumes of data, varying data structures and real-time processing include data security, data privacy policies that are in urgent need of reform and the rising quality expectations of the stakeholders. There is a widespread lack of suitable strategies to respond to the digital revolution.

Key words: **Existence, convergence, competitiveness.**

1. INTRODUCTION

Data is as a collection of large dataset that cannot be processed using traditional computing techniques. Big Data is not merely a data rather it has become a complete subject which involve various tools, techniques and framework. Big Data is a term that refers to dataset whose volume (size), complexity and rate of growth (velocity) make them to difficult to captured, managed, processed or analyzed by conventional technology and tools such as relational databases. The term "Big Data" has recently been applied to datasets that grow so large that they become awkward to work with using traditional database management systems. They are data sets whose size is beyond the ability of commonly used software tools and storage systems to capture, store, manage, as well as process the data within a tolerable elapsed time. Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. Consequently, some of the difficulties related to big data include capture, storage, search, sharing, analytics, and visualizing. Hence, big data analytics is where advanced analytic techniques are applied on big data sets.

- ✓ Analytics based on large data samples reveals and leverages business change. However, the larger the set of data, the more difficult it becomes to manage. Naturally, business benefit can commonly be derived from analyzing larger and more complex data sets that require real time or near-real time capabilities; however, this leads to a need for new data architectures, analytical methods, and tools.

2. THE CHALLENGES OF BIG DATA

2.1. Volume:

Volume refers to amount of data. volume represent the size of the data how the data is large. The size of the data is represented in terabytes and petabytes.

2.2. Variety:

Variety makes the data too big. The files comes in various formats and of any type, it may be structured or unstructured such as text, audio, videos, log files and more.

2.3. Velocity:

Velocity refers to the speed of data processing. The data comes at high speed. Sometimes 1 minute is too late so big data is time sensitive.

2.4. Value:

The potential value of Big data is huge. Value is main source for big data because it is important for businesses, IT infrastructure system to store large amount of values in database.

2.5. Veracity:

Veracity refers to noise, biases and abnormality. When we dealing with high volume, velocity and variety of data, the all of data are not going 100% correct, there will be dirty data.

3. HADOOP: SOLUTION FOR BIG DATA PROCESSING

Hadoop is an Apache open source framework written in Java that allows distributed processing of large dataset across cluster of computers using simple programming model. Hadoop creates cluster

of machines and coordinates work among them. It is designed to scale up from single servers to thousands of machines, each offering local computation and storage. Hadoop consists of two component Hadoop Distributed File System(HDFS)[1] and MapReduce Framework.

After the big data storage, comes the analytic processing. there are four critical requirements for big data processing. The first requirement is fast data loading. Since the disk and network traffic interferes with the query executions during data loading, it is necessary to reduce the data loading time. The second requirement is fast query processing. In order to satisfy the requirements of heavy workloads and real-time requests, many queries are response-time critical. Thus, the data placement structure must be capable of retaining high query processing speeds as the amounts of queries rapidly increase. Additionally, the third requirement for big data processing is the highly efficient utilization of storage space. Finally, the fourth requirement is the strong adaptivity to highly dynamic workload patterns.

As big data sets are analyzed by different applications and users, for different purposes, and in various ways, the underlying system should be highly adaptive to unexpected dynamics in data processing, and not specific to certain workload patterns. Map Reduce is a parallel programming model, inspired by the “Map” and “Reduce” of functional languages, which is suitable for big data processing. It is the core of Hadoop, and performs the data processing and analytics functions. The fundamental idea of MapReduce[4] is breaking a task down into stages and executing the stages in parallel in order to reduce the time needed to complete the task. The first phase of the MapReduce job is to map input values to a set of key/value pairs as output. The “Map” function accordingly partitions large computational tasks into smaller tasks, and assigns them to the appropriate key/value pairs.

Thus, unstructured data, such as text, can be mapped to a structured key/value pair, where, for example, the key could be the word in the text and the value is the number of occurrences of the word. This output is then the input to the “Reduce” function. Reduce then performs the collection and combination of this output, by combining all values which share the same key value, to provide the final result of the computational task. The MapReduce function within Hadoop depends on two different nodes: the Job Tracker and the Task Tracker nodes. The Job Tracker nodes are the ones which are responsible for distributing the mapper and reducer functions to the available Task Trackers, as well as monitoring the results. The MapReduce job starts by the Job- Tracker assigning a portion of an input file on the HDFS to a map task, running on a node. On the other hand, the Task Tracker nodes actually run the jobs and communicate results back to the Job Tracker. That communication between nodes is often through files and directories in HDFS[4], so inter-node communication is minimized.

Figure 1 shows how the MapReduce nodes and the HDFS work together. At step 1, there is a very large dataset including log files, sensor data, or anything of the sorts. The HDFS stores replicas of the data, represented by the blue, yellow, beige, and pink icons, across the data nodes. In step 2, the client defines and executes a map job and reduce job on a particular dataset, and sends them both to the job Tracker. The job Tracker then distributes the jobs across the Task Trackers in step 3. The Task Tracker runs the mapper, and the mapper produces output that is then stored in the HDFS file system. Finally in step 4, the reduce job runs across the mapped data in order to produce the result.

4. MOTIVATION: OUR DATA-DRIVEN WORLD

Advances in digital sensors, communications, computation, and storage have created huge collections of data, capturing information of value to business, science, government, and society. For example, search engine companies such as Google, Yahoo!, and Microsoft have created an entirely new business by capturing the information freely available on the World Wide Web and providing it to people in useful ways[2]. These companies collect trillions of bytes of data every day and continually add new services such as satellite images, driving directions, and image retrieval. The societal benefits of these services are immeasurable, having transformed how people find and make use of information on a daily basis. Some examples include:

Wal-Mart recently contracted with Hewlett Packard to construct a data warehouse capable of storing 4 petabytes (4000 trillion bytes) of data, representing every single purchase recorded by their point-of-sale terminals (around 267 million transactions per day) at their 6000 stores worldwide. By applying machine learning[8] to this data, they can detect patterns indicating the effectiveness of their pricing strategies and advertising campaigns, and better manage their inventory and supply chains.

Understanding the environment requires collecting and analyzing data from thousands of sensors monitoring air and water quality and meteorological conditions, another example of eScience. These measurements can then be used to guide simulations of climate and groundwater models to create reliable methods to predict the effects of long-term trends, such as increased

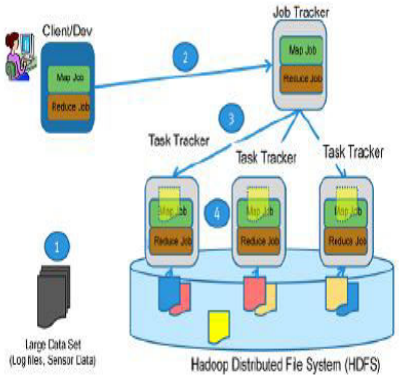

Fig. 1. MapReduce and HDFS

CO2 emissions and the use of chemical fertilizers.

Our intelligence agencies are being overwhelmed by the vast amounts of data being collected through satellite imagery, signal intercepts, and even from publicly available sources such as the Internet and news media. Finding and evaluating possible threats from this data requires "connecting the dots" between multiple sources[6], e.g., to automatically match the voice in an intercepted cell phone call with one in a video posted on a terrorist website.

The collection of all documents on the World Wide Web [6](several hundred trillion bytes of text) is proving to be a corpus that can be mined and processed in many different ways. For example, language translation programs can be guided by statistical language models generated by analyzing billions of documents in the source and target languages, as well as multilingual documents, such as the minutes of the United Nations. Specialized web crawlers scan for documents at different reading levels to aid English-language education for first graders to adults. A conceptual network of noun-verb associations has been constructed based on word combinations found in web documents to guide a research project at Carnegie Mellon University in which fMRI[5] images are used to detect how human brains store information.

These are but a small sample of the ways that all facets of commerce, science, society, and national security are being transformed by the availability of large amounts of data and the means to extract new forms of understanding from this data.

5. BIG-DATA TECHNOLOGY: SENSE, COLLECT, STORE, AND ANALYZE:

The rising importance of big-data computing stems from advances in many different technologies:

Sensors: Digital data are being generated by many different sources, including digital imagers (telescopes, video cameras, MRI machines), chemical and biological sensors (microarrays, environmental monitors), and even the millions of individuals and organizations generating web pages.

Computer networks: Data from the many different sources can be collected into massive data sets via localized sensor networks, as well as the Internet[7].

Data storage: Advances in magnetic disk technology have dramatically decreased the cost of storing data. For example, a one-terabyte disk drive, holding one trillion bytes of data, costs around $100. As a reference, it is estimated that if all of the text in all of the books in the Library of Congress could be converted to digital form, it would add up to only around 20 terabytes.

Cluster computer systems: A new form of computer systems, consisting of thousands of "nodes," each having several processors and disks, connected by high-speed local-area networks, has become the chosen hardware configuration for data-intensive computing systems. These clusters provide both the storage capacity for large data sets, and the computing power to organize the data, to analyze it, and to respond to queries about the data from remote users. Compared with traditional high-performance computing[7] (e.g., supercomputers), where the focus is on maximizing the raw computing power of a system, cluster computers are designed to maximize the reliability and efficiency with which they can manage and analyze very large data sets. The "trick" is in the software algorithms – cluster computer systems are composed of huge numbers of cheap commodity hardware parts, with scalability, reliability, and programmability achieved by new software paradigms.

Cloud computing facilities: The rise of large data centers and cluster computers has created a new business model, where businesses and individuals can rent storage and computing capacity, rather than making the large capital investments needed to construct and provision large-scale computer installations. For example, Amazon Web Services (AWS) provides both network-accessible storage priced by the gigabyte-month and computing cycles priced by the CPU-hour. Just as few organizations operate their own power plants, we can foresee an era where data storage and computing become utilities that are ubiquitously available.

6. DATA ANALYTICS SERVICE MODELS

The SaaS model offers complete big data analytics applications to end users, who can exploit

the cloud's scalability in both data storage and processing power to execute analysis on large or complex datasets. The PaaS model provides data analytics programming suites and environments in which data mining developers can design scalable analytics services and applications. Researchers can exploit the IaaS model to compose a set of virtualized hardware and software resources for running data analysis frameworks or applications. Developers can implement big data analytics services within each of these three models:

• Data analytics software as a service—provides a well-defined data mining algorithm or ready to-use knowledge discovery tool as an Internet service to end users, who can access it directly through a Web browser;

• Data analytics platform as a service—provides a supporting platform that developers can use to build their own data analytics applications or extend existing ones without concern about the underlying infrastructure or distributed computing

Issues;

• Data analytics infrastructure as a service—provides a set of virtualized resources that developers can use as a computing infrastructure to run data mining applications or to implement data analytics systems from scratch.

CONCLUSION

Big data is the new and that follows a series of logical stages in the development of the internet, such as individualization, the relocation of data to the cloud and the rapidly growing demand for digital mobility. It bridges the gap to what has evolved before. In principle, the idea is to combine different volumes of data with new data sets and to identify any patterns in this aggregated data using intelligent software, with the ultimate aim of drawing the right or possible conclusions from the findings. Once they have been compiled, primary data sets can be analyzed any number of times for different purposes and for different stakeholders. The data functions as a driver of innovation, creativity and out-of-the-box thinking, and in an ideal world results in new business ideas, products or services.

REFERENCES

[1] Andrew Pavlo, "A Comparison of Approaches toLarge-Scale Data Analysis", SIGMOD, 2009.
[2] Apache Hadoop: http://Hadoop.apache.org
[3] Dean, J. and Ghemawat, S., "MapReduce: a flexible data processing tool", ACM 2010.
[4] DeWitt & Stonebraker, "MapReduce:A major step backwards", 2008
[5] Hadoop Distributed File System, http://hadoop.apache.org/hdfs
[6] Hadoop Tutorial http://developer.yahoo.com/hadoop/tutorial/module1.html
[7] Big Data Analytics: Nada Elgendy, Ahmed Elragal
[8] Scalable system for processing Big Data, Dawei Jiang, Gang Chen, Sai Wu