International Journal of Latest Trends in Engineering and Technology
Vol.(8)Issue(4), pp.143-146
DOI: http://dx.doi.org/10.21172/1.84.22
e-ISSN:2278-621X

Big Data Analytics: Survey Paper

Haval Mohammed Sidqi[1], Rzgar Sirwan Raza[2], Dlshad Jaafr Hussin[3]

Abstract- Big data in simple terms is extremely huge sets of data which can reveal trends, interests & patterns and classify data upon computational analysis. Big data refers to a data set that not only large, but also in the creation of a high speed, which marks it difficult to compact with traditional and technological tools. This is owed to the rapid growth of data, you must to learning and to run the knowledge to deal with it and extract value from this group solution. In addition, it would be the decision-makers and can get the value of this information is different and rapidly changing tasks, everything from data transferences daily social network clients business.

enhance security, and preventing loss of data, and computational cost as well.

Keywords – big data, data mining, analytics.

I. INTRODUCTION

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 can be analyzed for insights that lead to better decisions and strategic business moves. World without data storage, where all the details about the person or the performance of each transaction, or every feature, which can be documented is to be lost nearly after use. Therefore, of loss of the ability to extract the value of information and knowledge, and careful analysis, as well as a new opening and the use provided. Strips anything that starts with the name and address of the customer, the goods, and thebuying made, the employee who was hired, and therefore it was necessary to continue day after day. Information cornerstone of any organization this growth condition. Currently on how much detail and wave data and information are now offering think, through advances in technology and the Internet with increasing storage size and data collection method, and the huge amounts of data it was readily-available. Made every second of data, more and more, the need to store and analyzed in order to extract value. in addition, it becomes known low-cost to store, so the association need to get the most out of the possible huge amounts of stored data. To quickly requires alter the size and diversity of this data a new type of analysis of large data, as well as store and analyze as different. It has a share of great this should be properly analyzed and the information must be extracted data.Literature based on news and talk about an issue that was significantIt is linked to a large data, in order to serve the purpose of our research. Year From 2008-2013, with many items on a large data ranging from 2011-2013.This due to large data recently focused on the issue. Furthermore it, Blog consists mostly of some of the top research journal, conference, and White book by leading companies in the industry. Due to the process a long review.

[1] *Sulaimani Polytechnic University, Phd Candidate in Computer Science*
[2] *Department of Computer Science, College of Science and Technology University of Human Development Qaradagh, Sulaymaniyah, Kurdistan Region, Iraq*
[3] *Sulaimani Polytechnic University*

II. CHALLENGES WITH BIG DATA SERVICES

The traditional data processing techniques are not competent enough to proficiently use these terabytes and petabytes of data pouring in organizations today. Every digital process produces big data which can be used to solve large problems with competitive strategy by unifying the underlying architecture. Almost all organizations today need to exploit the power of big data. Big data is not about the size, but it can offer insights into various aspects of data being generated rapidly and constantly. It offers opportunities to harness it for business growth. IT teams are burdened with ever-growing requests for data analyses and reports. Data visualization is becoming an increasingly important element of analytics in the age of big data. Organizations need to meet the need for speed. They need to explore huge data volumes and gain insights in the data. This huge data is irrelevant if not understood. The quality of this data need to be maintained for the appropriate audiences and needs. Data is valuable for decision making and analysis.

III. BIG DATA SERVICES

1. Big Data Implementation : We offer end to end installation, administration and configuration of Hadoop and other big data tools, developing map reduce programs according to your business needs, providing SQL query like interface to analyze and visualize your consolidated data.
2. Big Data Optimization : We provide you Big Data Applications with minimum costs and improved resource utilization to improve the efficiency of analytics algorithm.
3. Big Data Analytics : Gain Hadoop-based big data analytics platform, real time analytics for better decision making, report Visulization and dashboards, Big Data Analytics Automation and much more.
4. Custom Big Data Solutions : We proficiently utilize the power of Big Data to provide with customized solutions for gaining deeper insights into the data.
5. Big Data Cluster Management : We provide automated web based big data cluster management and monitoring services. AppPerfect's QueryIO is an open source tool for big data analytics and cluster management.
6. Data Mining & Data Aggregation : With Data Mining & Data Aggregation, you can collect data, integrate and extract metadata, transform and aggregate data, extract useful knowledge from data and evaluate and visualize the knowledge.

V. BIG DATA STORAGE

The volatile growth of data has more strict requirements on storage and management. In this section, we focus on the storage of big data. Big data storage refers to the storage and management of large-scale datasets while achieving reliability and availability of data accessing. We will review important issues including massive storage systems, distributed storage systems, and big data storage mechanisms. On one hand, the storage infrastructure needs to provide information storage service with reliable storage space; on the other hand, it must provide a powerful access interface for query and analysis of a large amount of data.

Traditionally, as auxiliary equipment of server, data storage device is used to store, manage, look up, and analyze data with structured RDBMSs. With the sharp growth of data, data storage device is becoming increasingly more important, and many Internet companies pursue big capacity of storage to be competitive. Therefore, there is a compelling need for research on data storage. Storage system for massive data.Various storage systems emerge to meet the demands of massive data. Existing massive storage technologies canbe classified as Direct Attached Storage (DAS) and network storage, while network storage can be further classified into Network Attached Storage (NAS) and St orage Area Network (SAN). In DAS, various harddisks are directly connected with servers, and data anagement is server-centric, such that storage devices are peripheral equipments, each of which takes a certain amount of I/O resource and is managed by an individual application software. For this reason, DAS is only suitable to interconnect servers with a small scale. However, due to its low scalability, DAS will exhibit undesirable efficiency when the storage capacity is increased, i.e., the upgrade ability and expandability are greatly limited. Thus, DAS is mainly used in personal computers and small-sized servers. Network storage is to utilize network to provide users with a union interface for data access and sharing. Network storage equipment includes special data exchange equipments, disk array, tap library, and other storage media, as well as special storage software. It is characterized with strong expandability. NAS is actually an auxillary storage equipment of a network. It is directly connected to a network through a hub or switch through TCP/IP protocols. In NAS, data is transmitted in the form of files. Compared to DAS, the I/O burden at a NAS server is reduced extensively since the server accesses a storage device indirectly through a network.

While NAS is network-oriented, SAN is especially designed for data storage with a scalable and bandwidth intensive network, e.g., a high-speed network with optical fiber connections. In SAN, data storage management is

relatively independent within a storage local area network, where multipath based data switching among any internal nodes is utilized to achieve a maximum degree of data sharing and data management. From the organization of a data storage system, DAS, NAS, and SAN can all be divided into three parts:

1. disc array: it is the foundation of a storage system and the fundamental guarantee for data storage;
2. connection and network sub- systems, which provide connection among one or more disc arrays and servers;
3. storage management software, which handles data sharing, disaster recovery, and other storage management tasks of multiple servers.

IV. BIG DATA ANALYTIC PROCESSING

After the big data storage, comes the analytic processing. According to [10], 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. Since the rapid growth in user activities can demand scalable storage capacity and computing power, limited disk space necessitates that data storage be well managed during processing, and issues on how to store the data so that space utilization is maximized be addressed. Finally, the fourth requirement is the strongadaptively 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 [10]. 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 [6]. Accordingto EMC, the MapReduce paradigm is based on adding more computers or resources, rather than increasing the power or storage capacity of a single computer; in other words, scaling out rather than scaling up [9]. The fundamental idea of MapReduce is breaking a task down into stages and executing the stages in parallel in order to reduce the time needed to complete the task [6]. 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 intosmaller tasks, and assigns them to the appropriate key/value pairs [6]. 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 [9]. 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 [6]. 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 [9]. 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 [13]. 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, so inter-node communication is minimized [9]. Figure 1 shows how the there is a very large dataset The HDFS stores , across the Data Node a reduce job on a particular Tracker then distributes the runs the mapper, and the m system. Finally, in step 4, the result. Hadoop is a MAD system data as files into the distribtations on the data. Hadoop loaded into Hadoop simply MapReduce interprets the data is capable of attracting all d tions that may occur in such After big data is stored

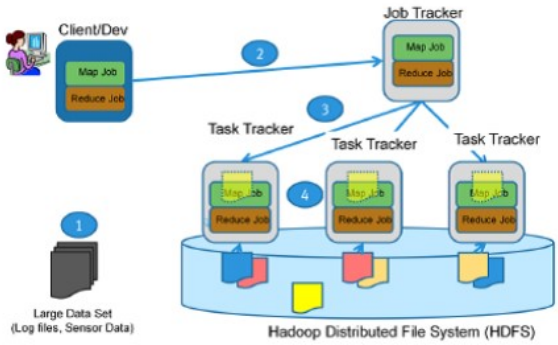

Fig. 1. MapReduce and HDFS

V.COMPARISONS OF STRUCTURED AND NON STRUCTURED

Structured data sets are those where the activity of processing and output is predetermined and highly organized. Structured systems are designed. Payroll, Inventory control systems, point of sale systems, airline reservations are all forms of structured systems since they are using structured data- the data which is stored and displayed as a set of rows and tables. In contrast, unstructured data sets are the data that have little or no predetermined form or structure. Unstructured data sets include email, contracts, blogs, and other communications. A person who performs a communications activity in an unstructured system has wide latitude to structure the message in whatever form is desired. The rules of unstructured systems are fewer and less complex [9, 10, and 13]. The structured and nano structured data can be different from technical, organizational, structural, functional point of view. Each has its own environment and needs to be treated and used accordingly. The structured and nano structured data can be different from technical, organizational, structural and functional point of view. [4, 11, 15, 18, 19]. Relational databases are highly structured: all the data in the table are stored as rows and columns. Each column has a data type which is mostly normalized. The SQL is suitable to relational databases to store and retrieve data in a structured way. Queries are Plain English commands. There are always fixed number of columns although additional columns can be added later. Most of the tables are related to each other with primary and foreign keys thus providing "Referential Integrity" among the objects. The major vendors are ORACLE, SQL Server, MySQL, PostgreSQL, etc. [7, 9].

VI. REFERENCES

[1] "Big Data: Volume, Velocity, Variability, Variety", http://nosql.mypopescu.com/post/6361838342/bigdata-volume-velocity-variability-variety, Accessed April 2015.
[2] B. Wiederhold, "18 essential Hadoop tools for crunching big data", Network World, www.googletagmanager.com/ns.html. Accessed April, 2015.
[3] A. K. Zaki, "NoSQL Database: New Millennium Database for Big Data , Big Users, Coud Computing and Its Security Challenges," http://esatjournals.org/Volumes/IJRET/2014V03/I15/IJRET20140315080.pdf, Accessed May 2015.
[4] L. Arthur, "What is Big Data", htpp:://www.forbes.com/sites/liasaarthur/2013/08/15/what-is-big-data/, Accessed May 2105
[5] S. Penchikala, "Virtual Panel: Security Considerations in Accessing NoSQL Databases‖, Nov. 2011. http://www.infoq.com/articles/nosql-data-security-virtual- panel., Accessed May 2015.
[6] Oracle Databases from web: http://www.oracle.com/us/products/database/overview/index.h tml, Accessed May 2015.
[7] Relational Database Management System (RDBMS) vsnoSQL. http://openproceedings.org/html/pages/2015_edbt.html, Accessed April 2015.
[8] "Relational -database-management-system-rdbms-vs-nosql/, http://www.loginradius.com/engineering/relational-database-management-system-rdbms-vs-nosql/ Accessed April 2015.
[9] M. Ramachandran "Relational Vs Non-Relational databases", http://bigdata-madesimple.com/relational-vs-non-relational-databases, Accessed May 2015.
[10] L. P. Issac "SQL vsNoSQL Database Differences Explained with few Example DB", http://www.thegeekstuff.com/2014/01/sql-vs-nosql-db/, Accessed May 2015.
[11] Sherpa Software. "Structured and Unstructured Data: What is It? http://www.sherpasoftware.com/blog/structured-and-unstructured-data-what-is-it/, Accessed May 2015.
[12] F. Chang, et al. "Bigtable: A distributed storage system for structured data.", ACM Transactions on Computer Systems (TOCS) 26.2 (2008): 4.
[13] D. Gosain, "A survey and comparison of relational and non-Relational Databases‖. IJERT,Vol 1,Issue 6,2012.
[14] *L. Okman, , N. Gal-Oz, Y. Gonen, E. Gudes, and J. Abramov, , "Security Issues in NoSQL*Databases," Trust, Security and Privacy in Computing and Communications (TrustCom), 2
[15] IEEE 10th International Conference on , vol., no., pp.541-547, 16-18 Nov. 2011 doi: 10.1109/TrustCom.2011.70
[16] Find cloud security alliance https://cloudsecurityalliance.org/research/big-data