International Journal of Computer Science and Engineering Communications
Volume.4, Issue.6 (2016): Page.1481-1487
www.scientistlink.com/ijcsec

Deliberation on Big Data

Reneesh TR[1] ,S.Thilagavathi[2], Kavitha.K[3]
PG Scholar, Assistant Professor
Department of Computer Applications
Hindusthan College of Engineering and Technology, Coimbatore, India.
18eneesh20@gmail.com,thilaga.hicet@gmail.com

Abstract: *Big data isn't just a technology it's a business strategy for capitalizing on information resources. Big Data is data whose scale, diversity, and complexity require new architecture, techniques, algorithms, and analytics to manage it and extract value and hidden knowledge from it."A massive volume of both structured and unstructured data that is so large that it's difficult to process with traditional database and software techniques." Big Data has become a valuable resource and mechanism for the practitioners and researchers to explore the value of data sets in all kinds of business scenarios and scientific investigations. New computing platforms such as cloud computing, mobile Internet, social network are driving the innovations of big data. From industry perspective, IBM, SAP, Oracle, Google, Microsoft, Yahoo, and other leading software and internet service companies have also launched their own innovation initiatives around big data.*

Keywords: *OLTP,OLAP,HDFS,RTAP*

INTRODUCTION

In 2010, enterprises and users stored more than 13 Exabyte's of new data; this is over 50,000 times the data in the Library of Congress. The potential value of global personal location data is estimated to be $700 billion to end users, and it can result in an up to 50% decrease in product development and assembly costs. Increasingly large numbers of embedded sensors, Smartphone's, PCs, and tablet computers connected to network are generating enormous amounts of data. This data creates new opportunities to "extract more value" for the areas that it is needed. Today the world enters the age of "Big Data." Just as this data is generated by people in real time, it can be analysed in real time by high performance computing networks, thus creating a potential for improved decision-making. The International Data Corporation (IDC) believes organizations that are best able to make real-time business decisions using Big Data solutions will thrive, while those that are unable to embrace and make use of this shift will increasingly find themselves at a competitive disadvantage in the market and face potential failure while the potential benefits of Big Data are real and significant, and some initial successes have already been achieved (such as the Sloan Digital Sky Survey), there remain many technical challenges that must be addressed to fully realize this potential. The sheer size of the data, of course, is a major challenge, and is the one that is most easily recognized. However, there are others. Industry analysis companies like to point out that there are challenges not just in Volume, but also in Variety and Velocity, and that companies should not focus on just the first of these. By Variety, they usually mean heterogeneity of data types, representation, and semantic interpretation. By Velocity, they mean both the rate at which data arrive and the time in which it must be acted upon. While these three are important, this short list fails to include additional important requirements such as privacy and usability. The analysis of Big Data involves multiple distinct phases as shown in the figure below, each of which introduces challenges. Many people unfortunately focus just on the analysis/modelling phase: while that phase is crucial, it is of little use without the other phases of the data analysis pipeline. Even in the analysis phase, which has received much attention, there are poorly understood complexities in the context of multi-tenanted clusters where several users' programs run concurrently. Many significant challenges extend beyond the analysis phase. It may need to figure out good questions based on the data.

Doing this will require smarter systems and also better support for user interaction with the analysis pipeline. In fact, currently have a major bottleneck in the number of people empowered to ask questions of the data and analyze it [NYT2012]. Solutions to problems such as this will not come from incremental improvements to business as usual such as industry may make on its own. Rather, they require us to fundamentally rethink how to manage data analysis.

2. DEFINITION

Big data usually includes data sets with sizes beyond the ability of commonly used software tools to capture, curate, manage, and process data within a tolerable elapsed time .Big data "size" is a constantly moving target, as of 2012 ranging from a few dozen terabytes to many petabyte of data. Big data is a set of techniques and technologies that require new forms of integration to uncover large hidden values from large datasets that are diverse, complex, and of a massive scale. In a 2001 research report and related lectures, analyst Doug Laney defined data growth challenges and opportunities as being three-dimensional, i.e. increasing volume (amount of data), velocity (speed of data in and out), and variety (range of data types and sources). Gartner, and now much of the industry, continue to use this "3Vs" model for describing big data .In 2012, updated its definition as follows: "Big data is high volume, high velocity, and/or high variety information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization." Additionally, a new V "Veracity" is added by some organizations to describe it.If Gartner's definition (the 3Vs) is still widely used, the growing maturity of the concept fosters a more sound difference between big data and Business Intelligence regarding data and their use: Business Intelligence uses descriptive statistics with data with high information density to measure things, detect trends etc.; Big data uses inductive statistics and concepts from nonlinear system identification to infer laws (regressions, nonlinear relationships, and causal effects) from large sets of data with low information density to reveal relationships, dependencies and perform predictions of outcomes and behaviours. A more recent, consensual definition states that "Big Data represents the Information assets characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value".

3.CHARACTERISTICS

Big data can be described by the following characteristics:
Volume – The quantity of data that is generated is very important in this context. It is the size of the data which determines the value and potential of the data under consideration and whether it can actually be considered Big Data or not. The name 'Big Data' itself contains a term which is related to size and hence the characteristic.
Variety - The next aspect of Big Data is its variety. This means that the category to which Big Data belongs to is also a very essential fact that needs to be known by the data analysts. This helps the people, who are closely analyzing the data and are associated with it, to effectively use the data to their advantage and thus upholding the importance of the Big Data.
Velocity - The term 'velocity' in the context refers to the speed of generation of data or how fast the data is generated and processed to meet the demands and the challenges which lie ahead in the path of growth and development.
Variability - This is a factor which can be a problem for those who analyse the data. This refers to the inconsistency which can be shown by the data at times, thus hampering the process of being able to handle and manage the data effectively.
Veracity - The quality of the data being captured can vary greatly. Accuracy of analysis depends on the veracity of the source data.

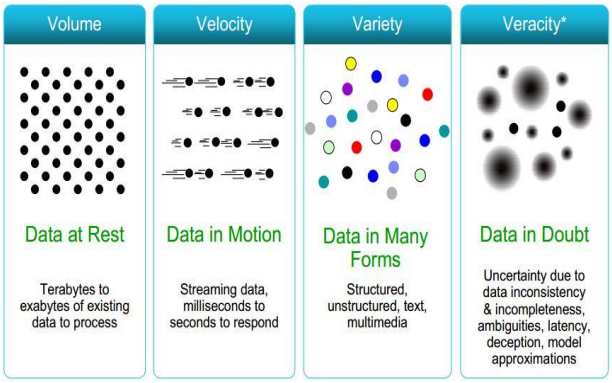

Figure 3.1 Characteristics of big data

4. ADOPTION OF BIG DATA

Storage capacity has doubled roughly every three years since the 1980s. Every day in the world generate 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone

1. Twitter Generate approximately 12 TB of data per day
2. New York Stock Exchange 1 TB of data everyday
3. 3 Billion RFID tags today (1.3B in 2005)
4. 76Million smart meters in 2009…200M by 2015
5. 6billion camera phones world wide
6. 100s of millions of GPS enabled devices sold annually
7. 2+ billion people on the Web by end 2011

There is a great challenge not only to store and manage the large volume of data, but also to analyze and extract meaningful information from it. There are several approaches to collecting, storing, processing, and analyzing big data. Unstructured data that could provide a real-time business decision support remains unused as they cannot be stored, processed or analyzed.

5.IMPLEMENTATION OF BIG DATA

HADOOP: Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines or racks of machines) are commonplace and thus should be automatically handled in software by the framework. The Hadoop framework itself is mostly written in the Java programming language, with some native code in c and command line utilities written as Shell script. For end-users, though Map Reduce Java code is common, any programming language can be used with "Hadoop Streaming" to implement the "map" and "reduce" parts of the user's program. Other related projects expose other higher-level user interfaces. The core of Apache Hadoop consists of a storage part (Hadoop Distributed File System (HDFS)) and a processing part (Map Reduce). Hadoop splits files into large blocks and distributes them amongst the nodes in the cluster. To process the data, Hadoop Map Reduce transfers packaged code for nodes to process in parallel, based on the data each node needs to process. This approach takes advantage of data locality nodes manipulating the data that they have on hand to allow the data to be processed faster and more efficiently than it would be in a more conventional supercomputer architecture that relies on a parallel file system where computation and data are connected via high-speed networking.

Hadoop Common: The common utilities that support the other Hadoop modules.
Hadoop Distributed File System (HDFS™): A distributed file system that provides high-throughput access to application data.
Hadoop YARN: A framework for job scheduling and cluster resource management.
Hadoop MapReduce: A YARN-based system for parallel processing of large data sets.
Hive: The Apache Hive ™ data warehouse software facilitates querying and managing large datasets residing in distributed storage. Hive provides a mechanism to project structure onto this data and query the data using a SQL-like language called HiveQL. At the same time this language also allows traditional map/reduce programmers to plug in their custom mappers and reducers when it is inconvenient or inefficient to express this logic in HiveQL

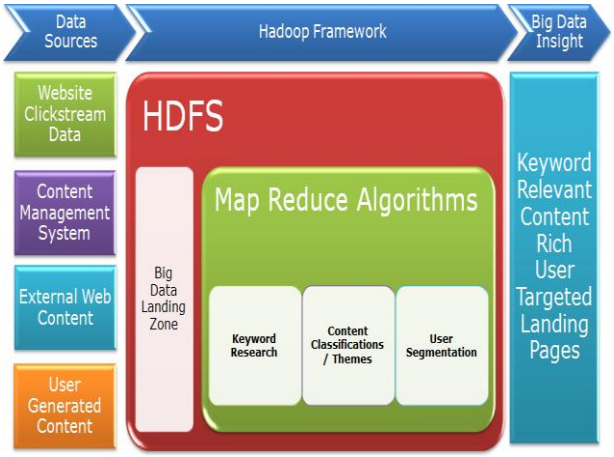

Figure 5.1.Architecture of Hadoop

Hbase:Use Apache HBase when you need random, real time read/write access to your Big Data. This project's goal is the hosting of very large tables -- billions of rows X millions of columns -- atop clusters of commodity hardware. Apache HBase is an open-source, distributed, versioned, non-relational database modelled after Google's Bigtable: A Distributed Storage System for Structured Data by Chang et al. Just as Bigtable leverages the distributed data storage provided by the Google File System, Apache HBase provides Bigtable-like capabilities on top of Hadoop and HDFS.
Pig:Apache Pig is a platform for analyzing large data sets that consists of a high-level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. The salient property of Pig programs is that their structure is amenable to substantial parallelization, which in turns enables them to handle very large data sets.
At the present time, Pig's infrastructure layer consists of a compiler that produces sequences of Map-Reduce programs, for which large-scale parallel implementations already exist (e.g., the Hadoop subproject). Pig's language layer currently consists of a textual language called Pig Latin, which has the following key properties:
Ease of programming: It is trivial to achieve parallel execution of simple, "embarrassingly parallel" data analysis tasks. Complex tasks comprised of multiple interrelated data transformations are explicitly encoded as data flow sequences, making them easy to write, understand, and maintain.
Optimization opportunities: The way in which tasks are encoded permits the system to optimize their execution automatically, allowing the user to focus on semantics rather than efficiency.
Extensibility: Users can create their own functions to do special-purpose processing.

MAP REDUCE : MapReduce is a programming model for processing and generating large data sets with a parallel, distributed algorithm on a cluster. Conceptually similar approaches have been very well known since 1995 with the Message Passing Interface standard having reduce and scatter operations. A MapReduce program is composed of a Map() procedure that performs filtering and

sorting (such as sorting students by first name into queues, one queue for each name) and a Reduce() procedure that performs a summary operation (such as counting the number of students in each queue, yielding name frequencies).
The "MapReduce System" (also called "infrastructure" or "framework") orchestrates the processing by marshalling the distributed servers, running the various tasks in parallel, managing all communications and data transfers between the various parts of the system, and providing for redundancy and fault tolerance. The model is inspired by the map and reduce functions commonly used in functional programming, although their purpose in the MapReduce framework is not the same as in their original forms. The key contributions of the MapReduce framework are not the actual map and reduce functions, but the scalability and fault-tolerance achieved for a variety of applications by optimizing the execution engine once. As such, a single-threaded implementation of MapReduce will usually not be faster than a traditional (non-MapReduce) implementation, any gains are usually only seen with multi-threaded implementations. The use of this model is beneficial only when the optimized distributed shuffle operation (which reduces network communication cost) and fault tolerance features of the MapReduce framework come into play. Optimizing the communication cost is essential to a good MapReduce algorithm.

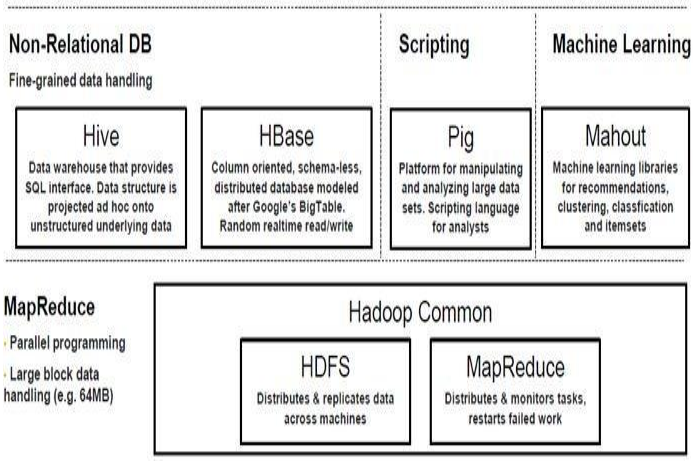

Figure 5.2 Architecture of map reduce

MapReduce libraries have been written in many programming languages, with different levels of optimization. A popular open-source implementation that has support for distributed shuffles is part of Apache Hadoop. The name MapReduce originally referred to the proprietary Google technology, but has since been genericized.

6. APPLICATIONS

Big data applications solve and analyze real world problems using Hadoop and associated tools. Internet users and machine-to-machine connections are causing the data growth. Real time areas are defined following in which big data is used:

Big data in healthcare: Healthcare practices and policies differ tremendously around the world, there are three objectives regarding healthcare system [10]. The first objective is to improve the patient experience (including quality and satisfaction). Second, improving overall population health and reducing the cost of health care and third is traditional methods have fallen short to manage healthcare and create modern technology to analyze large quantities of information. It is time consuming for clinical staff to Collecting massive amounts of data in healthcare. High-performance analytics are new technologies making easier to turn massive amounts of data into relevant and critical insights used to provide better care. Analytics helps to predict negative intervene and reactions.

Network Security: Big data is changing the landscape of security technologies. The tremendous role of big data can be seen in network monitoring, forensics and SIEM . Big data can also create a world where maintaining control over the revelation of our personal information is challenged constantly. Present analytical techniques don't work well at large scales and end up producing false positives that their efficacy is undermined and enterprises move to cloud architectures and gather much more data, the problem is becoming worse. Big data analytics is an effective solution for processing of large scale information as security is major concern in enterprises. Fraud detection is uses for big data analytics. Phone and credit card companies have conducted large-scale fraud detection for decades. Mainly big data tools are particularly suited to become fundamental for forensics and ATP.

Sports: Sport, in business, an increasing volume of information is being collected and captured. Technological advances will fuel exponential growth in this area for the foreseeable future, as athletes are continuously monitored by tools as diverse as sports daily saliva, GPS systems and heart rate monitors tests. These statistics and many more like them are high performance in Big Data. These numbers there is a massive amount of potential insight and intelligence for trainers, administrators, coaches, athletes, sports medics and players. Statistics can be analyzed and collected to better understand what are the critical factors for optimum performance and success, in all facets of elite sport. Injury prevention, competition, Preparation, and rehabilitation can all benefit by applying this approach. Recruitment, Scouting and retention can also be enhanced by these powerful principles. Keeping an eye on various information a coach or a manager can easily and quickly understand which athletes and players need additional support, training, and guidance. Areas for reasons for success and improvement will be understood more clearly. Used consistently this is a powerful measure of progress and performance.

Education Systems: By using big data analytics in field of education systems, remarkable results can be seen [9]. Data on students online behaviour can provide educators with important insights, such as if a student requires more attention, the class understanding of a topic is not clear, or if the course has to be modified. Students are required to answer accompanying questions as they go through the set of online content before class. By tracking the number of students that have completed the online module, the time taken and accuracy of their answers, a lecturer can be better informed of the profile of his students and modify the lesson plan accordingly. The analysis of data also clarify about the interest of student looking at time spent in online textbook, online lectures, notes etc. As result instructor can guide choosing the future path effectively.

Gaming industry: The amount of data that video game players are generating on a daily basis is growing rapidly. Video game developers are using variety of IT techniques such as Hadoop to keep up the massive amount of gaming data that's generated every day. People are playing video game and generated lot of data in separate areas: game data, player data and session data. In order to improve their game development, game experience, studios are turning to commercial Hadoop distributions such as MapR to analyze, collect and process data from these massive data streams. Armed with this valuable insight from big data, video game publishers are now able to enhance game player engagement and increase player retention by analyzing gamers' social behaviour activity and tracking players' statistics, calculating rewards, quickly generating leader boards, changing game play and mechanics and delivering virtual prizes, so that experienced players will continue to play the game. By using advanced analytics to uncover rich player insights, developers can now focus on creating meaningful gaming experiences for their customers.

Telecommunication Industry: Telecommunications companies have unique advantage in marketplace by controlling the communication infrastructure. Today big challenges for telecommunication are volume, variety and complexity. Current data systems based on batch processing and traditional relation technology, they process big data in real time. Telcos combine ETL and traditional relational databases with big data technologies on a single platform . Telcos technology parses, transforms and integrates the vast amount of data generated by location sensors, IPv6 devices, clickstream, CDRs, 4G networks and machine to machine monitors' information.

Telcos parse and transforms from multiple formats and source including unstructured mobile, media, web and machine monitor provide data. Telcos masking, managing and identifying sensitive data for regulatory compliance. Cloud data integration helps to control over off-premise data managed in the cloud.

7.FUTURE SCOPE

The new applications are generating vast amount of data in structured and unstructured form. Big data is able to process and store that data and probably in more amounts in near future. Hopefully, Hadoop will get better. New technologies and tools that have ability to record, monitor measure and combine all kinds of data around us, are going to be introduced soon. New technologies and tools are needed for anonymising data, analysis, tracking and auditing information, sharing and managing personal data in future. So many aspects of life health, education, telecommunication, marketing, sports and business etc that manages big data world need to be polished in future.

8.CONCLUSION

Big data provides an opportunity for "big analysis" leading to "big opportunities" to gain a competitive edge, to advance the quality of life, or to solve the mysteries of the world. Fast forward to today and these worlds are rapidly coming together. The ability to analyse and store massive amount of structured, unstructured and semi-structure data promises ongoing opportunities for academic institutes, businesses and government organizations. However, a common horizontal big data analytics platform is necessary to support these varieties of real time applications that include healthcare, security, market and business, sports, education system, gaming industry, telecommunications and probably many others in future. Furthermore, challenges of big data, 5 V's volume, velocity, variety, value, veracity and cloud enabled bigdata with models and types are also described in this paper. The main goal of this paper is to make a survey of various big data applications that are use in IT industries or organisation to store massive amount of data using technologies (Hadoop, HIVE, NoSQL, Mapreduce and HPCC).

REFERENCE

[1] Dr. Ajay Rana, Sangeeta Bansal, Transitioning from Relational Databases to Big Data . Volume 4, Issue 1, January 2014
[2] Sam Madden From Databases to Big Data – Article published in IEEE Internet Computing magazine
[3] Kapil Bakshi Considerations for Big Data: Architecture and Approach – Paper published in IEEE
[4] Sachidanand Singh Big Data Analytics – Paper published in 2012 International Conference on Communication Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India
[5] Introduction to Big Data: Infrastructure and Networking", White Paper Juniper Networks 2012.
[6] "Big Data A New World of Opportunities",NESSI White Paper, December 2012
[7] Sabia, and Sheetal Kalra Applications of big Data: Current Status and Future Scope
[8] http://hadoop.apache.org/