IJCSE International Journal of Computer Sciences and Engineering Open Access
Review Paper Volume-4, Issue-5 E-ISSN: 2347-2693

Applications of Big Data in various Domains

M. Kumarasamy[1*] and G. N. K. Suresh Babu[2]

[1]Department of Computer Science, Villa College, Male, Maldives
[2]Department of Computer Applications, Acharya Institute of Technology, Bangalore, India

Available online at: www.ijcseonline.org

Received: Apr/28/2016 Revised: May/09/2016 Accepted: May/22/2016 Published: May/31/2016

Abstract -- The term Big data is very popular recently in all the domains. Every where and every body talking about big data numerously. The goal of this paper is to describe what is big data and how it can be used in various applications. The rising number of applications serving millions of users and dealing with terabytes of data need to a faster processing paradigms. Recently, there is growing enthusiasm for the notion of big data analysis. Big data analysis becomes a very important aspect for growth productivity, reliability and quality of services. Processing of big data using a powerful machine is not efficient solution. So, companies focused on using Hadoop software for big data analysis. This is because Hadoop designed to support parallel and distributed data processing. Hadoop provides a distributed file processing system that stores and processes a large scale of data. The author tries to give the introduction about Hadoop and Map Reduce architecture. The main goal of this paper is applications of big data in various domains and how to build decision support system using big data. Big data have applications in many fields such as Business, Technology, Health Care, Smart cities etc. These applications will allow people to have better services, better customer experiences, and also to prevent and detect illness much easier than before.

Keywords : Big Data, Cloud Computing, Data Mining, Business, Hadoop and Map Reduce.

I INTRODUCTION

The purpose of big data is build decision support system for the organizations to take the correct decisions. Data is easier to capture and access through third parties such as Facebook, Twitter, LinkedIn and others. Geo location data, social graphs, user-generated content, user's personal information, machine logging data, and sensor-generated data are just a few examples of the array of data captured [1]. It's not surprising that developers find increasing value in leveraging this data to enrich existing applications and create new ones made possible by it. The use of the data is rapidly changing the nature of communication, shopping, advertising, entertainment, and relationship management. Applications that don't find ways to leverage it quickly will quickly fall behind. Scientists regularly face problems because of large data sets in many areas, including meteorology, genomics; complex physics simulations, biological environmental research, internet search, and finance and business informatics. Data sets grow in size in part because they increasingly gathered by widespread information-sensing mobile, remote sensing, software logs, cameras, microphones, radio frequency identification readers, and wireless sensor networks. Big data is difficult to work with using relational databases and desktop statistics, requiring instead massively parallel software running on tens, hundreds, or even thousands of servers. For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management alternatives. Big data usually includes data sets with sizes beyond the ability of commonly-used software tools to capture, curate, manage, and process the data within a tolerable elapsed time. Big data sizes are a constantly moving target, from a few dozen terabytes to many petabytes of data in a single data set. With this difficulty, a new platform of big data tools has arisen to handle sense making over large quantities of data, as in the Apache Hadoop Big Data Platform.

II ABOUT BIG DATA

The rapid development of Internet and mobile technologies has an important role in the growth of data creation and storage. Since the amount of data is growing exponentially, improved analysis of large data sets is required to extract information that best matches user interests. Big data refers to the use of large data sets to handle the collection or reporting of data that serves businesses or other recipients in decision making. The data may be enterprise specific or general and private or public. Big data are characterized by 3 V's: Volume, Velocity, and Variety[2].

Volume -the size of data now is larger than terabytes and peta bytes. The large scale and rise of
size makes it difficult to store and analyse using traditional tools.

Velocity – big data should be used to mine large amount of data within a pre defined period of
time. The traditional methods of mining may take huge time to mine such a volume of data.
Variety – Big data comes from a variety of sources which includes both structured and unstructured data. Traditional database systems were designed to address smaller volumes of structured and consistent data whereas Big Data is geospatial data, 3D data, audio and video, and
unstructured text, including log files and social media. New technologies are required to store unstructured large data sets and processing methods such as Hadoop and Map Reduce have greater importance in big data analysis. To process large volumes of data from different sources quickly, Hadoop is used. Hadoop is a free, Java-based programming framework that supports the processing of large data sets in a distributed computing environment. It allows running applications on systems with thousands of nodes with thousands of terabytes of data. Its distributed file system supports fast data transfer rates among nodes and allows the system to continue operating uninterrupted at times of node failure. It runs Map Reduce for distributed data processing and is works with structured and unstructured data.

Hadoop
Hadoop is a scalable, open source, fault tolerant Virtual Grid operating system architecture for data storage and processing. It runs on commodity hardware, it uses HDFS which is fault-tolerant high bandwidth clustered storage architecture[3][4]. It runs MapReduce for distributed data processing and is works with structured and unstructured data. For handling the velocity and heterogeneity of data, tools like Hive, Pig and Mahout are used which are parts of Hadoop and HDFS framework. Hadoop and HDFS by Apache is widely used for storing and managing big data. Hadoop consists of distributed file system, data storage and analytics platforms and a layer that handles parallel computation, rate of flow and configuration administration. HDFS runs across the nodes in a Hadoop cluster and together connects the file systems on many input and output data nodes to make them into one big file system. The present Hadoop ecosystem, consists of the Hadoop kernel, Map Reduce, the Hadoop distributed file system (HDFS) and a number of related components such as Apache Hive, HBase, Oozie, Pig and Zookeeper and these components are explained as below:

- HDFS: A highly faults tolerant distributed file system that is responsible for storing data on the clusters.
- MapReduce: A powerful parallel programming technique for distributed processing of vast amount of dataon clusters.
- HBase: A column oriented distributed NoSQL database for random read/write access.
- Pig: A high level data programming language for analyzing data of Hadoop computation.
- Hive: A data warehousing application that provides a SQL like access and relational
- model.
- Sqoop: A project for transferring/importing data between relational databases and Hadoop.
- Oozie: An orchestration and workflow management for dependent Hadoop jobs.

MapReduce

MapReduce is a programming model for processing large data sets with a parallel, distributed algorithm on a cluster. Hadoop MapReduce is a programming model and software framework for
writing applications that rapidly process vast amounts of data in parallel on large clusters of compute nodes. The MapReduce consists of two functions, map() and reduce(). Mapper performs the tasks of filtering and sorting and reducer performs the tasks of summarizing the result. There may be multiple reducers to parallelize the aggregations. Users can implement their own processing logic by specifying a customized map() and reduce() function. The map() function takes an input key/value pair and produces a list of intermediate key/value pairs. The MapReduce runtime system groups together all intermediate pairs based on the intermediate keys and passes them to reduce() function for producing the final results. Map Reduce is widely used for the Analysis of big data.

III GOALS OF BIG DATA

Big data helps to achieve various goals, which are following:

Cost Reduction

Hadoop is a framework for storing huge amount of data on distributed clusters. In Hadoop cluster, one year storage cost for one terabyte is Rs.120000. That is 800 times less than the traditional relational databases.

Time Reduction
Macy's merchandise pricing optimization application calculates data sets in seconds or in minutes which actually can take hours for calculation.

Support in Internal Business Decisions
The main idea of big data is to assist in the interior company decisions like, what kind of new products should be offered to people? , How much stock should be detained? And what must be the cost of our item?

Developing New Big Data-Based Offerings

Big data must be used to create new products and offerings. LinkedIn is the top example, which has used big data to develop products and offerings, including jobs you may be interested in, who have viewed my profile, people you may know, and numerous others. These ideas have pulled people to LinkedIn.

IV BIG DATA FOR BUSINESS

Organizations are grappling with what big data is and how it effects their organizations and how it makes benefits to their organizations. A survey is conducted in which found that the only 12 percent organizations are implementing or executing the big data strategy and 71 percent organizations are going to begin the planning stage[5]. It is clear that organizations need good knowledge of customers, goods and rules, with the help of big data organizations can find new ways to compete with other organizations. The organizations of the world are using the big data for their future decisions. Types of decisions that organizations can make from big data are smarter decisions, future decisions and decisions that make the difference. Organizations are making business decisions on the basis of the transactional data in past and in present but there is another kind of data which are non-traditional, less structured data for example weblogs, social media, Email and photographs that can be used for effective business decisions making. Oracle offers the products to acquire and organize these data types and analyze them to find new insights. Oracle's big data solution have 4 steps which are acquire big data, organize big data, analyze big data and decide on the basis of these analyses. Three models are also described for extracting value from big data. First model is ETL Extract, Transform, and Load. Second model is Interactive Queries. Third model is Predictive Analytics. Intel is taking advantage from big data and it has helped to speed up the innovation process. Organizations which built around big data from start are Google, eBay, LinkedIn, and Facebook. These organizations did not need to integrate big data with their existing sources of data. The author proposes that once if the organizations using big data analytics the organizations can get better business opportunities.

V DATA MINING AND BIG DATA

Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information - information that can be used to increase revenue, cuts costs, or both[6]. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational database. Data mining as a term used for the specific classes of five activities or tasks as follows: 1. Classification 2. Estimation 3. Prediction 4. Association rules 5. Clustering

- Classification is a process of generalizing the data according to different instances. Several major kinds of classification algorithms in data mining are Decision tree, k-nearest neighbor classifier, Naive Bayes, Apriori and AdaBoost. Classification consists of examining the features of a newly presented object and assigning to it a predefined class. The classification task is characterized by the well-defined classes, and a training set consisting of reclassified examples.
- Estimation deals with continuously valued outcomes. Given some input data, we use estimation to come up with a value for some unknown continuous variables such as income, height or credit card balance.
- Prediction It‟s a statement about the way things will happen in the future , often but not always based on experience or knowledge. P*rediction* may be a statement in which some outcome is expected.
- Association Rules An association rule is a rule which implies certain association relationships among a set of objects in a database.
- Clustering can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data.

VI BIG DATA IN CLOUD COMPUTING

Cloud Computing as an important application environment for big data has attracted tremendous attentions from the research community. Remarkable progress of big data networking has also been reported in this cloud computing area. Resource management plays a fundamental role in big data applications in the cloud[7]. A general introduction to resource management and allocation in multi-cluster clouds were introduced in virtualization planning and cloud computing methods in IBM data center networking. Key operational challenges such as support cost-saving technologies, rapid deployment, support for mobile and pervasive access, development of enterprise-grade network design has been discussed extensively. Despite existing efforts taking care of these challenges, an open question remains for making these objectives possible in a real-time and scalable fashion. Representative problems such as large datasets versus limited computational resources, data complexity versus limited knowledge, varying data structures/formats versus the need to integrate different tools. Key idea in this work is to make better use of computational and storage resources with the help of componentized software and cross-layer communications, which is as expected[8].

VII BIG DATA AND MOBILE COMPUTING

Mobile networking is becoming a more and more important counterpart of traditional Internet and big data. The mobile networking is becoming larger and larger due to releasing of hundreds of thousands of cell phones and pads. Moreover, the evolution of cellular network has enables mobile devices to be connected fast and reliably. A number of big data efforts have also been reported regarding mobile networking. One interesting improvement for the work in big data is to study daily behaviors of users based on usage of mobile maps on their cellphones or GPS, for which Apple Map and Google Map on cellphones are two important representatives. Mobile networking is by fact an important counterpart of traditional Internet[9]. More importantly, benchmarks and case studies have reflected usefulness of studying mobile big data. Moreover, considering the fast and reliable requirement of mobile networking requirements, effective interactions of the cloud and end users might be another interesting research direction.

VIII CHALLENGES OF BIG DATA

All Big Data projects begin with the selection and acquisition of data.

Storing the data for processing

Once the data arrives, it must be processed into a format that can be read by the analysis tools. Many collections are stored in proprietary or discipline-specific formats, requiring preparation and data reformatting stages. Applying Big Data analytics to the fuel of development faces several challenges. Some relate to the data—including its acquisition and sharing, and the overarching concern over privacy. Others pertain to its analysis[10][11].

Privacy is the most sensitive issue, with conceptual, legal, and technological implications. In its narrow sense, privacy is defined by the International Telecommunications Union as the right of individuals to control or influence what information related to them may be disclosed. Privacy can also be understood in a broader sense as encompassing that of companies wishing to protect their competitiveness and consumers and states eager to preserve their sovereignty and citizens. In both these interpretations, privacy is an overarching concern that has a wide range of implications for anyone wishing to explore the use of Big Data for development—vis-à-vis data acquisition, storage, retention, use and presentation.Privacy is a fundamental human right that has both intrinsic and instrumental values[12]. Focusing on individual privacy, it is likely that, in many cases, the primary producers—i.e. the users of services and devices generating data—are unaware that they are doing so, and/or what it can be used for. For example, people routinely consent to the collection and use of web-generated data by simply ticking a box without fully realising how their data might be used or misused. It is also unclear whether bloggers and Twitter users, for instance, actually consent to their data being analysed[13][14]. In addition, recent research showing that it was possible to 'de-anonymise' previously anonymised datasets raises concerns. The wealth of individual-level information that Google, Facebook, and a few mobile phone and credit card companies would jointly hold if they ever were to pool their information is in itself concerning[15]. Because privacy is a pillar of democracy, we must remain alert to the possibility that it might be compromised by the rise of new technologies, and put in place all necessary safeguards.

Access and Sharing

Although much of the publicly available online data has potential value for development, there is a great deal more valuable data that is closely held by corporations and is not accessible for the purposes described in this paper. One challenge is the reluctance of private companies and other institutions to share data about their clients and users, as well as about their own operations. Obstacles may include legal or reputational considerations, a need to protect their competitiveness, a culture of secrecy, and, more broadly, the absence of the right incentive and information structures. There are also institutional and technical challenges—when data is stored in places and ways that make it difficult to be accessed, transferred, etc.

Properly analysed, Big Data offers the opportunity for an improved understanding of human behaviour that can support the field of global development in three main ways:
1) **Early warning:** early detection of anomalies in how populations use digital devices and services can enable faster response in times of crisis;
2) **Real-time awareness:** Big Data can paint a fine-grained and current representation of reality which can inform the design and targeting of programs and policies;
3) Real-time feedback: the ability to monitor a population in real time makes it possible to understand where policies and programs are failing and make the necessary adjustments.

IX CONCLUSION

In this paper we have seen that what is big data, Hadoop, Map Reduce and how the big data can be used in various areas like business, data mining, cloud computing and mobile computing.Big Data constitutes an historic opportunity to advance our common ability to support and protect human communities by understanding the information they increasingly produce in digital forms. We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and

improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. The author strongly believes that fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

REFERENCES

[1] Anchalia, P.P.; Koundinya, A.K.; Srinath, N.K., "MapReduce Design of K-Means Clustering Algorithm," International Conference onInformation Science and Applications (ICISA), pp.1,5, 24-26 June 2013, doi:10.1109/ICISA.2013.6579448.

[2] "Big Data, Big Impact: New Possibilities for International Development." World Economic Forum (2012): 1-9. Vital Wave Consulting. Jan. 2012

[3] Big data: The next frontier for innovation, competition, and productivity. James Manyika, Michael Chui, Brad Brown, Jacques Bughin, Richard Dobbs, Charles Roxburgh, and Angela Hung Byers. McKinsey Global Institute. May 2011.

[4] C.Lam, "Hadoop in Action", Manning Publications Co., USA,ISBN:9781935182191, Dec. 2010.

[5] King, Gary. "Ensuring the Data-Rich Future of Social Science." Science Mag 331 (2011) 719-721. 11 Feb, 2011 Web.

[6] Keim, Daniel, Huamin Qu, and Kwan-Liu Ma. "Big-Data Visualization." Computer Graphics and Applications, IEEE 33.4 (2013): 20-21.

[7] Lakew, Ewnetu Bayuh. Managing Resource Usage and Allocations in Multi-Cluster Clouds. 2013, http://www8.cs.umu.se/~ewnetu/papers/lic.pdf

[8] Monga, Inder, Eric Pouyoul, and Chin Guok. Software-Defined Networking for Big-Data Science-Architectural Models from Campus to the WAN. High Performance Computing, Networking, Storage and Analysis (SCC), 2012 SC Companion:. IEEE, 2012.

[9] Russom, " Big Data Analytics", TDWI Research, 2011.

[10] Richa Gupta, Sunny Gupta, Anuradha Singhal, (2014), "Big Data:Overview", IJCTT, 9 (5). [11]S.Perera, T.Gunarathne, "Hadoop MapReduce Cookbook", Packt Publishing, ISBN:1849517282,Jan. 2013.

[11] Turn Big Data into Big Value, A Practical Strategy, Intel White Paper,2013.

[12] T. H. Davenport and J. Dyche, "Big Data in Big Companies," May 2013, 2013.

[13] Wei Fan and Albert Bifet " Mining Big Data:Current Status and Forecast to the Future",Vol 14,Issue 2,2013

[14] Wu, Xindong, et al. "Data mining with big data." Knowledge and Data Engineering, IEEE Transactions on 26.1 (2014): 97-107.

Authors Profile

Dr.M.Kumarasamy working as Professor in Department of Computer Science, Villa College, Male, Maldives. He is having more than 25+ years of experience in Teaching and Industries. His areas of interest is in real time applications, artificial super intelligence, Machine learning, Big Data and also Optimization in port sector and power sector.

Dr.G.N.K.Suresh Babu working as Associate Professor in Department of Computer Applications, Acharya Institute of Technology, Bangalore, India. He is having more than 25 years of experience in teaching and industries. His areas of interest in Data mining, Cloud computing, Big data and Network Securities.