Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 3, Issue. 4, April 2014, pg.446 – 453

REVIEW ARTICLE

A Review on the Role of Big Data in Business

Jafar Raza Alam[1], Asma Sajid[2], Ramzan Talib[3], Muneeb Niaz[4]

[1,2,3,4] College of Computer Science and Information Studies, Government College University, Faisalabad, Pakistan
[1] jafar.raza35@yahoo.com; [2] asmasajid936@gmail.com, [3] ramzan.talib@gcuf.edu.pk, [4] muneb.niaz@gmail.com

Abstract— Big data is a game changing thing. Successful organizations are achieving business advantages by analyzing big data. It has received significant attention in recent years but some challenges are one of the major causes in diminishing the growth of organizations. The main issues why these organizations are not begin their planning stage to implement the big data strategy because they don't know enough about the big data and they don't understand the benefits of big data. In this study, an attempt is made to review the role of big data in the business.

Keywords— Big data; Cloud computing; Data Mining

I. INTRODUCTION

The aim of big data is to provide better usage of resources and storage, reduce the time of computation and good business decision making. The term "Big data" indicates data, but in huge or enormous form, which cannot be processed by the conventional database systems.

Its basic characteristics are 3v's which are volume, velocity and variety. Here volume means data in large quantities. Velocity is just like social media data streams in which data is increasing on social networks for example posts on facebook and tweets on twitter. Last thing variety means data in many formats like structured, unstructured and semi structured. Every day world create 2.5 quintillion bytes of data; 90% of the data in the world today has been created in the last two years alone [1]. Data is growing at exponential rate and the experts of the data analytics technology do not have enough knowledge to analyze that enormous amount of data. Big data presents three main aspects to any organization for interest, first one is lacking of arrangement. Second it produces new opportunities. Third, technologies used for big data having low cost.

This review is an attempt to investigate the role of big data in business. The main aim of this research is to discuss the impact of using data mining techniques and frameworks of big data as the optimal strategy in achieving accuracy and timing in fast and reliable business decision making. Key challenges also discussed in this study, which big data are facing now.

II. MATERIALS AND METHODS

This paper is based on the randomly selected articles in the field of big data. Twenty five articles were positively identified for the study and read completely for the review. Calculations included in these articles were carefully rerun to guarantee information accuracy. In this review a question is posed

Question: What is the role of big data in business?

The required data is extracted from the papers to answer the question posed above.

III. 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 [2]. 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 [3]. 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[1]. 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 [4].

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 [5]. A process has been described (Figure 1) for the organizations that are interested in adopting Big Data [6].

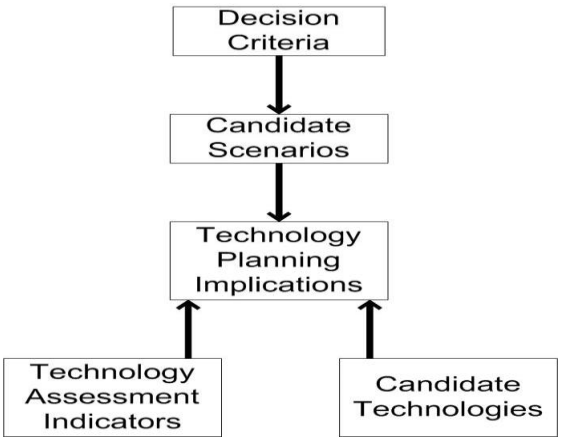

Figure 1: Research Framework of Technology Strategy Planning [6]

Steps of this process are following

- Decision criteria are dependent on the decision factors which are social factors, technological factors and economical factors.
- Candidate Scenarios, There are different scenarios which organizations can select for big data e.g. big demand and cautiously optimistic.
- Candidate Technologies are data warehouse, cloud analytics, embedded analytics and big data visualization.
- Technology Assessment Indicators are global market size, enterprise adoption ratio, entrance barrier and strength of industry.
- Technology Planning Implications, two types of implications are here which is technology planning implications for scenario big demand and technology planning implications for scenario cautiously optimistic.

By adopting this strategy organizations can get the fruits of big data.

IV. BIG DATA GOALS

Big data helps to achieve various goals [5], which are following.

4.1 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 $2,000. That is 800 times less than the traditional relational databases.

4.2 Time Reduction

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

4.3 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?

4.4 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.

V. DATA MINING WITH BIG DATA

In production environment big data mining process does not end. A good big data analytics platform has factors like speed of development, robustness, easily analyze huge amount of data. Growth of data in terms of size and number of users has increased in past few years. In 2010 there were only 4 data analysts working at twitter but now there are thousands of employees working at hadoop cluster node data centers of twitter. Twitter increased the analytics power before the exact time and if an organization does not do this before it needed then these issues will become nightmares [7]. Extracting information from the stream data at real time is the good way to come to know what is happening at the spot. Stream data arrive at very high speed and it is very difficult to analyze stream data at real time, stream data requires very efficient algorithms for mining , that algorithm should be accurate. [8].Online

news, social media and micro blogs are the examples of streams created by the users. Solutions to deal with these streams were not designed. Samoa was a platform for mining these streams. This is a tool for online mining in the cloud environment. Samoa can be run on different distributed stream processing engines like storm. In future Samoa will be open source and that will be evolution in the research area of the big data stream mining. Samoa also provides API for algorithm developers [9]. A data driven model named Hace is also proposed which aggregate the multiple sources of information; analyze data from the data mining perspective [10].

Big data helping Intel for improving their business intelligence; large portion of their data was unstructured which was 90 percent of their enterprise data. Aims of the Intel were make better decisions, increase business velocity, discover and tap new markets. Intel wants to increase their Business intelligence strength from the descriptive analytics to predictive analytics by using data mining from big data [11].

VI. BIG DATA APPLICATIONS

Big data applications are being used in different industries. Fed BP Disaster response was an application for the government's [USA government] response in the disaster situation. This was built in 2010 when oil rate flow was a key issue. BP and independent groups presented changeable estimates preventing efforts to manage the level and range of the U.S. Government's response. NIST analyzed the estimates and formed actionable intelligence on which to support the final reaction [12].

Applications in oil and gas business could be Equipment maintenance to prevent failure, production optimization, price optimization, safety and compliance. Oil and gas companies have big competition in their field, and facing regular change. Firms need to increase their production volumes and at the other hand they also want the healthy, safety and low risk environment. From exploration and production of the oil, leading companies are using big data for new business values, reduce cost and increase production [13].

Website of Whitehouse consists of all speeches of Barrack Obama. Aim was to find the influence of these speeches on the election. A scrapper was made to collect all speeches from the website. Hadoop and Map reduce both are used for parallel processing of these speeches. Study finds the most spoken or referred words, most referenced countries, personalities and also focus on internal and foreign affairs. According to [14], these all things are the basis for the influence on the elections.

VII. BIG DATA FRAMEWORKS

Organizations daily process large amount of data, this generates high network traffic. Researchers are trying to design a data analytics system which supports complex analysis from high network traffic. CLAaaS stands for Cloud-based Analytics-as-a-Service. CLAaaS was a conceptual architecture for the big data analytics in the cloud environment. It has features, which are customization, collaboration and assistance. Data privacy can be created by implementing CLAaaS in a private cloud [15]. Camcube is a cluster design and it used a topology to connect servers directly with one another. Camdoop is used to increase the capability like processing of packets in networks to perform aggregation of data. In the network Output size is small as compare to the input size. To overcomes this issue we adopted a new technique by decreasing traffic instead of increasing bandwidth. Camdoop have property that camcube uses to forward traffic to perform in network aggregation of data [16].

Bigtable is used to store structured data having size in petabytes. In [17] a data model of Bigtable has described. Bigtable store data of Google applications. Web indexing, Google Finance and Google earth are applications of the Google. These applications have different requirements for storage. The storage, collection and use of data can also create new vulnerabilities and risks. After analyzing these risks a framework has been proposed to help the effective use of data. In this framework few domains are considered which are ethics, governance, science and technology. By using these all domains together organizations can be more effective while making their decisions and avoid the failures of future projects [18].

The method which is used for the fast and accurate analysis on the large data sets are sampling, sampling is the data set which almost represents the all data set. In this article [19] an Earl framework was proposed which is an accuracy inference method and increasing big data analytics. It works by choosing the appropriate sample size. EARL is used for mining algorithms to calculate their outcome and errors. Accuracy of earl system will be same by increasing the samples sizes.

VIII. BIG DATA CHALLENGES

8.1 Security and Privacy

Cloud security alliance big data working group identify top security and privacy problems that need to capture for making the big data computing and infrastructure more secure. Most of these issues are related to the big data storage and computation. Some of the challenges are secure data storage [20]. Various security challenges related to data security and privacy are discussed in [21] which include data breaches, data integrity, data availability and data backup.

8.2 Dynamic Provisioning

A service of the cloud computing is infrastructure as service in which it provides computation resources on demand, many cloud related companies are implementing this concept and to making it easy for customers to access these services. Current frameworks do not have the property of the dynamic provisioning. Here is an issue that Compute resources can be insufficient for the submitted job, some process may requires more resources. Another issue is scheduling and protection algorithm, current algorithms does not consider these aspects [22].

8.3 Algorithms

Organizations were granting the papers by capturing key words from the abstract and titles. Analyzing the science with hand was difficult task. After that work was done by program analyst. They use algorithms to do this work. These algorithms can be varying from each other. This difference can reduce the effectiveness and reliability of the final result. Improvement in the data management will result in better technology but it will face many issues. [23].

8.4 Misuse of Big Data

Challenges including potential misuse of big data are here, because information is power. Types of the data which people will produce in the future are unknown. To overcome these challenges we have to strengthen and increase our intent and capacity [24].

8.5 Data Management

Data management is also a critical issue for corporation and industries. Data warehouse has efficient data management techniques. In [25], two data warehouse management strategies are discussed; which are Immediate Incremental Management (IIM), Deferred Incremental Management (DIM) but the favor is given to IIM because of its algorithmic implementation.

IX. BIG DATA FACTS AND FIGURES

In an organization everything is dependent on the decisions of the policymakers and their decisions are dependent on the data mining techniques, data mining algorithms and frameworks of big data. By integrating data mining with big data frameworks we can get more accurate business decision making. To find true business worth from big data, organizations require the tools to analyze and arrange different data types from different sources. If, organizations have power to analyze data of any size, any type

and from many sources then outcome is in form of deeper and reliable information about business trends, values and patterns. Data types which are used for business decisions [2] are shown in figure 2.

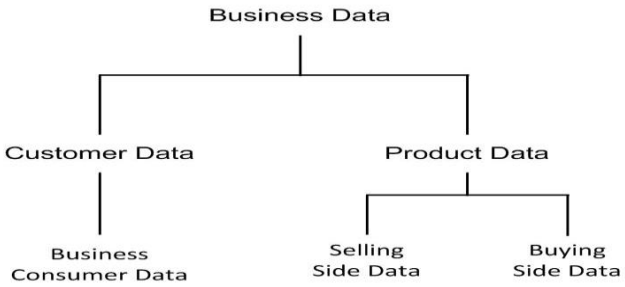

Figure 2: Business Data Types for Decision support

Figure 3 indicates fraction of business data which is used for business decisions. An organization's, 76 percent of customer data and 70 percent of product data are required for decision support. In the customer data 66 percent of business consumer data are used to make decisions and from the product data 62 percent of selling side and 61 percent of buying side data are used to make decisions [2].

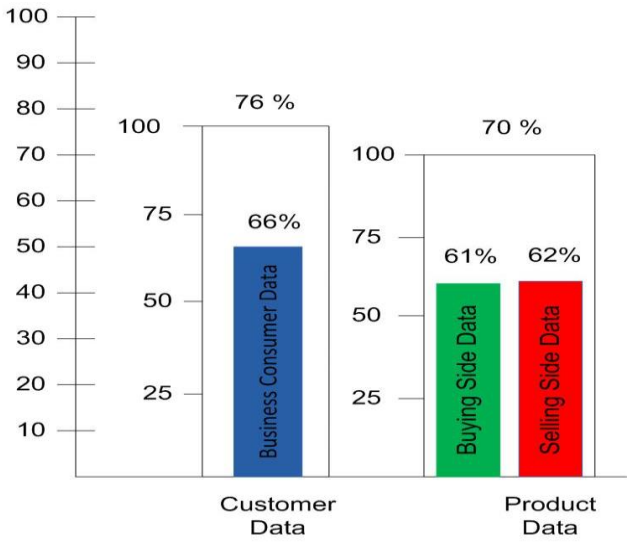

Figure 3: Percentage of Business Decision Support Data

The 63 percent organization reports that the use of big data is beneficial for their companies and organizations [3]. Analytic techniques go further than data mining to determine not only what has happened , but to predict what is going to happen based on all data.

X. CONCLUSION

Big data must be integrated in the organization's architecture, even the organization have their well established and large businesses. Countries in the world, IT companies and the relevant departments have started working on big data. Organizations which built

around big data are Google, eBay, LinkedIn, and Facebook. Large organizations are joining the data economy and combining the big data analytics with traditional analytics. This will effect on the organization's skills, leadership, structures and technologies. The 63 percent organization reports that the use of big data is beneficial for their companies and organizations. Organization's more than 70 percent of customer and product data are used for the business decisions making. Key challenges which appear are designing big data sampling and building prediction models from the big data streams. Challenges including potential misuse of big data are also here, because information is power. Types of the data which people will produce in the future are unknown.

REFERENCES

[1] J. P. Dijcks, "Oracle: Big data for the enterprise," *Oracle White Paper*, 2012.

[2] "Big Data Survey Research Brief," *Sas White Paper*,2013.

[3] M. Schroeck, R. Shockley, J. Smart, D. Romero-Morales, and P. Tufano, "Analytics: the real-world use of big data: How innovative enterprises extract value from uncertain data, Executive Report," *IBM Institute for Business Value and Said Business School at the University of Oxford*, 2012.

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

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

[6] W.-H. Weng and W.-T. Lin, "A Scenario Analysis Of Big Data Technology Portfolio Planning," in *International Journal of Engineering Research and Technology*, 2013.

[7] J. Lin and D. Ryaboy, "Scaling big data mining infrastructure: the twitter experience," *ACM SIGKDD Explorations Newsletter*, vol. 14, pp. 6-19, 2013.

[8] A. Bifet, "Mining Big Data in Real Time," *Informatica (03505596)*, vol. 37, 2013.

[9] G. De Francisci Morales, "SAMOA: A platform for mining big data streams," in *Proceedings of the 22nd international conference on World Wide Web companion*, 2013, pp. 777-778.

[10] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 26, pp. 97-107, 2014.

[11] "Mining big data in the enterprise for better business inteligence."*Intel White Paper*,2012.

[12] "The Compelling Economics and Technology of Big Data Computing, For Big Data Analytics There's No Such Thing as Too Big." *4syth White Paper*,2012.

[13] A. Hems, A. Soofi, and E. Perez, "How innovative oil and gas companies are using big data to outmaneuver the competition," 2013.

[14] A. Bensrhir, "Big data for geo-political analysis: Application on Barack Obama's remarks and speeches," in *Computer Systems and Applications (AICCSA), 2013 ACS International Conference on*, 2013, pp. 1-4.

[15] F. Zulkernine, P. Martin, Y. Zou, M. Bauer, F. Gwadry-Sridhar, and A. Aboulnaga, "Towards Cloud-Based Analytics-as-a-Service (CLAaaS) for Big Data Analytics in the Cloud," in *Big Data (BigData Congress), 2013 IEEE International Congress on*, 2013, pp. 62-69.

[16] P. Costa, A. Donnelly, A. Rowstron, and G. O'Shea, "Camdoop: Exploiting in-network aggregation for big data applications," in *USENIX NSDI*, 2012.

[17] F. Chang, J. Dean, S. Ghemawat, W. C. Hsieh, D. A. Wallach, M. Burrows, *et al.*, "Bigtable: A distributed storage system for structured data," *ACM Transactions on Computer Systems (TOCS)*, vol. 26, p. 4, 2008.

[18] K. Crawford, G. Faleiros, A. Luers, P. Meier, C. Perlich, and J. Thorp, "Big Data, Communities and Ethical Resilience, A Framework for Action," *Big Data, Communities and Ethical Resilience White paper,2013.*

[19] N. Laptev, K. Zeng, and C. Zaniolo, "Very fast estimation for result and accuracy of big data analytics: The EARL system," in *Data Engineering (ICDE), 2013 IEEE 29th International Conference on*, 2013, pp. 1296-1299.

[20] "Top ten big data security and privacy challenges," *Cloud Security Alliance White paper,2012.*

[21] A. A. Soofi, M. I. Khan, R. Talib, and U. Sarwar, "Security Issues in SaaS Delivery Model of Cloud Computing," 2014.

[22] M. N. Vijayaraj, M. D. Rajalakshmi, and M. C. Sanoj, "Issues and challenges of scheduling and protection algorithms for proficient parallel data processing in cloud."

[23] G. Halevi and H. Moed, "The evolution of big data as a research and scientific topic: overview of the literature," *Research Trends, Special Issue on Big Data*, vol. 30, pp. 3-6, 2012.

[24] U. G. Pulse, "Big Data for development: challenges & opportunities," *Naciones Unidas, Nueva York, mayo,* 2012.

[25] U. Rasheed, M. U. Sarwar, and R. Talib, "A Review on Data Warehouse Management."

