I J C T A, 10(8), 2017, pp. 249-255 ISSN: 0974-5572

Big Data Challenges: A Software Engineering Perspective

Malathi S.[1],* and Kannan Balakrishnan[2]

ABSTRACT

The term "Big Data" solidified around year 2008, even though big bulk data management was identified as a need of the hour prior to this. Big Data is dynamically growing data, which cannot be managed using the normal techniques for capture, storage, analysis and retrieval. Real-time processing techniques are required to manage Big Data. In this paper, we make a study of the literature regarding the recent challenges and opportunities faced in Big Data processing and Big Data Projects, starting from the challenge of defining Big Data and propose a few novel areas of research in Big Data relating to Software Engineering. The Software Engineering perspective of Big Data still remains vastly uncovered and hence our study focuses on such aspects which are important from a Software Engineering viewpoint, like process improvement, Requirements, Design, and Cost Estimation

Keywords: Big Data, Challenges, Software Engineering, Variety, Velocity, Volume

1. INTRODUCTION

Big Data refers to extremely large data sets which needs special analysis techniques to reveal patterns, relationships or other useful information from them. The challenges in big data processing include capture, *analysis, data curation, storage, search, querying, updating, sharing, visualization, transfer and information privacy[1]*. This is with respect to the technical, implementation view point. From a broader perspective, we need to look into the management aspects of Big Data Projects as such and the architecture, processes, methods and standards that go into making it. Huge volumes of data and the dynamic growth in data pose a big challenge with individuals and teams working in different locations and across cultures. Big Data Analytics is a promising area of research now and rapid progress is being made regarding automation and optimization of already available techniques.

2. DEFINITIONS OF BIG DATA

Definition of the term Big Data itself pose a challenge to the researchers. Many scientists and professional groups have defined it with progressive refinement, considering different aspects. The definition given by Doug Laney at Gartner in 2012 was based on the three V's, Volume, Velocity and Variety of the data [2].Volume refers to exabytes of data; the large bulk of it. Velocity is the high rate of data and information flow in the system. Variety means the heterogeneity of the data; different types of it, different formats, different platforms, different languages. Authors of [3] remarks that since the origin of the definition of Big Data can be attributed to the academia, industry and media there is no single unified definition for the term. It is defined by different stakeholders according to their environment of perception and when compared are contradictory at times.

Chula Engineering [4] defines Big Data as a term applied to data sets whose size is larger than that could be managed by the ability of commonly used software tools to capture, manage, and process the data.

[1,2] Department of Computer Applications, Cochin University of Science and Technology, Kerala, India, *Emails: malathisujith@gmail.com, mullayilkannan@gmail.com*

Emphasis is given to the data considered within a tolerable elapsed time. They also mention the growth in Big Data size and dynamic growth as contributing factors in the definition.

3. CHALLENGES IN THE IMPLEMENTATION SCENARIO

In [5], Dr. Kirk Borne refers to 10 big data 'V' challenges. They are: Volume, Velocity, Variety, Veracity, Validity, Value, Vagueness, Variability, Vocabulary and Venue. Veracity is the conformity to facts, data to test different hypothesis. Validity means the data quality, Value is the business value to the organization, Variability is the dynamic, evolving nature of Big Data, Venue means the source of data – it can be from different owners, from different platforms. Vocabulary indicates any representations of the data's structure. Vagueness is the confusion over the meaning of Big Data and the tools used. Ji, Changqing, et al. In paper [6] identifies the key issues in big data processing as big data service models, big data management platform, data storage, distributed file system, data virtualization platform and distributed applications. They also suggest optimization strategies for the Map Reduce framework. The work in [7] presents data challenges, Process Challenges, and Management Challenges. This covers Technology Challenges too. The data challenges include Volume, Variety, Velocity, Veracity, Data Availability, Data Quality, Data Discovery, Data Comprehensiveness, Quality and Relevance, Personally Identifiable Information, Data Dogmatism and Scalability. Process challenges include capturing data, aligning data from different sources, transforming the data to suit analysis, modeling the data and understanding the output. Management Challenges include data privacy, security, governance and ethics.

The authors in [8] identifies the various issues and challenges in Big data projects as issues of scale, heterogeneity, lack of structure, privacy, timeliness, error-handling, provenance and visualization. In [9], the three V's - Volume, Velocity and Variety are identified as key terms in the Big Data challenge. In this paper, the obstacles in Big Data implementation are considered to be lack of expertise in Big Data management, choice of an appropriate big data platform, the cost implications of Big Data, data democratization, encryption and securing Big Data, speed of analysis, vulnerability and security of big data, protection of data storage location and providing real time security. Open problems for research mentioned are visual data discovery tools, cloud based data analytics, shortage of skilled staff-how to overcome, providing a unified data architecture advancements in predictive analysis, value-addition of data, consistency in decision making and the decision making process, adoption of new technology, cognitive computing for access to services.

The authors of [10] suggest that the major challenge is in data cleansing, acquisition and capture, scalability issues, storage, sharing and transfer of data, analysis of results andcollection of results. Ethical considerations like issues of identity, privacy, ownership and reputation are also considered significant. Authors of [11] give a social and behavioral science viewpoint of Big Data. They identify the concept of network modelling as a key and the significance of relations thus represented. The complexity arising out of the relationships and the complexity measures are also mentioned as factors to be considered. Storage, data access, sharing and integration are considered the big challenges in Big Data by [12].

Many of the identified challenges are already being worked upon and techniques are evolving to solve them. But there are still new concerns coming up like the problems of data processing, data storage, data representation, visualizing, tracking data and how data can be used for pattern mining and analyzing user behaviors [13]. These are only quite few of the upcoming challenges.

4. THE SOFTWARE ENGINEERING PERSPECTIVE

In the work in [14], a context model of Big Data Software Engineering is developed. This model contains various elements and the relationships among these elements. The involvement of stakeholders is also included in this context model. A few challenges mentioned are requirements, architectures, testing and

maintenance. Scalability is also considered a factor of significant challenge. The author of [15] presents his findings based on the EPIC project. TheEPIC project is a $4M NS grant that goes into exploring how members of the public use social media during times of mass emergency [16]. Significance of multidisciplinary teams, use of highly iterative life cycles, lack of developer support tools,difficulty in understanding of frameworks and technologies used scalability issues and importance of modeling in producing scalable, robust, efficient systems are identified as the major challenges in Big Data processing, based on the findings from the EPIC project. Collection of ever growing data and its analysis are also taken care of in this project. Performance and reliability are two other concerns mentioned by the author.

In [17], software related data is collected about so many factors like customers, execution behavior, team work practices etc. The analysis results can be used for decision making and improvement. The research areas identified in this position paper are productivity, correctness, communication and collaboration.

Architectural qualities need be balanced for obtaining consistency, availability and network partition. Scalability is a major issue. The distribution, write-heavy workloads, varying request loads andintense computation-based analytics contribute to this. Additionally, there are Software Engineering challenges like architecture, testing, planning, management of large teams and coordination [18].

In [19], Software Engineering Challenges of Big Data are identified as scalable software architectures for Big Data, techniques for quality assurance of data-intensive software, monitoring and ensuring the qualityduring operation, developing new algorithms for real time storage and clusteringof data objects, Big Data engineering techniques and frameworks, using Big Data analytics to solve current Software Engineering problems etc.

The role of human and machine intelligence in Big Data Analytics has also been studied. A conceptual architecture is developed for framing requirements for the analysis, including the human and machine intelligence factors. Collaborative sense making is the focus. Sense making is an iterative cognitive process that the humans perform so as to build up a representation of an information space that is useful in accomplishing his/her goal [20]. As defined by [21], Sense making is seeking a representation which encodes data in it such that task-specific questions can be answered.

The major break in Cost and Effort estimation of Big Data has come with Barry Boehm and his team revising the Constructive Cost Model (COCOMO) II to suit the scenario of Big Data [22]. There are five scale drivers identified along with seventeen cost drivers.

In this paper, we propose the following factors also as being major Software Engineering challenges to Big Data projects:

a) Requirements collection, finalization and analysis are of utmost importance in big data projects because if a mistake is made in requirements, the cost would multiply over the phases and an enormous overhead thrust upon the organization.

b) Design needs care because constraints for Big Data Analysis projects are tough and the scope is limited most of the times due to limits imposed on the data.

c) Proper definition of Software Process models and methodologies for managing Big Data Analytics projects are not proposed yet.

d) The Big Data era marks the presence of process improvement scope. Optimization of processes for big data handling is a growing area of research. With this, the Software Process also needs improvement.

e) The continuous improvement models of CMM (Capability Maturity Model) and its variants and Six Sigma, Lean Sigma etc. needs refinement to suit Big Data. The process itself is liable to be submitted to continuous improvement.

f) Cost and Effort estimation is a crucial area in Big Data Projects. As being on time and quality are important, sticking to the budget is also important. Hence cost and effort estimation has to evolve to find more appropriate techniques to make precise estimations to incorporate the features of Big Data projects.

g) New variants of verification and validation techniques have to be developed, considering the characteristics of Big Data. A different set of metrics suitable for measuring Big Data Analytics Project operation and performance has to be deployed. In the literature Survey, we found that most of the Big Data Projects are Analysis Projects.

Table 1.1 below is a consolidation of the major contributions in the Big Data Challenge proposed by different researchers.

Table 1
Big Data challenges as proposed by different researchers

Sl. No.	*Authors*	*Year*	*Identified Challenges*
1	Ji, Changqing, Yu Li, Wenming Qiu,Yingwei Jin, Yujie Xu, UchechukwuAwada, Keqiu Li, and Wenyu Qu [6]	2012	data storage, Big Datamanagement platform, distributed file system,big data service models,data virtualization platform, distributed applications
2	Agrawal, Divyakant, Philip Bernstein, Elisa Bertino, Susan Davidson, UmeshwarDayal, Michael Franklin, Johannes Gehrke et al [8]	2012	scale, lack of structure, heterogeneity, time liness, error-handling, provenance, privacy, visualization
3	Snijders, Chris, Uwe Matzat, and Ulf-Dietrich Reips [11]	2012	The challenge is identified form a social and behavioral science viewpoint of Big data and the network view and complexitymeasures proposed as a different viewpoint
4	Dr. Kirk Borne [5]	2014	Volume, Velocity, Variety, Veracity, Validity, Value, Vagueness, Variability, Vocabulary and Venue
5	Zicari, Roberto V [7]	2014	Data challenges - Volume, Variety, Velocity, Veracity, Data Availability, Data Quality, Data Discovery, Data Comprehensiveness, Quality and Relevance, Personally Identifiable Information, Data Dogmatism, Scalability. Process challenges - capturing data, aligning data from different sources, transforming the data to suit analysis, modeling the formatted data and understanding the output. Management Challenges - data privacy, security, governance, ethics.
6	Metzger, A [19]	2014	scalable software architectures for Big Data, techniques for quality assurance of data-intensive software, monitoring and ensuring the quality during operation, developing new algorithms for real time storage and clustering of data objects, Big Data engineering techniques and frameworks, using Big Data analytics to solve current Software Engineering problems etc.
7	Lau, Lydia, Fan Yang-Turner, and Nikos Karacapilidis [20]	2014	Architecture for framing requirements, human and machine intelligence, Sense- making
8	Madhavji, Nazim H., AndriyMiranskyy, and Kostas Kontogiannis [14]	2015	Requirements, architectures, testing and main-tenance, Scalability, involvement of stakeholders
9	Anderson, Kenneth M [15]	2015	Significance of multidisciplinary teams, use of highly iterative life cycles, lack of developer support tools, difficulty in understanding of

(contd...Table 1)

Sl. No.	*Authors*	*Year*	*Identified Challenges*
			frameworks and technologies used scalability issues and importance of modeling in producing scalable, robust, efficient systems
10	DeLine, Robert [17]	2015	Productivity, correctness, communication and collaboration
11	Mukherjee, Samiddha, and Ravi Shaw [9]	2016	Volume, Velocity, Variety along with research openings in Visual data discovery tools, Cloud based data analytics, shortage of skilled staff-how to overcome, providing a unified data architecture advancements in predictive analysis, Value-addition of data, consistency in decision making and the decision making process, adoption of new technology, cognitive computing for access to services.
12	Anagnostopoulos, I., S. Zeadally, and E. Exposito [10]	2016	data cleansing, acquisition and capture, scalability issues, storage, sharing and transfer of data, analysis and collection of results, ethical considerations like issues of identity, privacy, ownership and reputation
13	Mardis, Elaine R [12]	2016	Storage, data access and sharing and integration
14	Bello-Orgaz, Gema, Jason J. Jung, and David Camacho [13]	2016	data processing, data storage, data representation and how data can be used for pattern mining, analyzing user behaviors and visualizing and tracking data
15	Gorton, Ian, AyseBasarBener, and Audris Mockus [18]	2016	Scalability, architecture, testing, planning, management of large teams and coordination

In general, the major challenges in Big Data Projects can be summarized as in Fig. 1 below:

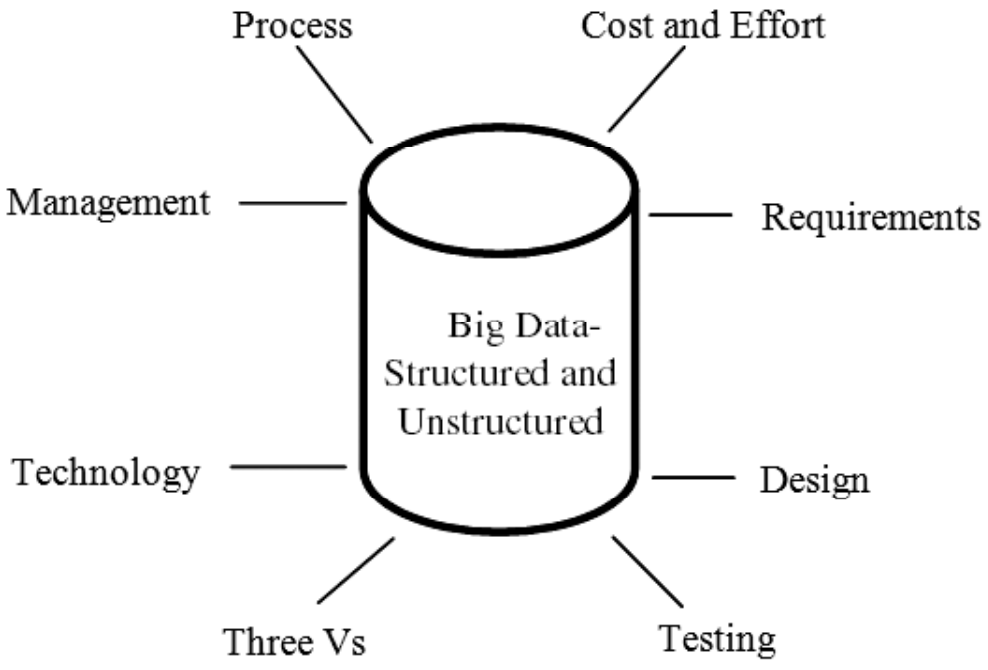

Figure 1: Major challenges in Big Data Projects

5. CONCLUSION

Big Data is special due to its bulk now reaching exabytes, growing dynamically and posing a challenge to how it can be handled efficiently and effectively. The three V's, Volume, Velocity and Variety need focus as basic areas of care whereas managing the other V's like veracity, value etc. provides more effectiveness in Big Data handling. Major work in the area is going on with storage, retrieval, analysis, transferring, visualization and interpretation of results of Big Data Analysis. Big Data Architecture and other representations and the complexity of data to be handled require special attention. From a Big Data project perspective, requirements management and designing the project for growing real time data is also very crucial. Other open areas of focus in research relating to Big Data and Software Engineering are Requirements Management, Design, Software Process Improvement, Cost and Effort estimation and Verification and Validation.

REFERENCES

[1] Big data [Internet]. En.wikipedia.org. 2016 [cited 23 November 2016]. Available from: https://en.wikipedia.org/wiki/Big_data

[2] Douglas, Laney. "The Importance of'Big Data': A Definition." Gartner (June 2012) (2012).

[3] Ward, Jonathan Stuart, and Adam Barker. Undefined by data: a survey of big data definitions.arXiv preprint arXiv:1309.5821 (2013).

[4] [Internet]. 2016 [cited 23 November 2016]. Available from: https://www.cp.eng.chula.ac.th/~fyta/213/ReadingList/Wiki - Big Data.pdf

[5] Top 10 Big Data Challenges – A Serious Look at 10 Big Data V's | MapR [Internet]. Mapr.com. 2016 [cited 23 November 2016]. Available from: https://www.mapr.com/blog/top-10-big-data-challenges-serious-look-10-big-data-vs

[6] Ji, Changqing, Yu Li, WenmingQiu, YingweiJin, Yujie Xu, UchechukwuAwada, Keqiu Li, and Wenyu Qu. Big data processing: Big challenges and opportunities.Journal of Interconnection Networks 13, no. 03n04 pp. 1250009, (2012).

[7] Zicari, Roberto V. Big data: Challenges and opportunities.Big data computingpp. 564, (2014).

[8] Agrawal, Divyakant, Philip Bernstein, Elisa Bertino, Susan Davidson, UmeshwarDayal, Michael Franklin, Johannes Gehrke et al. Challenges and Opportunities with Big Data. A community white paper developed by leading researchers across the United States.Computing Research Association, Washington (2012).

[9] Mukherjee, Samiddha, and Ravi Shaw. Big Data–Concepts, Applications, Challenges and Future Scope.International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 2, February (2016).

[10] Anagnostopoulos, I., S. Zeadally, and E. Exposito. Handling big data: research challenges and future directions.The Journal of Supercomputing 72.4, pp. 1494-1516 (2016).

[11] Snijders, Chris, Uwe Matzat, and Ulf-Dietrich Reips. "Big Data": big gaps of knowledge in the field of internet science.International Journal of Internet Science 7.1, pp.1-5 (2012).

[12] Mardis, Elaine R. The challenges of big data.Disease Models and Mechanisms 9.5, pp. 483-485, (2016).

[13] Bello-Orgaz, Gema, Jason J. Jung, and David Camacho. Social big data: Recent achievements and new challenges. Information Fusion 28, pp. 45-59, (2016).

[14] Madhavji, Nazim H., Andriy Miranskyy, and Kostas Kontogiannis. Big picture of big data software engineering: with example research challenges.Proceedings of the First International Workshop on BIG Data Software Engineering. IEEE Press, 2015.

[15] Anderson, Kenneth M. Embrace the challenges: software engineering in a big data world. Proceedings of the First International Workshop on BIG Data Software Engineering. IEEE Press, 2015.

[16] Anderson K. Home | Kenneth M. Anderson [Internet]. Cs.colorado.edu. 2016 [cited 23 November 2016]. Available from: http://www.cs.colorado.edu/~kena/

[17] DeLine, Robert. Research opportunities for the big data era of software engineering.Big Data Software Engineering (BIGDSE), 2015 IEEE/ACM 1st International Workshop on.IEEE, 2015.

[18] Gorton, Ian, AyseBasarBener, and AudrisMockus. Software Engineering for Big Data Systems.IEEE Software 33.2, pp. 32-35, (2016).

[19] Metzger, A. Software engineering: Key enabler for innovation. NESSI White Paper (2014).

[20] Lau, Lydia, Fan Yang-Turner, and Nikos Karacapilidis. Requirements for big data analytics supporting decision making: A sensemaking perspective.Mastering Data-Intensive Collaboration and Decision Making.Springer International Publishing, pp. 40-70, 2014.

[21] Russell, D.M., Stefik, M.J., Pirolli, P., Card, S.K.: The cost structure of sensemaking. In: Proceedings of SIGCHI. ACM Press, New York, pp. 269–276 (1993)

[22] [Internet]. 2016 [cited 23 November 2016]. Available from: http://csse.usc.edu/new/wp-content/uploads/2013/08/2013-10-23_COCOMO_BigData_Rachchabhorn_Wongsaroj.pdf