VOL. 10, NO. 8, MAY 2015 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences


www.arpnjournals.com

A SURVEY ON BIG DATA AND ITS RESEARCH CHALLENGES

S. Justin Samuel[1], Koundinya RVP[2], Kotha Sashidhar[3] and C.R. Bharathi[4]
[1]Department of IT, Faculty of Computing, Sathyabama University, Chennai, India
[2,3]B.Tech, Faculty of Computing, Sathyabama University, Chennai, India
[4]Department of ECE, Vel Tech University, Chennai, India
E-Mail: dr.s.justin@gmail.com

ABSTRACT

There has been an ever-increasing interest in big data due to its rapid growth and since it covers diverse areas of applications. Hence, there seems to be a need for an analytical review of recent developments in the big data technology. This paper aims to provide a comprehensive review of the big data state of the art, conceptual explorations, major benefits, and research challenging aspects. In addition to that, several future directions for big data research are highlighted.

Keywords: big data, research challenges, big data architecture, big data open problems.

1. INTRODUCTION

Big data is a term encompassing different types of complicated and large datasets that is hard to process with the conventional data processing systems. Numerous challenges are in place with big data like storage, transition, visualization, searching, analysis, security and privacy violations and sharing. The exponential growth of data in all fields demands the revolutionary measures required for managing and accessing such data. In [1], the authors have highlighted the need for the research in big data, in order to manage the online bio-logical data avenue. They have foreseen the importance of big data in the biological and biomedical research. It has exploded in such a way that it has marginalized a regulatory schema for personally identifiable information [2]. This is possible by analyzing the meta data and by using the predictive, aggregated findings thereby combining the previous discrete data sets. The significance of big data analytics comes when enterprises choose a technical stack, which dictates the type of data to store and to process. Relational Data Base management Systems are doing fine with structured data and continue to be the choice for many requirements. But for the exponential growth of unstructured data in terabytes or even peta bytes, derived from social networks, sensor networks and other federated data with replications, big data is the answer for handling such data.

In general, cloud based big data seems cost effective, speedy to build and scalable. The adverse effect for the administrators lies in riding the data. If a system administrator has a private cloud, the unstructured data look like conventional structured data stack where IaaS (Infrastructure as a Service) is in the bottom, middle layer is the database and applications on the top. But in public cloud services, the concern about data security and managing privacy are on the rise. Storing NoSQL data in systems like MangoDB, Amazon SimpleDb, Windows Azure Tables, started the era of Big data science. This paper ensembles a detailed study on big data technology with characteristics, behavior and classified categories. The main focus is on throwing light into the research issues, opportunities and open problems.

2. FIVE V'S OF BIG DATA

There are many properties associated with big data. The prominent aspects are Volume, Variety, Velocity, Variability and Value.

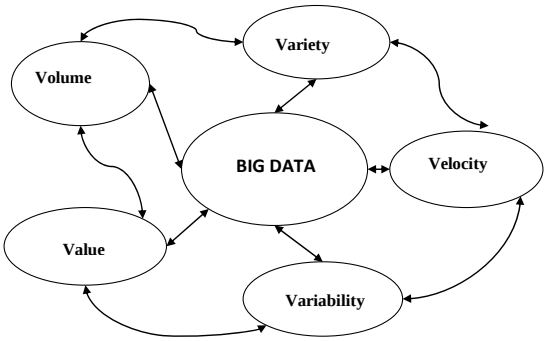

Figure-1. Five V's of Big Data.

There are many properties associated with big data. The prominent aspects are Volume, Variety, Velocity, Variability and Value.

Volume: The volume of big data is exploding exponentially day to day. The data accumulated through social websites and sensor networks going to cross from petabytes to Zetabytes.

Variety: Data produced are from different categories, consists of unstructured, standard, semi-structured and raw data which are very difficult to be handled by traditional systems.

Velocity: This is a concept which indicates the speed at which the data generated and become historical. Big data is able to handle the incoming and outgoing data rapidly.

Variability: It describes the amount of variance used in summaries kept within the data bank and refers how they are spread out or closely clustered within the data set.

Value: All enterprises and e-commerce systems are keen in improving the customer relationship by

providing value added services. For that, study on customer attitudes and trends in the market are to be analyzed. Moreover, users can also query the data store to find business trends and accordingly they can change their strategies. By making big data open to all, it creates transparency on functional analysis. Supporting real time decisions and experimental analysis in different locations datasets can do wonderful things for enterprises.

3. BIG DATA CLASSES

Categorization of big data falls with major aspects, since this technology involves with multiple diversified fields and ir-related types of information handling. Some of the classes can be framed like storing, sourcing, formatting, staging and processing [11]. Each class instantiates many entities through which the actions carried out. This is depicted as a technical view in Figure-2.

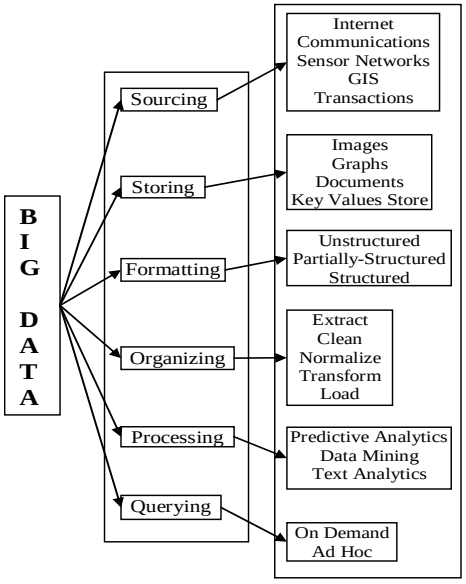

Figure-2. Technical view of Big Data.

Sourcing: Data sources identified are Internet Web Pages, Discussion Forums, Chats and message shared in and among social networks, Remote Sensing Networks, All kinds of day to day transactions done through internet based applications.

Formats: Unstructured, partially structured, and structured.

Storing: Image based, Graph based, documents, Key Value Stores (Key Values Store is a way of storing application's data with null schema. It doesn't require a static data model. Unique keys are used to represent values stored in it.)

Organize: Extract, Clean, normalize, transform, Load

Process: Online, Offline

Query: On demand, ad-hoc

4. RESEARCH AREAS AND THE CHALLENGES

There are six major research areas identified as shown in Figure-3. They are,

a) Applied ontology
b) Security
c) Storage and Transport
d) Accessibility
e) Inconsistencies
f) Mobility

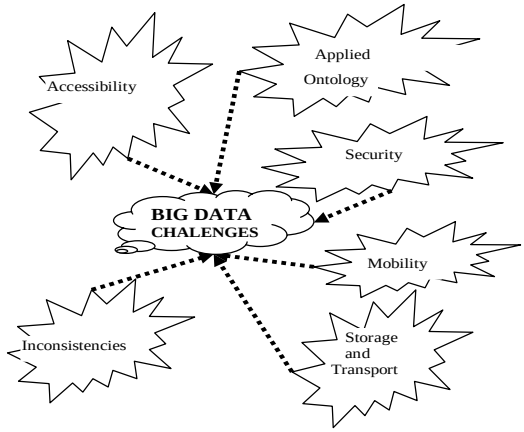

Figure-3. Major research areas in Big Data.

4.1 Applied Ontology

Applied Ontology is the way of applying the ontological resources to the domains like Geography as well as Bio-medicine. These works can be done within the semantic web framework. Applied ontology involves in looking the relationship between a person's world and his actions. The 9th Ontology Summit held on Jan'16-2014 under the theme "Big Data and Semantic Web Meet Applied Ontology" [4] has concluded the following;

- To build bridges between Semantic web, big Data, Linked Data and Applied Ontology.
- Performance issues, Challenges in scalability while combining big data and Semantic web. Technically, automated reasoning tools to be developed to make use of ontologies. Large common reusable ontologies and ontological analytical techniques to overcome the engineering bottlenecks are the need of the hour.

4.2 Security

Big data involved with many use cases like Staging, Pre-processing, Processing, Meta data storage and to store short term as well as long term fact data. For serving each use case multi-facets of infrastructure required.

Safe and private transactions are the two major concerns of IT. But the safety and privacy becomes a question mark as the data volume of big data fast grows. When we consider the safety aspect, the existing cryptography standards can not meet the demands of big

data [5]. Hence; effective mechanisms to handle structured, semi-structured, unstructured data have to be investigated and to be developed. Data acquisitions of users like habits, personal interests and the like through websites may happen with the permissions of the users or maybe when the users are not aware of. But the same may be leaked while storing, transmitting or handling. According to report [6], researcher Ron acquired 2.8 GB of Facebook user's data and made it available to download on the internet. Hence, Privacy protection is another challenging problem in big data.

4.3 Storage and transport

Big data stores and handles data in different way from traditional data warehouses. Big data comprises massive sensor data, raw and semi-structured log data of IT industries and the exploded quantity of data from social media. As per the examples given in [7], current disk technologies are limited to store 4 TB per disk. For storing 1 exa byte, it requires 25, 000 disks which will overwhelm the existing communication technologies. That means such phenomenon demands for a revolution in storage and communication technologies.

4.4 Accessibility

The rapid growth of data on the internet, challenges the research community to move towards innovating efficient algorithms and processing technologies. The access technique for big data has two manifestations. First, process the data in the source side and transmit results. That is, scripting technologies on the browser side should be improved to bring the necessary code from the server. Second, transmit only the critical data after performing through valid filters.

For enabling this scenario, the following recommendations are made.

a) High performance convexed clustering algorithms to be developed to improve the performance in the distributed and parallel architectures.
b) Efficient data analytic techniques required to handle large amount of data and to filter the data by applying proper constraints.
c) The Natural Language Processing techniques should become applicable over all kinds of enterprise data, politics, bio-medics and all other domains.
d) Conventional tools to handle Big data are outdated and become inefficient. So a lot of research required in developing tools for Big Data and platforms for deployment.
e) Improved algorithm for crawling data from multiple platforms is needed. Also, powerful algorithms to visualize random data from multi-disciplinary segments are to be generated to get accurate results as input for crucial applications.

4.5 Inconsistencies

The Big data research area is embedded with multi-dimensional technical and scientific spaces. The objectives of big data analytics differ with the stakeholders. As this big data analysis is the next frontier for innovation and advancement of technology, one should not underestimate the impact of it on the society. Soon the big data cloud is going to cover all domains and sectors like manufacturing, spacial data science, life and physical sciences, communication, finance and banking etc., As big data comprises all domains, definitely inconsistencies arise. Inconsistencies in data level, information level and knowledge level have to be addressed. There are four types of inconsistencies like temporal, textual, spatial and functional dependency. These inconsistencies are well addressed by Zhang in [8].

4.6 Mobility

Enterprises are poised to extend more investment in applications which support mobile devices. Very great potential to increase productivity is on the way for businesses when they combine enterprise process automation and mobile computing technologies. Increasing location based datasets, influx of data from mobile applications, their size and variety exceeds the capacity of spatial and mobile computing technologies. Mobile users contribute a lot to big data analytics through their online activities. The convergence of traditional routing services (including GPS and spatial data) into the big data paradigm has to face major challenges. First, it increases the computational cost because it magnifies the impact of routing queries to mobile devices. Second, it uses geographical reasoning in remote sensing and inference over time and space. The built-in motion detectors in mobile phones derive a huge amount of data from every user's life. How efficiently utilize these data and how to carry and share through limited bandwidth mobile stations, are the other challenges.

5. TECHNICAL CHALLENGES IN BIG DATA

Whenever new technologies evolve, they meet with new challenges in all the aspects. Once the functional challenges are in place, the next kin is the technical challenges. Big data faces many technical challenges which are on the roadway of the research.

a) Failure handling

Devising 100% reliable systems on the go is not an easy task. Systems can be devised in such a way that the probability of failure must fall within the permitted threshold. Fault tolerance is a technical challenge in big data. When a process started it may involve with numerous network nodes and the whole computation process becomes cumbersome. Retaining check points and fixing the threshold level for process restart in case of failure, are greater concerns.

b) Data heterogeneity

Big data deals with unstructured, semi-structured and structured data. Linking unstructured data with structured data, converting data from one form into another required form needs a lot of research.

c) Data quality

Huge amount of data pertaining to a problem is undoubtedly a big asset for both Business as well as IT leaders [9]. For predictive analysis or for better decision making amount of relevant data helps a lot. But the quality of such data is based on the source through which they are derived. Though big data stores large relevant data, the accuracy of data is completely dependent on the source domains. Hence, there is a question of how far the data can be trusted and it definitely requires appropriate trust agent filters.

6. OPEN PROBLEMS

a) Data management in a single perfect manner is remaining an open problem for both cloud and big data.

b) Big data handles unstructured data also. Hence, the data structures differ from conventional SQL databases. The data structures for NoSQl databases are Graph, Documents and Key value stores [10]. Though there are technologies to manage graphics and documents, Key value store needs a lot of research. The functionalities of handling key value stores to be improved to support on demand queries and also extended key value stores are required to manage diverse data oriented rich internet applications.

c) Elasticity for effective usage of unstructured as well as structured resources with consistent semantics, operating with those resources at a minimum cost and enabling autonomy in multi-tenant data systems are some of the open problems.

7. PROGRESS IN BIG DATA ANALYTICS

An insight about the ongoing research in Big data analytics has been presented in this section.

Table-1. Recent evolvements in Big Data research.

Ref.	Used technology	Findings
[12]	Hadoop and map reduce	Solution for Optimizing Job Scheduling, Organization of indexes and Layouts rendered.
[13]	Hive, Hadoop and Mahout	Have built a Random Forest based Decision Tree model to detect botnet in peer to peer network.
[14]	Rapidminder and Hadoop	Proposed an architecture called Radoop to scale the data and network size.
[15]	Hadoop and IBM smart analytic system	Introduced new architecture to support the analytical system
[16]	Hadoop	Proposed a self-tuning system called Starfish.

8. CONCLUSIONS

The data size in all areas is exploding day to day. The velocity and variety of data growth is increasing due to the proliferation of sensor and mobile devices with internet connection. Data generated by this way, is the greatest asset for enterprises in developing and defining business strategies. Cloud services were used to process and analyze huge amount of data and it has turned into the new Big Data model to meet the on-demand services. In this paper, we have done an elaborated study on Big Data and its research challenges. We have highlighted the existing problems and have presented the research opportunities. We propose the technical view of big data comprising the various classes. Also, the key research areas have been identified and the insights into those areas are discussed. Furthermore, key issues to be addressed by both industry and academia, are brought into light.

Even though a conclusion may review the main results or the contributions of the paper, do not duplicate the abstract or the introduction. For a conclusion, you might elaborate on the importance of the work or suggest the potential applications and extensions.

REFERENCES

[1] Howe AD, Costanzo M, Fey P, et al. 2008. Big data: The future of biocuration, Nature. 455(7209): 47-50. Doi: 10.1038/455047a.

[2] Crawford Kate and Jason Schultz. 2014. Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms, Boston College Law Review. 55(93): 93-128.

[3] http://www.techrepublic.com/blog/the-enterprise-cloud/cloud-computing-and-the-rise-of-big-data/.

[4] http://ebiquity.umbc.edu/blogger/2014/01/14/2014-ontology-summit-big-data-and-semantic-web-meet-applied-ontology/.

[5] Min Chen, Shiwen Mao, Yunhao Liu. 2014. Big Data: A Survey, Mobile Networks and Applications. 19(2): 171-209.

[6] Tasevski P. 2011. Password attacks and generation strategies, Tartu University: Faculty of Mathematics and Computer Sciences.

[7] Stephen Kaisler, Frank Armour, J. Alberto Espinosa, William Money. 2013. Big Data: Issues and

Challenges Moving Forward, Proceedings of 46th Hawaii International Conference on System Sciences, IEEE, (pp. 995-1004 Year of Publication: ISBN: 978-1-4673-5933-7).

[8] Du Zhang, Inconsistencies in big data, Proceedings of the 12th IEEE International Conference on Cognitive Informatics, New York City, NY (Page: 61-67 Year of Publication: 2013 ISBN: 978-1-4799-0781-6).

[9] A. Katal, M. Wazid, and R. H. Goudar, Big data: Issues, challenges, tools and Good practices, Proceedings of Sixth International Conference on Contemporary Computing (IC3), (Page: 404-409 Year of Publication: 2013 ISBN: 978-1-4799-0190-6).

[10] http://en.wikipedia.org/wiki/NoSQL.

[11] Ibrahim Abaker Targio Hashema, Ibrar Yaqooba, Nor Badrul Anuara, Salimah Mokhtara, Abdullah Gania, Samee Ullah Khanb. 2015. The rise of "big data" on cloud computing: Review and open research issues, Information Systems, Elsevier. 47: 98-115.

[12] Dittrich Jens and Jorge-Arnulfo Quiané-Ruiz. 2012. Efficient big data processing in Hadoop MapReduce, Proceedings of the VLDB Endowment. 5(12): 2014-2015. http://vldb.org/pvldb/vol5/p2014_jensdittrich_vldb2012.pdf.

[13] Singh Kamaldeep, et al. 2014. Big Data Analytics framework for Peer-to-Peer Botnet detection using Random Forests. Information Sciences, Elsevier. 278: 488-497.

[14] Prekopcsák Zoltán, et al. 2011. Radoop: Analyzing big data with rapidminer and hadoop. Proceedings of the 2nd RapidMiner Community Meeting and Conference. http://prekopcsak.hu/papers/preko-2011-rcomm.pdf.

[15] Ferguson Mike. 2012. Architecting a Big Data Platform for Analytics, a Whitepaper Prepared for IBM. http://public.dhe.ibm.com/common/ssi/ecm/en/iml14333usen/IML14333USEN.PDF.

[16] Herodotou Herodotos, et al. 2011. Starfish: A Self-tuning System for Big Data Analytics, CIDR. http://www.cs.duke.edu/~gang/documents/CIDR11_Paper36.pdf.