IJCSE International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.6, Special Issue.11, Dec 2018 E-ISSN: 2347-2693

Big Data Analytical Tools, Challenges and Applications: A Survey

M. Saraswathi

Sri Meenakshi Vidiyal Arts and Science College, Paluvanchi, Tiruchirappalli, India

*Corresponding Author: sarasjjmca@yahoo.com

Available online at: www.ijcseonline.org

Abstract— In this paper presents the term what exactly means "Big data". we first investigate how big data is, and what are the recent technologies developed for big data. Due to this, we identify the big data applications including enterprise management, Internet of Things, online social networks, media and entrainment and healthcare. The various Challenges faced in large data management include scalability, unstructured data, accessibility, real time analytics, fault tolerance and many more. This survey is concluded with problems to be identified and future directions.

Keywords— big data, Internet of Things, Challenges.

I. INTRODUCTION

Big data is incredibly giant sophisticated that's in adequate to process the data with ancient systems and tools.IT might be needed a parallel computer code running on tens, tons of or perhaps thousands of severs. Big data will be found in 3 forms like organized, unstructured and semi-structured with ample technologies and services developed. Big data consists of terribly giant volumes of varies knowledge that is being made often, at high speeds. Big data experiments capturing knowledge, knowledge storage, knowledge analysis, search, sharing, transfer, mental image, querying, updating,data privacy and knowledge supply.
Big data features are

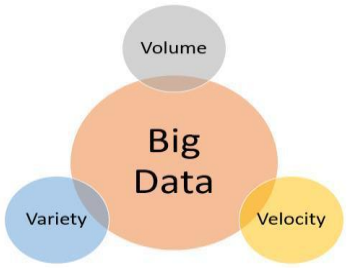

Fig.1

Volume-The name 'Big data' itself is expounded to a size that is huge. Size of data plays terribly crucial role in decisive worth out of data. Also, Whether or not a specific knowledge will really be thought-about as an enormous knowledge or not, depends upon volume of data. Hence, 'Volume' is one characteristic that has to be thought-about where as handling 'big data'.

Variety-Consequent side of 'Big data' is its selection. Variety refers to heterogeneous sources and therefore the nature of knowledge, each structured and unstructured. Throughout earlier days, spreadsheets and databases were the sole sources of data thought-about by most of the applications. Now days, data within the type of emails, photos, videos, observance devices, PDFs, audio etc. is additionally being thought-about within analysis applications. This kind of unstructured knowledge poses bound problems for storage, mining and analysing data.

Velocity– The term 'speed' alludes to the speed of age of learning. how energetically the data is created and handled to fulfil the pressure, decides genuine potential inside the information. Huge learning speed manages the speed at that information streams in from sources like business forms, application logs, systems and internet based life destinations, sensors, Mobile gadgets, and so forth. The stream of information is vast and nonstop.

Variability – This alludes to the irregularity which may be appeared by the data from time to time, so hampering the strategy for being able to deal with and deal with the data successfully.

II. LITERATURE SURVEY

D. P. Acharjya etal.(2016) explored impact of big data challenges, open research issues and necessary tools are associated with big data.KuchipudiSravanthiet al.(2015) highlighted the applications of big data especially for government and corporate sectors.J.Archenaa et al.(2015) to offer data analytics for health care and government systems. Hadoop plays an important role of real time analysis on large volume of data.

Chun‐Wei Tsai1 etal.(2015) designed a suitable mining algorithms to find useful things from big data. It is mainly focus on briefly discussed on data analytics and their open issues.

Amir Gandomi etal.(2015) mainly focused on data analytics associated with unstructured data. Highlighted the need of developing suitable and efficient analytical methods to influenced volumes of heterogeneous data in unstructured text, audio, and video format. It is mainly supports new tools for predictive analytics for structured big data.Min Chen et al. (2014) reviewed the related technologies such as could computing, Internet of Things, data centres, and Hadoop. We then emphasis on the four stages of the value chain of big data, i.e., data generation, data acquisition, data storage, and data analysis. Several applications can be examined.

Nawsher Khan et al. (2014)big data characteristics, tools for analysis and management related to big data. Open issues and challenges related to big data. CheikhKacfahEmani et al. (2015) presented models and features of big data. Stages for processing big data and what are the Problemsfaced in big data management. Then, the problem of merging Big Data architecture in an already present information system.

Ahmed Oussous a et al. (2017) recent technologies developed for big data. It helps to select and implement the right combination of dissimilar Big Data technologies according to their technical needs and specific applications' requirement. RaguseoElisabetta(2018)examines the adoption levels of big data technologies in companies, and the big data sources used by them. It also ideas out the most frequently recognized strategic, transactional, transformational and informational profits and risks related to the usage of big data technologies by companies. In order to accomplish these at the differences that exist among companies of dissimilar sizes, by comparing medium-sized and large companies, and the differences among companies of different industrial sectors, and provides evidence that only in a few cases these differences are important.

Jinchuan CHEN et al. (2013) reviewed challenges in big data from perspective of data management. The various types of big data and how it can be applied to data processing techniques, Data integration methods and techniques for data reduction, processing big data by using indexing mechanism and finally research opportunities available for data mining in the perspective of big data. Harshawardhanet al. (2014) to process large amount of data in efficient manner use parallelism. Hadoop is an open source software that ensures distributed processing from clusters of servers. Justin Samuel et.al (2015) categories of big data, what are all open problems in research areas to be identified and its technical challenges.

Ms. VibhavariChavan et al. (2014) presented concepts of big data and challenges of the big data to be discussed. whereas traditional methods are hard for processing big data. The new techniques and technologies are used to discover the hidden data from large data sets. Manjit Kaur et al. (2013) the definition of big data and how the big data can be categorized in terms of volume, velocity and variety. A well-known method for processing big data such as Hadoop which uses map reduce paradigm.

III. BIG DATA ANALYTICS-METHODOLOGY

In terms of methodology, big data analytics variesconsiderably from the traditional statistical approach of experimental design. Analytics begins with data. Typically, we model the data in a way to describe a response. The aims of this approach is to estimate the response behaviour or understand how the input variables communicate to a response. Generally, in statistical experimental designs, an experiment is established and data is retrieved as a result. This allows to generate data in a way that can be used by a statistical model, where certain expectations hold such as independence, normality, and randomization.

In big data analytics, we are presented with the data. We cannot design an experiment that fulfils our favourite statistical model. In large-scale applications of analytics, a large amount of work (normally 80% of the effort) is needed just for cleaning the data, therefore it can be used by a machine learning model.

We don't have a distinct methodology to follow in real large-scale applications. Typically, once the business problem is well-defined, a research stage is required to design the methodology to be used. Though general guidelines are related to be stated and apply to almost all problems.

One of the most important job in big data analytics is statistical modelling, meaning supervised and unsupervised classification or regression problems. Once the data is cleaned and pre-processed, available for modelling, care should be taken in assessing different models with reasonable loss metrics and then once the model is implemented, further evaluation and results should be reported. A common pitfall in predictive modelling is to just implement the model and never measure its performance.

IV. TECHNIQUES AND TECHNOLOGIES

Apache Hadoop

Apache Hadoop is a java based free programming structure that can successfully store expansive measure of information in a group. This system keeps running in parallel on a bunch

and has a capacity to enable us to process information over all hubs. Hadoop Distributed File System (HDFS) is the capacity arrangement of Hadoop which parts huge information and circulate crosswise over numerous hubs in a group. This likewise imitates information in a bunch hence giving high accessibility.

Microsoft HDInsight
It is a giant information answer from Microsoft powered by Apache Hadoop that is on the market as a service within the cloud. HDInsight uses Windows Azure Blob storage because the default filing system. This conjointly provides high accessibility with low price.

NoSQL
While the normal SQL may be effectively accustomed handle great amount of structured knowledge, we'd like NoSQL (Not solely SQL) to handle unstructured knowledge. NoSQL knowledge bases store unstructured data with no specific schema. Every row will have its own set of column values. NoSQL offers higher performance in storing huge quantity of knowledge. There are a unit several ASCII text file NoSQL DBs accessible to analyse huge knowledge.

Hive
This is a distributed information management for Hadoop. This supports SQL-like question choice HiveSQL (HSSQL) to access huge information. This will be primarily used for data processing purpose. This runs on prime of Hadoop.

Sqoop
This is a tool that connects Hadoop with numerous relative knowledge bases to transfer data. this will be effectively wont to transfer structured knowledge to Hadoop or Hive.

PolyBase
This works on high of SQL Server 2012 parallel information warehouse(PDW) and is employed to access information hold on in PDW.PDW could be an information warehousing appliance designed for process associate volume of relative information and provides an integration with Hadoop permitting U.S.A to access non-relational data yet.

Big data in EXCEL
As many folks square measure snug in doing analysis in surpass, a well-liked tool from Microsoft, you'll be able to additionally connect knowledge hold on in Hadoop Victimization surpass 2013.Hortonworks, that is primarily operating in providing Enterprise Apache Hadoop, provides associate degree choice to access huge knowledge hold on their Hadoop platform Victimization surpass 2013.you'll be able to use power read feature of surpass 2013 to simply summarise the information.

Presto
Facebook has developed and recently open-sourced its Query engine (SQL-on-Hadoop) named Presto which is built to handle petabytes of data. Unlike Hive, Presto does not depend on MapReduce technique and can quickly retrieve data.

V. APPLICATIONS

Health care
Now massive knowledge analytics have improved attention by providing personalised medication and prescriptive analytics. Researchers area unit mining the info to visualize what treatments area unit simpler for explicit conditions, determine patterns associated with drug aspect effects, and gain different vital data that may facilitate patients and scale back prices.

With the additional adoption of mHealth, ehealth associated wearable technologies the amount if information is increasing at an exponential rate. This includes electronic health record knowledge, imaging knowledge, patient generated knowledge, sensing element knowledge and different types of knowledge. By mapping attention knowledge with geographical knowledge sets, it's do able to predict wellness which will intensify in specific areas based mostly of predictions, it's easier to strategize medical specialty and arrange for stocking serums and vaccines.

Fig.2

Manufacturing
Predictive producing provides near-zero period and transparency. It needs a massive quantity of knowledge and advanced prediction tools for a scientific method of knowledge into helpful information.
Major edges of victimization massive information applications in producing business are:

- Product quality and defects chase
- Supply designing
- Manufacturing method defect chase
- Output prognostication
- Increasing energy potency
- Testing and simulation of recent producing process
- Support for mass-customization of producing

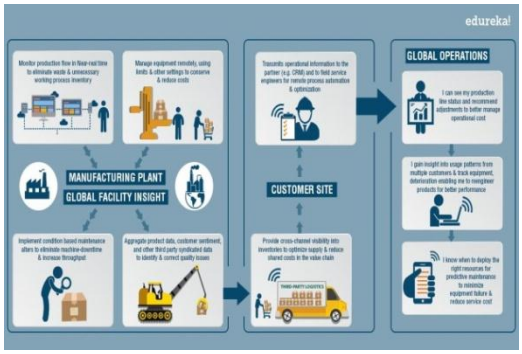

Fig.3

Media and Entertainment

Various corporations within the media and industry face new business models, for the means they-produce, market and distribute their content. This is often happening attributable to current consumer's search and also the demand for accessing content anyplace, any time on any device.

Big information provides unjust point of knowledge regarding countless people.

Now, business environments square measure craft advertisements and content to charm shoppers. These insights square measure gathered through numerous data-mining activities. Huge information applications edge media and industry by:

- Predicting what the audience needs
- Scheduling improvement
- Increasing acquisition and retention
- Ad targeting
- Content validation and new development

Internet of Things(IOT)

Data extracted from IoT devices provides a mapping of device inter-connectivity. Such mappings are employed by numerous corporations and governments to extend potency. IoT is additionally more and more adopted as a way gathering sensory knowledge, and this sensory knowledge is employed in medical and producing contexts.

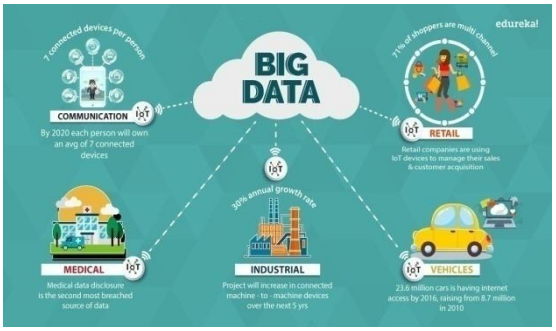

Fig.4

Government

The use and adoption of massive knowledge among governmental processes permits efficiencies in terms of price, productivity and innovation. In government use cases, an equivalent knowledge sets area unit typically applied across multiple applications and it needs multiple departments to figure together. Since government majorly acts all told the domains, so it plays a very important role in innovating huge knowledge applications in every and each domain.

VI. CHALLENGES

There is a number of challenges relating to its volume and complexity.

Scalability

Big information storage and analytics demand square measure attending to increase. Considering a giant information to understand that the answer they select should be to proportion and own on-demand to resolve these drawback adding additional physical servers which will be costlier and time overwhelming. With Hadoop deals quicker and easier measurability to accommodate growing information demands.

Unstructured data

Datasets square measure very giant, it's laborious to manage, and doesn't have predefined schema. Unstructured information is on the market within the style of complicated information formats, like emails, text files, web pages, digital pictures, multimedia system content, navigation details and social media post ancient systems like (RDBMS) weren't creator to store and retrieve unstructured information. Rather distributed systems solve several of the challenges associated with storing and retrieving unstructured information.

Accessibility

One of its challenges of massive information square measure inconvenience of information sets from external sources. Considering the degree of information can't be hold on in an exceeding single machine, rather than requiring super computers and high value. For managing massive information, tools and techniques square measure like Hadoop, Map reduce. However, several corporations should develop tools and technologies for access and analyse great deal of information.

Real time analytics

Real time analytical process might embody single or multiple integrated services. Storing massive information is typically difficult. Hence it is often abstracted into smaller information sets specified it will considerably enhance the execution time of analytical process. Period of time analytical services got to install quick algorithms that give various choices among a

restricted time. These choices are often optimum or semi-optimal thanks to the restricted time and handiness of the resources. This might be required for process intensive, high performance platforms. These platforms are often equipped with special hardware to execute special algorithms or with cluster computing systems.

Fault tolerance

Fault tolerance is very laborious to work out for victimisation complicated algorithms.it is merely uphill for 100 percent reliable fault tolerance machines or package. therefore, the most task is to scale back the chance of failure to associate "acceptable" level. sadly, we tend to attempt to scale back the chance, the upper the price

VII. CONCLUSION

The main objective of this paper is to examine the role of big data. An overview of what are the technologies used. In this paper we explored the different applications like Health care, Manufacturing, Media and Entertainment, Internet of Things(IoT) and Government. Challenges of big data can be reviewed. Researchers may get some ideas related to big data and also it is helpful to their research areas. In future work we can consider necessary tools and technologies used to overcome the challenges.

REFERENCES

[1] D. P. Acharjya,andKauser Ahmed *P "A Survey on Big Data Analytics: Challenges, Open Research Issues and Tools*" (IJACSA) International Journal of Advanced Computer Science and Applications, Vol. 7, No. 2, 2016

[2] KuchipudiSravanthi,,and TatireddySubba Reddy," *Applications of Big data in Various Fields*" (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (5) , 2015, 4629-4632

[3] J.Archenaa and E.A.Mary Anita *"A Survey Of Big Data Analytics in Healthcare and Government"* Procedia Computer Science 50 (2015) 408 – 413

[4] Chun‐Wei Tsai1, Chin‐Feng Lai, Han‐Chieh Chao and Athanasios V. Vasilakos "*Big data analytics: a survey*" Tsai et al. Journal of Big Data (2015) 2:21

[5] Amir Gandomi, MurtazaHaider"*Beyond the hype: Big data concepts, methods, and analytics*" International Journal of Information Management 35 (2015) 137–144

[6] Min Chen · Shiwen Mao · Yunhao Liu"*Big Data: A Survey*" Mobile NetwAppl (2014) 19:171–209

[7] Nawsher Khan,1,2 Ibrar Yaqoob,1 Ibrahim AbakerTargio Hashem,1Zakira Inayat,1,3 Waleed KamaleldinMahmoud Ali,1 Muhammad Alam,4,5Muhammad Shiraz,1 and Abdullah Gani1 *"Big Data: Survey, Technologies, Opportunities, and Challenges*", The Scientific World JournalVolume 2014, Article ID 712826, 18 pages

[8] CheikhKacfahEmani, Nadine Cullot, Christophe Nicolle "*Understandable Big Data: A survey*" Computer Science Review 17 (2015) 70-81

[9] *Information Management* 35 (2015) 137–144

[10] Ahmed Oussousa , Fatima-Zahra Benjelloun a , Ayoub Ait Lahcena,b, , Samir Belfkih *a "Big Data technologies: A survey*" Journal of King Saud University – Computer and Information Sciences xxx (2017) xxx–xxx

[11] ElisabettaRaguseo *"Big data technologies: An empirical investigation on their adoption, benefits and risks for companies"* International Journal of Information Management 38 (2018) 187–195

[12] Ibrahim AbakerTargio Hashem a,n , IbrarYaqoob a , Nor BadrulAnuar a , Salimah Mokhtar a , Abdullah Gani a , Samee Ullah Khan b*"The rise of "big data" on cloud computing: Review and open research issues* "Information Systems 47 (2015)98–115

[13] Jinchuan CHEN, Yueguo CHEN, Xiaoyong DU, Cuiping LI, JiahengLU,Suyun ZHAO, Xuan Zhou *"Big data challenge: adata management perspective*" front comput. Sci.,2013,7(2),:157-164

[14] Harshawardhan S. Bhosale1 ,Prof.Devendra P. Gadekar2 *"A review paper on big data and Hadoop"* international journal of scientific and research publications, volume 4, issue 10, october 2014 756 ISSN 2250-3153.

[15] S. Justin Samuel1 ,Koundinya RVP2 , Kotha Sashidhar3 and C.R. Bharathi4 *"A survey on big data and its research challenges"* ARPN Journal of Engineering and Applied Sciences Vol. 10, no. 8, may 2015 ISSN 1819-6608.

[16] Ms.VibhavariChavan, Prof.Rajesh.N.Phursule*"Survey Paper On Big Data*" International Journal of Computer Science and Information Technologies, Vol. 5 (6) , 2014, 7932-7939

[17] Manjit Kaur and Shilpa "*BIG Data and Methodology-A review*" © 2013, IJARCSSE , Volume 3, Issue 10, October 2013

