WWJMRD 2018; 4(1): 25-28
www.wwjmrd.com
International Journal
Peer Reviewed Journal
Refereed Journal
Indexed Journal
UGC Approved Journal
Impact Factor MJIF: 4.25
e-ISSN: 2454-6615

Ankur Gupta
Assistant Professor in Computer Science
RSD College, Ferozepur City, India

Correspondence:
Ankur Gupta
Assistant Professor in Computer Science
RSD College, Ferozepur City, India

Big Data: Survey, Technologies, Characteristics, Challenges and Opportunities

Ankur Gupta

Abstract
Today various fields are experiencing associate large increase within the volume, speed and complexness of information. There's a flood of information in nearly each field. Antecedently selections were created in line with estimation. However these days in each field, before creating any selections, information analysis is finished. This large quantity of information is thought as massive data. There are several challenges to handle analysis of massive information as a result of ninety percent of information is unstructured and solely ten percent is structured. Information is being generated by users, applications, systems, sensors and lots of a lot of. Most Organisations are creating analysis of massive information for remaining in competition and for up quality in their merchandise. This paper presents the challenges baby-faced by information Scientists for doing analysis of information. It conjointly highlights opportunities for researchers.

Keywords: Big Data, Hadoop, Map Reduce, Volume, Velocity etc.

Introduction
Big knowledge has gained attention from the majority industries, institutes, IT business etc. virtually two Billion folks are connected with net and as technology is advancing, the majority the devices are generating large quantity of information. For instance Air bus generates ten TB each half-hour. concerning 640 TB is generated in each flight, good meters reads the usage each quarter-hour, records 350 billion transactions in an exceedingly year, three hundred billion emails are sent on a daily basis, facebook generates twenty five TB of information daily, Twitter generates twelve TB of information daily[1] etc. There are several challenges to traumatize large quantity of information. Analysis is extremely promising for business purpose. Increasing quantity of information has created several problems and challenges like transfer speed, information and security problems etc. To see the power of analysis of big data consider the figure [2].

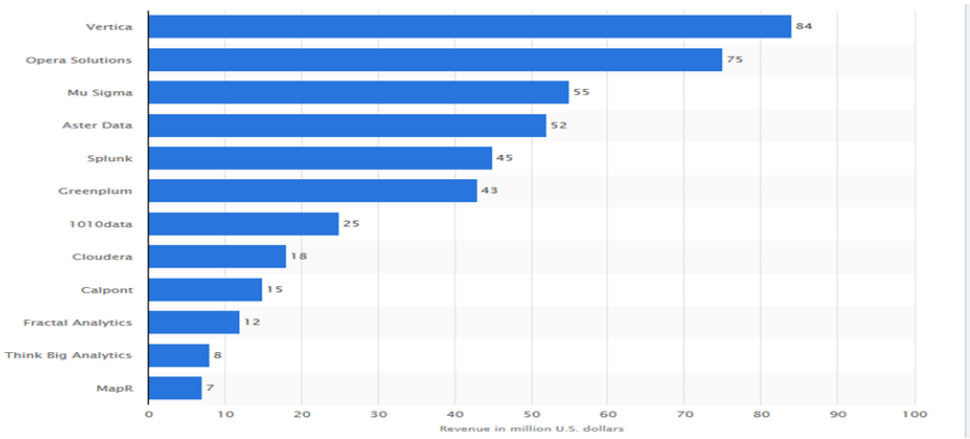

Fig.1: It shows the amount of revenue (in million U.S. dollars) generated by many companies named in the figure around the world.

There are many phases in data analysis process. Foremost is data acquisition then information extraction. After that data analysis is done, which is more challenging than identifying, locating and understanding, having heterogeneity, incompleteness, timeliness etc? There square measure heaps of definitions of huge knowledge because of its wide usage in various fields. There are the merchandise directed, method directed, noesis directed and movement perspective (7). The merchandise directed perspective appearance at options of the data; particularly its volume, velocity and selection. Huge knowledge imply modern technologies and architectures that are designed to with efficiency deduce get pleasure from vast and style of datasets (8). Another definition states huge knowledge involves vast volume, heterogeneous, localized management and finds tortuous and dynamic correlations between knowledge (9). A 3rd definition states as a results of exponential increase in world knowledge, huge knowledge signifies vast datasets. Relative to traditional knowledge, it includes vast unstructured knowledge that requires real time analysis. It conjointly comes with new opportunities for dissecting price and deeper understanding (10). The second perspective of huge data leverages the processes concerned in its operation.

Background

After every five years, information is increasing at the rate of 10 times. From 1986 to 2007, the capacity of storage devices, processing etc. has been increasing. According to a study [3], 60 analog and digital technologies has been used. In 2007, 2.9 X 10^20 compressed bytes were stored, the communication rate was 2.0 × 10^21 bytes for general purpose computers. It was also found that annual data storage growth was 23% percent between 1986 and 2007 and that of computing capacity, it was 58%. As digitisation occurs, it was not possible to store Big Data and also computing, managing, processing etc was not easy as it creates many challenges with respect to data management and analysis. Five common issues are volume, velocity, variety, value and complexity. Each of these issues represents a challenge for industry, institutions etc.

Data not require memory only in terabytes. Actually it requires memory in Exabyte. Instead space is managed by purchasing memory online. Most of the data is from users of social media which constitute 80% of the data and it is unstructured. Today demand of digital storage is quite high. In 2007, 52% of the data is stored in hard disk drives, 28% is stored in optical storage devices etc.

Characteristics

Characteristics of Big Data can be depicted in following figure [4]-

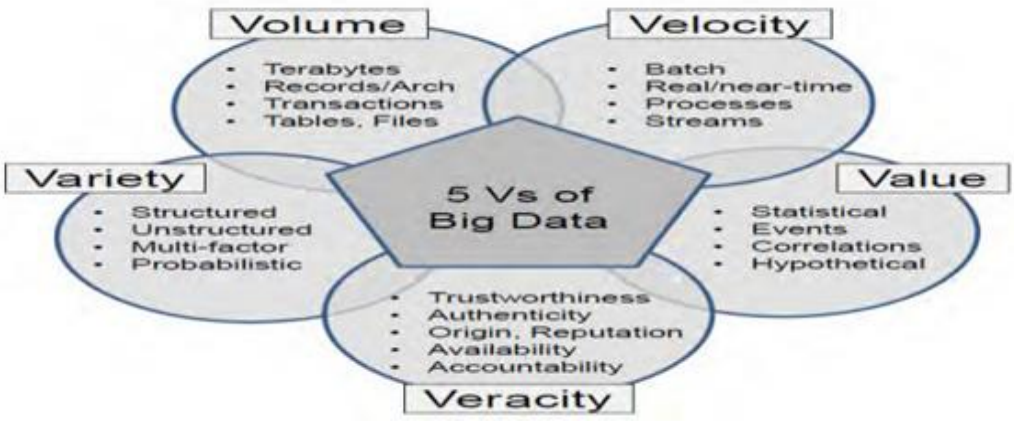

Fig.2: Characteristics of Big Data.

Volume- It represents enormous amount of Data. It is the first term that comes to the mind of people when they hear the term Big Data. Is a most important for industries, companies because the benefit achieved from analysis of data is much more important. Other challenges include availability of storage devices to store Big Data.

Velocity- It represents how fast data can be processed. It also means how data comes in from different sources like networks, sensors, airline reservations etc. This data can help the researchers to make decisions that can help both private and government organisations to provide better facilities.

Variety- It means collected data can be structured or unstructured. It can contain data which is not useful to the problem being analyzed. There must be a strategy to remove dirty data from accumulating in the system.

Veracity- It means that believability of massive information. after we square measure addressing immense quantity of information, we tend to can't make certain whether or not information is dirty or not, we tend to watch out of the origin from wherever the information is generated. We tend to should create into consideration concerning the name of the business from wherever information is collected.

Value- It is the most important term in context of Big Data. We know data is huge. It has to be converted into a form from where it can be used to make analysis. Otherwise it will be useless.

Challenges

1. **Need to find Results quickly-** In today's setting there's a great deal of competition and each trade needs to compile the info quickly thus on maximize profit.

The most challenge is to handle huge volume of data; conjointly what proportion level of detail required is needed at a high speed. Hardware support is required, having parallel processors, quantity of storage devices etc. offered thus on analyse the info quickly.

2. **Understanding Data-** All the data is collected is required for analysis. It may be possible that data required for one problem may not be required for other one. So there must be a procedure to clean the data and removing the useless data.
3. **Quality of data-** Even if the data is properly understood. How can we be sure whether it is accurate or not? If it is not, then it will not help in making correct decisions. Hence there is no use for such data. Data cleaning process must be applied in every industry. For this proper procedure should be there.
4. **Displaying Results-** As the data is huge, plotting results on graph will become be very difficult for all the data. There must be a procedure to view the data on high level i.e. to summarize findings based on some groups that can be created according to similarity in data, then those groups can be easily shown up on graph.
5. **5. Process Challenges-** There must as proper model for doing all the work. It includes capturing data, combining data from different sources, transforming it into a form suitable for analysis and then deriving useful results.
6. **Management Challenges-** Data can be sensitive that is it can be personal, secured etc. There are legal concerns with most of the data. So those things must be kept in mind while using that data.
7. **Complexity of Data-** Mainly data is of three types- Structured, Unstructured and Semi-Structured. Structured data is simple to process because this type of data is having similar formats, similar lengths etc. These are generated by sensors, networks etc. This data is easy to process using Query Languages like SQL etc. Unstructured data is not easy to process for example video, images etc all are examples of unstructured data. For processing this type of data software like Hadoop can be using. Semi structured is also processed through Hadoop because it is also difficult to handle such data.
8. **Storage Devices for Big Data-** Availability of Storage Devices is also a major challenge for industries. Because size of big data is huge in volume. So space must be made available to store all the data.

Techniques

There square measure several techniques for analyzing massive information, the techniques cut across disciplines and overlap in some cases. The disciplines vary from data processing, machine learning, neural networks and pattern recognition (11). A method from a discipline or combination of them is used to gain valuable insight from massive information. These techniques within the context of massive information square measure called analytics. Machine learning techniques square measure identified for his or her prophetic capability, given a previous instance of knowledge, the technique is ready to infer applicable response to resulting occurrence. With this ability at hand, machine learning techniques have resolved ancient computing tasks. There square measure various learning techniques for numerous drawback sorts. There is no "one size fits all" technique; this can be thanks to distinctive ability, limitations of every technique and nature of issues (12). Machine learning techniques square measure primarily hindered by scale of knowledge and computation in massive information. We'd cross-check some improved learning techniques employed in massive data context. A MapReduce framework for real time traffic processing was designed in a very study (13); the challenges were distributed process, distributed native learning and model fusion. The framework used Expectation Maximization formula for native learning method, the global model is compared to the distributed models so as to merge into a final model for traffic forecasting (13). Support Vector Machine was additionally improved with a probabilistic distribution for a large scale information application; the matter needed an energetic learning mechanism thanks to sample quality (14). Cross domain learning incorporating Support Vector Machine was used to develop a brand new and economical methodology for learning classification of pictures across domains, the method works despite a tiny low coaching information and completely different distributions. The strategy was evaluated against different cross domain ones on massive video datasets (15). Different scaled machine learning algorithms square measure quick learning for ranking (16) and distributed approximations through boosting (17). Machine learning square measure utilized by another set of techniques known as data processing in massive data issues. Data processing techniques square measure used for deducing data from information. There are used for agglomeration, regression and association rule mapping. Data processing techniques equally face procedure and information quality inherent in massive information. Sampling techniques are used to augment their weaknesses to meet big data challenges (11). Neural networks are one of the most used machine learning techniques.

Conclusion

This paper presents the most problems with massive knowledge. Organisations ought to affect huge quantity of knowledge, thus storage demand is the main issue. Co jointly the info is out there in several formats i.e. Unstructured and structured. Completely different methods should be adopted to handle unstructured knowledge. Extraction of helpful knowledge from huge quantity of knowledge is additionally an issue need to be considered. In step with security concern providing privacy, integrity, availableness, confidentiality etc. is additionally a difficulty; a lot of analysis is needed to be performed in all these areas.

References

1. Stats of Big Data Available at, https://www.youtube.com/watch?v=PlaJsseTgk4
2. Stats available at, http://www.statista.com
3. A study, http://www.computerworld.com/article/2513110/data-center/scientists-calculate-total-data-stored-to-date--295--exabytes.html
4. Ishwarappa, J. Anuradha, "A Brief Introduction on Big Data 5Vs Characteristics and Hadoop Technology", International Conference on Computer, Communication and Convergence (ICCC 2015), Volume 48, 2015, Pages 319–324.

5. Review Article, Big Data: Survey, Technologies, Opportunities, and Challenges, The Scientific World Journal, Volume 2014 (2014), Article ID 712826, available at, http://www.hindawi.com/journals/tswj/2014/712826/
6. Challenges of big data, available at, http://www.computerweekly.com/feature/Big-data-and-analytics-a-large-challenge-offering-great-opportunities
7. McKerlich R, Ives C, McGreal R. Measuring use and creation of open educational resources in higher education. *Int Rev Res Open Distance Learn*. 2013; 14(4): 90–103.
8. Gantz BJ, Reinsel D. Extracting Value from Chaos State of the Universe: An Executive Summary. *IDC iView*. 2011: 1–12.
9. Wu X, Zhu X, Wu G-Q, Ding W. Data mining with big data. *IEEE Trans Knowl Data Eng*. 2014; 26(1):97–107.
10. Chen M, Mao S, Liu Y. Big data: A survey. *Mob Networks Appl*. 2014; 19(2): 171–209.
11. Philip Chen CL, Zhang CY. Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. *Inf Sci*. 2014; 275: 314–347.
12. Steve O. *Machine Learning, Cognition, and Big Data*. CA Technologies. 2012
13. Chen C, Liu Z, Lin WH, Li S, Wang K. Distributed modeling in a mapreduce framework for datadriven traffic flow forecasting. *IEEE Trans Intell Transp Syst*. 2013; 14(1): 22–33.
14. Mitra P, Murthy CA, Pal SK. A Probabilistic Active Support Vector Learning Algorithm. *IEEE Trans Pattern Anal Mach Intell*. 2004; 26(3): 413–418.
15. Jiang W, Zavesky E, Chang SF, Loui A. *Cross-domain learning methods for high-level visual concept classification*. Proc - Int Conf Image Process ICIP. San Diego. 2008; 161–164.
16. Raykar VC, Duraiswami R, Krishnapuram B. A fast algorithm for learning a ranking function from large-scale data sets. *IEEE Trans Pattern Anal Mach Intell*. 2008; 30(7): 1158–1170.
17. Sun P, Yao X. Sparse approximation through boosting for learning large scale kernel machines. *IEEETrans Neural Netw*. 2010; 21(6): 883–894.