International Journal of Advanced Scientific Technologies ,Engineering and Management Sciences (IJASTEMS-ISSN: 2454-356X) Volume.3,Special Issue.1,March.2017

A Brief Survey on Big Data Analytics- Concepts, Challenges, Privacy Concerns and Future Scope.

[1]S.Sandhya kumari, [2]C.Sushma ,Academic Consultant, Dept. of Computer Science, SPMVV, Tirupati.
sridharamsandhya@gmail.com,c.sushma51@gmail.com

Abstract: In Today's environment, a huge repository of terabytes
data is generated everyday from various information systems and digital technologies such as Internet of Things and cloud computing. Analysis of these voluminous data requires a lot of efforts at different levels to extract knowledge and information for decision making.Because of globalization, emergence of social networks, more and more businesses are more interested in utilizing Big Data. Therefore, big data analysis is a current area of research which provides a platform to explore big data at various stages. The term "Big Data" refers to the bulk of data that cannot be handled with traditional data-handling techniques. We provide a brief survey study of 7V's in order to understand a big data. The existing data processing system is gradually decreasing whereas the capabilities of the new techniques for capturing, storing, visualizing and analyzing data are evolving has a given a confidence in the data engineering. The privacy concerns are raised due to unauthorized data extraction, collection and sharing of information about user. For privacy preserving, a general framework is discussed. Hence, this survey is to elucidate the study of Big Data concepts, challenges, Privacy concerns and it's Future Scope.

Keywords:
Big Data, 7V's, Data Visualization, Integration, Data Democratization, Privacy concerns, Privacy Preserving methods, Challenges, Future scope.

1. INTRODUCTION:

Big data is the term for a collection of data sets are complex and so large it becomes difficult to process using on-hand database management tool or traditional data processing applications. Big data is useful for the exponential growth and availability of data,both structured, unstructured and semi structured. Big data has new issues to consider such as discovery, iteration, mining, flexibility capacity, and predicting and decision management. IDC defines Big Data as a new generation of technologies and architectures which are designed to economically obtain value from very huge variety of data by enabling high-velocity capture, analysis and discovery. There are three main characteristics of Big Data: the data itself, the presentation of the output of the analytics and the analytics of the data. The digital universe consists of all kinds of data. The large amount of new data being generated is unstructured. This means that data is characterized that results in metadata. By 2020, more than 13,000 Exabyte's of the data in the digital universe will have Big Data value, but only if it is tagged and analyzed. Through data integration concept the IT departments may found their work load, skill sets and budgets over stretched by the need to manage terabytes to petabytes of data in a single dataset that delivers accurate value to business users.

II. THE 7V'S OF BIG DATA:

Big Data is a facilitator for value creation to remain competitive in a global environment by controlling data mining, ingestion and visualization of all the geophysical, seismic and related Exploration &Production data sets. Here we discussed the 7Vs' that describes the Big Data – volume, velocity, variety, veracity, virtual, variability and value.

2.1. V – VOLUME

Volume of Big data refers to the size of data is being created from all the sources including text, audio, video, social networking, research studies, medical data, space images, crime reports, weather forecasting and natural disasters, etc. [1]The Input data to big data systems could be chatter from social networks, web server logs, traffic flow sensors, satellite imagery, broadcast of audio streams,

banking transactions, MP3's of rock music, the content of web pages, scans of government documents, GPS trails, telemetry from automobiles, financial market data etc,. However such volume of data being, disorganized and unknown, cannot be handled or processed or queried with traditional ways. we are dealing with petabyte of unstructured data here. SQL based approach simply does not work.

2.2. V – VELOCITY

Velocity represents the speed or velocity of data that makes it too much to work with. The speed we are generating this kind of data using our Smartphone's and World Wide Web. Businesses must be focused with technology and database engines to process them of which the data coming in with this high velocity.That means that main thing is the speed of feedback loop that takes data from input through to the decision. Therefore, not only velocity to incoming data, that matters, but to stream the fast moving data into big storage for later processing and analysis.There are two important reasons for such data processing considerations. 1)The arrival of data is too fast to store input data.It needs some special analysis at time of data occurrence on the fly. 2) Application forces response to data as it arrives.

2.3. V – VARIETY

Data appears in many formats like Audio, video, text, images. This Creates the real complexity to the mix. That is why we can't specify it relational database any more. It is a great challenge to build a system so such data mix can be integrated into it directly. On the World Wide Web, people use different software's, browsers and they send data differently to the cloud. Not to ignore, most data is coming directly from real human interface and errors are unavoidable in data. Variety of data directly affects the integrity of data. In other words, more variety in data is more errors it will contain.

2.4. V – VERACITY

By Veracity, we mean the truthfulness of data.In other words, how certain we are about this data? Or how much data is the kind, it claims to be of that kind? We are discussing the meaningfulness of results from data for given problem space, people are working on or exploring. Now we are dealing with unstructured and bigdata here, which might be coming from Facebook posts, tweets, LinkedIn posts, etc. We cannot trust whatever the data we see out there and we may not realize that we cannot rely on it for our sales and business transactions. In my opinion this V is of highest concern to the processing of big data and related analysis and results outcome. Generally, we do normalization to our relational and traditional database to obtain the integrity of data and to avoid duplicates in data.Hence, the cleansing of big data with some great tools and algorithms are to be considered.

2.5. V – VALIDITY

Validity of data is similar to veracity of data. However, they are not the same concept but similar. By validity, we mean the correctness and accuracy of data with regard to the intended usage. Data may not have any veracity issues but may not be valid if not properly understood. Same set of data may be invalid for another application or usage and then valid for one application or usage. Even though, we are dealing with data where relationship may not be defined easily or in initial stages, but it is very important to verify relationship to some extent between elements of data, which we are dealing with to validate it against intended consumption, as possible. For example[2], Can a physician simply take a data from clinical trial that is related to patient's disease symptoms without validating them? The answer is No.

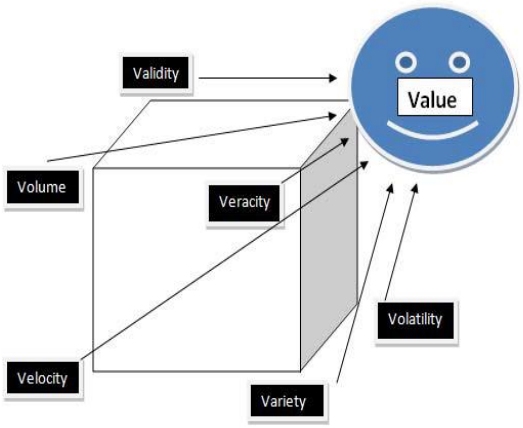
2.6. V – VOLATILITY

Speaking the topic of volatility of big data, we can easily recall the retention policy of structured data that we implement every day in our businesses. Once retention period expires, we can easily destroy it. For an example: an online ecommerce company may not want to keep a 1 year customer purchase history. Because after one year and default warranty on their product expires so there is no possibility of such data restore ever. Big data is no exception to this rule and policy in real world data storage. Such issue is very much magnified in big data world and not as easy as we have discussed with it in traditional data world. Big data retention period may exceed and security and strage may become expensive to implement.Volatility becomes significant due to Volume, Variety and Velocity of data.

2.7. SPECIAL V – VALUE

This V as Special for a reason. Comparing with other V's of big data, this V is the desired outcome of big data processing. We always put an eye to extract maximum value and true value from any big data set we are given to work with. Data value must exceed its cost or ownership or management. Every One must pay attention to the investment of storage for data. Storage may be cheaper and cost effective at time of purchase but such under investment

may affect highly valuable data, for example storing clinical trial data for new drug on cheap and unreliable storage may save money today but can put data on risk tomorrow[3]. Value of data greatly depends on governance mechanism as well. Writing of policies and structures that will eventually bring balance between reward and risk of the data. If these policies and structures are not carefully written and implemented may restrict businesses to extract true value of data. Hence, it will make data undervalued. With economy getting worse, IT budgets are being shrunk. Storage is always expensive. About 47 percent of IT budget to maintain IT infrastructure, 40 percent to information and transaction processing and about 13 percent to strategic IT investments[4]. Often data can move between various tiers. Higher the tier is, higher the value is. Data at higher tiers will have lower risk of catastrophe. Some organization can accept high cost with storage associated at higher tiers as protection is better guaranteed at those levels and thus value to cost ratio is higher[5][6].



7 V'S OF BIG DATA

III. DATA VISUALIZATION:

Data visualization is an important concept in big data. It is the presentation of data in a pictorial or graphical format. It allows users to see analytics presented visually, so that they can understand difficult concepts or identify new patterns.Giants like Google, Facebook,Twitter, EBay, Wal-Mart etc., adopted data visualization to handle the complex data very easily.Data visualization has shown good positive results in such business organizations.Advanced analytics can be integrated in the methods to support creation of interactive and animated graphics on mobile devices like smart phones, desktops, tablets and laptops. Implementing data analytics and data visualization, enterprises can finally begin to tap into the large potential that big data possesses and ensure greater return on investments and business stability.

IV. INTEGRATION:

The handling of big data is too complex. Data integration includes uncertainty of data Management, evaluate across data sources, retrieving data into a big data structure, getting useful information out of the big data, volume, skill availability, solution cost etc. Integrating digital capabilities in decision-making of an organization has changed the enterprises. By transforming the processes, such companies are developing agility, flexibility and precision that enables new growth. By making use of this concept, enterprises are finding solutions to gain an advantage of the data better either to increase revenues or to cut costs even if most of it is still focused on customer-centric outcomes using social and mobile technologies to make the people connect and interact with the organizations and incorporating big data analytics in this process is proving to be a benefit for organizations implementing it.

V. DATA DEMOCRATIZATION:

The current business scenario has brought forward several small and medium sized organizations who are trying to handle big data. Anyhow not all data can always be accessed. If we don't have thousands of engineers to work with big data then it is very difficult to find business solutions for the information. The ability to have real-time analytics is something that's becoming more prevalent, as is the ability to not just run a batch process for lot of hours on petabytes of data, but have a chart or a graph or some sort of report in real time. Interacting with it and making decisions at a time is becoming main stream. This often involves OLTP data that needs to run in memory or on hardware that's extremely fast, to create a data stream that can ingest all the different information that's coming in. When incorporated with attainable analytics capabilities from the onset, organizations are authorized with focus and the ability to reduce the time required knowing where opportunities, issues and risks locate in voluminous data.

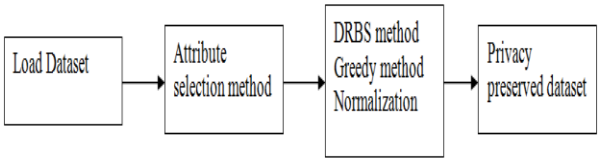
VI. PRIVACY CONCERNS:

The extraction of information of policies of organization have focused on the concerns of privacy of users. The huge amount of information coming from GPS, sensors, location trackers, clickstream, and log data can be treated as big data. Capturing and sharing such information may be the concern of users. While collecting the user related data there are large number of privacy pitfalls and

privacy related data is extracted in social media[7][8]. It is easy to identify or show the location of user from the tweets made by user. The basic machine learning and geotagged information is used for that and also[9] from geotagged twitter information the geographic coordinates can be extracted and it can be extended up to city of user or zipcode of location. In[10] proposed that image and structural analysis combined with content analysis on geotagged photos with textual tags collected from flicker can be used for finding location .The data analysis can be used for short term prediction. Some organizations cannot handle the volume, velocity and complexity of big data because they don't have capability to store the big data.This data which is produced at particular time and must be outsourced compulsorily. Based on demand the cloud service providers are providing scalable storage capability. while handing over this data to cloud service providers the privacy constraints should be applied at the same time. Variety characteristic of big data suggest that data comes in different formats such as images, videos, signals, instant messages. This unstructured and structured information may contain personally identifiable information and intellectual property. Such information capturing and sharing may leads to privacy violations. Many organizations use traditional databases as the main tools of handling data. The consumers have expressed deep concern about dishonesty among the businesses and misuse of personal information. So consumers are reluctant to give the correct information. Many consumers have taken actions such as turning off information collecting system such as location tracking feature.

7. PRIVACY PRESERVING METHODS

To apply the privacy preserving techniques we have to consider the different dimensions. In multidimensional dataset to find sensitive attributes, quasi identifiers and non sensitive attributes; different attribute selection methods should be applied[11]. These methods include Information Gain, Gain ratio, Pearson Correlation, Gini Index. After selection of key identifiers; these identifiers should be modified such that information will not be released to unauthorized user but at the same time utility of data will remain unchanged.



Privacy Preserving Flow graph

The methods available for perturbation of key identifiers are data relocation based sub clustering (DRBS), Greedy method, Normalization. In clustering based method; the clusters are found with centroid. Again clustering is applied to find subclusters. Then distance between the centroid of cluster and parent cluster is found and based on distance subclusters are arranged. The elements are rotated to neighbor cluster until last element is visited. In normalization method for perturbation the key identifier values normalized.

8. CHALLENGES

With Big Data, Users has to face sufficient attractive prospects and also some challenges. Such complications lie in storage, capturing data, searching, analysis, sharing and visualization. Big Data challenges include incompleteness and data inconsistency, scalability, timeliness and security. Data must be well constructed before doing data analysis. By understanding the method is to improve data quality and the analysis results, it is important to know which data can be preprocessed. Privacy is significant factor in outsourced data. Recently, some arguments have identified how some security agencies are using data generated by individuals for their personal benefits without permission.

9. FUTURE SCOPE AND DEVELOPMENT

In now a days, Big Data has got a significant role in IT industry than other technologies. The bulk data generated from sensor-enabled machines, mobile devices, cloud computing, social media, satellites help different companies improve their decision making and take their business to another level. "Big data absolutely has the potential to change the way academic institutions ,organizations, governments and conduct business and make discoveries, and its likely to change how everyone lives their day-to-day lives," - Susan Hauser, corporate vice president of Microsoft. As mentioned earlier in this paper, every day data is generated in such an increasing manner that, traditional database and other data storing system will gradually decreasing in storing, retrieving, and finding relationships among data. Big data technologies have found out the problems related to this new big data revolution through the use of commodity hardware and distribution. Companies like Google, General Electric, Yahoo!, Cornerstone, Microsoft, Kaggle , Amazon, Facebook, that are investing a lot in Big Data research and projects. IDC estimated the value of Big Data market to be ―about $ 6.8 billion in 2012 growing almost 40 percent every year to $17 billion by 2015.‖ By 2017, Wikibon's Jeff Kelly predicts the Big Data market will top $50 billion. There is high demand for big data strategies in companies to get exact solutions.

The problem is that the companies are not having good internal expertise and best practices. The side effect is that there are services and consulting boom in big data. It's a perfect storm of product and services says Wikibon's Jeff Kelly. Recently it was announced that, Big Data analytics technology is using in Indian Prime Minister's office in order to understand Indian citizen's thoughts and ideas through crowd sourcing platform www.mygov.in and social media to get a picture of common people's thought and opinion on government actions. Google is launching the Google Cloud Platform, which provides developers to develop a range of products from simple websites to complex applications. It enables users to project virtual machines, store bulk amount of data online, and plenty of other things [54]. Basically, it will be a one stop platform for cloud based applications, online gaming, mobile applications, etc. All these required huge amount of data processing where Big Data plays an immense role in data processing. The forecast from the IDC Future Scope for Big Data and Analytics are:

1. Visual data discovery tools will be growing 2.5 times faster than rest of the Business Intelligence (BI) market. By 2018, end-user self-service investment will become a significant requirement for all enterprises.
2. Over the next five years spending on cloud-based Big Data and analytics (BDA) solutions will grow three times faster than spending for on-premise solutions. Hybrid on/off premise deployments will become a significant requirement.
3. Shortage of skilled staff will persist. In the U.S. alone, the deep analytics roles in 2018 will be around 181,000 and five times that many positions requiring related skills in data management and interpretation.
4. By 2017, the foundation of BDA strategy is the unified data platform architecture. The unification will occur in different areas like information management, analysis, and search technology.
5.By 2015, Growth in applications incorporating advanced and predictive analytics, including machine learning, will accelerate. These apps will grow 65% faster than apps without anticipating functionality.6. By 2019,70% of large organizations already purchased external data and it will reach 100% also. In parallel more organizations will begin to fabricate their data by selling them or providing value-added content. 7.By 2015, Adoption of technology to continuously analyze streams of events will be accelerated as it is applied to Internet of Things (IoT) analytics, which is expected to grow at a five-year compound annual growth rate (CAGR) of 30%. 8. The idea of expanding the Decision management platforms at a CAGR of 60% through 2019 in response to the need for greater consistency in decision making and decision making process knowledge retention. 9. In 2015,the Rich media (video, audio, image) analytics will be tripled and emerge as the key driver for BDA technology investment. 10. By 2018, 50% of all consumers will be interacting with services based on cognitive computing on a regular basis.[56]

People wanted to digitize their lives hence Big data has become a significant technology. "People are walking sensors," said Nicholas Skytland, project manager at NASA within the Human Adaptation and Counter measures Division of the Space Life Sciences Directorate [57]. By the significant use of Big data, the market analyst and research firms took an average of all the figures and they concluded that approximately 15 percent of all IT organizations will move to cloud-based service platforms, and between 2015 and 2021, this service market is expected to grow about 35 percent.

CONCLUSION:

The literature Survey presents the fundamental concepts of Big Data. These concepts comprise the role of Big Data in the current environment of enterprise and technology. To augment the efficiency of data management, we have discussed the Data are also generated in different formats, which unfavorably affect data analysis, management, and storage. This deviation in data is accompanied by complexity and the development of further means of data acquisition. Data growth is promoting lot of technology innovation and creation. By understanding 7 Vs of Big Data, we can utilize its power for specific research and real world problems. As a result, additional research is obligatory for the efficient display, analysis, and storage of Big Data.Advancement in big data analytics is useful for drawing inferences; at the same time it is main reason for increasing privacy concerns of user.Framework for privacy preserving is discussed and challenges and how the big data will be used in different applications.

REFERENCES

[1] Edd Dumbill (O'Reilly Media), "Volume, Velocity, Variety: What You Need to Know About Big Data".
[2] Big Data Now: Current Perspectives from O'Reilly Radar.
[3] Judith Hurwitz, Alan Nugent, Fern Halper, Marcia Kaufman – Big Data for Dummies. ISBN: 978-1-118-50422-2
[4] Paul P. Tallon, Lyala Universtiy Maryland – Corporate Governance of Big Data: Perspectives on Value, Risk, and Cost, IEEE Computer Society 2013.
[5] P. Weill, S.L. Woerner, and H.A. Rubin, "Managing the IT Portfolio (Update Circa 2008): It's All about What's New," MIT Center for Information Systems Research (CISR), vol. 8, no. 2B, 2008, pp. 1-4.

[6] P.P. Tallon, R.V. Ramirez, and J.E. Short, "The Information Artifact in IT Governance: Towards a Theory of Information Governance," to appear in J. MIS, Loyola Univ. Maryland, 2013.

[7] Ferrara E, De Meo P, Fiumara G. Robert Baumgartner, Web data extraction, applications and techniques: A survey. Knowledge-Based Systems. 2014; 70:301–23.

[8] Hecht B, Hong L, Suh B, Chi E. Tweets from Justin Bieber's heart: the dynamics of the location field in user profiles. Proc International Conference on Human Factors in Computing Systems; 2011; British Columbia, Canada. pp. 237–46.

[9] Kinsella S, Murdock V, O'Hare N. I'm eating a sandwich in Glasgow: modeling locations with tweets. Proc International Workshop on Search and Mining User-generated Contents, ACM; 2011; Glasgow, UK. pp. 61–8.

[10] Crandall D, Backstrom L, Huttenlocher D, Kleinberg J. Mapping the world's photos. Proc 18th International Conference on World Wide Web, ACM; 2009; Madrid, Spain. pp. 761–70.

[11] Sudha M, Kumaravel A. Performance Comparison based on Attribute Selection Tools for Data Mining. Indian Journal of Science and Technology. 2014 Nov; 7(S7):1–5.