REVIEW OF BIG DATA SCIENCE AND ITS RELATION WITH CLOUD TECHNOLOGY

Ram Milan, Dr. Harisingh Gour Vishwavidyalaya, Sagar India (rammilan.in@gmail.com)
Kamlesh Kumar Pandey Dr. Harisingh Gour Vishwavidyalaya, Sagar India (kamleshamk@gmail.com)
Diwakar Shukla , Dr. Harisingh Gour Vishwavidyalaya, Sagar India (diwakarshukla@rediffmail.com)

ABSTRACT

Big data science is the new emerging term in the field of computer technology. During the emergence of computer science the database is used to store the data but with the emergence of social networking sites the data is generated exponentially that is not able to store in the conventional method of storage so the new technology is derived that is used to store data in cloud technology. Cloud technology is a technology that is used to store data in a safe and secure manner. And it also gives the opportunity to use the data when required. We can mine the data with data mining algorithm and take decisions on the basis of that data.

KEYWORDS: *Cloud Technology, Big Data, Virtualization.*

I. INTRODUCTION

Big data science is a science related to big data that is used to deal with the big data. Nowadays big data science is a new term that is becoming popular because we are leaving in a digital world where there is lot of networking social sites is present like Face book, Twitter, LinkedIn, Instagram etc where the members of these sites are uploading their status and also upload their photos, and friends of these members also comments on their status or they are liking their photos. In digital world many people send good morning messages to their friends and their loved ones or their bosses so lots of data is generated by every second and this data is beyond the normal limit that is not able to process by the conventional set of processors or not able to store the data in the conventional storage of media. These types of conversations generate the big data. Following statistics shows the use of social media in daily life- Google has more than 1 billion queries per day, Twitter has more than 250 tweets per day, Face book has more than 800 millions updates per day, and YouTube has more than 4 billion views per day. The data produced nowadays is estimated in the order of Zettabytes, and is growing around 40% every year by Fan Wei et al. (2012).

Big data is the collection of unstructured, semi structured and structured data whose volume, complexity, velocity, veracity make them difficult to process in the normal processing devices. Varieties of data can be there like text, video, audio, images, graphics etc. The sources of these data are also many by Jha Anupama et al.(2016).

Cloud technology- In cloud technology a service provider lends the services of IT to the users like (software, hardware devices, storage, network etc.). Cloud technology is related to Virtualization. Virtualization is managed by the hiding lower level details and make lower levels of hardware are unknown to higher level details. This makes portability of higher level functions and sharing of hardware resources. Figure 1 shows the Cloud Technology where the data of mobile phones, personal computer, laptops and Remote sensing devices data are stored in cloud databases.

Relation between Big Data and Cloud Technology:- Big data generates large amount of data every second and it is both structured, unstructured and semi structured and these large amount of data must be store some storage device so we need a cloud technology where we can store these excessive amount of data and when we need to access these data we can access it.

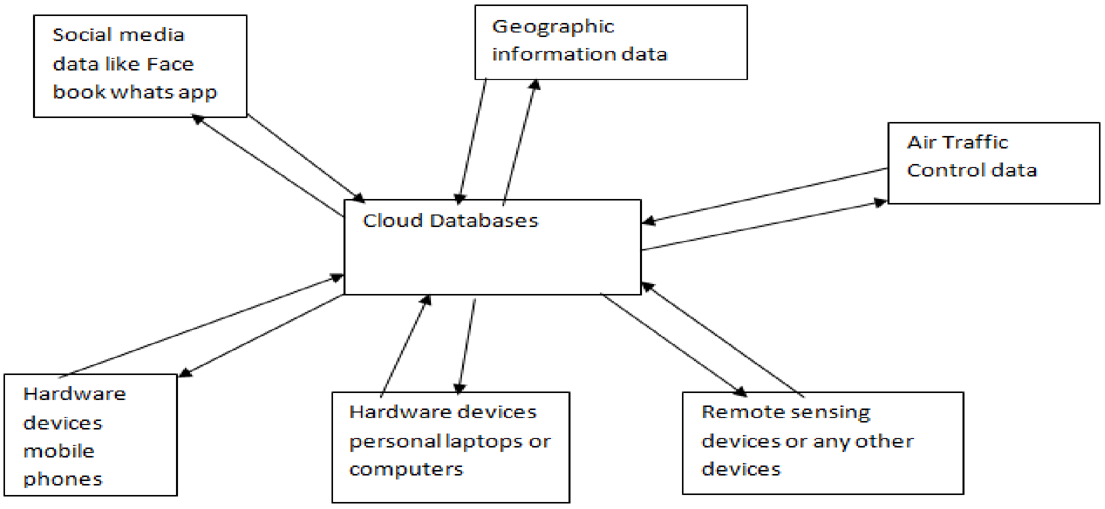
Social media data like Face book whats app
Geographic information data
Air Traffic Control data
Cloud Databases
Hardware devices mobile phones
Hardware devices personal laptops or computers
Remote sensing devices or any other devices

Figure 1: Cloud Technology

II. Literature Review

Related to Big Data- Jha A. et al (2016) describes that big data is a new concept that describes emerging techniques and technologies to analyze large volume of complex datatsets that are exponentially growing from various sources and with various rates.

Dobre and Xhafa (2014) report that every day the world generates around 2.5 quintillion bytes of data, with 90% of these data generated is unstructured .Gantz and Reinsel (2012) assert that 2020, over 40 Zettabytes of data will have been generated, imitated, and consumed. With these large amount of data is generated exponentially anytime, anywhere, and any device there is an era of big data that is referred to as Data Deluge by Sivarajah Uthayasankar (2017).

Five Exabyte (10^{18} bytes) of data were generated by human until 2003. Nowadays this amount of data is generated in two days. In 2012, digital world of data was expanded to 2.72 Zettabytes (10^{21} bytes). It is assumed to be double every two year. IBM predicted that every day 2.5 Exabyte's of data created also 90% of the data produced in last two years. Face book has 955 million monthly active accounts using 70 languages, 140 photos uploaded, 125 billion friend connections, every day 30 billion pieces of content and 2.7 billion likes and comments have been posted. Every minute, 48 hours of video are uploaded and every day 4 billion views performed on YouTube. 1 billion Tweets every 72 hours from more than 140 million active users on Twitter. 571 new websites are created every minute of the day. Within the next decade, number of information will increase by 50 times by SAGIROGLU (2017)

Uthayasankar S. et al. (2017) gives the properties of big data like Volume, Veracity, Velocity, Variability, Visualization, and Value.

1. Volume (large data sets like in terabytes, petabytes, Zettabytes, Exabyte's or even more) large sets of data is generated by the social networking sites like Face book, Twitter etc. face book generates over 500 terabytes of data, and Wal-Mart collects more than 2.5 petabytes of data every hour form its customer transactions.

2. Veracity (eg. Complex data structure, anonymities, imprecision or inconsistency in large data sets): the data collected from the different customer sites is not very much clear it is not consistent. To mining the data it must be consistent, we have to apply some techniques to make it consistent.

3. Velocity (high rate of flow of data with non homogenous structure): for instance Wal-Mart processes more than a million transactions each hour. These data also provides information about customer buying behavior and patterns. And these patterns can be used for taking decisions for improving business.

4. Variability (data whose meaning is changing very rapidly): variability of data is also used in sentiment analysis. For example the same tweets can have different meaning. And for analyzing these tweets we have to analyze the contexts of the tweets. And this is a very challenging task.

5. Visualization (shows the data which is readable) : presenting the data in the way which is more readable and easily understandable in the manner of graphs and figures. For example eBay has millions of customers and eBay wants to represent these customers in the graph so eBay has developed the BD visualization tool- tableau, which is used to transforming large and complex datasets into spontaneous depictions.

6. Value (extracting knowledge/ value from the large amount of structured and unstructured data without loss:

Related to Cloud Technology- Vouk A. Mladen (2008) is a new term, related to Virtualization, distributed computing, high speed network and utility computing and software services **.**Cloud computing stands up on the cyber infrastructure and made up on research in Virtualization, distributed computing, grid computing, utility computing, high s peed network, web and software services. It means that it is service oriented architecture, reduced infrastructure cost, greater flexibility, and provide on demand services and many other things by Fan Wei et al. (2012) and Spoorthy et al (2014) . Divides the cloud computing in 3 types of services**.**

1. Software as a Service (SaaS) is the most popular form of cloud computing and it is easy to use. SaaS provides the services to the clients with the help of third party vendor and whose interface is used by the client's side. SaaS does not require the need to install and run applications on individual computers. With the help of SaaS it is easy for enterprises to provide maintenance and support, because everything can be managed by third party vendors: applications, runtime, data, middleware, o/s, virtualization, servers, storage, and networking. Examples of SaaS are Google Apps, Microsoft Office 365, Google+, Face book, Yahoo, Gmail etc.

2. Platform as a Service (PaaS) provides the computational resources through a platform. PaaS provides the testing; development and deployment of applications simple, quick and cost effective and it eliminate the need of money to buy the layers of hardware and software. With PaaS vendors manage runtime, middleware, O/S, Virtualization, servers, networking and storage but users manage applications and data. Examples of PaaS are AWS Elastic Beanstalk, Heroku, Windows Azure and Google App Engine.

3. Infrastructure as a Service (IaaS) provides computer infrastructure (such as a platform virtualization), networking and storage. Instead of buying servers, software, or network equipment, users can buy these as a outsourced service that is usually billed according to the amount of resources consumed. For the payment of some rental fees, a third party allows you to install a virtual server on their IT infrastructure. Compared to PaaS, SaaS and IaaS users are responsible for managing: application, data, runtime, middleware, and O/S. Vendors still are responsible for manage virtualization, servers, and storage, hard drives, and networking. Examples of IaaS are Windows Azure, Amazon EC2, Google Compute Engine D and Rack Space.

4. Storage as a Service (StaaS), it provides cloud applications to scale beyond their limited servers. Staas provides users to store their data at remote disks and access them anytime form any place. Cloud storage are used to maintain user's data and information, including high availability, reliability, performance, replication and data consistency.

5. Amazon S3 is used for storage for internet. It is designed to make web computing easier for developers. Amazon S3 provides a simple web services interface can be used to store and retrieve any amount of data, at any time, from anywhere on the web.

III. CHALLENGES OF BIG DATA IN IMPLEMENTATION WITH CLOUD COMPUTING

Large and massive data is gone now we are taking about the enormous and very huge data and I have no word to define that term this much of data is generated every moment in a world. How to store that data is a real and first challenge in big data. And second most challenge is how to process that data and how to take the fruitful information to solve the real life problems. These are the real challenges of the big data related to cloud technology. Third challenge is the affordability that means if a company wants to store all data in own ownership so that company has to invest large amount of money in that purpose.

Proposed solution –store this data in the form of clustering technique where similar data is grouped together to store in a one location so that processing time is also reduced and storage area is also minimized. And for searching we use the hashing technique so the time complexity is also minimized. Solution of affordability is via the cloud technology where the cloud technology companies have the infrastructure and they rent the infrastructure to the serving company with minimal cost.

IV.CONCLUSION

Today's generations is the digital generation we are living in the digital age where most of the persons want to be connected with the internet in the form of face book, Twitter, We- Chat and other social networking sites and with these connections users are generated large amount of data and that data is known as Big Data, because that data is not able to process with the normal processing of computer we require advanced processing techniques and more storage space to store these exponential data. For storing these amounts of data we require cloud technology to store data. Cloud technology is a rental service that is provided by the third party to the client and they have to pay according to the use cloud technology is divided into 3 categories like PaaS (Platform as a service), IaaS (Information as a service), SaaS (Software as a service) and Storage as a service. But using cloud technology we have to put more focus on security and privacy of Big data because data can be compromised. We have to use encryption technique to secure the data.

REFERENCES

Vouk A. Mladen (2008). Cloud Computing - Issues research and implementation, Journal of Computing and information technology-CIT 16 (2008).

Fan Wei et al. (2012). Mining Big Data: Current Status, and Forecast to the Future, SIGKDD Explorations, 14(2).

Jha A. et al. (2016) , A Review on the Study and Analysis of Big Data using Data Mining Techniques. International Journal of Latest Trends in Engineering and Technology (IJLTET), 6(3).

Uthayasankar S. (2007). Critical Analysis of Big Bata challenges and analytical methods, Journal of Business of Research.

SAGIROGLU S. (2013). Big Data: A Review, IEEE.

Spoorthy V. et al. (2014). A Survey on Data Storage and Security in Cloud Computing, International Journal of Computer Science and Mobile Computing,*306 – 313.*

Kumar Santosh et al. (2012). Cloud computing- Research Issues, Challenges, Architecture, Platforms and Applications: A Survey, International Journal of Future Computer and Communication, 1(4).