INTERNATIONAL JOURNAL OF ADVANCE RESEARCH, IDEAS AND INNOVATIONS IN TECHNOLOGY

ISSN: 2454-132X
Impact factor: 4.295
(Volume3, Issue6)
Available online at www.ijariit.com

Analysis on Big Data Concepts and Applications

Sandhiya .B
Sri Shakthi Institute of Engineering and Technology, Coimbatore, Tamil Nadu
sandhiyab@siet.ac.in

Ramya Devi .N
Sri Shakthi Institute of Engineering and Technology, Coimbatore, Tamil Nadu
ramyadevi@siet.ac.in

A. Deepalakshmi
Sri Shakthi Institute of Engineering and Technology, Coimbatore, Tamil Nadu
deepalakshmi@siet.ac.in

Abstract: ***The term, Big Data 'has been referred as a large amount of data that cannot be handled by traditional database systems. It consists of large volumes of data which is been generated at a very fast rate, these cannot be handled and processed by traditional data management tools, so it requires a new set of tools or frameworks to handle these types of data. Big data works under V's namely Volume, Velocity, and Variety. Volume refers to the size of the data whereas Velocity refers to the speed that the data is being generated. Variety refers to different formats of data that is generated. Mostly in today's world thee average volumes of unstructured data like audio, video, image, sensor data etc. One can get these types of data through social media, enterprise data, and Transactional data. Through Big data analytics, one can able to examine large data sets containing a variety of data types. Primary goals of big data analytics are to help the organizations to take important decisions by appointing data scientists and other analytics professionals to analyses large volumes of data. Challenges one can face during large volume of data, especially machine-generated data, is exploding, how fast that data is growing every year, with new sources of data that are emerging. Through the article, the authors intend to decipher the notions in an intelligible manner embodying in text several use-cases and illustrations.***

Keywords: Big Data, 3 V's, Sentiment Analysis, Organisational Decisions.

I. INTRODUCTION

Every day, we create 2.5 quintillion bytes of data — so much that 90% of the data in the world today has been created in the last two years alone. This data comes from everywhere: sensors used to gather climate information, posts to social media sites, digital pictures and videos, purchase transaction records, and cell phone GPS signals to name a few. Such colossal amount of data that is being produced continuously is what can be coined as Big Data. Big Data decodes previously untouched data to derive new insight that gets integrated into business operations. However, as the amounts of data increase exponential, the current techniques are becoming obsolete. Dealing with Big Data requires comprehensive coding skills, domain knowledge, and statistics.

Despite being Herculean in nature, Big Data applications are almost ubiquitous- from marketing to scientific research to customer interests and so on. We can witness Big Data in action almost everywhere today. From Facebook which handles over 40 billion photos from its user base to CERN's Large Hydron Collider (LHC) which generates 15PB a year to Walmart which handles more than 1 billion customer transactions in an hour. Over a year ago, the World Bank organized the first WBG Big Data Innovation Challenge which brought forward several unique ideas applying Big Data such as big data to predict poverty and for climate smart agriculture and fore user- focused Identification of Road Infrastructure Condition and safety and so on.
Big Data can be simply defined by explaining the 3V's – volume, velocity, and variety which are the driving dimensions of Big Data quantification. Gartner analyst, Doug Laney introduced the famous 3 V's concept in his 2001 Metagroup publication, 3D data management: Controlling Data Volume, Variety, and Velocity'.

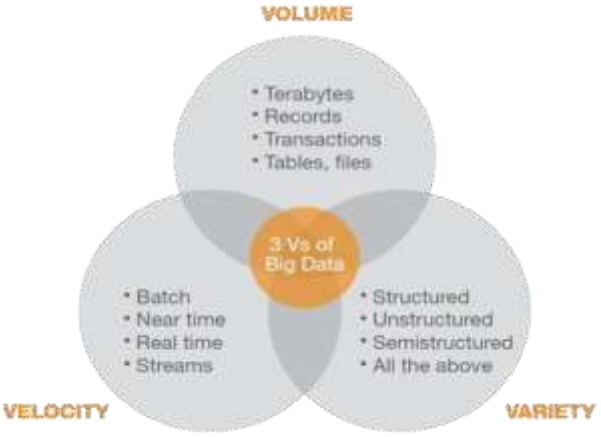

Image-1: schematic representation of the 3V's of Big Data

a. Volume: This essentially concerns the large quantities of data that is generated continuously. Initially storing such data was problematic because of high storage costs. However, with decreasing storage costs, this problem has been kept somewhat at bay as of now. However, this is only a temporary solution and better technology needs to be developed. Smartphones, E-Commerce, and social networking websites are examples where massive amounts of data are being generated. This data can easily distinguish between structured data, unstructured data, and semi-structured data.

b. Velocity: Velocity refers to the speed at which the data is being generated. In different fields and different areas of technology, we see data getting generated at different speeds. A few examples include trading/stocks at present times, the speed at which such colossal amounts of data are being generated is unbelievably high. Take Facebook for example – it generates 2.7 billion like actions/day and 300 million photos amongst others roughly amounting to 2.5 million pieces of content in each day while Google Now processes over 1.2 trillion searches per year worldwide.

c. Variety: It refers to the different formats in which the data is being generated. Different applications generate he data in different formats. Apart from the traditional flat files, spreadsheets, relational databases etc. We have a lot of unstructured data stored in the form of images, audio files, spreadsheets, relational databases etc., Implementing Big Data is a mammoth task given the large volume, velocity, and variety. Big Data is a term encompassing the use of techniques to capture, process, analyze and visualize potentially large datasets in a reasonable timeframe not accessible to standard IT technologies. By extension, the platform, tools , and software used for this purpose are collectively called ―Big Data technologies‖. Currently, the most commonly implemented technology is Hadoop. Hadoop is the culmination of several other technologies like Hadoop Distribution File Systems, Pig, Hive , and HBase. Etc. However, even Hadoop or other existing techniques will be highly incapable of dealing with the complexities of Big Data in the near future. The following are few cases where standard processing approaches to problems will fail due to Big Data-

Large Synoptic Survey Telescope (LSST): –Over 30 thousands gigabytes (30TB) of images will be generated every night during the decade –long LSST survey sky.

There is a corollary to Parkinson's Law that states:

Data expands to fill the space available for storage.

This is no longer true since the data being generated will soon exceed all available storage space.

72 hours of video are uploaded to YouTube every minute.

There are at present two general approaches to big data:

a. Divide and Conquer using Hadoop: The huge data set is broken into smaller parts and processed in a parallel fashion using many servers.
b. Brute Force using technology from the likes of SAP HANA: One very powerful server with massive storage is used to compress the data set into a single unit.

II. APPLICATIONS

Big Data is slowly becoming ubiquitous. Every arena of business, health or general living standards now can implement big data analytics. To put simply, Big Data is a field which can be used in any zone whatsoever given that this large quantity of data can be harnessed to one's advantage. The major applications of Big Data have been listed below.

The Third Eye- Data Visualization

Organizations worldwide are slowly and perpetually recognizing the importance of big data analytics. From predicting customer purchasing behavior patterns to influencing them to make purchases to detecting fraud and misuse which until very recently used to be an incomprehensible task for most companies big data analytics is a one-stop solution. Visualizing your data can be fun and challenging. If you are working with big data, it is easier to understand

Integration- An exigency of the 21st century

Integrating digital capabilities in the decision-making of an organization is transforming enterprises. By transforming the processes, such companies are developing agility, flexibility, and precision that enables new growth. Gartner described the confluence of mobile devices, social networks, cloud services and big data analytics as the as the nexus of forces. Using social and mobile technologies to alter the way people connect and interact with the organizations and incorporating big data analytics into this process is proving to be a boon for organizations implementing it. Using this concept, enterprises are finding ways to leverage the data better either to increase revenues or to cut costs even if most of it is still focused on customer-centric outcomes. Such customer-centric objectives may still be the primary concern of most companies, a gradual shift to integrating big data technologies into the background operations and internal processes.

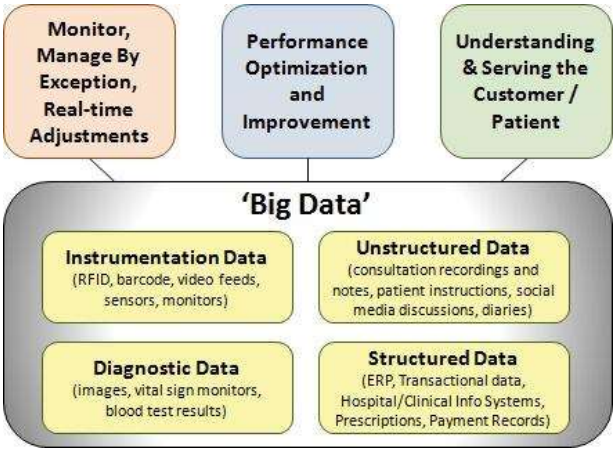

Image-2: Types of Data handled on Big Data

Big Data in Medical field

Healthcare is one of those arenas in which Big Data ought to have the maximum social impact. Right from the diagnosis of potential health hazards in an individual to complex medical research, big data is present in all aspects of it. Devices such as the Fitbit, Jawbone and the Samsung Gear Fit allow the user to track and upload data. Soon enough such data will be compiled and made available to doctors, which will aid them in the diagnosis. Several partnerships like the Pittsburgh Health Data Alliance have been established. The Pittsburgh Health Data Alliance is a collaboration of the Carnegie Mellon University, University of Pittsburgh and the UPMC. In their website, they state , ―The health care field generates an enormous amount of data every day. There is a need, and opportunity, to mine this data and provide it to the medical researchers and practitioners who can put it to work in real life, to benefit real people……The solutions we develop will be focused on preventing the onset of disease, improving diagnosis and enhancing the quality of care…….Further, there is the potential to lower health care costs, one of the greatest challenges facing our nation. And the Alliance will also drive economic growth in Pittsburgh, attract hundreds of companies and entrepreneurs, and generate thousands of jobs, from around the world…‖The patients diagnosis will be analyzed and compared with the symptoms of others to discover patterns and ensure better treatment. IBM has taken initiative on a large scale to implement big data in healthcare systems be in its collaboration with healthcare giant Fletcher Allen or with the Premier healthcare alliance to change the way unstructured but useful clinical data is made available to more medical practitioners so as to improve population health. Big Data can also be used in major clinical trials like cure for various forms of cancer and developing tailor-made medicines for individual patients according to their genetic makeup. To summarize, Sunder Ram of Oracle stated, ―Big Data solutions can help the industry acquire organize & analyze this data to optimize resource allocation, plug inefficiencies, reduce cost of treatment, improve access to healthcare & advance medicinal research.‖

Big Data and the World of Finance:

Big Data can be a very useful tool in analyzing the incredibly complex stock market moves and aid in making global financial decisions. For example, intelligent and extensive analysis of the big data available on Google

Trends can aid in forecasting the stock market. Though this is not a fool-proof method, it definitely is an advancement in the field. A research study by the

Warwick Business School drew on records from Google,

Wikipedia and Amazon Mechanical Trunk in the time period of 2004-2012 and analyzed the link between Internets searches on politics or business and stock market moves. In the paper, the author states, ―We draw on data from Google and Wikipedia, as well as Amazon Mechanical Turk. Our results are in line with the intriguing possibility that changes in online information-gathering behavior relating to both politics and business were historically linked to subsequent stock market moves….Our results provide evidence that for complex events such as large financial market moves, valuable information may be contained in search engine data for keywords with less-obvious semantic connections to the event in question. Overall, we find that increases in searches for information about political issues and business tended to be followed by stock market falls.‖

Big Data is also being implemented in a field called Quantitative Investing 'where data scientists with negligible financial training are trying to incorporate computing power into predicting securities prices by drawing ideas from sources like newswires, earning reports, weather bulletins, Facebook and Twitter

Image-3: Wall Street Journal summarizes the above concept.

One very interesting avenue of using Big Data in finance is the sentiment extraction from news articles. Market sentiment refers to the irrational belief in investors about cash-flow returns. The Heston-Sinha's Application of the Machine Learning algorithm provides us with the probability of an article being _positive', _negative 'and neutral'using two other popular methods, one being with the use of the Harvard IV Dictionary.
In general, big data is set to revolutionize the landscape of Finance and Economy. Several financial institutions are adopting big data policies in order to gain a competitive edge. Complex algorithms are being developed to execute trades through all the structured and unstructured data gained from the sources. The methods adopted so far has not been completely adept, however, extensive research ensures growing dependence of the stock markets, financial organizations and economies on big data analytics.

Big Data in Fraud Detection

Forensic Data Analytics or FDA has been an intriguing area of interest in the past decade. However, very few companies are actually using FDA to mine big data. The reasons for this unfortunate situation vary from the deficit of expertise and awareness, developing the right tools to mine big data to lack of appropriate technology and inability to handle such humungous quantities of data. Now-a-days financial industries are approaching data analytics for fraud detection from 2014, so this helps to prevent fraud becomes clearer, to make business dcase for data analytics.

A 10 step approach has been suggested by Infosys to implement analytics for fraud detection:

1. Perform SWOT analysis of existing fraud detecting paradigms.
2. Assign a dedicated fraud management team.
3. Developing or purchasing appropriate data analytics software.
4. Integrate siloed data and clear inefficiencies in the processes.
5. Establish rules relevant business obligations.
6. Determine thresholds for detection of error or discrepancies.
7. Implement predictive analysis to determine potential discrepancies and frauds.
8. Use Social Network Analysis or SNA to determine fraudulent activities.
9. Develop an integrated case management system.
10. Continue with extensive research to integrate existing systems of fraud detection with a new set of techniques developed.

Big Data in Sentiment Analysis

Sentiment Analysis is by far the most extensively used application of big data. Presently, zillions of conversations are occurring on the social media, which when harnessed to one's advantage can aid any company in determining new patterns, protecting their brand image and segmenting consumer base to improve product marketing and the overall customer experience. Whenever one collecting the data from various sources, we need to set up a Hadoop cluster and Hadoop distributed file system to store one's data, the best sentiment analysis includes:
a) Twitter feed
b) RSS fees
c) A mobile application.

Twitter, a micro logic site ha number of API's through that one can able to retrieve and manipulate the tweets. First of all, one wants to implement twitter oath framework. With the help of this framework, one can able to log in to twitter website. Big data tools can generate unbiased tools insight into generated data from any source for proper decision making, one can readily realize their return on investment by implementing big data tools.

Developing tools for efficient sentiment analysis.
IBM has developed IBM Social Media Analytics which is a powerful SaaS solution. It captures structured and unstructured data from social networking sites to develop a comprehensive understanding of attitudes, opinions, and trends. It then applies tools of predictive analysis to determine customer behavior and improve customer experience. This can aid the company to create personalized campaigns and promotion to increase the consumer base. It has presented their framework as the following:

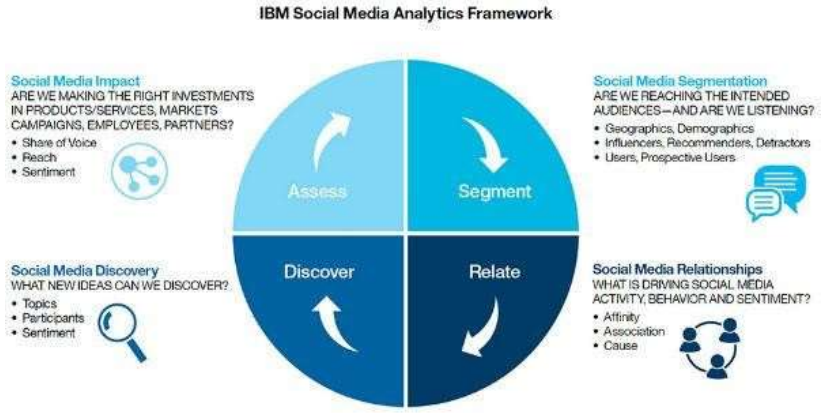

Image-4: IBM's Social Media Analytics [29] framework

Similarly, SAP has developed an SAP-HANA based application known as Social Contact Intelligence which monitors and develops insights from social media at real-time, determines the primary influencers thus determining new opportunities and improving the overall customer satisfaction.

Big Data in Food Industry:
The impact of Big Data on the food industry is increasing exponentially. Be it for tracking the quality of products or presenting recommendations to the customer or developing marketing strategies for better customer experience, the presence of Big Data analytics in the food industry is slowly becoming ubiquitous.
IBM collaborated with The Cheesecake Factory to analyze structured data like restaurant's location and unstructured data such as flavours to increase customer satisfaction. In a news article, it stated, –N2N has teamed up with IBM to provide The Cheesecake Factory with a technology that can communicate critical supply chain data instantly, so thousands of food items won't need to be recalled and tested. Nardone said they have initiated a conversation with the Centers for Disease Control and Prevention, as it may be easier to track the culprit if a food-related scandal occurs.‖
Similarly, apps such as the Food Genius [33] applies big data to predict specific recommendations to the customers. The company accumulates menu-level data parsed with ingredients, preparation methods, spices etc. and then analyzes them with individual customer preferences to determine trends and aid food giants make marketing strategies. Companies such as Starbucks, Dominos, and Subway take advantage of big data analytics to track individual customer preferences and present customers with personalized offers so as to increase customer base and improve customer satisfaction.

Big Data for the Telecom Industry
In order to improve customer service and satisfaction, concepts of Big Data and Machine Learning are being progressively implemented. Call detail records, web and customer service logs, emails to social media as well as geospatial and weather data are the few examples of data being accessible to telecom operators. Handling such massive amounts of data can be a daunting task.

Developing deep insights with the aid of Machine Language running on Apache Hadoop can help operators to economically take advantage of the ever-increasing datasets so as to enhance their quality of service and customer experience as well as to increase the customer base with ad targeting and promotions and reduce the operational costs. The benefits of using such technologies are immense. Predictive maintenance ensures that operational disruptions are predicted, prevented and recovered. Real-time processed data can be used to dynamically allocate the bandwidth to reduce congestion and outages.

III.BARRIERS IN IMPLEMENTING BIG DATA

In the 1990's Big Data became a hyped topic of interest in the world of distributed systems when the rapidly increasing impact of the world-wide Web and the exponential growth of the content. None of the then available resources were sufficient or cost-efficient to handle this task. At the turn of the millennium, in response to this issue, Google created the Google File System (GFS) which provided consumers with OS-level byte stream operations on data spanning several machines in clusters using rather expensive hardware. Later, Google developed the Map Reduce paradigm which was identical to the partitioned parallelism used in shred- nothing parallel query processing. Following this trend, multi-national giants like Yahoo and Facebook developed their own software. Yahoo! Developed Pig while Hive was developed by Facebook, Jaql by IBM and Dryad and Scope by Microsoft.

Practical Challenges facing Big Data

Despite the extensive hype around Big Data in the industry today, very few companies have actually been able to implement the concept of Big Data. A survey published in

2013 by SAS (_2013 Big Data Survey Research Brief') analyzed the reasons for why most industries are still delaying or refusing to pursue a big data strategy. It states, -A little more than one-fifth of the respondents is still trying to learn more about big data, while others are still trying to understand the benefits of big data. Even though the industry has written countless articles, blogs and white papers about big data, there is still a significant contingent of data management professionals trying to understand the basics.‖

The obstacles that limit the implementation of big data by any industry are aplenty. The _Big Data Talent Gap'which distinctively exists even though a lot of research has gone into this field in the past decade is a massive issue.

The following visual aid further explains the situation.

Image 5- what is the primary reason your organization is not considering or exploring the use of external big data to help make business decisions.

There are several big data experts, however, most of their expertise is limited to the implementation of one paradigm (usually one using the applications in Hadoop) rather than big data management skills. Most of these data scientists continue to remain oblivious to the practical zones of data- handling. A report from 2012 stated the following:

Gartner analysts predicted that by 2015, 4.4 million IT jobs globally will be created to support big data with 1.9 million of those jobs in the United States. However, while the jobs will be created, there is no assurance that there will be employees to fill those positions.‖

There are, at this present moment a diverse variety of tools available that are available in the market to implement operational and analytical processing of big data. Most of these are lumped together into a category called NoSql. A survey held in 2014 summarizes the data management options available.

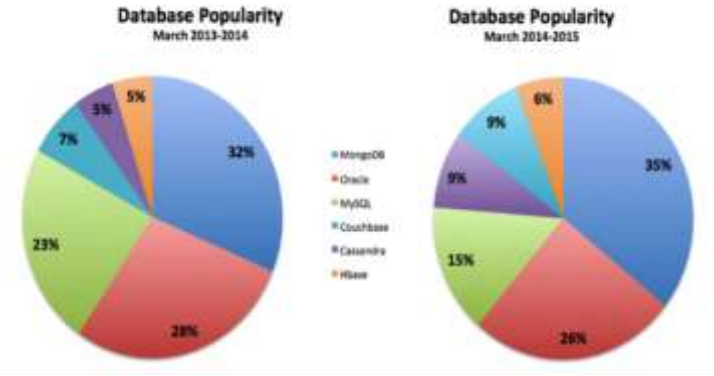

Image 6- Current adoption of relational database technologies with projected two year growth.

Such varied options have created a sense of confusion amongst the industry data experts making it difficult for them to zero in on one particular strategy. Choosing an appropriate Big Data Platform is a very complex task given the immeasurable amount of data that needs to be accessed, transmitted and delivered from the numerous sources and then accumulated in data sets. Finally, synchronization of such vast quantities of data coming from numerous sources with its originating systems is one massive job as rampant inconsistencies and asynchrony in the big data environment can have a disastrous effect.

One of the crucial practical challenges faced by Big Data is the cost implications of it. Even though the implementation of Big Data analytics has been there for about a decade now, the cost implications of storing such humungous amounts of data still remains a matter of serious concern. It is not only the quantity of data but also the complex processing techniques which make

its applications so expensive. An article by Forbes [45] states, ―A Petabyte Hadoop cluster will require between 125 and 250 nodes which costs ~$1 million. The cost of a supported Hadoop distribution will have similar annual costs (~$4,000 per node), which is a small fraction of an enterprise data warehouse ($10-$100s of millions).‖ This tumultuous situation requires that new technologies and algorithms be developed that will ensure that the financial challenges that face big data analytics today are made minimal so that an increasing number of enterprises can implement big data analytics in their regular operations.

Data Democratization: The present business scenario has brought forward several small and medium sized organizations who are trying to harness Big Data. However, not all data can always be accessed. As said by Paul Kent, the vice-president of Big Data with SAS, ―So if you're not Google or LinkedIn or Facebook, and you don't have thousands of engineers to work with Big Data, it can be difficult to find business answers in the information‖. In an IDG Research study, it was discovered that amongst all the organizations who claim to be effective at Big Data analysis, only about 58% have already implemented or in the process of implementing a data visualization solution while another 40% have concrete plans of implementing them. Tammi Kay George, the manager of R&D Program
& Project Management at SAS concisely summarizes the whole concept, ―A crucial element in minimizing the amount of time needed to understand data, visualization tools are imperative [to] realizing the value from a Big Data initiative, When incorporated with approachable analytics capabilities from the onset, organizations are empowered with focus and the ability to reduce the time required to know where opportunities, issues, and risks reside in voluminous data.‖

Encryption- Securing Big Data: With such massive amounts of data being generated, ensuring that the data doesn't fall at risk is quintessential. Such data left unsecured may put organizations or the general human race at risk. Sans the correct security solutions and encryption techniques, Big Data can imply big problems. The characteristics which make Big Data valuable to the market also make it valuable to various anti-social elements like cyber criminals. The number of encryption techniques available is aplenty. However, they mostly tackle one specific aspect and this is what makes it challenging. To make it easier to understand, one could consider a certain transparent encryption technique that is provided by a certain database vendor. They might be applicable to a particular database nut may not be suitable for implementation in a big data platform. There are a few organizations that offer encryption technology implementable on big data. However, most of the times they can only ensure the security of specific big data nodes and does not protect the original data that is fed into the big data platform. With such incompatible approaches in securing Big Data, IT industry has to make do with fragmented key and policy management, which increases administrative effort and makes it almost impossible to apply them consistently. Though several large organizations are taking their own initiatives to protect the data that they are generated, a mass awareness of the implications of unsecured data need to be initiated and smaller organizations need to step up to ensure that the world is a safe place for the data to reside

Theoretical Challenges facing Big Data

One of the key set of challenges faced in today's tight market is the need to find and analyze the required data at the least speed possible. However, with the exponentially growing amount of data, speed becomes a major issue as analyzing such sheer volumes of data in detail to find out required output becomes more and more tedious. It is not only the quantity of data but also discovering the data according to the appropriateness of the project which is a Herculean task. Elimination of out-of-context data is an essential objective. Even if in-context data retrieved at a high speed is achieved, the quality of data may be compromised if it is not accurate or timely. As a result of this, appropriate results of the project may not be published.

Another zone of challenges involves those relating to the vulnerability and security of Big Data. Breaches of privacy, especially with data relating to individuals and organizations have been a topic of serious concern. One solution has been to anonymize data by removing identifiers which could be used to pinpoint particular individuals thus compromising their privacy. However, this has been largely unsuccessful as it is possible to de- anonymize the data. One very popular example of this came out in 2007 when Arvind Narayanan and Vitaly Shmatikov of the University of Texas, Austin identified particular people who had given IMDB ratings with their names and were later anonymized in a Netflix dataset of movie ratings which was built for a data-mining competition. They stated, ―Our third contribution is a practical analysis of the Netflix Prize dataset, containing anonymized movie ratings of 500,000 Netflix subscribers (section 5). Netflix—the world's largest online DVD rental service—published this dataset to support the Netflix Prize data mining contest. We demonstrate that an adversary who knows a little bit about some subscriber can easily identify her record if it is present in the dataset, or, at the very least, identify a small set of records which include the subscriber's record. The adversary's background knowledge need not be precise, e.g., the dates may only be known to the adversary with a 14-day error, the ratings may be known only approximately, and some of the ratings and dates may even be completely wrong. Because our algorithm is robust, if it uniquely identifies a record in the published dataset, with high probability this identification is not a false positive.‖ The confidentiality of data, that is, the assurance that regardless of whether the anonymity of data is maintained, the data is not visible to anyone beyond the trusted and the allowed zone is also an important aspect. Protecting data so that confidential data is not made available to anyone who is unauthorized is a very complex task and no concrete solutions have yet been developed in this field.
Organizations dealing with big data need to take this issue in their stride and make sure that the data storage and location be made heavily protected so that it is not misused. They could do so by using unique database tables, having dedicated database servers, encrypting the data, having multiple security levels, having separate authentication and authorization modules and ensuring secure system operations, data transmission and data flow control.

Three key areas of security threats have been identified in the implementation of BigData using software such as Hadoop- Breach of privacy by the unauthorized release of data, manipulation of data in the database and denial of information. In particular, in Hadoop the following areas of threat have been recognized.

1. Unauthorized access of an HDFS client via RPC or via HTTP protocols.
2. Manipulation of data in a file at a DataNode through pipeline-streaming data-transfer protocol.
3. Adding/deleting/changing the priority of a job in a queue.
4. Unauthorized access of intermediate data of Map job via its task trackers HTTP shuffle protocol.
5. An executing task may use the host operating system interfaces to access other tasks, access local data which include intermediate Map output or the local storage of the DataNode that runs on the same physical node.
6. Masquerading as Hadoop service component.
7. Submitting a workflow to Oozie as another user.

Real time security or compliance monitoring is also a challenge that is faced by Big Data analysts. Due to the copious amounts of data involved, the number of alarms triggered by the security devices is so large that several of these alarms tend to be overlooked as humans cannot cope with the shear amount.

The above challenges that are faced by Big Data need to be addressed and solutions of these problems need to be determined so that industries can start implementing big data analytics in their business strategies.

CONCLUSION

File systems that manage the storage across a network of machines are called distributed file systems. Hadoop comes with a distributed file system called HDFS. It is a framework that allows for the distributed processing of large data sets across clusters of commodity computers using a simple programming model.it is designed to scale up from single servers to thousands of machines, each providing computation, and storage. The primary goal of bid data analytics is to help companies make more informed business decisions by enabling data scientists.

REFERENCES

1. *Apac*he Hive. Available at http://hive.apache.org
2. http://blogs.worldbank.org/voices/meet-winners-and-finalists-first- wbg-big-data-innovation-challenge *Apac*he Hive. Available at http://hive.apache.org
3. http://blogs.worldbank.org/voices/meet-winners-and-finalists-first- wbg-big-data-innovation-challenge.
4. http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big- data-definition-consists-of-three-parts- not-to-be-confused-with- three-vs/
5. http://www.exist.com/wp- content/uploads/2014/10/3Vsbigdata.png
6. http://www.internetlivestats.com/twitter-statistics/
7. http://www.internetlivestats.com/google-search-statistics/
8. Grand Challenge: Applying Regulatory Science and Big Data to
9. Improve Medical Device Innovation, Arthur G. Erdman∗, Daniel F. Keefe, Senior Member, IEEE, and Randall Schiestl, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 60, and NO. 3, MARCH 2013
10. http://lsst.org/lsst/google
11. http://en.wikipedia.org/wiki/Parkinson‘s_law
12. http://www.economist.com/node/15557443
13. http://www.youtube.com/t/press_statistics/?hl=en
14. http://www.forbes.com/sites/bernardmarr/2015/04/21/how-big- data-is-changing-healthcare/
15. http://www.forbes.com/sites/bryanpearson/2015/04/10/exercise-in- service-fitbit-omni-channel-begs-for-omni-prescience/
16. http://www.engadget.com/2015/04/10/jawbone-up3-shipping-april-20th/http://www.samsung.com/uk/consumer/mobile- devices/wearables/gear/SM-R3500ZKABTU
17. http://healthdataalliance.com/
18. http://www.ibm.com/software/data/bigdata/industry-healthcare.html
19. http://www.firstpost.com/business/big-data-booster-shot- healthcare-industry-needs-2160271.html
20. Chester Curme, Tobias Preis, Eugene Stanley, Helen Susannah Moat, –Quantifying the semantics of search behavior before stock market moves‖; CrossMark, December 2013
21. http://www.wsj.com/articles/how-computers-trawl-a-sea-of-data- for-stock-picks-1427941801
22. Nitish Sinha, –Using Big Data in Finance: Example of sentiment- extraction from news articles‖; FEDS notes, March 2014
23. Baker, Malcolm and Jeffrey Wurgler, 2007. "Investor Sentiment in the Stock Market", Journal of Economic Perspectives, vol. 21(2) pages 129-152.
24. Heston, Steven L. and Sinha, Nitish Ranjan, 2013. "News versus Sentiment: Comparing Textual Processing Approaches for Predicting Stock Returns", Robert H. Smith School Research Paper. Available at SSRN: http://ssrn.com/abstract=2311310 or http://dx.doi.org/10.2139/ssrn.2311310

25. http://www.ey.com/GL/en/Services/Assurance/Fraud-Investigation---Dispute-Services/EY-Global-Forensic-Data-Analytics-Survey- 2014
26. http://www.ey.com/CA/en/Newsroom/News-releases/2014-Global- forensic-data-analytics-survey
27. http://www.ikanow.com/how-can-i-use-big-data-analytics-for- fraud-detection/
28. http://www.pactera.com/resources/blog/how-big-data-is- revolutionizing-fraud-detection-in-financial-services/
29. Ruchi Verma, Sathyan Ramakrishna Mani, –Using Analyrtics for Insurance Fraud Detection‖; FINsights, Infosys, Issue 10
30. http://www-01.ibm.com/software/analytics/solutions/customer- analytics/social-media-analytics/
31. http://www.news-sap.com/sentiment-analysis-with-big-data/ [31] https://datafloq.com/read/big-datas-impact-food-industry/96
32. http://venturebeat.com/2013/03/01/ibm-brings-big-data-tech-to- food-to-prevent-the-next-horse-meat-scandal/
33. http://www.forbes.com/sites/daniellegould/2012/09/24/food- industry-understand-trends-big-data-tools/
34. Puneet S i n g h D u g g a l , Sanchita P a u l , — Big D a t a Analysis: Challenges.
35. Samidha Mukarjee, Ravi shaw,-Big data Concepts, Applications, challenges and future scope.

B. SANDHIYA[1].Assistant Professor in Sri Shakthi Institute Of Engineering and technology under the Department Information Technology, Coimbatore. I did my Masters in Kathir College of Engineering in Computer Science and Engineering. (sandhiyab@siet.ac.in)

N. RAMYA DEVI[2]. Assistant Professor in Sri Shakthi Institute Of Engineering and technology under the Department Information Technology, Coimbatore.I did my Masters in PSG College of Technology in Information Technology. I had 3 years of Industrial Experience in Wipro Technologies Bangalore (nramyadevi@siet.ac.in)

A. **DEEPALAKSHMI**[3].Assistant Professor in Sri Shakthi Institute Of Engineering and technology under the Department Information Technology, Coimbatore. I did my Masters in Karpagam College of Engineering in Information Technology. I had 2 years of teaching experience(deepalakshmi@siet.ac.in

