IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

Big Data – Concepts, Applications, Challenges and Future Scope

Samiddha Mukherjee[1], Ravi Shaw[2]

Information Technology, Institute of Engineering & Management, Kolkata, India [1, 2]

Abstract: The term, 'Big Data' has been coined to refer to the gargantuan bulk of data that cannot be dealt with by traditional data-handling techniques. Big Data is still a novel concept, and in the following literature we intend to elaborate it in a palpable fashion. It commences with the concept of the subject in itself along with its properties and the two general approaches of dealing with it. The comprehensive study further goes on to elucidate the applications of Big Data in all diverse aspects of economy and being. The utilization of Big Data Analytics after integrating it with digital capabilities to secure business growth and its visualization to make it comprehensible to the technically apprenticed business analyzers has been discussed in depth. Aside this, the incorporation of Big Data in order to improve population health, for the betterment of finance, telecom industry, food industry and for fraud detection and sentiment analysis have been delineated. The challenges that are hindering the growth of Big Data Analytics are accounted for in depth in the paper. This topic has been segregated into two arenas- one being the practical challenges faces whilst the other being the theoretical challenges. The hurdles of securing the data and democratizing it have been elaborated amongst several others such as inability in finding sound data professionals in required amounts and software that possess ability to process data at a high velocity. 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, Data Visualization, Integration, Data Democratization, Encryption.

I. CONCEPTS

"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."[1]. 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 increases 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 [2].

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 [3] introduced the famous 3 V's concept in his 2001 Metagroup publication, '3D data management: Controlling Data Volume, Variety and Velocity'.

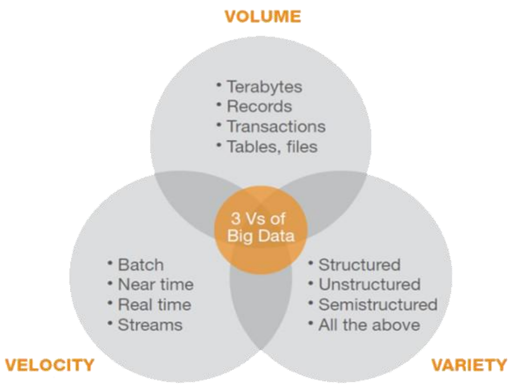


Image-1: schematic representation of the 3V's [4] of Big Data

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 be easily distinguishes between structured data, unstructured data and semi-structured data.

b. Velocity: In what now seems like the pre-historic times, data was processed in batches. However this technique is

DOI 10.17148/IJARCCE.2016.5215

only feasible when the incoming data rate is slower than the batch processing rate and the delay is much of a hindrance. At present times, the speed at which such colossal amounts of data are being generated is unbelievably high. Take Facebook [5] 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.[6].

c. Variety: Documents to databases to excel tables to pictures and videos and audios in hundreds of formats, data is now losing structure. Structure can no longer be imposed like before for the analysis of data. Data generated can be o any type- structures, semi-structured or unstructured. The conventional form of data is structured data. For example text. Unstructured data can be generated from social networking sites, sensors and satellites.

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". [7] 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." [8]
- There is a corollary to Parkinson's Law that states: "Data expands to fill the space available for storage."[9]
- This is no longer true since the data being generated will soon exceed all available storage space.[10][8]
- 72 hours of video are uploaded to YouTube every minute.[11]

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 on 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. Business experts should have the opportunity to question and interpret data according to their business requirements irrespective of the complexity and volume of the data. In order to achieve this requirement, data scientists need to efficiently visualize and present this data in a comprehensible manner. Giants like Google, Facebook, Twitter, EBay, Wal-Mart etc., adopted data visualization to ease complexity of handling data. Data visualization has shown immense positive outcomes in such business organizations. Implementing data analytics and data visualization, enterprises can finally begin to tap into the immense potential that Big data possesses and ensure greater return on investments and business stability.

• Integration- An exigency of the 21st century

Integrating digital capabilities in 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 nexus of forces. Using social and mobile technologies to alter the way people connect and interact with the organizations and incorporating big data analytics in 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.

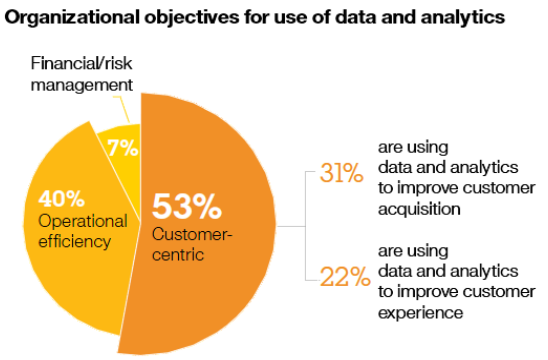

Organizational objectives for use of data and analytics

Image-2: Analysis as generated by IBM institute of Business Value 2014 Analytics Study

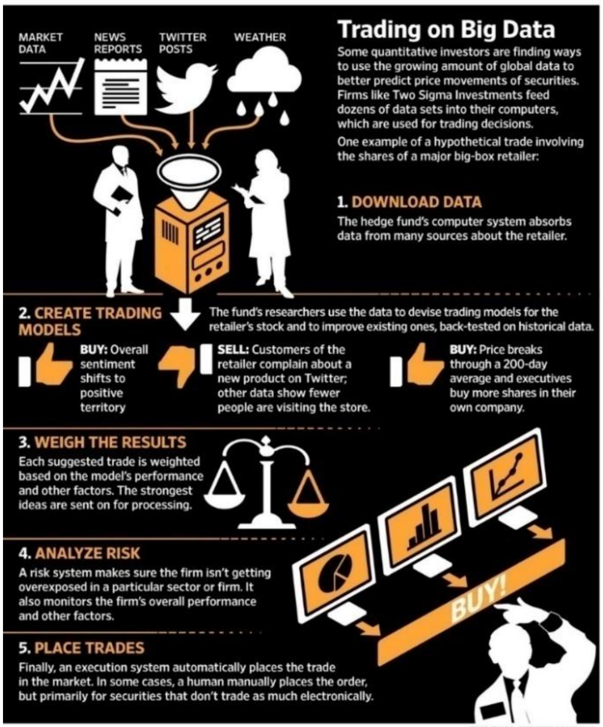
• Big Data in Healthcare:

Healthcare is one of those arenas in which Big Data ought

DOI 10.17148/IJARCCE.2016.5215

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 [12]. Devices such as the Fitbit [13], Jawbone [14] and the Samsung Gear Fit [15] 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 [16] is a collaboration of the Carnegie Mellon University, University of Pittsburgh and the UPMC. In their website, they state [16], "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 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, attracting hundreds of companies and entrepreneurs, and generating 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 [17] has taken initiative in 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 [12] for individual patients according to their genetic makeup. To summarize, Sundar Ram of Oracle stated [18], "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 [19] 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 Internet 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' [20] 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.



Trading on Big Data

Some quantitative investors are finding ways to use the growing amount of global data to better predict price movements of securities. Firms like Two Sigma Investments feed dozens of data sets into their computers, which are used for trading decisions.

One example of a hypothetical trade involving the shares of a major big-box retailer:

1. DOWNLOAD DATA
The hedge fund's computer system absorbs data from many sources about the retailer.

2. CREATE TRADING MODELS
The fund's researchers use the data to devise trading models for the retailer's stock and to improve existing ones, back-tested on historical data.

BUY: Overall sentiment shifts to positive territory

SELL: Customers of the retailer complain about a new product on Twitter; other data show fewer people are visiting the store.

BUY: Price breaks through a 200-day average and executives buy more shares in their own company.

3. WEIGH THE RESULTS
Each suggested trade is weighted based on the model's performance and other factors. The strongest ideas are sent on for processing.

4. ANALYZE RISK
A risk system makes sure the firm isn't getting overexposed in a particular sector or firm. It also monitors the firm's overall performance and other factors.

5. PLACE TRADES
Finally, an execution system automatically places the trade in the market. In some cases, a human manually places the order, but primarily for securities that don't trade as much electronically.

THE WALL STREET JOURNAL.

Image-3: Wall Street Journal [20] summarizes the above concept.

One very interesting avenue of using Big Data in finance is the sentiment extraction [21] from news articles. Market sentiment refers to the irrational belief in investors about cash-flow returns [22]. The Heston-Sinha's Application of the Machine Learning algorithm [23] 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 [24] 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. Ernst & Young undertook the Global forensic data analytics survey [25] in 2014 and found that, "Our survey finds that 42% of companies with revenues between US$100 million to US$1 billion are reviewing less than 10,000 records. And 71% companies with more than US$1 billion in sales report examining just one million records or fewer….Companies know there are high risk numbers in book entries, such as round thousands or duplicates, but they're only just starting to analyze descriptions for those book entries. Looking at both the numbers and words can mean the difference between uncovering fraud, and falling victim to it." The combination of appropriate data and big data analytics can help combat fraudulent activities. Though several companies are mining big data for this purpose there are still limitations [26] in their approach. They are either keeping the data siloed, limiting the analysis to be performed or only taking into consideration the structured data thus only giving a subset of information. A more holistic approach to the implementation of big data analytics is required. Companies such as Pactera [27] is developing solutions which will process massive amounts of structured and unstructured data and develop varied models and algorithms to find patterns of fraud and anomalies and predict customer behavior.

A 10 step approach has been suggested by Infosys [28] 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 new set of techniques developed.

• Big Data and 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. Several giants are presently developing tools for efficient sentiment analysis.

IBM has developed IBM Social Media Analytics [29] 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:

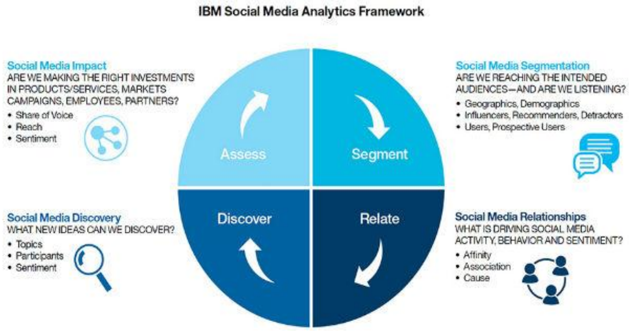

IBM Social Media Analytics Framework
Social Media Impact
ARE WE MAKING THE RIGHT INVESTMENTS IN PRODUCTS/SERVICES, MARKETS CAMPAIGNS, EMPLOYEES, PARTNERS?
• Share of Voice
• Reach
• Sentiment
Social Media Segmentation
ARE WE REACHING THE INTENDED AUDIENCES—AND ARE WE LISTENING?
• Geographics, Demographics
• Influencers, Recommenders, Detractors
• Users, Prospective Users
Assess
Segment
Discover
Relate
Social Media Discovery
WHAT NEW IDEAS CAN WE DISCOVER?
• Topics
• Participants
• Sentiment
Social Media Relationships
WHAT IS DRIVING SOCIAL MEDIA ACTIVITY, BEHAVIOR AND SENTIMENT?
• Affinity
• Association
• Cause

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

Similarly SAP has developed a SAP-HANA based application known as Social Contact Intelligence [30] 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 and the Food Industry:

The impact of Big Data on the food industry [31] 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 on 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, [32] 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 [31] to track individual customer preferences and present customers with personalized offers so as to increase customer base and improve customer satisfaction.

DOI 10.17148/IJARCCE.2016.5215

- 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.

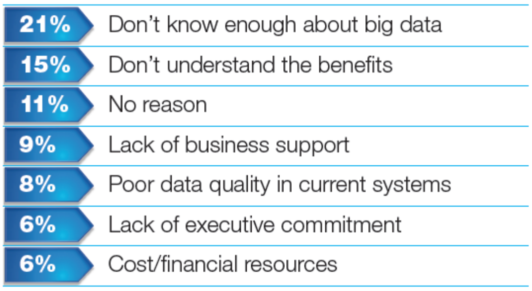
III. OBSTACLES IN BIG DATA INMPLEMENTATION

In the 1990's Big Data became a hyped topic of interest in the world of distributed systems [34] 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 [35] to handle this task. At the turn of the millennium, in response to this issue, Google created the Google File System (GFS) [36] which provided consumers with OS-level byte stream operations [35] on data spanning several machines in clusters using rather expensive hardware. Later, Google developed the MapReduce paradigm [34] which was identical to the partitioned parallelism used in shred-nothing parallel query processing. Following this trend [37], multi-national giants like Yahoo and Facebook developed their own software. Yahoo! Developed Pig while Hive was developed by Facebook [38], Jaql by IBM [39] and Dryad and Scope by Microsoft [40].

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 on why most industries are still delaying or refusing to pursue a big data strategy. It states, [41] "A little more than one-fifth of the respondents are 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' [42] 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.



21%	Don't know enough about big data
15%	Don't understand the benefits
11%	No reason
9%	Lack of business support
8%	Poor data quality in current systems
6%	Lack of executive commitment
6%	Cost/financial resources

Image 5- [41] 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. … [43] 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 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 [44] summarizes the data management options available.

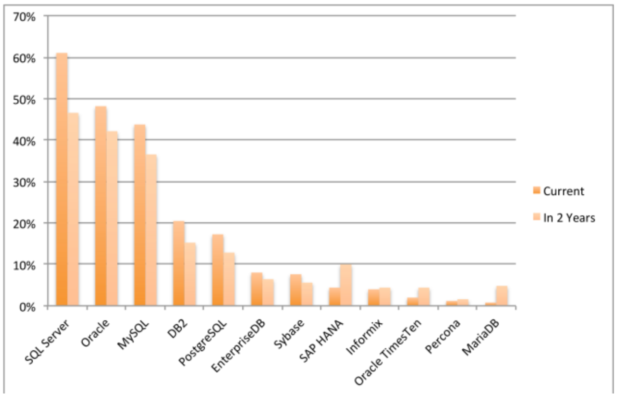


Image 6- Current adoption of relational database technologies with projected two year growth [44].

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 [42] 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 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 is 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 are 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 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 [46] faced in today's tight market is the need to find and analyze the required data at the least speed possible. However with 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, [47] "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

DOI 10.17148/IJARCCE.2016.5215

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 [48] have been identified in the implementation of BigData using software such as Hadoop- Breach of privacy by unauthorized release of data, manipulation of data in the database and denial of information. In particular, in Hadoop the following areas of threat [49] have been recognized.

- Unauthorized access of an HDFS client via RPC or via HTTP protocols.
- Manipulation of data in a file at a DataNode through pipeline-streaming data-transfer protocol.
- Adding/deleting/changing priority of a job in a queue.
- Unauthorized access of intermediate data of Map job via its task trackers HTTP shuffle protocol.
- 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.
- Masquerading as Hadoop service component.
- Submitting a workflow to Oozie as another user.

Real time security [50] 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 [51].

The above challenges that are faced by Big Data needs 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.

IV. FUTURE SCOPE AND DEVELOPMENT

Today, Big Data is influencing IT industry like few technologies have done before. The massive data generated from sensor-enabled machines, mobile devices, cloud computing, social media, satellites help different organizations improve their decision making and take their business to another level.

"Big data absolutely has the potential to change the way governments, organizations, and academic institutions 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.

Data is the biggest thing to hit the industry since PC was invented by Steve Jobs. As mentioned earlier in this paper, every day data is generated in such a rapid manner that, traditional database and other data storing system will gradually give up in storing, retrieving, and finding relationships among data. Big data technologies have addressed the problems related to this new big data revolution through the use of commodity hardware and distribution. Companies like Google, Yahoo!, General Electric, Cornerstone, Microsoft, Kaggle, Facebook, Amazon 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. [52]

“Demand is so hot for solutions that all companies are exploring big data strategies. The problem is that the companies lack internal expertise and best practices.. the side effect is that there is a 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, Indian Prime Minister’s office is using Big Data analytics to understand Indian citizen’s sentiments 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. [53]

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 launch virtual machines, store huge amount of data online, and plenty of other things [54]. Basically, it will be an one stop platform for cloud based applications, online gaming, mobile applications, etc. [55]. All these required huge amount of data processing where Big Data plays an immense role in data processing.

The predictions 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, investing in this enabler of end-user self-service will become a 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 requirement.
3. Shortage of skilled staff will persist. In the U.S. alone there will be 181,000 deep analytics roles in 2018 and five times that many positions requiring related skills in data management and interpretation.
4. By 2017 unified data platform architecture will become the foundation of BDA strategy. The unification will occur across information management, analysis, and search technology.

DOI 10.17148/IJARCCE.2016.5215

5. Growth in applications incorporating advanced and predictive analytics, including machine learning, will accelerate in 2015. These apps will grow 65% faster than apps without predictive functionality.
6. 70% of large organizations already purchase external data and 100% will do so by 2019. In parallel more organizations will begin to monetize their data by selling them or providing value-added content.
7. Adoption of technology to continuously analyze streams of events will accelerate in 2015 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. Decision management platforms will expand 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. Rich media (video, audio, image) analytics will at least triple in 2015 and emerge as the key driver for BDA technology investment.
10. By 2018 half of all consumers will interact with services based on cognitive computing on a regular basis.[56]

Big data isn't new, but now has reached critical mass as people digitize their lives. "People are walking sensors," said Nicholas Skytland, project manager at NASA within the Human Adaptation and Countermeasures Division of the Space Life Sciences Directorate [57].

Taking an average of all the figures suggested by leading big data market analyst and research firms, it can be 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.

V. CONCLUSION

This literature survey discusses Big Data from its infancy until itscurrent state. It elaborates onthe concepts of big data followed bythe applications and the challenges faced by it. Finally we have discussed the future opportunities that could be harnessed in this field. Big Data is an evolving field, where much of the research is yet to be done. Big data at present, is handled by the software named Hadoop. However, the proliferating amounts of data is making Hadoop insufficient. To harness the potential of Big Data completely in the future, extensive research needs to be carried out and revolutionary technologies need to be developed. Summarising, Peter Sondergaard, Senior Vice President of Gartner Research famously stated, "Information is the oil of the 21st century and analytics is the combustion engine."

REFERENCES

[1] Apache Hive. Available at http://hive.apache.org

[2] http://blogs.worldbank.org/voices/meet-winners-and-finalists-first-wbg-big-data-innovation-challenge

[3] http://www.forbes.com/sites/gartnergroup/2013/03/27/gartners-big-data-definition-consists-of-three-parts-not-to-be-confused-with-three-vs/

[4] http://www.exist.com/wp-content/uploads/2014/10/3Vsbigdata.png

[5] http://www.internetlivestats.com/twitter-statistics/

[6] http://www.internetlivestats.com/google-search-statistics/

[7] Grand Challenge: Applying Regulatory Science and Big Data to Improve Medical Device Innovation, Arthur G. Erdman∗, Daniel F. Keefe, Senior Member, IEEE, and Randall Schiestl, IEEE TRANSACTIONS ON BIOMEDICAL ENGINEERING, VOL. 60, NO. 3, MARCH 2013

[8] http://lsst.org/lsst/google

[9] http://en.wikipedia.org/wiki/Parkinson's_law

[10] http://www.economist.com/node/15557443

[11] http://www.youtube.com/t/press_statistics/?hl=en

[12] http://www.forbes.com/sites/bernardmarr/2015/04/21/how-big-data-is-changing-healthcare/

[13] http://www.forbes.com/sites/bryanpearson/2015/04/10/exercise-in-service-fitbit-omni-channel-begs-for-omni-prescience/

[14] http://www.engadget.com/2015/04/10/jawbone-up3-shipping-april-20th/

[15] http://www.samsung.com/uk/consumer/mobile-devices/wearables/gear/SM-R3500ZKABTU

[16] http://healthdataalliance.com/

[17] http://www.ibm.com/software/data/bigdata/industry-healthcare.html

[18] http://www.firstpost.com/business/big-data-booster-shot-healthcare-industry-needs-2160271.html

[19] Chester Curme, Tobias Preis, Eugene Stanley, Helen Susannah Moat, "Quantifying the semantics of search behavior before stock market moves"; CrossMark, December 2013

[20] http://www.wsj.com/articles/how-computers-trawl-a-sea-of-data-for-stock-picks-1427941801

[21] Nitish Sinha, "Using Big Data in Finance: Example of sentiment-extraction from news articles"; FEDS notes, March 2014

[22] Baker, Malcolm and Jeffrey Wurgler, 2007. "Investor Sentiment in the Stock Market", Journal of Economic Perspectives, vol. 21(2), pages 129-152.

[23] 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

[24] http://www.ey.com/GL/en/Services/Assurance/Fraud-Investigation--Dispute-Services/EY-Global-Forensic-Data-Analytics-Survey-2014

[25] http://www.ey.com/CA/en/Newsroom/News-releases/2014-Global-forensic-data-analytics-survey

[26] http://www.ikanow.com/how-can-i-use-big-data-analytics-for-fraud-detection/

[27] http://www.pactera.com/resources/blog/how-big-data-is-revolutionizing-fraud-detection-in-financial-services/

[28] Ruchi Verma, Sathyan Ramakrishna Mani, "Using Analyrtics for Insurance Fraud Detection"; FINsights, Infosys, Issue 10

[29] http://www-01.ibm.com/software/analytics/solutions/customer-analytics/social-media-analytics/

[30] 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 Singh Duggal, Sanchita Paul ," Big Data Analysis: Challenges and Solutions", International Conference on Cloud, Big Data and Trust 2013, Nov 13-15, RGPV

[35] MarcinJedyk, MAKING BIG DATA, SMALL, Using distributed systems for processing, analysing and managing large huge data sets, Software Professional‟s Network, Cheshire Data systems Ltd.

[36] S. Ghemawat, H. Gobioff, and S. Leung, "The Google File System." in ACM Symposium on Operating Systems Principles, Lake George, NY, Oct 2003, pp. 29 – 43.

[37] Jefry Dean and Sanjay Ghemwat, MapReduce:A Flexible Data Processing Tool, Communications of the ACM, Volume 53, Issuse.1,January 2010, pp 72-77.

[38]PIGTutorial,YahooInc.,http://developer.yahoo.com/hadoop/tutorial/pigtutorial.html

[39] IBM-What.is.Jaql, www.ibm.com/software/data/infosphere/hadoop/jaql/

[40] Dryad - Microsoft Research, http://research.microsoft.com/en-us/projects/dryad/

DOI 10.17148/IJARCCE.2016.5215

[41] "2013 Big Data Survey Research Brief", SAS, The power to know, 2013

[42] David Loshin, "Addressing Five Emerging Challenges Of Big Data"; Progress Software, 2014

[43] Eric Lundquist, "Gartner: 2013 Tech Spending To Hit $3.7 Trillion" October 23, 2012

[44] 2014 Data Connectivity Outlook paper"; Progress Software, January 2014.

[45] http://www.forbes.com/sites/ciocentral/2012/04/16/the-big-cost-of-big-data/

[46] SAS, The power to know, "Five Big Data Challenges And how to overcome them with visual analytics"

[47] Arvind Narayanan, Vitaly Shmatikov, "Robust De-anonymization of Large Sparse Datasets"; The University of texas at Austin, 2007

[48] Victor L. Voydock and Stephen T. Kent. Security mechanisms in high-level network protocols. ACM Comput. Surv 15(2):135–171, 1983.

[49] Devaraj Das, Owen O'Malley, Sanjay Radia and Kan Zhang, "Adding security to Apache Hadoop", Hortonworks Technical Report 1

[50] Disha H. Parekh, Dr. R. Sridaran ,‖An Analysis of Security Challenges in Cloud Computing‖ in International Journal of Advanced Computer Science and Applications, Vol. 4, No.1, 2013

[51] Rashmi N, Uma K M, Jayalakshmi K, Vinodkumar K P, "Big Data Security Challenges: Dealing with too many issues"; International Journal of Recent Development in Engineering and Technology, Volume 3, Issue 2, August 2014

[52] http://www.forbes.com/sites/siliconangle/2012/02/29/big-data-is-creating-the-future-its-a-50-billion-market/

[53] http://dataconomy.com/indian-government-using-big-data-to-revolutionise-democracy/

[54] https://en.wikipedia.org/wiki/Google_Cloud_Platform

[55] https://cloud.google.com/

[56] http://www.idc.com/getdoc.jsp?containerId=prUS25329114

[57] http://www.zdnet.com/article/30-big-data-project-takeaways/

DOI 10.17148/IJARCCE.2016.5215