Survey on Big Data Using Data Mining

[1]Siddharth Singh, [2]Tuba Firdaus, [3] Dr. A.K. Sharma
[1]M.TECH Scholar, [2]M.TECH Scholar, [3]Associate Professor
[1,2]Information Technology, [3]Computer Science Department
[1]Madan Mohan MalaviyaUniversity of Technology, Gorakhpur, Uttar Pradesh, 273001, INDIA

***Abstract* - Extracting useful information from large data-set like in all science and engineering domain, There will be most exciting opportunity in upcoming years for big data. This paper includes big data, Data mining, Data mining with big data, Challenging issue and survey papers of various companies related to big-data. Every organization focused on how to manage large set of data and how much companies invested in big-data as well as what type of return they get. Many technical challenges like implementations and visualizations are to be taken into consideration in future. To manage and analyze edge data explore business opportunities deriving from the analytics of edge data. Collaborate with the business to understand existing edge system and the potential use for data. It concluded from the findings that Enterprise are still looking for the right infrastructure tools that will enable them to effectively handle their big-data with their business needs.**

***Index Terms* - Data mining Algorithm, Big-Data, Data mining, big- data challenges, Big-data characteristics, Data mining trends.**

I. INTRODUCTION

Anything that we requires in this hi-tech generation or you can say that is unaware to us then we go for Google and within few second we got several results according to entered queries. This may be a better example of big Data. We cannot manage big data by various data mining tools or software's that we have. In 2011 when India won world cup then it triggered numbers of tweets within 1 - 2 hour and among these tweets that appear to be special comments that can reveal interest in public. Such online discussion provides a new way to sense public interest and generating feedbacks in real time. This example demonstrates the rise of big Data application. Because of regularly increasing of data collection we cannot use software tools to capture and managing it within a tolerable time. Big data is a buzzword, catchphrase, used to describe a massive volume of structured and unstructured data that is so large and it's very difficult to process using traditional database like RDBMS, ORDBMS etc and various software techniques. If we consider example of Facebook where lot of people upload images, videos and text etc daily and also keep updating these data continuously. So due to large in size, we cannot control centralized and different data sources with different size as well as types, data becomes hard to access and create complexity. When data changes time to time it stores in data warehouse and it creates large amount of data that will require to large space and storage for actual implementation. Because of large size of data it is impossible to control data individually so it may be dividing into groups. The tool that we used for managing the data regularly, we cannot use it for big-data in real time. In this paper section 2 dictates a formal understanding about big-data and data mining. Section 3 illustrates about various key features of big-data in mining platform. In section 4, we discuss about challenges in big-data mining platform and section 5contains the literature review or survey of various companies after that comparison between various data mining trends includes in section 6 and finally conclusion and future work are discussed in section 7.

II. BIG DATA AND DATA MINING

Data stored at the server of Facebook that is used by people in daily life where we upload various types of information like pictures, videos and all of these data stored on the warehouse of data at the Facebook servers, we called it big- data due to its complexity. Big-data is nothing but a data available at autonomous and heterogeneous sources in extreme large amount which gets updated within a fraction of second. Another example of big- data we can take like reading taken from an electronics microscope of the universe. Now the term Data mining can be defined as extraction of useful information from the collected or gathered data or we can say extraction of knowledge from database. So big-data mining is a close up view that contains a lot of useful detailed information of big-data.

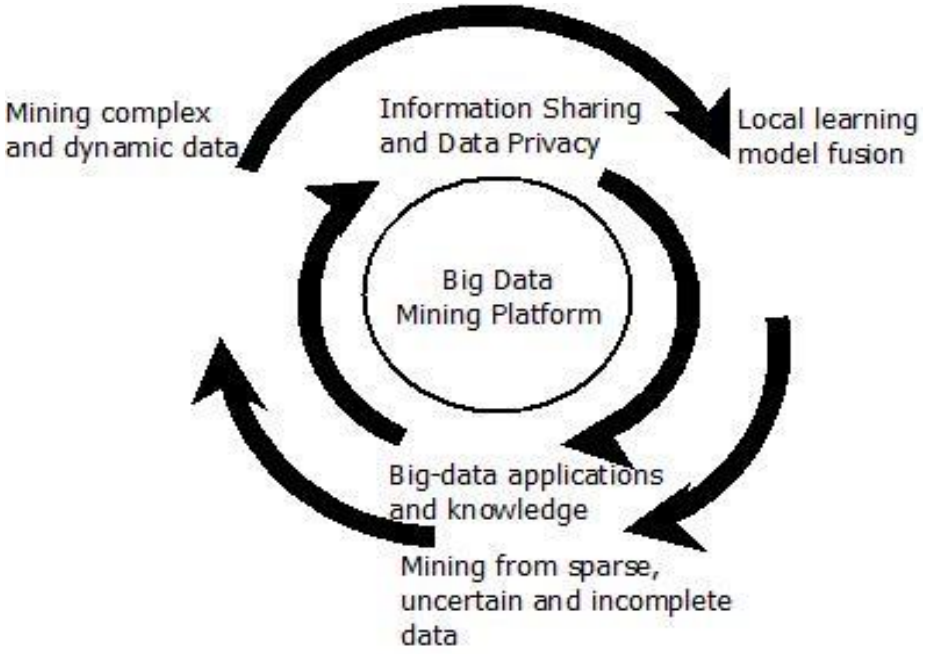

Fig: 1 Cycle of big-data mining platform

Big-data included: Enterprise data
Transaction data
Social media
Public data
Sensor data

Processing of data by various companies each year as follows:

- Facebook has 2.5 PB of user data+ 15 TB /day.
- E-Bay has 6.5 PB of user data+50 TB/day.
- Google processes 20 PB a day.
- Way-back machine has 3PB+100 TB/month.
- CERN's large Hydron Collider (LHC) generates 15 PB a year.

***Types of data consisted such as*: -**

(1) Relational data (Tables/Transactions/Legacy data).
(2) Text data (Web).
(3) Semi structured data (XML).
(4) Graph data (Social network, Semantic web).
(5) Streaming Data (You can only scan the data once).

***Operations performed with these data*:-**

- Aggregations and Statistics (Data warehouse and OLAP).
- Indexing, Searching, Querying (Keyword based searching and pattern matching).
- Knowledge Discovery (Data Mining and Statistical modeling).

Data Mining:

(1) Discovery of useful, possibly unexpected, patterns of data.
(2) Non-trivial extraction of implicit previously unknown and potentially useful information from data.
(3) Exploration and analysis by automatic or semiautomatic means of large quantities of data in order to discover meaningful pattern.
(4) The goal of data mining is to extract knowledge from dataset in human understandable structures.
(5) In recent years data mining has been used in various science and engineering fields such as medicine, bio-informatics, genetics, education and engineering.

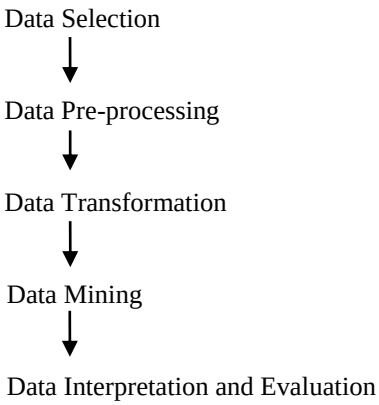

Fig: 2 (Knowledge Discovery Process)

***Data Mining Task*:**

- Classification (Predictive)
- Clustering (Descriptive)
- Association Rule Discovery(Descriptive)
- Sequential Pattern Discovery (Descriptive).
- Regression (Predictive).
- Deviation Detection (Predictive).
- Collaborative Filter (Predictive).

Advantage of Data Mining in Various Applications:-

(1) Banking
(2) Marketing
(3) Health Care
(4) Manufacturing and Production
(5) Insurance
(6) Law
(7) Government and Defense
(8) Computer hardware and software
(9) Airlines
(10) Brokerage and Securities trading.

***Challenges Faced By Data Mining*:-**

- Data quality
- Privacy preservation
- Network Setting
- Data Ownership and distribution
- Complex and Heterogeneous data
- Scalability
- Streaming Data
- Dimensionality.

(1) Key Features Of Big-Data Mining Platform

(1) Hard to handle because of its complexity.
(2) Continuously changing of data time to time.
(3) Big-data are huge in size.
(4) Big-data is free from control and guidance of anyone.
(5) Data sources of big-data are different from different phases.

If we consider example of Facebook where lot of people upload images, videos and text etc daily and also keep updating these data continuously. So due to large in size, we cannot control centralized and different data sources with different size as well as types, data becomes hard to access and create complexity. When data changes time to time it stores in data warehouse and it creates large amount of data that will require to large space and storage for actual implementation. Because of large size of data it is impossible to control data individually so it may be dividing into groups. The tool that we used for managing the data regularly, we cannot use it for big-data in real time.

(2) 6 'V' Challenges in Big-Data

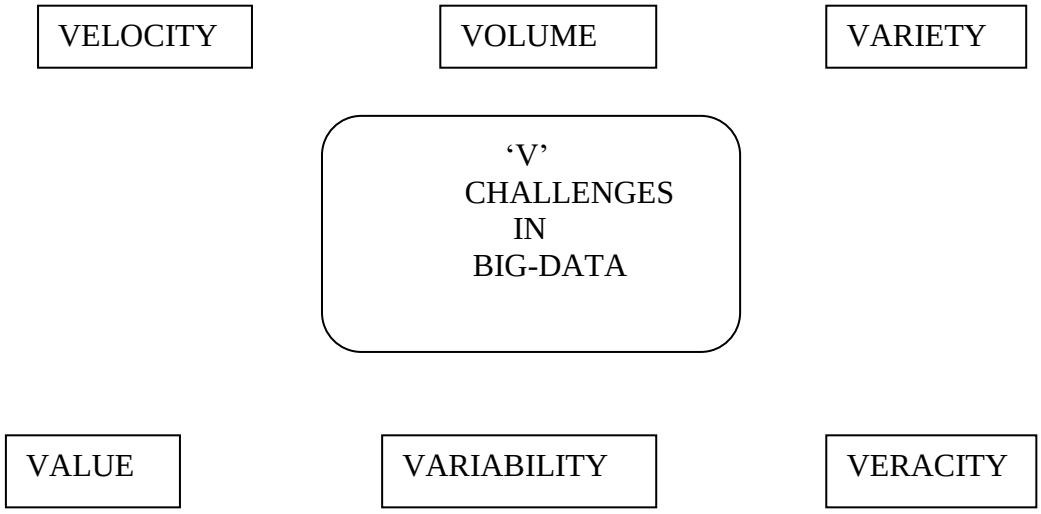

Fig: 3 Challenges in Big-data

1) ***Velocity*:** When we use applications on any android devices or phone, we expect a response from mobile devices as soon as possible i.e. immediately. As a hardware engineer or application developer we want to give best user experience to smart phone users. So mobile applications are now able to fetch sensor data in real time as well as continuously. Velocity of mobile data generation is far quicker than before. More and more data are produced and must be collected in shorter time frames or we can say velocity refers to the speed of generation of data i.e. how fast the data is generated and processed to meet the demand and challenges which lie ahead in the path of growth and development.
2) ***Volume:*** Most visible aspect of big-data referring fact that amount of generated data has increased tremendously from the past years. Monthly data traffic will surpass 15 Exabyte's by 2018.According to Wikipedia one Exabyte could hold 100,000 times the printed material. Mobile device now generate unprecedented data volume. No SQL database approach is a response to store and query huge volumes of data heavily distributed. It is the size of data which determines the value and potentials of data under consideration and whether it can actually be considered as big-data or not? The term 'Big-data' itself refers the size of data.
3) ***Variety:*** All forms of data incorporated by big-data like audio, video and mobile sensor data, post and updates from all social networking sites. Creating more business models, new sources of data will be added and ability to handle data with so many sources is the key features of big- data. Range of variety becomes from structured text to free text. This means that category to which big-data belongs to is also a very essential fact that needs to be known by data analysts. This helps the people who are closely analyzing the data and are associated with it to use the data effectively and thus upholding the importance of big-data.
4) ***Value:*** The large amount of data volume extremely rapidly increase in velocity and the number of forms of data make big-data unique from previously storage of data. We can find correlations of data with real world incidents which will help us to predict the future and develop strategies for future using data with the help of data mining, machine learning and big-data. The challenges is to find the way to transform raw data into information that has value either internally or for making a business out of it.
5) ***Variability:*** Variability may be a big factor which can be a problem for those who analyze data. This refers to the inconsistency which can be shown by data time to time, thus hampering the process of being able to handle and manage the data effectively.
6) ***Veracity:*** The quality of data being captured can vary greatly and accuracy of analysis depends on the veracity of the source data.

One more challenges can be included that is **Complexity** of data. Data management can become a very complex process when large volume of data comes from multiple sources. These data need to be linked, connected and correlated in order to be able to grasp information that is supposed to be conveyed by these data. Big-data analytics consists of 6 **'C'** system

- Connection (Sensor and network).
- Cloud (Computing and data on demand).
- Cyber (Model and memory).
- Content/Context (Meaning & correlation).
- Community (Sharing & collaboration).
- Customization (Personalization & value)

III. THE PAPER SURVEY

(1). The 2014 **IDG** Enterprise big data research was completed with the goal of gaining a better understanding of organizations big data initiatives, investments and strategies. Key findings include as:

(ii) Organizations are investing in developing or buying software applications, additional server hardware and hiring staff with analytical skills in preparation for big data initiatives.

(iii) Organizations are facing numerous challenges with big data initiatives and limited availability of skilled employees to analyze and manage data tops the list.

(iv) Half of the respondent indicated there is no clear thought leader in the big-data solution space.

(v) Organizations are seeing exponential development in the amount of data managed with an expected increase of 76% within the next 12-18 months.

(vi) CEO's are centered around on the value of big data and are partnering with IT executive that will purchase/manages/executes on the strategies.

(2). TCS launched its own study on big-data that focused on following issues-

(i) What is the current state of technology and where is it going?

(ii) What kind of digitized data are they finding to be most important?

(iii) What are the biggest challenges turning big data into insights that enable the company to make far better and faster decisions?

(iv) How much are companies investing in big -data and what kinds of returns they achieving on their spending?

(v) How are they organizing the professionals who process and analyze big-data?

(3). ACCENTURE engaged today with the practical reality of helping make big-data work across large, complex enterprises in many different industries. To get the most from their big-data projects organizations focused on:-

(i) ***Explore the entire big-data eco-system*****: -** The big-data landscape is in a consistent state of flux with new data sources and emerging big-data technologies .Explore all data available and be prepared to explore a broad range of technology options when developing a big-data strategy with a focus towards business actions and outcomes that can be differentiating in the market.

(ii) ***Start small then grow*****: -** Focus resources around proving value quickly in one area of the business first via a pilot program or proof of concept. Build internal consensus and then grow big data programs organically.

(iii) ***Be nimble***: - Stay flexible, adapt and learn as technologies evolve and new opportunities can be explored.

(iv) ***Focus on building skills*****: -** In addition to staffing up when possible, builds skills of existing employees with training and development and tap outside expertise. If we talk about research, more than 4,300 targets were screened, 36 percent have not completed nor or currently pursuing a big-data installation while nearly 4 percent were currently implementing their first big-data project. Among those who have completed their big-data project more than half did not meet our demographic criteria. A total of 1007 respondents completed the survey.

(4). According to **BARC** Big-data describes methods and technologies for the highly scalable loading, storage and analysis of structured data. Big-data technology can help companies to manage large data volumes, complex analysis and real time integration of data from a variety of data structures and sources. There are following key findings of survey:-

(i) ***Drivers of Big-data*****: -** The main drivers for big-data were new or better possibilities for data analysis (75 percent), large volumes of data (72 percent), poly-structured data sources (66 percent) as well as faster data integration (43 percent). With regard to larger volume of data, 49 percent of respondents anticipated growth rates exceeding 25 percent in 2013.

(ii) ***Organization of big-data: -*** In most companies, the topic of big data fell under the responsibility of a BI team or competency center (47%) compared to IT department (23 %). Best in class companies and those located in the UK often assigned the topic of big data to a BI team or competency center.

(iii) ***Big-data strategies*****: -** 14 % of the companies surveyed have already developed a specific strategy for big-data. While 63 % did not have set a big data strategy at the time of this survey, 23% of respondents intended to implement one. Merely having a big-data strategy however was no guarantee that companies handled their data successfully.

(iv) ***Usage of big-data*****: -** Companies uses big-data technologies in finance and controlling (24 %), marketing (19 %), Sales (18%), IT (18%) and production 17%.Participant in this survey saw a broad range of benefits, the top two being better strategic decision making and improved operational processes.

(v) ***Problems using big-data*****: -** Problems in using big-data were inadequate knowledge of both technical (46%) and business (44%) issues. No clear business case (36%), technical problems (34%), and cost (33 %) were also commonly cited problems.

(vi) ***Using different types of data*****: -** Organization utilizes different types of data like log (55%), sensor (44%) and unstructured data (40%).Social media data has the largest potential.

(5). DELL survey yields the surprising results for big data. To get a grasp on how companies plan to implement data management into their business strategies, DELL software conducted a survey of 300 DBA to see what type of system they use, how they use them and where they expect to invest in future. With all of the hype around big-data and new analytics platform like HADOOP and No SQL, john Whittaker, executive director of information management at DELL software found it surprising that small structured data is still the focal point for up to 75% of companies surveyed.

(i) The survey reminds us that these technologies are not wholly the future of data analytics. There is still room for traditional database platform such as MS-SQL, ORACLE and IBM DB2.

(ii) The survey said structured data has not yet drowned in the ever-deepening data pool. For all the interest in how to capture and managed unstructured and semi -structured data, structured data remains the bedrock of the information infrastructure in most companies.

(iii) The most important driver for the growth of unstructured data is internally generated documents, followed by e-mail.

(iv) Cloud computing and virtualization are bigger priorities according to the survey.

(v) Ten percent of the respondent said that currently use No SQL, while 20% said they are current HADOOP users and 57% said that they have no plans to implement HADOOP in future.

(3) . Big-data research in IBM: -

(i) ***New varieties of data:*** **-** Text/Social media, Network, Multimedia, Machine data/Sensor.

(ii) ***Big-Data Performance*****: -** In memory, Benchmark, New Architectures.

(iii) ***Information Integration*****: -**Integrating enterprise and public data**,** Linking data/context**,** Entity extraction and integration.

(iv) ***Industry Applications*****: -** Healthcare, Telco, Retail and marketing, Energy, Water/Agriculture, Public safety, Smarter Workforce.

Big-data definition given by IBM in form of 3'V' characteristics-

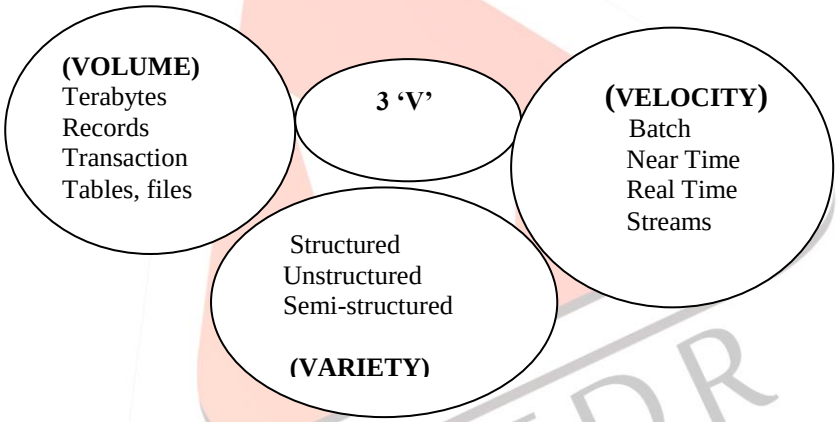

Fig: 4 (Big-data characteristics given by IBM)

Structured Data: - Data from enterprise system (ERP, CRM), Relational Database, Spreadsheets.

Unstructured Data: -Word documents, PDF files, Text Files, E-mail body, Twitters messages, books and non language based data such as Pictures, slides, Audio, and Videos.

Semi-structured Data: - XML files, Traffic signs posted along highways or WebPages.

(4) . NVP (New Vantage Partner) survey: -The 2014 big-data executive survey is the beneficiary of a record number of senior executive respondents. This year, 59 companies participated with 125 individual executive respondents. Executive participants spanned senior business and technology roles. However, wherever these executives sat within the organization, they shared a common interest and common objective to see their company effectively utilize data and analytics in support of corporate business goals. The 2014 highlighted some new as well as some familiar themes such as:

(i) Big-data investment is significant and growing.

(ii) Big-data initiatives are in production or fast on the way.

(iii) Big-data initiatives are being driven from the top, with senior executive sponsorship and commitment.

(iv) The chief data officer (CDO) role is becoming a corporate standard.

(v) Business and technology partnership is growing and critical to big-data success.

(vi) Big-data is being integrated in to main stream.

(vii) Executives are wary of the term big-data.

(viii) The survey focused on organizations that traditionally make the largest investment in data initiatives, particularly financial service firms as well as emerging new sectors such as healthcare and life-sciences firms, which are making major new investment in data initiatives.

(8). Capgemini Consulting Survey: - Capgemini consulting conducted a global survey of senior big-data executives in November 2014. The survey covers 226 respondents across Europe, North America and APAC, and spanned multiple industries including retail, manufacturing, financial services, energies and utilities and pharmaceuticals. The survey targeted senior executives across the analytics, business and IT functions that are responsible for overseeing big-data initiatives in their organization. Respondents were asked questions around their organization's approach to big-data governance, data management, skill development and technology infrastructure.

Worldwide distribution of Respondents	**Europe 50%**	**North America 39%**	**APAC 11%**
Function wise distribution of Respondents	Analytics 36%	Business 36%	IT 26%

I. Nearly 60% of the senior executives believe that big-data will disrupt their industry within the next three years.
II. Only 27% of the executives surveyed described their big-data initiatives successful.
III. Lack of strong data management and governance mechanisms and the dependence on legacy system, are among the top challenges that organizations face.
IV. Key challenges of big data included as: -
  i. Absence of clear business case for funding and implementations.
  ii. Ineffective co-ordinations of big-data and analytics team across the organizations.
  iii. Dependency on legacy system for data processing and management.
  iv. Ineffective governance model for big-data and analytics.
  v. Lack of sponsorship from top management.
  vi. Lack of big-data and analytical skills.
  vii. Lack of clarity on big-data tools and technology.
  viii. Cost of specific tools and infrastructure for big-data and analytics.
  ix. Data security and privacy concerns.
  x. Resistance to change within organizations.

(9). Applications of big-data in governmental processes: -

(i) United States of America

i. In 2012, the OBAMA administrations announced the big-data research and development initiative, to explore how big-data could be used to address important problems faced by the government. The initiative is composed of 84 different big-data programs spread across six departments.
ii. Big-data analysis played a large role in BARACK OBAMA's successful 2012 re-election campaign.
iii. The United States federal government owns six of the 10 most powerful supercomputers in the world.
iv. The Utah data center currently being constructed by the United States national security agency. When finished the facility will be able to handle the large amount of information collected by the NSA over the internet.

(ii) India

i. Big-data analysis was in parts, responsible for the BJP and its allies to win a highly successful Indian general election 2014.
ii. The Indian government utilizes numerous techniques to ascertain how the Indian electorate is responding to government actions, as well as ideas for policy augmentations.

***(iii)* United Kingdom**

i. Data on prescription drugs: By connecting origin, location and time of each prescription, research units were able to exemplify the considerable delay between the release of any drug, and a UK wide adaption of the National Institute of Health and care excellence guidelines. This suggests that new/ most up to date drugs take some times to filter through to the general patient.
ii. Joining up Data: The weather challenges in winter 2014 a local authority blended data about services, such as road gritting rotas, with services of people at risk, such as 'meals on wheels'. The connection of data allowed the local authority to avoid the any weather related delay.

(10). Analytics @ Twitter: -

Table 1. Twitter analysis

	Features	**Time Dimension**	**Data Resolutions and Processing models**
COUNTING	How many request/ day?	Real Time	Mostly event driven.

	What is average latency? How many Signups, SMS, tweets?	(M sec/sec)	High Resolution- every tweets counts
COR-RELATING	Desktop vs. Mobile users? What devices fail at same time? What features get user hooked?	Near Real Time (min/ hours)	Ad-Hoc Queries Mid Resolution – Aggregated counters
RESEARCH	What features get Re-Tweeted? Duplication detection Sentiment Analysis	Batch (days)	Pre- Generated Reports Cross gain Resolutions- trends

(6) Comparison Made Between Data Mining Trends:-

Table 2. Comparison between Data Mining Trends

Trends of Data Mining	Techniques/ Algorithm used	Data Formats	Computing Resources
Past	Statistical and Machine Learning Techniques	Structured data stored in traditional database and numerical data	Evolution of 4G PL and various related techniques
Present	AI, Pattern Reorganization, Statistical and Machine Learning techniques	Structured Semi-structured And Unstructured data formats	Parallel Distributed computing, High Speed Networks, High end storage devices
Future	Fuzzy logic Neural network Genetic Programming	Complex data objects like high speed data streams Noise in the time series Graph, Multi represented objects and Temporal data	Cloud computing and Multi-agent technologies

IV. CONCLUSION AND FUTURE WORK

Big data is going to continue growing during the next years and each data scientist will have to manage much more amount of data every year. The data is going to be larger, diverse and faster. Many technical challenges like implementations and visualizations are to be taken into consideration in future. This is just the survey paper which shows the demand of big data and how big companies are taking interest in it. We are at the beginning of a new era where big data mining will help us to discover knowledge that no one has discovered before. To manage and analyze edge data explore business opportunities deriving from the analytics of edge data. Collaborate with the business to understand existing edge system and the potential use for data. It can be concluded from the findings that Enterprise are still looking for the right infrastructure tools that will enable them to effectively handle their big-data, in line with their business needs. Most companies are already using dedicated big-data tools but all still see gaps in capabilities or have concern regarding the fit between these tools and their current and expected needs.

V. REFERENCES

[1] Accenture Big Success with Big Data Survey, (April2014).
[2] Anderson, J .Rainie, L. (July, 2012): The future of big data, the Pew Research Center's Internets American Life Project Series Pew internet.
[3] BARC_BIG_DATA_SURVEY_EN_final.
[4] Bernstein Philip et al. (1-1-2011): "Challenges and opportunities with big data".
[5] Bharti, Ramageri,"Data Mining Techniques and Applications, "Indian Journal of Science and Engineering, Vol.1 no-4, PP.301-305, Available: http://www.ijcse.com/docs/IJCSE10-01-04-51.pdf.
[6] Big-Data-Survey-Executive- Summary-110314-2014
[7] Big Data Fatigue (June23, 2014).
[8] Big Data Executive Survey (2013) Summary Report.
[9] Bottega, john (2014) Former CDO, Bank of America: "FinancialInformationSummit.com".
[10] Capgemini Consulting (November 2014)" Big Data Survey".
[11] Computer_Weekly.com (December2013)" Big Data, big legal trouble?
[12] Financial Services Companies See Results from Big Data Push (January 27, 2014).
[13] Goele, sangeeta, Nishachanana (2012):" Data Mining Trend in Past, Current and Future", International Journal of Computing & Business Research in Proc.I-Society2012.
[14] How Business Culture Defines Data Success (October 7, 2014).
[15] http://sites.tcs.com/big-data-study/industries-big-data-investment/TechAmerica Foundation.(2012).Demystifying big-data. Washington, DC.
[16] http://www.rcrwireless.com/2015 0415/big-data-analytics/dell-survey-yields-surprising-results-for-big-data-tag204/5.
[17] https://www.idgenterprise.com/report/big-data-2.
[18] http://www.gigaspaces.com

[19] IDG ENTERPRISE RESEARCH REPORTS, (JAN 6, 2014).
[20] Journal of Organization Design," Big Data and organization Design", (2014).
[21] Milan Big Data Keynote HMESSATFA Final, (2013).
[22] Nessi White Paper (December 2012)"Big Data a new world of opportunities".
[23] New Vantage Partner (NVP), (January 2013).
[24] Organizational Alignment is key to Big Data success, (January 28, 2013).
[25] The Emerging Big returns on big data, TCS-Big –Data-Global –Trend-study-2013, (March 21, 2013).
[26] The legacy of big data, (September 9, 2014).