Scientific Journal of Impact Factor (SJIF): 3.134
ISSN (Online): 2348-4470
ISSN (Print) : 2348-6406

International Journal of Advance Engineering and Research Development

Volume 1, Issue 11, November -2014

EVOLUTION AND CHALLENGES OF BIG DATA SYSTEMS

Meruva Srinadh[1], Majeti Naveen[2], Anuradha.T[3]

[1] BTech Student, Dept. of Electronics and Computers. K L University, Guntur, AP.
[2] BTech Student, Dept. of Electronics and Computers. K L University, Guntur, AP.
[3]Assoc. Professor, Dept. of Electronics and Computers. K L University, Guntur, AP
srinadh.meruva@gmail.com, naveenmajeti1994@gmail.com, anuradha_ecm@kluniversity.in

Abstract— *Exponential data growth from internet, low cost sensors, and high fidelity instruments has fueled the advanced analytics operating on vast data repositories. the reviews of the work, evolution and current state of Big data industry. The components and opportunities of today's business world are analyzed in many dimensions for big data and analytics i.e., Evolution of big data systems, big data collection, big data analysis, data visualization*

Index Terms— *Evolution of big data systems, big data collection, big data analysis, big data visualization.*

I. EVOLUTION OF BIG DATA SYSTEMS

1.1 First technology enabled "Snapshot" Data collection

This is fist big data problem in North America in 1890, it is first complied by using the methods invented by Herman Hollerith. In this method, data is entered on a machine-readable manner, a card which is punched and evaluated and tabulated by a machine.

In 1881 Herman Hollerith started to design a machine which can compile census data efficiently than the existing hand methods, and by 1880s, a technology had been introduced as punched card tabulating machine that could be read by electrical sensing. Every pin which is passed through the hole makes an electric contact with a small amount of cup of mercury which closes the circuit and increments the dail counter. This reduced the time required to manipulate the census of eight years can be done within one year. A total population census of 62,947,714 calculation is done within six weeks of time by using this technology. Later in 1924 changed its name to **International Business Machines Corporation (IBM).**

1.2 Social Security Act

Franklin D.Roosevelt the President of U.S launched the Social security Act in 1935. The act states that a Social Security number is to be provided and store employee records of 26 million Americans and 2 million employers. This program is started in 19th century by a European social welfare tradition**[2]**. This Social Security was first implemented in 1889 in Germany.

According to Houston chronicle**[3],** the job of collecting the 26 million applications is given to the Post Office Department. For these applications, Post office established 1,072 type institutes all over the country. The process is done as follows:

- A copy was sent to Internal Revenue Bureau
- One copy is sent to Wage Record Office.
- One copy was kept for the Post Office Records.

1.3 Colossus

During the World war II, British engineers developed a series of mass data-processing machines, culminating in the first programmable electronic computer: **The Colossus in 1943[1].**

These computers are developed and used during the World War II for cryptanalysis of Lorenz Cipher. This technology has given the British army to decode huge quantity of encrypted high-level telegraphic messages between German High Command and their troops. Colossus used vacuum tubes to perform Boolean operations and calculations and was the first electronic digital machine with programmability. The Technology of Colossus, had a significant influence on the development of early computers as helped to confirm the feasibility of reliable high-speed electronic digital computing devices**[4]**.

1.4 The U.S National Security Agency Backlog

During the 1950's, there are most challenges and concerns with the data related to the volume, later in 1960's there is development in automated data collection unveiled a second key component in big data i.e., Velocity.

In 1961, the U.S National Security Agency (NSA) faced a huge information overload due to automated intelligence collection and process during the saturation cold war. Besides having more than 12,000 cryptologists the agency

continuously struggled to digitize a backlog of records stored on analog magnetic tapes, just in July 1961, the agency received about 17,000 reels of tape**[1]**.

1.5 how to cope with the information explosion

In April 1964 Harry J. Gray and Henry Ruston publishes "Techniques for coping with the Information Explosion," in which they offer the following advice in order to cope with the increasing volume of information produced**[6]**.

- No one should publish any new papers. **[6]**
- If the above is not feasible, only short papers should be published. "Short" means not more than 2500 characters counting "space," punctuation marks, etc. as characters. **[6]**
- If the above is considered the following restriction should be applied: "Only those papers should be published which delete one or more existing papers which combined length is 2,501 characters or more", **[6]**

1.7 First Data Centre

In 1965, there is a growing problem of where to keep more than 742 million tax returns and 175 sets of fingerprints, the US federal government considered a plan to consolidate its data centers into a single mega-center. **[7]**

In 1965, a report was submitted to the office of statistical standards od the bureau of the budget entitled as "A Review of Proposals for a National Data Center." This analyzed the challenges which prevents the more effective usage of the resources of the Federal Statistical System in the establishment of public policy, the management of public affairs, and the research conduct.

Data Science [6]

In 1974 Peter Naur, a Danish innovator in computers and an award winner published "Conscise Survey of Computer Methods in Sweden and United States." This book inspects existing data process methods which are used in wide range of applications. This method defines data as "a Representation of facts or ideas in a formalized manner capable of being communicated or manipulated by some process." **[6]**

IASC [12]

The Inter-Agency Standing Committee (IASC) is the primary mechanism for inter-agency coordination of humanitarian assistance. It is a unique forum involving the key UN and non-UN humanitarian partners. It is founded with the mission of linking Traditional statistical methodology, latest computer technology, and the knowledge of experts in order to convert data to information and knowledge.

The Internet is Born [13]

In 1989, British scientist Tim Lee invented World Wide Web to provide data to share by "Hypertext".

Data Mining and Big Data

In March 1997, a journal named "Knowledge Discovery and Data Mining was started. This journal gives the new practice of data mining techniques and descriptions of appropriate applications.

Hadoop

Hadoop is an open-source software framework for distributed storage and distributed processing of Big Data on clusters of commodity hardware. Its Hadoop Distributed File System (HDFS) splits files into large blocks (default 64MB or 128MB) and distributes the blocks amongst the nodes in the cluster.(Wikipedia). Hadoop was originally developed to support for the Nutch search engine project by Micheal J. Cafarella and Doug Cutting who it after his son's toy elephant. **[8]**

Watson's Constraints

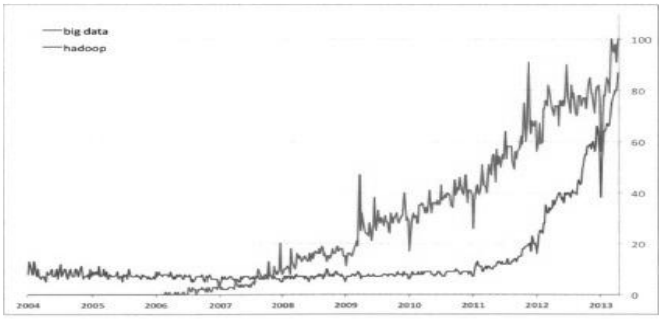
The most important person in the evolution of big data and technology is Watson. In 2011, Watson competed jeopardy on Brad Rutter, and Ken Jennings, winning the first prize of $1 million. **[10]**

To acquire the challenge, Watson concentrated on three key constraints: **[11]**

- Natural language processing.
- Hypothesis generation.
- Evidence-based learning.

What is Big Data?

Big data is collection of large data sets storage and complex that it is difficult to process using traditionalize data processing applications. This include analysis, capture, curation, search, sharing, storage, transfer, visualization, and privacy violations.



Google search frequency for Hadoop and Big Data

Source: Google Trends

II. BIG DATA COLLECTION

The Big Data Collection gives the information about the definitions and challenges that are involved in the big data collection and which considerations these big data aspects to have successful initiatives in big data organizations.
There are three essential features that defines the big data systems they are

- High Volume
- High Velocity
- High Variety

These three essential features makes the correct decision on how to data is to be collected and how the data is to be stored for the use of organization to make processed and analyzed .
Big data organizations have to provide the correct processes and correct technology implementation and skills for collecting the data and to provide the real time information in the big data systems so that the people can analyze the respective business environment in order to take right decisions. By this process the time can be saved for decision making process and we can get improved efficiency than our own decisions.
For Example, In tax auditing process the techniques of data mining are used i.e., regression trees. In this if we want to analyze and identify the irregularities of tax payers. In order to make decisions that whether to audit the individual or not, the automated decision is made by the system that the individuals to be audited and the final decision is left for the officials/humans for the decision to make.
Data shelf life**[14]** is the another important key point of data collection, by companies may be able to understand the some data is useful for short span of time.
To understand the big data systems and analyze the big data systems in business organizations we have to go deeper into the sources that are available with organizations and collection techniques used by organizations not only these we have to be aware of collection of right kind of data according to business requirement/ Organization requirements. The projects which costs Billions had to mention clearly the information and model of the data to be structured before the developing of project using data collection techniques.
Latest data collection techniques/technologies give a great attraction for business organizations and the data collection members have to keep in mind that the data collected is for the sake of the business organizations will provide only the technical capability but that process is the poor business strategy/judgment. The data that is collected is to be worth for the organization and it should be relevant between the both business terms and technical terms i.e., to make the correct data collections decisions.
In the future the data analysis will make the business process more and more competitive to design analytics which are not available in the past and it will be so hard for managers and business processors/organization processor to make this opportunity as an great advantage.
According to Gartner**[15]**, " The most important hurdle in order to get succeed in big data systems is to ask the right/correct question and therefore make use of right/correct technologies to get the answers". So we have to keep in mind that the big data collection is mean and the data that is collected is to be against the business objective.
Many organizations have lack of skills for exploiting the big data systems, so the organizations/business firms had to make required investments for hiring of the talented persons and also with advanced data management and analytical skills. Not only the hiring of the talented persons with analytical skills they also must train them with subject matter to become more attractive data and analytical consumers.
Gartner**[15]** said, that the big data systems are not the alternatives for the information technologies and they are not suitable for the fundamentals of information management. We have to keep in mind that the fundamentals that are existing in information managed programs are mostly challenged by the big data systems features i.e., volume, velocity and variety of big data systems.
If the companies/organizations have to get good quality of the information management system the organization should keep keen attention on basic principles of data quality and data management and information management which will cause the extra weight to data which is processed. The organizations/companies have to think alternative for the

expectations they require in quality and manage the context over the information and quality requirements must be used well in use case.

In the big data systems the solutions of these the big data systems are related to data and organized/managing is the key point to understand how the data will be resided in the databases. There are three basic forms of the data organization in databases they are:

The organizations/companies normally uses the databases like Non relational databases and structured databases.

Unstructured data[16]:

The data which is not in the fixed fields i.e., with specific format if the data is available that data is unstructured data. For example, untagged audio, image, video data etc.

Semi Structured data[16]:

The data which don't contain the fixed fields, it contains the tags and markers for separating the data elements . HTML and XML languages are examples for the semi structured data.

Structured Data[16]:

The data which will be resided in the fixed fields are known as the structured data . the structured data includes the relational databases or data that is in the spreadsheets. Relational databases are example of the structured data.

Relational Database[16]

Relational database consists of collection of tables which are stored in the form of the rows and columns. Initially these relational databases are developed at IBM, SQL(structured Query Language) is widely used for relational databases and it consists of Data definition and data manipulation languages which include insertion, updation, deletion, creation of schema and modification of data .

Non Relational databases[16]

These non relational databases are the databases that the data that is stored in this databases are not stored in the tables. social networking sites contain multi structured data and system created files which are sources of data but they don't have any relations between them so they are example for the relational database capabilities.

NoSQL databases/technologies provide access for the more easy usage of high volume applications like web applications in any circumstances. These technology had to go long way of development in small start-up too.

There are four solutions for the NoSQL databases**[23]** they are:

- Key value store
- Document style stores
- Table style databases
- Graph databases

The above four solutions of NoSQL databases are used in the particular use cases of specific capabilities that suited. We have to mind that the nosql products doesn't support classical applications and the nosql databases doesn't contain ACID[Atomcity, consistency, isolation, durability] properties that contain in Relational database management system.

The use of Nosql solutions may advance the performance of number of variety of applications**[23]**. They are:

- Internet commerce
- Mobile computing
- Social networking
- Machine to Machine communications

Vendors that design Nosql space are: Amazon, Neo Technology, Oracle, 10 gen etc are many companies that vendors the Nosql databases.

Various types of data has to be generated in big data systems like multimedia, video, image, audio, text/information**[16]**.

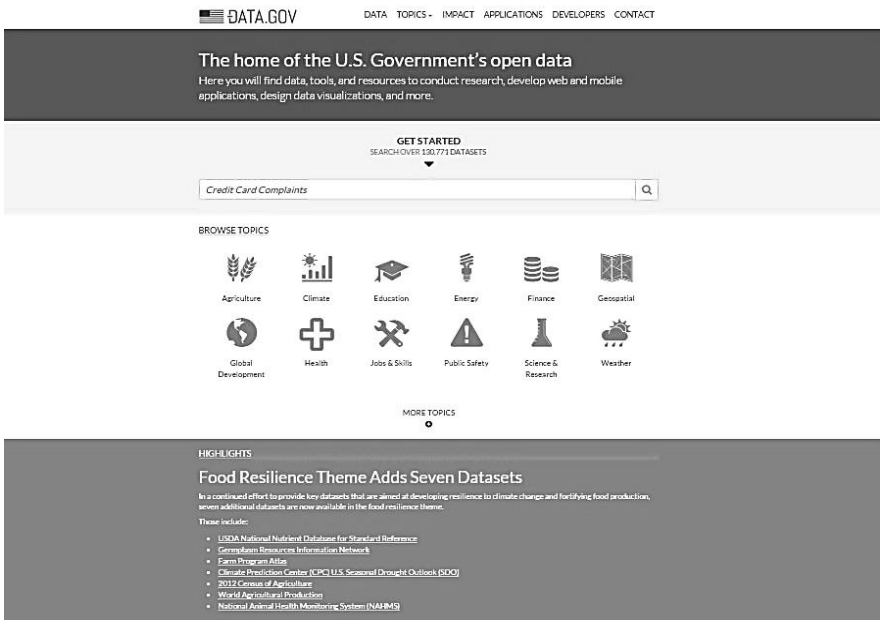
According to Deloitte**[17]** , there are three main key challenges that the data management have are:

- Quality
- Governance
- Privacy

Example of the big data collection systems

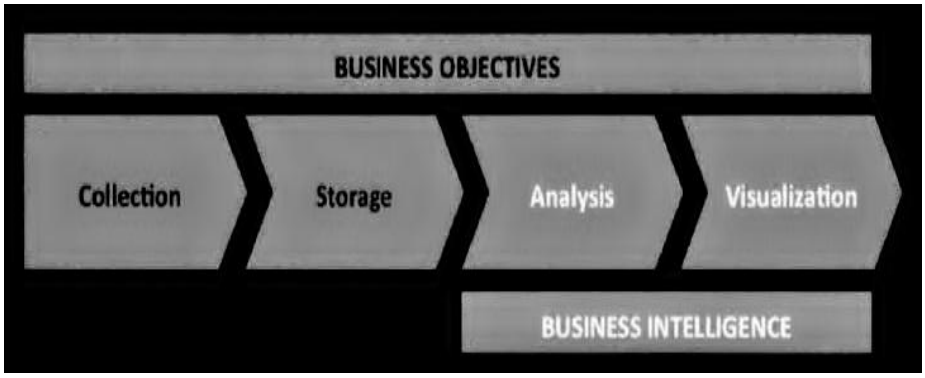
Public Sources of Data i.e., Data.gov

- Data.gov is the public depository of high value and datasets that are created by united states government.
- It increases the ability of people to search and download and use the metadata descriptions through datasets.
- The main aim of the Data.gov is to extend the creative usage of data on the walls of the government by appreciating the ideas and applications



III. BIG DATA ANALYSIS

So far we have discussed about big data collection and how to initialize the big data systems to fulfil the specific business objective of any organization**[15]**.
We have discussed that the challenges and opportunities that are with the data collection and storage of information. So far we have gone through just the one side of issue i.e., data collection and we have two more key components in the big data systems they are, Analysis and Visualization.



The Analysis and visualization will provide access to the organizations/business firms to provide correct insights and also to make understand and communicate these searching's to the respective decision makers.
Data analysis can be split into two main applications. They are

Descriptive analysis: The classical business reporting and dashboards presenting means etc will be more in the visualization.
Prediction and classification: The models which allow to classify or predict the value of variable which is based on other predictive variables.
These techniques are used in various sectors to make the specific insights, for example., micro targeting, this is a technique which is used for the bulk analysis and this technique is widely used in the marketing because it plays a key role in political campaigns for identifying the voters.

SEMMA Methodology[21]
Every technique has to follow the development and validation process in big data analysis, this methodology is developed by the SAS corporation and it is a better alternative for the development of the classification models. This methodology has to be given with sample data and visualization technique is applied and statistical techniques are also

applied on the data and to we have to select the data and transform the data into most predictive variables, model that variables predict the outcomes and affirm the model accuracy.
This semma methodology has 5 components they are:

Sample Sample is from data sets and make partition into training validation and test datasets. Training set is used for designing and construction of the model . validation set is used for the validating the ability of prediction of the model and make the adoption to model may get the required validation of model. Test sets are used to confirm the models predictive ability.
Explore The data sets contains the statistically and graphically data in order to get relationship among the variables.
Modify In Modify the transformation of the variables and assign the missing values in an order to make the data in correct format before modelling.
Model The model that is designed for the big data systems are the model i.e., classification etc.
Assess In this the models are compared using partition and databases are tested for this purpose.

Techniques used for analyzing Big data systems

There are many techniques which are provided access for the development of the data science. Data mining is one of the biggest contributors of the development of data science. This is more important to intimate that these techniques doesn't require big data sets.
Some of the techniques that are used for analyzing of big data systems are:

- K-Nearest Neighbors**[18][20]**
- Naïve Bayes**[18][20]**
- Classification and Regression Trees**[18][20]**
- Discriminant analysis**[18][20]**
- Logistic Regression**[18][20]**
- Neural Networks**[18][20]**
- Regression**[16][20]**
- Association Rules**[20]**
- Time series analysis**[20]**
- A/B testing**[16]**
- Spatial Analysis**[16]**
- Hierarchical cluster analysis **[20]**

IV. VISUALIZATION

Visualization is the conversion of data into a required format i.e., which can be visual or tabular format etc. so that the characteristics of the data can't be changed and the relationships among the attributes can be analyzed**[18]**.

There are three reasons for which the visualization of data is very strong and winsome techniques for the data exploration**[19]**. They are:

- It demonstrates complex concepts and representation of data in graphical mode.
- Generally Humans have a good ability to analyze the large amount of data by visual rather than any other way.
- Generally Humans can find the patterns that are visualized. So visualization provides the searching of relationships and to find the unusual patterns also.

Data visualization has several challenges for organizations. The challenges are**[22]**:
Speed To provide the visualization in high speed, the delivery speed will increase as complexity of relationship increases in business firms
Understanding the data The data that is provided in the systems are to be understandable. The visualization tool makers had to make sure that the data that is visualized is in right format and to make decision in organizations.
Displaying meaningful results Displaying the very high amount of data in various types can make the data processing and delivery very complex. Experts with team/business leaders must analyze and deliver the specific levels of granularity required for decision making.
The visualization of data process is most challenging issues in business organization/firms because it converts at very high levels of speed, processing capability, graphic communication and granularity.
Here we have presented the most different forms of data visualization, and it will be possible to search that the techniques of visualization are more complex than any other that is in given here. We have to notice that the new forms of data visualization that is available are being derived from the basic forms.

- Bar charts**[22]**
- Histogram
- Box plots**[18][22]**
- Scatter plots**[18][22]**
- Bubble plot**[22]**
- Correlation Matrix**[24]**

- Run chart**[18]**
- Star plots**[18]**
- Geo Map**[24]**

V. CONCLUSION

The evolution of the big data systems started from 1880 and till now it has not ended. Thus the big data collection and analysis of the big data systems has drastic effect on the business organizations/ firms in every day life . The 3 V's of the big data systems i.e., velocity, volume and variety has their major effect on the big data systems as the higher volumes of data are increasing the complexity of the data also increasing.. Data Analysis and Data visualization provide the advantage to the companies for big data systems. The analysis will provide the solutions to the big data questions and data that is visualized should be understandable to all the users.

REFERENCES

[1]. Uri Friedman,2012,Big Data: A short History, Foreign policy http://www.foreignpolicy.com/articles/2012/10/08/bigdata?print=yes&wploginredirect=0
[2]. The official website of U.S.Social Security Administration 2012, The social Insurance Movement, http://www.ssa.gov/history/briefhistory3.html
[3]. Luthe A.Huston, Dec 27,1936, Huge Machine Busy on Social Security, The Newyork times. http://query.nytimes.com/gst/abstract.html?res=F50C11FE305D12728DDDAEOA94DA415B8FF1D3#
[4]. Wikipedia,2013,colossus computer http://en.wikipedia.org/colossus(computer)
[5]. Wikipedia,2013,Fremont Rider http://en.wikipedia.org/wiki/FremontRider>
[6]. Gilpress,June-Sept 2012, A very short history of Bigdata, what's Big data? http://whatsthebigdata.com/2012/06/06/a-very-short-history-of-big-data/
[7]. Joe Mckendrick, June 2012, The Federal Governments fiat foray into cloud computing circa 1965,Forbes http://www.fobes.com/sites/jowmckendrick/2012/16/29/the-federal-government-first-foray-into-cloud-computing-circa-1965/
[8]. Wikipedia,2013, Apache hadoop, http://en.wikipedia.org/wiki/Apache-Hadoop#History
[9]. IBM,2013. What is Hadoop? http://www.0l.ibm.com/software/data/infosphere/hadoop/
[10]. Wikipedia, 2013, Watson http://en.wikipedia.org/wiki/watson(computer)
[11]. IBM 2013, The Science behind Watson http://www.03.ibm.com/innovations/us/watson/science-behind-watson.shtml
[12]. Gil press, April 2012, A very short history of data science, what is the big data? http://watsthebigdata.com/2012/04/06/a-very-short-story-of-data-science/
[13]. Mark van Rijmenam, Jan 2013, Big data history http://www.big-data-startups.com/big-data-history/
[14]. http://public.dhe.ibm.com/comman/ssi/ecm/en/iml14296usen/iml14396usen.pdf
[15]. Gartner Research, July 2012, Hype Cycle for Big data, 2012.
[16]. Mckinsey Global Institute 2011, Big data: The next frontier for innovation competition & productivity, Mckinsey &co
[17]. D.Steier, D.Marthias, Nov27,2012, The Backstory on Bigdata: what TMT Executives show know Deloitte.
[18]. Lecture Notes 16.062/ESd.7545 Data mining: Finding the data & models that create value, R. Welsch MIT fall 2012.
[19]. S.Rosenbush, M.Totty, March11,201, How bigdata is changing the education for business. The wall street journal.
[20]. G.Shueli, N.Patel, D.Bruce, 2011, Data mining for Business intelligence wiley.
[21]. SAS Institute Inc.,2012, SAS Enterprise Miner: SEMMA http://www.sas.com/offices/europe/uk/technologies/analytics/datamining/miner/semma.html
[22]. SAS Institite Inc., 2012, Five Big data challenges http://www.SAS.com/resources/asset/106008_5Bigdata_Final.pdf
[23]. Meru Adrian, June 7,2012, who's who in NOSQL DBMS, Gartner, Inc
[24]. SAS Institite Inc., 2012, Data Visualization: common charts and Graphs http://www.sas.com/data-visualization/common-charts.html

