A Statistical Survey on Imperial Data Handling in Big Data

T. Princess Raichel[1], S.Kokila2, N.Sowmya3
Assistant Professor[1], Assistant Professor[2], M.Tech Scholar[3]
Department of Computer Science and Engineering
Sreenivasa Intitute of Technology and Management Studies
Chittoor, Andhra Pradesh,India

ABSTRACT: In 2010 the world has generated 1ZB of data and expected that it will generate 7ZB in 2014.So that it will generate devices of network that includes Embedded Sensors, Smart Phones & Tablet computers it will be an opportunity in human genomics,healthcare,economical,cultural,oil and gas,political stage,surveillance,finance.Big Data technologies describe a new generation of Technologies & Architecture.Big data requires a change in computing Architecture to customers so that they can handle data storage and server processing heavily. Most of the companies reply on applications for communication and to provide service to the customers. So big data is a big challenge for companies which deal with large data and fast growing information.Big data has a direct impact on Applications, Services and Software technologies in the view of Technical,Legal,Social & market related aspects.The data storage is a major issue for owners, so efficient and scalable technology for data management and storage is no longer issue in big data. The data is generated on daily basis by all the sectors in the whole world.

Index Terms: - Big Data, Natural Compositions, Survey of Data

I. INTRODUCTION

Big information is a sweeping term for any gathering of information sets so expansive and complex that it gets to be hard ProcessThem utilizing customary information preparing applications.

The difficulties incorporate examination, catch, curtain, hunt, offering, stockpiling, exchange, visualization, and protection infringement. The pattern to bigger information sets is because of the extra data logical from examination of a solitary substantial set of related information, as contrasted with particular littler sets with the same aggregate sum of information, permitting connections to be found to "spot business patterns, counteract infections, battle wrongdoing.

Big data is difficult to work with using most relational database management systems and desktop statistics and visualization packages, requiring instead "massively parallel software running on tens, hundreds, or even thousands of servers". What is considered "big data" varies depending on the capabilities of the organization managing the set, and on the capabilities of the applications that are traditionally used to process and analyze the data set in its domain. Big Data is a moving target; what is considered to be "Big" today will not be so years ahead. "For some organizations, facing hundreds of gigabytes of data for the first time may trigger a need to reconsider data management options. For others, it may take tens or hundreds of terabytes before data size becomes a significant consideration

II. SURVEY METHODOLOGY IN VARIOUS SECTORS

I) POLITICAL SECTOR

In the present scenario Big Data has become a popular in major countries, also Big Data has business opportunities and research challenges. It also act as boundary between the countries also it has become potential for competition and growth of individual companies.

Big Data can increase the productivity, innovation and competition for all the sectors. The benefits can be known if we know how to use Big Data like intelligence, Managerial activity and reusability of data like private and public sectors. The process is very easy still we need to analyze both side because it is an open data it can also be a threat for privacy.

II) RESEARCH

During the research the processing of data is vast like creation of content and data processing for Analytics and Real Time processing.

In 2020 Big Data will be in lead in industries like Technology, Information Management, and Social to structure the data. The data has to be utilized efficiently.

III) ECONOMIC & SOCIAL IMPACT

The most developed countries have the big potential to create value through the use of Big Data. Big Data has created the impact on all the organizations because it extracts the embedded knowledge from large amount of data in a powerful way. Technologies produce data in Real Time & Space. The data's are gathered by sensors which integrate Transport Data, Financial Transactions, use location, Social Network interaction.

IV) PRIVACY

Big Data has systematic collection, storage and analysis is increases because of personal data. From the internet details of the user can be accessed through surveillance. There are tools which are used to identify and track the information about user.

Using Mobile phones the user location is stored in private. Also using Debit and Credit card payment the amount spent will be known. Using loyalty cards consumer details are stored. Through social media user to user contact and their pictures, videos, movie can be accessed.

Also Data Mining patterns are searched in large collection of personal data to match the individuals to predict preferences and interests. In the in depth privacy by design scheme, Big Data applications can bring strong safeguards for data protection.

This must be non-abusing,liberality and notorious non-compliant with data protection rules.

V) THREATS

Globally there is a fast grow in knowledgeabout the use of data.

III. INDUSTRY SERVICES ON THE CONTEXT OF BIGDATA STRENGTH

A) STRENGTHS

There may be a deep knowledge about markets and customers locally. There may be problem and ability to develop customized products such as language dependent problems, legislation dependent products.

B) WEAKNESS

Google, Yahoo, Twitter are widely recognized their activities on Big Data.There is no knowledge about available data.

C) OUTCOME

Rise in demand for small products increase the individualism and customs. The benefits can be enhanced in products and services with Business Development Administration and secrecy by design. Also benefits can be reduced with networked products and services. The data has to be available freely. Using the standards the new products and service should be developed with Business Development Administration

D) BUSINESS ECO SYSTEM

It is a community supported by an organization and individual. The Business Development Administration does not exist in most of the industries. The data in the world is generated and processed so quickly. Also data science and skills associated with that are in huge demand. The education system not to teach students basic skills also programming and design skills.

The machine learning is needed to create Business Eco System environment. Also effective machine learning requires academic background in order to convert mathematical knowledge in to marketable knowledge.

The Big Data plays important role in commodity like the improvement in Technology, Infrastructure, Database, and Communication. The Big Data is accessible for developers by making it as easy to create applications.

IV. IMPORTANCE OF BIG DATA

The Big Data is combined with traditional enterprise data it can develop more understanding of the business which can develop the productivity like Healthcare services, manufacturing companies, Automotive Industry reveals usage patterns, failure rates for product improvement that reduces the development

V. ORGANIZATION OF BIG DATA

Data integrity is called as organization because there is high volume of Big Data also there is a tendency to organize data initially so that Time and Money is saved. There should be ability to organize the data that process and manipulate data in original location. So using Hadoop Tool the data's are organized and processed by keeping the original data storage in cluster. Hadoop Distributed File System (HDFS) are used to store weblogs for long time.

VI. CHALLENGES IN BIG DATA

The best strategies for a company are the data must be properly analyzed. The time span is important because some analysis need to be performed very frequently in order to determine fast change in business.

Also new technologies are developed every day. Nowadays the term Big Data is new to the companies. Also it is necessary for the organization to learn how to use the newly developed technologies when they are about to enter the market. This will bring competitive advantage to the business. There is a need for an IT specialist as a challenge. The company has to take Big Data initiative by either hiring the experts or training the existing employees in the new era.

VII. ISSUES IN BIG DATA

In fundamental area need to be addressed in dealing with Big Data like Storage Processing & Management. Each represents a large set of technical research problems.

A) STORAGE & TRANSPORT ISSUES

The data quantity is blows up each time, so new invention is found for new storage medium. The difference about the most recent explosion is because of large data from social media that has been no new storage medium.

Moreover data's are generated by everyone & everything like professionals, scientist, journalist & writers etc. The data could be processed on a single computer system, so it is not possible to directly to attach the required number of disks. To access the data would defeat the current communication. Always sustained transfer could be maintained, it takes larger time to transmit the data from a collection point to the processing point then it actually processes it. Always process the data in right place and transmit only the resulting information. The code has to be brought out to the data. Also has to perform triage up on the data and transmit only the data which is crucial to downstream analysis. In few cases integrity and provenance Meta data has to be transmitted in along with the actual data.

B) MANAGEMENT ISSUES

In the management the most difficult problems is to identity the Big Data like data will be distributed geographically by owning it or by managing it with multiple entities by resolving the issues of access,update,metadata,utilize,governance and reference has proven.

During the collection of e-data in manual methods the unpleasant protocols are often followed in order to ensure accuracy and validity,digitalized data represents a methodology for data collection. The qualified data focuses more on missed data than the available data item. Always the data is finely grained like click stream.

It is impossible to validate every data item. The sources for these kind of data's are varied both temporary manner and spatial manner. Individual person contribute digital data in some medium like document,drawings,pictures,sounds,video recording,models,software behavior's,UI designs.

All these data's are available for inspection and analysis.

C) PROCESSING ISSUES

There is no perfect Big Data management solution yet. This yields important gap in the research literature on Big Data that has to be fulfilled. The effective processing of data requires extensive parallel processing and new analytics algorithms in order to provide make that data readily available timely and actionable information.

VIII.CHALLENGES IN BIG DATA DESIGN

There are so many challenges for long term in research to work with Big Data. The design for systems and components that works with Big Data will require an understanding of user and technology need to solve the problem for some instance. Since the data is newly created is truly known or understood designers need to consider graphical interfaces and icons.

Organized application models using conceptual, metaphors and functionality. Most of the end users will not be a system designer that will be the design challenge. The unknown challenge will increase in scale and development of new products. So the challenges will increase the size data sets.

I) INPUT AND OUTPUT PROCESS CHALLENGE

The major issue in Big Data design is in output process. Like it is easy to get the input data the output data. The data entry and storage can be handled with processes currently used for RDBMS.

The tools designed for transaction processing that Add, Update, Search, and Retrieve small to large amounts of data that capable of extracting the huge volumes and also it cannot be executed in seconds. To access very large quantities of semi or unstructured data and utilization tool is not designed yet. The problem may neither be solved by dimensional modeling and Online Analytical Processing (OLAP) which is slow and limited functionality. Technical factors must be considered in to design to include the speed of random memory access.

II) QUALITY VS QUANTITY

The big challenge for Big Data user is quality vs. quantity users,users have to access more data (i.e.) quantity also want more. Some users acquire data additionally, because they believe that enough data will be able to explain in which they are interested. Mostly Big Data use may focus on quality which means that data's are not available but having very large quantity of high quality data.

III) GROWTH OF DATA VS EXPANSION OF DATA

Many organization expect that their data to grow in lifetime. So that organization increases its Client, Business, Services, Partners, Projects and its employees. So some business sufficiently requires data expansion which will happen when the data's grow in richness.

When the world evolves over time we need some information additionally like new techniques, process and evolve information demands. Most of the data varies in time same type of data can be collected with different values by means of Time Stamp.

Most of the data is required for analysis particularly which is used to estimate predictive analysis.

IV) SCALE VS SPEED

When the volume of the data grows the scale of the data to the amount of data that can be processed in a given period. Insight of problem being analyzed it is more important than the processing of the data. An organization must find out how much data is needed in setting its processing because this will drive the processing system architecture, the characteristics of the computation engines and the structure of the algorithm and implementation. Data dissemination is the communication middleware because hardware speeds are increasing with new technologies, handling message speed decreases slowly.

V) STRUCTURES VS UNSTRUCTURED DATA

Translation between data with data definitions as stored in RDBMS and unstructured data suitable for analysis can prevent processing performance. To come out of non-relationaldistributed, analytics oriented databases such as NoSql,Mongo DB,SciDB,linked data DBS provides dynamic flexibility in representing and organizing information. Data source has no Beginning and no ending. Data streams are not so well behaved and often experience is unrelated to the primary data of interest.

VI) DATA OWNERSHIP

Data ownership is a critical and a challenge in social media with social media sites, there is a trend in **Big** Data analytics towards the first party with verified data for public and third part external data with not been validated and verified by any methodology. The addition of unverified data compromises the data set with non-relevant entity and may lead linkage among entities. The accuracy is drawn from processing and mixed data varies widely.

VII) SECURITY & COMPLIANT

In certain domains the data is accumulated about the individuals in few organizations will know more about the individuals. Developing an algorithm will randomize personal data among large data set to ensure privacy. The biggest problem to personal security is unregulated accumulation of data by some companies. The data represents a concern especially individuals surrender such information.

Big Data must be secured with respect to privacy and security laws and regulations. The research is required clearly to define the security levels and map them against current laws and analysis.

VIII) VALUES OF THE DATA

Data is created equal some data is more valuable the other data like temporally,spatially,contextually etc.storage limitation require Data Filtering and deciding to declare what data to keep. The new mechanism for converting unstructured test.

Latent, image or audio information in to numerical indicators to make them to traceable with Big Data that analyze will emerge over time.

At any point of time the amount of data need to analysesespecially that represent only a small fraction of all data source else most data will go analyzed.

IX) DISTRIBUTED DATA & PROCESSING

The replication of hardware and system expandable is represented on cloud computing along with the map reducer and Message Passing Interface. The significant performance can still occur just because of the nodes communication.

If distributed processing is an alternative approach, the system reliability will increase to assure that no simple node fails. We need to ensure fault tolerance for hardware, system software up to some extent for the algorithms in user application software.

COMPARISON TABLE FOR CHARACTERISTICS OF BIG DATA WITH OTHER TECHNOLOGIES

Spec/name	Big data	Apache accumulo	Apache hadoop	Cloud era	Map reduce	Tuple space
Volume	Varies	Constant	Constant	Constant	Constant	Constant
Velocity	Speed ≥Magnitude	Speed ≥Magnitude	Speed= Magnitude	Speed=Magnitude	Speed ≤Magnitude	Speed= Magnitude
Variety	All types of Data	All Types of Data	Restricted	All Type of Data	Restricted	Restricted
Value	Large	Medium	Medium	Medium	Large	Large
Complexity	O(n)	O(n)	$O\sqrt{n}$	$O\sqrt{n}$	$O\sqrt{n}$	$O(n)^2$

Table 1. Comparison table for Characteristic of Big Data with Other Technologies

STATISTICAL TABLE FOR THE BIG DATA COMPARED WITH OTHER COMPLIANCES

Spec/Name	Big Data	Apache Accumulo	Apache Hadoop	Cloud Era	Map Reduce	Tuple Space
Input & Output Design	√	×	√	√	√	×
Quality vs. Quantity	×	√	√	√	×	√
Data vs. Expansion of Data	√	√	×	∞	√	×
Scale vs. Speed	√	∞	×	√	∞	∞
Structured vs. Unstructured Data	∞	∞	×	∞	∞	∞
Data Ownership	√	√	∞	√	√	√
Security & Compliant	×	√	√	√	∞	√
Values of Data	√	×	√	∞	√	×
Distributed & Data Processing	√	×	√	√	×	√

Satisfactory - √

Unsatisfactory - ×

No Result - ∞

Table 2. Statistical Survey of Big data with other Compliances

IX. CONCLUSION

In 2010 the world has produced 1zb of information and expected that it will create 7zb in 2014.so that it will produce gadgets of system that incorporates Embedded Sensors, Smart Phones & Tablet computers it will be an open door in human genomics,healthcare,economical,cultural,oil and gas,political stage,surveillance,finance.big Data innovations portray another era of Technologies & Architecture.big information obliges a change in figuring Architecture to clients so they can deal with information stockpiling and server transforming vigorously. The vast majority of the organizations answer on applications for correspondence and to give administration to the clients. So enormous information is a huge test for organizations which bargains with extensive information and quickly developing information. Big information has an immediate effect on Applications, Services and Software innovations in the perspective of Technical,legal,social & business sector related aspects.the information stockpiling is a significant issue for owners,so effective and adaptable innovation for information administration and capacity is no more issue in Big information. The information is created on everyday schedule by all the parts in the entire world

X. REFERENCES

[1] MIKE 2.0, Big Data Definition, http://mike2.openmethodology.org/wiki/Big Data Definition

[2] P. Zikipoulos, T. Deutsch, D. Deroos, Harness the Power of Big Data, 2012, http://www.ibmbigdatahub.com/blog/harness-power-big-data-book-excerpt

[3] Gartner, Big Data Definition,http://www.gartner.com/it-glossary/big-data/

[4]E. Dumhill, "What is big data?", 2012 ,http://strata.oreilly.com/2012/01/what-is-big-data.html

[5] A Navint Partners White Paper, "Why is BIG Data Important?" May 2012, http://www.navint.com/images/Big.Data.pdf

[6] Greenplum. A unified engine for RDBMS and Map Reduce, 2009. http://www.greenplum.com/resources/mapreDuce/.

[7]"Data Analysis Challenges" by the JASON study group (JSR- 08- 142, December 2008), available at http: // www .fas .org / irp / agency / dod / jason / data .pdf

[8] Hastie, Trevor, Robert Tibshirani, and Jerome Friedman. 2008. The Elements of Statistical Learning: Data Mining, Inference and Prediction. New York: Springer-Verlag

[9] Piketty, Thomas, and Emmanuel Saez. 2003. "Income Inequality in the United States, 1913–1998." Quarterly Journal of Economics118 (1): 1–39.

[10]Scott, Steve, and Hal Varian. 2013. "Bayesian Variable Selection for Nowcasting Economic Time Series." ASSA Annual Meeting, San Diego, CA, Presentation Overheads

AUTHOR PROFILE

T. Princess Raichel has recieved her B.E (CSE) in 2006 from Anna University, Chennai. She finished her M.Tech (CSE) from VelTech University, Chennai, and currently she works as Asst.Professor in SITAMS, Chittoor. Her Area of Interest is Big Data and Software Engineering.

S.Kokila has received her B.Tech (CSE) in 2008 from JNTU University, Ananthapur. She received her M.Tech (CSE) in 2011 from the Same University. Currently she works as Asst. Professor in SITAMS, Chittoor. Her area of interest is Web Technology and DBMS

N.Sowmya has received her B.Tech (CSE) in 2010 from JNTU University, Ananthapur. Currently she is a M.Tech Scholar in the same university. Her area of Interest is Big Data and Mongo DB