BIG DATA CHALLENGES IN GENERATING RIGHT METADATA

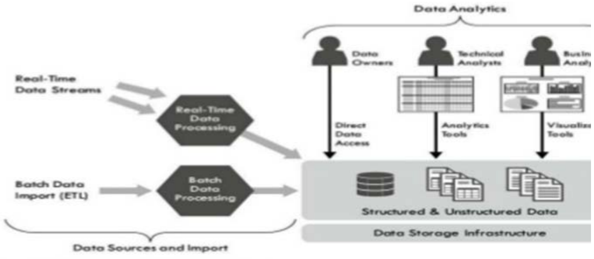
1.K.Hema,Assistant professor, ,MCA Dept., KMM College, 2.K.Hima Bindhu, Btech-IInd Year, CSE Dept., KMMITS,Tirupathi 3. B Bindhu Madhavi,. Btech-IInd Year, CSE Dept., KMMITS,Tirupathi.

Abstract: Big data is a term for massive data sets having large, more varied and complex structure with the difficulties of storing, analyzing and visualizing for further processes or results. The process of research into massive amounts of data to reveal hidden patterns and secret correlations named as big data analytics .These useful information's for companies or organizations with the help of gaining richer and deeper insights and getting an advantage over the competition. It refers to the data sets that are too big to be handled using the existing database management tools, are emerging in many important applications, such as Internet search, business informatics, social networks, social media, genomics, and meteorology. Big data presents a grand challenge for database and data analytics research. The primary purpose of this paper is to provide an in-depth analysis of different platforms available for performing big data analytics. This paper surveys different hardware platforms available for big data analytics and assesses the advantages and drawbacks of each of these platforms based on various metrics such as scalability, data I/O rate, fault tolerance, real-time processing, data size supported and iterative task support.For this reason, big data implementations need to be analyzed and executed as accurately as possible.

Keywords: Big data; MapReduce; graphics processing units; scalability; big data analytics; big data platforms; k-means clustering; real-time processing.

1. INTRODUCTION:

Big data is defines as large amount of data which requires new technologies and architecture to make possible to extract value from it by capturing and analysis process. New sources of big data include location specific data arising from traffic management, and from the tracking of personal devices such as smart phones. Big data has emerged because we are living in a society which makes increasing use of data intensive technology. Due to such large size of data it becomes very difficult to perform effective analysis using the existing traditional technology. Since big data is a recent upcoming technology in the market which can bring huge benefits to the business organization, it becomes necessary that various challenges and issues associated in bringing and adapting to these technology are needed to be understood. Big data concepts means a datasets which continues to grow so much that it becomes difficult to manage it using existing database management concept & tools. The difficulties can be related to data capture, storage, search, sharing, analytics and visualization etc.



Example of Big Data Architecture (Aveksa Inc., 2013)

Big data due to its various properties like volume, velocity, variety, variability, value and complexity put forward many challenges. The various challenges faced in large data management include-scalability, unstructured data, accessibility, real time analytics, fault tolerance and many more. In addition to variations in the amount of data stored in different sectors, the types of data generated and stored-i.e encoded video, images, audio, or text/numeric information; also differ markedly from industry to industry.

Big data is a relatively new phenomenon. As with any new adoption, the adoption of big data depends on the tangible benefits it provides to business. Large data sets which are considered as information overload are invariably treasure troves for business insights. The volume of data sets has immense value that can improve the business forecast, help in decision making, deciding business strategies over the competitors. For instance, facebook, blogs and twitter data gives insights on current business trends.

The data sets are beyond the capabilities of humans to analyze manually. Big data tools have the ability to run ad-hoc queries against the large data set in less time with a responsible performance. For instance, in retail domain understanding what makes a buyer to look into a product online, sentiment analysis of a product based on the facebook, twitter and blogs are of great value to the business. This will enable the business to improve their services for customers.

Big data analysis enables the executives to get the relevant data in less time for making decisions. Big data can pave way for fraudulent analysis, customer segmentation based on the store behaviour analysis, loyalty programs that identifies and targets the customers. This enables us to perform innovative analysis which indeed changes the way we think about data.

2. EXPLORING BIG DATA SPECTRUM:

With unstructured data dominating the world of data, the way to exploit it is just becoming clearer. Information proliferation is playing a vital role in leveraging the opportunities and is also presenting a plethora of challenges.

The industry opportunities presented by the plethora of data are plenty. To understand how to leverage big data opportunities is clear need to a business. Big data spectrum covers use case from five different industries Retails, Airlines, Auto, Financial services and Energy.

All opportunities come with a set of challenges. The way to know and address these challenges is discussed in the key challenges section. To name a few: Data privacy, Data security, integrating various technologies, catering to real time floe of data and leveraging cloud computing.

3. ANALYSIS OF BIG DATA:

The analysis of big data involves multiple distinct phases as shown in the figure below, each of which introduces challenges. Many people unfortunately focus just on the analysis/modelling phases: while that phase is crucial, it is of little use without the other phase of the data analysis pipeline. Even in the analysis phase, which has received much attention, there are poorly understood complexities in the context of multi-tenanted clusters where several users programs run concurrently. Many significant challenges extend beyond the analysis phase.

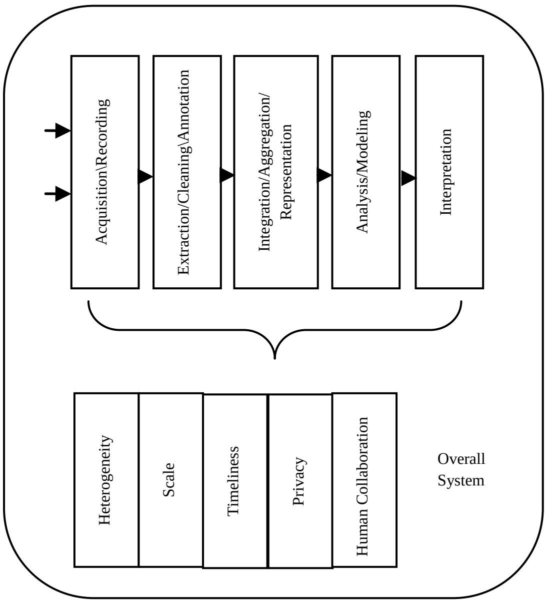


Fig: System Analysis of Big Data

4. PHASES IN THE PROCESSING PIPELINE: DATA ACQUISITION AND RECORDING:

Big data does not arise out of a vacuum: it is recorded from data generating source. For example, consider our ability to sense and observe the world around us, from the heart rate of an elderly citizen, and presence of toxins in the air we breathe, to the planned square kilometre array telescope, which will produce up to 1 million terabytes of raw data per day. Similarly scientific experiments and simulations can easily produce of petabytes of data today.

Much of this data is of no interest, and it can be filtered and compressed by orders of magnitude. One challenge is to define these filters in such a way that they do not discard useful information.

The second big challenge is to automatically generate the right metadata to describe what data is recorded and how it is recorded and measured.

Information extraction and cleaning:

Frequently, the information collected will not be in a format ready for analysis. For example consider the collection of electronic health records in a hospital, comprising transcribed dictations from several physicians, structured data from sensors and measurements and images data such as x-rays. We cannot leave data in this form and still effectively analyze it.

Rather we require an information extraction process that pulls out the required information from the underlining sources and express it in a structured form suitable for analysis. Doing these correctly and completely is a continuing technical challenge. Note that this data also includes images and will in the future include video; such extraction is often highly application dependent. In addition, due to the ubiquity of surveillances camera and popularity of GPS enabled mobile phones, cameras and other portable devices, rich and high fidelity location and trajectory data can also be extracted.

Data Integration, Aggregation and representation:

Given the heterogeneity of the flood of data, it is not enough merely to record it and throw it into a repository. Data analysis is considerably more challenging than simply locating, identifying, understanding, and citing data. For effective large-scale analysis all of this has to happen in a completely automated manner. This requires differences in data structure and semantics to expressed in forms that are computer understandable, and then "robotically" resolvable. There is a strong body of work in data integration that can provide some of the answers. However, considerable additional work is required to achieve automated error-free difference resolutions.

Even for simpler analyses that depend on only one data set, there remains an important question for suitable database design. Usually, there will be many alternative ways in which to store the same information. Certain design will have advantages over others for certain purposes, and possibly drawbacks for other purposes.

Query processing, Data modelling, and Analysis:

Methods for querying and mining big data are fundamentally different from traditional statistical analysis on small sample. Big data is often noisy, dynamic, heterogeneous, inter-related and untrustworthy. Nevertheless, even noisy big data could be more valuable than tiny samples because general statistics obtained from frequent patterns and correlation analysis usually over power individual fluctuations and often disclose more reliable hidden patterns and knowledge. Further interconnected big data forms heterogeneous information network, with which information redundancy can be explored to compensate for missing data, to crosscheck conflicting cases, to validate trustworthy relationships, to disclose inherent cluster, and to uncover hidden relationships and models.

Mining requires integrated, cleaned, trustworthy, and effectively accessible data declarative query and mining interfaces, scalable mining algorithms, and big data computing environments. At the same time, data mining itself can also be used to help improve the quality and trustworthiness of the data, understand its semantics, and provide intelligent querying function.

Interpretation:

Having the ability to analyze big data is of limited value if users cannot understand the analysis. Ultimately a decision maker, provided with the result of analysis, has to interpret these results. This interpretation cannot happen in a vacuum. Usually, it involves examining all the assumptions made and retracing the analysis. Furthermore, as we saw above, there are, many possible sources of error: computer system can have bugs, models almost always have assumptions, and result can be based on erroneous data. For all of these reasons, no responsible user will cede authority to the computer system. Rather she will try to understand, and verify, the results produced by the customers. The computer system must make it easy for her to do so. This is particularly a challenge with big data due to its complexities. There are often crucial assumptions behind the data recorded. Analytical pipeline can often involve multiple steps, again with assumptions built in. The recent mortgage-related shock to the financial system dramatically underscored the need for such decision –making diligence rather than accept the stated solvency of a financial institution at a face value; a decision-maker has to examine critically many assumptions at multiple stages of analysis.

In short, it is rarely enough to provide just results. Rather, one must provide supplementary information that explains how each result was derived, and based upon precisely what inputs. Such supplementary information is called the Provenance of the data.

5. BIG DATA CHARACTERISTICS:

Data Volume:

The big word in big data itself defines the volume. At present the data existing is in petabytes (10^{15}) and is supposed to increase to zettabytes (10^{21}) in nearby future. Data volume measures the amount of data available to an organization, which does not necessarily have own all of it as long as it can access it.

Data Velocity:

Velocity in big data is concept which deals with the speed of the data coming from the different sources. This characteristic is not being limited to the speed of incoming data but also speed at which the data flows and aggregated.

Data Variety:

Data variety is a measure of the richness of the data representation – text, images, video, audio etc. Data being produced is not of single category as it not only includes the traditional data but also the semi structured data from various resources like web pages, web log files, social media sites, E-mail, documents.

Data value:

Data value measures the usefulness of data in making decision. Data science is exploratory and use in getting to know the data, but “analytic science” encompasses the predictive power of big data. User can run certain queries against the data stored and thus can deduct important results from the filtered data obtained and can also rank it according to the dimension they required. These reports help these people to find the business trends according to which they can change their strategies.

Complexity:

Complexity measures the degree of interconnectedness (possibly very large) and interdependence in big data structures such that a small change (or combination of small changes) in one or a few elements can yield very large changes or a small change that ripple across or cascade through the system and substantially affect its behaviour , or no change at all.

6. CHALLENGES IN BIG DATA:

The challenges in big data are usually the real implements hurdles which require immediate attention. Any implementation without handling these challenges may lead to the failure of the technology implementation and some unpleasant result.

Privacy and Security:

It is the most important challenges with the big data which is sensitive and includes conceptual, technical as well as legal significance.

- The personal information of a person when combined with external large data set, leads to the interface of new facts about that person and it’s possible that these kinds of facts about the person are secretive and the person might not want the data owner to know or any person to know about them.
- Information regarding the people is collected and used in order to add value to the business of the organization. This is done by creating insights in their lives which they are unaware of.
- Another important consequence arising would be social stratification where alliterate person would be taking advantage of the big data predictive analysis and on the other hand underprivileged will be easily identified and treated worse.
- Big data used by the law enforcement will increase the change of certain tagged people to suffer from adverse consequences without the ability to fight back or eve having knowledge that they are being discriminated.

Data Access and Sharing of information:

If the data in the companies information systems is to be used to make accurate decisions in time it becomes necessary that it should be available in accurate, complete and timely manner. This makes the data management and governance process bit complex adding the necessity to make data open and make available to government agencies in standardized manner with standardized API’s, metadata and format that leading to better decision making, business intelligence and productivity improvements.

Expecting sharing of data between companies is awkward because of the need to get an edge in business. Sharing data about their clients and operations threatens the culture of secrecy and competitiveness.

Analytical challenges:

The main challenging questions are:

- What if data volume get so large and varied and it is not known how deal with it?
- Does all the data needed to be stored?
- Does all the data to be analyzed?
- How to find out which data points are really important?
- How can the data be used to best advantage?

Big data brings along with it some huge analytical challenges. The type analysis to be done on this huge amount of data which can be unstructured, semi structured and structured requires a large number of advance skills. More over the type of analysis which is needed to be done on the data depends highly on the result to be obtained i.e. decision making. This can do by using one of two techniques: either incorporates massive data volumes in analysis or determine upfront which big data is relevant.

Human Resources and Manpower:

Since big data is at its youth and an emerging technology so it needs to attract organizations and youth with diverse new skill sets. These skills should not be limited to technical ones but also should extend to research, analytical, interpretive and creative ones. These skills need to be developed in individual hence requires training programs to be held by the organization. Moreover the universities need to introduce curriculum on big data to produce skilled employees in these enterprise.

Technical Challenges:

Fault Tolerance:

With the incoming of new technologies like cloud computing and big data it is always intended that whenever the failure occurs the damage done should be within acceptable threshold rather than beginning the whole task from the scratch.

Fault-tolerance computing is extremely hard involving intricate algorithms. It is simply not possible to device absolutely foolproof. 100% reliable fault tolerance machine or software. Thus the main task is to reduce the portability of failure to an acceptable level. Unfortunately, the more we strive to reduce this portability, the higher the cost.

Two methods which seem to increase the fault tolerance in big data are as:

- First is to divide whole computation being done into tasks and assign these tasks to different nodes for computation.
- Second is, one node is assigned the work of observing that these nodes are working properly. If something's happens that particular task is restarted.

But sometimes it's quite possible that the whole computation can't be divided into such independent tasks. There could be some tasks which might be recursive in nature and the output of the previous computation of task is the input to the next computation. Thus restarting the whole computation becomes cumbersome process. This can be avoided by applying checkpoint which keeps the start of the system at certain intervals of the time. In case of any failure, the computation can restart from the last checkpoint maintained.

Scalability:

The scalability issue of big data has lead towards cloud computing, which now aggregates multiple disparate workloads with varying performance goals into very large clusters. This requires a high level of sharing of resources which is expensive and also brings with it various challenges like how to run and execute various jobs so that we can meet the goals of each workload cost effectively. It also required dealing with the system failure in an effective manner which occurs more frequently if operating on large clusters. These factors combined put the concern on how to express the programs, even complex machine learning tasks. There has been a huge shift in the technologies being used. Hard disk drives (HDD) are being replaced by the solid state drives and phases change technology which is not having the same performance between sequential and random data transfer. Thus, what kinds of storage device are to be used; is again a big question for data storage.

Quality of Data:

Collection of huge amount of data and its storage comes at a cost. More data if used for decision making or for predictive analysis in business will be definitely leads to better result. Business leaders will always want more and more data storage whereas the IT leaders will take all technical accepts in mind before storing all the data. Big data basically focus on quality data storage rather than having very large irrelevant data so that better result and conclusion can be drawn.

This further leads to various questions like how it can ensured that which data is relevant, how much data would be enough for decision making and whether the stored data is accurate are not to drawn conclusion from it etc.

Heterogeneous Data:

Unstructured data represents almost every kind of data being produced like social media interaction, to recorded meetings, to handling of PDF documents, fax transfer, to email and more. Working with unstructured data is cumbersome and of course costly too. Converting all this unstructured data into structured one is also not feasible.

Structured data is always organized into highly mechanized and manageable way. It shows well integration with database but unstructured data is completely raw and unorganized.

7. FEATURES OF A BIG DATA:

The big data platform should give a solution which is designed specially with the needs of the enterprise in the mind. The following are the basic features of a big data platform offering

- Comprehensive - It should offer a broad platform and address all the three dimensions of the big data challenges – volume, variety and velocity.
- Enterprise-ready – It should include the performance, security, usability and reliability features.
- Integrated – It should be simplified and accelerates the introduction of big data technology to enterprise. It should enable integration with information supply chain including database, data warehouses and business intelligence applications.
- Open Source Based – It should be open source technology with the enterprise-class functionality and integration.
- Low latency reads and updates
- Robust and fault-tolerance
- Scalability
- Extensibility
- Allows adhoc queries
- Minimal maintenance

8. APPLICATIONS:

Applications of big data analytics are given below:

1. Smarter health care
2. Home land security
3. Traffic control
4. Manufacturing
5. Multi-channel sales
6. Telecom
7. Search quality

9. BENEFITS OF BIG DATA:

- Real-time big data isn't just a process for storing petabytes or Exabyte of data in a data warehouse, it's about the ability to make better decision and take meaningful actions at the right time.
- Fast forward to the present and technologies like hadoop give you the scale and flexibility to store data before you know how you are going to process it.

- Technologies such as MapReduce, Hive and Impala enable you to run quires without changing the data structure underneath.
- Our newest research finds that organizations are using the big data to target customer-centric outcomes, tap into internal data and build a better information ecosystem.
- Big data is already an important part of the $64 billion database and data analytics market.
- It offers commercial opportunities of a comparable scale to enterprise software in the late 1980s.
- And the internet boom of the 1990s, and the social media explosion of today.

11. BIG DATA IMPACT ON IT:

- Big data is a troublesome force presenting opportunities with challenges to IT organizations.
- In 2015 4.4 million IT jobs in Big data; 1.9 million is in US itself.
- India will require a minimum of 1 lakh data scientist in the next couple of years in addition to data analyst and data mangers to support the big data space.

12. FUTURE OF BIG DATA:

1. $15 billions on software firm only specialized in data management and analytics
2. This industry on its own is worth more than $100 billon and growing at almost 10% a year which is roughly twice as fast as the software business has the whole.
3. In February 2012, open source analyst form wikibon released the first market forecaster for a big data, listing $ 5.1 revenue in 2012with the growth to $53.4B in 2017
4. The Mckinsey global institution estimates that data volume is growing 40% per year, and will grow 44x between 2009 and 2020.

13. CONCLUSION:

In this seminar report some of the important issues are covered that are needed to be analysed by the organization while estimating the significance of implementing the big data technology and some direct challenges to the infrastructure of the technology

The commercial impacts of the big data have the potential to generate significant productivity growth for the number of vertical sectors. They should also create healthy demand for the talented individual who is capable to help organization in making sense of this growing volume of raw data. In short, big data present opportunity to create unprecedented business advantages and better service delivery.

10. RISKS OF BIG DATA:

- Will be so overwhelmed
- Need the right people and solve right problems
- Costs escalate too fast
- Isn't necessary to capture 100%
- Many sources of big data is privacy
- Self regulation
- Legal regulation.

A regulatory framework for big data is essential. That framework must be constructed with a clear understanding of the ravages that have been wrought on personal interests by the reduction of the information to data, its centralization, and its expropriation but the biggest gap is the lack of the skilled managers to make decision based on analysis by a factor of 10x. Growing talent and building teams to make analytic-based decision is the key to realize the value of big data

References:

1. www.computereducation.org
2. www.wikipedia.com
3. www.seminarsonly.com
4. www.seminarproject.net
5. www.authorsstream.com
6. www.slideshare.com
7. Big Data: A Revolution That Will Transform How We Live, Work, and Think(Hardcover) by Viktor Mayer-Schönberger.
8. Big Data (ebook) by Nathan Marz
9. Learning Spark: Lightning-Fast Big Data Analysis by Holden Karau.
10. Big Data at Work: Dispelling the Myths, Uncovering the Opportunities By Thomas H. Davenport.
11. Big Data: Using SMART Big Data, Analytics and Metrics To Make Better Decisions and Improve Performance (Kindle Edition) by Bernard Marr.