Fretfulness and Distinctiveness of Big Data: A Review

A.Vinothini, K. MuthuLakshmi, K. Lalitha

Abstract— Big data is the term for data sets so large and thorny that it becomes difficult to process using traditional data management tools or processing applications. Bid data problems can't be solved with available computer resources in today's business. They require clusters of computer with special applications and might take long period to complete. The key enablers for the escalation of big data are 1.increase of storage capacity 2.increase of processing power 3.availability of data. The fundamental challenge of Big Data is not collecting data but making sense of it. This paper reveals recent progress on big data, issues of big data in the field of e-commerce, mobile, healthcare and cloud.

Index Term— Big data, Big data analytics , Data traffic, Ecommerce, Health care, Storage, Zetta bytes.

1 INTRODUCTION

Big data is a collection of collections of data that is totally in different formats. It is too complex to store and compute those heterogeneous data in a mode so as to dig out information from it. Every fraction of second, data is evolving from various sources in the range of Exabyte's (1024^6) ,Zettabytes(1024^7) and above . Data has the power to transform business decisions, enhance customer relationships, and create new opportunities – but only if you can trust and make sense of it. Big Data starts with large-volume, diverse, independent sources with disseminated and decentralized control, and seeks to look at complex and embryonic relationships among data. Such large volumes surpass the capacity of current on-line storage systems and processing systems. Data, information, and knowledge are being shaped and accumulated at a rate that is rapidly approaching the Zettabytes/year range.

Volume is only one imperative aspect of big data; other attributes are variety, velocity, value, and complexity. Storage and data transport are technology issues that represent long-term challenges that require research and new paradigms. Traditional WAN-based transport methods cannot move terabytes of data at the speed dictated by businesses; they use a fraction of available bandwidth and achieve transfer speeds that are unsuitable for such volumes, introducing unacceptable delays in moving data into, out of, and within the cloud. We analyze the issues and challenges for big data analysis and design.

- *A.Vinothini, K.Muthulakshmi, K.Lalitha are Currently working as Assistant Professor in department of Information Technology at Panimalar Engineering College, India.*
- *E-mail: vinoforsundar@gmail.com akmuthube@gmail.com lalithatrends@gmail.com*

2 DISTINCTIVENESS OF BIG DATA

Big data is characterized by 3V's: Volume of information that systems must ingest process and disseminate; the velocity at which information grows or disappears; the variety in the diversity of data sources and formats. These challenges include collecting, storing, transferring, and visualizing all kinds of big data. We need effective ways of tuning "big data" into "big insights". The true value of big data lies in the implicit valuable knowledge derived from the analysis of a group of interrelated data sets, which allow deep correlations and hidden principles to be found for business trends prediction, health hazard analysis, terrorist threats detection, search engine optimization, biological and environmental research, etc. That is where new theories, novel methods and right analytics tools are needed to help scientists and business leaders make sense of the volumes of data. Volume describes the absolute magnitude of the data being analyzed. Terabyte is becoming a relatively small amount, as petabyte becomes the metric being used to quantify data sets in industries across the board. It is easy to revel in the astronomical numbers around data in this new world, a world measured in 1,180,591,620,717,411,303,424 bytes. Large numbers like these and the amount of processing power needed to digest this amount of information make it easy to capture the attention of early adopters.

Variety in big data refers to the differing types of data. This is where the discussion turns to Social Media. Buried in the

monotony of social data, nuggets of advertising gold may be found. The problem with this is the digging process. The information advertisers and B2C businesses want, is buried in mountains of non-structured data. The information one can mine from social media is unlike traditional structured information and as a result, does not fit well into a standard database, where you can run tried and true analytical tools. The information is found in the metadata of photos and videos. It is found in statuses about forgotten anniversaries and in the well-meant birthday wishes of individuals. Velocity is the sheer speed of data. Statistics such as every minute, users watch over 138,000 hours of video on YouTube. Every minute, 27,778 new blog posts go live on Tumblr. Every minute, 100,000 Tweets are shared. Every minute, 208,000 pictures are posted on Facebook.[10] Sometimes, these authors will glaze over more industrial statistics that are equally amazing, such as every flight a Virgin Atlantic Boeing 787 takes collects 500 gigs of data. Since Big data systems are expected to help analysis of structured and unstructured data and hence are drawing huge investments. Analysts have estimated enterprises will spend more than US$120 billion by 2015 on analysis systems.[10] The success of Big data technologies depends upon natural language processing capabilities, statistical analytics, large storage and search technologies .Big data analytics can help cope with large data volumes, data velocity and data variety. Enterprises have started leveraging these Big data systems to mine hidden insights from data.

The four most important questions every senior enterprise executive, especially CIOs and CMOs, should be asking about their big data infrastructure right now are: Are we ready to handle the massive data volumes expected over the next 3 years?;are we ready to accept and process a huge growth in sources of data that our business will experience?;Are we able to reliably deliver the data that the business needs at the speeds required?;So we have big data, but is it 100% trustworthy data that will reveal reliable and actionable insight? Everyone is anticipating the explosion to come in the next few years.

IDC predicts that the market for big data will reach $16.1 billion in 2014, growing 6 times faster than the overall IT market. IDC includes in this figure Infrastructure (servers, storage, etc., the largest and fastest growing segment at 45% of the market), services (29%) and software (24%). IDC commented that the benefits of big data are not always clear today (indeed, BNY Mellon recently asked its 50,000 employees "for ideas about how to harness the power of Big Data"). IIA predicted that companies will want to see demonstrable value in 2014 and will focus on embedding big data analytics in business processes to drive process improvement[11].The fourth V as newly identified in Big data is Value. Value is the only reason to work on big data. This value must be seen in better business outcomes such as: Higher Customer Profitability; Faster Time to Market; Reduced Cost; Improved Risk Management; Better Compliance; Greater Business & IT Agility.

3 CHALLENGES

3.1 Data traffic growth

All network data is set to grow with a compound annual growth rate (CAGR) of 23% through to 2018[12].One of the key drivers is mobile data traffic. Global mobile data traffic reached 1.5 Exabyte's per month at the end of 2013, up from 820 Petabytes per month at the end of 2012. This represents a staggering 81% growth rate of global mobile traffic in 2013. Over half a billion (526 million) mobile devices and connections were added in 2013. Global mobile devices and connections in 2013 grew to 7 billion, up from 6.5 billion in 2012. Smart phones accounted for 77 % of that growth, with 406 million net additions in 2013.[5]

3.2 Mobile data in 2014 equals whole Internet in 2001

Mobile data traffic was nearly 18 times the size of the entire global Internet in 2000. One Exabyte of traffic traversed the global Internet in 2000, and in 2013 mobile networks carried nearly 18 Exabyte's of traffic. Considering that between 2013 and 2018, global mobile data traffic is estimated to increase nearly 11 times, the scope of the challenge becomes clear. Machine-to-machine(M2M) connections are set for a massive growth, increasing data volume and complexity. AT Kearney estimates that M2M connections will reach 1.2 billion in 2017, up from only 200 million in 2012. A good example is the installed base of smart meters – a component part and first step of the 'smart grid' – represented 18% of all meters in Europe at the end of 2011. This is estimated to increase to 56% by 2017 due to large-scale rollouts in the UK, France and Spain achieving 100% coverage by 2022.

3.3 Security threats

As data volume and complexity increase, the challenges that security teams face are getting increasingly difficult and they grow in numbers. The small and medium businesses (up to 250 employees) are targeted for around 30% of the attacks, as they prove to be the path of least resistance for attackers. In 2013, over 342 million identities were exposed. Of the reported breaches so far, the top three types of information exposed are a person's real name, government ID number (e.g. Social Security), and birth date. Targeted attacks increased by 42% from 2011 to 2012. 40% of all targeted attacks were aimed at organizations that have more than 2500 employees[12].

3.4 Universal E-commerce

Growth within ecommerce will come primarily from the rapidly expanding online and mobile user bases in emerging markets, increases in m-commerce sales, advancing shipping and payment options, and the push into new international markets by major brands. Worldwide business-to-consumer (B2C) ecommerce sales are estimated to increase by 20.1% in 2014 to reach $1.5 trillion. With a 2012 – 2017 compound annual growth rate (CAGR) of more than 17%, ecommerce sales are targeted to reach $2.3 trillion by 2017, according to eMarketer[12]

3.5 Data Transport

With the large-scale deployment of traffic sensors and communication systems in road network during the last decades, the Intelligent Transportation Systems (ITS) have collected a tremendous amount of structured/unstructured traffic data. This large amount of data have been used to develop new paradigms and strategies in system design, system development, information processing, and performance evaluation in Intelligent Transportation Systems. Extensive research efforts have been dedicated to computational models that analyze and process these large-scale data, but effective tools to manipulate them are still at their infancy. A few key technical challenges are as follows: 1) the difficulty to efficiently and effectively discover low-level and high-level visual features for large-scale traffic data analysis; 2) the challenge to implement a real-time surveillance system that accurately identifies different vehicles and pedestrians; 3) the necessity to build an intelligent system that dynamically visualizes the statistics of the large-scale traffic data; and 4) efficient means to control the flow of the traffic data in the network.

4 BIG DATA ANALYTICS

Big data analytics enables organizations to analyze a mix of structured, semi-structured and unstructured data in search of valuable business information and insights**.** Big data analytics is the process of examining large data sets containing a variety of data types -- i.e., big data -- to uncover hidden patterns, unknown correlations, market trends, customer preferences and other useful business information. The primary goal of big data analytics is to help companies make more informed business decisions by enabling data scientists, predictive modelers and other analytics professionals to analyze large volumes of transaction data, as well as other forms of data that may be untapped by conventional business intelligence (BI) programs. That could include Web server logs and Internet click stream data, social media content and social network activity reports, text from customer emails and survey responses, mobile-phone call detail records and machine data captured by sensors connected to the Internet of Things. According to Gartner – "Big Data & Analytics is one of the top 10 strategic technologies for businesses and there would be 4.4 Million Big Data jobs by 2015"[13].

5 STORAGE AND WAREHOUSE

Data storage is the basis for big data networking. Representative technologies are Relational database and Not Only SQL (NoSQL) databases and data warehouse. Although considerable progresses have been made in database research, much remains to be done: firstly, handling streaming high-rate data in relational models remains as an open problem; second, statistical analysis and machine learning algorithms for big data need to be more robust and easier to use; lastly but more importantly, an ecosystem-alike mechanism should be built around the devised big data algorithms such that data management and usage can evolve sitting on top of the proposed algorithms.

Another important aspect in big data related database is data placement structures. Traditional data placement structures such as row-stores, column-stores and hybrid-stores are no longer suitable in large data analysis using MapReduce on distributed systems. Instead, RCFile (Record Columnar File) and its implementation in Hadoop, meets fast data loading, query processing, efficient storage space utilization, and strong adaptability to dynamic workload patterns. RCFile have multiple HDFS blocks, and each HDFS block is organized with basic units of row groups and all groups have the same size. This clustering idea enables RCFile to more efficiently manage data rows. Data areas of RCFile tables are divided as sync marker, metadata and table data sections. More importantly, RCFile has adopted RLE (Run Length Encoding) algorithm to compress metadata while using the Gzip compression algorithm for independently column data compression, which takes advantage the columnar storage of data. Moreover, because of the lazy decompression, RCFile does not need to decompress all columns while processing a row group. Decompression overhead can thus be reduced.

6 RECENT PROGRESS

Apache Hadoop, open-source data-processing platform first used by Internet giants including Yahoo and Facebook, leads the big-data revolution. Cloudera introduced commercial support for enterprises in 2008, and MapR and Hortonworks piled on in 2009 and 2011, respectively. Among data-management incumbents, IBM and EMC-spinout Pivotal each has introduced its own Hadoop distribution. Microsoft and Teradata offer complementary software and first-line support

for Hortonworks' platform. Oracle resells and supports Cloudera, while HP, SAP, and others act more like Switzerland, working with multiple Hadoop software providers. Storm, which is now owned by Twitter, is a real-time distributed computation system. It works the same way as Hadoop provides batch processing as it uses a set of general primitives for performing real-time analyses. Storm is easy to use and it works with any programming language. It is very scalable and fault-tolerant. Greenplum HD ,allows users to start with big data analytics without the need to built an entire new project. Greenplum HD is offered as software or can be used in a pre-configured Data Computing Appliance Module. SAMOA, is a platform for mining on big data streams. It is a distributed streaming machine learning (ML) framework that contains a programing abstraction for distributed streaming ML algorithms. Ikanow, focuses on developing products to enable uninhibited fusion and analysis of Big Data using open source technology. They have created an open source analytics platform.

6.1 IBM products for Big data

Big data represents a new era in data exploration and utilization. IBM is uniquely positioned to help clients design, develop and execute a big data strategy that will enhance and complement existing systems and processes. IBM delivers a big data and analytics infrastructure with data mining, data warehousing and business intelligence capabilities. So you can share access more securely, speed decision making in real time, and leverage insight for competitive advantage. Info Sphere Streams-enables continuous analysis of massive volumes of streaming data with sub-millisecond response times. Info Sphere Big Insights-An enterprise-ready, Apache Hadoop-based solution for managing and analyzing massive volumes of structured and unstructured data.IBM Netezza Data Warehouse- High-performance data warehouse appliances, purpose-built to make advanced analytics on exploding data volumes simpler, faster and more accessible[11]

6.2 Big data and cloud

Cloud Computing as an important application environment for big data has attracted tremendous attentions from the research community. Remarkable progress of big data networking has also been reported in this area. The rises of cloud computing and cloud data stores have been a precursor and facilitator to the emergence of big data. Cloud computing is the co modification of computing time and data storage by means of standardized technologies.

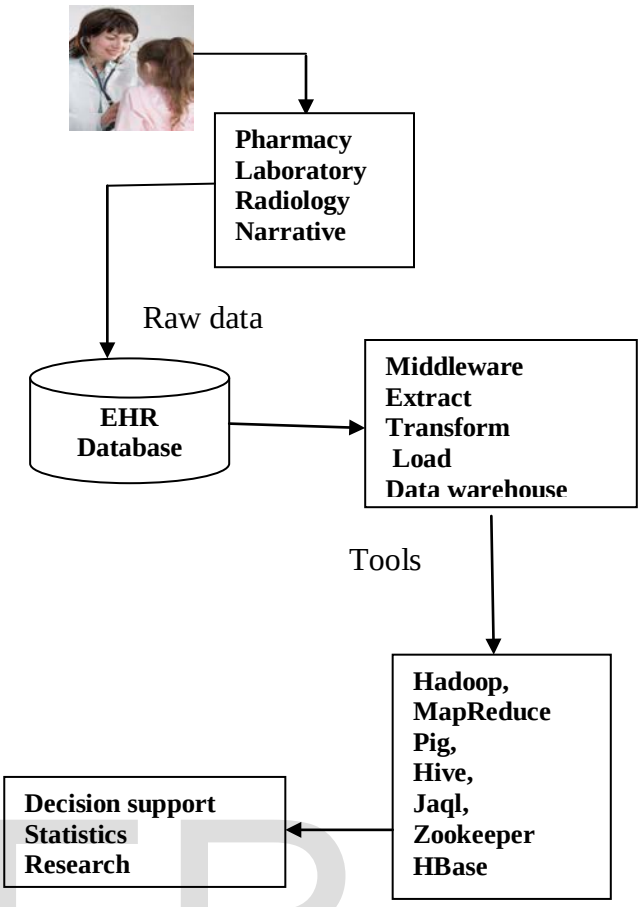

Fig1:Big data analysis in health care

It has significant advantages over traditional physical deployments. However, cloud platforms come in several forms and sometimes have to be integrated with traditional architectures. This leads to a dilemma for decision makers in charge of big data projects. How and which cloud computing is the optimal choice for their computing needs, especially if it is a big data project? These projects regularly exhibit unpredictable, bursting, or immense computing power and storage needs. At the same time business stakeholders expect swift, inexpensive, and dependable products and project outcomes.[7]

Big data analytics in Healthcare Sources and Techniques for Big Data in Healthcare: Structured EHR Data; Unstructured Clinical Note; Medical Imaging Data; Genetic Data[8][9].The healthcare industry historically has generated large amounts of data, driven by record keeping, compliance & regulatory requirements, and patient care. As Shown in the Fig 1 the raw data that is been collected from pharmacy, laboratory, radiology and doctor prescription are loaded into the Electronic Health Record(EHR).From HER the data can be transformed to either middleware, data warehouse etc. The big data from these storages are then processed using the big data mining tools. The output of these tools is the knowledge

extracted. These knowledge can be used for decision support, statistics and research.

7 CONCLUSION AND FUTUREWORKS

As Big Data challenges and necessity grows day by day, updation of tools and techniques must be keep on growing to meet the requirements. Analysis of big data is required in every field like mobile devices, e-commerce, social networking and healthcare. We need effective ways of tuning "big data" into "big insights". The true value of big data lies in the implicit valuable knowledge derived from the analysis of big data. This paper reveals only the issues and progress of big data. In future the work can be continued on development of techniques for an efficient knowledge discovery in big data.

REFERENCES

[1] Xingquan Zhu, Senior Member, IEEE, "Data Mining with Big Data"Gong Qing Wu, and Wei Ding, Senior Member, IEEE –IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING,VOL. 26, NO. 1, JANUARY 2014.
[2] Fox, B. 2011. "Leveraging Big Data for Big Impact", Health Management Technology, http://www.healthmgttech.com/
[3] http://www.cse.wustl.edu
[4] http://www.techrepublic.
[5] http://www.infosys.com/infosys-labs/publications/Documents/bigdata challenges-opportunities.pdf
[6] http://www.qubole.com/big-data-cloud-database-computing/
[7] http://bigdataanalyticsnews.com/16-top-big-data-analytics-platforms/
[8] http://www.dataversity.net/16-top-big-data-analytics-platforms/
[9] Cyrusone "EXECUTIVE REPORT + Big Data and the 3 v's:Volume, Variety and Velocity" http://www.cyrusone.com/pdf/c1-exec-report-big-data-and-the-vs-volume-variety-and-velocity.pdf
[10] $16.1 Billion Big Data Market: 2014 Predictions From IDC And IIA http://www.forbes.com.
[11] The challenge: Increasing data volume and complexity.http://www.napatech.com
[12] http://www.simplilearn.com.
[13] Viktor Mayer, Schonberger Kenneth Cukier "Big Data A Revolution that will transform how we Live, Work and Think",2013.