Volume 3, Issue 8, August 2013 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Significant Trends of Big Data Analytics in Social Network

Pradeepa. A*, Dr. Antony Selvadoss Thanamani
Department of Computer Science, NGM College
India

Abstract*— A massive volume of both structured and unstructured data that is so large are becoming difficult to process and manage with traditional databases and software techniques. These data sources may include prepared data such as databases, device capable, click streams and location information, as well as unstructured data like email, HTML, social data and images. The social network database (Eg: Facebook, Twitter, Google+, YouTube, Flicker) represents millions PB of data and the databases are doubled during every three months. The analysing of such data is very challenges issues. The overall goal of Big Data is to provide a scalable solution for vast quantities of data (Terabyte/ Petabytes / Exabyte's) while maintaining reasonable processing times. Big Data forms a way through which it becomes easy to scale, diversify, and interactively analyse huge amount of data that has hundreds of billions of rows within the tables. To accomplish efficient processing of huge amounts of data, companies will need to intelligently incorporate big data into their existing information management systems and take advantage of the developing ecosystem of integration and analysis tools. This study gives an overview of Big Data and their importance.

Keywords*— Big Data Analytic Tools, Data Mining, Hadoop and MapReduce, HBase and Hive tools, User-Friendly tools.

I. INTRODUCTION

Big data is evolving into viable, cost-effective way to store and analyse large volumes of data across many industries. Some of Big Data technologies like Apache Hadoop provide a better framework for large scale, distributed data storage and processing across clusters of hundreds or even thousands of networked computers. Industries ranging from supply chain, Logistics, Manufacturing, online services, web analysis, financial services, energy and utilities will be largely benefited through Big Data analytics. Big Data helps in analysing terms used to identify the datasets that due to their large size, that cannot be managed them with the typical data mining software tools. Instead of defining "Big Data" as datasets of a existing large size, for example in the order of magnitude of petabytes, the definition is related to the fact that the dataset is too big to be managed without using new algorithm or technology. Big Data analytics is becoming an important tool to improve efficiency and quality in organizations, and its importance is going to increase in the next years. The most popular distributed systems used nowadays are based in the map-reduce framework dealing with datasets in the order of terabytes or even petabytes is a reality [2, 3, 4]. Therefore, processing such big datasets in an efficient way is a clear need for many users. In this context, Hadoop and MapReduce [6] is a Big Data processing framework that has quickly become in both industries [9, 7, 15, 10, 5 and 13]. The map-reduce methodology started in Google, as a way to perform overflowing the web in a faster way. Hadoop is an open-source implementation of map-reduce started in Yahoo! and is being used in Big Data analysis. Nowadays, in business, more than size and scale, it is important is speed and agility.

Global Pulse is a United Nations initiative, launched in 2009, that functions as a modern lab and that is based in mining Big Data for developing many research and jobs in many aspect countries [14, 10]. This framework arise in the analysis of large social networks and etc.., there are hundreds of millions of nodes with billions of conversations [19]. Big Data uses inductive data with low information density, whose huge volume allows to regressions and therefore giving with the limits of implication reasoning to Big Data some predictive capabilities [8]. In his invite talk at the KDD active significant data numbers about internet usage, among the following: each day Google has more than 1 billion queries per day, Twitter has more than 250 million tweets per day, Facebook has more than 800 million updates per day, and you tube has more than 4 billion views per day [11].The data produced nowadays is estimated in the order of zetabyte, and it is growing around 40% every year. The latest large source of data is going to be generated from mobile devices, and big companies as Google, Apple, Facebook, Yahoo, and Twitter are starting to look carefully to this data to find useful patterns to improve user experience. We need new algorithms, and new tools to deal with all of this data [19]. The absolute size of data being collected is more than traditional compute infrastructures can be handle; exceeding the capacities of databases, storage, networks and everything in between, extract actionable intelligence from big data requires handling large amounts of data and processing it very quickly. We introduce Big Data mining and its applications in Section 3. We point the importance of open-source software tools in Section 5 and give some challenges and forecast to the future in Section 6. Finally, we give some conclusions in Section 7.

II. RELATED WORK

Data mining is a process of finding value from volume, data processing and prediction. In any enterprise, the high level of transactional data generated during its day-to-day operation is massive in volume, velocity, variety. The more information assets that require new forms of processing to enable improved decision making, approaching discovery and process optimization. Even though these traditional every instance of any activity. The terms terabytes and petabytes were unknown data capacity was measured in megabytes and gigabytes. The issue of storing and measuring large volumes of data is so important the first and most important technological requirement for the data warehouse is ability to manage large Big Data sets. Data mining attempts to extract smaller pieces of valuable information from this massive database. Mining requires integrated, cleaned, trustworthy, and efficiently 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, and provide intelligent querying functions. Big Data is also enabling the next generation of interactive data analysis with real-time answers. In the future, queries towards Big Data will be automatically generated for content creation on websites, and to provide an ad hoc analysis of the value of a data set to decide whether to store or to discard it. Scaling complex query processing techniques to terabytes while enabling interactive response times is a major open research problem today. There are many situations in which the result of the analysis is required immediately. The privacy of data is another huge concern, and one the context of Big Data. For example, a user's location information can be tracked through several stationary connection points (e.g., cell towers).

III. BIG DATA MINING AND ITS APPLICATION

Big data is a relative term describing a situation where the volume, velocity and variety of data exceed an organization's storage or compute capacity for accurate and timely decision making [13]. Some of this data is held in transactional data stores the by product of speed growing online activity. Machine-to-machine interactions, such as metering, call detail records, environmental sense and RFID systems; generate their own tidal waves of data. All these forms of data are expanding, and that is coupled with fast growing streams of unstructured and semi structured data from social media.

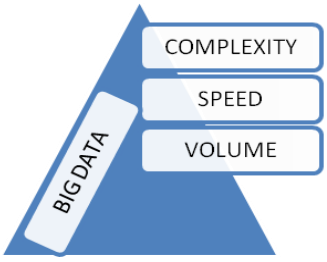

Fig: 1 big data organization.

That's a lot of data, but it is the reality for many organizations. By some estimates, organizations in all sectors have at least 100 terabytes of data, many with more than a petabytes. Big data has been used to convey all sorts of concepts, including: huge quantities of data, social media analytics, next generation data management capabilities, real-time data, and much more [12]. There are (3Vs) Big Data vectors for volume, velocity, and variety [15]:

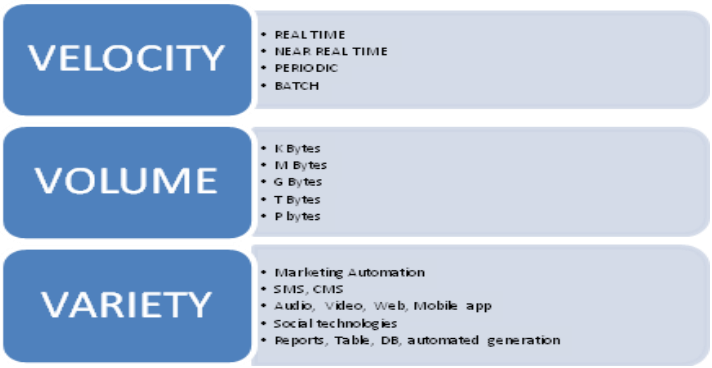

Fig: 2 Three V's of Big Data Vectors.

VOLUME (SCALE):

The amount of data in addition to data volume is increasing exponentially. There are 44x increases from 2009 to 2020 from 0.8 zetabyte to 35zb.

VELOCITY (SPEED):

Speed rate in collecting or acquiring or generating or processing of data. The data is begin generated fast and need to be processed fast, online data analytics, late decisions missing opportunities. Eg: e-promotions, healthcare monitoring.

VARIETY (COMPLEXITY):

The different data types along with latest insights are found when analysing together. This are various formats, types, and structures Text, numerical, images, audio, video, sequence, time series, social media data, multi-dimensional arrays, etc.. and Static data vs. streaming data, a single application can be generating and collecting many types of data, to extract knowledge-all these types of data need to correlate

Iv. MEASURING THE VALUE AND POTENTIAL YIELD OF BIG DATA

The people are capturing and digitizing more information than ever before. According to IDC, the world produced one zetabyte 1,000,000,000,000 gigabytes of data in 2010. this data explosion are over five billion mobile phones, 30 billion pieces of content shared on Facebook per month, 20 billion Internet searches per month, and millions of networked sensors connected to mobile phones, energy meters, automobiles, shipping containers, retail packaging and more. Big Data is a platform for transforming all of this data into actionable items for business decision making. Commercial vendor support from companies like cloudera can speed development and deliver more value from Big Data projects. Finally, modular data centre designs are emerging as a way to efficiently manage hardware and scale-out speedily and cost-effectively [17]. Value of Big Data is more real-time in nature than traditional DW applications, traditional DW architectures (e.g. Exadata, Teradata) are not well-matched for Big Data apps, shared nothing, massively parallel processing, scale out architectures are well-matching for Big Data apps[20,21]. Potential values of Big Data utilize to $300 billion potential annual value to US health care. Need to focus on the important pieces of data. It makes Big Data easier to handle. The biggest value in big data can be driven by combing Big Data with other corporate data: browsing history, knowing how valuable a customer.

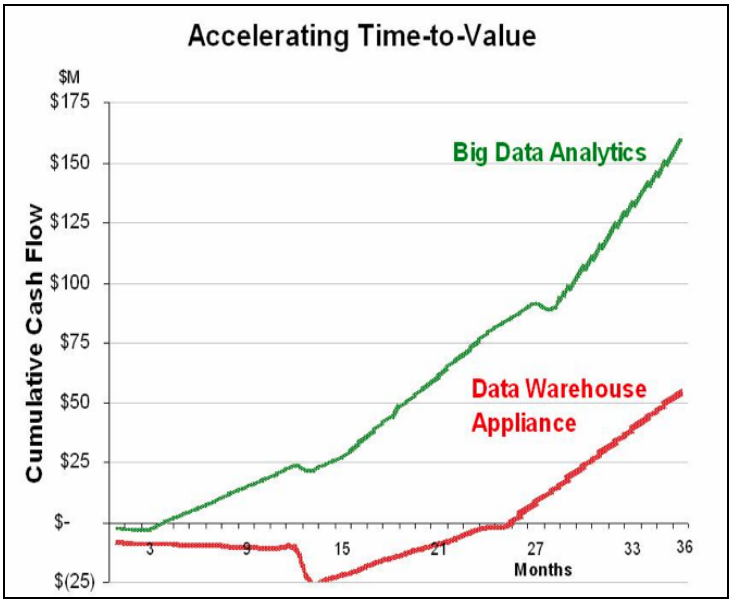

Fig 2: Graph showing Accelerating Time – to – value

The most widely used frameworks for big data processing are MapReduce, developed by Google, and its open source implementation from Yahoo, Hadoop. Based on Java, it offers a complete infrastructure for reliable, scalable and distributed computing. Even though new sets of tools continue to be available to help you manage and analyse your Big Data framework more effectively, you may not be able to get what you need. In addition, a range of technologies can support Big Data analysis and requirements such as availability, scalability, and high performance. Some of these include Big Data appliances, columnar databases, in-memory databases, non-relational databases, and massively parallel processing (MPP) engines [22]. Some important considerations as select a Big Data application analysis framework include the following: support for multiple data types, handle batch processing and real time data streams, overcome low latency, provide cheap storage, and integrate with conventional deployments.

v. FORECAST TO THE FUTURE OF BIG DATA

There are 84 different big data programs spread across six departments [23]. Private Sector: Wal-Mart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data. In Facebook handles 40 billion photos from its user base. Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide [24]. Science: Large Synoptic Survey Telescope will generate 140 Terabyte of data every 5 days. The Large Hardon Colider 13 Petabytes data produced in 2010. Medical computations like decoding human Genome.

VI. TOOLS: OPEN SOURCE REVOLUTION

Parallel DBMS technologies

To implement the proposed in late eighties, matured over the last two decades, multi-billion dollar industry: Proprietary DBMS Engines intended as data warehousing solutions for very large enterprises, relational data model, indexing, familiar SQL interface, advanced query optimization and well understood and studied.

Map Reduce

The pioneer by Google popularized by Yahoo! (Hadoop), data-parallel programming model, an associated parallel and distributed implementation for commodity clusters [25]. The following table shows a difference between parallel Database Management Systems and MapReduce.

TABLE I
DIFFERENCE BETWEEN PARALLEL DBMS VS. MAPREDUCE

	Parallel DBMS	MapReduce
Schema support	✓	Not out of the box
Indexing	✓	Not out of the box
Programming model	Declarative (SQL)	Imperative(C/C++, , JAVA...) Extensions through pig and hive
Optimizations (compression, query optimization)	✓	Not out of the box
flexibility	Not out of the box	✓
Fault tolerance	Coarse grained technique	✓

Hadoop

Apache Hadoop provides an open source software and framework designed to manage enormous data volumes. Hadoop consist of common APIs programmers can access its supported file system [3, 7]. The framework can also database that provide more structured data management and query functionality. Consider the Apache Foundation database described below:

***HBase*:**

HBase is an open source, distributed, versioned column-oriented store modelled after Google Big table. HBase can serve as both the input and output for MapReduce jobs run in Hadoop.

***Hive*:**

Hive is an open sources data warehouse infrastructure that facilitates easy data summarization, queries and the analysis of large datasets stored in Hadoop complete file system.

User-friendly tools for Big Data

Tools like Apache Pig and Apache Hive provide SQL-like frameworks for advanced data analysts to run queries directly against data stored in Hadoop. This is an effective way to do targeted, one-time analysis, perform exploratory data mining, or develop queries that may later be automated and loaded into a data warehouse. However, these tools require technical expertise and do not cater to end users. Tools that enable end users to slice, dice and visualize data in Hadoop will become increasingly important components of a company's Big Data analytics arsenal over the coming years.

VII. CONCLUSION

We have entered an era of Big Data. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. Technical challenges include not just the observable issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, attribution, and visualization, at all stages of the analysis direct from data achievement to result understanding. These technical challenges are common across a large variety of application domains, and therefore not cost-effective to address in the context of one domain alone. Furthermore, The challenges with Big Data are limited compared to the potential benefits, which are limited only by our creativity and ability to make connections among the trillions of bytes of data we have access. The influence of big data will likely inspire when it comes to processing of huge amounts of data, detailed analysis of data, obtaining results based on various conditions in large number for huge number of data than traditional databases.

REFERENCES

[1] www.bigdatauniversity.com, www.bigdatatraining.in.
[2] C. C. Aggarwal, Managing and Mining Sensor Data. Advances in Database Systems, Springer 2013.
[3] Hadoop, http://hadoop.apache.org/mapreduce/.
[4] Woody, Todd, "Automakers, Tech Companies Mining Electric Car Big Data to Plot Industry's Future." Forbes. June 18, 2012.http://www.forbes.com/sites/toddwoody/2012/06/18/automakers-tech-companiesmining-electric-car-big-data-to-plot-industrys-future/.
[5] Jim Wyllie, 'How to sort a terabytes quickly'.
[6] J. Dean and S. Ghemawat, 'MapReduce: A Flexible Data Processing Tool', CACM, 53(1):72–77, 2010.
[7] A. Thusoo et al. Hive – A Petabytes Scale Data Warehouse Using Hadoop, ICDE pages 996–1005, 2010.
[8] http://www.sas.com/resources/whitepaper/wp_46345.pdf• http://wikibon.org/blog/big-data-statistics/• http://wikibon.org/blog/big-data-infographics/•
[9] Jindal, J A. Quian –e – Ruiz, J. Dittrich, "Trojan Data Layouts: Right Shoes for a Running Elephant", In SOCC, 2011.

[10] http://www.forbes.com/sites/davefeinleib/2012/07/24/big-data-trends/•
[11] U. Fayyad. Big Data Analytics: Applications and Opportunities in On-line Predictive Modelling. http:// big-data-mining.org/keynotes/#Fayyad, 2012.
[12] Pittman, David. *"Lords of the Data Storm: Vestas and IBM Win Big Data Award."* The Big Data Hub: Understanding big data for the enterprise, September 28, 2012. http://www.ibmbigdatahub.com/blog/lords-datastorm-vestas-and-ibm-win-big-data-award.
[13] http://en.wikipedia.org/wiki/Big_data• http://mike2.openmethodology.org/wiki/Big_Data_Definition•
[14] United Nations Global Pulse, http://www.unglobalpulse.org.
[15] Gartner, http://www.gartner.com/it-glossary/bigdata.
[16] The Age of Big Data. Steve Lohr. *New York Times*, Feb 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html.
[17] Intel. Big Thinkers on Big Data, 2012 http://www.intel.com/content/www/us/en/bigdata/ big-thinkers-on-big-data.html/
[18] Y. Lin et al. Llama: Leveraging Columnar Storage for Scalable Join Processing in the MapReduce Framework. In SIGMOD, pages 961–972, 2011.
[19] D. Laney. 3-D Data Management: Controlling Data Volume, Velocity and Variety. META Group Research Note, February 6, 2001.
[20] Pattern-Based Strategy: Getting Value from Big Data. Gartner Group press release. July 2011. http://www.gartner.com/it/page.jsp?id=1731916
[21] A. Pavlo et al. A Comparison of Approaches to Large-Scale Data Analysis. In SIGMOD, pages 165–178, 2009.
[22] S. Wu, F. Li, S. Mehrotra, and B. C. Ooi, 'Query Optimization for Massively Parallel Data Processing', In SOCC, 2011.
[23] B. Brown, M. Chuiu and J. Manyika, "Are you ready for the era of Big Data?" McKinsey Quarterly, Oct 2011, McKinsey Global Institute.
[24] C. Bizer, P. Bonez, M. L. Bordie and O. Erling, "The Meaningful Use of Big Data: Four Perspective – Four Challenges" SIGMOD Vol. 40, No. 4, December 2011.
[25] JA. Quian – e – Ruiz, C. Pinkel, J. Schad, and J. Dittrich. Rafting, "MapReduce: Fast Recovery on the RAFT", ICDE, pages 589–600, 2011.

