Big Data Technology: Opportunities and Challenges

Anil Kumar Mishra
Senior Technical Assistant 'B',
CVRDE, DRDO, Avadi, Chennai
Email id: anilmlis@gmail.com

Satish Kumar
Information Scientist,
ARIES, U.K.
Email id: sklisc@gmail.com

Abstract

The increasing volume of information captured by enterprises, upsurge of multimedia and social media, and the Internet of Things (IoT) is fueling exponential growth in Big Data analytics. Every day close to With the volume and variety of data expanding exponentially, organizations that are able to capture quintillion bytes of data is being created and according to IDC report on Digital Universe, from now until 2020, the digital universe will about double every two years. Big Data by itself is meaningless, unless it translates to or helps us drive relevant outcomes.

With the volume and variety of data expanding exponentially, organizations that are able to capture and analyze it are bound to have a much higher competitive advantage over those that don't. But choosing how to go about using Big Data is a tough puzzle to crack.

This paper deals with such opportunities and challenges of big Data. In this paper we have mentioned some open source tools and technologies to create, manipulate, and manage such 'Big Data'.

Keywords: Big Data; Open Source Software; Data Analysis; Data Management; Big Data Tools

1. Introduction

The growth of data will never stop. According to the 2011 IDC Digital Universe Study, 130 exabytes of data were created and stored in 2005. The amount grew to 1,227 exabytes in 2010 and is projected to grow at 45.2% to 7,910 exabytes in 2015.3 The growth of data constitutes the "Big Data" phenomenon – a technological phenomenon brought about by the rapid rate of data growth and parallel advancements in technology that have given rise to an ecosystem of software and hardware products that are enabling users to analyse this data to produce new and more granular levels of insight[1,]

The promise of data-driven decision-making is now being recognized broadly, and there is growing enthusiasm for the notion of ``Big Data.'' While the promise of Big Data is real -- for example, it is estimated that Google alone contributed 54 billion dollars to the US economy in 2009 -- there is currently a wide gap between its potential and its realization[5].

Much data today is not natively in structured format; for example, tweets and blogs are weakly structured pieces of text, while images and video are structured for storage and display, but not for semantic content and search: transforming such content into a structured format for later analysis is a major challenge. The value of data explodes when it can be linked with other data, thus data integration is a major creator of value. Since most data is directly generated in digital format today, we have the opportunity and the challenge both to influence the creation to facilitate later linkage and to automatically link previously created data. Data analysis, organization, retrieval, and modeling are other foundational challenges[2].

A major investment in Big Data, properly directed, can result not only in major scientific advances, but also lay the foundation for the next generation of advances in science, medicine, and business[6].

2. What is Big Data?

According to McKinsey, Big Data refers to datasets whose size are beyond the ability of typical database software tools to capture, store, manage and analyse. There is no explicit definition of how big a dataset should be in order to be considered Big Data. New technology has to be in place to manage this Big Data phenomenon. IDC defines Big Data technologies as a new generation of technologies and architectures designed to extract value economically from very large volumes of a wide variety of data by enabling high velocity capture, discovery and analysis. According to **O'Reilly**, "Big data is data that exceeds the processing capacity of conventional database systems. The data is too big, moves too fast, or does not fit the structures of existing database architectures. To gain value from these data, there must be an alternative way to process it.[3]"

3. Big Data Technologies

Big Data technology can be broken down into two major components – the hardware component and the software component, as shown in the figure below. The hardware component refers to the component and infrastructure layer. The software component can be further divided into data organisation and management software, analytics and discovery software, and decision support and automation software. The layered architecture of Big Data consists of following elements[17]:-

3.1 Data Organisation and Management: This layer refers to the software that processes and prepares all types of structured and unstructured data for analysis. This layer extracts, cleanses, normalises and integrates data. Two architectures – extended Relational Database Management System (RDBMS) and the NoSQL database management system – have been developed to manage the different types of data. Extended RDBMS is optimised for scale and speed in processing huge relational data (i.e., structured data) sets, adopting approaches such as using columnar data stores to reduce the number of table scans (columnar database) and exploiting massively parallel processing (MPP) frameworks. On the other hand, the NoSQL database management system (NoSQL DBMS) grew out of the realisation that SQL's transactional qualities and detailed indexing are not suitable for the processing of unstructured files[4].

3.2 Data Analytics and Discovery: This layer comprises two data analytics software segments – software that supports offline, ad hoc, discovery and deep analytics, and software that supports dynamic real-time analysis and automated, rule-based transactional decision making. The tools can also be categorised by the type of data being analysed, such as text, audio and video.

3.3 Decision support and automation interface: The process of data analysis usually involves a closed-loop decision making model which, at the minimum, includes steps such as track, analyse, decide and act. To support decision making and to ensure that an action is taken, based on data analysis, is not a trivial matter. From a technology perspective, additional functionalities such as decision capture and retention are required to support collaboration and risk management. There are two decision support and automation software categories: transactional decision management and project-based decision management software[5].

Fig. 1 Big Data Technology Stack

4. Big data Tools and Technology

Tools typically used in Big Data scenarios[22] -

- **NoSQL** – Databases MongoDB, CouchDB, Cassandra, Redis, BigTable, Hbase, Hypertable, Voldemort, Riak, ZooKeeper
- **MapReduce** - Hadoop, Hive, Pig, Cascading, Cascalog, mrjob, Caffeine, S4, MapR, Acunu, Flume, Kafka, Azkaban, Oozie, Greenplum.
- **Storage** - S3, Hadoop Distributed File System.
- **Servers** - EC2, Google App Engine, Elastic, Beanstalk, Heroku.
- **Processing -** R, Yahoo! Pipes, Mechanical Turk, Solr/Lucene, ElasticSearch, Datameer, BigSheets, Tinkerpop.

5. **Big Data Analysis Platforms, Databases/Data Warehouses and Tools**

Today, organizations face lot of difficulty in managing data, because of the sheer size of datasets. It's coming from so many different mediums, be it social media, sensors, e-mail, etc. These are all termed as unstructured data and therefore cannot be managed by traditional database systems. In order to create, manipulate and manage such 'Big Data', you need specialized tools. Here are some of the most useful, open source Big Data tools to handle and store vast amounts of data and analyze it on low-cost commodity hardware[20].

- **Hortonworks** – Hortonwork's product named Hortonworks Data Platform (HDP) includes Apache Hadoop and is used for storing, processing and analyzing large volumes of data. The platform is designed to deal with data from many sources and formats. The platform includes various Apache Hadoop projects including the Hadoop Distributed File System, map Reduce, Pig, Hive, HBase and Zookeeper and additional components. ***Requirement*** **– Virtual machine**
- **Cassandra** – Developed by Facebook, and built on Amazon Dynamo and Google BigTable, its designed to handle large amounts of data across many commodity servers, while providing highly available service and no single point of failure. The Apache Cassandra database offers continuous availability, linear scale performance, operational simplicity and easy data distribution across multiple data centres and cloud availability zones. ***Requirement*** **– Linux**
- **Apache Pig** – Initially developed at Yahoo! to allow people using Apache Hadoop to focus more on analyzing large data sets and spend less time having to write mapper and reducer programs. Pig programming language is designed to handle any kind of data. It is a platform for analyzing large datasets that consists of a high level language for expressing data analysis programs, coupled with infrastructure for evaluating these programs. Pig is made up of two components: the first is the language itself, which is called **PigLatin** and the second is a runtime environment where PigLatin programs are executed. ***Requirement*** **– Linux**.
- **Flume** – Flume is a distributed, reliable and available service for efficiently collecting, aggregating and moving large amounts of log data. Requirement – Linux.
- **Gridgain** – A Java-based tool for real-time big data processing, offers an alternative to Hadoop's MapReduce that is compatible with the Hadoop Distributed File System. It offers in-memory processing for fast analysis of real-time data. You can download the open source version from GitHub or purchase a commercially supported version. ***Requirement*** **– Cross platform.**
- **HBase** – Hbase is the non-relational, distributed database written in Java for Hadoop. It is NoSQL database written that runs on top of Hadoop. Features include linear and modular

scalability, strictly consistent reads and write, automatic failover support and much more. ***Requirement*** **– OS independent.**

- **MongoDB** – A cross-platform document-oriented database that supports dynamic schema design, allowing the documents in a collection to have different fields and structures. It's a NoSQL database with document oriented storage, full index support, replication and high availability, and more. MongoDB can be used as a file system, taking advantage of load balancing and data replication features over multiple machines for storing files. ***Requirement*** **– Windows/linux/Mac**
- **CouchDB** – It is a document-oriented NoSQL database. Designed for the Web, CouchDB stores data in JSON documents that you can access via the Web or query using JavaScript. It offers distributed scaling with fault-tolerant storage.CouchDB implements a from of MultiVersion Concurrency Control (MVCC) in order to avoid the need to lock the database file during writes. ***Requirement*** **– Webserver**
- **Zookeeper** – Formely a Hadoop sub-project, Zookeeper is a centralized for maintaining configuration, inforamtion, naming, providing distributed synchronization and providing group services. APIs are available for Java and C, with Python, Perl and REST interfaces planned. ***Requirement*** **– Apache**
- **Hive** – Hive was initially developed by facebook Hadoob's data warehouse, Hive promises easy data summarizaton, ad-hoc queries and other analysis of big data. For queries, it uses a SQL-like language known as HiveQL. ***Requirement*** **– OS independent**
- **Sqoop** – Apache sqoop is a tool designed for efficiently transferring bulk data between Apache Hadoop and Structured datastores such as relational databases & data warehouses. It's a command-line interface tool.

< 3 years	3-5 years	> 5 years
• Hadoop MapReduce and HDFS • NoSQL DBMS • Text Analytics • Visualisation-based data discovery tool • In-Memory Analytics • Predictive Analytics • SaaS-based Business Analytics • Master Data Management	• Data Federation • Audio Analytics • Video Analytics (consumer marketing) • Complex Event Processing • Mobile Business Analytics • Non-Volatile Memory: PC-RAM • Improved Analytics Algorithms	• Quantum Computing

Fig. 2 Big Data Technology Radar

For example:

– Manufacturing companies deploy sensors in their products to return a stream of telemetry. Sometimes this is used to deliver services like OnStar, that delivers communications, security and navigation services.

– Other widely-cited examples of the effective use of big data exist in the following areas:

- Using information technology (IT) logs to improve IT troubleshooting and security breach detection, speed, effectiveness, and future occurrence prevention.
- Uses of voluminous historical call center information more quickly, in order to improve customer interaction and satisfaction.
- Use of social media content in order to better and more quickly understand customer sentiment about you/your customers, and improve products, services, and customer interaction.
- Fraud detection and prevention in any industry that processes financial transactions online, such as shopping, banking, investing, insurance and health care claims.
- Use of financial market transaction information to more quickly assess risk and take corrective action.

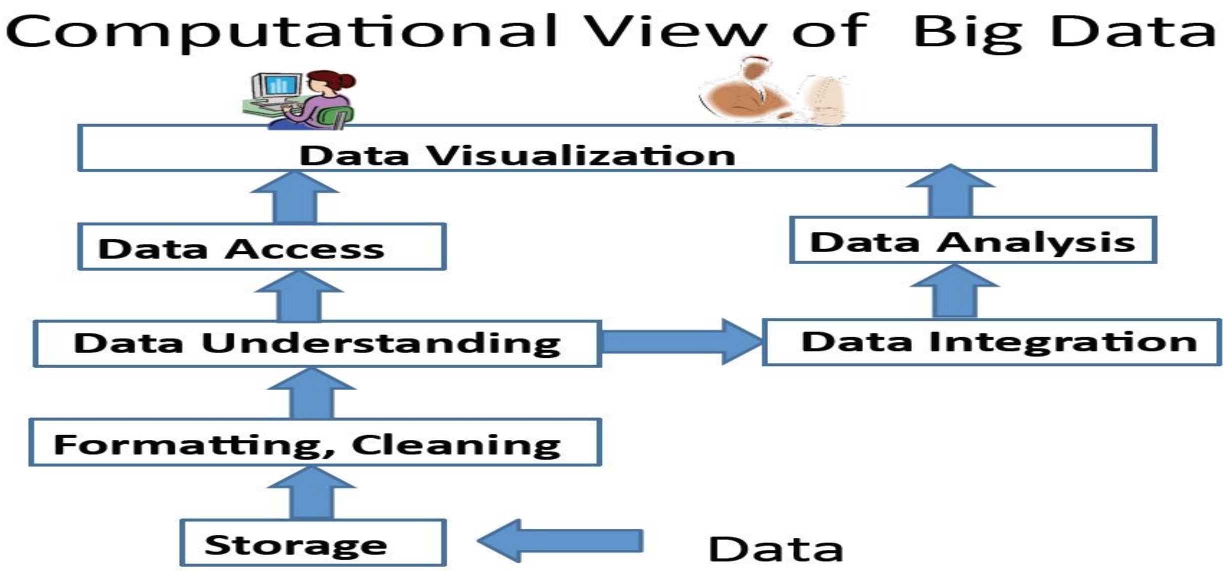

Fig. 3 Computational view of Big data

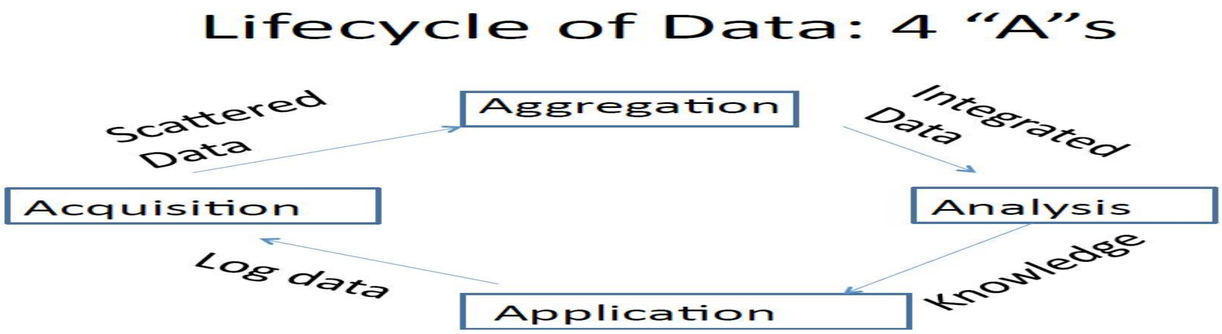

Fig. 4 Lifecycle of Big Data – Acquisition, Aggregation, Analysis, Application (4A)

5.1 BIG Data – Growth and Size Facts (MGI Estimates)

- There were 7 billion mobile phones in use in 2014.
- There are 40 billion pieces of content shared on Facebook each month.
- There is a 40% projected growth in global data generated per year vs. 5% growth in global IT spending.
- There were 235 terabytes of data collected by the US Library of Congress in April 2011.
- 15 out of 17 major business sectors in the United States have more data stored per company that the US Library of Congress.

5.2 Big Data – Value Potential

- $300 billion annual value to US healthcare – more than twice the total annual healthcare spending in Spain.
- $600 billion – potential annual consumer surplus from using personal location data globally.
- 60% - potential increase in retailers' operating margins possible via use of big data.

5.3 Big Data – Industry Examples

- Major utility company integrates usage data recorded from smart meters in semi real-time into their analysis of the national energy grid.
- Pay television providers have begun to customize ads based on individual household demographics and viewing patterns.
- A major entertainment company is able to analyze its data and customer patterns across its many and varied enterprises – e.g. using park attendance, on-line purchase, and television viewership data.
- The security arm of a financial services firm detects fraud by correlating activities across multiple data sets. As new fraud methods are detected and understood, they are used to encode new algorithms into the fraud detection system.

6. Opportunities of Big Data

6.1 Technology-

Big Data can be used to create value across sectors of the economy, bringing with it a wave of innovation and productivity gains. The discussion on the impact of Big Data focuses very much on the application of Big Data analytics rather than on the middleware or the infrastructure. Therefore, the adoption of Big Data technologies always comes from the analytics perspective which in turn drives the adoption of the underlying supporting technologies.

6.2 Healthcare -

Big Data has a huge application in healthcare, particularly in areas where analysis of large data sets is a necessary pre-condition for creating value. Possible adoption of Big Data analytics could be done in a few specific areas. One of them is comparative effectiveness research (CER). CER is designed to inform healthcare decisions by providing evidence on the effectiveness, benefits and harm of different treatment options.

6.3 Retail–

The retail sector is built on an understanding of the consumers' retail habits. Top retailers are mining customer data and using Big Data technologies to help make decisions about their marketing campaigns, merchandising and supply chain management. Retailers are using more advanced methods in analysing the data they collect from multiple sales channels and interactions.

One of these applications is to enable cross-selling which uses all the data that can be known about a customer, including the customer's demographics, purchase history and preferences, to increase the average purchase amount. Online retailer, Amazon, is a good example.

6.4 Education -

In the education sector, learners are creating information at the same time as they are consuming knowledge. Students are faced with increasingly demanding curricula where they are no longer expected to regurgitate facts from hard memorising but are required to learn the subjects with deep understanding. Creating a profile for each of the students would require disparate sets of information and this is where the opportunity lies for Big Data analytics[17].

6.5 Transport -

Big Data analytics offers the opportunity for public transport service operators to obtain critical insights on passenger demand trends so as to implement more effective measures in their service provisions.

6.6 Finance -

Big Data plays a significant role in the finance sector, especially with regard to fraud detection with the application of Complex Event Processing (CEP). CEP is typically done by aggregating data from distributed systems in real time and applying rules to discern patterns and trends that would otherwise go unnoticed[21].

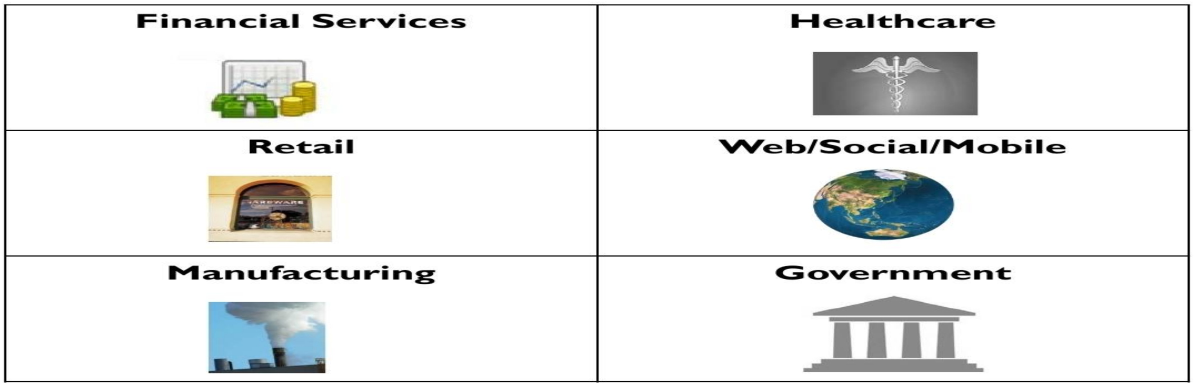

Fig. 5a. Opportunity of Big Data

Fig. 5b. Opportunity of Big Data

7. Key Big Data Challenges

7.1 Understanding and Utilizing Big Data – It is a daunting task in most industries and companies that deal with big data just to understand the data that is available to be used, determining the best use of that data based on the companies' industry, strategy, and tactics. Also, these types of analyses need to be performed on an ongoing basis as the data landscape changes at an ever-increasing rate, and as executives develop more and more of an appetite for analytics based on all available information.

7.2 New, Complex, and Continuously Emerging Technologies – Since much of the technology that is required in order to utilize big data is new to most organizations, it will be necessary for these organizations to learn about these new technologies at an ever-accelerating pace, and potentially engage with different technology providers and partners than they have used in the past. Like with all technology, firms entering into the world of big data will need to balance the business needs associated with big data with the associated costs of entering into and remaining engaged in big data capture, storage, processing, and analysis.

7.3 Cloud Based Solutions – A new class of business software applications has emerged whereby company data is managed and stored in data centers around the globe. While these solutions range from ERP, CRM, Document Management, Data Warehouses and Business Intelligence to many others, the common issue remains the safe keeping and management of confidential company data. These solutions often offer companies tremendous flexibility and cost savings opportunities compared to more traditional on premise solutions but it raises a new dimension related to data security and the overall management of an enterprise's Big Data paradigm.

7.4 Privacy, Security, and Regulatory Considerations - Given the volume and complexity of big data, it is challenging for most firms to obtain a reliable grasp on the content of all of their data and to capture and secure it adequately, so that confidential and/or private business and customer data are not accessed by and/or disclosed to unauthorized parties. The costs of a data privacy breach can be enormous. For instance, in the health care field, class action lawsuits have been filed, where the plaintiff has sought $1000 per patient record that has been inappropriately accessed or lost. In the regulatory area, for instance, the proper storage and transmission of personally identifiable information (PII), including that contained in unstructured data such as emails can be problematic and necessitate new and improved security measures and technologies. For companies doing business globally there are significant differences in privacy laws between the U.S. and other countries. Lastly, it will be very important for most forms to tightly integrate their big data, data security/privacy, and regulatory functions.

7.5 Archiving and Disposal of Big Data – Since big data will lose its value to current decision-making over time, and since it is voluminous and varied in content and structure, it is necessary to utilize new tools, technologies, and methods to archive and delete big data, without sacrificing the effectiveness of using your big data for current business needs.

7.6 The Need for IT, Data Analyst, and Management Resources – It is estimated that there is a need for approximately 140,000 to 190,000 more workers with "deep analytical" expertise and 1.5 million more data-literate managers, either retrained or hired. Therefore, it is likely that any firm that undertakes a big data initiative will need to either retrain existing people, or engage new people in order for their initiative to be successful.

The convergence of mobile device adoption, the mobile Internet and social networking provides an opportunity for organisations to derive competitive advantage through an efficient analysis of unstructured data. Businesses that were early adopters of Big Data technologies and that based their business on data-driven decision making were able to achieve greater productivity of up to 5% or 6% higher than the norm12. Big Data technology early adopters such as Facebook, Linkedln, Walmart and Amazon are good examples for companies that plan to deploy Big Data analytics.

According to Cisco's Internet Business Solutions Group (IBSG)13, 50 billion devices will be connected to the Web by 2020. Meanwhile, Gartner reported that more than 65 billion devices were connected to the internet by 2010. By 2020, this number will go up to 230 billion Regardless of the difference in estimation, these connected devices, ranging from smart meters to a wide range of sensors and actuators continually send out huge amounts of data that need to be stored and analysed. Companies that deploy sensor networks will have to adopt relevant Big Data technologies to process the large amount of data sent by these networks[23].

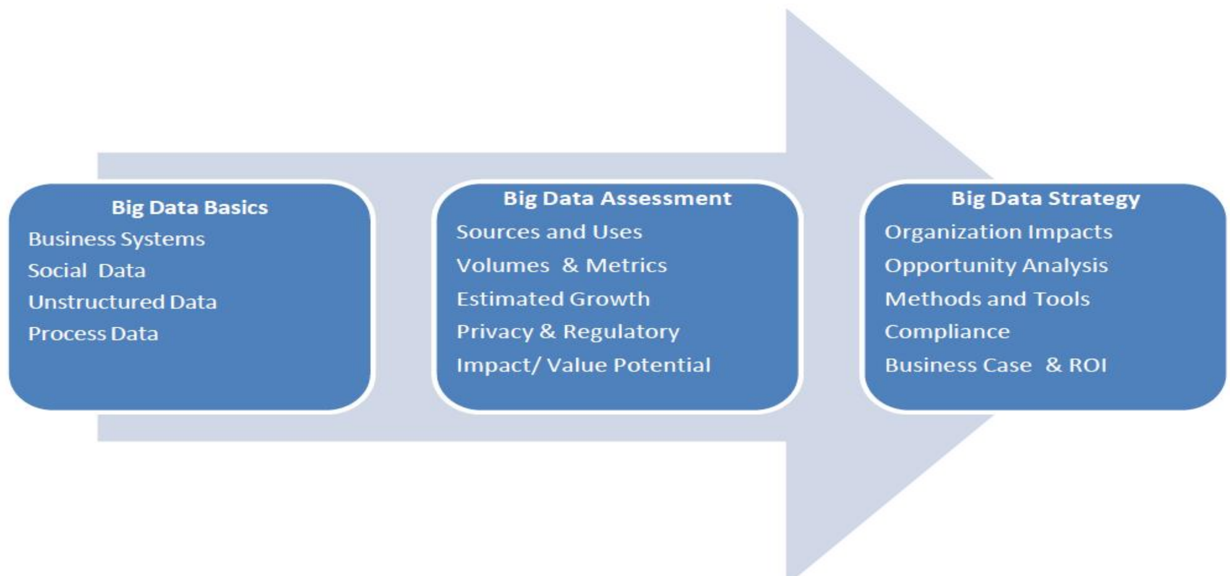

Fig. 6 Developing a Big Data Strategy[1]

8. 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. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. The challenges include not just the obvious issues of scale, but also heterogeneity, lack of structure, error-handling, privacy, timeliness, provenance, and visualization, at all stages of the analysis pipeline from data acquisition to result interpretation. 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, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

Analyzing new and diverse digital data streams can reveal new sources of economic value, provide fresh insights into customer behavior and identify market trends early on. But this influx of new data creates challenges for IT departments. To derive real business value from big data, you need the right tools to capture and organize a wide variety of data types from different sources, and to be able to easily analyze it within the context of all your enterprise data.By using big data analysis tools, enterprises can acquire, organize and analyze all their enterprise data – including structured and unstructured – to make the most informed decisions.

Reference

1. Why is Big data important? www.navint.com [Accessed 5th July 2015].
2. The Economist. Building with big data. [Online] Available from: http://www.economist.com/node/18741392 [Accessed 9th July 2015].
3. Edd Dumbill. What is big data? [Online] Available from: http://radar.oreilly.com/2012/01/what-is-big-data.html [Accessed 9th July 2015].
4. Peter Buneman. Semistructured Data. [Online] Available from: http://homepages.inf.ed.ac.uk/opb/papers/PODS1997a.pdf [Accessed 10th July 2015].
5. IDC. IDC's Worldwide Big Data Taxonomy, 2011. [Online] Available from: http://www.idc.com/getdoc.jsp?containerId=231099 [Accessed 9th July 2015].
6. IDC. Worldwide Big Data Technology and Services 2012-2015 Forecast. [Online] Available from: http://www.idc.com/getdoc.jsp?containerId=233485 [Accessed 12th July 2015].
7. Erik Brynjolfsson, et al. Strength in Numbers: How Does Data-Driven Decisionmaking Affect Firm Performance. [Online] Available from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1819486 [Accessed 9th July 2015].
8. Cisco. The Internet of Things: How the Next Evolution of the Internet is Changing Everything. [Online] Available from: http://www.cisco.com/web/about/ac79/docs/innov/IoT_IBSG_0411FINAL.pdf [Accessed 9th July 2015].
9. John Mahoney, Hung LeHong. The Internet of Things is Coming. [Online] Available from: http://www.gartner.com/id=1799626 [Accessed 9th July 2015].
10. Arnold, K. E., and Pistilli, M. D. (2012). Course signals at Purdue: Using learning analytics to increase student success. In S. Buckingham Shum, D. Gašević, & R. Ferguson (Eds.), Proceedings of the 2nd International Conference on Learning Analytics and Knowledge (LAK 2012) (pp. 267-270). New York: ACM.
11. Diebold, Francis X. (2012). A personal perspective on the origin(s) and development of 'Big Data': the phenomenon, the term, and the discipline. Second version. PIER Working Paper No. 13-003. Retrieved from http://dx.doi.org/10.2139/ssrn.2202843 on 9.7.2015

12. Digital Preservation Network. (2015). Retrieved from http://www.dpn.org/ [Accessed 13th July 2015].
13. GSV Capital (2014). Market commentary. Retrieved from http://gsvcap.com/market-commentary/san-francisco-37/ [Accessed 9th July 2015].
14. Hathi Trust Digital Library. (2015). Partnership Community. Retrieved from https://www.hathitrust.org/community [Accessed 13th July 2015].
15. Laney, Doug. (2001). 3D data management: controlling data volume, variety and velocity. Retrieved from http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf [Accessed 22nd July 2015].
16. Rijmenam, Mark van. (2013). Why the 3 Vs are not sufficient to describe big data. Retrieved from https://datafloq.com/read/3vs-sufficient-describe-big-data/166 [Accessed 22nd July 2015].
17. Soares, Louis. (2011). The 'Personalization' of Higher Education: Using Technology to Enhance the College Experience. Retrieved from https://www.americanprogress.org/issues/labor/report/2011/10/04/10484/thepersonalization-of-higher-education/ [Accessed 10th July 2015].
18. Sharma, Adeesh. (2015). Making the Right Moves on Big Data. PC Quest. July 2015 37p.
19. Ganesan, Vinod. (2015). How Big Data is going to impact Businesses in the near Future. PC Quest, July 2015; 40-41p.
20. Maurya, Raj Kumar. (2015). 15 Big Data Analysis Platforms, Databases/Data Warehouses and Tools. PC Quest, July 2015; 52-53p.
21. Varshney, Rashi. (2015). Big Data: Analytics provides key opportunities for growth to startups. PC Quest, July 2015; 48-50p.
22. Murthy, Narayana. (2015). Key Open Source Tools for SMEs to jumpstart Big Data. PC Quest, July 2015; 44-45p.
23. Lohr, Steve. (2012). The Age of Big Data. New York Times, Feb 11, 2012. http://www.nytimes.com/2012/02/12/sunday-review/big-datas-impact-in-the-world.html