A Review on Big Data Challenges and Opportunities

Varsha Jambunathan[1], S. Venkatesan[2]

[1]*Pre Final Year, Department of Computer Science & Engg., Dayananda Sagar College of Engineering, Bangalore, Karnataka, India*

[2]*Professor, Department of Computer Science & Engg., Dayananda Sagar College of Engineering, Bangalore, Karnataka, India*

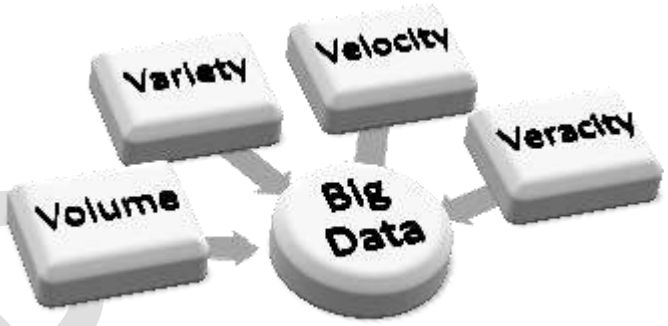
Abstract: **Data is increasingly cheap and ubiquitous. We are now digitizing analog content that was created over centuries and collecting myriad new types of data from web logs, mobile devices, sensors, instruments, and transactions. IBM estimates that 90 percent of the data in the world today has been created in the past two years. At the same time, new technologies are emerging to organize and make sense of this avalanche of data. We can now identify patterns and regularities in data of all sorts that allow us to advance scholarship, improve the human condition, and create commercial and social value. The rise of "Big Data" has the potential to deepen our understanding of phenomena ranging from physical and biological systems to human social and economic behaviour. This paper briefs the evolution of Big Data, describes some of its management challenges, opportunities in business sectors, and the tools used to manipulate it.**

Keywords: **Big Data, Data Mining, Machine Learning, Analytics, Business Intelligence.**

I. INTRODUCTION

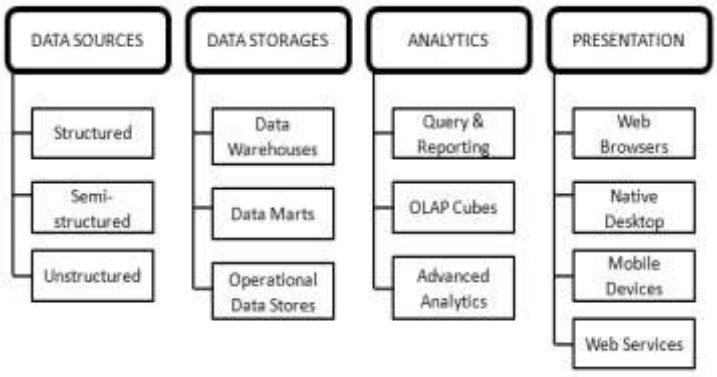
Every day, 2.5 quintillion bytes of data are created. These data come from digital pictures, videos, posts to social media sites, intelligent sensors, purchase transaction records, cell phone GPS signals, to name a few. While the concept of data is commonly associated with scientific research, raw data is unprocessed facts and statistics collected together for reference or analysis, whereupon it can be visualized using graphs, images or other analysis tools.

Data, both structured and unstructured, becomes Big Data when its volume, velocity, variety, or veracity exceeds the abilities of IT systems to ingest, store, analyze, and process it [8]. Many organizations have the equipment and expertise to handle large quantities of structured data—but with the increasing volume and faster flows of data, they lack the ability to "mine" it and derive actionable intelligence in a timely way. Not only is the volume of this data growing too fast for traditional analytics, but the speed with which it arrives and the variety of data types necessitates new types of data processing and analytic solutions [6].



Big Data: 3 V's

Volume refers to amount of data stored in enterprise repositories that range from gigabytes and terabytes to zettabytes. Data variety exploded from structured and legacy data stored in enterprise repositories to unstructured, semi structured, audio, video, XML etc. Velocity refers to the speed of data accumulation. For time-sensitive processes such as catching fraud, Big Data must be used as it streams into your enterprise in order to maximize its value [5]. Big Data Veracity refers to the biases, noise and abnormality in data. This data should be meaningfully mined before processing.



Big Data: Relational Architecture

Big Data Analytics is really about two things—Big Data and Analytics. First, there's Big Data for massive amounts of detailed information. Second, there's Analytics, which is an arsenal of different tool types, based on SQL queries, data mining, statistical analysis, fact clustering, data visualization, predictive analytics, artificial intelligence, natural language processing, text analytics, and so on. This, they yield advanced analytical techniques to operate on Big Data sets [1].

A problem with current Big Data analysis is the lack of coordination between Database systems, which host the data and provide SQL querying with analytics packages that perform various forms of non- SQL processing such as data mining and statistical analyses.

II. CHALLENGES

Big Data analysis involves making "sense" out of large volumes of varied data that in its raw form lacks a Data Model to define what each element means in the context of the others. The following challenges arise in doing so.

1. *Data Acquisition*: Today, scientific experiments and simulations can easily produce petabytes of data. 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. Metadata acquisition systems can minimize the human burden in recording metadata. Another important issue here is data provenance. Recording information about the data at its birth is not useful unless this information can be interpreted and carried along through the data analysis pipeline. Thus, we need research both into generating suitable metadata and into data systems that carry the provenance of data.
2. *Variety:* Data, not only structured, but raw, semi-structured, unstructured data come from sensors, smart devices, and social collaboration technologies like web pages, web log files, search indexes, e-mails, documents, sensor data, etc. Semi-structured Web data such as A/B testing, sessionization, bot detection, and pathing analysis all require powerful analytics. The challenge is how to handle multiplicity of types, sources, and formats.
3. *Integration, Aggregation and Representation:* The real value of data is when data sets are integrated and cross-correlated. Integration and cross-correlation among data sets from different sources allows us to uncover information and trends that often cannot be uncovered by looking at a data set in isolation [3]. The challenge comes with figuring out which data elements relate to which other data elements, and in what capacity. The process of discovery not only involves exploring the data to understand how you can use it but also determining how it relates to your traditional enterprise data [6].
4. *Automated Analysis:* 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 be expressed in forms that are computer understandable, and then robotically resolvable. Even for simpler analyses that depend on only one data set, there remains an important question of suitable database design. Usually, there will be many alternative ways in which to store the same information. Certain designs will have advantages over others for certain purposes, and possibly drawbacks for other purposes.
5. *Query processing and Analysis:* Query processing, and analysis methods suitable for Big Data need to be able to deal with noisy, dynamic, heterogeneous, untrustworthy data and data characterized by complex relations. However despite these difficulties, Big Data even if noisy and uncertain can be more valuable for identifying more reliable hidden patterns and knowledge compared to tiny samples of good data. Also the (often redundant) relationships existing among data can represent an opportunity for cross-checking data and thus improve data trustworthiness. Supporting query processing and data analysis requires scalable mining algorithms and powerful computing infrastructures [3].
6. *Privacy & Security:* The privacy of data is another huge concern, and one that increases in the context of Big Data. There is great public fear regarding the inappropriate use of personal data such as buying preference, healthcare records, and location-based information, particularly through linking of data from multiple sources. Managing privacy is effectively both a technical and a sociological problem, which must be addressed jointly from both perspectives to realize the promise of Big Data [5]. While there have been significant studies on protecting data centres from being attacked, the privacy and security loopholes when moving crowd sourced data to data centres remain to be addressed. There is an urgent demand on technologies that endeavour to enforce security in data transmission. Given the huge data volume and number of sources, this requires a new generation of encryption solutions (e.g., homomorphic encryption) [9].
7. *Scalability:* Managing large and rapidly increasing volumes of data has been a challenging issue for many decades. Large data processing systems had to worry about parallelism across nodes in a cluster. Now, one has to deal with parallelism within a single node. Unfortunately, parallel data processing techniques that

were applied in the past for processing data across nodes don't directly apply for intra-node parallelism, since the architecture looks very different. These unprecedented changes require us to rethink how we design, build and operate data processing components [5].

8. *Human Collaboration:* In spite of the tremendous advances made in computational analysis, there remain many patterns that humans can easily detect but computer algorithms have a hard time finding. Ideally, analytics for Big Data will not be all computational rather it will be designed explicitly to have a human in the loop. The new sub-field of visual analytics is attempting to do this, at least with respect to the modelling and analysis phase in the pipeline. In today's complex world, it often takes multiple experts from different domains to really understand what is going on. A Big Data analysis system must support input from multiple human experts, and shared exploration of results. These multiple experts may be separated in space and time when it is too expensive to assemble an entire team together in one room. The data system has to accept this distributed expert input, and support their collaboration [5].

III. OPPORTUNITIES

1. *Business Intelligence (BI):* Business intelligence is a practice that uses software applications to analyze an organization's raw data and monitors the performance of business operations frequently. Business organizations are emerging with new insights to use key methods for data mining and learning analytics on Big Data. Predictive analysis is an area of statistical analysis that deals with extracting information using various technologies to uncover relationships and patterns within large volumes of data that can be used to predict behaviour and events. It is used in log analytics, fraud detection, social media and sentiment analysis, risk modelling and management, and energy sectors to name a few [1]. For example, companies can study consumer purchasing trends to target better marketing. In addition, near-real-time data from mobile phones could provide detailed characteristics about shoppers that help reveal their complex decision-making processes as they walk through malls [10]. Thus, the more data we have about more people, the more we can improve services to offer more customized, personalized choices to help them meet their goals. We must support and encourage fundamental research towards addressing these technical challenges if we are to achieve the promised benefits of Big Data.

2. *Education:* Big Data has the potential to revolutionize not just research, but also education. Imagine a world in which we have access to a huge database where we collect every detailed measure of every student's academic performance. This data could be used to design the most effective approaches to education, starting from reading, writing, and math, to advanced, college-level, courses. With algorithms it will be possible to determine the strengths and weaknesses of each individual student based on the way a student learned online, how and which questions were answered, the social profile etc. This will create stronger groups that will allow students to have a steeper learning curve and deliver better group results. We are far from having access to such data, but there are powerful trends in this direction. In particular, there is a strong trend for massive Web deployment of educational activities, and this will generate an increasingly large amount of detailed data about students' performance [7].

3. *Health sector:* In the scientific domain, secondary uses of patient data could lead to the discovery of cures for a wide range of devastating diseases and the prevention of others [4]. By revealing the genetic origin of illnesses, such as mutations related to cancer, the Human Genome Project, completed in 2003, is one project that's a testament to the promises of big data. Consequently, researchers are now embarking on two major efforts, the Human Brain Project (EU; www.humanbrainproject.eu/vision.html) and the US BRAIN Initiative (www.whitehouse.gov/the-press-office/2013/04/02/fact-sheet-brain-initiative) in a quest to construct a supercomputer simulation of the brain's inner workings, in addition to mapping the activity of about 100 billion neurons in the hope of unlocking answers to Alzheimer's and Parkinson's. Other types of big data can be studied to help solve scientific problems in areas ranging from climatology to geophysics to nanotechnology [10].

4. *Security:* Big Data techniques can also be used to address the security challenges in networked systems. Network attacks and intrusions usually generate data traffic of specific patterns in networks. By analyzing the Big Data gathered by a network monitoring system, those misbehaviours can be identified proactively, thus greatly reducing the potential loss [9].

IV. TOOLS

A normal system can't handle very large dataset calculation and data size is increasing day by day, thus the obtained model should be adapted accordingly. To obtain this we have to implement distributed computing using Big Data technologies like Apache Mahout, Spark, R-Hadoop or initial analytics processing in projects like hive/ pig and feed output to machine learning algorithms for model generation.

1. AWS, Microsoft Azure, Oracle: BI tools are important for reporting, analysis and performance management, primarily with transactional data from

data warehouses and production information systems. BI Tools provide comprehensive capabilities for business intelligence and performance management, including enterprise reporting, dashboards, ad-hoc analysis, scorecards, and what-if scenario analysis on an integrated, enterprise scale platform.

2. Hadoop: Hadoop is framework developed by the Apache foundation to solve the storage and distributed processing problems of Big Data. Hadoop is architected using the Hadoop Common libraries that contains the java libraries required by other modules of Hadoop. The Yarn module acts like a cluster resource manager for Hadoop and separates the management and processing functions of Hadoop, and provides a platform for other processing tools. The Hadoop Distributed File System well known as HDFS enables interaction among the nodes. The Map-Reduce framework is then used to solve the problem of big data analysis using the distributed file systems [8]. The prevalent architecture that people use to analyze structured and unstructured data is a two-system configuration, where Hadoop is used for processing the unstructured data and a relational database system or an NoSQL data store is used for the structured data as a front end. This is termed as Big Data Dichotomy [7].
3. R: R is a highly extensible open source software package to perform statistical analysis. It provides a wide variety of built-in as well as extended functions for statistical computing, machine learning, graphical techniques and visualization tasks [2].
4. Machine Learning: Machine Learning is about discovering patterns buried in the data to help manufacturers bring about operational and business transformation. Recent developments in advanced computing, analytics, and low cost sensing have the potential to bring about a transformation in the industry. The implementation of Machine Learning and Big Data may drive the next wave of innovation and may soon prove to be an unavoidable tactical move in achieving higher levels of optimization.

V. CONCLUSION

To put it in a nutshell, large volumes of varied data are being accumulated at an unpredictable rate. When properly captured and analyzed, Big Data can provide unique insights into market trends, equipment failures, buying patterns, maintenance cycles and many other business issues, lowering costs, and enabling more targeted business decisions.

Data analysis is a clear bottleneck in many applications, both due to lack of scalability of the underlying algorithms and due to the complexity of the data that needs to be analyzed. Presentation of the results and its interpretation by non-technical domain experts is crucial to extracting actionable knowledge. While many organizations have achieved proficiency in exploiting their data through data analysis, some are still at the early stages of creating an analytic model that can deliver real business value from Big Data. The main obstacles are these slow and arcane processes for enabling direct and timely access to corporate data. However, new technologies are collapsing the old walls between IT and data analysts by enabling advanced analytics within the database itself, alleviating the need to move large volumes of data around.

There has been a tremendous increase in the number of individuals necessary for organizations to take advantage of Big Data. Thus, 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.

REFERENCES

[1]. Phillip Russom. Big Data Analytics. TWRI Research; 2011.
[2]. Vignesh Prajapathi. Big Data Analytics with R and Hadoop. Packt Publishing.
[3]. Elisa Bertino. Big Data - Opportunities and Challenges Panel Position Paper. IEEE 37th Annual Computer Software and Applications Conference;2013.
[4]. Agrawal D., Bernstein P., Bertino E., Davidson S., Dayal U., Franklin M., ...Widom J. Challenges and Opportunities with Big Data: A white paper prepared for the Computing Community Consortium committee of the Computing Research Association; 2012.
[5]. Harshawardhan S. Bhosale, Prof. Devendra P. Gadekar. A review paper on Big Data and Hadoop. International Journal of Scientific and Research Publications; 2014.
[6]. Big Data Analytics Advanced Analytics in Oracle Database. An Oracle White Paper; 2013 Mar.
[7]. Roberto V. Zicari. Big Data: Challenges and Opportunities. ODBMS.org; 2013 Jul 26.
[8]. Shobhit Srivastava, S. Venkatesan, S. Amutha. Big Data Analysis and Its Tools – A Review. International Journal for Research in Applied Science & Engineering; 2016.
[9]. Kalyani Shirudkar, Dilip Motwani. Big-Data Security. International Journal of Advanced Research in Computer Science and Software Engineering; 2015 Mar.
[10]. K. Michael, K. W. Miller. Big Data: New Opportunities and New Challenges. IEEE;2013.