International Journal for Modern Trends in Science and Technology, 6(8S): 31-41, 2020

ISSN: 2455-3778 online
DOI: https://doi.org/10.46501/IJMTSTCIET08
Available online at: http://www.ijmtst.com/ncracse2020.html

Fragmentation of Data beyond Traditional Databases and Opportunities using Big Data

N Phanindra[1] | N Junnu Babu[2]
[1]Computer Science and Engineering, Bapatla Engineering College, Bapatla, AP, India
[2]Computer Science and Engineering, AIMS College of Engineering, India

To Cite this Article
N Phanindra and N Junnu Babu, "Fragmentation of Data beyond Traditional Databases and Opportunities using Big Data", *International Journal for Modern Trends in Science and Technology*, 6(8S): 31-41, 2020.

Article Info
Received on 16-July-2020, Revised on 15-August-2020, Accepted on 25-August-2020, Published on 28-August-2020.

ABSTRACT

Big Data has gained much attention from the academia and the IT industry. In the digital and computing world, information is generated and collected at a rate that rapidly exceeds the boundary range. Currently, over 2 billion people worldwide are connected to the Internet, and over 5 billion individuals own mobile phones. By 2020, 50 billion devices are expected to be connected to the Internet. At this point, predicted data production will be 44 times greater than that in 2009. As information is transferred and shared at light speed on optic fiber and wireless networks, the volume of data and the speed of market growth increase. However, the fast growth rate of such large data generates numerous challenges, such as the rapid growth of data, transfer speed, diverse data, and security. Nonetheless, Big Data is still in its infancy stage, and the domain has not been reviewed in general. Hence, this study comprehensively surveys and classifies the various attributes of Big Data, including its nature, definitions, rapid growth rate, volume, management, analysis, and security. This study also proposes a data life cycle that uses the technologies and terminologies of Big Data. Future research directions in this field are determined based on opportunities and several open issues in Big Data domination. These research directions facilitate the exploration of the domain and the development of optimal techniques to address Big Data.

KEYWORDS: Big Data, Big Data Analytics, Challenges, Methods, Systematic literature review

INTRODUCTION

Big data services have created a new computational paradigm shift in data system architecture across horizontally-coupled resources to achieve the scalability needed for the efficient processing of extensive datasets. This is exactly how big data services differs from typical information system (IT). It is a technological innovation where complex unstructured and structured data are parallel distributed, stored and direct queries could be applied to these stored data . Through big data services, an enterprise could better monitor the acceptance of products/services in the marketplace and in understanding its business environment, potentially fueling competitive advantages. Big data services have potential to unleash major impacts on reducing business costs, kindling business insights, and unraveling strategic information, and subsequently boosting quality and effectiveness of corporate decision making. Service providing sectors like telecommunication, banking and finance, IT companies, e-commerce and more have quickly adopted this big data bandwagon (Chen and Zhang, 2014). However, many manufacturing firms are still sitting on the fence and are contemplating whether to move or not to adopt the big data trend (Dubey et al., 2015). This may be

due to a lack of understanding of benefits of big data services in manufacturing sector, skills and experience in handling the big data. The situation points to the need for more research to comprehend issues (e.g., data quality, perceived costs) pertaining to big data services adoption in manufacturing sector. Big Data is proficient for business application and is rapidly escalating as a segment of IT industry. It has generated significant interest in various fields including the manufacture of health care machines, Banking transactions, Social media, Satellite imaging. Traditionally data is stored in a highly structured format to maximize its informational contents. Conversely, current data volumes are driven by both unstructured and semi structured data. Billions of individuals are various mobile devices and as a result of this technological revolution. These people are generating tremendous amounts of data through the increased use of such devices. Particularly remote sensors continuously produce much heterogeneous data that are either structured or unstructured. This data also is termed as Big Data. Big Data is characterized by 3 aspects namely; (a) Numerous data (b) Categorization of data cannot be done in regular relational data bases (c) Processing of data can be generated and captured quickly. Consequently end to end processing can be obstructed by the translation between structured data in relational systems of DBMS and unstructured data for analytics. Big Data is a compilation of very huge data sets with a great diversity of types so that it becomes complicated to process by using state of the art data processing approaches. In general a data set can be called a big data if it is formidable to perform capture, visualization, analysis at current technologies. The Essential characteristics of Big Data are followed with 7 V's namely, Variety, Velocity, Volume, Virality, viscosity, visualisation. Volume is described as the relative size of the data to the processing capability. Viscosity measures the resistance to flow in the volume of data. Here the resistance may come from different data sources. Virality is described as faster distribution of information across B2B networks (Business to Business).Variety depicts the spread of data types from machine to machine and adds new data types to traditional transactional data. Velocity is described as a frequency at which the data is generated, Shared and captured.

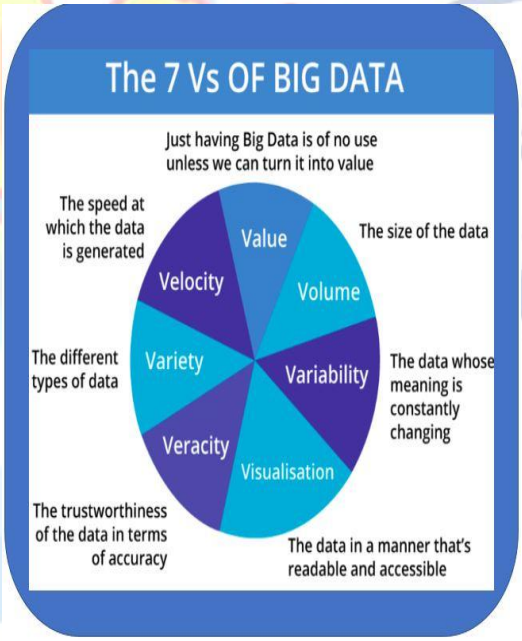
Applications of Big Data

Industry influencers, academicians, and other prominent stakeholders certainly agree that Big Data has become a big game-changer in most, if not all, types of modern industries over the last few years. As Big Data continues to permeate our day-to-day lives, there has been a significant shift of focus from the hype surrounding it to finding real value in its use.

While understanding the value of Big Data continues to remain a challenge, other practical challenges, including funding and return on investment and skills, continue to remain at the forefront for several different industries that are adopting Big Data. With that said, according to Research and Market reports, in 2017 the global Big Data market was worth $32 billion and by 2026 it is expected to reach by $156 billion.

With this in mind, having a bird's eye view of Big Data and its application in different industries will help you better appreciate what your role is or what it is likely to be in the future, in your industry or across various industries.

Banking and Securities



The Securities Exchange Commission (SEC) is using Big Data to monitor financial market activity. They are currently using network analytics and natural language processors to catch illegal trading activity in the financial markets.

Retail traders, Big banks, hedge funds, and other so-called 'big boys' in the financial markets use Big Data for trade analytics used in high-frequency trading, pre-trade decision-support analytics, sentiment measurement, Predictive Analytics, etc.

This industry also heavily relies on Big Data for risk analytics, including; anti-money laundering,

demand enterprise risk management, "Know Your Customer," and fraud mitigation.

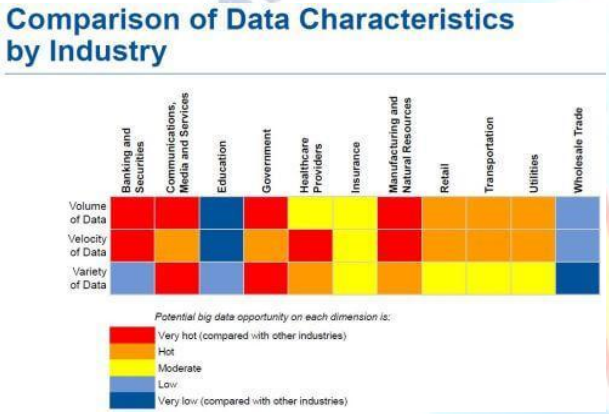
Big Data providers are specific to this industry includes 1010data, Panopticon Software, Streambase Systems, Nice Actimize, and Quartet FS.

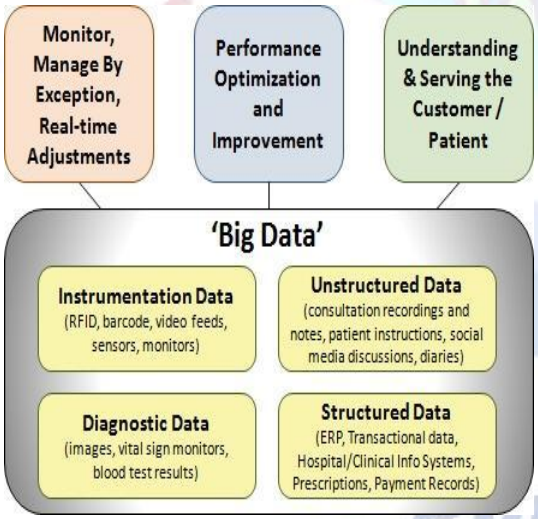
Communications, Media and Entertainment

Organizations in this industry simultaneously analyze customer data along with behavioral data to create detailed customer profiles that can be used to:

- Create content for different target audiences
- Recommend content on demand
- Measure content performance

Healthcare Providers





Some hospitals, like Beth Israel, are using data collected from a cell phone app, from millions of patients, to allow doctors to use evidence-based medicine as opposed to administering several medical/lab tests to all patients who go to the hospital. A battery of tests can be efficient, but it can also be expensive and usually ineffective.

Free public health data and Google Maps have been used by the University of Florida to create visual data that allows for faster identification and efficient analysis of healthcare information, used in tracking the spread of chronic disease.

Big Data Providers in this industry include Recombinant Data, Humedica, Explorys, and Cerner.

Education

Big data is used quite significantly in higher education. For example, The University of Tasmania. An Australian university with over 26000 students has deployed a Learning and Management System that tracks, among other things, when a student logs onto the system, how much time is spent on different pages in the system, as well as the overall progress of a student over time.

In a different use case of the use of Big Data in education, it is also used to measure teacher's effectiveness to ensure a pleasant experience for both students and teachers. Teacher's performance can be fine-tuned and measured against student numbers, subject matter, student demographics, student aspirations, behavioral classification, and several other variables.

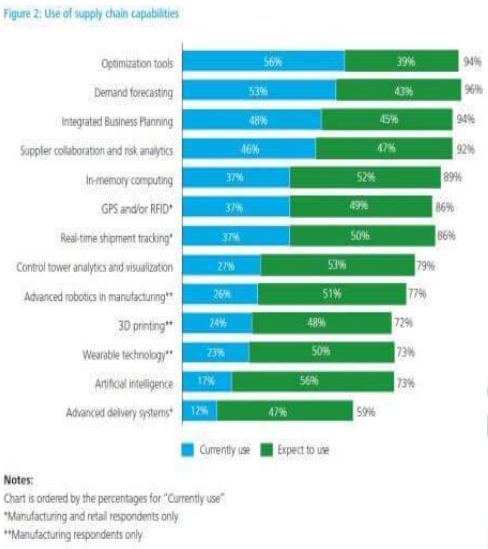
On a governmental level, the Office of Educational Technology in the U. S. Department of Education is using Big Data to develop analytics to help correct course students who are going astray while using online Big Data courses. Click patterns are also being used to detect boredom.

Manufacturing and Natural Resources

In the natural resources industry, Big Data allows for predictive modeling to support decision making that has been utilized for ingesting and integrating large amounts of data from geospatial data, graphical data, text, and temporal data. Areas of interest where this has been used include; seismic interpretation and reservoir characterization.

Big data has also been used in solving today's manufacturing challenges and to gain a competitive advantage, among other benefits.

In the graphic below, a study by Deloitte shows the use of supply chain capabilities from Big Data currently in use and their expected use in the future.



Big Data Providers in this industry include CSC, Aspen Technology, Invensys, and Pentaho.

Government

In public services, Big Data has an extensive range of applications, including energy exploration, financial market analysis, fraud detection, health-related research, and environmental protection.

Some more specific examples are as follows:

Big data is being used in the analysis of large amounts of social disability claims made to the Social Security Administration (SSA) that arrive in the form of unstructured data. The analytics are used to process medical information rapidly and efficiently for faster decision making and to detect suspicious or fraudulent claims.

The Food and Drug Administration (FDA) is using Big Data to detect and study patterns of food-related illnesses and diseases. This allows for a faster response, which has led to more rapid treatment and less death.

The Department of Homeland Security uses Big Data for several different use cases. Big data is analyzed from various government agencies and is used to protect the country.

Insurance

Big data has been used in the industry to provide customer insights for transparent and simpler products, by analyzing and predicting customer behavior through data derived from social media, GPS-enabled devices, and CCTV footage. The Big Data also allows for better customer retention from insurance companies.

When it comes to claims management, predictive analytics from Big Data has been used to offer faster service since massive amounts of data can be analyzed mainly in the underwriting stage. Fraud detection has also been enhanced.

Through massive data from digital channels and social media, real-time monitoring of claims throughout the claims cycle has been used to provide insights.

Big Data Providers in this industry include Sprint, Qualcomm, Octo Telematics, The Climate Corp.

Retail and Wholesale trade

Big data from customer loyalty data, POS, store inventory, local demographics data continues to be gathered by retail and wholesale stores.

In New York's Big Show retail trade conference in 2014, companies like Microsoft, Cisco, and IBM pitched the need for the retail industry to utilize Big Data for analytics and other uses, including:

- Optimized staffing through data from shopping patterns, local events, and so on
- Reduced fraud
- Timely analysis of inventory

Social media use also has a lot of potential use and continues to be slowly but surely adopted, especially by brick and mortar stores. Social media is used for customer prospecting, customer retention, promotion of products, and more.

Big Data Providers in this industry include First Retail, First Insight, Fujitsu, Infor, Epicor, and Vistex.

Transportation

Some applications of Big Data by governments, private organizations, and individuals include:

- Governments use of Big Data: traffic control, route planning, intelligent transport
- t systems, congestion management (by predicting traffic conditions)
- Private-sector use of Big Data in transport: revenue management, technological enhancements, logistics and for competitive advantage (by consolidating shipments and optimizing freight movement)
- Individual use of Big Data includes route planning to save on fuel and time, for travel arrangements in tourism, etc.

Energy and Utilities

Smart meter readers allow data to be collected almost every 15 minutes as opposed to once a day the old meter readers. This granular data is being used to analyze the consumption of utilities better,

which allows for improved customer feedback and better control of utilities use.

In utility companies, the use of Big Data also allows for better asset and workforce management, which is useful for recognizing errors and correcting them as soon as possible before complete failure is experienced.

Big Data Providers in this industry include Alstom Siemens ABB and Cloudera.

Challenges

a) Data challenges

Data challenges are the group of the challenges related to the characteristics of the data itself.

Different researchers have distinct understandings towards the data characteristics – such as some say 3Vs, others reported 4Vs [volume, velocity, variety, and variability] of data and 6Vs [volume, velocity, variety, veracity, variability, and value] of data. In analysing the different articles reviewed we identified 7Vs – seven characteristics of data [volume, variety , veracity , value , velocity , visualization and variability and discussed as follows:

- *Volume* (e.g. *large data-sets consisting of terabytes, petabytes, zettabytes of data – or even more*): Large scale and the sheer volume of data is a big challenge in its own right. The latter argument is also supported by Barnaghi et al. (2013) that state the heterogeneity, ubiquity, and dynamic nature of the different data generation resources and devices, and the enormous scale of data itself, make determining, retrieving, processing, integrating, and inferring the physical world data (e.g. environmental data, business data, medical data, surveillance data) a challenging task. This colossal increase of large-scale data (e.g. *Facebook daily generates over 500 terabytes of data, and Walmart collects more than 2.5 petabytes of data every hour from its customer transactions*) sets brings new challenges to data mining techniques and requires novel approaches to address the big-data problem (Zhao, Zhang, Cox, Duling, & Sarle, 2013).
- *Variety* (e.g. *multiple data formats with structured and unstructured text/image/multimedia content/audio/video/sensor data/noise*): Data challenges related to the variety (i.e. diverse and dissimilar forms) of data are also deemed a challenge. These articles revealed that the enormous volume of data is not consistent nor does it follow a specific template or format – it is captured in diverse forms and diverse sources e.g.: messages (text, email, tweets, blogs) – user generated content, transactional data (e.g. web logs, business transactions), scientific data (e.g. data coming from data-intensive experiments – genome and healthcare data), web data (e.g. images posted on social media; sensor data readings), and much more (Chen, et al., 2012a, Chen, et al., 2013). These different forms and quality of data clearly indicate that heterogeneity is a natural property of BD and it is a big challenge to comprehend and manage such data (Labrinidis & Jagadish, 2012). For instance, during the Fukushima Daiichi nuclear disaster, when the public started broadcasting radioactive material data, a wide variety of inconsistent data, using diverse and uncalibrated devices, for similar or neighboring locations was reported – all this add to the problem of increasing variety of data.
- *Veracity* (e.g. *increasingly complex data structure, anonymities, imprecision or inconsistency in large data-sets*): This is not merely about data quality – it is more about understanding the data, as there are integral discrepancies in almost all the data collected. IBM came up with this characteristic of data, which represents the untrustworthiness inherent in many sources of structured as well as unstructured data. Akerkar (2014) and Zicari (2014) refer veracity to as coping with the biases, doubts, imprecision, fabrications, messiness and misplaced evidence in the data. Veracity feature measures the accuracy of data and its potential use for analysis (Vasarhelyi, Kogan, & Tuttle, 2015). For instance, every customer opinion on different social media networks and web is different and unclear in nature, as it involves human interaction (Sivarajah, Irani, & Weerakkody, 2015). Moreover, the web, more specifically, is a soft medium to publish and broadcast fabricated information across multiple sources and, so it is essential to isolate the wheat from the chaff when presenting quality data. Thus, the necessity to deal with inaccurate and ambiguous data is another facet of BD, which is addressed using tools and

analytics developed for management and mining of unreliable data (Gandomi & Haider, 2015).

- *Velocity* (*e.g. high rate of data inflow with non-homogenous structure*): The challenge of velocity comes with the requisite to manage the high influx rate of non-homogenous data, which results in either creating new data or updating the existing data (Chen et al., 2013). This mainly applies to those datasets that are generated through large complex networks including data generated by the proliferation of digital devices, which are positioned ubiquitously resulting in driving the need for real-time analytics and evidence-based planning (Lu, Zhu, Liu, Liu, & Shao, 2014). For instance, Wal-Mart processes more than a million transactions each hour (Cukier, 2010). The data stemming from mobile devices and flowing through mobile apps or by using store cards (e.g. Sainsbury's card for collecting nectar points) generates floods of information that can be brought to use through producing real-time, *personalized* offers for customers. These data also provide sound information about customers, such as their geospatial location, buying behaviour and patterns, which can be analyzed in real-time to generate value for customers (Gandomi & Haider, 2015).

- *Variability* (*e.g. data whose meaning is constantly changing*): Among the seven pillars of BD, variability is another extremely essential feature but is often confused with variety. For instance, Google or Facebook repository stores and generates many different types of data. At the same time, if from these different types of data, one of them is brought to use for mining and making sense out of it but every time the data offers a different meaning – this is variability of data – whose meaning is constantly and rapidly changing. The volumes of machine and human-generated data constitute much greater and their rates of change and variability higher than process-mediated data. Variability is also related in performing sentiment analyzes. For example, in (almost) the same tweets a word can have a totally different meaning. In order to perform a proper sentiment analyzes, advocates assert that algorithms need to be able to understand the context and be able to decipher the exact meaning of a word in that context (Zhang, Hu et al., 2015). Nevertheless, this is yet still very challenging.

- *Visualization* (*e.g. presenting the data in a manner that is readable*): Visualising data is about representing key information and knowledge more instinctively and effectively through using different visual formats such as in a pictorial or graphical layout (Taheri, Zomaya, Siegel, & Tari, 2014). For instance, eBay has millions of users and from these many million users, even more millions of goods are sold every month – this generates a lot of data. To make all these data explicable, eBay considered the BD visualization tool – Tableau, which is capable of transforming large and complex datasets into spontaneous depictions. Based on these interactive results, eBay employees can visualize search relevance and quality, to monitor the latest customer feedback and conduct sentiment analysis. Chen and Zhang (2014) argue that for many existing BD applications that have poor performances in functionalities, scalability and response time, it is mainly problematic when conducting data visualization. This reason for this is a consequence of large sizes and high dimension of BD.

- *Value* (e.g. *extracting knowledge/value from vast amounts of structured and unstructured data without loss, for end users*): Storing BD is complex. For instance, significant values can be extracted from the stream of clicks left behind by the internet users – and this is becoming a backbone of the internet economy. Big data researchers consider value as an essential feature, as somewhere within that data, there is valuable information – extracting golden data (high-valued data), though most of the pieces of data independently may seem insignificant (Zaslavsky, Perera, & Georgakopoulos, 2012). Regardless of the number of dimensions used to describe BD, organizations are still faced with challenges of storing, managing and predominantly extracting value from the data in a cost effective manner (Abawajy, 2015).

b) Process challenges

Process challenges are the group of challenges encountered while processing and

analysing the data that is from capturing the data to interpreting and presenting the end results. As large datasets are usually non-relational or unstructured, thus processing such semi-structured data sets at scale poses a significant challenge; possibly more so than managing BD (Kaisler, Armour, Espinosa, & Money, 2013). In analysing the different articles reviewed the authors identified several data processing related challenges that can be grouped into 5 steps that is data acquisition and warehousing (PC_DAW) → C = 97 (42.7%), data mining and cleansing (PC_DMC) → C = 38 (16.7%), data integration and aggregation (PC_DAI) → C = 29 (12.8%), data analysis and modelling (PC_DAM) → C = 25 (11%) and data interpretation (PC_DI) → C = 15 (6.6%). As illustrated in Fig. 5, data mining and cleansing appears to be a vital step during processing the large scale unstructured data, as 97 articles out of 227 specifically discussed and highlighted the importance of this step.

- *Step 1 – Data Acquisition and Warehousing*: This challenge is related to acquiring data from diverse sources and storing for value generation purpose. The integral complexity of BD and exponentially growing demands develop unprecedented problems in BD engineering such as data acquisition and storage (Wang & Wiebe, 2014). The latter argument is supported by Paris, Donnal, and Leeb (2014) who assert that one of the prime barriers to the analysis of BD arises from a lack of data provenance, knowledge and discrepancies of scale inherent in data collection and processing. This further restricts the speed and resolution at which data can be captured and stored. As a result, this affects the capability to excerpt actionable information from the data (Chen & Zhang, 2014). To capture related and valuable information, smart filters are required that should be robust and intelligent to capture useful information and discard useless that contains imprecisions or inconsistencies – this is a challenge in itself. For the latter, efficient analytical algorithms are required to understand the provenance of data and process the vast streaming data and to reduce data before storing (Zhang, X., et al., 2015b, Zhang, F., et al., 2015a).

- *Step 2 – Data Mining and Cleansing*: This challenge relates to extracting and cleaning data from a collected pool of large scale unstructured data. Advocates of BD and BDA perceive that in identifying a better way to mine and clean the BD can result in big impact and value (Chen, Chen et al., 2012). Due to its strident, vibrant, diverse, inter-related and unreliable features, the mining, cleansing and analysis proves to be very challenging (Chen et al., 2013). For instance, in the UK National Health Service (NHS) there are many millions of patients' records comprising of medical reports, prescriptions, x-ray data, etc. Physicians make use of such data – if for instance incorrect information is stored this may lead to physicians wrongly diagnosing conditions, resulting in inaccurate medical records. In order make use of this huge data in a meaningful way, there is a need to develop an extraction method that mines out the required information from unstructured BD and articulate it in a standard and structured form that is easy to understand. According to Labrinidis and Jagadish (2012) developing and maintaining this extraction method is a continuous challenge.

- *Step 3 – Data Aggregation and Integration*: This process challenge relates to aggregating and integrating clean data mined from large unstructured data. BD often aggregates varied online activities such as tweets – retweets, microblogging, and *likes* on Facebook that essentially bear diverse meanings and senses (Edwards & Fenwick, 2015). This characteristically amorphous data naturally lacks any binding information. Aggregating these data evidently goes beyond the abilities of current data integration systems (Carlson et al., 2010). According to Karacapilidis, Tzagarakis, and Christodoulou (2013), the availability of data in large volumes and diverse types of representation, smart integration of these data sources to create new knowledge – towards serving collaboration and improved decision-making – remains a key challenge. Halevy, Rajaraman, and Ordille (2006) assert that the indecision and provenance of data are also a major challenge for data aggregation and integration. Another challenge relates to aggregated data in warehouses – in line with this argument, Lebdaoui, Orhanou, and Elhajji (2014) report that to enable decision systems to efficiently respond to the real world's demands,

such systems must be updated with clean operational data.

- *Step 4 – Data Analysis and Modelling*: Once the data has been captured, stored, mined, cleaned and integrated, comes the data analysis and modelling for BD. Outdated data analysis and modelling centers around solving the intricacy of relationships between schema-enabled data. As BD is often noisy, unreliable, heterogeneous, dynamic in nature; in this context, these considerations do not apply to non-relational, schema-less databases (Shah et al., 2015). From the perspective of differing between BD and traditional data warehousing systems; Kune, Konugurthi, Agarwal, Chillarige, and Buyya (2016) report that although these two have similar goals; to deliver business value through the analysis of data, they differ in the analytics methods and the organization of the data. Consequently, old ways of data modelling no longer apply due to the need for unprecedented storage resources/capacity and computing power and efficiency (Barbierato et al., 2014). Thus, there is a need for new methods to manage BD for maximum impact and business value. It is not merely knowing about what is currently trendy, but also need to anticipate what may happen in the future by appropriate data analysis and modelling (Chen et al., 2013).

- *Step 5 – Data Interpretation*: This step is relatively similar to visualising data and making data understandable for users that is the data analysis and modelling results are presented to the decision makers to interpret the findings for extracting sense and knowledge (Simonet, Fedak, & Ripeanu, 2015). The astounding growth and multiplicity of unstructured data have intensely affected the way people process and interpret new knowledge from these raw data. As much of these data both instigate and reside as anonline resource, one open challenge is defining how Internet computing technological solutions have evolved to allow access, aggregate, analyze, and interpret BD (Bhimani & Willcocks, 2014). Another challenge is the shortage of people with analytical skills to interpret data (Phillips-Wren & Hoskisson, 2015).

c) Management challenges

Management challenges related to BD are a group of challenges encountered, for example while accessing, managing and governing the data. Data warehouses store massive amounts of sensitive data such as financial transactions, medical procedures, insurance claims, diagnosis codes, personal data, etc. Organizations and businesses need to ensure that they have a robust security infrastructure that enables employees and staff of each division to only view relevant data for their department. Moreover, there must be some standard privacy laws that may govern the use of such personal information and strict observance to these privacy regulations must be applied in the data warehouse. In analysing the different articles reviewed in this SLR, the authors identified several data management related challenges that can be grouped into seven areas (Fig. 6) such as privacy (MC_P) → C = 23 (10.1%), security (MC_S) → C = 17 (7.5%), data and information sharing (MC_D&IS) → C = 10 (4.4%), cost/operational expenditures (MC_C&OE) → C = 7 (3.1%), data governance (MC_DG) → C = 4 (1.8%), and data ownership (MC_OG) → C = 3 (1.3%).

- *Privacy*: BD poses big privacy concerns and how to preserve privacy in the digital age is a prime challenges. Huge investments have been made in BD projects to streamline processes; however, organizations are facing challenges in managing privacy issues, and recruiting data analysts, thus hindering organizations in moving forward in their efforts towards leveraging BD (Krishnamurthy & Desouza, 2014). In a smart city environment where sensory devices gather data on citizen activities that can be accessed, several government and security agencies pose significant privacy concerns (Barnaghi et al., 2013). Among such privacy related challenges, location-based information being collected by BD applications and transferred over networks is resulting in clear privacy concerns (Yi et al., 2014). For example, location-based service providers can identify subscriber by tracking their location information – which is possibly linked to their office or residential information. Then there is the challenge of protecting privacy – Machanavajjhala and Reiter (2012) report that failure to protect citizens' privacy is illegal and open to relevant Government oversight bodies.

- *Security*: Security is a major issue and is identified by Lu et al. (2014) who argue that if security challenges are not appropriately addressed then the phenomenon of BD will not receive much acceptance globally. Securing BD has its own distinctive challenges that are not profoundly different to traditional data. Among the several BD related security challenges are the distributed nature of large BD which is complex but equally vulnerable to attack (Yi et al., 2014), malware has been an ever growing threat to data security (Abawajy, Kelarev, & Chowdhury, 2014), lack of adequate security controls to ensure information is resilient to altering (Bertot, Gorham, Jaeger, Sarin, & Choi, 2014), analysing logs, network flows, and system events for forensics and intrusion detection has been a challenge for data security (Cárdenas, Manadhata, & Rajan, 2013), lack of sophisticated infrastructure that ensures data security such as integrity, confidentiality, availability, and accountability, and data security challenges become magnified when data sources become ubiquitous (Demchenko, Grosso, De Laat, & Membrey, 2013).

- *Data Governance*: As the demand for BD is constantly growing, organizations perceive data governance as a potential approach to warranting data quality, improving and leveraging information, maintaining its value as a key organizational asset, and support in attaining insights in business decisions and operations (Otto, 2011). According to Intel IT Centre (2012), IT managers highly support the presence of a formal BD strategy, this especially makes sense, since the issue of data governance for describing what data is warehoused, analyzed, and accessed is termed as one of the three top challenges they face (besides data growth and data centre infrastructure and the ability to provide scalability). du Mars (2012) state that a significant challenge in the process of governing BD is categorizing, modelling and mapping the data as it is captured and stored, mainly due to the unstructured and complex nature of data. Moreover, effective BD governance is essential to ensure the quality of data mined and analyzed from a pool of large datasets (Hashem et al., 2015).

- *Data and Information Sharing*: Sharing data and information needs to be balanced and controlled to maximise its effect, as this will facilitate organizations in establishing close connections and harmonisation with their business partners (Irani, Sharif, Kamal, & Love, 2014). However, where organizations store large scale datasets that have potential analysis challenges, it also poses an overwhelming task of sharing and integrating key information across different organizations (OSTP, 2012). Al Nuaimi et al. (2015) also state that sharing data and information between distant organizations (or departments) is a challenge. For instance, each organization and their individual departments typically own a disparate warehouse (developed based on different technological platforms and vendors) of sensitive information and several departments are often reluctant to share their patented data governed by privacy conditions. According to Khan, Uddin, and Gupta (2014) the challenge here is to ensure not to cross the fine line between collecting and using BD and guaranteeing user privacy rights. The is also related to a smart city environment that entails a plethora of sectors and in such context, smart city technological systems will need to reduce the barriers to achieve seamless information sharing and exchange among different entities (Su, Li, & Fu, 2011).

- *Cost/ Operational Expenditures*: The constantly increasing data in all different forms has led to a rising demand for BD processing in sophisticated data centers. These are generally dispersed across different geographical regions to embed resilience and spread risk, for example Google having 13 data centers in eight countries spread across four continents (Gu, Zeng, Li, & Guo, 2015). The significant resources have been allocated to support the data intensive operations (i.e. acquisition, warehousing, mining and cleansing, aggregation and integration, processing and interpretation) – all this lead to high storage and data processing *big costs* (Raghavendra, Ranganathan, Talwar, Wang, & Zhu, 2008). Researchers assert that cost minimization is an emergent challenge (Irani, Z., et al., 2006, Irani, Z., 2010), with Gu et al., 2015 explaining the challenges of processing BD across geo-distributed data centers. Advocates of BD search for cost-effective and efficient ways to handle the massive amount of complex data (Sun, Morris, Xu, Zhu, & Xie,

2014). The cost of data processing and other operational expenditures of the data center are a sensitive issue that may also impact in the way organizations adopt and implement technological solutions (Al Nuaimi et al., 2015).

- *Data Ownership*: Besides privacy, Web (2007) asserts that ownership of data is a complex issue – as big as the data itself – while sharing real time data. Kaisler et al. (2013) also claim that data ownership presents a critical and continuing challenge, specifically in the social media context such as who owns the data on Facebook, Twitter or MySpace – are the users who update their status or tweet or have any account in these social networks (Sivarajah et al., 2015, Sivarajah et al., 2014). It is generally perceived that both view they (the users and the social media provider) own the data. Kaisler et al. (2013) argues that this dichotomy still needs to be settled. With ownership arise the issue or controlling and ensuring its accuracy. For instance, Web (2007) states that sensor data is too sensitive and can result in mounting errors – this may further result in capturing and revealing inconsistent data – but then who owns that data. Data ownership is a much deeper social issue. These concerns are beyond the focus on several applications, for example SensorMaps by Web (2007) requires more research since they may have deep implications.

The following are the other challenges need to be considered before implementing a Big Data & Analytics solution:

1. Data Quality – In a credit union, data is coming from many disparate sources from all facets of the organization. In order to overcome this, a data warehouse is essential. However, when a data warehouse tries to combine inconsistent data from disparate sources, it encounters errors. Inconsistent data, duplicates, logic conflicts, and missing data all result in data quality challenges. Poor data quality results in faulty reporting and analytics necessary for optimal decision making.

2. Understanding Analytics – The powerful analytics tools and reports available through integrated data will provide credit union leaders with the ability to make precise decisions that impact the future success of their organizations. When implementing a Big Data & Analytics solution, analytics and reporting will have to be taken into design considerations. In order to do this, the business user will need to know exactly what analysis will be performed. Envisioning these reports will be difficult for someone that hasn't yet utilized a Big Data & Analytics solution and is unaware of its capabilities and limitations.

3. Quality Assurance – The end user of a Big Data & Analytics solution is using reporting and analytics to make the best decisions possible. Consequently, the data must be 100 percent accurate or a credit union leader will make ill-advised decisions that are detrimental to the future success of their business. This high reliance on data quality makes testing a high priority issue that will require a lot of resources to ensure the information provided is accurate. The credit union will have to develop all of the steps required to complete a successful Software Testing Life Cycle (STLC), which will be a costly and time intensive process.

4. Performance – Implementing a Big Data & Analytics solution is similar to building a car. A car must be carefully designed from the beginning to meet the purposes for which it is intended. Yet, there are options each buyer must consider to make the vehicle truly meet individual performance needs. A Big Data & Analytics solution must also be carefully designed to meet overall performance requirements. While the final product can be customized to fit the performance needs of the organization, the initial overall design must be carefully thought out to provide a stable foundation from which to start. Major customizations are extremely expensive.

5. Designing the Solution – People generally don't want to "waste" their time defining the requirements necessary to properly design Big Data & Analytics solution. Usually, there is a high level perception of what is wanted out of a Big Data & Analytics solution. However, they don't fully understand all the implications of these perceptions and, consequently, they have a difficult time adequately defining them. This results in miscommunication between the business users and the technicians developing a Big Data & Analytics solution.

6. User Acceptance – People are not keen to changing their daily routine especially if the new process is not intuitive. There are many challenges to overcome to make a Big Data &Analytics solution that is quickly adopted by an organization. Having a comprehensive user training program can ease this hesitation but will require planning and additional resources.

7. Cost – A frequent misconception among credit unions is that they can develop a Big Data & Analytics solution in-house to save money. As the foregoing points emphasize, there are a multitude of hidden problems in developing a Big Data & Analytics solution. Even if a credit union adds a data "expert" to their staff, the depth and breadth of skills needed to deliver an effective result is simply not feasible with one or a few experienced professionals leading a team of non-BI trained technicians. The harsh reality is an effective do-it-yourself effort is very costly.

CONCLUSION

This paper presents the elementary concepts of Big Data. These concepts comprise the role of Big Data in the current environment of enterprise and technology. To augment the efficiency of data management, we have devised a data-lifecycle that uses the technologies and terminologies of BigData. The stages in this life cycle include collection, filtering, analysis, storage, publication, retrieval, and discovery. Data are also generated in different formats(unstructured and/or semi structured), which unfavorably affect data analysis, management, and storage. This deviation in data is accompanied by complexity and the development of further means of data acquisition. Big Data has developed such that it cannot be harnessed separately. Big Data is characterized by large systems, profits, and challenges. As a result, additional research is obligatory to address these issues and advance the efficient display, analysis, and storage of Big Data. To improve such research, capital investments, human resources, and pioneering ideas are the basic requirements.

REFERENCES

[1] C.L. Philip Chen, Chun-Yang Zhang, "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data", Information Sciences, www.elsevier.com/locate/ins, January 2014.

[2] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data", IEEE Transactions On Knowledge and Data Engineering, vol. 26, no. 1, January 2014,pp.97-107.

[3] "IBM What Is Big Data: Bring Big Data to the Enterprise," http://www01.ibm.com/software/data/bigdata/, IBM, 2012.

[4] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data", IEEE Transactions On Knowledge and Data Engineering, vol. 26, no. 1, January 2014,pp.97-107.

[5] unping Zhang, Fei-Yue Wang, Kunfeng Wang, Wei-Hua Lin, Xin Xu, Cheng Chen, Data-driven intelligent transportation systems: a survey, IEEE Trans. Intell. Trans. Syst. 12 (4) (2011) 1624–1639.

[6] D. Che, M. Safran, and Z. Peng, "From big data to big data mining: challenges, issues, and opportunities," in Database Systems for Advanced Applications, B. Hong, X.Meng, L. Chen,W. Winiwarter, and W. Song, Eds.,

[7] Z. Sebepou and K. Magoutis, "Scalable storage support for data stream processing," in Proceedings of the IEEE 26th Symposium on Mass Storage Systems and Technologies (MSST '10), pp. 1–6, Incline Village, Nev, USA, May 2010.

[8] A. Katal, M. Wazid, and R. H. Goudar, "Big data: issues, challenges, tools and good practices," in Proceedings of the 6th International Conference on Contemporary Computing (IC3 '13), pp. 404–409, IEEE, 2013.

[9] A. Azzini and P. Ceravolo, "Consistent process mining over big data triple stores," in Proceeding of the International Congress on Big Data (BigData '13), pp. 54–61, 2013.

[10] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'Big data', Hadoop and cloud computing in genomics," Journal of Biomedical Informatics, vol. 46, no. 5, pp. 774–781, 2013.

[11] Y. Demchenko, P. Grosso, C. de Laat, and P. Membrey, "Addressing big data issues in scientific data infrastructure," in Proceedings of the IEEE International Conference on Collaboration Technologies and Systems (CTS '13), pp. 48–55, May 2013.

[12] Y. Demchenko, C. Ngo, and P. Membrey, "Architecture Framework and Components for the Big Data Ecosystem," Journal of System and Network Engineering, pp. 1–31, 2013.

[13] M. Loukides, "What is data science? The future belongs to the companies and people that turn data into products," AnOReilly Radar Report, 2010.

[14] A.Wahab,M. Helmy, H.Mohd, M. Norzali, H. F. Hanafi, and M. F. M. Mohsin, "Data pre-processing on web server logs for generalized association rules mining algorithm," Proceedings of World Academy of Science: Engineering & Technology, pp. 48–53

[15] http://healthdataalliance.com/

[16] http://www.firstpost.com/business/big-data-booster-shot healthcare-industry-needs-2160271.html

[17] Chester Curme, Tobias Preis, Eugene Stanley, Helen Susannah Moat, —Quantifying the semantics of search behavior before stock market moves ; CrossMark, December 2013

[18] Nitish Sinha, —Using Big Data in Finance: Example of sentimentextraction from news articles ; FEDS notes, March 2014

[19] Baker, Malcolm and Jeffrey Wurgler, 2007. "Investor Sentiment in the Stock Market", Journal of Economic Perspectives, vol. 21(2), pages 129-152.

[20] Heston, Steven L. and Sinha, Nitish Ranjan, 2013. "News versus Sentiment: Comparing Textual Processing Approaches for Predicting Stock Returns", Robert H. Smith School Research Paper. Available at SSRN: http://ssrn.com/abstract=2311310or http://dx.doi.org/10.2139/ssrn.2311310