UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

Real-Time Big Data Analytics: Investigating Different Application Domains

Eman A. Alhazmi[1], Walaa A. Bajunaid[2], Fahd S. Alotaibi[3]

Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Saudi Arabia[1,2,3]

Abstract: Big Data Analytics is one of the primary focuses for data science. Its importance lies in its application in gathering information by many public and private organizations. The domains' specific information may have critical content concerning cyber security, national intelligence, health information, economics, and fraud detection. The current operations in technology and common activities revolve around the transfer of large data, making them predominate in the operations and the handling of real-time data. This paper aims at analyzing different application domains in the information technology architecture that involve the visualization of the massive data systems such as the X-SimViz. The framework is necessary for making analytics and visualization of real-time big data. The paper will address the issue of the big data major challenges facing design, implementation \& operation through describing the methodology, result, discussion and concluding remarks on devising approaches and strategies that could assist in reconciling security with privacy in Real-time Big data analytics.

Keywords: Big Data Analytics, Real-Time Analytics, Big Data Challenges.

I. INTRODUCTION

Big data is universal and has become popular in the bioinformatics, scientific, intelligence and internet domains. The data has been applicable to areas with scarce, missing or unrepresentative information in the domain. Scientists and enterprise can get usable information from datasets that are involved, enormous and interconnected. Such aspects as exponential growth, availability and reliability are the key components described under real-time big data. To achieve this, data used must either be semi-structured, unstructured, or even structured. The four dimensions that are useful include volume, velocity, variety and veracity. Real-time big data analytics focus mainly on type because it is an aspect that has allowed unstructured and semi structured data. The 1990's development of grid technologies primary focus was on selecting, sharing and aggregating some resources that were distributed geographically (Ellis, 2014).

It was the peer-to-peer systems that were responsible for the concept of using as a resource. The fundamental notion of service has made cloud computing environments to target clients that are accessible, reliable and with robust services. The greatest challenge that users of information technology are facing in the 21st century is the way to handle large and complex data collection.

Nevertheless, real-time big data makes it easy for persons in the science and business industry to gather extensive information from many sources quickly. People often access information that is conflicting, dynamic and disparate from the various sources resulting in an overload problem with the information gathered. Real-time big data analytics are essential especially in application areas that require processing and analyzing of large information space. This paper is organized in a way that challenges and complexities that come with big data handling will be analyzed using different domains (Ruan, 2013).

II. A BACKGROUND ON APPLICATION OF REAL-TIME BIG DATA ANALYTICS

Over the past few years, real-time processing or time-critical processing along with big data analytics have received a considerable degree of focus from the global business community. The term “Big Data” refers to a massive dataset that encompasses a large variety of characteristics and attributes while also being fairly complex in structure. In this regard, Big Data Analytics refers to the procedures or methods that can assist people, organizations, and industries to make sense out of these huge chunks of data sets to facilitate decision making.

On the other hand, the term “Real-Time” refers to the processing and completion of a particular task within milliseconds; or simply put, it refers to an immediate completion of a specific task (Merriam-Webster, 2016). However, from the computing perspective, real-time refers to the degree of responsiveness of a computing program such that it processes information almost immediately without having to wait for even a minute (Tech Target, 2016). This implies that Real-time Big Data Analytics refers to methods or procedures that have been developed to instill efficiency into

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

big data analytics, so that massive volumes of highly complex data sets could be analyzed almost immediately, and is adequate to make sense out of it to empower decision-making. Because if decisions fail to be made within the given timelines, the whole process is rendered not useful (Nader & Al-Jaroodi, 2014).

Applications of real-time big data analytics include, but are not limited to business operations, organizational decision-making, management, general information processing, disaster management response and recovery, and others. Due to its ability to accurately perform time-critical information processing faster than traditional big data analytics, real-time big data analytics has gained the spotlight across multiple application domains.

III. BIG DATA CHARACTERISTICS

In the field of information technology, the term Big Data is called on a group of packages of massive and complex data which's hard to deal with them by traditional database management systems (DBMS) in terms of storage space, search, representation, and analysis (Nader & Al-Jaroodi, 2014). In fact, it is difficult to define Big Data accurately, that; the difference of huge data could be considered from one organization to another due to the different capabilities of each of them individually.

As indicated earlier, big data comprises of large datasets with a variety of characteristics that must be grouped together to make sense out of the entire dataset. From a general perspective, big data is characterized by structured, unstructured and semi-structured datasets that cannot be processed as effectively by traditional database management systems. In general, when we mention Big Data, we are talking about multiple types and sources of data size up to hundreds of terabytes or even petabytes. However, the issue is not only the size of the data; but there are also other dimensions of the Big Data known as the "5 Vs", including volume, variety, velocity, veracity and value are as shown in the figure 1; (Bello-Orgaz et al.,2016).

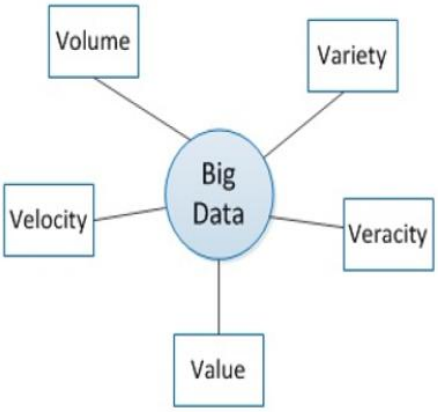

Fig. 1. 5 V's of Big Data

A. Volume:
As the name implies, volume means the overall size of datasets involved, which may be in Exabyte. It refers to the large amounts of any data type from a variety sources, such as digital devices and mobile digital data creation devices. Take advantage of the collection, processing and analysis of these large amounts of data it generates many challenges in obtaining valuable information for people and organizations (Bello-Orgaz et al.,2016).

B. Variety:
This refers to the variety of information entailed in big data sets, including but not limited to file formats, file types, and the availability of structured, unstructured and semi-structured data. In other word (Bello-Orgaz et al.,2016). With the increasing numbers of the Internet users, smartphones, and different social networks, the form of the common data has been changed from structured data in databases to unstructured data which includes a large number of formats such as images, audio, video, text and data of GPS and documents multiple-format such as MS Word and PDF (Nader & Al-Jaroodi 2014).

C. Velocity:
This characteristic specifically relates to the rate at which big data is being produced, processed and updated. It intended to speed the production and extraction of data to cover the demand. For example, to get the data in real time, or for varying periods (Bello-Orgaz et al., 2016). With the huge volume of data and the increasing of speed frequency,

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

the need to ensure high-speed system in the massive data analysis in "Real Time" or the speed of convergence of real-time more urgent. That need led to technology innovation, such as Apache Hadoop and SAP HANA.

D. Veracity:
It refers to the correctness and accuracy of information that relates directly to the extent to which data in big datasets is genuine. Behind any information management practice lies the fundamental doctrines of data quality, and management of metadata, data governance, as along with considerations of privacy and legal concerns (Bello-Orgaz et al.,2016).

E. Value:
It refers to extract valuable information from large data sets. This attribute involves the overall quality of data entailed within the dataset that can assist in making better, more informed decisions (Jony, 2016). Value is the most important characteristic of any big-data-based application because it allows generation of useful business information. (Bello-Orgaz et al.,2016).

IV. REAL-TIME BIG DATA APPLICATIONS

We now know that the big data around us everywhere! Extensive areas of application include various sectors from health to education to social networks and other sectors. While Real-time applications vary from regular applications; that real-time applications depend on immediate input and quick analysis to make the decision within the timeline. Thus, if the decision failed to be made within the given timelines, it becomes useless. Consequently, it is important making all the required data for such decision available on proper time, and analyzing in a fast and reliable way (Nader & Al-Jaroodi, 2014).

A. Real-time Big Data Analytics and Visualization Architecture
Data and information that is held and contained in the vast data networks are massive. A critical role is played by the interactivity and analysis of networks that deal with large data. Big data visualization is particularly important in the process of analyzing any data more particularly in the current emergence of the social network. More attention should be focused on maintaining visualization and interactivity that is superior to scale the rapid evolution of social networks that is always and incessantly evolving. It is complex to understand a network that is multifaceted and one that involves calculations and computations of a variety of network metrics. The enormous nature of the system is time-consuming because it involves utilizing the device memory and numerous resources in computing required factors.

Evidently, the emergence large scale big data demanded the technological field to develop more advanced and capable systems for the managing this form of data. In that case, the architecture in the network analytics and data visualization had to be advanced in terms of the technological capacity. In the process, they could be equal and capable of handling the task of managing the large-scale data volumes particularly in the transfer of social-technical data.

According to Chopade (2015), the rapid increase in the social-technical data resulted in a systematic challenge that affected the functioning of the system. In that case, there was a need to enhance the visualization and the analytical aspects of the systems. Therefore, such technological aspects must be faced out or improved to be relevant in the modern day Through this understanding, researchers and technological experts agree that the technological systems must be upgraded with time to be relevant and effective in solving modern problems. In that regard, the X-SimViz development had offered a solution of clearly understanding the data system, and the process of visualization requires that the raw data be transformed into a format that can be easily understood by the user. The system was effective because it had been an upgrade of the former systems. In fact, its designs considered the modern analytical challenges and developments that were not considered in the former analytical systems (Chopade et al., 2015a).

Over time, the complexity of the technological systems in data management continues to increase due to the increasing data volumes they serve. Rather than simply executing their roles in simplicity, they are more complex to offer a wide range of roles. In that case, the field is experiencing an evolution that aims to enhance the entire data visualization field (Selim et al., 2014). With the changing dynamics of data management, Chopade (2014) demonstrates how the complexity of the data management systems has increased. In this case, he formulated an interface component that allowed the user to access Visualizer and it exists at different complexity levels that at some point are characterized as an intelligent part. Irrespective of the complexities involved, the system must achieve its intended purpose. In this case, for example, the Visualizer results showed by providing communication between the interface and the user (Chopade et al., 2015b).

Furthermore, the large-scale data has triggered the increase in network requirements. This means that data cannot only be managed by the current network systems but complex network systems for efficient management. Despite this, the

UGC Approved Journal
IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 4, Issue 7, July 2017

successful application of the X-SimViz suggests that the large data networks must have adequate basic requirements to be successful (Chopade et al., 2015a).

B. Big Data Stream Computing in Healthcare Real-Time Analytics

With the advanced technology, health facilities like all other institutions have invested heavily in technology to enhance their performance. This has been done to enhance the operations while offering quality services to the people. The recent health care system changes have, however, resulted in difficult processes that are utilized in the health service provision. In the process, the care and service delivery have been influenced by these processes due to the acquisition of extensive data from the non-clinical and clinical environments (Ta et al., 2016). In the long run, the process has impacted both the caregiver and the medical practitioners who are evident in the care and service delivery output.

With the adoption of the big data stream systems, the medical field has become more liberal regarding access to patient information by the medical officers. For instance, patients' information can be gathered from multiple sources like clinical tests, imaging, social media, EHRs, genomic and biomedical data. Therefore, real-time big data is a necessity than an opportunity and should be treated as so.

Furthermore, the big data stream has resulted in various benefits in the medical field. According to (Raghupathi, 2014), big data stream computing has improved health care services and decreased medical costs through allowing for real-time decisions. By using common database systems, the health care system can benefit significantly from the efficiently in data management by these programs. For instance, the use of NoSQL and Apache Kafka databases like Cassandra, HBase or Apache Storm as mainstream computing technologies for medical operations can lead to effective and efficient data management. This is because the databases ensure that the medical records are safely stored and easily retrieved whenever needed. Also, the databases offer the security of medical records in case of disaster. As such, the data can easily be retrieved from the online database with ease (Ta et al., 2016). In addition, the use of database facilitates data consistency of medical records. Also, these technologies are important in the elimination of fraud, and pervasive health. They are also helpful in the support system for clinical decision. The expanding big data analytical field is playing a vital role in the health care evolution especially in research and practices (Raghupathi, 2014).

Furthermore, real-time big data allows for accumulation, management, analysis and assimilation of massive volumes of unstructured, disparate and structured data produced present in the current health care systems. This was tested through a study by (Archenaa, 2015), which involved the collection of data in the healthcare system and spreading it amongst hospitals, clinical centers, research labs, health insurers and government entities. The findings were data collection results that streamed in rapid and overwhelming speeds. This was an indicator of high big data velocity in the health care (Archenaa, 2015). In practical applications, real-time data analytics in health care helps in prediction and prevention of epidemic, decision making in medicine, and system recommendation for health (Raghupathi, 2014).This demonstrates that there is an improvement in healthcare quality and a decline in accompanying costs.

C. Real Time Big Data Analytics for Predicting Terrorist Incidents

Terror has become a common phenomenon in the modern society. The rapid rise in terrorism incidents which has become extremely lethal in the last decade demonstrates the volatility of the human society. However, the phenomenon is evolving and complicated because there has been a massive use of the technology in executing these heinous attacks. Despite the threat being minimal in America in the recent past, there is an alarming growth of foreign terrorist's cells like Al-Qaeda and ISIS that have radicalized unsuspecting individuals (Strang and sun, 2015).

Unlike in the past, the modern terrorists use the sophisticated modern weapons in their massive attacks. This tends to get many by surprise and hence the fatalities mushroom over days due to the following effects of attacks. However, the use of real-time big data has significantly enhanced the environment for war against terror. This is because the experts can narrow the gap between the terrorist groups and the anti-terrorism experts through the development of methodologies that are novel and proven together with systems that are sophistically designed to address the rampant terrorism issues (Strang and sun, 2015).

Numerous efforts have been taken in the collection and analysis of terrorism incident data in the world. According to (Toure and Gangopadhyay, 2016), the use of big real time data can enhance the war on terror. This is whereby the experts can collect real-time terror data from reliable sources and sue it to calculate the terrorism risk levels for particular locations. In this case, the extraction task is driven by the new event and the following strategies of document selection that are based on the pattern. The aim of the method is to extract terrorism events information timely and choosing minimal materials for construction of entity and relation instances that is relevant to the terrorism event (Toure and Gangopadhyay, 2016).

According to Toure and Gangopadhyay (2015) study, a model was developed and software to collect terrorism incidents and then analyze the data to calculate the terrorism risk for the different locations. The models gathered data

IARJSET
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
ISO 3297:2007 Certified
Vol. 4, Issue 7, July 2017

automatically from news sources and analyze the data to create a risk factor indicator. The model was developed based on frequency, time and normalization factors. After testing with the Baltimore shooting, the model was concluded to have a 1.5-mile accuracy from the incident location.

D. Bandwidth Availability in Cellular Networks

The great data creation in the collection of digital data extremely increased emerging pressures on cellular networks, and the result indicated the requirement for the analysis and estimation of link status and capacity. In the process, the appropriate and applicable data monitory and analysis approaches must be identified based on the available bandwidth. Therefore, it is possible to analyze the cellular network regardless of the used bandwidth (Rebecchi et al., 2015). According to Arjun Sahni et al. (2015), the bandwidth data was collected through the active which involves traffic probing (BART, Bandwidth available in Real-Time), and passive methods which included collecting data from SNMP (Simple Network Management Protocol). In most cases, however, the standard methodologies used in the collection is SNMP. Furthermore, it is necessary to have a reliable monitory mechanism in handling large data. This can, however, be realized through the establishment of a complete maintenance and operation of digital data. Evidently, the rapid increase in the rate of data flows on the Internet results in putting and subsequent downloading of significant information on the internet. More than ever, the internet is currently a tremendous tool for research and countless applications. The increase in Internet usage has led to providers offering extensive services to both users and corporations, and all web end users demand a service that is of high speed and quality in transmission.

There is the uneven distribution of network topologies where there is significant variation in Internet use regarding speed and access. Example, some users have the internet connection that is commonly shared and the result is congestion of the network following high usage. Evidently, the bandwidth congestion sets back transmission rates of data and the overall independent destination quality are affected (Sahni et al., 2015). Therefore, it necessary to measure and analyzes the application in an active manner to avoid all negative consequences associated with bandwidth.

Basically, a standard Internet system does not have the adequate capacity to handle vast and diverse data, therefore real-time big data analytics is the appropriate solution that provides for growing demand for internet content. By estimating the number of traffic based on the real-time big data, the management of the web service providers can then weigh the level or speed of internet service they can provide to their customers. By categorizing the trend of the data request by end users for the big data cache in a particular base station, the results depict a decline in time consumed for the application (Russom, 2011).

E. Improving Worker Safety at the Workplace

One of the domains that call for the application of real-time data processing is manufacturing. Other than the massive production optimization trends that are based on improving resource consumption or quality control on the real-time floor, there is desperate need of making working space safe, user-friendly and ergonomic. The relevance of this requirement is in the facilities that utilize modern manufacturing facilities that have modern cameras and other environmental or personal monitoring systems. There is a challenging task on proper control in production based on the real-time processing. Industrial companies should use big data analytics to improve their business. This will make big data famous because the manufacturing businesses and other organizations will familiarize with the capabilities that come with big data analytics (Stojadinovic et al., 2015).

As a developing technology, real-time big-data analytics in the construction industry will result in the discovery of greater benefits and uses. From a real-time point of view, it is quite a challenge for manufacturing companies to use personal monitoring in their processes. This ranges from the situation complexity to data quality that the control system can acquire for safety (Jagadish et al. 2014).

F. Achievements and New Challenges of Social Big Data

There has been an increase in the use of big data in most research areas in the recent past. The importance of big data is in the field of computational intelligence, data mining, and information fusion, data mining and social networks. Big data comprises of structured and unstructured diverse data that moves at high speed making it difficult or impossible for traditional software and database technology to process. Most enterprises are faced with quick streaming enormous data that has variability in speed that at times it exceeds the average processing capacity (Bello-Orgaz et al. 2016). The high volume and high velocity for big data demand innovative and profitable methods to enhance decision making and insights. The primary challenge in business analytics is big data because it lacks standard procedures and tools designed for searching and analyzing massive sets of data.

The challenge in big data management results from velocity, volume, and variety. Big data management may lead to the introduction of new management systems where managers will be required to capture better and more accurate data

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

because it is gradually becoming a way of life for many businesses. Technologies for big data analysis on their own are not sufficient enough to handle the necessary operations. There is the need for well-planned human resource and analytical processes with appropriate skills and talent that are required for leveraging essential technologies (Bello-Orgaz et al. 2016).

G. Educational data virtual world

The digital educational policy and governance, promote the integration and collaboration that represents the influential social reality of education globally. The digital education system manipulation was initially mobilized by government agencies such that their actions were streamlined and consistent and increased contemporary dominance. Today, educational data has been cascaded using the enumeration of logical operations that are more efficient and able to handle educational data. Manufacturers can reduce variability that is brought about by wasteful production processes. The result will be a dramatic improvement in yield and quality of output per unit input (Williamson, 2016).The components further improve educational governance through researching behavior and mapping out emerging issues. In addition, there were visible functioning schools through evidence by both central government and public as relates to the publicly available technological instruments like school websites. The data quality is developed by government agencies and the public as end users will demonstrate the way they can utilize data and transfer it into usable knowledge for better management services. In turn, the governments can focus on fears by schools concerning relevant, transparent online and database technologies (Williamson, 2015).

H. Importance of Big data in smart city

The focus on the implementation of big data application which has support for smart city components of sustainability level has been driven by utilization of multiple technologies. The fact that citizens are more actively involved as a core target of smart cities has the corresponding advantage in the reduction of the cost of living and the rate at which resources might be depleted. Previous projects have been carried out using smart city big data components (Ellis, 2014). The element has been used by government and agency administration in the management of the real-time process of analyzing large streaming data in the USA. The smart city component can develop infrastructure that is massive and clustered. This is expedient in the visualization and unearthing of vital information from various real-time sources that incorporate system organization and application progress based on the Hadoop, stream computing, as well as data warehousing. The induction of data.gov in 2009 by the US government administration was a decisive move towards fostering transparency and accountability within the various government branches. An approximate 400,000 datasets are housed in the warehouse covering diverse aspects ranging from healthcare, transportation, human services to economy (Al Nuaimi et al., 2015).

The various components of smart cities promote the integration and collaboration between differential government agencies such that their actions are streamlined and consistent. The result will be operations that are more efficient and able to handle shared data. The components further improve business decisions through researching behavior and economic growth of a firm. In addition, there will be tremendous benefits to the various government agencies and individual citizens especially as relates to the publishing of new policies. The data quality is developed by government agencies and the citizens as end users will demonstrate the way they can utilize data and transfer it into usable knowledge for better management services. In turn, the governments can concentrate on apprehensions by citizens concerning relevant educational, social, policing, housing, or health care issues (Al Nuaimi et al., 2015).

V. MAJOR CHALLENGES FACING DESIGN, IMPLEMENTATION & OPERATION

Data management and analysis are the biggest challenges for many organizations of all sizes and in all industries. Companies have always struggled to find a practical approach to capture information about their customers and their products and services. In the past and with a few clients and a limited number of services, the process may not be a great deal of difficulty and challenge, but over the time a lot of companies began to grow, making things more complicated. Now we have information related to the consist brand and social media. As we have data about things that are bought and sold over the Internet.

There is a wealth of literature available centered on the challenges facing design, implementation, and operation of real-time big data analytics. Sivarajah et al. (2017) indicate that challenges about Big Data analytics are associated with the characteristics of big data, including the 5 V's as discussed earlier by Jony (2016) with the addition of Variability and Visualization as other important Big Data characteristics. They segregate key challenges in terms of data, process and management. Process challenges include data mining and cleaning, data acquisition and warehousing, data aggregation / integration, analysis and modeling and data interpretation (Sivrajah et al. 2017).

On the other hand, management challenges encompass challenges about privacy, security, governance, sharing, ownership and cost / operational expenditures. Others have identified the challenges of data complexity, computational complexity and system complexity as major challenges about real-time big data analytics design, implementation and

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

operation (Jin et al., 2015). Wang et al. (2016) on the other hand, have identified the involvement of non-technical challenges related more to management problems of users and service suppliers as a key challenge. This is because human expertise plays a critical role in designing, managing, implementing and operating real-time big data analytics to make the most sense out of big data sets efficiently.

However, considering the relatively lesser level of attention given to management challenges as identified earlier, a viable approach would be to conduct further research to develop a greater understanding of the underlying management challenges. Thus, there is no doubt that we need to come up with different solutions, as there are a lot of different data that make up the real assets of our dealings and that caused a lot of complexity when it comes to their administration and management to get out and see the meaning clear as it represents this data. The big data technology is important because it helps organizations to collect and store massive amounts of data and handle in proper speed and real-time (Nader & Al-Jaroodi, 2014).

As shown in Figure 2, there is some untapped value in big data. Different types of data coming from different resources such as satellite imagery, e-commerce and television, sensors, video, and that needs to be a mechanism linking all these data. We need to use analytics, predictive analytics, and deep analysis to come up with answers to help improve the decision-making process, return on investment, understand customer trends, or students and knowing their habits and anticipating their actions in the future. The idea here is that the value of untapped big data brings all this hype and interest in technologies and tools that we can use to get an insight into our data.

The question here is can we manage this data and find a sense of its value? The answer is: Yes. Big data technology has made the world be able to use the difficult part. It is really not a single technology. We should think about big data as a combination of old and new technologies that help companies to gain additional knowledge about the data. The aim would be to come up with viable strategies that could help eliminate or at least mitigate these challenges as much as possible. This can be done by making use of a systematic review approach, whereby industry-wide best practices are sourced and utilized to build an integrated strategy. In particular, challenges about security and privacy are of extreme significance. Hence, the focus would be on devising approaches and strategies that could assist in reconciling security with privacy in Real-time Big data analytics. To be specific, the following strategies would be considered useful; - Privacy-preserving data matching. - Privacy-preserving collaborative data mining. - Biometric authentication with privacy preserving integration. - Use of a Multi-objective Optimization Framework to enhance data privacy.

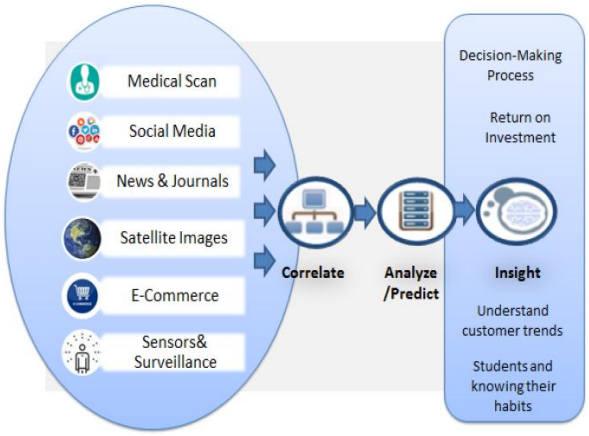

Fig. 2. 5 V's of Big Data

VI. RESULTS

Results of the systematic review suggest the four strategies — identified earlier — to be of prominence in preserving both security and privacy with regards to data management.

A. Privacy-preserving data matching:

Systematic review revealed that privacy-preserving data matching is an effective strategy to help manage security and privacy with regards to big data effectively. Of prominence here is the research by Scannapieco et al (2007) proposes a protocol based on this strategy that can assist in ensuring privacy and security of big data at both data and schema levels (Scannapieco et al). To be specific, their study proposes a protocol based on secure data matching and secure schema matching, which allows the system to hide sensitive data that is not meant to be shared by others, while specifically showing that data that is to be shared (Scannapieco et al., 2007).

UGC Approved Journal

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

B. Privacy-preserving collaborative data mining:

With regards to this specific strategy, the research by Zhan (2008) is of specific significance. Here, the researchers have proposed a secure framework for data mining, bearing in mind the fact that organizational domains often have multiple parties / departments involved who require collaboration yet still have the need to keep their sensitive data private. Based on Secure Number Product Protocol and Secure Two-party Product Protocol, their research revealed that by using a semi-honest algorithm, multiple parties can engage in collaborative data mining while having the ability to keep their sensitive data private and secure from any vulnerabilities or breaches underlying collaboration (Zhan, 2008).

C. Biometric authentication with privacy preserving integration:

Systematic review revealed the use of multi-factor authentication with biometrics as yet another essential strategy towards ensuring security and privacy with regards to management of big data. In particular, the research by Bhargav-Spitzel et al (2007) proposes a two-phase mechanism for authentication purposes considering the federated identity management systems. The first phase made use of tow factors (2F) biometric authentication on the basis of zero knowledge proof of knowledge ZKPK, whereas the second phase utilized biometric authentication in combination with biometric key generation and passwords(Bhargav- Spitzel et al, 2007).

D. Use of a Multi-objective Optimization Framework to enhance data privacy:

With regards to this strategy, the only research study that may be of – limited – relevance was the case-study research conducted by Altiparmak et al (2006) is of certain relevance and value. They considered Supply Chain Network as a case study since it involves big data along with the underlying management challenges of security and privacy. A genetic algorithm is proposed in their research study in combination with a weight approach administered in two phases; however, relevance of the proposed approach by Altiparmak et al (2006) with big data management may only be confirmed upon completion of future studies that apply this approach to big data management with respect to security and privacy.

VII. ANALYSIS & DISCUSSION

From the results of systematic review presented earlier, it can be stated that the first three strategies can be utilized for effective management of big data with regards to security and privacy if implemented in three specific organizational scenarios, i.e. data matching where organizations need to identify sensitive information that must be secured and homogenous data that can be shared, data mining where organizations involve multiple parties to collaborate on big data so that sensitive data is hidden from others, and biometric multi-factor authentication to restrict access to parts of big data to only the concerned organizational members (Sivarajah et al, 2017). On the other hand, the fourth approach appears to be rather confusing, as no evidence is available that speaks of its effectiveness. This implies that organizations need to employ effective security infrastructures that could restrict unwanted / unauthorized access while also preventing access to data by employees who are unrelated to the respective department. Furthermore, organizations need to develop and deploy standardized security laws that are present specifically for the purpose of improving governance, i.e. to ensure that the sensitive data is not used for purposes other than the ones that have been already delineated by the standard operating procedures and protocols.

VIII. CONCLUSION

Big data collection and analysis requires extended periods of time to complete. Consequently, real-time applications cannot immediately benefit from big data analytics. However, many applications should deal with big data and generate results in real-time. This paper investigated the different application domains that could benefit from big data applications operating in real time as - for example, but not limited to - visualization architecture of social networks, health care, predicting terrorist incidents, workplace safety, education, and smart cities (Nader & Al-Jaroodi, 2014).

To build efficient real-time big data applications; several challenges need to be addressed. As indicated earlier, management challenges concerning to big data essentially encompass challenges that are experienced while accessing, governing or managing big data. Since big data implies the presence of huge volumes of highly sensitive data in organizational data warehouses, major challenges associated with big data management include assurance of security and privacy of the data being maintained.

We conclude by devising four approaches and strategies that could assist in reconciling security with privacy in Real-time Big data analytics, Privacy-preserving data matching, Privacy-preserving collaborative data mining, Biometric authentication with privacy preserving integration, and Use of a Multi-objective Optimization Framework to enhance data privacy. These effective approaches will highly improve the privacy and security of real-time big data applications. While first three —out of the four suggested strategies — are applicable in specific organizational settings

IARJSET

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology

ISO 3297:2007 Certified

Vol. 4, Issue 7, July 2017

involving big data management, the fourth strategy needs to be explored in greater detail while applying it to big management so as to identify effectiveness in assuring security and privacy of big data.

REFERENCES

[1] Eiman Al Nuaimi, Hind Al Neyadi, Nader Mohamed, Jameela Al-Jaroodi: "Applications of big data to smart cities", Journal of Internet Services and Applications, pp. 1, 2015.
[2] Fulya Altiparmak, Mitsuo Gen, Lin Lin, Turan Paksoy: "A genetic algorithm approach for multi-objective optimization of supply chain networks", Computers & Industrial Engineering, pp. 196—215, 2006.
[3] J Archenaa, EA Mary Anita: "A survey of big data analytics in healthcare and government", Procedia Computer Science, pp. 408—413, 2015.
[4] Gema Bello-Orgaz, Jason J Jung, David Camacho: "Social big data: Recent achievements and new challenges", Information Fusion, pp. 45—59, 2016.
[5] Abhilasha Bhargav-Spantzel, Anna C Squicciarini, Shimon Modi, Matthew Young, Elisa Bertino, Stephen J Elliott: "Privacy preserving multi-factor authentication with biometrics", Journal of Computer Security, pp. 529—560, 2007.
[6] Pravin Chopade, Kenneth Flurchick, Justin Zhan, Marwan Bikdash: "Visualization Techniques for Large-Scale Big Data Networks: Smart Power Grid Survivability in a Complex Operating Environment", 2015.
[7] Pravin Chopade, Justin Zhan, Kaushik Roy, Kenneth Flurchick: "Real-time large-scale big data networks analytics and visualization architecture", Emerging Technologies for a Smarter World (CEWIT), 2015 12th International Conference & Expo onchopade2015real, pp. 1—6, 2015.
[8] Byron Ellis: Real-time analytics: Techniques to analyze and visualize streaming data. John Wiley & Sons, 2014.
[9] Anika Gupta: "Analysis of Real-Time Big Data: Its Applications and Challenges", Journal of Data Mining and Management, 2016.
[10] HV Jagadish, Johannes Gehrke, Alexandros Labrinidis, Yannis Papakonstantinou, Jignesh M Patel, Raghu Ramakrishnan, Cyrus Shahabi: "Big data and its technical challenges", Communications of the ACM, pp. 86—94, 2014.
[11] Xiaolong Jin, Benjamin W Wah, Xueqi Cheng, Yuanzhuo Wang: "Significance and challenges of big data research", Big Data Research, pp. 59—64, 2015.
[12] Akinul Islam Jony: "Applications of Real-Time Big Data Analytics", , 2016.
[13] Marriam-Webster: "Real Time". 2016. URL https://www.merriam-webster.com/.
[14] Nader Mohamed, Jameela Al-Jaroodi: "Real-time big data analytics: Applications and challenges.", HPCS, pp. 305—310, 2014.
[15] Wullianallur Raghupathi, Viju Raghupathi: "Big data analytics in healthcare: promise and potential", Health Information Science and Systems, pp. 1, 2014.
[16] Filippo Rebecchi, Marcelo Dias De Amorim, Vania Conan, Andrea Passarella, Raffaele Bruno, Marco Conti: "Data offloading techniques in cellular networks: a survey", IEEE Communications Surveys & Tutorials, pp. 580—603, 2015.
[17] Keyun Ruan: "Cybercrime and Cloud Forensics: Applications for Investigation", , 2013.
[18] Philip Russom, others: "Big data analytics", TDWI Best Practices Report, Fourth Quarter, pp. 1—35, 2011.
[19] Arjun Sahni, Divyanshu Marwah, Raman Chadha: "Real time monitoring and analysis of available bandwidth in cellular network-using big data analytics", Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on, pp. 1743—1747, 2015.
[20] Monica Scannapieco, Ilya Figotin, Elisa Bertino, Ahmed K Elmagarmid: "Privacy preserving schema and data matching", Proceedings of the 2007 ACM SIGMOD international conference on Management of data, pp. 653—664, 2007.
[21] Haysam Selim, Pravin Chopade, Justin Zhan: "Statistical Modeling and Scalable, Interactive Visualization of Large Scale Big Data Networks", , 2014.
[22] Uthayasankar Sivarajah, Muhammad Mustafa Kamal, Zahir Irani, Vishanth Weerakkody: "Critical analysis of Big Data challenges and analytical methods", Journal of Business Research, pp. 263—286, 2017.
[23] Aleksandar Stojadinović , Nenad Stojanović , Ljiljana Stojanović : "Dynamic monitoring for improving worker safety at the workplace: use case from a manufacturing shop floor", Proceedings of the 9th ACM International Conference on Distributed Event-Based Systems, pp. 205—216, 2015.
[24] Kenneth David Strang, Zhaohao Sun: "Analyzing relationships in terrorism big data using Hadoop and statistics", Journal of Computer Information Systems, pp. 67—75, 2015.
[25] Van-Dai Ta, Chuan-Ming Liu, Goodwill Wandile Nkabinde: "Big data stream computing in healthcare real-time analytics", Cloud Computing and Big Data Analysis (ICCCBDA), 2016 IEEE International Conference on, pp. 37—42, 2016.
[26] Tech Target: "Real-Time". 2016. URL http://whatis.techtarget.com/.
[27] Ibrahim Toure, Aryya Gangopadhyay: "Real time big data analytics for predicting terrorist incidents", Technologies for Homeland Security (HST), 2016 IEEE Symposium on, pp. 1—6, 2016.
[28] Hai Wang, Zeshui Xu, Hamido Fujita, Shousheng Liu: "Towards felicitous decision making: An overview on challenges and trends of Big Data", Information Sciences, pp. 747—765, 2016.
[29] Ben Williamson: "Digital education governance: data visualization, predictive analytics, and real-time policy instruments", Journal of Education Policy, pp. 123—141, 2016.
[30] Ben Williamson: "Governing methods: policy innovation labs, design and data science in the digital governance of education", Journal of Educational Administration and History, pp. 251—271, 2015.
[31] Justin Zhan: "Privacy-preserving collaborative data mining", IEEE Computational Intelligence Magazine, pp. 31—41, 2008.