Mining Big Data
Challenges and Opportunities

Zaher Al Aghbari
Department of Computer Science, University of Sharjah, Sharjah, U.A.E.

Keywords: Big Data, Mining Big Data, Big Data Challenges, Big Data Research Directions in U.A.E..

Abstract: Nowadays, the daily amount of generated data is measured in exabytes. Such huge data is now referred to as Big Data. Big data mining leads to the discovery of the useful information from huge data repositories. However, this huge amount of data hinders existing data mining tools and thus creates new research challenges that open the door for new research opportunities. In this paper, we provide an overview of the research challenges and opportunities of big data mining. We present the technologies and platforms that are required for mining big data. A number of applications that can benefit from mining big data are also discussed. We discuss the status of big data mining, current efforts and future research directions in the UAE.

1 INTRODUCTION

The amount of data is growing exponentially as it is doubling in size every two years. By 2020, as indicated by Fig. 1, the size of the world's digital data will reach 44 zettabytes (44,000 petabytes), according to a recent IDC study, (IDC, 2014). This dramatic increase in the amount of data is paralleled with an increase in the various data-generating devices, such as sensors, mobile devices and cameras, and data-generating applications such as social media, location-based services and the Internet. For example, in 2014 Facebook warehouse stored 300 petabyte of data and processed data at an incoming daily rate of 0.6 petabytes (Ching, 2013). While Google currently stores about 15 zettabytes of data in its warehouses and processes data at a daily rate of 100 petabytes. The massive data increases is beyond the capability of current technologies to store, analyze and extract useful information from big data (Fan et al., 2012).

Researchers refer to the huge datasets that are unmanageable by current technologies and software tools as Big Data. Big data may come in different forms such as text, images, videos, sounds, or their combinations (Che et al., 2013). In addition to the large size and variety of big data, its incoming rate is often very high. The author of (Beyer, 2012) argued that the main characteristics of big data are the three V's (Volume, Velocity, and Variety). These characteristics pose serious challenges to big data management and mining. The capabilities of current DBMSs can no longer handle the increasing demands of big data satisfactorily (Madden, 2012).

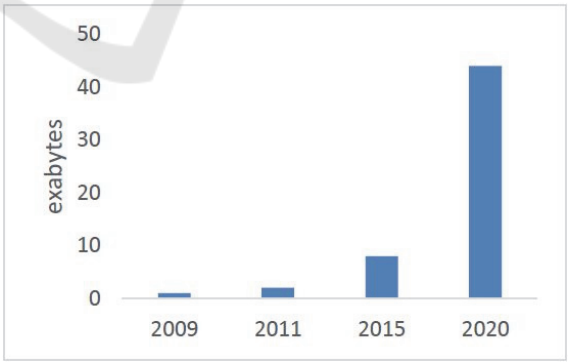

Figure 1: Expected big data growth.

To discover useful knowledge from big data, research should focus on developing data-intensive distributed storage systems, parallel processing architectures and efficient mining techniques. Recent efforts in this direction, Google's programming model, called MapReduce, for processing big data using parallel and distributed algorithms (Dean et at., 2014) and Yahoo's distributed file system, called Hadoop Distributed File System, that was created as an open source system (Apache, 2009). Some of the applications that could benefit from big data and related technologies are healthcare, finance, retail business and transportation.

In this paper, we review the big data concepts,

Al Aghbari Z..
Mining Big Data - Challenges and Opportunities.
DOI: 10.5220/0005463803790384
In *Proceedings of the 17th International Conference on Enterprise Information Systems* (ICEIS-2015), pages 379-384
ISBN: 978-989-758-096-3

mining and required technologies. We also discuss the characteristics of big data, its issues and challenges, benefits to the industry and community at large. We also explore some future research directions. The main applications in the UAE that could benefit from big data are also presented. The aim of our effort is to shed light on some of the benefits and challenges of big data research in the UAE and help researchers in data mining to explore new research directions to solve emerging challenges brought about by big data.

The rest of this paper is organized as follows. Section 2 presents the characteristics of big data, emergence of big data, and the motivation to analyze big data. In Section 3, we discuss issues related mining big data. Section 4 presents the hardware and software technologies required for collecting, storing, and analyzing big data. We discuss the applications of big data in the UAE in Section 5. Finally, the future research opportunities in the UAE are presented in Section 6.

2 BIG DATA EVOLUTION

The introduction of the Internet and World Wide Web in 1990s have increase the number of web applications and web services, which in turn increased the amount of generated data tremendously. Companies and governmental agencies are now dealing with large databases whose sizes are in petabytes and in the near future the sizes of these databases will reach exabytes.

2.1 Characteristics of Big Data

The size of big data is beyond the capability of traditional database algorithms to store, manage and analyze it. However, it is not just the volume of data but also the variety and velocity of the data. These three attributes form the three Vs of Big Data (Beyer, 2012). Organizations are now storing massive amount of data in the form of textual, numerical, or multimedia data. The size of such data in each individual organization will soon reach exabytes. That is big data may come from different sources such as sensors, mobile devices, social networks, etc. Moreover, big data may consist of various types and different levels of complexities. Big data could be structured such as relational databases, unstructured such as multimedia data, and/or semi-structures such as xml and social media data. These different type of data often generated and shared at high velocities. Therefore, the three Vs have made it difficult for current technologies to handle big data.

2.2 Emergence of Big Data

The wide spread of Web 2.0 with various popular applications including the various forums, newsgroups and social media contributed to the increase of content of data. Advances in digital sensors, mobile devices, communications, computers, and storage devices have contributed to the massive growth of data. Some organizations are trying to extract valuable information from these massive collections of data. Google and other search engines have create profitable businesses by collecting the information posted on the Internet and presenting it to people in a useful way (Bryant et al., 2008). The use of big data by search engines have transformed the way people access information. Other big organizations and companies are now collecting huge amounts of their daily transactions and are trying (or hoping) to process it and extract valuable information to give them a competitive edge in the market.

2.3 Motivation of Big Data

As the size of collected and stored data is increasing tremendously, the chances of extracting useful information to gain business advantage is also increasing. Furthermore, the cost of new technologies to store this huge amount of data has fallen dramatically allowing even average-sized companies to collect and manage big data. The peer pressure of competitor organizations forces individual organizations to collect and process big data to extract valuable information to allow them to remain competitive in the market. Decision making will learn from big data to find patterns, relationships between data, correlations between different types of data, groupings of data, etc.

3 MINING BIG DATA

Traditional data mining discovers useful information, interesting groupings, valuable patterns and relationships hidden in the data. The discovered results help make valuable predictions and help formulate decisions in the real world. Various applications have benefited from data mining such as medicine, business, and science. Recently, data mining algorithms have been facing many challenges when applied to huge amount of data due to the limitations of these algorithms in handling the characteristics of big data. Despite of these

limitations, big data brings new opportunities for extracting valuable insights from the complex and heterogeneous contents.

It is believed that big data will play a critical role in the future and affect the way businesses and services are conducted. For example, governments may make use of big data in the form of social media and other sources of online information to gauge public satisfaction about governmental services, identify the need for new government facilities, detect suspicious criminal groups, or predict future threats. However, the capabilities of existing database technologies cannot extend to the requirements of managing big data.

To meet the requirements of big data, Google introduced MapReduce (Dean et al., 2014), which is a new programming model. This programming model requires a distributed platform; therefore, a distributed file system, called GFS (Google File System), was used by Google to divide the huge the datasets among thousands of computers that are referred to as a cluster (Madden, 2012). On another front, Yahoo created Hadoop MapReduce, which uses the Hadoop Distributed File System (HDFS). Both of Yahoo's technologies are open source version of the Google's technologies.

3.1 MapReduce

With MapReduce, the input data is partitioned into large sets of key-value pairs. These sets are then processed by map() functions in parallel. Each map() function processes the local data, and writes the output to a temporary storage (Ghemawa et al, 2003). This mapping task is then followed by applying a reduce() function to the result of the map() functions. The reduce() functions merges each group of output data based on common keys, in parallel, to obtain the final result .

3.2 Hadoop

Hadoop is a distributed and scalable file-system written in Java (Apache, 2009). It is an open-source framework for distributed storage and distributed processing of Big Data on clusters of computers. Hadoop stores huge amounts of data and processes them efficiently via distributed processing. Hadoop also replicates the sets of data on several computer nodes to achieve reliability.

4 REQUIREMENTS OF MINING BIG DATA

Extracting useful information from the huge volume of interconnected heterogeneous big data has the potential to maximize our knowledge and insights on a particular domain and thus, leads to better decision making. To gain these insights, a new breed of hardware and software technologies are required to cope up with the challenges of big data.

4.1 Hardware Requirements

The foundation of the Big Data technology is the infrastructure.

Computer Networks: efficient computer networks are essential to collect massive amounts of data incoming from many different sources, such as sensors, cameras and the Internet, at high rates (Matti, 2012).

Storage Systems: Storage systems have flexible designs to enable the scaling of system computation and capacity. A storage system should store data in such a way to allow quick access to the data when required. In addition to scalability of the storage systems, they should be fast, support reliability and availability of the data (Woods, 2012).

Cluster Computing Systems: Cluster computing incorporates a software architecture implemented on thousands of commodity computers to provide high-performance processing for big data applications. These clusters are composed of large numbers of cheap commodity computing devices that support scalability, reliability, and programmability achieved by new software paradigms (Bryant et al., 2008). Typically, these clusters provide both the storage capacity and the data-intensive computing power.

Cloud Computing Systems: Cloud computing uses virtualization to run several standardized virtual servers on the same physical machine. Cloud computing created a new billing-based business model, where organizations and individuals can rent storage and computing resources from the cloud service providers instead of making large investments on buying and maintaining their own IT facilities. Cloud computing facilitates the following services:

- IaaS: Infrastructure-as-a-Service for storage and back up.
- SaaS: Software-as-a-Service for software tools, data analytics, etc.
- PaaS: Platform-as-a-Service for providing tools and libraries to build, test, and run applications on

cloud infrastructure. Typically, it provides hardware and software solutions.

4.2 Software Requirements

The rise of big data has demanded a set of new algorithms and tools.

Distributed Computing Programs: processing huge amounts of data in a reasonable amount of time requires distributed computing models. To achieve this goal, MapReduce was created as a distributed programming model that allows programs to partition big data, process the partitions, and merge their results in parallel. Furthermore, distributed algorithm should support scalability of the data since the size of data is exponentially increasing annually.

Cloud Computing Tools: cloud service providers utilize software tools for data analytics to help customers analyze the big data stored in the cloud using the Software-as-a-Service business model. The output of these data analytics can be shown to the customers through a graphical interface. Data analytics allow organizations to build and diagnose models, and interpret analytic insights for better decision making.

Security and Privacy Tools: Big data, typically, consists of sensitive data and thus tools to prevent unauthorized access to such data are of great importance. For example, cloud computing gave rise to a new business mode called database outsourcing, in which an organization's database is stored at the cloud service provider. However, the organization would not want the cloud service provider to access, uncover, or understand the stored data. At the same time, the organization's authorized users should be allowed to access the data freely 0Yui et al. 2010). Such a model raises privacy concerns that require stringent security tools to prevent unauthorized access.

Interactive Tools: To help gain insight and extract valuable information from big data, interactive tools are required to help users interact with the results and receive feedback. Such interactions allow users to better interpret the mining results and discover accurate knowledge.

Visualization Tools: Data visualization helps users make better sense of the analysis delivered by machine learning tools. Visualization tools converts big data of the organization to an easier visual format to understand and reduces significantly the time to make decisions based on that data.

5 APPLICATIONS OF MINING BIG DATA TO UAE

UAE is now a hub of new technologies in the Middle East and its cities are striving to become "smart cities" in which online technologies are integrated into everyday life. Many governmental and business organizations across the UAE are capturing massive amounts of data daily. Therefore, several applications can benefit from big data and its related technologies.

5.1 Healthcare

Captured healthcare data is increasing in volume; however, with the existing technologies, healthcare organizations cannot make full use of such huge volumes of data to improve the quality of human health. Most of healthcare big data is unstructured, complex, and heterogeneous, such as physician notes, X-rays, MRIs, lab analysis data, medical correspondence, claims, etc. By leveraging these datasets, healthcare organizations can identifying the right treatment for individual patients (Bizer, 2011). Another benefit of big data comes from capturing and analyzing big data in real-time from medical sources can alert hospitals of potential infection diseases before patients show signs of these diseases (Chawla, 2013).

In the UAE, healthcare organizations are adopting data analytics solutions to harness their big datasets to improve patient care and make better health-related decisions. Another example is Smart Health, which is aligned with H.H. Sheikh Mohammed Bin Rashid initiative for SMART GOVERNMENT and SMART HEALTH. Using big data analytics in real time, health organizations can identify which patients are more at risk of life threatening complications.

5.2 Education

Students are expected to gain deep understanding of the subjects they learn and at the same time teachers are expected to provide personalized teaching and mentoring. Therefore, teachers should have access to complete profiles of their students. However, manual analysis of such these students' profiles is a daunting task. The opportunity lies in Big Data analytics to harness the huge number of students' profiles to provide flexible, adaptive and personal learning. For example, in the UAE big data can allow teachers to identify patterns of student behaviors during the semester. That will help teachers identify the level of engagement and understanding of their students and

accordingly give them appropriate directions to improve their learning experience.

5.3 Retail Business

By understanding the buying behavior of the consumers, retailers can make targeted advertising to reduce their marketing cost (Al-Khouri, 2014). Therefore, retailers can exploit Big Data to help make decisions about their discounts, marketing campaigns and supply chain management. Top retailers in the UAE can capture and analyze massive heterogeneous datasets of their customer's buying transactions, buying history, demographics and preferences, to provide product recommendations relevant to each of its customers and increase the average purchase amount. Moreover, retailers can also use the data and consider future promotion events and identify causes that may drive sales (Jham, 2012).

5.4 Transport

Big Data analytics can offer the UAE transport authorities a valuable insight on customer demands for new services and effective routing of traffic especially in big cities of the UAE like Dubai. Big data analytics could help the urban planners in the UAE cities to predict and timely construct new roads in certain areas of the city to prevent future traffic congestions. By analyzing a massive history of data collected from sensors placed on roads of the city, UAE transport authorities can provide travelers in Dubai, for example, the opportunity to select a travel option during rush hours to save time and cost. Additionally, one of the public services could be to compare the different means of travel from a certain position to a destination point at any day, time, weather, season, etc. in terms of travel period and cost (Brown et al., 2011; Jagadish, 2014).

During major festivals, such EXPO 2020, a large number of visitors are expected to come into Dubai and thus road traffic is expected to be a concern. Analyzing history of huge traffic datasets collected from previous big events will help the transport authorities provide services to re-route traffic to prevent road congestions, speed up the travel time and predicting the best time to conduct maintenance on each of its vehicles to prevent disruption of transport services.

5.5 Finance

In Finance, big data analytics can help in detecting fraud. By analyzing financial transactions in real time from different sources, financial companies will have the ability to identify and predict opportunities and threats (LaValle et al., 2011). For example, fraud events should be detect such as purchasing expensive jewelry worth of thousands of Euros in Germany minutes after paying for two tickets to enter a movie theatre in the UAE. To make use of big data, one financial services firm in the USA uses data from 17 million customers and 19 million daily transactions as an early warning system to help retain its customers (SAS, 2012).

5.6 Smart Cities

Innovation, creativity and entrepreneurship drive the economy and governance of smart cities (Kitchin, 2014). Typically, sensors, cameras, CCTV, fingerprints, iris scans and other digital data collectors are placed in every corner of a smart city to monitor and manage in real-time every day events of the city. Gathered data in smart cities would also contain information about activities of the people living in the city streamed from their smart devices, e.g. phones, tablets, laptops, etc. (Kitchin, 2013). This information include people's locations, time, visited websites, transactions, etc. Big data analytics can be used to process collected city data and extract knowledge that can be used in directing traffic to avoid congestions, adjust road speed limits, dispatching ambulances, post weather warnings, etc.

6 POTENTIAL BIG DATA RESEARCH IN UAE

The following research areas are of potential benefit to the UAE. The data mining research team at University of Sharjah may collaborate with other research teams in the UAE, other countries, and/or the industry to delve into the following possible areas of big data analytics:

- Capturing and analyzing big data in real-time to predict potential infection diseases before patients show signs of these diseases.
- Capturing and analyzing big data in real-time to predict potential infection diseases before patients show signs of these diseases.
- Using big data analytics in real time to identify which patients are more at risk of life threatening complications, such as cancer.
- Analysis of healthcare big data to identify the level of engagement and understanding of the students during the course to improve their learning experience.

- Analyze big data of retail customer to provide product recommendations relevant to each customer and increase the average purchase amount.
- Detection of fraudulent customers and merchants by mining credit card transaction data.
- Analyze big data to allow UAE transport authorities provide travelers with new services, such as travel options during rush hours to save time and cost, compare different transportation means at any time, day, weather condition, etc.
- Leveraging of data mining to conduct preventive maintenance to transportation vehicles.
- Analyzing big social networks to identify influential people, criminal groups, and public sentiments.
- Analysis of financial datasets to detect customer disengagement to help retain customers.
- Analysis of big datasets of smart city to help re-direct traffic to avoid congestions, adjust road speed limits, post weather warnings, etc.

REFERENCES

IDC, 2014, http://www.emc.com/leadership/digital-universe/2014iview/executive-summary.htm

Ching, A., 2013. Scaling Apache Giraph to a trillion edges. https://www.facebook.com/notes/facebook-engineering/scaling-apache-giraph-to-a-trillion-edges/10151617006153920

Fan, W., Bifet, A., 2012. Mining Big Data: Current Status, and Forecast to the Future. *ACM SIGKDD Exploration Newsletter, vol. 14, no. 2*.

Che, D., Safran, M., Peng, Z., 2013. From Big Data to Big Data Mining: Challenges, Issues, and Opportunities. *DASFAA Workshops, LNCS 7827, pp. 1–15*.

Beyer, M. A., Laney, D., 2012. The Importance of Big Data: A Definition. *Gartner*.

Madden, S., 2012. From Databases to Big Data. *IEEE Internet Computing, vol.16, no.3, pp. 4–6*.

Gama, J., 2010. *Knowledge discovery from data streams*. Chapman & Hall/CRC.

Dean, J., Ghemawat, S., 2014. MapReduce: simplified data processing on large clusters. *In 6th Symposium on Operating System Design and Implementation, pp. 137–150*.

The Apache Hadoop Project, 2009. .http://hadoop.apache.org/core/,

Bryant, R. E., Katz, R. H., Lazowska, E. D., 2008. Big-Data Computing: Creating revolutionary breakthroughs in commerce, science, and society. *pp. 1-15, at* http://www.cra.org/ccc/files/docs/init/Big_Data.pdf.

Ghemawat, S., Gobioff, H., Leung, S.T., 2003. The Google File System. *In 19th ACM Symposium on Operating Systems Principles*, pp. 29–43.

Woods, D., 2012. Ten Properties of the Perfect Big Data Storage Architecture. http://www.forbes.com/sites/danwoods/2012/07/23/ten-properties-of-the-perfect-big-data-storage-architecture/

Matti, M., Kvernvik, T., 2012. Applying big-data technologies to network architecture. *Ericsson Review*.

Yang, Y., Papadopoulos, S., Papadias, D., Kollois, G., 2009. Authenticated indexing for outsourced spatial databases. *VLDB Journal, vol. 18, pp. 631-648.*

Yiu, M. L., Ghinita, G., Jensen, C. S., Kalnis, P., 2010. Enabling search services on outsourced private spatial data. *VLDB Journal, Vol. 19, no. 3, pp. 363-384.*

Chawla, N. V., 2013. Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework. *Journal of General Internal Medicine, vol. 28, no. s3.*

Bizer, C., Boncz, P., Brodie, M. L., Erling, O., 2011. The Meaningful Use of Big Data: Four Perspectives – Four Challenges. *SIGMOD Record, vol. 40, no. 4*.

Al-Khouri, A., 2014. Identity Management in the Retail Industry: The Ladder to Move to the Next Level in the Internet Economy. *Journal of Finance and Investment Analysis, vol. 3, no.1, PP. 51-67*.

Jham, V., 2012. Change management in retail banking in the UAE: an assessment of some key antecedents of customer satisfaction and demographics. *vol. 4, no. 3.*

Brown, B., Chui, M., Manyika, J., 2011. Are you ready for the era of big data? *McKinsey Quarterly*.

Jagadish, H. V., 2014. Big Data and Its Technical Challenges. *Communications of the ACM, vol. 57, no. 7, pp. 86-94.*

LaValle, S., Lesser, E., Shockley, R., Hopkins, M. S., Krushcwits, N., 2011. Big Data, Analytics and the Path from Insights to Value. *MITSloan Management Review, Winter ed.*

SAS. 2012. Banking on Analytics: How High-Performance Analytics Tackle Big Data Challenges in Banking. *July ed.* http://www.sas.com/resources/whitepaper/wp_42594.pdf.

Kitchin, R., 2014. The real-time city? Big data and smart urbanism. *GeoJournal, vol. 79, pp. 1–14*.

Kitchin, R., 2013. Big data and human geography: Opportunities, challenges and risks. *Dialogues in Human Geography.*