STUDY ON APPLICATIONS OF BIG DATA ANALYTICS

Dr. G.BABU
Professor & Head of Department, Department of Computer Applications
Adhiparasakthi Engineering College, Melmaruvathur,Tamilnadu India..

Abstract
Data is being continuously generated from various sources which in turn can be collected to carry out useful analysis and derive a feasible output. Big Data Analytics has become a hot topic in academics, industry and everywhere. Big Data Analytics is the process of examining large amounts of data (big data) to discover hidden patterns or unknown correlations. Data mining techniques and extracting patterns from large datasets play a vital role in knowledge discovery. Most of the decision makers encounter a large number of decision rules resulted from association rules mining. Moreover, the volume of datasets brings a new challenge to extract patterns such as the cost of computing and inefficiency to achieve the relevant rules. Big data encompasses large size of data which has the traits of being complex and are amalgamated from an array of autonomous sources.

Keywords: Analytics, Big Data, Data Security, Hadoop, Threats, Methodologies, Applications.

I. INTRODUCTION

Big data refers to data sets that are not only big, but also high in variety and velocity, which makes them difficult to handle using traditional tools and techniques. Big data sizes are constantly increasing, currently ranging from a few dozen terabytes (TB) to many petabytes (PB) of data in a single data set. With the evolution of technology and the increased multitudes of data flowing in and out of organizations daily, there has become a need for faster and more efficient ways of analyzing such data [15]. Big Data Analytics (BDA) can also enable the construction of predictive models for customer behavior and purchase patterns, therefore raising overall profitability [13]. Big data analytics is a technology-enabled strategy for gaining richer, deeper, and more accurate insights into customers, partners and the business and ultimately gaining competitive advantage. By processing a steady stream of real-time data, organizations can make time-sensitive decisions faster than ever before, monitor emerging trends, course-correct rapidly and jump on new business opportunities. Big data can create transparency, and make relevant data more easily accessible to stakeholders in a timely manner. Big Data Analytics holds much potential for customer intelligence, and can highly benefit industries such as retail, banking, and telecommunications [13]. In Big Data Analytics, Data Security is a challenging task to implement and needs strong support in terms of security policy formulation, techniques, tools and mechanisms. Hadoop is one of the tools and is a highly scalable storage platform, because it can store and distribute very large data sets across hundreds of inexpensive servers that operate in parallel [8,13].

II. APPLICATIONS OF BIG DATA

Big data applications solve and analyze real world problems using Hadoop and associated tools. Internet users and machine-to-machine connections are causing the data growth [6]. Real time areas are defined following in which big data is used:

1. **Big data in healthcare**: High-performance analytics are new technologies making easier to turn massive amounts of data into relevant and critical insights used to provide better care. Analytics helps to predict disease history and its trends. Unstructured data can be captured through text mining from patient records. It means information can be collected without causing additional work for clinicians. A

massive amount of data collected from different sources provides the best practices for today, and will help healthcare providers identify trends so they can achieve better results to improve medical facilities all around the world.

2. Network Security: Big data is changing the landscape of security technologies. The tremendous role of big data can be seen in network monitoring. Big data analytics is an effective solution for processing of large scale information as security is major concern in enterprises. Fraud detection is done by using big data analytics. Phone and credit card companies have conducted large-scale fraud detection for decades. Mainly big data tools are particularly suited to become fundamental for forensics.

3. Market and business Big Data is the biggest game-changing opportunity for sales and marketing, since 20 years ago the Internet went main stream, because of the unprecedented array of insights into customer needs and behaviours. Big data reveals customers' behaviour and proven ways to elevate customer experiences. These insights ensure your business's success.

4. Sports Sport, in business, an increasing volume of information is being collected and captured. Technological advances will fuel exponential growth in this area for the foreseeable future, as athletes are continuously monitored .Statistics can be analyzed and collected to better understand what are the critical factors for optimum performance and success, in all facets of elite sport. Injury prevention, competition, Preparation, and rehabilitation can all benefit by applying this approach. Used consistently this is a powerful measure of progress and performance.

5. **Education Systems** By using big data analytics in field of education systems, remarkable results can be seen. Data on students online behaviour can provide educators with important insights, such as if the course has to be modified or not based on students reception. This modification can be done by making students answer set of online questionnaire and track the accuracy and time taken to answer those questions.

III. DIMENSIONS OF BIG DATA

Big data is characterised by the following five dimension .The 5Vs that define Big Data are Variety, Velocity and Volume, Variability and Veracity

1) Volume: There has been an exponential growth in the volume of data that is being dealt with. Data is not just in the form of text data, but also in the form of videos, music and large image files. Data is now stored in terms of Terabytes and even Petabytes in different enterprises. With the growth of the database, we need to re-evaluate the architecture and applications built to handle the data.

2) Velocity: Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.

3) Variety: Today, data comes in all types of formats. Structured, Numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. We need to find ways of governing, merging and managing these diverse forms of data.

4) Variability: Variability. In addition to the increasing velocities and varieties of data, data flows can be highly inconsistent with periodic peaks. Daily, seasonal and event-triggered peak data loads can be challenging to manage. Even more so with unstructured data involved [2]

5) Veracity: Today's data comes from multiple sources. And it is still an undertaking to link, match, cleanse and transform data across systems. However, it is necessary to connect and correlate relationships, hierarchies and multiple data linkages or your data can quickly spiral out of control. A data environment can lie along the extremes on any one of the following parameters, or a combination of them, or even all of them together.

IV. TYPES OF BIG DATA:

There are two types of big data. Structured Data and Unstructured Data.

1. **Structured Data**: Structured Data are numbers and words that can be easily categorized and analyzed. These data are generated by things like network sensors embedded in electronic devices, smart phones, and global positioning system (GPS) devices. Structured data also include things like sales figures, account balances, and transaction data.

2. Unstructured Data: Unstructured Data include more complex information, such as customer reviews from commercial websites, photos and other multimedia, and comments on social networking sites. These data cannot easily be separated into categories or analyzed numerically. The explosive growth of the Internet in recent years means that the variety and amount of big data continue to grow. Much of that growth comes from unstructured data. When making an attempt to understand the concept of Big Data, the words such as ―maps reduce‖ and ―Hadoop‖ cannot be avoided.

Hadoop is a free, Java-based programming frame work, supports the processing of large sets of data in a distributed computing environment. It is a part of the Apache project sponsored by the Apache Software Foundation. Hadoop cluster uses a Master/Slave structure [6]. Using Hadoop, large data sets can be processed across a cluster of servers and applications can be run on systems with thousands of nodes involving thousands of terabytes. Distributed file system in Hadoop helps in rapid data transfer rates and allows the system to continue its normal operation even in the case of some node failures. This approach lowers the risk of an entire system failure, even in the case of a significant number of node failures. Hadoop enables a computing solution that is scalable, cost effective, fault tolerant and flexible. Hadoop Framework is used by popular companies like Google, Yahoo, Amazon and IBM etc., to support their applications involving huge amounts of data.

- **Hadoop has two main sub projects**

Map Reduce & Hadoop Distributed File System (HDFS).

Map Reduce Hadoop Map Reduce is a framework [10] used to write applications that process large amounts of data in parallel on clusters of commodity hardware resources in a reliable, fault-tolerant manner. A Map Reduce job first divides the data into individual chunks which are processed by Map jobs in parallel. The outputs of the maps sorted by the framework are then input to the reduce tasks. Generally the input and the output of the job are both stored in a file - system. Scheduling, Monitoring and re-executing failed tasks are taken care by the framework.

Hadoop Distributed File System (HDFS) HDFS [8] is a file system that spans all the nodes in a Hadoop cluster for data storage. It links together file systems on local nodes to make it into one large file system. HDFS improves reliability by replicating data across multiple sources .

V. STAGES INVOLVED IN BIG DATA

1. Data Acquisition: The first step in Big Data is acquiring the data itself. With the growing medium the rate of data generation is rising exponentially. With the introduction of smart devices which are used with a wide array of sensors continuously generate data. The Large Haudron Collider in Switzerland produces petabytes of data. Most of this data is not useful and can be discarded, however due to its unstructured form; selectively discarding the data presents a challenge. This data becomes more potent in nature when it‘s merged with other valuable data and superimposed. Due to the interconnectedness of devices over the World Wide Web, data is increasingly being collated and stored in the cloud.

2. **Data Extraction:** All of the data generated and acquired is not of use. It contains a large amount of redundant or unimportant data. For instance, a simple CCTV camera, constantly polls sensor to gather information of the user‘s movements. However, when the user is in a state of inactivity, the data generated by the activity sensor is redundant and of no use.. Due to wide variety of data that exists, bringing them under a common platform to standardize data extraction is a major challenge.

3. Data Collation: Data from a singular source often is not enough for analysis or prediction. More than one data sources are often combined to give a bigger picture to analyze. For example a health monitor application often collects data from the heart-rate sensor, pedometer, etc. to summarize the health information of the user.Likewise, weather prediction software take in data from many sources which reveal the daily humidity, temperature, precipitation, etc. In the scheme of Big Data convergence of data to form a bigger picture is often considered a very important part of processing.

4. Data Structuring: Once all the data is aggregated, it is very important to present and store data for further use in a structured format. The structuring is important so queries can be made on the data. Data structuring employs methods of organizing the data in a particular schema.

5. Data Visualization: Once the data is structured, queries are made on the data and the data is presented in a visual format. Data Analysis involves targeting areas of interest and providing results based on the data that has been structured. **6. Data Interpretation**: The ultimate step in Big Data processing includes interpretation and gaining valuable information from the data that is processed.

VI. BIG DATA SECURITY

Big data can create transparency, and make relevant data more easily accessible to stakeholders in a timely manner. Big Data Analytics holds much potential for customer intelligence, and can highly benefit industries such as retail, banking, and telecommunications . In Big Data Analytics, Data Security is a challenging task to implement and needs strong support in terms of security policy formulation, techniques, tools and mechanisms.[12]

1 **Secure Computations in Distributed Programming Framework**: Distributed programming framework utilize parallelism in computations and storage to process massive amounts of the data .A popular example is map reduce framework, which splits an input file into multiple chunks in the first phase of map reduce, a mapper for each chunk reads the data, perform some computation, and outputs a list of key/value pairs.In the next phase, a reducer combines the values belonging to each distinct key and outputs the result. There are two major attack prevention measures: securing the mappers and securing the data in the presence of an untrusted mapper.

2. Security Best Practices for Non-Relational Data Stores: Non-relational data stores popularized by NoSQL databases are still evolving with respect to security infrastructure..NoSQL databases do not provide any Support for Enforcing it explicitly in the database. However, clustering aspect of NoSQL databases poses additional challenges to the robustness of such security practices.

3. Secure Data Storage and Transaction Logs: Data and transaction logs are stored in multi-tiered storage media manually moving data between tiers gives the it manager direct control over exactly what data is moved and when. However as the size of data set has been and continues to be, growing exponentially, scalability and availability have necessitated auto-tiring for big data storage management. Auto-tiering solutions do not keep track of where the data is stored, which poses new challenges to secure data storage.

4. End Point Input Validation/Filtering: Many big data use cases in Enterprise settings require data collection from many sources, such as end point devices for example, a security information and event management system (SIEM) may collect event logs from millions of hardware devices and software application in an enterprise network.

5. Real –Time Security/Compliance Monitoring: Real time security monitoring has always been a challenge, given the number of alerts generated by (Security) devices.These alerts (correlated or not) lead to many false positive, which are mostly ignored or simply‖ clicked away‖, as humans cannot cope with the shear amount. This problem might even increase with the bid data given the volume and velocity of data streams however, big data technologies might also provide an opportunity, in the sense that these technologies do allow for fast processing and analytics of different types of data.

6. Granular Access Control: The security Property that matters from the perspective of access control is secrecy-preventing access to data by people that should not have access .The problem with course-grained access mechanisms is that data that could otherwise be shared is often swept into a more restrictive category to guarantee sound security granular access control gives data managers a scalpel instead of a sword to share data as much as possible without compromising secrecy.

7. Granular Audits: With real time security monitoring, we try to be notified at the moment an attack takes place. In reality, this will not always be the case (e.g., new attacks, missed true positives). In order to get to the bottom of the missed attack, we need audit information. This is not only relevant because we want to understand what happened and what went wrong ,but also because compliance, regulation and forensics reasons .in that regard ,auditing is not something new, but the scope and granularity might be different. For example, we have to deal with more data objects, which probably are (but not necessarily) distributed.

VII. CONCLUSION

As an organization collects more data at this scale, formalizing the process of big data analysis will become paramount. As the data is becoming bigger and bigger, there is a need to store this data in an efficient manner. The paper is a systematic study of various security issues and challenges of Big Data analytics. Big Data is a very challenging research area. Through better analysis of the large volumes of data that are becoming available, there is the potential for making faster advances in many scientific disciplines and improving the profitability and success of many enterprises. However, many technical challenges described in this paper must be addressed before this potential can be realized fully. Furthermore, these challenges will require transformative solutions, and will not be addressed naturally by the next generation of industrial products

REFERENCES

[1] Big Data: science in the petabyte era‖, Nature 455 (7209):1, 2008

[2] Douglas and Laney, ―The importance of ‗Big Data: A definition‖ ,2008

[3] Agrawal, Amr El Abbadi et al.,‖Big data and cloud computing: current state and future opportunities‖, Proceedings of the 14th International Conference on Extending Database Technology, ACM, Sweden, March 21-24, 2011

[4] Bharti Thakur, Manish Mann, Volume 4, Issue 5, May 2014, “Data Mining For Big Data”, International Journal of Advanced Research in Computer Science and Software Engineering: 469-473.

[5] Manish Kumar Kakhani, Sweeti Kakhani and S.R. Biradar, Volume 2 Issue 8, August 2013. “Research Issues in Big Data Analytics”, International Journal of Application or Innovation in Engineering and Management: 228-232

[6] Lu, Huang, Ting-tin Hu, and Hai-shan Chen. "Research on Hadoop Cloud Computing Model and its Applications‖, Hangzhou, China: 2012,pp. 59 – 63, 21-24 Oct. 2012.

[7] Wie, Jiang, Ravi V.T and Agrawal G., "A Map-Reduce System with an Alternate API for Multi-core Environments‖, Melbourne, VIC: 2010,pp. 84-93, 17-20 May. 2010.International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, May 2014

[8] K, Chitharanjan, and Kala Karun A. "A review on hadoop — HDFS infrastructure extensions‖, JeJu Island: 2013, pp. 132-137, 11-12 Apr.2013.

[9] F.C.P, Muhtaroglu, Demir S, Obali M, and Girgin C. "Business model canvas perspective on big data applications", IEEE International Conference, Silicon Valley, CA, pp. 32 – 37, Oct 6-9, 2013.

[10] Zhao, Yaxiong , and Jie Wu. "Dache: A data aware caching for big-data applications using the Map Reduce framework." INFOCOM, 2013 Proceedings IEEE, Turin, pp. 35 – 39, Apr 14-19, 2013.

[11] Xu-bin, LI , JIANG Wen-rui, JIANG Yi, ZOU Quan "Hadoop Applications in Bioinformatics." Open Cirrus Summit (OCS), 2012 Seventh,Beijing, pp. 48 – 52, Jun 19-20, 2012.

[12]Venkata Narasimha Inukollu , Sailaja Arsi1 and Srinivasa Rao Ravuri, ―SECURITY ISSUES ASSOCIATED WITH BIG DATA IN CLOUDCOMPUTING‖, International Journal of Network Security & Its Applications (IJNSA), Vol.6, No.3, pp. 45-56, May 2014.

[13]Sachchidanand Singh, Nirmala Singh,‖Big Data Analytics‖, International Conference on Communication, Information & Computing Technology (ICCICT), Oct. 19-20, Mumbai, India,2012.

[14] Changqing Ji, Yu Li, Wenming Qiu, Uchechukwu Awada, Keqiu Li,‖Big Data Processing in Cloud Computing Environments‖, 12th International Symposium on Pervasive Systems, Algorithms and Networks (ISPAN), IEEE, pp. 17–23, 2012.

[15] Dan Garlasu, Virginia Sandulescu, Ionela Halcu,Giorgian Neculoiu, Oana Grigoriu, Mariana Marinescu and Viorel Marinescu, ―A big data implementation based on Grid computing‖, 11th RoEduNet international Conference, pp. 1-4, 2013

[16] Xiaoxue Zhang, Feng Xu, "Survey of Research on Big Data Storage",12th International Symposium on Distributed Computing and Applications to Business, Engineering & Science (DCABES), IEEE, pp.76-80, SEPT.2013.

[17]Kapil Bakshi, ―Considerations for Big Data: Architecture and Approach‖, IEEE Aerospace conference, pp. 1-7, March 2012.

[18] C. Byun, W. Arcand, D. Bestor, B. Bergeron, M. Hubbell, J. Kepner, A. McCabe, P. Michaleas, J. Mullen, D. O'Gwynn, A. Prout, A. Reuther,A. Rosa & C. Yee, ―Driving Big Data With Big Compute‖, IEEE High Performance Extreme Computing (HPEC), Sep 10-12, 2012

[19]Top Ten Big Data Security And Privacy Challenges‖, CLOUD SECURITY ALLIANCE, NOV. 2012, https://cloudsecurityalliance.org/

[20] Tyson Condie , Paul Mineiro , Neoklis Polyzotis , Markus Weimer,‖Machine Learning on Big Data‖, 29TH IEEE INTERNATIONAL CONFERENCE ON DATA ENGINEERING (ICDE Conference), pp. 1242-1244, 2013

[21] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding,‖Data mining with Big Data‖, IEEE Transactions on Knowledge & Data Engineering, vol.26, Issue No.01, Jan 2014.

[22] Dr. Sun-Yuan Kung,‖From Green Computing to Big-Data Learning: A Kernel Learning Perspective‖IEEE 24th International Conference on Application-Specific Systems, Architectures and Processors (ASAP), USA, JUNE 2013

[23] V. Mayer-Schonberger and K. Cukier, ¨ Big Data: A Revolution that Will Transform how We Live, Work, and Think. Eamon Dolan/Houghton Mifflin Harcourt, 2013.

[24] James Manyika, Michael Chui, Peter Bisson, Jonathan
Woetzel, Richard Dobbs, Jacques Bughin, Dan Aharon, “The Internet of things: Mapping the value beyond the hype,” tech. rep., McKinsey Global Institute, 06 2015

[25] R. Eynon, “The rise of big data: what does it mean for education, technology, and media research?,” Learning, Media and Technology, vol. 38, no. 3, pp. 237–240, 2013.