International Journal of Distributed and Cloud Computing
Volume 5 Issue 1, June 2017
ISSN.: 2321-6840

BigData Analytical Challenges with IOT

Lalitha Balla[1*], Chavva Ravi Kishore Reddy[2], A V L N Sujith[3]
[1]Professor, Department of Computer Science and Engineering, JNTUAC, India.
[2]Assistant Professor, Department of Computer Science and Engineering, VLITS, Vadlamudi, India.
[3]Assistant Professor, Department of Computer Science and Engineering, JNTUP, India.
*Corresponding Author E-mail: lalitha_balla@yahoo.co.in

Abstract: Data in today's world is much more complex than ever before. With the technological advancements businesses are able to easily gather data both at the organizational level as well as from the external data sources. The accumulated data is huge and diverse with structured, unstructured components or data generated by Internet-of-Things (IOT). Businesses are in dire need to analyze these sets of data to derive a better value to the organizations. With analytics becoming central to all the business strategies, this paper presents a review of the challenges which the organizations have to take into account while dealing with these complex data residing in the data stores. Apart from the volumes and complexity of data, IOT brings in new challenges in the form of security to the BigData systems as a whole and data in particular. This paper also reviews the conceptual studies which have attributed to the growth of Bigdata technologies to provide business analytics by ensuring security to the data.

Keywords: BigData, BigData Analytics, Challenges, Internet-of-Things, Security

I. Introduction

BigData is being generated by almost all the components around us, but to derive a meaningful value from this data we need have processing power and analytical capabilities. Organizations have to carefully manage their business model around the key processes as shown in Fig 1.

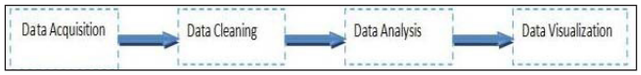

Fig. 1: BigData Processes

Insights from BigData will help the organizations to make better decisions, provide better customer service and get better operational efficiency and new sources of revenue. The importance of BigData is realized only it is able solve the business-related tasks such as failure detections in real-time, performing risk-management and detecting fraudulent behaviors before they can affect the organization [3].

A. Data Acquisition

Data acquisition is the process of gathering data from different data sources. It is commonly governed by the three V's- volume, variety and velocity. The core of data acquisition comes down to acquiring data from distributed information sources to store them in a BigData-capable data storage. To achieve the organizations have to keep in mind of the protocols that help in information gathering from these distributed sources and technologies that provide the persistent storage. The nature and source of this data is very diverse and complex. Tools and methods are used deal with such data for improved performance of the BigData stores.The main goal of a data acquisition strategy employed at an organization must understand the needs of the system and take the right decision on which tool is best to ensure the acquisition.

B. Data Cleaning

Data cleaning is the process of preparing the data for analysis by dealing with corrupt or inaccurate records. The process is mainly in data stores which usually have the inaccurate, incomplete or irrelevant data. A typical data cleaning process involves removing of errors or validating and correcting values of records. Raw data which is usually present in several formats is transformed by the data cleaning process. High quality data need to have high accuracy, completeness, uniformity and consistent behavior across the system. Data preparation i.e. Data acquisition, data cleaning and data management constitute major part of the work for a data scientist. Understanding these steps improves the efficiency of the data acquisition activities there by streamlining the business activities which improves the productivity and revenue.

C. Data Analytics

BigData analytics deals about using advanced analytical algorithms and methods against very large and complex data sets. It is about discovering patterns and other useful information and knowledge that comes with analyzing the data. The important analytical techniques include text analysis, data mining, machine learning, natural language processing and

many others. Data analytics is the core of BigData to obtain the business intelligence. Regardless of the way BigData is generated from and how it is stored the main motive of the data lies with analysis. With data processing and management, we acquire new knowledge which helps in responding to the emerging challenges in a timely manner [4].

D. Data Visualization

Data visualization is about presenting the results of the data analysis in a pictorial or graphical format. It helps the decision makers to look at the analytical results visually and identify patterns more easily. Data visualization helps the businesses to identify areas that needs focus, improve the users experience and predict the trends easily. The visual results have to provide quickly with meaningful outcomes which help the users to gain insights into data.

E. Internet of Things (IoT)

The term Internet of Things (IoT) refers to situations where internet connectivity and computing capabilities are present for day to day items and sensors, making these devices generate and consume data without any human influence. IoT provides real time alerts and insights, which helps us in making intelligent and real time actions to address issues and get productivity. Various communication models are available for the IoT devices to connect and provide value to the user. IoT is expected to spread immeasurably in the coming years and this will provide new services that will improve quality of life of consumers and productivity of the organizations. Some findings forecasted that 20.8 billion connected things will in use around the world by 2020. The International Data Corporation (IDC) predicted that by 2018, there will be 22 billion installed IoT devices and the number of wearable deviceswill be around 214.6 million by 2019 [5]. So keeping with these numbers in mind there is a lot of scope in developing key technologies which will help in maintaining the data in BigData stores.

II. Challenges

Even though BigData provides innumerous benefits, a large number of challenges are to be addressed to make use of these benefits. The challenges of BigData can broadly classified into three [6, 7].

- Data Challenges
- Process Challenges
- Management Challenges

This paper will review the BigData challenges as shown in fig 2, with the growth of IoT and how these issues will affect the organizations.

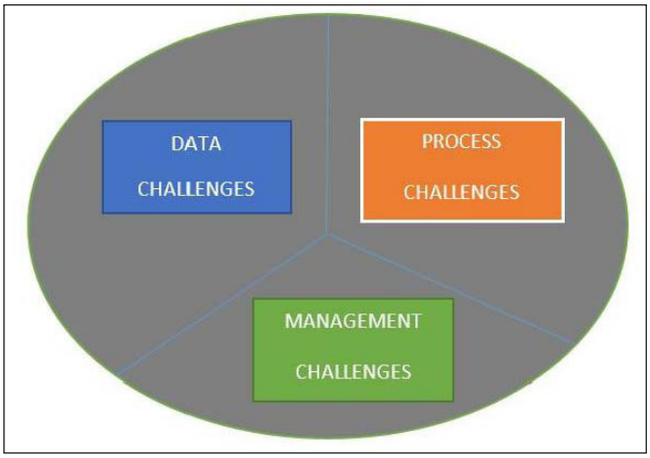

Fig. 2: BigData Challenges

A. Data Challenges

The data challenges relate to the characteristics of the data. Different researchers have different understanding towards the characteristics of data.Some specify 3 V's (Volume, Variety, Velocity) [8], 4 V's (Volume, Variety, Velocity, Variability) [9] and 6 V's (Volume, Variety, Velocity, Variability, Veracity, Value) [10], while others specify 7 V's (Volume, Variety, Velocity, Variability, Veracity, Value, Visualization) [11].

a) Volume

Volume refers to the large data sets consisting of terabytes, petabytes of data or even more. IoT will account formost part of these datasets in the future. Cisco [12] estimates that the total volume of data generated by IoT will reach 600Zetabytes per year by 2020. This huge amount of data being generated is bigger than the BigData stores. So it will be a challenge for the storage management.

IoT threatens to generate massive amounts of data from sources that are globally distributed. Transferring the entire data into a single location for processing will not be good option both technically and economically. So the organizations have to look for options of distributed mini data centers, where initial processing can occur then relevant data can be forwarded to a central site for additional processing.

b) Variety

Variety deals with the numerous data formats which the BigData has to store and process. The data could be structured, semi-structured or unstructured. The volume of data generated by the sources of BigData and IoT does not follow a specific format. The data captured into the data stores could be diverse from user's data, transactional data to web data and many more such formats [13].

IoT adds to these complex data types, which includes customer's data, customer's relationship management data, business documents, Omni-channel or cross-device data, location-

based data, population data, production and manufacturing data. These different formats of data are a big challenge to the BigData systems.

c) Velocity

Velocity refers to rate at which data is getting accumulated into BigData stores. The main challenge of velocity to BigData is the requirement to manage the large amounts of unstructured data getting into the stores at rates which the normal databases are not capable to store and maintain. IoT applications add to the complexity of these data elements and rate at which the IoT devices generate data is very higher than their BigData counterparts. One of the prominent features of IoT is its real time communication of information about the connected things. When concerned with velocity, IoT applications present challenges of storage of data at higher rates and to run analytical queries on these stored data items. To perform the real time analytics is a huge challenge to BigData in context of IoT.

d) Variability

Variability of data refers to those data elements whose meaning is constantly and rapidly changing. The data which is generated by the IoT devices is huge and the rates of change and variability are very high. Analyzing data with high variability is a challenge and gaining insights from this data is further a challenge as each analysis may provide different results due to variability in the data. Algorithms which are to be developed to deal with such data should be able to understand the context of the data and derive its meaning, their by gain valuable insights [14].

e) Veracity

Veracity refers to the quality of captured data in the stores. Even though there is a potential value in using the BigData, the data in stores is worthless if it is not accurate. This is true in cases where the data is being used for automated decision making. The accumulated data may be full of biases, abnormalities and imperfections. This data may largely affect the analytics of BigData and these results will be ineffective in decision making process.The important factors which define the quality of data are the accuracy and reliability of the data source. The reliability of the data source will be a concern of BigData when dealing with IoT devices.

f) Visualization

Visualization is about presenting the data in a form that is easy for the user to comprehend easily. Many BigData applications which have poor performance in functionality, scalability and response time is a problem when conducting data visualization. The main reason for this is as a result of large sizes of data and high dimensionality of the data elements. Given the amount of BigData with IoT is dealt, the need for enormous parallelization is a challenging task in visualization. To decompose a problem into independent tasks and to run concurrent queries on them is a challenge for visualization algorithms. With visualization, real users can easily explore and interact with data. It plays an enormous role in assisting the user in deriving better insights from the existing data.

As IoT brings in huge volumes of data at higher rates with additional complexity, not only providing storage but also developing visualization tools will be challenge to the BigData. Existing BigData visualization tools used with IoT shows poor performance in terms of functionality and response time. To provide effective visualization for IoT several important issues such as visual noise, variance data and information loss should be taken into consideration [15, 16].

g) Value

Value is about extracting knowledge or information from the huge amounts of structured or unstructured data. Value is an essential feature of BigData as most of the pieces of data independently may seem insignificant, but when this data is correlated, value can be derived which might be helpful to the organization. As IoT is involved with the connected things which are both personal and industrial, deriving value from data collected from these devices will be highly helpful in improving industrial efficiency and also to provide better user experience.

In order to derive value from such complex data, BigData faces challenges while storing and performing analytical queries by keeping in mind the volume and complexity of the data items accumulated in the BigData stores. So developing analytical methods and algorithms which takes into consideration of the IoT data is the key challenge for the growth of BigData.

B. Process Challenges

Process challenges are those set of challenges which deals with data acquisition, data storage, data analysis and data visualization.The various process challenges include Data Acquisition and warehousing, Data cleaning and Integration, Data Analysis and Modelling and Data Interpretation.

Data Acquisition and Storage is related to gathering data from diverse sources and storing them in BigData stores to generate value from it. Researchers [17] points that one of the key hurdles to the analysis of BigData is the issues in scaling of data collection and storage. IoT promises the availability and generation of vast amounts of data. If this data is easily accessible, services will be able to leverage things efficiently and provide valuable information.

The key challenge for BigData in context with IoT is the acquiring data from large number of heterogeneous devices and storing them. This involves developing a common API for IoT devices to interface with BigData systems and involving additional steps such as data cleaning and data integration to store the data in common formats for easy analysis.

The new architecture of BigData with IoT will present significant challenges such as backing up these mini data stores, or backing up the entire volumes of data generated by IoT devices and shifting and sorting the data elements. These operations further increase the processing loads on storage and network resources that are need to be managed. The storage infrastructure is also

heavily influenced with arrival of IoT data. Organizations are starting to prefer cloud platforms over specific infrastructure, as existing infrastructure has to be expanded to handle the load of BigData. These solutions are able to provide scalability and flexibility to store the IoT data. But the cloud based solutions have to provide better privacy to user's data.

Data Integration and Cleaning refers to maintaining a homogeneous view of different formats of data acquired from different data sources. Data Integration involves the data acquisition from different sources like social media, IoT and other communication devices and organizing and storing these complex data formats in a uniform format [18]. These unique formats help in acquiring value from the stored values [19].

This step is very important in the BigData cycle as effectiveness of these processes decides how effective the analytical results are produced. Dealing with structured and unstructured data by adjusting into a unique structure is a challenge to BigData. New technologies have been developed to extract images, videos and other information from unstructured data [20]. With large data sets comprising more abnormalities and ambiguities additional processing steps are needed such as cleaning, reduction and transmission [21, 22].

Data Mining deals with identifying new patterns and correlations between the data items to get valuable insights about the data. The size and complexity of BigData with IoT imposes new requirements on data mining and also the diversity in the data sources poses another challenge [23, 24, 25]. Another important task of data mining is acquiring information from complex data which needs the analysis of data properties and finding associations among different data elements.

Many parallel and sequential programming models for querying large and complex data sets have been developed. However devising parallel algorithms compatible with the latest IoT architecture is a challenging task for researchers.

C. Management Challenges

Management challenges tackle privacy, security, governance, data and information sharing, cost/operational expenditure and data ownership. This section presents various data management challenges of BigData in context with IoT.

Privacy deals with policies of the organization of to what data can be shared with the third parties. It mainly deals with protecting user's data from unauthorized access. In BigData from the context of IoT, security and privacy are key challenges in storing and analyzing large amounts of data. The security risk associated with IoT data is the presence of heterogeneous types of devices and the nature of data generated by them. It is a challenge to authenticate these vast varieties of devices. This gives rise to new architectures. As a result developing security protocols and implementing them would be a challenging task to developers.

Security should be provided to the data in the BigData stores. Data generated through IoT incurs various security problems which include keeping the data consistent, identifying suspicious traffic patterns and interoperability between the devices. Existing security solutions are no longer applicable to provide complete security for BigData with IoT in place. Existing solutions are not enough to deal with dynamic data sets as most of the existing solutions involve static data sets. So developing a security model to implement on the IoT devices will be the most challenging task of BigData as the IoT devices have constraints in terms of computational capabilities, memory size etc. So the security model has to keep in mind all these heterogeneous nature of the devices.

III. Conclusion

The rate at which the data is being generated has increased enormously over last few years with the explosion in smart and sensor devices. The combination of IoT and BigData technologies involves processing, transforming and analyzing large amounts of data at great speeds is necessary. In this paper we broadly listed out the challenges of BigData in the context of IoT. So, there is a lot of scope for research to overcome many of these challenges. The existing solutions are still at their early stages of development and there will be scope in improving many of these solutions. So real time analytical solutions would be necessary to gain faster insights.

References

[1] J. Gantz, and D. Reinsel, "The Digital Universe in 2020: Big data, bigger digital shadows, and biggest growth in the Far East," IDC- EMC Corporation, 2012. Available: http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf

[2] C. Dobre, and F. Xhafa, F. "Intelligent services for big data science," *Future Generation Computer Systems*, vol. 37, pp. 267-281, 2014.

[3] A. McAfee, and E. Brynjolfsson, (2012). "Big data: The management revolution," *Harvard Business Review*.

[4] J. Chen, Y. Chen, X. Du, C. Li, J. Lu, S. Zhao, and X. Zhou, (2013). "Big data challenge: A data management perspective," *Frontiers of Computer Science*, vol. 7, no. 2, pp. 157-164, 2013.

[5] G. Press, Internet of Things (IoT) "Predictions from Forrester, Machina Research, WEF, Gartner, IDC," 2016. Available: http://www forbes.com/sites/gil-press/2016/01/27/internet-of-things-iot-predictions-from-forrester-machina-research-wef-gartner-idc/#4b1601546be6.

[6] R. Akerkar, "*Big Data Computing*," Florida, USA: CRC Press, Taylor & Francis Group, 2014.

[7] R. V. Zicari, R. V. (2014). "Big Data: Challenges and Opportunities," In R. (Ed.), *Big data computing*. Florida, USA: CRC Press, Taylor & Francis Group, pp. 103-128, 2014.

[8] T. Shah, F. Rabhi, and P. Ray, "Investigating an ontology-based approach for Big Data analysis of inter-dependent medical and oral health conditions," *Cluster Computing,* vol. 18, no. 1, pp. 351-367, 2015.

[9] Z Liao, Q. Yin, Y. Huang, and L. Sheng, "Management and application of mobile big data," *International Journal of Embedded Systems*, vol. 7, no. 1, pp. 63-70, 2014.

[10] A. Gandomi, and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics," *International Journal of Information Management*, vol. 35, no. 2, pp. 137-144, 2015.

[11] U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, "Critical analysis of Big Data challenges and analytical methods," *Journal of Business Research*, vol. 70, pp. 263-286, 2017.

[12] J. Bradley, J. Barbier, and D. Handler, "Embracing the Internet of Everything to Capture Your Share of $14.4 Trillion," Available: http://www.cisco.com/c/dam/en_us/about/ac79/docs/innov/IoE_Economy.pdf

[13] H. Chen, R. H. Chiang, and V. C. Storey, V. C. "Business intelligence and analytics: From Big Data to big impact," *MIS Quarterly*, Vol. 36, no. 4, pp. 1165-1188, 2012.

[14] X. Zhang, Y. Hu, K. Xie, W. Zhang, L. Su, and M. Liu, "An evolutionary trend reversion model for stock trading rule discovery," *Knowledge-based Systems*, vol. 79, pp. 27-35, 2015.

[15] C. P. Chen, and C. Y. Zhang, "Data Intensive applications, challenges, techniques and technologies: A survey on Big Data," *Information Sciences*, 275, pp. 314-347, 2014.

[16] E. Y. E. Gorodov, and V. V. E. Gubarev, "Analytical review of data visualization methods in application to Big Data," *Journal of Electrical and Computer Engineering*, pp. 22, 2013.

[17] J. Paris, J. S. Donnal, and S.B. Leeb, D. B. Nilm, "The non-intrusive load monitor data-base," *IEEE Transactions on Smart Grid*, vol. 5, no. 5, pp. 2459-2467, 2014.

[18] B. B. Ahamed, T. Ramkumar, and S. Hariharan, "Data integration progression in large data source using mapping affinity," *In 7th International Conference on Advanced Software Engineering and Its Applications (ASEA)*, 2014.

[19] J. Liu, and X. Zhang, "Data integration in fuzzy XML documents," *Information Sciences*, vol. 280, pp. 82-97, 2014.

[20] D. Agarwal, P. Bernstein., E. Bertino, S. Davidson, U. Dayal, and M. Franklin, "Challenges and opportunities with big data: A community white paper developed by leading researchers across the United States," *Whitepaper, Computing Community Consortium,* 2012.

[21] A. Gani, "A survey on indexing techniques for big data: taxonomy and performance evaluation," *Knowledge and Information Systems*, vol. 46, no. 2, pp. 241-284, 2016.

[22] M. Chen, V. C. Leung, and S. Mao, "Directional controlled fusion in wireless sensor networks," *Mobile Networks and Applications,* vol. 14, no. 2, pp. 220-229, 2009.

[23] Y. Sun, "Mining knowledge from interconnected data: A heterogeneous information network analysis approach," *Proceedings of the VLDB Endowment*, vol. 5, no. 12, pp. 2022-2023, 2012.

[24] M. Chen, "Itinerary planning for energy-efficient agent communications in wireless sensor networks," *IEEE Transactions on Vehicular Technology*, vol. 60, no. 7, pp. 3290-3299, 2011.

[25] D. Zhang, "A taxonomy of agent technologies for ubiquitous computing environments," *TIIS*, vol. 6, no. 2, pp. 547-565, 2012.