Big Data Analytics in IOT: Challenges, Open Research Issues and Tools

Fabián Constante Nicolalde[1,2], Fernando Silva[1], Boris Herrera[2], and António Pereira[1,3(✉)]

[1] School of Technology and Management, Computer Science and Communications Research Centre, Polytechnic Institute of Leiria, Leiria, Portugal
2162316@my.ipleiria.pt, {fernando.silva, apereira}@ipleiria.pt
[2] Universidad Central del Ecuador, Quito, Ecuador
bherrera@uce.edu.ec
[3] Information and Communications Technologies Unit, INOV INESC Innovation-Delegation Office at Leiria, Leiria, Portugal

Abstract. Terabytes of data are generated day-to-day from modern information systems, cloud computing and digital technologies, as the increasing number of Internet connected devices grows. However, the analysis of these massive data requires many efforts at multiple levels for knowledge extraction and decision making. Therefore, Big Data Analytics is a current area of research and development that has become increasingly important. This article investigates cutting-edge research efforts aimed at analyzing Internet of Things (IoT) data. The basic objective of this article is to explore the potential impact of large data challenges, research efforts directed towards the analysis of IoT data and various tools associated with its analysis. As a result, this article suggests the use of platforms to explore big data in numerous stages and better understand the knowledge we can draw from the data, which opens a new horizon for researchers to develop solutions based on open research challenges and topics.

Keywords: Big Data Analytics · Internet of Things · Hadoop
Massive data · Structured data · Unstructured data

1 Introduction

The development of Big Data and the Internet of Things (IoT) is growing, affecting all technological areas and companies by increasing the benefits for organizations and individuals. The growth of data generated through IoT has played an important role in the large data landscape. The collection of these data is difficult to process using database management tools or data processing applications, since they are usually available in semi-structured, and unstructured fashion [1, 2].

The large data are classified according to three aspects: (a) volume, (b) variety, and (c) velocity [3]. This classification was introduced by Gartner to describe the elements of large data challenges [4]. The ability to analyze and use huge amounts of IoT data,

© Springer International Publishing AG, part of Springer Nature 2018
Á. Rocha et al. (Eds.): WorldCIST'18 2018, AISC 746, pp. 775–788, 2018.
https://doi.org/10.1007/978-3-319-77712-2_73

including applications in smart cities, smart transport systems, smart energy meters and remote monitoring devices for patient care, offer immense opportunities.

The key problem in Big Data's analysis is the lack of coordination between databases as well as analysis tools, such as mining and statistical analysis. These challenges grow when the objective is to realize discovery and representation knowledge for specific applications. The research was carried out by several studies on Big Data and its trends [5]. Large data analysis is defined as the steps in which a variety of IoT data are examined to reveal trends, unseen patterns, hidden correlations and new information. IoT Big Data Analytics aims to help business associations and other organizations to achieve better understanding of data, and efficient decision making [6].

Big Data Analytics allows data mining and scientists to analyze large amounts of unstructured data that can be harnessed using traditional tools [7]. The goal of Big Data Analytics is to immediately extract informed information using data mining techniques that help make predictions, identify recent trends, find hidden information and make decisions [8].

The IoT data are different from the normal Big Data collected through systems in terms of characteristics, due to the various sensors and objects involved during data collection, which include heterogeneity, noise, variety and rapid growth. Statistics show that the number of sensors will increase by 1 trillion by 2030 [2].

The introduction of Big Data Analytics and IoT in Big Data requires huge resources, and IoT has the ability to offer an excellent solution. Implementing IoT and integrating Big Data into solutions can help solve problems related to storage, processing, data analysis and visualization tools. Areas of application, such as smart green environments, smart traffic, smart networks, intelligent buildings and logistics management, can benefit from the aforementioned provision [9, 10].

In this paper we focus mainly on the processing of data generated by IOT using current technological solutions, mainly associated with Big Data and not at the business level. The rest of the paper is divided into the following sections: Sect. 2 presents an overview of Big Data and IoT. In Sect. 3 the challenges during Big Data Analytics are addressed. Section 4 presents Big Data Analytics' open-ended research problems in IoT, which will help on processing Big Data and extracting useful insights from it. Section 5 provides an overview of the main technical tools used to process Big Data. Section 6 presents observations and suggestions for future work within this line of research. Finally, Sect. 7 presents the conclusions of the research carried out.

2 Overview of Big Data and IoT

Current IoT offers several opportunities for data analysis for big data analytics, this section provides an overview of these technologies.

2.1 Big Data

Big data is no more than an enormous amount of data (structured, unstructured and semi-structured) that exceeds the ability of conventional software to be captured, managed and processed in a reasonable time. In 2012, it was estimated that its size

should be between dozen terabytes to several petabytes of data in a single data set [1]. The MIKE2.0 methodology, dedicated to investigating issues related to information management, defines Big Data in terms of useful permutations, complexity and difficulty in erasing individual records [11]. McKinsey Global Institute defined Big Data as the size of data sets that are a better database system tool than the usual tools for capturing, storing, processing, and analyzing such data [12]. Traditional database systems are inefficient when increasing data or Big Data is quickly stored, processed and analyzed [12]. "The Digital Universe" writes about data technologies as a new generation of technologies and architectures that aim to extract the value of a massive volume of data in various formats enabling high-speed capture, discovery and analysis [13]. This previous study also characterizes Big Data in three aspects: (a) data sources, (b) data analysis, and (c) presentation of analysis results. Beyer defines a model which uses the 3Vs (volume, variety, velocity) for describing Big Data [4]. Volume refers to the huge amount of data that are being generated daily, whereas velocity is the rate of growth and how fast the data are gathered for analysis. Variety provides information about the types of data such as structured, unstructured and semi structured [4]. Figure 1 refers to this latter definition of Big Data.

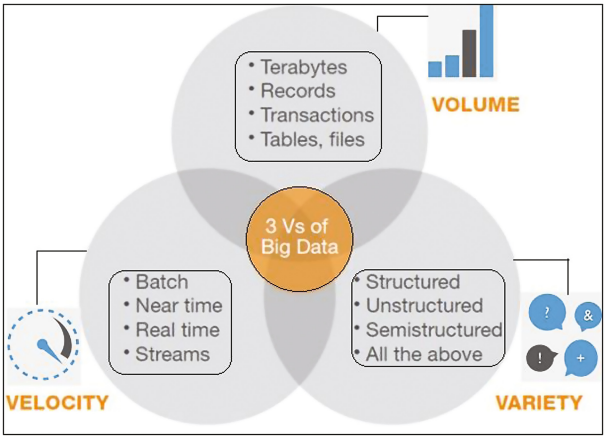

Fig. 1. Definition of big data as the three vs. Adapted from [4]

2.2 IoT

IoT provides a platform for sensors and devices to communicate seamlessly within an intelligent environment and enables the exchange of information between platforms in a convenient manner [2]. The recent adaptation of different wireless technologies, places IoT as the next revolutionary technology to take advantage of all the opportunities offered by Internet technologies. IoT has been adopted in the development of intelligent systems, such as smart offices, intelligent retail, intelligent agriculture, intelligent water, intelligent transportation, smart health and intelligent energy [14].

IoT has emerged as a new trend in recent years, since mobile devices, transport facilities, public facilities and appliances can be used as data acquisition equipment.

All surrounding electronic equipment to facilitate daily operations, such as wristwatches, vending machines, emergency alarms and garage doors, as well as appliances such as microwave ovens, air conditioners and water heaters are connected to Internet and can be remotely controlled [2].

Data collection devices detect data and transmit data using integrated communication devices through a variety of communication solutions such as Bluetooth, WI-Fi, ZigBee and GSM. These communication devices transmit data and receive commands from remotely controlled apparatus that allow direct integration with the physical world through computer systems to improve living standards.

More than 50 billion devices, from smartphones, notebooks, sensors and game consoles, are expected to be connected to the Internet through several heterogeneous access networks enabled by technologies such as radio frequency identification (RFID) and wireless sensor networks [2]. IoT could be recognized in three paradigms: Internet-oriented; sensors; and knowledge. The recent adaptation of different wireless technologies places IoT as the next revolutionary technology to benefit from all of the opportunities offered by Internet technology [15].

2.3 Big Data Analytics

Big Data Analytics (BDA) involves the processes of examining large datasets containing a variety of data types [6] to reveal invisible patterns, hidden correlations, market trends, customer preferences, and other useful business information [7]. It involves the processes of searching in a database, mining, and data analysis, with the purpose of improving the performance of organizations [7, 16]. BDA emerges as a response to these needs, as it focus on the study to improve the ability to obtain, store, analyze and visualize millions of data that would be inaccessible to conventional analysis processes or tools [17]. The main objective of BDA is to help any area of research to improve the understanding of the data and, therefore, to make efficient and informed decisions. BDA allows to analyze a large volume of data that cannot be exploited with traditional tools [7].

BDA requires technologies and tools that can transform a large amount of structured, unstructured and semi-structured data into a more comprehensive format of data and metadata for analytical processes. The algorithms used in these analytical tools must discover patterns, trends and correlations over a variety of time horizons in the data. After analyzing the data, these tools visualize the findings in tables, graphs and spatial graphs for efficient decision making. The challenge focuses on the performance of current algorithms used in BDA, which is not increasing linearly with the rapid increase in computational resources [18].

3 Challenges in BDA

IoT and BDA have been widely accepted by many organizations. However, a number of existing research problems have not yet been addressed. It is necessary to understand several computational complexities to handle the challenges, such as information

security and computational methods, to analyze great data. For example, many statistical methods that work well for small data sizes do not scale to voluminous data. Similarly, many computational techniques that work well for small data face significant challenges when analyzing large data [19]. Here, the challenges of large analytical data fall into four general categories, namely: data storage and analysis; discovery of knowledge and computational complexities; scalability and data visualization; and information security.

3.1 Data Storage and Analysis

Due to the high cost of storage, the first BDA challenge is the storage media and higher input/output speed. In this case, the accessibility of the data must prioritize for the discovery and representation of knowledge.

In past decades, hard drives were used to store data, but their random input/output performance is slower than sequential input/output. However, the available storage technologies cannot have the performance required to process Big Data [5].

Another challenge with Big Data analysis is attributed to the diversity of data. The reduction of data, the selection of data and the selection of features are important tasks, due to the large size of the data sets [5]. This happens because existing algorithms do not always respond at an appropriate time when dealing with this high-dimensional data. Automation of this process and development of new machine learning algorithms for ensuring consistency is a great challenge in recent years [20].

Recent technologies, such as Hadoop [21] and MapReduce [22], allow the collection of a large amount of semi-structured and unstructured data in a reasonable amount of time. The challenge focuses on how to effectively analyze these data to obtain better knowledge. A standard process for this purpose is to transform semi-structured or unstructured data into structured data, and then apply data mining algorithms to extract knowledge. A framework for analyzing the data was discussed by Das and Kumar [23].

3.2 Knowledge Discovery and Computational Complexities

Knowledge discovery and representation is a prime issue in big data and it includes several secondary fields such as authentication, archiving, administration, preservation, information retrieval and representation. Due to the increase in size of Big Data the existing tools may not be efficient to process this data for meaningful information. The most popular approach in the case of data management are Data Warehouses and Datamarts [5]. A Data Warehouse is primarily responsible for storing the data that is obtained from the operating systems, while a Datamart is based on a Data Warehouse and facilitates the analysis. The basic objective of these researches is to minimize the processing of computational costs and complexities [5]. Nevertheless, today's BDA tools perform poorly in handling computational complexities, uncertainties, and inconsistencies. This leads to a great challenge for developing techniques and technologies that can handle computational complexity, uncertainty and inconsistencies in an effective way [5].

3.3 Scalability and Visualization of Data

The most important challenge for BDA's techniques is its scalability and security. In the last decades researchers have paid attention to speed up data analysis and speed up processing, followed by Moore's Law [5]. Data scalability has become necessary for many organizations dealing with explosive datasets, precisely when performance issues arise. A scalable data platform accommodates rapid changes in data growth, whether in traffic or volume, using hardware or software aggregate to increase data production and storage [24].

The objective of data visualization is to present the data in a more appropriate way, using some techniques of graphic theory. The graphical display provides the link between the data with an appropriate interpretation. However, online marketplaces, like Flipkart, Amazon or e-bay, have millions of users and billions of products to sell each month, this generates a lot of data. Some companies use a Tableau tool to display big data [5]. Tableau is a centralized analytical platform for data discovery and exploration that combines the two most important assets of a company: its people and their data (both big data and data of lesser volume) [25].

3.4 Information Security

In BDA, a lot of data is correlated, analyzed and used to extracted significant patterns. All organizations have different policies to protect their confidential information. Preserving sensitive information is a major problem in BDA. There is a great risk of information security that is becoming a major data analysis problem. Big Data security can be improved by using authentication, authorization, and encryption techniques. Attention should be given to developing a model of security policy and a multilevel prevention system [5].

4 Open Research Problems in IoT for BDA

IoT has an imperative economic and social impact for the future construction of information, network and communication technologies. It presents challenges in combinations of volume, speed and variety. Several diversified technologies such as computational intelligence and large data can be incorporated to improve data management and discovery of knowledge of large-scale automation applications [26].

The biggest challenge presented by Big Data is the acquisition of knowledge from IoT data. It is essential to develop infrastructures to analyze IoT data. Numerous IoT devices generate continuous flows of data and researchers can develop tools to extract meaningful information from these data using automated learning techniques [26].

Understanding data streams and analyzing them for meaningful information is a challenge and leads to BDA. Machine learning algorithms and computational intelligence techniques are the only solution to handle large IoT prospective data, the key technologies that are associated with IoT are also discussed in many research papers [27]. Figure 2 shows an overview of the process of data discovery and knowledge in IoT.

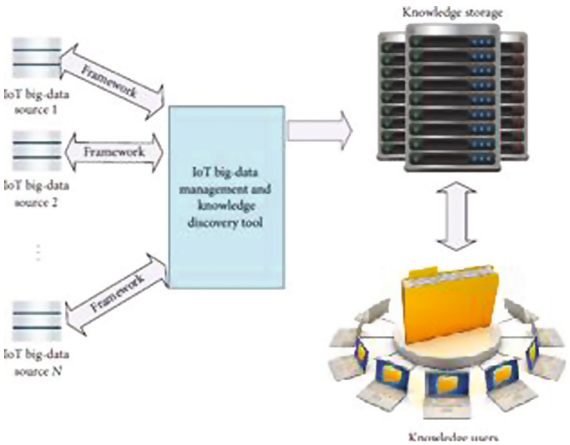

Fig. 2. IoT Big Data knowledge discovery. Adapted from [26].

The knowledge exploration system, illustrated in Fig. 3, consists of four segments: knowledge acquisition; knowledge base; knowledge diffusion and knowledge application [26].

Acquisition of knowledge is where the knowledge is discovered through the use of various traditional and computational intelligence techniques. Databases of knowledge are used to store the discovered knowledge and expert systems are generally designed based on the knowledge discovered. Dissemination of knowledge is important in order to get meaningful information from the knowledge base. The extraction of knowledge is a process that looks for documents, knowledge within documents, as well as knowledge bases. Application of knowledge applies knowledge discovered in several applications. There are many topics, debates and research in this area of knowledge exploration [26].

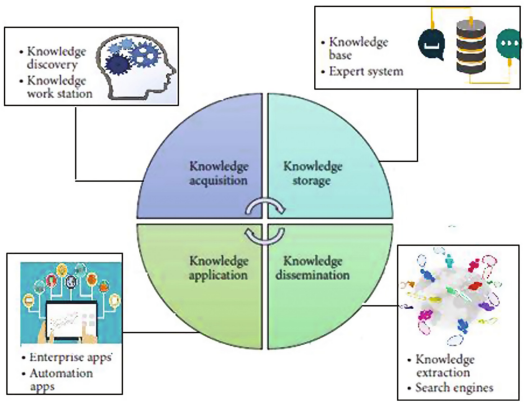

Fig. 3. IoT knowledge exploration system. Adapted from [5].

5 Tools for Big Data Processing

There are several tools available for processing Big Data. In this section, we discuss some current techniques for analyzing Big Data with a focus on three important emerging tools, such as MapReduce [22], Apache Spark [28], and Storm [29]. Most of the tools available focus on batch processing, flow processing, and interactive analysis. Most batching tools are based on the Apache Hadoop infrastructure [30], such as Mahout [31] and Dryad [32]. Flow data applications are mainly used for real-time analytics. An example of large-scale streaming platform is Splunk [33].

Dremel and Apache Drill [34] are the big data platforms that support interactive analysis. These tools are very useful for the development of Big Data projects. The typical workflow of Big Data projects is discussed by Huang et al. [35]. The workflow of a Big Data Project is shown in Fig. 4.

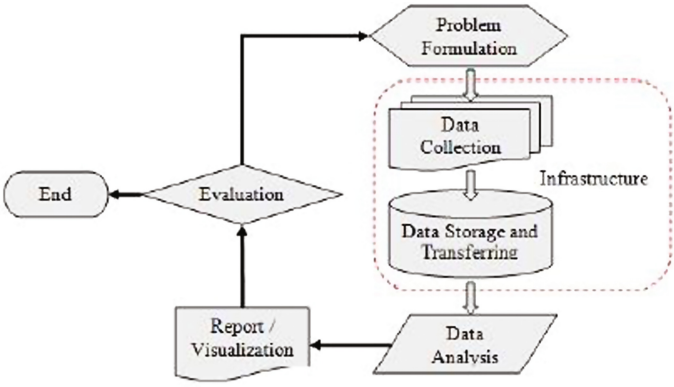

Fig. 4. Workflow of a Big Data project [35]

5.1 Apache Hadoop and MapReduce

Standard frameworks are considered for storing large volumes of data, consist of Hadoop kernel, MapReduce, Hadoop distributed file system (HDFS) and Apache Hive. Map Reduction is a programming model for processing large datasets based on split and conquer method. They are also used to analyze and process data, and are used by companies such as Facebook and Yahoo [30].

The Hadoop library uses simple programming models for the storage and distributed processing of large clustered data sets, giving redundancy so that nothing is lost and at the same time taking advantage of many processes. It has a distributed file system in each cluster node: the HDFS, and it is based on the two-stage MapReduce process [30]. The combination of these frameworks allows the data to be replicated and distributed by N nodes, benefiting from the capacity of access to large volumes. To execute an operation with distributed data, Hadoop is responsible for processing each part of the data in the node that contains them. If there is a need to grow in capacity, it is possible to add more nodes. Storage is handled by HDFS and MapReduce processing [36]. Figure 5 shows the High-Level Architecture of Hadoop.

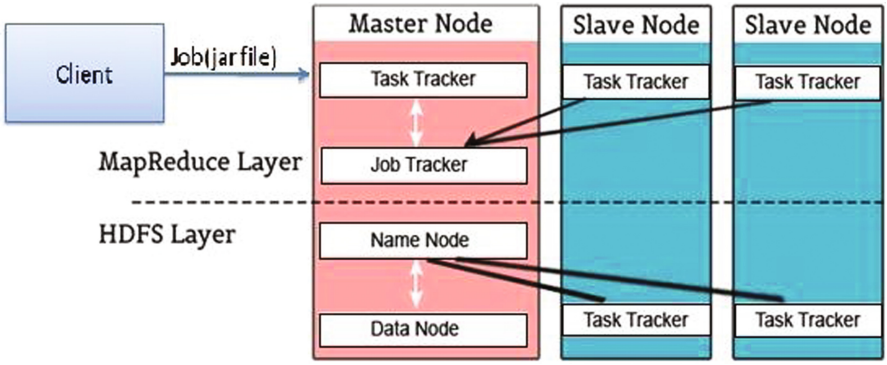

Fig. 5. High level architecture of hadoop. Adapted from [5].

Hadoop works on two types of nodes: the master node and the worker node. The master node divides the input into smaller sub-problems and then distributes them to the worker nodes in the map step. The master node then combines the outputs for all the sub problems in the reduction step [30].

5.2 Apache Spark

Apache Spark is a powerful open source processing engine built around speed, ease of use, and sophisticated analytics. It was originally developed at UC Berkeley AMP Lab in response to limitations in the MapReduce processing framework based on Hadoop's two-stage disk, maintaining the linear scalability, fault tolerance of MapReduce, and expanding processing capabilities in four important areas: In-memory analytics; Data federation; Iterative analytics; and Near real-time analytics [28].

In-memory analytics, allows in-memory access to intermediate results in a multi-stage processing pipeline through its Resilient Distributed Dataset (RDD) abstraction increasing performance, in some cases, up to 100 times faster than MapReduce. Data federation has an extensive set of libraries and APIs that allow developers to quickly create analytic workflows more efficiently for access to any data source, from HDFS or object storage, to NoSQL and relational databases.

The analysis of highly iterative algorithms is used for automatic learning and graph analysis, Spark's GraphX libraries unify the analysis of iterative graphs with ETL and interactive analysis. Provides Near Real-Time Analysis for integration with tools in the Hadoop ecosystem, and facilitates a unified platform that can be used to process Flume or Kafka data streams in micro-packages using Spark Streaming [28].

Driver Program is the start point of the execution of an application in the Spark Cluster. The Cluster Manage assigns resources and worker nodes to perform data processing in the form of tasks. Each application will have a set of processes, called executors that are responsible for executing tasks. Its main advantage is the support to deploy Spark applications in an existing Hadoop Cluster [37]. Figure 6 shows the Apache Spark Architecture diagram.

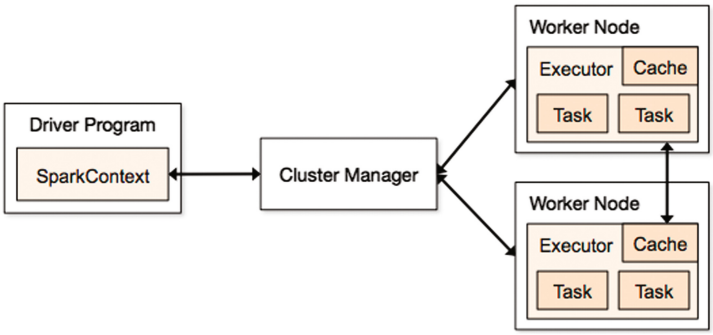

Fig. 6. Apache spark architecture diagram. Adapted from [37].

5.3 Dryad

Popular programming model to implement parallel and distributed programs to handle large context bases in the data flow graph. It is a set of computer nodes and it allows using the resources of a computer cluster to execute a program in a distributed way. A Dryad user uses thousands of machines, each with multiple processors or cores [32]. The main advantage of this model is that users do not need to know anything about simultaneous programming. A DRYAD application runs a computational directed graph that is composed of computational vertices and communication channels. Therefore, Dryad provides a wealth of functionality including task graph generation, scheduling of machines for available processes, handling transition faults in the cluster, collecting performance metrics and visualizing the job [32].

5.4 Apache Drill

Distributed system for interactive analysis of Big Data. It has more flexibility to support many types of query languages, data formats and data sources. It is also specially designed to exploit nested data. It also has a target of expanding to 10000 servers or more and reaches the ability to process petabytes of data and trillions of records in seconds. Drill uses HDFS for storage and MapReduce to perform batch analysis [34].

5.5 Storm

It is a distributed and fault-tolerant real-time computing system for processing large transmission data. The Storm Cluster is similar to the Hadoop Cluster. In Storm, different topologies are executed for different Storm tasks, while the Hadoop platform implements MapReduce jobs for the corresponding applications [29].

A Storm Cluster consists of two types of nodes, such as the Master Node and Worker Node. Master Node and Worker Node implement two types of functions, such as Nimbus and supervisor, respectively. The two roles have similar functions according to Jobtracker and Tasktracker of the MapReduce framework. Nimbus distributes the code through the Storm Cluster, schedules and assigns tasks to the Worker Nodes and

supervises the entire system [29]. The supervisor fulfills the tasks assigned by Nimbus. In addition, the process is started and terminated as needed based on Nimbus instructions. All of the computational technology is partitioned and distributed to a series of work processes and each work process implements a part of the topology [29].

5.6 Splunk

It is a real-time and intelligent platform developed for Big Data generated machine exploitation. It combines technologies in the cloud and Big Data. It helps the user to search, monitor and analyze their data generated by the machine through the web interface. The results are presented intuitively, such as graphs, reports and alerts [33].

Splunk is different from other flow processing tools. One of its features is the indexing of structured and unstructured data generated by machine, real-time search, analytical results reports and dashboards. It provides matrices for many applications, diagnoses problems for systems and information infrastructures, and intelligent support for business operations [33].

5.7 Jaspersoft

It is an open source software for scalable data analysis with real time analysis capabilities, database column reports and has a fast data visualization capability on popular storage platforms, including Mongo DB, Cassandra and Redis [38].

One of its most important features is the exploration of Big Data without extraction, transformation and load (ETL). It has the ability to build powerful HTML reports and panels interactively and directly from large data stores without the need for ETL. These generated reports can be shared with anyone inside or outside the user's organization [38].

5.8 Apache Mahout

It is an open source project that is mainly used to create scalable algorithms for automatic learning. Implements machine learning techniques such as clustering, sorting, pattern extraction, regression, dimensional reduction, evolutionary algorithms and batch-based collaborative filtering run at the top of the Hadoop platform through the MapReduce framework. Its basic objective is to provide a tool to face great challenges. Companies such as Google, IBM, Amazon, Yahoo, Twitter and Facebook have used it to implement scalable machine learning algorithms [31].

6 Suggestions for Future Work

Data transformation into knowledge is not an easy task for large-scale, high-throughput data processing, including leveraging the parallelism of current and future computer architectures for data mining. In addition, these data may imply uncertainty in many different ways.

Expressing access requirements to application data and designing programming language abstractions to exploit parallelism is an immediate necessity [39].

Machine learning concepts and tools are gaining popularity among researchers to facilitate meaningful results from these concepts. Each of the tools has its own advantages and limitations, the development more efficient tools to deal with the problems inherent to large data. Efficient tools to be developed must have provision for handling noisy data and imbalances, uncertainty and inconsistency, and missing values.

7 Conclusions

With the rise of smart devices and sensors, data production has increased in recent years. The interaction between Big Data and IoT is currently in its early stages of development, where it is necessary to process, transform and analyze large amounts of data with high frequency.

We discussed the relationship between BDA and IoT, examined various research topics, various opportunities generated by data analysis in the IoT paradigm, challenges and tools used for BDA.

From this research, it is understood that each Big Data platform has its individual approach; some are designed for batch processing, while some are good at real-time analytics.

The different techniques used for the analysis include statistical analysis, automatic learning, data mining, intelligent analysis, cloud computing and data flow processing. In the future, it is desirable that researchers pay more attention to these techniques for solving large data problems effectively and efficiently.

Acknowledgments. This work was possible thanks to Senescyt of Ecuador for the financing of research studies at the Polytechnic Institute of Leiria, Portugal and to the FCT project UID/CEC/4524/2016.

References

1. Tavana, M., Puranam, K.: Handbook of Research on Organizational Transformations through Big Data Analytics, p. 109 (2012)
2. Marjani, M., et al.: Big IoT data analytics: architecture, opportunities, and open research challenges. IEEE Access **PP**(99), 1 (2017)
3. Tiainen, P.: New opportunities in electrical engineering as a result of the emergence of the Internet of Things. AaltoDoc, Aalto Univ. (2016)
4. Beyer, M.: Gartner says solving 'Big Data' challenge involves more than just managing volumes of data. AaltoDoc, Aalto Univ. (2011)
5. Acharjya, D.P., Ahmed, K.: A survey on Big Data analytics: challenges, open research issues and tools. Int. J. Adv. Comput. Sci. Appl. **7**(2), 511–518 (2016)
6. Mital, R., Coughlin, J., Canaday, M.: Using Big Data technologies and analytics to predict sensor anomalies. In: Proceedings of the Advanced Maui Optical and Space Surveillance Technologies Conference, p. 84 (2014)
7. Golchha, N.: Big data-the information revolution. Int. J. Adv. Res. **1**, 791–794 (2015)
8. Tsai, C.-W.: Big Data analytics: a survey. J. Big Data **2**, 1–32 (2015)
9. Russom, P.: Big Data Analytics. TDWI Best Pract. Rep., pp. 1–35 (2011)

10. LaValle, S., Lesser, E., Shockley, R., Hopkins, M.S., Kruschwitz, N.: Big Data, analytics and the path from insights to value. MIT Sloan Manag. Rev. **52**, 21 (2011)
11. CollaB, O.: Big Data Definition, Open Framework, Information Management Strategy & Collaborative Governance| Data & Social Methodology - MIKE2.0 Methodology, 2015. http://mike2.openmethodology.org/wiki/Big_Data_Definition
12. Gantz, J., Reinsel, D.: The digital universe in 2020: Big Data, bigger digital shadows, and biggest growth in the far east. IDC Anal. Future (2012)
13. U. S. Profile, The Digital Universe in 2020: Big Data, Bigger Digital Shadows, and Biggest Growth in the Far East — United States, pp. 1–7 (2013)
14. Atzori, L., Iera, A., Morabito, C.: The Internet of Things: a survey, pp. 2787–2805 (2010)
15. Hsieh, H.-C., Lai, C.-H.: Internet of Things architecture based on integrated PLC and 3G communication networks. IEEE Access, pp. 853–856
16. Kwon, O., Lee, N., Shin, B.: Data quality management, data usage experience and acquisition intention of big data analytics. Int. J. Inf. Manag. **34**, 387–394 (2014)
17. Alvarado, J.C.: Estudio descriptivo de técnicas aplicadas en herramientas Open Source y comerciales para visualización de …, January 2017, 2016
18. Hashema, I.A.T., Yaqoob, I., Anuara, N.B., Mokhtara, S., Gania, A., Khanb, S.U.: The rise of 'Big Data' on cloud computing: review and open research issues. Inf. Syst. **47**, 98–115 (2015)
19. Kuo, M.-H., Sahama, T., Kushniruk, A.W., Borycki, E.M., Grunwell, D.K.: Health Big Data analytics: current perspectives, challenges and potential solutions. Int. J. Big Data Intell. **1**, 114–126 (2014)
20. Huang, Z.: Extensions to the k-Means algorithm for clustering large data sets with categorical values. Data Min. Knowl. Discov. **2**(3), 283–304 (1998)
21. T. A. S. Foundation., Apache Hadoop (2014). http://hadoop.apache.org/
22. Dean, J., Ghemawat, S.: MapReduce: simplified data processing on large clusters. In: Proceedings of the 6th Symposium on Operating Systems Design and Implementation, pp. 137–149 (2004)
23. Kaisler, S., Armour, F., Money, W., Espinosa, J.A.: Big Data issues and challenges, vol. 5, no. 2013, pp. 2013–2015 (2015)
24. E. Consulting: The importance of scalability in big data processing. http://blog.eccellaconsulting.com/the-importance-of-scalability-in-big-data-processing
25. Hanrahan, P.: Tableau (2017). https://www.tableau.com/es-es/resource/business-intelligence
26. Mishra, N., Lin, C.C., Chang, H.T.: A cognitive adopted framework for IoT Big-Data management and knowledge discovery prospective. Int. J. Distrib. Sens. Networks **2015** (March), 1–13 (2015)
27. Chen, X.-Y., Jin, Z.-G.: Research on key technology and applications for internet of things. Phys. Procedia **33**, 561–566 (2012)
28. Spark – A modern data processing framework for cross platform analytics Deploying Spark on HPE Elastic Platform for Big Data
29. A. S. Foundation: Apache Storm (2015)
30. T. O. Center: Introducción a Hadoop y su ecosistema. http://www.ticout.com/blog/2013/04/02/introduccion-a-hadoop-y-su-ecosistema/
31. Ingersoll, G.: Introducing apache mahout: Scalable, commercial friendly machine learning for building intelligent applications, pp. 1–18. White Paper, IBM Developer Works (2009)
32. Isard, M., Budiu, M., Yu, Y., Birrell, A., Fetterly, D.: Dryad: distributed data-parallel programs from sequential building blocks. ACM SIGOPS Oper. Syst. Rev. **41**, 59–72 (2007)
33. Chen, C.L.P., Zhang, C.Y.: Data-intensive applications, challenges, techniques and technologies: a survey on Big Data. Inf. Sci. **275**, 314–347 (2014)

34. Kelly, J.: Apache Drill Brings SQL-Like, Ad Hoc Query Capabilities to Big Data (2013). http://wikibon.org/wiki/v/Apache_Drill_Brings_SQL-Like,_Ad_Hoc_Query_Capabilities_to_Big_Data
35. Huang, T., Lan, L., Fang, X., An, P., Min, J., Wang, F.: Promises and challenges of big data computing in health sciences. Big Data Res. **2**, 2–11 (2015)
36. Castella: Introducción a Hadoop y su ecosistema (2013). http://www.ticout.com/blog/2013/04/02/introduccion-a-hadoop-y-su-ecosistema/
37. A. S. Foundation: Spark 0.8.0: This document gives a short overview of how Spark runs on clusters, to make it easier to understand the components involved. (2014). https://spark.apache.org/docs/0.8.0/cluster-overview.html
38. I. d. i. d. Conocimiento: 7 Herramientas Big Data para tu empresa (2016). http://www.iic.uam.es/innovacion/herramientas-big-data-para-empresa/
39. Acharjya, D.P., Dehuri, S., Sanyal, S. (eds.): Computational Intelligence for Big Data Analysis (2015)