Survey on Trends in Big Data: Data Management, Integration and Cloud Computing Environment

Washington Henrique Carvalho Almeida[1], Luciano de Aguiar Monteiro[1], Anderson Cavalcanti de Lima[1], Raphael Rodrigues Hazin[1] and Fernando Escobar[2]
[1]Recife Center for Advanced Studies and Systems
Recife, Brazil
[2]PMI-DF
Brasília, Brazil
E-mail: {washington.hc.almeida, lucianoaguiarthe, andclima, raphaelhazin}@gmail.com
Email: fernando.escobar@pmidf.org

***Abstract* — The evolution of the processing power of computers has increased the applicability of Big Data, reinforced by the advent of Internet of Things and Industry 4.0. This article conducts a literature review in order to address aspects of Big Data related to Data Management, Integration, Processing, and Cloud Computing Environment. This paper presents a perspective on the major conceptual foundations of this technology in cloud environments. Also, the survey presents trends and concerns related to this subject.**

Keywords- Big Data; Cloud; Architecture; Hadoop.

INTRODUCTION

This paper discusses trends in Big Data, challenges and opportunities. The motivation for this work is to identify trends through a survey in the most recent publications on the subject, as well as challenges and opportunities.

Recently, there has been increased interest in Big Data, mainly driven by a widespread number of research problems strongly related to real-life applications and systems, such as representation, modeling, processing, querying and mining massive, distributed, large-scale repositories. The term 'Big Data' identifies specific kinds of data sets, mainly of unstructured data, which populate the data layer of scientific computing applications [1].

Big Data can be understood as "datasets whose sizes are beyond the ability of typical database software tools to capture, store, manage, and analyze" [2]. Also, the term is generally used to describe the collection, processing, analysis and visualization associated with very large data sets. Although it is difficult to define Big Data, it can be described in terms of the data characteristics of Big Data (the 'what' of Big Data); the architectures and processing for Big Data (the 'how' of Big Data); and the applications of Big Data (the 'why' of Big Data) [3]. Figure 1 illustrates this approach.

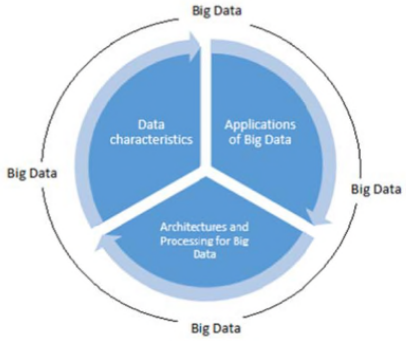

Figure 1. The components of a Big Data Definition [3].

'Big Data' is used mostly as an umbrella term to cover a range of data, technologies, and applications. This contrasts with previous data management approaches, which are typically based on data models that define the structure and operations on a database and specify elements, such as data structures and data operators [3].

The process of collecting and organizing raw data to discover patterns and draw conclusions about the information is called data analytics. It differs from data mining in three aspects – scope, purpose and focus of analysis. Data mining sorts through Big Data to identify patterns that are undiscovered and to identify hidden relationships, whereas data analysis focuses on the conclusion and process of deriving it based only on information already known by the researcher. Organizations can better understand the content of data and help them to identify the data, which will be useful for future scope in business [4].

The approach known as the 3V's (Volume, Velocity, Variety) is widely used, particularly in the practitioner and technical literature. Volume, Velocity and Variety are not by themselves regarded as sufficient to define Big Data and these terms also require definition. 'Volume', for example, is understood differently in different contexts. The 3Vs approach focuses on the characteristics of data and does not consider the wider Big Data environment [3][5].

Big Data is characterized by what is often referred to as a multi-V model. As depicted, Variety represents the data types, Velocity refers to the rate at which the data is

 ISBN: 978-1-61208-701-6

produced and processed and Volume defines the amount of data. In addition, expanding the multi-V model, Veracity refers to how much the data can be trusted, given the reliability of its source, whereas Value corresponds to the monetary worth that a company can derive from employing Big Data computing [6]. Below, we summarize the definitions of the 5V's:

Variety - Data types
Velocity - Data production and processing speed
Volume - Data size
Veracity - Data reliability and trust
Value - Worth derived from exploiting Big Data

Figure 2 shows the information applied to travel and transportation companies, but it can be used to exemplify data complexity; it illustrates the multi-V model.

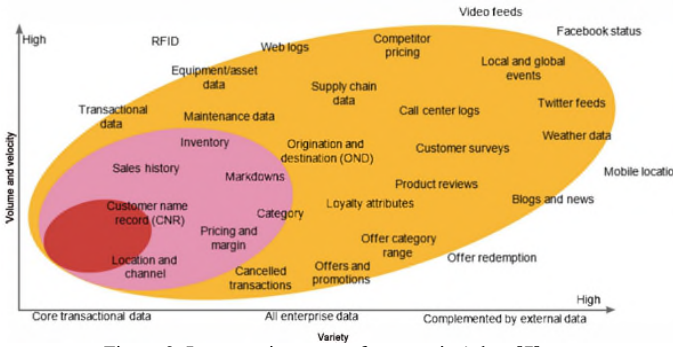

Figure 2. Large variety type of companies' data [7].

With Big Data, it is evident that many of the challenges of cloud analytics concern data management, integration, and processing [6]. The overview of the analytics workflow for Big Data is presented in Figure 3.

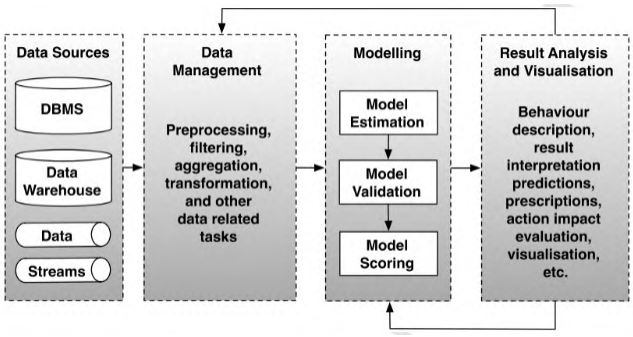

Figure 3. Overview of the analytics workflow for Big Data [6].

As shown in Figure 3, 'Data Sources' represents the Variety of sources and types. 'Data Management' describes the transformation process, because the large Volume can demand it; after that, on 'Modelling', the processed data is used to train a model and to estimate the parameters; finally, on 'Results Analysis and Visualization', the results are analyzed and evaluated, generating Value [6].

Regarding researches related to Big Data, in the first stage, they were primarily focused on the technology requirements that companies needed in order to correctly process the huge amount of data [8]. To extract knowledge from Big Data, various models, programs, software, hardware, and technologies have been designed and proposed. They tried to ensure more accurate and reliable results for Big Data applications. However, in such environment, it may be time-consuming and challenging to choose among numerous technologies [9].

The trends in Big Data are focused on the challenges for the popularization of its use in corporations, as well as in the predictive analysis. The future of the technology is being called Big Data 3.0.

The main contributors of Big Data 3.0 are the Internet of Things (IoT) and applications that generate data in the form of images, audio, and video. The IoT refers to a technology environment in which devices and sensors have unique identifiers with the ability to share data and collaborate over the Internet even without human intervention. With the rapid growth of the IoT, connected devices and sensors will surpass social media and e-commerce websites as the primary sources of Big Data [10].

Also, the fourth revolution, Industry 4.0, is mainly based on the IoT, Cyber-Physical-Systems (CPS), Internet of Services (IoS), Internet of People (IoP), and Internet of Energy (IoE) [11]. Big Data will integrate all technologies.

In this scenario, a big concern is related to human resources, especially data scientists. As the need to manipulate unstructured data, such as text, video, and images increases rapidly, the need for more competent data scientists grows. According to Kearney's survey [10] of 430 senior executives, despite the prediction that firms will need 33% more Big Data specialists over the next 5 years, roughly 66% of firms with advanced analytics capabilities were not able to obtain enough employees to deliver insights into their Big Data [10].

Challenges and opportunities will be addressed in this article and in the following sections: Section II introduces the data management. Section III presents integration and Section IV shows the processes adopted in solutions to Big Data. Also, Section V describes the cloud environment and Section VI lists the standards and solutions; finally, Section VII presents our conclusion.

II. Data Management

Data management is one of the great challenges of this approach. Trends for the future are related to the volume of data growing exponentially bringing a series of advantages and, at the same time, handicaps.

In this aspect, data complexity is a fundamental problem related to how to formulate or quantitatively describe the essential characteristics of the complexity of Big Data. The study on complexity theory of Big Data will help understand essential characteristics and formation of complex patterns in Big Data, simplify its representation, get better knowledge abstraction, and guide the design of computing models and algorithms on Big Data. To do so,

 ISBN: 978-1-61208-701-6

we will need to establish the theory and models of data distribution under multi-modal interrelationships. We will also need to sort out intrinsic connections between data complexity and spatio-temporal computational complexity [12].

Over time, key challenges were related to storage, transportation, and processing of high throughput data. This is different from Big Data challenges to which we have to add ambiguity, uncertainty and variety. Consequently, these requirements imply an additional step where data is cleaned, tagged, classified and formatted [2].

Also, social media and streaming sensors generate massive amounts of data that need to be processed. Few firms would be able to invest in data storage for all Big Data collected from their sources [10]. For example, in recent years, cloud computing has rapidly evolved from a vague concept at the beginning to a mature technology. Many big companies, including Google, Microsoft, Amazon, Facebook, Alibaba, Baidu, Tencent, and other Information Technology (IT) giants, are working on cloud computing technologies and cloud-based computing services. Big Data and Cloud Computing are seen as two sides of the same coin: Big Data is a killer application of Cloud Computing, whereas Cloud Computing provides the IT infrastructure to Big Data. The tightly coupled Big Data and Cloud Computing nexus are expected to change the ecosystem of the Internet, and even affect the pattern of the entire information industry [13].

Focusing on proper security aspects of Big Data, query processing and, in particular, on confidentiality of data during such a critical task are the major concern. Here, the main problem consists in defining models which protect cloud private data via the widely-accepted solutions: (i) encryption approaches and (ii) trust computing. Both aim at trading-off strong confidentiality and high efficiency of query processing over Big Data [14].

In this paper, we will cover the following Big Data main elements:

A. Data Quality

Data quality refers to the fitness of data with respect to a specific purpose of usage. Data quality is critical to confidence in decision making. As data are more unstructured and collected from a wider array of sources, the quality of data tends to decline. For firms adopting data analytics for their supply chain, data quality is paramount. If the data are not of high quality, managers will not use the data, let alone want to share the data with their partners. Streaming analytics use data generated by interconnected sensors and communication devices [10].

B. Data Security

Weak security creates user resistance to the adoption of Big Data. It also leads to financial loss and damage to a firm's reputation. Without installing proper security mechanisms, confidential information could be transmitted inadvertently to unintended parties. Also, cloud infrastructure has become an appealing target for cyber attackers [15]. Blockchain, an underlying technology behind the Bitcoin cryptocurrency, is a promising future technology for Big Data security management [10].

C. Data Storage

Service Oriented Architecture (SOA) and virtualization altered the whole paradigm of Information and Communication Technology (ICT) resources management from traditional computing to Cloud Computing. Storage, computing power, infrastructures, platforms, and software are provided as a service in the form of Pay-as-you-go on demand usage. Infrastructure as a Service (IaaS), Platform as a Service (PaaS), and Software as a Service (SaaS) are the three main service models of Cloud Computing. IoT Cloud Computing architecture plays a tremendous role in IoT data. IoT data and applications are stored in the cloud to make it accessible from anywhere with any web browser or client software [11][16].

These Bid Data's main elements are fundamental to the organized implementation of Big Data solutions and have been adopted; hence they were considered in this survey.

Information, privacy and security are the most concerning issues for Cloud Computing due to its open environment with very limited user-side control posing great challenges. Especially on cloud-based platforms, where there are two important aspects of Big Data security. One is how to protect data. The other is how to use Big Data analytic techniques to enhance security of the whole system. Current work on Big Data focuses on information processing, such as data mining and analysis [17].

Big Data are valuable because they are a treasured source of knowledge that turns to be useful for decision making and prediction purposes. Analytics are exploited to this end, but they expose the underlying knowledge discovery process to challenging research issues, due to the fact that analytics process huge volumes of (big) data; hence privacy of target data sets is not preserved [18].

Another relevant data management context for Big Data research is represented by the issue of effective and efficiently supporting analytics over Big Data, a collection of models, algorithms and techniques oriented to extract useful knowledge from cloud-based Big Data repositories for decision making and prediction purposes, e.g., by means of multidimensional data analysis paradigms [14].

III. INTEGRATION

Data integration is the cornerstone of Big Data, due to distributed and heterogeneous data sources and data types to provide data discover and predictive analysis. The big difference is that the software must go to the data rather than the traditional approach to data warehouse.

Concurrent with the success of the regional integration of computers and advances in fixed computers everywhere, smartphones have gained a significant contract rate capacity and resources, particularly movement and awareness related to a sensor's unique location-based services and multimedia data. The data generated through heterogeneous resources

are unstructured and cannot be stored in traditional databases [19].

Alternatives appear, like the Data Vault, a persistent staging area, which advocates for less structured repositories. This trend reaches the extreme in the form of Data Lake as a repository where raw data is stored in waiting for an analytical resolution [20].

For analysis of Big Data, database integration and cleaning are much harder than the traditional mining approaches. Manipulation of large datasets possesses problems of computational speed and error recovery. In this survey, the issue of speed has been addressed by distributing the computation over several nodes each of which works in parallel on a subset of the complete dataset and maintains coherence for producing appropriate result [21].

Through effective integration and accurate analysis on multisource heterogeneous Big Data, better predictions of future trends of events can be achieved. It is possible for Big Data analysis to even promote sustainable developments of society and economy and further give birth to new industries related to data services [13].

To overcome the delays in storing Big Data on distributed cloud storage, public storage is used as a solution. However, using public storage will make the data vulnerable to transmission and storage. Therefore, there is a need for a security algorithms to provide tradeoffs during time delay, security strength, and storage risks with encryption techniques based on flexible key [22]. The data which is being stored onto the cloud is most likely to be infiltrated with. This data can be best protected using encryption [23].

IV. PROCESSING / ANALYTICS

Big Data analytics are used in many areas, such as machine learning, computer vision, Web statistics, medical applications, Deoxyribonucleic Acid (DNA) analysis, data classification and clustering [7].

Related to this, *data validation is a major concern and is applied to define* data validity, data completeness and data consistency, as well as to validate if data are trustworthy, accurate, and meaningful. It has been reported [24] that more than half of the time spent on Big Data projects goes towards data cleansing and preparation. This section discusses the validation process for Big Data. Data collection, data cleaning, data transformation, data loading and results report are the necessary data validation processes [24].

Further, in visual analytic applications for Big Data, it is often the case that one or more models are used to calculate or to transform data prior to visualizing the results [25]. Algorithmic intelligence has gained popularity along with the rise of Big Data and current advancements in technology and organizations are increasingly able to rely on such intelligence to analyze Big Data [26].

However, the adoption of Big Data technologies is complex. The deployment and setup of an implementation of Big Data solution are time-consuming, expensive, and resource-intensive. Companies need tools and methodologies to accelerate the deployment of Big Data analytics. For this reason, Cloud Computing is becoming a mainstream solution to provide large clusters on a pay-per-use basis [27].

V. CLOUD ENVIRONMENT

As stated, Big Data is intrinsically linked to Cloud Computing; hence, its expansion will require the adoption of cloud environments due to the various aspects presented in this work. Several privacy and security discussions are covered when talking about the cloud environment, but the big-time trend is the adoption of this kind of solution. Figure 4 shows the exponential growth in digital data during the current decade.

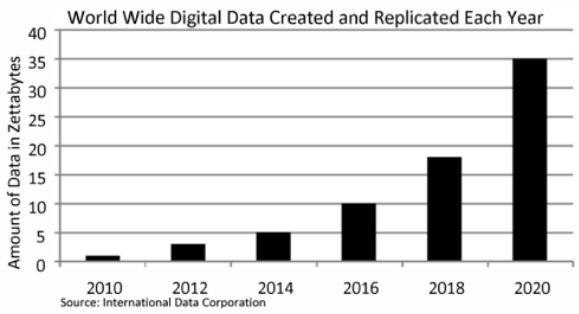

Figure 4. Exponential growth in digital data during current decade [7].

Cloud Computing provides an interesting model for analytics, where solutions can be hosted on the Cloud and consumed by customers in a pay-as-you-go fashion. For this delivery model to become reality; however, several technical issues must be addressed, such as data management, tuning of models, privacy, data quality and data currency [6].

Reinforcing the approach, cloud appears as hot topic to Big Data, which explains the importance of their 'relationship'. Table I presents a research [28] of more important topics in Big Data, where (A) represents the abstracts found, (K) for keywords, and (T) for title of the papers analyzed.

TABLE I. NUMBER OF PAPERS DEDICATED TO BIG DATA AND TO ONE OF HOT TOPICS [28]

Numbers of Papers Dimensions	with chains in A, K or T	with chains in K or T	with chains in T
Cloud	154	66	25
Analytics	136	49	36
Social	99	33	11
Mobility	78	15	5
Internet of things	37	15	4

In order to setup their Big Data on cloud environment, organizations would need to setup a master cloud to achieve the performance against the Big Data requests. All incoming client requests are submitted to the master cloud, which analyses the Big Data request size and detects the availability of suitable slave clouds (private/public),

 ISBN: 978-1-61208-701-6

according to the priority set on the cloud table. Master cloud diverts incoming request intelligently to those slave clouds, which contain big size clusters so that it takes less time to fulfill their computational needs [29].

Related to security issues, Architectural Security includes various parameters like distributed nodes, shared data, data ownership, inter-node communication, etc. The security measures elaborated in are related to the architecture of Big Data. Mostly, the security and privacy concern of Big Data arises due to its distributed file system and large volumetric data. The capabilities of the architecture need to use the data generated for mapping [23]. Moreover, we can classify the security issues as a hierarchy of security weaknesses with challenges on the present Cloud Computing models, specifically deployment and service models [30].

There are many solutions for Big Data related to Cloud Computing. Depending on the level of volume, velocity, and variety, it is important to choose appropriate Big Data tools. Thanks to the cloud, the tendency is towards Big Data as a Service and Analytics as a Service. Thus, customer and provider's staff are much more involved in the loop [28].

VI. BIG DATA STANDARDS AND SOLUTIONS

Several solutions and standards were found in this research. In this section, we present fundamental elements that have become trends.

The pioneer in managing Big Data was Google. In order to be able to store up to petabytes, they moved away from Relational Database (RDBMSs) and created a distributed file system that could scale to thousands of machines [5][20].

Recently, Big Data platforms are supported by several processing analytical tools as well as dynamic visualization [31].

Highly Archived Distributed Object Oriented Programming (HADOOP) [24][31][32][33] was created by Cutting and Cafarella, in 2005, for supporting a distributed search Engine Project. It is an Open source Java Framework technology that helps to store, access, and gain large resources from Big Data in a distributed fashion at lower cost, high degree of fault tolerance and high scalability [34].

A key component of HADOOP is the Hadoop Distributed File System (HDFS), which is used to store all input and output data for applications [4][31].

HADOOP Architecture [15][35][36] should be implemented within the slave clouds registered in the existing stack. HADOOP is a map/reduce framework that works on HDFS, which provides high throughput access to application data and has the capability to store large data across thousands of servers. In the context of the HADOOP architecture, a job is split into smaller identical tasks that can be executed closer to the data node in two phases. In map phase each task is distributed and parallelized. After map phase, all intermediate results are combined into one result, which is called reduced phase [29].

MapReduce [2][5][31][37] is a framework for writing applications that can handle large volumes of structured and unstructured data in parallel on a cluster of thousands machines, reliable and fault tolerant. The distribution of data across multiple servers allows parallelized processing of multiple tasks, each bearing on separate pieces of files. The Map function performs a specific operation on each item. The Reduce transaction combines the elements according to a particular algorithm, and provides the result [38].

Data Nodes are responsible for storing the blocks of files as determined by the Name Node. Data file to be stored is first split into one or more blocks internally. Data Nodes serve the read/write requests from file system's client data. These are also responsible for creating, deleting and replicating blocks of file after being instructed by the Name Node [39].

Hive [31][32][40] is a data warehousing solution built on top of HADOOP. It provides SQL-like query language named HiveQL. The Apache Mahout free machine learning library's goal is to build scalable machine learning tools and data mining framework for use on analyzing Big Data on a distributed manner [41].

Apache Spark [9][31][42] is an open source distributed processing framework that was created at the UC Berkeley AMPLab. Spark is like HADOOP, but it is based on in-memory system to improve performance. It is a recognized analytics platform that ensures fast, easy-to-use and flexible computing. Spark handles complex analysis on large data sets. Indeed, Spark runs programs up to 100x faster than Hive and Apache Hadoop via MapReduce in memory system. Spark is based on the Apache Hive codebase. In order to improve system performance, Spark swaps out the physical execution engine of Hive. In addition, Spark offers APIs to support a fast application development in various languages, including Java, Python, and Scala. Spark [40][43][44] is able to work with all files storage systems that are supported by HADOOP [9].

Supported Database. All of these selected tools support different types of databases, including both relation databases and non-relational databases. The most popular supported relational databases include MySQL, DB2, Oracle, PostgreSQL, Vertica and Teradata. The commonly used non-relational databases include Hive and Hbase [9]. In addition, Datameer also supports Windows Azure Blob Storage and Amazon Redshift [24]. In order to remove the scalability limit of index searching and have a fast searching speed simultaneously, HBase, the Hadoop NoSQL database, is often exploited to store chunk index table in current Hadoop-based deduplication system [45].

Supported File Format. All of the listed tools have a wide range support for different types of data files formats. The commonly supported file formats include: CSV/TSV, TXT Files, Fixed Width Text, HTML, and Server Log File [24].

VII. CONCLUSION

This survey presents elements of Big Data in the Cloud Computing environment. In the search carried out, references were searched for the construction of a framework for better understanding Big Data trends.

The models proposed increasingly based on this concept were found in the most recent research on the subject. The

main issues run into problems faced in the Cloud Computing environment, such as security and privacy.

The standards adopted have been found in many studies with the implementation of increasingly flexible solutions. Regardless of the technology used, this work presents the main features of Big Data and some barriers to the use of the resources.

A more detailed analysis of privacy problems can be made due to the exposure of the Bid Data in cloud, since only the use of the researched model does not guarantee the implementation of this architecture reliably. Some issues that we can also highlight is the investigation of new patterns that may arise, since technological changes have been increasingly fast and can be investigated and analyzed for the discovery of new techniques.

Trends identified in this paper include the growing contribution of the Internet of Things adoption, promoting exponential increasing in the volume of data analyzed as a data source for Big Data solutions. This trend also highlights the growth in the demand for qualified professionals for the data scientist profile. In the context of data management, we verified in the study as a trend the need for the definition of a data distribution model that allows a multi-modal interrelationship, with the adoption of less structured repositories.

In the processing and analysis of Big Data, it was also shown as a tendency the adoption algorithmic intelligence to create models to calculate and transform data in order to facilitate the visualization of the analysis results. In this sense, the use of distributed file systems to facilitate access and manipulation of the large volume of data analyzed by Big Data solutions was also a trend. Finally, a trend strongly identified in the study was the use of Cloud Computing environments for the deployment of Big Data solutions, allowing the use of large solutions in the form of clusters that promote significant gains in the treatment of the data maintained by the solution.

Many problems will undoubtedly arise, as for monitoring, deployment, elastic scheduling and runtime adaptation when the architecture of solutions based on Big Data will replace the data warehouse solutions and, in this context, their adoption can become complementary, since in this research a complete study on the post-migration reality was not closed.

REFERENCES

[1] A. Cuzzocrea, D. Saccà, and J. D. Ullman, “Big Data: A Research Agenda,” Proc. 17th Int. Database Eng. Appl. Symp. - IDEAS ’13, pp. 198–203, 2013.

[2] C. Kacfah Emani, N. Cullot, and C. Nicolle, “Understandable Big Data: A survey,” Comput. Sci. Rev., vol. 17, pp. 70–81, 2015.

[3] E. Dumbill, “Defining Big Data,” Forbes, 2014.

[4] S. Garion, “Big Data Analytics Hadoop and Spark,” pp. 1–55, 2016.

[5] U. Kazemi, “A Survey of Big Data : Challenges and Specifications,” no. May, 2018.

[6] M. D. Assunção, R. N. Calheiros, S. Bianchi, M. A. S. Netto, and R. Buyya, “Big Data computing and clouds: Trends and future directions,” J. Parallel Distrib. Comput., vol. 79–80, pp. 3–15, 2015.

[7] A. Ben Ayed, M. Ben Halima, and A. M. Alimi, “Big data analytics for logistics and transportation,” 2015 4th IEEE Int. Conf. Adv. Logist. Transp. IEEE ICALT 2015, pp. 311–316, 2015.

[8] J. F. Aldana, “Big Data. New approaches of modelling and management,” Comput. Stand. Interfaces, vol. 54, pp. 61–63, 2017.

[9] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, “Big Data technologies: A survey,” J. King Saud Univ. - Comput. Inf. Sci., 2017.

[10] I. Lee, “Big data: Dimensions, evolution, impacts, and challenges,” Bus. Horiz., vol. 60, no. 3, pp. 293–303, 2017.

[11] R. M. Ward, R. Schmieder, G. Highnam, and D. Mittelman, “Big data challenges and opportunities in high-throughput sequencing,” no. March, pp. 1–6, 2013.

[12] X. Jin, B. W. Wah, X. Cheng, and Y. Wang, “Significance and Challenges of Big Data Research,” Big Data Res., vol. 2, no. 2, pp. 59–64, 2015.

[13] S. Yu, “The Role of Big Data Analysis in New Product Development,” 2016.

[14] A. Cuzzocrea, “Warehousing and protecting big data: State-of-the-art-analysis, methodologies, future challenges,” ACM Int. Conf. Proceeding Ser., vol. 22–23–Marc, pp. 1–7, 2016.

[15] T. Y. Win, H. Tianfield, and Q. Mair, “Big Data Based Security Analytics for Protecting Virtualized Infrastructures in Cloud Computing,” IEEE Trans. Big Data, vol. 7790, no. c, p. 1, 2017.

[16] M. Dastbaz, H. Arabnia, and B. Ahgkar, “Technology for smart futures,” Technol. Smart Futur., pp. 1–363, 2017.

[17] Q. Liu, A. Srinivasan, J. Hu, and G. Wang, “Preface: Security and privacy in big data clouds,” Futur. Gener. Comput. Syst., vol. 72, pp. 206–207, 2017.

[18] A. Cuzzocrea, “Privacy and Security of Big Data : Current Challenges and Future Research Perspectives,” pp. 45–47, 2014.

[19] I. Yaqoob et al., “Big data: From beginning to future,” Int. J. Inf. Manage., vol. 36, no. 6, pp. 1231–1247, 2016.

[20] A. Abelló, “Big Data Design,” Dol. ’15 Proc. ACM Eighteenth Int. Work. Data Warehous. Ol., pp. 35–38, 2015.

[21] A. Saldhi, “Big Data Analysis Using Hadoop Cluster,” 2014 IEEE Int. Conf. Comput. Intell. Comput. Res., pp. 0–3, 2014.

[22] P. Adluru, S. S. Datla, and X. Zhang, “Hadoop eco system for big data security and privacy,” 2015 Long Isl. Syst. Appl. Technol., pp. 1–6, 2015.

[23] S. Bahulikar, “Security measures for the Big Data , Virtualization and the Cloud Infrastructure .,” pp. 0–3, 2016.

[24] C. Xie, J. Gao, and C. Tao, “Big data validation case study,” Proc. - 3rd IEEE Int. Conf. Big Data Comput. Serv. Appl. BigDataService 2017, pp. 281–286, 2017.

[25] A. Endert, S. Szymczak, D. Gunning, and J. Gersh, “Modeling in Big Data Environments,” Proc. 2014 Work. Hum. Centered Big Data Res., p. 56:56--56:58, 2014.

[26] W. A. Günther, M. H. Rezazade Mehrizi, M. Huysman, and F. Feldberg, “Debating big data: A literature review on realizing value from big data,” J. Strateg. Inf. Syst., 2017.

[27] M. Ciavotta, E. Gianniti, and D. Ardagna, “Capacity Allocation for Big Data Applications in the Cloud,” Proc. 8th ACM/SPEC Int. Conf. Perform. Eng. Companion - ICPE ’17 Companion, pp. 175–176, 2017.

[28] J. Akoka, I. Comyn-Wattiau, and N. Laoufi, “Research on Big Data – A systematic mapping study,” Comput. Stand. Interfaces, vol. 54, no. April 2016, pp. 105–115, 2017.

 ISBN: 978-1-61208-701-6

[29] M. Adnan, M. Afzal, M. Aslam, R. Jan, and A. M. Martinez-Enriquez, "Minimizing big data problems using cloud computing based on Hadoop architecture," 2014 11th Annu. High Capacit. Opt. Networks Emerging/Enabling Technol. (Photonics Energy), pp. 99–103, 2014.

[30] G. J.-W. Communication and undefined 2017, "Cloud Security Issues and Privacy," Ciitresearch.Org, pp. 499–514.

[31] A. Oussous, F. Z. Benjelloun, A. Ait Lahcen, and S. Belfkih, "Big Data technologies: A survey," J. King Saud Univ. - Comput. Inf. Sci., vol. 30, no. 4, pp. 431–448, 2018.

[32] C. Dincer, G. Akpolat, and E. Zeydan, "Mobil Operatörler Tarafından Servis Edilen Büyük Veri Uygulamalarında Güvenlik Sorunları Security Issues of Big Data Applications Served by Mobile Operators," pp. 0–3, 2017.

[33] J. Eckroth, "Teaching Future Big Data Analysts : Curriculum and Experience Report," 2017.

[34] B. Saraladevi, N. Pazhaniraja, P. V. Paul, M. S. S. Basha, and P. Dhavachelvan, "Big data and Hadoop-A study in security perspective," Procedia Comput. Sci., vol. 50, pp. 596–601, 2015.

[35] A. O'Driscoll, J. Daugelaite, and R. D. Sleator, "'Big data', Hadoop and cloud computing in genomics," J. Biomed. Inform., vol. 46, no. 5, pp. 774–781, 2013.

[36] Y. Yetis, R. G. Sara, B. A. Erol, H. Kaplan, A. Akuzum, and M. Jamshidi, "Application of Big Data Analytics via Cloud Computing," 2016 World Autom. Congr., pp. 1–5, 2016.

[37] M. M. Rathore, A. Paul, and A. Ahmad, "Big Data Analytics of Geosocial Media for Planning and Real-Time Decisions," 2017.

[38] K. Abouelmehdi, A. Beni-Hssane, H. Khaloufi, and M. Saadi, "Big data emerging issues: Hadoop security and privacy," Int. Conf. Multimed. Comput. Syst. -Proceedings, pp. 731–736, 2017.

[39] K. Singh and R. Kaur, "Hadoop: Addressing challenges of Big Data," Souvenir 2014 IEEE Int. Adv. Comput. Conf. IACC 2014, pp. 686–689, 2014.

[40] zhihan Lv, H. Song, P. Basanta-Val, A. Steed, and M. Jo, "Next-Generation Big Data Analytics: State of the Art, Challenges, and Future Research Topics," IEEE Trans. Ind. Informatics, vol. 3203, 2017.

[41] J. P. Verma, B. Patel, and A. Patel, "Big data analysis: Recommendation system with hadoop framework," Proc. - 2015 IEEE Int. Conf. Comput. Intell. Commun. Technol. CICT 2015, pp. 92–97, 2015.

[42] K. S and I. Bodrušić, "A Big Data Solution for Troubleshooting Mobile Network Performance Problems," pp. 472–477, 2017.

[43] I. Sorić, D. Dinjar, and D. Oreščanin, "Efficient Social Network Analysis in Big Data Architectures," pp. 1397–1400, 2017.

[44] J. Eickholt, "Teaching Big Data and Cloud Computing with a Physical Cluster," Proc. 2017 ACM SIGCSE Tech. Symp. Comput. Sci. Educ., pp. 177–181, 2017.

[45] Q. Liu, Y. Fu, G. Ni, and R. Hou, "Hadoop Based Scalable Cluster Deduplication for Big Data," Proc. - 2016 IEEE 36th Int. Conf. Distrib. Comput. Syst. Work. ICDCSW 2016, pp. 98–105, 2016.

 ISBN: 978-1-61208-701-6