The Era of Big Data: A Thorough Inspection in the Building Blocks of Future Generation Data Management

Zeinab Lashkaripour

Abstract— Data as one of the main assets in any organization, is generated at a constantly increasing pace from various sources of network devices such as smart appliances and embedded sensors. This high pace in device expansion and data generation indicates the dawn of Big Data (BD) era. Thus, this paper is aimed at providing an extensive knowledge on this ever increasing pool of data. Accordingly, a variety of events leading to BD and definitions given for it through the years are demonstrated and analyzed based on different factors. Furthermore, the infrastructures and architectures for storing, processing, manipulating, and analyzing such large-scale scheme-free datasets are compared with respect to criteria such as usage, performance, flexibility, scalability, and complexity. Moreover, for better understanding of BD, the related technologies named Cloud Computing (CC) and Internet of Things (IoT) and the broad sources of data generation are also presented. Finally, the challenges that rise beside all the gains are discussed and to conclude, a novel summarize of the issues in CC, IoT, and BD is also given. This paper would be of great value to those who seek to study, research, and work in this scientific field and demand a full dimensional perspective.

Index Terms— Big Data (BD), Big Data Analytic (BDA), Cloud Computing (CC), future generation, Internet of Things (IoT), Machine Learning (ML), storage infrastructure, technology.

1. INTRODUCTION

The Latin word "data" is the plural of "datum" that is the value of qualitative or quantitative variables. According to [1] mobile sensors, social media, video surveillance and rendering, smart grids, geophysical exploration, medical imaging, and gene sequencing are driving the data deluge. Based on the report presented in [2] by 2025, 1-2 ExaBytes (EBs) of video data would be uploaded per year in YouTube and for DNA sequencing, this amount would approach 1 ZettaByte (ZB) of sequence per year. Moreover, surprisingly a single flight from London's Heathrow Airport to John F. Kennedy in New York would generate about 650 TeraByte (TB) of data [3]. Last but not least, real world examples of BD in health are given in [4] of patients with different conditions. These are only part of the constant data generation referred to as BD (please refer to [5] and [6] for more information), where the sources could be humans, machines, and/or objects [7]. The term "Big Data" with the context understood today appears to have been first used in the late 1990s by John Mashey. The first academic paper was also presented in 2000, and published three years later, by Francis X. Diebolt titled as "Big Data Dynamic Factor Models for Macroeconomic Measurement and Forecasting" [8]. Data, which is expanding faster than the Moore's law [9], is just a valuable raw material that can be converted into information that eventually creates knowledge [8]. This increase in data has led to providing even better solutions in how to store, manage, and analyze the data we produce due to the fact that traditional operational abilities would not be enough. Accordingly, Machine Learning (ML) [10] has been used to manage BD, which includes Deep Learning (DL) [11], Online Learning (OL), Local Learning (LL), Transfer Learning (TL), and Ensemble Learning (EL) [10]. Variety of useful techniques are demonstrated, named Support Vector Machine (SVM), Bayesian Network (BN), Decision Tree (DT), Random Forest (RF), Ant Colony Optimization (ACO), Fuzzy Logic (FL), Convolutional Neural Network (CNN), Autoencoder (AE), Restricted Boltzmann Machine (RBM), and Deep Belief Network (DBN). Thus, the term BDAs is of great importance in order to process and extract meaningful data so that the organizations can make informed business decisions. While the amount of data is constantly increasing this would be an even harder challenge. Intel [12], in a survey of 200 Information Technology (IT) professionals about BDA, reports that the most common data type is business transactions stored in relational databases, followed by documents, email, sensor data, imaging, blogs, and social media. Furthermore, [13] and [14] have systematically demonstrated the techniques and challenges faced in BDA for healthcare and personalized medicine respectively, and [15] has indicated the BDA actualization mechanisms that are performed in different levels. This case study of PremiumCar could be of great use for practitioners in BDA. Moreover, the tensor completion algorithms in [16] are applicable for BDAs in a variety of applications such as data mining, computer vision, signal processing, and neuroscience. The structure of the paper is as follows: in the first section BD is introduced by demonstrating the milestones that have had an impact on it and also presenting and analyzing various definitions given for this concept. After that, traditional and BD specific storage infrastructures are compared with respect to applications and specifications, to provide a good insight into the existing systems. Related technologies named CC and IoT, the diverse sources of BD, and also the grand challenges in this field are discussed respectively in the next sections. Finally, after presenting a novel summarize of the issues in the mentioned related technologies, conclusion and future work are presented in the last section.

2. BD TIMELINE AND DEFINITION

This section provides the initial information required to understand the concept of BD through how it has been affected, the definitions given for it, and its variety of forms.

2.1 Timeline

Table 1 demonstrates the different milestones of the BD era from 1991 to 2019 that have had an impact on, and also prepared the way for it. In other words, this table is a brief overview of the events that directly or indirectly have led to

- *Zeinab Lashkaripour is a Instructor at the department of Computer Engineering, faculty of Engineering, Velayat University, Iranshahr, Iran. E-mail: z.lashkaripour@velayat.ac.ir*

this data era. It should be mentioned that the milestones from 1991 to 2013 are stated based on [8], and the ones after 2013 are stated by the author of the current paper.

TABLE 1
MILESTONES OF THE BD ERA

Year	Milestones
1991	• Internet or the World Wide Web (www) is born with the protocol of Hyper Text Transfer Protocol (HTTP) as a standard for transmission and sharing.
1995	• Sun releases the Java (invented in 1991) platform. Java, which is the second most popular language after C is used for middle-tier applications that record and store web traffic. • Global Positioning System (GPS) becomes fully operational. It was originally developed by Defense Advanced Research Projects Agency (DARPA) for military purposes in the early 1970s. This technology is used for determining location, navigaiton, tracking, mapping and timing.
1998	• The open-source relational database named NoSQL is developed by Carlo strozzi. After ten years there was a movement towards developing NoSQL databases in order to work with large and unstructured datasets. • Google is founded by Larry Page and Sergey Brin, who worked nearly a year on a project called BackRub.
1999	• The term " Internet of Things" (IoT) is invented by Kevin Ashton, cofounder of the Auto-ID Center at the Massachusetts Institute of Technology (MIT). IoT is an integration of several technologies and communications solutions and could be viewed as a convergence of three different visions named "Things" oriented, "Internet" oriented and "Semantic" oriented. This paradigm is related to different fields of knowledge, such as telecommunications, informatics, electronics and social science [17].
2001	• Wikipedia, the free internet encyclopedia is launched. It is the largest and most popular general reference work on the internet.
2002	• Version 1.1 of the Bluetooth is released by the Institute of Electrical and Electronics Engineers (IEEE). Bluetooth is a wireless technology that can transfer data over short distances.
2003	• The studies by IDC and EMC indicate that "the amount of data created in 2003 surpasses the amount of data created in all of human history before then." • LinkedIn, the popular social networking website for professionals, launches. In the third quarter of 2016, the site had about 467 million users [18].
2004	• Wikipedia reaches 500,000 articles in February and seven months later it tops 1 million articles. • Mark Zuckerberg and others in Cambridge, Massachusetts founded Facebook, the social networking service. According to [19] in October 2019, this social media web site had 2.41 billion users.
2005	• The Apache Hadoop project is created by Doug Cutting and Mike Caferella. • In order to manage the growth of digital information the National Science Board recommends creating a career path for "a sufficient number of high□ quality data scientists" by the NationalScience Foundation (NSF).
2007	• IPhone is released by Apple which creates a great consumer market for smartphones.
2008	• The number of devices connected to the internet exceeds the world's population.
2011	• "IBM's Watson computer scans and analyzes 4 TeraBytes (TBs) of data in seconds to defeat two human players on the television show Jeopardy!" • Work begins in UnQL which is a query language for NoSQL databases. • The available pools in the IPv4 32 bit address space have all been assigned. This shows the growing number of devices connects to the global network.
2012	• The BD Research and Development Initiative consisting of 84 programs in six departments is announced. • IDC and EMC estimate that 2.8 ZB of data will be created in 2012 and 40 ZB by the year 2020. • The job of data scientist is called "the sexiest job of the 21st century" by Harvard Business Review.
2013	• The democratization of data begins. With Wi-Fi and devices capable of using it everyone generates data at a tremendous rate.
2014	• According to Gartner, there were 3.7 billion connected "things" in use in 2014 and by the year 2020 that number will rise to 25 billion [20].
2016	• The bitcoin blockchain grew from 53 GB to 96 GB in size [21], which is the highest inter annual increase until the third quarter of 2019.
2018	• Emerging jobs such as BD specialists, data analysts and scientists [22].
2019	• The fifth generation (5G) of wireless technology for digital cellular networks that has begun wide deployment would result fast and vast amount of data transfer.

2.2 Definition

Various definitions have been given for this scientific concept and a few are presented in this section. McKinsey & Company [23], a global consulting agency announced BD as "the next frontier for innovation, competition, and productivity. Big data shall mean such datasets which could not be acquired, stored, and managed by classic database software." Moreover, National Institute of Standards and Technology (NIST) [24] states that "big data consists of extensive datasets-primarily in the characteristics of volume, variety, velocity, and/or variability-that require a scalable architecture for efficient storage, manipulation, and analysis." And also defines the BD paradigm as "the distribution of data systems across horizontally coupled, independent resources to achieve the

scalability needed for the efficient processing of extensive datasets." On the other hand, Apache Hadoop [6] defines BD as "datasets which could not be captured, managed, and processed by general computers within an acceptable scope." Finally, in [25] BD is defined as "a term describing the storage and analysis of large and or complex datasets using a series of techniques including, but not limited to: NoSQL, MapReduce and machine learning."

These definitions mention different factors and indicate that although BD is an evolution in IT but, due to the data volume and complex structure, working with it will require new technologies that the traditional methods will not support. As a result, the Relational DataBase Management Systems (RDBMSs) could not manage this high amount of distributed data, as mentioned in section 3.2. It can also be concluded that as stated in [25], the definitions surveyed here encompass at least one of the factors named size, complexity, and technology. Therefore, the factors mentioned in the above definitions are summarized by the author as indicated in Table 2.

TABLE 2
FACTORS OF THE GIVEN DEFINITIONS

Definition	Size	Complexity	Technology
McKinsey & Company [23]			✓
NIST [24]	✓	✓	✓
Apache Hadoop [6]		✓	
Ward and Barker [25]	✓	✓	✓

From another perspective, BD could also be defined as a number of characteristics known as 3Vs-7Vs [3, 6, 11, 26, 27]. Characteristics of data known as volume, variety, velocity, value, veracity, variability, and visibility (for more definitions on BD please refer to [28]). The four main factors of BD named volume, variety, velocity, and value (4Vs) are given bellow:

1. Volume: Datasets can contain billions of rows and millions of columns. Several examples were given in Section 1.
2. Variety: BD reflects the variety of data sources and structures. Sound classification [3] is a great example for this characteristic of BD.
3. Velocity: Data is generated and processed at a high speed which increases each year. For instance, consider an enterprise network where threats come in microseconds so you need a technology that can respond, keep pace, and analyze the packets to identify the emerging signatures and patterns on them as they flow across the network infrastructure [3].
4. Value: Valuable information could be extracted from pieces of data that may seem valueless individually. The examples given in [4] shows how patients conditions could be compromised if the value of the gathered information is not considered.

Furthermore, four types of data structures exist, where BD is mostly unstructured or semi-structured in nature demanding different techniques and tools for processing [1]. These four forms of data are as follows:

1. Structured: This type of data has a defined data type, format, and structure such as the data stored in traditional RDBMSs and simple spread-sheets.
2. Semi-Structured: This type of data is also known as a self-describing structure. Textual data files with a pattern that enables parsing such as Extensible Markup Language (XML). □
3. Quasi-Structured: This type of textual data can be formatted with effort, tools, and time. An example could be web click stream data that may contain inconsistencies in data values and formats.
4. Unstructured: This type of data has no pre-defined data model which can contain text documents, PDFs, images, and video. □

We have illustrated BD with the specified characteristics and structures as Fig. 1.

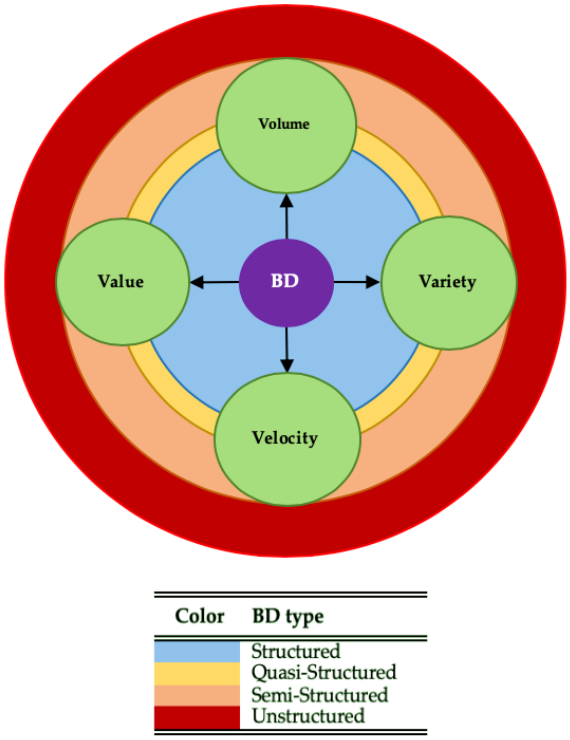

FIG 1. BD IN ONE PICTURE

3 MASSIVE DATA STORAGE AND MANAGEMENT INFRASTRUCTURES

The ability to retrieve and change the data according to your needs is known as data manipulating, which consists of selecting, deleting, inserting, and updating. Tools for performing such manipulations could be spreadsheets and RDBMSs. Spreadsheets are suitable when dealing with data that has less detail and is also low in the amount. If the data contains more than millions of rows, managing it in the spreadsheet is difficult therefore, RDBMSs become the best option. They help store, manipulate, and manage data via the usage of Structured Query Language (SQL). Despite the fact that RDBMSs are easy to use, they can only support a specific amount of structured data and more than that is not applicable. For instance, SQL Server has the maximum capacity of 10 GigaByte (GB) per database. As a result, they are not suitable for storing and managing large-scale distributed data in various formats such as multimedia and text. Therefore, special storage infrastructures are required that should be able to cope with the scale and generation speed of data. Furthermore, high efficiency, availability, and

reliability are demanded while dealing with BD that requires specific technologies. The architectures that have been developed for massive data can be categorized into three groups named Direct Attached Storage (DAS), Network Attached Storage (NAS), and Storage Area Network (SAN) [6]. For data transfer, NAS uses file level, whereas DAS and SAN use the efficient approach of block level transfer. Moreover, from the perspective of business size DAS due to its limited upgradability and expandability [6], and NAS are mainly used in small to medium size, while SAN is a network used in medium to large size businesses.

3.1 CAP theorem

Handling various aspects of BD would require horizontal scaling rather than vertical scaling. Scaling horizontally means either increasing (scale out) or decreasing (scale in) the number of nodes inside a network. Whereas, scaling vertically means either adding (scale up) or removing (scale down) resources in a single node inside a network. Therefore, scaling up would have to be done with respect to various factors such as the ones in CAP theorem [29]. In this theory, it is stated that a distributed data system can only have two out of the three factors listed below simultaneously:

1. Consistency (C): equivalent to having a single up to date version of the data.
2. Availability (A): all valid requests are always responded.
3. Partition tolerance (P): the system continues to operate although an arbitrary number of messages are dropped or delayed [30].

Brewer [29] states that Atomicity, Consistency, Isolation, and Durability (ACID) and Basically Available, Soft state, and Eventually consistent (BASE) are two design philosophies at opposite ends of the consistency-availability spectrum. ACID with focus on consistency is used for relational databases and BASE (created by Brewer and his colleagues in the late 1990's) focuses on availability. Consequently, a mixture of both is used in modern large-scale wide-area systems.

3.2 Storage Infrastructures

In this section a thorough inspection is performed on a variety of database types known as relational, NoSQL (Not only SQL that has four types demonstrated in Table 3), and NewSQL. The aim is to provide a deep understanding of the existing BD storage infrastructures so that the reader could choose the best possible solution according to his needs. A taxonomy and summarize of the existing database models is presented based on [30], in Table 3. This table would provide users with a suitable reference to choose among the options, based on their requirements. These databases have different policies in storing and retrieving data. It should be stated that based on data complexity, the databases are ordered from highest to lowest as graph, document, column, key-value and finally relational. Furthermore, based on the data size, the order would be relational, graph, document, column and key-value. Thus, relational databases have the least complexity and smallest size, graph databases have the highest complexity, and key-value databases have the biggest size. Other types of databases are located in between.

Based on the various factors mentioned in Table 3 the suitable storage infrastructures vary depending on the goal of the system. For instance, if horizontal scaling is of importance, column oriented would be a better choice since these databases have very wide horizontal scale capabilities that would make them a good option in many cases. On the other hand, if availability is of importance graph oriented is the best choice, while in case of consistency all the other type of databases could be chosen. Last but not least, if the best specifications are required, key-value has the highest performance, scalability, and flexibility with no complexity which makes it the best choice in case of efficiency factors.

4 Related Technologies and Data Sources of BD

One of the approaches for better understanding a technology is studying the technologies that are closely related to it. Therefore, BD is compared with CC and IoT in this section.

4.1 CC

CC is closely related to BD that provides the infrastructure and solutions for storing, processing and managing the data. Therefore, BD depends on CC for operations, while the dawn of BD speeds up the development of CC. There is no standard definition for CC therefore, among them all here we present the definition given by NIST [35]: "Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction."

According to this definition some characteristics [35] of CC are:

1. On-demand self-service: Ordering and managing cloud services can be done on demand without human interaction with the provider.
2. Broad network access: Accessing the services is through the internet therefore, any platform with internet connectivity either a mobile, a Personal Digital Assistant (PDA), a laptop or a Personal Computer (PC) could gain benefit from them.
3. Resource pooling: Computing resources either physical or virtual such as processor, memory and network bandwidth are pooled to serve different demands of consumers via multi-tenant model. In this information system consumer has no control or knowledge over the exact location of the sources but may be able to determine it at a higher level of abstraction like country, state, or datacenter.

TABLE 3
STORAGE INFRASTRUCTURE COMPARISON

Relational	Document oriented	Key-value	Column oriented	Graph oriented	NewSQL

Definition	Application	Example	Horizontal scaling	CAP theorem	Specifications [34]
Stores data in the form of rows. Some capacity for vertical scaling but poor capacity for horizontal scaling.	Structured data with known storing content.	Oracle MySql SQL Server SQLite	Possible via replication.	Prefers consistency over availability.	Performance: variable Scalability: variable Flexibility: low Complexity: moderate
Stores all data of a given object in a document that simplifies object mapping into a database. Retrieving could be performed via key-document lookup, document content or metadata.	When the data is semi-structured and records have relatively bounded growth. It can store all of their related properties in a single doc.	MongoDB CouchDB BigCouch Cloudant	Provided via replication, or replication and sharding.	Generally, prefers consistency over availability.	Performance: high Scalability: high Flexibility: high Complexity: low
Stores data at a unique key. This data or record is treated as a single vague collection that may have different fields which results flexibility. This type of database is very simple and very fast.	When distributed hash tables are required, schemas are very simple, or for scenarios with extreme speed.	CouchBase Redis HStore InfinityDB	Provided via sharding.	Generally, prefers consistency over availability.	Performance: high Scalability: high Flexibility: high Complexity: none
Stores data in the form of column, which is inspired by Google's BigTable paper [31]. This approach boosts performance in comparison to RDBMSs, due to data access precision.	When the data is semi-structured, you need consistency and write performance that scales past the capabilities of a single machine.	BigTable Hbase Cassandra	Very wide horizontal scale capabilities.	Prefers consistency over availability.	Performance: high Scalability: high Flexibility: moderate Complexity: low
Stores data in the form of graph that contains nodes, edges (also known as relationship), and properties. It is simple and fast for retrieving complex hierarchical structures.	When collections of objects that lack a fixed schema, and are linked together by relationships, are required. They can scale to large datasets as they do not typically need costly join operations.	Neo4j OrientDB Giraph Titan	Poor horizontal scaling so far, except for Titan.	Prefers availability over consistency.	Performance: variable Scalability: variable Flexibility: high Complexity: high
Supports the relational data model and uses SQL as the primary interface, with high performance and scalability. They are capable of high throughput online transaction processing.	When a scalable NoSQL version of a relational database with strong ACID guarantees, is required. Used in applications having a large number of transactions that (1) are short-lived, (2) touch a small subset of data using index lookups, and (3) are repetitive [32].	HStore VoltDB SQLfire Infobright ClearDB [33]	Provided via sharding.	Prefers consistency over availability.	Performance: high Scalability: high

4. Rapid elasticity: Scaling the resources up and down could be done rapidly and elastically [54] and from the consumer's point of view allocating them could be done in any quantity at any time.
5. Measured service: Monitoring, controlling and reporting resource usage provides transparency for both the provider and the consumer. This measurement is performed on pay-per-use or pay-as-you-go basis which if the consumer wanted to provide the service directly by itself it would have cost higher.
6. Service oriented [37]: Offering service in a cloud is based on the Service Level Agreement (SLA) negotiated with the customer which clarifies the responsibilities of the parties.

Despite the positive characteristics of CC, some disadvantages exist such as latency, issues related to sharing, security, and privacy. First of all, latency could be an issue due to the fact that services are remote. Second, because multiple consumers are served by a cloud, issues related to sharing can arise. For example, if the system is compromised by one consumer it can affect other consumers that share the same system. Finally, the main issue is the security and privacy of the data which is

accessible to third parties such as a service provider [38].

4.2 IoT

The BD generated by IoT has different characteristics compared to the general BD, due to the variety of types in the collected data. It is stated that by 2030 IoT would be the dominant part of BD. IoT is of great use in distinct environments such as smart cities [39], health [40], and astronomy [41]. Various definitions have also been given for IoT and here we present the one from [42]: "Competing visions agree that it relates to the integration of the physical world with the virtual world-with any object having the potential to be connected to the Internet via short-range wireless technologies, such as radio frequency identification (RFID), near field communication (NFC), or wireless sensor networks (WSNs). This merging of the physical and virtual worlds is intended to increase instrumentation, tracking, and measurement of both natural and social processes."

According to this definition some characteristics of IoT are:

1. Integration of physical and virtual world: An ever increasing number of sensors and actuators, including those embedded in smart devices have blended in different parts of our daily life.
2. Broad network access: The interconnected things have the potential to be connected to the internet via technologies like RFID, NFC, or WSN. This broad network of objects could communicate and exchange data whenever required.
3. Wide purpose and aim: This concept of interconnected devices is used with the purpose to increase instrumentation, tracking, and measurement of both natural and social processes that facilitate various aspects in our life and the surrounding environment.

Although IoT has simplified our lives in a variety of applications, it has some issues that due to the nature of its environment with numerous limited resource devices, they are of great importance. As a result, one key problem would be how to resolve the contradiction between such a large scale, heterogeneous and dynamic network, and the necessity of efficient data exchange [43].

4.3 Data Sources

Each day a huge amount of data is generated over the internet and this vast amount of data or BD has a variety of data sources that the main ones are listed below:

1. Social media: It is known as newspapers of the younger generation [44]. In this data source, users share or exchange information via virtual communication, such as LinkedIn [45], Twitter, Facebook, and Instagram. In a large-scale where billions of users are socializing, the amount of generated data would be significant.
2. Internet of Things (IoT) and Sensing: The devices in an IoT network provide services that support various needs and generate a huge amount of data [45]. IoT networks contain sensors that detect events in various fields, such as traffic and congestion control [46], and farming [47]. The high number of sensors used for various purposes, would result a huge stream of data that has redundancy.
3. Genomics: The genome sequencing data is very large in amount. The European Bioinformatics Institute (EBI) in Hinxton, UK, part of the European Molecular Biology Laboratory and one of the world's largest biology-data repositories, currently stores 20 PetaBytes (PBs) of data and backups related to genes, proteins and small molecules. BGI (formerly the Beijing Genomics Institute) in Shenzen, China, is one of the largest producers of genomic data in the world. It generates 6 TBs of genomic data from people, plants, animals and microbes each day [5]. This genomic data can be useful for human health and the future of our planet via identifying patterns of normality and anomaly.
4. Neuroscience: Many important diseases such as Alzheimer have been shown to be related to brain connectivity networks. Therefore, it would require producing massive amounts of high resolution brain image that would lead to a great deal of data [48].
5. Economics and Finance: Analyzing the huge amount of financial data in order to reduce the organization risks, would require suitable infrastructure and professionals. Due to the importance of financial issues one of the main sources of BD is related to economics [48].
6. Astronomy: This scientific field is an essential link for understanding the origin of life on Earth and its possible emergence on planets orbiting other stars that could result 20 billion rows of data, or even the data flow of 5–10 TB each night through photographing the sky via Large Synoptic Survey Telescope (LSST) [41]. It is stated that by 2025, 25 ZBs of data would be acquired annually [2].
7. Smart cities: The vast services given in a city could also be another source of high amount of data generation. Services such as education, transportation, manufacturing, and migration flows. All the above mentioned groups are a subset of the services that a smart city would provide to facilitate life, therefore their data is also a subset of the data that such a city would produce [49].

5. Grand Challenges of BD

Each novel concept aside from having a distinct number of benefits, has its own shortcomings. Thus, a variety of challenges we face in the era of BD are discussed in this section. We could refer to data as the new oil with all of the same challenges: it is plentiful but at the same time difficult and sometimes messy to extract [8]. Some of these issues are caused by the characteristics of BD (data complexity), some by its current analysis methods (system complexity), and finally, some by the limitations of current processing systems (computational complexity) [50].

5.1 Confidentiality, Integrity and Availability (CIA)

When it comes to data, CIA [45] has always been an issue and the characteristics of BD makes it an even harder challenge. Due to the limited resources, data owners rely on service providers to maintain and analyze the huge amount of data. Therefore, having a third party manage the data is a potential risk that could affect its confidentiality. On the other hand, it is necessary to have data modified by authorized

users only, to preserve its integrity. Furthermore, resources should only be available for authorized users on demand. This is also challenging when there is a huge amount of data and a high number of users.

5.2 Security and Privacy

Security and privacy are one of the main issues in any field for individuals, organizations and governments. In BD environments, weak security creates user resistance to its adoption and also leads to financial loss and damage to a corporations' reputation [51]. Furthermore, as stated in [52], managing privacy effectively is both a technical and a sociological problem that demands careful considerations from both perspectives. In the context of BD, where we deal with sensitive and private data, maintaining security and privacy is cumbersome for several reasons. Firstly, the extreme size of BD channels the protection approaches. Secondly, it also leads to much heavier security workload. Finally, the threats from the distributed networks used for processing BD can also aggravate the issues [53]. For instance, in the fields of medical science [4], which includes gathering data, discrimination, and sharing of private information with those whom the individual does not want, privacy is a significant issue. Thus, part of the security and privacy related issues mentioned in [54] are merging and integrating of access control policies, authorization management, valid usage of data, and data ownership. Therefore, establishing strong security management protocol, along with intrusion prevention and detection systems, encryptions, firewalls built into BD systems [51], and ML as used in [44] to deal with anomaly detection, could be used for security challenge alleviation.

5.3 Analytics

Processing BD contains four components as indicated in [46]: acquisition (data capture), access (data indexing, storage, sharing and archiving), analytics (data analysis and manipulation), and application (data publication). Being able to analyze the large-scale distributed data with appropriate analysis methods and tools is critical. Therefore, step three among the mentioned steps is a more challenging issue that [55] has divided it into three groups named descriptive, predictive, and prescriptive, which help in understanding, anticipating, and responding respectively. However, existing algorithms are inefficient in case of BD and in order to make sense out of the huge amount of meaningless stored data for further use, analytic techniques are required. Data Mining (DM), ML, and statistical techniques [45] are used for this purpose. DM is a computing process for identifying patterns of large datasets. ML according to Arthur Samuel in 1959, is when computers have the ability to learn. And last but not least, statistical techniques contain mathematical formulas, and models for analyzing raw data. It could be argued that statistical techniques are a subset of ML techniques and ML techniques are a subset of DM techniques. Aside from the nature of BD, the speed in which data is generated or even its format, increases the issues in this context [45].

5.4 Representation

The complexity in type and structure of data could lead to complexity in understanding, computing, and analyzing. Thus, data representation which means presenting data in a form that could be interpreted, is of great importance. Representation is a challenge and improper form of it reduces the value of the original data [6]. The most commonly used representation techniques are spatial layout, abstract or summary, and interactive or real-time, where spatial layout simplifies human interpretation (tree maps and arc diagrams), abstract or summary representation is used for summarizing large-scale data before rendering it (data cubes, and histogram binning), and the last group of techniques have to adapt to user interactions in real-time (Microsoft pivot viewer and Tableau) [45]. Other known techniques for representation as mentioned in [44] are heat maps, scatter plots, parallel coordinates, and node-link graphs. For instance, eBay with hundreds of million active users and selling billions of goods each month, generates a lot of data. EBay used Tableau to make all that data understandable and visualize the search relevance and quality to monitor customers' feedback [53].

5.5 Redundancy

With a vast number of network sensors that are the sources of data in real-time, managing such a high volume specially due to the constant repetition of previously generated data, is a great challenge. Duplication arises when at least two data samples represent the same entity [56]. Although, the existing BD processing technologies, such as Hadoop and Spark frameworks have been developed for handling data replication across multiple clusters, still they are inadequate in addressing the challenges related to data redundancy, data quality, inconsistency and cost of maintaining storage. Thus, it is essential to design a framework that is capable of addressing and minimizing redundancy issues in order to satisfy the present and future needs [44].

5.6 Compression

Traditional computation methods could not deal with the vast volume and fast changing nature of BD; therefore new approaches are needed in order to make use of it. It is a fact that redundant data exists in the pool of BD specially when the data is generated by sensors. Thus, redundancy reduction and compression although a challenge would be effective [6], which can not only reduce the storage cost, but also increase analysis efficiency and accuracy.

5.7 Scalability

Scalability is known as the ability to properly manage the ever increasing amount of data which its volume is increasing faster than CPU speeds and other computation resources [52]. Having scalable storage systems is one of the main specifications of CC environments that deal with BD. As a result, the RDBMSs are not suitable for large-scale applications, whereas NoSQL and NewSQL databases are designed to store, retrieve and manage high volume of distributed data [45] (Section 3.2). In addition, the users increasingly demand real-time streamed data which is beyond human imagination [57] and therefore adds to the significance of scalability.

5.8 Heterogeneity

Having virtually unlimited variety of data sources in a distributed BD network is known as heterogeneity. Due to this variety, data types could be structured, semi-structured, Quasi-Structured, or unstructured, which includes images, audio and video, text documents, 3D models, geographical and sensor data [57] so that efficient conversion between these formats would result more efficient values [45]. This

heterogeneity issue will be addressed through the incorporation of hybrid ML algorithms and real-time BD technologies that help clustering the incoming data into different categories, which eventually helps identifying data types easily [44].

5.9 Legal/Regulatory Issues

In order to gain benefit from BD or any other new technology, specific laws and regulations are required. Without these proper legal regulations, any technology could be used against human rights. These laws differ from one country to another or even the states of one country itself [45] that every organization needs to obey. Moreover, these laws are related to various factors such as data sharing and ownership, and as new issues arise they would need update too.

5.10 Energy Management

Managing energy consumption is vital from both economic and environmental point of views. As the amount of data increases various factors such as storage, transmission, and analysis would require more energy. Thus, managing power consumption at a reasonable level as well as maintaining quality of service is necessary [45].

Based on the studies and experience in the fields of CC, IoT, and BD, the summarize of the main related challenges are given in Table 4. The issues related to BD are mostly discussed in this paper. As it is indicated in Table 4, most of the challenges are joint issues since these fields are closely related. On the other hand, there are also some challenges that are technology specific. For instance, mobility is an IoT specific issue since its interconnected devices are capable of moving in the network, and multi tenancy is a CC specific challenge since multiple tenants can be served by a cloud and share its resources depending on the service that it provides.

6. Conclusion

The deluge of data generated each day through the expanded number of network devices such as smart phones, laptops, tablets and embedded sensors, indicates the era of BD. This trend in research and technology requires a thorough inspection in order to understand the concept and extract the demanded meaningful information from this ever increasing pool of data. Consequently, the aim of this paper is to fulfil this purpose. As a result, definitions given for BD through the years from various perspectives and diverse events that have prepared the way for it, are analyzed. Moreover, effective massive data management is not possible via relational databases, thus special technologies and storage infrastructures are required.

TABLE 4
CHALLENGES IN CC, IOT, AND BD

No	Challenge	CC	IoT	BD
1	Energy management	✓	✓	✓
2	Security and Privacy	✓	✓	✓
3	CIA	✓	✓	✓
4	Latency	✓	✓	✓
5	Performance	✓	✓	✓
6	Scalability	✓	✓	✓
7	Reliability	✓	✓	✓
8	Complexity	✓	✓	✓
9	Heterogeneity	✓	✓	✓
10	Ownership	✓	✓	✓
11	Standardization	✓	✓	✓
12	Legal regulations	✓	✓	✓
13	Data management	✓	✓	✓
14	Data analysis	✓	✓	✓
15	Data representation		✓	✓
16	Data compression		✓	✓
17	Device management	✓	✓	
18	Software/hardware architecture	✓	✓	
19	Interoperability	✓	✓	
20	Service provisioning	✓		
21	Multi tenancy	✓		
22	Mobility		✓	

Therefore, the variety of existing options are deeply analyzed so that the user could easily select the best option based on his requirements. Furthermore, the technologies that are closely connected to BD such as IoT and CC and the sources that generate data at an enormous tempo are discussed. Last but not least, a brief overview on the challenges that we face in case of BD and a novel summarize of the challenges in the three related technologies is given to sum up the issues in this field of science. This paper would be very useful for those who seek a thorough inspection on this concept in order to study, research, and work further in this field.

Since the demand of real-time streamed data from distinct, constant increasing sources, is enhancing beyond our imagination; this prominent field would constantly demand novel strategies, infrastructures, and technologies to manage such an environment as new obstacles arise. Therefore, due to the significance of bioinformatics as one of main sources of BD, as a future work we are planning to develop and deploy a sound and secure system in order to store, manage, and manipulate the sensitive data of dynamic healthcare environments.

References

[1] EMC Education Services., *Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data*, Indiana: John Wiley & Sons, pp. 3-9, 2015.

[2] Z.D. Stephens, S.Y. Lee, F. Faghri, R.H. Campbell, C. Zhai, M.J. Efron, R. Iyer, M.C. Schatz, S. Sinha, and G.E. Robinson, "Big Data: Astronomical or Genomical?," *PLoS biology*, vol. 13, no. 7, p.e1002195, 2015.

[3] P. Zikopoulos, D. Deroos, K. Parasuraman, T. Deutsch, J. Giles, and D. Corrigan, *Harness the Power of Big Data The IBM Big Data Platform*, McGraw Hill Professional, pp. 5-14, 2012.

[4] W.N. Price and I.G. Cohen, "Privacy in the Age of Medical Big Data," *Nature medicine,* vol. 25, no. 1, p. 37, 2019.

[5] V. Marx, "Biology: The Big Challenges of Big Data," *Nature*, vol. 498, no. 7453, pp. 255-260, 2013.

[6] M. Chen, S. Mao, and Y. Liu, "Big Data: A Survey," *Mobile Networks and Applications*, vol. 19, no. 2, pp. 171-209, 2014.

[7] D. Zhang, October. "Big Data Security and Privacy Protection," *In 8th International Conference on Management and Computer Science (ICMCS 2018)*. Atlantis Press, 2018

[8] J. Dean, *Big Data, Data Mining, and Machine Learning: Value Creation for Business Leaders and Practitioners*, New Jersey: John Wiley& Sons, pp. 5-8, 2014.

[9] R.L. Villars, C.W. Olofson, and M. Eastwood, "Big Data: What it is and Why You Should Care," White Paper, MA, USA, IDC, 2011.
[10] A. L'heureux, K. Grolinger, H.F. Elyamany, and M.A. Capretz, "Machine Learning With Big Data: Challenges and Approaches," *IEEE Access*, vol. 5, pp. 7776-7797, 2017.
[11] M. Mohammadi, A. Al-Fuqaha, S. Sorour, and M. Guizani, "Deep Learning For IoT Big Data and Streaming Analytics: A survey," *IEEE Communications Surveys & Tutorials*, vol. 20, no. 4, pp.2923-2960, 2018.
[12] Intel IT center, "Peer Research on Big Data Analysis," https://www.intel.com/content/dam/www/public/us/en/documents/reports/data-insights-peer-research-report.pdf. 2012.
[13] N. Mehta and A. Pandit, "Concurrence of Big Data Analytics and Healthcare: A Systematic Review," *International journal of medical informatics*, vol. 114, pp.57-65, 2018.
[14] D. Cirillo and A. Valencia, "Big Data Analytics For Personalized Medicine," *Current opinion in biotechnology*, vol. 58, pp.161-167, 2019.
[15] C. Dremel, M.M. Herterich, J. Wulf, and J. Vom Brocke, "Actualizing big data analytics affordances: A revelatory case study," *Information & Management,* vol. 57, no. 1, pp. 103121, 2020.
[16] Q. Song, H. Ge, J. Caverlee, and X. Hu, "Tensor Completion Algorithms in Big Data Analytics," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 13, no. 1, p. 6, 2019.
[17] L. Atzori, A. Iera, and G. Morabito, "The Internet of Things: A Survey," *Computer networks*, vol. 54, no. 15: pp. 2787-2805, 2010.
[18] Statista Research Department, "Numbers of LinkedIn Members From 1st Quarter 2009 to 3rd Quarter 2016 (in Millions)," https://www.statista.com/statistics/274050/quarterly-numbers-of-linkedin-members/. 2017, (accessed 24 Jan 2020).
[19] Statista Research Department, "Most Popular Social Networks Worldwide as of October 2019, Ranked by Number of Active Users (in Millions)," https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. 2019, (accessed 24 Jan 2020).
[20] Gartner, "Gartner Says 4.9 Billion Connected "Things" Will Be in Use in 2015," https://www.gartner.com/en/newsroom/press-releases/2014-11-11-gartner-says-nearly-5-billion-connected-things-will-be-in-use-in-2015. 2014, (accessed 24 Jan 2020).
[21] Statista Research Department, "Size of the Bitcoin Blockchain From 2010 to 2019, by Quarter (in Megabytes)," https://www.statista.com/statistics/647523/worldwide-bitcoin-blockchain-size/. 2019, (accessed 28 Jan 2020).
[22] World Economic Forum, "The Future of Jobs Report 2018," World Economic Forum, Geneva, Switzerland, 2018.
[23] M. James, M. Chui, B. Brown, J. Bughin, R. Dobbs, C. Roxburgh, and A.H. Byers, "Big Data: The Next Frontier For Innovation, Competition, and Productivity," 2011.
[24] W.L. Chang and N. Grady, "NIST Big Data Interoperability Framework," vol. 1, Big Data Definitions(No. Special Publication (NIST SP)-1500-1), 2015.
[25] J.S. Ward, and A. Barker, "Undefined by Data: A Survey of Big Data Definitions," arXiv preprint arXiv:1309.5821, 2013.
[26] P. Goswami and S. Madan, "A Survey on Big Data & Privacy Preserving Publishing Techniques," *Advances in Computational Sciences and Technology,* vol. 10, no. 3, pp. 395-408, 2017.
[27] J.J. Seddon, and W.L. Currie, "A Model For Unpacking Big Data Analytics in High-Frequency Trading," *Journal of Business Research*, vol. 70, pp. 300-307, 2017.
[28] P. Mikalef, I.O. Pappas, J. Krogstie and M. Giannakos, "Big Data Analytics Capabilities: A Systematic Literature Review and Research Agenda," *Information Systems and e-Business Management*, vol. 16, no. 3, pp. 5 47-578, 2018.
[29] E. Brewer, "CAP Twelve Years Later: How the 'Rules' Have Changed," *Computer*, vol. 2, pp. 23-29, 2012.
[30] P. Murthy, A. Bharadwaj, P.A. Subrahmanyam, A. Roy, and S. Rajan, "Big Data Taxonomy," Big Data Working Group, Cloud Security Alliance, 2014.
[31] F. Chang, J. Dean, S. Ghemawat, W.C. Hsieh, D.A. Wallach, M Burrows, T. Chandra, A. Fikes, and R.E. Gruber, "Bigtable: A Distributed Storage System For Structured Data," *ACM Transactions on Computer Systems (TOCS)*, vol. 26, no. 2, pp. 1-26, 2008.
[32] M. Stonebraker, S., Madden, D.J. Abadi, S. Harizopoulos, N. Hachem, and P. Helland, September. "The End of an Architectural Era: (It's Time For a Complete Rewrite)," *In Proceedings of the 33rd international conference on Very large data bases,* pp. 1150-1160, VLDB Endowment, 2007.
[33] A. Pavlo and M. Aslett, "What's Really New With NewSQL?," *ACM Sigmod Record*, vol. 45, no. 2, pp. 45-55, 2016.
[34] B. Scofield, "NoSQL–Death to Relational Databases,"CodeMash Presentation, January, 2010.
[35] P. Mell and T. Grance, "The NIST Definition of Cloud Computing," 2011.
[36] B. Grobauer, T. Walloschek, and E. Stocker, "Understanding cloud computing vulnerabilities," *IEEE Security & privacy*, vol. 9, no. 2, pp. 50-57, 2010.
[37] Q. Zhang, L. Cheng, and R. Boutaba, "Cloud Computing: State-of-the-Art and Research Challenges," *Journal of internet services and applications*, vol. 1, no. 1, pp. 7-18, 2010.
[38] R.L. Grossman, "The Case For Cloud Computing," *IT professional*, vol. 11, no. 2, pp. 23-27, 2009.
[39] Z. Allam and Z.A. Dhunny, "On Big Data, Artificial Intelligence and Smart Cities," *Cities*, vol. 89, pp. 80-91, 2019.
[40] S.K. Lakshmanaprabu, K. Shankar, M. Ilayaraja, A.W. Nasir, V. Vijayakumar, and N. Chilamkurti, "Random Forest For Big Data Classification in the Internet of Things Using Optimal Features," *International Journal of Machine Learning and Cybernetics*, pp. 1-10, 2019.
[41] E.D. Feigelson and G.J. Babu, "Big Data in Astronomy," *Significance*, vol. 9, no. 4, pp. 22-25, 2012.
[42] J. Winter, "Algorithmic Discrimination: Big Data Analytics and the Future of the Internet," *In The future internet*, (pp. 125-140). Springer, Cham, 2015.
[43] H.D. Ma, "Internet of Things: Objectives and Scientific Challenges," *Journal of Computer science and Technology*, vol. 26, no. 6, pp. 919-924, 2011.
[44] R.A.A. Habeeb, F. Nasaruddin, A. Gani, I.A.T. Hashem, E. Ahmed, and M. Imran, "Real-time Big Data Processing For Anomaly Detection: A Survey," *International Journal of Information Management*, vol. 45, pp. 289-307, 2019.
[45] I.A.T. Hashem, I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani, and S.U. Khan, "The Rise of "Big Data" on Cloud Computing: Review and Open Research Issues," *Information Systems*, vol. 47, pp. 98-115, 2015.
[46] C. Dobre, and F. Xhafa, "Intelligent sServices For Big Data Science," *Future Generation Computer Systems*, vol. 37, pp. 267-281, 2014.
[47] S. Wolfert, L. Ge, C. Verdouw, and M.J. Bogaardt, "Big Data in Smart Farming–A Review," *Agricultural Systems*, vol. 153, pp.

69-80, 2017.
[48] J. Fan, F. Han, and H. Liu, "Challenges of Big Data Analysis," *National science review*, vol. 1, no. 2, pp. 293-314, 2014.
[49] R. Iqbal, F. Doctor, B. More, S. Mahmud, and U. Yousuf, "Big Data Analytics: Computational Intelligence Techniques and Application Areas," *Technological Forecasting and Social Change*, p. 119253, 2018.
[50] X. Jin, B.W. Wah, X. Cheng, and Y. Wang, "Significance and Challenges of Big Data Research," *Big Data Research*, vol. 2, no. 2, pp. 59-64, 2015.
[51] I. Lee, "Big Data: Dimensions, Evolution, Impacts, and Challenges," *Business Horizons*, vol. 60, no. 3, pp. 293-303, 2017.
[52] H.V. Jagadish, J. Gehrke, A. Labrinidis, Y. Papakonstantinou, J.M. Patel, R. Ramakrishnan, and C. Shahabi, "Big Data and its Technical Challenges," *Communications of the ACM*, vol. 57, no. 7, pp. 86-94, 2014.
[53] C.P. Chen and C.Y. Zhang, "Data-Intensive Applications, Challenges, Techniques and Technologies: A Survey on Big Data," *Information sciences*, vol. 275, pp. 314-347, 2014.
[54] E. Bertino and E. Ferrari, "Big Data Security and Privacy," *In A Comprehensive Guide Through the Italian Database Research Over the Last 25 Years*, (pp. 425-439). Springer, Cham, 2018.
[55] 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.
[56] Zhou, L., Pan, S., Wang, J. and Vasilakos, A.V., "Machine Learning on Big Data: Opportunities and Challenges," *Neurocomputing*, vol. 237, pp. 350-361, 2017.
[57] S.M. Idrees, M.A. Alam, and P. Agarwal, "A study of Big Data and its Challenges," *International Journal of Information Technology*, vol. 11, no. 4, pp. 841-846, 2019.