Big Data, Cloud Computing, and IoT (BCI) Amalgamation Model: The Art of "Reinventing Yourself" to Analysis the World in which we live

Kareem N. Areed, Mahmoud Badawy, Amira Y. Haikal, and Mostafa A. Elhosseini

***Abstract*—The spread of omnipresent sensing technology brings with it an increasing number of innovative models. The smart mobility initiatives offer new opportunities for Intelligent Systems to maximize the utilization of real-time data that are streaming out of different sensory resources. In recent years, the convergence trend of Big Data, Cloud and IoT has received considerable attention in industry and academia. A huge amount of data is generated every day from information systems and modern digital technologies such as the Internet of things (IoT) and cloud computing. The analysis of these massive data requires a lot of effort at multiple levels to extract knowledge to facilitate decision-making. Big data analysis is therefore a topical area of research and development. The main objective of this survey is to propose Big Data, Cloud Computing, and IoT (BCI) Amalgamation Model. Additionally, this paper explores the big data characteristics, challenges, analysis techniques, and various tools associated with it. The recommendation of the suitable analysis techniques of big data that could reduce the time and increase efficiency is discussed.**

***Index Terms*— Analysis; Big Data; Decision Making; Information Retrieval; Metrics; Tools.**

I. INTRODUCTION

Big Data is defined as the collection of various organized and randomized data in a larger manner to perform a particular task and enhance various developments in the software field to manage large data storage. From the beginning of time till 2003 there were 5 billion gigabytes, the same amount was created every two days in 2011, and every ten minutes in 2013, The amount of data in 2015 are 7.9 zettabytes (10^{21} bytes, or 1,000 exabytes) and 35 zettabytes in 2020. Generally, 90% of the world's data was generated in the last few years [1]. There are main functionalities that provide different corporations to enhance quicker access and smart decision in the improvement of big data management [1]. The speed of this big data increasing leads to enhance processing or accessing speed and economy of the enterprise. As shown in figure 1, there are potential challenges to the growth of big data applications. To address these challenges, there are several fundamental technologies that are closely related to big data, including cloud computing, the Internet of Things (IoT), and Hadoop.

Cloud Computing has been established as relatively stable environments for proving a wide number of tackled solutions. However, the limitations of network bandwidth as well as the rapid expansion into the data transfer rate are still the bottlenecks. Internet of Things (IoT) is a paradigm of telecommunications, bringing with it an increasing number of Internet-connected devices, smart objects, automotive systems in a complex global infrastructure characterized by self-configuration, smart actions, and interoperable communication. IoT is seen as a promising solution for the development of domestic services and business processes.

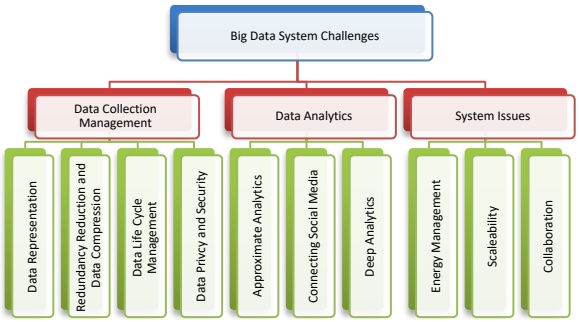

Fig 1: The Growth of Big Data Applications

Ericsson and Cisco forecast estimates the connected devices will reach approximately 50 billion devices in 2025, and the data produced by these devices will reach 500 zettabytes. 97% of organizations feel there are challenges to creating value from IoT-related data. Because of IoT resources constraints and rising consumer demands for efficient services. It has become important to integrate IoT and computing technologies such as cloud, edge, and fog computing. IoT can use cloud computing as an interface for the use of virtual resources. Thus, it could say that nowadays, efficient information dissemination has become the new production factor, notably in terms of improving network resource utilization, mitigating traffic congestion, and reducing transmission power consumption over the network. The convergence of Big data, cloud computing and the Internet of things (IoT) has difficulties to satisfy the user's complex demands.

Big Data is increasing the benefits of both industry and research, such as healthcare, commercial and finance advice [2]. The Economist says that data is becoming a new business raw material. Economic input is nearly equivalent

Published on February 14, 2020.
Kareem N. Areed, MET Academy, Misr Higher Institute of Engineering and Technology in Mansoura, Egypt.
(email: ek8819@gmail.com)
Mahmoud Badawy, Mansoura University, Egypt.
(email: badawy_mm@hotmail.com)
Amira Y. Haikal, Mansoura University, Egypt.
(email: amirayh@gmail.com)
Mostafa A. Elhosseini, Taibah University, Saudi Arabia.
(email: melhosseini@gmail.com)

to labor and capital. The data to be analyzed nowadays are dynamic and enormous in volume, as well as being grouped into different types of data. Such data come from various sources of data, such as social networks, GPS signals for mobile phones, instant messaging apps and more. Big Data, therefore, has unique characteristics such as heterogeneous, unstructured, semi-structured, incomplete, high-dimensional. According to industrial data analyst Doug Laney [3], big data is articulated in the year 2000 as the following three Vs:

- **Volume (Data in Rest):** Agencies collect data from a variety of data sources, including business transactions, social media data, and machine to machine information or sensor information.
- **Velocity (Data in Motion):** The data streams join at unrivaled speed and should be distributed accordingly. Different types of IoT sensors, RFID tags, and intelligent metering push the need to manage real-time data flows.
- **Variety (Data in Many Shapes):** Data comes in various formats for unstructured text documents, audio, video, email, stock and financial transactions, such as structured, numerical data in traditional databases.

But these V's are extended to seven V's later. The new four V's are as follow:

- **Variability (Highlighted Data):** Data set inconsistency can hamper and manage processes.
- **Veracity (Completeness of Data):** Refers to the data's messiness. The quality of the data collected will vary greatly, affecting the accuracy of the analysis.
- **Validity (Data Correctness):** Refers to the range of data correctness and accuracy.
- **Volatility (Data Using Time):** Refers to the duration that needed to store this data.

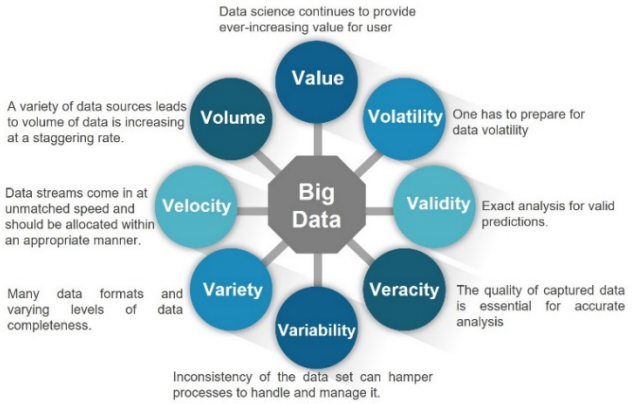

Fig 2: Big Data Characteristics

All major IT companies have already started their big data projects, including EMC, Microsoft, Google, Amazon, and Facebook, etc. Extracting information or data from big data requires optimum processing power, analytics capabilities and skills [4]. Therefore, effectively dealing with big data involves generating interest against big data length, variety, and veracity [5].

The main objectives of this paper are: (i) proposing Big Data, Cloud Computing, and IoT (BCI) amalgamation model, (ii) discuss big data analytics steps and analytics techniques and (iii) constructing a comparison between most important techniques that help on choosing the most suitable one for your analysis mission.

The rest of this paper is organized as follows: The BCI amalgamation model is introduced in Section 2. Section 3 discusses big data challenges, Section 4 will discuss what is big data analytics and what is the main four steps of the analytics process, Section 5 shows the layers of analytics infrastructure, Section 6 lists the performance metrics items, Section 7 explores nine important analytics techniques, Section 8 previews many important tools that are using in big data analytics, and Section 9 shows the challenges that force big data analysis.

II. THE BCI AMALGAMATION MODEL

The concept of a network of smart devices was discussed as early as 1982, with a modified Coke vending machine at Carnegie Mellon University. The content delivery network (CDN) idea was introduced in 1990 for the delivery of cached content such as images and videos. In 1991 the ubiquitous computing concept is introduced. Between 1993 and 1997, The field gained interest when the "Six Webs" framework introduced. The major influencer in history is Cloud computing attracted particular attention in 2006 when Amazon first promoted its "Elastic Compute Cloud". Then, the term cloudlet has appeared in a two-tier architecture. The first tier a high latency layer, and the second is lower latency.

Generally speaking, many researchers considered that the first announcement on cloud computing and IoT integration in 2009 was the integration of cloud computing with Wireless Sensor Networks (WSN), which is considered to be the core of IoT. In 2012, Cisco introduced the term fog computing for dispersed cloud infrastructures. The aim was to promote IoT scalability, i.e., to handle a huge number of IoT devices and big data volumes for real-time low-latency applications. The convergence of big data, cloud computing with either WSN or IoT today provides many useful features such as scalability, virtualization, unlimited resources, and infinite storage.

Over the past few years, the amalgamation trend of Big data, Cloud and IoT have received considerable attention in both academia and industry. The future of Internet services has centered on "the ability to build a network or a platform has the ability to deliver service effectively to a large number of users". The proposed BCI amalgamation model, as shown in figure 3, consists of six layers:

1) The perception layer includes sensors and smart devices to collect data from the IoT environment.
2) The network layer provides connecting to other servers, smart things such as sensors and network devices. The other important feature of this layer is processing and transferring the data gathered from the perception layer.
3) The fog Layer: Remote IoT scalability, i.e., to handle a huge number of IoT devices and big data volumes for real-time low-latency applications.

4) The cloud layer provides various sub-services by several private or public clouds.
5) The service composition layer is responsible for composing a number of sub-services together regarding the user's functional and non-functional requirements.
6) The application layer drives particular composited services to the end-users according to their requests.

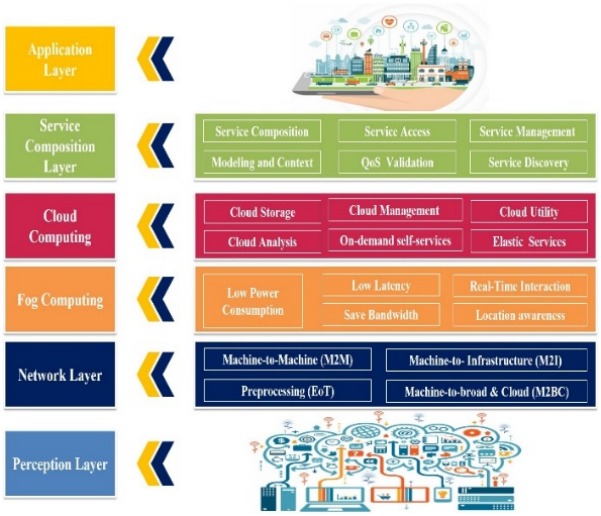

Fig 3: The BCI amalgamation model

III. BIG DATA CHALLENGES

Just move together opportunities and challenges with each other. Big data, on the one hand, offers different opportunities for society and business but, on the other, it also poses many challenges [6]. Big data issues such as storage, transportation, management and processing [7], variety and heterogeneity, scalability, speed, accuracy [8], privacy and security, data access and sharing, skill requirements, technical or hardware-related challenges, analytical challenges [9], have been identified and addressed by various researchers. The following subsections explore some of the most pertinent challenges which need immediate attention from researchers.

Fig 4: Big Data Challenges

A. The scarcity of big data professionals

Big Data processing tools and algorithms that have recently been developed include MapReduce, Hadoop, Dryad, Apache Spark, Apache Mahout, Tableau, etc. [13],[6]. But in addition to developing this high-processing big data, complex processing technologies, organizations need highly skilled professionals to handle and use these tools according to an organization's needs. There is no doubt about the big data as well, but looking at the current scenario, these naive experts should be given a special kind of training so that they are able to handle the big data from different dimensions, including data architecture, data modeling, data integration, etc.[9]. According to a McKinsey & Company report [10], the US could realize the data analysis requirement of 140,000 to 190,000 qualified individuals, as well as more than 1,000,000 analysts and managers with specialized analytical knowledge and skills to make accurate and correct decisions.

And finally, it could conclude that for business organizations engaged in big data analytics and frameworks, Professionals, namely data scientists/engineers, are in huge demand, for efficient decision-making to tackle challenges such as data architecture and management, etc. Each business enterprise needs to recruit big data analysts to succeed in a competitive market. According to a study [11], to deal with big data, each company should have a Special Data Force (SDF), with great analytical skills. Big data analysts along with business intelligence were seen as one of the factors for the company's exponential growth [10].

B. Interactiveness (or designing)

A data mining system's interactivity is the capability that enables users to effectively interact with the system [8] including guidance, suggestions, user feedback, etc. Interactivity is seen as one of the main challenges for system designers, especially in business organizations, when it comes to big data mining. Better interactivity will address the challenges surrounding the 3Vs (velocity, variety, and volume) [7]. Optimal user engagement gives users a better way to identify their area of interest from the large volume of data that also helps marketing experts to easily obtain the mining results. Because now the consumer is concerned only with his or her subspace, which increases the processing speed (velocity) as well as the system's scalability. In addition, big data's heterogeneity (or variety) adds uncertainty in the data that can also complicate the mining outcomes. Better interactivity, however, offers opportunities to visualize the effects of big data analysis and mining with ease.

It can be inferred that the mining results of a system with better interactivity would possibly result in poor or inappropriate mining results being acknowledged by potential users that lack the interactivity of the data mining system. According to authors in [12],[13] the big data organization must design the systems in such a way that they can understand both the customer's needs and the technology to be used to solve the issue. It has also been shown that due to poor contact processes, numerous valued consumers have been given the least attention by business organizations. For this purpose, Designers should be responsible for interfaces, conceptual models, designs, user feedback systems, etc. [7] in order to capture important customers and understand the needs of each individual customer from thousands of customers.

C. Synchronization and loading

Software loading includes getting software into a single data repository from various heterogeneous data sources [14]. Loading process undergoes various issues that require

close attention from researchers and practitioners, e.g. multiple data sources should be mapped into a unified structural framework, tools and infrastructure should be available that cooperate with the size and speed of big data and should transfer data in a timely manner.

In addition to these loading problems, synchronization through various data sources is also seen as one of the key challenges. Once the data is loaded from different data sources into a big data network, with a different velocity at different time intervals, it is likely to get away from synchronization. Synchronization between data sources refers to the process of maintaining continuity between the different data sources and the standard repository over time. In other words, in terms of time and sequence, data from different sources will suit each other [15]. If the big data processing system cannot guarantee consistency, inconsistent or even invalid data would likely result in poor and/or incorrect mining performance. Therefore, the integration of data sources should be given a lot of attention to help business organizations avoid the risks in the research process and draw accurate and relevant conclusions as a result. Hence, the heterogeneous nature of the data makes the processing and cleaning of businesses more difficult before they are loaded into a warehouse for analysis [16]. Hadoop and MapReduce employed by many companies offer different ways to effectively process unstructured data.

D. Visualization

Data visualization is the method of obviously representing information so as to make decisions more efficient [17]. When big data grows exponentially with unbridled velocity and huge volume, extracting the hidden information becomes extremely difficult due to the unavailability of scalable visualization software. There is no doubt that big data visualization tools are used by online marketplaces (e.g. eBay) to transform their large, complex data sets into picture formats to make all the data clearly understandable. But as the big data tsunami hits researchers at very high speed [15], in the near future these current visualization techniques are likely to be of no use. Researchers have been paying attention in the current scenario to the visualizing challenges of big data, but there is a need to be prepared to face the future challenges of big data. In other words, there is a need to have ready tools to carry and visualize the Zettabytes (1024) or even Yottabytes (1024) of data, or beyond. Since the data is generated by everyone from anywhere, such as online social networks, medical science, geostationary satellites, sensors, etc. [7], big data becomes "bigger data". So there are more chances of facing challenges in the future, so it should be armed with the new technologies and resources in advance to address obstacles that threaten us horrifyingly and uninformed. Big data visualization techniques have the task of showing the analytical findings visually, using various graphs for decision making [16]. A visual report speaks twice as much as a text report and the visualization techniques for complex customer data have also been proven sophisticated. Visualization tools such as Tableau, QlikView, etc. [18] have been used by organizations to improve the visibility of big data, even include business-specific visualizations and ensure meaningful data exploration.

Table 1 below provides a glimpse into the comprehensive conclusion of identified challenges and their associated solutions applied so far. But still, there are the shortcomings in the available approaches that need researchers ' great attention.

Table 1: Big Data Challenges

Challenge	Possible approaches	Limitations
Big Data Professionals	Establishment of special data force (SDF) with advanced analytical skills	Expensive but necessary to survive
Interactiveness	Design of systems by taking user needs and technology under consideration	User interactive designs satisfy customers, and a satisfied customer is itself an advertisement
Loading and Synchronization	Hadoop and MapReduce to load various formats of data in a distributed and synchronous manner	The heterogeneous nature of data is the reason which raised the challenge.
Visualization	Tableau, QlikView, etc.	Businesses use visualization tools to increase the throughput over big data.

IV. BIG DATA ANALYTICS

In order to develop more clearly the Knowledge Discovery in Databases (KDD), it should be inferred that the KDD method has a preprocessing, transformation, collection, data mining, and interpretation. With these operations, a complete data analytics platform can be built that gathers the data and then identifies data information and visualizes the user's knowledge.

In essence, data processing is seen as collecting, processing and managing data for the purpose of generating new information for end users [19]. Big Data analysis involves four steps: Acquisition, Assembly, Analyze and Action. These steps are named as the 4 A's.

Acquisition: Big Data architecture needs to get high-speed data from different data sources and deal with various access control protocols. It is where only data that could be beneficial or underdone with a smaller degree of uncertainty could be identified from a filter [20]. In some applications, data generation conditions are important, so capturing these metadata and storing them with the corresponding data might be useful for further study.

Assembly: At this stage, the architecture has to deal with various data formats and must be able to parse them and extract the actual information, such as named entities, their relationships, etc. [20]. This is also the point where data must be cleaned, placed into a computable format, structured or semi-structured, organized and processed in the right place. Therefore, it had to do some kind of extract, turn, and load. Successful cleaning is not guaranteed entirely in Big Data architecture. Nonetheless, Big Data's

volume, volume, variety, and variability can prevent from taking the time to thoroughly clean it all out.

Analyze: Here you will find queries running, modeling and developed algorithms to find new insights. Mining demands clean, detailed, and reliable data. At the same time, data mining can also help improve data quality and confidence, understand its semantics and provide intelligent querying functions [20].

Action: Valuable judgments must be in a position to evaluate the outcomes of the research. Consequently, the interpretation and verification of outputs are very relevant for the user [20]. In addition, the root of the data should be given to assist the user in understanding how he obtains.

V. BIG DATA ANALYTICS LAYERS

According to the big data analytics steps depicted in the previous section, it discussed the implementation of big data Layers as shown in figure 5 [23].

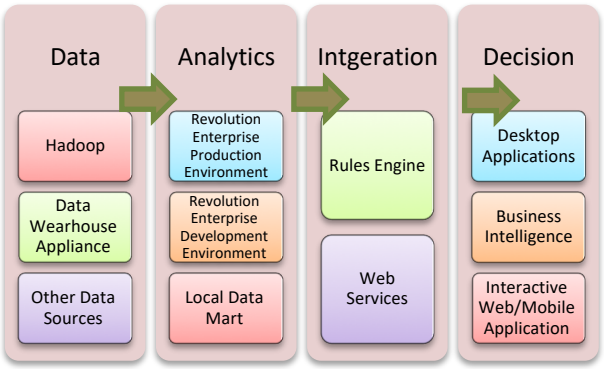

Fig 5: Big Data Analytics Implementation Layers

- **Data Layer:** This layer has structured data, semi-structured and unstructured data based on Relational Database Management System (RDBMS). NoSQL databases are used for storing unstructured information. Cassandra and MongoDB, for example, NoSQL databases. Sources of unstructured and semi-structured data include streaming data from the web world, social media area, data from IoT sensors, and operating systems. At this layer are also software tools, such as HBase, Hive, HBase, Spark, and Storm. That layer is also supported by Hadoop and MapReduce.
- **Analytics Layer:** The analytics layer has the infrastructure for applying dynamic data analytics and delivering values in real time. It has building models to create an area, and frequently modifies local data. That also increases analytical engine performance.
- **Integration Layer:** This layer integrates analytics engine and end-user applications. This usually includes a rules engine and a dynamic data analytics API.
- **Decision Layer:** This layer is where the end product meets demand. It includes end-user apps, such as desktop apps, mobile apps, interactive web apps, and software for business intelligence. It is the layer that connects people with the device in. Each and every layer mentioned above is connected in real-time to different end-user sets and makes a critical step of real-time data analytics implementation.

VI. PERFORMANCE METRICS

Big data analytics algorithms must be subject to a number of criteria that listed as following:

Fig 6: Big Data Analysis Techniques Performance Metrics

- **Processing power:** this term is associated to measuring the processing power needed to accomplish big data analysis process, when it needs to measure this term and find the best analysis technique between many techniques there is a need to fix a time and put all terms under equal terms then it will give the processing power that needed to every technique.
- **Time:** with this term, it could easily get the duration needed for analysis technique to finish its process, this term can measures when putting all compared techniques under the same terms like using the same processing hardware and introduce the same datasets to every technique to analyze them.
- **Storage:** this term refers to the storage capacity needed to store the analysis result, and it is easy to measure this after every technique finishes itis a mission.
- **Data Efficiency:** this term refers to the efficiency of the resulted data that output after finishing the analysis process, In comparing many techniques using this term gives the advantage to the technique that produces more useful data that give information that has the ability to make known that helps in making decisions easily.

After comparing many techniques using all of these terms it is an ability to make some mathematical operations that can give in finally a single value that describe the total priority for every compared technique.

VII. BIG DATA ANALYSIS ALGORITHMS

In Big Data Analytics, data mining algorithms and their data analysis techniques play a vital role in terms of dimension reduction, computational cost, memory requirement, and end result management and precision.

This section provides a brief discussion to explain its importance from the perspective of the analysis and search algorithms [24].

- **Clustering Algorithm:** One of the most common clustering tools is CloudVista which is used in cloud computing to run the clustering process in parallel. BIRCH and other clustering approaches are used in CloudVista to prove that you can manage very large-scale data. GPU is yet another clustering tool used to enhance a clustering algorithm's performance and safety.
- **Classification algorithms:** Like the clustering algorithm for big data mining, classification algorithms were learned from the input data collected Through data sources and a heterogeneous group of learners in the plan and implementation.
- **Frequent Pattern Mining:** Most of the time, data mining researchers focus at the very beginning on managing the enormous volume of the dataset because some of their initial approaches Tried to check the data from transaction data from large companies and shopping malls.
- **C4.5:** This method generates a decision tree classifier. A classifier is a data mining tool that takes a dataset that specifies the things you want to classify and Make an effort to predict the class to which the new data belongs and how it belongs. Decision tree learning produces a flowchart that is roughly similar to classifying new data.
- **K-Means:** K-means algorithm produces k groups from a set of data or objects that make the members of a group look more similar. It is a popular technique for the clustering and analysis of data.
- **Apriori:** The Apriori algorithm learns association rules and is implemented into a database containing a very large number of transactions and their outcome. Learning the rule of association is one of the techniques of data mining to learn correlations and associations among variables in a database.
- **Expectation-Maximization (EM):** This algorithm is generally used for the discovery of knowledge in mining as a clustering algorithm. The EM algorithm iterates and optimizes the probability of seeing experimental data in statistics until the parameters or values of a non-experimental statistical variable model are calculated.
- **PageRank:** PageRank is another analytical algorithm which is an algorithm for the analysis of links designed to standardize the subjective meaning of some linked object within a data object network. This algorithm performs a form of network analysis, which seeks to explore and rank the connections between objects.
- **AdaBoost:** Adaboost algorithm constructs a classifier. It is a classifier bringing in the data and trying to predict which class belongs to a new data element. This algorithm has the purpose of forming and integrating a group of weak learners to create a single strong learner.

VIII. Big Data Processing Tools

Big data processing tools are available in large numbers. This segment is a summary of some current Big Data Analysis techniques. Most of the available tools include batch processing, stream processing, and interactive analysis. Most batch processing tools like Mahout, are based on the infrastructure of Apache Hadoop. Flow data systems are used mostly for real-time analytics.

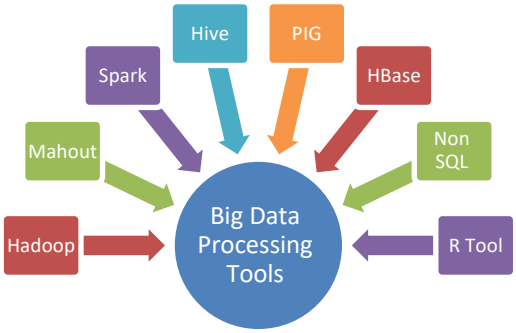

Fig 7: Big Data Processing Tools

A. *APACHE HADOOP And MapReduce*

Apache Hadoop and MapReduce are the most developed software platform for big data analytics. It consists of the Hadoop kernel, MapReduce, distributed file system Hadoop (HDFS), and hive apache, etc. Map-reduce is a programming model that is based on the divide and conquer approach for processing large datasets. The method of dividing and conquering is implemented in two steps such as step Map and step Reduce. Hadoop operates with two types of nodes including master node and worker node. In the map stage, the master node splits the input into smaller subproblems and then distributes them to worker nodes. Afterward, the master node combines the outputs in a reduced stage for all the subproblems. In addition, Hadoop and MapReduce serve as a powerful software platform for solving big data problems. It is also useful for storing fault-tolerant data and for high data processing performance.

B. *Apache mahout*

Apache mahout aims to provide scalable and commercial machine learning techniques for applications for large scale and smart data processing. Core mahout algorithms including clustering, classification, pattern mining, regression, dimensionality reduction, evolutionary algorithms, and batch-based distributed filtering run across the map-reduce system on the Hadoop platform. Mahout's aim is to build a vibrant, responsive, diverse community to facilitate project discussions and potential use cases. Apache mahout's basic goal is to provide a method to raise major challenges. The various companies with robust machine learning algorithms are Amazon, IBM, Yahoo, Google, Twitter and Facebook [25].

C. *Apache spark*

Apache Spark is an open-source big data processing framework built for speed processing and sophisticated analytics. It is easy to use and was originally developed in 2009 at UC Berkeley's AMPLab. It was open-sourced in 2010 as an Apache project. Sparklets you quickly write applications in Java, Scala, or python. In addition, to map-reduce operations, it supports SQL queries, streaming data, machine learning, and graph data processing. Spark runs on top of the existing Hadoop distributed file system (HDFS) infrastructure to provide enhanced and additional functionality. Spark consists of components namely driver program, cluster manager and worker nodes. The driver program serves as the starting point of execution of an application on the spark cluster. The cluster manager allocates the resources and the worker nodes to do the data processing in the form of tasks. Each application will have a set of processes called executors that are responsible for executing the tasks. The major advantage is that it provides support for deploying spark applications in an existing Hadoop cluster.

D. *Hive*

Hive is the SQL-like bridges that permit predictable business applications to run SQL queries against a Hadoop cluster. It was developed earlier by Facebook, then it has been made open-source software tool now, and it is a high-level perception of the Hadoop which allows all to make queries against data stored in a Hadoop storage medium just as if they were manipulating a conventional data store.

E. *PIG*

This is another analytical tool aiming to bring the Hadoop closer to developers and users of companies. PIG includes a Perl-like language that enables queries to be executed over data stored on a Hadoop, rather than a SQL.

F. *HBASE*

A non-relational database (NoSQL) that operates atop HDFS. Apache HBase is an open-source NoSQL database system which reads and writes in real-time access to those large databases. A linear HBase system with a few billion rows and millions of columns to manage very large data sets, and it conveniently blends data sources using a wide range of different structures and schemes [26]. HBase is native to Hadoop and works seamlessly with YARN access engine.

G. *NoSQL (non- relational databases)*

This NoSQL (not only SQL) database is a data administration and database design strategy which is useful for the large volume of data sets in the distributed background. Built using Apache Cassandra, the most popular NoSQL database is. In 2008, Cassandra, once the proprietary database of Facebook, was launched as an open-source. Certain implementations of the NoSQL database include Google BigTable, Cassandra, Map Reduce, SimpleDB, MemcacheDB, Voldemort, and MongoDB. Organizations using NoSQL include Amazon, LinkedIn, and Twitter as social media.

H. *R TOOL*

R is a well-known programming language and software tool for visualizing data using graphics and statistical computing. This is sponsored by the Statistical Computing Foundation, R. The R Tool is widely used for the development of statistical software and data analysis amongst statistical areas and data miners.

IX. BIG DATA ANALYTICS ISSUES AND CHALLENGES

To handle the challenges, in order to analyze big data, it is a need to know different computational complexities, information security, and computational method. Some statistical methods, for example, that perform well for small data size do not scale to voluminous data. Likewise, many statistical methods that work well for small data face substantial difficulties in analyzing big data. Many researchers have been investigating various challenges facing the health sector [27], [28]. Here, as shown in Figure 8, the big data analytics challenges are classified into five main categories.

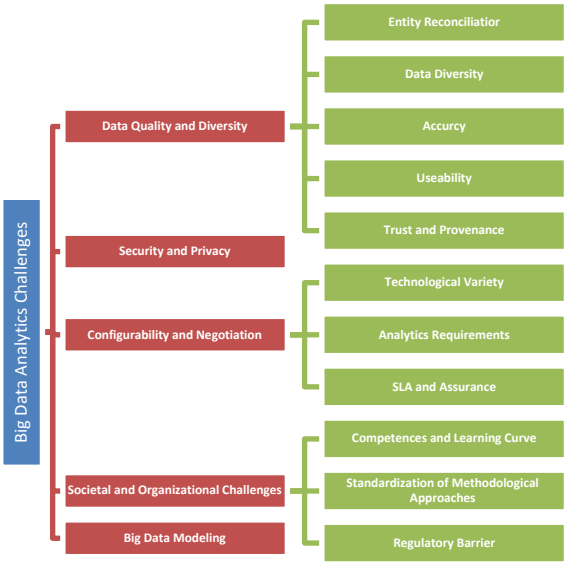

Fig 8: Big Data Analytics Challenges

A. *Data Quality and Diversity*

One of the most important characteristics of big data is diversity, and this diversity affects directly the analysis of big data, and this does not mean only the multiplicity in the type or nature of data, but also the accuracy of the data, its usability, and confidence in it, and this is what will be discussed in the following paragraphs:

Entity Reconciliation: Domain reconciliation, also known as record linking, refers to the identification of information across different data sources relating to the same domain. In data integration, this function is important where different sources may adopt different identifiers for the same entity or may refer more generally to related data, but there is no direct connection. Traditionally proposed techniques focus on a probabilistic assessment of the matching strength between two records that belong to different sources of data.

Issues and Challenges: The techniques adopted today must be revised in the context of parallel computation since the matching algorithm must be reorganized according to a

parallel approach too. This also implies the development of fault-tolerant evaluation of the results obtained from the matching functions, to prevent noisy data altering the reconciliation process. Top-down or up-front approaches can also be considered for managing the disambiguation between entries. However, as illustrated in [29], the application of a common model to data sources that do not originate from the same process may lead to heterogeneity in semance.

Possible Solutions: Big Data Analytics (BDA) should find entity reconciliation as a necessary step in data planning providing the user with a range of techniques from probabilistic to up-front approaches to tackle it.

Data Diversity: The data generated by Big Data technologies will present their structure in a variety of formats with different constraints. For example, centralized structures such as Enterprise Resource Planning (ERP) implement a predefined (structured) data structure. For distributed systems, no compliance at the object level is achieved then the content is typically encapsulated for meta-data. The format is unstructured when data is generated for human access, such as in files, documents, text, images, audio, and photographs, with the exception of a degree of metadata that may be introduced during the data preparation phase. Today the rise of cloud, loT and big data paradigms has led to the proliferation of any kind of sensors and probes. Furthermore, data is often multi-owner and thus displays multiple formats, quantities, granularities. In this case the data collected have an even greater degree of complexity and heterogeneity.

Issues and Challenges: In a BDA scenario, where the addition of a new and diverse data set has to be accomplished in an automatic way, the problem of data diversity is even exacerbated. Traditional approaches to normalization must be supported by techniques which adapt the deployed Big Data pipeline to the new source of data. The pipeline move will keep the pressure under control and the uncertainty that affects its operations.

Possible Solutions: Despite BDA's expected high level of automation, conventional techniques should be expanded to achieve a variety of intrinsic data. As a possible approach, expanded standardization of adaptation techniques may also be used. It is also important to prepare predefined patterns for on - the-fly incorporation of new and diverse datasets.

Accuracy: It is important to be able to produce accurate results as the output of the analytics to avoid missing the analytical objective completely. In general, however, this implies being aware of a dataset's level of accuracy before processing it. For non-stationary data series, this problem can be even more serious.

Issues and Challenges: The accuracy of Big Data is strongly linked to the need to calculate the difference between the perceptions of the consumer and the Big Data analysis received. The accuracy strongly depends on the specificity of the analytics criteria suggested by the users. The more precise the specifications are, the greater the accuracy. BDA may deploy various Big Data pipelines for the abstract requirements. Instead, choosing the correct pipeline leads back to the traditional scenario where selecting and installing the "best" analytics on the "right" computing infrastructure is more an art than a science. To get rid of such a situation, a BDA solution should reflect correctly the information that Big Data experts own in current solutions.

Possible Solutions: Adaptive optimization based on feedback on the results of previous BDA implementations will play a role in improving computational accuracy. For a single request, alternative solutions may be weighted based on the performance found in similar cases (i.e., for a different set of requirements).

Usability: It is a measure of the efficiency and effectiveness of achieving the objective. When talking about the usability of data, it is easy to look at two aspects. On the one hand, Data usability may increase with the availability of a more compact data set description, providing users with a clear understanding of the dimensions of the data affecting a phenomenon. On the other hand, the mapping between data and the corresponding analytics would become cumbersome if data are not stored using human-friendly representations.

Issues and Challenges: Given that a BDA system must be designed from high-level targets, it is of considerable importance to give the consumer an understanding of the data to be processed. Once a Big Data pipeline is issued, there is also a need for clear understanding of the interconnections between data format, data integration issues and analytics.

Possible Solutions: Summarization is a key concept in data mining that involves methods for finding a description of a compact data set. Summarization can be viewed as a compression to a smaller set of patterns for a given set of transactions while retaining as much information as possible. The opportunity to sample a BDA solution based on a trial that tests and shows the effectiveness of a solution is also an important line of action to be introduced.

Trust and Provenance: For proper Big Data management and based on a trustworthy Big Data Analytics, being able to verify data provenance is essential. In reality, the accuracy and usefulness of the results depend on the quality and origin of the data, which in turn helps to increase the confidence of the end-users in the process that produces these results. Additionally, trust in the processes and activities of third parties is mandatory when outsourcing Big Data computation to third parties. Third parties must provide evidence of their behavior and proper management of the data.

Issues and Challenges: BDA should provide means, similar to a traditional Big Data scenario, to verify the data and their origin. When it comes to trust, BDA needs stronger means for the establishment of trust. In fact, users of Big Data not only outsource their data and Big Data computing but also make all choices regarding the Big Data pipeline deployment. Components should also be deployed for monitoring and testing to verify the behavior of BDA solutions.

Possible Solutions: Current approaches to data integrity and non-repudiation along with assurance and SLA verification strategies should be applied. Trust and

provenance in BDA are not only linked to the verification of data but must also verify the behavior of the BDA pipeline as well as the provenance of the results obtained.

B. *Security and Privacy*

Security and privacy play an important role in impeding paradigms based on data outsourcing. Moving data as well as computing to external infrastructures further increases users ' concerns about their security and privacy, data loss, data breach, and data theft become critical threats to the organization's assets. The distributed and multi-source nature of big data actually poses many challenges. Compliance with the requirements and legal aspects of owners then becomes imperative to support Big Data outsourcing, particularly in essential security and privacy scenarios. In other words, all stakeholders who will use it must be made available to accredited compliance with Big Data analytics in their business model right from the start.

Issues and Challenges: BDA's introduction further increases security and privacy issues. In addition to traditional concerns related to data integrity safety, availability and confidentiality, new issues emerge due to the effect on security and privacy of a fully outsourced network and implementation strategy that would have. For instance, BDA gives an attacker the opportunity to implement inference attacks at a much lower cost.

Possible Solutions: In addition to traditional data security and privacy assurance testing, in a BDA environment, ensuring consistency with the BDA framework again becomes central to existing standards and user privacy and security specifications. Private data that the process configurations can infer must also be secured.

C. *configurability and Negotiation*

The efficiency of big data analysis depends on several factors, the ability of the data and its response to the analysis process, the strength and effectiveness of the analysis tools, and the availability of the analysis requirements. This will be mentioned as follow:

Technological Variety (Tools, Products): The advent of the Big Data model led to the development of several tools and products that managed the Big Data process in different ways. Among those items are NoSQL databases (e.g. MongoDB, Cassandra), parallel processing frameworks (e.g., Spark, Storm, Apache Tez), workflow management and execution systems (e.g., Apache Oozie, Azkaban, Luigi). Big Data's range of products makes it difficult for consumers to choose the solution that best suits their goals. Just Big Data experts have the expertise to compare similar products on the basis of their analytical criteria.

Issues and Challenges: BDA can be an appropriate approach to help Big Data users find a clear way through the Big Data labyrinth. However, the BDA model implies a detailed understanding of the characteristics of Big Data tools and products on the one hand, and user specifications on the other. If this information is less accurate, compensation techniques are needed which drive the selection of the best product set. Feedback on the success of previous implementations will play a part in improving selection efficiency.

Possible Solutions: BDA includes a clear description of tools and product features, with a priori mapping of the specifications of the users. Following technological evolution, proper taxonomies and vocabulary must be established. Binding between predefined Big Data pipelines and applicability scenarios can also enhance the quality of the analytics and the results obtained.

Analytics Requirements: Selection and configuration of analytics often requires the description of a complex set of parameters that depend on the data types and analytics needs. In some cases, requirements which drive a process of analytics which interfere with, or even be incompatible. Expert users who build on their expertise to handle conflicting requirements can easily manually solve those inconsistencies.

Issues and Challenges: BDA further exacerbates the problems posed by incompatibility in analytical parameters, since they are entirely in the hands of end users and are sometimes presented at an abstract level, increasing the likelihood of conflicts. However, BDA incorporates the need to evaluate interferences between requirements that can only materialize during the deployment period. For example, when data is anonymized, it can find that data visualization criteria can no longer be met.

Possible Solutions: A first approach to addressing the above issues is a modeling interferences-based conflict management solution which drives users in a consistent selection of requirements. There are two types of interference rules that can be foreseen: a priori interference modeling aspects common to any Big Data analytics, research-based modeling aspects of data-dependent interferences. An additional layer may specify a mix of capable disincentives.

SLAs and Assurance: To enhance the adoption of Big Data facilities, monitoring, and control of a Big Data analytics process is crucial. It is important to define and verify the Services Level Agreement (SLA) to ensure that the Big Data pipeline and its processes act as expected, such as exhibiting the performance required, ensuring minimum accuracy and preserving privacy and security properties. Assurance techniques support SLA assurance by offering "a way of gaining reasonable confidence that infrastructure and/or applications can reliably demonstrate one or more security issues and behave as planned amid faults and attacks"[30]. They provide a way for providers to receive assurances about the feasibility of the conducted process and to facilitate a post-auditing process. Organizations wishing to put BDA in their governance and/or command control chains can calculate clear ROis and risk assessments, thanks to SLAs and assurance.

Issues and Challenges: BDA needs to provide organizations with measurable factors at two different levels in order to bring organizational needs and technological solutions into contact. On the one hand, features must be mapped with specific measurable factors at the execution environment and architectural level. On the other hand, the fulfillment of the users defined specifications often needs to be observable. In this context, assurance techniques that measure the distance between the expectations of the user and the supplied Big Data pipeline become a pressing necessity.

Possible Solutions: A first approach to addressing the above issues is the concept of multi-layer assurance agents that collect information from the execution of the Big Data pipeline, verify user support specifications, and verify the proper behavior of the Big Data architecture. Inferencing and reverse engineering on the findings of analytics can also help to further improve assurance assessment.

D. Societal and Organizational Challenges

There are many social challenges facing the process of analyzing big data. These challenges include the availability of the competencies required to work on the analysis process and the availability of standardized methodologies for analysis that operate according to an organized framework and rules, and this will be discussed as follow:

Competences and Learning Curve: A Big Data gives unparalleled opportunities to the organisations. It supports advanced business intelligence applications and allows us to extract value from the huge amount of different data. The complexity of Big Data and the proliferation of Big Data solutions raise the bar on the need for costly in-house skills, which hinders its wide-ranging adoption. Outsourcing aspect of Big Data management to the outside does not minimize the need for in-house competences and, most significantly, prohibits SMEs from implementing Big Data solutions.

Issues and Challenges: BDA provides a rethink of current BDA methods, allowing users to announce and abstract their analytics, and leaving the responsibility of deploying the right pipeline and analytics to the analytics providers. However, BDA's quality relies heavily on the quality and precision of the user-specified requirements. The higher the BDA performance, the more BDA users have a clear understanding of their goals and Big Data technologies.

Possible Solutions: Specification of model-based, declarative Big Data analytics represents a possible first approach. In selecting declarative requirements, A wizard will help users in reducing the risks of conflicting requirements (see Section 9.3.2).

Standardization of Methodological Approaches: "Until very recently, the global IT community looked at Big Data in the same way that the elephant was examined by the six blind men in the fable. That is, each community member viewed the topic (Big Data) from a single perspective, at most" [31]. Just a few standardization programs have arisen today for these reasons. Due to the various legal and regulatory requirements for compliance in different countries, well-defined and internationally recognized standards can potentially reduce potential controversy.

Issues and Challenges: As stated in the European Big Data Value Partnership's Strategic Research and Innovation Agenda (SRIA),1 the lack of standards represents a major barrier to the spread of Big Data technologies.

Possible Solutions: Some initiatives to standardize can be taken as a reference. IEEE Big Data Technical Community held the 1st Big Data Initiative (BDI) IEEE Standards Workshop2 in 2015. Several standardization topics were established during the workshop, among which are essential for BDA Metadata Standard for Big Data Management and Data Representation in Big Data Management. ITU-T also began efforts to define Big Data standards3 and approved the first Big Data –Cloud computing-based requirements and capabilities standard in 2015 [32]. Finally, ISO began its efforts to define standard ISO / IEC CD 20546 Information Technology — Big Data— Definition and Vocabulary in 2016 [33].

Regulatory Barrier: Concerns about violating data access, sharing and custody regulations when using BDA, and the high cost of obtaining legal clearance for their specific scenario, deter companies, particularly small and medium-sized businesses, from taking BDA.

Issues and Challenges: As discussed above, data management comes with legal issues that need to be resolved through a proper Big Data analytics approach. This is especially true when, in addition to the data, the entire Big Data process is outsourced.

Possible Solutions: A first approach would rethink existing regulations to achieve peculiarities of the BDA scenario. It should, in particular, provide a consistent way for the management of data and infrastructures across different countries and regulations.

E. Big Data Modeling

Big-data developers have so far devoted little attention to modeling [34]. The traditional data modeling, which focused on solving the complexity of relationships between schema-enabled data, was discarded as no longer applicable to Big Data scenarios. Recent ideas have emerged from potential Big Data users ' expectations and criteria to establish the concept that a model comeback is needed to achieve the full potential of Big Data analytics.

Issues and Challenges: BDA adds a layer of sophistication to conventional BDA that needs to be managed using correct modeling techniques. Data modeling will not only be essential to BDA, but it will also include a model-based approach for process execution and architecture implementation

Possible Solutions: An appropriate approach to large-data modeling should provide a model-driven architecture for BDA. TOREADOR project4 is an H2020 project which aims to provide a specification for a fully declarative architecture and a model collection which supports Big Data Analytics. TOREADOR will enable users to set business-level goals (ii) describe and manage data and process diversity (iii) with a single learning curve for a variety of different fields of analytics and simulation-driven applications.

C. CONCLUSION

Data are produced at a dramatic pace during the last years. For a general guy, analyzing those data is difficult. In this survey, there are various challenges, and the tools used to analyze these big data. Every big data platform is focused on its own. Some are designed for batch processing while others are good for real-time analytics. Every big data platform also has functionality specific to it. Various analytical techniques used include statistical analysis, machine learning, data mining, smart analysis, cloud

computing, quantum computing, and data stream processing. There is a belief in these techniques that's able to give greater attention in future researchers to solve big data problems effectively and efficiently.

REFERENCES

[1] Katal, A. Wazid, M. Goudar, R.H "Big data: Issues, challenges, tools and Good practices, IEEE" Contemporary Computing (IC3), 2013 Sixth International Conference, pp:404 - 409.
[2] Shuhui Jiang, Xueming Qian, Tao Mei, Yun Fu, Personalized Travel Sequence recommendation on Multisource Big Social Media, 2016, IEEE Transactions on Big Data, Vol.2, Issue:1
[3] Gantz J, Reinsel D, Extracting value from chaos.IDC iView, 2011, pp 1–12
[4] Mayer-Schonberger V, Cukier K, Big data: a revolution that will transform how we live, work, and think. Boston: Houghton Mifflin Harcourt; 2013.
[5] Kitchin R. The real-time city? Big data and smart urbanism. Geo J. 2014, 79(1), pp: 1–14.
[6] Chen, C.P., Zhang, C.Y.: Data-intensive applications, challenges, techniques, and technologies: A survey on big data. Information Conference on. pp. 404-409. IEEE (2013)
[7] Kaiser, S., Armour, F., Espinosa, J.A., Money, W.: Big data: Issues and challenges moving forward. In: System Sciences (HICSS), 2013 46th Hawaii International Conference on. pp. 995-1004. IEEE (2013).
[8] Che, D., Safran, M., Peng, Z.: From big data to big data mining: challenges, issues, and opportunities. In: Database Systems for Advanced Applications. pp. 1-15. Springer (2013).
[9] Kata!, A., Wazid, M., Goudar, R.: Big data: Issues, challenges, tools, and good practices. In: Contemporary Computing (IC3), 2013 Sixth International Conference on. pp. 404-409. IEEE (2013).
[10] Manyika, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., Byers, A.H.: Big data: The next frontier for innovation, competition, and productivity (2011).
[11] Kim, G. H., Trim, S., & Chung, J. H. Big-data applications in the government sector. Communications of the ACM, 57(3) 78-85 (2014).
[12] Stonebraker, M., and J. Hong. Researchers 'Big Data Crisis; Understanding Design and Functionality, Communications of the ACM, 55(2), 10-11 (2012).
[13] Wu, X., Zhu, X., Wu, G.Q., Ding, W.: Data mining with big data. IEEE Transactions on Knowledge and Data Engineering, 26(1), 97-107 (2014).
[14] Zaharia, M., Chowdhury, M., Franklin, M.J., Shenker, S., Stoica, I.: Spark: cluster computing with working sets. In: Proceedings of the 2nd USENIX conference on Hot topics in cloud computing. vol. 10, p. 10 (2010).
[15] 0. Driscoll, A., Daugelaite, J., Sleator, R.D. big data, Hadoop and cloud computing in genomics. Journal of biomedical informatics 46(5), 774-781 (2013).
[16] Hashem, I. A. T., Yaqoob, I., Anuar, N. B., Mokhtar, S., Gani, A., & Khan, The rise of "big data" on cloud computing: review and open research issues. Information Systems, 4 7, 98-115 (2015).
[17] Simoff, S., Bohlen, M.H., Mazeika, A.: Visual data mining: theory, techniques, and tools for visual analytics, vol. 4404. Springer Science & Business Media (2008).
[18] Sawant, N., & Shah, H. Big Data Visualization Patterns. In Big Data Application Architecture Q & A 79-90 (2013).
[19] Katrina Sin and Loganathan Muthu, Applications of big data in education data mining and learning analytics – A literature Review, ICTACT Journal on soft computing special issue on soft computing models for big data, July 2015, Vol:05, Iss: 04, pp: 1035-1049
[20] Cheikh Kacfah Emani, Nadine Cullot, Christophe Nicolle, Understandable Big Data: A Survey, Computer Science Review, 2015, Vol: 17, pp: 71-80
[21] K. Krishnan, Data warehousing in the age of big data, in The Morgan Kaufmann Series on Business Intelligence, Elsevier Science, 2013.
[22] H.V. Jagadish, D. Agarwal, P.Bernstein, Challenges, and Opportunities in Big Data, The Community Research Association, 2015
[23] K. Davis, D. Patterson, "Ethics of Big Data: Balancing Risk and innovation", O'Reilly Media, 2012.
[24] Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufo Abdelaziz Bouras, A Survey of Clustering Algorithms for Big Data Taxonomy and Empirical Analysis, on Emerging Topics on Computing, IEEE, 11 June 2014.
[25] G. Ingersoll, Introducing apache mahout: Scalable, commercial friendly machine learning for building intelligent applications, White Paper, IBM Developer Works, 2009, pp. 1-18.
[26] Mike Barlow, Real-Time Big Data Analytics: Emerging Architecture, ISBN: 978-1-449-36421-2, 2013
[27] MH. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki, and D. K. Grunwell, Health big data analytics: current perspectives, challenges, and potential solutions, International Journal of Big Data Intelligence, 1 (2014), pp.114-126.
[28] R. Nambiar, A. Sethi, R. Bhardwaj, and R. Varghese, A look at challenges and opportunities of big data analytics in healthcare, IEEE International Conference on Big Data, 2013, pp.17-22.
[29] A. Azzini and P. Ceravolo, "Consistent process mining over big data triple stores," in 2013 IEEE International Congress on Big Data. IEEE, 2013, pp. 54-61 .
[30] C. Ardagna, R. Asal, E. Damiani, and Q. Vu, "From security to assurance in the cloud: A survey," ACM Computing Surveys (CSUR), vol. 48, no. 1, pp. 2:1-2:50, August 2015 .
[31] E. Gasiorowski-Denis, Big plans for big data, March 2014, http://www.iso.org/iso/home/news~index/news_archive/news.htm?refid=Refl821 .
[32] TUT-T, Big data - Cloud computing based requirements and capabilities, November 2015, http://www.itu.int/rec/T-REC-Y. 3600-201511-T .
[33] ISO/IEC, !SOI/EC CD 20546: Information Technology - Big Data - Definition and Vocabulary, 2016, http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=68305 .
[34] V. Markl, "Breaking the chains: On declarative data analysis and data independence in the big data era," Proc. of VLDB Endowment, vol. 7, no. 13, pp. 1730–1733, August 2014.