Big Data Strategy

Alicia Valdez[1], Griselda Cortes[2], Sergio Castaneda[3], Laura Vazquez[4], Angel Zarate[5], Yadira Salas[6],
Gerardo Haces Atondo[7]
Research Center, Autonomous University of Coahuila, Coahuila, Mexico[1, 2, 3, 4, 5, 6]
Research Center, Autonomous University of Tamaulipas, Tamaulipas, Mexico[7]

***Abstract*—The importance of data analysis in companies grows every day, with a global market that generates large amounts of transactions. Industry 4.0 is one of the technological trends, which is a set of diverse technologies whose objective is the digitalization and technological connectivity of the entire value chain of organizations. Data analysis and decision making in real time have a positive impact on efficiency. One of the technologies that support this concept is big data, which can support companies to use and manage large volumes of data as support in decision making. In this research project, the computational environment of Apache Hadoop software has been analyzed to create a technological strategy that supports companies in creating a roadmap to know and implement big data technology; as a result, a computer laboratory for big data has been created at the Autonomous University of Coahuila, Mexico to support medium-sized manufacturing companies in their data analysis strategy for decision making.**

Keywords—Technological strategy; big data; Hadoop; data analysis

I. Introduction

The global economy is in a phase that is characterized by digitalization and connectivity, the trend towards the automation of processes in the manufacturing industry.

Technologies such as the internet of things, cloud computing, big data, artificial intelligence and 3D printing, among others; emphasizing the importance of the manufacture of personalized and intelligent products [1]. The analysis of data, the exchange of information and decision making in real time have a positive impact on the efficiency of the entire value chain.

These technologies can support companies to reduce costs, as well as, other factors linked to competitiveness, such as infrastructure, logistics and the digital connectivity system, the cost of energy and the talent of the people.

The digitalization of the economy allows companies to have more information about their customers, at the same time as the entry of new competitors; therefore they face the challenge of increasing and escalating competition, and of making decisions on a large amount of data that they sometimes do not have the capacity to interpret; four main effects are identified in business in all industries: customers' expectations are changing, products are improving with data, new forms of collaboration between companies, and business models are being transformed into digital models [2].

Concepts such as data science and data scientist have emerged as a growing need for data analysis, combining the knowledge of statistics with the design of information and communication technologies, mathematics, operations research and applied sciences in order to extract knowledge derived from the processing, analysis and interpretation of data [3].

Some universities have incorporated the study of data science into their programs, as a need to train students with data science and analytics (DSA) competencies, so that they can successfully face the challenges of the future that require people with critical thinking [4]; because when organizations increase their ability to store data, having staff that can extract valuable information from these data, will be the differentiator of the organization.

Purdue University considers the study of data science in all its careers to be essential, through an initiative called "Integrative Data Science Initiative" [5], they are working in establishing an educational ecosystem of data fluency to prepare students for data-driven knowledge economy, developing infrastructure support for data science research and teaching.

A report elaborated by the Business Higher Education Forum (BHEF), mentions that by the year 2020, 2.72 million new jobs will require information analysis and data management skills; so they are looking to create academic strategies to align the needs of the industry with the development of DSA skills in students, and thus not affect economic growth [6].

Therefore, it is necessary to create technological strategies for companies and universities that allow them to investigate and assimilate new technologies based on Industry 4.0, especially in the analysis of data for the improvement of their processes and decision making; coupled with a college education that provides students and teachers with the foundation for data science competition.

In this project was developed a computational strategy for big data based on the Apache Hadoop Framework, limited to medium sized manufacturing companies.

The software was used as a tool in a computer lab for students and teachers at the Faculty of Mechanical and Electrical Engineering at Autonomous University of Coahuila and tested with real data from a manufacturing company in the raw material inventory processes.

Basically, this study has four sections.

In Section I, the introduction was shown. In Section II, the fundamental concepts are described. Also, the strategy elements such as the framework for big data was also stated. Thus, Section III describes the methodology, and Section IV describes the principal findings of the project.

II. Fundamental Concepts

A. Big Data

The increasing availability of data and information in organizations, because of the use of new services such as cloud computing, internet of things, and social network among others; it has meant the learning and use of new technologies for the storage and handling of large amounts of data. Among these technologies, big data stands out.

Watson quoting Mills mentions: “Big data is a term that is used to describe data that is high volume, high velocity, and/or high variety; requires new technologies to capture, store, and analyze it; and is used to enhance decision making, provide insight and discovery, and support and optimized processes” [7].

Russom [8], define a perspective for characterize big data, named the three V’s:

- High volume: the amount or quantity of data.
- High velocity: the rate at which, big data is created.
- High variety: the different types of data, structured and unstructured.

This data can be reported by machines, equipment, sensors, cameras, microphones, mobile phones, production software, among others; and can come from different sources such as companies, suppliers, customers and social networks.

The analysis of these data is key to making decisions in real time [9], allows to achieve better quality standards in products and processes, in addition to facilitating access to new markets.

Big data is creating a new generation of decision support data management, other factors has been considered for a successful big data project as: organizational culture, data architecture, analytical tools, and personnel issues.

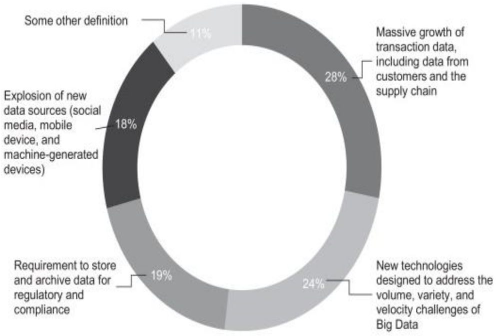


Fig. 1. Definitions of Big Data based on Executive Online Survey, Source [10].

The term big data is not yet fully understood by technology executives, a study that shows the application of an online survey to 154 global executives in April 2012 [10], mentions a variety of concepts and technologies, shown in Fig. 1.

The figure shows different understandings of big data, some definitions were focused on what is it, while others tried to answer what it does; focused on large volumes and varieties of data.

Some of the relevant techniques for structured and unstructured data analysis include [11]:

- Text analytics: extract information from textual data as corporate documents, emails, among others.
- Audio analytics: Analyses and extract information from unstructured audio data as human spoke language.
- Video analytics: involves a variety of techniques to monitor, analyze, and extract meaningful information from video streams.
- Social media analytics: Analysis of structured and unstructured data from social media channels.
- Predictive analytics: Techniques that predict future outcomes based on historical and current data.

Once the big data concept is understood, the related work and the software components are carried out.

B. Related Work

Sivarajah et al., [12] have analyzed the impact of the big data technology, defined as the new raw material of the 21st century; appropriate data processing and management could expose new knowledge, and facilitate in responding to emerging opportunities and challenges in a timely manner. Database and data warehouse technologies are becoming inadequate to manage the amount of data that the world is generating; advanced big data analyzing technologies (e.g. NoSQL Databases, BigQuery, MapReduce, Hadoop, WibiData and Skytree) can be better attained to enable in improving business strategies and the decision-making process in critical sectors such as healthcare, economic productivity, energy futures, among others.

Elgendy and Elragal [13], have investigated the question about how to integrate the big data analytics into the decision making process; they proposed the “B-DAD framework” or Big Data, Analytics and Decision framework for organizations who want to manage the big data technology. The framework has four stages: Intelligence, design, choice, and implementation. The intelligence stage has the activities for identifying big data sources and technologies to manage it; design has model planning, data analytics and analysis; the choice evaluate and decide activities; and the implementation activities are monitoring and feedback.

C. Hadoop Framework

Hadoop is an open source software framework that supports data intensive for distributed storage and distributed processing of very large data sets on computer clusters; the base Apache Hadoop[1] framework is composed of several modules as: Apache Hadoop MapReduce application tool for the programming aspect and Hadoop Distributed File System (HDFS) for an infrastructural point of view, the Fig. 2 shows the components of the framework.

Hadoop has an ability to move the processing or computing logic to the data where it resides as opposed to traditional systems, which focus on single server [14]. The core of Hadoop is it storage systems and its distributed computing model with its basic components:

- Name node: Head node or master node of the cluster, contains the metadata for HFDS during processing of data, which is
- Data node: These are the systems across the cluster which store the actual HDFS data blocks, these blocks are replicated on multiple nodes to provide high solutions.
- Job tracker: Service running on the Name node, which manages MapReduce jobs and distributed individual tasks.
- Task tracker: Service running in the Data nodes, which monitors individual MapReduce tasks that are submitted.
- Distributed across the nodes.

There are supporting projects for Hadoop, having different roles in the system, these are:

- Apache Hive: is a data-warehouse software that facilitates reading, writing, and managing large datasets residing in distributed storage using structured query languages (SQL) through a Java Data Base Connectivity (JDBC) driver [15], allow users to query the data without developing MapReduce applications.
- Apache HBase: Is the Hadoop database, a distributed, scalable, big data store, hosting of very large tables [16].
- Apache Mahout: Is a distributed linear algebra framework designed to implement algorithms [17].
- Apache Sqoop: Is a tool designed for efficiently transferring data between Hadoop and relational databases management systems (RDBMS) [18]; is a command line tool that controls the allocation between the tables and the data storage layer, translates the tables into a configurable combination for HDMS or Hive [19].

Fig. 3 shows the supporting software for Hadoop.

D. Strategy

The strategy is composed of five phases that involve different activities, being these:

1) Analysis of the hardware that is required for the installation of the software and the data to be analyzed, recommending a data server with a large storage capacity.

2) Selection of the processes of the company that will be analyzed can be customer sales processes, production data, equipment failures, among others; this process selection collects the necessary information and data that will be the raw material for the subsequent activities.

3) Installation and configuration of the Hadoop platform for distributed data processing, as well as the software to support the Hadoop System.

4) Extraction, Transformation and Loading (ETL) activities with analysis services.

5) Big data analytics, tools for analyzing reports (reporting), queries and visualization (dashboards) that will lead to data analytics. Fig. 4 shows the proposed strategy.

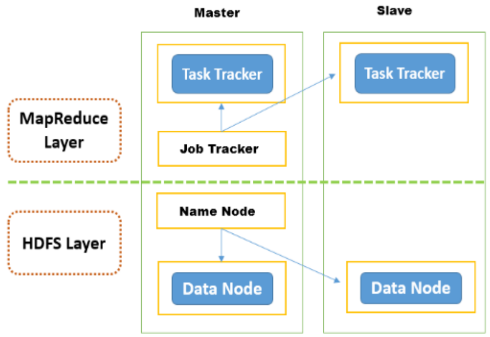


Fig. 2. Hadoop Framework Components, Source [14].

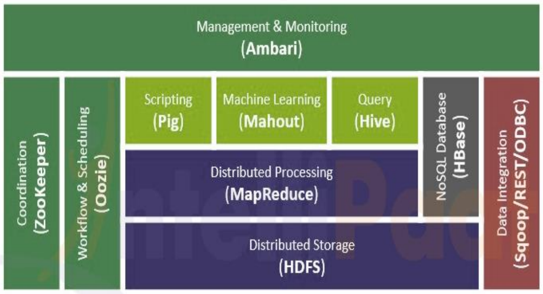


Fig. 3. Supporting Software Projects for Hadoop, Source [14].

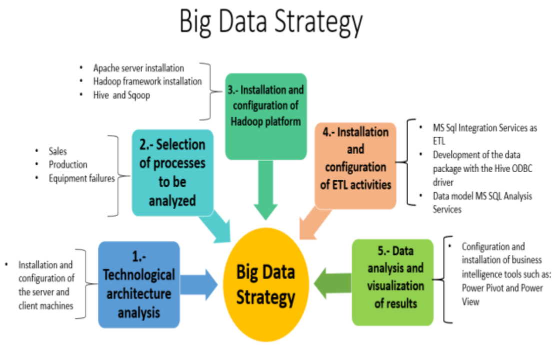

Big Data Strategy
Apache server installation
Hadoop framework installation
Hive and Sqoop
3.- Installation and configuration of Hadoop platform
Sales
Production
Equipment failures
2.- Selection of processes to be analyzed
4.- Installation and configuration of ETL activities
MS Sql Integration Services as ETL
Development of the data package with the Hive ODBC driver
Data model MS SQL Analysis Services
Installation and configuration of the server and client machines
1.- Technological architecture analysis
Big Data Strategy
5.- Data analysis and visualization of results
Configuration and installation of business intelligence tools such as: Power Pivot and Power View

Fig. 4. Strategy for Big Data.

[1] https://hadoop.apache.org

For the strategy development, a computer laboratory has been installed and configured at the Faculty of Mechanical and Electrical Engineering; for testing before its later use in a medium-sized company in the metalworking sector of the industrial region.

III. Methodology

The methodology phases were: Analysis of big data technology where a relevant importance has been found to this technology, which forming part of Industry 4.0; with the information of the technology, the software options for its management were found; defining the analysis for the software components and the strategy components. Subsequently, the Hadoop environment was downloaded, installed and configured at the data lab; this phase was the one that took most time; finally testing and implementation with real data. Fig. 5 displays the methodology phases.

The execution mode of Hadoop was a semi-distributed cluster to simulate a cluster of several nodes running on the same machine with its environment variables configured. Fig. 6 shows the execution of Hadoop.

The biggest challenge was the installation and configuration of the Hadoop environment as Hive, Pig and Sqoop software projects for the data analysis, point 3 of the strategy.

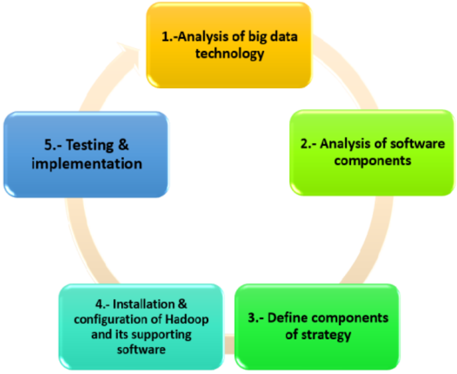
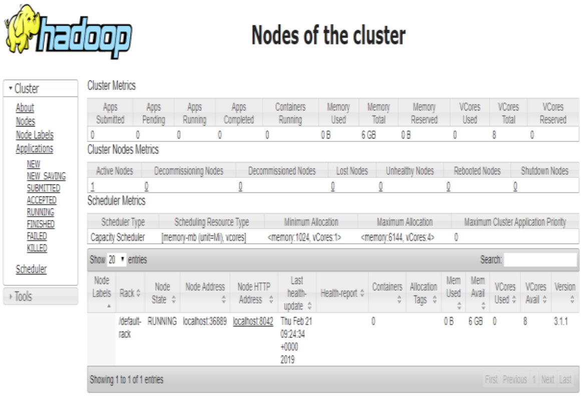

1.-Analysis of big data technology
2.- Analysis of software components
3.- Define components of strategy
4.- Installation & configuration of Hadoop and its supporting software
5.- Testing & implementation

Fig. 5. Methodology Phases.



Nodes of the cluster

Cluster Metrics

Apps Submitted	Apps Pending	Apps Running	Apps Completed	Containers Running	Memory Used	Memory Total	Memory Reserved	VCores Used	VCores Total	VCores Reserved
0	0	0	0	0	0 B	6 GB	0 B	0	8	0

Cluster Nodes Metrics

Active Nodes	Decommissioning Nodes	Decommissioned Nodes	Lost Nodes	Unhealthy Nodes	Rebooted Nodes	Shutdown Nodes
1	0	0	0	0	0	0

Scheduler Metrics

Scheduler Type	Scheduling Resource Type	Minimum Allocation	Maximum Allocation	Maximum Cluster Application Priority
Capacity Scheduler	[memory-mb (unit=Mi), vcores]	<memory:1024, vCores:1>	<memory:6144, vCores:4>	0

Show 20 entries Search:

Node Labels	Rack	Node State	Node Address	Node HTTP Address	Last health-update	Health-report	Containers	Allocation Tags	Mem Used	Mem Avail	VCores Used	VCores Avail	Version
	/default-rack	RUNNING	localhost:36889	localhost:8042	Thu Feb 21 09:24:34 +0000 2019		0		0 B	6 GB	0	8	3.1.1

Showing 1 to 1 of 1 entries First Previous 1 Next Last

Fig. 6. Hadoop Execution.

A. Technological Architecture

For this project was acquired an HP Proliant server 10th generation with 6TB of storage and three computers functioning as clients, localized at data lab in university. The Linux Red Hat operative system was installed as virtual machine, because the Windows Server is main operating system in the server.

B. Selection of Processes to be Analyzed

Several studies, including IBM, have analyzed the large number of applications for big data; the study shows five preferred orientations to apply big data in organizations in which 49% prefer apply it to focus on the customer, 18% in operational optimization, 15% in financial and risk management, 14% in the new business model and 4% in business collaboration [20].

The client centered processes of manufacturing companies can be considered as: Sales, distribution, market analysis, digital marketing, among others. For this project we have analyzed the processes of inventories management of raw material to produce harnesses and air bags for the automotive industry of a metal-mechanical company.

C. Installation and Configuration of Hadoop Platform

The process began with the installation and configuration of the Hadoop platform as virtual machine using Linux Red Hat version 2.6.18.

The installation and configuration of the software that is required for handling the data in Hadoop is very complex, especially in that each software must install and configure the environment variables separately, to give the user an integral management of the solution; so, this step has been the one that has consumed more time and resources.

The high complexity of the configuration of each software environment as Java, Hadoop, MapRed, Hive, and Scoop has required training to start the operation of the Linux operating system, so training in the management of this operating system will be included in the recommendations.

D. Installation and Configuration of ETL Activities

The activities of extraction, transformation and loading of the data have been made from a relational database (RDBMS) of MS SQL Server, being the data of a medium-sized company of the metalworking industry, belonging to the processes of material inventory management premium for the manufacture of harnesses, armchairs and air bags for the automotive industry of northern Mexico.

E. Data Analysis and Visualization of Results

A part of the design of the database is shown in Fig. 7, where the main entities that are represented were: articles, orders, providers, among others. The visualization of the results has been made with Excel's Power Pivot software for data processing indicator boards.

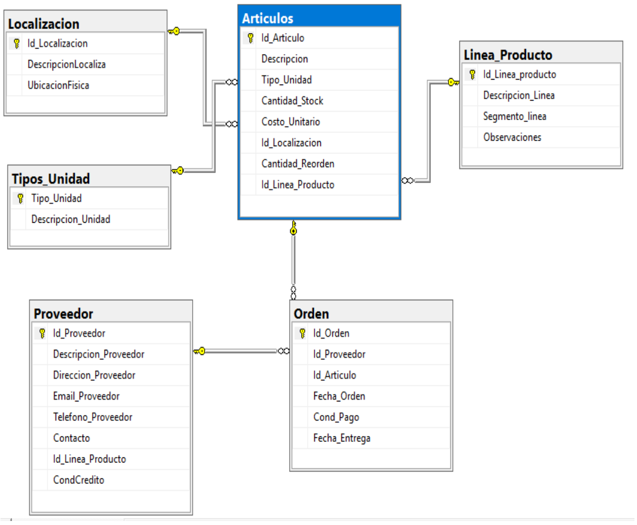


Fig. 7. Database Diagram for Inventory.

With these activities, the development of the strategy is presented, so the findings will be discussed in the results section.

IV. Results

The development of the proposed strategy has resulted in the following findings, also recommended as best practices [21]:

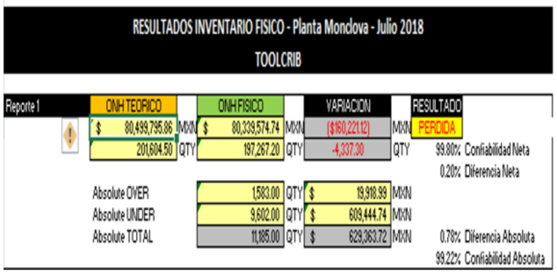
- Assess the true need for a big data processing, depending on the size of the company, real-time data management, types of analysis of the information required and the processes that support the data.
- Consider that the firm can have both types of processing on servers, RDBMS and big data.
- Identify the big data sources and map the big data types into workflow data types.
- Ensure the hardware for processing speed and storage.
- The operational data in companies are stored mostly in RDBMS.
- Consider interfaces for the different data sources that may exist in the company.
- Design an architecture for big data, considering the needs for this technology.
- Consider training in the operation of the Linux operating system and all the software of the Hadoop environment to obtain a real benefit.
- Some companies consider big data technology for the storage of non-SQL data.
- The software to perform the ETL activities is not integrated in the Hadoop platform, it is required to install and configure another type of software for these activities.

The results obtained have shown that developing and implementing a strategy based on big data involves several software, hardware and training activities; mentioning as principals: Emphasize the true need to have a big data processing in the company or use solutions with business intelligence techniques and relational databases.

The RDBMS still have a large proportion of the business operation, so it is recommended to have both types of big data processing and RDBMS if necessary; there are several solutions in the market for big data, finding in the Hadoop platform an open source software that can be installed from a virtual machine for tests and implementations; training in the operative Linux system and all the software that forms the Hadoop environment for communication between the RDBMS and the distributed processing of Hadoop.

Storing large volumes of data may require purchasing extra space or storage in the cloud. It is also important to identify the sources of big data such as social networks, geolocation, GitHub, market data, among others.

The results show the processing of the physical inventory at Monclova plant for the month of July 2018, making a comparative physical analysis against theoretical, divided into 4 reports, Fig. 8 shows the report 1 results where, the company has lost money for almost $8,000.00 USD in that month.



RESULTADOS INVENTARIO FISICO - Planta Monclova - Julio 2018

TOOLCRIB

Reporte 1	ONH TEORICO		ONH FISICO		VARIACION		RESULTADO
!	$ 80,499,795.86	MXN	$ 80,339,574.74	MXN	($160,221.12)	MXN	PERDIDA
	201,604.50	QTY	197,267.20	QTY	-4,337.30	QTY	99.80% Confiabilidad Neta
							0.20% Diferencia Neta
Absolute OVER			1,583.00	QTY	$ 19,918.99	MXN	
Absolute UNDER			9,602.00	QTY	$ 609,444.74	MXN	
Absolute TOTAL			11,185.00	QTY	$ 629,363.72	MXN	0.78% Diferencia Absoluta
							99.22% Confiabilidad Absoluta

Fig. 8. Results, Report1.

Hardware resources have also been obtained to have a small computation laboratory for big data in the Faculty of Mechanical and Electrical Engineering as part of the Autonomous University of Coahuila, where students of computer systems in the specialty of data management can learn from the development of this project.

V. Conclusion

In this project, a strategy for big data technology has been designed for support data analysis in the medium sized companies, in inventory processes of raw materials to produce various goods for the automotive industry of northern Mexico.

Several states of the strategy propose different goals, each of which involves an important number of activities such as: analysis of the technological architecture, selection of processes of the company to be analyzed, installation and configuration of the Hadoop platform software; installation and configuration of the activities of extraction, transformation and loading of the data, and the final data analysis and visualization of results in Excel Power Pivot.

Managing different operating systems may require training activities for the company's data administrator.

Developing a strategy for big data in a medium company is a difficult task for all the activities involved to develop successfully.

The tendency of managing large volumes of data, speed and variety of data, may imply an opportunity area for big data technology.

A lot of theoretical contribution was found on the subject.

As future work at the University, include students training for learn DSA skills, as mentioned in the educational trends of higher education.

This project has been a great learning experience for students, authors and teachers in the database area and granted by the PRODEP authority in Mexico.

VI. Discussion

In this project, a technological strategy was developed to manage big data in medium sized companies, was a large set of activities to get the objective. The main activities were: Install and configure the ethernet network with 1 server and 3 nodes, virtualization in Windows or Linux environment, to avoid configuration and implementation failures to create the virtualized environment, a bootable pen drive with Linux operating system was elaborated, and from there the operating system was taken to work with the implementation of the node.

The Java JDK software was installed to be able to operate the tools needed by Hadoop and Hadoop itself.

It has been a great technical challenge of hardware and software for the team responsible of the project.

The evaluation of the project has been successful in the first place for the configuration and installation of the hardware, the network, the server, and the nodes; second, the configuration and installation of the Hadoop software and the environment that works with Hadoop, as well as the administration and management of the data of the Hadoop nodes.

References

[1] A. Basco, G. Beliz, D. Coatz and P. Garnero , Industria 4.0 Fabricando el futuro, 2018, Buenos Aires, Argentina: BID, pp. 14-18.

[2] K. Schwab, The Fourth Industrial Revolution, 2016, Switzerland:The World Economic Forum, pp. 52-55.

[3] C. Weihs and K. Ickstadt, "Data Science:The impact of statistics". International Journal of Data Science and Analytics, 2018. 6(3): p.189-194.

[4] T. Chamorro, "3 Ways to Build a Data Driven Team". Harvard Business Review, 2018. 10(3): p.1-4.

[5] J. Akridge, "Purdue University launches robust collaborative Integrative Data Science Iniatitive", 2018.

[6] BHEF, Investing in America's data science and analytics talent, 2017, Business Higher Education Forum:U.S.A.

[7] H. Watson, "Tutorial:Big Data Analytics:Concepts, Technologies, and Applications". Communications of the Association for Information Systems, 2014 34(65): p. 1247-1268.

[8] P. Russom, "Big Data Analytics", TDWI Best Practices Report, TDWI, Editor 2011, The Data Warehousing Institute: Seattle, U.S.A.

[9] C. Ynzunza, J.Izar, J. Bocarando, F. Aguilar and M. Larios, The environment of Industry 4.0: Implications and future perspectives. Conciencia Tecnologica, 2017. 54(1): p. 1-8.

[10] A. Gandomi and M. Haider, "Beyond the hype: Big data concepts, methods, and analytics". International Journal of Information Management, 2015. 35(2):p. 137-144.

[11] S. Erevelles, N. Fukawa, and L. Swayne, "Big Data consumer analytics and the transformation of marketing", Journal of Business Research, 2016. 69(2):p. 897-904.

[12] U. Sivarajah, et al., "Critical analysis of big data challenges and analytical methods", Journal of Business Research, 2017. 70:p. 263-286.

[13] N. Elgendy and A. Elragal, "Big data analytics in support of the decision making process", Procedia Computer Science, 2016. 100(1071): p. 1071-1084.

[14] D. Sarkar, Microsoft SQL Server 2012 with Hadoop, 2013, Mumbai, India:Packt Publishing.

[15] Apache Software Foundation. Apache Hive TM. 2019 [cited 2019 10/01/2019]; Available form: https://hive.apache.org

[16] Apache Software Foundation. Apache Hbase TM. 2019 [cited 2019 10/01/2019]; Available form: https://hbase.apache.org

[17] Apache Software Foundation. Apache Mahout TM. 2019 [cited 2019 10/01/2019]; Available form: https://mahout.apache.org

[18] Apache Software Foundation. Apache Sqoop TM. 2019 [cited 2019 10/01/2019]; Available form: https://scoop.apache.org

[19] T. White, Hadoop:The Definitive Guide, 4th Edition, 2015, U.S.A.: O'Reilly Media Inc.

[20] D. Lopez, Analisis de las posibilidades de uso de Big Data en las organizaciones, Negocios y tecnologias de la informacion, 2012, Universidad de Cantabria: España, p.73.

[21] M. Nathan and W. James, Big data: principles and best practices of scalable real-time data systems, 2015, U.S.A.: New York; Manning Publications Co.