Big Data Analysis Proposal for Manufacturing Firm

Alicia Valdez, Griselda Cortes, Laura Vazquez, Adriana Martinez, and Gerardo Haces

Abstract — The analysis of large volumes of data is an important activity in manufacturing companies, since they allow improving the decision-making process. The data analysis has generated that the services and products are personalized, and how the consumption of the products has evolved, obtaining results that add value to the companies in real time.

In this case study, developed in a large manufacturing company of electronic components as robots and AC motors; a strategy has been proposed to analyze large volumes of data and be able to analyze them to support the decision-making process; among the proposed activities of the strategy are: Analysis of the technological architecture, selection of the business processes to be analyzed, installation and configuration of Hadoop software, ETL activities, and data analysis and visualization of the results. With the proposed strategy, the data of nine production factors of the motor PCI boards were analyzed, which had a greater incidence in the rejection of the components; a solution was made based on the analysis, which has allowed a decrease of 28.2% in the percentage of rejection.

Index Terms — Big data, Data analytics, Hadoop, Power BI.

I. INTRODUCTION

The massive analysis of data has become a process adopted by the leading companies in any economic sector and in any industry, particularly in the production area; it can be a competitive advantage that allows timely decisions to be made for the benefit of sales strategies [1].

Effective decision making is a fundamental part of the actions that a company can take to stay alive and based on current big data solutions, it is possible to adapt to change efficiently and in some cases in advance. A competitive advantage that allows timely decisions to be made for the benefit of sales strategies [2].

Database systems are very efficient in handling data, but when multiple queries are required that involve a lot of data or perhaps added functions to the records (such as the count or the average of a certain field) can present certain difficulties. These difficulties were detected because within the business world, companies began to notice the benefit of analyzing the data in their systems, grouping them by dates, by type of product, among others [3].

This has created a new challenge for data processing: getting useful information about a business to starting from the data generated on a day-to-day basis such as sales notes or customer records in order to carry out decision-making, planning and coordination processes [4].

The needs to apply various operations to the data stored in a database to obtain certain information, it is necessary to count how many records exist that meet a certain condition, add the total of certain data for records that meet some other criteria. As said before, the databases are optimized for data storage and it is more expensive (in processing time) to obtain the stored data. This handling of the data when is applied to large volumes of data, can be slow, which has generated new models and new structures to store the data [5].

Continuing with the importance of data analysis skills in organizations, the Business Higher Education Forum (BHEF) has produced a report called "Data science and analytics skills", where they analyze the importance of data analysis skills in students of higher education, the demand is for people with analytics skills in all areas of knowledge [6].

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 [7].

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.

The objective of this paper is to apply the concepts of big data for the analysis of large volumes of data, using Power BI software; to generate an improvement in productivity in a manufacturing company, based on the analysis performed.

Basically, this study has five sections.

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

II. FUNDAMENTAL CONCEPTS

A. Big Data

Sruthika and Tajunisha [8] explain that the concept varies from company to company, where for some, managing Tera Bytes (TB) is considered big data and for others big data

Submitted on February 2, 2021.

Published on February 15, 2021.

Alicia Valdez, Autonomous University of Coahuila, Mexico.

(e-mail: aliciavaldez@uadec.edu.mx)

Griselda Cortes, Autonomous University of Coahuila, Mexico.

(e-mail: griselda.cortes.morales@uadec.edu.mx)

Laura Vazquez, Autonomous University of Coahuila, Mexico.

(e-mail: laura_vazquez@uadec.edu.mx)

Adriana Martinez, Autonomus University of Coahuila, Mexico.

(e-mail: adrianaarmenta@uadec.edu.mx)

Gerardo Haces, Autonomous University of Tamaulipas, Mexico.

(e-mail: ghaces@uat.edu.mx)

DOI: <http://dx.doi.org/10.24018/ejece.2021.5.1.298>

would be talking about peta bytes (PB). These authors mention that “Big Data analytics refers to the process of collecting, organizing and analyzing large amounts of data; that are important to a business. To simplify our understanding of Big Data, we can support in three great characteristics: Volume, variety and speed”.

When it comes to defining the concept of big data, the first thing that comes to mind is the feature large data, coupled with other features such as variety and speed.

Other features have recently been added for which up to 5 V's can be counted, which are: Veracity and Value [9].

The Tech America Foundation defines big data as: “Big data is a term that describes large volumes of high velocity, complex and variable data that require advanced techniques and technologies to enable the capture, storage, distribution, management, and analysis of the information” [10].

Fig. 1. The Big Data V's, source [11].

The "Volume" refers to the size of the data, which grows exponentially. These are generated at different speeds, hence the "Velocity" parameter. "Variety" is given for two main reasons: the first is because the data is generated from different geographically distributed sources, and the second is due to the existence of different types of data: structured, semi-structured, and unstructured. "Visualization" is a very important part of any big data environment. The proper use of data visualization methods can lead to obtaining excellent results applicable to the objectives set [11].

The "Veracity" refers to be able to intelligently process and analyze the large volume of data in order to obtain true and useful information, that allows to improve decision-making process. "Validity" is a concept that tends to be confused with "Veracity". The data may be reliable and not present errors, but if they are not correctly understood, they are not valid, and finally, the final objective of using big data is to generate "Value" from the stored information that is obtained efficiently and at the lowest possible cost through different processes.

Given the great potential that big data technologies have for the processing and management of large volumes of data, great interest has been generated in applying this technology in many sectors of the economy [12].

The increasing availability of data and information in organizations, as a result 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 [13].

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, allows to achieve better quality standards in products and processes, in addition to facilitating access to new markets [14].

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 [15].

In general, the tools offered for big data solutions take advantage of the customer information by following the steps below:

1. The collection and classification of data. As a first approach, what kind of information the client handles and what indicators they can get by analyzing the information. The objective of this step is to know and understand the data that is handled and what information they can offer.

2. Analysis of the data. After defining what value can offer us data that is handled, the data is analyzed and the information that allows strategic decisions to be made, in other words, the results to be presented.

3. The unification and presentation of information. At the end it is identified how it is connected the information of the whole company and a visualization is presented easy to digest and understand for customers (graphics, boards, colors, among others) [16].

Theoretically, one of the advantages of big data is that the more data that is analyzed, the greater the breadth of visions that can be established around the objectives pursued by its use by an organization or company. The first question to ask when selecting data analysis tools is: what or what problems are trying to solve. In addition, the selection of analysis tools must also take into account the level of complexity of the problem to be solved [13].

Table 1 shows the use of some analytical tools for various types of analysis to be performed [13].

TABLE 1: ANALYSIS LEVEL, SOURCE [13]	
Type of analysis	Use
Basic analytics	Selection and division of data, reporting, simple visualizations and basic monitoring.
Advanced analytics	Complex analysis such as predictive models or pattern-setting techniques.
Operationalized analytics	Business process analysis.
Monetized analytics	Monetary analysis.

In this case study, the type of basic analytics was applied to the company of this project, described in the methodology section.

B. Power Business Intelligence (BI) Software

Microsoft Power BI is a business intelligence and analytics platform that supports both self-service data visualization and exploration as well as enterprise BI deployments. Power BI consists of cloud services, mobile applications, data modeling and report authoring application [17].

A Power BI dataset is a semantic data model composed of data source queries, relationship between dimensions and fact tables, and measurement calculations.

A Power BI Project supporting the extract-transform-load (ETL) process. For the case study, a database has been used that shows information on the production of engines for the automotive industry, with 15 tables each representing approximately more than 7000 data records.

Power BI software has been used, since it is a user-friendly tool and has a great capacity to show graphs of the results obtained.

III. METHODOLOGY

This research was carried out based on the big data strategy methodology, which presents a series of steps with which the most effective data treatment is given within the Hadoop ecosystem, in this strategy the following steps are presented: Analysis of the technological architecture, selection of the processes to be carried out, installation and configuration of the Hadoop platform, installation and configuration of the ETL activities with the Power BI software and finally, the data analysis and visualization of results. Fig. 2 displays the methodology phases.

Fig. 2. Methodology phases.

The first step is to analyze the technological architecture of the case study company, verifying the characteristics of the data servers where the Linux operating system will be installed as a virtual machine, and the Hadoop platform with its HDFS file system, to which they will be transferred the datasets.

The company is a manufacturing company dedicated to the manufacture of peripheral component interfaces (PCI), which include: interface modules, communications, inputs / outputs, power supplies, processors and complete PLC's racks, boards for the control of three-phase motors (servo motors); being this last production process the one that presents the highest rejections in production (42.7%), so in phase 2 of the methodology, the production process of boards for three-phase motors was selected. The Fig. 3 shows the programming parameters of a manufactured PCI.

Fig. 3. Programming parameters of a PCI.

For phase 3, the Hadoop software with its components has been installed and configured; once the Linux Debian [18] operating system has been installed as a virtual machine on the company's data server, highlighting the configuration of the main Hadoop [19] files such as: Core-site.xml, hdfs-site.xml, Mapred-site.xml and Yarn-site.xml, as well as the configuration of the environment variables [20].

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. The Fig. 4 shows the execution of Hadoop.

For phase four of the methodology used, several ETL activities have been carried out, for the preparation of the data package. Once the corresponding adjustments are made, the data extraction procedures are executed; obtaining the data from the source, operational step stores can be installed here that work as a bypass between the source of the data and the final node of the data; in situations where the data comes from different sources and are not the same format, we proceed to transform them to avoid duplication of information or data without any correlation and to be able to structure them. Once structured, they are loaded into the final condensation system for their treatment and classification in such a way that we can extract them into different strata to facilitate future studies, when the data is classified, they are concentrated in a single group to be processed and classified in a common format. then they are kept in a control process until their analysis.

Among the activities to prepare the data to be analyzed, the identification of the data source stands out.

This means that the data can come from different sources such as text files, Oracle databases, Access, Sql Server, Excel or any other source of data; so, the first activity is to prepare the data package consists of unifying them in a single format to be transferred to the final node, providing them with a structure to be loaded into the system, processing and analyzing them.

In this case, the data source is a database of the company described in the case study; the database is in a single format in Access, so it has been converted to a text format separated by semicolons, in this case a file in .CSV (Comma Separated Values) format, which are a type of document in open format, simple to represent data in table form, in which the columns are separated by commas or semicolons. Fig. 6 shows the CSV format of the database converted from Access to .CSV format, in this case each column is separated by semicolons. Fig. 5 shows a part of the engine database in ACCESS.

Nodes of the cluster

Fig. 4. Hadoop execution.

```
1;"18140508";"374";";";"06/04/2018";"09:03:00 a.m."; "FAIL"
2;"18140508";"3522";";";"06/04/2018";"09:03:41 a.m."; "PASS"
3;"18140512";"1172";";";"06/04/2018";"09:04:33 a.m."; "FAIL"
4;"18140512";"2914";";";"06/04/2018";"09:05:20 a.m."; "FAIL"
5;"18140494";"3398";";";"06/04/2018";"09:06:06 a.m."; "FAIL"
6;"18140408";"3784";";";"06/04/2018";"09:06:51 a.m."; "PASS"
7;"18140525";"3773";";";"06/04/2018";"09:07:35 a.m."; "PASS"
8;"18140445";"3763";";";"06/04/2018";"09:08:20 a.m."; "PASS"
9;"18140500";"3797";";";"06/04/2018";"09:09:04 a.m."; "PASS"
10;"18140522";"3769";";";"06/04/2018";"09:09:47 a.m."; "PASS"
11;"18140496";"3787";";";"06/04/2018";"09:10:31 a.m."; "PASS"
12;"18140454";"3802";";";"06/04/2018";"09:11:15 a.m."; "PASS"
13;"18140540";"3789";";";"06/04/2018";"09:11:58 a.m."; "PASS"
14;"18140516";"2919";";";"06/04/2018";"09:12:43 a.m."; "FAIL"
15;"18140516";"3802";";";"06/04/2018";"09:13:33 a.m."; "PASS"
16;"18140493";"3813";";";"06/04/2018";"09:14:13 a.m."; "PASS"
```

Fig. 5. Database tables.

Fig. 6. Database file in .CSV format.

Once the dataset to be processed has been prepared and reviewed, the next step is to transfer the CSV format file

from the user's physical directory to the Hadoop HDFS directory, since they are independent systems.

To start using Hadoop, the file system must be formatted, that is, the HDFS from the sbin folder with the -format instruction.

Hadoop was enabled with the start-all.sh command, then proceeding with loading the processed data from the local Windows file system to the Hadoop or HDFS file system, with the hdfs -include File command C:\User\User\Documents\Database.CSV; all these instructions have been executed from the Linux terminal; with this operation, the database was copied to the Hadoop file system and the data was analyzed creating and starting a session in the Hive software. Once the information has been integrated, queries were made in the database and the Hadoop environment was analyzed to integrate high-volume data.

To complete the proposed methodology, the final phase has been carried out, which is the analysis of the processed data and the visualization of the results obtained. It should be noted that for this case study, the data from the manufacture of the servomotor boards were used, which have nine stored parameters, which are: ID, serial number, RPM, voltage, amperage, watts, date, time, and state.

The visualization of the results has been made with Power BI software, for data processing indicator boards.

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

IV. RESULTS

The analysis of the information developed with big data technology, has provided data on how to improve the process and decrease the rejection points of PCI boards. Fig. 7 shows the rejection graph before big data analysis.

90378002	90378003	90378004	90378005	90378024	90378203	90380002
ID	ID	ID	ID	ID	ID	ID
Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie
RPM	RPM	RPM	RPM	RPM	RPM	RPM
VLT	VLT	VLT	VLT	VLT	VLT	VLT
AMP	AMP	AMP	AMP	AMP	AMP	AMP
WATT	WATT	WATT	WATT	WATT	WATT	WATT
90380004	90380007	90387001	90387004	90387005	90387007	90387015
ID	ID	ID	ID	ID	ID	ID
Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie
RPM	RPM	RPM	RPM	RPM	RPM	RPM
VLT	VLT	VLT	VLT	VLT	VLT	VLT
AMP	AMP	AMP	AMP	AMP	AMP	AMP
WATT	WATT	WATT	WATT	WATT	WATT	WATT
90387020	90387025	90387031	90387032	90387033	90387034	90387040
ID	ID	ID	ID	ID	ID	ID
Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie
RPM	RPM	RPM	RPM	RPM	RPM	RPM
VLT	VLT	VLT	VLT	VLT	VLT	VLT
AMP	AMP	AMP	AMP	AMP	AMP	AMP
WATT	WATT	WATT	WATT	WATT	WATT	WATT
90387053	90387055	90387057	90387058	90387059	90387061	90387070
ID	ID	ID	ID	ID	ID	ID
Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie
RPM	RPM	RPM	RPM	RPM	RPM	RPM
VLT	VLT	VLT	VLT	VLT	VLT	VLT
AMP	AMP	AMP	AMP	AMP	AMP	AMP
WATT	WATT	WATT	WATT	WATT	WATT	WATT
90387072	90387075	90387077	90387080	90387081	90387083	90387087
ID	ID	ID	ID	ID	ID	ID
Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie	Numero_de_serie
RPM	RPM	RPM	RPM	RPM	RPM	RPM
VLT	VLT	VLT	VLT	VLT	VLT	VLT
AMP	AMP	AMP	AMP	AMP	AMP	AMP
WATT	WATT	WATT	WATT	WATT	WATT	WATT

Fig. 7. Status of the PCI manufacturing process.

Fig. 8. Data analysis with Power BI.

One of the results has been the identification of the manufacturing parameters of the PCI boards, where the highest percentage of failures was concentrated, once the analysis was performed with Hadoop. The parameters of RPM, VLT, AMP and WATT are decisive to know the status of the PCI board, and each one of them has a tolerated error range; therefore, when processing the production data, it has allowed us to know which parameters must be improved to decrease the rejection percentage.

Considering the necessary parameters, the information has been analyzed, as well as each statistical representative of the process, the results obtained are shown; Fig. 8 shows the information with the analyzed process.

Fig. 8 shows the percentage of engines that have passed the test and are therefore in a condition to be sold. Of the data obtained, 15.36% of the total number of records processed are at a level of fails, while 84.64% have passed the tests that the engines have been subjected to for verification.

As a next step, the screen has been processed to be viewed and its structure was designed to appear in a mobile presentation according to the monitored parameters, as shown in Fig. 9.

Fig. 9. Data visualization on mobile platform.

Fig. 10. Changes in the physical test station.

In other words, the wiring and assembly process in the machine is carried out by the operator, who in turn operates the test station and notes the resulting parameters that are thrown by the machine, consequently the information is compared with the model standards in function to go to the acceptance or rejection process.

The accumulation of pieces with some fault in the parameters was observed, which take time to reach the rework station and therefore it is not possible to reduce the high rate of failure in the pieces.

In order to reduce rejection and rework rates, a more versatile reading mode was developed and in order to enhance the traceability of the product, it was decided to facilitate the reading process, which in turn yields real-time information to production and quality controls.

In this process, the diagram and protocols of the workstation connected to the information transfer node were

designed in order to improve the reading, remaining as in Fig. 10.

Fig. 10 shows the different points that were optimized in order to decrease the rejection rate in the manufactured parts. The changes and modifications are described below:

In the reading of measured parameters, an input and output module were installed with the purpose of converting these signals to a digital protocol. When obtaining these signals, a server was installed for the treatment, conversion, process and storage of information through the database. Access.

In the same way, a processor with peripherals was installed for the manipulation and programming of the reading and interpretation of the data, then a terminal for the operator was installed, which visually alerts (by color) the operator of the test result and, at in turn, it communicates it to the traceability of the product system.

The information management at this point was configured for controlled access by authorized personnel, the remote communication system was installed through ethernet for remote access of any device with prior access granted.

Once the operative has made each change, the observation was made to collect the new data.

It can be seen that the comparison time was reduced by thirty-eight percent (from forty seconds to fifteen seconds), achieving that the piece is passed instantly to the rework station, which likewise has the parameter failed and is repaired more efficiently. As for the parameter that presents more dysfunctionality, it accumulates in the current flow, which is caused by the tolerance variation with respect to the flow of the anode to the cathode of the SCR'S in the trigger control part.

Regarding the rejection rate at the station, based on the data collected, the reduction to fourteen points was found, five percent, which integrates us within the parameters allowed by the quality area, as shown in Fig. 11.

With the implementation carried out, the operator develops the acceptance or rejection process more quickly and efficiently, allowing the continuous flow of materials to the next station. The results show the decrease of the percentage of rejection of the manufacture of the PCI by 28.2%.

With the above, once exposed the analysis and improvement that has been implemented in the case study company, derived from the analysis and processing of large volumes of data; the big data tools for data processing proved to be effective and can be implemented in manufacturing companies backed by a big data strategy.

Other results were focused on preparing and configuring the hardware equipment necessary for handling big data, self-training in the Linux operating system, since the entire Hadoop environment is developed to work and run on Linux.

Fig. 11. improved process state.

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.

V. CONCLUSION

In this project, a strategy for big data technology has been designed for support data analysis in the medium sized companies, in manufacturing processes.

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 Power BI.

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.

In this case study, the analysis performed on the data processed with big data in the manufacturing process of the PCI boards of the engines, where nine production factors were analyzed, which had very high rejection percentages and the PCI board had to be reprocessed, the factors that had the highest incidence in rejection were detected and a solution based on the analysis was performed, which has allowed a decrease of 28.2% in the rejection percentage.

REFERENCES

- [1] K. Schwab, *The Fourth Industrial Revolution*, 2016, Switzerland: The World Economic Forum, pp. 52-55.
- [2] R. LopezAraiza, *Data analytics systems for large data volumes using agile methodology in computer engineering*, 2017, U.N.A.M.: Mexico City. p. 111.
- [3] R. Elmasri and S. Navathe, *Fundamentals of Database Systems*. 6th ed. 2010: Addison-Wesley.
- [4] I. Bojicic, et al, *Domain/Mapping Model: A Novel Data Warehouse Data Model*. *International Journal of Computers Communications & Control*, 2017. 12(2): p.166-182.
- [5] I. Sommerville, *Ingenieria del Software*, 2011, Mexico City: Addison-Wesley.
- [6] BHEF, *Investing in America's data science and analytics talent*, 2017, Business Higher Education Forum: U.S.A.
- [7] C. Weihs and K. Ickstadt, *Data Science: The impact of statistics*. *International Journal of Data Science and Analytics*, 2018. 6(3): p. 189-194..
- [8] S. Sruthika and N. Tajunisha, *A study on evolution of data analytics to big data analytics and its research scope*, in *International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS)*, ICIIECS, Editor 2015, IEEE: United States.
- [9] T. Erl, W. Khattak, and P. Buhler, *Big Data Fundamentals: Concept, Drivers & Techniques* 2016, United States: Prentice Hall Press.
- [10] TechAmericaFoundation, *Demystifying bigdata: A practical guide to transforming the business of Government.*, T.F.s.F.B.D. Commission, Editor 2012, The Tech America Foundation: U.S.A..
- [11] L. Plasencia and C. Calderon, *Architecture of Big Data for the management of telecommunications*. *Revista chilena de ingeniería*, 2017. 25(4): p. 1-7..

- [12] R. Chiang, et al., *Strategic value of Big Data and business analytics*. Journal of Management Information Systems, 2018. 35(2): p. 383-387.
- [13] B. Schoenborn, *Big Data Analytics Infrastructure for Dummies*. First ed, 2014, Indianapolis, Indiana U.S.A.: John Wiley & Sons Inc.
- [14] C. Ynzunza, et al., *The environment of Industry 4.0: Implications and future perspectives*. Conciencia Tecnológica, 2017. 54(1): p. 1-8.
- [15] J. Cabrera-Sanchez and A. Villarejo-Ramos, *FACTORS AFFECTING THE ADOPTION OF BIG DATA ANALYTICS IN COMPANIES*. Revista de Administração de Empresas, 2020. 59(6): p. 1-10.
- [16] B. Schmarzo, *Big Data: Understanding How Data Powers Big Business* 2013, Indianapolis, Indiana, U.S.A.: John Wiley & Sons Inc.
- [17] B. Powell, *Mastering Microsoft Power BI* 2018, Birmingham, U.K.: Packt Publishing LTD.
- [18] Software in the Public Interest, I. *Installing Debian via the Internet*. 2019 [cited 2019 11/05/2019]; Available from: <https://www.debian.org/distrib/netinst>.
- [19] TheApacheFoundation. *Apache Hadoop*. 2019 [cited 2019 10/01/2019]; Available from: <https://hadoop.apache.org/>.
- [20] TheApacheFoundation. *Hadoop Characteristics*. 2019 [cited 2019 11/05/2019]; Available from: <https://cwiki.apache.org/confluence/display/HADOOP2/GettingStartedWithHadoop>.

Alicia Valdez was born in Coahuila, Mexico in December 1965. Earned the title of Computer Systems Engineer in 1988, at the Autonomous University of Coahuila, Mexico. The master's degree was obtained at the Monterrey Institute of Technology and Higher Studies in Information Technology Management in 2000, and the Doctoral studies were made at the Popular Autonomous

University of the State of Puebla in Strategic Planning and Technology Management in 2014. Major field of study is data base management, big data, enterprise architecture, and software engineer.

She has had work experience in manufacturing firms in Coahuila, Mexico in Information Technology Management. Is a research professor at the Faculty of Mechanical and Electrical Engineering. She has published research papers at International Journal of Advanced Computers Science and Application, ECORFAN Journals, IARIA, between others.

Dra. Valdez was awarded by the honorable mention in doctoral studies by the research developed in the Enterprise Architecture area.

Griselda Cortes was born in Coahuila, Mexico in December 1967. Earned the title of Computer Systems Engineer in 1990, at the Autonomous University of Coahuila, Mexico. The master's degree was obtained at the Autonomous University of Coahuila in Computer Sciences in 2006, and the Doctoral studies were made at the Popular Autonomous University of the State of Puebla in Strategic Planning and Technology Management in 2015. Major field of

study is systems simulation, statistics, and software engineer.

She Is a research professor at the Faculty of Mechanical and Electrical Engineering. She has published research papers at International Journal of Advanced Computers Science and Application, ECORFAN Journals, IARIA, between others.

Dra. Cortes is a member of the Multidisciplinary Mexican Academy.

Laura Vazquez was born in Coahuila, Mexico in December 1975. Earned the title of Computer Systems Engineer in 1998, at the Autonomous University of Coahuila, Mexico. The master's degree was obtained at the Autonomous University of Coahuila in Computer Sciences in 2006, and the Doctoral studies were made at the Popular Autonomous University of the State of Puebla in Strategic Planning and Technology

Management in 2015. Major field of study is programming, data bases, and software engineer.

She Is a research professor at the Faculty of Mechanical and Electrical Engineering. She has published research papers at International Journal of Advanced Computers Science and Application, ECORFAN Journals, IARIA.

Dra. Vazquez is a member of the Multidisciplinary Mexican Academy.

Adriana Martinez was born in Coahuila, Mexico in 1980, she is a student of the Faculty of Mechanical and Electrical Engineering at the Computer Systems Engineer program.

She is doing a professional internship in the big data area and collaborating in the research project.

Gerardo Haces has a Bachelor's Degree in Administrative Computing from the Autonomous University of Tamaulipas. He has a Master's degree in Teleinformatics from the Autonomous University of Nuevo León and a Doctor's degree in Strategic Planning and Technology Management from the Popular Autonomous University of the State of Puebla.

He is currently a Full Time Professor in the Bachelor of Information Technology career at the Autonomous University of Tamaulipas attached to the Victoria Faculty of Commerce and Administration. He is a member of the Academic Body of Information Technologies and Strategy. In his work of scientific dissemination, he has written several academic articles published in important national and international journals, such as: Revista de Administração FACES Journal, Journal of Technology Management & Innovation, International Business Research, among others. He was a member from 2017 to 2019 of the National System of Researchers of Conacyt, level C. Similarly, some of the research projects in which he has participated derived from his line of research, the Strategic Administration of Information Technologies in Organizations are: Information Technology Capabilities and their impact on the generation of value, entrepreneurship and innovation, Government and Management of Information Technology and Electronic Commerce in SMEs.