Design of technological strategy for Big Data with Hadoop software

Diseño de estrategia tecnológica para Big Data con el software Hadoop

VALDEZ-MENCHACA, Alicia†*, VAZQUEZ-DE LOS SANTOS, Laura, CORTES-MORALES, Griselda and PAIZ-RIVERA, Ana

Universidad Autónoma de Coahuila, Facultad de Ingeniería Mecánica y Eléctrica

ID 1st Author: *Alicia, Valdez-Menchaca* / **ORC ID:** 0000-0002-3494-4830, **Researcher ID Thomson:** S-4551-2018, **Scopus ID:** 571110051800, **CVU CONACYT ID:** 292172

ID 1st Coauthor: *Laura, Vazquez-De Los Santos* / **ORC ID:** 0000-0002-0291-7774, **Researcher ID:** S-6543-2018, **CVU CONACYT ID:** 615088

ID 2nd Coauthor: *Griselda, Cortes-Morales* / **ORC ID:** 0000-0002-2567-7056, **CVU CONACYT ID:** 617827

ID 3rd Coauthor: *Ana, Paiz-Rivera*

DOI: 10.35429/JCSI.2020.16.6.7.14

Received January 18, 2020; Accepted June 25, 2020

Abstract

The objective of this research project is the design and implementation of a technological strategy for the use of big data technologies as Apache Hadoop, as well as its supporting software projects that allows to prepare medium-sized companies in new innovative technologies. As part of the methodology, an analysis of the best big data practices, analysis of the software for design and configure big data in a linux server for the technological proposal. As a first result, a roadmap for the installation and configuration of Hadoop software running on a Linux virtual machine has been obtained, as well as the proposal of the technological strategy whose main components are: analysis of the technological architecture, selection of processes or data to be analyzed and installation of Hadoop, among others.

Technological strategy, Big data, Hadoop

Resumen

El objetivo de este proyecto de investigación es el diseño e implementación de una estrategia tecnológica para el uso de tecnologías de big data como Apache Hadoop, así como sus proyectos de software de soporte que permitan preparar a las empresas medianas en nuevas tecnologías innovadoras. Como parte de la metodología, un análisis de las mejores prácticas de big data, análisis del software para diseñar y configurar big data en un servidor Linux para la propuesta tecnológica. Como primer resultado, se obtuvo una hoja de ruta para la instalación y configuración del software Hadoop que se ejecuta en una máquina virtual Linux, así como la propuesta de la estrategia tecnológica cuyos componentes principales son: análisis de la arquitectura tecnológica, selección de procesos o datos para ser analizado e instalación de Hadoop, entre otros.

Estrategia tecnológica, Big data, Hadoop

Citation: VALDEZ-MENCHACA, Alicia, VAZQUEZ-DE LOS SANTOS, Laura, CORTES-MORALES, Griselda and PAIZ-RIVERA, Ana. Design of technological strategy for Big Data with hadoop software. Journal of Computational Systems and ICTs. 2020. 6-17: 7-14.

*Correspondence to Author (E-Mail: aliciavaldez@uadec.edu.mx)
† Researcher contributing as first author.

Introduction

Companies today require strategic solutions to improve their capabilities and respond to the business or technological challenges that today's markets demand. The global economy is focused on a phase characterized by digitization, connectivity, and the trend towards process automation (Basco, Beliz, Coatz, & Garnero, 2018).

Technologies such as internet of things, cloud computing, big data, artificial intelligence and 3D printing, among others; They reinforce the importance of the manufacturing industry through the manufacture of personalized and intelligent products. Data analysis, information exchange and real-time decision making have a positive impact on the efficiency of the entire value chain (Basco et al., 2018).

Technologies such as cloud computing, IoT and big data, among others, further reduce coordination costs. Therefore, other factors linked to competitiveness, such as infrastructure, logistics and the digital connectivity system, the cost of energy and the talent of people according to the requirements of Industry 4.0, once again occupy a important place in location decisions of global companies.

The digitization of the economy changes the rules of the market: companies have more and more information about their customers, but at the same time, they allow new competitors to enter. Therefore, they face the challenge of facing increasing and scalable competition, and making decisions about a large amount of data that they sometimes do not have the capacity to interpret.

Four main business effects have been identified across industries: customer expectations are changing, products are being improved with data, new forms of collaboration between companies, and operating models are being transformed into digital models (Schwab, 2016). Therefore, it is necessary to create new technological strategies for SMEs that allow them to research and assimilate new technologies based on Industry 4.0 to improve competitiveness and productivity.

Fundamental concepts

The term Industry 4.0 refers to a new model of organization and control of the value chain, through the product life cycle and throughout the manufacturing systems supported by information technologies, it is also called "factory smart ”or“ industrial internet ”(Román, 2018); The technologies that support this term are known as pillars of Industry 4.0, among which are:

- Simulation.
- Additive manufacturing.
- Integration systems.
- Cybersecurity.
- Augmented reality.
- Cloud Computing.
- Robotization.
- Industrial Internet of Things.
- Big data and data analysis.

Big data

Since the presentation of the term by the MGI (McKinsey Global Institute) in June 2011, there have been various attempts to limit the concept. MGI define define it as the data set whose size goes beyond the ability to capture, store, manage and analyze database tools (Mayinka et al., 2011).

One of the most complete approaches to Big Data is the one provided by Gartner (Beyer & Laney, 2012): “They are information assets characterized by their high volume, speed and variety, which demand innovative and efficient processing solutions to improve knowledge. and decision making in organizations.”

Big data refers to data characterized by its volume (large quantity), speed (at which it is generated, accessed, processed and analyzed) and a variety of structured and unstructured data (OECD, 2016).

This data can be reported by machines and equipment, sensors, cameras, microphones, mobile phones, production software, and can come from various sources, such as companies, suppliers, customers and social networks. The analysis of this data through advanced algorithms is key to making decisions in real time, allowing to achieve better quality standards of products and processes, and facilitating access to new markets. Big data analysis plays a fundamental role in the decision-making process (Lescano, Lot, & Vasquez, 2020). Another use of this tool is to control and improve commercial and manufacturing planning. These data can provide information on hidden patterns, trends, associations, especially for human decision making; The term includes three concepts: volume, speed and variety (Deepa, Zongwei, Shan, Thanos, & Rameshwar, 2017).

Now, how can SMEs benefit from this technology to position themselves at a level that allows them to compete globally?

Various studies, including IBM, have analyzed the large number of big data applications, the scope of this technology is very broad, however, the analysis carried out by IBM shows the 5 preferred guidelines when applying big dates in organizations where 49% of organizations prefer to 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 (López, 2012).

The customer-focused processes of manufacturing companies can be considered as: Sales processes, distribution, market analysis, digital marketing, among others.

Based on the foregoing, a strategy is designed to be considered by manufacturing companies, highlighting that certain authors propose the use of an intensive data management platform such as Hadoop, which is a framework that supports applications distributed under a free license. (Sarkar, 2013).

Apache Hadoop software

Hadoop is an open source software framework that supports data intensive use for distributed storage and distributed processing of very large data sets in computer clusters; The Apache Hadoop database (TheApacheFoundation, 2019) is made up of several modules such as: the Apache Hadoop MapReduce application tool for programming and the Hadoop Distributed File System (HDFS) for infrastructure management, Figure 1 show frame components.

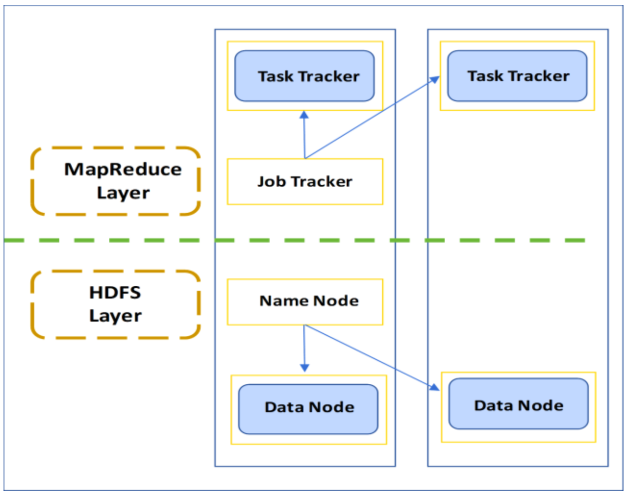


Figure 1 Components of the Hadoop framework
Source: (Sarkar, 2013)

Hadoop building blocks

Name Node: The main node or main node of the cluster, contains the metadata for HFDS during the processing of the data that is distributed among the nodes.

Data Node: These are the systems in the cluster that store the real HDFS data blocks, these blocks are replicated on various nodes to provide high quality solutions.

Job Tracker: Service running on Name node, which manages MapReduce jobs and distributed individual tasks.

Task Tracker: Service running on the data nodes, which monitors the individual MapReduce tasks that are submitted.

There are support projects for Hadoop, which have different roles in the systems, these are:

- Apache Hive: is a data warehouse software that makes it easy to read, write and manage large data sets residing in distributed storage using structured query languages (SQL) through a Java Database Connectivity (JDBC) driver.), which allows users to query data without MapReduce Application Development (ApacheSoftwareFoundation, 2019b).

- Apache HBase - This is the Hadoop database, a scalable, distributed big data warehouse that hosts very large tables (ApacheSoftwareFoundation, 2019a).

- Apache Mahout: is a distributed linear algebra framework designed to implement algorithms (ApacheSoftwareFoundation, 2019c).

- Apache Sqoop: is a tool designed to efficiently transfer data between Hadoop and relational databases; is a command-line tool that controls the mapping between tables and the data warehouse layer, translates the tables into a configurable join for HDMS or Hive (ApacheSoftwareFoundation, 2019d). Figure 2 shows the Hadoop support software.

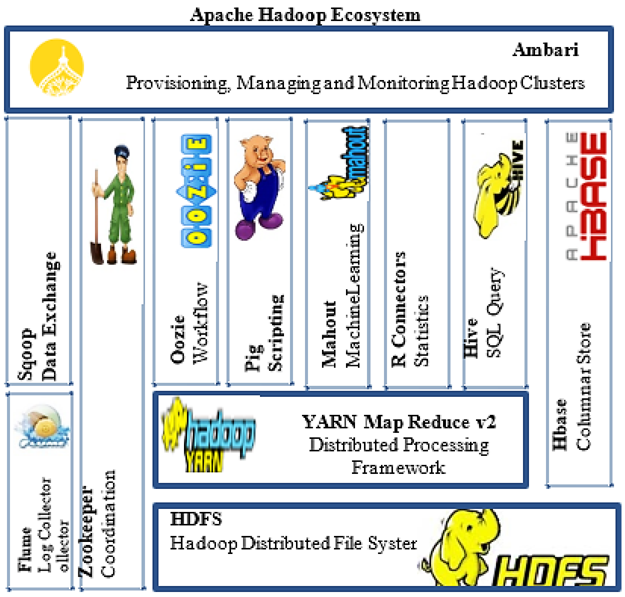


Figure 2 Hadoop Support Software
Source: (Sarkar, 2013)

Once the fundamental concepts on which big data technology is based and the software with which it is managed and managed are known, we proceed to the proposal and discussion of the technology strategy for big data based on Hadoop.

Strategy

The strategy is made up of 5 complex parts that involve different activities in each one. The first part refers to the analysis of the hardware technology required for the installation of the software and the data to be analyzed; A data server with the latest storage capacity and memory is recommended, as well as the computers that will be the client machines.

The second part refers to the selection of company processes that will be analyzed, they can be customer sales processes, production data, equipment failures, etc. From these selected processes, the necessary information and data that will be the raw material for the extraction, transformation and loading (ETL) activities are collected. In this case, it is a manufacturing company with a worldwide turn of suppliers of electronics and robots for industrial automation, located in the city of Acuña, Coahuila.

Subsequently, the strategy focuses on the Hadoop distributed processing platform, on this platform the data will be processed. Emphasizing that the Apache server and later Hadoop must be installed and configured; There are different components of this platform, among which Hive and Sqoop stand out, for the connector between the platform and the MS SQl Server database, which is where the data resides. The next activity deals with ETL activities which will be managed by the MS SQL Integration Services software, a data package will be developed with the Hive ODBC driver and the data model with Analysis services.

Once the activities of components 1 to 4 have been carried out, the following activity deals with data analysis and visualization of the results through Power BI software. This strategy is currently under development by computer systems engineering students and is planned to be applied to a medium-sized manufacturing company using data from the production of three-phase motors. Figure 3 shows the strategy.

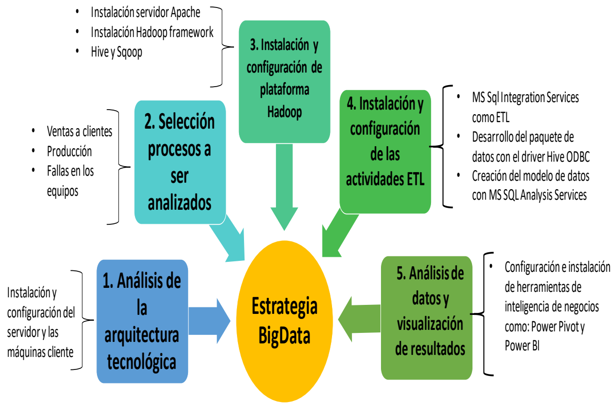

Instalación servidor Apache
Instalación Hadoop framework
Hive y Sqoop
3. Instalación y configuración de plataforma Hadoop
Ventas a clientes
Producción
Fallas en los equipos
2. Selección procesos a ser analizados
4. Instalación y configuración de las actividades ETL
MS Sql Integration Services como ETL
Desarrollo del paquete de datos con el driver Hive ODBC
Creación del modelo de datos con MS SQL Analysis Services
Instalación y configuración del servidor y las máquinas cliente
1. Análisis de la arquitectura tecnológica
Estrategia BigData
5. Análisis de datos y visualización de resultados
Configuración e instalación de herramientas de inteligencia de negocios como: Power Pivot y Power BI

Figure 3 Technological strategy for big data
Source: own elaboration

Methodology

The phases of the methodology for the development of the project, in which the activities proposed in the strategy were carried out, these being:

- Analysis of the technological architecture: In this phase the server for big data and the client machines are installed and configured.

- Selection of processes to be analyzed: In this phase, the company's data is collected, which are feasible to be analyzed with big data, and may be structured as well as unstructured data. For the project, the manufacturing process of three-phase motor components has been selected, among which are the plates (Peripheral Component Interconnect), which are assembled and tested according to the client's parameters; This is one of the processes that presents the most failures (42.7%), so when managing the production data with big data, it will provide information on how to improve the process and decrease the rejection points of the tablets. Figure 4 shows the graph of rejections.

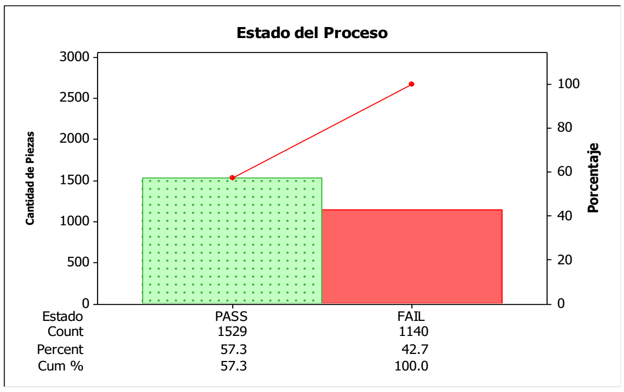


Figure 4 Status of the manufacturing process in a PCI
Source: case study company

- Installation and configuration of the Hadoop platform: This phase has been the one that has consumed the most time and resources since it is required to install the Linux operating system as a virtual machine using Centos Red Hat, virtualization of each node on the network, installation and configuration of Hive, Hbase and Sqoop as part of the software projects supporting Hadoop. Figure 5 shows part of the Hadoop configuration file.

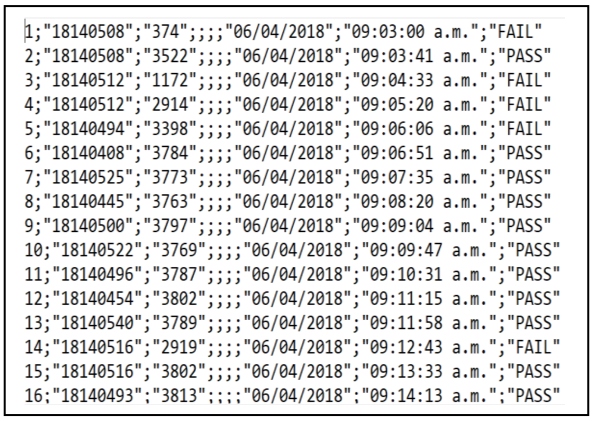
Part of the Hadoop configuration file

```
<configuration>
<property>
<name>
yarn.resourcemanager.opportunistic-container-allocation.enabled
</name>
<value>false</value>
<final>false</final>
<source>yarn-default.xml</source>
</property>
<property>
<name>yarn.ipc.rpc.class</name>
<value>org.apache.hadoop.yarn.ipc.HadoopYarnProtoRPC</value>
<final>false</final>
<source>yarn-default.xml</source>
</property>
<property>
<name>mapreduce.job.maxtaskfailures.per.tracker</name>
<value>3</value>
<final>false</final>
<source>mapred-default.xml</source>
</property>
<property>
<name>mapreduce.job.speculative.retry-after-speculate</name>
<value>15000</value>
<final>false</final>
<source>mapred-default.xml</source>
</property>
<property>
<name>yarn.client.max-cached-nodemanagers-proxies</name>
<value>0</value>
<final>false</final>
<source>yarn-default.xml</source>
</property>
<property>
<name>
```

Figure 5 Hadoop configuration
Source Hadoop configuration program

- Installation and configuration of Data Extraction, Transformation and Loading (ETL) activities using a JDBC and ODBC driver.

- Data analysis and visualization of results: An Access database with 7000 engine production records was used, in Figure 6 a part of the engine data is displayed in a file in .CSV format (Comma Separated Value).



```
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"
```

Figure 6 Data file part
Source: own elaboration

Each of the phases is a sequence of the previous one, so the activities of each phase were carried out in the proposed order. Figure 7 shows a PCI engine board.

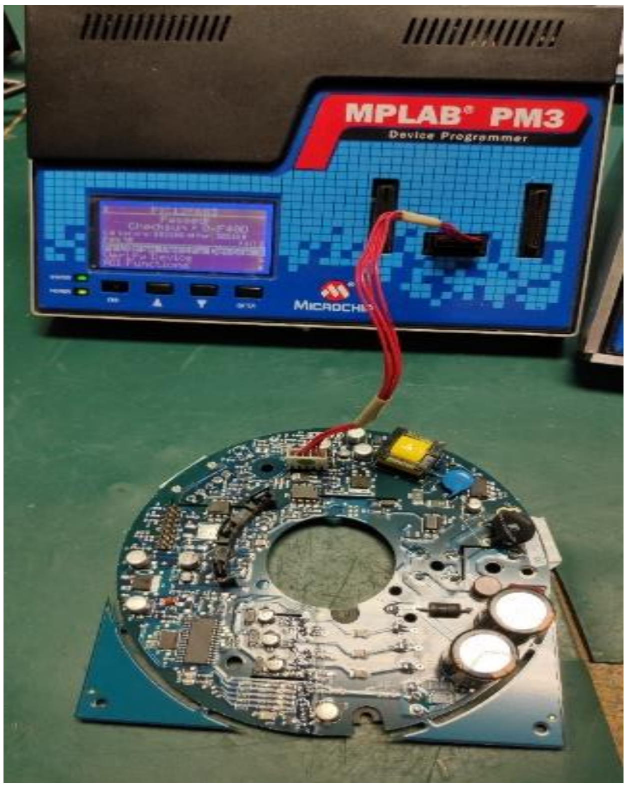


Figure 7 Parameter programming in a PCI
Source: company case study

Subsequently, the data extraction procedures are executed, obtaining the data from the source, where operational step stores can be installed that function as a bridge between the data source and the final node.

Among the activities to prepare the data package, the identification of the data source stands out, since they can come from different sources such as Oracle databases, Access, SQL Server, Excel or any other data source; so a very important activity is the unification in a single format to be transferred to the final node, provide them with a structure, process and analyze them. In this case it is an Access database, which was converted to CSV format, which can be read by Hadoop.

Once the data packet has been prepared and reviewed, the file is transferred from the user's physical directory to the Hadoop HDFS directory, since they are independent systems.

Once the information is integrated, queries are made on the database, creating and starting a session.

Results

The results obtained from this project have been the following:

- Training in the handling of the Linux operating system, since all the supporting software, as well as Hadoop work in the Linux environment, installation and configuration of Linux on the Proliant Gen 10th server.

- Installation and configuration of Hadoop, with the Linux command wget and configuration of the environment variables associated with Hadoop.

- Hive, Hbase and Sqoop installation and configuration; At this point each software has a different function and configuration, as well as the configuration of the environment variables for each type of software.

- Tests with the entire environment installed and configured with the server and 3 nodes.

- Identification of the manufacturing parameters of the slabs where the highest percentage of failures were concentrated, once the analysis was performed with Hadoop.

Figure 8 shows the graph of the improvement in the production of tablets, after making the changes detected by analyzing the data with big data.

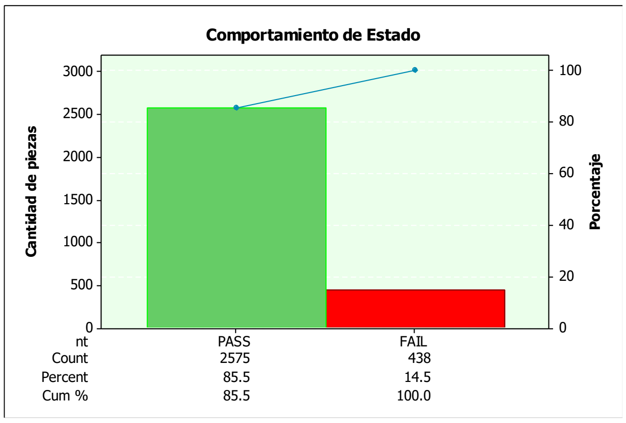
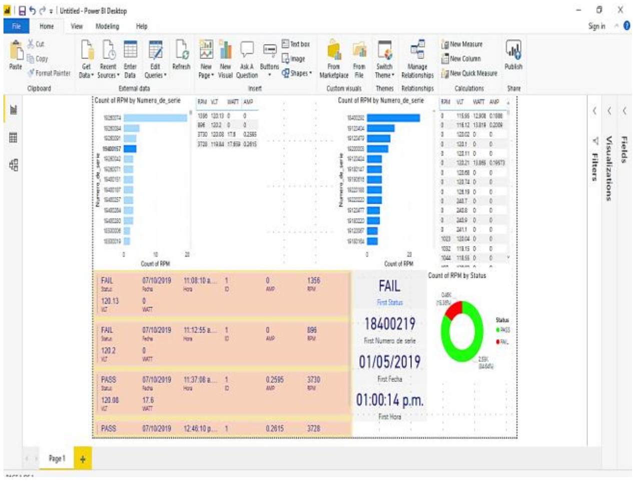


Figure 8 PCI parts state behavior
Source: case study company

Figure 9 shows the analysis processing with Power BI.



FAIL 07/10/2019 11:08:10 a.... 1 0 1356
Status Fecha Hora ID AMP RPM
120.13 0
VLT WATT

FAIL 07/10/2019 11:12:55 a.... 1 0 896
Status Fecha Hora ID AMP RPM
120.2 0
VLT WATT

PASS 07/10/2019 11:37:08 a.... 1 0.2595 3730
Status Fecha Hora ID AMP RPM
120.08 17.6
VLT WATT

PASS 07/10/2019 12:46:10 p.... 1 0.2615 3728

FAIL
First Status
18400219
First Numero de serie
01/05/2019
First Fecha
01:00:14 p.m.
First Hora

Figure 9 Data analysis with Power BI
Source: own elaboration

Conclusions

The current needs of the manufacturing industry are increasingly influenced by the adoption of new technologies based on Industry 4.0, which will allow them to improve the processes and products they manufacture.

One of the great needs of this industry is to have trained personnel who can design and implement customized solutions based on the new technologies of Industry 4.0.

SMEs are also facing an investment in hardware and software equipment that allows them to implement new technologies, which represents an extra expense in operating costs.

It is expected that in the coming years this technology will be more available to SMEs and that solutions will be focused on the main processes of SMEs.

The challenge is to achieve a roadmap with more detailed activities, especially in the technical aspect of the installation and configuration of the Hadoop and its supporting software projects to achieve effectiveness in the analysis of company data.

This project has been an exhaustive learning process between research professors and students since it works with two different software platforms, Windows and Linux, so making the data and technologies of both platforms compatible has been quite a challenge.

References

ApacheSoftwareFoundation. (2019a). Apache HBase TM Retrieved 10/01/2019, 2019, from tttps://hbase.apache.org

ApacheSoftwareFoundation. (2019b). Apache Hive TM Retrieved 10/01/2019, 2019, from https://hive.apache.org

ApacheSoftwareFoundation. (2019c). Apache Mahout TM Retrieved 10/01/2019, 2019, from https://mahout.apache.org

ApacheSoftwareFoundation. (2019d). Apache Sqoop Retrieved 10/01/2019, 2019, from https://sqoop.apache.org

Basco, A., Beliz, G., Coatz, D., & Garnero, P. (2018). Industria 4.0 Fabricando el Futuro. In B. I. d. Desarrollo (Ed.). Buenos Aires, Argentina: BID.

Beyer, M., & Laney, D. (2012). The Importance of 'Big Data': A Definition. In G. Group (Ed.). U.S.A.: Gartner.

Deepa, M., Zongwei, L., Shan, J., Thanos, P., & Rameshwar, D. (2017). A bibliographic study on big data: concepts, trends and challenges. *Business Process Management Journal, 23*(3), 555-573.

Lescano, A., Lot, D., & Vasquez, S. (2020). Analítica de datos para el soporte en la toma de decisiones en el área de distribución y ventas de la distribuidora farmacéutica la libertad SRL utilizando Microsoft Azure y la metodología LARISSA MOSS.

López, D. (2012). *Análisis de las posibilidades de uso de Big Data en las organizaciones.* Maestría, Universidad de Cantabria, España. (2012-1013)

Mayinka, J., Chui, M., Brown, B., Bughin, J., Dobbs, R., Roxburgh, C., & Hung, A. (2011). Big Data: The next frontier for innovation, competition and opportunity. U.S.A.: Mckinsey Global Intitute.

OECD. (2016). Enabling the Next Production Revolution: the Future of Manufacturing and Services In O. Council (Ed.). Paris, France: OECD.

Román, J. (2018). *Industria 4.0: la transformación digital de la industria*. Paper presented at the CONFERENCIA DE DIRECTORES Y DECANOS DE INGENIERÍA INFORMÁTICA, España.

Sarkar, D. (2013). *Microsoft SQL Server 2012 with Hadoop*. Mumbai, India: Packt Publishing.
Schwab, K. (2016). *The Fourth Industrial Revolution*. Switzerland: The World Economic Forum.

TheApacheFoundation. (2019). Apache Hadoop Retrieved 10/01/2019, 2019, from https://hadoop.apache.org/