v. 3, n. 3, 2023

Scientific Journal of
Applied Social and Clinical Science

QUALITY IN BIG DATA ICT ECOSYSTEM: BIG DATA AS A RESEARCH PROPOSAL APPLIED TO DATA QUALITY FOR THE CORRECT GOVERNMENT OF AN ICT ECOSYSTEM

Pedro Elizardo Donis del Cid

Abstract: The term "data" derives from the Latin DATUM and its meaning is what is given in the sense of what happens. Big Data is the terminology used to designate large volumes of data that are generated with a certain speed in the generation of information and variety in the format in which it is stored. The object of the article is to determine the factors that define the global quality of the information in terms of variables of information units by means of marking according to the typologies of errors presented in the methodology of this research. It is concluded that for the purposes of adequate data governance, it is necessary to consider small errors in the data sources so that the data can be used by the industry in different applications through professionals trained for this purpose. Being the valid research hypothesis where, there is the following proportional relationship: when the value tends to 1 there is high quality of the information and when it tends to 0 there is low quality, according to the observations made, this being a directly proportional relationship between the measurable variables and the categorical quality.

Keywords: Big Data, Hadoop, MapReduce, NoSQL, Data analysis and modeling.

INTRODUCTION

The term Big Data has been used in different technological contexts due to advances, however, in recent years it has had great value in terms of research related to this approach. Big Data exceeds the scope of commonly used hardware environments and software tools to capture, manage, and process data within a tolerable elapsed time (Teradata Magazine, 2011) for its user population. This is data that grows exponentially and that traditional systems, despite advances in technology, cannot support it (store, process and visualize) in order to obtain effective value in time from data analysis. Big Data refers to data sets whose size is beyond the capabilities of typical database software tools to capture, store, manage, and analyze. This definition is, according to McKinsey Global Institute (2011). According to Gartner, Big Data is large data sets that have three main characteristics: volume (quantity), velocity (speed of creation and use), and variety (types of unstructured data sources, such as social interaction, video, audio, anything that can be classified in a database).

In this article, the veracity as a property of Big Data is studied based on its quality as data and availability in a system with a constant integration speed, with both timestamps and volume or information units that are stored in electronic devices.

The method used is the use of Hadoop, MapReduce, NoSQL and BigData tools to generate important indicators that minimize the number of errors based on error typologies related to input variables and outputs of a specific processing in Big Data in different periods or cases. study. With variables that measure the relationship of the input volume measured in bytes with the output volumes with the same unit of measurement and the relationship that exists between each measurement and its antecedent.

The process works directly on top of the open source Apache Spark data processing engine created by Matei Zaharia at UC Berkeley. Programming in the Scala language.

THEORETICAL SECTION AND GENERAL CONCEPTS

IGN Group (2017) Erick Larson used the Big Data concept for the first time and made it a publication on a topic related to marketing and customer data. The word data derives from the Latin DATUM and means what is given in the sense of what happens, that is, background necessary to arrive at the exact knowledge of a thing.

There is an ISO standard for data quality, it is the ISO 8000 standard.

Important industries make decisions based on massive data, among which are transportation, food, communication, banking, etc. Thus, the technological platforms and the data that can be stored in historical warehouses manage to carry out correlational, predictive and historical analysis through statistical methods represented by algorithms and tools to compare effectiveness. According to research by Réda, et al. (2020), in Industry 4.0, code generation can be obtained from dynamic statistical processes from data without the need for these actions to be previously programmed.

Some of the areas that use Big Data that can be mentioned are banks, health, education, commerce, government, among others. According to da Silva (2021), it is confirmed that the sportswear and footwear company Nike offers garment services that monitor sports activity data while the garments are worn, which by evaluating the amount of data that consumers they enter per second, they can determine how many consumers make use of the garment with the service they offer. They also determine in which countries these garments are used the most through the input that consumers make in their social networks and at the same time they can determine what hours of the day is where information is most shared in real time. And at the same time it is verified that the data is true according to the data and the garments that have been sold.

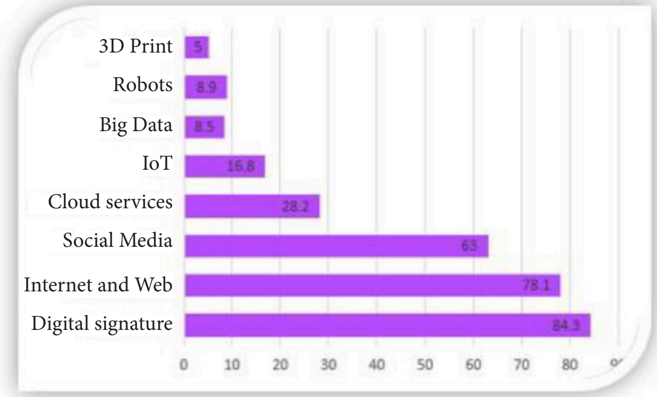
In the educational field, for example, a large volume of data is produced through online courses, teaching and learning activities (Oi, Yamada, Okubo, Shimada, & Ogata, 2017). With the advent of big data, teachers can now access student academic performance, learning patterns, and provide instant feedback (Black & Wiliam, 2018). Timely and constructive feedback motivates and satisfies students, which has a positive impact on their performance (Zheng & Bender, 2019). Academic data can help teachers analyze their teaching pedagogy and affect changes according to the needs and requirements of students. Many online educational sites have been designed and multiple courses have been introduced based on individual student preferences (Holland, 2019). Improvement in the education sector depends on acquisition and technology. Large-scale administrative data can play a tremendous role in managing various educational problems (Sorensen, 2018). Therefore, it is essential that professionals understand the effectiveness of big data in education to minimize educational problems.

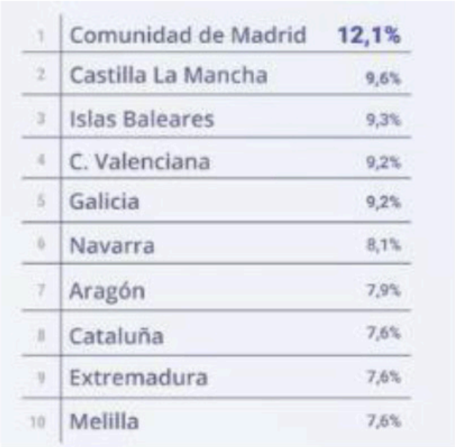
By 2021, 8.5% of Spanish companies use Big Data. According to INE data, 8.5% corresponds mainly to industry, services and construction. In addition, interest in Big Data according to Google Trends has grown since 2004. But in recent years it has skyrocketed. Among the technologies that this term enhances we have: Internet of things, industrial robots, service robots, analysis of large volumes of data, chat bots. But there is a need for specialized professionals. In Spain, for example, only 18.4% have ICT specialists and only 8.5% use Big Data analysis.

Big Data ICT specialists by sectors and regions (Graph 1).

INDUSTRIALIZATION

Galvão et al. (2022) Osmana and Ghirana (2019) define automation as the basis of industry 4.0 combined with Big Data. For which sensors and other specialized technologies are needed for machine-to-machine communication. Qaffas et al. (2021) the IoT constitutes an important advance for this process. Franke et al., (2016) Qaffas et





1	Comunidad de Madrid	12,1%
2	Castilla La Mancha	9,6%
3	Islas Baleares	9,3%
4	C. Valenciana	9,2%
5	Galicia	9,2%
6	Navarra	8,1%
7	Aragón	7,9%
8	Cataluña	7,6%
9	Extremadura	7,6%
10	Melilla	7,6%

Graph 1: Image adapted from: https://bigdatamagazine.es/2021-el-ano-del-big-data#comment10566

al. (2021) Tang et al. (2019) van Evert et al. (2017) conceptualize Big Data as an emerging technology given exponential growth in data volume and hardware limitations. Multiple nodes, cores, and memory work on the same task.

Qaffas et al. (2021) IoT automatically collects data for research only by configuring equipment and parameterizing systems.

Galvão et al (2022) have used the automation of an industrial process using Big Data techniques with Spark and Python. Modeling data, aggregation functions, and JVM objects with maps and reducers. The application of APIs of this type are used for the acquisition, processing and storage of data in order to achieve communication between machines.

It is now possible to predict (Galvão el al., 2022) using a variety of data from Big Data and IoT technologies, through ETL. Computer science makes it possible to collect, store and process the information that arrives from machines such as the IoT. Given current technology. The most important thing is the data because they are used for decision making, through the transformation into information and then into knowledge. Retrospective and advanced descriptive (Mohamed et al., 2019) uses information with unstructured and structured data to generate correlations with greater objectivity.

Qaffas et al (2019) The Internet of Things and Big Data Analytics for Chronic Disease Monitoring in Saudi Arabia, The Steps of Support Vector Machine Classifier, shows a case study where there are input variables which are defined as Hypertension disease data set and output variables comprising a classification of the SVM. SVM or support vector machine shows how it can be applied by means of the method: Loading the data set, randomly aggregating a training set of 20% and a test data set of 80%, in the case study method, a generation of classifiers based on the training data set is included, later, the classifier is trained, and the prediction applied to the sample is built by means of the classifier that has been trained.

The case study includes an evaluation of the classifiers based on the parameters, as well as a selection of their characteristics according to their weights; In order to apply this type of biotechnology in the resolution of computer-assisted problems in the classification of parameters or factors that affect hypertension, it is necessary to use IoT.

At the end of the investigation it was determined that the data based on IoT is very effective and promising, however, some optimization is necessary; Among the relevant results of the study, older people with diabetes are more likely to develop hypertension. Smoking plays a less important role than diabetes and older people must consume less salt. SVM produces better results than the C4.5 algorithm. Kumar et al (2021) conducted a study using a scalable intrusion detection approach using a Big Data Framework using classification methods such as KMeans, RUSBoost, and DT.

In the industry, frameworks such as Spark and Storm have been used, which are used in combination with ATCS. In these frameworks there are many configuration parameters in the case of Spark more than 180 parameters for applications, so in some cases it is inefficient or unavailable due to the handling of so many configurable parameters for different scenarios. Therefore, the need for autotuning for general Big Data processing frameworks arises. After the study, the automatic tuning of parameters improves the performance of the system to find the optimal configurations Li et al (2020).

Kibria et al (2018) describe the network elements in an interconnected environment where Micro BS, Drone BS, Macro BS, Massive MIMO, D2D, Cloud-RAN, BBU Pool, MTC, Multi-Access Connectivity, Ultra-Dense Network participate, V2X, 3D Beamforming, AP. All of these networks interconnect data servers, people, devices such as network-connected cars, smartphones, Wi-Fi networks, businesses, and multiple Smart City devices. For example, Massive MIMO is for connecting industries and Ultra-Dense Network for connecting multiple devices gathered in a small environment like a concert. V2X can use a network of devices on electric lighting poles to be able to control autonomous or driven cars without human intervention, through sensors and networks.

Qi and Tao (2018) made a study of industry 4.0 where cyber-physical systems are integrated for manufacturing. In the study, design, production planning, manufacturing and predictive maintenance are carried out through Big Data and Digital Twin. Taking into account that industrial automation is already a reality, it is now possible to generate information on these, both production and consumption that help manufacturers make better decisions regarding the product.

EXPLOTATION

Abdessatera et al (2020) have used a multivariate data analysis through logistic regression and a three-day survey to determine the physical and psychological health of young people, 55.5% of the 495 members were surveyed, of which 90% responded that they felt more stressed during the pandemic. This had a major impact on the quality of their work.

Van Evert et al (2017) carried out a weed control study using Big Data systems and within the methods used, NoSQL environments were used (‘not only’ Structured Query Language) because traditional systems are difficult to manage when partitioning the system across multiple machines. SVMs techniques were used and the data set supported by Big Data for threat control includes spatial information, landscape position, soil and climate characteristics.

Within the results obtained, it was obtained that, if it can be controlled by the factors of time, severity and location, it was also determined that in the growth seasons of the weeds is when these data must be captured in real time to obtain advances. In agricultural science in ICTs for the capture, storage of data, analysis and contributions to it, to implement a precision agriculture system it is necessary to have automatic ways to obtain or inject data, of which the most important are : remote sensing, mapping, soil analysis data, data from field explorers; the latter can be carried out through mobile technology; In addition to this information, information can be obtained in real time on the conditions of the weather season, whether it is sunny, cloudy, rainy, etc.

After these important data, the evaluation system must consider a data processing phase, it is here where computing plays a very important role, through automated statistical operations, thematic prescription maps are built. And for the implementation in the field through the input data to VRA, a construction of performance and data quality predictors was used that are stored in databases that support a large amount of historical information from various periods and also Reference libraries are built. In summary, this case study of large volumes of data requires a correct acquisition of the data for its subsequent processing.

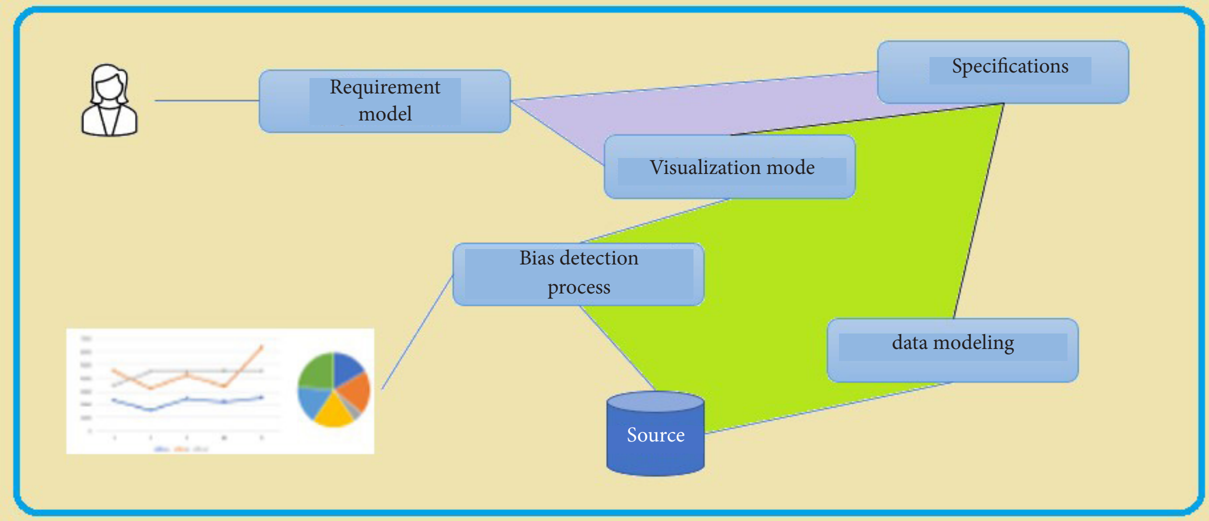
The implementation in the field obtained important findings: invasive, parasitic and herbicide resistant weeds must be controlled especially, but an effort of experts in computer science, data science and agricultural experts is necessary so that the entire system is automatic, functional and efficient in agricultural production. It was also noted that there are other organizational, ethical and legal aspects for the correct administration of the data. Finally, this case: data acquisition, data processing and implementation; Big Data historical data repositories or NoSQL databases need Big Data methodologies to make it platform-independent, scalable and open source, that is, it can be replicated to many demographic sectors; always in the context of agriculture as part of the modeling of the needs to implement Big Data systems.

Lavalle (2021) describes the visual analytics of large volumes of data as a crucial part of decision-making about data that requires a visual data model accompanied by requirements and a detailed description of the data sources. Among the agents involved, the requirements for the model given by the user, the visualization specifications, the data visualization model, the data sources and their corresponding model, and the bias detection process. All these implemented steps lead to the implementation and periodic monitoring of the results obtained, which are basically data visualizations that can be interpreted by the user. The visual elements will be the result of a controlled system of processing and ingesting information to the data sources, modeling, processing and specifications.

In order to understand the visualization of this data, it is necessary to understand the context of the data and the data model. Have a reliable source of Big Data data and the processes that facilitate the periodic availability of information.

DATA SCIENCE

Jamshidi et al (2020) have used DL for the diagnosis and treatment of COVID19 disease. In addition, they have used GAN, ELM and LSTM techniques. Using a bioinformatics approach where unstructured data is abundant and requires use by clinicians and researchers. The main advantage of the platform is to speed up diagnosis and treatment. Big Data is used as a repository of information for analysis. After using RNN, LSTM, GAN and ELM, it has been possible to identify propagation patterns, in addition to improving the speed and accuracy of the diagnosis. He assists subject matter experts in the development of effective new therapeutic



Graph 2: Image adapted from: http://hdl.handle.net/10045/119626.

approaches including the discovery of genetic and physiological traits that make a person vulnerable. It is important to note that the system is capable of collecting data from different sources such as airline ticketing systems, which enables Big Data volume ownership. The data sets are both clinical and non-clinical.

Heart failure is mainly studied due to the inflammation caused by the disease. Above all, in patients with cardiovascular problems, within the main results and through the RNN, LSTM, GAN and ELM techniques; there is a correct identification of high-risk people by analyzing large volumes of data and making this information available to experts, however, like everything, it has its advantages and limitations, so it is important to stock up on an arsenal of platforms, methods, approaches and tools to achieve the objectives.

Network operators have access to large volumes of data from the network and from subscribers, which is why this study aims to make network management selfadaptive, proactive and prescriptive for next-generation communication systems for BSs (macro, micro, femto, pico) given the current challenges in applying coverage with scarcity of both capital and spectrum. In this study, by expanding the variety of data sources, NWDA has been used, which requires more optimization effort. This achieves improvements in network planning and control, thus adding value to analytics. With big data analytics, better insight and understanding can be conveyed across multiple sources to reveal patterns and correlations. Achieving QoE. Kibria et al (2018).

INVESTIGATION METHODOLOGY

Varl is the variation of the data entry set for the process identified with percentage values and which responds to the name of reading variation, the number of information units of the previous brand is divided with respect to the current one.

Vare is the variation of the data output set for the process identified with percentage values and which responds to the name of writing variation, the number of information units of the previous brand is divided with respect to the current one.

Var is the variation of the set of the data output with respect to the data input. The process is identified with percentage values and responds to the writing and reading variation, the number of information units of writing is divided with respect to reading. The latter depends directly on varl and vare.

In addition to the variables, you must understand how to typify (classify) the errors, for which we have as initial values:

- Operational problems, error 400 missing data, 401 duplicate data.
- Quality errors identified as moderate: 402 low (very low values) and 403 high (values above normal).
- Severe quality errors: 404 outliers, 405 variation with zero value.
- Unclassified errors, lack of sample data Error406 which are excluded from the research analysis because they are implicit.

This same criterion applies to errors with the range 300 to 306 and 200 to 206; which correspond to other measurable variables from different perspectives.

So the problem lies when a job can be lifted indefinitely, for which it can have two final states suspended or removed from the ecosystem, in the case the Big Data computational tool throws an exception when unsubscribing or canceling a job, because an unexpected event occurred.

And the objective of the investigation is to study, classify and determine both the reasons for the unexpected event and the overall quality of the data set of the experiment that served the case study.

The data set resulting from Big Data processing generates some important indicators of network information, among them the client identified by its unique identifier in the mobile network stands out, from this indicator another arises as the number of different clients for a certain period of time to what is called the park or set of active clients. Obtaining this information directly from the main source would be a process with a high consumption of Big Data resources, given the frequency of use. The most important indicator is the amount of navigation traffic and its operational variations. Among them the application, the server and its identification, the classifying groups, additional rules, the communication protocol.

The operation of the variable and parametric definition of the case study for varl its range is [<=0.8] errors with low values and [>=1.2] for errors with high values, that is, the variable is in this acceptance range [0.8<varl<1.2]. For the other variable (write variation) we have the ranges: vare [<=0.8] [>=1.2].

DISCUSSION OF THE METHODOLOGICAL PATH

The hypothesis indicates that the values that represent an error in the quality of the information are represented by the values of varl, vare and var that tend to zero. That is, when the variables tend to zero, it is an error in the processing engine, and inversely when they tend to 1, the data presents greater fidelity.

To start the analysis of case studies we have the following: (502), this study tries to determine which are the experiments with the highest incidence of errors during loads in an information system based on Big Data technologies in a distributed file environment. (hdfs). For this, twenty-six experiments (represented by weeks of the annual calendar year) were analyzed, where each experiment is assigned seven similar dimensional groups (days of the week from Sunday to Saturday) and a set of jobs, processes or jobs, which are run in the Big Data environment each day and interact with the input data sets and generate an output data set. These experiments (Jobs) are executed sequentially, in total twenty-four for each dimensional group of the experiment, each dimensional group follows similar patterns based on the day of the week and each one represents a measurement that corresponds to input and output data for a given day following the methodological description. For this purpose, and following the Big Data methodology, each job has executors who carry out tasks with a set of input and output records. This is done through tasks (stages) of the job.

In the Big Data system, analyzing the error logs to determine the quality of the information; In the present case study, the behavior of the variable that determines the level of measurement or relationship between an input and an output of data in an information system is analyzed. The information system includes a set of data measured in input bytes and another output, this variable consists of the ratio value of these and some cases where there are apparently errors are filtered. The data set is identified with the year, experiment, day of the experiment, and time of the error occurrence.

The analysis of the following case study is carried out, this is part of the typology of errors in the Big Data system that consists of an input data set and an output data set, for which several variables that influence the performance are analyzed.

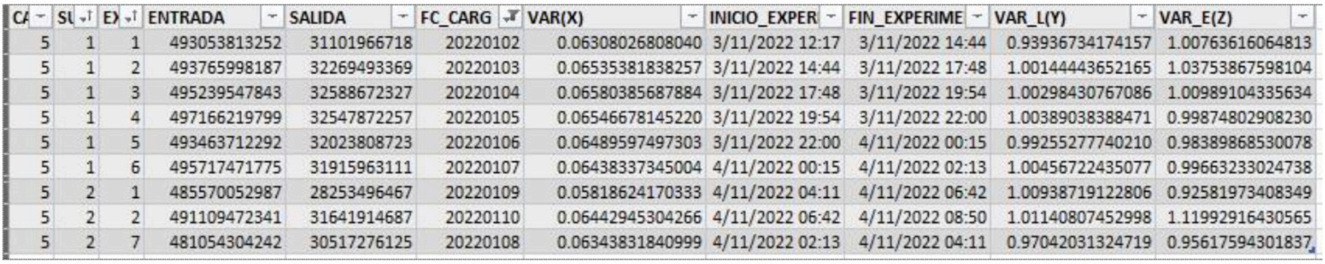
Behavior of this These variables have to do with the volume of the data, that is, the number of bytes, mb, gb or tb that correspond. These data volume measures are related to each other to construct the ratio variables discussed in this article.

When we speak of an experiment, we refer to a temporary space where the variable is studied with which the magnitude or measurement of the error or possible error can be evaluated. These temporary spaces are given by cycles, there are fifty-three cycles per year. This case study (502) has the following error and acceptance ranges: varl[<=0.8][>=1.2] vare[<=0.8][>=1.2]. That is, the acceptance range is the following: 0.8<=varl<=1.2 and 0.8<=vare<=1.2.

In this graph we can see the history of errors for each temporary experiment carried out, all the values that are given upwards correspond to ranges where the variable is typified in this error, it does not necessarily have to be an error, but the hypothesis affirms that it has occurred an error, however, the lower the value of this variable, the greater the possibility that it is an effective error and there is a significant loss of information in the job due to the causes presented in this investigation.

RESULTS AND DISCUSSION OF THE HYPOTHESIS

Below is a table of the measurements taken into consideration. Being the most important, entry and exit, date, and the identifier of the experiment; with other derived variables (x, y, z).



CA	SU	EX	ENTRADA	SALIDA	FC_CARG	VAR(X)	INICIO_EXPER	FIN_EXPERIME	VAR_L(Y)	VAR_E(Z)
5	1	1	493053813252	31101966718	20220102	0.06308026808040	3/11/2022 12:17	3/11/2022 14:44	0.93936734174157	1.00763616064813
5	1	2	493765998187	32269493369	20220103	0.06535381838257	3/11/2022 14:44	3/11/2022 17:48	1.00144443652165	1.03753867598104
5	1	3	495239547843	32588672327	20220104	0.06580385687884	3/11/2022 17:48	3/11/2022 19:54	1.00298430767086	1.00989104335634
5	1	4	497166219799	32547872257	20220105	0.06546678145220	3/11/2022 19:54	3/11/2022 22:00	1.00389038388471	0.99874802908230
5	1	5	493463712292	32023808723	20220106	0.06489597497303	3/11/2022 22:00	4/11/2022 00:15	0.99255277740210	0.98389868530078
5	1	6	495717471775	31915963111	20220107	0.06438337345004	4/11/2022 00:15	4/11/2022 02:13	1.00456722435077	0.99663233024738
5	2	1	485570052987	28253496467	20220109	0.05818624170333	4/11/2022 04:11	4/11/2022 06:42	1.00938719122806	0.92581973408349
5	2	2	491109472341	31641914687	20220110	0.06442945304266	4/11/2022 06:42	4/11/2022 08:50	1.01140807452998	1.11992916430565
5	2	7	481054304242	30517276125	20220108	0.06343831840999	4/11/2022 02:13	4/11/2022 04:11	0.97042031324719	0.95617594301837

Table 1: Own elaboration [2022], Data source: Field work [2022].

In order to guarantee the quality of the output without depending on the data input, the latter represents a much larger volume of information, it is necessary to verify the conditions that cause failures in the experiment. For the case [5,1] experiment [1] according to the estimates the data is complete. This quality can be observed through the trends for this observation.

In case study 4 and experiment 3, an incidence in the job of this experiment is shown. The problem is caused by a collapse in the Big Data system. Produced between 0 hours 14 minutes and 03 hours 25 minutes. The process was maintained for a time of 191.5 minutes. In the end, the following is concluded: there was a saturation in the processing queue, this is verifiable through the log or event history (H10.log). On the other hand, the case study 2 and experiment 7, and the case study 3 experiment 1, have failed for the following reasons: The queue (t) was saturated between 20 hours 46 minutes and 01 hours 06 minutes; and 01 hours 06 minutes and 05 hours 26 minutes.

When the waiting time expires, the job is eliminated and causes an absence of information. That is, the application has been finished. There were problems in the block manager. (Graph 3)

The mark 20220122 that corresponds to experiment 07 of case 04 has an operation error which was cleaned before starting the analysis. Remaining as follows. (Graph 4).

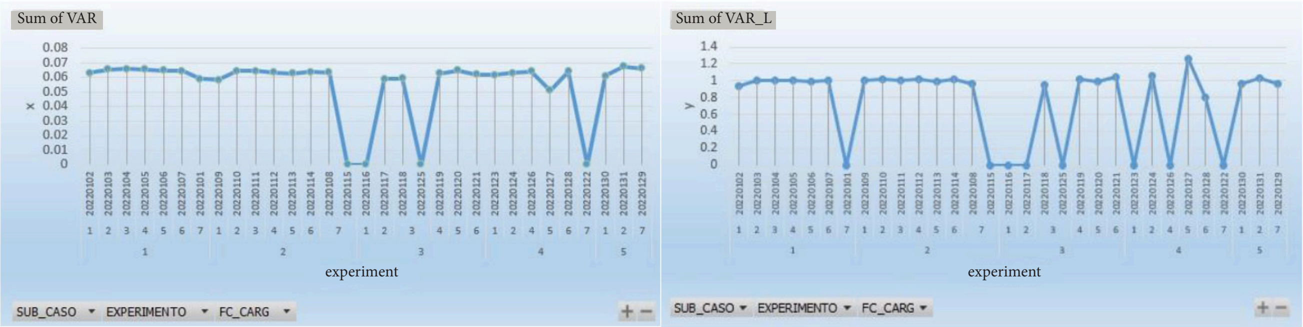
You can see the drops in the measurements made, these originate the typical cases of errors that are being studied and classified, in order to measure the quality of the data. (Graph 5)

In the graph on the left side, it shows an R square of 0.4416, lower, in relation to the R square on the right side of 0.9075, the first is related to the relationships that exist between the write-read variation and the reading variations in relation to the previous mark. The second is related to the write-read variation and the writing variation in relation to the previous mark.

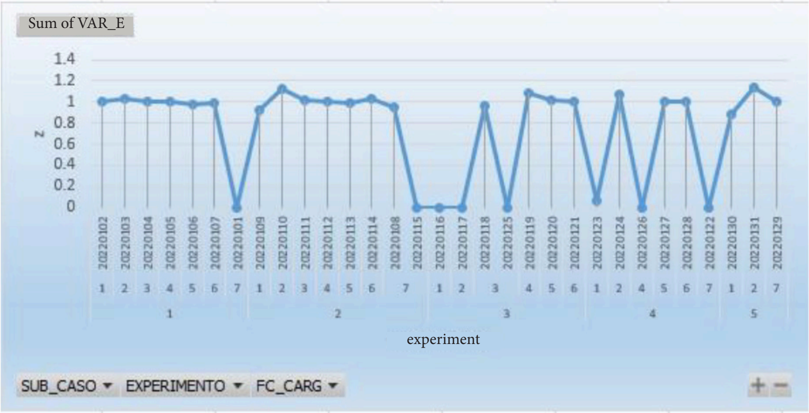
Below is a summary of the findings found for the errors that were categorized in each variable (x, y, z) in each case (p, q).

	X	Y	Z
1	7	6	6
2	7	7	7
3	5	4	4
4	5	4	4
5	6	5	5
6	7	7	7
7	7	7	7
8	4	2	5
9	7	6	7
10	7	7	7
11	7	7	7
12	6	6	7
13	7	6	7
14	6	5	5
15	7	7	7
16	6	5	7
17	7	7	7
18	7	7	7
19	5	5	5
20	5	5	5
21	7	7	7
22	7	7	7
23	7	7	7
24	7	7	7
25	7	7	7
26	6	6	6

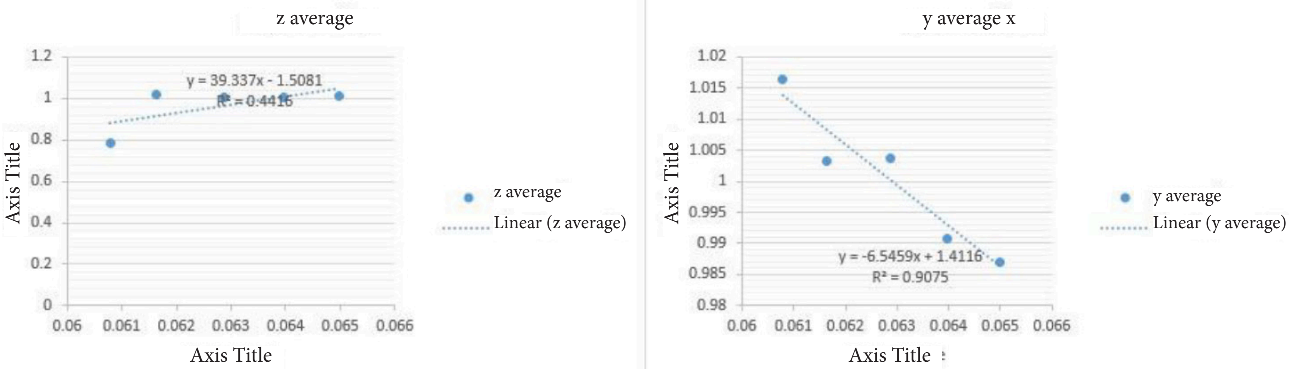
Table 2: Own elaboration [2022], Data source: Field work [2022].



Graph 3: Own elaboration [2022], Data source: Field work [2022].



Graph 4: Own elaboration [2022], Data source: Field work [2022].



Graph 5: Own elaboration [2022], Data source: Field work [2022].

Graph 6.

In each case study where the value of the variable is 7, there is no error, in the other cases if there is, according to the verification of the hypothesis and the results, it is concluded that the variables are representative in order to determine the typologies of errors presented in this study for the quality of the information.

Effective experiments for each case study. (Graph 7)

In the previous graph, the value of the variable [var] is placed on the x-axis. And on the y-axis the variable [varl] for this case study (p = {502}) where the sub-cases are q = {9,10,11} dates between March 01 and March 18 (partial sample). When carried out in full analysis, the following graph is presented: Graph 8.

The points that are grouped are those that tend to maintain the quality of the information and the scattered ones, on the contrary, represent typical errors classified in the methodological section, both of the variable var (x) and of the variable varl (y). (Graph 9)

The graph above shows the variable [var] on the x-axis and [vare] on the y-axis.

The equation of the line is denoted by y=23.561x-05374 with an R square of 0.4231. Taking into account the case study p = {502} and sub cases q= {9,10,11}, dates from March 1 to March 18, 2022. After carrying out a general analysis, the following graph is presented: Graph 10.

A correlation coefficient close to 0 indicates that there is no correlation between variables. This is because the values of vare are much more atypical than those of varl, this happens in a general way throughout the case study for this period of time.

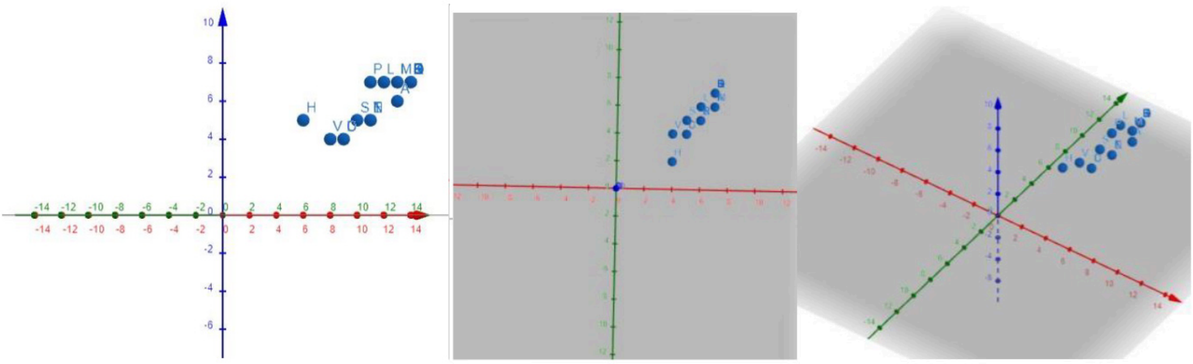
CONCLUSION

All research is based on facts or events which have been recorded or placed in a reliable source, therefore, it is important to verify their quality. Likewise, in the present study the quality of the information has been classified by means of a dichotomous variable, this manifests an important failure that triggers a saturation in the Yarn (Big Data) queues, this administrator tries to allocate the resources, but it is not possible. possible because it has reached the limit. Consequently, saturation, forced resource release, process kills, waste is almost inevitable.

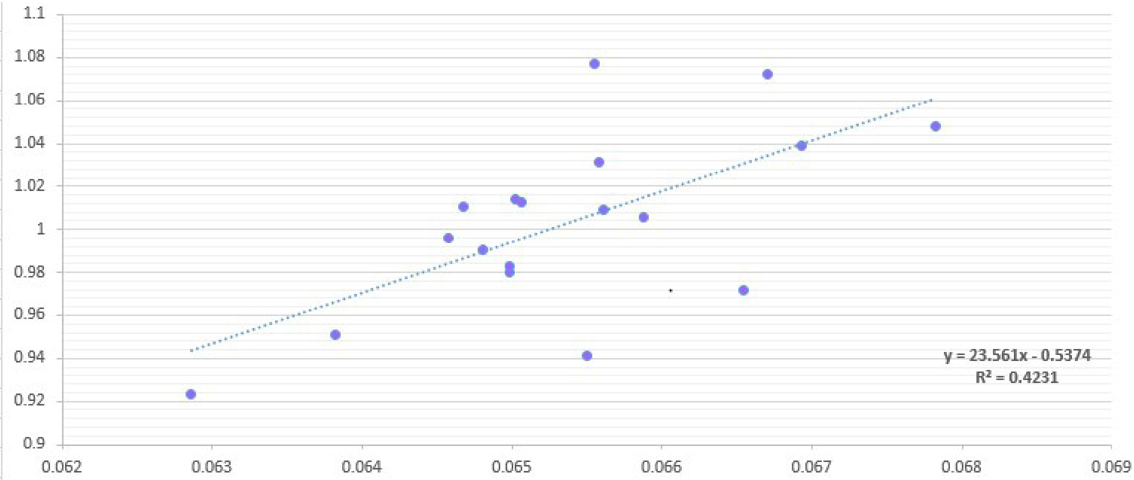
Data governance and data science require validation and purification of data sources so that they can be used by the industry through big data methodologies.

Tools are needed to measure the reliability of these data sets.

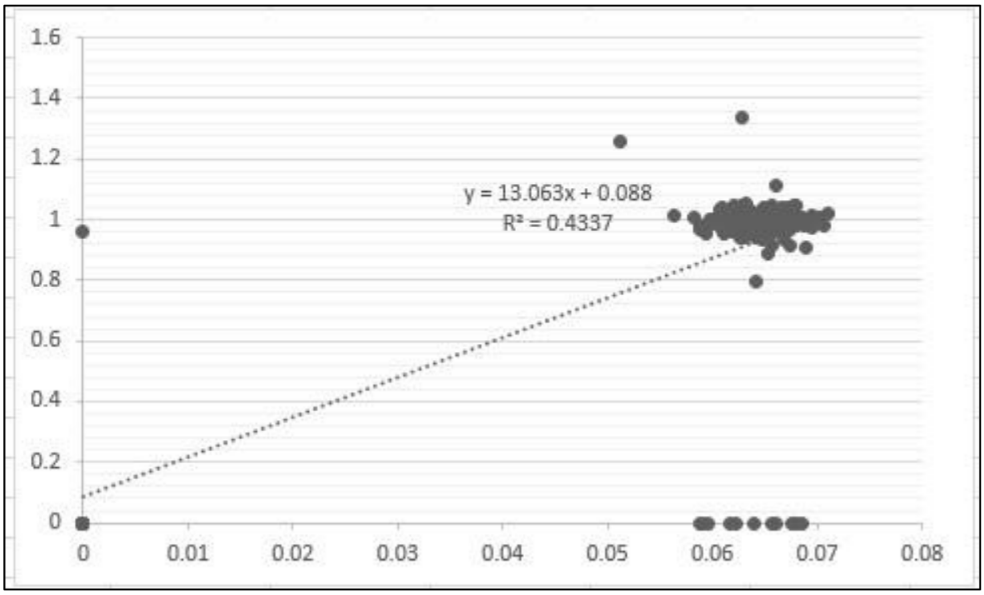
The hypothesis is verified, according to the twenty-six case studies and the analysis of results, it is possible to determine the quality of the information in Big Data environments through the analysis of the variables presented in this investigation, measurable and according to the graphs, these experiments tend to to maintain their quality if they are grouped at the same point. And those with low quality are grouped into one of the axes. With this, the reasons for the events that occur randomly and that affect the quality of the data set of the experiments in this case study were studied, classified and determined in a global way, it was classified into four categorical groups, being the main challenge the working time of the nodes.



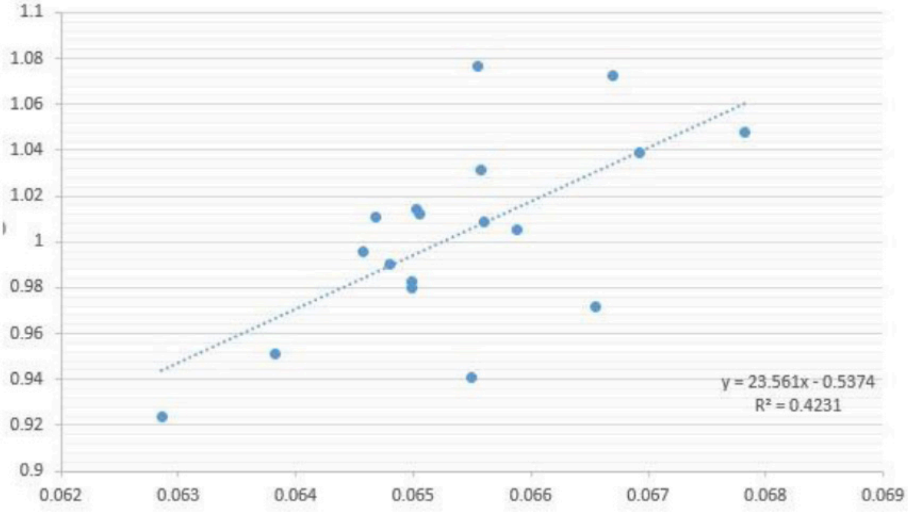
Graph 6: Own elaboration [2022], Data source: Field work [2022] Red x Green y Blue z**.**



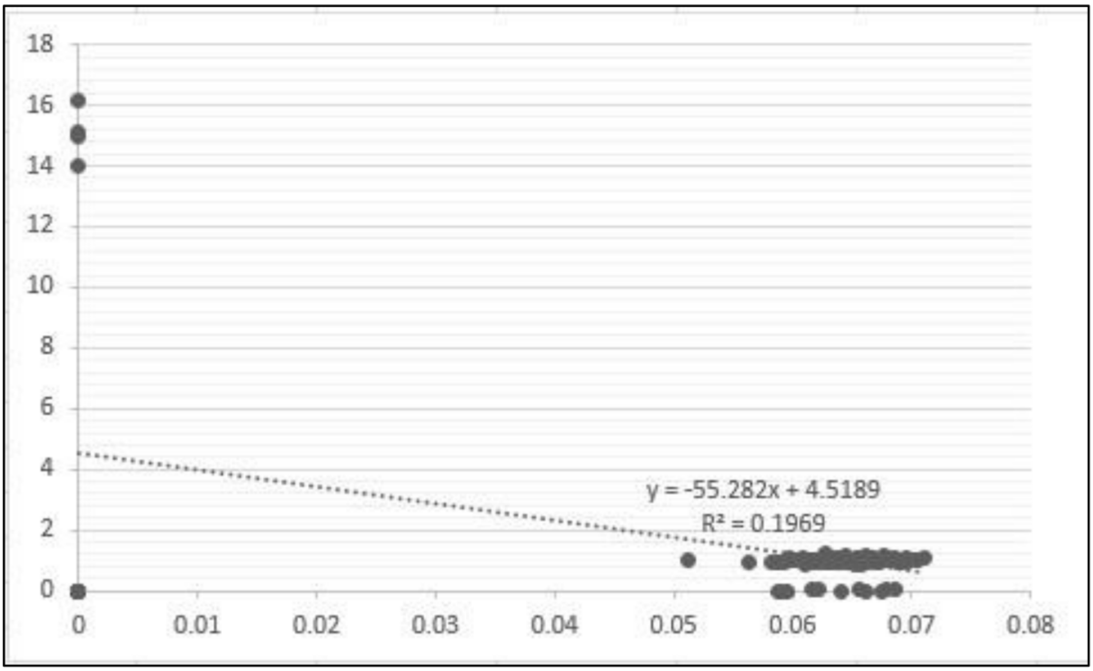
Graph 7: Own elaboration [2022], Data source: Field work [2022].



Graph 8: Own elaboration [2022], Data source: Field work [2022].



Graph 9: Own elaboration [2022], Data source: Field work [2022].



Graph 10: Own elaboration [2022], Data source: Field work [2022].

REFERENCES

Black, P., & Wiliam, D. (2020). **Classroom assessment and pedagogy. Assessment in Education: Principles, Policy & Practice**, 25(6), 551–575. Recuperado de: https://doi.org/10.1080/0969594X.2018.1441807.

Da Silva (2021) ¿Qué es el big data y para qué sirve? Zendesk, México. Recuperado de: https://www.zendesk.com.mx/blog/big-data-que-es/

Franke, B., Plante, J.-F., Roscher, R., Lee, E. A., Smyth, C., Hatefi, A., Chen, F., Gil, E., Schwing, A., Selvitella, A., Hoffman, M. M., Grosse, R., Hendricks, D., & Reid, N. (2016). **Statistical Inference, Learning and Models in Big Data. International Statistical Review**, 84(3), 371–389. https://doi.org/10.1111/insr.12176

Galvão J.; Ribeiro, D.; Machado, I.; Ferreira, F.; Gonçalves, J.; Faria, R.; Moreira, G.; Costa, C.; Cortez, P.; Santos, M. (2022) **Bosch's Industry 4.0 Advanced Data Analytics: Historical and Predictive Data Integration for Decision Support.**

Holland, A. (2020). **Effective principles of informal online learning design: A theory-building metasynthesis of qualitative research**. Computers & Education, 128, 214–226. Recuperado de: https://doi.org/10.1016/j.compedu.2018.09.026.

Kumar, S.; Prasad, D.; Kumar, J.; Sagar, K.; and Ashish Kr. (2021). **An EnsembleBased Scalable Approach for Intrusion Detection Using Big Data Framework**. Big Data. Aug 2021.303-321. Volume: 9 Issue 4: August 16, 2021 Recuperado de: https://doi.org/10.1089/big.2020.0201

Mohamed, A.; Nahafabadi, M.; Wah, Y.; Zaman, E; Maskat, R. (2019). **The state of the art and taxonomy of big data analytics: View from the new big data framework.**

Oi, M., Yamada, M., Okubo, F., Shimada, A., & Ogata, H. (2020). **Reproducibility of findings from educational big data. In Paper presented at the proceedings of the Seventh International Learning Analytics & Knowledge Conference**, (pp. 536–537). New York: ACM. Recuperado de: https://doi.org/10.1145/3027385.3029445.

Qaffas A.; Hoque R.; Almazmomi N. (2021). **The Internet of Things and Big Data Analytics for Chronic Disease Monitoring in Saudi Arabia**

Sorensen, L. C. (2020). **"Big data" in educational administration: An application for predicting school dropout risk. Educational Administration Quarterly**, 45(1), 1–93. Recuperado de: https://doi.org/10.1177/0013161x18799439.

Tang, Y., Tang, Y., Peng, Y. et al.(2020). **Automated abnormality classification of chest radiographs using deep convolutional neural networks. npj Digit**. Med.3,70. https://doi.org/10.1038/s41746-020-0273-z.

Wang, L., y Wong, A. (2020). COVID-Net: **A Tailored Deep Convolutional Neural Network Design for Detection of COVID-19 Cases from Chest X-Ray Images**, 1-12.

van Evert, F. K., Fountas, S., Jakovetic, D., Crnojevic, V., Travlos, I., & Kempenaar, C. (2017). **Big Data for weed control and crop protection. Weed Research**, 57(4), 218–233. https://doi.org/10.1111/wre.12255

Zheng, M., & Bender, D. (2020). **Evaluating outcomes of computer-based classroom testing: Student acceptance and impact on learning and exam performance**. Medical Teacher, 41(1), 75–82. Recuperado de: https://doi.org/10.1080/0142159X.2018.1441984.