 doi:10.20944/preprints202103.0640.v1

DBMS AND ORACLE DATA MINING

Bernal John Nicolás, Rodríguez Johanna Patricia, Portella Jorge
johnbernala@cun.edu.co, johanna.rodriguezmo@cun.edu.co, Jorge_portella@cun.edu.co

***Summary*- Databases are by far the most valuable asset of companies. Since the need was seen not only to count but also to have some type of record of elements such as crops, animals, money, properties and that this record could be consulted and modified according to the situation, that is where the first database was born. , and after that, these databases cannot be disorganized, they also need to be managed and administered under established standards that facilitate their understanding and management not only by their creators but by the other people who subsequently administer them. Databases and database management systems have an interesting evolutionary history that deserves to be analyzed and this is the objective of this document, where it is sought to understand,**

Along with databases and their management systems, data mining or Data mining arises that in order not to extend ourselves so much, it is the job of finding common patterns in various data sources and in what way they can be used to predict situations or results of various circumstances; We also focus on the other topic that we will present, Oracle data mining, which roughly is to merge data mining with Oracle, which makes it a powerful tool for obtaining information and predicting results based on statistics.

In this article we will study and analyze the ideas, concepts and basic examples that make up SGBD and Data Mining and, we will try to go deeper into this topic, the use of decision techniques such as advanced statistical algorithms. We also present a fictitious example of the application of these techniques: predicting which products can be sold based on their relationship with others. we will give a brief explanation of association rules, data mining cycle and the types of learning and the evolution that data mining has had

***Keywords*- Databases, database administration, database management systems, counting, storage, structure, search, No SQL, SQL, Oracle, relational databases, non-relational databases, magnetic tapes, punched tapes, relational model, Datamining, BigData, Datawarehouse**

I. GLOSARY

No SQL - Database management systems looking for a non-SQL approach

SQL: Structured Query Language

Oracle: database management system most used today, multiplatform

Query: Query

IoT: Internet of Things) or Internet of Things

GPL: General Public License

ODS (operational data store)

Data Smart: intelligent transformation of Big Data data

ODM: Oracle Data Mining

OLAP: (online analytical processing) allows users to easily and selectively extract data and view it from different points of view

BigData: Large-scale and more complex data set that is being generated and that requires both more storage and more specific processing to "discard" trivial data and retain relevant data

Datawarehouse: Dataware is nothing more than a large data warehouse and can have a very diverse architecture, can be based on an ODS or have multiple data smarts and has multiple layers.

II. INTRODUCTION

This document is our third deliverable of the subject, where it is desired to explore in the study materials presented by the teacher the different themes that will allow us to deepen our knowledge and internalize according to our particular interest, this at an academic level.

Data management has become an essential discipline in the modern world that is intertwined with other areas such as calculation and statistics, since factors such as the exponential growth of a database are considered, this is done by time to start creating this; the key is the organization and vision for the future, that the base allows flexibility of changes to new technologies and forms of administration. Likewise, there is an important evolution in the management of databases that seek non-traditional forms of management thinking about the best management of it and that currently has incredibly huge volumes of information that can be exploited as a good and This is how the companies that own this information are seeing it; but it is not about storing it because yes,

The objective of this work is to analyze the evolution of the concept of Database Management Systems, Data Mining and Oracle Data Mining and determine some factors that must be considered to evaluate the feasibility of using this technology in organizations that already They use conventional Data Mining processes. For this, Data Mining is the main tool, of Computing, thus demonstrating the benefits that it can provide to the organization as its good practices that must be taken into account for the use of Data Mining.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2021 doi:10.20944/preprints202103.0640.v1

I. Conceptualization on DBMS and Oracle Data Mining

Any piece of data can be information (a document number, an email, a telephone number, an address) but if this information is not related to something or someone, it will be empty, meaningless information that has no relevance; On the other hand, if it is ordered within a set with other data duly refined, classified and organized, it will be useful information and with a specific objective: data of a student or worker, of a borrower, of the population of a city , of the financial movements of a company, etc.

A database is a set of data related to each other and organized in structures, plus some programs that facilitate its administration and interaction with its users and that allow both to obtain in a simple way the data that is stored there as well as its management: previously they existed in physical formats (paper, punched tapes and magnetic tapes), today they are mostly handled in digital form (disks, USB, physical servers and web servers).

The databases as such began to be created at the end of the 50's of the last century. They started as hierarchical databases, moving to relational ones, with structured language queries, which led to the creation and strengthening of the SQL language; Currently another approach is being sought, creation of non-relational databases or No SQL (No only SQL), although relational databases still predominate and it is considered very important to know about their management, there is still high demand in the labor market to know about the handling of databases in SQL.

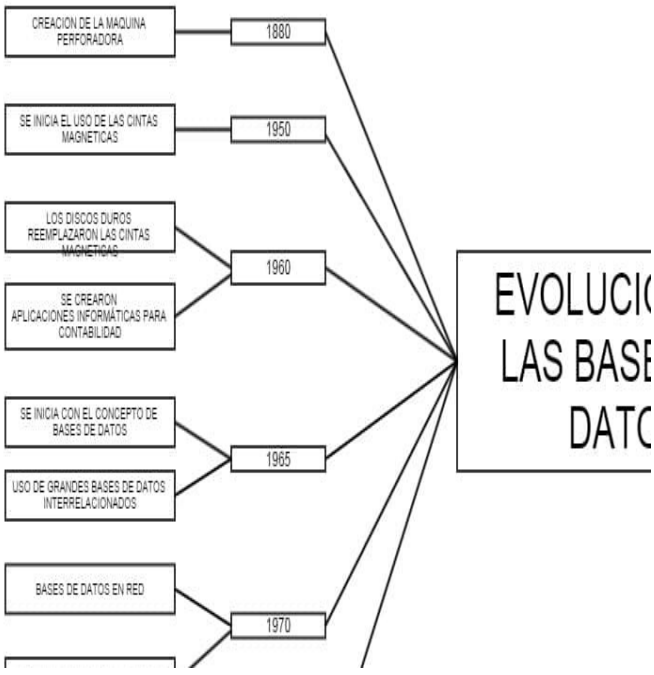


Fig 1. Evolution of databases

II. Evolution of Database Management Systems

[1] A database management system can be defined as "a coordinated set of programs, procedures, languages, etc. that provides, both non-computer users and analysts, programmers or the administrator, the necessary means to describe, retrieve and manipulate the data stored in the database, maintaining its integrity, confidentiality and security. " (accessed February 16/2021)

CHARACTERISTICS OF THE BD

A. INTEGRITY: That the storage of the data is properly managed and its conservation is guaranteed

B. SECURITY: Access and manipulation restrictions according to the type of user, that unauthorized users cannot access the data

C. CONFIDENTIALITY: data must be available for access, but within the limits for each user.

DBMS ARCHITECTURE

The architecture of database management systems (hereinafter DBMS) in 1975 was defined in four levels, which seek to separate the physical DB from the application components

1. Internal or physical level: or the closest to it is the data that is stored in the computer.

2. External or display level: it is the interface where the user interacts with the DB or with the part of it to which access permission has been granted.

3. Conceptual level; where the conceptual model is described with its details such as structures, entities and relationships, restrictions, is the information contained in the DB, but outside the computerized approach

4. Logical level: Represents entities and relationships from the computing point of view, already in the application used for it.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2021 doi:10.20944/preprints202103.0640.v1

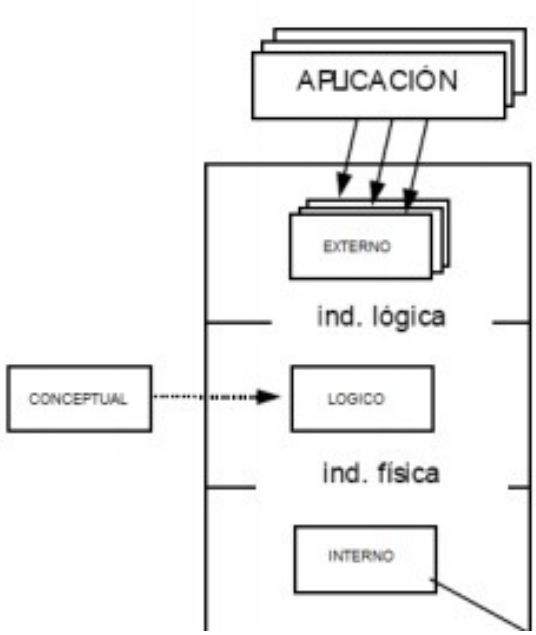


Fig 2 Architecture representation of a DBMS

COMPONENTS OF DBMS

There are three main components in DBMS:

1. Data dictionary: they consist of metadata that define at a physical level the characteristics of the objects, their use restrictions and modifications, tables, space allocated and used by the objects.
2. Data manipulation language: o DML (Data Manipulation Language) allows, as its name says, to manipulate the data directly with its modifications, queries and exports.
3. Data definition language: o (Data Definition Language) is used to structure the content of the database, read and update the data.

DBMS: MOST POPULAR SYSTEMS

Among the many database management systems that exist, these have been some of the most popular:

Microsoft Access (relational): it is a database manager included in the professional version of Office

Microsoft SQL Server (relational): Microsoft alternative to the most used systems

MySQL (relational): open source database manager

Oracle Database (relational): another of the most used systems

OrientDB (document-oriented) is non-SQL, supports graphs and documents and is developed in java

CouchDB (document oriented): No open source SQL based on web compatibility

IBM Db2 (Relational) Developed for high performance with multiple features and is cross-platform

IBM IMS (hierarchical) the first database manager in history

IBM Informix (relational): scalable and self-manageable database focused on the IoT (Internet of Things)

MariaDB (relational): based on MySQL, but GPL licensed

Sybase ASE (relational). Star database engine from Sybase company

MongoDB (document oriented) No SQL based on open source and offering great scalability

PostgreSQL (combines relational and object-oriented) has an open source license, it is one of the most powerful on the market

Firebird (relational) SQL-based, cross-platform

Databases date back to ancient times, as we cited in the abstract when human beings saw the need to document the quantities of various elements; With the emergence of numbers, writing and the first representative symbols, it was only necessary to have a place to record these counts and keep them; This was the first way to manage a database, logically it was done manually and rudimentary, according to the knowledge and tools available at the time. These processes were later made simpler with the invention of paper and printing, although they continued to be counted and recorded manually;

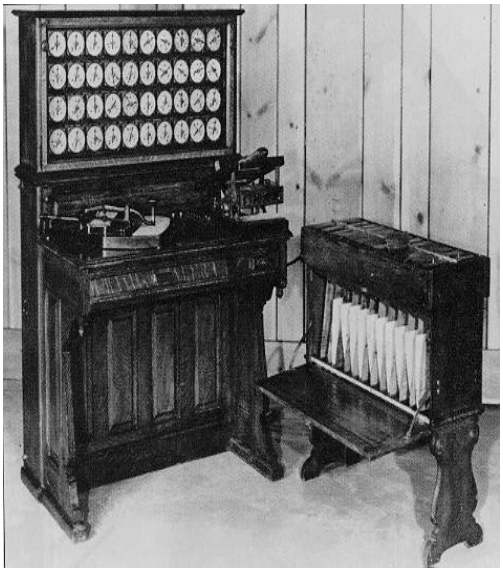


Fig 4 Hollerith tab

The electronic punch card tabulator exists thanks to Herman Hollerith who invented it in 1884 and was used until the 20th century, being a great support for accounting. Until then, the databases had their records on punched tapes.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2021 doi:10.20944/preprints202103.0640.v1

Fig 3 Magnetic tapes

In the twentieth century, we then went from 1950 to 1960 to the transition from punched tapes to the work of converting and using magnetic tapes for data storage. The processing consists of taking data from two tapes and writing them on a third tape; This is where the automation of the information really begins, with the disadvantage of not being able to modify data since it was only done sequentially, that is, new records had to always be created if it was necessary to make changes to the data and replace the tapes. old ones to avoid data redundancy (this was the "first standardization of the databases") According to the need, a file was rewritten with the new instructions of the processes that had to be carried out as who says to rewrite the code.

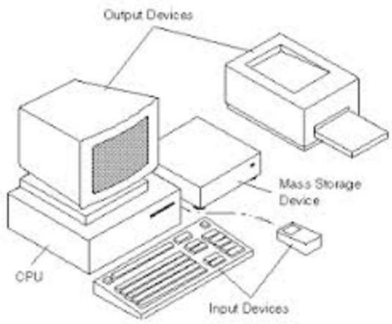


Fig 5 first computer for sale in 1963

In 1960 the definitive step towards the computer age was taken due to the fact that the prices of computers fell and they were "more accessible" although initially only for large companies and these same large companies were the first to encourage the creation of management systems of databases: the first hard disk was created in 1956 and over the years the use of hard disks whose stored information could last over time was an undeniable commercial attraction. Work also began on network databases and the creation of information languages to structure and standardize these databases. Here arises the CODASYL (Conference on Data Systems Language) entity in charge of creating a standardized computer system for all computers. They worked in the COBOL language, but the final standardization came with ANSI. In 1961 the first DBMS emerged as such, created by William Bachman in General Electrics between 1965 and 1969, the IBM IMS / 1 emerged, which has a great capacity for processing information, although they force the database manager to search record by record .

In the 70's Edgar Cood, an IBM worker, founded the concept of relational databases with the idea of working with tables of records with fixed sizes, this was the second generation of DBMS. Based on this, Larry Ellis develops the Relational Software System, now known as Oracle Corporation, creating the database management system we know and emulates the name of his company, Oracle. This system allowed for greater physical independence. logic and flexibility. Oracle is still in force today as another pillar of information management systems considered one of the most complete and reliable and can be used with the SQL language that allows direct access to objects and is multiplatform (it can be used in Windows , Linux, Mac, Unix etc.).

In 1980, being standardized in 1986, the main SQL programming language or structured query language based on tables of columns and rows emerged that allows us to perform relevant data searches, simple modifications on the database or declarative queries, displacing existing network and hierarchical models. SQL is still the most used tool in DBMS today.

In 1986 Microsoft also entered the database market: with Access and Excel, they started an object-oriented database model where SQL relational databases cannot have a field of action, these systems also keep up to date. date, being the third generation of DBMS.

In the 90`s with the arrival of the Internet now publicly (the World Wide Web) the database market grew much more than in previous years; Now it was necessary to integrate the databases with applications and services on the web because many more users had their computer and required access for consultation and modifications to these databases (client-server architecture) and the office applications it requires for this end. Internet access will greatly facilitate access to these databases and will further simplify people's work and force computer manufacturing companies to create equipment that can meet the requirements of these users in terms of storage, capacity, performance and speed and telecommunications companies to improve and expand their communication network system. In 1992 Microsoft launched SQL Server.

At the end of the 90's ANSI modified its codes focusing on the SQL language, adding improvements and new features; With object-oriented programming, it is necessary for DBMS to be merged and integrated with the Web, therefore, it is necessary to include SQL with languages such as C ++, Java and in HTML codes. MySQL then arises as a way of integrating databases with the web environment in XML, which is a language used for encoding documents.

Already in the 21st century, three large companies in the database industry are strengthened: Microsoft, Oracle and IBM, and on the Web side Google has a great advantage over its competitors in the creation and management of information. Freebird is released, and other Open Source models. There is also concern about the limitations given in the SQL language due to the accelerated growth of IP traffic and the volume of information, making its administration difficult: SQL and its relational model do not always adapt to the growing needs of the market and then the No model appears. SQL in which companies are becoming interested and beginning to invest.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2021 doi:10.20944/preprints202103.0640.v1

III. DBMS NO SQL

[6]Among the most notable characteristics of the No SQL architecture cited in the online document "Redalyc.Utility and operation of No SQL databases" (consultation of February 20/2021) we can mention:

• *Horizontal scalability*: referring to the easy to add, delete or perform operations with elements (hardware) of the system, without affecting performance.
• *Distribution ability:* has to do with horizontal scalability, but emphasizing on its support; for this, the ability to replicate and distribute data on servers.
• *Efficient use of resources:* take advantage of the new technologies such as solid state drives, efficient use of resources such as RAM and distributed systems in general.
• *Scheme freedom:* not having a scheme rigid allows more freedom to shape the data; also facilitates integration with object-oriented programming languages, which avoids the mapping process.
• *Weak concurrency model:* does not implement ACID (Atomicity, Consistency, Isolation and Durability), which has the characteristics necessary for a series of instructions can be considered a transaction, without However, some are taken into account considerations to ensure these aspects, but they are not so strict.
• *Simple queries*: queries require less operations and are more natural, therefore, wins in simplicity and efficiency.

IV. WHAT IS DATA MINING?

In this article we will study and analyze the ideas, concepts and basic examples that make up Data Mining and, we will try to delve into this topic, the use of decision techniques such as advanced statistical algorithms. We also present a fictitious example of the application of these techniques: predicting which products can be sold based on their relationship with others. We will give a brief explanation of association rules, the data mining cycle and the types of learning and the evolution that data mining has had. Also called data mining, I will try to give a brief explanation about what it is, what it consists of and what it is used for.
Data mining arises from the current need for companies to extract information from the large amount of data they obtain from their users and customers, but not just any information; As we said before in DBMS, any data is information but if it does not have a specific purpose or if it is not associated with something, it is no more than a simple data. Data mining is based on a set of techniques and technologies that are used to explore large databases automatically in order to find patterns, trends and rules that try to explain this data to be used in a specific context and also power understand it; is based on the use of statistical techniques and highly advanced algorithms approximated to artificial intelligence
Said in this way, what is learned from the data fits and is right with them as a person would do, all this arises as a result of the very well-called fourth industrial revolution that refers to the term as such of mining but not mining Traditional, such as looking for precious metals or stones, on this occasion they seek to find information that is attractive for companies where such patterns and behaviors can be found, which in this case would be of great value for companies in financial terms.

Now, after briefly explaining what data mining is about, we will enter another topic called association rules.

V. ASSOCIATION RULES:

They are techniques used in data analysis that use analysis for classification as a tool. regression, prediction, selection and detection of characteristics, anomaly detection, these rules allow us to find the most frequent combinations of items, for example when buying an item online the algorithm suggests another similar type of popular product or that already others people who also bought or consulted said product.
Association rules start from a database and can be represented in 3 ways: list, vertical representation and horizontal representation.

A list basically represents each row as a transaction where each row lists the products bought by the consumer also each row can have a different number of columns
The second way (vertical representation) tells us that it should be represented in the following way: first that it should only occupy 2 columns in the first column should be the number or the ID of the transaction and in the other column the product as such is the most efficient way to store large databases with lots of data.
In the third way of representing (horizontal representation) it tells us that it must be represented as if it were a binary matrix
Each row of the matrix represents a transaction, each column represents an item, for example if an item is present in a transaction it should be represented as a 1 and if the item is not present it is represented as a ° 0 ° following the binary system. This is the most orderly way but at the same time it is not the most efficient the problem is that too many 0 are used, it works very well for small databases but for larger databases a lot of space is wasted if we use this way there are too many blank fields continuing with this topic we are going to explain how in (1994) Agrawal discovers an efficient method to find these rules
But at the same time the problem arises that with this method too many rules arise, therefore we must find out how to limit this number of rules that we can discover for this purpose we will use a series of metrics which are:
Medium. this metric tells us the following (given a rule “if A = <B, support this rule is defined as the number of times or relative frequency with which A and B are displayed together in our transaction database. Support can also be used for individual items but can also be defined for a rule.
The first requirement that we are obliged to use to limit the number of rules is that they have a minimum support
For example, suppose that we have a database with 10 thousand transactions where 4 thousand of them include an item that could in this case be an item that for this example it can be a pencil in this case we denote it as follows: support (pencil) = 4000 or support (pencil) = 4000/10000 = 0.4
Given this notation, we can realize that this speaking in probability language where P (pencil) is a marginal probability, that is, it is the possibility that the pencil item appears in a transaction.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2021 doi:10.20944/preprints202103.0640.v1

Continuing with the example we have another individual item that in this case we will call "notebook" which in our 10,000 transactions appears 500 times so we will calculate support for this new individual item in the following way support (notebook) = 500; support (notebook) = 500/10000 = 0.05; P (notebook) = 0.05 is another marginal probability now, we realize that when the 2 items appear together, what would be for the case in 400 transactions looks like a "pencil" with a "notebook" where that 400 will become our support (pencil and notebook) = 400 any rule that we build from this point will have 400 transactions as support; support (pencil and notebook) = 400/10000 = 0.
Now the metric called confidence for which we will continue to use our example of (Pencil And Notebook)
What is the confidence of "Notebook => Pencil"?
What is the confidence of "Pencil => Notebook"?
When trying to understand what trust means, the implication that this entails can be denoted as follows: Trust (Pencil => Notebook) = Support (Pencil => Notebook) / Support (Pencil) = 400/500 = 0.8 what we are trying to analyze here is how important these transactions are with respect to the antecedent of the size of the ruler intuitively there is an 80% chance that these 2 items will appear since typically the pencil goes with the bone notebook that typically people buy pencil with notebooks
Let's analyze the other case the trust of a notebook with a pencil in this case we want to measure the relationship between those 400 transactions but now against the 4000 transactions that pencil is present, which we will denote as follows
Confidence (Notebook => pencil) = Support (pencil => Notebook) / Support Notebook = 400/4000 = 0.1 is 10% much lower if we compare it with that of the previous rule which was 80%
The third metric that we are going to study is called lift which is denoted as follows
Lift: Quantifies the relationship between X and Y: - lift> 1: positively correlated X and Y - lift <1: negatively correlated X and Y - lift = 1: independent X and Y. li () () ()
Once this topic has been briefly explained, we will go directly to the software developed by Oracle, which is nothing more than a powerful software that is incorporated into the Oracle data base that allows us to get more knowledge from our databases, which are hidden, this software helps companies to find more and better customers to detect and prevent fraud

VI. CYCLE DATAMINING

Data mining has been adapting to the day-to-day life of companies, governments of all countries in the world, educational entities, hospitals and many types of organizations that are constantly looking to explore their databases. to which we can say that "in data mining each case is a case". However, broadly speaking, the process is made up of five main stages:
This data mining has 5 stages or phases which are:

1. Determining objectives
2. data processing
3.determination of the model
4. the analysis of results
5.model update

We will explain what these stages consist of:

1. Determination of objectives: in the determination of objectives we will focus on what type of information we want to obtain, for example we have a shirt store and we want to know what style and colors would sell the most and what segment of the population would like to buy this type of goods.

This is the stage with which the process opens and is focused on understanding and analyzing the objectives and project requirements starting from the business perspective. Due to which, it is necessary to acquire this knowledge of the data (we insist, always from the business point of view) and turn it into the definition of a data mining problem, drawing up a preliminary plan according to the planned objectives.
2. Data processing: in this step, once we have the information, the matter is complicated since it is the most complicated phase since we must take a sample of this information from which we must carry out an analysis where we take into account the different variables that the possibilities and the forecast can be given
In this data preparation phase, we want to cover and take into account all the activities necessary to adapt our raw source data and approximate them and thus take them to the final data set (the data that will be the source and object of analysis of the modeling tools).
The data preparation or cleaning tasks will be done repeatedly and not in any order. Among these tasks we have the selection of tables, records and attributes, as well as the transformation and cleaning of data in preparation for modeling tools.

3. **Determination of the model**: it is closely related to the previous point since, based on the results we obtain, we must create our algorithm that satisfies our need, taking into account the different algorithm models such as: **linear regression,** decision tree, time series, neural network, etc. As we will see in the next paragraph, there are multiple data modeling techniques, being in this phase of the process when, after the knowledge acquired, the appropriate ones are selected (always according to the business and project objectives and needs) and apply. In this phase, the following four types of relationships are sought:
Lessons- Observations are assigned to default groups. Clusters: groups of similar observations are built according to a predetermined criterion.
Associations: observations are used to identify associations between variables.
Sequential patterns: it is about identifying behavior patterns and trends.
Among these techniques we can find: characterization or summary, discrimination or contrast, association analysis, classification, prediction, clustering or clustering detection, anomaly detection, evolution and deviation analysis. There are also different techniques for solving the same type of data mining problem. Some techniques have specific requirements on the form of data. Therefore, return to the data preparation phase to adapt these to the technique as necessary.

4.. **Analysis of results**: At this point, it is about analyzing the results and validating if it gives us a logical explanation against our need and guides us to make a logical decision that is easily applicable to our need, as a result of the previous stage, at this stage in the project a model has already been built. To ensure that the quality standards proposed for the project are met, it is necessary to evaluate it from a data analysis perspective. That is, before proceeding to the final deployment and its putting into production, it is important to carry out a test table together with the review of each step implemented in the creation of the model, which helps to compare the model obtained with the business objectives. or project.
5. model update: in this step we must see our previous model that we relied on to make decisions and constantly update ourselves with the results obtained with our algorithms since if it is not done they can become obsolete due to the constant change in the information collected
In this phase, the exploitation and use of the results of the Data Mining process that we have carried out will be carried out, depending on the requirements, it can be as simple as the completion of a report or as complex as the repeated performance of a cross-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2021 doi:10.20944/preprints202103.0640.v1

mining process of data across the company. Therefore, in many cases, it is the client himself, and not the data analyst, who performs the exploitation. Since he knows your needs more thoroughly, Data Mining is an iterative process, therefore, the creation of the model does not imply that the project is concluded. The knowledge acquired with the result of the process is perfectly usable again as input information to carry out another cycle of the process. That is, once the found knowledge is presented to the user, Evaluation measures can be constantly improved, mining can be refined, new data can be selected or transformed again and new data sources can be added. all this in order to obtain different results or more focused on the needs of our project.

VII. LEARNING TYPES

Supervised and unsupervised learning

Algorithms are separated into 2 categories that are supervised learning and unsupervised learning which we will describe below.

Supervised learning: this kind of analysis

You need a data analyst to identify a target attribute or dependent variable (for example, customers who purchased a specific product or item), then meticulously examine the data to find patterns and relationships between other attributes and the target attribute (for example, the characteristics that indicate whether a potential customer is likely to purchase a specific product or item). Supervised learning algorithms with Oracle Data Mining include Natïve Bayes, Decision Tree, Generalized Linear Models, and Support Vector Machines.

Unsupervised learning:

It's the other big category of data extraction algorithms that he calls "unsupervised learning." In these cases, there is no "target attribute"; instead, data extraction algorithms seek to find associations and clusters in the data independent of any previously defined business objective. Oracle Data Mining includes Oracle Data Miner, a graphical user interface which the user uses for data analysis that is used to create, evaluate, and implement data mining models. Oracle Data Miner guides data analysis through the data extraction process with complete flexibility and presents the results in graphical and tabular formats. Oracle Data Miner can generate the PL / SQL code associated with a data extraction activity.

Data Mining tools can analyze very large databases in minutes. The faster Processing speed means that users can automatically experiment with more models to understand more complex data. This high speed makes it practical and profitable for users to analyze huge amounts of data. Large databases, in turn, produce better predictions and results

As we have already seen briefly at the beginning of the Introduction to Machine Learning, supervised learning models are those in which functions are learned, relationships that associate income with outputs, so they conform to a set of examples that we know of. the relationship between the input and the desired output. This fact even gives us an idea of the most common classifications in the type of algorithms that are developed, thus, depending on the type of output, there is usually a subcategory that distinguishes it between classification models, if the output is a categorical value (for example, an enumeration, or a finite set of classes), and regression models, if the output is a continuous space value.

Additionally, unsupervised learning models are those with which we do not want to work on matching pairs (input, output), on the contrary the objective is to increase the structural knowledge of the available data (and possible data that we may find in the future that come from of the same phenomenon), for example, giving a database according to its similarity (clustering), simplifying the structure of the same while maintaining its most important characteristics (as in the processes of dimensionality reduction), or extracting the internal structure with the that the data is distributed in its original space (topological learning).

Typically, most of the most important common definitions, theoretical results, and algorithms are classified as supervised algorithms and, especially in earlier times, many of the unsupervised algorithms were used exclusively for data preprocessing tasks embedded in more advanced methodologies. spacious and complex. This fact is predominantly due to a chain of factors. On the one hand, the objective to which supervised learning is directed is much more broadly defined and oriented, while the unsupervised one is more subtle and diffuse. This also not only affects a more extensive and complex development by having a large number of better defined applications available, It also allows the use of metrics that allow the advantages of the learning performed (the performance of the algorithm) to be evaluated much more clearly. On the other hand, and perhaps as a result of the above, unsupervised algorithms turn out to be of a very high value because they require many more trial-and-error tests, causing them to have a greater use or consumption of a theoretical framework and much physical resources. larger which considerably raises your costs

Recently, new unsupervised algorithms have been appearing related to what is known as Representation Learning, which has proven to be the core of Machine Learning, and where lines of work such as the now famous Deep Learning which is taking over and positioning itself as the the most powerful tool, one of the most interesting advances that are being obtained, is the future of Artificial Intelligence, which is closer to unsupervised than supervised learning.

In this post we will not delve into these last lines of work, for which we will dedicate future posts, and as an example and to fix the main ideas we are going to focus on two specific algorithms, one of each type, which can serve as simple and clear representatives to understand their more general behavior. In addition, along with other paradigmatic algorithms that we will see in other posts, we will use them later to introduce the general concepts of confusion matrix, error, over-learning, performance metrics, etc.

VIII. DATA MINING TECHNIQUES

As we have already commented previously in this article, data mining techniques are derived from Artificial Intelligence and statistics, these techniques, not algorithms, with a certain level of sophistication that are applied to a set of databases to find results. The most representative techniques are:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2021 doi:10.20944/preprints202103.0640.v1

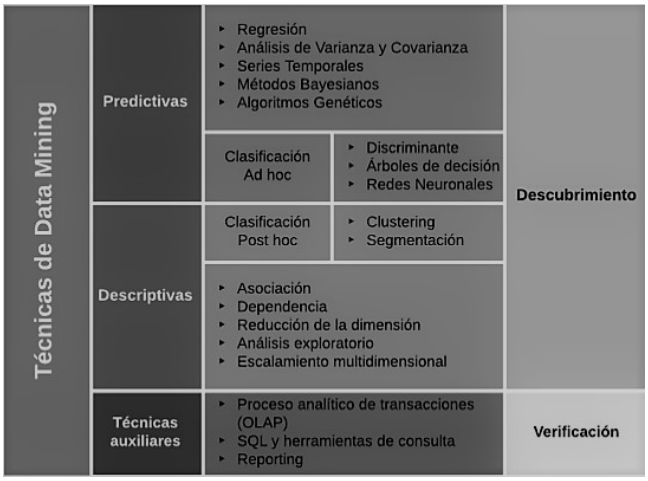


FIG 4 Source: Ing. Cathy Pamela Guevara Vega, University of the Armed Forces, Ecuador

Predictive analysis: It is an area of data mining that is based on the extraction of information already existing in the data and its use to predict trends and behavior patterns, in order to apply them to any case or unknown need, regardless of whether we are talking about the past, present or future. Predictive analysis is based on the analysis and identification of relationships between variables in past events, to then take advantage of and exploit these relationships and thus be able to predict possible results in future situations or needs. then, it must be taken into account that the degree of precision of the collected results varies greatly from how the data was analyzed, as well as the quality and veracity of the assumptions. At first, predictive analysis can be confused with a forecast (that makes predictions at a macro level), but it has nothing to do with each other, it is something completely different. While a forecast can predict how many loaves will be sold next month, predictive analytics can indicate which population is most likely to eat bread. This information, if used in the correct way, represents a radical and substantial change in the game, since it allows us to direct our efforts to be more productive in favor of the objectives. In order to perform predictive analysis, it is necessary to have a considerable amount of data, both current and past, to be able to define behavior patterns with certainty and thus obtain knowledge. For example, in the case used in the previous paragraph, about who is more likely to eat a bread, if data is crossed about the time of year, festivities, religious celebrations or gastronomic trends and if it is a weekend or a holiday, it can be inferred which profile of the person will eat bread. This process is carried out thanks to computational learning. Computers can "learn" autonomously and in this way obtain new knowledge and capacity, for this it is enough to provide them with the most powerful and great natural resource of modern society: called data.

Regression: These regression models are the backbone of predictive analytics. With one approach, it is based on the implementation of a mathematical equation as a model in order to represent the interactions of all the variables that are being taken into account. Depending on the situation and needs, there are a large number of models that can be used during predictive analysis.

Analysis of variance and covariance. The analysis of covariance (ANCOVA) is a technique used in statistics, and specifically it speaks of a parametric test. Parametric tests within statistics allow you to analyze factors within a population. In addition, they allow quantifying the extent to which two variables are independent. The acronym ANCOVA comes from “ANalysis of COVAriance”. Actually, ANCOVA combines two types of strategies: Analysis of Variance (ANOVA) together with Regression Analysis. Here we must bear in mind that ANOVA is another statistical technique that segregates from the total variability of our results, the part due to sources of error; thus, in addition to being an error control technique, it discovers the influence of the treatments. For its part, the analysis of covariance is also a statistical technique, but more complete than ANOVA;

Bayesian methods. thismethod what it does is relate the most influential variables of an event through that same characteristic. This method is based on the Bayes theorem which explains the following: probabilities of a series of events Ai occurring. To this is added an event B whose occurrence provides certain information, because the probabilities of occurrence of B are different depending on the event Ai that has occurred. The naive Bayesian method is based on Bayes' conditional probability rule, which is used for the classification task. The Bayesian classifier assumes that the predictors are statistically independent, which
makes it an effective classification tool that is easy to interpret. It is best used when faced with the problem of the 'curse of the
dimensionality ”, that is, when the number of predictions is very high.

Genetic algorithms. This algorithm extracts prediction rules commonly used to position web pages in the different search engines that exist on the web. It uses a Classic Algorithm with three important and highly relevant changes in order to obtain good results; one of these modifications is to use a uniform population, (instead of random) to overcome problems of the genetic search. From a uniform population, a broader search spectrum will be achieved so that the algorithm expands uniformly in its search through the space of more possible solutions.

Classification and regression trees. This type of Hierarchial Optical Discriminat Analysis (HODA) is a generalization of the Optimal Discriminatory Analysis that is used to identify statistical models that have the highest precision to predict the value of a dependent categorical variable for a set of data that It consists of categorical variables and continuous variables. The output of HODA is a tree that combines categorical variables and cut points for continuous variables that provides maximum predictive precision and an assessment of potential cross-generalizability of the statistical model. Optimal discriminant analysis is an alternative to ANOVA (ANalisis Of VAriance or analysis of variance) and regression analysis, that attempt to explain a dependent variable as a linear combination of other characteristics or measures. However, ANOVA and regression analysis give a dependent variable that is a numerical variable, while hierarchical optimal discriminant analysis gives a dependent variable that is a class variable. Classification And Regression Trees (CART) are a nonparametric decision tree learning technique that produces classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively. Decision trees are made up of a collection of rules based on variables in the modeling data set: ANOVA and regression analysis give a dependent variable that is a numeric variable, while hierarchical optimal discriminant analysis gives a dependent variable that is a class variable.

• Rules based on variable values are selected to obtain the best division to differentiate observations based on the dependent variable.
• Once a rule is selected and divides a node in two, the same process is applied to each "child" node, that is, it is a recursive procedure.
• The split stops when CART detects that no more profit can be made or some preset stop rules are met.

Each branch of the tree ends at a terminal node. Each observation falls on a terminal node, and each terminal node is uniquely defined by a set of rules.

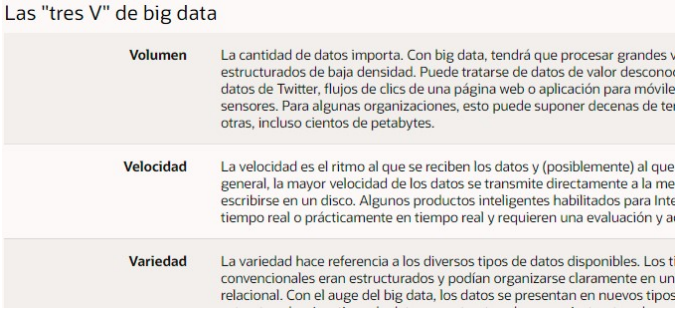
Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 25 March 2021 doi:10.20944/preprints202103.0640.v1

Computational learning it is a fundamental part of a predictive analytics process. Computational learning provides data analysis techniques through which relationships between variables can be discovered that at first may seem like little, but after the implementation of these techniques the importance of these can be discovered Once correlations between variables have been established the work of the human being comes into play, which consists of knowing how to interpret them and make the appropriate assumptions.

Descriptive analysis: Descriptive models quantify the relationships between data in a way that is most commonly used to classify users or contacts into groups. Unlike predictive models that focus on predicting the behavior of a particular user, descriptive models identify different relationships between users and items. Descriptive analytics provides much simpler summaries about the subjects in the sample and about the observations that have been made. Such summaries can form a basis for the initial description of the data as part of a more complex statistical analysis, or they may also be sufficient in themselves for a particular investigation. Descriptive models do not rank or rank users by their probability of performing a particular action in the same way that predictive models do. However, descriptive models can be used for example to assign categories to users according to their taste in products, items or their age range. The uses of descriptive models can be used to develop additional new models that can copy a large volume of individual agents and make predictions. Descriptive models include simulation models, queuing theory, or forecasting techniques. Descriptive analysis calculates descriptive statistics to summarize them into data. The vast majority of social analytics fall into this category of analysis.

BIGDATA

The concept of bigdata is relatively new, however, it is handled since the 60s-70s with early data centers and relational databases- Big data accumulates more information as now not only humans generate data but also the devices used in the IoT; The fact that cloud services are being developed makes the possibilities of big data even wider as the cloud allows for even greater scalability.



[42] Fig. 5 the three v's for Bigdata

DATAWAREHOUSE

[20] Dataware can have a very diverse architecture, it can be based on an ODS or have multiple data smarts and has several layers among which we mention:

1. Data sources layer: as mentioned, they are the sources from which the data comes regardless of the format.
2. Data extraction layer: the data leaves the data source layer and reaches the datawarehouse system, it may have a little debugging
3. Testing area: this layer is where the data is actually purified and transformed into a datasmart and datawarehouse.
4. ETL layer: The logic is applied so that the data acquires its analytical nature, it may be the most delayed stage due to the data transformation process.
5. Data storage layer: this is where the transformed data is already
6. Logical data layer: Business rules that do not have to do with the previous data transformation process are applied
7. Data presentation layer: it is the user interface layer
8. Metadata layer: here is the data of the data, worth the redundancy and it is usually also managed by an application specialized in metadata
9. System operations layer: This layer shows how the datawarehouse system works, user access data, status and performance of the ETL tool.

IX. ORACLE DATAMINING

The existing tools for data analysis have two drawbacks: either they are high-cost, sometimes not accessible to small companies, or when they are integrated into the databases, they become slow and do not process at the required speed, which is also a computer security risk. that data travels on the network; That is why companies such as Microsoft and Oracle devised to integrate data mining techniques into their database management systems, which greatly simplifies data processing, analysis and extraction in terms of both costs and personnel requirements. Another quality is that it can be used with SQL functions.
Oracle Data Mining is a component of Oracle Advanced Analytics that provides powerful algorithms for data mining without having to export the database to other tools. Oracle Data mining processes are integrated with the Oracle Database kernel to maximize its efficiency and scalability; among its advantages we find:
to. No need to export database data to mining tools
b. Security: For the same reason that the data does not need to be exported or transformed, therefore the mining process is less dense or prone to failures.
c. Data preparation: since it allows to automatically manage the data and adapt it to the needs of the client using the administrative tools of Oracle Database
d. Ease of updating data: since it can deliver results for analysis quickly.
and. Analysis: since it integrates for this with Oracle Database and OLAP

X. ORACLE DATA MINER

It is the graphical interface of Oracle Data Mining through which data analysts and scientists can manage through a graphical text editor (GUI) the data within the same database and can follow the processes of the users, discovering common patterns and relationships in data, groups, segments and profiles, allowing changes

 doi:10.20944/preprints202103.0640.v1

as it documents workflows, its interface resembles those of SQL databases

XI. CONCLUSIONS

In conclusion, DBMS have had an interesting evolution since they must adapt to the need to control the processes and the information that is handled in them, in addition to ensuring that it is stable, reliable, multiplatform and flexible for exponential growth. Data mining techniques enable consultation and analysis that are aimed at discovering patterns, trends, profiles or other relationships that are of interest and that, being present in the information, remained hidden.
This treatment has had to adapt to the new needs, being in both cases the basis of business knowledge, since it is now in the full digital context, we have gone from confirmation or verification to finding. That is, to the detection of something hidden through the implementation of predictive models.
Thanks to an adequate process, this objective is achieved with data that is stored in conventional relational systems (Data Warehouse or data warehouse) or, if it is varied or unstructured information that comes from different sources, then it is necessary use Big Data technologies.
The subject of data mining is interesting since we have obtained some knowledge by working with Data Mining tools although we have not delved into it; Nowadays, companies to survive in the market and be competitive, must know how to optimize their resources in the best way and be very agile when making decisions in the face of everything that happens to them day by day and minute by minute. In order to make accurate decisions, it is necessary to have timely data about the business and the ability to analyze this data to extract the most knowledge that is not deductible at first glance. With the large amount of data that businesses generate, it is increasingly difficult to process and analyze this data manually. For this reason, Data Mining techniques have acquired great importance today as a competitive factor for organizations. There are a large number of Data Mining techniques that allow us to improve and even automate some deductive or predictive analyzes that any human being can do. There are many software tools, paid platforms and open source to perform Data Mining, the vast majority of them abstract the mathematical and statistical complexity of the algorithms, thus allowing users to perform advanced analysis on the data without the need for this knowledge. . There are also Data Mining tools that offer graphical data modeling interfaces that allow you to manipulate large amounts of data in a very simple way. For these reasons, Data Mining today has been in great demand in companies. At first, it was difficult to find information about Distributed Algorithm Mining for Data Mining, there are very few programs that can be obtained that provide this functionality. Although there are currently several Data Mining tools that develop the concept, some could no longer be downloaded and others only work in UNIX-like environments and the concept of both generation and execution of the models is still not fully developed. This is one of the greatest evidence that more research is still required on the subject and it would be an interesting area in which to invest resources given the importance of Data Mining today. It was difficult to find information about Distributed Algorithm Mining for Data Mining, there are very few programs that can be obtained that provide this functionality. Although there are currently several Data Mining tools that develop the concept, some could no longer be downloaded and others only work in UNIX-like environments and the concept of both generation and execution of the models is still not fully developed. This is one of the greatest evidence that more research is still required on the subject and it would be an interesting area in which to invest resources given the importance of Data Mining today. It was difficult to find information about Distributed Algorithm Mining for Data Mining, there are very few programs that can be obtained that provide this functionality. Although there are currently several Data Mining tools that develop the concept, some could no longer be downloaded and others only work in UNIX-like environments and the concept of both generation and execution of the models is still not fully developed. This is one of the greatest evidence that more research is still required on the subject and it would be an interesting area in which to invest resources given the importance of Data Mining today. Although there are currently several Data Mining tools that develop the concept, some could no longer be downloaded and others only work in UNIX-like environments and the concept of both generation and execution of the models is still not fully developed. This is one of the greatest evidence that more research is still required on the subject and it would be an interesting area in which to invest resources given the importance of Data Mining today. Although there are currently several Data Mining tools that develop the concept, some could no longer be downloaded and others only work in UNIX-like environments and the concept of both generation and execution of the models is still not fully developed. This is one of the greatest evidence that more research is still required on the subject and it would be an interesting area in which to invest resources given the importance of Data Mining today.
All companies that are able to use data mining in their favor have a powerful tool for predicting market and customer behavior and can make the decision-making process more effective based on the results of the analyzes made in Data mining.

XIII. GRATEFULNESS

In addition to strengthening both individual work and discipline and the need to meet the deadlines established for the delivery of this first ACA, the knowledge acquired during the semesters prior to this period of academic training was also strengthened where perhaps for time we could not know about other types of database managers and their history. We thank Jorge Portella, our Business Intelligence professor who encourages self-learning and research, and the other colleagues in our training environment, who in one way or another have become familiar with these topics either through work or previous studies, collaboration that they gave us and the opportunity to enrich our autonomous learning with their knowledge.

.

XIV. REFERENCIAS

.

[1] https://virtual.cun.edu.co/contenidos/migracion2020/sistemas/s9/business-intelligence/u4/recurso7.pdf
[2] https://platzi.com/tutoriales/1183-bd/1520-breve-resena-de-los-origines-de-las-bases-de-datos/
[3] https://rmorenopovedano.files.wordpress.com/2014/10/evolucion-de-los-sgbd.pdf
[4] https://anmacosa200.webnode.com.co/curso-web-de-bases-de-datos-por-angela-sandoval/concepto-y-origen-de-base-de-datos/
[5] https://docs.google.com/presentation/d/1w52Nw6yBs_rPh3jwfF7QNTsBIP9LmIsL/edit#slide=id.p5
[6] https://www.redalyc.org/pdf/4139/413940772003.pdf
[7] https://histinf.blogs.upv.es/2011/01/04/historia-de-las-bases-de-datos/
[8] https://www.arsys.es/blog/programacion/mariadb/
[9] https://hostingpedia.net/postgresql.html

[10] https://www.ecured.cu/Sybase_ASE
[11] https://www.ibm.com/co-es/products/db2-database
[12] http://gplsi.dlsi.ua.es/bbdd/bd1/lib/exe/fetch.php?media=bd1:0910:trabajos:aimsgbd.pdf
[13] https://es.slideshare.net/AlejandraCaballeroQu/lnea-del-tiempo-sobre-los-sistemas-gestores-de-bases-de-datos-60942174
[14] https://www.ionos.es/digitalguide/hosting/cuestiones-tecnicas/sistema-gestor-de-base-de-datos-sgbd/#:~:text=Un%20sistema%20de%20gesti%C3%B3n%20de,lenguaje%20de%20manipulaci%C3%B3n%20de%20datos.
[15] https://avbravo-2.gitbook.io/jmoordb/capitulo-1/introduccion/orientdb
[16] https://searchdatacenter.techtarget.com/es/definicion/CouchDB
[17] https://conocelahistoria.com/historia-de-la-base-de-datos/
[18] https://www.timetoast.com/timelines/historia-y-evolucion-de-los-sistemas-gestores-de-las-bases-de-datos
[19] https://www.oracle.com/co/big-data/what-is-big-data/
[20] https://blog.powerdata.es/el-valor-de-la-gestion-de-datos/lo-que-necesitas-saber-acerca-de-la-arquitectura-de-un-datawarehouse
[21] https://www.predictiveanalyticstoday.com/oracle-data-mining-odm/
[22] https://docs.google.com/document/d/1Iw3IEqOJwLdzMrmHtRDCbxj9aujBCgEAXYeiaIbK1aw/edit
[23] https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/intro_concepts.htm
[24] https://www.oracle.com/database/technologies/advanced-analytics/odm.html
[25] http://www.oracle.com/technetwork/es/database/enterpri se-edition/documentation/database-11g-warehousing-ybi-426655-esa.pdf
[26] https://www.youtube.com/watch?v=mB2V0BXH608&feature=emb_logo
[27] https://tentulogo.com/oracle-padre-las-empresas-software-mundo/
[28] https://bbvaopen4u.com/es/actualidad/el-ranking-de-lasmejores-soluciones-de-analisis-predictivo-para-empresas
[29] https://www.youtube.com/watch?v=xlqkKEZoO9s
[30] https://www.youtube.com/watch?v=i9-UfF2a38Q
[31] https://www.ceupe.com/blog/tecnicas-y-aplicaciones-de-data-mining.html
[32] https://virtual.cun.edu.co/contenidos/migracion2020/sistemas/s9/business-intelligence/u7/recurso4.pdf
[33] https://docs.oracle.com/cd/E55747_01/doc.41/e58114/introodmr.htm
[34] https://docs.oracle.com/cd/B28359_01/datamine.111/b28129/intro_concepts.htm
[35] https://www.oracle.com/database/technologies/datawarehouse-bigdata/dataminer.html
[36] https://www.ibm.com/support/knowledgecenter/es/SS3RA7_18.1.0/modeler_mainhelp_client_ddita/clementine/oracle_dbmining_examples.html
[37] https://www.youtube.com/watch?v=8Km-UG2LDZw
[38] https://virtual.cun.edu.co/contenidos/migracion2020/sistemas/s9/business-intelligence/u8/recurso3.pdf
[39] https://virtual.cun.edu.co/contenidos/migracion2020/sistemas/s9/business-intelligence/u8/recurso4.pdf
[40] https://virtual.cun.edu.co/contenidos/migracion2020/sistemas/s9/business-intelligence/u8/recurso5.pdf
[41] https://searchdatacenter.techtarget.com/es/definicion/Definicion-de-OLAP-procesamiento-analitico-en-linea
[42] https://www.oracle.com/mx/big-data/what-is-big-data/
[43] https://blog.mdcloud.es/que-es-data-warehouse-ralacion-data-mart/