Case Study (Neptuno)

February 7, 2024

Jonathan Bargiela

Content

Introduction

Neptuno is a company that markets Gourmet-style food and drink products. Despite its success in its domain and its great relevance, the company encounters a series of pain points in various aspects concerning data, among which are the following:

- Distrust of clients and suppliers due to lack of data quality
- Slowness, inefficiencies and delays in decision-making due to lack of clarity regarding the accessibility of information
- Lack of delimited authority in data governance
- Slow systems and processes in data updates, and consequent difficulties in resolving reported incidents
- Expensive and ineffective reports due to lack of clear processes in their preparation

In my capacity as Chief Data Officer—hereinafter CDO—of the company, in this report I will show a set of holistic solutions related to the company's data area to eliminate or reduce the incidence of pain points in its processes, as well as monitor the evolution through metrics functional to the situation of Neptuno in the given context.

In this documentary piece, the sections of the project are indexed, within which are their respective subsections indexed equally.

The sections are divided as follows, and fulfill the following functions:

Data Governance

This section shows in a global way the structural changes to be implemented, from the organizational and operational, through the roles necessary to carry out the work to the key performance indicators—KPIs—for subsequent monitoring.

Data Classification

Given the importance in data governance of safeguarding the integrity of the data of the main actors of the organization, this section shows a distinction between personal and sensitive data, as well as the justification of the practical need to catalog them. So.

Data Quality

The improvement in the quality of the company's data represents in itself the entire final objective in taking the measures set out in this documentary. In this section you can see the quality rules to apply to business data, as well as the reasons for doing so.

Data Modeling

The resulting data modeling is presented in this section, since all the relevant data of the company and its relationships, as well as all the changes made to them, can be visualized here, which will simplify the understanding of all the changes to them. carry out, as well as the motivation for its implementation.

Conclusions

Due to the extensive and intensive nature of the changes to be applied to Neptuno's data governance, it is important to understand the implications of the modifications to be implemented, as well as the added value resulting from employing a sensible data policy in a company of the dimensions that Neptuno possesses.

In this section, for explanatory purposes, the conclusions derived from the significance of applying these measures not only for the cleanup of present data, but also for the refinement of quality standards for future data ingestion and other protocols to be considered, are seen in detail. with all the administrative, operational and economic benefits that this entails.

1. Data governance

1.1 Organizational Model

Regarding the organizational model, given the size of Neptuno's business structure—with 4 operational areas which include Sales, Logistics, Human Resources and Information Systems—it is prudent to adopt a federated organizational model.

The federated model consists primarily of an organizational scheme which allows each business unit to make specific decisions for its area independently and according to its needs, but each area will be uniformly supervised by a unified and centralized data governance committee. .

The benefits that structure provides consist of greater freedom for daily operational actions and high adaptability to contingencies, but without giving up a uniform business guideline in the procedures of each department.

To illustrate, the sales area can decide the date for issuing a routine report, but the format and information required will have to comply with what is stipulated by the committee, as well as what is required by ad hoc norms and standards.

1.2 Operating Model

In reflection of the operating model, a pyramidal structure will be chosen, which is composed of the following items, going from the base to the top:

- Guides: They provide procedural definitions of routine events that occur in the company; can be understood as recommendations
- Procedures: They outline the way in which norms and standards are materialized, as well as tasks to be carried out, and the parties involved and their time frame.
- Standards: Provide requirements for decision making, improving interdepartmental communication by eliminating assumptions and arbitrary interpretation of processes.
- Policies: They explain the reason for leading the proposed policies and demarcate rules to be followed by the members of the established data governance.

1.2.1 Guides

Inspired by the Data Management Body of Knowledge —a book also colloquially known as DAMA-DMBOK—referential bibliography in the data area, each business unit will be able to establish its guides for internal processes.

These guides will be mostly designed by Project Managers and Data Stewards from each area and will be designed to successfully deal with the needs of each Vice Presidency, but they will also be subject to review and subsequent approval by the committee body.

The partial sums of the internal departmental process guides endorsed by a unified supervisory committee must be a first accurate and tangible step towards ratifying Neptuno's commitment to improving the quality of its data.

1.2.2 Procedures

The procedures to be carried out in Neptuno will have the objective of, on the one hand, improving the organizational structures of the information, whether from the ingestion, storage or processing of information, as well as its quality and accessibility facilitated efficiently by the parties of interest in Neptuno's business processes.

In the procedural part, a strong emphasis will be placed on carrying out an organization of the stored information with a unified and federated Data Warehouse, which will have Data Marts divided by the 4 Vice Presidencies in question.

Access to these will be given by user roles and profiles, to ensure that the company's sensitive information, whether personal data of subjects directly or indirectly related to Neptuno or other types of sensitive information such as financial reports, is effectively safeguarded from harmful leaks that could jeopardize the integrity of the data.

Every time a Vice Presidency requires information from another, it must request the relevant permission from the committee, indicating reasons and use of the required information, attaching a person responsible for said request, declared in a Data Log with a precise date. All of this also applies to requests for information from actors external to the company such as suppliers or customers.

The lack of speed in the processes should not be a problem, since there will be a specific section of the committee to guarantee or deny interdepartmental permissions for information, making this improve the problem of slowness.

1.2.2.1 Data Storage

- Data Warehouse with Data Marts by Vice Presidencies (Sales, Logistics, Human Resources, Computer Systems)
- Data updating through incremental daily ingestions in batches with special emphasis on the Sales and Logistics sectors
- Access (by roles and user profiles) managed and guaranteed by departmental Data Stewards for intradepartmental requests, while for interdepartmental requests the Chief Data Steward of the committee will be the one who will facilitate these
- Dictionaries for company data and metadata approved by committee

1.2.2.2 Security Resources

- Identity verifications through authentication for data access
- Centralized DBA for efficient permissions management and troubleshooting

- Logs of actions regarding data, indicating the subject making the request, as well as the reason and use of the information
- Daily backups in the cloud of various processes throughout the entire ETL process to safeguard data, with quarterly backup backups in On Premise systems

1.2.2.3 Progress Measurement

- Application of KPIs to be able to do benchmarking—using as a reference point the moment prior to the application of this data policy—to quantify progress and incidence of data governance in improving business performance
- Vice Presidency reports generated by their Data Users to compile reports based on KPIs
- Quarterly external IT audits (Q1, Q2, Q3, Q4) of all Vice Presidencies for process analysis and subsequent comparison with internal audit records
- Quarterly external accounting audits (Q1, Q2, Q3, Q4) of Sales for analysis of financial statements and subsequent comparison with internal audit records

1.2.3 Standards

Neptuno is a company whose operational base is located in Argentina, but since the company will implement cloud solutions whose servers are hosted in parts of Europe and North America, it must also comply with the regulations in these sectors.

This implies that, when the information is hosted on these servers, the company must carry out Compliance policies consistent with the main places where the servers are located: Europe, California and Virginia, the latter being 2 States belonging to the United States (Europe has a unified continental data protection policy).

In addition to this, since the company carries out its operations in Argentina, it will also need to comply with local data protection laws.

This is why Neptuno will adhere to the following standards in its data governance:

- General Data Protection Regulation (GDPR)
- California Consumer Privacy Act (CCPA)
- Virginia Consumer Data Protection Act (VCDPA)
- Law 25,326
- Resolution 47/2018

1.2.4 Policies

The declaration of data governance policies for Neptuno is based on the following precepts:

- Improve data privacy and security, including the confidentiality and protection of sensitive data, both for clients and Neptuno employees.

- Establish appropriate responsibility for data management as an organizational asset and that members can access information from the other Vice Presidencies with the appropriate permissions.
- Improve ease of access and ensure that once data is found, users have enough information about it to interpret it correctly and consistently
- Improve data quality, which will result in greater precision, timeliness and integrity of information for decision making through the application of data quality rules, decisively seeking customer and supplier satisfaction
- Ensure that all data-related issues are resolved through the data governance structure adopted under this policy

1.3 Structure

Regarding the structure, that of a federated government will be used at the organizational level, where the government office, those responsible, and the Data Users are located on both the business side and the IT side.

This section shows all the members of data governance, divided into the aforementioned subsections by the role they occupy in said sector of the organization, detailing the functions they occupy.

1.3.1 Government Office

The governance office is the one that will be in charge of directing the company on the path of correct data governance practices. They deal with the strategic part of data governance, and they are the ones who ensure that policies are executed and that the results are as expected.

1.3.1.1 Chief Data Officer (CDO)

- Develop a data strategy that aligns with Neptuno's objectives in terms of inventory management, customer preferences and market trends
- Ensure that data-related requirements, such as information about customers, suppliers and employees, are aligned with the information technology resources and business resources available in Neptuno
- Establish standards and policies to ensure the quality of Neptuno-related data
- Provide advice on how to use data to improve initiatives such as business analytics, data-driven inventory management and data-related technologies in the context of Gourmet food and beverage sales
- Highlight the importance of efficient information management to improve decision making and the satisfaction of both customers and suppliers in the Neptuno area
- Oversee the use of data in analytics and business intelligence to optimize operations, marketing strategies, and Gourmet product offerings

1.3.1.2 Chief Information Officer (CIO)

- He will work alongside Neptuno's CDO, but more closely with the technical implementation of the strategies proposed by him.
- Develop and supervise an IT budget that includes systems that facilitate the quick and clear obtaining of key information for decision making. For example, cloud resources for Neptuno databases do not exceed the budget for the IT area.
- Plan, implement and maintain efficient and updated IT systems to improve incident resolution and optimize efficiency at Neptuno, working closely with the DAO in this regard.

1.3.1.3 Data Governance Officer (DGO)

- Develop, implement and review informed and appropriate systems, procedures and controls to ensure that continuous improvement is integrated into the delivery of governance and risk management functions. An example of this is the quality rules for the parameters of the tables of the various Vice Presidencies.
- Improve efficiency and effectiveness, paying special attention to the quality of the reports of the Vice Presidencies so that the reports accurately reflect the health of the sector at the time of being delivered.
- Establish high levels of responsibility, governance and data reliability, seeking to ensure that the appropriate quality rules are used in all business units
- Ensure Compliance with all regulations and standards that must govern this data governance

1.3.1.4 Data Architect Officer (DAO)

- Execution of the organizational data strategy proposed by the CDO to avoid reputational losses with clients and suppliers due to low quality data
- Guarantee the integrity and consistency of data throughout the organization, seeking above all things that the Sales and Logistics areas have fast and pristine processes which facilitate the extraction of information in an agile and secure manner.
- Improve the existing structure with the CIO to align data requirements with resources, using an optimal network of resources, appropriately combining the Cloud and On Premise products available, and thus improve efficiency and speed in obtaining key information for decision-making. decisions
- Provide advice in areas of your specialty such as business analysis, Big Data, data quality, and data technologies

1.3.1.5 Data Quality Officer (DQO)

- Evaluate and guarantee the quality of data related to gourmet products to prevent loss of reputation with clients and suppliers, working with the DGO for this

- Collaborate in the analysis of root causes of data problems at Neptuno, contributing to identifying improvements in business processes and establishing the quality rules necessary to achieve this
- Address the lack of clarity about who to contact and where to look for information related to other Vice Presidencies. This is achieved by constantly monitoring the condition of the data and identifying technical and business process improvements.
- Contribute to the efficient resolution of reported incidents, thus improving the efficiency and quality of work at Neptuno
- Develop, together with the DGO, clear and responsible processes to evaluate and present reports on data quality to the President of Neptuno, reducing losses in the form of time and superfluous costs in the preparation of reports and presentations.

1.3.2 Government Officials

Given the intensive knowledge that those responsible have regarding the business, they are a fundamental link between the implementation of strategy and daily operability of the company's IT area.

1.3.2.1 Data Owners

- 1 for Vice Presidency
- It is your responsibility to have detailed knowledge of the data related to the products that Neptuno sells, as well as all the surrounding information that may be sensitive.
- They will be in charge of verifying that the data life cycles are respected in their respective Vice Presidencies, starting from the ETL processes, working closely with the DGO and the Data Engineers.
- Establishes quality validation criteria and supervises the maintenance of data in the catalog, ensuring that these processes align with Neptuno's business objectives

1.3.2.2 Data Stewards

- 1 single Database Administrator or DBA who acts as Chief Data Steward
- 1 for each vice presidency in the case of other Data Stewards of lower hierarchy
- It will be a highly relevant link between Data Users and access to information, which is why they will work closely with the committee—mostly with the DGO—to generate fluid access to information for all actors, but without sacrificing data security.

1.3.3 Data Users – IT

1.3.3.1 Business Analysts

- They use data to understand the needs and objectives of the business, providing valuable insights to improve efficiency and decision making in the Vice Presidencies

1.3.3.2 Data Analysts

- By collecting, cleaning, designing, exploiting and interpreting data, they ensure that information is relevant, accurate and useful to avoid reputational losses and improve decision making at higher levels within Neptuno.

1.3.3.3 Data Engineers

- They design, build, maintain and optimize data management systems and pipelines within their performance areas, ensuring that information flows are not interrupted by technical failures, working with Data Owners to achieve this successfully.

1.4 Key Performance Indicators (KPIs)

Given that the main pain points in the company are located in the Sales sector, this is where greater efforts will be made to ensure that the quality of the data is not compromised.

This is why some general KPIs will be indicated here that will apply to all Vice Presidencies, since problems such as slow and inefficient processes affect several spheres of influence within it.

Given the urgency of the situation, the committee will temporarily decide the KPIs to implement indefinitely, once all pain points are corrected and the KPIs yield results within desired ranges. Once this happens, each Vice Presidency may choose to use the KPIs whose sector leaders consider appropriate, once previously approved by the committee and that the COO—Chief Operations Officer—does not have specific KPI requirements for a given area.

1.4.1 General KPIs

1.4.1.1 Data Incident Rate

This rate calculates the number of data incidents taking into account hours worked, assuming a given number of employees.

This KPI will be useful to determine the number of incidents per hour worked, helping to accurately determine the frequency with which there are incidents related to data in the 4 Vice Presidencies.

1.4.1.2 Average Response Time

This KPI indicates the variation between the time taken to respond to a data incident and the initial time of the event. The total sum of these times over the number of incidents gives the average response time.

This metric is very useful to be able to evaluate response times to data problems and know how long it takes to respond depending on the severity of the problem.

1.4.1.3 Satisfaction Rate

The satisfaction rate calculates how many users are satisfied with the information acquisition process of a total sample of its users.

“Users” should be understood as the clients and suppliers who request information from Neptuno. They must fill out a brief interactive survey to clarify this.

It should also be understood that this KPI differs from the one set out in point 1.4.2.1, which inquires about the quality of data from the Sales sector. This KPI quantitatively measures the satisfaction of the entire process and does not make a judgment on the quality of the information itself.

The satisfaction rate must provide a more solid foundation to know the true efficiency of the data policy established, in addition to accounting for whether the problems of accessibility to information and lack of clarity in the assignment of roles are alleviated by the procedures. and policies implemented.

1.4.1.4 Reporting Transparency Rate

The reports issued by the Vice Presidencies of Neptuno will be contrasted with the data filtered by the quality rules of point 3.1 by the COD of the committee. Any report whose information is not consistent with the government's Trusted Data will be declared non-transparent.

The transparency rate determines the number of transparent reports over the total population of reports.

This KPI will have two effects: On the one hand, it will show how effective the quality rules implemented are being, and on the other hand, how much the quality of the reports improves over time.

1.4.2 KPIs Exclusive to the Sales Vice President

1.4.2.1 Data Quality Score

As the data most requested by users is that of this Vice Presidency, those users who request this information will see an additional section added to the questionnaire proposed in point 1.4.1.3 which asks users to rate the quality of the data from 1 to 10. the requested data, with 1 being the lowest and 10 being the highest.

The purpose of this KPI is to see if the quality policies implemented are having a positive reception among the most conflicted information users.

1.4.2.2 Return on Investment (ROI)

This KPI does not necessarily calculate in quantitative terms the success of Neptuno's business feat, but rather it calculates the added benefit to this sector of the implementation of data governance.

It will be very useful to scrutinize the future and long-term viability of data governance as well as the possibilities of scalability.

2. Data Classification

The classification of data in a governance context is of utmost vitality. Considering that much of the customer and employee data that is manipulated is intrinsically related to their sphere of privacy, it is prudent and sensible, not only in ethical terms, but in economic matters for Neptuno as well, to catalog them to avoid problems of data filtering.

This is why functions will be created that examine the data present in the databases to determine the presence of personal data and sensitive data in them.

The created functions return 'True' when in the presence of personal or sensitive data, or both. Otherwise, returns 'False'.

To know precisely what specific attribute may have caused this result, use this documentation as a reference.

2.1 Personal Data and Sensitive Data

The most sensitive data is the so-called “personal data” and “sensitive data”. It should be noted that there is data that can be both personal and sensitive. Those attributes absent in the categories do not belong to it.

This is the GDPR definition regarding these:

2.1.1 Personal Data

“Personal data are those that allow the identification of a specific person. They are assets that, isolated or crossed, allow an individual to be found and referred to. In that sense, they are unique references, which distinguish a human being from a mass of people, to allow a concrete action.”

2.1.1.1 List of Personal Data from Neptuno Databases

These are some of the parameters present in the Neptuno databases that are considered “personal”:

2.1.1.1.1 “Customers” Database

- Client ID
- Contact name
- Address
- Phone
- Fax

2.1.1.1.2 “Employees” Database

- Employee ID
- Last name

- Name
- Birthdate
- Address
- Home telephone
- Extension

2.1.2 Sensitive Data

“The following personal data are considered “sensitive” and are subject to specific processing conditions:

- *Personal data that reveals racial or ethnic origin, political opinions, religious or philosophical convictions,*
- *Union membership,*
- *Genetic data, biometric data processed solely to identify a human being,*
- *Data relating to health,*
- *Data relating to the sexual life or sexual orientation of a person.”*

2.1.2.1 List of Sensitive Data in Neptuno Databases

These are some of the parameters present in the Neptuno databases that are considered “sensitive”:

2.1.2.1.1 “Customers” Database

- This database does not contain sensitive data according to the definition provided by the GDPR

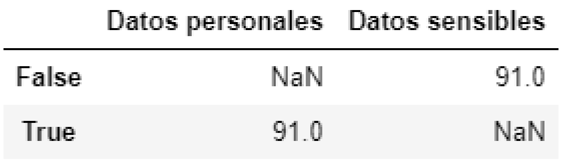
2.1.2.1.2 “Employees” Database

- Photo (biometric data)
- Notes (attributes of the attribute reflect or infer orientations of various kinds)

2.1.3 Results

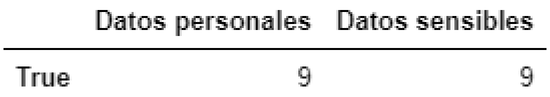
By virtue of the above, it will be shown that the functions effectively return in the case of the 'Employees' database the presence of personal data and sensitive data, while in the case of the 'Clients' database only the presence of personal information.

2.1.3.1 Results from the “Customers” Database



	Datos personales	Datos sensibles
False	NaN	91.0
True	91.0	NaN

2.1.3.2 Results from the “Employees” Database



	Datos personales	Datos sensibles
True	9	9

3. Data Quality

3.1 Quality Rules

The establishment and exercise of quality rules are one of the key points of healthy and effective data governance. The format of the rules and their main components are detailed below.

3.1.1 General Format of Quality Rules

The rules developed in point 3.2 have the following format:

- Criticality
- Dimension
- Column(s) affected
- General description

3.1.1.1 Criticality

Describes how important the rule is for proper functioning of data governance. They have a descending numerical hierarchy, with 3 values assigned:

- 3: Maximum criticality rule
- 2: Average criticality rule
- 1: Minimum criticality rule

3.1.1.2 Dimensions

The dimensions qualitatively indicate the way in which the data must be filtered, in order to increase its quality and be able to classify it as reliable data. Below, the definitions of each dimension will be detailed as established by the DAMA-DMBOK:

3.1.1.2.1 Accuracy

The degree to which the data correctly describes the "real-world" object or event being described.

3.1.1.2.2 Completeness

The proportion of data stored versus 100% potential.

3.1.1.2.3 Consistency

The absence of difference, when comparing two or more representations of a thing with a definition.

3.1.1.2.4 Validity

Data is valid if it conforms to the syntax (format, type, range) of its definition.

3.1.1.2.5 Uniqueness

No entity instance will be registered more than once depending on how the entity is identified.

3.1.1.2.6 Opportunity

The degree to which the data represents reality from the moment required.

3.2 Rules and Justifications Thereof

3.2.1 Rule 1: get_uniqueness_client_id(client_id)

3.2.1.1 Description

- Criticality: 3
- Dimension: Uniqueness
- Affected column/s: 'Id. of customer' from the database 'customers'
- Overview: Checks that there is no other identical customer ID in the 'customers' database. If the value is repeated, the function returns False, since the statement negates the 'keep=False' parameter, which sets all values to the default value False

3.2.1.2 Justification

If the customer ID value is repeated and the uniqueness principle is not respected, there may be problems when establishing relationships with other databases, especially when locating orders.

3.2.2 Rule 2: get_completeness_client_id(client_id)

3.2.2.1 Description

- Criticality: 3
- Dimension: Completeness
- Affected column/s: 'Id. of customer' from the database 'customers'
- Overview: Parses the 'ID' attribute. customer' from the 'customers' database. The function returns True if the value in question is neither null nor an empty string

3.2.2.2 Justification

Likewise, if a value of the attribute 'Id. of customer' of the database 'customers' is empty or has a null value, it will cause problems when establishing important relationships between the other Neptuno databases.

3.2.3 Rule 3: get_customer_id_consistency(entered_customer_id_value, contrasted_df_column)

3.2.3.1 Description

- Criticality: 3
- Dimension: Consistency
- Affected column/s: 'Customer' from the 'orders' database and 'Id. of customer' from the database 'customers'
- Overview: If a value of the 'Customer' attribute in the 'orders' database is not present among the values of the 'Id.' attribute. 'customer' from database 'customers', returns False

3.2.3.2 Justification

If there is no consistency in this data, customer tracking for a given order number becomes unnecessarily difficult and inefficient, since it would have to be done by other means and using filters to arrive at a sample of results that may or may not match. the desired client. This would bring serious difficulties to daily operations at Neptuno.

3.2.4 Rule 4: get_order_id_consistency(entered_order_id_value, contrasted_df_column)

3.2.4.1 Description

- Criticality: 2
- Dimension: Consistency
- Affected column/s: 'Id. order' from the 'orders' database and 'Id. order' from database 'order_details'
- General description: If a value of the attribute 'Id. order' in the database 'order_details' is not present among the values of the attribute 'Id. order' from database 'orders', returns False

3.2.4.2 Justification

Although it is not a rule of equal criticality as the previous ones, order details are an important extension of the original orders, from which information such as the product identifier, purchased product quantities and discounts can be extracted, so it is important that These data are consistent to determine this.

3.2.5 Rule 5: get_precision_customer_id(customer_id)

3.2.5.1 Description

- Criticality: 2
- Dimension: Precision
- Affected column/s: 'Id. of customer' from the database 'customers'
- General description: Using a regular expression, the attribute 'Id. 'customer' from the 'customers' database is an alphabetic data type of about 5 letters, all uppercase. If this does not happen, it returns False

3.2.5.2 Justification

Since customer IDs in the database follow a pattern of 5 uppercase letters as identifiers, it is sensible to maintain this quality rule to avoid future complications.

3.2.6 Rule 6: get_order_quantity_validity(entered_quantity)

3.2.6.1 Description

- Criticality: 2
- Dimension: Validity
- Affected column(s): 'Quantity' from 'order_details' database
- Overview: The values of the 'Quantity' attribute of the 'order_details' database must be non-null and greater than 0. If this is not the case, it returns False

3.2.6.2 Justification

If this data does not follow this quality rule, no aggregate functions or arithmetic operations can be performed on the database. This implies that the daily operations of the company would be affected by this.

3.2.7 Rule 7: get_price_validity_per_order_unit(price_entered)

3.2.7.1 Description

- Criticality: 2
- Dimension: Validity
- Affected column/s: 'Price per unit' from database 'order_details'
- Overview: The values of the 'Unit Price' attribute in the 'order_details' database must be non-null and greater than 0. Otherwise, it returns False

3.2.7.2 Justification

If this data does not follow this quality rule, no aggregate functions or arithmetic operations can be performed on the database. This implies that the daily operations of the company would be affected by this.

3.2.8 Rule 8: get_opportunity_order_date(entered_date)

3.2.8.1 Description

- Criticality: 2
- Dimension: Opportunity
- Affected column(s): 'Order date' from 'orders' database
- Overview: The order date values for the 'Order Date' attribute in the 'orders' database must be greater than or equal to May 25, 1995, the creation date of Neptuno. Returns False if this condition is not met

3.2.8.2 Justification

It would be absurd if the order dates were prior to the year the company was created, which is why this quality rule was created.

3.2.9 Rule 9: get_product_image_validity(input_file)

3.2.9.1 Description

- Criticality: 1
- Dimension: Validity
- Affected column/s: 'Image' from 'categories' database
- General description: Verifies that the values of the files entered as images in the 'Image' attribute of the 'categories' database are not null and that the file format is '.BMP'. Returns False if the file is null or does not have the desired format

3.2.9.2 Justification

Although it is not a vital quality rule for the essential operation of the business, having clear quality rules regarding image format helps improve Neptuno's internal processes.

3.2.10 Rule 10: get_validity_product_category(entered_category)

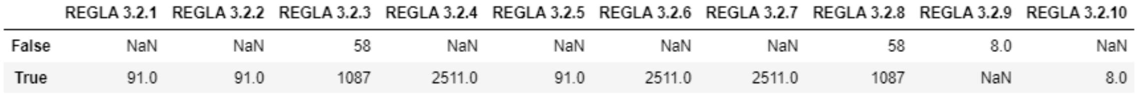
3.2.10.1 Description

- Criticality: 1
- Dimension: Validity
- Affected column/s: 'Category name' from the 'categories' database
- Overview: Checks that the values of the 'Category Names' attribute names are not null. Returns False if they are

3.2.10.2 Justification

Although it is not a vital quality rule for the essential operation of the business, having clear quality rules regarding image format helps improve Neptuno's internal processes.

3.3 Report



	REGLA 3.2.1	REGLA 3.2.2	REGLA 3.2.3	REGLA 3.2.4	REGLA 3.2.5	REGLA 3.2.6	REGLA 3.2.7	REGLA 3.2.8	REGLA 3.2.9	REGLA 3.2.10
False	NaN	NaN	58	NaN	NaN	NaN	NaN	58	8.0	NaN
True	91.0	91.0	1087	2511.0	91.0	2511.0	2511.0	1087	NaN	8.0

4. Data Modeling

4.1 Table Normalization

To normalize the database, it was decided to use the snowflake model, where there is a main database, commonly known as a “fact table,” and annexes known as “dimensions” which can be accessed and extracted. data through the relationships established between tables, shown in figure 4.1.2.

The main benefit of this model is that it takes up less disk space, facilitating a large volume of data storage.

4.1.1 Modifications to the Databases

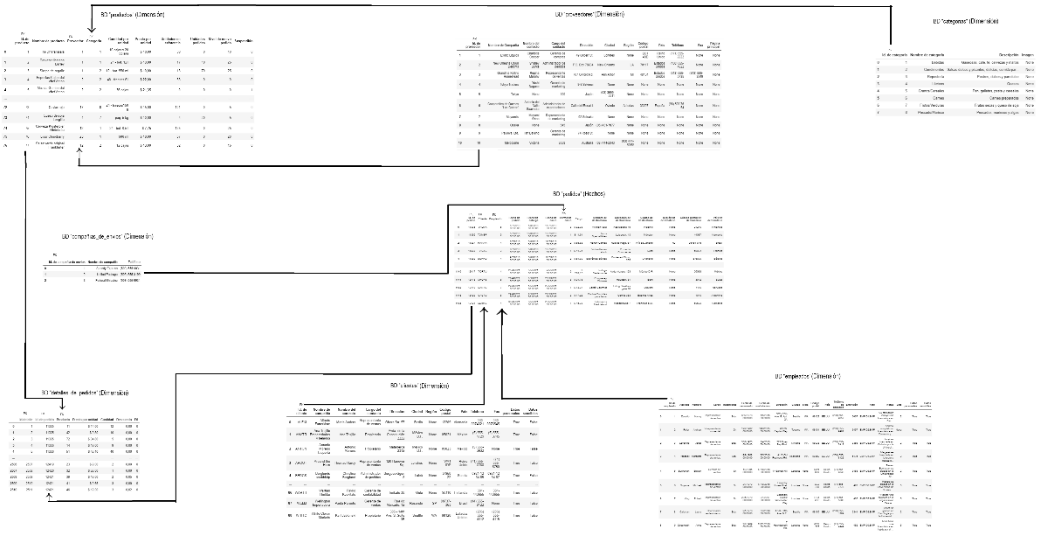
4.1.1.1 Modification of the 'Id' attribute. supplier' from the 'suppliers' database

Given certain inconsistencies in the identifiers, the 'Id' attribute was normalized. 'provider' to avoid identification problems.

4.1.1.2 Creation of the 'Id' attribute. detail' from database 'order_details'

This attribute was created in order to be able to identify with a unique identifier the value assigned to each order identifier in the 'orders' table in the aforementioned table. This will make it easier to identify details in the mentioned database.

4.1.2 Final Result


BD "productos" (Dimensión)
BD "proveedores" (Dimensión)
BD "categorías" (Dimensión)
BD "pedidos" (Hechos)
BD "compañías_de_envíos" (Dimensión)
BD "detalles de pedidos" (Dimensión)
BD "clientes" (Dimensión)
BD "empleados" (Dimensión)

Given possible difficulties in viewing the normalization, a copy of the image with higher resolution has been uploaded to the following link: https://drive.google.com/file/d/1pp2TGTQ7QznSGrnn61c9m9HbMxEG_893/view?usp=sharing

5. Conclusions

5.1 A Structured and Well-Planned Data Governance Is Necessary for Neptuno

In general terms, based on what the results of the proposed quality rules show, Neptuno has an acceptable organization of its information at the time of preparation of this report.

However, if we think in terms of business scalability, it is essential to propose a federated organization with assignment of roles where good practices and cleanliness of information are prioritized in order to grow in sustainable terms, both IT and administratively speaking.

Furthermore, this will also bring positive economic benefits to the organization measurable in the forms of KPIs over a period of time.

5.2 An Adequate Classification of Data Can Avoid Present and Future Economic Losses for Neptuno

In light of the results of the reports of the presence of personal data and sensitive data in Neptuno databases, it is sensible to consider a data governance that regulates the accessibility of these types of data.

Failure to do so can lead the company to face financial losses greater than any implementation costs that data governance may have.

5.3 Data Governance Means Greater Business Efficiency for Neptuno

Actions such as the establishment and exercise of quality standards and standardization of tables involve the elimination or correction of superfluous data with little added value, making Neptuno a more efficient company and acquirer of an asset in the form of information.

This gives it a competitive advantage over its counterparts in the field and gives it the potential to obtain direct economic benefits from the data governance practices suggested in this report.