Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

INTERNSHIP REPORT

OPTIMIZING STATISTICAL PRODUCTION AT A CENTRAL BANK

SARA ALEXANDRA DA CUNHA E SILVA PENA MADEIRA

JANUARY - 2025

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK
INTERNSHIP REPORT

OPTIMIZING STATISTICAL PRODUCTION AT A CENTRAL BANK

SARA ALEXANDRA DA CUNHA E SILVA PENA MADEIRA

SUPERVISION:
PROF. JESUALDO CERQUEIRA FERNANDES
MS. PAULA A. SILVA

JANUARY - 2025

GLOSSARY

AI – Artificial Intelligence

API – Application Programming Interface

BPA – Business Process Automation

CSV – Comma-Separated Values

CTE – Common Table Expression

DW – Data Warehouse

IES – Informação Empresarial Simplificada (Simplified Business Information)

ML – Machine Learning

RPA – Robotic Process Automation

SQL – Structured Query Language

VAT – Value Added Tax

ABSTRACT

The quick evolution of data analysis and manipulation tools over the recent decades has transformed how organizations manage and process information. However, many organizations, particularly those with more years of experience, continue to rely on systems built with the tools and knowledge available at the time. While innovative when introduced, these systems have become inefficient as technology advances, business needs evolve, and data volume grows. As a result, updating and improving these processes is now essential to meet current demands.

At Banco de Portugal, particularly in the Central Balance-Sheet Division of the Statistics Department, these challenges are evident. Although still operational, the annual statistical production process requires reformulation, as it was developed incrementally over time to address specific needs, with now outdated methodologies and tools. These processes currently present inefficiencies such as excessive manual tasks and redundant code. Therefore, to address these issues and align with the department's goal of integrating the systems within the Data Warehouse, the Central Balance-Sheet Division initiated several optimization efforts, including this project.

This project involved five optimization actions to address inefficiencies and modernize the statistical production process. Initially, efforts were directed towards developing an automated drill-down tool, and analyzing the current production workflow while documenting programs developed using SAS software. These insights led to a proposed redesign of the production process, which addressed inefficiencies and facilitated integration with the Data Warehouse. Building on this new process design, subsequent tasks included creating an alert analysis program and migrating the programs used to define the universe of active companies to SQL.

Although the full optimization is projected to take two years, this project represents a vital first step, delivering tangible improvements and establishing the foundation for a more efficient and integrated statistical production system at Banco de Portugal.

KEYWORDS: Data Analysis and Manipulation Tools; Data Warehouse; Process Automation; Process Optimization.

JEL CODES: C61; C80; D22; E58; Y10.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

First and most importantly, I would like to express my deepest gratitude to my family. To my parents, for your unconditional support and for giving me every opportunity I could have ever hoped for. To my brother, for always bringing light and laughter when I needed it most. To my grandmother, for your constant love and for always being my safe place. Thank you for always being there.

To my friends, especially Matilde, Renata, and Faria, your friendship and support over the years have meant so much to me. No matter the distance or circumstances, you have always been there and, for that, I am deeply thankful.

I am also very grateful to Professor Jesualdo for his guidance and support during this project. Your advice and expertise have been invaluable in completing this work.

My appreciation also goes to the Statistics Department of Banco de Portugal for giving me the opportunity to expand and apply my knowledge in such an exceptional environment. To the Central Balance-Sheet Division team, especially Carolina, Paula, and Ana Bárbara, thank you for your guidance and support during both my internship and this project. Your contributions were crucial in making this experience so rewarding.

Lastly, to my boyfriend, thank you for your endless patience, for always listening, and for supporting me through the highs and lows of this journey. I couldn't have done it without you.

1. INTRODUCTION

The Central Balance-Sheet Division of the Statistics Department at Banco de Portugal is responsible for producing a wide range of statistics, including annual data regarding non-financial companies. These statistics are essential for economic analysis and informed decision-making. However, the current system used to produce them relies on multiple SAS programs that, while functional, suffer from significant inefficiencies due to manual tasks and lengthy processing times. Alongside these challenges, the department has adopted a new vision focused on centralizing and integrating information within its Data Warehouse (DW), making it evident that a transformation of the annual production system was needed.

Therefore, this project aimed to improve the efficiency of the existing annual production system while supporting the department's goal of integration. The approach involved structuring and implementing improvement actions to initiate the optimization process, which was particularly relevant considering the project's focus on preparing for long-term improvements while delivering tangible short-term results.

The report is structured into three main sections. The first section presents a literature review, offering valuable context and insights into relevant topics such as process automation, the role of DWs in process optimization, and data analysis and manipulation tools. The second section outlines the methodology, detailing the framework guiding the project's execution. The final section focuses on the implementation phase, providing the necessary contextualization and describing the optimization actions taken. These actions include the development of a drill-down tool, analyzing and documenting existing SAS processes, designing a new production workflow, creating an alert analysis program, and reformulating a process for migration to SQL.

During the development of this report, AI tools including Grammarly and ChatGPT were used to improve the quality of writing. These tools supported the enhancement of sentence construction, ensuring clarity, coherence, and precision throughout the work documentation.

2. LITERATURE REVIEW

This literature review explores a crucial method for collecting financial information from companies, the IES, while highlighting the importance of guaranteeing data quality in this type of data. Additionally, it explores some relevant key topics for the project, including process automation, data warehousing, and three essential data analysis and manipulation tools: SAS, SQL, and Python.

2.1. IES (Simplified Business Information)

In the late 20th century, the demand for comprehensive statistical data on Portuguese enterprises increased significantly (Pereira, 2011). This highlighted the need for a solution enabling the collection of information from companies through a unified and efficient process while minimizing data collection costs and reducing the burden on companies, that were previously required to submit the same information to the Ministry of Justice, Ministry of Finance, Banco de Portugal and Statistics Portugal separately (David et al., 2014).

To address this challenge, the Decree-Law No. 8/2007 was introduced and implemented in 2007, establishing the IES, an acronym for the Portuguese term "Informação Empresarial Simplificada", which translates to "Simplified Business Information" in English (Decree Law No.8/2007, 2007). Developed collaboratively by the four entities mentioned above, IES enables companies to attend to their annual reporting obligations through a single electronic submission, consolidating accounting, fiscal, and statistical information (Pereira, 2011).

IES comprises 13 annexes including the official information defined in the Portuguese Official Accounting Plan, such as the Balance Sheet and Income Statement, along with some additional information necessary for statistical and fiscal purposes (Banco de Portugal, 2008; Comissão de Normalização Contabilística, 2005).

After submission, the information provided by the company goes through an initial consistency and quality check. Subsequently, each public entity may conduct further tests considered necessary to guarantee data quality. For example, Banco de Portugal compares the information submitted through IES with other sources and databases available to

guarantee consistency between the submitted data and these other sources (Banco de Portugal, 2008).

This highlights the importance of high-quality data in financial reporting, since the success of IES depends not only on effectively collecting data and reducing administrative burden, but also on the accuracy and reliability of the information submitted (Rathnayake et al., 2021). Consequently, maintaining high data quality standards is essential for the effective use of the IES.

2.2. Data Quality in Financial Reporting

The quality of financial reporting refers to how accurately the financial information reported reflects an entity's true financial situation (Bai et al., 2023), and, according to the International Accounting Standards Board (2010), it is defined by two key characteristics: faithfulness and relevance.

Ensuring high-quality financial data is crucial for a wide range of stakeholders, including investors, analysts, and researchers, as they rely on that information for different purposes, whether it is evaluating a company's financial health to make informed decisions, producing statistics, or conducting academic research (E'leimat et al., 2023; Redman & Hoerl, 2022).

However, poor data quality is a prevalent issue that can lead to incorrect conclusions and flawed decision-making, particularly in fields like statistics, where there is a high reliance on accurate information (Redman & Hoerl, 2022). Therefore, it is essential to have robust methods for assessing data quality and try to mitigate this risk (Herath & Albarqi, 2017).

Data validation is one of the most effective processes for ensuring data quality, involving a systematic verification of data against predefined standards and rules. This process enhances the detection and correction of errors, as well as the identification of anomalies, being particularly important when data is sourced from multiple entities, as the complexity and potential for inconsistencies increase (Charles, 2024). According to the same author, there are several types of data validation techniques: semantic validation, which guarantees that the data is meaningful within its context; range validation, that analyzes if numerical values fall within an acceptable range; null/not null validation,

which ensures that all required fields contain data; cross-validation, that compares the data from different sources to guarantee consistency and accuracy; among others.

While data validation techniques can help mitigate inaccuracies, the effective management and utilization of this data is equally crucial, particularly in today's world of increasing data volume, where the likelihood of human errors increases (Breton & Bossé, 2003). To mitigate this risk, automation has emerged as a key solution, reducing manual intervention and enhancing operational reliability.

2.3. Process Automation

According to IBM (2024), automation involves leveraging technological tools, systems, or methods to achieve results with minimal human intervention. Automation is becoming increasingly relevant due to the competitive business environment, that forces organizations to continuously seek new strategies for enhancing their efficiency and productivity (Adesina et al., 2024).

Process automation offers a wide range of benefits, such as reducing employee workload, enabling them to focus on more complex tasks, increasing the stability of the task execution, and reducing the occurrence of human errors (Breton & Bossé, 2003).

Automation can be categorized into several types, including task automation, workflow automation, and business process automation (BPA).

Task automation involves automating single and repetitive tasks, such as clicking on a button to send an email (Bataev & Davydov, 2020). This type of automation is generally done using Robotic Process Automation (RPA), which is a software solution designed to execute repetitive tasks usually performed by humans based on simple rules (Hartley & Sawaya, 2019). It often represents the initial step of a company's digitalization transformation, not only enabling employees to focus on more complex tasks but also serving as a cost-effective investment that generates quick returns (Hartley & Sawaya, 2019).

Workflow automation refers to the process of automating a specific sequence of tasks within a larger process, usually to improve efficiency within a department or a specific process (Stohr et al., 2001). Workflow automation often employs several tools, including artificial intelligence (AI) and rule-based logic programs (IBM, 2021).

Lastly, BPA is a broader concept, that involves the automation of complex and multi-step business processes, often by integrating and optimizing multiple workflows within an entire process to increase its efficiency, reduce errors, and enhance organizational outcomes (Mohapatra, 2009). Given its broad application, BPA often leverages a wide range of technologies such as RPA, AI, and cloud platforms (IBM, 2024b).

However, while automation tools can function across multiple data sources, their efficiency is enhanced when those sources are centralized, as centralized systems are easier to maintain and ensure consistency, making data warehousing crucial for seamless integration (Samad et al., 2007).

2.4. The Role of Data Warehousing in Process Optimization

According to Lechtenbörger et al. (2000, p.1) a DW is "*an integrated and time-varying collection of data* (…) *that stores integrated, often historical, and aggregated information extracted from multiple, heterogeneous, autonomous, and distributed information sources*". In other words, a DW serves as a centralized repository that consolidates data from different sources (Gardner, 1998).

The use of a DW offers several advantages to an organization. First, it is time-saving once it centralizes data, allowing users to complete tasks faster than they would with different data sources (Watson et al., 2002). Additionally, it aggregates high-quality and consistent data from across the organization, ensuring that decision-makers are working with accurate and reliable information (Watson et al., 2002).

List et al. (2002) identified several key reasons for the effectiveness of DWs in optimizing business processes. One big advantage is that DWs allow access to extensive historical data, unlike process logs, which typically capture data only for shorter periods. Additionally, while process logs are often stored in normalized databases – an efficient format for storing data – this structure slows down analysis due to the need to join multiple tables. Data warehouses, by contrast, are designed to store data in an optimized format for faster analysis, eliminating this type of constraint. Furthermore, DWs integrate data from several source systems regardless of format or structure, a task difficult to achieve in process logs due to design constraints. Moreover, DWs store information at multiple levels of granularity, allowing them to meet the informational needs of different

types of decision-makers across the organization. Finally, Gardner (1998) also stated that DWs enhance cross-functional data analysis by consolidating data from various departments into a standardized format, enabling seamless access and use of this information across the organization.

In addition to that, and according to Watson et al. (2002), a DW is particularly valuable for organizations undergoing significant strategic shifts, as it facilitates the redesign and coordination of processes to align with new objectives, ensuring cohesive support for the updated strategy.

2.5. Data Analysis and Manipulation Tools

Having explored methods to ensure data quality, implement automation, and leverage data warehousing for process optimization, it is now essential to examine some tools that enable data analysis and manipulation: SAS, SQL and Python.

2.5.1. SAS for Data Processing

In 1976, SAS, originally representative of Statistical Analysis System, was developed as an integrated software system that brought together several standalone programs under a unified framework (Herzberg, 1990). It is a powerful software that uses a fourth-generation programming language, enabling users to perform a wide range of tasks including data management, report generation or statistical analysis (Li, 2013).

Peng (2019) outlines several advantages of using SAS for data analysis. First, SAS can import data from other statistical software packages, such as SPSS and Stata, facilitating data integration across platforms. Additionally, SAS supports various data formats, such as Oracle, enabling users familiar with these formats to continue using them. Moreover, SAS is versatile, powerful, and flexible, allowing users to work with diverse input and output formats while offering extensive options for a wide range of statistical analyses. Finally, SAS is not only user-friendly for data entry and management, but also benefits from a robust community of users who provide support and resources.

However, there has been a noticeable trend of migration from closed-source systems, such as SAS, to open-source options (Batarseh et al., 2020). According to Dr. Ralf Wintergerst, president of Bitkom, open-sources systems offer numerous advantages

including increased flexibility, cost savings, enhanced security, and the promotion of technological advancement (Bitkom, 2023). Consequently, an increasing number of companies are transitioning from SAS to alternative software solutions such as Python, R, and SQL.

2.5.2. SQL for Data Processing

Structured Query Language (SQL) is a standardized programming language initially developed to manage and manipulate relational databases (Silva et al., 2016). Originally created by IBM in the 1970s, SQL became the first commercially available relational database management system, reaching the public market in 1981 (Hoffer et al., 2016).

Today, SQL is one of the most widely used programming languages globally, being essential in data-driven environments due to its capacity to efficiently handle large volumes of structured data (DeBarros, 2018; Silva et al., 2016). To do so, SQL supports a wide range of operations, with some of the most used being the SELECT queries to retrieve columns, JOIN operations to merge data from several tables, and the GROUP BY clause to group results by specific columns (DeBarros, 2018).

These functions are extremely important to data management and can be optimized for enhanced performance, particularly in large datasets, through options such as indexing – which allows faster access and selection of rows within tables – or partitioning – which divides a dataset into multiple tables, improving data organization and retrieval (Hoffer et al., 2016).

SQL is often used along with computationally more complete languages, such as Java, C, and Python to develop applications, functions or procedures (Hoffer et al., 2016; Silva et al., 2016). Additionally, SQL's compatibility with data warehousing solutions enables centralized storage and processing of data, which is essential for efficient data workflows within organizations (Hoffer et al., 2016; Watson et al., 2002).

2.5.3. Python for Data Analysis and Automation

Introduced in 1991, Python is a versatile programming language known for its readability and simplicity, qualities that have contributed to its widespread adoption in the programming community (Mckinney, 2017). Often referred to as a scripting language,

Python is particularly useful to quickly create small programs or scripts to automate repetitive tasks, enabling significant productivity and workflow improvements (Mckinney, 2017; Tupsakhare, 2019).

Python's impact on the technology industry is extensive due to its versatility across fields such as web development, data analysis, AI, and machine learning (ML) (Mckinney, 2017). This versatility comes from Python's open-source ecosystem, which includes a vast collection of packages that distinguish it from a standard programming language. In what concerns data analysis, key libraries such as NumPy, Pandas, Matplotlib, Seaborn, SciPy, and Statsmodel empower users to manipulate data, perform statistical analysis, and visualize results (Haslwanter, 2016).

Python's capabilities in automation further enhance its appeal, making it an ideal tool for both simple and complex scripting tasks. According to Tupsakhare (2019), this happens not only because Python is accessible and easy to use, but mainly because it has several other characteristics that make it particularly suitable for automation. First, it is cost-effective once it is an open-source platform. Additionally, its compatibility with multiple operating systems and its scalability ensure it is a reliable choice for automation tasks of varying complexity (Tupsakhare, 2019). Python's ability to interact directly with databases through libraries such as SQLAlchemy, as well as its effectiveness in managing Application Programming Interface (API) interactions, is essential for automating data exchanges across platforms, further supporting its role in integrated data workflows (Tupsakhare, 2019).

3. METHODOLOGICAL APPROACH

The methodology adopted in this project follows a unique framework that does not fully align with any established methodologies. However, it shares several characteristics with Scrum, despite diverging in some other key aspects.

Similarly to Scrum, this projects' framework has an iterative and incremental approach organized into sprints that follow the same structure as Scrum's – planning, building, testing, and reviewing (Srivastava et al., 2017). Nevertheless, unlike Scrum's predefined fixed-length sprints (Sachdeva, 2016), the sprints in this project did not have specific deadlines. Instead, the progress was monitored through periodic meetings where the status of the tasks was assessed. Regarding retrospection, this project followed Scrums' principle of reflecting at the end of each sprint to evaluate progress and plan for the next iteration (Srivastava et al., 2017).

However, a key difference lies in the roles and structure within the framework. While Scrum emphasizes collaboration among defined roles – Scrum Master, Product Owner, and Scrum Team (Morandini et al., 2021) – this project did not incorporate such role distinctions. Additionally, Scrum relies on structured artifacts, such as the Product Backlog, to manage and prioritize tasks from the beginning of the project (Morandini et al., 2021), whereas in this project, tasks and requirements were identified and adjusted dynamically without a formal backlog.

In conclusion, while the methodology used in this project does not align entirely with any established framework, it resembles Scrum in its focus on flexibility and adaptability to change. Its iterative nature, emphasis on feedback, and ability to adjust plans align with Scrum's principles, even as certain structural elements and formal roles are absent.

4. PROJECT IMPLEMENTATION

4.1. Project Phases

As mentioned earlier, the annual production system of the Central Balance-Sheet Division of the Statistics Department at Banco de Portugal, where this project was conducted, faces some efficiency challenges. These challenges, combined with the department's vision to integrate processes within the DW, highlighted the necessity to optimize the process and motivated the initiation of this project. Recognizing these priorities, the project was designed with a clear structure, divided into three key phases:

Phase 1: Analysis of the Current Production Process

The first phase of the project focused on gaining a comprehensive understanding of the current production process. This involved reviewing the annual production manual and participating in several training sessions to become familiar with key statistical concepts, the various divisions within the department, and the different information produced and analyzed by the department. These sessions also included an overview of the department's DW, which plays a critical role in the optimization process.

As an initial step towards improvement, a manual process was automated to enable the disaggregation of economic and financial indicators. This task also provided an opportunity to become more familiar with the data and information. Additionally, some existing SAS codes used in the production process were analyzed and documented to better understand the system's structure and workflow.

Phase 2: Identifying Improvement Areas and Proposing a New Production Process

After gaining a solid understanding of the existing process, the second phase involved identifying inefficiencies and designing a new production process to address these issues while aligning with the department's vision. This was done by applying the knowledge gained in the previous phase and integrating insights from colleagues.

The redesigned process was then presented to the team, providing an opportunity for discussion, collecting feedback, and defining the next steps for implementation.

Phase 3: Improvements Implementation

After completing the second phase, it was determined that two key tasks would be carried out to begin reformulating the annual production system: developing a program to facilitate the alert analysis, and reformulating the processes used to generate the universe of active companies.

4.2. Contextualization

4.2.1. Current Production Process

The existing annual production process of the Central Balance-Sheet Division follows a detailed and multi-step workflow, summarized in a simplified and condensed format illustrated in Figure 1.

The production process begins with the annual receipt of the IES, typically submitted by companies until July 15th, which serves as the primary data source. The first step of the process involves an initial quality control phase to ensure the consistency and accuracy of the microdata, which consists of granular, company-level records reported in the IES. This involves, for example, cross-validating the IES information with other departmental sources, like the Central Credit Register, to ensure data coherence. This initial verification is performed using SQL. However, the data is subsequently copied to SAS, the main platform used for the remainder of the processing workflow.

The next step involves defining the universe of active companies. Companies that have submitted the IES are automatically included in this universe. However, it is also necessary to investigate the ones who have not submitted the IES to determine if they are truly inactive or have simply delayed their submission. This is done by consulting other data sources, such as the Balance of Payments, or the Monetary and Financial Information, to assess if the company has shown any signs of activity. If a company is confirmed to be active despite not submitting the IES, it is included in the universe, and a non-response treatment is applied to handle the missing data.

After defining the universe of companies, the data is aggregated to generate the desired statistical outputs and then undergoes an additional quality control phase to ensure its accuracy and coherence. During this step, a list with priority issues is produced, identifying anomalies or inconsistencies that require individual analysis and correction.

These issues are addressed and, once all corrections are made, the aggregated series are concluded.

The final statistical data is published on BPstat, Banco de Portugal's official statistics website, and stored in the department's DW. After publication, the data goes through a last quality control check using a quality control platform. This platform ensures the accuracy of the data published by generating alerts for unusual patterns or anomalies, such as unexpected negative values or deviations from accounting standards. These alerts are then manually reviewed by technicians, who investigate and address any identified issues to guarantee data quality.

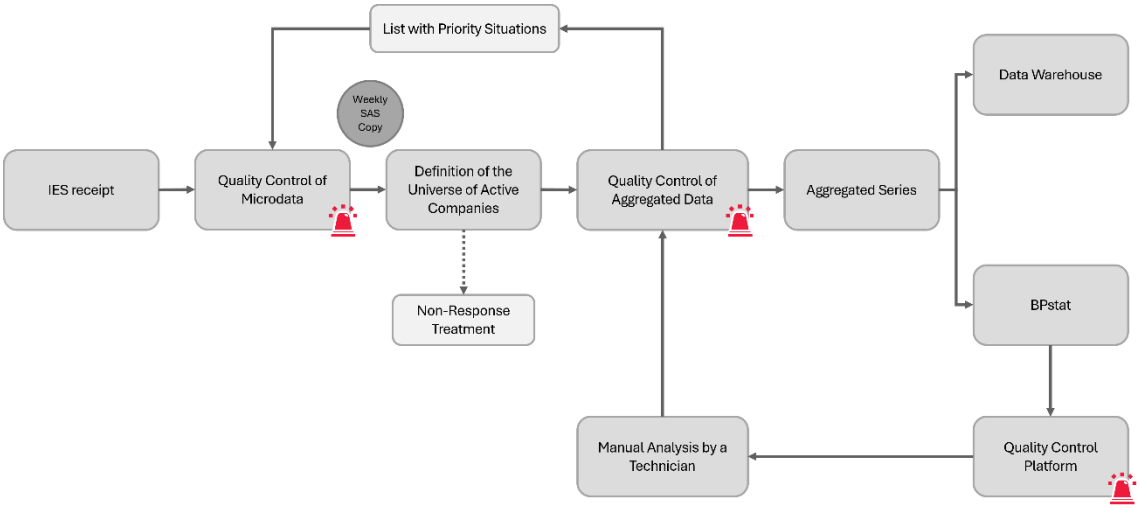

FIGURE 1 - Current Production Process

4.2.2. Future Vision for the Department's Systems

In 2021, the department introduced a new vision for its systems, emphasizing the centralization and harmonization of data. This vision aims to ensure that all tables and databases within the department follow a unified format, improving efficiency, reducing inconsistencies, and establishing best practices for managing production systems. By harmonizing data structures, the department seeks to create a seamless and consistent framework for data management.

As part of this initiative, two centralized data repositories were created to provide a unified structure for storing and managing data. These repositories simplify data sharing

across systems without requiring direct connections, improving collaboration and access to information.

The vision also includes department-wide tools, such as the quality control platform, which performs quality tests on data, and the MAP tool, which facilitates the mapping of information to BPstat.

Therefore, it is essential to align with this vision to benefit from its resources and achieve the department's new goals.

4.3. Optimization Actions

The reformulation process involves multiple steps, including automating several tasks, re-evaluating certain programs, transitioning to SQL efficiently, and documenting processes. As a result, it is estimated that the whole reformulation will take approximately two years to be completed. Consequently, the objective of this internship project was not to conclude the reformulation but to initiate the optimization process.

Therefore, to optimize the annual production system, five tasks were completed within the internship timeframe. These tasks were organized into the three phases described in section 4.1. and are summarized in Figure 2.

These tasks were carried out independently, with direct support from a colleague and the supervisor at Banco de Portugal. Other team members were also consulted on specific occasions to provide clarification or additional input when necessary. This collaborative approach facilitated the effective identification of solutions and the efficient completion of tasks.

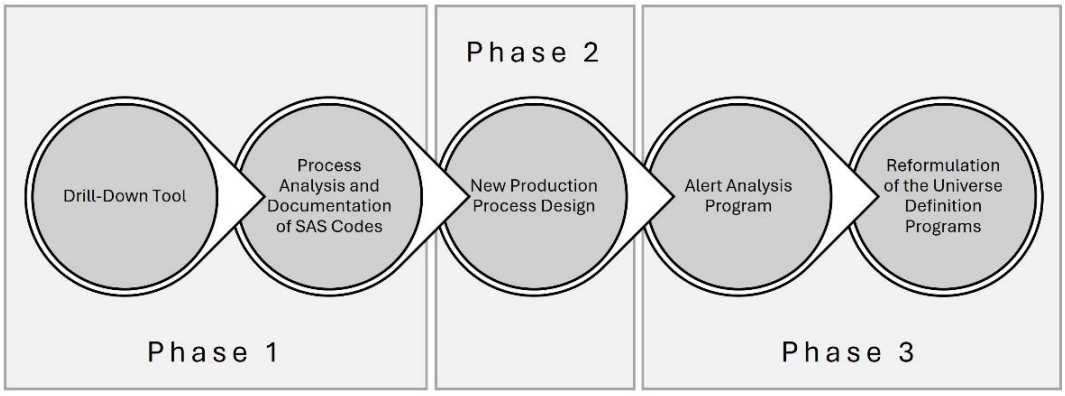

FIGURE 2 - Tasks performed in each project's phase

4.3.1. Drill-Down Tool

To optimize a process, it is essential to first understand it, a task that requires time and interaction with the data. Consequently, the initial step in this project was the development of a drill-down tool. While this tool was not directly integrated into the annual production process, it was based on the same data and needed to be developed.

The primary objective of this new tool was to automate the drill-down process from codes associated with economic and financial indicators (subsequently referred to as codes) into the IES items used for their calculation. This tool was developed to address a challenge since technicians were previously required to manually disaggregate codes to identify the corresponding IES items. This task could be particularly demanding due to the complexity of the indicators, that can have up to nine levels of aggregation and decompose into multiple components.

The manual disaggregation process relied on an Excel spreadsheet containing all economic and financial indicators calculated by the Central Balance-Sheet Division. Each indicator was represented by a code of one letter followed by three digits, stored in a column labeled “Code”. Additional key columns in the spreadsheet included:

- Designation: The name of the indicator corresponding to the code;
- Level: The aggregation level of the indicator, ranging from 0 (indicative of an IES item) to 9 (the highest level of aggregation);
- Formula: The formula used to calculate each indicator:
    - For indicators with an aggregation level above 0, it contains operations with the indicator codes influencing its calculation (e.g. B001 + B002);
    - For level 0 items, the formula specifies the operations between IES items based on their locations within the IES document. This includes detailed references to the sheet, table, field, and column in an abbreviated format (e.g. Q04A-A5101-1 + Q04A-A5102-1);
- Components: Provides the breakdown of the components of the formula;
- Coefficient: Coefficient of the component (if positive or negative).

For understanding purposes, a representative and simplified example of the drill-down process performed using the Excel spreadsheet is illustrated in Figure 3. For instance, consider a scenario where a technician needed to identify the IES items used to

calculate the indicator represented by the code B001. The technician would first examine the "Level" column for B001. If the level was 0, this would indicate that B001 directly corresponds to an IES item. However, in this example, the level is 2, meaning further disaggregation is necessary. The technician would then consult the "Components" column to identify the codes into which B001 is disaggregated and then locate those same codes in the "Code" column. This process would be repeated until the IES items were identified. Once located, the "Components" column would provide the paths to the precise location of each of these items within the IES document.

Code	Designation	Level	Formula	Components	Coefficient
B001	DsgB001	2	[B002]+[B003]	B002	1
B001	DsgB001	2	[B002]+[B003]	B003	1

↓

Code	Designation	Level	Formula	Components	Coefficient
B002	DsgB002	1	[B005]+[B006]+[B007]	B005	1
B002	DsgB002	1	[B005]+[B006]+[B007]	B006	1
B002	DsgB002	1	[B005]+[B006]+[B007]	B007	1
B003	DsgB003	1	[B010]-[B011]	B010	1
B003	DsgB003	1	[B010]-[B011]	B011	-1

↓

Code	Designation	Level	Formula	Components	Coefficient
B005	DsgB005	0	[Q04A-A5101-1]+[Q04A-A5102-1]	[Q04A-A5101-1]	1
B005	DsgB005	0	[Q04A-A5101-1]+[Q04A-A5102-1]	[Q04A-A5102-1]	1
B006	DsgB006	0	[Q04A-A5104-1]-[Q04A-A5105-2]	[Q04A-A5104-1]	1
B006	DsgB006	0	[Q04A-A5104-1]-[Q04A-A5105-2]	[Q04A-A5105-2]	-1
B007	DsgB007	0	[Q03A-A5002-1]	[Q03A-A5002-1]	1
B010	DsgB010	0	[Q05A-05071A-A5566-5]	[Q05A-05071A-A5566-5]	1
B011	DsgB011	0	[Q03A-A5006-1]	[Q03A-A5006-1]	1

FIGURE 3 - Drill-Down Process

This manual process was both time-consuming and prone to errors, highlighting the need for automation. To address this, the drill-down tool was developed using Python and three data sources: the same Excel spreadsheet already described, and two SQL tables. These SQL tables were joined to obtain detailed path names for accessing the IES items, converting the abbreviated codes into more descriptive labels. For example, while an abbreviated code might appear as "Q04A-A5125-1", the SQL tables provide a clearer description, such as "Balance Sheet – Current Assets – Cash and Bank Deposits".

However, before proceeding, some data preparation was required to address inconsistencies found in the Excel spreadsheet. Missing values were identified in both the "Level" and "Components" columns. To address these issues, missing aggregation levels were assigned a default value of 9, as this would not affect the path to reach the IES item,

it would simply indicate that the indicator was not an IES item. Missing components were handled using a Python script detecting patterns of one capital letter followed by three digits in the "Formula" column and added them to the "Components" column. Figure 4 illustrates examples of these issues and their corresponding corrections. After this step, the dataset was free of null values and ready for use.

Code	Designation	Level	Formula	Component	Coefficient
R001	DsgR001		[D002]/[D003]	D002	1
R001	DsgR001		[D002]/[D003]	D003	1
B003	DsgB003	6	[B020]+[B030]		
D001	DsgD001		[D010]-[D030]		

→

Code	Designation	Level	Formula	Component	Coefficient
R001	DsgR001	9	[D002]/[D003]	D002	1
R001	DsgR001	9	[D002]/[D003]	D003	1
B003	DsgB003	6	[B020]+[B030]	B020	1
B003	DsgB003	6	[B020]+[B030]	B030	1
D001	DsgD001	9	[D010]-[D030]	D010	1
D001	DsgD001	9	[D010]-[D030]	D030	-1

FIGURE 4 - Missing Values and Corrected Output

The drill-down tool was then implemented as a Python function to automate the previously manual process. The function developed takes a single input – the indicator code the user wants to analyze – and then identifies the corresponding rows in the Excel dataset. It automatically performs the steps that were once done manually, starting by determining the level of aggregation. If the level is 0, indicating an IES item, the function returns it as result, otherwise, the function identifies the components and recursively performs the drill-down process until all level 0 items are found.

Additionally, the script establishes a connection with the two already mentioned SQL databases that are joined and help identifying the precise location of each IES item, facilitating its interpretation and use.

An example of the output of this function is illustrated in Figure 5.

```
IES('B001')
```

Python

	Code	CodeDesignation	Component	ComponentDesignation	Coefficient
395	B056	Activo corrente - Diferimentos (exclui activos por impostos diferidos)	Q04A-A5120-1	Balanço - Activo corrente - Diferimentos	1
361	B049	Activo corrente - Caixa e depósitos bancários	Q04A-A5125-1	Balanço - Activo corrente - Caixa e depósitos bancários	1
325	B043	Outros activos correntes	Q04A-A5119-1	Balanço - Activo corrente - Outras contas a receber	1
328	B043	Outros activos correntes	Q04A-A5121-1	Balanço - Activo corrente - Activos financeiros detidos para negociação	1
329	B043	Outros activos correntes	Q04A-A5122-1	Balanço - Activo corrente - Outros activos financeiros	1
...	...	...	...	...	...
85	B013	Activo não corrente - Terrenos e edifícios	Q05A-05121A-A5738-6	Propriedades Invest. - Modelo do custo - Quantia líquida escriturada final em Terrenos e recursos naturais	1
86	B013	Activo não corrente - Terrenos e edifícios	Q05A-05121A-A5738-7	Propriedades Invest. - Modelo do custo - Quantia líquida escriturada final em Edifícios e outras construções	1
87	B013	Activo não corrente - Terrenos e edifícios	Q05A-05081A-A5640-1	Activos Fixos Tangíveis - Quantia líquida escriturada final em Terrenos e recursos naturais	1
88	B013	Activo não corrente - Terrenos e edifícios	Q05A-05081A-A5640-2	Activos Fixos Tangíveis - Quantia líquida escriturada final em Edifícios e outras construções	1
89	B014	Activo não corrente - Equipamento básico	Q05A-05081A-A5640-3	Activos Fixos Tangíveis - Quantia líquida escriturada final em Equipamento básico	1

92 rows × 5 columns

FIGURE 5 - Example of Output of the Drill-Down Tool

Results and Contribution to Efficiency

The implementation of the drill-down tool significantly improved the efficiency of the disaggregation process. Before automation, the manual process of identifying IES items through recursive filtering was highly time-consuming and error-prone, often requiring several minutes to complete for each indicator. By automating this task, the tool reduced processing time to mere seconds, resulting in a drastic reduction of manual intervention.

This automation not only enhanced speed but also minimized the risk of human errors, leading to more consistent and reliable data outcomes. The faster response times enabled by the tool allowed technicians to focus on higher-value tasks, rather than repetitive manual steps, and helped improve the accuracy of data analysis.

By the conclusion of this project, the drill-down tool has facilitated the response to three different user requests, each asking what IES items were used to calculate specific economic and financial indicators.

4.3.2. Process Analysis and Documentation of SAS Codes

The next task focused on a detailed review of the annual production process manual, which outlines all the necessary steps, programs, and associated paths required for execution.

A key challenge identified by the Central Balance-Sheet Division team was the limited knowledge of the programs used in the process. These SAS programs were developed years ago and lacked proper documentation, leaving technicians unclear about their operations and the tables they created or modified.

To address this issue, as part of the project, several SAS programs within the production process were analyzed and documented to clarify their functionality. To do so, documentation detailing the primary objectives of each program was created, and an Excel spreadsheet was developed to catalog the tables associated with each program. This catalog detailed the input tables, the tables modified during execution, and the output tables generated. Furthermore, during the analysis, the main bottlenecks in the programs were identified, highlighting areas that could benefit from future revisions. Potential

improvements were also suggested, which could enhance the program's efficiency and performance in subsequent iterations.

Results and Contribution to Efficiency

This task was critical in addressing issues faced during the previous annual production cycle, where errors in certain tables were difficult to trace due to the absence of clear documentation. The structured documentation created as part of this task provided a transparent and easily accessible overview of the programs, enabling quicker identification and resolution of potential errors.

Furthermore, as the programs are scheduled for future updates, the identification of bottlenecks and the suggestion of improvements will provide essential support for the technician responsible for the revisions. These insights will help optimize the program's performance and guide the implementation of more efficient solutions during the update process.

This task was essential in enhancing experience with SAS programming, a widely used tool in the area, and in developing a deeper understanding of the annual production process – two critical skills necessary for completing the remaining tasks.

4.3.3. New Production Process Design

The next step towards the optimization of the annual production process was to develop a proposal for a new system that would address the inefficiencies of the existing process while aligning with the department's long-term vision for system optimization. Therefore, the first phase of this optimization task was to identify the inefficiencies of the existing process using data-driven insights.

Process Inefficiencies

Despite the structured nature of the existing workflow, detailed in section 4.2.1., the current process suffers from several inefficiencies. These inefficiencies were identified using the knowledge gained during the project's first phase – the two previous tasks – and the insights provided by the team, benefiting from their expertise to identify key bottlenecks. It was concluded that there were several areas for enhancement, with the most significant ones being:

- Program Optimization: The current system has a large number of lengthy programs that could be optimized and consolidated to reduce computational time and improve the overall efficiency of statistical production;
- Manual Processes: There is a high volume of manual tasks, including adjustments, analyses, and other processes that could benefit from automation;
- Documentation and Data Models: There is insufficient documentation and data models for the programs, making it challenging to understand and maintain the system or adapt it to new data sources or analytical methods;
- Alert Analysis: Several challenges were identified in the alert analysis process, including the late generation of some alerts, their large volume, and the manual analysis, that complicates the prioritization and resolution of critical issues.

While the current process meets its core purpose, these inefficiencies are the result of an ad-hoc development process over time. The system was developed incrementally to address immediate needs, often prioritizing quick fixes instead of long-term data-driven solutions. Consequently, the system now relies on manual, inefficient, and time-consuming processes that could be optimized through automation and advanced data processing techniques.

New Production Process

To address these inefficiencies, a new production system was proposed. The new system, in Figure 6, builds upon the existing process but introduces several key improvements for optimization.

The new process begins similarly to the existing one, with the receipt of the IES and quality control of the microdata. A key improvement in the new design is the elimination of the SQL-to-SAS transition step, as all tasks would be completed directly using SQL. This change aligns with the department's objective to reduce its reliance on SAS, which is gradually being deprecated in the department.

Following this, the universe of active companies is defined, as was previously done, and a non-response treatment is applied to companies confirmed to be active through alternative data sources but that did not submit their IES. The granular data is then stored in the DW and serves as the basis for the aggregated series, which are also stored in the

DW. This marks the beginning of the DW's first layer, the staging phase. During this phase, data is incrementally stored in the DW as production progresses, capturing multiple execution cycles and being only accessible to the responsible division.

Subsequently, the quality control platform performs tests on the aggregated data in the DW, generating alerts that, instead of being manually reviewed, as it was previously done, would be reviewed using an alert analysis program, allowing technicians to prioritize and address issues efficiently

Once the data stabilizes or is ready to be shared with other divisions, it will move to the DW's second layer, the storage phase. There, the data will be accessible for cross-divisional analysis, allowing teams to compare results across multiple execution cycles.

Finally, the MAP tool would be applied to this data to generate vectors for publishing the information on BPstat and, once the data is published, it enters the final layer of the DW, known as the stable phase. In this phase, the data is finalized, consistent, and no longer subject to changes.

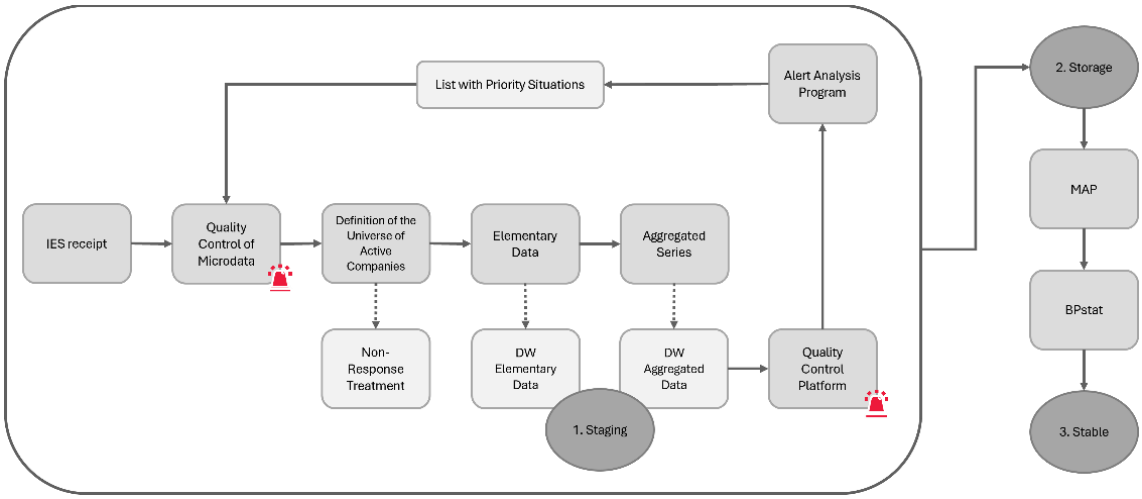

FIGURE 6 - New Production Process Design

Results and Contribution to Efficiency

The new production process design significantly contributes to improving system efficiency by organizing tasks in a more logical and optimized way. The elimination of the SQL-to-SAS transition step, for example, reduces unnecessary complexity and computational time, improving overall performance. By simplifying data flow and

minimizing transitions between different systems, this change ensures a more consistent and efficient process.

This design also represents a critical shift toward a more structured, data-driven approach, in line with the department's goal of centralizing and integrating data. It also played a crucial role in establishing the foundation for subsequent tasks and redirecting the project's focus towards achieving this new objective.

4.3.4. Alert Analysis Program

As previously mentioned, the alert analysis in the existing production process was identified as a significant challenge due to its reliance on manual reviews, which was particularly inefficient when dealing with large volumes of alerts. To address this limitation, an automated alert analysis program was developed as part of the new production process design, outlined in the previous section. This program automatically identifies the companies responsible for at least 30% of the series' annual variation. Since the main alerts typically highlight anomalies in variation associated with specific companies, the program allows technicians to detect potential issues more quickly. By adopting a data-driven approach, it improves the speed and efficiency of the analysis while minimizing the need for human intervention.

This program was developed in Python and it integrates data from multiple sources, including the DW, accessed through SQL queries, and comma-separated values (CSV) files essential to map business codes to their corresponding DW codes. Using the DW was crucial for enhancing project efficiency and aligning with the department's future vision for system integration, described in section 4.2.2.

The program was developed using a Python function that requires two inputs: the series number and the year the user is interested in analyzing. Then, the function performs the following steps:

1. Series Identification in CSV Mapping Files

The program begins by searching the series number provided by the user in two distinct CSV files. These files map the DW series to BPstat series, providing the necessary column combinations to locate a BPstat series within the DW. Each series corresponds to

a variable, and the separation into two files reflects the distinction between the two types of variables, which differ in their calculation methods:

- Simple indicators: Calculated using arithmetic operations such as summation or subtraction of several items;
- Ratios: Derived from the division of one item by another.

Despite their similarities in data structure, these variables are stored in separate tables and require different processing methods. Once the series number is located in one of the CSV files, the program flags the series type (simple indicator or ratio) and stores the corresponding row data in a dictionary for subsequent analysis.

2. Aggregated Value and Annual Variation

To optimize data processing, the program excludes irrelevant columns in the CSV mapping file by comparing them with the columns in the DW table. It then uses the dictionary of common columns to extract the corresponding aggregated value of the series, which corresponds to the value published in BPstat.

To provide additional insights, the program also calculates the annual variation of the series and provides a historical overview by displaying both aggregated values and yearly variations for the three most recent years. For example, if the user inputs 2024, the output will display the 2022, 2023, and 2024 values and the annual variations for 2023 and 2024.

3. Variable Identification

The program proceeds to identify the variable associated with the input series through a two-step process involving two CSV files:

- Relevant Columns Compilation: The first CSV file specifies the DW columns used to identify a variable. The program extracts these column names and stores them in a dictionary. The corresponding values from the row associated with the series provided by the user are then added to the dictionary;
- Variable Mapping: Using the dictionary created, the program filters this second CSV file to locate the variable whose mapping matches the dictionary, representing the variable of the series.

These variables are the economic and financial indicator codes analyzed in the drill-down tool, composed by a letter followed by three digits. The prefix of the code indicates whether the variable represents a ratio, which always begins with an "R" (e.g., "R456"), or a simple indicator (e.g., "B123"). This classification is essential, as the analysis process differs significantly between simple indicators and ratios.

For simple indicators, the analysis involves identifying the entities reporting the variable and determining the ones with the highest or lowest contribution to its variation, depending on the specific case. In contrast, the analysis of ratios involves a more complex process.

4. Treatment of Ratios

Variables that begin with an "R" are classified as ratios and require a specific analysis. This analysis involves decomposing the ratio into its numerator(s) and denominator(s) to identify the component contributing most significantly to the ratio by calculating their individual effects. The companies reporting the component with the greatest impact on the variation are then identified, along with their corresponding values.

To simplify the implementation of this analysis, four helper functions were developed. These functions extract the ratio formula, identify its numerator(s) and denominator(s), and organize all formula components into a list.

These functions were based on the "Formula" column in the Excel spreadsheet used in the drill-down tool. This column often contains complex conditional expressions, such as "IIF(ISEMPTY([B010]) or [B010]<0 or ABS(([B077]+[B082])/[B010]) > 100, null, IIF(ISEMPTY(([B077]+[B082])), 0, (([B077]+[B082])/[B010])))", which made it essential to create functions to simplify it.

The approach for the formula extraction function involved isolating the final part of the expression by finding the last comma, extracting the contents after it, and removing all brackets. The division symbol ("/") was then located, and brackets were inserted around the numerator and denominator to ensure the correct precedence of operations. For instance, the formula used as an example above would be simplified to (B077+B082)/(B010). A similar logic was applied to the remaining three functions.

After identifying the numerator(s) and denominator(s), their individual effects were calculated and stored in a dataframe for analysis. The script evaluated these effects to determine the component with the greatest contribution to the ratio, referred to as the "variable of impact".

5. Individual Data and Variation

The final step involves identifying the companies that contributed most significantly to the aggregated series. This process begins by locating the companies in the DW that reported the variable associated with the series (or the variable of impact in the case of ratios) for the specified year, along with their corresponding values. Since the microdata is also stored into two DW tables, according to the type of variable, the previously assigned flag was used to help efficiently extract the individual values.

After retrieving the microdata, the one-year change ($y_n - y_{n-1}$) is calculated for each company. The code then measures each company's contribution to the total variation in the aggregated series by using the formula:

(1) $$Contribution\ of\ Company\ X = \frac{OneYearChange_{\ Company\ X}}{\sum OneYearChanges}.$$

As output, only the companies accounting for 30% of the total variation in the annual series are displayed. As an example, Figure 7 illustrates the privacy-censored output of the alert analysis program for a series representing a ratio in 2020.

This approach effectively identifies key contributors by focusing on the top 30% of the total variation, ensuring the most significant impacts are captured without imposing arbitrary constraints, such as limiting the analysis to a fixed number of companies.

```
alert_program(12597001, 2020)
```

Temporal evolution of the series 12597001:

	2018	2019	2020
Value	-1.161474	-1.048397	-4.890121
Annual Variation (%)	None	-9.735634	366.437848

Formula: (D301-D316-D029)/(D020)

Individual effects of the components of the ratio:

	2020
Numerator effect (D301)	-0.054813
Numerator effect (D316)	0.006736
Numerator effect (D029)	-0.025687
Denominator effect (D020)	-0.033978

Variable of impact: ['D316']

Companies contributing to, at least, 30% of the total variation of the series in 2020:

	EntityID	Value_N-1	Value_N	One-Year_Change	Variation_Weight (%)
5823	[redacted]	[redacted]	[redacted]	[redacted]	11.99
14300	[redacted]	[redacted]	[redacted]	[redacted]	7.84
161211	[redacted]	[redacted]	[redacted]	[redacted]	7.83
14353	[redacted]	[redacted]	[redacted]	[redacted]	5.84

FIGURE 7 - Example of Output of the Alert Analysis Program

Results and Contribution to Efficiency

The automated alert analysis program offers significant advantages by aligning with the department's future goals of data integration and centralization through its direct connection to the DW. This connection ensures that the analyzed data is always current, minimizing the risk of working with outdated information.

This program significantly accelerates the alert analysis process. Traditionally, technicians would spend significant time preparing data and using inefficient tools to identify potential issues. In contrast, this program automatically processes the data in seconds, providing the same insights much faster. This not only reduces the time spent on routine tasks but also enables more timely analysis and issue resolution.

Furthermore, by focusing on the companies contributing to at least 30% of the total variation in the series, the program helps to differentiate between genuine errors and anomalies. If multiple companies are identified as contributing to 30% of the variation, yet their variation weights are low, it suggests that the alert may represent a typical

anomaly. However, if only a single company is identified, it indicates a higher probability of a data error, enabling technicians to prioritize their attention accordingly. This approach optimizes the workflow and enhances the overall effectiveness of the alert analysis process.

4.3.5. Reformulation of the Universe Definition Programs

The final task focused on redesigning the universe definition programs, which consisted of six separate SAS programs used to identify the universe of active companies based on data from the IES and other sources. These programs were selected as the next step in the optimization process once they represent the first phase of the SAS workflow, which is set to be deprecated and replaced with SQL, as outlined in the new design in section 4.3.3. Additionally, these programs had not been reviewed for an extended period, making it essential to reevaluate their logic and criteria to ensure they remain suitable for classifying companies as active. Furthermore, while documentation existed, it was outdated and lacked clarity, further highlighting the need for reformulation. Therefore, the main objective was to migrate these programs to SQL while improving their efficiency and ensuring comprehensive documentation. The reformulation process was divided into multiple steps:

- Program Analysis

The reformulation process began with a detailed analysis of the six programs and a review of the existing manual. While parts of the manual were outdated, it provided valuable context for understanding the overall workflow and logic. This initial analysis was crucial for building familiarity with the programs.

- Development of Relational Data Models

Following that, relational data models were developed for each program to visually represent their workflows. A unified relational data model including all programs was also created to illustrate the interactions among the six programs and their corresponding tables. This approach was essential not only for gaining a clear understanding of the program's operations but also for identifying potential areas for improvement and providing a foundation for the development of the new program.

- Criteria for Defining Active Companies

Then, a list of criteria and data sources used to classify companies as active was made. These sources included IES, Balance of Payments, Central Credit Register, and others, summarized in Table I along with brief explanations.

TABLE I - Sources used to determine the activity of a company

Sources	*Brief Explanation*
IES	Annual mandatory report for non-financial corporations.
ITENF	Quarterly survey of a sample of non-financial corporations conducted by Banco de Portugal and Statistics Portugal.
Balance of Payments	Records of operations with foreign entities.
Monetary and Financial Information	Data on monetary and financial aspects derived from companies' financial positions.
Central Credit Register	Information on loans granted to non-financial corporations by the resident financial sector.
Integrated System for Issuance of Securities	Information about the issuance of corporate bonds and securities.
VAT (Value Added Tax)	VAT-related information submitted by companies to Tax Authority.

- Assessment of Criteria Relevance

The next step involved assessing the relevance of each criterion. This was achieved by counting the number of companies classified as active under each criterion from 2020 onward. The results are displayed in the graph in Figure 8.

The analysis revealed that, as expected, the IES and ITENF were the most significant sources for classifying companies as active. Together, they accounted for 89.55% of the active classifications. This dominance is largely due to the mandatory nature of the IES, which all companies must submit, although some fail to do so occasionally.

When IES and ITENF data were insufficient to confirm a company's activity, the Central Credit Register combined with VAT information emerged as the second most important criterion, accounting for 10.10% of the active classifications. Following this,

the Balance of Payments data contributed 0.27% of classifications, while the Central Credit Register combined with Monetary and Financial Information accounted for 0.07%.

Finally, the Integrated System for Issuance of Securities provided the least information on company activity, classifying only 0.01% of active companies.

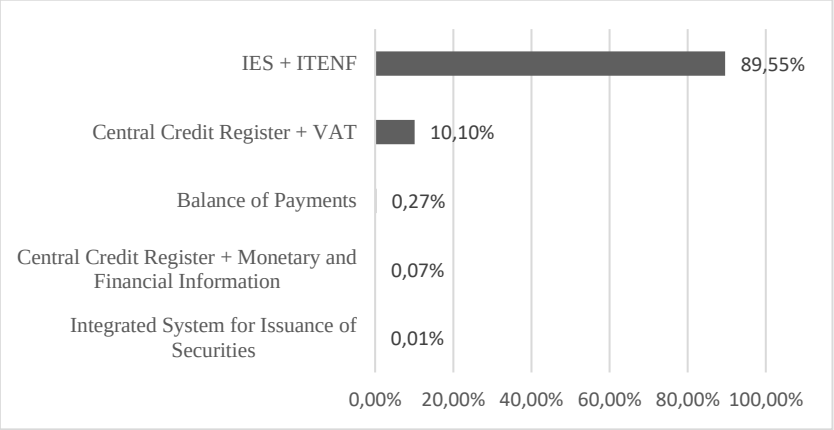

FIGURE 8 - Relevance of the criterion to classify the activity of companies from 2020 onwards

- Identification of Possible Improvements

Following the assessment of each criterion's relevance, the next step involved identifying potential areas for improvement during the transition to SQL. Several improvement opportunities were identified:

- Consolidation of Programs: It was identified that all six programs needed to be executed sequentially to produce the final output. This led to the recommendation of merging all six programs into a single SQL program that incorporated all the steps currently handled by SAS, therefore improving efficiency and manageability;
- Reduction of Redundancy: The first three programs performed identical processes for different indicators, resulting in unnecessary repetition and inefficiency. To address this, it was proposed to consolidate these programs into a single process to reduce redundancy and optimize the code;
- Automation of Manual Processes: The analysis also revealed several manual processes within the programs, increasing the risk of human error. To minimize this risk, it was proposed to automate as much as possible of the process, reducing the reliance on manual interventions and ensuring consistent and accurate outputs;

- Incorporation of Additional Sources: Another area for improvement involved the integration of a new data source: the e-Fatura. Managed by the Portuguese Tax Authority, e-Fatura tracks invoices and receipts issued by companies, providing detailed insights into their economic transactions. This data source could significantly enhance the assessment of a company's activity. Therefore, it was decided to evaluate the potential contribution of e-Fatura to the classification process and discuss its integration.

- Transition to SQL

After completing these steps, the final phase was the implementation, where the programs were reformulated and transitioned to SQL. This phase required identifying and mitigating limitations while implementing improvements to enhance performance and functionality.

Identified Limitations

One notable limitation encountered during this transition was the absence of specific built-in SQL functions commonly used in SAS, such as the transpose function. To achieve equivalent outputs, the SQL JOIN function was adapted and employed effectively.

Another significant challenge was SQL's lack of a direct functionality to export data to Excel, a critical requirement in this workflow. Therefore, the data had to be extracted using alternative tools such as Python, although this approach introduced additional complexity to the overall process.

Additionally, the unavailability of specific tables in SQL that were essential to the SAS program was another challenge. To address this, alternative tables were identified and used to provide equivalent information, ensuring the continuity of the process.

The most significant limitation, however, arose from SQL's inability to connect to multiple servers within a single query. This was particularly impactful once the required data was distributed across the division's server and the department's DW, which required a connection through another server. Therefore, to resolve this issue, Python was employed to extract the necessary tables from the DW and transfer them to the division's server. Although this solution successfully addressed the connectivity challenge, it introduced another layer of complexity to the workflow.

Improvements

During the transition to SQL, several improvements were implemented that resulted in substantial performance enhancements. These improvements were designed to optimize efficiency, reduce execution time, and optimize the overall process.

The first key improvement was code consolidation. Initially, the first three programs repeated identical steps for three different indicators. By consolidating these programs, the repetitive code was eliminated, allowing the process to be executed once and applied to all three indicators. This change not only improved efficiency but also simplified future modifications, as any necessary updates now only need to be made once, rather than three times.

Another important enhancement involved the use of Common Table Expressions (CTEs). Instead of creating intermediate tables during the process, CTEs were employed to perform complex queries in a simplified format. This approach increased the efficiency of the process while maintaining readability and reduced the need for additional data storage, preventing potential slowdowns.

Code simplification was also a key focus. Unnecessary parts of the code, particularly those that were computationally expensive and inefficient, were removed. This had a significant impact on execution time, as it eliminated operations that were not contributing to the output.

The transition also eliminated manual steps in the previous SAS programs. These manual tasks, such as adding the execution date or identifying the last available period in a table, were automated through built-in SQL functions. For instance, the GETDATE() function was implemented to automatically insert the current date, and logic was incorporated to automatically retrieve the most recent period, eliminating manual intervention and improving process consistency.

Year filtering optimization was another improvement. The original program was designed to run for all years starting from 2006, even if only a subset of years required updating. The new implementation ensures that the program runs only for the years that need updating, significantly reducing processing time by avoiding unnecessary operations.

Finally, filtering processes were optimized by replacing inefficient string-based conditions with numerical filtering. This change led to more efficient queries, resulting in reduced computation time and overall performance improvement.

Integration of e-Fatura Data

In addition to the previously mentioned improvements, a key enhancement in the new process for defining the universe of active companies was the integration of e-Fatura as a new data source. While the existing sources were already reliable and accurate, the addition of e-Fatura data provided an additional layer of precision. By using transactional data, this integration allowed a more accurate classification of companies, ensuring that the results better reflected their actual operational status.

With the new classification criteria, companies are considered active if they have issued any value of invoices, or if they have purchased goods or services worth over one thousand euros – this threshold was set to exclude non-relevant costs, such as those related to building expenses. As a result, approximately 3000 non-financial companies that would have previously been classified as inactive in 2023 were reclassified as active based on the e-Fatura data, highlighting the relevance of this new criterion in the classification process.

A current limitation is that the e-Fatura data is only available in SAS. However, given that the data is expected to be integrated into the department's DW in the near future, a temporary solution was implemented. A query was developed within SAS to extract and export e-Fatura data from SAS to the SQL server, ensuring continued access for analysis.

Results and Contribution to Efficiency

The implemented improvements led to a substantial reduction in execution time across multiple programs, resulting in a significant overall increase in process efficiency.

The first three SAS programs, which previously required a combined runtime of approximately 45 minutes, were reduced to less than 4 minutes, yielding an efficiency gain of over 91%.

The fourth program, which previously required approximately 20 minutes to execute, now completes in about 10 minutes using the new SQL program, resulting in a 50% reduction in execution time.

The fifth program, which originally took around 1 hour to execute, underwent significant improvements. Following its reformulation, the execution time has been reduced to approximately 15 minutes, resulting in an efficiency gain of 75%.

The sixth program, originally taking 20 minutes, was optimized to be completed in less than 3 minutes, achieving an efficiency gain of around 85%.

In summary, these improvements collectively led to a remarkable efficiency enhancement. Instead of running six separate programs, the workflow was consolidated into a single program. The total execution time, previously approaching 2 hours, was reduced to approximately 30 minutes, representing a reduction of 75% in overall runtime. Furthermore, the migration to SQL not only facilitated this reduction but also eliminated the need to transfer data to SAS for execution, resulting in additional efficiency gains.

5. Conclusions and Next Steps

This project successfully initiated the optimization of the annual production system in the Central Balance-Sheet Division, establishing a strong foundation for achieving substantial efficiency gains in future statistical production cycles. Several key optimization actions were implemented, focusing on leveraging data-driven tools and methodologies to enhance system performance.

The first significant contribution was the development of a drill-down tool using Python, that automated a previously manual and repetitive process. This innovation enabled a quicker and more efficient analysis of economic and financial codes, particularly in responding to user requests. By the conclusion of this project, the tool had facilitated responses to three distinct user requests, highlighting its utility and time-saving potential. Additionally, the tool contributed to the overall understanding of the data and supported the development of the alert analysis program.

Another key achievement was the creation of updated and organized documentation for the SAS programs in use. This step not only supported current operations but also identified bottlenecks and inefficiencies for future improvements. This ensured that knowledge transfer and future updates could occur more seamlessly.

However, the key optimization action during the project was the development of a new production process for the annual statistics. By applying knowledge of the data fields and identifying inefficiencies, the new workflow was designed to address these challenges effectively. This process was developed with a focus on aligning with the department's objective of integrating the DW into operations. The resulting workflow introduced a logical and efficient structure, creating a foundation to guide further optimization and ensure a more optimized production process.

As part of the proposed workflow design, an alert analysis program was also developed using Python, with the DW data as the primary source of information. The program focused on addressing the challenge of efficiently analyzing and prioritizing the large volume of alerts generated annually. Therefore, the tool enables the quick identification of the companies responsible for at least 30% of the total annual variation in the series, reducing the process to just seconds. By providing technicians with an

optimized method to focus on high-priority alerts, this program significantly improves the efficiency and effectiveness of the alert management process.

The final optimization action within the project's timeframe was the reformulation and migration of programs defining the universe of active companies from SAS to SQL. Prior to the transition, relational data models were created in order to document the programs and facilitate the future analysis and transition.

Subsequently, the new program was developed in SQL, addressing several identified inefficiencies, such as redundant code and manual steps, and adding a new data source, the e-Fatura. This restructured program resulted in significant efficiency improvements, reducing the total execution time from nearly 2 hours to around 30 minutes.

The optimization actions completed have already delivered measurable efficiency gains to the annual production system. However, this project represents only the first step of a broader, two-year reformulation process. The next steps aim to fully implement the new production workflow, addressing all the inefficiencies identified and integrating the DW.

Key future activities will include the revision and optimization of the SAS codes where bottlenecks were identified, ensuring an optimized performance and transition to SQL. Additionally, the entire production process will also be reorganized and updated to work seamlessly with the DW. These enhancements aim to create a system that is more reliable, efficient, and transparent for users, while also being adaptable to future needs and ensuring the production of higher-quality statistics on non-financial corporations.

REFERENCES

Adesina, A. A., Iyelolu, T. V., & Paul, P. O. (2024). Optimizing Business Processes with Advanced Analytics: Techniques for Efficiency and Productivity Improvement. *World Journal of Advanced Research and Reviews*, *22*(3), 1917–1926. https://doi.org/10.30574/wjarr.2024.22.3.1960

Bai, J., Tang, X., & Zheng, Y. (2023). Serving the truth: Do directors with media background improve financial reporting quality? *International Review of Financial Analysis*, *85*. https://doi.org/10.1016/j.irfa.2022.102452

Banco de Portugal. (2008). *Simplified Reporting: inclusion of IES in the Statistics on Non-Financial Corporations from the Central Balance-Sheet Database* (Supplement 1/2008 to the Statistical Bulletin, May 2008).

Bataev, A. V., & Davydov, I. S. (2020). The role of automation in improving the quality of enterprise business processes. *IOP Conference Series: Materials Science and Engineering*, *986*(1). https://doi.org/10.1088/1757-899X/986/1/012015

Batarseh, F. A., Kumar, A., & Eisenberg, S. (2020). *The History and Future Prospects of Open Data and Open Source Software*. https://doi.org/https://doi.org/10.48550/arXiv.2108.01592

Bitkom. (2023). *Open Source Monitor: Research Report 2023*. Retrieved October 26, 2024, from https://www.bitkom.org/sites/main/files/2023-11/Bitkom-Open-Source-Monitor-2023-EN.pdf

Breton, R., & Bossé, É. (2003). *The Cognitive Costs and Benefits of Automation*. OTAN.

Charles, E. (2024). *Data Validation Techniques for Ensuring Data Quality*.

Comissão de Normalização Contabilística. (2005). *Plano Oficial de Contabilidade*.

David, F., Abreu, R., & Carreira, F. (2014). *Simplified Business Information: a technical position in accounting and taxation*.

DeBarros, Anthony. (2018). *Practical SQL: a beginner's guide to storytelling with data*. No Starch Press.

E'leimat, D., Ebbini, M. M. Al, Aryan, L. A., & Al-Hawary, S. I. S. (2023). The effect of big data on financial reporting quality. *International Journal of Data and Network Science*, 1775–1780. https://doi.org/10.5267/j.ijdns.2023.7.015

Gardner, S. R. (1998). Building the Data Warehouse. *Communications of the ACM*, *41*(9), 52–60.

Hartley, J. L., & Sawaya, W. J. (2019). Tortoise, not the hare: Digital transformation of supply chain business processes. *Business Horizons*, *62*(6), 707–715. https://doi.org/10.1016/j.bushor.2019.07.006

Haslwanter, T. (2016). *An Introduction to Statistics with Python: With Applications in the Life Sciences* (W. K. Härdle, Ed.). Springer. https://doi.org/10.1007/978-3-319-28316-6

Herath, S., & Albarqi, N. (2017). Financial Reporting Quality: A Literature Review. *International Journal of Business Management and Commerce*, *2*. Retrieved October 19, 2024, from www.ijbmcnet.com

Herzberg, P. (1990). *How SAS Works: A Comprehensive Introduction to the SAS System* (Springer-Verlag, Ed.; 2nd ed.).

Hoffer, J., Venkataraman, R., & Topi, H. (2016). *Modern Database Management* (12th ed.). Pearson Education Limited.

IBM. (2021). *What Is Workflow Automation?*. Retrieved October 27, 2024, from https://www.ibm.com/think/topics/workflow-automation

IBM. (2024a). *What is automation?*. Retrieved October 26, 2024, from https://www.ibm.com/topics/automation

IBM. (2024b, April 5). *What is business process automation (BPA)?*. Retrieved October 27, 2024, from https://www.ibm.com/topics/business-process-automation

International Accounting Standards Board. (2010). *Conceptual Framework for Financial Reporting*.

Lechtenbörger, J., Vossen, G., & Hüsemann, B. (2000). *Conceptual Data Warehouse Design*.

Li, A. (2013). *Handbook of SAS DATA Step Programming* (CRC Press, Ed.).

List, B., Schiefer, J., Tjoa, A. M., & Quirchmayr, G. (2002). Multidimensional business process analysis with the process warehouse. In *Knowledge discovery for business information systems* (Vol. 600, pp. 211–227). Springer, Boston, MA.

Mckinney, W. (2017). *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython* (2nd ed.). O'Reilly Media, Inc.

Ministry of Justice. (2007). "Decree Law No.8/2007". *Republic Diary No. 12/2007*. Série I of 2007-01-17. 378-388. Retrieved October 10, 2024, from https://diariodarepublica.pt/dr/detalhe/decreto-lei/8-2007-522813

Mohapatra, S. (2009). *Business Process Automation*. PHI Learning Private Limited.

Morandini, M., Coleti, T. A., Oliveira, E., & Corrêa, P. L. P. (2021). Considerations about the efficiency and sufficiency of the utilization of the Scrum methodology: A survey for analyzing results for development teams. *Computer Science Review*, *39*. https://doi.org/10.1016/j.cosrev.2020.100314

Peng, C. Y. J. (2019). *Data Analysis Using SAS*. SAGE.

Pereira, H. J. (2011). *Simplified Business Information (IES)-Is coordination between public entities really possible?* (pp. 177-188). Statistics Portugal

Rathnayake, R. M. S. S., Rajapakse, R. P. G. S. N., & Lasantha, S. A. R. (2021). The Impact of Financial Reporting Quality on Firm Performance. *Journal of Business and Technology*, 53–67. https://doi.org/10.4038/jbt.v5i0.53

Redman, T., & Hoerl, R. (2022). *Data Quality and Statistics: Perfect Together?*

Sachdeva, S. (2016). Scrum Methodology. *International Journal Of Engineering And Computer Science*. https://doi.org/10.18535/ijecs/v5i6.11

Samad, T., McLaughlin, P., & Lu, J. (2007). System architecture for process automation: Review and trends. *Journal of Process Control*, *17*(3), 191–201. https://doi.org/10.1016/j.jprocont.2006.10.010

Silva, Y. N., Almeida, I., & Queiroz, M. (2016). *SQL: From Traditional Databases to Big Data*.

Srivastava, A., Bhardwaj, S., & Saraswat, S. (2017). SCRUM Model for Agile Methodology. *International Conference on Computing, Communication and Automation*, 864–869.

Stohr, E. A., Howe, W. J., & Zhao, J. L. (2001). Workflow Automation: Overview and Research Issues. In *Information Systems Frontiers* (Vol. 3). Kluwer Academic Publishers.

Tupsakhare, P. (2019). Python for Automation and Scripting: Streamlining Operations and Increasing Efficiency. *Journal of Scientific and Engineering Research*, *6*(9), 222–227. https://doi.org/10.5281/zenodo.13918609

Watson, H. J., Goodhue, D. L., & Wixom, B. H. (2002). The benefits of data warehousing: Why some organizations realize exceptional payoffs. *Information and Management*, *39*(6), 491–502. https://doi.org/10.1016/S0378-7206(01)00120-3