Ministério da Fazenda

Receita Federal

RECEITA DATA: THE EVOLUTION OF ANALYTICS IN RECEITA FEDERAL

ABSTRACT

In the 1970s, the importance of data for organizations began to gain prominence, with the increase in the volume of information generated and the search for simplified methods of storing and consuming this data. At the same time, the concept of the data warehouse began to form, fundamental to data storage technology, consolidated in the 1990s as a pioneer in the paradigm of building repositories suitable for manipulating information and extracting knowledge.

In the case of the Receita Federal do Brasil (RFB), which at the beginning of the 2000s already had a relevant set of systems and databases, the need to subsidize internal work involving extraction and crossing of data from different sources led to the emergence, in 2001, of the Data Warehouse Corporativo, considered the initial milestone of the RFB's analytical environment, which, since then, has significantly impacted several of the agency's work processes.

In this context, this article aims to demonstrate the history of RFB's analytical environment, as well as the current scenario of the data lake ecosystem. Additionally, tools and processes developed to maximize the results obtained with the use of the environment will be highlighted, as well as relevant work and projects that were made possible thanks to the increased maturity of both RFB users and the technological environment itself.

INTRODUCTION

The Receita Federal do Brasil (RFB), agency responsible for activities relevant to the Federal Public Administration and, consequently, to society, such as management and execution of tax collection activities, administrative tax collection, inspection, research and tax investigation, within the scope of internal taxes and foreign trade [1]. To carry out these activities, the agency manages a rich collection of data comprising a diversity of accessory obligations, electronic documents, registration data and history of national and international operations.

Considering only the systems developed and maintained by the Serviço Federal de Processamento de Dados (Serpro), the RFB's ecosystem of transactional systems is made up of more than 600 services/systems, distributed across hundreds of databases and bringing together hundreds of thousands of attributes. An additional complexity lies in the fact that this ecosystem has different technologies, both development and storage (relational and non-relational databases, documents, text files, etc.).

For example, Table 1 presents statistics on some of the most relevant topics for RFB's final work processes, in terms of number of files received and disk space occupied.

Source/System	Number of Files		File Size	
	Annualy	Total	Annualy	Total
Eletronic Invoice (NF-e)	4 billion	38 billion	13 TB	132 TB
Financial Transaction Events (e-Financeira)	3 billion	18 billion	17 TB	112 TB
Digital Tax Bookkeeping (EFD) - Contributions	15 billion	145 billion	7 TB	64 TB
Digital Tax Bookkeeping (EFD) - IPI/ICMS	15 billion	133 billion	10 TB	85 TB

Table 1: Volumetrics of some of RFB's critical systems [1].

In this context, over the last few years, RFB has needed to invest resources and people in methodologies and processes that enable the manipulation of these large volumes of data, with the aim of improving its decision-making process, improving its work processes and offering better services to society. Thus, the RFB's analytical environment emerged and developed, whose history, evolution and current scenario will be presented in the next sections, highlighting the main results recently obtained with the massive use of data.

2. HISTORY OF REFERENCE TECHNOLOGIES

In the literature, it is common to find references to decision-making support systems from the 1960s onwards, but they generally comprised expensive and specific solutions, which were hampered by the lack of architecture necessary to carry out optimized research and the absence of historical data. However, with the advent of Database Management Systems (DBMS) in the mid-1980s, better ways of accessing data, formatting and formulating queries made data manipulation less complex, but the focus remained on the structure of the processes at work rather than the business vision [2].

In the 1990s, new programming languages, added to the growing need to process large volumes of data, which never stopped increasing, led to the development of more modern alternatives to support decision making, based on Data Warehouse, OLAP and Data Mining. As a result, queries and reports began to be prepared by the users themselves, without the need for deeper knowledge of information technology [2].

In this context, data processing systematization processes were also developed, highlighting the process known as Knowledge Discovery in Database (KDD). KDD fundamentally consists of: structuring the database; data selection, preparation and pre-processing; transformation, adaptation and dimensionality reduction of data; and data mining. Such systematization enabled the popularization of new data analysis technologies, highlighting the concept of Data Warehouse [3].

The Data Warehouse (DW) can be understood as a centralized and optimized repository of historical data that will be used for analysis and creation of dashboards, providing support for decision making, generally based on sources such as transactional databases and spreadsheets, and with the

[1] Information obtained through the ContÁgil tool, with data available in Receita Data, July/2023 (annual estimates reported according to 2022 results).

help of applications that implement the concept of Business Intelligence. To this end, an extraction, transformation and load (ETL) process is carried out with the aim of sanitizing and ensuring data consistency, enabling the construction of denormalized schemes, which favor the execution of more efficient queries quickly [3].

With the emergence of big data, DW were no longer able to meet the needs of some organizations, leading to the emergence of the second generation of data analysis platform, known as Data Lake (DL). DL use low-cost forms of storage, such as Apache Hadoop's HDFS, and generally open storage formats, such as Apache Parquet and ORC. The architecture is generally based on schema-on-read, that is, data is not stored in rigid pre-defined structures, but in its original format, providing flexibility and agility in storing and accessing data, as illustrated in Figure 1 It should be noted that this flexible scenario requires investment in data quality and governance mechanisms, so that DL can be used in a solid and permanent manner [4].

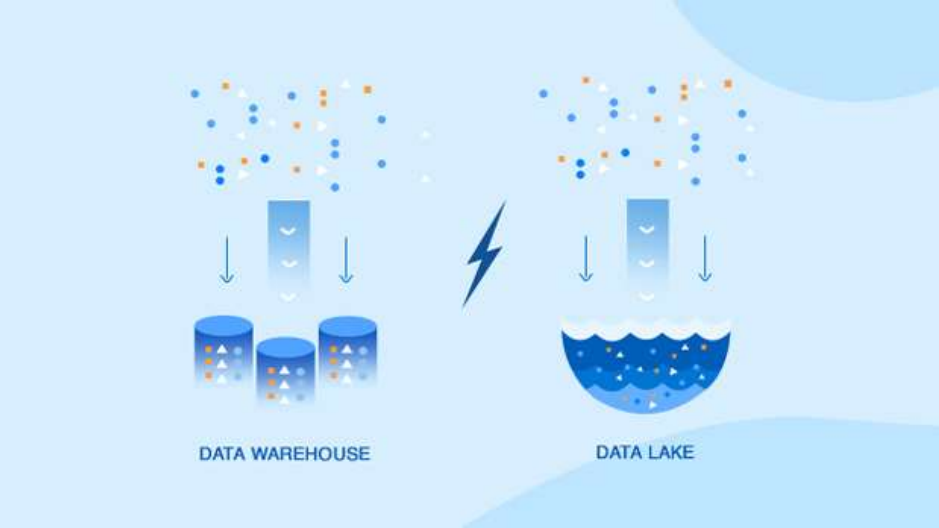

Figure 1: *Data Warehouse vs Data Lake*[2].

It is important to highlight that, despite initially imagining that DL would be alternatives to the use of DW, the most common form currently adopted is the implementation of so-called analytical environments, composed of both solutions, in addition to other technologies and tools that assist in the processing, analysis and governance of data necessary for organizations' work processes.

3. RFB ANALYTICAL ENVIRONMENT

RFB was a pioneer within the public service in electronically materializing various types of accessory obligations required by tax legislation. For example, in 1991 the annual adjustment declaration for Personal Income Tax (IRPF) began to be made electronically. In 1993, another system focused on foreign trade emerged - SISCOMEX (Integrated Foreign Trade System), starting with the export module, starting to also deal with imports from 1997 onwards. In 1998, Normative Instruction 127

[2] *Data Lake vs Data Warehouse: Key Differences*, disponível em https://sarasanalytics.com/blog/data-lake-vs-data-warehouse-differences/.

established the Integrated Declaration of Information Economic-Tax of Legal Entities (DIPJ), in addition to several other files and data collection systems [5].

3.1. RFB Corporate Data Warehouse

At a given time, there were several isolated systems in the RFB, each related to specific sets of data, without direct communication with each other. To support the internal work of selecting contributors, additional efforts were needed to extract data from different systems and cross-reference data, with a lot of manual work. The need to build a place to gather all this data - the Data Warehouse - became increasingly evident.

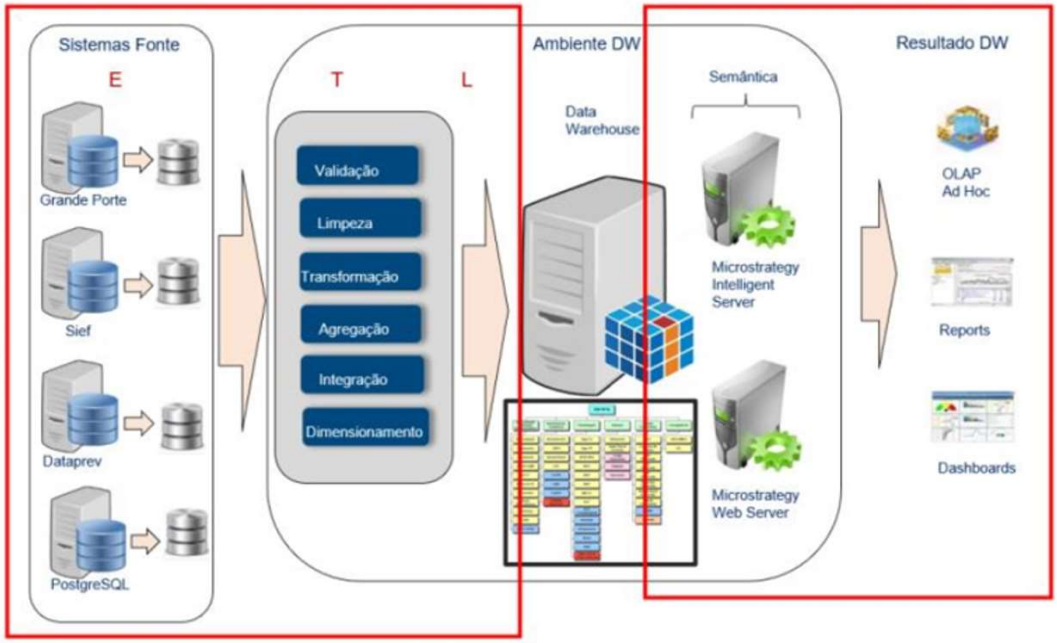

Figure 2: RFB Corporate DW Architecture.

RFB Corporate DW was created in 2001, with the purpose of being the means through which information collected from different systems was stored in a structure more suitable for carrying out queries. With DW it is now possible for each user to build and share their reports flexibly, bringing together data from different sources. The implementation of DW was an important step in the journey towards analytics at RFB, becoming the first milestone on this journey. The architecture adopted for the DW is illustrated in Figure 2.

3.2. ContÁgil

In 2007, a system initially focused on accounting was developed by RFB employees, called "ContÁgil". The name of the system derives from the expression "Contabilidade Ágil" (agile accounting in English), as its initial objective was to accelerate the work of processing and correcting accounting information, also including functionalities designed to facilitate the work of monitoring

this data. For example, in one of ContÁgil's innovative features it was possible to graphically visualize, through the so-called “Accounting Flow Chart”, all the movements that existed between all the accounts, using algorithms that reorganized the data in a more appropriate way.

The use of ContÁgil has become increasingly dynamic, starting to introduce other data in addition to accounting, such as: data from invoices, payroll, bank statements, foreign trade, among others. Many of the features initially implemented in ContÁgil had specific purposes for different work processes, but there were also some features that could be applied to any work context, such as the “Dynamic Analytical Model” (abbreviated as MAD). Through MAD, the user can build their own reports to cross-reference any data sources, just like DW.

An important characteristic of ContÁgil, in this context of analytics, is that initially only data stored locally on the user's machine was subject to manipulation, while DW only worked with data collected from server systems (such as SISCOMEX and DIPJ). Later, ContÁgil also started to use data stored on server systems, including direct integration with DW, in a move that can be considered the second milestone in the trajectory of analytics at RFB.

3.3. Public Digital Bookkeeping System (SPED)

The Public Digital Bookkeeping System (SPED), established in 2007 and implemented over the following years, initially defined layouts for accounting files, and years later began to define additional layouts for electronic invoices, tax records, financial transactions, and several others. Furthermore, SPED was innovative in determining the electronic format as the official format for presenting tax books. Before SPED, what was officially authenticated were paper books, with archives being mere representations of these. From SPED onwards there is a reversal in this scenario: what has legal value is the digital file, with the paper book being a mere representation of its content.

Over the following years, the mandatory and regular rules for delivering SPED files were established and expanded for all taxpayers, and with these large volumes of data were being provisioned in the national environment, which changed the panorama of the SPED's analytical environment. RFB, in training, and leveraged the use of ContÁgil as an essential tool in the daily lives of RFB employees. Currently, SPED accounts for most of the information stored by the RFB, with billions of files being received each year, as exemplified in Table 1. The creation of SPED, with the assimilation by ContÁgil and integration with the RFB's analytical environment, can be considered as the third milestone on this trajectory.

3.4. Receita Data (Corporate Data Lake)

Despite the great wealth of information, the process of structuring and making data available for internal consumption in the RFB work processes was still very time-consuming. File layouts evolved, with the addition of new fields and new structures, but it took a long time to map these data structures and make them available for use in DW. Furthermore, the technological infrastructure

behind DW was not scaled to make SPED's large volume of data available in full detail. As a result, it was necessary for the RFB to take another important step in the analytics journey: the implementation of a big data environment based on Data Lake architecture - Receita Data.

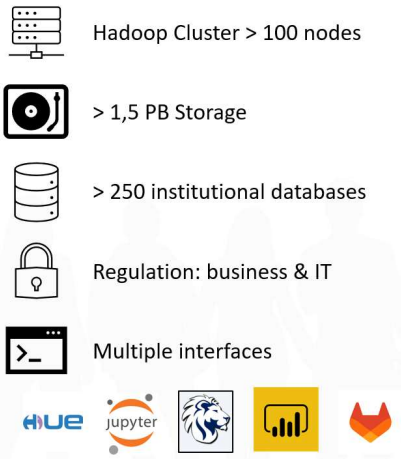

Figure 3: Receita Data overview [6].

Started in 2018, Receita Data has become the main environment for data analysis and data crossing. Based on “HADOOP” technology, Receita Data consists of a cluster of hundreds of machines, with high memory, CPU and disk capacity. Currently the total raw storage space already in use in Receita Data is about 4 petabytes (PB). To deal with such a large volume of data, the tools need to make use of distributed processing and be executed within the Receita Data computing environment itself. Receita Data is not a large database, but rather a platform with several services that include storage, processing and query activities, distributed across a cluster of several machines, which can be scaled horizontally as needed, by adding more machines to the cluster. An architecture that supports continuous growth in the volume of data and its use.

With the advent of Receita Data, it became possible to carry out consultations that would previously have been technically unfeasible. For example, in Receita Data it is possible to run a query that traverses trillions of electronic invoice item records, covering a period of several years and tens of millions of taxpayers, to find and summarize invoices that satisfy some searched criteria, examining the contents in the descriptions provided in these items. Depending on other criteria, queries like this can be carried out in Receita Data in a relatively short time, in the order of minutes or seconds.

Furthermore, in Receita Data the data transformation process becomes faster, making the process of making new data structures available for analytics use faster. It becomes possible for a greater number of people to participate in the modeling process of these data structures and develop different research strategies, and with this the importance of a data governance process within the institution also becomes more evident, establishing different roles and responsibilities.

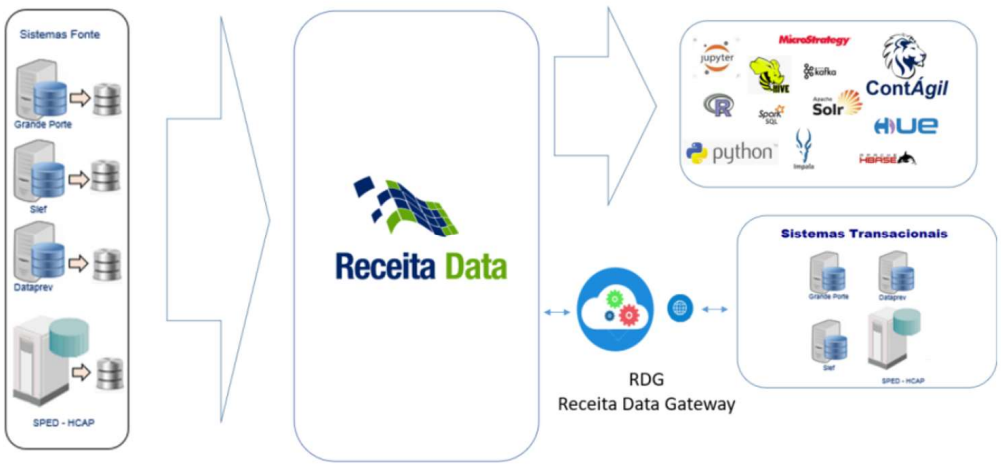

Figure 4: Receita Data architecture.

4. DATA GOVERNANCE FOR THE ANALYTICS ENVIRONMENT

The growing use of the analytical environment, as well as recent standards and regulations that began to demand, as a priority, the focus of public institutions on compliance risk, motivated RFB to start its Data Governance program. Started in 2019, the Data Governance work, initially focused on data from the analytical environment, was structured in order to produce artifacts that supported servers to assimilate best practices in the use and processing of data, in line with institutional guidelines. The structure of the RFB Analytical Environment Data Governance Policy is illustrated in Figure 5.

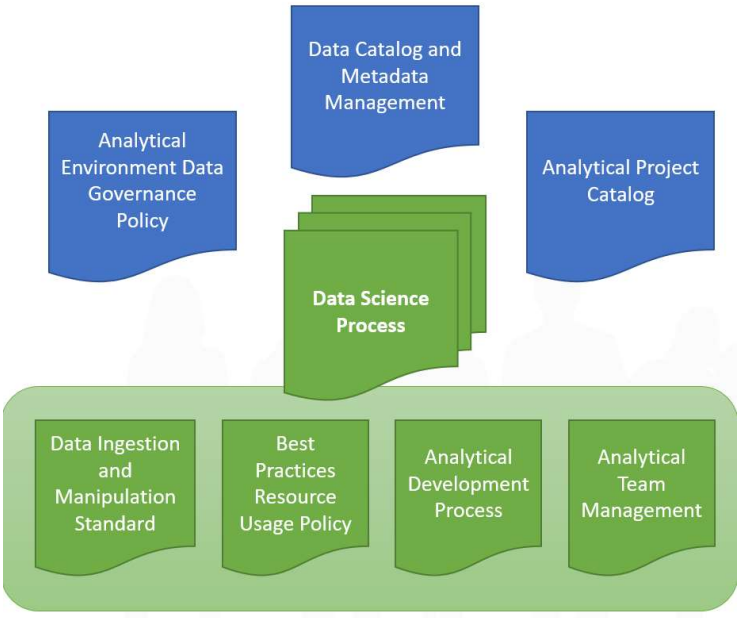

Figure 5: Artifacts of the RFB Analytical Environment Data Governance Policy.

Among the already published artifacts of the Data Governance Policy, the following stand out: the initial regulation that establishes the policy and defines initial guidelines and concepts; the

nomenclature standardization manual, which seeks to homogenize the designation that is assigned in the analytical environment and; more recently, the detailing of roles that, in addition to dealing with the most different levels of Data Governance Policy roles (data management roles, audit and control roles, technical implementation roles, auxiliary roles and data consumption roles), lists the participating internal committees, responsible for harmonizing both the development and application of the Data Governance Policy.

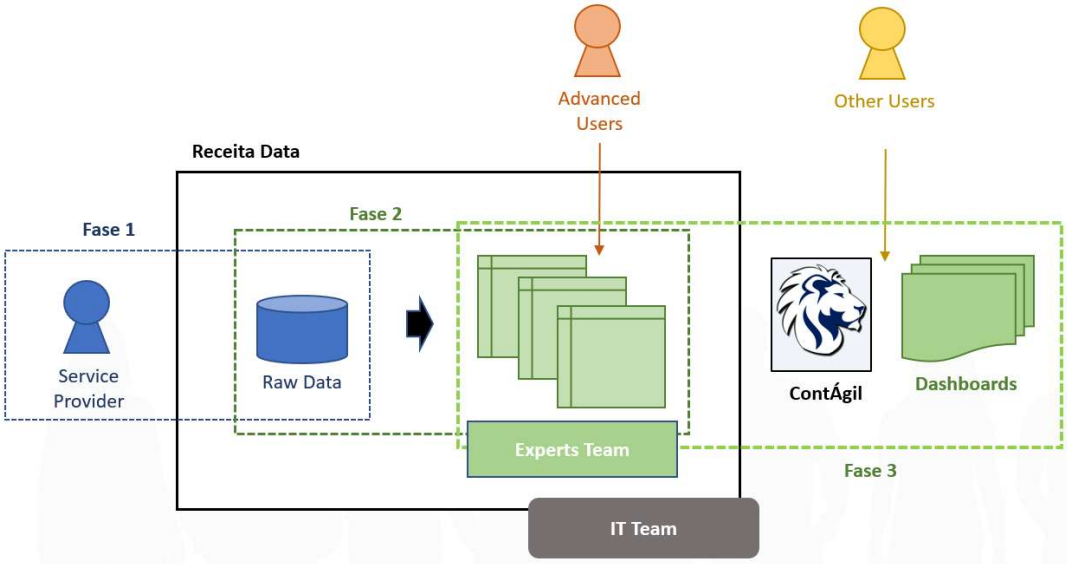

Figure 6: Vision adopted in the Data Governance Policy for the data lifecycle in the RFB analytical environment.

The RFB Analytical Environment Data Governance Policy adopts a vision for the data life cycle based on phases and iterations between teams, as illustrated in Figure 6. In this vision, data is categorized according to its degree of maturity, distributed into 3 phases distinct: in phase 1, the data is in its raw form, that is, as it is stored at the source, generally transactional, after going through an ingestion process in the DL; in phase 2, the data went through a data engineering process that comprises transformations and semantic analyzes of the raw data, producing a refined set of homogenized tables, which facilitates the work of data scientists and, finally; In phase 3, the data structured in phase 2 is mapped into graphical analytical tools, the main one being ContÁgil, enabling the consumption of information by servers who are not familiar with information technology concepts.

5. APPLICATIONS AND RESULTS

The RFB's analytical environment, over the last two decades, has brought numerous benefits to the agency. More recently, the Data Lake has stood out as an important source of information and insights for different strategic projects, among which the following stand out:

- **Electronic Invoice Bulletin (NF-e)**: in the early days of the Coronavirus pandemic, which most directly impacted Brazil at the beginning of the second half of March 2020, the need arose

to develop mechanisms to monitor and evaluate the impacts of sanitary measures on the national economy. Due to Receita Data, the RFB team was able to produce, in just a few days, daily monitoring reports on the issuance of Electronic Invoices which, when compared with previous data, provided an important thermometer, both in general terms and by activity and/or unit federation of the country. The production of this report required the development of scripts with the capacity to manipulate terabytes of information, the result of which would hardly have been achieved without using the potential made available by the RFB's analytical environment [7].

Figure 7: NF-e Bulletin, produced from Receita Data in order to support decisions due to the Coronavirus pandemic.

- **Tax Gap**: project that massively uses data and processing capacity from the RFB's analytical environment, with the aim of structuring the quantification of economic tax bases in a manner reconciled (coherent) with the National Accounts (CN) and enabling the production of matrix product-input (MIP) with a tax focus and Tax Gap calculation. These technical advances will complement the instruments used by the RFB's tax and customs studies area, enabling the production of revenue forecasts and estimates of the effects of tax policies and tax expenditures with macro and microeconomic consistency in terms of CN and MIP. They will also contribute to knowledge of effective tax incidence on economic sectors and income, which is essential to guide tax policy decision-making [8].

- **Sharing data with external bodies:** the analytical environment, especially Receita Data, has been fundamental in the production of knowledge to be shared with different bodies, aiming to support the definition of public policies with the most diverse approaches, such as structured registration bases for granting social benefits, monitoring economic indicators, among others. Since the start of Receita Data, more than a hundred demands of this nature have been met by RFB employees.

- **Pre-filled Personal Income Tax Declaration (IRPF):** the pre-filled declaration consists of a facility offered by the RFB to taxpayers, who need to fill out the IRPF Annual Adjustment Declaration. By crossing and consolidating data available in the analytical environment, the RFB provides several pre-filled fields of the declaration, such as: information on paying sources, originating from the DIRF database; real estate information, from DIMOB; notary information obtained from the DOI; data from healthcare providers, extracted from DMED; among others. It is estimated that, this year, more than 20% of the declarations received by the RFB relied on pre-filling [9].

The topics listed above are some examples of several successful cases in the use of RFB's analytical environment by its servers. Such cases were highlighted because they had a direct impact on the taxpayer's life. However, daily, several internal work processes are impacted by routines that make use of information extracted from the RFB's analytical environment.

6. CONCLUSION

In this work we sought to present a summary of the history of investments made by RFB, relating to resources and people in methodologies and processes that enabled the manipulation of these large volumes of data, over the last few years. Some milestones of the RFB's analytical environment were highlighted, which began in 2001 with the entry into production of the Data Warehouse, and consolidated more recently, in 2018, with the beginning of the use of the Data Lake.

Daily, new developments linked to the analytical environment, whether related to the treatment of new bases, or the development of programs through Jupyter[3], or even ContÁgil scripts, bring significant improvements to RFB processes, reflected in the form of operation rationalization, optimization of resources, increased transparency, among other benefits.

The analytical environment is dynamic, and new milestones are on the short and medium-term horizon. In the next few days, the integration of MS Power BI[4] with Receita Data should be in production, enabling the creation of numerous panels and reports from the crossing of different data sources. In parallel, additional studies involving cloud technologies, such as Databricks[5], may provoke the adoption of new paradigms in this trajectory of great changes experienced in recent years.

Finally, some important examples of successful use cases were cited that were only put into practice thanks to the RFB's analytical environment, with direct impacts on the lives of taxpayers and the quality of services provided by the Receita Federal do Brasil.

[3] https://jupyter.org/
[4] https://powerbi.microsoft.com/pt-br/
[5] https://www.databricks.com/

6. REFERENCES

1. Receita Federal. *Acesso à informação – Institucional*. Disponível em https://www.gov.br/receita federal/pt-br/acesso-a-informacao/institucional. 2023.

2. Gabriela Netto Guerra. *A Importância do Data Warehouse no Processo de Tomada de Decisão*. Fundação Getúlio Vargas. 2005.

3. Ramon Modesto Pacheco. *Guia de Boas Práticas para Implantação de Data Warehouse*. Universidade de Taubaté. 2019.

4. Flávio Lopes de Morais. *Data Lakehouse*. Universidade Federal de Lavras. 2022.

5. Receita Federal. *Tecnologia da Informação e Fluência em Dados*. 2023.

6. Receita Federal. *Government, quality, and information security of data within a CRM environment*. 2023.

7. Receita Federal. *Boletim da NF-e*. Disponível em https://www.gov.br/receitafederal/pt-br/centrais-de-conteudo/publicacoes/boletins/boletim-nfe/boletim-da-nfe-edicao-10.pdf/view. 2023.

8. Receita Federal. *Tax Gap*. Disponível em https://www.gov.br/receitafederal/pt-br/acesso-a-informacao/tax-gap. 2023.

9. Receita Federal. *Meu Imposto de Renda - Declaração - Declaração Pré-Preenchida*. Disponível em https://www.gov.br/receitafederal/pt-br/assuntos/meu-imposto-de-renda/preenchimento/declaracao-pre-preenchida. 2023.