MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

INTERNSHIP REPORT

DATA WAREHOUSE AND BUSINESS INTELLIGENCE APPLICATION DEVELOPMENT FOR IMPROVED DECISION-MAKING

BRUNA RAFAELA CASIMIRO DE MATOS

MARCH - 2023

MASTER
DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK
INTERNSHIP REPORT

DATA WAREHOUSE AND BUSINESS INTELLIGENCE APPLICATION DEVELOPMENT FOR IMPROVED DECISION-MAKING

BRUNA RAFAELA CASIMIRO DE MATOS

SUPERVISION:
PROFESSOR WINNIE NG PICOTO
JOSÉ PEDRO ANTUNES

MARCH - 2023

GLOSSARY

.PT – DNS.PT Association

BI – Business Intelligence

CDC - Change Data Capture

DDL - Data Definition Language

DML - Data Manipulation Language

DW – Data Warehouse

ETL – Extract, Transform and Load

IS – Information Systems

IT – Information Technology

OLAP - Online Analytical Processing

SCD - Slowly Changing Dimensions

TLD – Top-Level Domain

ABSTRACT

In the last few years, data volume has grown tremendously, and so has the importance of its analysis, given the potential knowledge emerging from the information it provides. Business Intelligence (BI) projects are getting a larger portion of enterprises' IT budgets, as they allow the improvement of decision-making and enhanced value creation (Farrokhi & Pokorádi, 2012).

This Master's Final Work describes the internship at DNS.PT Association, where a Data Warehouse and a BI application were developed for this organization. DNS.PT Association, referred to as .PT, is a non-profit association responsible for Internet domain management, maintenance, and registration under the TLD .pt, that is, the top-level domain corresponding to Portugal.

This report details, along with a theoretical background, the development of a Data Warehouse, with all its steps, including the ETL – Extract, Transform and Load – process. Subsequently, a BI visualization tool is used for the creation of dashboards to provide a better analysis of the data.

Given the importance of documentation in any IS project, a brief overview of the Data Management and Governance Model is presented.

Keywords: Business Intelligence, Data Warehouse, ETL Processes

RESUMO

Nos últimos anos, o volume de dados tem aumentado substancialmente, assim como a importância da sua análise, dado o conhecimento que surge da informação que proporciona. As empresas têm aplicado cada vez mais do seu orçamento em TI, em projetos de *Business Intelligence* (BI), sendo que permitem uma melhor tomada de decisão e maior criação de valor (Farrokhi & Pokorádi, 2012).

Este Trabalho Final de Mestrado descreve o estágio na Associação DNS.PT, onde uma *Data Warehouse* e uma aplicação em BI foram desenvolvidos para a organização. A Associação DNS.PT, abreviadamente designada por .PT, é uma associação sem fins lucrativos responsável pela gestão, manutenção e registo de domínios de Internet sob o TLD .pt, ou seja, o domínio de topo correspondente a Portugal.

Desta forma, este relatório explica, acompanhado de um contexto teórico, a implementação de uma *Data Warehouse* e o seu processo de *ETL – Extract, Transform and Load*. Para além disso, uma ferramenta de visualização é utilizada para a criação de *dashboards*, de forma a obter uma melhorar análise dos dados.

Por fim, devido à grande importância de documentação em qualquer projeto de SI é apresentada uma breve descrição do Modelo de Gestão e Governação de Dados.

Palavras-chave: *Business Intelligence*, *Data Warehouse*, Processos de ETL

TABLE OF CONTENTS

TABLE OF FIGURES

1. INTRODUCTION

The present Master's Final Work consists of a description of the internship carried out, concerning the Master's Degree in Data Analytics for Business, in a collaboration between ISEG – Lisbon School of Economics and Management and DNS.PT Association, referred to as .PT.

.PT is a non-profit association responsible for domain management, maintenance, and registration under the TLD .pt, which is the top-level domain corresponding to Portugal.

In the recent years, data has grown immensely, and so has the importance of its analysis given the knowledge emerging from the information it provides. This development has led to the rise of a data-driven decision-making era (Vassakis et al., 2018).

Data is constantly being produced through internet browsing, travel booking, or online shopping. Given this fact, it is anticipated that the volume of data will be multiplied by five between 2018 and 2025 (European Commission, 2022).

This has led to an increased need for Business Intelligence projects, where enterprises are allocating a larger portion of their IT budgets to such projects. This allows the application of information for the support of processes and decisions. It also helps identify the areas in which the organization should apply its resources and where it should focus on (Farrokhi & Pokorádi, 2012).

Data is obtained from the company's operational system and placed in the DW/BI system, which allows its analysis and improves decision-making (Kimball, 2013).

The management and analysis of data allow its transformation into knowledge, which gives organizations the ability to use it as a competitive advantage and to consider it an asset (Kopanakis et al., 2016).

The internship had a duration of 6 months, from September 2022 to March 2023, and it had the purpose of developing a Data Warehouse, a data management and governance model, and the data visualization.

In this BI project, the objectives were for the intern to understand the ETL process as well as to gain knowledge and help in the development of the data model and structure.

1

It follows a Kimball approach, starting with the project planning and definition of business requirements, and only after that the dimensional modelling, ETL process, and implementation of the physical design.

The work focuses on just one core business event of the company, being considered a prototype. Once finished, the same should be performed for other events.

This report is divided into six chapters: Introduction, Literature Review, Company and Internship Plan, Project Development, Discussion and Conclusions.

The introduction explains the internship's matter and goals, as well as its relevance.

The second chapter regards the literature review, which presents a brief explanation of each topic considered important in the comprehension of this work. It is separated into five subchapters: The Importance of Data in Organizations, focusing on the growth of data and its significance in the decision-making process; Data Quality Management and Governance, on the decisions and implementations that need to be made to ensure the quality of the data; Data Warehouse, understanding what it consists of and its advantages; ETL Process, this is the extraction, transformation, and loading of the data that will then be inserted in the DW; and to finish Business Intelligence, that allows the understanding of the data to make decisions.

It follows the third chapter about the company and internship plan which elucidates the company and its work, as well as a brief description of the internship plan and objectives. After that, there's the chapter about the project development, namely the creation of the Data Warehouse, management and governance model, and visualization of the data. To finish, there is a critical analysis of the activities performed, and some conclusions are drawn.

2. LITERATURE REVIEW

2.1. Importance of Data in Organizations

"The goal is to turn data into information and information into insights" – Carly Fiorina

Over the years, as organizations have seen an increase in the amount of data available, it has become more and more important to analyse it to get insights and, this way, help the decision-making process.

Thus, data analysis has become a fundamental part of managing a business, as it has become a huge asset and source of knowledge. Not only that but it is also anticipated that the volume of data will be multiplied by five between 2018 and 2025 (European Commission, 2022).

Moreover, Statista reports that data all over the world is rising exponentially, being set to grow from 33 zettabytes in 2018 to 2100 zettabytes in 2035. This has inevitably led companies to ponder how to gain value from the increasing volumes of data (Zhang et al., 2022).

Quality data contributes to the discovery of evidence regarding different issues, which will then lead to informed and calculated conclusions, as basing decisions solely on assumptions and abstract observations may result in wasted resources (The Council on Quality and Leadership, 2019).

Managers and experts of companies aim for better organizational performance. However, as Peter Drucker said: "You can't manage what you don't measure" (Wang et al., 2022).

This way, data, and its analysis will allow companies to productively point out the source of the problems they may be facing, as it also provides the visualization of relationships between the current situation and different locations, departments, and systems. It improves efficiency as it will support the discovery of the areas that should be prioritized, directing its resources to it (The Council on Quality and Leadership, 2019).

As much as there are advantages to data analysis and its use for decision-making, it is also important to point out that there are some challenges associated with it. These difficulties may not necessarily have to do with technology, as they might be caused by

the need for cultural change inside the company since people and organizations need to adapt to change (Harvard Business Review, 2022).

2.2. Data Quality Management and Governance

As organizations gather higher amounts of data, managing it becomes even more crucial every day. This way, data quality management is key when it comes to making sense of the data, as the goal is to create insights (SAS Institute, 2021).

Data quality management combines the organizational culture, technology, and data to ensure that the results are correct and helpful, building support for all business initiatives, as poor-quality data can lead to unnecessary costs and mistakes (SAS Institute, 2021).

There are multiple dimensions regarding data quality management, including accuracy, completeness, consistency, age, and uniqueness. Moreover, it is also fitting to mention some features used to help refine the credibility of the data, being data cleansing, which helps correct duplicates, nonstandard representations, and unknown data types; data profiling which uncovers relationships and validates the data; and validation of business rules and creation of a business glossary (SAS Institute, 2021).

Data quality is very relevant when it comes to the implementation of a data governance framework (SAS Institute, 2021).

Data governance specifies the framework for decision rights and accountabilities to stimulate preferable behaviour in the valuation, creation, use, and control of data (Gartner, 2019).

The key difference between 'governance' and 'management' is that the first one focuses on the decisions that are necessary to be made, as well as who makes them so that in the end, not only effective management but also the use of resources are secured. On the other hand, the second one concerns the implementation of those decisions (Alhassan et al., 2016).

Data governance is emerging in organizations all over the world, as they're becoming more and more focused on the management of data as an asset (Cheong & Chang, 2007).

Panian (2010) also mentioned that the key driver of data governance is data being considered an enterprise asset (Alhassan et al., 2016).

Many large companies such as IBM, Google, and SAP allocate plenty of attention to data governance, as they see it as a way of fast-tracking decision-making, rising legitimacy, and expanding collaboration (Zhang et al., 2022).

On the other hand, there are still gaps in how enterprises implement it, particularly given that the amount of data continues to grow exponentially, and the data governance processes are still vague (Zhang et al., 2022).

2.3. Data Warehouse

A Data Warehouse is a storage architecture designed to hold data extracted from operational and transactional data systems, as well as external sources. It is then combined to enable analysis and reporting for predetermined business needs (Gartner, 2019).

The operational data originates from business activities, whereas the external data can be from newspapers, websites, and customers. Given the importance of ensuring that its storage allows for smooth retrieval and use, Data Warehouses are a fundamental component of Business Intelligence for organizations, collecting data from different sources into one place for refined analytics and decision support (IBM, 2020).

Robert Inmon and Ralph Kimball are the authors of the two most common architectural approaches. Although they have some similarities such as the use of time-stamped data and ETL processes, there are also major differences (Brelin, 2004).

Inmon describes the term 'data warehousing' as a subject-oriented, integrated, nonvolatile, and time-variant collection of data that sustains decisions oriented to the major areas of the company (Inmon, 2005). This author recommends a top-down approach starting with the construction of the Data Warehouse and only after that it is separated into individual data marts that focus on different departments (Breslin, 2004).

As for Kimball, he suggests a bottom-up approach that uses dimensional modelling. It starts by understanding and documenting the business needs and creating one database

or data mart for each core business process, and only after that, the building of the Data Warehouse occurs (Breslin, 2004).

There are five components of a DW according to Gartner (2019), including the production of data sources, data extraction and conversion, Data Warehouse database management system, and administration and Business Intelligence tools.

As mentioned previously, there has been fast progress in technology, which has led to an increase in the amount of data available. This has created a challenge for companies in terms of organizing and accessing data with accuracy, relevance, and being up to date. If not properly gathered, it may cause some difficulties in decision-making, as well as the disappearance of files, loss of customers, misuse of resources, and absence of consistency (Chinyere, 2022).

In terms of its architecture, DWs have three tiers. Starting with the bottom tier, which is a server that collects, cleans, and transforms the data from various sources, through an ETL process, which will be explained in the next chapter. Then, there's the middle tier that contains the OLAP server – online analytical. Finally, the top tier is represented by a front-end user interface or reporting tool that allows users to conduct data analysis on their business data (IBM, 2020).

A DW logical model, depicts how the model is constructed in the database, that is it contains the attributes' names, data types, and constraints, as well as all the relationships, including primary key and foreign key (IBM, 2021).

There are two commonly used types of schemas, the star schema – which is one fact table connected to several denormalized dimension tables – and the snowflake schema – where the fact table, unlike the first one, is linked to various normalized dimension tables (IBM, 2020).

As briefly discussed, there are many benefits to creating a Data Warehouse in an organization. It allows better data quality by cleaning, eliminating duplicates, and standardizing it to have a single source. Not only that, but it also grants quick business insights, as it allows for consideration of all the company's data in each business decision (IBM, 2021).

It also concedes brighter decision-making as it supports data mining, artificial intelligence, and machine learning tools, providing the achievement of competitive advantage (IBM, 2021).

According to Forbes (2022), Data Warehouse projects are more likely to fail if treated as purely technological projects as they're not focusing on the end users and returning value.

2.4. ETL Process

The extract, transform, and load – ETL – process, briefly mentioned in the previous chapter, of the Data Warehouse and Business Intelligence environment, that is composed by a work area, data structures, and a set of processes, acting as an intermediary between the operational source systems and the DW/BI presentation area (Kimball, 2013).

The first step of this operation is extraction, which consists of acquiring the data from internal and external sources of an organization (Awiti & Zimányi, 2020). This involves reading and understanding the source data and replicating the data into the ETL system for additional manipulation (Kimball, 2013).

The second step would be the transformations that can take multiple forms, such as the cleaning of the data, which includes the amendment of misspellings or resolving conflicts, as well as, handling missing elements or eliminating duplicate data. Finally, the last step is the loading of the data into the dimensional models (Kimball, 2013).

This process enhances the value of the data, as it contributes to the metadata analysis, which can be beneficial for improving the quality of the sources (Kimball, 2013).

2.5. Business Intelligence

Business Intelligence is a term that emerged in 1989 at Gartner Group and was first used by Howard Dressner.

BI was introduced as a tool, product, and system to enable accurate and intelligent business decisions in the shortest possible amount of time, incorporating applications and analytics built on operational and analytical databases (Wang et al., 2022).

It focuses on analysing large quantities of data about a company and its operations, incorporating data gathering, data storage, and knowledge management, to obtain inputs to aid the decision-making process (Negash & Gray, 2008).

BI enables enterprises to make greater and faster business decisions, as it can extract and present information rightfully and purposefully, which is essential for any business. The Data Warehouse, which was explained in the previous chapter, is the core of the Business Intelligence environment, being a vast database that contains all the data needed (Balaceanu, 2007).

This way, it allows the organization to collect up-to-date information, present it through comprehensible tables and graphs, and deliver it fast as well as support the questions that will make sense of the who, what, where, when, and why of the market (Oracle, 2017).

As Business Intelligence helps collect crucial information from various unstructured data sources and transform it into usable information, it allows us to make enlightened policy decisions and enhance business efficiency and productivity (Niu et al., 2021).

These days, countless organizations have used BI to reach certain business needs. On the other hand, some still haven't been able to maximize the insights provided by Business Intelligence processes, mainly due to the absence of good guiding BI architecture as a support for the implementation of the system (Ong et al., 2011).

Moreover, it has recently evolved to comprise additional processes and activities to help refine performance, that includes data mining, which uses statistics and machine learning to uncover trends; reporting, for sharing data analysis to stakeholders; descriptive analytics, for preliminary data analysis; querying, that asks data-specific questions; data preparation, combining multiple data sources; among others (Tableau, 2022).

As enlightened before there are several benefits to the use of Business Intelligence, being important to highlight some, including data clarity, higher efficiency, greater customer experience, and superior employee satisfaction (Tableau, 2022).

3. COMPANY AND INTERNSHIP PLAN

The purpose of this section is to give a brief description of the company where the internship took place, as well as go over its plan and the project's methodology.

3.1. .PT Description

.PT is a private non-profit association responsible for domain management, maintenance, and registration under the TLD .pt, meaning the top-level domain corresponding to Portugal. It has multiple members including FCT, IP – Foundation for Science and Technology, IP, - ACEPI – Digital Economy Association – and DECO – Portuguese Association for Consumer Protection (.PT, 2022).

A domain is a name that can be quickly retained and is used to detect and identify computers on the Internet, assembled so that it is easier to memorize computer addresses. That is, when we visit a website, our computer requires to know the location of the server of the destination page to ensure the right address is being indicated, to provide the requests sent by our computer to the Internet (.PT, 2022).

When it comes to registration and management of domain names, .PT works with Registrars, accredited by the organization through a protocol, which remarks reciprocal rights and obligations, and this way it allows the management of domain names to be faster and more flexible (.PT, 2022).

.PT's vision is to promote the use of the internet and digital skills in Portugal, ensuring their service is a reference to the community, counterparts, and partners (.PT, 2022).

With this in mind, it has various initiatives and programs to achieve this goal. Starting with INCoDe.2030, which is a governmental program that focuses on the build-up of digital skills. Similarly, there is MUDA that aims to boost the participation of Portuguese citizens in the digital space, as well as Portugal Digital which plans to transform the country, through the digital empowerment of people and companies (.PT, 2022).

There is also ENH – Empresa na Hora or 'company on the spot' – which encourages the existence of companies, associations, or branches created on the spot with a .pt domain, following a protocol signed by the Ministry of Justice – RNPC (.PT, 2022).

In terms of data protection, as .PT aims to guarantee the privacy and safety of the data made available by the users, it uses a WHOIS policy, which is a free public directory that allows the identification of data linked with the registration and technical maintenance of a domain name (.PT, 2022).

3.2. Internship Plan

The internship carried out at .PT had the duration of 6 months, from September 2022 to March 2023, in which the intern performed the role of a data analyst at the innovation department of the company.

The project was developed with the guidance of two people inside the company, a data analyst and the director of the department, with daily meetings to discuss any doubts and/or determine the next steps, as well as for the analysis and evaluation of the status of the project.

The main objectives of the internship were for the intern to gain knowledge and help in the development of the data model and structure of .PT. This way, it focused on the analysis, definition, and implementation of a Data Warehouse, following an ETL process.

To achieve the goals mentioned above, the internship was organized into three major tasks/phases. The first one was for the intern to give support in the building of the Data Warehouse of .PT, in its different stages, that is requirement gathering, data model design, and development. Secondly, it was to help in the development of the Data Management and Governance Model of the company. Finally, the last activity was to assist in the creation of data visualization models in Power BI.

3.3. Project Methodology and Tools

Given the fact that the company didn't have a Data Warehouse, it was necessary to develop one from scratch.

It was decided that, firstly, the work would be focused on one operational core business event of the company, creating a prototype, which will allow an end-to-end

implementation, that is from the database to the dashboard. It also enables to quickly review and tune the plan, if necessary, as well as an earlier assessment of risks and issues.

Consequently, the work would begin with *Domain Registration*, considered the most relevant, as .PT is responsible for domain management, maintenance, and registration.

As for the creation of the data model design and development, it mostly followed a Kimball approach. It starts with the project planning as well as the definition of business requirements and after that comes the dimensional modelling, ETL process, and implementation of the physical design.

All the operational data of .PT is obtained from its primary source system, SIGA, which is a domain registry and management system settled in the Outsystems technology, capturing business transactions. This data was then made available in SQL Server.

After the development of the first core business event, the same practice should be replicated for the other events, which combined would deliver the final Data Warehouse.

The work on the construction of the DW is performed in two different environments: development, and quality.

The first one – development – is where the construction of the Data Warehouse occurred, which only had monthly updates of new information. On the other hand, the second – quality – mirrored the production environment, having new information around every 15 minutes, and this way, allowing the performance of tests, for quality assurance, to ensure that there aren't bugs or errors.

The final step will be to use a Business Intelligence visualization tool – Power BI – to get insights from the data available, through the creation of dashboards.

4. PROJECT DEVELOPMENT

This chapter aims to give a detailed explanation of the activities that were performed for the internship, such as the creation of a Data Warehouse and its subsequent visualizations developed in Power BI.

After the initial period of integration at the organization, with an understanding of its core business, the intern began to understand the current state of that project at .PT and was given access to the source system, SIGA, and to all the documents necessary to start to work on the project. The intern used multiple tools, such as Excel, Microsoft SQL Server Management Studio, and Power BI.

4.1. Data Warehouse Design

4.1.1. Dimensional Modelling

.PT did not have a Data Warehouse created, meaning that everything had to be done from the start. Justified by the multiple core business events of the company, the plan was to begin with the development of a prototype for one business event, that would then be replicated throughout the remainder.

The first step was to identify and analyse the core business events of the company, and after that it was decided that the prototype would be developed for the *Domain Registration* process, as it was considered the most important one, given that the company is responsible for the domain management, maintenance, and registration. Secondly, concerning the granularity, it was decided to be daily.

Thirdly, the dimensions for this fact were defined, and the star schema was created, which can be visualized in figure 1 below.
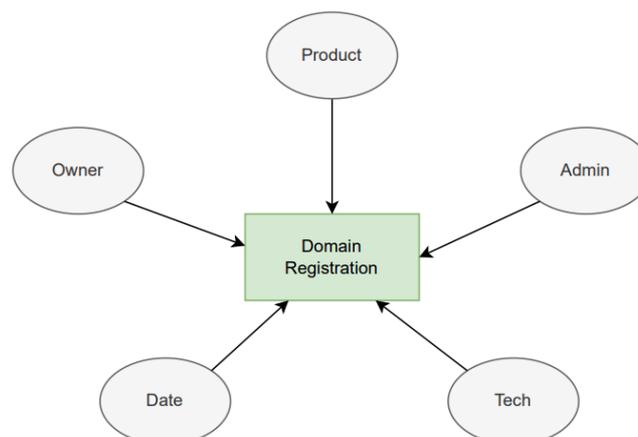

Figure 1 - Conceptual Model of the Star Schema

The five dimensions chosen, comprise the storage of all performance indicators found to be relevant to be included.

In this case, it was important to not only have information about the domain's owner - *Owner* -, administrator - *Admin* -, and technical responsible – *Tech* – but also on the details of the domain itself - *Product* - and certain dates from different occurrences along with the registration of the domain - *Date* -.

Finally, as the source system has a vast number of tables, it was then crucial to perform an intensive study of the data it provided, that is, analysing the utility of each attribute and determining what made the most sense to include in our Data Warehouse, which would then be used to populate the dimensions just shown.

Some tables that were obtained from the source system are *Domain*, *Entity*, *Address*, *Domain Purpose*, *Domain Status*, and *Entity Legal Status*, among others.

4.1.2. Physical Model

After determining some crucial first steps, followed the development of the physical model of the Data Warehouse. Our architecture is separated into two main areas: the staging area and the DW area, which can be seen in figure 2.
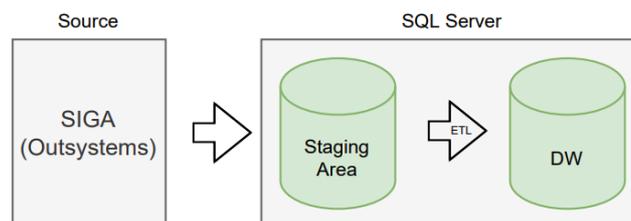

Figure 2 - DW Architecture

The staging area supplies a buffer between the data sources and the Data Warehouse. These are intended to be tables that will include the new inserts or updates that need to be implemented in the Data Warehouse.

This way, a staging table was created for each source system table the team found necessary. Some examples of the staging tables created are *Domain*, *Address,* and *Entity*.

For these, it was decided to keep the same table names and column names as the operational system ones. Nevertheless, it is important to mention that not all columns present in the source's tables were imported.

On the other hand, some of the tables only needed to be extracted to the staging area once, as they were purely informational tables that do not depend on the event in question and do not change over time. This way, they were just loaded during the historical data population phase and not reloaded afterwards. Some examples of those tables are *Country*, *Entity Status*, and *Domain Status*.

Additionally, a temporary table was created for the dimensional tables of the star schema to help ease the loading process.

After establishing a connection to the data source and having all the necessary data extracted to the BI environment, the next step would be to perform the ETL process, meaning, extracting, transforming, and loading the data from the staging tables to the Data Warehouse.

Regarding the Data Warehouse area, it includes the fact table and dimensional tables for the event.

To create those tables, it was necessary to start with the DDL (Data Definition Language) in SQL Server, which is the construction of the script that will not only be describing the data but also will establish the primary key and foreign key relationships between the different tables in the Data Warehouse. It was decided to use surrogate keys as the primary keys, which will be explained more in detail in the next few chapters. This was performed using the command *CREATE TABLE*, shown in figure 3.

```
CREATE TABLE dbo.Domainregistration_facttable(
    CREATEDAT_Date_key bigint NOT NULL,
    EXPIRATIONDATE_Date_key bigint NOT NULL,
    DOMAIN_ID bigint NOT NULL,
    OWNERID bigint NOT NULL,
    ADMINISTRATIVEID bigint NOT NULL,
    TECHNICALRESPONSIBLEID bigint NOT NULL,
    SK_PRODUCT int NOT NULL,
    SK_OWNER int NOT NULL,
    SK_ADMIN int NULL,
    SK_TECH int NULL,
    Quantity float(53) NOT NULL,
    LASTUPDATED datetime NOT NULL
     constraint DF_Domainregistration_facttable_LastUpdated default getdate(),
    UPDATEDBY varchar(50) NOT NULL
     constraint DF_Domainregistration_facttable_UpdateBy default suser_sname(),
    PRIMARY KEY (CREATEDAT_Date_key, DOMAIN_ID, OWNERID),
    CONSTRAINT FK_Domainregistration_facttableProduct FOREIGN KEY (SK_PRODUCT) REFERENCES Product(SK_PRODUCT),
    CONSTRAINT FK_Domainregistration_facttableOwner FOREIGN KEY (SK_OWNER) REFERENCES Owner(SK_OWNER),
    CONSTRAINT FK_Domainregistration_facttableAdmin FOREIGN KEY (SK_ADMIN) REFERENCES Admin(SK_ADMIN),
    CONSTRAINT FK_Domainregistration_facttableTech FOREIGN KEY (SK_TECH) REFERENCES Tech(SK_TECH),
    CONSTRAINT FK_Domainregistration_facttableCREATEDAT_date FOREIGN KEY (CREATEDAT_Date_key) REFERENCES CREATEDAT_Date(CREATEDAT_Date_key),
    CONSTRAINT FK_Domainregistration_facttableLASTREGISTRATIONDATE_Date FOREIGN KEY (EXPIRATIONDATE_Date_key) REFERENCES EXPIRATIONDATE_Date(EXPIRATIONDATE_Date_key)
    );
```

Figure 3 - Create Table Example

Finally, following the creation of the DDL and its execution, we get the fact table and the dimensional tables. The star schema can be visualized in figure 4 below.
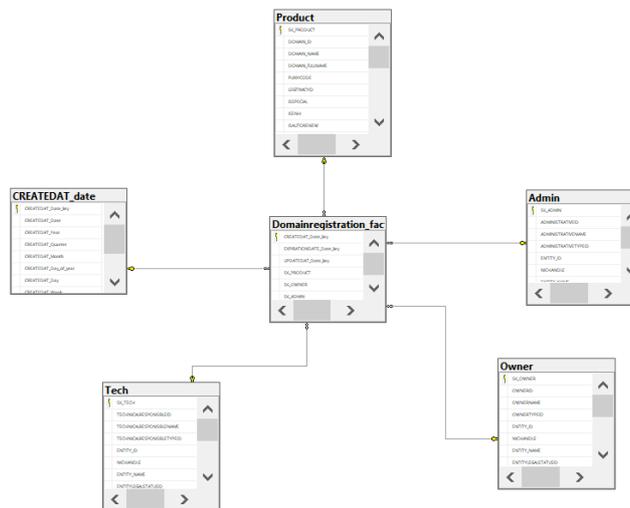
Figure 4 - Star Schema

4.1.3. Historical Data

After the definition of the architecture and the creation of the tables, it came a very important step which is defining and inserting the historical data and performing the first loading into all the tables in the DW.

For this, it was necessary to develop the DML (Data Manipulation Language) scripts of code for the data to be inserted into the corresponding tables. This was done through

15

an *INSERT INTO* command, followed by SELECT and FROM clauses that gathered the necessary columns from the multiple tables used. An example can be seen in figure 5.

```
INSERT INTO dbo.Product (
DOMAIN_ID, DOMAIN_NAME, DOMAIN_FULLNAME, PUNYCODE, LEGITIMICYID, ISSPECIAL, ISENH, ISAUTORENEW, HIERARCHYID, HIERARCHY_ID, HIERARCHY, TLDID, DESCRIPTION, ISDEFAULT,
DOMAINDOMAINPURPOSE_ID, DOMAINPURPOSEID, DOMAINPURPOSE_ID, LABEL, LEGITIMICY_cat, DOMAINLEGALSTATUSID, DOMAINSTATUSID, DOMAINSTATUSID_cat, DOMAINLEGALSTATUSID_cat
)
SELECT
d.DOMAIN_ID, d.DOMAIN_NAME, d.DOMAIN_FULLNAME, d.PUNYCODE, d.LEGITIMICYID, d.ISSPECIAL, d.ISENH, d.ISAUTORENEW, d.HIERARCHYID, h.HIERARCHY_ID, h.HIERARCHY,
h.TLDID, h.DESCRIPTION, h.ISDEFAULT, ddp.DOMAINDOMAINPURPOSE_ID, ddp.DOMAINPURPOSEID, dp.DOMAINPURPOSE_ID, dp.LABEL, l.LEGITIMICY_EN, dls.DOMAINLEGALSTATUS_ID,
ds.DOMAINSTATUS_ID, ds.LABEL as DOMAINSTATUSID_cat, dls.LABEL as DOMAINLEGALSTATUSID_cat
FROM dbo.Domain d
LEFT JOIN Hierarchy h
ON d.HIERARCHYID = h.HIERARCHY_ID
LEFT JOIN dbo.Domaindomainpurpose ddp
ON d.DOMAIN_ID = ddp.DOMAINID
LEFT JOIN dbo.Domainpurpose dp
ON ddp.DOMAINPURPOSEID = dp.DOMAINPURPOSE_ID
LEFT JOIN dbo.Legitimicy l
ON d.LEGITIMICYID = l.LEGITIMICY_ID
LEFT JOIN dbo.Domainstatus ds
ON d.DOMAINSTATUSID = ds.DOMAINSTATUS_ID
LEFT JOIN dbo.Domainlegalstatus dls
ON d.DOMAINLEGALSTATUSID = dls.DOMAINLEGALSTATUS ID
```

Figure 5 - Insert Into Example

After its execution, it is possible to populate all the necessary tables and finally the fact table *Domain Registration*.

4.1.4. Staging Area

Since all the tables in the DW area are populated with the first load of historical data, it was possible to start the implementation phase for populating the staging tables.

For the tables present in the staging area, it is important to mention that they will be loaded with two types of new rows. On the one hand, the first one is rows that correspond to new IDs that aren't already present in the Data Warehouse, and on the other hand the rows regarding IDs that are in the DW but have suffered a change in at least one of the columns, which also needs to be incorporated.

With this second being a more complex process, that involves SCD – Slowly Changing Dimensions, which will be further described in the next chapter.

For this phase of the project, when placing these lines into the staging tables, it was essential to identify the changes between the data currently in the DW and the new one from the source system, namely knowing what tables have new inserts to be performed and/or have updates to be made in its existing rows, so that they can, then, be included in the DW rightfully.

This is done through an additional table, which follows a CDC (Change Data Capture) method, meaning detecting and extracting the new or updated registers present in the source and only loading that into the DW. This identification is possible using a *BINARY_CHECKSUM* function, which computes a different value over each of the tables' rows.

The value obtained is compared between a certain row in the DW and its corresponding line in the data source. This way, if it is different, it means that some disparity was found on at least one of its columns, and this way there is an update to be performed. On the other hand, if the row is not present in the DW, there is a row that needs to be inserted.

Finally, to populate the staging tables an *INSERT INTO* command was used. For each of these tables, there is an additional column that indicates if a certain row is an 'INSERT' or 'UPDATE', populated according to the CDC table mentioned earlier, to help the following phase.

4.1.5. ETL Process

With the staging tables loaded, the team started the ETL process to populate the DW. Through this process, all necessary corrections and transformations are performed to improve the quality of the data. This process involves getting the information from the staging area into the DW area, which comprises the insert of new rows as well as the update part of the process, which will be explained shortly, as it focuses on the SCD.

It is important to mention that, for this phase of the project, even though an ETL tool was not used, the ETL workflow was created with clear steps.

The scripts used for this portion of the process transformed the data present in the staging area by cleaning it, making sure there are not unwanted null values, avoiding duplicate records, or filtering the rows we want to use, among other inconsistencies.

These also included the creation of additional columns, that we deemed relevant for improving the structure of the information.

An example of a transformation performed was converting a numerical value into a categorical value, which we did to improve the comprehension of status attributes.

Also, a calculated column was created which identifies if a certain domain is active or not, in order to improve efficiency. For this, we used a *CASE WHEN* statement, where if it were in a specific status that would be considered active a '1' would be inserted in that cell, otherwise it would be '0'.

Finally, after all the transformations the data is loaded into the corresponding tables in the DW. Making sure all tables are populated with the newly inserted rows based on the information present in the staging area, the remaining steps of the ETL process are performed. They include incorporating the changes in pre-existing lines in our Data Warehouse, that is, the ones that suffered alterations in the operational database, since the last load, and this way focus on the Slowly Changing Dimensions.

It is important to mention that we only focused on two types of SCD, that is types 1, and 2. Particularly, for type 1 the changes are incorporated by overwriting the existing data; whereas type 2 keeps the history, and this way it adds a line instead of replacing it.

The first step was to determine which columns should incorporate the updates, and for those, select the type of SCD, from the ones mentioned just now. For this, a *BINARY_CHECKSUM* function was used once again to establish a comparison, and this way see which needed the updates. That is, a checksum was calculated for each one of the rows and separated into types 1 and 2, which would respectively include the columns specified for each type of SCD. An example of this function can be seen in figure 6, below.

```
UPDATE dbo.Product set DIMENSION_CHECKSUM_SCD2 =
    BINARY_CHECKSUM(DOMAIN_ID, DOMAIN_NAME, DOMAIN_FULLNAME, LEGITIMICYID, ISENH, ISAUTORENEW, HIERARCHYID,
    DOMAINLEGALSTATUSID, DOMAINSTATUSID)

UPDATE dbo.Product set DIMENSION_CHECKSUM_SCD1 =
    BINARY_CHECKSUM(DOMAIN_ID, PUNYCODE, ISSPECIAL, TLDID, DESCRIPTION)
```

Figure 6 - Binary Checksum Example

If a different value was found for the same ID in the source – staging tables – and target – DW tables -, then an update needed to be performed.

After finding which rows needed to be updated as well as the types of SCD, it was time to incorporate those updates, in the designated tables, which would occur for all the lines that had the word 'UPDATE' in the column previously mentioned and present in the staging tables.

The SCD type 1 overwrites the current row, while SCD type 2 inserts a new line with the updates. For these IDs where it was necessary to keep its history, it was crucial to have an additional column, which we named 'CURRENTRECORD', which enables us to distinguish the most recent row – using a 'Y' – and the old one(s) – with an 'N'.

It is important to mention that for this process to be possible, we needed to create Surrogate Keys – SK -, as Primaries Keys for each of our tables, instead of having it as the Domain ID or Entity ID, for example, to allow us to have multiple rows for each of those IDs, given the updates that could be needed to incorporate.

4.1.6. Quality Assurance

The quality assurance phase is performed in the quality/testing environment which provides more regular loads of data. This environment allows the performance of tests to validate what was previously developed, and to implement corrections or improvements found necessary to ensure that the code is correct and there are no errors.

As to keep track of the tests and results, it was necessary to have some documentation, which in this case were two Excel files.

The first one was used to maintain a record of the results of all the tests performed in each run, for each table of the DW, as well as its order of usage. Some examples of tests that were executed were duplicated lines, number of inserted rows, updates performed on preexisting rows, and ensuring the right information was passed from the source.

The second one is a summary of the test results as well as a brief explanation of each test executed, including the type of test and when it should be implemented, among others.

4.2. Data Management and Governance Model

With the objective of not only understanding the data but also ensuring its accurate creation, use, control, and quality, it is important to develop a data management and governance model, creating and organizing the documentation and metadata necessary for the purpose of this project.

Given the importance of documentation for any IS project, it was crucial to start by developing a data dictionary, meaning a metadata repository that collects names, definitions, and attributes used in the database.

This is done to understand the different variables we were handling and use this knowledge to identify what should or shouldn't be incorporated. The existing data dictionary was incomplete, so it required some work.

It consisted of creating an Excel sheet, listing every table from the source system we were using and all the columns that compose it, followed by the explanation of each of those attributes, data type, and the possibility of null values.

Another piece of documentation created was the Source to Target Mapping, in Excel, which is a blueprint for the ETL process.

The mentioned document consists of registering the information regarding the source system tables and the attributes used in the project and mapping them to the fields of the DW, considered the target. This should be done for each business event included in the development of the Data Warehouse.

Some of the essential pieces of information to keep a record of, both from the source and the target tables, are names and data types for all the attributes, if null values are allowed or not, and any transformations performed, as well as SCD attributes and logic.

Furthermore, an Excel document was created with all the business questions the team found relevant to be answered through the data visualization model in Power BI, including the indicators and calculations used to create the respective measures.

Lastly, as files change over time, given the continuous developments that occur throughout a project, it was fundamental to have a version management system, which includes not only file versions, but also logging changes, and dates.

4.3. Data Visualization Model

The data visualization phase of the project comprises creating dashboards through a visualization tool, in this case, Power BI, so that it is possible to answer specific questions found relevant in a more appealing and clear way using the data we have on the DW.

The first step was to create a list of questions we found important to be answered based on the data we have, as well as determine which departments of the company that information would be the most meaningful to.

This division was relevant, given the several departments present in the organization, including marketing, finance, customer service, and IT. While some information is significant across all departments, other pieces of information cater certain needs of a specific one.

After the questions were selected and categorized, the team decided to start with a few more general dashboards and afterward focus on the marketing department, as an example.

For the implementation phase, a link was established between Power BI and the Data Warehouse, so that it was possible to access the star schema created and subsequently start developing the dashboards. As the connection is made, the data is automatically updated and this way so are the reports, by clicking the option *Refresh*.

With the data imported to Power BI, it was possible to start creating some measures with DAX (Data Analysis Expressions) functions.

Starting with the function *CALCULATETABLE* which allowed us to filter the *Product* dimension to only obtain the domains that are currently active and after that, a *COUNTROWS* function was used to count those domains. The second function was also used to count the total amount of registered domains, which then allowed us to calculate an active domain rate, through a *DIVIDE* function.

Moreover, a conditional column was created to generate intervals for the number of years that domains are registered for.

Finally, a quick measure was used to generate the evolution of the registration of domains through the years, that is, the running total. After calculating the necessary measures, and as we had the required indicators, the creation of the visuals was started.

Additionally, we included slicers in all dashboards, thereby enabling the data to be filtered. These slicers included filters for both the date and/or the country, allowing users to adjust the visuals and better comprehend them.

This way, the first two dashboards focused on the analysis of registered and active domains, so that it is possible to not only understand its evolution throughout the years, months, and quarters but also the countries where they are present the most. Furthermore, it is possible to realize the proportion of active domains in relation to the total number of registered domains.

The finished dashboards can be visualized in figures 7 and 8, for the registered and active domains, respectively.
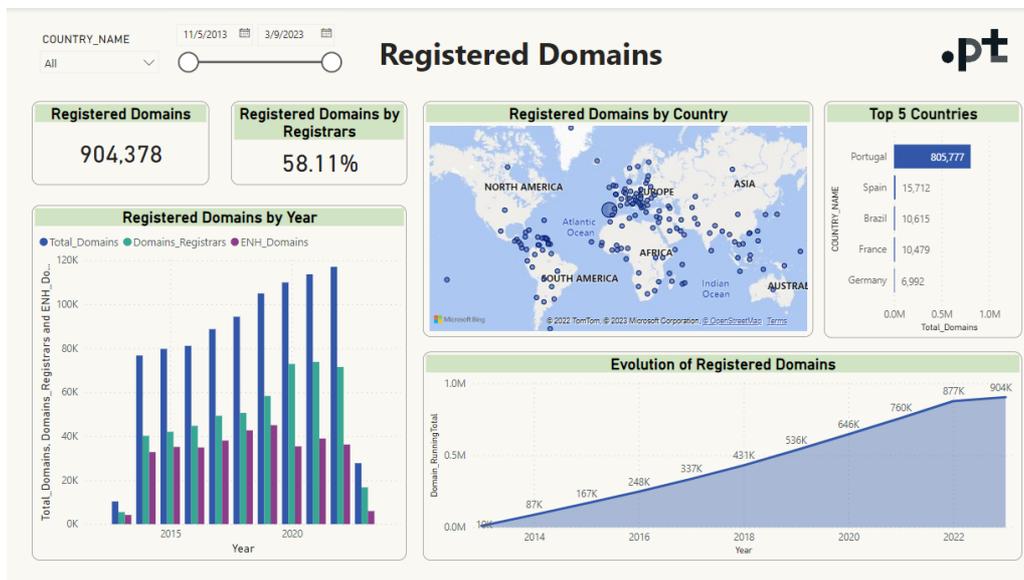

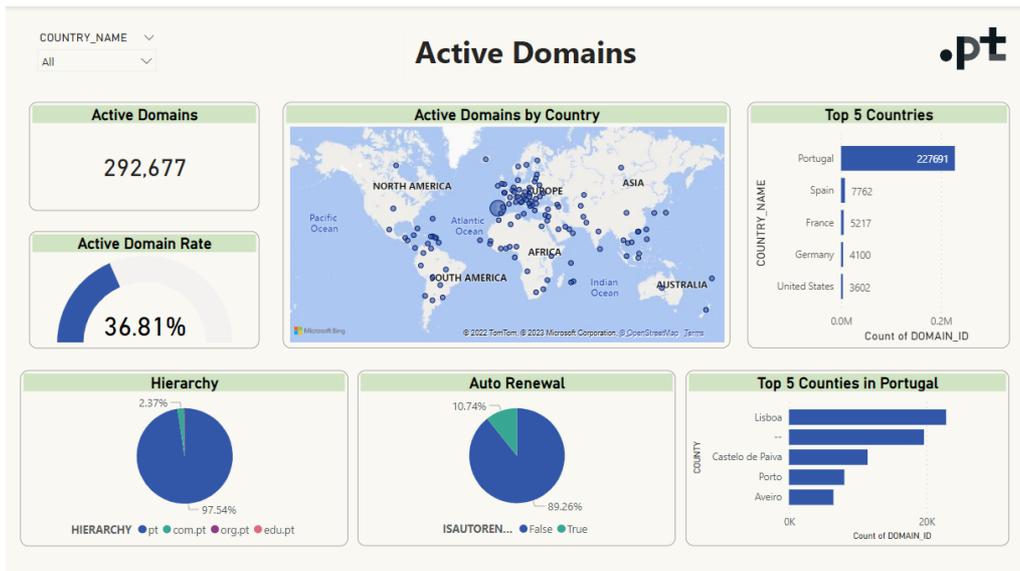Figure 7 - Registered Domains Dashboard

Figure 8 - Active Domains Dashboard

Next, the two dashboards that address questions more specifically about the marketing department were created. For the first one, the team decided to focus on the clients, that is, understanding the entities that register their domains as well as their purpose. Regarding the second dashboard, the emphasis was on the number of years for which clients registered their domains.

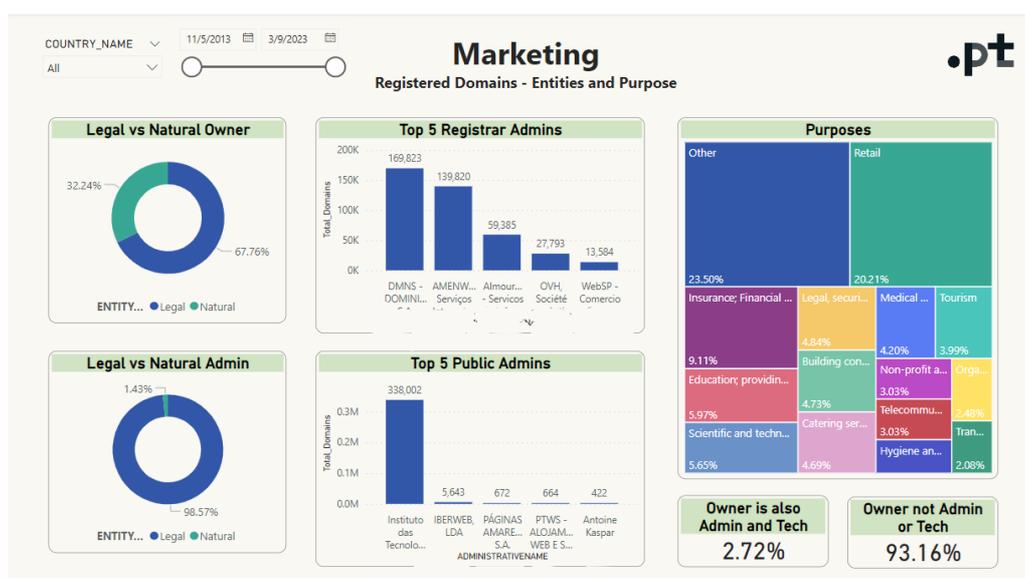The results of both dashboards can be seen in figures 9 and 10.

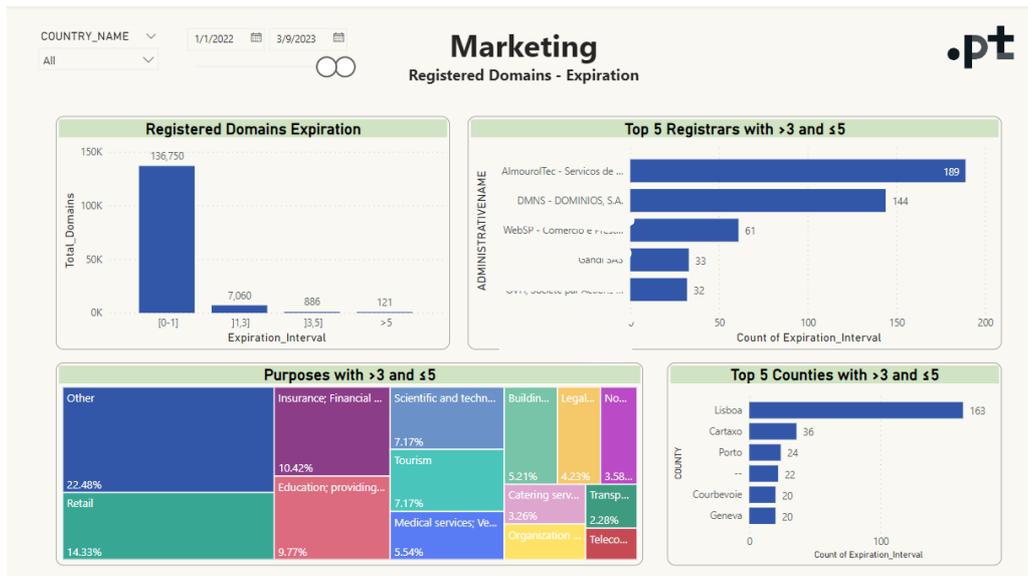Figure 9 - Entities and Purpose Dashboard

Figure 10 - Expiration Dashboard

As the Data Warehouse developed during the internship was created as a prototype, that is, for only the *Domain Registration* event, these were just a few examples of dashboards, using the star schema created. On the other hand, the same should be performed, not only, for the other departments of the company, but also, with all the business events created, so that it is possible to visualize all the relevant information and this way, help decision-making.

5. DISCUSSION

This internship allowed the consolidation of the insights obtained during the Master's Degree and its application in a business context, and thus increasing knowledge of different tools such as SQL and Power BI. The objectives defined in the beginning were accomplished, making the internship a very positive experience for the intern.

On the one hand, a Data Warehouse was created, which wasn't present in the company before, providing improved data quality. This was developed for one of the business events of .PT as a prototype, which should then be replicated to the remainder, giving the company a great advantage in understanding the current large quantities of information.

On the other hand, throughout the project, a lot of knowledge on different topics was gained, which allowed for the comprehension that some matters could have been dealt with differently, such as the inclusion of certain columns from the source system that weren't present in the beginning, change in data types, and relationships between tables, among others.

For the ETL process part of the project, although the workflow was well established, a limitation might have been the selection of the ETL tool, which was chosen more towards the end of the development of the DW for the business event in question and, therefore, was not used during the internship. This could have eased some of the steps in terms of time and documentation.

Furthermore, a data management and governance model was established to provide clear and structured documentation on the different sources of information available at .PT.

For that, a data dictionary was developed to ensure a better understanding of the data, containing all the tables from the source system that were found important to include along with a description of all attributes. In addition, the source-to-target mapping was created, incorporating a detailed description of all attributes from both the source and target tables used. This way, a data management and governance model was found to be very important to help fast-track decision-making and understanding.

A limitation that was found in the beginning was understanding what each attribute meant from all the tables used. This way, it was essential to develop a data dictionary, as

what had already been created was very incomplete, and needed some improvement. It was important to know this in order to define what attributes should or shouldn't be included. To do this, it was necessary to ask a few people within the company from different departments to get that information.

Following the development of the data management and governance model and ETL process, the visualization phase began, using Power BI. One limitation of this phase might have been that, as we only developed one business event, there were some restrictions in terms of possible analyses.

6. Conclusions

The internship undertaken at .PT was a great experience and an amazing opportunity to learn. This was the intern's first professional encounter, allowing for an understanding of the business world and the acquisition of knowledge on various topics.

Being a company focused on the registration and management of domains, it was a unique chance to not only get insights about this business but also understand its importance, especially in the most recent years, given the amount of data which has been increasing over time.

It had the objectives of creating a Data Warehouse and comprehending its ETL process, as well as developing a data management and governance model, which was accomplished, allowing the improvement of the data quality as well as providing organized storage of data, that will help decision-making across the different departments of the company.

Furthermore, the visualization of the cleansed and organized data, provided a clear view of the work developed throughout the majority of the course of the project, which wasn't yet possible before, as the company didn't have a Data Warehouse to start with.

Additionally, this project enabled the intern to merge her academic knowledge with the practical skills developed during the project, which included SQL and Power BI, and the ability to turn information into valuable insights.

The first phase of the project focused on establishing the project methodology, which followed a Kimball approach. The second one was a great choice as it allowed for the refinement of business requirements throughout the development process.

Next, was the construction of the logical and physical models for one of the core business events of the company, *Domain Registration*. However, the remaining events, which were not covered during the internship, will need to be developed in the future. On the other hand, the ETL process was the most challenging part, given that it includes the use of SCD for updating the different tables across the DW, using different commands.

Subsequently, the task to build the Data Warehouse of .PT was effectively completed, serving as a prototype for the remaining business events to be developed in the DW, and this way allow fast and easy access to data for better decision-making.

The second phase focused on the development of the Data Management and Governance Model of .PT, which included the enhancement of the already existing data dictionary and the creation of a source-to-target mapping, was also well-concluded.

Finally, the third phase involved the development of a data visualization model, through a Business Intelligence application in Power BI. This model includes different views of the data present in the Data Warehouse, improving its overall understanding, and giving it a more visually clear and appealing representation.

Lastly, there is still work to be done regarding the construction of the Data Warehouse, given that as mentioned previously, there are many others core business events across the various departments of .PT, which will allow for better decision-making, and this way should then be complemented with data visualizations.

REFERENCES

Alhassan, I., Sammon, D. & Daly, M., 2016. Data governance activities: an analysis of the literature. *Journal of Decision systems*.

Awiti, J., Vaisman, A., Zimányi, E., 2020. Design and implementation of ETL processes using BPMN and relational algebra. *Data & Knowledge Engineering*.

Balaceanu, D., 2007. Components of a Business Intelligence software solution. *Informatica Economica*.

Breslin, M., 2004. Data Warehousing Battle of the Giants: Comparing the Basics of the Kimball and Inmon Models. *Business Intelligence Journal*.

Cheong, L., Chang, V., 2007. The Need for Data Governance: A Case Study. *Association for Information Systems.*

Chinyere, J., 2022. Data Warehouse: A Tool For Organizational Efficiency. *Nigerian Journal Of Management Sciences*.

European Commission, 2022. *Data*. [Online]. Available at: https://digital-strategy.ec.europa.eu/en/policies/data. [Accessed 07-01-2023].

Farrokhi, V., Pokorádi, L., 2012. The necessities for building a model to evaluate Business Intelligence projects- Literature Review. *International Journal of Computer Science & Engineering Survey*.

Forbes, 2022. *Why So Many Data Warehouse Projects Fail.* [Online]. Available at: https://www.forbes.com/sites/forbestechcouncil/2022/06/07/why-so-many-data-warehouse-projects-fail/?sh=8778a4012bc8. [Accessed 07-12-2022].

Gartner, 2019. Data Governance. *Gartner Glossary*. [Online]. Available at: https://www.gartner.com/en/information-technology/glossary/data-warehouse. [Accessed 08-12-2022].

Gartner, 2019. Data Warehouse. *Gartner Glossary*. [Online]. Available at: https://www.gartner.com/en/information-technology/glossary/data-warehouse. [Accessed 08-12-2022].

Harvard Business Review, 2022. *Why Becoming a Data-Driven Organization Is So Hard*. [Online]. Available at: https://hbr.org/2022/02/why-becoming-a-data-driven-organization-is-so-hard. [Accessed 08-12-2022].

IBM, 2020. *Data Warehouse*. [Online]. Available at: https://www.ibm.com/cloud/learn/data-warehouse. [Accessed 07-12-2022].

IBM, 2021. *Logical data models*. [Online]. Available at: https://www.ibm.com/docs/en/data-studio/4.1.1?topic=SS62YD_4.1.1/com.ibm.datatools.logical.ui.doc/topics/clogmod.html. [Accessed 07-12-2022].

IBM. *What is Business Intelligence?*. [Online]. Available at: https://www.ibm.com/topics/business-intelligence. [Accessed 08-12-2022].

Inmon, W., 2005. *Building the Data Warehouse*. Indiana, Indianapolis: Wiley.

Kopanakis, I., Vassakis, K., Mastorakis, G., 2016. *Big Data in Data-driven Innovation: The Impact in Enterprises' Performance.*

Kimball, R. & Ross M., 2013. *The Data Warehouse Toolkit: The Definitive Guide to Dimensional Modeling*. Indiana, Indianapolis: Wiley.

Matillion, 2019. *STTM (Source to Target Mapping) document Definition.* [Online]. Available at: https://www.matillion.com/data-glossary/sttm-source-to-target-mapping-document/. [Accessed 23-12-2022].

Negash, S., Gray, P., 2008. Business Intelligence. Handbook on Decision Support Systems 2: Variations, pp 175–193. *International Handbooks Information System.* Springer, Berlin, Heidelberg.

Niu, Y., Ying, N., Yang, J., Bao, M. & Sivaparthipan, C., 2021. Organizational business intelligence and decision-making using big data analytics. *Information Processing & Management.*

Ong, I., Siew, P., Wong, S., 2011. A Five-Layered Business Intelligence Architecture. *IBIMA Publishing.*

Oracle, 2017. *What is Business Intelligence (BI)?.* [Online]. Available at: https://www.oracle.com/pt/what-is-business-intelligence/. [Accessed 08-12-2022].

SAS Institute, 2021. *Data quality management: What you need to know.* [Online]. Available at: https://www.sas.com/en_us/insights/articles/data-management/data-quality-management-what-you-need-to-know.html. [Accessed 08-12-2022].

Tableau, 2022. *Business intelligence: A complete overview.* [Online]. Available at: https://www.tableau.com/learn/articles/business-intelligence#what-is. [Accessed 08-12-2022].

The Council on Quality and Leadership, 2019. *12 Reasons Why Data Is Important.*

Towards Data Science, 2021. *Why Organizations Need To Be Data-Driven.* [Online]. Available at: https://towardsdatascience.com/why-organizations-need-to-be-data-driven-98ade3ca53a. [Accessed 08-12-2022].

Vasques, D., 2019. *O Statistical Data Warehouse do INE : uma observação participativa*. [Online]. Available at: https://www.repository.utl.pt/bitstream/10400.5/19176/1/DM-DLSV-2019.pdf. [Accessed 07-12-2022].

Vassakis, K., Petrakis, E., Kopanakis, I., 2018. *Big Data Analytics: Applications, Prospects and Challenges*.

Wang, J., Omar, A., Alotaibi, F., Daradkeh, Y., Althubiti, S., 2022. Business intelligence ability to enhance organizational performance and performance evaluation capabilities by improving data mining systems for competitive advantage. *Information Processing & Management*.

Wende, K., 2007. A Model for Data Governance – Organising Accountabilities for Data Quality Management. *Association for Information Systems*.

Yin, J., Li, C., 2022. Data governance and green technological innovation performance: A curvilinear relationship. *Journal of Cleaner Production*.

Zhang, Q., Xinbo, S., Zhang, M., 2022. Data Matters: A Strategic Action Framework for Data Governance. *Information & Management*.

.PT, 2022. *About domains*. [Online]. Available at: https://www.pt.pt/en/domain/about-domains/. [Accessed 08-12-2022].

.PT, 2022. *About Registrars*. [Online]. Available at: https://www.pt.pt/en/registrar/about-registrars/. [Accessed 08-12-2022].

.PT, 2022. *About us*. [Online]. Available at: https://www.pt.pt/en/pt/about-us/. [Accessed 08-12-2022].

.PT, 2022. *Digital Inclusion*. [Online]. Available at: https://www.pt.pt/en/pt/digital-inclusion/. [Accessed 08-12-2022].

.PT, 2022. *.PT WHOIS Policy*. [Online]. Available at: https://www.pt.pt/en/data-protection/pt-whois-policy/. [Accessed 08-12-2022].