Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

INTERNSHIP REPORT

BUSINESS INTELLIGENCE PROJECT IMPLEMENTATION: FRAMEWORK

JOANA MARIA LOPES GUEIFÃO

MARCH - 2022

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK
INTERNSHIP REPORT

BUSINESS INTELLIGENCE PROJECT IMPLEMENTATION: FRAMEWORK

JOANA MARIA LOPES GUEIFÃO

SUPERVISION:

PROF. WINNIE PICOTO

FLÁVIO ROMÃO

MARCH - 2022

GLOSSARY

BI – Business Intelligence.

BPO – Business Process Outsourcing.

DW – Data Warehouse.

ETL – Extract, Transform and Load.

IQS - Internal Quality Score.

IT – Information Technology.

IVR - Interactive Voice Response.

KPI – Key Performance Indicator.

MFW – Master's Final Work.

ODS – Operational Data Storage.

OLE DB - Object Linking and Embedding Database.

SSIS – SQL Server Integration Services.

ABSTRACT

In recent years, businesses have experienced a substantial increase in data availability from which they recognize a source of value. As a result, business intelligence (BI) projects are in great demand as they offer a way to transform raw data into information that may assist decision-making and hence provide corporate value. This study provides a practical view of the implementation of a BI project into a business.

Link Redglue is a company that helps clients get the most out of their data through an expertise approach. The internship described in this master's final work integrates a BI project developed by Link Redglue, the purpose of which was to implement an existing BI solution for two new clients. The data that was to be analysed was related to these clients contact-centre processes. This report details, along with a theoretical background, the BI solution's implementation, from the contact centre's source data to the Extract, Transform and Load (ETL) processes that brought data to the BI environment and were responsible for populating the already existing data warehouse, and finally, to the reporting layer which was responsible for representing the most important indicators related to the clients' contact-centre processes using Microsoft Excel.

The client's BI solution implementation resulted in a data warehouse that provides a single view of all available data as well as reporting capabilities that can be used on a daily basis to assist decision-making regarding contact-centre operations of these clients.

KEYWORDS: Business Intelligence, Data, Reporting, Business value

RESUMO

Nos últimos anos, um aumento substancial na disponibilidade de dados nas empresas tem vindo a ser notado, o qual é reconhecido por estas como uma fonte de valor. Consequentemente, os projetos de *business intelligence* (BI) têm vindo a ser cada vez mais procurados pelas empresas, dado que estes oferecem uma forma de transformar dados em informações relevantes que podem auxiliar na tomada de decisões e, consequentemente, gerar valor de negócio. O presente relatório apresenta uma visão prática de como foi implementado um projeto de BI num negócio.

A empresa Link Redglue tem como objetivo ajudar os seus clientes a tirar o máximo valor dos seus dados. O estágio descrito neste trabalho final de mestrado foi parte de um projeto da Link Redglue que tinha como objetivo implementar uma solução de BI já existente para dois novos clientes. Os dados utilizados para esta implementação consistiam nos dados de processos de *contact-centre* destes dois clientes. Este relatório descreve, acompanhado por um enquadramento teórico, a implementação da solução de BI desde os dados operacionais de contact-centre do cliente, passando pelos processos de extração e transformação dos dados responsáveis por mover os dados até à *data warehouse* e finalmente ao processo de reporte dos resultados ao cliente utilizando Microsoft Excel para representar os indicadores mais relevantes nos seus processos de *contact-centre*.

A implementação da solução de BI para o cliente resultou numa *data warehouse* que proporciona uma visão unificada de todos os dados disponíveis,

bem como a disponibilização de relatórios que fornecem informação relevante diariamente para auxiliar na tomada de decisões.

PALAVRAS-CHAVE: *Business Intelligence*, Dados, Relatórios, Valor de negócio

TABLE OF CONTENTS

TABLE OF FIGURES

ACKNOWLEDGMENTS

First, I wish to thank Professor Winnie Picoto for her encouragement and guidance. Moreover, I wish to thank Flávio Romão, COO of Link Redglue, for presenting me the opportunity to participate in this project and for providing guidance and support.

I am also grateful to my master's colleague Ana for numerous discussions and support.

Finally, I am also thankful to my family for their support while I pursued this project.

1. INTRODUCTION

1.1.Relevance

The internship described in this master's final work document started as a collaboration between Lisbon School of Economics and Management and Link Regdlue, as part of the master's degree in Data Analytics for Business.

The twenty-first century has ushered in a new age of data science and analytics. The huge growth of data has contributed to a data-driven era, in which data analytics is used in every industry (agricultural, health, energy and infrastructure, economics and insurance, sports, food, and transportation) and in every world economy (Kopanakis, et al., 2018).

The volume of data produced worldwide is rapidly increasing, rising from 33 zettabytes in 2018 to an expected 175 zettabytes in 2025, corresponding to a 530 percent rise in global data volume, therefore, by 2025, this growth in data availability will have increased the value of the European Union's data economy to 829 billion euros, up from 301 billion euros in 2018 (European Commission, 2020).

In this data-intensive landscape, the availability of additional data streams is an appealing differentiator for businesses (Hermann, et al., 2016). Therefore, enterprises must embrace business intelligence efforts to extract value from their data. Because data availability does not imply value, business intelligence is an approach traditionally used by firms to derive value from data (Bordeleau, et al., 2020). Business intelligence is used for sifting through large amounts of data, discovering significant facts, and turning that data into actionable knowledge to assist decision-making (Ranjan, 2009).

By recognizing data as an asset, businesses have turned to business intelligence solutions as a way of achieving a more informed decision-making process that leads to improved organizational outcomes based on the available data (García &

Pinzón, 2017). Moreover, as sources, and volume of data grow so too has the data analytics techniques to extract knowledge and ultimately produce insights (Papadopoulos, et al., 2017). Business intelligence tools enable querying, statistical analysis, reporting, data visualization, and dashboarding so that raw data can be transformed into meaningful business knowledge (DAMA International, 2017).

1.2. Company description

1.2.1. Link Group

Link group was founded in 2000 as a spin-off of INESC, Lisbon's foremost IT technology institute (Link, 2019). INESC - Institute for Systems Engineering and Computers- is a non-profit association founded in 1980 dedicated to education, science, research activity and technological consulting (INESC, 2020). INESC develops science and technology that competes and leads in national and international markets as well as high-quality human resources that are driven to contribute to Portugal's national capacities and modernization, moreover, due to the strong links to universities and businesses INESC contributes to better adapting the system of scientific and technological education to the economic and social needs (INESC, 2020).

Link's group main goal is to add value to its customers by providing technology innovation in the fields of information and communication technologies. Link provides solutions for several industry areas such as Retail, Healthcare, Financial Services and Logistics and Transportation and most of the large national companies are among Link's main customers, from sectors such as Telecommunications, Bank and Insurance, Logistics and Distribution (Link, 2019).

1.2.2. Link Redglue

Responding to the high demand on implementation of business intelligence projects on organizations to differentiate their businesses, Link Redglue was founded in 2017 to help its clients get the most out of their data by leveraging its value through an expertise approach.

Link Redglue's areas of expertise consist of artificial intelligence comprising data science, predictive analytics and deep learning; advanced analytics comprising data visualization and BI and finally data foundations comprising data engineering, architecture, governance and security (Link Redglue, 2017). The project developed during the internship took place in the area of advanced analytics which develops BI projects that aim to provide decision-makers the ability to navigate data from all its perspectives in a manner that is understandable to business.

1.3.Objectives

This internship took place between October 2021 and February 2022, in which I integrated a data-specialized team from Link Reglue that developed a business intelligence project for a customer hereby mentioned in this document by company A, for confidentiality purposes. Company A provides business process outsourcing (BPO) for contact-centre services to clients in a variety of industries, including two insurance companies that were part of this project and referred to as the final clients. The project emerged from these insurance business clients need to generate value from their available data, specifically for their contact-centre operational data. The main goal of this internship, as part of a business intelligence project, was to produce insights to improve decision-making regarding the contact-centre operations of these clients by analysing their contact-centre operational data.

As in a previous project that addressed another BPO client a solution was developed by company A, the goal of this project was to replicate the high-level approach for the new clients. Company A's solution consisted of the design and

creation of a data warehouse and the development of reporting in Microsoft Excel by creating excel files that contained the definition of the key performance indicators (KPIs) that were relevant for business in the context of the contact-centre environment. As a result, for these two new clients, the goal was to apply the previously designed solution yet adapting it to their own rules. The project was divided into two phases, each including the implementation of a single client.

When implementing the solution for a client the considered steps were the following:

First, populating the data warehouse for the client we were working on – this included performing Extract, Transform and Load (ETL) processes in order to migrate data from the operational data storage to a staging area and finally to the data warehouse already created. To develop the above mentioned ETL processes, we used SQL Server and SQL Server Integration Services (SSIS) for transforming and cleaning data as well as migrating from its different stages.

Having the data warehouse populated for this client, we moved on to the next stage of reporting which was performed using Microsoft Excel. By connecting Microsoft Excel to the data warehouse, it was possible to gather the relevant data to report the defined KPIs to the business.

After successfully performing these tasks, company A would be responsible to approve them and communicate to the final client the results. In the next sections, this report aims to describe the pursue of the ultimate goal to generate value from data for this specific real case. For that, this report covers the use of the above-mentioned BI tools as well as the data migration process and finally how the reporting was achieved.

2. LITERATURE REVIEW

2.1. Business Intelligence

As more businesses embrace digital transformation, the amount of data acquired by them has increased significantly, therefore BI efforts are required to extract value from their data (Bordeleau, et al., 2020). Business intelligence has been defined as a combination of tools, technologies, and solutions that enable end users to extract relevant business information from vast amounts of data (Zeng, et al., 2006). These set of techniques include ETL processes, data warehousing, database querying and reporting, on-line analytical processing data analysis, data mining, and visualization (Gangadharan & Swami, 2004). More recent literature is defining BI rather than a set of technologies, as an integrated solution for businesses, in which the business requirement is unquestionably the driving force behind technological progress (Ranjan, 2009). A broader definition for business intelligence is that it consists of an automated process for gathering raw data from a variety of sources and analysing it in a manner so that models and insights can be created from it to improve business processes (Vo, et al., 2018).

Moreover, BI systems are specialized tools for data analysis, querying, and reporting that enable organizational decision-making and potentially improve the performance of a variety of business operations, and for their deployment and effective usage, BI systems require specialized IT infrastructure such as data warehouses, data marts, and ETL tools (Elbashir, et al., 2008).

According to the 2015 MIT Sloan Management Review survey of 2719 managers in organizations around the world, the greatest barrier to create business value from data is translating analytics into business actions, that is, performing efficient data-driven decision making (Ransbotham, et al., 2015). In this sense, BI intends to close the gap between data availability and actual business actions. Therefore, the term BI suggests a thorough understanding of all the aspects that influence business operations. To make effective and high-quality business

decisions, companies must have a thorough understanding of aspects such as consumers, competitors, business partners, economic environment, and internal processes (Kiron, 2017).

Many companies have been recognizing data as an important asset and hence the importance of BI to generate value from it. Although digital transformation has boosted the appearance of BI and data driven cultures, Walmart is an example of a company that has pioneered the usage of data to create value. Since the 1970's, it was one of the first organizations to use data warehouses to manage its inventory and therefore the first company to reach one billion in sales in its first 17 years (Patil & Mason, 2015). In more recent years, with the amount of data that is created, captured, processed and stored growing exponentially, there are more opportunities to broaden the scope of BI and achieve more benefits (Vo, et al., 2018). Consumer internet companies such as Google, Amazon, Apple, Facebook, Netflix, and LinkedIn have become some of the most well-known data-driven enterprises in recent years, owing to their heavy reliance on data assets to the point where data is at the heart of their operations and has altered the nature of competition in their respective industries (Mahanti, 2021). It is reasonable to conclude that BI principles emerged with pioneers such as Walmart, however the increase in data volume and access in recent years has shown organizations how data can be a valuable asset. All these companies have realized the importance of their data therefore BI is required for them to properly manage and generate value from it.

Companies that utilize BI can reap a variety of benefits. BI enables businesses to make decisions based on timely and accurate information which in turn improves its performance in different aspects, moreover, it can help companies respond quickly to changes in financial conditions, customer preferences, and supply chain operations, improving communication between departments, and allowing companies to respond quickly to changes in financial conditions, customer preferences, and supply chain operations (Ranjan, 2009). Furthermore,

BI can help organizations improve information quality in a variety of ways by enabling fast access to information, quick querying and analysis, a higher level of interaction and improved data consistency as a result of data integration operations, and other data management activities including data cleansing, metadata management and data integration (Popovič, et al., 2012).

2.1.1. BI components

BI components include the source systems, ETL processes, ODS, the data warehouse, and reporting (Balaceanu, 2007).

2.1.1.1. Source Systems

In the BI environment, data sources may include operational databases, historical data, external data, and information from an existing data warehouse system and can contain both structured and unstructured data, furthermore, relational databases or any other data structure that supports line of business applications can be used as data source (Ranjan, 2009).

2.1.1.2. ETL processes

The main goal of ETL processes is to transform operational data to subject data in the data warehouse (Zeng, et al., 2006).

First, data is extracted and brought to the BI environment from the data sources to the staging area which is an intermediate storage area located between the data sources and the data warehouse that simplifies data cleansing and consolidation coming from different sources (Kimball & Caserta, 2014).

Then, the transform step comprises applying a set of rules to transform the data in the staging area to the targeted destination and finally, as the data is transformed and cleaned it is loaded into the data warehouse (Kadadi, et al., 2014).

2.1.1.3. ODS

An ODS is a replica of one or more source systems, which is built to support operational reporting, it respects the data model of the source system and adds the possibility to store historical versions of the data (Balaceanu, 2007).

2.1.1.4. Data Warehouse

The data warehouse is the input to the BI's analytic environment, and most businesses use it as their primary source for BI data, it is defined as a data collection that is subject-oriented, integrated, time-variant, and non-volatile that is used to support the decision-making process (Negash & Gray, 2008). According to Negash & Gray (2008), the data warehouse is defined as follows:

- Subject-Oriented: Data is organized around key subjects of the business, these may include customers, products, or time.
- Integrated: Data is defined using consistent definitions, formats and naming conventions to create a single version of the truth.
- Time-variant: Data contains a time dimension to study its history and current status.
- Non-volatile: Data cannot be changed over time.

The data warehouse is a significant component of BI. By handling the multiple corporate records for integration, cleansing, aggregation, and query operations, it assists in the propagation of data (Ranjan, 2009).

In the data warehouse, data is understood in terms of tables and columns (Mahanti, 2021). The data model of the data warehouse provides a visual representation that captures the nature and relationships among data (Hoffer, et al., 2016). The dimensional model is a data structure technique developed to optimize the performance and ease of use of the data warehouse which comprises fact and dimensional tables (Mahanti, 2021).

The fact table rows correspond to an actual measurement and are numeric, such as amounts, quantities or counts while dimensional tables hold descriptive data about important objects of the business (DAMA International, 2017).

The data in the warehouse can be designed according to different models such as star or snowflake schema, which according to Hoffer et al (2016) are defined as follows:

- The star schema consists of one fact table and one or more-dimensional tables in the format of a star.
- The snowflake schema is a variation of the star schema in which dimensional tables are normalized into several related tables.

2.1.1.5. Reporting

Reporting tools are highly useful in the BI environment since they provide a quick way to access data in a way that assists decision-making for management purposes (Ong, et al., 2011).

The reporting component of a BI solution is the tip of the iceberg as it is the only visible part of the solution to most end users, therefore, failing to pay attention to this aspect of the BI project may result in the project's failure, even if the other BI components are done effectively (Balaceanu, 2007). As a result, collaboration between business and IT is essential throughout the report development process in order to deliver adequate reporting that end-users can clearly comprehend.

Data visualization tools such as dashboards and score cards are used to provide an overall view of business performance (Ong, et al., 2011). A dashboard, which is the most relevant for this internship, enables end users to assess current and past status of the business using visuals such as charts and tables trough a web browser interface to improve the communication of BI results (Negash & Gray, 2008). According to Gowthami & Kumar (2017), a dashboard should combine multiple data visualizations in a single interface, in a manner that it is clear to visualize and monitor its KPIs.

KPIs are frequently used in BI to assess the current state of business and prescribe a course of action (Ranjan, 2009). KPIs with target values that must be met within a given time frame are used to specify performance standards for business processes. Rio-Ortega et al (2009) recommends KPIs that satisfy the SMART criteria which is an abbreviation for five characteristics: specific, measurable, achievable, relevant and time bounded which he defines as follows:

- Specific: the KPI is clear in what it describes or measures.
- Measurable: it is possible to measure the KPI value and to compare it with a target.
- Achievable: The KPI targeted value is realistically possible to meet.
- Relevant: The KPI is aligned with business needs and affects overall performance.
- Time bounded: The KPI has an associated time-period when it is measured.

2.2.Data Management

Kiron (2017) states that the success of a BI project depends foundationally on well-governed data, partnerships, and long-term commitment from both leaders and employees and that the organization can only move forward with the creation of value from its data after this foundation is in place.

Data management consists of the processes, architectural techniques, and tools to manage the full data lifecycle needs of an organization in order to meet the data consumption requirements of business (Gartner, 2019). The purpose of data governance is to ensure that data is managed properly as it is the oversight of the data management effectiveness and enterprises are expected to manage their data activities in accordance with the governance direction and structure that were primarily established (Gorball, 2016). Data governance addresses issues such as who has ownership of the data, who has access to what data, how and with whom they may share it, what security measures are in place and if data compliance policies are being achieved (Kiron, 2017). While the ultimate goal of data

management is to guarantee that an organization derives value from its data, data governance focuses on how data decisions are made and how people and processes are expected to act in regard to data, data governance is therefore a discipline of data management (DAMA International, 2017).

DAMA International (2017) has identified 11 areas of data management activity that it refers to as knowledge areas, the most relevant for the present context of this internship being data quality, metadata, data integration and data modelling. These activities cross paths with one another and with other organizational activities as data moves horizontally inside organizations.

2.2.1. Data Quality

The importance of high-quality data in data management cannot be underestimated. According to Hoffer et al (2016), high quality data is characterized by being accurate, complete and consistent. Data accuracy refers to whether the values stored for an object are correct, that is, whether they are the right value and are expressed in an unambiguous manner (Olson, 2003). Data completeness refers to the degree to which all data required for current and future business activities is available in the company's data repository (Kwon, et al., 2014). Completeness is an important requirement for data quality because if data is missing, it may be unusable (Hoffer, et al., 2016). Data consistency refers to ensuring that data is uniform within and across data sets and that it is in sync across the enterprise applications and systems (Mahanti, 2021).

According to Friedman & Smith (2011), poor data quality is the key reason why 40% of business initiatives fail to achieve their targeted results. Therefore, if well-managed data represents value, on the other hand, poor quality data depicts risk and cost.

2.2.2. Metadata

To improve its usage and application, data should be associated with metadata that identifies its origin, quality, provenance, language, and semantics

(Devarakonda, et al., 2015). Metadata may be defined as the who, what, when, where, and how of every aspect of the data (Michener, 2006). Metadata is also referred to as the set of instructions or documentation that describes the content, context, quality, structure, and accessibility of a data set, being the most significant reason to invest time and effort into metadata development that human memory is limited, hence adequate metadata will be necessary if data is to be reused (Michener, et al., 1997). Capturing, categorizing, storing, maintaining, integrating, regulating, controlling, and disseminating metadata are all aspects of metadata management (Mahanti, 2021).

2.2.3. Data Modelling

The goal of data modelling is to create a data model that will serve as a link to provide key insights gathered via the analysis of data along with business rules (Patel, 2019). Data modelling is about describing business rules and requirements that govern data and then representing these in the form of a data model, being the entity-relationship (E-R) model the most used model, it is expressed in terms of entities of the business and its attributes as well as the relationships between these entities (Hoffer, et al., 2016). In the E-R model, an entity is a business object and a focus point, and the relation expresses the relationships between these entities, while its attributes are what characterize the entity (Xu, et al., 2010).

3. METHODS

3.1.Project Organization

The project team was composed by me as a data engineer trainee, by two data engineers (a junior and a senior one) and a project manager. The project manager had the responsibility to evaluate the status of the project, delegate tasks and perform communications with company A regarding deadlines establishment and resolving inconveniences. On the other hand, the other team members had the responsibility of implementing the solution for the two clients. Our interactions with company A were mainly through their project manager who was in charge of evaluating the project status, give feedback and clarify business related requirements.

The project followed some of the agile principles which are described next. Communication, collaboration and the welcoming of change in requirements during the project are all emphasized in the Agile Manifesto (Beck, et al., 2001). The agile approach promotes continuous improvement and feedback throughout the project, as seen by its principle "At regular intervals, the team reflects on how to become more effective, then tunes and adapts its behaviour accordingly." (Beck, et al., 2001). Throughout the project, these agile principles were followed. There were established two daily meetings during the project. One of these meetings was with the internal team in which the status of the project was shared with the project manager and technical doubts were clarified between the data engineers. The second meeting was with both the internal team and the project manager from company A in which the communication was mainly between the two project managers. In this meeting we received feedback from company A as well as it was discussed new requirements to the project, establishment of deadlines and possible doubts regarding project requirements. These meetings allowed for collaboration and constant communication as well as gathering feedback and adjusting based on prospective improvements, all of which are examples of agile principles.

3.2.Implementation Steps

Notice that the solution was to be implemented for two clients, as a result, the project was divided into two phases, each of which included one client's implementation and therefore, the implementation of the solution was conducted twice. For reporting purposes, the solution implementation will be described for one of these clients as the implementation was identical.

The steps conducted to implement the solution were the following:

- Populating the data warehouse:
    - From January 2021 to January 2022, we populated the data warehouse with operational data from the source database replica. These data would be referred to as “historical data”.
    - From the current month of the solution implementation (February 2022) onwards we populated the data warehouse by ETL processes of data cleaning and transformation from its source database replica to the data warehouse.
- Reporting in Microsoft Excel by connecting the existent files to the data warehouse and gathering the relevant data to get the values for each defined KPI.

Throughout the project, the work was divided by tables in the data warehouse, therefore each team member was responsible for both the stages of populating the table in the data warehouse and then, to report on Microsoft Excel the KPIs that were related to that table. As a result, the data warehouse would be populated once all team members had finished populating their tables. Furthermore, after the reporting phase was also completed for all team members, the excel file would be complete and ready to be delivered.

4. ACTIVITIES

The purpose of this section is to describe the activities that were performed during the internship, particularly regarding the solution implementation and its phases which were outlined in the methods section.

4.1.Training

To be able to adequately contribute to the project, the first two weeks of the internship consisted of online training. The training had two online courses regarding Microsoft Power Bi and SSIS from Udemy.com.

4.2.Architecture of the solution

After completing the training, I was welcomed into the project team from Link Redglue. Throughout the first weeks of the project, we met with the solution's developers from company A to gain a deeper knowledge of how the solution was built and implemented for the previous clients. Following, figure 1 represents the solution's architecture from the source layer till the reporting phase.

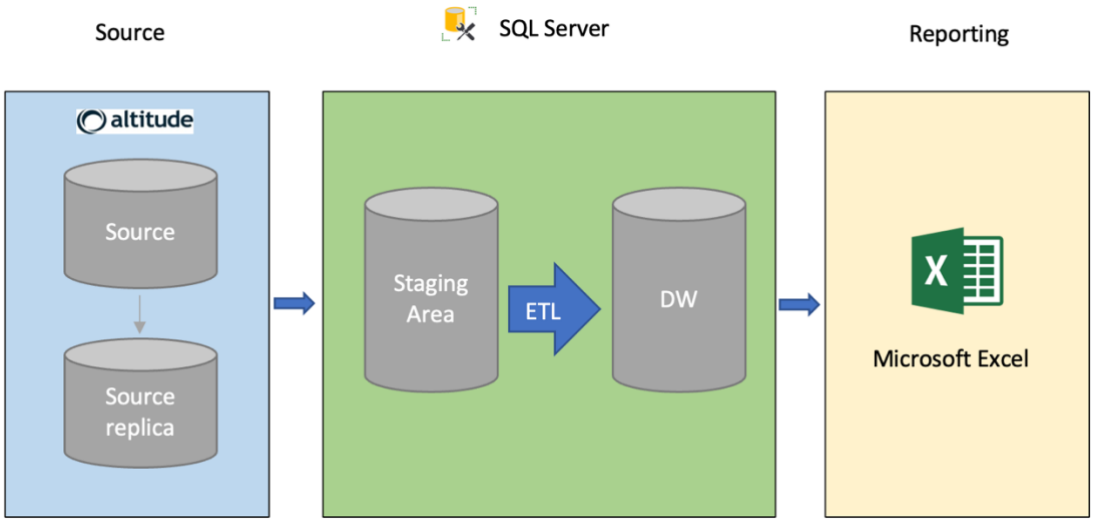

Figure 1 - Architecture of the solution

The source layer represents the operational environment of the contact-centre data that uses Altitude which is a contact-centre software. The source data

remained unchanged for performance reasons, instead there was a replica of the source data consisting of an exact copy of the source data referred to as the source data base replica.

From the source data base replica, data was extracted and brought to the staging area which is located between the source and the data warehouse and simplifies data cleaning and transformation.

Having the data in the BI environment, ETL processes were performed from the staging area to the data warehouse. These processes were built on top of the staging area and aimed to clean the data and improve its quality before loading it into the data warehouse. Once in the data warehouse, data was cleaned, consistent and transformed according to the business rules for the client.

Finally, having the data warehouse loaded, it is possible to build reporting on top of it. At this step Microsoft Excel was used and connected to the data warehouse. Using pivot tables to aggregate the retrieved data and define the values of the KPIs, it was possible to represent the KPIs values and then reporting it to the client.

4.3.Populating the data warehouse

Having completed the context phase of the project, our team began with the implementation of the solution being the first step to populate the data warehouse. Since the tables in the data warehouse were already created our goal at this point was to populate them with data referred to the client. In this case it will be described the populating of the data warehouse for client identified by the number 8.

The tables created in the data warehouse referred to the most important subjects of the contact-centre environment. One example of fact table was regarding the agents in the contact-centre and was referred to as “Agents table”. This table characterized the agent’s contacts, these agents are the professionals that work in the contact-centre and communicate with income callers. This table contains

columns that characterize each call performed by an agent: agent name and username characterize the agent and the conversational time, subject, and so on characterize the call performed. The *Agents* table population process in the data warehouse is the one that will be described next as it was my responsibility.

Notice that, when populating the data warehouse, data was split into two subsets: past data which included data from before the project's starting month and data from the project's starting month onwards. As it was a requirement from the client to keep all the data from the past unchanged, this data should be copied exactly as it was in the source to the data warehouse and would be referred to as "historical data". Moreover, the client defined that the relevant historical data to be included in the data warehouse was data from the past year, as a result, since the implementation for the client occurred in February 2022, the relevant historical data to be brought to the data warehouse was data from January 2021 to January 2022. The second subset of data consisted of data from the starting month of the project implementation onwards, that data would be subject to ETL processes to improve its quality. Therefore, data from February 2022 onwards was the subject of these ETL processes. This split of data was a requirement, as data from the past was already acknowledged by business, they wanted to keep it exactly like it was in the source. From the project implementation onwards, they required that ETL processes were conducted so that future data will have improved quality.

4.3.1. Historical Data

This sub-phase consisted of copying the data between January 2021 and January 2022 from the source database replica to the staging area and then to the data warehouse. Since the source data base replica and the staging area were hosted in different servers, we used SSIS to perform this copy and only then load the data from the staging area into the data warehouse. Next the steps that were taken are described:

1. Create a temporary table in the staging area using the command *CREATE TABLE*. This table would have the same columns as the *Agents* fact table in the data warehouse and would serve as an intermediate area to move data from the source replica to the data warehouse. The script to create this table had to consider the data types of the source database replica – these would have to be the same in both stages to perform a copy when using SSIS to extract the data from the source replica to this table.
2. Use SSIS to extract data from the source data base replica and load it to the created temporary table in the staging area. In SSIS it was used a data flow task which allows to move data from source to destination. Both source and destination were defined as Object Linking and Embedding Database (OLE DB) connections which allow the use of a connection manager to extract and load data.

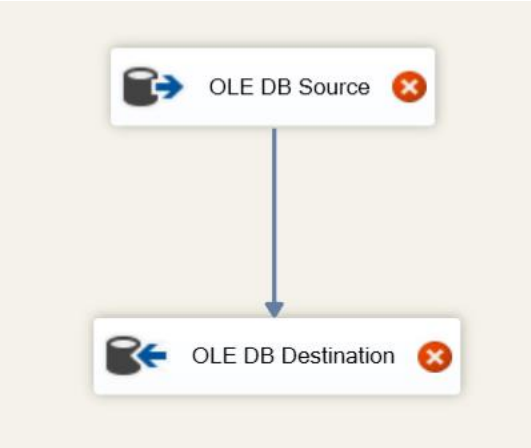

Figure 2 - SSIS data flow task from source to destination

The source consisted of the source data base replica and the destination of the staging area temporary table already created. Before performing the data flow task, both the source and destination were configured. First, the source was defined as the source database replica, and using the connection manager, an SQL statement would be defined to gather the relevant data to be extracted. In this statement, the *SELECT* clause gathered the columns present in the *Agents* table, moreover, since the source data base replica

contained data for multiple clients and for the last 5 years, there was the need to implement specific filters using the *WHERE* clause so that the data to be extracted would be the relevant. The data should be filtered in a manner so that only data regarding the specific client we were dealing at that time and for the relevant time period would be gathered, that is, only data from January 2021 to January 2022 and for the client identified by the number 8. Notice that in the source data base replica the client was identified by its name rather than by the number 8 as it would be in the data warehouse to improve performance. An example of filter created in the SQL Statement used to gather data from the source data base replica in SSIS were:

WHERE CLIENT = "NameClient8" and Day between '2021-01-01' and '2022-01-31'.

Having configured the source connection, next, the destination was configured by selecting the server of the staging area as well as the temporary table created before. As in the source configuration only the columns present in the temporary *Agents* table were gathered in the SQL Statement, the SSIS would automatically assign the columns from source to destination when performing the data flow task.

Having completed the data flow task, the *Agents* temporary table in the staging area was populated with the historical data for client 8. We could then move on with the loading of this data to the data warehouse *Agents* fact table.

3. Insert the data from the staging area temporary table to the *Agents* fact table in the data warehouse. Having the staging area *Agents* temporary table populated, we could proceed and copy its data to the data warehouse. As the staging area and the data warehouse were hosted in the same server, an *INSERT TO* command using SQL Server was enough to copy the data from the staging area to the data warehouse. At this point, we would perform the relevant transformations so that data would be copied accurately according

to the data types defined in the *Agents* table in the data warehouse that for some columns were not the same. An example of transformation that had to be done at this phase was the following: the column "TP_CONV" referring to the conversation time of a call was structured in time *hh: mm: ss* in the source replica and hence in the staging area as well. However, in the data warehouse this column was defined as an integer and intended to represent the conversational time in seconds, therefore, when inserting the data from the staging area to the data warehouse there was the need to perform the following transformation:

*DATEPART (hour, TP_CONV) * 3600 + DATEPART (minute, TP_CONV) * 60 + DATEPART (second, TP_CONV) as TP_CONV*

The *DATEPART* function returns a specified part of the date – the first argument is the part of the date to be returned, the second is the date to be used. As a result, the above expression would populate the column *TP_CONV* in the *Agents* table of the data warehouse in seconds as desired. Since the historical data copied at this step referred to client 8, when populating the table in the data warehouse we assured that the *CLIENT* column was populated with the number 8, using *CAST (8 as smallint) as CLIENT*. Smallint is a datatype that uses integer data that ranges from (-32768) to 32767, moreover, arithmetic operations are conducted effectively since this data type only requires two bytes of storage per value (IBM, 2021). The *CAST* specification returns the first operand to the data type specified. (IBM, 2021). As a result, rather than the client's name as in the source replica, in the data warehouse the *CLIENT* column was identified by its number for performance purposes.

This type of transformations regarding data types were the only performed at this step since this data is historical and as a result should be as it was in the source data base replica. Finally, when running the script *INSERT TO*, we would have populated the data warehouse *Agents* fact table with the historical data from January 2021 to January 2022 for client 8.

4.3.2. ETL processes

Data from February 2022 onwards was subject to ETL processes to improve data quality. As mentioned before, the ETL scripts were already developed by company A for previous clients, therefore, our goal was to perform the necessary changes and adaptations in the scripts according to the business rules of each new client.

These scripts transformed the data from the staging area by applying rules, cleaning the data and dealing with inconsistencies and then loaded the transformed data into the data warehouse. In the staging area, there were already created tables by company A which were referred to as ODS tables. The presence of current data in these tables was also assured by company A as they developed and implemented jobs that extracted data from the source replica and loaded it into the staging area tables and hence the operational data was structured in an organized manner in the staging area. These jobs are a series of actions that are performed by SQL Server on a defined period of time, in this case daily. As a result, the staging area contained current operational data but arranged in tables to facilitate the creation of ETL processes which were built on top of these tables.

The ETL scripts gathered data from the ODS tables in the staging area and applied transformations and cleaning of the data. These scripts also defined new columns that intended to improve the structure and consistency of the data. An example of a column that was created in the ETL scripts was the column *CLIENT* to identify the specific client, we used: *CAST (8 as smallint) as CLIENT* when producing ETL processes for client identified by the number 8. Another example of a column that was created in the ETL scripts was the *Subject Code* which in the source environment context would be identified by its name. However, in the data warehouse we identified this column by a code to improve efficiency. *CASE* statements were used to perform this, following there is an illustrative example:

CASE

WHEN SUBJECT = 'name1'

THEN 1

WHEN SUBJECT = 'name2'

THEN 2

ELSE NULL

END AS SUBJECT_CODE

The *CASE* statement allows to return a value based on the evaluation of one or more conditions (IBM, 2021).

Therefore, the ETL processes implementation consisted of analysing the script for previous clients and changing it accordingly to each new client business rules. Having completed the script implementation, we delivered it to company A to be validated and then implemented as a job to run every day and load the data warehouse.

4.4.Reporting

After populating the table *Agents* in the data warehouse, it was possible to move on to the next activity: reporting the indicators related to this table. Some examples of KPIs to be reported consisted of the total number of received calls and the average conversational time.

At this point, we connected the existing excel files to SQL Server where the data warehouse was hosted using the option "Get Data from SQL Server database" in excel. Using this option, we identified the server and the name of the database and then configured the SQL statement responsible for querying the data to be brought to excel. An illustrative example of SQL statement used to gather data from the data warehouse is the following:

SELECT Day, Month, TP_CONV, Agent

FROM [DW].[Agents_Fact_Table]

WHERE CLIENT = 8

When completing the connection to the data warehouse, a pivot table was automatically created, and it was possible to aggregate data as desired. For example, the average of *TP_CONV* as value and *Month* and *Agent* as rows. As a result, this pivot table would contain the average conversational time for each agent by month which was an indicator to be filled in the excel sheets. These pivot tables served as an intermediate step before filling the actual sheets that were presented to the client and therefore, the sheets that contained the pivot tables were hidden before delivering the excel to the client.

The next step was to fill the sheets to be presented to the client with the KPI values. As the sheets were already designed and contained the agents' names and the relevant dates, we needed to fill in the indicator's values. To complete the sheets with the values of the indicators, the function "GETPIVOT" was used, which can obtain specific data from a pivot table by name based on the structure. This function uses the syntax *=GETPIVOTDATA (data_field, pivot_table, [field1, item1], ...).* As a result, this function would make possible to gather the KPI values from the pivot table and hence complete the sheets with the indicators defined already by company A.

KPIS	feb/2022	march/2022
Average Conversational Time		
Agent1	02:30:40	01:25:32
Agent2	00:01:50	00:50:16
Total number of calls		
Agent1	310	120
Agent2	230	123

Figure 3 - Illustrative example of KPIs used in excel

The sheets contained in the existing excel files contained tables similar to the one shown in figure 3. There, the KPIs, agent's name and dates were already

specified, hence the KPI values were filled using the function *GETPIVOT* to gather data from the created pivot table and by referencing the respective month, agent name and KPI it would automatically fill the value.

This process was conducted for all the KPIs defined in the sheets until it was complete – by doing a data refreshment every day, the customers would have current data on the KPIs defined for their business.

4.5.Documentation

As we were working for company A, one of the requirements for the project were to produce documentation of how we implemented the solution. Once again, each team member was responsible to describe the process of the tables he/she was responsible for during the project. Moreover, we characterized the client, the KPIs and the objectives of the project.

This document was then handled to company A so that if in the future there is the need to implement the solution for a new client by another team it would be as easy as possible.

5. DISCUSSION

The internship allowed for the application and consolidation of knowledge gained during the master's degree program in a real business context, allowing for an increased understanding of tools such as SQL, SSIS and a greater sensitivity in terms of concern for the quality of information in the reports prepared.

At this point, it is possible to consider what may have been done differently and what would have benefited the project, therefore, following are some aspects of the project that in my perspective would have improved its relevance to the master's degree as well as its performance.

One component of the project that, in my perspective, could be improved is the reporting phase. During this phase, Excel was used as a reporting tool to represent the defined KPIs for the client. To enhance performance, it is important to use tools that allow to quickly analyse data and translate it into information, in this regard, choosing a reporting tool that improves performance is critical. When retrieving data from the data warehouse, Excel showed slow processing, which indicates that when the client refreshes the files on a regular basis to show current data it will not be as efficient as it would be with a more suitable tool. Furthermore, the created files showed mainly tables containing the KPIs which is not as appealing as using data visualization. Therefore, the adoption of a tool that allows for more efficient performance as well as sophisticated data visualization would have improved the reporting phase for both the project and internship experience. Power BI is an example of this type of tool because it enables for improved data extraction performance from the data warehouse having faster processing than Excel. Furthermore, Power BI data visualization is more sophisticated, being able to provide a clearer and more appealing representation of the business state to the final client.

Another aspect that could have been improved was business alignment during the project. While there were background meetings at the start of the project

regarding the technological architecture of the solution, it was not given context regarding contact-centre environment. Contact-centre environment background would have been beneficial for understanding the context of the data we were dealing with. It would have helped in the comprehension of the tables in the data warehouse as these ones consisted of the major entities related to contact-centre environment including Inbound, Outbound, IVR, IQS and Agents as fact tables. However, because this background is important for understanding the data used during the project, some context of the contact-centre environment is provided below.

Each contact was characterized as whether inbound or outbound initiative. Inbound handles incoming contacts whereas outbound contacts occurs when an agent reaches out to the customer. Moreover, the contact centre also comprises IVR and IQS. IVR is an automated voice response system from which incoming callers can obtain information without speaking to an agent, as well as use menu options to have their call routed to specified departments. IQS enables the customers to evaluate their calls from a scale 0-10 after the call.

Another aspect that could have been enhanced is related to the choice of the data types used in each column of the tables in the data warehouse. As the development of the project allowed for a deep understanding of the data warehouse design it was possible to analyse possible improvements that could be made in order to save up space in the database and increase its efficiency. For example, the *CLIENT* column that was present in all fact tables to identify the client by a defined number was established as a *smallint.* To save space in the database and make it as efficient as possible, the smallest data type that can reliably contain all possible values should be used. Therefore, as in the project context, the client number ranged from 1 to 30, this column should consist of a *tinyint* data type which takes up less space in the database than the *smallint* datatype and can accurately contain all possible values.

6. CONCLUSION

The completion of this internship at Link Redglue was a unique opportunity, as it allowed the combination of relevant dimensions of my academic background with the first work experience.

The proposed objectives of the project were accomplished. The creation of the data warehouse and the implemented ETL processes for each client will allow for improved data quality as well as a unified place to store all their data in an organized manner. Furthermore, having access to KPIs on a daily basis will allow both clients to have a clear knowledge of their business status. Summing up, as it is the objective of BI projects, the successful implementation of this project will enable these clients to extract value from their data regarding their contact-centre processes in the future.

The fact that the project focused on end-to-end development from data source to final reporting was beneficial and essential in understanding how raw data can generate value for business, which is what business intelligence is all about. Moreover, working for a customer allowed me to better understand business requirements, timelines, and commitment, as well as the importance of business alignment when implementing a BI project.

Despite the internship's overall positive experience and academic relevance, there were certain limitations during its course. As the project consisted of the implementation of a solution already established by company A it would have been relevant to provide documentation as well as business related context such as contact-centre related background.

To improve the future performance of the proposed solution implementation, I suggest enhancing reporting with a more powerful tool such as Power BI to allow for sophisticated visualizations and increased performance as well as improving business alignment by providing documentation for the developed solution available, as well as taking a closer look at the contact-centre environment context.

REFERENCES

Balaceanu, D., 2007. Components of a Business Intelligence software solution. *Informatica Economica*, pp. 67-73.

Beck, K. et al., 2001. *Manifesto for Agile Software Development.* [Online]. Available at: http://agilemanifesto.org/. [Accessed 10 03 2022].

Bordeleau, F.-E., Mosconi, E. & Santa-Eulalia, L. A. d., 2020. Business intelligence and analytics value creation in Industry 4.0: a multiple case study in manufacturing medium enterprises. *Production Planning & Control*, pp. 173-185.

DAMA International, 2017. *DAMA-DMBOK: Data Management Body of Knowledge.* Bradley Beach, New Jersey: Technics Publications.

Devarakonda, R. et al., 2015. *Use of a Metadata Documentation and Search Tool for Large Data Volumes: The NGEE Arctic Example.* Santa Clara, CA, IEEE, pp. 2814-2816.

Elbashir, M. Z., Collier, P. A. & Davern, M. J., 2008. Measuring the effects of business intelligence systems: The relationship between business process and organizational performance. *International Journal of Accounting Information Systems*, p. 135–153.

European Commission, 2020. *European data strategy*. [Online]. Available at: https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/european-data-strategy_en. [Accessed January 2022].

Friedman, T. & Smith, M., 2011. *Measuring the Business Value of Data Quality*, s.l.: Gartner.

Gangadharan, G. & Swami, S. N., 2004. *Business Intelligence Systems: Design and Implementation Strategies'.* Cavtat, Croatia, IEEE, pp. 139-144.

García, J. M. V. & Pinzón, B. H. D., 2017. Key success factors to business intelligence solution implementation. *Journal of Intelligence Studies in Business*, pp. 5-86.

Gartner, 2019. *Gartner Glossary.* [Online]. Available at: https://www.gartner.com/en/information-technology/glossary/dmi-data-management-and-integration. [Accessed 11 01 2022].

Gorball, J., 2016. *Data Governance vs. Data Management.* [Online] Available at: https://blog.kingland.com/data-governance-vs.-data-management. [Accessed 15 02 2022].

Gowthami, K. & Kumar, M. P., 2017. Study on Business Intelligence Tools for Enterprise Dashboard Development. *International Research Journal of Engineering and Technology*, pp. 2987-2992.

Hermann, M., Pentek, T. & Boris, O., 2016. *Design Principles for Industrie 4.0 Scenarios.* Koloa, HI, IEEE, pp. 3928-3937.

Hoffer, J. A., Venkataraman, R. & Topi, H., 2016. *Modern Database Management.* 12 ed. New York City, NY: Pearson.

IBM, 2021. *SMALLINT data type*. [Online]. Available at: www.ibm.com/docs/en/informix-servers/12.10?topic=types-smallint-data-type. [Accessed 10 02 2022].

INESC, 2020. About INESC. [Online]. Available at: https://inesc.pt/pt/inesc-pt/o-insec. [Accessed 24 01 2022]

Kadadi, A., Agrawal, R., Nyamful, C. & Atiq, R., 2014. *Challenges of Data Integration and Interoperability in Big Data.* Washington, DC, IEEE, pp. 38-40.

Kimball, R. & Caserta, J., 2014. *The Data Warehouse ETL Toolkit: Practical Techniques for Extracting, Cleaning, Conforming, and Delivering Data.* Hoboken, New Jersey: Wiley.

Kiron, D., 2017. Lessons from becoming a data-driven organization. *MIT Sloan Management Review*.

Kopanakis, I., Vassakis, K. & Petrakis, E., 2018. Big Data Analytics: Applications, Prospects and Challenges. In: *Mobile Big Data.* New York City, New York: Springer, pp. 3-20.

Kwon, O., Lee, N. & Shin, B., 2014. Data quality management, data usage experience and acquisition intention of big data analytics 34. *International Journal of Information Management*, pp. 387-394.

Link Redglue, 2017. *Leveraging Data*. [Online]. Available at: https://linkredglue.com/leveraging-data/. [Accessed 24 01 2022]

Link, 2019. *About Link*. [Online] Available at: https://linkconsulting.com/about-link/#about-link-group. [Accessed 24 01 2022]

Mahanti, R., 2021. Data and Its Governance. In: *Data Governance and Data Management.* New York City, NY: Springer, pp. 5-82.

Michener, W. K., 2006. Meta-information concepts for ecological data management. *Ecological informatics*, pp. 3-7.

Michener, W. K. et al., 1997. Nongeospatial metadata for the ecological sciences. *Ecological Applications*, pp. 330-342.

Negash, S. & Gray, P., 2008. Business Intelligence. In: *Handbook on Decision Support Systems 2.* New York City, NY: Springer, pp. 175-193.

Olson, J., 2003. *Data Quality: The Accuracy Dimension.* Amsterdam: Elsevier.

Ong, I. L., Siew, P. H. & Wong, S. F., 2011. A five-layered business intelligence architecture. *Communications of the IBIMA.*

Papadopoulos, T., Gunasekaran, A., Dubey, R. & Wamba, S. F., 2017. Big data and the transformation of operations models: a framework and a new research agenda. *Production Planning & Control*, pp. 873-876.

Patel, J., 2019. *An Effective and Scalable Data Modeling for Enterprise Big Data Platform.* Los Angeles, CA, IEEE, pp. 2691-2697.

Patil, D. & Mason, H., 2015. *Data Driven.* Sebastopol, CA: O'Reilly Media, Inc..

Popovič, A., Hackney, R., Coelho, P. S. & Jaklič, J., 2012. Towards business intelligence systems success: Effects of maturity and culture on analytical decision making. *Decision Support Systems*, pp. 729-739.

Ranjan, J., 2009. Business intelligence: Concepts, components, techniques and benefits. *Journal of theoretical and applied information technology*, pp. 60-70.

Ransbotham, S., Kiron, D. & Prentice, P. K., 2015. *Minding the analytics gap*, Cambridge, MA: MIT Sloan Management Review.

Rio-Ortega, A. d., Resinas, M. & Cortés, A. R., 2009. Towards Modelling and Tracing Key Performance Indicators in Business Processes. *II Taller sobre Procesos de Negocio e Ingeniería de Servicios, PNIS*.

Vo, Q. D. et al., 2018. *Next generation business intelligence and analytics.* New York, USA, Association for Computing Machinery, pp. 163-168.

Xu, L., Lee, S. & Kim, S., 2010. *E-R model based RDF data storage in RDB.* Chengdu, IEEE, pp. 258-262.

Zeng, L. et al., 2006. *Techniques, process, and enterprise solutions of business intelligence.* Taipei, Taiwan, IEEE.