NOVA IMS Information Management School

# MGI

Master Degree Program in

**Information Management with a specialization in Knowledge Management and Business Intelligence**

## Business intelligence-centered software as the main driver to migrate from spreadsheet-based analytics

Tomás de Oliveira Saramago Brito Rebelo

Internship Report

presented as partial requirement for obtaining the Master Degree Program in Information Management with a specialization in Knowledge Management and Business Intelligence

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**
Universidade Nova de Lisboa

# BUSINESS INTELLIGENCE-CENTERED SOFTWARE AS THE MAIN DRIVER TO MIGRATE FROM SPREADSHEET-BASED ANALYTICS

by

Tomás Rebelo

Internship report presented as partial requirement for obtaining the Master's degree in Information Management with a specialization in Knowledge Management and Business Intelligence

**Supervisor / Co Supervisor:** Prof. Doutor Flávio Luís Portas Pinheiro

October 2022

# STATEMENT OF INTEGRITY

I hereby declare having conducted this academic work with integrity. I confirm that I have not used plagiarism or any form of undue use of information or falsification of results along the process leading to its elaboration. I further declare that I have fully acknowledge the Rules of Conduct and Code of Honor from the NOVA Information Management School.

*Tomás Rebelo*

*Lisbon, October 21th, 2022*

# DEDICATION

To my grandmother Teresa Oliveira, who tirelessly believed in me from the beginning of the Master's Degree and made all of this possible. To my twin brother Rodrigo Rebelo and my mother Cláudia Santana, for the love and passion they're always sharing with me. To Serena Velez and her parents Gianna Rivara and Jorge Velez, for their constant support and for teaching me how to think, act and decide the best way possible. To Miguel Lopes and his wisdom, sharing countless pieces of advice that made (and make) me a stronger professional and human being. To my uncle David Oliveira, for teaching me so many personal things and for making me smile at every moment. To my uncles Mafalda Azevedo and Miguel Pais and my cousins Carlota Pais and Francisca Pais, who make me a better family player and, above all, a happier man. To my grandmother Rosa Melo and my grandfather João Melo for the immeasurable love and compassion they have for me. Finally, to my father Miguel Rebelo, for educating me and teaching me fundamental values that shaped the person I am today.

# ACKNOWLEDGEMENTS

I would like to express my appreciation to Professor Doutor Flávio Luís Portas Pinheiro for his constant feedback and patience throughout the development of this Report, as well as for his knowledge-sharing during his classes.

I am also grateful to my teammates Alexandre Vassilenko, Nuno Belo, Gonçalo Silvestre, and Manuel Beijinho for all the late-night sessions coding in Python and discussing business topics. The projects we developed wouldn't be the same without you.

# ABSTRACT

Nowadays, companies are handling and managing data in a way that they weren't ten years ago. The data deluge is, as a mere consequence of that, the constant day-to-day challenge for them - having to create agile and scalable data solutions to tackle this reality.

The main trigger of this project was to support the decision-making process of a customer-centered marketing team (called Customer Voice) in the Company X by developing a complete, holistic Business Intelligence solution that goes all the way from ETL processes to data visualizations based on that team's business needs. Having this context into consideration, the focus of the internship was to make use of BI, ETL techniques to migrate their data stored in spreadsheets — where they performed data analysis — and shift the way they see the data into a more dynamic, sophisticated, and suitable way in order to help them make data-driven strategic decisions.

To ensure that there was credibility throughout the development of this project and its subsequent solution, it was necessary to make an exhaustive literature review to help me frame this project in a more realistic and logical way. That being said, this report made use of scientific literature that explained the evolution of the ETL workflows, tools, and limitations across different time periods and generations, how it was transformed from manual to real-time data tasks together with data warehouses, the importance of data quality and, finally, the relevance of ETL processes optimization and new ways of approaching data integrations by using modern, cloud architectures.

# KEYWORDS

ETL; Data Warehouse; Business Intelligence; Dashboards; Data Visualization; Data Engineering

# INDEX

# LIST OF FIGURES

# LIST OF TABLES

# LIST OF ABBREVIATIONS AND ACRONYMS

| | |
|---|---|
| **API** | Application Programming Interface |
| **ARR** | Annual Recurring Revenue |
| **BI** | Business Intelligence |
| **CV** | Customer Voice |
| **DBMS** | Database Management System |
| **DW** | Data Warehouse |
| **ETL** | Extraction, Transformation and Load |
| **OLAP** | Online Analytical Processing |
| **OLTP** | Online Transaction Processing |
| **OWB** | Oracle Warehouse Builder |
| **RDBMS** | Relational Database Management System |

# 1. INTRODUCTION

With the need to stay ahead of the ever-changing market needs, marketing teams have been using and consuming more and more data over time. Indeed, informed decision-making is a key requirement for competitiveness in a global marketplace characterized by uncertainty and fast technological changes (Esmail, 2014). According to a report by McKinsey (2021), Marketing has been on the front lines of digital transformation and revolution. Still, the landscape has become much more complex — forcing a focus on data, growth, and new ways of creativity. Additionally, the incremental use of first-party data is now part of the day-to-day life of modern marketing teams, a reality that was not true five years ago. Hence, marketing teams must deal with an ever-increasing customer database and need to ensure the right data, with the right quality, and the adequate consistency of processes so that the specific data storage(s) system(s) they're using continue offering the necessary responsiveness and latency while achieving to be scalable. High data quality may be critical for any data warehouse success project (Loshin, 2003). On the other hand, poor data quality can lead to unfavorable consequences on the decision-making process (Huang et al., 1999; Clikeman, 1999).

According to Inmon (1996), a data warehouse is a "subject-oriented, integrated, time-varying, non-volatile collection of data in support of the management's decision-making process." Since marketing teams have a strategy as their core work engine, they're resorting to data architectures such as data warehouses to integrate multiple data sources and perform drill downs and drill-throughs for better data analysis (Kimball, 2003). To conceptually build them, it is necessary to make use of correct ETL techniques to make sure that the data loaded in the data warehouse is made of purity and robustness, and the quality of the data is directly dependent on the ETL process efficiency (Kakish & Kraft, 2012). In addition, today, companies can also take advantage of real-time data processing tools and data lakes to form a data warehouse prepared for event-driven, massive data entries – ensuring timely insights. Having that in consideration, only after this data engineering process it is indeed possible and recommendable to develop data visualizations for an easier and more intuitive data interpretation.

With that said, this internship report, consequently addressed the historical evolution of ETL and what should be done to optimize it over time, covered the main characteristics of data quality and new ways of performing data integrations beyond the traditional ETL architecture, and described the evolution from enterprise data warehouses to data lakes as well as the differences between data warehouses deployed on-premises versus on cloud environments. With this, the creation, development, and deployment of the BI solution I implemented was not only based on the current practices of Company X but also supported by the scientific literature.

## 1.1. CONTEXT AND MOTIVATION

To provide an overview of the context and motivation of the BI project I was assigned to, the Customer Voice team is a department fully focused on the customer and marketing oriented. From the publication of case studies on the company's web properties to helping internal teams such as Sales, Field Enablement and Field Marketing by sharing rich customer information to improve internal processes and leverage external impact (to the market), this department was making use of a "mega" spreadsheet with years of use, thousands of rows, and hundreds of columns to store customer-related data. As there were always data entries every day, the team's data source was becoming slower and less intuitive to use. As a cost, the latency was getting higher and it was consuming time and effort that could be allocated to focus on the business, strategic sides. Additionally, as they also performed data analysis in spreadsheets, they were obtaining limited, static insights instead of unlimited, dynamic information. Therefore, there were a lot of counterparts in the way the Customer Voice team managed and leveraged the power of data. With that context in mind, they asked the Data Engineering team (which I was part of as an intern) to help them migrate from Excel to a scalable BI solution so they could see different business scenarios covered in the form of real-time dashboards.

## 1.2. PROJECT DESCRIPTION

The goal of this report was to help a customer-centered marketing team, called Customer Voice, to migrate from spreadsheets, where they essentially stored data about their customers, to a Business Intelligence platform running on Power BI. Having joined the Data Engineering team in the Company X, I was responsible for developing ETL processes and the relational data model directly into Power BI to populate the requested/necessary dashboards. This process was intended to help the client team to shift from manual to automatic data entries as well as to improve their decision-making process by creating easy-to-understand and meaningful data-driven visualizations that could serve as actionable items to whatever strategies they wanted to make in the future.

# 2. LITERATURE REVIEW

## 2.1. ETL HISTORICAL EVOLUTION

ETL, which stands for Extract, Transform and Load, is a data engineering process responsible for the integration of business data from a number of different sources, its transformation into strategic, actionable information and the subsequent storage of the transformed data in a format that can facilitate business analysis (Petrovic et al., 2019).

The ETL process is the most complex, time-consuming and expensive phase of a data warehouse development, and given it is estimated that approximately 70% of the time and effort invested in its development is allocated on the ETL process development (Kimball & Caserta, 2004; Kimball et al. 2010) it is crucial to take a thorough methodological approach when developing it. When defining an ETL process for a data warehouse (DW), its implementation depends on the types of the DW and also between department data marts within a data warehouse. According to Zode (n.d.) back in the early 1990's, most of the organizations developed custom code to extract and transform data from OLTP systems and load it into OLAP systems – such as data warehouses. In the middle of the 1900's, the vendors found an opportunity and they started to ship ETL tools optimized for the reduction or even elimination of the labor-intensive process of writing code for custom ETL development programs. But in the end, ETL systems have the common purpose of moving data from one database to another. The figure below represents the traditional workflow of an ETL process.

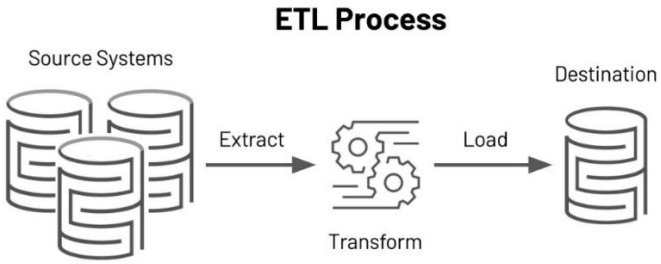



Figure 1 - Technical Architecture of ETL

(Databricks, n.d.)

The ETL process has evolved since its existence, and it's separated into three generations of evolution. *First Generation* of ETL According to Kakish and Kraft (2012), ETL tools of the First Generation ran on proprietary engines and offered good performances due to the inheriting performance of

natively compiled code. However, they were complex to use for the developers because the ETL tools required high technical knowledge of programming in COBOL or C. But with the need of ingesting and processing data generated from different transactional sources, the Second Generation of ETL emerged. Graphical user interfaces and data integration capabilities became available for the ETL processes – requiring lower coding capacities -, and the ETL functions became more automated compared to First Generation's. However, in terms of processing, they were only capable of processing data at apace it ), which led to major impacts on the time required to perform ETL tasks.

The need to overcome the limitations of having high technical knowledge and slow processing tasks performance of the First and Second ETL Generations respectively, led to the most recent *Third Generation*. ETL tools of this Era have a relational architecture that can generate native SQL, removing the hub server between the original sources and the destination systems (Kakish and Kraft, 2012). The tools that are under the *Third-Generation* umbrella reduce the network traffic to improve their overall performance because they have a distributed architecture implemented, which means ETL tasks can be performed in parallel. In addition, the data load is distributed among database engines to raise the levels of scalability and are responsive to all types of data sources. These *Third Generation* ETL stack makes use of relational DBMS to transform the data, so the "row by row" process disappeared. "In the ETL architecture, all database engines can potentially participate in a transformation - thus running each part of the process where it is the most optimized. Any RDBMS can be an engine, and it may make sense to distribute the SQL code among different sources and targets to achieve the best performance. For example, a join between two large tables may be done on the source" (De Montcheuil, Y., 2005). With that in consideration, as Relational Database Management Systems have the ability to integrate data, ETL tools are taking advantage of this capability to improve their performance.

| GENERATION | ADVANTAGES | | LIMITATIONS | |
|---|---|---|---|---|
| ***FIRST*** | i) | Tools are good at extracting the data from legacy systems. | i) | These tools required in-depth knowledge of programming in COBOL or C; |
| | ii) | Performance was good because of inheriting performance of native compiled code. | ii) | Proven less successful on relational database for handling large volume of data; |
| | | | iii) | Transformations require manual coding. |
| ***SECOND*** | i) | Graphical interface ETL tools and available transformations features; | i) | Row by row data transformations when passing the engine; |
| | ii) | ETL functions are highly integrated and automated; | ii) | Engines performing all the transformations became a bottleneck in the transformation process. |
| | iii) | Engine-based approach is fast, efficient and multi-threaded. | | |
| ***THIRD*** | i) | ETL tools support parallelism, enabling developers to perform complex processing jobs faster | i) | Many ETL tools don't support integration at the metadata level with end-user tools |
| | ii) | ETL tools can handle most complex transformations faster; | ii) | Some tools are still limited to the source database like OWB |
| | iii) | The code generated by the code-based tools can run on various platforms at very high speed and also enables organizations to distribute loads across multiple platforms to optimize the performance. | | |

Table 1 - ETL Generations summary

(Zode, M., n.d.)

## 2.2. REAL-TIME DATA WAREHOUSING AND ETL TECHNIQUES

Data warehouses and the underlying ETL techniques are changing in performance, in order to provide updated data all the time. Increasingly, there's the need to support business decisions in near real-time based on the transactional data, having the ETL processes to "move away from periodic refreshments to continuous updates" (Esmail, 2014). According to Santos & Bernardino (2009), "to cope with real-time requirements, the DW must be able to enable continuous data integration, to deal

with the most recent business data", and that is of paramount importance if businesses want to keep up with the ever-changing market needs. A survey of over 300 BI and Information managers identified access to relevant, timely, and reliable data as the highest impediment to information management success, as shown in the figure below (Henschen, 2010). That is proof that data warehouses need to be more agile and designed for real-time data storage, but that only comes with a modern, future-proof ETL approach.

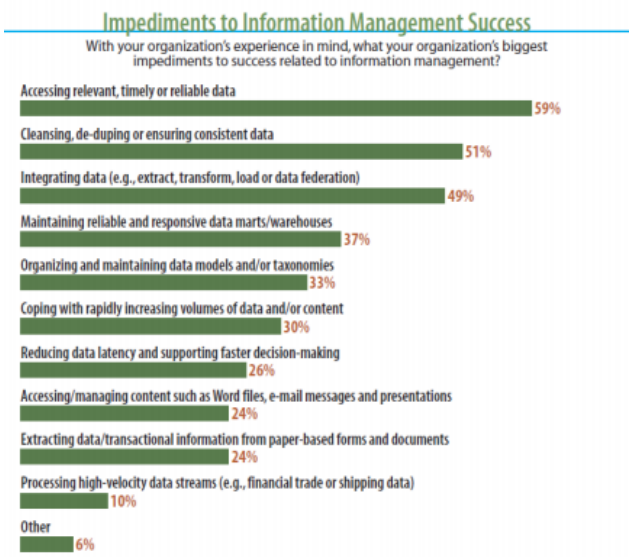



Figure 2 - Impediments to Information Management Success

(Henschen, D., 2010)

Traditionally, static ETL tools load the data periodically during the inactivity time where, during this period, no one can access the data in the data warehouse. Still, real-time DW loads the data continuously as opposed to the traditional approach (Esmail, 2014), improving the loading frequency. If this can be defined as real-time data BI techniques, Agrawal (2009) proposes that real-time Business Intelligence architecture requires that the data stored in OLTP systems or any operational data sources move to the data warehouse in real-time in the format of data streams of events. This shift in the BI architecture introduces a middleware technology component – referred to as the stream analysis engine -, that provides services or capabilities such as data management, application services and API management. This engine performs an accurate analysis of the incoming data just before it can be integrated into de DW to identify data patterns and outliers (Esmail, 2014).

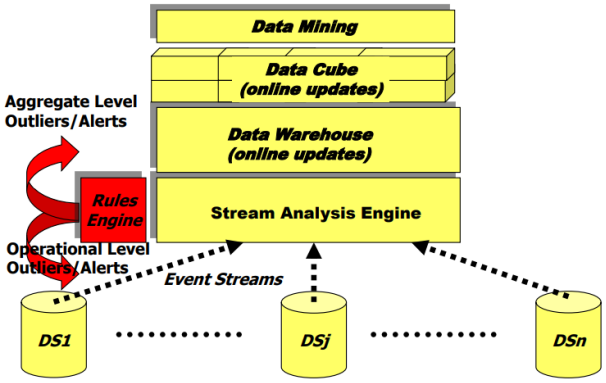



Figure 3 - Real-time BI architecture

(Agrawal, D., 2009)

## 2.3. DATA QUALITY

According to Ruiz (2017), most data scientists spend only 20% of the time on data analysis and the remaining 80% is for data cleaning – "The 80/20 data science dilemma", he calls. This is the prove that ensuring data quality is a critical success factor for any data-related job. In fact, going back to ten years ago, a research program conducted by Bloomberg BusinessWeek Research Services (2011) identified data quality, integrity and consistency as the biggest challenge that companies are facing in their adoption of Business Intelligence and analytics projects.


Figure 4 - Business Intelligence Roadblocks

(Bloomberg BusinessWeek Research Services, 2011)

For a deeper understanding, Singh and Singh (2010) refer that the Data Quality Dimension encompasses six critical factors that need to be taken in consideration:

i) Completeness – which ensures the attributes of data are provided and all required information is available. It is important to mention that even if data is not available, it still could be considered completed – but this only happens when the data meets the needs and expectations of the user;

ii) Consistency – for the data to be consistent, data on the enterprises should be harmonious with each other so that it doesn't conflict with other sets of data;

iii) Validity – refers to the correctness of data;

iv) Conformity – it means that data values are consistent across specific formats. Conformance maintenance is important;

v) Accuracy – it is considered accurate when data is the real-world object or event being described. For instance, incorrect spellings of person and product names can affect the operational and consequent analytical operations;

vi) Integrity – refers to the data's trustworthiness. If the data is missing relationships linkages and is unable to link related records together, it may introduce duplication across the systems.

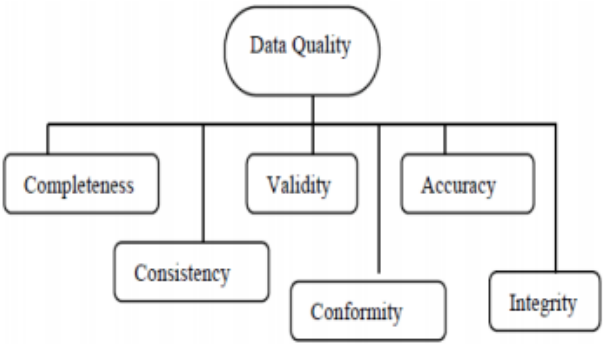



Figure 5 - Data Quality Dimensions

(Singh, R., Singh, K., 2010)

In addition to the previous traditional Data Quality Dimensions, there are other factors that are equally important, such as confidentially, availability and security (Kakish and Kraft, 2012). Confidentially ensures that the use of data is for authorized-only people and availability is providing the information requested by authorized users (Jesan, 2006). Data and information might be the most important asset of any enterprise and data security policies must protect these information assets

(Gerber, 2008). With all the mentioned factors in consideration, data quality and data warehouse effectiveness must be together to ensure the success of a data-related project.

As important as knowing what the six critical factors are to establish data quality, it is also relevant to understand what the cost is of not implementing them. As such, according to a case study published by StreamSets (n.d.), they helped a telecommunications company called BT Group on consolidating data as this customer was running into several challenges that were delaying their ability to deliver. Contextually, they had different versions of the same data from different systems (violating the "Consistency" dimension) that was leading to incorrect operations reporting and analysis. A lag in data (violating the "Availability" dimension) effectively led to an inefficiency of resource allocation, which meant that without all the data they just couldn't put the right people in the right place, at the right time. Additionally, BT Group lacked a 360-degree view of the data environment which ultimately led to bad decision-making. It's clear that by having this context, they needed a way to centralize all the data, ensure that users were using the same source of truth and democratize data to explode user adoption and excitement around data. Having said this, this was a good example of the underlying different implications of not having data quality practices, all the way from operational efficiency to decision-making processes.

## 2.4. ETL PROCESSES OPTIMIZATION

The ETL processes have predefined steps that need to be followed to build a consistent data pipeline. Firstly, the business requirements must be specified so that they can be put down in a concept. Secondly, the concept model needs to be transferred into a logical, architectural model so that, in the end, all the underlying code can be developed successfully according to that logic (Hahn, 2019). However, an ETL process is not a process that should stand still once it's completed. It needs optimization processes and iterations, quality objectives, and measures that can evaluate if the ETL itself is accomplishing the quality objectives defined beforehand. And this can be done in the business requirements definition or concept model, in the logic phase, and in the code (Castellanos et al., 2014).

To check if an optimization had, or not, an impact, the quality objectives need to be defined upfront, otherwise it'll not be possible to measure the optimization contributions for the ETL. According to Hahn (2019), it is a common practice that an ETL workflow is first designed and developed without any additional optimization. The optimization is only made afterward if the quality objectives are not fulfilled.

To measure the optimizations of an ETL process, Simitsis et al. (2009) did a benchmark to study what could be the main elements used to assess the ETL engines and design methods concerning their

standard behavioral properties over a broad range of ETL workflows. As such, the authors were not interested in providing performance measures for every single ETL task nor even enumerating all the possible alternatives for specific operations. The purpose of the benchmark developed by the authors was not to facilitate the comparison of different methods for duplicate data - since it doesn't take the tuning of all the possible parameters for that task into account - but to assess the integration of such methods in complex ETL workflows.

Whether the ETL operates in batch or in real-time, the two critical goals that should be considered and achieved are effectiveness and efficiency (Simitsis et al., 2009). Those are the goals that should be evaluated.

Regarding effectiveness and based on long discussions that these authors had with ETL practitioners and experts, they have identified that real-life ETL projects' performance is not the only goal. It is the other way around, as other optimization qualities are very interesting and important as well, naming them as QoX (Dayal et al., 2009). This is a group of metrics that is fundamentally inside all stages of the design process, from more high-level specifications to the implementation itself: performance, recoverability, reliability, freshness, maintainability, scalability, availability, flexibility, robustness, affordability, consistency, traceability, and auditability.

There are quantitative measures (e.g., reliability, freshness, and others) that might be more difficult to measure (e.g., maintainability and flexibility). In addition, it's important to mention that there are trade-offs or opportunity costs that should be considered, which means that by improving one particular objective another one may be hurt (Simitsis et al., 2009). Regardless of that, the main goal is to have data that respects both database and business needs simultaneously. Simitsis et al. (2009) believed that all the following questions should be considered when creating an ETL benchmark:

i) "Does the workflow execution reach the maximum possible level of data freshness, completeness, and consistency in the warehouse within the necessary time (or resource) constraints?"
ii) "Is the workflow execution resilient to occasional failures?"
iii) "Is the workflow easily maintainable?"

But effectiveness is not the only goal that should be accomplished when an ETL workflow operates. Efficiency is another very important variable of the ETL design. Performance is not everything but plays a critical role since ETL processes should typically run within strict time windows (Simitsis et al. 2009). In fact, achieving high-performance levels can serve as a means for enabling or achieving other qualities as well. Taking that into consideration, Simitsis et al. (2009), have formulated the following questions regarding ETL processes' efficiency:

i) "How fast is the workflow executed?"

ii) “What degree of parallelization is required?”

iii) “How much pipelining does the workflow use?”

In addition, Simitsis et al. (2009) proposed Experimental Parameters and a set of measures to be monitored so they could assess the accomplishment of their benchmark goals.

| | |
|---|---|
| **Experimental Parameters** | P1) The size of the workflow (i.e., the number of nodes contained in the graph) |
| | P2) The structure of the workflow (i.e., the variation of the nature of the involved nodes and their interconnection as the workflow graph) |
| | P3) The size of input data originating from the sources |
| | P4) The workflow selectivity, based on the selectivities of the workflow activities |
| | P5) The values of probabilities of failure |
| | P6) The latency of updates at the warehouse |
| | P7) The required completion time (i.e., this reflects the maximum tolerated execution time window) |
| | P8) The system resources (e.g., memory and processing power) |
| | P9) The “ETL workload” that determines an execution order for ETL workflows and the number of instances of the workflows that should run concurrently |

Table 2 - Experimental Parameters

(Simitsis et al., 2009)

For each Experimental Parameter, measures need to be evaluated in order to determine the fulfillment of the benchmark goals. For that to happen, the authors classified the specific measures according to the assessment question they are responsible to answer.

| Question | Measures |
|---|---|
| Measures for data freshness and data consistency | • Percentage of data that violate business rules<br>• Percentage of data that should be present at their appropriate warehouse targets, but they are not |
| Measures for the resilience to failures | • Percentage of successfully resumed workflow executions<br>• MTBF, the mean time between failures<br>• MTTR, mean time to repair<br>• Number of recovery points used<br>• Resumption type: synchronous or asynchronous<br>• Number of replicated processes (for replication). |

| | |
|---|---|
| | • Uptime of ETL process. |
| Measures for maintainability | • Length of the workflow (lengh of its longest path<br>• Complexity of the workflow refers to the amount of relationships that combine its components<br>• Modularity (or cohesion) refers to the extent to which the workflow components perform exactly one job<br>• Coupling captures the amount of relationship among different recordsets or activities |
| Measures for the speed of the overall process | • Total completion time<br>• Throughput of workflow execution including a specific percentage of failures and their resumption<br>• Average latency per tuple in regular execution. |
| Measures for partitioning | • Partition type which should be chosen according to the characteristics of the workflow (round-robin, follow-database-partitioning and so on)<br>• Number and length of workflow parts that use partitioning<br>• Number of partitions<br>• Data volume in each partition (this is related to partition type too) |
| Measures for pipelining | • CPU and memory utilization for pipelining flows or for individual operation run in such flows<br>• Min/Max/Avg length of the largest and smaller paths containing pipelining operations<br>• Min/Max/Avg number of blocking operations. |
| Measured Overheads | • Min/Max/Avg/ timeline of memory consumed by the ETL process at the source system<br>• Time needed to complete the processing of a certain number of OLTP transactions in the presence of ETL software at the source<br>• Min/Max/Avg/ timeline of memory consumed by the ETL process at the warehouse system<br>• Time needed to complete the processing of a certain number of decision support queries in the presence of ETL software at the warehouse |

Table 3 - Measures by assessment question

(Simitsis et al., 2009)

## 2.5. ELT – A NEW DATA INTEGRATION APPROACH

According to an article published by Fivetran (Wang, C., 2021), when computation, storage and bandwidth were very scarce (being consequently expensive) and the data volume and variety were

limited, the labor intensiveness was acceptable. This means that the ETL process itself is a solution of a time with a lot of technological constraints compared to today.

The cost of storage has decreased from approximately $1 million “to a matter of cents per gigabyte” (Wang, C., 2021), the cost of computation has also decreased by millions since the 1970s and, in addition, the internet transit cost has also plummeted by a factor of thousands.

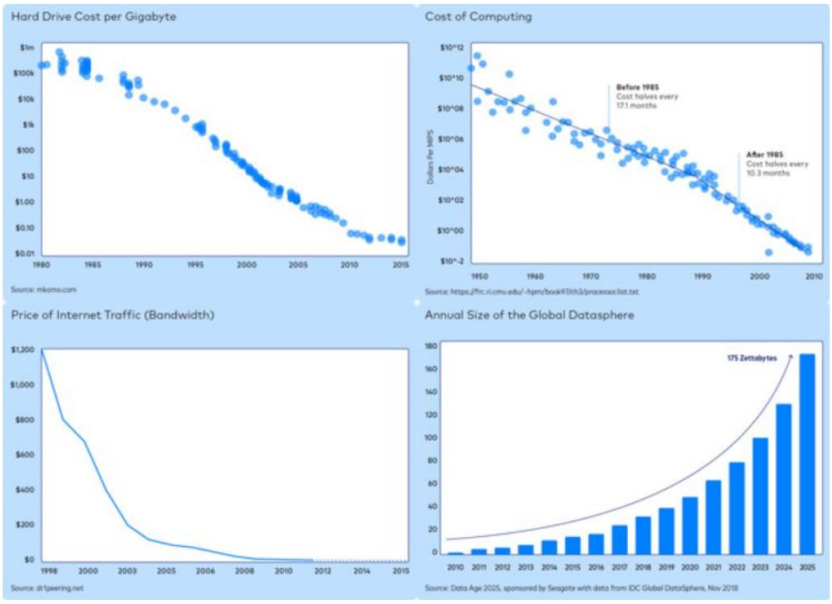



Figure 6 - Technology trends for data cloud integrations

(Wang, C., 2021)

The trends above have made the traditional ETL process almost obsolete, as the affordability of computation, internet bandwidth (with the 5G appearance) and storage led to an exponential growth of cloud-based services. And because the cloud has grown, so has the volume, variety and complexity of data. As cloud services provide more storage capacity, businesses have the ability to store untransformed, event data in data warehouses with a new integration architecture - the ELT (Extract, Transform and Load). This modern alternative for handling data, enables data to be immediately loaded from the source to a destination system and the transformation step moved forward to the end of the workflow.

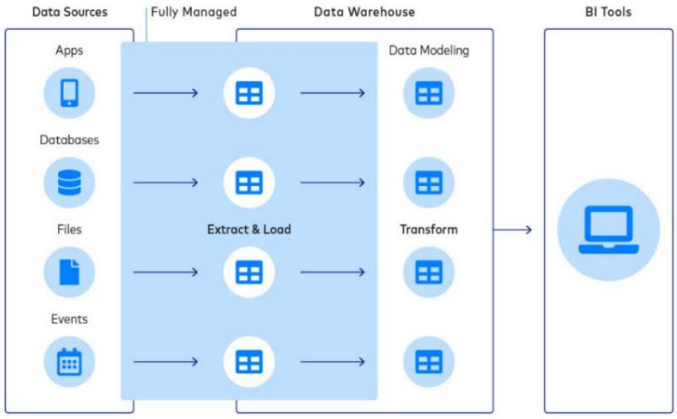



Figure 7 - ELT workflow

(Wang, C., 2021)

A priori, this process prevents two failures of ETL (i.e. changing upstream schemas and downstream data models) from directly impacting extraction and loading, leading to a simpler, more efficient and robust approach to data integration projects. Challenging the traditional ETL processes, here's the workflow features for ELT:

i) Identity the data sources to work on;
ii) Perform automated extraction and loading;
iii) Outline the analytics needs the project is meant to solve;
iv) Create the data models by building transformations;
v) Conduct actual analytics work and extract insights.

The fundamental advantage of using this architecture is the fact that although the Transformation layer may still fail as upstream and downstream schemas change, the data will still be loaded into the destination system - and that is of critical importance for businesses, as they need fresh data every day due to markets pressure and changings.

An organization that matches automation with ELT workflows, "stands to dramatically simplify its data integration workflow" (Wang, C., 2021), leading to faster information delivery capacity to the stakeholders.

## 2.6. DATA ARCHITECTURE EVOLUTION

According to Dataversity (Foote, 2022), the data architecture can be defined as the set of rules, models and policies that determine what kind of data gets collected and how it is used, processed, and stored within a decision support system.

From the 1940s to the early 1970s, computer programs were exclusively created and designed to cope with specific types of computer problems, where processing capacity was the primary concern. As such, the concepts of data integration and data architecture were not even considered (Foote, 2022). The main focus of a programmer was on getting a computer to perform specific types of actions that fundamentally supported an organization's short-term goals, and not for long-term data storage. If there was the need to recover data, it would require the ability to write programs capable of retrieving specific information, which was extremely time-consuming and expensive. With that reality in consideration, there was the need to shift from a Programming Paradigm to a Database Architecture Paradigm.

According to Edgar F. Codd (2002), a relational procedure for organizing data was the step to scale data storage and data management, which resulted in the creation of databases structures that streamlined the efficiency of computers. The relational approach developed by Edgar F. Codd in 1970, allowed users to store data in a more organized, more efficient way using two-dimensional tables, replacing the COBOL programs where the data was arranged hierarchically. A significant advantage of this relational view was that it formed a basis for treating redundancy, derivability, and consistency of relations, allowing for a "clearer evaluation of the scope and logical limitations of formatted data systems" (Codd, 2002).

As an evolution of Edgar F. Codd's work, Peter Chan (1976) introduced the "entity-relationship model" (commonly known as "data modeling"), where the data structures were represented graphically. As a matter of fact, Oracle announced the first relational database management system (RDBMS) designed for business, inspired by the work done by Peter Chan. Consequently, people working with computers began to realize that these data structures were more reliable than program structures.

Codd's relational approach and Chan's data modeling resulted in the Structured Query Language (SQL), which became the standard query language in the 1980s. As such, relational databases became extremely popular and boosted the database market, causing major loss of popularity for COBOL, hierarchical database models. In the beginning of the 1990s, many computer companies tried to sell expensive, complicated database products, which led to a highly competitive database market – as more businesses began releasing data-related tools and software to improve the systems data architecture.

## 2.7. ENTERPRISE DATA WAREHOUSE IN CLOUD VS ON-PREMISES

The cloud has become an integral part of the IT environment modernization, enabling the digital transformation of large and small companies. According to an article published by McKinsey (2018), cloud-based computing and storage platforms provide business with advantages over conventional on-premises systems, "from lower operating costs to better compatibility with the working styles of digital enterprises".

With the increasing number of data-driven organizations, the need for data warehouses that can handle the high number of users and high volumes of data has become a priority. The dilemma is which data warehouse model companies should choose: on-premises or cloud. The difference between these two models is that on-premises hardware, software and applications are on-site, which means that a business manages its own data center without a third-party (EM360 Tech, 2020). On the other hand, in the case of the cloud it's all performed off-side, and external entities are responsible for monitoring and maintaining a data center, with which the data warehouse implementations can be more agile when they are deployed on that type of environment (Golec et al., 2021). For example, like any other "as-a-service" solution, business leaders can add and/or remove features to cope with the changing needs of their organization (Kaur et al., 2012). But other parameters can be compared between the cloud and on-premises data warehouse implementation, such as deployment, cost, security, maintenance, and flexibility (Kurunji et al., 2012). In the case of on-premises, companies take ownership on maintaining and handling deployment processes whereas, in the cloud environment, resources are provided at the service provider's end and accessed by the public (Golec et al., 2021). At a cost level, the cost of operations is substantially higher in an on-premises environment compared to the cloud model. In terms of security, businesses are still skeptical about it in the cloud environment. For instance, government and healthcare organizations, or financial institutions, have highly sensitive information that they need to maintain secure (Shaikh et al., 2021; Alouffi et al., 2021). This way, companies need a dedicated cloud provider so that they can move from on-premises to the cloud in a secure way, storing and maintaining their data in the environment. Regarding maintenance, organizations take responsibility for maintaining their servers, software, data backup and storage devices if they are using an on-premises model. Compared to the cloud, organizations don't need to worry about regular maintenances, since they'll have automatic maintenance set up periods and then the solution provider itself takes full care of version upgrade on their own (Ying et al., 2009). Last, but not least, cloud environments offer more scalability and flexibility compared to the on-premises models, as changes at the infrastructure level or upscaling the server are more time-consuming in the on-premises environments (Agrawal et al., 2011).

## 2.8. From Enterprise Data Warehouse to Data Lakes

Enterprise data warehouses have long been the solution to provide decisional, actionable information for end-users. According to Madera & Laurent (2016), data warehouses are built based on indicators and analysis dimensions that must be pre-defined. Then, the data is collected through data integration processes and finally aggregated to deliver the pre-defined indicators. However, it's not always the case that such indicators are known prior to the data collection. As such, implementing a data warehouse without them (indicators) is not recommended. And that's when the data lakes come part of the equation in a data-related project. For instance, with the increasing use of the Internet of Things and smartphone applications by everyone, it produces an extremely large volume of data that is not automatically linked with information requirements. This means that the data itself is not yet known or sufficiently explored, at a first level, to define specific use cases – not being able to pre-define the mandatory indicators and analysis dimensions that data warehouses require. Therefore, to avoid a data swamp, which is nothing but an outcome of undefined, unstructured data sets from multiple sources (Data Eaze, 2019), and to add all the generated data into a specific information system, the data will not be integrated before knowing in advance what is going to be the specific use of it (Madera & Laurent, 2016). And this was the scope of the new data architecture step evolution: the Big Data.

According to Madera & Laurent (2016), a new concept appeared in 2014 from the Big Data and the Apache Hadoop waves: the data lake. Many vendors were stepping towards this concept to highlight all their Hadoop solution, without technically knowing what was really behind the term of data lake. Since the literature about this topic is still in early stages, there's not a specific definition for it, and that's what Madera & Laurent (2016) worked on. Furthermore, the authors covered the evolution of the decision support systems under the data governance's influence.

By default, the purpose of a Decision Support System is to collect relevant data sources from an organization and structure or process them in a way that is fundamentally useful and easier for the end-user to make decisions (Madera & Laurent, 2016).

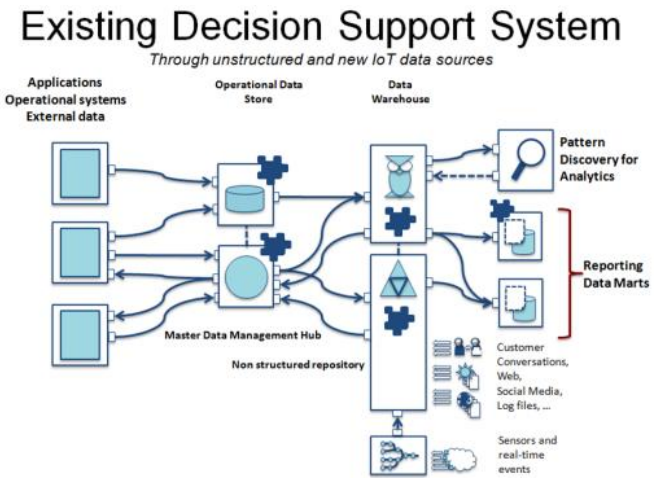



Figure 8 - Decision Support System's Architecture

(Madera, C., Laurent, A., 2016)

It was the very first information architecture evolution step, followed by a wave of technological tools such as ETL tools to build the DSS (acronym for Decision Support System), OLAP stacks to analyze data in a better way, and relational databases-dedicated analytics features. For the Decision Support System, the use cases or the information must be, as already mentioned, pre-defined in order to apply a scalable, structured and optimized design. Later in 2012, the Big Data wave was indeed the second information architecture evolution step, according to Madera and Laurent (2016). Challenges regarding the volume, the variety, the velocity, and the veracity of the data started to become important to deliver information. With the Big Data ecosystem and its challenges, a technology revolution emerged: the Hadoop technology. It is defined as an open-source framework that is used to efficiently store and process large datasets, capable of storing gigabytes to petabytes of data (Amazon, n.d.). Consequently, instead of using on large computer (such as a mainframe) to store and process the data, Hadoop allows clustering multiple computers connected to a common server to analyze massive datasets in parallel in a fast way. Considering this, the goal was not to replace Decision Support Systems with Hadoop, but to improve them and extend their scope by embedding big data technology. For example, with the increased data volume generated by the Internet of Things, big data technologies are needed to deliver information, together with privacy, compliance and security – which also play a big role. To ensure veracity when delivering information, data governance principles are fundamental, and that's why the next information architecture wave needs to integrate data governance into its core (Madera & Laurent, 2016). But that's not the only contributive, influential variable for the information architecture evolution. The end-users of a Decision Support System are increasingly requesting for help and solutions to find new insights, in order to take advantage of the available data and correlate with data governance principles. They don't know which type of

information they can retrieve from this data, so they essentially need a data repository to give them the new insight they don't have thought yet: the data lake. According to Madera and Laurent (2016), the data lake is defined as "a methodology to approach the raw data, structured and non-structured within an enterprise and seen as an evolution of existing data architecture". The data is loaded into a physical place to propose them for future insight, and stored in its native format, being possible to process any variety of it, regardless of size limits (Google, n.d.).

As any technology, there are underlying risks of data lakes that are important to mention:

i) The data quality;
ii) The data security;
iii) The access control risks.

Those are significant risks, and they are fully embraced into the data governance principles, which means that data lakes are more than decisional-only focus data repositories. In fact, they are more of a data governance concept (Madera & Laurent, 2016).

Now that the Decision Support Systems and data lakes are defined and their scopes are known, it's important to mention that the goal of the data lake methodology is to propose a new environment for the DSS themselves. Thus, the positioning between them is completely different. Whereas a Decision Support System, or Data Warehouse, follows a more traditional approach in the sense that it embeds practices related to relational databases, and structured data tasks, the data lake is more agile in the way that is associated with mixed types of data (volume and variety). This means, in short, that the DSS or Data Warehouses are more mature but more expensive at the same time, while the data lakes are cheaper but less mature (Madera & Laurent, 2016). As such, there's always a trade-off between these two types of data architectures, but they fundamentally differ on the use cases in the first place.

## 2.9. The Importance of BI Businesses and Marketing

It's of extremely importance to understand what's the value that the Business Intelligence area can generate to businesses and marketing teams, in addition to knowing the ETL evolution, Data Warehousing transformations and all the other technical information covered above.

Having said this, according to Vasarla (2021) Business Intelligence is an umbrella concept that covers different methods of collecting, storing, and analyzing the data from business operations. As such, BI experts can effectively find relevant business insights and help the decision-makers on making

better, data-driven decisions. As the human mind has a better performance at understanding pictures and visuals, compared to text and numbers, the data visualization plays a key role in empowering data interpretation – which ultimately increases know-how. The fact is that the topic of data visualization in an integral part of Business Intelligence and it helps business and marketing teams to easily digest the data and trigger actionable items on top of that (Vasarla, 2021).

The Business Intelligence importance is massive, and below are presented seven different advantages that business and marketing teams can benefit from:

i) Ability to gain customer insights
ii) Get actionable insights
iii) Real-time data availability
iv) Better marketing efforts
v) Higher competitive advantage

When it comes to gaining customer insights, marketing teams can better understand their customers by analyzing their profiles, buying patterns, calculate churn risk and even prevent what could be their next move. In addition, marketing teams can define a more robust customer segmentation due to higher customer knowledge, which can lead to better, differentiated customer experiences (Vasarla, 2021). Regarding the advantage number two ("Get actionable insights"), the best way to make a business decision is by using data as the foundation. However, using traditional static reports is not enough to extract actionable insights, but with Business Intelligence dashboards it is – because they are dynamic reports. In addition, Business Intelligence systems provides real-time data all the time (advantage: "Real-time data availability"), which allow businesses and marketing teams to stay informed and constantly monitor the performance and health of their strategies. What's more, BI enables marketing teams to define, create, and optimize marketing campaigns that drive ROI (Return on Investment) by providing them with a convenient way to access the data regarding past and current campaigns. On top of that, Business Intelligence can also provide marketing professionals the ability to control critical metrics such as Customer Acquisition Cost (CAC), Cost per Lead (CPL) and Click-Through Rates (CTR) of the campaigns themselves. Last but not least, BI gives businesses the ability to get higher competitive advantage by delivering deep competitive analysis and market trends.

## 2.10. DATA VISUALIZATION AND DASHBOARDS ASSESSMENT

Literature regarding data analytics has not been appropriately addressed empirical investigation of dashboards, as frameworks that measure dashboard dimensions and usefulness are yet to be

processed (Lawson-Body et al., 2022). The evaluation of dashboards is defined as the collection, analysis, reporting of data and measurement about their specific contexts. According to studies conducted by many authors regarding the measurement and evaluation of dashboards, their frameworks contained both internal and external factors that influence the decision-making of enterprises. However, there's a lack of consensus among the authors regarding the choice of these dashboards' evaluation factors, which means there are very different explanations about the content of the dashboard evaluation frameworks (Lawson-Body et al., 2022). On one hand, one group of dashboard assessment theorists says that dashboard measurements should have only quantitative and objective criteria to assess dashboard effectiveness in organizations. On the other hand, another group points out that, for measuring dashboards, it should be considered only qualitative and subjective criteria.

Lawson-Body et al. (2022), on his hand, developed a dashboard evaluation framework following Churchill's instrument development process (Figure 9) with the goal of developing measures for dashboard dimensions and usefulness, as well as creating a more solid assessment framework to help public organizations on the decision-making processes. As such, and methodologically, the authors collected 160 dashboards of public institutions to validate their measurement model using structural equation modeling and using Partial Least Squares techniques and the SPSS software to assess the model itself and its measures.

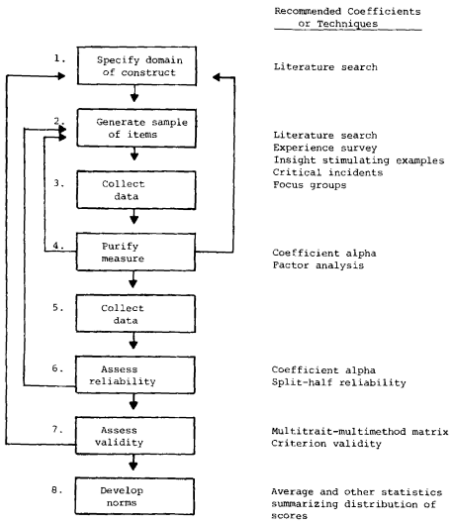



Figure 9 - Churchill's Instrument Development Process

(Churchill Jr, G. A., 1979)

To identify the measures that should be integrated into the dashboard evaluation framework, the authors did an extensive review of relevant literature and built a survey questionnaire. As a result,

the core measurement instruments were Complexity, Visual Design, Information Delivery, Alerting, and Perceived Usefulness. In addition, new items were developed from past literature to fundamentally measure interactivity: NLP and KPIs. To build the initial survey itself, the authors discussed with several expert academics and researchers in the field of data analytics and, in the end, a pretest was conducted also with expert academics who essentially provided information to minimize bias in the survey.

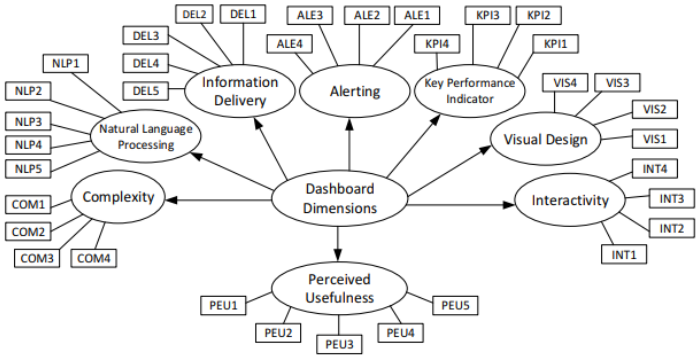



Figure 10 - Measurement model with dashboard dimensions and generated items

(Lawson-Body et al., 2022)

As a conclusion of this empirical study, the validation of the dashboard measurement instruments concludes that among the dimensions, some are more connected to the public sector than others (this study was targeting public institutions, as mentioned above). Therefore, one of this study's contributions was about identifying how important certain dashboard dimensions – NLP, Alerting, and Information Delivery – are for the public sector. The findings also tell us that researchers should carefully select indicators of dashboards usefulness, as the model validation scale really depends on where they're trying to validate (i.e., public sector vs. private sector). What's more, as the field of data analytics is still not mature, the authors also concluded that there was also a lack of reliable, validated measurement instruments related to dashboard usefulness, due to inadequate scales – and this was the reason why they resorted to expert academics and researchers in data analytics in the first place.

# 3. METHODOLOGY

To ensure the success of this internship, I had the possibility to use and leverage my existing Business Intelligence knowledge and techniques by making use of Power BI data visualization tool and running complex SQL queries in Snowflake, as well as performing ETL tasks in the back-end. More concretely, I was responsible for the integration of different data sources (the extraction phase), development of a relational model from scratch (the transformation/data engineering phase), and building BI dashboards (the load/front-end phase) so that the Customer Voice team could easily understand the data and extract valuable insights. In terms of integrations, Snowflake, Google Spreadsheets, Asana, Application Portfolio, and Customer Information System (these last two sources were internal systems used by the Company X to store product usage data and customer information) were the main data sources to develop this project. However, and with the direct help of my colleagues from the Data Engineering team, we ended up centralizing all the data from Google Spreadsheets, Asana, Application Portfolio and Customer Information System into Snowflake itself using APIs, creating the respective databases for querying purposes. This was a critical process to improve the overall performance of the Power BI solution itself, as the more different data sources this software gets the slower the performance. In that sense, by using Snowflake as the unique data source and interface to extract the data, the analytical engine running on Power BI was much faster.

For a tangible understanding, here's the nature of each database that was generated by Snowflake and that were used to develop this project:

- *Customer_Voice_Google_Spreadsheet* - stored information related to the customers that already had one or more case studies published (or any other types of marketing projects with the Company X) and the underlying use cases;
- *Application_Portfolio* – stored information about the Company X's product usage by the customers;
- *Customer_Information_System* – same as *Application_Portfolio*, but had more historical information
- *All_Accounts* – this was a different, yet very important, database that stored all the core customer-related information (from their name and location to ARR and Churn Risk). Living in Snowflake as well, but directly extracted from the Company X's CRM platform.
- *Asana* – a database generated from a project management tool that stored information of ongoing case studies initiatives running at the time.

After querying the data to extract the relevant information to the destination system (Power BI), the next stage was to transform and process the data to ensure its quality and consistency. That ended up being the most important and time-consuming step of this project. Changing data types, renaming and creating custom columns, replacing values and creating measures were some of the mandatory tasks to complete the transformation phase. Later on, when the previous step achieved a good level of maturity and scalability, the last stage was to load the processed data and create the relationships between the different tables so that the data warehouse could be deployed in a form of a relational model. Last, but not least, it was the time for the front-end phase: the Reporting. On that step, dynamic data visualizations were created in the form of BI dashboards to foster a good user experience and easy data interpretability for the end-users.

For a better understanding, there's a high-level representation below showing how this Business Intelligence project pipeline worked in the back-end.

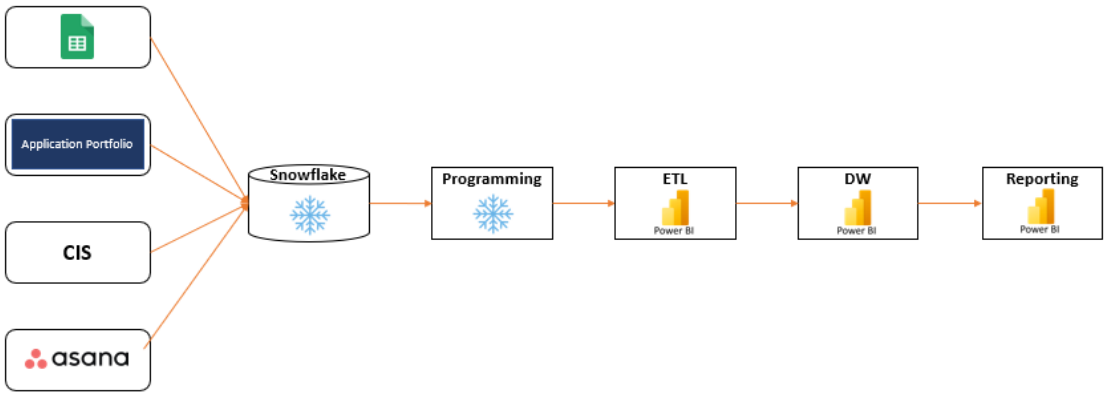



Figure 11 - Project's BI architecture

(Made by the author)

As for the data visualizations that were created in the end of the pipeline above (in the "Reporting" phase"), the goal was to promote a low cognitive cost for the Customer Voice team members when seeing them. That means the dashboards were fundamentally developed to foster an easy and intuitive interpretation so that the visuals could be turned into valuable and, most importantly, actionable items for future decision-making.

To assess the effectiveness of the dashboards themselves, I ran a survey to collect feedback from users. In this sense, I asked the Customer Voice team members nine questions about the data visualizations that I developed. It's important to note that the Data Engineering team played a key role in helping me define the best questions to ask. As soon as I collected their answers, I asked them to score the data plots from 1 to 10 according to the Visualization Wheel framework created by Alberto

Cairo (2011), which is a tool for evaluating tradeoffs in visualizations. The reason why I followed this methodology was that I wanted to have both qualitative (survey) and quantitative (Visualization Wheel assessment) insights to help me draw better conclusions. It's also important to mention that, while making the nine questions, I had no active role in biasing their answers.

Having said this, and in a practical way, I used Microsoft Excel to build the framework with the questions and the respective answers by each participant, so that I could register their insights for future analysis (note: the answers to the questions are placed in the "Appendix" section of this report).

| | Participants' answers | | | |
|---|---|---|---|---|
| **Question** | **Participant 1** | **Participant 2** | **Participant 3** | **Participant 4** |
| **1)** What is the first impression when looking at the data visualizations on this tab? | | | | |
| **2)** Are the data plots easy to understand? (complex vs simple) | | | | |
| **3)** Do you easily understand what are the metrics on each graph? | | | | |
| **4)** Can you easily extract insights from the dashboards? | | | | |
| **5)** Do the data visualizations answer your business needs? | | | | |
| **6)** With these graphs, how simple is it to tell a story from the data? | | | | |
| **7)** Do the data plots help you figure out what you need to improve on a business perspective? | | | | |
| **8)** Do these graphs help you be more data-driven to take decisions? | | | | |
| **9)** How easy is to play and navigate with the data visualizations? (i.e., filters, drill-downs, drill-throughs) | | | | |

Table 4 - Framework with the nine questions

(Made by the author)

The Visualization Wheel shown in the figure below is divided into two halves that represent a spectrum on which data visualizations may be placed. The top half represents visualizations that contain deep, complex data, whereas the bottom half represents plots that provide accessible, simpler data. However, it's important to mention that different audiences and professions prefer different types of visualizations. For instance, scientists and engineers are likely to prefer visuals that are dense, multidimensional, and that have high functionality, while artists, graphic designers, and journalists are likely to prefer visualizations that include decoration, lightness, and figuration (Wingate, 2019). This is a relevant point because even if the data visuals are complex and dense, it doesn't necessarily mean that the visuals are not fit-for-purpose – it can actually be the other way around, depending exclusively on the audience. In the case of this project, the Customer Voice team preferred data plots that include decoration, lightness, and figuration – to foster easy interpretability.


Figure 12 - Visualization Wheel

(Cairo, A., 2011)

Having said that, as soon as I collected all the scores from the team, I presented the results in a form of radar plots by each tab, in order to understand the variables with higher weight and classify them as more complex or more intelligible accordingly.

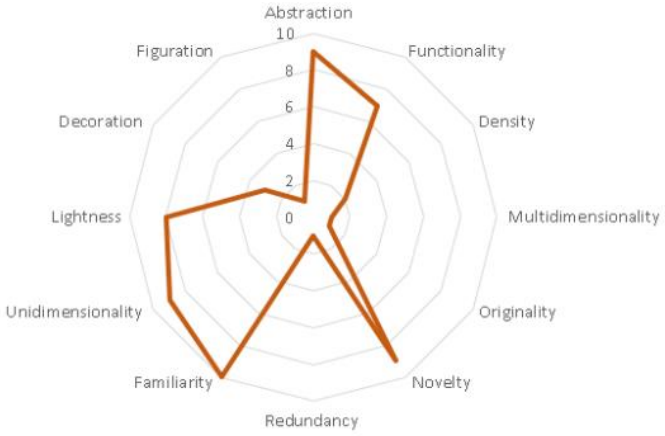



Figure 13 - Visualization Wheel radar plot example

(Made by the author)

In the end, the radar chart provided me with solid information about whether I achieved the goal of creating easy-to-understand, intuitive dashboards for the team or if there was a need to optimize them.

# 4. SOLUTION

Contextually, moving from spreadsheet-based analysis to dynamic, intuitive dashboards required the development of a robust data solution. When implementing it, it was of primary importance to only extract the most relevant data from the source systems (tables and attributes) to follow the business needs of tasks that needed to be performed during the transformation phase to increase the data's robustness and contextuality the Customer Voice team and to also avoid unnecessary consumption of computing resources (e.g., latency, memory capacity, CPU, etc.). Secondly, once the extraction phase was consolidated, transforming the data was the next task of the project, which turned out to be fundamental to increase the quality and consistency of the data coming from the source systems, as well as to format it into business-ready information. More precisely, replacing or removing blank values, changing data types, checking for duplicate records, and creating new customized columns were tasks that needed to be performed during the transformation phase to increase the robustness, and contextuality of the data. Without this step, all the data would not have any business meaning. As soon as this stage was completed, I built a data warehouse to load the transformed data in a contextual format, developing different dimensions and a fact table for that purpose. Finally, I managed to develop the data visualizations to increase data literacy levels for the Customer Voice team so they could make better, data-driven decisions.

Focusing on the details of the project, the first technical task was to import all the data directly from Snowflake to Power BI. However, as there was no direct connection to the Snowflake application in the Power BI itself, there was the need to install the ODBC driver from Snowflake and use the ODBC connector in Power BI. After the successful integration between both ODBCs, the data was ready to be extracted.

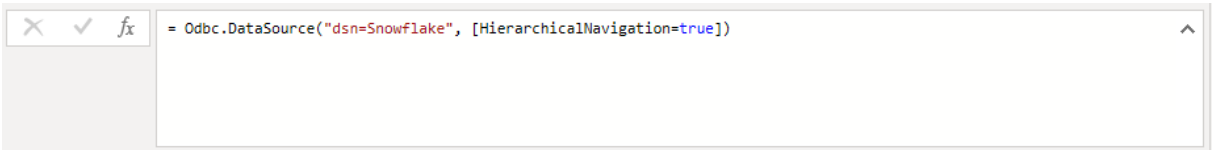



Figure 14 - Command for connecting Snowflake's ODBC to Power BI

(Made by the author)

Subsequently, as soon as the data was imported to Power BI, it was possible to check all the query dependencies, which illustrates how the queries are linked together inside Power BI. As shown in figure 12, there are 5 queries that are directly dependent on Snowflake:

i.) *Application Portfolio*;

ii.) *CIS*;

iii.) *Stories*;

iv.) *All Accounts*;

*v.)* *Stories Pipeline*

To enhance the analysis with other angles, business needs and measures, all the other tables were customized using one of the 5 tables above. The only exception was the "*Ambassadors and Use Cases*" table, which was created by joining specific attributes from both "*Use Cases*" and "*Ambassadors*" tables.

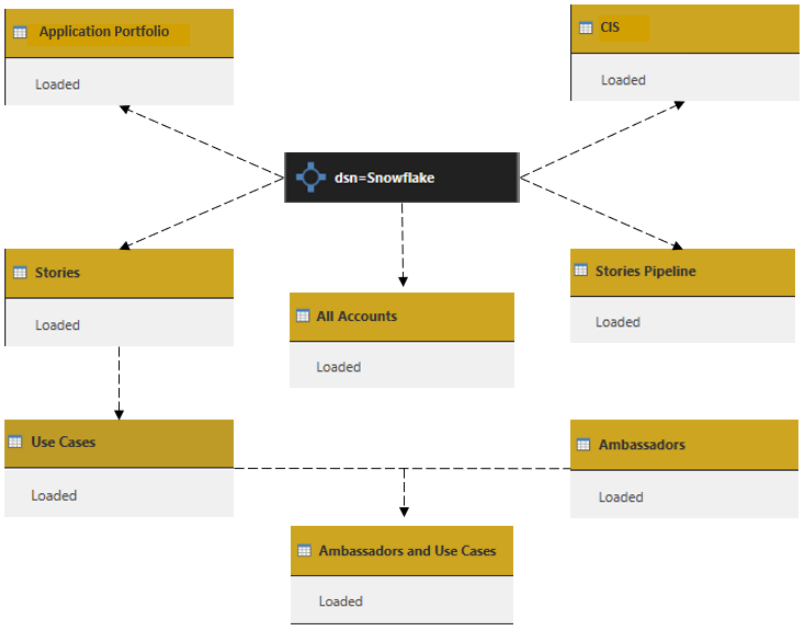



Figure 15 - Query Dependencies in Power BI

(Made by the author)

It's important to mention that some of the calculations, measures, tables and data plots were added throughout the development of the project with different workflows, hence having, for instance, some of the tables that were not directly imported from Snowflake itself but rather created from tables that were indeed imported from Snowflake.

Following the line of the project, every single table had a lot of data engineering tasks that were needed to process the data correctly and to add more information to the table (i.e., creating new columns). In addition, this step was crucial in the overall performance of the solution, optimizing the speed of data processing and scalability, being the most time-consuming part of the project. As there

were a wide variety of data transformations, below there's a pivot table that compiles what were the data tasks that were needed for the solution and the respective number of times I needed to perform that specific task - all before loading the data and building the data warehouse. In addition, and of extreme importance, I validated the quality and relevance of all the data engineering tasks with the Data Engineering team directly, which highly contributed to strengthening this phase of the project. As a result, the data turned out to be accurate, which means that the transformations generated business information for the Customer Voice team.

| Data Engineering Tasks | # Events |
|---|---|
| Replaced Values | 90 |
| Changed Data Type | 46 |
| Filtered Rows | 12 |
| Removed Columns | 10 |
| Renamed Columns | 9 |
| Concatenation | 3 |
| Promoted Headers | 3 |
| Custom Columns | 2 |
| Reordered Columns | 1 |
| Removed Duplicates | 1 |
| Removed Blank Values | 1 |
| **Grand Total** | **178** |

Figure 16 - Number of Events by Data Engineering Task

(Made by the author)

With the above tasks already deployed and automated in the back-end, building the data warehouse was the next step. The basis for the entire development of the DW was the business needs of the Customer Voice team, which means that the project followed a ground-up approach from the beginning. Regarding the model itself, there were five dimensions providing context and one fact table that contained both metrics and a few attributes.

- Dimensions: *Stories*, *Stories Pipeline*, *Use Cases*, *Ambassadors* and *Ambassadors* and *Use Cases* – storing attributes
- Fact table: *All Accounts* – storing measures or calculations (and a few attributes)

Every single dimension had a Primary Key (being the *Account_ID*), which directly matched with the Foreign Key of the fact table – also called *Account_ID*. This ensured that all customers within each dimension table were also stored in the fact table, which allowed for data integrity and that no information was lost during this process. As such, every single calculation worked for every customer involved in the analysis. In addition, it's also important to mention that the *Stories*, *Stories Pipeline* and

*Ambassadors* tables had a one-to-many relationship with the fact table, which means that, for instance, one customer (Fact Table: *All Accounts*) can more than one story (Dimension Table: *Stories*) and that one customer (Fact Table: *All Accounts*) can have more than one ambassador (Dimension Table: *Ambassadors*). This is a simple illustration of what the connections mean in the business case specifically.

However, and as seen in the data model below, some tables were not directly connected to the fact table itself, but to the dimension tables instead. On those entities, it was only possible to make many-to-many relationships, which means that multiple records in one table are associated with multiple records with another table. In this project particularly, many use cases (Dimension Table: *Use Cases*) can have multiple ambassadors (Dimension Table: *Ambassadors*) and vice-versa. Another business example could be that many stories (Dimension Table: *Stories*) can have multiple use cases (Dimension Table: *Use Cases*) and the other way around too. Even though Power BI was not recommending many-to-many relationships due to lack of scalability at the report filters level, it was the only way possible to connect these tables. What's more, and by default, the many-to-many relationships can bring data redundancy and updating difficulties, but the Data Engineering team validated this part of the model, and no impact was found in the final report itself.

Besides this, the *CIS* and the *Application Portfolio* tables were not loaded to the official data warehouse due to different natures. On the one hand, the *CIS* table was overlapping information with the *All Accounts* table, so it would bring data redundancy unnecessarily; on the other hand, throughout the development of the project, the Customer Voice team reported that they didn't need to have the information stored in the *Application Porfolio* table, making no sense to add it on the data model. On the one hand, it was good for the report itself in terms of performance and speed of the data engineering tasks, but on the other side, the model had to be changed a few times because of the constant iterations and changing needs of the Customer Voice team. But in the end, after collecting all the insights from that team and with the help of the Data Engineering team, the data warehouse was officially built and robust enough to start building the dashboards.

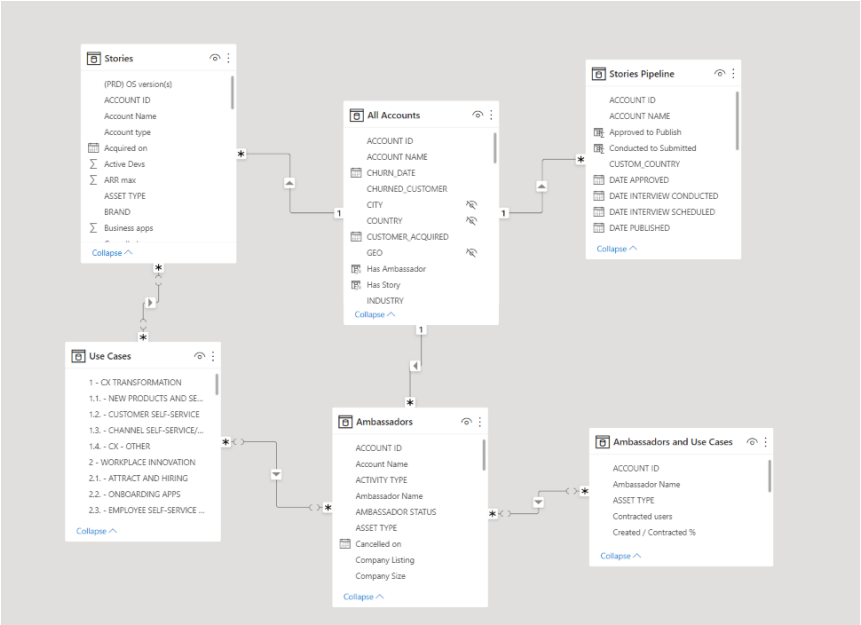



Figure 17 - Data Warehouse

(Made by the author)

It's important to reinforce that the origin of this entire project was the fact that the Customer Voice team wanted to have dynamic, intuitive data plots to foster better decision-making strategies and improve their own processes. This was critical for the development of the front-end part of the report because it guided me to create dashboards that had a low cognitive cost. This means that if there were hard-to-understand data visualizations, I would not fulfill their needs; it would probably create entropy, which would be a setback *per si*. Contrarily, another goal of this project was to develop greater data literacy for that team, to allow them to extract insights in seconds, rather than hours of data exploration with low actionable items. Therefore, the dashboards required a lot of iterations and validations before every deployment.

As their needs were this project's top priority, the first page of the report, called "*CV Team – Introduction*", was their company profile pictures with their names on it, creating the best user experience in their first exploration of the report. In addition, and a very important point, the report was not only published for the Customer Voice team to consume. There were another marketing teams that wanted to have access to it for exploration and usage as well. As such, having their pictures on the report's first page would allow those teams to easily reach out to a member of the Customer Voice team if needed. From the second tab onwards, all the dashboards were indeed about their business needs.

The report followed a general-to-particular approach so that the end-user could have a logical experience and data understanding while exploring the solution. Consequently, the second tab called "*CV Universe*" presented a more high-level vision, composed of the core information about Company X's customers and ambassadors. As such, the Customer Voice team could easily check how many customers they had by location (geography, country, region), as well as by industry, company size, tier (customer level), and journey stage (from "*Starting*" to "*Scaling*"). In addition, as the CV team's main operations were about creating and sharing customer assets in their marketing channels and generating more ambassadors for the company, they could see in a form of gauge plots the number of customers with those public assets as well as the number of ambassadors they had, versus the total number of customers. This provided them with a simple view of how far or close they were in terms of their goals.

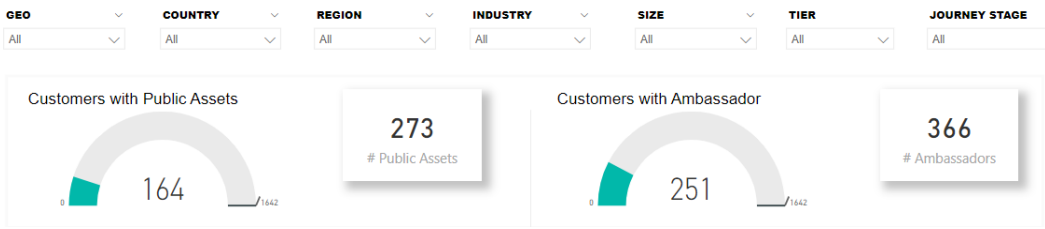
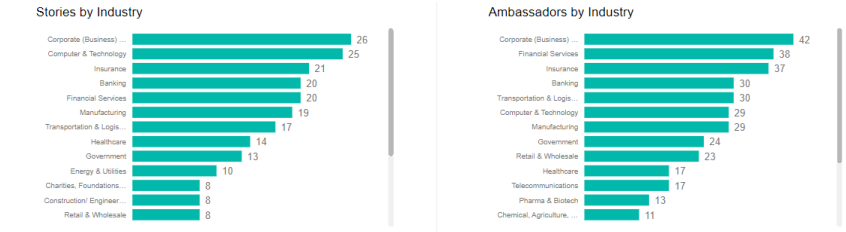



Figure 18 - Filters and gauge plots representations

(Made by the author)

What's more, they could also see how many customer stories and ambassadors they had by industry and what were the customers themselves and their characteristics, using the "drill-through" capability that I built. This ended up fostering high customer intelligence and helping them find strategic gaps and act on them immediately.




| CONTENT TYPE (core) | Account type | Is ISV / MSP | Industry | Journey Stage | Geo | Region | Country |
|---|---|---|---|---|---|---|---|
| Blog Post | Customer | False | Construction/ Engineering | Evolving | Americas | North America Central | United States |
| Short Written Case Study | Customer | False | Construction/ Engineering | Evolving | Americas | North America Central | United States |
| Written Case Study | Customer | False | Construction/ Engineering | Evolving | EMEA | Northern Europe | Netherlands |
| Blog Post | Customer | False | Construction/ Engineering | Optimizing | APAC | Southeast Asia | Singapore |
| Blog Post | Customer | False | Construction/ Engineering | Evolving | EMEA | United Kingdom | United Kingdom |
| Written Case Study | Customer | False | Construction/ Engineering | Evolving | EMEA | United Kingdom | United Kingdom |
| Written Case Study | Customer | False | Construction/ Engineering | Optimizing | APAC | Northeast Asia | Japan |
| Press Release | Customer | False | Construction/ Engineering | Optimizing | APAC | Northeast Asia | Japan |

Figure 19 - Number of stories and ambassadors by industry and customer drill-through

(Made by the author)

Related to the strategic gaps was the third tab of the report, called "*Case Studies – Main Coverage Matrix*". As mentioned above, since one of the CV's main activities was to publish customer stories - which is the same as Case Studies -, this tab was a cross-analysis involving the number of case studies published by use case. It's important to note that the use cases are called "*DT Priorities*" in this business context, which means "*Digital Transformation Priorities*". Company X as a whole was steering its path based on those priorities and, as such, it was crucial for the CV to have their analysis based on them too. By definition, there were four "*DT Priorities*":

1. *CX Transformation*;
2. *Workplace Innovation*;
3. *Process Automation*;
4. *Application Modernization*.

Each of them had their own Business Initiatives ("*Biz Initiatives*", they call), which were the specific use cases inside a specific DT Priority.

This worked as a two-level hierarchy, which means that, for instance, a case study falling in the "*CX Transformation*" DT Priority could be categorized as "*1.1. New Products and Services*" or "*1.2. Customer Self-Service*" as its respective "*Biz Initiative*" for that DT Priority. For a more tangible understanding, figure 16 shows a representation of the three DT Priority-based matrices.


Figure 20 - "DT Priorities" case studies' matrices

(Made by the author)

The matrices followed a heat map-like approach so that the Customer Voice team could interpret what were the less or more representative "DT Priorities" by location, strategic markets, and top industries. This would, again, help them find gaps and act or decide based on that data and its colors. For example, the CV team could create more customer stories working with APAC (Asia and Pacific) or German customers in the "Retail & Wholesale" industry.

Regarding the fourth tab, called "*Coverage Matrix for Ambassadors*", it was the same as the previous one, but this time counting the number of ambassadors by DT Priority and their underlying "*Initiatives*". The ultimate goal of this part of the report was to also help them identify where they had a greater or smaller number of ambassadors so they could decide where to invest based on that data. For example, they could generate more ambassadors from APAC customers, especially in Japan, coming from the "Education" industry.

It's this type of data analysis and insights extraction that this team couldn't previously do.

Ambassadors by DT Priorities, Biz Initiatives and Geo

| Use Cases | Americas | APAC | EMEA | Total |
|---|---|---|---|---|
| 1. CX TRANSFORMATION | 37 | 8 | 49 | 94 |
| 1.1. - New Products and Services | 21 | 1 | 20 | 42 |
| 1.2. - Customer Self-Service | 20 | 5 | 40 | 65 |
| 1.3. - Channel Self-Service/Operations | 4 | | 20 | 24 |
| 1.4. - CX - Other | 7 | 1 | 5 | 13 |
| 2. WORKPLACE INNOVATION | 14 | 7 | 11 | 32 |
| 2.1. - Attract and Hiring | 6 | 3 | 1 | 10 |
| 2.2. - Onboarding Apps | 2 | | | 2 |
| 2.3. - Employee Self-Service Apps | 5 | 4 | 11 | 20 |
| 2.4. - Workplace Innovation - Other | 5 | 3 | | 8 |
| 3. PROCESS AUTOMATION | 37 | 13 | 62 | 112 |
| 3.1. - Complex Case Management | 14 | 2 | 23 | 39 |
| 3.2. - Field and Customer Service Operations | 7 | 7 | 13 | 27 |
| 3.3. - Supply Chain Transparency and Efficiency | 5 | 1 | 6 | 12 |
| 3.4. - Workflow Automation | 16 | 12 | 24 | 52 |
| 3.5. - Process Automation - Other | 2 | | 10 | 12 |
| 4. APPLICATION MODERNIZATION | 30 | 8 | 50 | 88 |
| 4.1. - COTS Replacement | 10 | | 11 | 21 |
| 4.1.1. - Oracle Siebel CRM | | | 5 | 5 |
| 4.1.1. - Other | 10 | | 6 | 16 |
| 4.2. - EOL SW Replacement | 8 | 2 | 20 | 30 |
| 4.2.1. - Adobe Flex | 2 | | | 2 |
| 4.2.1. - Java Platform | 1 | 1 | 1 | 3 |
| 4.2.1. - Lotus Notes | 3 | | 5 | 8 |
| Total | 58 | 15 | 83 | 156 |

Ambassadors by DT Priorities, Biz Initiatives and Strategic Markets

| Use Cases | Germany | Japan | United Kingdom | United States | Total |
|---|---|---|---|---|---|
| 1. CX TRANSFORMATION | | | 14 | 32 | 46 |
| 1.1. - New Products and Services | | | 8 | 19 | 27 |
| 1.2. - Customer Self-Service | | | 9 | 13 | 22 |
| 1.3. - Channel Self-Service/Operations | | | | 4 | 4 |
| 1.4. - CX - Other | | | 2 | 7 | 9 |
| 2. WORKPLACE INNOVATION | 1 | | 4 | 11 | 16 |
| 2.1. - Attract and Hiring | | | | 6 | 6 |
| 2.2. - Onboarding Apps | | | | 2 | 2 |
| 2.3. - Employee Self-Service Apps | 1 | | 4 | 5 | 10 |
| 2.4. - Workplace Innovation - Other | | | | 2 | 2 |
| 3. PROCESS AUTOMATION | | 3 | 14 | 30 | 47 |
| 3.1. - Complex Case Management | | | 8 | 14 | 22 |
| 3.2. - Field and Customer Service Operations | | 2 | 1 | 4 | 7 |
| 3.3. - Supply Chain Transparency and Efficiency | | 1 | 2 | 4 | 7 |
| 3.4. - Workflow Automation | | 3 | 3 | 13 | 19 |
| 3.5. - Process Automation - Other | | | 3 | 2 | 5 |
| 4. APPLICATION MODERNIZATION | 1 | 3 | 14 | 21 | 39 |
| 4.1. - COTS Replacement | | | 6 | 10 | 16 |
| 4.1.1. - Other | | | 6 | 10 | 16 |
| 4.2. - EOL SW Replacement | 1 | 1 | 8 | 8 | 18 |
| 4.2.1. - Adobe Flex | | | | 2 | 2 |
| 4.2.1. - Java Platform | | 1 | 1 | 1 | 3 |
| 4.2.1. - Lotus Notes | | | | 3 | 3 |
| 4.2.1. - Microsoft Sharepoint | 1 | | 3 | | 4 |
| Total | 1 | 3 | 22 | 42 | 68 |

Ambassadors by DT Priorities, Biz Initiatives and Top Industries

| Use Cases | Banking | Education | Financial Services | Government | Healthcare | Insurance | Retail & Wholesale | Total |
|---|---|---|---|---|---|---|---|---|
| 1. CX TRANSFORMATION | 17 | 1 | 7 | 5 | 7 | 13 | 2 | 52 |
| 1.1. - New Products and Services | 10 | | 5 | | 5 | 5 | | 25 |
| 1.2. - Customer Self-Service | 9 | 2 | 6 | 4 | 6 | 12 | 3 | 42 |
| 1.3. - Channel Self-Service/Operations | 2 | | | | 1 | 12 | | 15 |
| 1.4. - CX - Other | | | 1 | 2 | | | 1 | 4 |
| 2. WORKPLACE INNOVATION | | 1 | 1 | 2 | 3 | | | 7 |
| 2.1. - Attract and Hiring | | 1 | | | 1 | | | 2 |
| 2.2. - Onboarding Apps | | | | 2 | | | | 2 |
| 2.3. - Employee Self-Service Apps | | | | | 3 | | | 3 |
| 2.4. - Workplace Innovation - Other | | | 1 | | | | | 1 |
| 3. PROCESS AUTOMATION | 13 | 2 | 7 | 8 | 5 | 14 | 4 | 53 |
| 3.1. - Complex Case Management | 13 | | 6 | 4 | | 8 | | 31 |
| 3.2. - Field and Customer Service Operations | | | | | 1 | | 2 | 3 |
| 3.4. - Workflow Automation | 4 | 2 | 2 | 4 | 4 | 2 | 1 | 19 |
| 3.5. - Process Automation - Other | | | 1 | 1 | 1 | 5 | 1 | 9 |
| 4. APPLICATION MODERNIZATION | 14 | | 5 | 5 | 1 | 13 | 2 | 40 |
| 4.1. - COTS Replacement | 8 | | 3 | | | 5 | | 16 |
| 4.1.1. - Oracle Siebel CRM | | | | | | 5 | | 5 |
| 4.1.1. - Other | 8 | | 3 | | | | | 11 |
| 4.2. - EOL SW Replacement | 3 | | 2 | 5 | 1 | 6 | 1 | 18 |
| 4.2.1. - Java Platform | | | 1 | | | 1 | | 2 |
| Total | 18 | 3 | 11 | 10 | 9 | 15 | 7 | 73 |

Figure 21 - DT Priorities ambassadors' matrices

(Made by the author)

The fifth and sixth tabs were more operational, as opposed to the previous ones – which were more strategic-oriented. Having said this, the fifth tab was called "*Customer Voice Assets Behavior*". This section was about measuring the teams' performance from 2014 to 2022 in terms of how many

assets they published, drilling down that information by location (hierarchy of *Geography*: "*Geo*", "*Region*" and "*Country*") and *Architecture Type* (Cloud, On-Premises, Hybrid, Azure Cloud, and Amazon Cloud). The data visualizations provided the team with a simple, concise view of how their entire work evolved over time, which was key for deciding whether they wanted to change processes in the future, keep their best-practices, or even both.

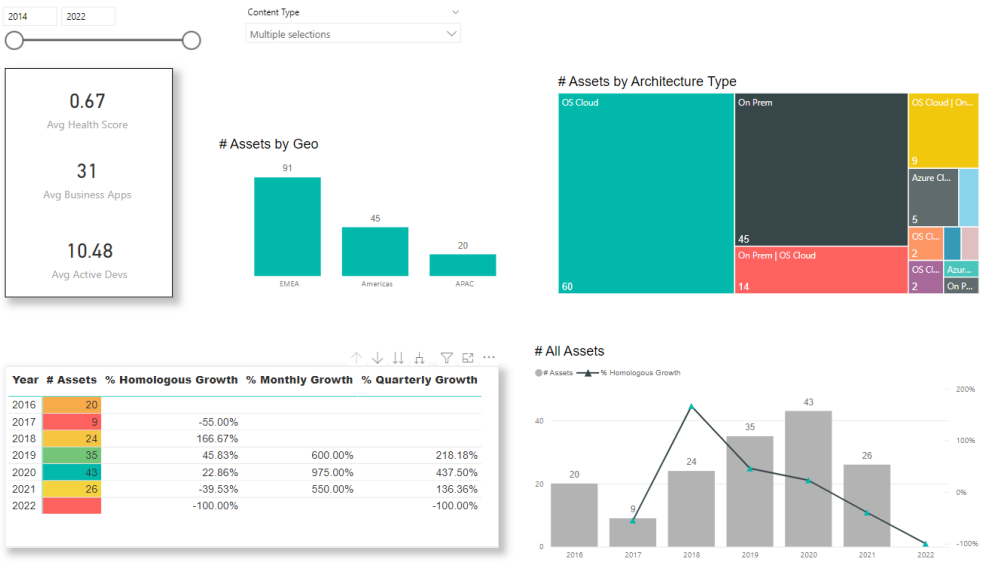



Figure 22 - "Customer Voice Assets Behavior" tab

(Made by the author)

For instance, if they used the drill-down capability on the "Date" column from the *"# All Assets"* plot, they could check that the quarter that was more successful in terms of published assets was the Quarter 1 – with January being the most representative month, respectively. At an operational level, this simple, yet important, insight could raise a lot of questions for the team, such as:

- "Why do we have a bigger number of assets published in January?"
- "Why is the distribution of public assets 2x to 3x times higher in Quarter 1 compared with the other Quarters?"
- "How can we have approximately the same number of assets between all Quarters and Months?"
- "In terms of locations, where should we invest more?"

Answering all these questions could be enough for the Customer Voice team to start defining ways of improving operational efficiency and start scaling their marketing activities in a more balanced way.

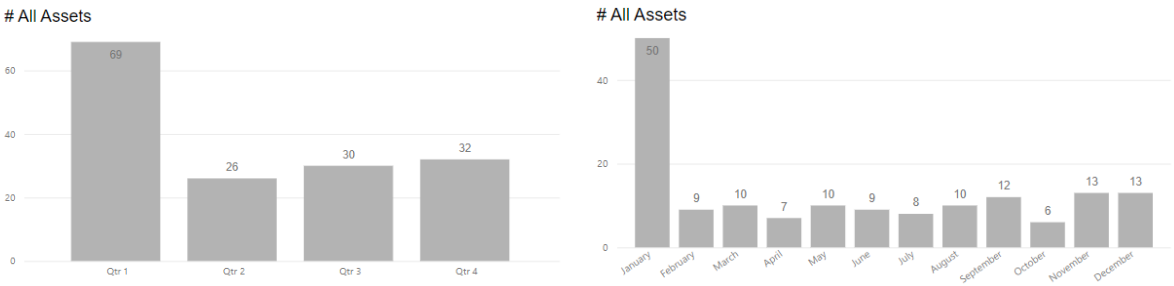



Figure 23 - "# All assets" by Quarter and Month

(Made by the author)

The sixth tab, called "*Stories Timeline*", also had an operational-centered analysis. But as opposed to the fifth section, tab number six was about measuring marketing initiatives that were on their pipeline, instead of published assets. As such, these were projects that were ongoing, and that the CV was working on. Using in the "*Stories Pipeline*" table for the analysis of this section, this was an entity that had an API integration between Asana (the project management tool the CV was using) and Snowflake – which, afterward, the Data Engineering team integrated it as a Snowflake table. This means that the "*Stories Pipeline*" in the Power BI was an aggregated data view of the Asana application itself, which was something the CV team wanted for years. Their goal was to obtain a simplified view of how their projects were behaving by stages and how to improve them over time. In their specific business context, their marketing initiatives had 7 stages by default:

1) Qualifying;
2) Scheduling;
3) Scheduled;
4) Conducted;
5) Submitted for Approval;
6) Official Approved;
7) Published.

As the team needed to know the transitioning time between the different stages of their projects, I managed to create the average time (in days) from one stage to another, using the *DATEIFF* Power BI function for the purpose.

```
Scheduled to Conducted = DATEDIFF('Stories Pipeline'[DATE INTERVIEW SCHEDULED], 'Stories Pipeline'[DATE INTERVIEW CONDUCTED], DAY)
Qualy to Scheduling = DATEDIFF('Stories Pipeline'[DATE QUALIFIED], 'Stories Pipeline'[DATE INTERVIEW SCHEDULED], DAY)
Conducted to Submitted = DATEDIFF('Stories Pipeline'[DATE INTERVIEW CONDUCTED], 'Stories Pipeline'[DATE SUBMITTED FOR APPROVAL], DAY)
Submitted to Official Approval = DATEDIFF('Stories Pipeline'[DATE SUBMITTED FOR APPROVAL], 'Stories Pipeline'[DATE APPROVED], DAY)
Approved to Publish = DATEDIFF('Stories Pipeline'[DATE APPROVED], 'Stories Pipeline'[DATE PUBLISHED], DAY)
```

Figure 24 - DATEIFF function for every stage transition

(Made by the author)

After running the calculations above, I decided to create cards as the data visualizations and order them by stage transition, following the logic and workflow of how their projects worked. By having that perspective of the data, the CV team could immediately understand that the most time-consuming stage transition was the "*Submitted to Approval*". In their language and context, this means that from the date they submitted their marketing initiative for the customer to validate and the customer's official approval date itself, it was taking 76 days on average, which is equal to 2,5 months. Just this insight by itself could trigger a lot of questions to the CV team, such as:

- "Is the marketing initiative too dense for the customer to validate?";
- "Should we invest in lighter formats to accelerate the time between the submission and the customer's official approval?";
- "Are we communicating well when submitting the marketing initiative for the customer?";
- "Are we using the right marketing channel to submit the initiative?".

These are the types of exercises that were not possible to do on spreadsheet-based analytics and that were of critical importance for the overall decision-making processes and strategies of the team, as well as the knowledge base they could create of their own department.

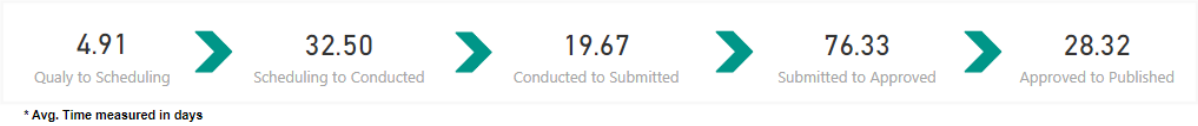



Figure 25 - Average time between stages

(Made by the author)

Besides having a high-level view of how their marketing initiatives were behaving, I managed to add another layer of information, but this time more specific. Having said this, I built a stacked bar chart based on the *DATEIFF* function and used it for the projects that were running at the time, so they could see which customers were in the pipeline from a project's marketing perspective. Therefore, by

having a customer view, the Customer Voice team could also act fast and be customer-targeted, as well as prioritize their projects based on the different stages' behavior by customer.

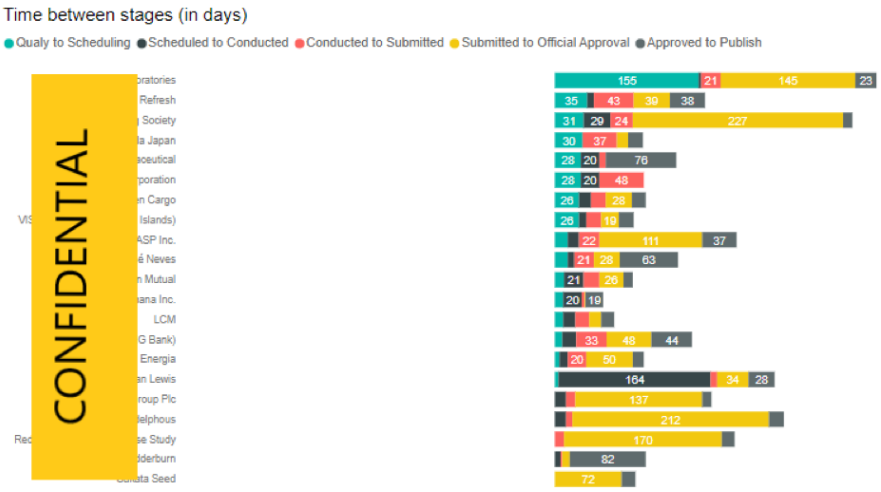



Figure 26 - Chart with the time between stages (in days) by customer

(Made by the author)

The seventh, and penultimate, tab was related to Asana as well – the project management application they were using. However, instead of measuring the CV's operational performance by stage and by customer, this tab followed the same architecture as the third tab ("*Case Studies – Main Coverage Matrix*"). The only difference was that the seventh section, called "*Stories Pipeline – Coverage Matrix*", was quantifying the number of stories (or marketing initiatives) in the pipeline itself, which means assets that were not published yet, but headed in that direction. As this type of analysis, again, was not available through excel/spreadsheet, the Power BI solution helped them exponentially in this angle.

The CV team's need was to have an aggregating view of the number of stories in the pipeline by stage, use cases and location. The purpose behind this was for them to know what was on their radar in terms of strategic assets, what other initiatives they could add, and what were the use cases and location distribution of those same initiatives – always based on the data.

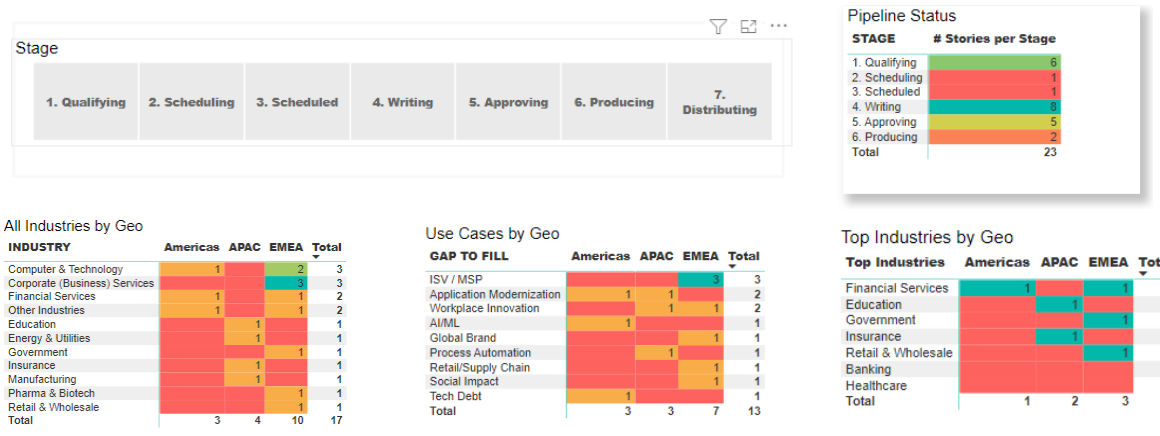



Figure 27 - “Stories Pipeline – Coverage Matrix” tab

(Made by the author)

For instance, based on the data charts above, the team had only three stories/case studies in “Americas” and four in “APAC”, as opposed to “Europe” - which had 10. Maybe it could make sense to balance this distribution by location so that the marketing efforts could be equally distributed. Another insight is that, at that time, the CV team had one story by each “Top Industries” (excluding *Banking* and *Healthcare*), which means that they were working towards the industries that add more value to them. Last, but not least, they only had 2 stories in the “Producing” stage, so it might make sense to keep pushing the entire pipeline to feed that stage. By using the filters, this tab can become even more dynamic in the sense that the team can check the different cross angles represented by the tables for a specific stage.

Finally, the last tab – called “*Asana – Stages Behavior*” -, represented a historical view of all the stories developed by the Customer Voice team until the present. For them, it was not only important to know the present (where they are) but also where they were in the past to learn in the future. Even though this tab might sound like the previous two tabs (because they all retrieve information from the Asana application), they do have very different angles. This seventh section helped the team understand their overall performance by year, quarter, and month, during the last 4 years. By default, historical information is the basis to identify patterns and behaviors, and that’s what this tab was all about. What’s more, the CV team never handled this type of information before, which means they were working, brainstorming, deciding, and deploying marketing initiatives (mainly customer stories) without any previous knowledge of how they behaved in the past. Even though this section had only one type of data visualization to perform the analysis, the “Decomposition Tree” from Power BI, the Customer Voice team approved it with distinction.

Partitioning the data visualizations by stage, which goes all the way from "*Qualified*" to "*Submitted*", was the best way to monitor their performance. Besides this, including the three-level "Date" hierarchy as a column to every chart, allowed for a better visibility and comparison between the different timings.


Figure 28 - "Asana – Stages Behavior" tab

(Made by the author)

By looking at the above data plot the team could see that, in the "Qualifying" stage, for instance, there were more stories that entered the pipeline in 2020 versus 2021 – hitting a decrease of 51%. As this stage marks new entries/marketing initiatives in the pipeline, this means that they had a higher number of marketing initiatives in 2020 compared to 2021. In addition, and jumping into the "*Approved*" and "*Published*" stages, they could also check that the number of approved marketing initiatives decreased from 39 to 23 (-41%) and from 51 to 27 (-47%), respectively. Actually, their overall performance by stage decreased from 2020 to 2021.

In short, these are examples of important data extractions and insights that the CV team should consider if they want to improve their operational efficiency and build a future-ready marketing roadmap to see growth in the numbers.

# 5. RESULTS AND DISCUSSION

After deployment, it was important to measure the effectiveness of the developed dashboards. From the beginning, the goal was about creating easy-to-understand, intuitive dashboards supported by meaningful data visualizations. As such, I ran a survey of nine questions to obtain the users' qualitative insights about the data visualizations, and I also asked for their quantitative assessment (ranging from 1 to 10) for each tab of the report, using Alberto Cairo's Visualization Wheel – a powerful tool for thinking about, and assessing tradeoffs in visualization.

Before presenting the results themselves, it's important to mention that this tool consisted of two halves that represent a spectrum on which data visualizations may be placed. The first half is about data visualizations that contain deep, and complex data. On the other hand, the second half of the wheel represents visuals that provide accessible, but shallower, data. Having said this, I collected their assessments by each tab of the report using a scale from 0-10 for every attribute of the Visualization Wheel, presenting the results in the form of radar plots to find the differences and tradeoffs.

Regarding the second tab of the solution called "*CV Universe*" (the first tab had no evaluation because it only had the profile pictures of the CV team members), the team evaluated it as complex and deeper, as well as simple and easy-to-understand. The majority assessed the charts as light, which means the visuals have less information but it's easier to get to the points, but in general, all the data charts are functional, which is a variable that belongs to the first half of the Wheel ("More complex and deeper"). In addition, they gave a score of 7 in "Novelty" and 3 in "Redundancy", which means that the visuals describe each phenomenon in only one way, rather than graphics that use multiple modalities to tell the same story. Additionally, they also assessed the plots as more abstractive, which represents more conceptual, and less real representations of the specific phenomenon they're analyzing. Last but not least, the majority of the team members said that the data visualizations of this tab are easy-to-understand, and a great way to understand how they are doing from an operational point of view, even though some of them think there's a little bit of complexity due to the unfamiliarity with the visuals: "*It looks really straightforward to read and easy to digest (…)*", "*(…) providing a simple, centralized way of what we do as a team (…)*" – Participant 1 and Participant 3; "*The dashboards on this page are user-friendly, even though there are some plots I'm not familiar with (…)".* – Participant 4.

With these results, it's possible to conclude that the visualizations covered the two goals of the solution: on the one hand, they assessed the visuals as simple and intelligible (covering the "easy-to-

understand data visualizations" goal) and, on the other hand, they also assessed the visuals as complex and deeper (covering the "enriching insights by building sophisticated dashboards" goal).

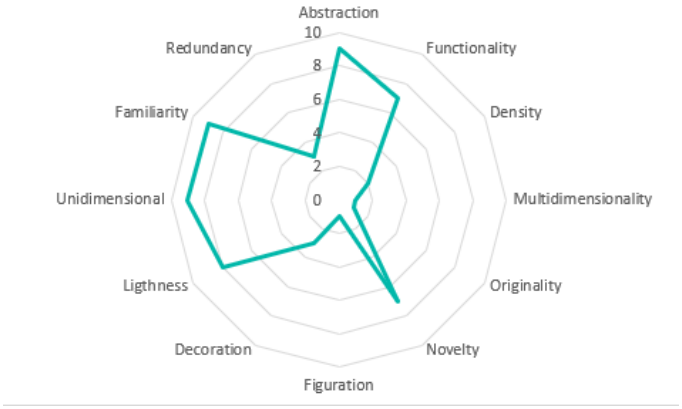



Figure 29 - Radar plot assessment for the "CV Universe" tab

(Made by the author)

As for tabs number 3 and number 4, called "*Case Studies – Main Coverage Matrix*" and "*Coverage Matrix for Ambassadors*" respectively, I collected the participants' answers in the same Excel tab because both these Power BI tabs shared the same data visualizations. With this in consideration, the Customer Voice members shared very similar opinions about the data visualizations – complex and time-consuming to extract insights. What's more, when I asked them if the data plots help them figure out what they need to improve from a business perspective, they all said they needed to dig deep in the data to find patterns, as there were hierarchies in the tables to show the different data points according to a specific filter value. For instance, the Participant 2 said that "(…) the data plots require a dig deep on the data to allow us to extract insights" and the Participant 4 also corroborated by mentioning that by having other levels in the data, they needed to "spend more time digging the data vs having a one-shot analysis and define action points". This clearly indicates that the plots helped them understand what they needed to improve from a business, and marketing perspectives, but it required more focus and time to see valuable findings out of the visuals.

Quantitatively, the team scored the Visualization Wheel's attributes in the exact same way because both sections followed the same style, layout, and information architecture. Having said this, the Customer Voice team clearly evaluated them as more complex and deeper, meaning that the data visualizations contained a higher cognitive cost. For example, the fact that they assessed the visuals as more multidimensional and less unidimensional illustrates that the dashboards were representing many different aspects and angles of analysis. It is an advantage if they want to be more precise in analyzing specific patterns or behaviors in the charts, but it can also become more time-consuming

and complicated. What's more, they evaluated the charts as more original and less familiar, which means they dealt with graphics that they were not used to, challenging their visualization patterns. Last, but not least, they gave a higher score on "Functionality" and less on "Decoration", which illustrates that the visuals were closer to a direct representation of the data.

With these results, I could have used other charts to foster better and faster interpretability for the CV team, balancing the complex half of the Visualization Wheel with the shallower, lighter half. However, another important insight to take from this output is that, even though they have classified this tab as more complex and deeper, they still gave an extremely high score in "Functionality" compared to "Decoration" - which fundamentally represents that the plots were indeed a close representation of the data.

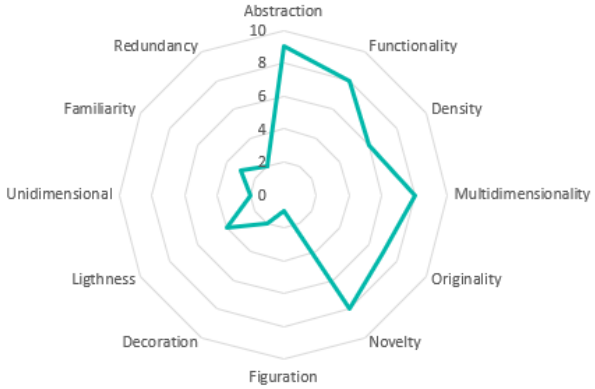



Figure 30 - Radar plot assessment for the "Case Studies – Main Coverage Matrix" and "Coverage Matrix for Ambassadors" tabs

(Made by the author)

As for the fifth tab, called "*Customer Voice Assets Behavior*", the participants' inputs were practically the same when answering the nine questions. One of the most important milestones from the data visualizations inside this section was that the great majority of them answered the team's business needs, which means they could see the business value out of the way the data was displayed. As Participant 1 mentioned, "the great majority of the data visualizations answer our business needs, which is very important for us (…)", in addition to Participant 3 that mentioned that "having a filter with the years as well as data plots with a date hierarchy turns everything very intuitive to find patterns of behavior (…), and this, per si, is very valuable from a business perspective, so the plots definitely answer our business needs and questions".

As for the CV team's quantitative assessment regarding this tab, the results showed a good balance between the complex, deeper part of the Visualization Wheel and the more intelligible and shallower

part. The data visualizations developed in this section were already known by the members of the team, due to the high score on “Familiarity” (which is represented by broadly understood visuals), but at the same time, the charts were dense in terms of information. This happened because most of the data plots had drill-down capabilities to allow them to navigate deeper into the data and get more insights. Consequently, the team also gave a higher score on “Multidimensionality” versus “Unidimensionality” (6 to 4, respectively) because the visuals showed many different angles of a specific phenomenon, covering it from end-to-end. As for the “Novelty” and “Redundancy”, they also scored 6 to 4, respectively, meaning that there were some charts describing each phenomenon in only one way and other charts using multiple modalities to share the same inputs.

In short, and resuming the outcomes of this tab, the two original goals were also achieved. The balance between having to navigate deeper into the data to get useful insights and consuming easy-to-understand visualizations to get faster data extractions was reached successfully with the development and implementation of different, but complementary, dashboards.

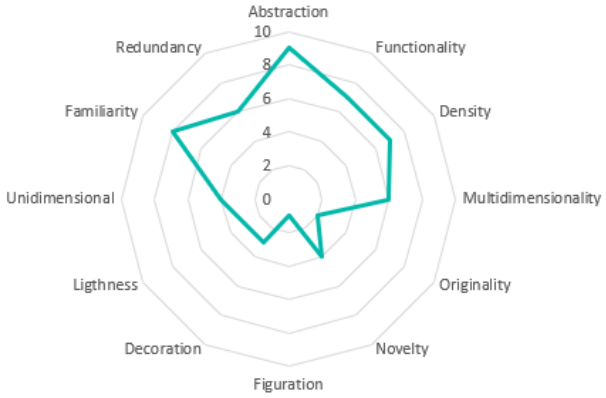



Figure 31 - Radar plot assessment for the “Customer Voice Assets Behavior” tab

(Made by the author)

Regarding the sixth section, called “*Stories Timeline*”, the balance between the complex part of the Visualization Wheel and the simpler, cleaner one, happened again. As for their qualitative assessment, they found the data plots very easy to understand and digest, also referring that the page itself was very clean or not dense at all, helping them on the data interpretation side. In addition, they mentioned that the visuals used were already familiar to them, assessing the visuals as “simple charts that trigger fast insights and, consequently, reduce the time to value” (inputs from Participant 1).

In the quantitative classification itself, the CV team clearly rated the charts as being familiar, unidimensional, and light (the simple half) as well as abstracted, functional, and novel (the complex half). With that being said, the great majority of the visualizations were already known by the team,

facilitating data interpretations *a priori*, they illustrated each phenomenon in a single, concise way, avoiding telling the same story in different modalities, and they were fast and easy to understand. In addition, and as per the team's assessment, the visuals had little embellishments, which means that they "met" with the data in a more straightforward way, and more conceptual, representing the phenomenon as it is. What's more, the average score of the complex, deeper attributes of the Visualization Wheel and the average of the second half's attributes were the same (average score = 5), showing the full balance between the two halves.

To conclude these results, it happened the same as in the previous assessment – both goals were accomplished successfully. This means that the team could extract immediate insights through the consumption of simple, high-level data visualizations and, at the same time, navigate to other charts to get deeper information and identify patterns. Therefore, the combination of complex and simple visualizations gave them the foundation to improve the team, strategically and operationally.

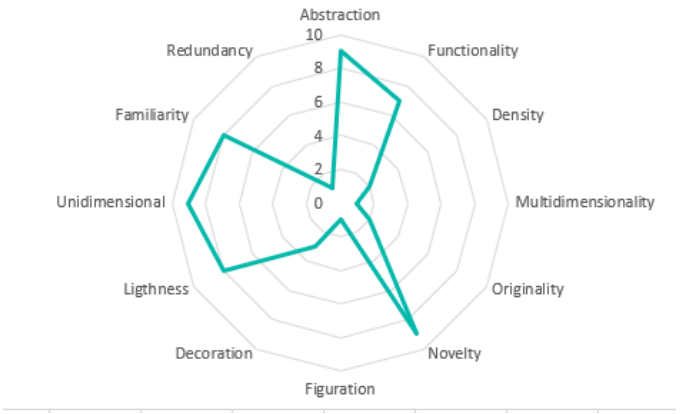



Figure 32 - Radar plot assessment for the "Stories Timeline" tab

(Made by the author)

The seventh and penultimate tab, called "*Stories Pipeline – Coverage Matrix*", had approximately the same results as tabs number 2 and 3. When asking the nine questions about the data visualizations, the team members of the Customer Voice team agreed that the information inside the visuals could easily trigger insights and better define next steps. As Participant 2 mentioned, "it is very easy to extract insights and define action points on top of them. And that is very positive for us at an operational level because it allows us to identify where we are investing more of our resources and especially where should we invest in the future". What's more, and not less important, Participant 3 mentioned that the tables in the section were giving the team a simple view of how their marketing pipeline was working, allowing them to act fast and in an agile way – which were things they just couldn't do before due to the lack of data analysis.

As the data visualizations were exclusively pivot tables using a heatmap style, the Customer Voice team quantitatively assessed this section as shallow and simple. Even though some of the visuals had drill-down capabilities, the data points were extremely easy-to-understand, which trigger fast insights instead of having to navigate deeper to find other data extractions. But as the Visualization Wheel has its tradeoffs by default, the disadvantage of this tab being so simple to understand was that the CV team was not able to dig deeper into the data and find other inputs for their strategies and operations. As such, this section proved to be a good asset for fast analysis and easy interpretability, consuming enough data to allow them to decide what to do next. As for the future, it would make sense to also add more complex data visualizations to enhance the data extractions, balancing the two halves of the Visualization Wheel.

With these results, one of the goals was accomplished – creation of easy-to-understand, low cognitive cost dashboards.

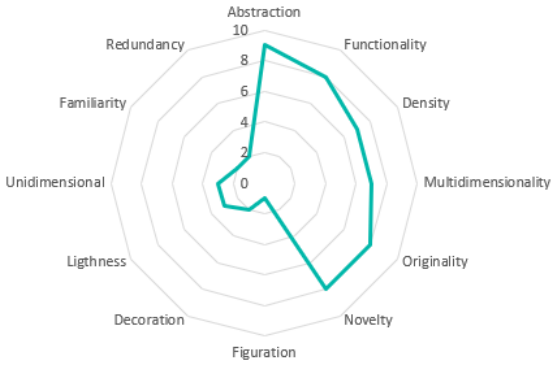



Figure 33 - Stories Pipeline – Coverage Matrix

(Made by the author)

Last, but not least, the eighth section. Named "*Asana - Stages Behavior*", this was the tab that had the most different data visualizations - the Decomposition Trees. As such it is no surprise that the participants mentioned that the style of that graph was "complex to understand", "not familiar" and "hard to extract quick insights". However, they very much liked the aspect of it, mentioning that the visuals were "very appellative" and even "intuitive to use". Consequently, from these inputs is possible to infer that, in terms of content, the Decomposition Trees are dense data charts - requiring deeper navigation to understand the data patterns and extract relevant business insights -, but at the same time they are good-looking plots.

In what concerns the Customer Voice's team quantitative classification of the charts, it's notorious that they rated them as complex and deeper. There were a lot of data points to be consumed – scoring

high on “Density” – and the visuals showed many different angles of the same phenomenon – scoring also high on “Multidimensionality”. What’s more, they evaluated the dashboards as originals, because they were not used to seeing those types of visuals (the Decomposition Trees). This means that, a priori, there was probably friction in understanding what the data visualizations were all about. In addition, the team rated high on “Novelty”, which means the graphics describe each phenomenon in only one way, being a positive result. Finally, the visuals contained more artistic embellishments than functionality, as they gave a score of 7 out of 10 in the “Decoration” attribute. Even though I thought that the Decomposition Trees implemented in the Power BI were functional enough for them to get the main insights of that tab, it was curious because they classified them the other way around. They found it difficult to understand, time-consuming, and too dense.

Having said this, it’s no surprise that this section was assessed as complex. It could be good for the identification of behaviors using the historical information available in the charts (possibility of navigating deep in the data), but it ended up being hard to understand and extract insights. For the future, it would make sense to add more simple, straightforward charts to raise the interpretation levels and to, consequently, decrease the cognitive cost associated with the consumption of data.

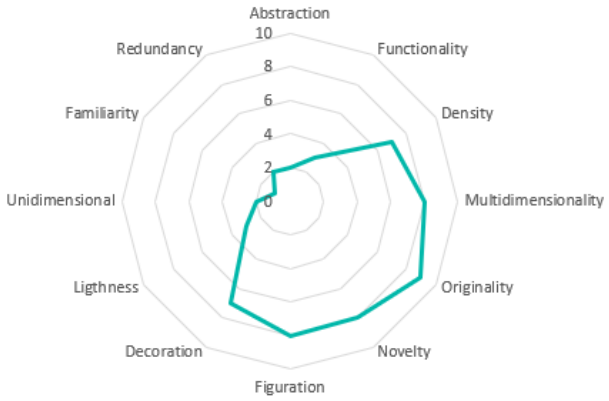



Figure 34 - Radar plot assessment for the “Asana - Stages Behavior” tab

(Made by the author)

# 6. CONCLUSION

Data projects are always complex to conceptualize, design, build and deploy. Because there's a lot of different practices, workflows and platforms involved in data solutions, it's hard to prove that a specific project was done using the best processes. However, there is one variable that should be transversal to any data-related solution – it needs to start with a business case. As such, the project I had the chance to develop and deploy was fundamentally based on real business needs that needed to be tackled fast. The goal, consequently, was to help a marketing team (called Customer Voice) migrating from spreadsheet-based analytics to a Business Intelligence-centered platform to scale up insights, information availability and to increase their data literacy by using dynamic dashboards.

Throughout the development of this project, I got a deeper knowledge of the ETL's nature - all the way from its evolution to how it should be optimized iteratively over time -, and I've also found other modern types of performing data integrations based on the literature (such as ELT). In addition, I learned the main differences between data warehouses deployed at on-premises environments versus cloud infrastructures, as well as the evolution from DWs to data lakes. All this historical context and information allowed me to build this project in a more thoughtful and knowledgeable way, beyond having the direct help of the Data Engineering team throughout the development of the solution.

Even though I couldn't perform all the ETL processes I managed to study from the literature – due to internal reasons -, I still got the chance to deploy critical data engineering tasks that helped automate data entries and raise data quality standards. It's important to reinforce that the internal customer involved in this project (the Customer Voice team) had their entire data stored on a mega spreadsheet full of rows (thousands) and hundreds of columns – creating higher latency as new data entries came in. Consequently, just the fact that I contributed with this migration successfully, it is a success by itself – even though, again, I didn't have the chance to perform all the tasks I wanted to.

In addition, a data warehouse was built, and seven Power BI tabs were created, covering the Customer Voice team's business needs and angles from end to end. The result of the data visualizations was extremely positive, based on the outputs taken from the Visualization Wheel framework. The dashboards contributed to higher data interpretability and, at the same, helped the team members get deeper, more complex insights – which allowed them to find data patterns and behaviors that were impossible to identify through static spreadsheet-based analytics. As such, the Power BI solution made use of both simple, easy-to-understand visualizations for immediate insights - promoting a low cognitive cost -, as well as deeper data charts to enable them to dig deeper into the data by making

use of the drill-downs, drill-throughs, and filters so they could extract more useful, specific insights too.

To finalize, and as a result of this project, the Customer Voice team has now the resources and the data to improve their overall decision-making processes, strategies and operational efficiencies.

# 7. REFERENCES


A. Simitsis, K. Wilkinson, M. Castellanos, U. Dayal. QoX-Driven ETL Design: Reducing the Cost of the ETL Consulting Engagements. In SIGMOD, 2009.

Agrawal, D. (2009). The Reality of Real-Time Business Intelligence. In M. Castellanos, U. Dayal, & T. Sellis, Proceedings of the 2nd International Workshop on Business Intelligence for the Real Time Enterprise (BIRTE 2008) (pp. 75-88). Heidelberg: Springer.

Agrawal, D., Abbadi, A. E., Das, S., & Elmore, A. J. (2011, April). Database scalability, elasticity, and autonomy in the cloud. In International Conference on Database Systems for Advanced Applications (pp. 2-15). Springer, Berlin, Heidelberg.

Alouffi, B., Hasnain, M., Alharbi, A., Alosaimi, W., Alyami, H., & Ayaz, M. (2021). A systematic literature review on cloud computing security: threats and mitigation strategies. IEEE Access, 9, 57792-57807.

Cairo, A. (2011). The Functional Art: an introduction to information graphics and visualization.

Castellanos, M. G., Dayal, U., Simitsis, A., & Wilkinson, W. K. (2014). *U.S. Patent No. 8,719,769*. Washington, DC: U.S. Patent and Trademark Office.

Chen, P. P. S. (1976). The entity-relationship model—toward a unified view of data. ACM transactions on database systems (TODS), 1(1), 9-36.

Churchill Jr, G. A. (1979). A paradigm for developing better measures of marketing constructs. *Journal of marketing research*, *16*(1), 64-73.

Clikeman, P. M. (1999). Improving information quality. Internal Auditor, 56(3), 32-33.

Codd, E. F. (2002). A relational model of data for large shared data banks. In Software pioneers (pp. 263-294). Springer, Berlin, Heidelberg.

De Montcheuil, Y. (2005) Lesson – Third Generation ETL: Delivering the Best Performance, What Works: Best Practices In Data Warehousing and Business Intelligence, 20, p. 48

Dyczkowski, M., Korczak, J., & Dudycz, H. (2014, September). Multi-criteria evaluation of the intelligent dashboard for SME managers based on scorecard framework. In *2014 Federated Conference on Computer Science and Information Systems* (pp. 1147-1155). IEEE.

EM360 Tech (2020). Top 10 Cloud Data Warehouse Solution Providers https://em360tech.com/data_management/tech-features-featuredtech-news/top-10-cloud-data-warehouse-solution-providers

Esmail, F. S. (2014). A Survey of Real-Time Data Warehouse and ETL. *Management*, *9*(3), 3-9.

Foote, K. D. (2022, February). A Brief History of Data Architecture: Shifting Paradigms. https://www.dataversity.net/brief-history-data-architecture-shifting-paradigms/#

Gerber, M., Von Solms, R. (2008) Information security requirements – Interpreting the Legal Aspects, Computers and Security, 27, 124 - 135.

Get out of the data swamp with a governed Data Lake. (2019). Dataeaze. https://www.dataeaze.io/get-out-of-the-data-swamp-with-a-governed-data-lake/#:~:text=Data%20swamp%20is%20nothing%20but,results%20in%20a%20data%20swamp.

Golec, D., Strugar, I., & Belak, D. (2021). The Benefits of Enterprise Data Warehouse Implementation in Cloud vs. On-premises. ENTRENOVA-ENTerprise REsearch InNOVAtion, 7(1), 67-76.

Hahn, S. M. L. (2019, April). Analysis of Existing Concepts of Optimization of ETL-Processes. In Computer Science On-line Conference (pp. 62-76). Springer, Cham.
Henschen, D. (2010), Agile Business: 2010 BI and Information Management Survey, Information Week.

https://www.infoworld.com/article/3228245/the-80-20-data-science-dilemma.html

Huang, K., T., Lee, Y., W., Wang, R., Y., (1999), Quality Information and Knowledge, NJ: Prentice-Hall.

Inmon, W. H., (1996), Building the Data Warehouse, 1 st edition, Indiana: Wiley Publishing Inc.

Jesan, J., P. (2006) Information security. Ubiquity, Article 3 (January 2006), 1 pages. DOI=10.1145/1117693.1117695 Retrieved May 14, 2012 from http://doi.acm.org/10.1145/1117693.11176 95

Jivet, I., Scheffel, M., Specht, M., & Drachsler, H. (2018, March). License to evaluate: Preparing learning analytics dashboards for educational practice. In *Proceedings of the 8th international conference on learning analytics and knowledge* (pp. 31-40).

Kakish, K., & Kraft, T. A. (2012). ETL evolution for real-time data warehousing. In *Proceedings of the Conference on Information Systems Applied Research ISSN* (Vol. 2167, p. 1508).

Karami, M., Langarizadeh, M., & Fatehi, M. (2017). Evaluation of effective dashboards: key concepts and criteria. *The open medical informatics journal*, *11*, 52.

Kaur, H., Agrawal, P., & Dhiman, A. (2012, September). Visualizing clouds on different stages of DWH-an introduction to data warehouse as a service. In 2012 International Conference on Computing Sciences (pp. 356-359). IEEE.

Kimball, R. (2003). The Soul of the Data Warehouse, Part 2: Drilling Across. https://www.kimballgroup.com/2003/04/the-soul-of-the-data-warehouse-part-two-drilling-across/

Kimball, R., & Caserta, J. (2004). The data warehouse ETL toolkit. John Wiley & Sons.

Kurunji, S., Ge, T., Liu, B., & Chen, C. X. (2012, December). Communication cost optimization for cloud Data Warehouse queries. In 4th IEEE International Conference on Cloud Computing Technology and Science Proceedings (pp. 512-519). IEEE.

Lawson-Body, A., Lawson-Body, L., & Illia, A. (2022). Data Visualization: Developing and Validating Dashboard Measurement Instruments. *Journal of Computer Information Systems*, 1-13.

Loshin, D., (2003), Business Intelligence. San Francisco: Morgan Kaufmann Publishers.

Madera, C., & Laurent, A. (2016, November). The next information architecture evolution: the data lake wave. In Proceedings of the 8th international conference on management of digital ecosystems (pp. 174-180).

McKinsey & Company (2018). Creating value with the cloud. https://www.mckinsey.com/~/media/McKinsey/Business%20Functions/McKinsey%20Digital/Our%20Insights/Creating%20value%20with%20the%20cloud%20compendium/Creating-value-with-the-cloud.ashx

McKinsey & Company (2019). Unlocking growth with data-driven marketing and creativity. https://www.mckinsey.com/business-functions/marketing-and-sales/our-insights/data-driven-marketing

Muddasir, M., Raghuveer, K., & Dayanand, R. (2021). Towards Comparative Analysis of Resumption Techniques in ETL. Indonesian Journal of Information Systems, 3(2), 82-93

Nogueira, V. G. C., & Fuscaldi, K. C. (2018). Painel de Especialistas e Delphi: Métodos complementares na elaboração de estudos de futuro. Documentos, 5, 58.

Ruiz, A. (2017). The 80/20 data science dilemma.

Santos, R.J., Bernardino, J. (2009). Optimizing data warehouse loading procedures for enabling useful-time data warehousing. In Proceedings of the 2009 International Database Engineering & Applications Symposium (pp. 292-299). New York: ACM

Shaikh, A. H., & Meshram, B. B. (2021). Security issues in cloud computing. In Intelligent Computing and Networking (pp. 63-77). Springer, Singapore.

Simitsis, A., Vassiliadis, P., Dayal, U., Karagiannis, A., & Tziovara, V. (2009, August). Benchmarking ETL workflows. In Technology Conference on Performance Evaluation and Benchmarking (pp. 199-220). Springer, Berlin, Heidelberg.

Singh, R., Singh, K. (2010), A descriptive classification of causes of data quality problems in data warehousing. IJCSI International Journal of Computer Science Issues, 7(3)

StreamSets (n.d.). BT Group's Openreach: Democratizing Data to Drive Business Results https://streamsets.com/dataops-case-studies/bt-group/

Tank, D. M., Ganatra, A., Kosta, Y. P., & Bhensdadia, C. K. (2010, October). Speeding ETL processing in data warehouses using high-performance joins for Changed Data Capture (CDC). In *2010 International Conference on Advances in Recent Technologies in Communication and Computing* (pp. 365-368). IEEE.

U. Dayal, M. Castellanos, A. Simitsis, K. Wilkinson. Data Integration Flows for Business Intelligence. In EDBT, 2009.

Vasarla, P. (2021). 7 Reasons Why Business Intelligence (BI) Is Crucial https://towardsdatascience.com/7-reasons-why-business-intelligence-bi-is-crucial-55e9d32833eb

Wang, C. (2021). ETL VS ELT: Choose the Right Approach for Data Integration https://www.fivetran.com/blog/etl-vs-elt

What is a data lake?. (n.d.). Google. https://cloud.google.com/learn/what-is-a-data-lake#:~:text=A%20data%20lake%20is%20a,of%20it%2C%20ignoring%20size%20limits.

What is Hadoop?. (n.d.). Amazon Web Services. https://aws.amazon.com/pt/emr/details/hadoop/what-is-hadoop/

Wingate, R. (2019). Alberto Cairo's Visualization Wheel. https://ryanwingate.com/visualization/guidelines/visualization-wheel/#:~:text=The%20Visualization%20Wheel%20is%20a,which%20contain%20deep%2C%20complex%20data.

Ying, Z., & Yong, S. (2009, May). Cloud storage management technology. In 2009 Second International Conference on Information and Computing Science (Vol. 1, pp. 309-311). IEEE.

# APPENDIX

## Tab 2 of the report: "*CV Universe*" – Participants' answers

| Question | Participants' answers | | | |
|---|---|---|---|---|
| | Participant 1 | Participant 2 | Participant 3 | Participant 4 |
| **1)** What the first impression is when looking to the data visualizations on this tab? | "It looks really straightforward to read and easy to digest, with great colors to catch our attention. In addition, the filters are a great way to drill-down any analysis we may want to make." | "As I see a few different data visualizations covering different angles, I feel like it's a little bit hard to extract insights in a few seconds. The visuals are simple, but I feel like it's a dense page with data." | "This page is providing a simple, centralized way of what we do as a team, and that is great. I can easily understand and interpret the graphs, which helps me take action points." | "The dashboards on this page are user-friendly, even though there are some plots I'm not familiar with. This means I can definitely understand what the data is telling us and where we are as a team and our operations, but I also have to take a few time to interpret some of the visuals. Finally, playing with the filters also tells me that are plenty angles to analyze, which is a good, but also a complex thing to do." |
| **2)** Are the data plots easy to understand? (complex vs simple) | "In my opinion, the data plots are indeed easy to understand, because they are familiar visuals and because it portraits what we do as a team." | "I think the data visualizations are more complex than simple, as they are trying to cover the same angles but in different ways. As I'm not a data-driven person, maybe this opinion is a little bit biased, but it is my interpretation." | "They are easy to understand not only because I'm familiar with the visualizations but also because the metrics are self-explanatory." | "The charts that are being used on this report, like bar charts, gauges and cards are very simple to understand and I can get fast insights out of it." |
| **3)** Do you easily understand what are the metrics inside each graph? | "Yes, because every plot as its label presented." | "I do. The metrics are very descriptive, allowing me to understand what each data visualization is measuring." | "As I ended up saying in the previous question, the metrics are labeled on each plot, providing us with a good interpretation, read experience." | "Yes. I easily understand what the metrics are because they are presented like a legend - so it's self-explanatory." |
| **4)** Can you easily extract insights from the dashboards? | "Yes, I can. The dashboards are extremely descriptive, in my opinion." | "The fact that I'm not too data-driven as of today, I can't easily extract insights from the dashboards." | "Yes, definitely. The familiarity with the types of visuals allows me to get insights with little effort." | "It's a mix. In some plots, I can immediately extract insights, but in others I can't. But my overall answer it's more positive than negative, which means getting insights out of this page it's easier than harder." |
| **5)** Do the data visualizations answers your business needs? | "Absolutely. This page is covering a lot of data angles of what we needed as a team." | "As I understand the metrics on this section of the report, I can see that the visuals are indeed matching our business needs." | "Yes, they do. The data visualizations are covering the metrics according to the business needs we raised beforehand." | "Yes, and this is something we never had before." |
| **6)** With these graphs, how simple is it to tell a story from the data? | "It's easy because the data is displayed in a sequential, general to particular way." | "I think it's a little bit complex because there are a lot of graphs in my opinion. Aggregating all that information into a single story would be hard for me." | "It's simple because the report goes from a more broad overview of the department to a more narrowed analysis as we scroll-down the page. If we follow this flow, it's going to be easy to tell a compelling story." | "As I'm not familiar with every plot, for me it requires a little bit of effort to share a story. But it's totally achievable." |
| **7)** Do the data plots help you figure out what you need to improve on a business perspective? | "Yes. For example, the gauge plots on top of the page give us a very tangible view of what we still need to do to achieve our best-in-class performance." | "I think so. But when I become more familiar with the data I can give a more credible answer." | "I have no doubts that, with these dashboards, we now have the answers we need. And it's important to mention that we never had this type of view in the past." | "As far as I can understand, and considering the calculated metrics, the data plots do help us figure out what we need to improve on a business point of view." |
| **8)** Do these graphs help you be more data-driven to take decisions? | "Yes. I think these graphs will change our mindset in the way we decide and define strategies." | "I really think this page is the first step to become a more data-driven person indeed, which will help me take decisions more objectively." | "This vision of our department will highly foster data-driven decisions, so yes." | "I believe in the power of data, in every occasion. And having this data about our team is basically a gateway for all of us to become more data-centric." |
| **9)** How easy is to play and navigate with the data visualizations? (i.e., filters, drill-downs, drill-throughs) | "It's very easy and user-friendly." | "They are easy to use thanks to the simple way the filters are displayed in the page." | "It is easy for filters, but for drill-downs and drill-throughs is more complex." | "They are very intuitive to use and open our minds to try new things and find new capabilities while we navigate the data." |

## Tab 3 & 4 of the report: "*Case Studies – Main Coverage Matrix*" and "*Coverage Matrix for Ambassadors*" – Participants' answers

| Question | Participants' answers | | | |
|---|---|---|---|---|
| | Participant 1 | Participant 2 | Participant 3 | Participant 4 |
| **1)** What the first impression is when looking to the data visualizations on this tab? | "Simple to understand. The idea of using heatmaps was great to help us interpret the data." | "Very simple page report with pivot tables having the same style. The colors in a form of a heatmap really give us a notion of what we need to improve, as well as what we are already doing well." | "These matrices are very good too identify where we are and what (and where) we need to do to improve." | "Basic view of the data that really is easy to analyze, even though there are a lot of data points to digest." |
| **2)** Are the data plots easy to understand? (complex vs simple) | "The pivot tables are easy to understand if we look at the very first level of the hierarchy. However, as there's more levels, it becomes complex to interpret and take final insights." | "Overall, it's easy to understand. As I mentioned in the previous question, the heatmap colors are key when interpreting the results." | "The plots are easy to understand indeed. However, I'm aware of the hierarchies, which means there are more data points to analyze beyond the ones in the first level of the hierarchy. This adds more complexity in the analysis." | "A pivot table is always straightforward to analyze, so I do think those plots are easy to understand." |
| **3)** Do you easily understand what are the metrics inside each graph? | "Yes. It's basically a count of all the stories by the different angles." | "Yes, because the graphs really match what we wanted to have as an output." | "Completely. The metric is the same and transversal to every matrix." | "Yes. The graphs are essentially showing the exact same metric, so it's very easy to understand." |
| **4)** Can you easily extract insights from the dashboards? | "It's very easy to get findings if the plots only had one hierarchy level. But as it's not the case, it can become more complex to extract final insights." | "I can, and I very much like this style of visualization. The only sensitive point are the hierarchies, which will add a little bit more complexity." But overall I would say it's easier than more complex." | "I definitely can. Even though there are hierarchies, they allow us to go deep and find insights that we couldn't if we didn't have them. And that, in my opinion, is a game-changer." | "Yes. As I answered in a few questions before, the pivot tables are really straightforward to analyze." |
| **5)** Do the data visualizations answers your business needs? | "They do. This is a simple view that contains the metric we wanted to track." | "Yes. As I mentioned before, the visuals are really matching what we wanted to have as a team." | "For a long time that we wanted to analyze how many marketing assets we were publishing, and these visuals are definitely answering that business need." | "They effectively answer our needs. We finally have a centralized view of what we have published over the years, as well as the use cases of those publications and regions." |
| **6)** With these graphs, how simple is it to tell a story from the data? | "It's extremely accessible because we're talking about the same metric, but covering different angles." | "Very easy. The tables follow a logical flow of the information, all the way from a wider perspective to a more narrowed vision of the same phenomenom." | "As they are easy to understand, I find no difficulties in telling a story using thses visuals." | "It is simple because it's dynamic. The hierarchies add dynamics to the data and that's really important if we were to tell a story." |
| **7)** Do the data plots help you figure out what you need to improve on a business perspective? | "The pivot tables used on these tabs definitely help us understand where we need to invest more in a marketing perspective and especially where we should allocate more resources. In addition, it also gives us a notion of what use cases we are covering more, allowing us to define strategies according to it." | "Even though I'm familiar with tables and pivot tables, I think the data plots require a dig deep on the data to allow us to extract insights. The drill-down capabilities are definitely a great way to deepen the analysis, but it requires more complexity, focus and time to take action points and decide on top of that." | "I think the data plots give us a snapshot of how we behave and what we need to do in terms of marketing initiatives, at least looking at the highest level of the hierarchy. But we do need to use the drill-down capability to understand the different realities for the same variables, meaning that we'll need to schedule a slot or something similar just to focus on interpretating the data using different angles. This means it'll be complex and time-consuming." | "On a first glimpse, the general view of the tables indeed shows us what we need to improve from a business perspective. However, there are other levels in the data that we need to address in order to find useful, targeted insights. This means that we need to spend more time digging the data vs having a one-shot analysis and define action points. This type of visualizations require a more technical view over the data points, so we can extract the insights and define a plan to improve our marketing efforts." |
| **8)** Do these graphs help you be more data-driven to take decisions? | "Of course. I think that these pivot tables are actually a starting point for us to be more data-oriented and strategize based on the available information in the visualizations." | "By having to dig deep in the data visualizations, it forces all of us to become more data-driven when making decisions." | "Same inputs as tab number 2, this vision of the data will highly contribute to more decisions based on data." | "I have no doubts that these pivot tables will help us take better, data driven decisions.. Because they are portraiting one of the most important metrics of our department." |
| **9)** How easy is to play and navigate with the data visualizations? (i.e., filters, drill-downs, drill-throughs) | "It's very easy to use the drill-drough capability." | "I really enjoy playing with the drill-through feature to get more data insights." | "Even though I never use a drill-through before, I can see the value of it and it's very intuitive to use." | "Navigating these pivot tables using the drill-through is way much better than using Excel." |

## Tab 5 of the report: "*Customer Voice Assets Behavior*" – Participants' answers

| | Participants' answers | | | |
|---|---|---|---|---|
| **Question** | **Participant 1** | **Participant 2** | **Participant 3** | **Participant 4** |
| **1)** What the first impression is when looking to the data visualizations on this tab? | "The main goal of this page is for us to access and analyze historical information of our team. The filters are there to help us throgh the analysis journey and the layout looks very organized." | "Well-organized page with data visualizations adressing historical information. It looks simple to analyze." | "It's a bit complex at first because I see different charts covering multiple data angles. However, I'm familiar with them." | "I feel like there's a lot of information to digest." |
| **2)** Are the data plots easy to understand? (complex vs simple) | "It's actually a mix. One one hand there are plots I'm familiar with, like the bar chart and the pivot table, but there are others I have no idea to analyze and its business context." | "In my opinion, the data plots are simple to read, and its clear what they're covering. But I also know that the filters will add more complexity when analyzing different views of the data." | "I think they are a little bit hard to understand at first. At least I feel that having a variety of different charts in this page adds more complexity." | "As I feel there's a lot of information to analyze, the data plots are not that easy to understand at first. Of course I can read a pivot table, but as there are plots that have multiple features to dig deep in the data, I feel that they can cause entropy in the interpretation of the data |
| **3)** Do you easily understand what are the metrics inside each graph? | "Yes, because every graph has the metrics labeled." | Yes, I can understand because the metrics are written in every data plot." | "It's labeled, so it's easy to see what the metric is inside each visualization." | "I easily understand the metrics and the underlying business value that they bring for our team, which is great." |
| **4)** Can you easily extract insights from the dashboards? | "Overall I can, but there are plots that it's harder because of the drill-downs. I mean, it's great to see more specific findings, but we need to navigate to obtain findings." | "It can be complex to get insights because almost every plot has different levels of information. Personally, I think the insights coming from the drill-down feature are great and relevant to our operation, but it can be a time-consuming experience. So yes, it's a little bit complex, as I | "Even though there are plots I'm not that familiar, I do think we can extract insights in a fast way about different business topics." | "The high concentration of data in one page doesn't allow me to extract insights in an easy way. I prefer simpler, homogeneous visualizations to help on the interpretations." |
| **5)** Do the data visualizations answers your business needs? | "The great majority of the data visualizations answer our business needs, which is very important for us. Although there are more complex data plots than others, requiring more critical thinking and focus, it's easy to understand what's the story behind." | "Having a filter with the years as well as data plots with a date hierarchy turns everything very intuitive to find patterns of behavior. This is, per si, very valuable from a business perspective, so the plots definitely answer our business needs and questions. It requires some navigation on the drill-down capabilities in the plots and the filters as well to understand the different phenomenas, so it can be complex to extract insights fast. But in terms of content, the visuals are solid." | "Yes they do. There are some visuals I'm not familiar with, adding a little bit of complexity when interpreting the data, but if you dig deep and invest some time analyzing the data, you can get the answers to our business questions." | "For sure that the visualizations cover our business needs. I only think that there's a lot of info to consume not only on a page level but also on a data visualization level. As far as I could understand, there are filters and hierarquies in the visuals, so every plot will change the data according to what we want to see - and that requires time to process all the information stored in this page. It's very good because we get a holistic view over this page subject, but it can also be hard to draw conclusions from a lot of data points." |
| **6)** With these graphs, how simple is it to tell a story from the data? | "Not easy at all, because this report's section is covering multiple scopes of our activities." | "As I see a lot of diferent metrics in just one page, it can be hard to tell a compelling data story at first. It requires time to understand the patterns about the different scenarious and then share in a form of a story." | "I feel that it's simple. We just need to invest more time on this page vs others, but we can tell a story. I mean, there are a few metrics to analyze, but it's feasible." | "It's difficult to tell a story. We have multiple angles to analyze, instead of having a more centralized analysis. I feel that I would be lost in the story." |
| **7)** Do the data plots help you figure out what you need to improve on a business perspective? | "Yes. Having historical information is always good in businesses and we can act on top of that to become more efficient." | "Although I think that the Tree Map plot adds no value to this page, I can figure out what we need to improve because we can have a lot of layers of analysis. That's powerful." | "Absolutely. It's complex to get the necessary insights for improving ourselves on a business perspective, but once we have them we can. The information is there, so we can improve." | "For sure. We have a lot of different data points about things we never thought possible to have, and that's the value." |
| **8)** Do these graphs help you be more data-driven to take decisions? | "Yes, undoubtedly." | "As some of the plots are complex to analyze, the idea of having to figure out different data insights by ourselves is the way to be more data-driven and impact our decisions according to that." | "Same as the other tabs, this data is the gateway to start being more data-driven, replacing the subjective way of making strategies." | "As I mentioned, because we never had this data before, this definitely help us be more data-driven when making decisions." |
| **9)** How easy is to play and navigate with the data visualizations? (i.e., filters, drill-downs, drilll-throughs) | "Once we deal with the drill-downs and filters, it's easy to use them." | "The user experience is not the best because playing with all the features of the page and the its visuals is confusing. I'm not questioning the value of them, I'm just talking about the experience of a user." | "It's straighforward, there's no much to say. The filters are simple to use and the drill-downs are just clicks, so I can navigate the data plots easily." | "It was hard in the first time, but now that I know how to use them, it's already easy." |

## Tab 6 of the report: "*Stories Timeline*" – Participants' answers

| | Participants' answers | | | |
|---|---|---|---|---|
| **Question** | **Participant 1** | **Participant 2** | **Participant 3** | **Participant 4** |
| **1)** What the first impression is when looking to the data visualizations on this tab? | "Very simple view of the data, being easy to digest." | "No complexity on this page, as there's only two types of data visualizations." | "I feel like the page and the visuals are not dense, meaning that they don't have too much content to analyze. That can be good to extract fast insights, as opposed to the previous | "By looking at the layout of the page and the data charts as well, we have a very simple view of the information." |
| **2)** Are the data plots easy to understand? (complex vs simple) | "I would say yes because they are well organized and presented. As I already dealed with these visuals in the past, I feel that it's easy to get insights." | "They are easy to understand because I'm familiar with these types of data plots." | "They are simple charts to analyze fast, so yes." | "There's no complexity at all associated with these data visualizations, so they are very accessible to read and retain the information." |
| **3)** Do you easily understand what are the metrics inside each graph? | "Following the logic of the other tabs, the metrics are written in the plots, so it's easy to understant them." | "Yes, it's very simple because they are labeled on every plot. And if we put the mouse over a data point, we also have the name of the metric and the specific value." | "Yes, totally." | "I can understand what metric the visualizations are using because it's placed on every plot." |
| **4)** Can you easily extract insights from the dashboards? | "Definitely. These are simple charts that trigger fast insights and, consequently, reduce the time to value. And that's what we wanted to have." | "Yes, and fast." | "As the page and visuals are not dense, as I mentioned before, I can extract findings really fast." | "The fact that there's no complexity in the visuals, allows me to get insights with little effort." |
| **5)** Do the data visualizations answers your business needs? | "Yes, the visuals have all that we need to track." | "Yes, absolutely." | "They do, and having a historical view over our operations is even better, because we can learn from previous | "Having this information in this type of visuals is great. And yes, the information stored in the plots, definitely answers our business |
| **6)** With these graphs, how simple is it to tell a story from the data? | "Telling a story using these graphs it's easy. As this page is simple and has only two chart types with the information we need, it'll be easy to tell a story for sure." | "Having very few data visualizations on this section, allows me to tell a story that's compelling enough. And the fact that I can play with the filters means that I can add specific insights for a specific scenario, | "As I can analyze the visualizations fast, it is simple for me to study the data and tell a story." | "The simple view of the data enables me to share a good story just by using theses plots." |
| **7)** Do the data plots help you figure out what you need to improve on a business perspective? | "Yes. These plots are telling me how much time we move from one stage to another, so we can extract insights of that time duration and iterate over our initiatives to improve the way we operate and manage them." | "Yes. The visuals really provides us with data points that we can act on top of them, which means they are not just data points, but data that translates into higher business knowledge." | "As I mentioned before, the fact that there's historical information on this report's page, allows us to know how we worked in the past and learn to better in the future." | "Absolutely. Because the plots don't demand for complex interpretations, we can easily understand what we did wrong and what we did the right way." |
| **8)** Do these graphs help you be more data-driven to take decisions? | "Absolutely." | "Of course. These plots are also very important for us to have a more data-oriented journey as we operate over time." | "Even though we just have two different plot types, it contributes to the way we do our work." | "Yes. For a long time that we wanted to have this data in a form of visualizations, so I'm sure this will help as a team on making decisions based on that information." |
| **9)** How easy is to play and navigate with the data visualizations? (i.e., filters, drill-downs, drilll-throughs) | "There's only one filter to use, so it's easy." | "It's simple to use." | "It's like any other filter, so I can easily navigate any of the plots while using it at the same time." | "Very easy to play with the filter." |

## Tab 7 of the report: "*Stories Pipeline – Coverage Matrix*" – Participants' answers

| | Participants' answers | | | |
|---|---|---|---|---|
| **Question** | **Participant 1** | **Participant 2** | **Participant 3** | **Participant 4** |
| **1)** What the first impression is when looking to the data visualizations on this tab? | "As the visualizations are very similar compared to the tab number 3 and 4, my opinion is the same. The tables in a form of heatmaps are great in helping us understand the data. | "It seems like this page follows the same style of the ones in the tab 3 and 4, so I keep saying that the heatmaps really give us an easy understanding of what we are not doing nor prioritizing, but they also provide us the things we are doing correctly." | "These matrices really tell us how we manage our pipeline operations in very straightforward way." | "This looks like, once again, a consolidated view of the things we're doing every day in our pipeline. And finally, we have all this data in the same place, displayed in an organized way." |
| **2)** Are the data plots easy to understand? (complex vs simple) | "Yes. The pivot tables are always easy assets to understand. Even though we can make use of the drill-down feature inside each plot, that doesn't reduce any understanding when interpreting the data." | "Yes. The key on these plots is the colors of the heatmap, which enables us to easily identify spots to invest, as well as replicating best practices." | "The plots are easy to understand, for sure. However, the presence of the hierarchies can possibly add more complexity in the interpretation. But I don't think it will make the tables more complex to read." | "I'm very familiar with pivot tables, as I mentions tabs before. So for me they are easy to understand." |
| **3)** Do you easily understand what are the metrics inside each graph? | "Absolutely. The metric is the same for every plot, but it's just covering 3 different angles of analysis." | "Yes. The metric is essentially a count of the number of initiatives we currently have on our pipeline, utilized in different x-axis." | "It's very easy because this was another important metric that we wanted to have to improve our operational efficiency." | "Yes, it's very clear wat the metric is on every plot." |
| **4)** Can you easily extract insights from the dashboards? | "Yes I can. As these types of visuals are extremely easy to read and digest, I can take insights really fast." | "Looking at the visualizations in this tab, which are simple pivot tables covering various data angles, it is very easy to extract insights and define action points. And that is very positive for us at an operational level, because it allows us to identify where we are investing more of our resources and especially where should we invest in the future." | "Not only I can extract insights fast, but I can define what moves should we move and work next. The tables give us a simple view of how our marketing pipeline is working, and that allow us to act fast and in an agile way - which were things we couldn't do before due to the lack of data analysis." | "I can. The familiarity I have with pivot tables from previous experiences helps me extract insights using these pivot tables too." |
| **5)** Do the data visualizations answers your business needs? | "Yes. These type of visualizations are also something we really wanted to have. And the data inside really give us the business answers we were looking for." | "From the very beggining that we wanted to measure how many initiatives we have on our pipeline, separated by different angles. And these pivot tables are the answer." | "Yes. And the fact that there are different hierarchic levels in the tables, really allow us to get deeper business answers." | "I have no doubts that the data visualizations do answer our business needs. We were actually waiting for something like this for a long time and here they are." |
| **6)** With these graphs, how simple is it to tell a story from the data? | "I think it is simple because of the way the data is presented. Easy-to-understand visualizations are always a trigger to tell a compelling story." | "It's really straighforward. The plots are all the same but they just cover different angles of analysis." | "I would have no difficulty in presenting insights in a form of a story." | "I believe it's no rocket science. I mean, the visuals are simple and the metric is very clear to me, that's all I need to share a story out of this data." |
| **7)** Do the data plots help you figure out what you need to improve on a business perspective? | "Yes. The heatmap style is the key, because it help us identify where we stand and where we need to invest more resources. It fundamentally allow us to prioritize better." | "For sure. The colors used on the tables are really a trigger to understand how we behave and what we can do to improve on a business level." | "Yes, absolutely. They are providing us with data about how our pipeline is working, so that's all we needed." | "The heatmap colors are there to help us identify our weak spots as well as the spots we're doing great. It's like a business decision matrix." |
| **8)** Do these graphs help you be more data-driven to take decisions? | "Yes they do. I think that these easy-to-digest types of visualizations are a good starting point for us all, in terms of being more data-driven and make better decisions." | "Yes, these graphs are really helpfull to make decisions based on data." | "The simple view of our pipeline presented, in a form heatmap pivot tables, is really helping us giving more credit to the data we have available, which was something we didn't have before." | For sure. Every time I dealed with pivot tables before had that goal in mind: to help us being more data-drive, and these ones are no exception at all." |
| **9)** How easy is to play and navigate with the data visualizations? (i.e., filters, drill-downs, drilll-throughs) | "It's really easy to use, and the filters are self-explanatory." | "The filters are very easy to use and visuals are also very simple to navigate." | "The user experience is great, because the filters and the drill-downs are super straightforward to use." | "I have no difficulties in making use of the filters and drill-downs at the same time to enrich my analysis. So it's very easy." |

Tab 8 of the report: "*Asana – Stages Behavior*" – Participants' answers

| | Participants' answers | | | |
|---|---|---|---|---|
| **Question** | **Participant 1** | **Participant 2** | **Participant 3** | **Participant 4** |
| **1)** What the first impression is when looking to the data visualizations on this tab? | "Every plot has the same style, so as far as I can understand the first one, the rest won't require any effort to interpret. In addition, I think the plots are intuitive to use. But I think, just by looking at them, it will require some time to understand what the plots are really showing to us in terms of | "The charts are like trees of information and they are displayed by stage. They are well distributed, so I feel there's no entropy on this page of the report, although it might require more time to understand how to navigate through them and extract insights. | "Very appellative plots on this page. However, I think they are confused to read and digest." | "My first impression is that the data visualizations contain a lot of information to read and a lot of data to digest, so it will be hard to get insights fast." |
| **2)** Are the data plots easy to understand? (complex vs simple) | "Concerning the data plots in this page of the report, I think they are complex to understand, as I'm not familiar to that type of visualizations." | "Even though I understand what the visual is showing to us, it's hard to get fast insights and find any patterns out of the data. I would use a less complex plot for easier interpretability." | "In my opinion, the data plots are really beautiful but it's hard to extract quick insights. So I think the plots are not that easy to understand.." | "I very much like new, not familiar plots to learn other angles of data. These plots are not the best ones to have an insight in seconds, but they are very good if we want to see information in a hierarchical way. So I would rate it complex, but valuable to get more specific insights." |
| **3)** Do you easily understand what are the metrics inside each graph? | "Yes, totally. The metric is written in the plots, so everything that comes within them is self-explanatory." | "Absolutely. The fact that the label is written on every chart allows us to understand what the plot is measuring. And that's really important." | "Yes. It's clear what the metric is and what angles it covers. I really like the fact that the same metric is divided by operational stages, giving us an overview of what we have in the pipeline as of today." | "Yes, I can see what the metric is just by looking at the plot. And as the metric is labeled below each bar, it's really straightforward." |
| **4)** Can you easily extract insights from the dashboards? | "I can't extract immediate insights, as I'm not familiar with those types of dashboards." | "No, I can't. As I told you in a few questions before, I feel that it's hard to get fast insights out of these charts." | "I can extract insights out of the data visualizations on this page, but it won't be that easy. I feel that I need to dig deep into them to really have a valuable insight for | "I can't. The visualizations are structured hierarchically, so this tells me that I need to navigate on a deeper level to be able to extract something out of them." |
| **5)** Do the data visualizations answers your business needs? | "They do, because they are divided by stage. So this tells us that the visuals are segmented by the things we asked for - and I think it will help us a lot in an operational point of view." | "I'm sure they answer to our business needs. Because they are portraiting the different stages of our pipeline, we finally have a holistic view of what we have and what we need to have and do in the future as well." | "Completely. We asked for charts that could analyze each stage of our marketing pipeline and here they are. " | "The fact that I can't extract insights in a fast way, I'm sure these dashboards answer to our needs." |
| **6)** With these graphs, how simple is it to tell a story from the data? | "It's hard, because there are hierarchy levels by each stage inside each plot. So, to tell a story from the data that is displayed, it will take a long time to tell the full story itself. | "It's not simple because there are a lot of different data points to have in consideration. The story can be good in the end in terms of content, but it won't be simple to perform that job." | "To tell a compelling story I feel that it won't be easy. Just by looking at the plots, there are a lot of angles to explore and that, by itself, won't simplify the construction of the story I want to tell in the end." | "As the nature of the plot is to show different levels of information inside each stage of our marketing pipeline, covering all that in a storytelling experience won't be definitely easy - even though it will be reach in content". |
| **7)** Do the data plots help you figure out what you need to improve on a business perspective? | "They do, if we dig deep on the visualizations." | "Yes. As the plots have information by year, quarter and month, we can understand our behavior today vs in the past - and that's really important to have." | "The visualizations are covering historical information, and that enables us to have a benchmark to compare ourselves and our operations over time." | "Yes. These visuals are not only contributing to optimize the way we do business in our department, as well as improving at every stage of our initiatives." |
| **8)** Do these graphs help you be more data-driven to take decisions? | "Yes, absolutely. I have no doubts that we as team, will became more data-oriented due to these types of dashboards." | "Absolutely. I believe that each one of these graphs have a lot of relevant findings and those will be critical for our operational efficiency." | "Of course. These plots, in particular, force us to explore the data in a deeper level in order to extract insights for decision-making. So we will be automatically more data-driven. | "As these data visualizations are divided by stage, as already mentioned, it will provide us with a lot of value in terms of information. And that's due to the data and the use we make out of it." |
| **9)** How easy is to play and navigate with the data visualizations? (i.e., filters, drill-downs, drill-throughs) | "It's a little bit complex at first, but it becomes like second nature over time. The plots are intuitive to use because they have different layers of information and that's good to obtain various types of information." | "I have to say it's not easy. It could be more straightforward for a new data user like me." | "The features in the visualizations are there indeed, but it took me some time to get used to it." | "It's difficult to be honest. I think creating filters outside the visuals would be better for user experience." |


**NOVA Information Management School**
**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa