NOVA IMS
Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

Data Centralization For Netflix

Abdallah Zaher

Internship report presented as partial requirement for obtaining the Master's degree in Advanced Analytics

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

Data Centralization for Netflix

by

Abdallah Zaher

Internship report presented as partial requirement for obtaining the master's degree in Advanced Analytics

Advisor / Co Advisor: Mauro Castelli

Co Advisor*:* Victor Juaregui

October 2021

ACKNOWLEDGEMENTS

Special thanks to stack overflow for making the impossible possible.

ABSTRACT

Presently, the amount of data produced every day is truly mind-boggling. There are 2.5 quintillion bytes of data created each day at this current pace and companies are actively generating and collecting it. This data has become of the utmost importance and companies are changing their behavior towards it. Many of these companies nowadays are giving more consideration and time investment into the data analytics area. They have started to figure out a way on how to use their customers'/clients' data to generate insights and strategies for better futuristic marketing approaches. In order to have a better understanding of the data, these corporations took the initiative to centralize it in an arranged manner so that the analysis will be more organized through the whole procedure. The following internship report studies the transformation of data from its very early stages to the final PowerBi part. This report will discuss specifically the ETL process that the data undergoes from how the data is collected via SSIS packages, treated through T-SQL, and provided as a clean final table to build a PowerBi report from it so that other employees and people who do not have much knowledge about data will be able to see through graphs that explain what the data means and what insights it holds within.

KEYWORDS

Centralize, Data transformation, ETL process, SSIS packages, T-SQL, PowerBi report.

INDEX

LIST OF FIGURES

LIST OF ABBREVIATIONS AND ACRONYMS

AD Group	Active Directory Group
KPI	Key Performance Indicator
TP	Teleperformance
DBA	Database Administrator
CSR	Customer Service Representative
CSV	Comma-Separated Values
SP	Stored Procedure
HR	Human Resources
SSIS	SQL Server Integration Services
LOB	Line of Business
PDR	Planning, Development & Review
DE	Data engineering

1. INTRODUCTION

“Big data is currently a hot research topic, with four million hits on Google scholar in October 2016. One reason for the popularity of big data research is the knowledge that can be extracted from analyzing these large data sets (B. Nelson 2016)”. With this progressive enhancement, scientists had to adjust to the way they treat big data. In the beginning, big data was just an analogy. There was no clear explanation to what makes data “BIG”. Because big data is everywhere, there is almost an urgent need to collect and preserve whatever data is generated. “In recent years, big data is flourishing, exceeding the traditional data processing methods with 5 'V' characteristics (M. Shabana & K. Sharma (2021)”. Throughout the years, there has been a developing comprehension of the job that huge information can play in conveying inestimable experiences to an association, uncovering qualities and shortcomings, and enabling organizations to improve their practices. Large information has no plan, is non-critical and non-hardliner – it essentially uncovers a depiction of action.

However, while numerous associations comprehend the significance of information, not many are yet seeing its effect. “Another examination entitled Broken Connections: Why investigation still cannot seem to be taken care of makes the case that 70% of business leaders recognize the significance of deals and advertising investigation, yet just 2% say that their examinations have accomplished a wide, positive effect (LINDELL, JIM 2020)”. This discovery focuses on the requirement for Enormous Information to be taken care of by re-appropriated firms who spend significant time in examining the information created by organizations and who can offer genuine, noteworthy experiences. In the foreword to his report, Dan Weatherill composes that "Our study and follow-up interviews with almost 450 U.S-based senior chiefs from enterprises including drugs, clinical gadgets, IT, monetary administrations, telecoms and travel and accommodation affirmed one thing that we definitely knew: hardly any associations have had the option to hit the nail on the head and to produce the sort of business sway that they had expected."

As mentioned before, this thesis took the inspiration of how organized data flow can ease the construction of integration processes. To start with, a plan should be constructed to know the targeted audience, the data providers, and how the staging organization will take place. As a result, the governance model was essential to provide accesses and limitations based on which AD group the user belongs to. After creating these AD groups, the time comes to go through the requirements. For example, the kind of data provided should be defined, as well as the creator, and enumerate to whom it will be appointed to. Based on that, one will have a clearer vision to acknowledge the grant access levels. An expected output will be drawn as a first draft for the final solution.

After going through all the previous points, the light should be shed on the data now. The more one understands the data they are willing to transform, the more the integration process will go smoother. The dimensions and the fact tables will be created accordingly, and then the results will be stored in the fact table to be provided for the PowerBi report.

With this thesis, one will be able to go through all the steps in detail for a better understanding of how the process was built to the delivery stage.

1.1.1. Company Overview

With the huge and advanced technological level companies have reached recently, their demand has increased and accordingly the maintenance part alongside with the customers' satisfactions. Nowadays, if a group wants to be dominant in the market, they must supply their customers with the best services to stay. In sum, customer service originates from sources of desperation (such as call center technology and marketing analytics), which will clearly continue to adapt, grow, and change in the future. The client's involvement begins right from the advancement of an item or benefit by the company till the buying handle and after deals benefit. A great client involvement can be assessed by rehash buying or rehashing deals. Companies have evaluated client encounter as one of the foremost imperative criteria for customer satisfaction, goodwill and to extend brand value within the advertisement.

Figure 1 – *Cycle of Customer Experience*

Many companies in Portugal started to focus more on retail from the data perspective and Teleperformance is one of them. It has seen the potential of creating a customer services platform to

provide the need that has been placed in the market. Briefly, Teleperformance is one of the global leaders in customer experience management that has been connecting customers with the world's most successful companies. Teleperformance's Digital Integrated Business Services combines human touch and high technology to deliver extraordinary customer experiences. While technology creates new and agile ways of working, its interaction

2. PROBLEM OVERVIEW

Being an employee for Teleperformance, a huge number of clients can be added to the list of suppliers of “Data” to work with. Each client differs by the way they operate, the kind of product they offer and what business plan they follow.

As mentioned before, this amount of data and their different data types will generate a problem in categorizing them and providing the ultimate solution. After what has been said, a series of problems rose that need to be discussed later:

1. What type of accesses will be granted to different groups.
2. The type of Raw Data that is provided.
3. KPIs needed for measurements.
4. Expected output.

2.1.1. Case Study

This section will go over the project that inspired the creation of this report. As previously mentioned, Teleperformance has a great set of partners such as Netflix. The partnership between Netflix and TP grew rapidly in the form of providing customer services and studying marketing strategies to evolve both parties. As a result, centralizing data for them became a priority. This process will benefit all associates to keep track of their performance on all aspects based on the KPIs that were built.

2.1.2. Governance Model

Before initiating the project, research had to be done regarding the governance model. The term governance has been extensively used in the last two decades, but its meaning is still ambiguous (Colebatch, 2014; Rhodes, 1996). Data Governance alludes to the requirement of specialist over information collection, announcing, and administration. Within the least complex terms, it is the control of information and the utilization of that data, where, for example, governance has been framed as "the multitude of actors and processes that lead to collective binding decisions (Van Asselt and Renn, 2011: 431)". While more data could be a great thing for the most part, it can be more troublesome to form a sense of that data, keep it clean, and utilize it in a compelling way. That is where an information steward comes in. Information Stewardship is the formalization of responsibility for information administration. The individual assigned to this part will be capable of all thing's information, counting information compliance, information judgment, and information objectives. In other words, information administration is almost the parts, duties, and forms to guarantee responsibility for and possession of information resources and may best be thought of as a work that underpins an organization's overarching information administration strategy.

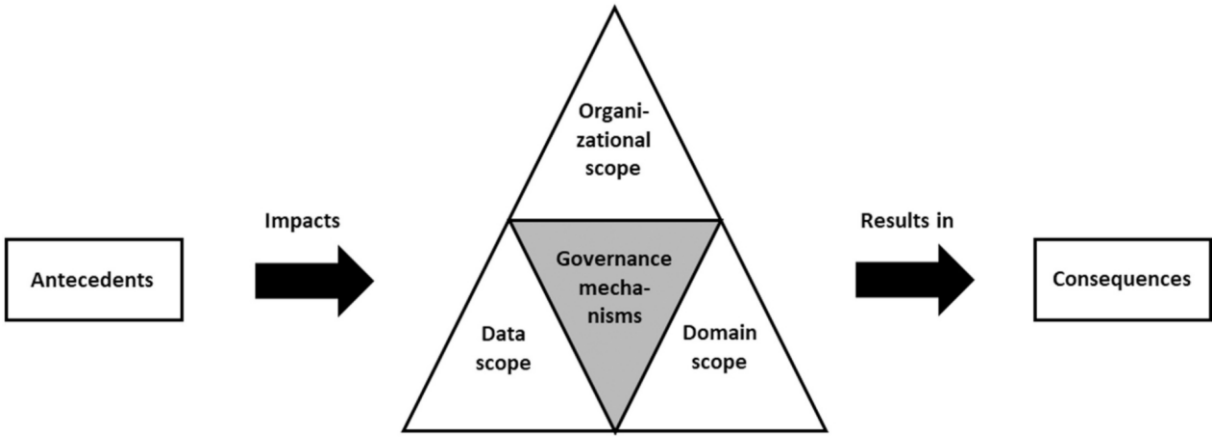


Figure 2 – *Governance model structure.*

Objectives by supporting such a technique may be characterized at all levels of the endeavor and doing so numerous helps in acknowledgment of forms by those who will utilize them. A few objectives include:

- Increasing consistency and certainty in choice making
- Decreasing the chance of administrative fines
- Improving information security, characterizing and confirming the necessities for information dissemination policies
- Maximizing the wage era potential of data
- Designating responsibility for data quality
- Enable superior arranging by supervisory staff
- Minimizing or killing re-work
- Optimizing staff effectiveness
- Establishing handle execution baselines to empower enhancement efforts
- Acknowledging and holding all gain

Regarding TP data structure, first must define the “Active Directory Security Groups”. These AD groupings will target the hierarchy and the permissions to be granted.

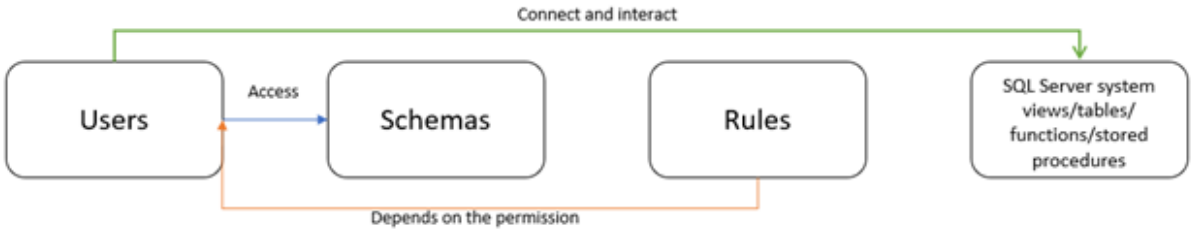


Figure 3 – *Introducing the interconnectivity between all the AD layers.*

Following the AD structure, users should be added to a specific group and should not be appended independently or separately from the group that they belong to. Then the access will be granted to that specific group. This assignment will ease the transition/addition to any new employee or the transfer of one from a project or group to another. By adding them to the new group they will directly have all the accesses needed without granting them manually.

Additionally, one can separate the users into 3 categories as shown in Figure 4:

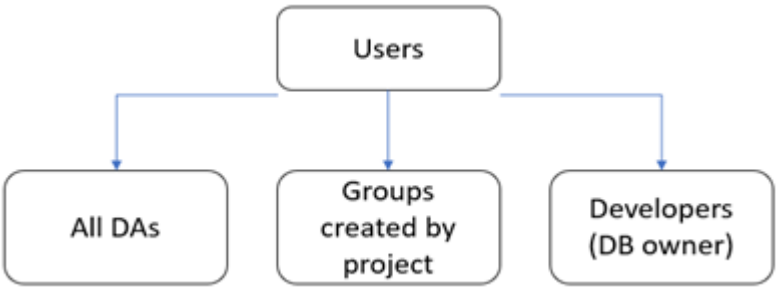


Figure 4 – *Showing different types of users regarding assignment employees to AD Groups.*

As specified before, these rules will be provided based on the group these employees belong to and to their hierarchy. So, for example, not all employees will have the same permissions to access the database but will be based on specificities; for example, a DBA is able to read, write, update… data in a database however a CSR will only be able to read.

Type	Read	Write	Update	Delete	Execute
DBA	X	X	X	X	X
DA	X				X
CSR	X				

Figure 5 – *Showing different accesses of users regarding their AD Groups.*

2.1.3. Raw Data

Before diving through the process, a definition for raw data is required. Raw data alludes to any information question that has not experienced careful preparation, either physically or through automated computer programs. Raw data may be assembled from different forms and IT assets.

As a result of this definition, understanding the data that is being issued is considered one of the primary keys before development. In this part, if a clear understanding of the type of data the clients or DAs supply, it will not only ease the development, but it will save an abundant amount of time and effort, so the same job will not be done twice and maybe more due to the redundancy of the data that is there. With a meeting or just viewing a portion of data can be considered as an option.

A meeting was scheduled with the DA responsible for this project which allowed for the review of this data. It was acknowledged that the data was coming from three different sources, Excel files, CSV files, and SPs from the current server or even through a linked server.

2.1.4. KPIs

"KPI stands for key performance indicator, a quantifiable measure of performance over time for a specific objective. KPIs provide targets for teams to shoot for, milestones to gauge progress, and insights that help people across the organization make better decisions. From finance and HR to marketing and sales, key performance indicators help every area of the business move forward at the strategic level (J. Morrow 2020)."

These KPI measurements are usually submitted by the DAs. Based on these formulas, the KPIs are calculated on the backend side. Later in this report, more details will be covered on how these KPIs were being computed.

$ Make Good Amt	ART	CSAT	KB Usage Skill	Shrinkage System
$ Make Good Amt Skill	ART Skill	CSR1 Avg. Wait Time	MG Rate	Shrinkage Training
Abandon Rate	ASA Secs	Customer Satisfaction	MG Rate Skill	SLA
Abandon Rate Skill	ASA Secs Skill	Escalation Rate	RCR	SLA Skill
Absenteeism	Avg. Concurrency	Escalation Rate Skill	RCR Skill	Utilization
Actual vs Locked €	Avg. MG $	FC Accuracy (Locked)	Service Quality	Volume Contribution
Actual vs Locked H	Avg. MG $ Skill	FC Accuracy (Revised)	Shrinkage	
Adherence	Bill to Pay	FTEs	Shrinkage Away	
AHT	Call Accuracy	IVR Release Rate	Shrinkage Break	
AHT Skill	CRES	KB Usage	Shrinkage Coaching	

Figure 6 – *KPIs that were built for this specific project.*

These KPIs are divided into 6 different group categories, each category takes insights from the data in a different perspective. Here is the list of the ones that were used:

- **Operational:** a discrete estimation that a company employs to screen and assess the effectiveness of its day-to-day operations.
- **WFM:** Workforce management reporting for modern businesses.
- **Forecast:** Forecasting KPIs and performance measures is about finding leverage. And that is why building a forecasting model for a KPI, or performance measure is so valuable: it makes a difference when an impact is found.
- **Shrinkage:** This KPI is utilized to measure the rate at which the esteem of stock has been diminished due to loss, burglary, or wrong record keeping.
- **Quality:** A quantitative measure of data quality. A data quality measurement system measures the values for the quality of data at estimation and focuses on a certain recurrence of measurement.
- **Financial:** A leading high-level measure of revenue, expenses, profits or other financial outcomes.

2.1.5. Expected Output

After going through all the initial steps to decide on the expected output, now is the time to move forward discussing it. But before continuing, a deliberation needed about integrating the data files first from their initial files to SQL to check how to store the data and in which format. Because there was a close deadline for the project, going with the most trivial solution was mandatory to transform the files into a table.

As a result, the significant and suitable proposition is to have a “Star Schema” to produce a final fact table with all the columns used to calculate the KPIs mentioned before.

This solution will provide all the data needed without affecting the deadline.

3. PRACTICAL SOLVING

In this part, two solutions will be provided. The first one is the one that was adapted primarily. Because some issues were faced, and a more generic solution was needed, another solution was established. A detailed explanation is supplied to reveal the hows and the whys.

3.1. FIRST SOLUTION

3.1.1. SSIS Flow

First, what does ETL mean? ETL stands for Extract transform and load. “An ETL device extracts the information from unique RDBMS source structures, transforms the data by making use of commercial enterprise common sense, concatenate, and so on (Raghuraman 2021)”.

In this part, it will be explained in detail the steps set up to transform the data from their initial state to the database. After creating the SSIS Solution, one can check the file, its location; if accessible or not, take a quick look at the file to have a rapid scan of what are the columns, if there exist multiple sheets, and importantly if it is corrupted or not.

After that, a server connection is created, and of course tested to check if it is built between the SSIS and the corresponding database. But because there is potential for more than one file to be executed, one is able to create a foreach loop to go through all the files placed in their respective folder.

In that loop, a data flow is placed. There are only two options in this case with these kinds of dataflows; one with a flat file source and this indicates that the file that is being processing are type csv or excel source file on which it has any of the excel extensions (xlsx, xls, xlsm …). Another source of data is existing data in other databases/servers (which will be discussed later in this thesis). As a result, after specifying the data source type, data should be loaded into its specific table. For governance and security reasons, a schema was created specifically for this project so the employees

who need admittance to these tables can be granted access to this schema and not the whole database. These tables will be created on spot after mapping the file in SSIS. The primary keys of these tables were created with the assistance of the DAs for their knowledge with these extractions. And after finalizing this step, the first part of the data flow is done.

After that, a SQL task is introduced to the control flow to make sure that the data is clean and grant credibility of the stored data. The aim of using such procedures is to ensure that the specified key do not overlap and store duplicate values in the tables and therefore wrong measures. These procedures operate by selecting the data first and inserting them into a processing table. Then check the key with the FACT table, if a match happens then update the data otherwise upload it as is.

After the procedure gets executed, the file is now removed to an archive file with the same format but with today's date added to its name to keep track over the files processed and to help re-track errors if it is the case and to prevent files from overlapping.

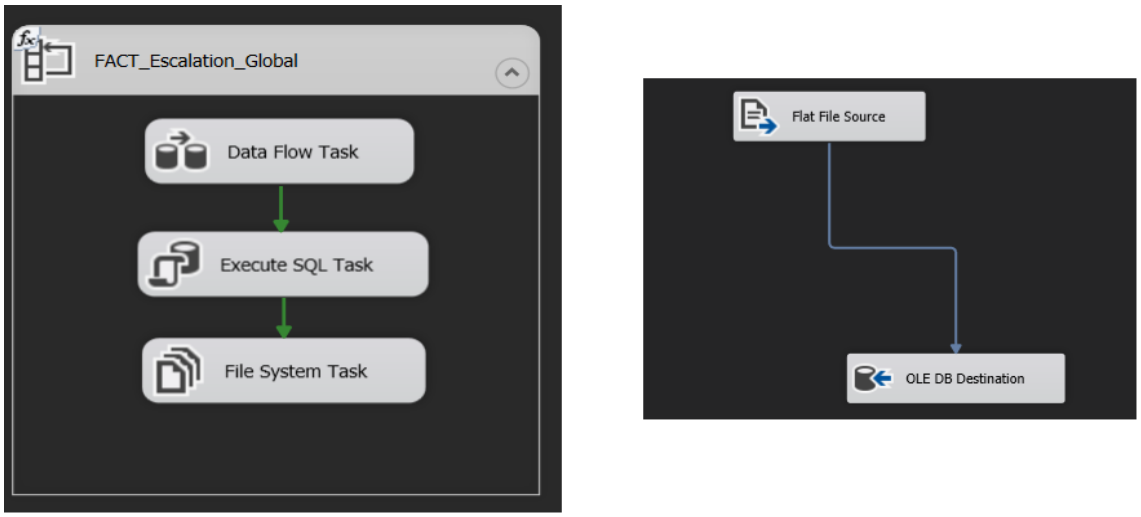


Figure 7 – *A scheme of the SSIS loop components from control flow to data flow.*

And here is an example of how the SQL task is configured:

```
UPDATE
    [DataAnalytics].[ProcessingNFX].[tb_FACT_GlobalIVRSurvey]
SET
       [ProcessingNFX].[tb_FACT_GlobalIVRSurvey].[UTC Date] = b.[UTC Date]
      ,[ProcessingNFX].[tb_FACT_GlobalIVRSurvey].[Skill] = COALESCE(b.[skill], 'NULL')
      ,[ProcessingNFX].[tb_FACT_GlobalIVRSurvey].[Site] = COALESCE(b.[Site], 'NULL')
      ,[ProcessingNFX].[tb_FACT_GlobalIVRSurvey].[Question] = COALESCE(b.[Question], 'NULL')
      ,[ProcessingNFX].[tb_FACT_GlobalIVRSurvey].[score_1_cnt] = b.[score_1_cnt]
      ,[ProcessingNFX].[tb_FACT_GlobalIVRSurvey].[score_2_cnt] = b.[score_2_cnt]
      ,[ProcessingNFX].[tb_FACT_GlobalIVRSurvey].[score_3_cnt] = b.[score_3_cnt]
      ,[ProcessingNFX].[tb_FACT_GlobalIVRSurvey].[score_4_cnt] = b.[score_4_cnt]
      ,[ProcessingNFX].[tb_FACT_GlobalIVRSurvey].[score_5_cnt] = b.[score_5_cnt]

FROM
    [ProcessingNFX].[tb_FACT_GlobalIVRSurvey] a
INNER JOIN
    [DataAnalytics].[processing].[ProcessingNFX_tb_FACT_GlobalIVRSurvey] b
ON
   a.[utc date] = b.[utc date]
      AND COALESCE(a.[skill], 'NULL') = COALESCE(b.[skill], 'NULL')
      AND COALESCE(a.[site], 'NULL') = COALESCE(b.[site], 'NULL')
   AND COALESCE(a.[Question], 'NULL') = COALESCE(b.[Question], 'NULL')

INSERT INTO [ProcessingNFX].[tb_FACT_GlobalIVRSurvey]
SELECT
a.[UTC Date],
COALESCE(a.[skill], 'NULL')                    [Skill],
COALESCE(a.[site], 'NULL')                     [Site],
COALESCE(a.[Question], 'NULL')     [Question],
COALESCE(a.[score_1_cnt], 'NULL')    [score_1_cnt],
COALESCE(a.[score_2_cnt], 'NULL')    [score_2_cnt],
COALESCE(a.[score_3_cnt], 'NULL')    [score_3_cnt],
COALESCE(a.[score_4_cnt], 'NULL')    [score_4_cnt],
COALESCE(a.[score_5_cnt], 'NULL')    [score_5_cnt]
FROM   [DataAnalytics].[processing].[ProcessingNFX_tb_FACT_GlobalIVRSurvey] a
       LEFT JOIN [DataAnalytics].[ProcessingNFX].[tb_FACT_GlobalIVRSurvey] b
              ON a.[utc date] = b.[utc date]
                 AND COALESCE(a.[skill], 'NULL') = COALESCE(b.[skill], 'NULL')
                 AND COALESCE(a.[site], 'NULL') = COALESCE(b.[site], 'NULL')
     AND COALESCE(a.[Question], 'NULL') = COALESCE(b.[Question], 'NULL')
WHERE  b.[utc date] IS NULL
       AND b.[skill] IS NULL
       AND b.[site] IS NULL
    AND b.[Question] IS NULL

TRUNCATE TABLE [DataAnalytics].[processing].[ProcessingNFX_tb_FACT_GlobalIVRSurvey]
```

Figure 8 – *Example of a code for duplicate check procedure.*

This process is done for all the folders that provide raw data files. After the data gets stored in the processing tables, then a procedure is executed and measurements are calculated.

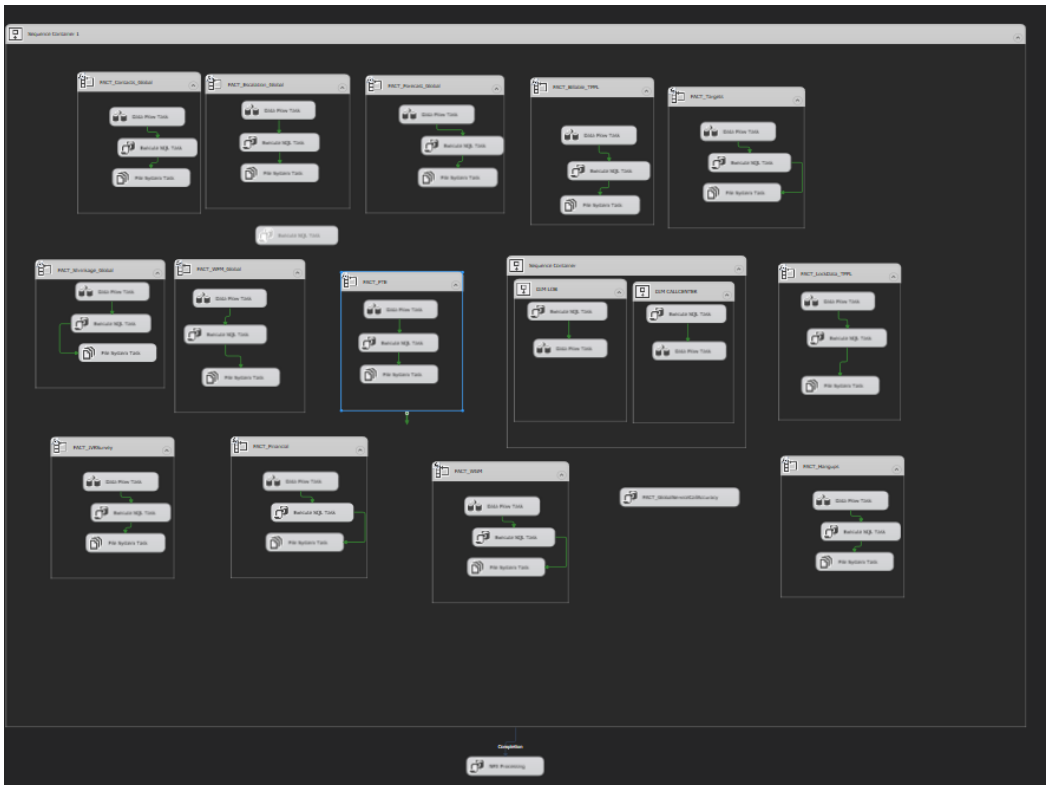


Figure 9 – *A scheme of the entire SSIS integration process.*

3.1.2. Mapping and Processing

A mapping file is then provided by the DA responsible for this project. In this file, one can determine what are the KPIs that need to be calculated and the components needed for these measurements to be generated. These mappings are then stored in a table to be used in the cube metrics.

In this definition table, the measurements are specified based on a type (addition, subtraction, division, and/or multiplication). So, the type is provided for the operation as well as the components to be calculated.

	type	idkpi	KPI Group	description	description1	description2	description3	component1	component2	component3	component4
1	1	1	Operational	Sum([# Make Good Amnt])	Sum(coalesce(cast(component1 AS DECIMAL(28,6)),0))	NULL	NULL	[# Make Good Amnt]	NULL	NULL	NULL

Figure 10 – *SQL table showing an example of how the data is being calculated.*

In Figure 10, one can see that the type of this operation (type 1) is summation for only one component.

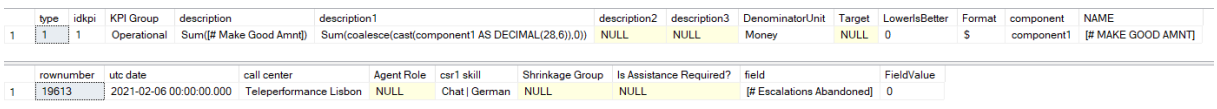
After the files have been integrated into their respective tables, a stored procedure is created to join all the data in one FACT table.

3.1.3. Fact table

This part will cover thoroughly how the process was built, how it functions, and what are the challenges that were faced with coding or data itself.

Starting with how this process was built, and after knowing how the dimensions look, the ideal strategy was to see how to combine all the data found into a global table that contains all the columns required to measure the requisite KPIs. As a first step, all the data was saved in temp tables, and because one needs to separate the columns that are needed for calculations and the others that will be used for joins later, pivoting must be used to achieve that goal. Pivoting is a mechanism used to transform columns into rows. So, with this, now multiple datasets have almost the same key but with different field values.

After this transformation is completed, one can unite all the tables based on the key and obtain fields that have the names of the columns and field values that possess the values stored in that column. As a result, one can obtained a table in which all the columns are separated with their respective values. Moving to the next step, the same is performed but with the definition table that was mentioned before (3.1.2. Mapping and Processing Figure 11). These two tables share the same structure, one for the data and the other for the component distribution.



	type	idkpi	KPI Group	description	description1	description2	description3	DenominatorUnit	Target	LowerIsBetter	Format	component	NAME
1	1	1	Operational	Sum([# Make Good Amnt])	Sum(coalesce(cast(component1 AS DECIMAL(28,6)),0))	NULL	NULL	Money	NULL	0	$	component1	[# MAKE GOOD AMNT]

	rownumber	utc date	call center	Agent Role	csr1 skill	Shrinkage Group	Is Assistance Required?	field	FieldValue
1	19613	2021-02-06 00:00:00.000	Teleperformance Lisbon	NULL	Chat | German	NULL	NULL	[# Escalations Abandoned]	0

Figure 12 – *A scheme to show pivoting transformations.*

After collecting the two tables, one is able to join based on the field column. Although it is known that this solution was not the optimal because the join is based on a “String” column, and indexing

cannot be performed over a varchar; but this was the most suitable one in this case and the only available column to do the join on. As a result, all the columns from idkpi are collected and needed to calculate each KPI separately, in addition to the components which will be used with their respective values. (The two rows in fig11 are now a combined row)

In the next step, pivoting is implemented to the table again on the components. The column component now will be transformed into multiple columns and each cell in that column will possess its particular value.

idkpi	description1	description2	utc date	call center	Agent Role	csr1 skill	Component1	component2	component3
15	Sum(coalesce(cast(component1 AS DECIMAL(28,6)),0))	Sum(coalesce(cast(component2 AS DECIMAL(28,6)),0))	2019-10-25 00:00:00.000	Teleperformance Athens	CSR1	Voice | German	0.0000	870.0000	NULL
16	Sum(coalesce(cast(component1 AS DECIMAL(28,6)),0))	Sum(coalesce(cast(component2 AS DECIMAL(28,6)),0))	2020-01-31 00:00:00.000	SKILL	NULL	Voice | English Asia	1497.0000	3345.0000	NULL

Figure 13 – *Table showing the distribution of components and their values per KPI.*

After the previous transformation, a query is executed to do the calculations per KPI. this query operates by selecting the components and their values. Once established, the query can be passed to the descriptions found in Figure 12. These are SQL statements that intend to swap the values of the components and sum them instead.

Later in this process, couple of tables are joined to append other ones needed to calculate more KPIs, additionally to provide option to calculate the extra columns requested by the DAs. With that being said, and after all the requirements are fulfilled to evaluate the final table, these tables are joined all together based on their [Date] and other components. One can then delete the values based on the minimum date in the table and push the data to the final fact table.

idkpi	Region	Country	TP Site	Client	LOB	SubLob	Role	Market/Language	channel	Billing Type	WAHA/BM	Date	Numerator	Denominator	Denominator Unit	KPI Group	KPI	KPI NAME
2	EMEA	Turkey	Teleperformance Turkey	Netflix	Chat	Chat | Turkish	CSR1	Turkish	Chat	NULL	NULL	2020-11-10	4.000000	1.000000	Money	Operational	Avg. MG $	Avg. MG $

Figure 14 – *Table showing each KPI with their Numerator and Denominator Values.*

3.1.4. Problem Solution

Regarding this method, a lot of problems were encountered with the files, data, and their values. Some of the files were uploaded with many problems, sometimes the sheet name changes, and this will affect the mapping because once an excel file is mapped in the data source, the excel sheet name is specific and if it is not there, the data flow will crash. Another problem was the delimiter; sometimes the delimiter either changes and messes up with the distribution of the data with their columns, or some cell values contains the delimiter; for example, if the delimiter is a ',' and a cell has some values with ',' like specifying a list, then the values for the consecutive columns will be absolutely wrong. Regarding the values, extra checking was performed when inserting them from the processing schema to their dimensions by replacing spaces by empty space and commas to dots.

Another problem was faced with date format, some files had different formats; sometimes dd/MM/YYYY and others MM/dd/YYYY and this would lead to KPIs' miscalculations, so unifying the date format with the DAs was a must.

Occasionally, some columns might share the same name in different tables/files, and because the join is based on the column name, the values might overlap and conduct false values. As a solution for that, an additional letter was added to the end of the columns per file which leads to distinctive names, and this will not be an issue anymore.

Integrating the whole file will be excessive use of space in the server so a decision was taken to go with the components that are needed to conduct these measures. This option was catastrophic for calculation, but it was acknowledged once the comparison between the raw data and the data that was stored in the database took place. The reason why is because now multiple columns might share the same values and on joining them, a lot more will be missed. As a proposition, an increment row number was added for every file to keep all the columns, and eventually this issue will not be faced anymore.

With that now the first solution was ready to go live.

3.2. SECOND SOLUTION

In this part, the criteria will not be discussed in detail regarding the creation of the basic steps because they are exactly the same and it has been already covered (3.2). On the other hand, the whole idea is different and that is what will be focused on here. This chapter will start by mentioning why the decision has been made to develop another solution, and what are the advantages that were achieved based on this change.

3.2.1. Problems and Prerequisites

First of all, a lot of problems were encountered regarding the first solution. The data was always changing from the client's side, and manual intervention was required on the SSIS part to perform the steps all over again to adapt to these changes. Sometimes the column names change, sheet names, columns were removed, and date formats might change as well so another component (Derived Column) is added in the SSIS data flow to do the changes needed. Some files used to come with duplicates and every time this happens, looking back to the SSIS was essential to examine what was the error and then go explore the raw data files and remove the duplicate because no duplicate values should be inserted in the processing tables based on the keys specified by the DAs.

In this approach, organizing the whole process was critical to create a more generic solution; that does not only work for this project but for any integration process that will be conducted in the future. For that reason, two databases were created, one for the live data and the other for historical one. By doing this, the server will not withhold massive amounts of data so the extraction lately will be smoother. After that, a file was handed to the DA that requested the new integration. This file will be filled by him/her and then handed to the data engineering department to change some hidden tabs that will allow us to create the necessary tables and values to fulfill the requirements. This file has all the info needed to move to the next stage which is creating all the essentials to start building the data flow.

REQUIREMENT	
* Project	
* Short description of the file	
* Schedule Start Date	
* Schedule Days of Week	
* Desired Schedule Time	
* Schedule End	
Point of Contact	
*Operative POC Full Name	
*POC Email	
SOURCE	
File format	
Sheet name (if applicable)	
Average rows per day	
Has sensitive Information	
Extra comments	
PROFILE	
*Required days of Available Data on Server	
Please justify If this requirement will need historical data that exceeds over three months :	
***** APPROVAL *****	
DE Coordinator / DS Manager	

Figure 15 – *Table to provide the DEs for all necessary factors to create the process.*

Source Name [redacted]

Source Column Name *	Destination Column Name *	Data Type *	Historical column	Unique columns	Date format
UTC Date	UTC Date	date	1	1	103 - dd/mm/yyyy
Call Center	Call Center	nvarchar(2550)		1	

Figure 16 – *Portion of the table structure.*

This file has another sheet that possesses a small portion of the data from the raw file, so in case a check over the data structure is needed for any reason, there is no need to open the original file, or if a comparison between the files is demanded to check if some changes took place.

This excel file is connected to the SQL server, and there is a tab related to extension id. This id will be the key identification between all the processes. Once this tab is refreshed a new unique

extension id will be spawned and a SQL code is generated based on all the sheets that were filled by the DA. This code holds within it the creation queries for the tables.

After running the query, a table is established for that file which is created on the current database and the historical one and the details of each file/table will be something like Figure 16.

Ext_ID	Data_Folder	Historical_Folder	Destination_Table	Client_abb	Description	Days_To_Save	File_Type	Sheet_Name	Extra_Columns	IsActive	deletion_mode	n_columns	unpivot
14	\\office.teleperformance.pt\pt\Lisbon-CIT\Depart...	\\office.teleperformance.pt\pt\Lisbon-CIT\Depart...	[DataAnalytics].[NETF].[tb_inbd_billable]	P.NETF.INBD	P.NETF.INBD Billable	365	xlsx	Sheet1$	NULL	1	0	11	NULL

row_id	ext_id	file_column	table_column	isActive	historical_column	unique_column	date_format	unpivot
171	14	Activity	Activity	1	0	1	0	NULL
172	14	Month	Month	1	0	0	0	NULL
173	14	Date	Date	1	1	1	110	NULL
174	14	Language	Language	1	0	1	0	NULL
175	14	Role	Role	1	0	1	0	NULL
176	14	Skill	Skill	1	0	1	0	NULL
177	14	Billable hours w/o cap	Billable hours w/o cap	1	0	0	0	NULL
178	14	Revenue w/o cap	Revenue w/o cap	1	0	0	0	NULL
179	14	Billable hours w/ cap	Billable hours w/ cap	1	0	0	0	NULL
180	14	Revenue w/ cap	Revenue w/ cap	1	0	0	0	NULL
181	14	Rate/h	Rate/h	1	0	0	0	NULL

Figure 17 – *Creation query result.*

Over here there is the extension id, file location, archive location, table where the data will be inserted in, client's name, days to keep in the database before moving the data to the historical one, file type, sheet name if applicable, check if it is an active solution or not, and finally the number of columns.

In the second table, all the column names with their extension id come across, specifying each ones' date type and what format the column has, and also, the key needed for duplicate check either single or combined one depends on the file itself.

3.2.2. SSIS Flow

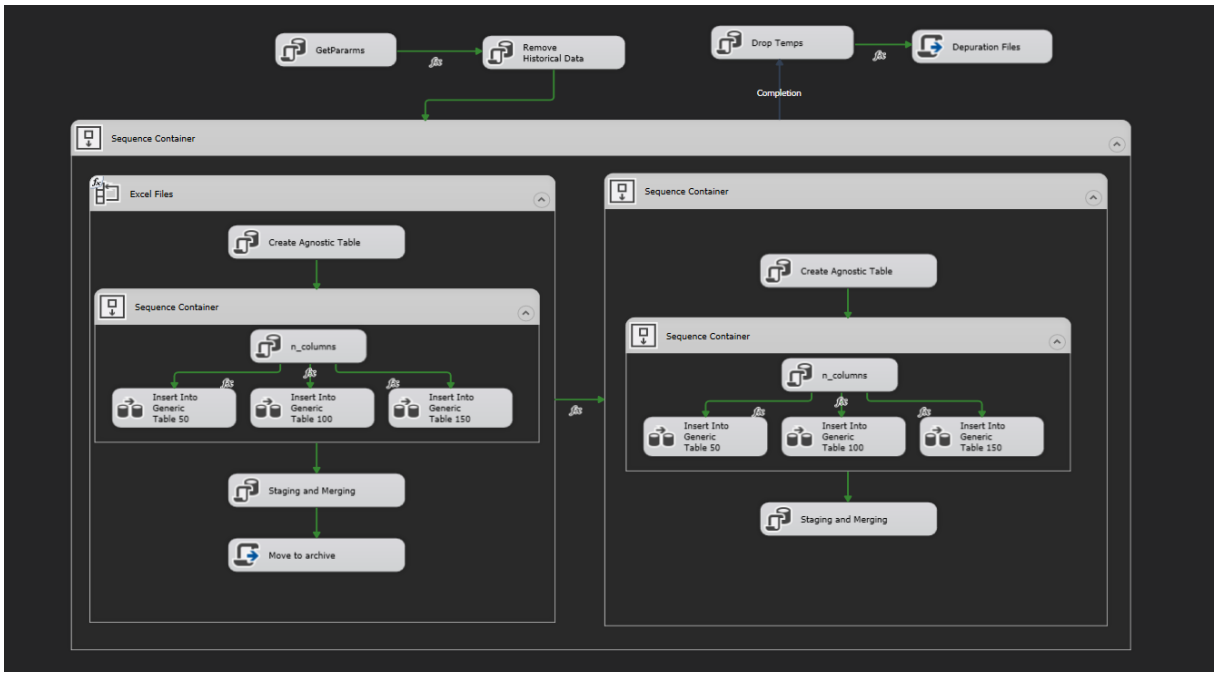


Figure 18 – *Showing the SSIS Data Flow.*

First, the integration process is established by getting the parameters needed. These parameters are [File_Type], [Data_Folder], [Historical_Folder], [Destination_Table], [Sheet_Name], and [IsActive]. After that, a check for how many days is decided to save in the current database is taken into consideration to remove the data that exceeds the number of days from the production database to the historical one. Moreover, and because there are multiple types of data that can go through integration, the type of ones used is either a file or SQL code. In case it was a file, it goes through the left loop of the process, otherwise it goes to the other.

3.2.3. Flat File Source

As introduced before, the aim of this project was to create a more generic solution. The next part is the most important key to achieve that goal. The agnostic table part is the creation of a temporary table with empty columns that will be later filled while fetching the file. Extraction of the number of columns occurs because this leads to minimizing the time the process takes to run. There

is no need to create a lot of columns for a 10 columns file. After that, insert to generic component is where all the magic takes place. Over here, a C# script runs to check first the type of the file, then will check for the sheet name to match the one that is in the database and check for the delimiter used in that file. Then the data in that sheet will be filled in an adapter and stored in a temporary table. After all these steps, a data set is provided with all the data with their respective columns, but the file header will be on the first row under the [column] column and then the data is stored in the agnostic table that was created before.

When this occasion is accomplished successfully, a stored procedure is executed to store the data from that temp table to the destination table specified in Figure16 based on the extension id supplied in the parameters. In this procedure, all the checking and validation needed to load the data are carried out. It is initiated by creating a table with all the required columns and is named dependent on that id to prevent having multiple tables that share the same name and consequently facing SQL issues, also to allow multiple executions to occur simultaneously. Then the data is inserted where row number is greater than one to have the original header for the file. The date format is examined for the dates that are stored, if any, and then check the data if contains duplicate values on the key. If a match is obtained, update the values for the remaining columns otherwise insert all the data. Subsequently, the file will then be removed to archive with today's date added to the end of it. The loop keeps going with the same process until the containing folder is empty.

3.2.4. SQL Source

After acknowledging that the type of the file is SQL, the behavior is a bit different in the data flow, but the steps are exactly identical. Instead of putting the location of the data folder in SQL, the query to be execute is provided then the data that was collected is added to a temporary table that was created and the process continues with the same strategy.

After storing the data in their destined tables and moving the files to the archive folder, these temporary tables that were created will be dropped, therefore the need for the processing schema that was necessary in the first part is not anymore needed which leads to less memory use. And finally, a C# script will run to remove the files that their date exceeds a week which will also lead to less load on the server. A week was chosen because in case there is problem in the file integration or data, then it should be detected within that time limit.

3.2.5. Job Creation

Because parameters are provided to the SSIS process, job creation is possible. This will prevent the user intervention by running these processes manually. Primarily, the creation of these jobs starts by constructing the steps demanded to perform all of them. In each of these steps, connection is established first to the SSIS solution and then the extension id related to this specific integration is supplied. After preparing everything, and based on the DA's request, this job is scheduled on the time agreed between the two parties.

With this, the process is finalized and ready to go live.

3.2.6. Advantages of The Second Solution

Regarding the second solution, a lot of benefits and advantages are acquired that were missed in the first one. For instance, this solution is a more generic one which allows multiple projects to use the same integration process. All what need to be done is specify the extension id and accordingly the file will be integrated. Using this method, the elimination of multiple data flows and sticking to only one generic flow is achieved. Moreover, by creating a temporary table and deleting it after looping through the files to be integrated, the server will have a lot of extra temp space because none of them are needed anymore. After removing the file to the archive folder, and

deleting the single temp table created, a script will be executed to check the dates of the files that were integrated and delete those that their dates exceed a week to reduce space on the shared server. And thanks to the PDR files, a significant reduction of errors was detected mostly because the DAs are now responsible for filling all the information needed for the file to be integrated. (They are more in contact with the client and more aware of the scenarios that might be face in the future). Not only that, but the margin of errors was minimized especially with the date formats, because some files might come with mm/dd/yyyy format and other might be dd/mm/yyyy. The fact table will therefore lack integrity. As a result, with these PDRs, the DAs specify the format of the date type, so in case any file shared different format, it should be checked.

A nice feature was adapted to the new solution which is the ability of adding extra columns that will be added manually to the data flow. For example, if someone wants to create a new column based on concatenating multiple columns, it is possible by specifying that in a SQL code and adding it in the table which will serve its purpose.

This diagnostic process allowed multiple files that share the same columns but in a different order to be processed together. The manual mapping will not be needed because the generic temp table created will match later with the column name which is in the header of the file. Not only that, but the process was finalized by eliminating the user intervention and creating scheduled jobs. All what needed is to monitor the job activities and make sure everything is working smoothly.

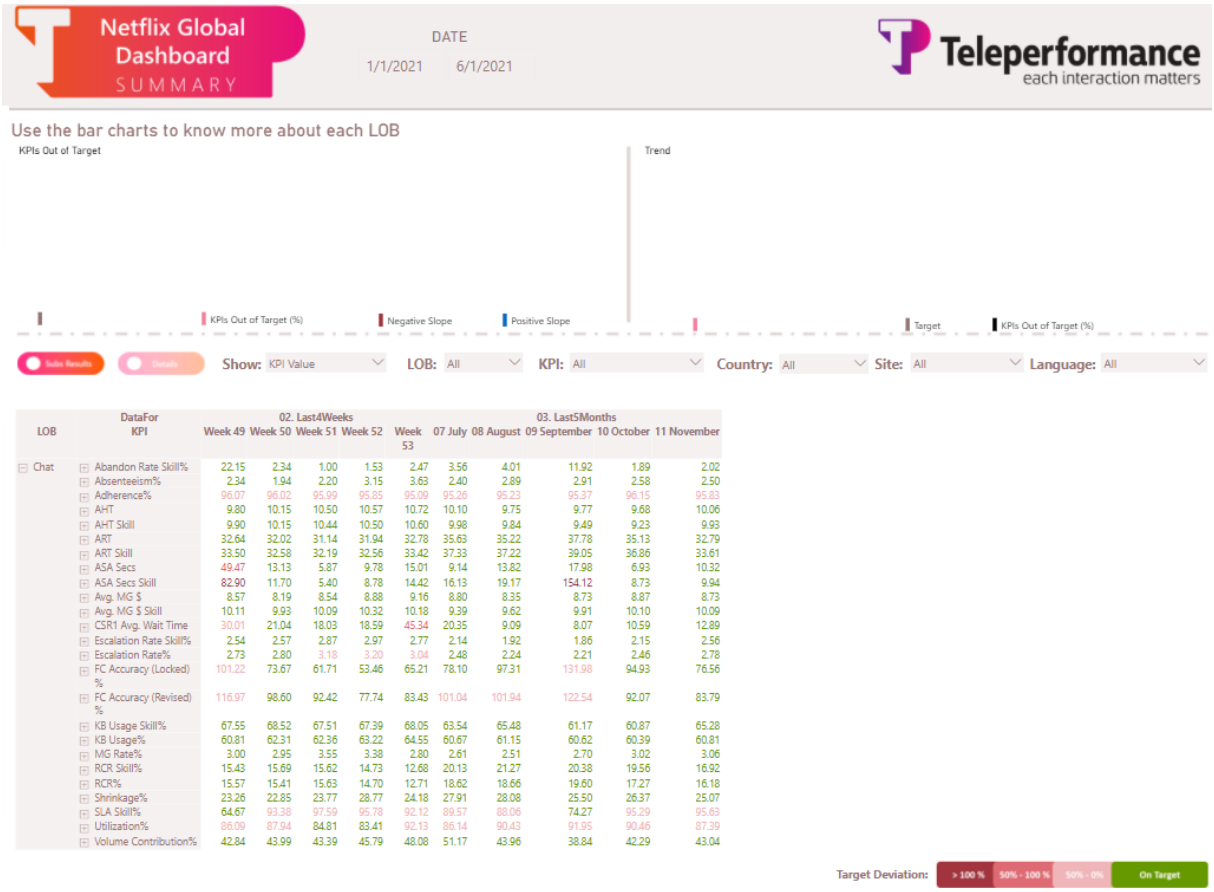
4. FEATURES TO ADD

After finalizing this solution, some features can be included to achieve more stability and integrity to the data. Adding a table to keep track of the rows and columns count with their integration date is a constructive idea. This will ensure the data integrity and will apply to an integration error if these numbers were not matching. This idea can be extended by creating a PowerBi report that will keep track of the jobs. While testing the integration process, a bug was found and needs to be fixed. This scenario occurs if the file was type csv. With this specific type, a delimiter should be defined. However, if the file has only 2 columns, this delimiter is hard to catch in the C# script which might lead to wrong data integration.

In addition, if a file has no key, a method should be created to handle this. A way of fulfilling that is by selecting the date range in the file, deleting data from the respective table with the same date range and merge the data from the file to the table. After merging and executing of the jobs, an email can be sent with the status either completed with success or failure, with the error and the extension id so follow up can be performed. In case of job failure, insert into a table the exception that was thrown while executing that job to have a sort of idea what are the exceptions that are being generated.

5. REPORTING

5.1. SUMMARY



LOB	DataFor KPI	02. Last4Weeks Week 49	Week 50	Week 51	Week 52	Week 53	03. Last5Months 07 July	08 August	09 September	10 October	11 November
Chat	Abandon Rate Skill%	22.15	2.34	1.00	1.53	2.47	3.56	4.01	11.92	1.89	2.02
	Absenteeism%	2.34	1.94	2.20	3.15	3.63	2.40	2.89	2.91	2.58	2.50
	Adherence%	96.07	96.02	95.99	95.85	95.09	95.26	95.23	95.37	96.15	95.83
	AHT	9.80	10.15	10.50	10.57	10.72	10.10	9.75	9.77	9.68	10.06
	AHT Skill	9.90	10.15	10.44	10.50	10.60	9.98	9.84	9.49	9.23	9.93
	ART	32.64	32.02	31.14	31.94	32.78	35.63	35.22	37.78	35.13	32.79
	ART Skill	33.50	32.58	32.19	32.56	33.42	37.33	37.22	39.05	36.86	33.61
	ASA Secs	49.47	13.13	5.87	9.78	15.01	9.14	13.82	17.98	6.93	10.32
	ASA Secs Skill	82.90	11.70	5.40	8.78	14.42	16.13	19.17	154.12	8.73	9.94
	Avg. MG $	8.57	8.19	8.54	8.88	9.16	8.80	8.35	8.73	8.87	8.73
	Avg. MG $ Skill	10.11	9.93	10.09	10.32	10.18	9.39	9.62	9.91	10.10	10.09
	CSR1 Avg. Wait Time	30.01	21.04	18.03	18.59	45.34	20.35	9.09	8.07	10.59	12.89
	Escalation Rate Skill%	2.54	2.57	2.87	2.97	2.77	2.14	1.92	1.86	2.15	2.56
	Escalation Rate%	2.73	2.80	3.18	3.20	3.04	2.48	2.24	2.21	2.46	2.78
	FC Accuracy (Locked) %	101.22	73.67	61.71	53.46	65.21	78.10	97.31	131.98	94.93	76.56
	FC Accuracy (Revised) %	116.97	98.60	92.42	77.74	83.43	101.04	101.94	122.54	92.07	83.79
	KB Usage Skill%	67.55	68.52	67.51	67.39	68.05	63.54	65.48	61.17	60.87	65.28
	KB Usage%	60.81	62.31	62.36	63.22	64.55	60.67	61.15	60.62	60.39	60.81
	MG Rate%	3.00	2.95	3.55	3.38	2.80	2.61	2.51	2.70	3.02	3.06
	RCR Skill%	15.43	15.69	15.62	14.73	12.68	20.13	21.27	20.38	19.56	16.92
	RCR%	15.57	15.41	15.63	14.70	12.71	18.62	18.66	19.60	17.27	16.18
	Shrinkage%	23.26	22.85	23.77	28.77	24.18	27.91	28.08	25.50	26.37	25.07
	SLA Skill%	64.67	93.38	97.59	95.78	92.12	89.57	88.06	74.27	95.29	95.63
	Utilization%	86.09	87.94	84.81	83.41	92.13	86.14	90.43	91.95	90.46	87.39
	Volume Contribution%	42.84	43.99	43.39	45.79	48.08	51.17	43.96	38.84	42.29	43.04

Figure 19 – *Revealing a summary for the calculated KPIs.*

According to Figure 18, the bar chart shows the distribution of targets per LOB. Moreover, as a user, one not only has the ability to drill down through these targets on a level more than LOB itself, but also through the KPIs that are dedicated to it. All of these data are selected based on dates so there is the potential to choose what to look at and check the target deviation. On the bottom right there is the scale that will assist the user to differentiate between these deviations for a better understanding. In addition, if one presses on any of these KPIs, it will show the distribution per language.

So, for example if the desired case is to study the absenteeism LOB, all what has to be done is click on it and it will show something like Figure 19.

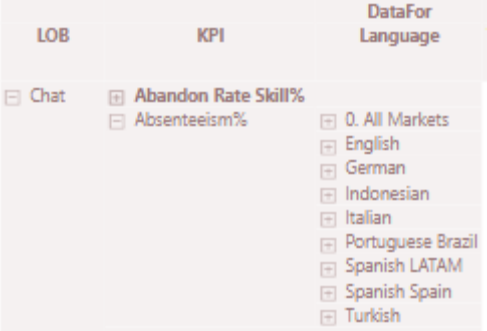


Figure 20 – *Drill down one of the KPIs.*

All of these data are provided by the final table that was already induced before.

5.2. TREND

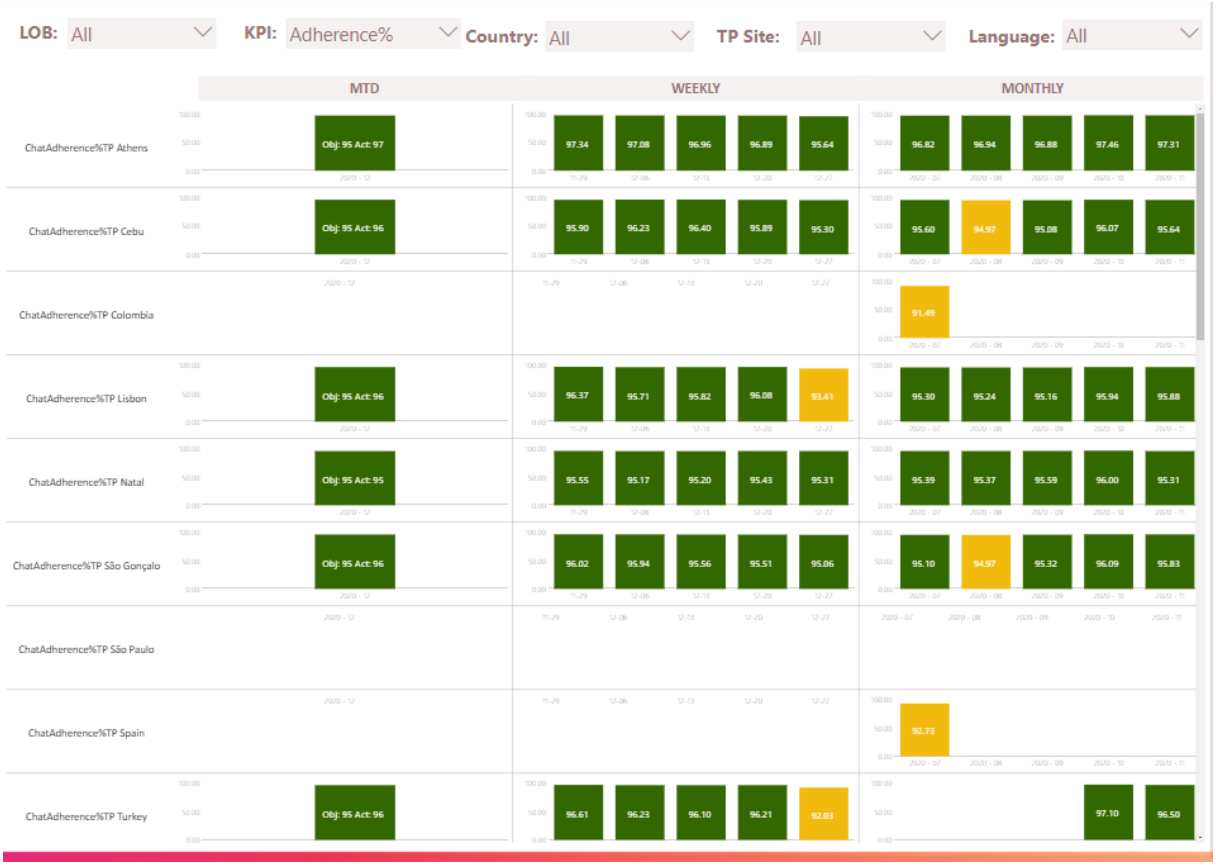


Figure 21 – *Bar chart showing the trend per LOB, KPI, Country, Site, and Language.*

5.3. KPI

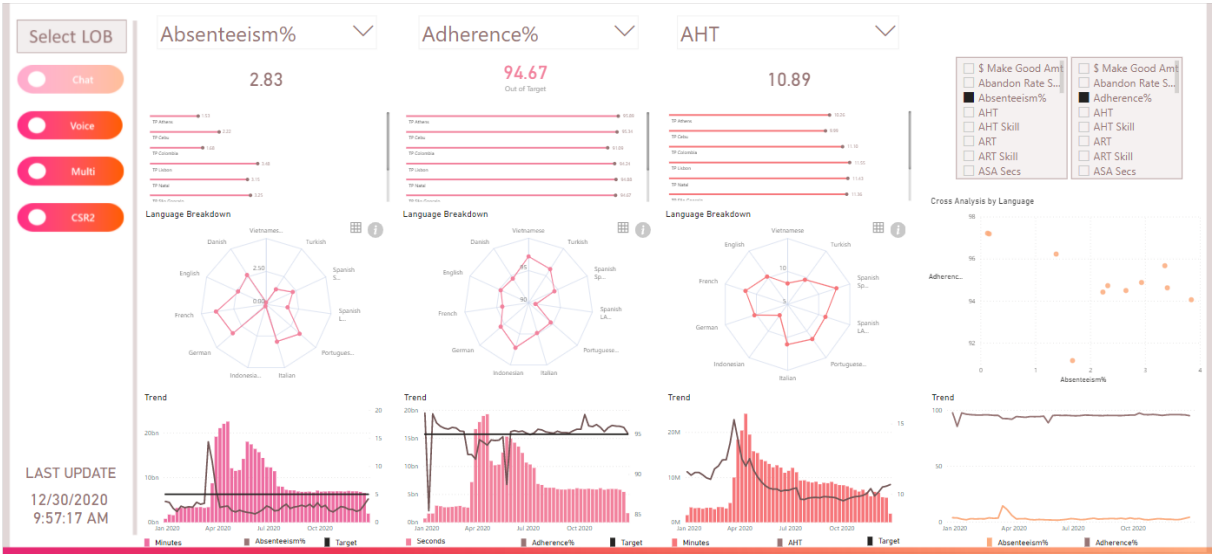


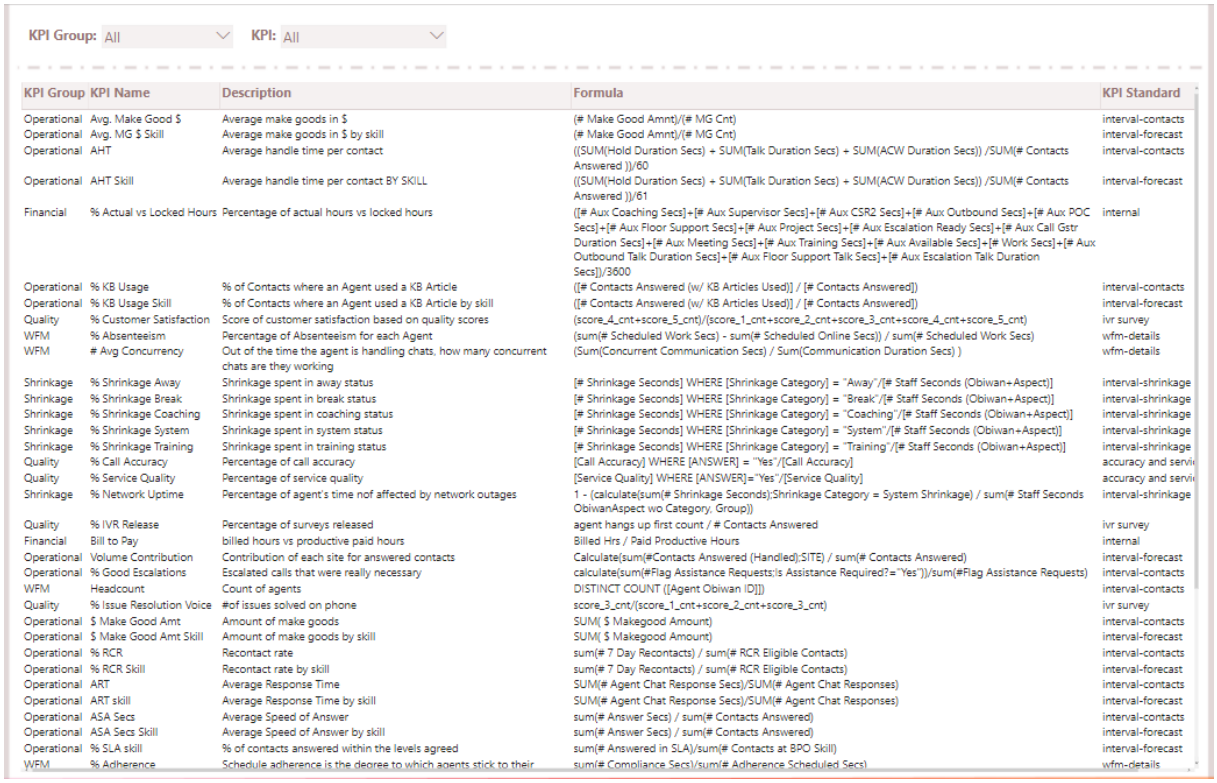
Figure 22 – *Graphs to compare multiple KPIs.*

In this part, a request was demanded to provide a way to compare multiple KPIs at once with specificities regarding languages and sites. This request was for the group to keep track of all these different sites at once and to check for problems or futuristic risks in any of the sites/languages. As seen in Figure 21 above, the comparison here is between absenteeism, adherence and AHT.

If pressed on any of the sites, the graphs will eliminate all other sites except the selected one and will show all the data regarding it. It is also feasible to choose a language and/or site to provide the users with a better understanding for whatever they are looking for.

5.4. GLOSSARY

In conclusion, a glossary was provided in case any DA that is new to the project or even a higher ranked position colleague would like to see the source of calculations and how they are done to answer any doubt. These glossaries are based on both the KPI group and the KPI itself.



KPI Group: All KPI: All

KPI Group	KPI Name	Description	Formula	KPI Standard
Operational	Avg. Make Good $	Average make goods in $	(# Make Good Amnt)/(# MG Cnt)	interval-contacts
Operational	Avg. MG $ Skill	Average make goods in $ by skill	(# Make Good Amnt)/(# MG Cnt)	interval-forecast
Operational	AHT	Average handle time per contact	((SUM(Hold Duration Secs) + SUM(Talk Duration Secs) + SUM(ACW Duration Secs)) /SUM(# Contacts Answered))/60	interval-contacts
Operational	AHT Skill	Average handle time per contact BY SKILL	((SUM(Hold Duration Secs) + SUM(Talk Duration Secs) + SUM(ACW Duration Secs)) /SUM(# Contacts Answered))/61	interval-forecast
Financial	% Actual vs Locked Hours	Percentage of actual hours vs locked hours	([# Aux Coaching Secs]+[# Aux Supervisor Secs]+[# Aux CSR2 Secs]+[# Aux Outbound Secs]+[# Aux POC Secs]+[# Aux Floor Support Secs]+[# Aux Project Secs]+[# Aux Escalation Ready Secs]+[# Aux Call Gstr Duration Secs]+[# Aux Meeting Secs]+[# Aux Training Secs]+[# Aux Available Secs]+[# Work Secs]+[# Aux Outbound Talk Duration Secs]+[# Aux Floor Support Talk Secs]+[# Aux Escalation Talk Duration Secs])/3600	internal
Operational	% KB Usage	% of Contacts where an Agent used a KB Article	([# Contacts Answered (w/ KB Articles Used)] / [# Contacts Answered])	interval-contacts
Operational	% KB Usage Skill	% of Contacts where an Agent used a KB Article by skill	([# Contacts Answered (w/ KB Articles Used)] / [# Contacts Answered])	interval-forecast
Quality	% Customer Satisfaction	Score of customer satisfaction based on quality scores	(score_4_cnt+score_5_cnt)/(score_1_cnt+score_2_cnt+score_3_cnt+score_4_cnt+score_5_cnt)	ivr survey
WFM	% Absenteeism	Percentage of Absenteeism for each Agent	(sum(# Scheduled Work Secs) - sum(# Scheduled Online Secs)) / sum(# Scheduled Work Secs)	wfm-details
WFM	# Avg Concurrency	Out of the time the agent is handling chats, how many concurrent chats are they working	(Sum(Concurrent Communication Secs) / Sum(Communication Duration Secs))	wfm-details
Shrinkage	% Shrinkage Away	Shrinkage spent in away status	[# Shrinkage Seconds] WHERE [Shrinkage Category] = "Away"/[# Staff Seconds (Obiwan+Aspect)]	interval-shrinkage
Shrinkage	% Shrinkage Break	Shrinkage spent in break status	[# Shrinkage Seconds] WHERE [Shrinkage Category] = "Break"/[# Staff Seconds (Obiwan+Aspect)]	interval-shrinkage
Shrinkage	% Shrinkage Coaching	Shrinkage spent in coaching status	[# Shrinkage Seconds] WHERE [Shrinkage Category] = "Coaching"/[# Staff Seconds (Obiwan+Aspect)]	interval-shrinkage
Shrinkage	% Shrinkage System	Shrinkage spent in system status	[# Shrinkage Seconds] WHERE [Shrinkage Category] = "System"/[# Staff Seconds (Obiwan+Aspect)]	interval-shrinkage
Shrinkage	% Shrinkage Training	Shrinkage spent in training status	[# Shrinkage Seconds] WHERE [Shrinkage Category] = "Training"/[# Staff Seconds (Obiwan+Aspect)]	interval-shrinkage
Quality	% Call Accuracy	Percentage of call accuracy	[Call Accuracy] WHERE [ANSWER] = "Yes"/[Call Accuracy]	accuracy and servi
Quality	% Service Quality	Percentage of service quality	[Service Quality] WHERE [ANSWER]="Yes"/[Service Quality]	accuracy and servi
Shrinkage	% Network Uptime	Percentage of agent's time nof affected by network outages	1 - (calculate(sum(# Shrinkage Seconds):Shrinkage Category = System Shrinkage) / sum(# Staff Seconds ObiwanAspect wo Category, Group))	interval-shrinkage
Quality	% IVR Release	Percentage of surveys released	agent hangs up first count / # Contacts Answered	ivr survey
Financial	Bill to Pay	billed hours vs productive paid hours	Billed Hrs / Paid Productive Hours	internal
Operational	Volume Contribution	Contribution of each site for answered contacts	Calculate(sum(#Contacts Answered (Handled):SITE) / sum(# Contacts Answered)	interval-forecast
Operational	% Good Escalations	Escalated calls that were really necessary	calculate(sum(#Flag Assistance Requests:Is Assistance Required?="Yes"))/sum(#Flag Assistance Requests)	interval-contacts
WFM	Headcount	Count of agents	DISTINCT COUNT ([Agent Obiwan ID]))	interval-contacts
Quality	% Issue Resolution Voice	#of issues solved on phone	score_3_cnt/(score_1_cnt+score_2_cnt+score_3_cnt)	ivr survey
Operational	$ Make Good Amt	Amount of make goods	SUM($ Makegood Amount)	interval-contacts
Operational	$ Make Good Amt Skill	Amount of make goods by skill	SUM($ Makegood Amount)	interval-forecast
Operational	% RCR	Recontact rate	sum(# 7 Day Recontacts) / sum(# RCR Eligible Contacts)	interval-contacts
Operational	% RCR Skill	Recontact rate by skill	sum(# 7 Day Recontacts) / sum(# RCR Eligible Contacts)	interval-forecast
Operational	ART	Average Response Time	SUM(# Agent Chat Response Secs)/SUM(# Agent Chat Responses)	interval-contacts
Operational	ART skill	Average Response Time by skill	SUM(# Agent Chat Response Secs)/SUM(# Agent Chat Responses)	interval-forecast
Operational	ASA Secs	Average Speed of Answer	sum(# Answer Secs) / sum(# Contacts Answered)	interval-contacts
Operational	ASA Secs Skill	Average Speed of Answer by skill	sum(# Answer Secs) / sum(# Contacts Answered)	interval-forecast
Operational	% SLA skill	% of contacts answered within the levels agreed	sum(# Answered in SLA)/sum(# Contacts at BPO Skill)	interval-forecast
WFM	% Adherence	Schedule adherence is the degree to which agents stick to their	sum(# Compliance Secs)/sum(# Adherence Scheduled Secs)	wfm-details

Figure 23 – *Glossary for all measurements used in the project.*

As a result of this work, the project was done. Some more KPIs might show in the future to be added but will follow the same procedure of calculation and visibility.

6. CONCLUSIONS

The data integration project took its first baby steps in February 2021 and began with a dive into the business strategies and a better understanding of what are the requirements the DAs need. Various meetings with them were done to align expectations, and to understand their goals.

Once these conditions were met, the production started with specifying the strategy it will overtake. First it began with checking the input file type, then establishing the SSIS process, and finalizing it with the final solution.

After having the final fact table, a report was generated to help visualizing the results in a friendly way, helping other untechnical colleagues filter and understand the data, allowing them to narrow down for a deeper dive through the data, and developing insights to improve the market and accordingly the profit.

This project permitted the utilization of numerous concepts introduced in the first year of the Master's in Data Science and Advanced Analytics. The process was built in ETLs and SQL, a software that became familiar in the Data Storage and Recovery course. Various business approaches as well were gained throughout the process especially the part related with the KPIs and their measurements and to what the imply to. All in all, the experience was fully related to the master's concepts and opened the gate for me to pursue a career as a data engineer.

7. LIMITATIONS AND RECOMMENDATIONS FOR FUTURE WORKS

During the development and after achieving the essential final fact table, a lot of limitations were faced and here are some of them. The first was in case the SQL agent went down for any reason, then this requires a user interference to run all the processes manually by providing the SSIS the extension id for each file. Additionally, the files are being processed in a C# script that goes through the file, checks the type and integrate it accordingly. In case the file was a csv, a delimiter should be defined and in case the file has only 2 columns, this delimiter is hard to catch.

On the other hand, some features can be included to achieve more stability and integrity to the data. A table can be added to keep track of the rows and columns count with their integration date. This will ensure the data integrity. By Sending an email with the status of the job once done, it will give the DAs a heads up about the issue so they will not have to get back to the DE department to know the status of the job and the error will be easier to be track and faster to be treated. In case of job failure, insert into a table the exception that was thrown while executing the job to have a sort of idea what are the exceptions that are being generated.

8. REFERENCE

- **B. Nelson and T. Olovsson, "Security and privacy for big data: A systematic literature review," *2016 IEEE International Conference on Big Data (Big Data)*, 2016, pp. 3693-3702.**
- **Colebatch HK (2014) Making sense of governance. Policy and Society 33(4): 307–316.**
- **H. Bhasin. (2020). What is Customer Experience Management and its Importance?**
- **J. Dinis-Carvalho. (2020). Presentation KPI. 10.13140/RG.2.2.32344.11521.**
- **J. Morrow (2020). Reporting Made Easy: 3 Steps to a Stronger KPI Strategy**
- **LINDELL, JIM. (2020). Big Data History – Big Data Sources and Characteristics. 2-1-2-22. 10.1002/9781119784692.ch2.**
- **M., Smallcombe. (2019) Top 5 Reasons to Centralize Data & Become a Data-Driven Business.**
- **M.T, Raghuraman. (2021). Assessment of ETL TOOLS ABSTRACT.**
- **M. Van Asselt, O. Renn (2011) Risk governance. Journal of Risk Research 14(4): 431–449.**
- **S. Kameliq (2021) Advantages of Data Mining for Digital Transformation of the Educational System.**
- **Shabana, Mahammad & Sharma, K. (2021). A STUDY ON BIG DATA ADVANCEMENT AND BIG DATA ANALYTICS.**