Business Intelligence Traineeship - KPMG Luxembourg

Submitted in Partial Fulfillment of the Requirements for the Degree of Masters in Data Science

by

Sahil Mohammad

0220993945

Department of Mathematics
UNIVERSITY OF LUXEMBOURG
Luxembourg
29th August, 2024

Abstract

This report provides a detailed explanation of the project completed as part of my internship within the Business Intelligence team at KPMG Luxembourg. This project focused on enhancing data accessibility and to provide a holistic view of the usage of different reports across KPMG. Being an audit firm, it was really crucial to take care of the data privacy standards and hence, data analysis and management was a crucial part of the project. To accomplish the same, SQL stored procedures were utilized to ensure efficient and accurate handling of information. As mentioned, significant aspect of the project was the implementation of data privacy measures. Sensitivity to regulatory requirements lead to decisions to mask certain kind of data and to randomize the other thereby balancing data utility with confidentiality. The report also emphasizes the role of visualization in data reporting. This project utilized PowerBI to create visualizations and reports after data preparation was done. Additionally, the project highlighted the automation of deployment pipelines using Azure DevOps. This platform facilitated the easy deployment of reports across different environments, which eliminated the need of any manual processes. As a member of the BI team, this internship provided firsthand experience in leveraging technical skills for database management, data privacy compliance, reporting, and automated deployment processes. The knowledge and experienced gained here contributes to a deeper understanding of effective data management practices within a professional environment, which lays emphasis on their significance in driving informed business decisions.

Contents

1 Introduction

KPMG is a brand name when it comes to Audit, Tax and Advisory. One of the renowned Big4 firms, it is located centrally in Luxembourg City's Kirchberg district. There is a lot of diverse workforce with over 1800 employees belonging to 70 different nationalities. This very fact makes it a great place to meet new people, and also gives a sense of belonging to every employee. The firm serves a diverse client base looking over important areas such as asset management, alternative investments, banking, insurance, corporates, and the public sector.

1.1 Mission and Values

KPMG Luxembourg operates under the umbrella of KPMG International Limited, which is a globally integrated network of independent firms. Their presence is known in 143 countries all over the world and KPMG International takes pride in a huge workforce comprising of over 273,000 partners and employees. Each member firm, just like KPMG Luxembourg, operates autonomously, providing specific and tailored services that can cater to the local clientele. But at the same time, they need to adhere to the global standards and maintain all aspects of excellence and integrity.

KPMG Luxembourg offers a comprehensive list of services designed to address the evolving needs of its clients. In addition to audit, tax, and advisory services, the firm provides specialized expertise in areas such as financial services, including asset management and banking, as well as insurance and public sector consulting. This multidisciplinary approach allows KPMG Luxembourg to offer integrated solutions that enhance operational efficiency, regulatory compliance, and strategic decision-making for its clients.

1.2 Organization Structure

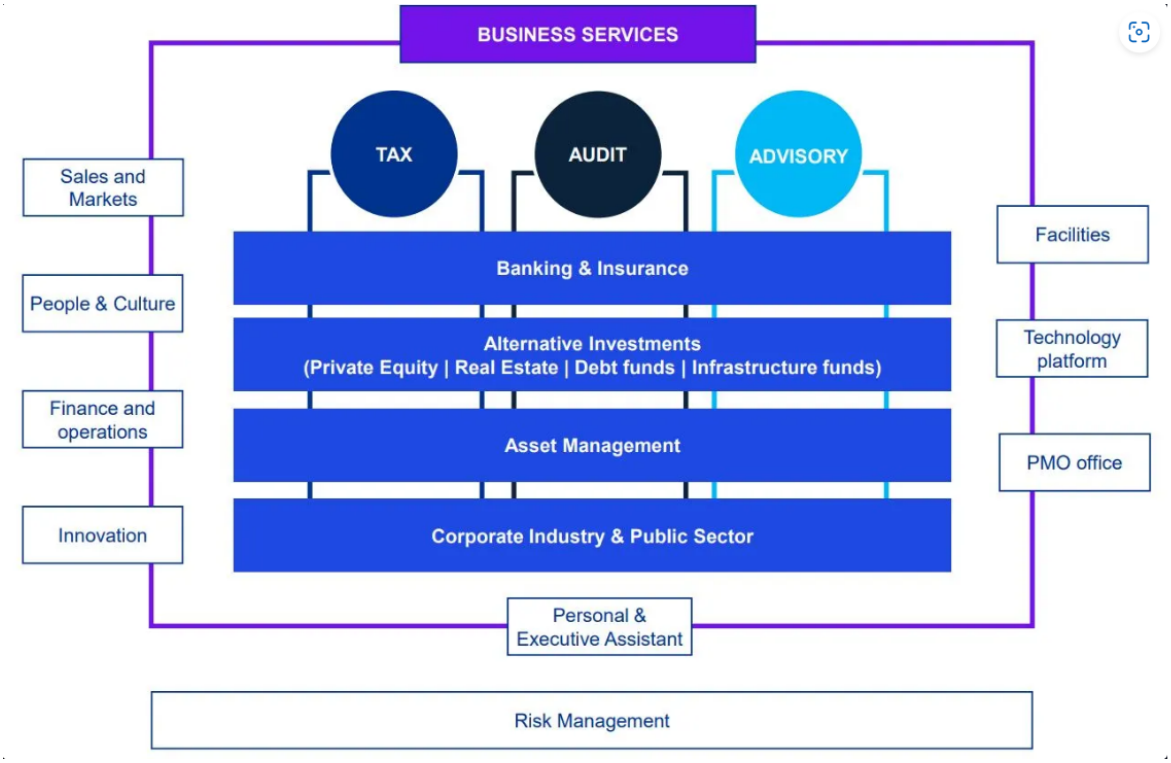

Figure 1: Organization Structure

Figure 1 describes the organizational structure of KPMG. It emphasizes on the core service areas and how they integrate along with each other in addition to providing comprehensive support functions to deliver services to the clients. In addition to the core areas, there are also specific sectors mentioned

within the core areas which emphasize on the importance and the various task that each core area undertakes within the organization. The structure can be broken down into the following areas as explained in Figure 1 :

- Tax : KPMG's Tax services provide detailed tax compliance, advisory, and planning solutions designed to help clients get through the complexities of national and international tax regulations, since KPMG is a global brand. By staying updated with the constantly evolving tax laws and policies, KPMG ensures that their clients remain compliant while enhancing their tax positions. This involves preparing and filing tax returns, advising on tax implications of various business decisions, and developing strategies to minimize tax liabilities. We will see in the upcoming pages, how I worked on a project which helps clients with F500 tax returns. The Tax team also offers specialized services such as transfer pricing, tax risk management, and dispute resolution to address specific client needs, ensuring a holistic approach to tax management.

- Audit : KPMG's Audit services are basic building blocks to ensure financial transparency and regulatory compliance for their clients. The Audit team conducts both statutory and internal audits, while examining financial statements and internal controls to provide assurance on their accuracy and reliability. Through strict testing and verification processes, KPMG helps clients maintain confidence in their financial reporting, while building trust among stakeholders. The audit process not only detects and prevents financial irregularities but also identifies opportunities for improving internal controls and operational efficiencies. Theses audits are not only for the clients but also within the teams. There have been audits on the BI team as well, which take notice of how the team handles senstitive data, and if KPMG policies are being respected.

- Advisory : KPMG's Advisory services offer a wide range of strategic support to help clients improve and grow their businesses. They provide business advice, help make operations more efficient, assist with financial restructuring, and develop policies. KPMG's advisory team works closely with clients to understand their specific challenges and goals, creating customized solutions that drive big changes. Whether it's improving supply chains, upgrading IT systems, or handling mergers and acquisitions, KPMG has the knowledge and insights needed to make smart decisions. Their comprehensive approach ensures that clients can keep up with changing market conditions, grow sustainably, and stay competitive.

- Business Services : The Business Services provide essential support functions that for the basis of the operational success of the clients. These services include financial management, IT solutions, and other critical operational functions that ensure smooth business operations. By offering specific solutions in areas such as finance, HR, and IT, KPMG helps clients improve efficiency, reduce costs, and focus on their main areas of work. All other teams apart from Tax, Audit and Advisory, fall under Business Services. The Business Intelligence team as expected, falls under business services too. Instead of dealing with clients directly, these teams mostly help the other teams by providing solutions for enhancing and optimizing their work.

1.3 Focus Areas

These core areas focus on quite a lot of sectors which are mentioned below :

Banking and Insurance: In addition to providing tailored services for financial institutions, this sector plays a very important role in ensuring compliance with both national and international regulations. By meticulously navigating regulatory landscapes, KPMG supports banks and insurance firms in adhering to stringent guidelines that govern their operations. By offering comprehensive risk assessment and mitigation strategies, KPMG enables financial institutions to enhance their risk management frameworks and operational resilience. This proactive approach not only safeguards the interests of clients but also strengthens their competitive edge in a dynamic and highly regulated financial environment.

Alternative Investments: These core areas perform valuation, due diligence, and advisory services for a wide range of investments such as private equity, real estate, debt funds etc. They evaluate potential investment opportunities and conducting financial due diligence. As far as market analysis is concerned, they perform the actual analysis, property valuation, and transaction support. Moreover,

they structure and manage debt funds and perform financial modeling and risk assessment for large-scale infrastructure projects.

Asset Management: This function deals with the investment strategy development, portfolio management, and performance analysis. Which means, that the areas crafting strategies to optimize returns based on market conditions and client objectives. They monitor and adjust investment portfolios to meet client goals. Additionally, evaluating the performance of various asset classes and investment funds is one of the many tasks for the core sectors.

Corporate Industry and Public Sector: Last but not least, another very important function of the sectors is to provide Business advisory, audit, and compliance services for corporate clients and public sector organizations. It comes with strategic planning, operational efficiency improvements, and financial restructuring. The audit services on the other hand conduct statutory audits and internal audits to ensure compliance and transparency. With the public sector outlook, KPMG looks forward to advising clients on policy development, public finance management, and program implementation.

1.4 Business Intelligence Team

The Business Intelligence (BI) team within the Business Services subsidiary at KPMG Luxembourg plays an essential role in enhancing data-driven decision-making across the firm's key sectors: Audit, Tax, and Advisory. The MyBI project is central to the BI team's mission, focusing on optimizing BI processes and workflows using Azure DevOps. This project aims to streamline data management tasks, from extraction and transformation to report generation and deployment, creating efficient and easy to understand solutions. The primary objective of the BI team is to develop a comprehensive data warehouse. This data warehouse serves as a virtual, integrated, and clean data environment that meets the firm's extensive reporting and analytical needs. The data warehouse ensures stakeholders have access to reliable and ready-to-use data, which helps them in making informed decisions. Built on an on-premise architecture, it is important for the BI team to adhere to the compliance regulations at KPMG and also to manage the license costs effectively.

In the Audit sector, the BI team enhances data accuracy and consistency by meticulously extracting and transforming data from various sources. This ensures auditors have reliable data for comprehensive compliance reporting and risk assessment, using advanced tools and analytics to identify and mitigate potential risks. The team's strict guidelines to maintain data quality helps uncover discrepancies, and ensure that financial reports reflect true financial performance. The BI team's efforts contribute to the robustness of audit practices, ensuring that clients adhere to regulatory requirements, maintain transparency, and build trust with investors and regulatory bodies. By integrating data from multiple systems and standardizing it, the BI team streamlines the audit process, reducing the time and effort required for manual data reconciliation. We will see further how the different sources are integrated together in the BI environment.

Within the Tax focus area, the Business Intelligence (BI) team leverages advanced data analytics to provide deep insights into tax compliance and planning. We produce detailed and comprehensive tax reports that help clients optimize their tax strategies and gain a clear understanding of their tax positions and obligations. These reports are tailored to highlight potential tax savings, identify risks, and offer strategic recommendations to improve overall tax efficiency. The BI team plays a crucial role in compliance monitoring by continuously analyzing tax-related data to ensure adherence to both local and international regulations. For the Advisory sector, the BI team delivers strategic, data-driven insights that enable advisory teams to craft well-informed plans for clients. The team's ability to track and analyze key performance indicators (KPIs) provides essential guidance for client advisory services. By leveraging data analytics and visualization techniques, the BI team helps uncover trends, identify opportunities, and pinpoint potential challenges within the client's business. The KPIs are important in any report that the BI teams creates. These KPIs give much needed information around any indicator which maybe used for a better performance in any areas that are lacking behind.

1.4.1 BI Architecture

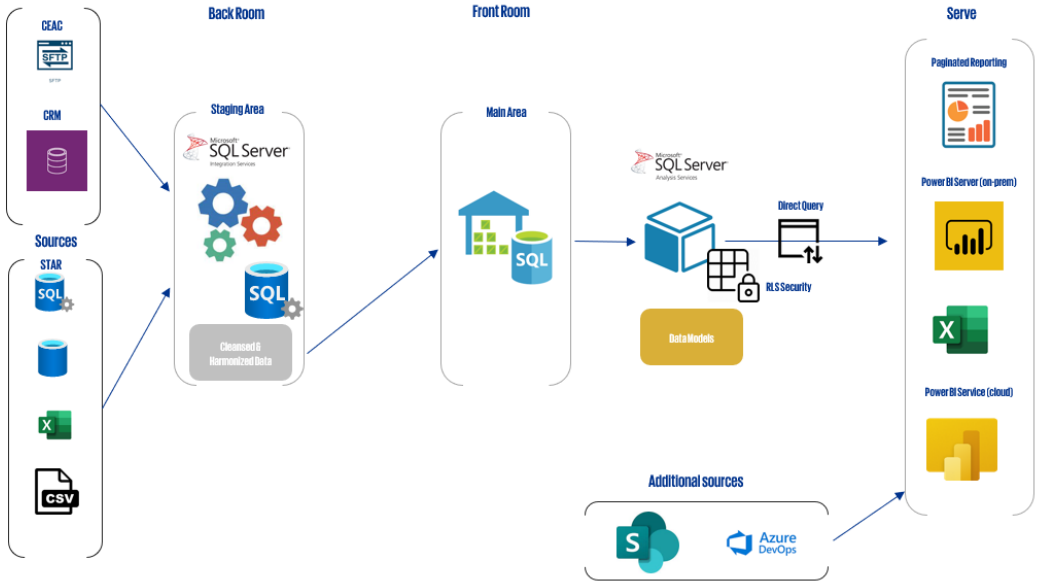

Figure 2: BI Team Architecture

The architecture of the BI team is designed in a way to ensure efficient data management and reporting across KPMG Luxembourg. The whole process begins with collection of data from various sources, including file transfers, CRM systems, SQL Databases, Excel and CSVs and more importantly, KPMG's internal centralized system : STAR.

1.4.2 STAR

At KPMG, the centralized repository, often referred to as STAR, is the heart of the company's data operations. It serves as a very comprehensive and a huge data center, centralizing critical information about staff, clients, jobs, engagements, and work in progress (WIP). This repository is very important and highly necessary for KPMG's operations, as it supports a wide range of functions from time entry and client billing to scheduling and reporting.

The data center contains detailed records of KPMG's staff, including their roles and assignments. This information is the baseline and important for managing client engagements and ensuring that the right resources are allocated to the right tasks. The repository also organizes clients into groups, reflecting the diverse and complex nature of KPMG's client relationships. For instance, a client group like 'University of Luxembourg' might consist of various client accounts such as 'University of Luxembourg - Belval', 'University of Luxembourg - Kirchberg', and 'University of Luxembourg - Limpetsberg'. All of these client accounts will have different charge account IDs but within the same client group. This hierarchical structure allows for precise management and billing of client engagements. Also, in the end you can calculate the total amount that was spent for a client group in general.

Within the STAR system, each client is associated with specific jobs and engagements. These records are essential for tracking the progress of work and ensuring that each task is accounted for. The WIP component of the repository provides real-time insights into the ongoing work, detailing which client and charge account each engagement belongs to. This information is crucial for accurate time entry in the timesheet system, ensuring that all billable and non-billable hours are correctly attributed.

One of the core functionalities of the repository is its role in client billing. Clients are billed based on different charge type IDs, which categorize the nature of the work performed. These charge types in-

clude categories such as Planning, Profit/Loss, Non-Chargeable IT, Client Meetings, and many others. By categorizing work in this way, the system ensures that clients are billed accurately and transparently for the services they receive. STAR accurately tracks which client's account was charged, what type of charge (identified by the charge type ID), who performed the work (based on staff ID), and the nature of the job. This level of detail ensures that every transaction is thoroughly documented, providing a clear and comprehensive record for both internal audits and client billing. For example, if a staff member works on a planning task for 'University of Luxembourg - Kirchberg', the repository will log the staff ID, the specific client account, the charge type ID for planning, and the details of the job performed.

Beyond billing, the STAR repository is also a vital tool for scheduling and planning. It provides capabilities for planning resources and scheduling tasks, helping KPMG manage its workforce efficiently. By using the repository to track WIP and plan future engagements, the company can optimize its operations and ensure that all projects are completed on time and within budget. The data stored in the STAR repository is not just for operational purposes; it also supports comprehensive reporting and analytics. By analyzing the data, KPMG can gain valuable insights into its operations, identify trends, and make informed decisions. This analytical capability is crucial for maintaining high standards of service and continuously improving the company's performance while providing client satisfaction.

Now coming back to the architecture as shown in Figure 2, the raw data is first processed in the staging area using Microsoft SQL Server Integration Services (SSIS), where it undergoes cleansing and harmonization to ensure accuracy and consistency. Accuracy is essential because it guarantees that the data reflects the true values and scenarios that are being analyzed, which is critical for making informed decisions. Consistency ensures that the data from different sources conforms to the same standards and formats, enabling meaningful integration and reliable analysis across various datasets. If we do not ensure accuracy and consistency, any analysis or reports generated could be misleading, or probably would lead to incorrect conclusions and potentially wrong business decisions.

Once this is done, the next step is to move the cleansed data to the main data warehouse, which is also managed by the Microsoft SQL Server. In this central repository, the data is stored in a structured format, facilitating easy access for further analysis. The BI team uses Microsoft SQL Server Analysis Services (SSAS) to create data models that organize the data into structures that make querying data very efficient. In order to ensure security of the data and compliance, an implementation of Role-Based Security (RLS) is implemented, which controls access based on user roles.

In the architecture, the final step involves serving the processed data through various reporting tools. PowerBI Server (on-premise) and PowerBI Service (cloud) provide very robust platforms for internal and easy to scale, cloud-based reporting respectively. Additionally, Excel is used for traditional data analysis, and paginated reporting is available for generating other reports in a detailed manner. The architecture can also integrate additional sources like SharePoint and Azure DevOps, which leads to enhanced collaboration and overall management of the project. As a whole, this BI Architecture supports KPMG Luxembourg's core areas of Audit, Tax, and Advisory by delivering clean, integrated, and reliable data which helps the various teams in the firm to make decisions based on reports which highlight nothing but factual information from the data. Thus, providing data-driven decisions and high quality services.

2 Project Discussion

The first phase in the project, like in a typical data analysis life cycle was comprehensive data preparation. This was a very important step which was aimed at ensuring compliance with the strict data privacy standards that are essential for KPMG, given the sensitive nature of the data that is handled by an audit firm. The process began with classifying the entire database into different privacy levels. Sensitive data such as names, emails and client names were identified for special handling to protect all confidentiality. For instance, certain columns containing the sensitive information were masked / anonymized to preserve the structure while obscuring the actual data. This approach was highly essential for the maintenance of data privacy without compromising the integrity that was needed for any further report development or analysis. In cases where masking the data would cause the data

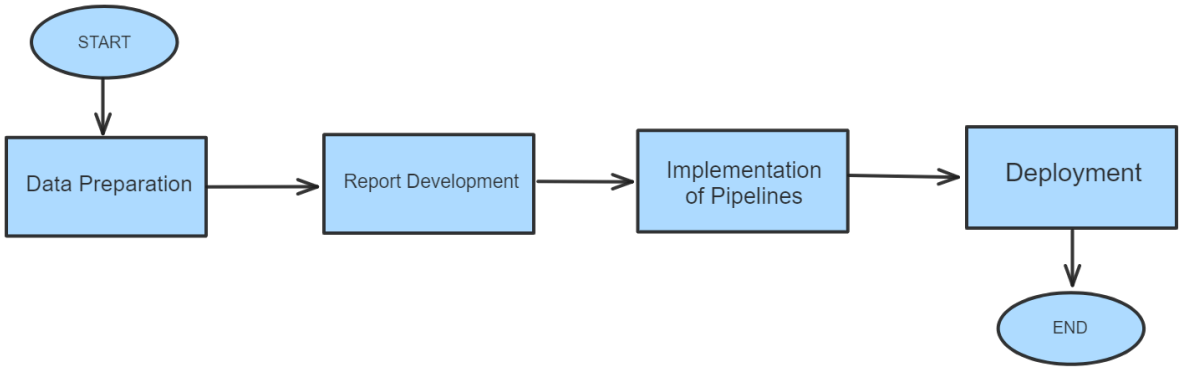

Figure 3: Project Overview

to be not much useful for reporting, such as with names and emails, randomization techniques were employed. This allowed the data to remain useful for analytical purpose while ensuring that the data or the report does not reveal any real information.

Now, in order to streamline and enhance the data preparation process, SQL stored procedures were utilized extensively. These procedures provided automation of various tasks, such as data masking and randomization, which in turn ensures consistency and reduces the chances and probabilities for any manual errors. By making use of the SQL procedures, the project achieved much better data handling and enhanced the performance in the data preparation phase of the life cycle. Additionally, these procedures facilitated faster data processing, allowing for more efficient data analysis and reporting. The automation also freed up valuable time for the team, enabling them to focus on more complex and strategic tasks rather than routine data preparation. Moreover, the use of stored procedures improved scalability, making it easier to manage larger datasets as the project grows, and ensured that best practices and data governance standards were consistently applied throughout the process.

Once the data was prepared, the next phase focused a lot more on the report development. The name of the report created is "myBI Usage" which provides a comprehensive overview of the utilization of the various BI reports within KPMG Luxembourg. The report includes a lot of key metrics, including the total number of subscriptions, reports and executions, offering a holistic view of the report usage across the organization. There are various sections in the report which allow the users to access details for various information. Different visualization were utilized which makes the report easy to understand while conveying a lot of information at the same time. Another dashboard was developed for the F500 - Yearly tax declaration. The object is to generate a financial statement that clients will append to myGuichet as part of their yearly tax declaration process. There were two reports developed for this project. The first report was developed for the partners and it would track the tax declaration process through its various stages by connecting to SharePoint online lists. It would also represent the percentage of time that was spent on each stage for all the CTRs (Company Tax Returns). The second report was for Timeframe calculations between different stages of the process. This report would calculate the time spent between different stages to identfiy and forecast any potential bottlenecks in the process. This was achieved by using the last modified information from the SharePoint lists. Further explanations on these reports will be provided in later sections.

Post report development, the final phase of the project involved the implementation of Deployment Pipelines using Azure Devops. This phase was critical to ensure that the developed reports can be easily and efficiently deployed across various environments without any manual effors. By automating this deployment process, Azure DevOps pipelines ensure consistency and reliability, which minimizes the risk of any potential errors and highly reduces the time required for deployments. The deployment pipelines were configured to handle the transition and flow of reports from development to the production environment, which helps the project to conform and adhere to the best practices of Continuous Integration and Continuous Deployment (CI/CD), which is highly sought after in Data Engineering as well.

As a member of the Business Intelligence team at KPMG Luxembourg, the project not only enhanced technical skills but also explained the importance of effective data management practices in driving informed business decisions. This report contains sections on each of the phases of the project which are mentioned above. In these sections we aim to tackle and explain in much more detail each phase, while highlighting the importance as well as the steps that were undertaken to implement all the same in the project.

3 Implementation

The section above gave a brief overview of the whole project cycle and the steps that were followed in the completion of the internship project. This section deals with detailed explanation on the various steps involved and tries to explain the flow of the project in a detailed manner. The steps involved are as follows :

3.1 Data Preparation

The data warehouse used by the BI team resides in SQL Server which acts as a robust and highly available database that supports a lot of applications and offers features like point-in-time restorations. This data warehouse, named 'DWH' in SQL Server Management Studio (SSMS), serves as the central repository for consolidating and organizing data from various sources, making it a crucial tool for analytical and reporting purposes. The data warehouse is carefully designed to provide a structured environment for both data storage and retrieval, ensuring that data is easily accessible and well-organized. At the center of this data warehouse are its 235 tables, which are essential for storing transactional data, dimensional data, and metadata. These tables include various dimension and factual tables, each holding a wide range of information necessary for comprehensive data analysis.

The preparation of data for projects primarily involved the classification of these various columns [1]. Given the large number of tables and added to that, the high number of columns within them, this classification process can be quite cumbersome and time-consuming. However, understanding why this classification is necessary is crucial before diving into the process itself [3]. Classification is essential for several reasons. First, it ensures that data is organized in a logical and consistent manner, which is critical for efficient data retrieval and analysis. By categorizing columns based on their type and purpose, the process of querying and analyzing data can be streamlined, making it easier to generate accurate and insightful reports.

In addition, classification aids in maintaining data quality and integrity. By clearly defining the types and constraints of each column, the BI team can implement validation rules and checks that prevent any data entries with potential errors and ensure consistency across the data warehouse. This is particularly important in a large-scale environment like the data warehouse, where the volume of data can make manual quality control not much useful and impractical. Overall, while the process of classifying columns in a vast data warehouse like DWH may be challenging, it is a vital step that lays the foundation for effective data management and analysis. Before getting to the process of classification, and how it was actually performed on the data warehouse, it is important to understand why classification is actually required.

What is Classification ?

Classification is the process of categorizing data based on its sensitivity and importance. Here we assign labels or tags to the data to indicate its level of confidentiality [3]. This complete process starts with organizing and listing out the data, which may include tables and various columns that need to be identified. Once the data is organized, the importance of each attribute is to be decided. This is done by considering several factors like how risky is the data, what would the business impact be if that data is accessible, and who should have access to it.

This step involves defining clear classification criteria that align with organizational policies and regulatory requirements. These criteria help in systematically assessing the sensitivity of data elements. After

setting the criteria, each data item is evaluated and classified accordingly. High-sensitivity data might include personal identifiable information (PII), financial records, or proprietary business information, while low-sensitivity data might include publicly available information. It's crucial to regularly review and update the classification as the data and its context can change over time. Implementing data classification policies helps in mitigating risks by ensuring that sensitive data is adequately protected and only accessible to authorized personnel. Training and awareness programs are also essential to ensure that employees understand the importance of data classification and adhere to the guidelines.

Why is Classification needed ?

The primary reason why classification is needed is to make sure that KPMG data is protected according to its level of sensitivity. It is clear the not all data is equally sensitive and some information may pose a higher risk if it is exposed or compromised. Based on these differences, we used three different sentivity labels, which are as follows:

- General : This information can be shared with external partners, as required.
- Confidential : Sensitive business data that could cause damage to the business if shared with unauthorized people.
- Confidential - GDPR : Sensitive data containing personal information associated with an individual, that could be misused.

The Classification procedure as mentioned above, begins with identifying the columns that contain sensitive information. For this, we need to go through all the columns in each table, in each schema and decide if they should be marked sensitive or not. This information is then updated in the SQL Server Management Studio. Once the columns are identified and labelled, there are two processes that need to be carried out, which are as follows:

- Masking : This process involved replacing sensitive data with anonymized values while preserving the data format [2]. It ensures that no one can read the data once it has been masked. In our case, we mask sensitive information by replacing the characters with 'x'. For example, a person's email such as john.doe@kpmg.lu will become joxxxxxxoe@kpmg.lu.
- Randomization : Here, we replace sensitive data with entirely new and random values. Unlike masking which hides the data, randomization replaces actual data with fake/simulated data. This is very helpful in reporting as we will have some values to display instead of words that have been masked with x's. It is clear that when randomization is done, a mapping table is required to make a link between the random data and the real data that is being randomized or else the information would be lost.

Moreover, in the case of new table being added to the database, the sensitive columns in the new table should also be added to the existing Data Classification in SQL Server Management Studio and masking and randomization queries should be run again so that the new sensitive data is also protected. Once all the data has been classified with the respective sensitivity labels, the stored procedure for masking of the data can be run on it. These processes are explained in much more detailed in the next subsection.

3.1.1 Masking Procedure

The process is carried out with the help of a stored procedure designed to protect sensitive information by masking data in the tables. It uses data cursors to sequentially process each row, extract table, column and schema names along with the column details from the system's extended properties table. The system's extended properties contains the information that has been classified in the SSMS. All the columns and the senstivity label that has been provided to that particular column is present in the system tables. SQL cursor is nothing but a database object that is particularly useful for operations that require row-level-processing, such as performing calculations and updates on each individual row.

The procedure iterates through each row, using the cursor to fetch and process data from the staging tables. For numeric fields, it assigns a fixed number, for date fields, it sets a standard date, and for string fields, it either reduces the string to a single character or masks part of it with 'x's. It is to

be noted that the procedure skips masking certain fields like names of people, their emails, windows logins and the client names so that randomization is possible on them later. Since there are cases where the PowerBi report has to show some names and data, it only makes sense to replace the information with fake names to allow reporting to be done successfully. This procedure, as a whole ensures that sensitive information remains confidential and secure during the staging process. Once the masking / anonymization is done, randomization of all the columns that have been skipped from masking has to be done for reporting purposes as explained above. This process is explained in the subsection below.

3.1.2 Randomization Procedure

The process of randomization is important because masking hides all the data and we would not be left with values that can be used for reporting purposes. However, randomization is not as direct as the masking procedure as here we need to first create a mapping table that would link together the new random values with an actual staff member or a client. We create two mapping tables. The first one contains information related to staff and the second one is for clients. To create the tables, we make use of the following IDs :

- KPMGGPID : These are unique identifiers for staff and hence, can be used to link actual staff with the randomized names.
- GISID : The clients have these unique identifiers which are used for mapping purpose with the newly generated fake business names.

However, we need to replace names of almost 12,000 staff members with a fake name. Therefore, creating a single name manually for each and every staff was not possible. This process was also automated using SQL [4]. Instead of thinking of every individual name, a list of 200 fake names was utilized which can be easily generated from a fake name generator present online. These names were in the format of 'FirstName' and 'LastName' which was inserted into a table. After inserting these names in a table called 'RandomNameBank', a cross join was performed between the FirstName and LastName columns to give a huge number of combinations of fake names.

The CROSS JOIN is used to show every possible combination between two or more sets of data. We can do a cross join with more than two sets of data, allowing for comprehensive combinations. Cross Joins are typically done without join criteria, unlike the inner, left, and right joins which require a condition to match rows. A Cross Join is also known as the Cartesian Join as it is nothing but the Cartesian product of the two sets of data. This means every row from the first set is combined with every row from the second set. It can be done between different tables or on two different columns of the same table, which is what we used in the project. For example, if you have a table with n rows, a cross join would produce n*n results. Hence, since the RandomNameBank contains 200 rows, a cross join gave us 40.000 results. This extensive combination of data pairs can be particularly useful for generating all possible pairs or scenarios that need to be tested or analyzed.

In our project, the cross join was utilized to create a comprehensive list of staff names by combining first names and last names from two separate columns within the RandomNameBank. By doing so, we ensured that we had a varied and exhaustive set of potential names for use in simulations or anonymization processes. The benefit of using a cross join in this context is that it provides a straightforward and efficient way to generate all possible name combinations, which can then be filtered or used as needed for further data processing tasks.

Moreover, while cross joins can be powerful, they should be used judiciously, as the resulting dataset size grows exponentially with the size of the input tables. This can lead to performance issues if not managed properly. Therefore, understanding the implications and ensuring that the system can handle the resultant data load is crucial when implementing cross joins in any project. Proper indexing and optimization techniques should be applied to handle large datasets effectively, making the process smoother and more efficient.

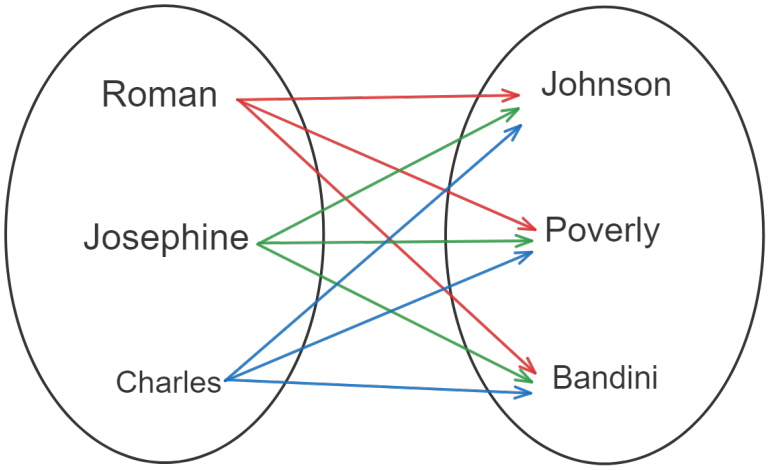

Figure 4: Working of a Cross Join

Following the generation of fake names, the next critical phase involves creating a mapping table. This table serves as a centralized repository linking each staff member's unique KPMGGPID which was mentioned above, to their corresponding fake name. Alongside the fake names, the mapping table includes additional attributes such as an email ID generated specifically for the fake name and a Windows login ID associated with the staff member. This comprehensive approach ensures that each staff member is effectively connected to their anonymized identity across various systems and applications.

With the mapping table in place, a stored procedure is employed to systematically identify and randomize names across the database. The procedure scans through all tables containing columns with names that match specific criteria, typically involving the term 'name' in their column headers [5]. Upon identifying such columns, the procedure retrieves the actual names stored within and initiates a lookup process. During the lookup, the procedure cross-references each actual name with the staff table to retrieve the corresponding KPMGGPID. This step is very important as it enables the procedure to pinpoint the exact staff member associated with each instance of an actual name within the database. By using the mapping table created before, the procedure then retrieves the corresponding fake name linked to the identified GPID.

Once the fake name is identified, an update statement is executed to replace the actual name with the fake name across the database. Given the huge size of the database, this process might be a bit time consuming as it replace all the instances of a person's name with the fake name accross all the tables. This automated process ensures uniformity and consistency in anonymizing personally identifiable and any information related to staff members throughout all data sets and tables. By systematically replacing actual names with fake names, this project and process mitigates the risk of exposing sensitive information while adhering to strict data protection regulations and compliance standards. Moreover, since there still exists a name instead of masked name, the fake name can be now utilised in the generation of reports and at least the reports will now have information to display instead of masked data. The following figure explains the name generation, mapping table creation, and the lookup processes in an efficient way:

This way, the names, windowslogins and emails for the staff members have been randomized [6]. Additionally as mentioned before, the client names were to be randomized as well. In this case, the client names were replaced by fake business names generated in a way similar to the fake person names. The only challenge with generating a mapping table for client names was that the client information was present in two different tables. Also, as mentioned before, the client GISIDs were used to map an actual client with a fake generated client. These IDs had to be combined from two different tables using the union operation while taking care of using only unique IDs and skipping any client that was repeated in the two data sources.

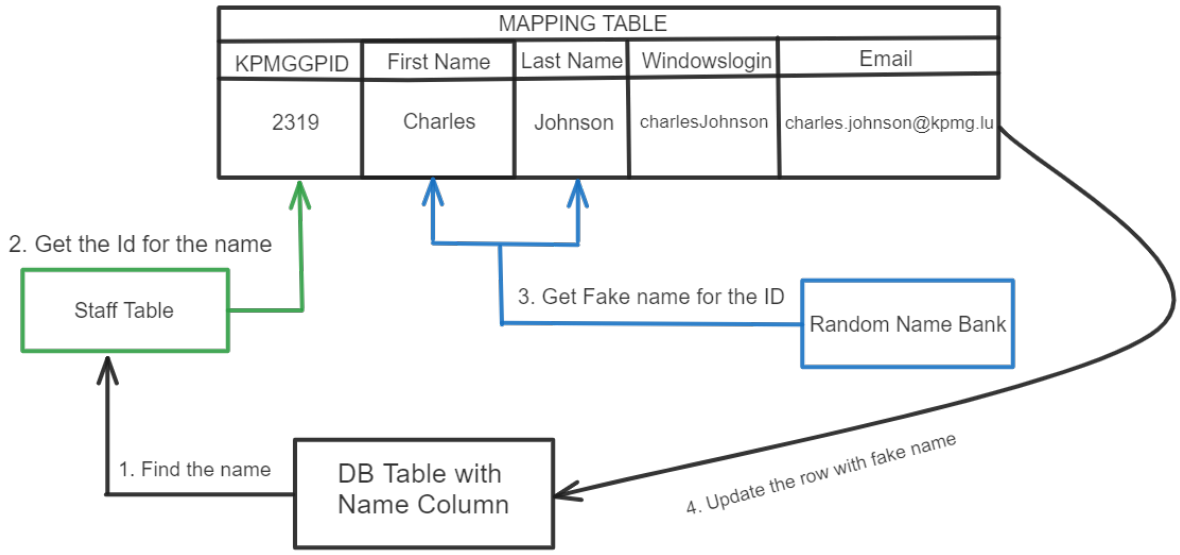

Figure 5: Lookup and Randomization

Once the mapping table was ready for the Client information, the stored procedure could be created. Just like person names, Client name was also searched for in all the tables and the corresponding ID was selected from the Client Source table [7]. Once the ID was retrieved, the same ID was searched for in the Mapping table and the corresponding fake client name was obtained. Then a simple update statement changed the actual client name with the fake name [8].

3.2 Reporting

Once the data preparation was done, the main task of the BI team is to provide dashboards and reports which in turn provide valuable insights to the different teams that requested the report. To prepare the data, analyse it, and to provide valuable findings from it is one of the most important aspects of Data Analysis life cycle. In section 3.1 we discussed how the data was prepared and senstive data was masked and randomized for efficient reporting purposes. In this section, we will discuss more about the reports that were created post data preparation and the information that they provided. The development process involved creating comprehensive and interactive dashboards using PowerBI. Since KPMG has a partnership with Microsoft, the tools used here are all provided by the latter and hence, PowerBI was utilised for reporting purposes. Moreover, PowerBI provides a lot of features and capabilities along with a user friendly interface which helps translate complex data sets into understandable insights.

3.2.1 myBI Dashboard

The primary dashboard, as shown in the myBI Usage report, provided a holistic view of report usage across KPMG. This report used about 18 different tables, some of which were taken from the SQL database and a couple of them were created in PowerBI. To connect these tables together, special care had to be taken to join the correct fields together. Moreover, the cardinality of the join was an important factor in joining the tables together. The most commonly used cardinalities are as follows [14] :

- 1:1 Cardinality : In a one-to-one (1:1) relationship, each entity from the first set is associated with exactly one entity from the second set, and vice versa. This is perfect choice in case of joining unique ids from one table to another table with unique IDs. Thus, for each ID, there is a unique ID which maintains the 1:1 order.
- 1:* Cardinality : In a one-to-many (1:*) relationship, a single entity from the first set can be associated with multiple entities from the second set but each entity from the second set is associated with only one entity from the first set. For example in case of the myBI reports, there are multiple reports in a folder on the PowerBI server. Moreover, in the opposite direction, a single folder holds multiple reports which maintains 1-many relationship.

Once the correct data was pulled into the report, the dashboard was designed with several key features and sections which are mentioned below :

- Activity Summary : This section provided a snapshot of overall report usage, highlighting the frequency and distribution of report executions. This section helped identify peak usage times and potential bottlenecks in report access.
- Workspaces : Power BI workspaces are collaborative environments within the Power BI service where users can work together on dashboards, reports, datasets, and dataflows. This section provides information about different workspaces to show how various departments or teams were utilizing the BI tools. This allowed for a better understanding of departmental engagement and collaboration.
- Report and User Lookup : These features enabled users to search for specific reports and users, providing details about which report was executed recently and the time details about each execution on each report. This was particularly useful for troubleshooting and for identifying power users or training needs.
- Report Inventory and Catalog History : A comprehensive list of all available reports and their folder details is maintained in this view. It also shows the last modification details for each report.

3.2.2 Time Window

The above list briefly mentions the various types of information present in the myBI report. However the whole report's execution tracking feature controls all the visuals present in it. This feature is designed to provide users with a easy to use and a very interactive way to monitor the frequency of report usage over various time periods. Users have the flexibility to select a time frame such as the past 7, 30, 60, or 90 days—using a slicer tool on the dashboard [15]. This simple selection mechanism allows the dashboard to dynamically update and display the number of times reports have been executed within the chosen period, offering valuable insights into report usage trends. Moreover, as mentioned before, the slicer selection controls all the other visuals as well to show the results from a given time window.

To implement this time window functionality, a specialized date table is created in PowerBI. This table is provided with markers that indicate whether each date falls within the last 7, 30, 60, or 90 days. These markers are like filters, that enable the visuals to accurately count only the relevant executions for the selected time frame. For example, if a user selects the 7-day period, the dashboard will count and display only those executions that occurred within the last 7 days. Similarly, selections for 30, 60, and 90 days will adjust the displayed data accordingly.

There are additional functionalities added to the date table which help in a lot of other visuals. These functionalities were implemented by addition of various additional columns, such as day of the week, month, year, and flags for weekends. This detailed implementation allows for more detailed data analysis, since this report is for monitoring the executions of various reports, it was important for implementation of various time window details. Separate measures are defined for each time period, ensuring that the correct calculations are performed based on the user's selection. These measures use the markers from the date table to filter and count the report executions accurately.

This information is used on the Activity Summary page which tracks the number of executions of the report according to the days and the time. Using this feature we can know how many reports are being executed and when is the busiest time for the executions. On the prod server, this report shows that the busiest days are Mondays and Fridays basically because they are the first and the last days of the week and hence there is more activity. This single page contains a lot of information like the percentage of success and failed executions, along with the above explained details by the week. Additionally, there is another visual called the 'Selected Execution Count' that combines all the individual calculations and displays the appropriate count which explains the count of executions along with the number of distinct users that ran the executions. This line chart provides quite a lot of information while using the Date table that was created earlier.

3.2.3 Workspaces

Power BI workspaces [16] are a fundamental part of the Power BI service. They are collaborative environments where teams can work together on Power BI content, such as dashboards, reports, datasets, and dataflows. There are mainly two different types of workspaces:

- Personal Workspace
- Shared Workspace

Considering the first one, as the name suggests it is a personal workspace available to every Power BI user. It is intended for individual use and is not typically shared with others. Users can create and store personal dashboards, reports, and datasets here. Secondly, a shared workspace is the one that are collaborative workspaces where multiple users can contribute. They are ideal for teams working on shared projects. In shared workspaces, users can collaborate on the development of reports and dashboards, ensuring consistency and enabling integration and combination of individual work. Workspaces have administrators that can control access to the workspace by assigning different roles to users. Common roles include Admin, Member, Contributor, and Viewer, each with varying levels of permissions. The sharing of data and organization of dashboards etc can be done on the workspace itself. Once this is done, the reports and dashboards can be published to the workspace. After they are published, they can be shared with other users, either within the workspace or even in a broader way by sharing it accross the organization.

The workspace owner plays an important role in managing and maintaining the workspace. The owner is responsible for creating the workspace, defining its purpose, and taking care of its overall administration. They manage access by assigning roles to other users, such as Admins, Members, Contributors, and Viewers, each with different permissions. The workspace owner also takes charge of content management, ensuring that dashboards, reports, datasets, and dataflows are organized and up-to-date. Any issues with the workspace be it access related or performance related, needs to be addressed by the owner or by someone who as relevant access. A page in the report has been implemented which shows all the details regarding the all the workspaces present in KPMG. There are different states that a workspace can reside in. These states are as follows :

- Orphan Workspaces : Orphan workspaces occur when a workspace no longer has an active owner. This can happen if the owner leaves the organization or their account is deactivated. Without an owner, the workspace lacks proper oversight, which can lead to issues in managing content and permissions. Organizations typically need to reassign ownership to another user to ensure that the workspace continues to function effectively and that its contents remain accessible and manageable.
- Suspended Workspaces : Suspended workspaces are those that have been temporarily deactivated, which mostly is because they have been inactive for quite a long time. In this state, users cannot access the workspace or its contents until the suspension is lifted. The suspension could be a result of not adhering to data governance policies, misuse, or other administrative reasons. Once the issues are resolved or the necessary actions are taken, the workspace can be reactivated and normal operations can resume.

The "Workspace Status" page shows all the details about workspaces and the number of workspaces in these particular states. It categorizes workspaces based on their orphaned status which is : active, inactive, suspended, and orphaned and presents a summary count for each category. Additionally, it lists workspaces by their respective owners, showing the number of workspaces managed by each individual. This page of the report highlights key metrics, including the total number of workspaces and unknown workspaces. A lot of detailed information for each workspace is displayed, including workspace names, statuses, owners, departure dates, and functions which is nothing but to which sector of KPMG does the workspace belong to. These sectors were discussed in a lot of detail in section 1.2 of this internship report. These enables efficient monitoring and management of workspaces, ensuring up-to-date information on workspace usage and ownership. The breakdown by orphaned status, WO (work order) status, and "Left On" dates further aids in identifying and managing workspace utilization and transition of owners. This report was very useful at an occassion when the BI Team received an email from KPMG global to delete all the workspaces that have been inactive for a long time. This

particular page of the dashboard was used and hence this particular feature of the report was very helpful.

3.2.4 Data Model Refreshes

The Power BI Data Model Refreshes View is a critical tool within the "myBI Dashboard" report, which offers a real-time visibility into the status of data model refreshes. This view provides users with detailed insights into the refresh process for various Power BI reports. It includes information such as the directory where each report is stored and the exact title of the report, which helps the team to quickly locate and manage their reports. The view categorizes reports by type, such as interactive or paginated, and provides detailed descriptions of refresh schedules, the last run times which is nothing but the last time a refresh occurred on a report. Users can also see the status of the most recent refresh attempt, whether it succeeded or failed, along with any error messages that might have been generated. The timestamp of the last refresh operation is also displayed to indicate how recent the data is, and the event type classification helps anyone in the team working on an issue to understand the context in which the refresh was triggered, whether it was scheduled or manually initiated.

The Power BI Reporting Services - Subscription Status provides detailed insights into the status of timed subscriptions for Power BI reports. It includes information on the report name, subscription name, event type, last run date and time, last status, and a description of the last run. This status report is crucial for understanding the execution history and current state of subscriptions, indicating if they are successful, pending, or disabled. Every day at 9 AM, a subscription status email is sent, summarizing the latest refresh statuses and subscription outcomes. This email ensures that stakeholders are aware of the most recent updates and any issues that need attention. For instance, if the "DQJobs" report refreshes successfully at 3 AM on July 19, 2024, this success would be reflected in the daily email, providing a comprehensive update on the data availability and report health. Together, the Power BI Data Model Refreshes View and the Daily 9 AM Subscription Status Email ensure a robust monitoring framework. The real-time insights provided by the Data Model Refreshes View allow users to address issues as they arise, while the daily email offers a broader perspective on refresh activities and issues over time. This combination enables the team to maintain up-to-date and accurate reports and manage their business intelligence operations more efficiently.

The daily subscription status email that is sent every morning at 9 AM, summarizes the refresh outcomes and provides the team with an everyday update on the health of the PowerBI Server. In case a report refresh is failed, it needs to be investigated as to what caused the failure and any issues need to be resolved so that the refresh can be done again and the report is functional as as result. Any of these reports might be used by any team within KPMG and a failed report means the other team cannot work on their tasks. Hence, this email serves as a regular checkpoint, highlighting any critical issues that need attention. By combining these two tools, the organization ensures both immediate, detailed monitoring and periodic, summarized updates. This approach in a dual manner enhances transparency, reliability, and the overall efficiency of the data reporting process. The team can rely on the email for daily updates and use the PowerBI view for in-depth analysis when necessary. This system supports proactive data management and helps maintain the integrity of the reports provided to any of the other teams and users.

While this report focuses on these pivotal aspects of the myBI Dashboard, it is important to acknowledge that the dashboard consists of a much a broader range of information and features designed to enhance data management and reporting efficiency. The information and features discussed in this internship report represent a summary of the most impactful elements, but the full dashboard offers additional functionalities that can provide even deeper insights and support more detailed analyses. All in all, the "myBI Dashboard" is an invaluable resource for optimizing our data reporting processes and ensuring the effective management of our reporting assets. The features outlined in this report highlight its capacity to deliver critical information and support informed decision-making, reinforcing its role as a cornerstone of our data management strategy.

3.2.5 Project F500 - Yearly Tax Declaration

Project F500 aims to facilitate the generation of a comprehensive financial statement that clients will use as part of their annual tax declaration process. This financial statement is a critical component of the yearly tax filings and is designed to be appended to the myGuichet platform, an essential tool for managing and processing tax-related submissions in Luxembourg. The primary objective of this project is to streamline and automate the tax declaration process, ensuring that the financial statements are accurate, timely, and compliant with relevant tax regulations. By achieving this, the project not only simplifies the tax declaration for clients but also enhances the overall efficiency and reliability of the tax reporting process.

In the context of the project, integrating with the MyGuichet platform is really important. MyGuichet is Luxembourg's official digital platform that provides a centralized, user-friendly interface for citizens, residents, and businesses to access and manage a wide range of government services online. Launched to enhance efficiency and accessibility, it allows users to submit documents, request information, and perform various administrative tasks electronically. The task for BI team is to create the reports as required by the Tax Team. However, connecting to MyGuichet and using the reports for filling in details on the portal is not the scope for the BI team. By ensuring that the financial statement produced and the insights provided by the F500 reports are valid and appropriate, the project aims to provide the tax team and then the clients that the tax team is dealing with, a streamlined and efficient means of fulfilling their tax obligations. This integration is designed to enhance the accuracy of submissions and reduce the manual effort involved in preparing tax documents.

The reports are designed to be a comprehensive tool for monitoring and managing the tax declaration process. It includes visualizations and metrics that track key performance indicators (KPIs), such as the number of declarations processed, the status of each declaration, the time taken on each stage and any potential issues or delays. By utilizing PowerBI features, the report can present complex data in an easily understandable format, allowing users to make data driven decisions and take timely actions. One of the core objectives of Project F500 is to enhance the efficiency of the tax declaration process. This is done with the help of two reports which meticulously monitor and manage the timing and status of various Client Tax Returns (CTRs) throughout their lifecycle. The project is designed to ensure that each CTR is tracked efficiently, from its initiation to its final submission, thereby enhancing the overall management of tax returns within the organization.

What is a CTR, and how it relates to the F500 Report ?

A Client Tax Return (CTR) is a crucial document that represents a tax return specific to an entity within a client's organization. Each CTR corresponds to a unique entity and encapsulates the financial information and tax details corresponding to that entity. In this context, a CTR is not just a generic tax return but a detailed report that is modified to the specific needs and circumstances of each client entity. Project F500 is centered around the detailed monitoring of the timing and status of these CTRs all of these CTRs for different clients and the entities for these clients. This involves several key activities:

- Tracking Progress: The project tracks the progress of each CTR through various stages of the tax declaration process. This includes monitoring when a CTR is created, when it is reviewed, and when it is finally submitted. By maintaining a close watch on these timelines, Project F500 ensures that no delays occur and that each CTR is processed within the required timeframes.
- Status Updates: The status of each CTR is continuously updated and recorded. This involves capturing various status indicators such as 'In Progress,' 'Under Review,' 'Pending Submission,' and 'Submitted.' Regular updates help in maintaining an accurate and up-to-date view of the CTR's lifecycle, facilitating better management and timely intervention if needed.
- Entity-Level Reporting : Each CTR is associated with a specific entity within a client's organization. Project F500 ensures that the returns are not only tracked at the client level but also at the entity level. This means that the project provides detailed insights into the status and timing of CTRs for each individual entity, allowing for a more granular and precise management of tax returns.

- Integration with Systems: To effectively track and manage CTRs, Project F500 integrates with various systems and platforms. This includes leveraging data from SharePoint Online Lists to monitor changes and updates, as well as using reporting tools to provide insights into the timing and status of each CTR. Integration with these systems ensures that the project has access to accurate and comprehensive data, which is crucial for effective monitoring.
- Identifying Bottlenecks: The main aspect of Project F500 is its ability to identify and address potential bottlenecks in the CTR process. By analyzing the timing and status data, the project can pinpoint areas where delays may occur or where additional support may be needed. This proactive approach helps in resolving issues before they impact the overall tax return process.
- Improving Efficiency : By focusing on the timing and status of CTRs, Project F500 aims to enhance the efficiency of the tax return process. This includes streamlining workflows, reducing delays, and ensuring that all CTRs are processed in a timely manner. Improved efficiency translates to better management of tax returns and a smoother experience for clients.

Accuracy is another fundamental goal of Project F500. Ensuring that the financial statement reflects the true financial position of the CTR and the correct stage for a client entity is critical to identify the exact bottleneck. In addition to tracking the progress of the tax declaration process, Project F500 also includes a time calculation component. This aspect of the project involves calculating the time spent between different steps in the tax declaration process. The goal is to identify and forecast potential bottlenecks or delays that could impact the overall efficiency of the process. To achieve this, the project leverages information from the version history available in SharePoint Online lists. This version history provides a detailed record of changes and updates made to the tax declaration data, including timestamps for each change. By analyzing this data, the project team can calculate the duration spent at each stage of the process and identify any areas where improvements can be made.

Report Development :

One of the particularly interesting steps in the F500 project was to integrate with a more dynamic and collaborative platform that acts as a data source rather than a traditional static database [17]. SharePoint Online serves not just as a plain old data repository but also as a collaborative tool where multiple users can interact with their data in real-time. The organization or the repository that the data resides under can also provide role-based accesses and this ensures data privacy standards within KPMG. Since it has access requirements, the complexity and the project development time was increased as it required handling various aspects like user permissions, version history, and live data updates. The process of establishing a connection to SharePoint involved leveraging its APIs and handling its unique data structures, which differ significantly from the relational tables typical of SQL databases.

Out of the many data sources supported by PowerBI, it also allows connection to a SharePoint Online list as well. Once the required access was provided, a connection to the SharePoint List was made using PowerBI directly and this gave access to all the tables present under the list. The required tables were then pulled in, which are as follows:

- Business Processes: This table contains information about the various business processes involved in the tax declaration process. It includes information about the Client, the entity, title of the CTR, the status and more information about Partners and Managers assigned to the same.
- EventStore: This table is the most important of all and contains information like the CTR name, the current status that the CTR is in, the time and date of modification and the person assigned to that particular CTR.
- User Information List: This table contains information about users involved in the tax declaration process, including user IDs, names, and roles. It helps in associating changes and updates with specific individuals.

There are specific relationships needed between all the tables for various visualization purposes, the tables mentioned above are imported into Power BI, where they are used to create a data model that supports the reporting and analysis needs of the project.

Report 1 : Partners

As mentioned before, there are two reports developed as part of this Project 500. The first one being the Partners Report has been developed to serve needs of KPMG Partners, which helps them in getting a comprehensive overview of all the Client Tax Returns across the organization. The report is designed to offer partners a deep insight into the status and progression of each CTR, which is combined with a detailed client and entity information.

When a partner accesses the dashboard, the system loads data related to all CTRs under their team and control. This data encompasses associated details for each client and their respective entities, providing a full view of the ongoing processes. The report does not only display data simply, but it serves as an important tool for partners to monitor progress, identify potential bottlenecks, and ensure timely completion of tax-related tasks. As mentioned before, the data source is the SharePoint Online list. The most important requirement from the Partners Report is an interactive bar chart, which graphically represents the percentage of the number of CTRs in each stage of the process. This visualization helps partners in quickly assessing the distribution of work and identifying areas that may require additional resources or attention. The bar chart is designed to be interactive, allowing partners to drill down into specific stages for more detailed information.

A detailed table in the Partners Report gives a complete list of key points, designed so that we can easily see how the Client Tax Return (CTR) is doing. This table is carefully designed to contain a number of important columns: the "Client" column, which states the name of the client for which the CTR is being processed, and the "Entity" column, which reveals the exact entity inside the organization of the client. The "CTR" column is the code that identifies each Client Tax Return that is shared to track and reference individual cases. The "Status" column denotes the current stage of the CTR within the tax declaration process, so it lets partners know what stage each return has reached currently. Furthermore, this table has the "KPMG Operator Name," who is the KPMG employee in charge of dealing with the records, and the "KPMG Manager Name," who is the manager in charge of the process respectively. The "KPMG Partner Name" column connects the partner with the client and the CTR, making sure that the person responsible is accountable and clear in managing the processes. The "Modified Date" column displays the date and time of the last modification to the CTR, so we get a timeline of the updates and changes. Finally, the "Year" column indicates the tax year for which the CTR is effective and, it helps in the organization and the historical tracking of returns. This configuration along with the details mentioned guarantee that partners can instantly retrieve and analyze the data they need, which in turn quickens the decision making and the execution of tasks as well as supports the firm's tax-related activities being carried out efficiently.

Report 2 : CTR TimeFrame Calculation

This is a specialized report designed for the Tax team, focusing on the details and stages involved in the Client Tax Return (CTR) processes. The primary objective of this report is to provide a comprehensive view of all CTRs, which shows their progress through various stages and calculate the time spent on each stage. This detailed information is very important for the Tax team to monitor efficiency, identify potential bottlenecks, and optimize the tax declaration workflow. The first report discussed above calculates the percentage of CTRs in each stage while this report calculates the time spent on each stage for each CTR.

The report is designed with an easy to understand user interface that allows users to effortlessly navigate through the data. Upon accessing the dashboard, the user from tax team will be presented with an overview of all CTRs, categorized by their current stages. This includes everything from initial submission to the final approval stages. The dashboard provides a holistic view of the time spent on each stage of the CTR, enabling the Tax team to understand the overall efficiency of the process. One of the key features of the dashboard is the ability to drill down into specific stages. By selecting a particular stage, users can view the cumulative time spent on that stage across all CTRs. This functionality is essential for pinpointing stages that may be causing delays or require additional resources, thereby facilitating improvements in the workflow.

Similar to the first report, the data source is a SharePoint Online List. The primary table used from this datasource is the EventStore table. This table is important and used as it records each event or action taken during the CTR process, including timestamps and details of changes, making it an important source for tracking the progression and time allocation of each CTR. A slicer is implemented to display the various stages of all CTRs currently in process. This slicer acts as a filter, allowing users to focus on specific stages by simply selecting them. Once a stage is selected, the report dynamically updates to display the total time spent on that stage, providing immediate visual results. This interactive feature not only enhances the user experience but also allows for quick identification of stages that may require attention due to processing times that are longer than expected.

In addition to the detailed view of individual Client Tax Returns, the report also provides multiple key performance indicators that offer a bigger and broader insight into the overall tax declaration process. These KPIs include the total number of CTRs, which helps to know the volume of work and the wide range of client engagements for the tax team. The total number of clients shown in the dashboard provides an overall idea of the client base covered, giving a sense of the report's reach. Moreover, the report tracks the time spent on the stage currently selected in the slicer, allowing the user from the tax team to focus on specific aspects of the process. This feature is particularly useful for identifying areas that may require additional resources or improvements. Finally, the report highlights the stage consuming the most amount of time, offering useful insights into potential bottlenecks or inefficiencies within the process of filing tax returns for the clients. This information can be very useful for the Tax team in prioritizing process improvements and optimizing their workflow.

An important step in development of this report is the time spent spent on each and every stage of the CTR. The SharePoint list has the modified time corresponding to the CTR but not the time spent. Therefore this is something that had to be calculated explicitly using PowerBI measures. To achieve this, the code first looks for the timestamp of the current event and identifies the business process it belongs to using the unique Business Process ID. It then searches through the dataset to find all events associated with the same business process. Among these events, it filters out those that occurred after the current event, focusing only on those that happened earlier. This filtering helps us in ensuring that the analysis remains relevant and corresponds only to the sequence of events leading up to the current stage. Once we find all the stages for one particular CTR, the code calculates the maximum timestamp from this group. This timestamp represents the latest stage that occured before the current stage within the same business process. By finding and recording this previous event, the code can then simply subtract the timestamp of the current event and the most recent previous event. This time difference is put in a new column of the table. The visual then uses this column to produce the visual and display the time spent on each stage.

After the time taken at each stage is calculated, another measure is created which would find out the stage out of all the currently available stages that takes the maximum time to complete. This stage is then populated in the KPI which was discussed above which calculates the stage on which most time is being spent. This information calculation and providing the same on the report is crucial for several reasons. It helps in analyzing the efficiency of the process, identifying any delays or bottlenecks that may exist between different stages. For example, if there is a significant gap between two events, it could indicate an area where the process is slowing down, suggesting a need for improvement or further investigation. Additionally, understanding these time intervals is essential for accurate reporting, allowing stakeholders to track the progress and status of various processes effectively.

This is how the reports are designed to be a comprehensive tool for monitoring and managing the tax declaration process. By utilizing the various PowerBI features, the reports present complex data in an easily understandable format, allowing users to make data driven decisions and take timely actions. One of the core objectives of Project F500 was to enhance the efficiency of the tax declaration process. This is done with the help of the two reports which were discussed above that monitor and manage the timing and status of various CTRs throughout their lifecycle. The project was designed to ensure that each CTR is tracked efficiently, from its initiation to its final submission, thereby enhancing the overall management of tax returns within KPMG. As mentioned before, these reports are used by the tax team to fulfill the tax obligations and then finally fulfill the obligations on myGuichet for the client.

3.3 Development of Deployment Pipelines

Deployment pipelines are very useful tools in the software development and the BI environments, as they help in the smooth transition of reports for BI purposes or any code when it comes to Software Development, across different stages and environments such as Development (DEV), User Acceptance Testing (UAT), and Production (PROD)[18]. The Pipelines were developed along with a Data Engineering team and implemented for the BI team for two types of reports, which are mentioned as follows :

- SSRS Reports: SQL Server Reporting Services (SSRS) reports are a feature of Microsoft's SQL Server that provide a detailed solution for creating, deploying, and managing reports. These reports are designed to be accessible across the entire organization. SSRS reports can be either paginated or mobile, that caters to different reporting needs.
- PowerBI Reports: Power BI reports are dynamic and interactive visual representations of data designed to provide insights and facilitate decision-making. They allow users to explore data through various visualizations like charts, graphs, and maps, enabling easy identification of trends and patterns. This makes them ideal for self-service business intelligence, empowering users to create, share, and collaborate on data insights efficiently.

These pipelines play a crucial role in ensuring that reports are developed and deployed in a structured, efficient, and secure manner. Utilizing deployment pipelines, especially with Azure DevOps, offers several advantages, including streamlined processes, reduced manual effort, and enhanced consistency across different environments. Earlier without the pipelines, a BI Engineer had to go through the manual effort of setting connection strings and uploading the reports to all the different environments. However, now with Deployment pipelines, these reports can be deployed as soon as they are uploaded to the repository. Azure DevOps is a suite of development tools and services that supports the entire application lifecycle, from planning and development to testing and deployment. It is very useful and important in setting up and managing deployment pipelines for both Power BI and SSRS reports. Azure DevOps integrates easily with Git to provide the required version control and management in a CI/CD Environment.

In this setup, Git repositories serve as the central place for all report development and updates. When a new report or an update is committed and pushed to the Git repository, Azure DevOps triggers a build pipeline. This automated process ensures that any changes are consistently and immediately compiled, tested, and prepared for deployment. There are two different kind of pipelines created. One is for PowerBi reports and the other is for SSRS Reports. The foundation of the pipeline lies in the structured architecture of the Git repository. The repository should be organized as follows:

```
– Project Reports
  – Project Reports
    – *.rdl
    – *.rdl.data
    – Project Reports.rptproj
  – Project Reports.sln
  – SQL
    – ...
– .gitignore
– README.md
```

This structure ensures that all necessary files, including the report definition files (.rdl), solution files (.sln), and project files (*.rptproj), are systematically organized. The .gitignore file helps ignore any unwanted files like build outputs or any other user specific settings, which can be ignored by simply mentioning the extensions in the .gitignore file. Secure handling of credentials and sensitive information is managed through the *Azure KeyVault*. Several information is passed to the keyvault like the service account username and password, the server URI and there should be a unique trigram assigned to each repository which uniquely identifies the artifacts deployed for that repository in the azure vault. Once the repository structure is present as mentioned above, another folder needs to be created at the root folder, called 'pipelines' which will have all the files required for actually setting up the pipeline [19]. These files are mentioned as follows :

- Variables.yml: This file is like the vault for implementation of the pipeline. As the name suggests, all the variables that are used throughout the pipeline are mentioned in this. It includes essential properties and key-value pairs like the folder name, which indicates the directory where the project report is stored. Additionally, it specifies the release folder name, which contains the report project files (.rptproj). The file also defines the artifact feed name, which is another repository where the built reports are stored and versioned. Furthermore, it includes the KeyVault trigram, a unique identifier that links the deployment pipeline to the specific Azure KeyVault holding sensitive credentials and server information, which was also mentioned above.
- Build-pipeline.yml: This configuration file outlines the steps for building the reports from the source files. It starts by defining the name and versioning strategy for the build, which helps in maintaining a systematic version control which is related to the major. minor and any patch updates. The build-pipeline.yml file also specifies the source repository from which the pipeline fetches the templates and scripts needed for the build process. It contains parameters for the build configuration, such as whether the build should be in Debug or Release mode, and a package description to briefly describe the update or changes made in that particular build. This is the file that builds the .rdl files and then deploys any of the generated artifacts to the artifact feed. It takes care of the correct versioning and the artifact is deployed along with the version that was mentioned.
- Deploy-pipeline.yml: The deploy-pipeline.yml is a crucial part of automating the deployment of SSRS reports within the MyBi project. This file contains the final steps necessary to take the reports, which have been built and versioned, and deploys them to the appropriate server environments. There are several variables which specify key settings like feedName, projectFolder, targetFolder, and keyVaultTrigram. feedName refers to the specific artifact feed (such as mybi-ssrs-feed) where the reports are stored post-build. The projectFolder specifies the directory containing the report files. targetFolder indicates where on the SSRS server the reports should be deployed. The keyVaultTrigram links the pipeline to Azure KeyVault, providing access to necessary credentials securely. If the keyvault has not been assigned, the deployments will not be possible. All in all, the deploy-pipeline.yml picks up the artifact from the artifact-feed which was uploaded earlier at the build-pipeline stage and deploys that artifact to the respective DEV, UAT and PROD environments. Currently, the deployment to DEV is automated but for the other two environments, it needs special approval from the Team Lead or Senior Engineers. This is to prevent any unneccessary deployments to the Production environment and maintains a second layer of security.

For a better understanding, the deployment workflow is explained below along with the figure:

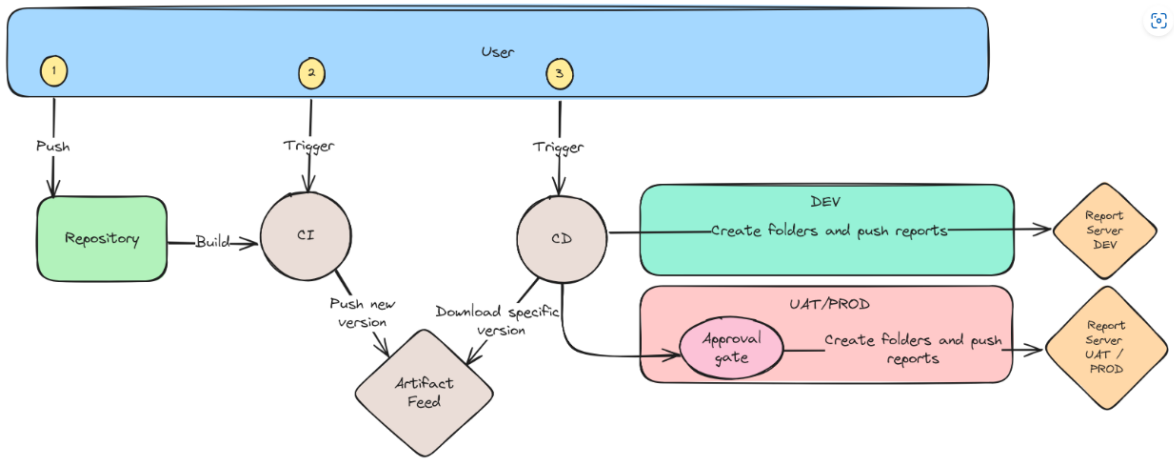

Figure 6: Deployment Workflow

The Figure 6 gives a detailed overview of how the Continuous Integration and Continuous Deployment process works for a typical workflow. The various stages and components of the pipeline are explained furthermore in detail below:

- Repository : The process begins with the Repository, which has all the source code, including the SSRS reports (.rdl files) and any associated project files. These repositories are hosted on Azure Repos which is a Git Platform.
- Build Process (CI) : When a new code change is pushed to the repository, it triggers the Continuous Integration (CI) pipeline. The CI pipeline performs several key functions:
  - Build: The CI pipeline builds the project, compiling the SSRS reports and other components which are relevant and necessary. This step ensures that the reports are correctly formatted.
  - Push New Version: After the build, the compiled reports and other artifacts are packaged and pushed to an Artifact Feed. This feed serves as a versioned storage for the build outputs, making it easy to retrieve specific versions for deployment.
- Artifact Feed : The Artifact Feed acts as a central repository for the compiled reports. It stores different versions of the reports, allowing the team to track changes over time and deploy specific versions as needed. This feed is particularly useful for ensuring consistency across different environments, such as DEV, UAT and Prod.
- Continuous Deployment (CD) : The next phase involves the Continuous Deployment (CD) pipeline, which automates the deployment of the reports to different environments. This stage is triggered by the user or automatically based on certain conditions:
  - Download Specific Version: The CD pipeline downloads the specific version of the reports from the Artifact Feed that needs to be deployed.
  - DEV Deployment: In the DEV environment, the pipeline creates the necessary folders and pushes the reports to the DEV Report Server. This environment is used for initial testing and development purposes.
  - Approval Gate: Before deploying to more critical environments like UAT or PROD, the pipeline includes an Approval Gate. This step requires manual approval, ensuring that only tested and validated reports are promoted to UAT/PROD environments. This was mentioned before regarding the approval from a Senior Engineer or the Team Lead.
  - UAT/PROD Deployment: Once approved, the pipeline creates folders and pushes the reports to the UAT or PROD Report Server. This final deployment step ensures that the reports are available for end-users or final testing in the UAT environment before going live in PROD.

This pipeline ensures a streamlined, automated process for deploying SSRS reports across different environments, significantly reducing the manual effort required and minimizing the risk of errors during deployment. By making use of Azure DevOps and Git repositories, the pipeline provides a detailed and easy to maintain framework for managing report versions and ensuring consistency across servers for Development, Testing and Production. A similar procedure is followed for deploying the PowerBI reports to the different servers [20]. The major difference is in the artifact feed. A separate feed is created for PowerBI reports so that all the PowerBI artifacts are stored in one place and these are kept separately from the SSRS reports. The deploy-pipeline.yml file in this case looks for any artifacts with the extension of '.pbx' instead of '.rdl' that was the case in SSRS reports. Once the repository structure is created and the files are added with the required paramaters, the pipeline can be set up explictly. Azure Devops provides a 'Pipeline' section which provides an easy way of setting up the pipeline once the required yml files are setup.

The implementation of the deployment pipelines for SSRS and PowerBI reports in the project, not only brought consistency in the deployment workflow but also streamlined the process of moving the reports throught the development, UAT and the production environment. As mentioned before, these deploymnets had to be done manually while taking care of all the versioning strategy, access rights and connection strings to all the different environments but now the pipelines have significantly reduced the manual effort that was typically associated with these deployments. By using Azure DevOps and the Git repositories, the pipelines ensured an easy and efficient transition of reports from the development to the deployment stages, which helps in maintaining consistency and reliability across the three different environments. Additionally, this project enhanced inter-team communication and collaboration, as the DevOps and the BI teams worked closely together to align the technical requirements with the Business Intelligence needs. This collaboration was crucial in ensuring that the deployment pipelines were well-integrated with existing systems and workflows maintained by the DevOps team.

4 Conclusion and Future Work

The completion of the whole project at KPMG Luxembourg marks a significant achievement in enhancing the BI team's data reporting and deployment capabilities, using multiple layers of innovation, and development across various domains. The project began with a critical focus on data security and privacy, specially through the implementation of Data Classification using data masking and randomization techniques. This was an essential step to make sure that any sensitive information belonging to a client was protected throughout the report development lifecycle. By using a cross join operation, we generated a large number of fake names and randomized the data, providing the stakeholders with realistic data for testing and development without compromising any sensitive client data. This approach only made sure that the BI team follows the KPMG global policies.

Once the data preparation was done, the focus was on the creation of comprehensive reports, which was crucial for both internal monitoring and client-specific deliverables. Among these reports, the server monitoring report provided valuable insights into the utilization of various reports across the organization which helped in the management of BI infrastructure. Additionally, the F500 tax declaration reports were developed to assist the tax team in maintaining the tax compliance obligations for the clients. These reports provided detailed analytics on the tax declaration process, along with the tracking of each Client Tax Return (CTR) through the various stages. This not only made the client engagament and transparency better but also imporved the internal processes by highlighting the bottlenecks and any inefficient behaviour.

After securing the data and developing the reports, the final phase of the project involved setting up deployment pipelines for all the reports within the BI space. These included all of the PowerBI reports and the SSRS reports. By using DevOps, I established automated pipelines that significantly reduced the manual effort that engineers were putting in for the deployment across different environments. The integration of Git Repositories through Azure Repos allowed for version control and continuous integration/continuous deployment (CI/CD) processes. This automation not only improved the efficiency of the deployment process but also minimized the risk of errors, thereby ensuring that the end-users received the highest quality reports.

Overall, the project has not only advanced the technical capabilities but also enhanced the inter-team collaboration. This collaboration was important regarding the complexities in report development, data security and the deployment. I had to make sure that everything works well together in a well integrated manner.

4.1 Future Work

As far as any future work is concerned, the foundations were laid down by this project and they will serve as a good platform for further innovations and improvement in the BI team architecture. There are some additional features that can be added to the project completed as part of the internship. This is mentioned as follows:

- Data Classification: The Data Classification which was done at the very initial stage can be enhanced further to support a script that would just ask for the columns needed to be masked, instead of going through the whole database and selecting the columns one by one.
- Monitoring Report: There can be real time alerting features implemented in the Monitoring report. For example, when there are multiple reports failing, the BI team can be notified about the same and any potential issues can be taken care of. This can be done with Power Automate.
- F500 Report : Currently, the different tables that are used for report development are filled in with details manually. There can be a script that would fill in the tables if the details have been provided in any of the tables used.
- Cloud Migration : As part of a future plan, all the PowerBI Reports need to be migrated to cloud and hence, there is a need for gateways to connect to the on-premise data sources while the reports are in the cloud.

- Pipelines for Cloud : The team does currently have PowerBI cloud reports, but there are no deployment pipelines for the same. These need to be created, much more now as there will be migration to the cloud. There is a requirement for a dedicated service account that would do the data model refreshes on the cloud.

The deployment of on-premise PowerBI reports has paved way to extend the deployments to cloud. There is additional work required for connections by using specific service accounts and for refreshing the data over the cloud. Moreover, the initial data classification

Throughout my time on the project in KPMG, I was actively engaged in various team meetings, including planning sessions, retrospectives and daily stand-ups. My contributions were often recognized, particularly during retrospectives where my efforts in data masking and randomization, and collaborating with KPMG Netherlands team for resolving bugs in some other reports received commendation. I was acknowledged for my ability to deliver high-quality results even when the deadlines were tight. Additionally, my internship tenure included comprehensive training sessions provided by KPMG, which gave me much needed important knowledge related to data management and maintaining and adhering to the strict data privacy guidelines set by an audit firm as big as KPMG. Beyond the main project, I also contributed to smaller projects, where I implemented enhancements and resolved issues with existing reports. Moreover, I was actively writing documentation for all my work and this was posted to the team's internal Wiki which can be referred to in the future whenever needed. The six-month period in a global firm like KPMG not only broadened my experience but also deepened my understanding of the organization's reporting needs and infrastructure. The extensive importance given to any task related to handling of data was very intriguing to me. These experiences were invaluable in refining my technical skills and enhancing my ability to work effectively within a team-oriented, agile environment.

5 References

1. Gupta, Rajendra. "SQL data classification – Add sensitivity classification in SQL Server 2019".SQLShack, Adding Sensitivity label-SQLShack.
2. Colley, Derek. "Automatically Create and Anonymize Downstream Databases from Azure DB". mssqltips, Anonymizing the Databases - Azure DB.
3. Tripathy, Madhumita, et al. "SQL Data Discovery and Classification". Learn Microsoft, SQL - Data Discovery.
4. Hutmacher, Daniel. "Fun with Random Names". SQL Sunday, Random Name Generation.
5. Dave, Pinal. "SQL SERVER – Find Column Used in Stored Procedure – Search Stored Procedure for Column Name". SQL Authority, Searching column names in a stored procedure.
6. Dave, Pinal. "SQL SERVER – Find Column Used in Stored Procedure – Search Stored Procedure for Column Name - Part 2". SQL Authority, Searching column names in a stored procedure - Part 2.
7. essentialSQL. "What is a SQL Server Data Dictionary?". CodeProject, Data Dictionary - Information Schemas.
8. Pollack, Edward. "Building a SQL Server data dictionary". red-gate, Building Data Dictionaries.
9. "SQL Server reporting services (SSRS) documentation". ApexSQL, SSRS Documentation.
10. Neugebauer, Niko, et al. "System catalog views (Transact-SQL)". Learn Microsoft, Catalog Views - Transact SQL.
11. K, Elazar, et al. "SQL information protection policy - Microsoft for Cloud". Learn Microsoft, Information Protection Policy.
12. Bar, Paulin, et al. "Dynamic Management Views (DMVs)". Learn Microsoft, Dynamic Management Views.

13. Murray, Scott. "SQL Server 2012 Analysis Services (SSAS) DMVs". mssqltips, SSAS DMVs.

14. Tanushree. "Cardinality in DBMS". GeeksforGeeks, DBMS Cardinality.

15. Myers, Miguel, et al. "Create a relative date slicer and filter in Power BI". Learn Microsoft, Relative Date Slicer.

16. Bar, Paulin, et al. "Workspaces in Power BI". Learn Microsoft, PowerBI Workspaces.

17. Zhou, Rico. "How to get version history from SharePoint". Microsoft Fabric Community, Sharepoint Version History.

18. Wright, Josh, et al. "Deployment jobs". Learn Microsoft, Pipelines with Azure DevOps.

19. D, Steve. "YAML schema reference for Azure Pipelines". Learn Microsoft, YAML Reference.

20. Romano, Rui. "Power BI Project (PBIP) and Azure DevOps build pipelines for validation". Learn Microsoft, Pipelines for PowerBI Projects.