MASTER

DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK

INTERNSHIP REPORT

LEVERAGING BUSINESS INTELLIGENCE THROUGH ETL PROCESSES AND POWER BI DASHBOARD DEVELOPMENT

LEONOR BARBOSA LEAL SIMÕES DOS SANTOS

MARCH - 2024

Lisbon School of Economics & Management
Universidade de Lisboa

MASTER
DATA ANALYTICS FOR BUSINESS

MASTER´S FINAL WORK
INTERNSHIP REPORT

LEVERAGING BUSINESS INTELLIGENCE THROUGH ETL PROCESSES AND POWER BI DASHBOARD DEVELOPMENT

LEONOR BARBOSA LEAL SIMÕES DOS SANTOS

SUPERVISION:
JESUALDO FERNANDES

MARCH - 2024

LIST OF ABBREVIATIONS

AWS – Amazon Web Services.

BI – Business Intelligence.

DAX – Data Analysis Expressions.

EC2 – Elastic Compute Cloud.

ER – Entity Relationship.

ETL – Extract, Transform and Load.

G&L – General Ledger.

IBM – International Business Machines Corporation.

JEL – Journal of Economic Literature.

KPIs – Key Performance Indicators.

MFW – Master's Final Work.

P&L – Profit and Loss.

ABSTRACT

This report aims to outline the tasks accomplished during the internship at LGG Advisors as part of the Master's in Data Analytics for Business program at ISEG. The main goal of the internship was to enhance the intern's skills in data analytics, with a particular emphasis on refining methods related to data acquisition, processing, and integration within a database framework, as well as improve the ability to create dynamic and informative dashboards. During the internship, the intern was placed with the Data Transformation team at LGG Advisors, where their primary duties involved ETL processes and generating Power BI reports for LGG clients.

The acquired skills in ETL processes streamlined data acquisition and preparation and laid the foundation for developing robust and insightful dashboards.

The developed dashboards allowed store directors and executives to effectively compare performance across different locations, identify areas for improvement, and implement data-driven decisions that boosted profitability.

The internship reinforced the invaluable role of data analytics and business intelligence principles, actionable insights that drive business success.

KEYWORDS: Business Intelligence; Power BI; Data Visualisation; ETL Processes.

TABLE OF CONTENTS

List of Figures

LIST OF TABLES

Acknowledgments

To my loving parents, Mafalda and Miguel, and my brother, Simão. Thank you for always believing in me, even when I doubted myself. Your love, sacrifices, and encouragement are the core of my success.

To my dearest Avó Bia, your love, wisdom and belief in my potential continue to inspire me. I miss you dearly.

To my cousin Inês, thank you for opening your home to me during my time in Lisbon. Your warmth transformed a potentially lonely experience into an enriching adventure filled with laughter and memories I will cherish forever.

To all my family and friends, I am forever grateful for your unwavering support.

To Professor Jesualdo Fernandes, I want to express my deepest gratitude for their invaluable guidance and support throughout the development of this thesis. Thank you for accepting this challenge.

I also extend my appreciation to LGG Advisors, particularly the Digital Transformation team that I had the pleasure of working with during my internship. The hands-on experience and knowledge gained with this internship have significantly contributed to the practical aspects of this report.

LEVERAGING BUSINESS INTELLIGENCE THROUGH ETL PROCESSES AND POWER BI DASHBOARD DEVELOPMENT

By Leonor Santos

1. INTRODUCTION

In today's data-driven world, businesses are awash in a sea of information, from customer transactions to operational metrics to market trends. However, more than simply having data is required; organisations must transform this data into actionable insights that drive strategic decision-making. This is where business intelligence (BI) comes into play (IMD, 2024).

BI encompasses a range of strategies, processes, applications, data, products, technologies and technical architectures that enable organisations to collect, analyse, and visualise data to gain valuable insights and improve decision-making (Kumar & Belwal, 2017). Business intelligence relies on a structured approach encompassing four key steps to convert raw data into easily understandable insights accessible to all within an organisation (Microsoft Power BI, 2020):

i. **Collect and transform data from multiple sources**: business intelligence tools often make use of the ETL (extract, transform, and load) method to gather and restructure data from multiple sources, including both structured and unstructured data. The transformed data is then stored in a central location, enabling more efficient data analysis and query processing by various applications.

ii. **Uncover trends and inconsistencies**: data mining is the automated analysis of data to find patterns and anomalies. BI tools include multiple forms of data modelling – process of visualising an information system or sections of it to communicate the relationships between the data points and structures; the purpose is to highlight the many forms of data used and stored inside the system, the relationship between the different data types, the various ways the data may be categorised and arranged, and its formats and attributes (IBM, 2020) – and analytics – exploratory, descriptive, statistical, and predictive – that allow users to analyse data, forecast patterns, and offer suggestions.

iii. **Use data visualisation to present findings**: data visualisations are used in business intelligence reporting to make results easier to understand and share.

iv. **Take action on insights in real-time**: the ability to analyse both current and historical data in the context of business activities allows companies to quickly move from insights to action. This enables them to make real-time adjustments and long-term strategic changes, empowering businesses to make better decisions and stay ahead of the competition. (Microsoft Power BI, 2020)

One key aspect of business intelligence is the integration of data from disparate sources, which can be a complex and time-consuming task, hence ETL (Extract, Transform, Load) processes. This integration process lays the foundation for the subsequent steps in the BI journey (Amazon, 2022).

ETL is a data integration methodology that involves three main steps – extracting data from various sources, transforming it into a consistent format, and loading it into a centralised repository – and ensures that the data is clean, accurate, and ready for analysis (Dhanda & Sharma, 2016).

Once the data is integrated using ETL, it can be used by business intelligence tools to analyse historical and current data and present findings in intuitive visual formats (Microsoft Power BI, 2020). One such tool is Power BI, a Microsoft business intelligence platform.

Power BI is a powerful cloud-base business analytics service for transforming raw data into actionable insights. It provides a user-friendly interface to create interactive dashboards and reports by connecting to diverse data sources and can also be used to further clean and transform data. Power BI's visual storytelling capabilities helps organisations to effectively share insights with their stakeholders (Ali et al., 2016).

By leveraging ETL processes and Power BI dashboard development, organisations can unlock the true potential of their data, transforming it into a powerful tool for driving business growth and success (Makara, 2023).

This internship report will go over the activities and work performed during the internship at LGG Advisors and is divided into four main sections. The first section, *Introduction*, briefly discusses the pivotal role of BI in using ETL processes and Power BI to transform raw data into actionable insights for driving business growth and success in today's data-driven environment. Section 2, *Literature Review*, is divided into two subsections: (i) *ETL Processes for Data Collection and Integration* which delves into the definition of ETL processes, explores various ETL tools and emphasises the importance of data integrity, and (ii) *Power BI Dashboard Development* which covers aspects like data preparation and the specifics of creating impactful dashboards to drive actionable insights in Power BI. Section 3, *Methodology*, summarises the main activities conducted during the internship and explores the created ETL process and its implementation. The report concludes with Section 4, *Work performed during the Internship*, providing a comprehensive understanding of the gained practical experiences.

1.1 Objectives of the Internship

The primary objectives of the internship were to apply the knowledge gained during the Master's program and gain practical insights into how data analytics works in real-world scenarios.

The internship was driven by the core objective to foster a comprehensive skill set in data analytics, with a focused approach to refining techniques related to data acquisition, processing, and integration within a database framework, as well as perfecting the ability to construct dynamic and informative dashboards. This was facilitated through hands-on experience with various technologies, including Microsoft Power Automate, Power Query M, DAX, Power BI, and MySQL Database.

Integral to the internship was active participation in the department's day-to-day operations, involving regular briefings and meetings. Key objectives achieved during the internship included:

i. Deep engagement in case studies, providing insights into real-world scenarios: analysis of each case, aiming to understand the presented challenges and create viable solutions based on theoretical knowledge and practical application.
ii. Creation of impactful dashboards using Power BI: transformation of raw data into visual insights crucial for the decision-making processes.
iii. Participation in the formulation and structuring of databases: contribution with insights into database design and organisation, ensuring efficiency and usability.
iv. Contribution to Extraction, Transformation, and Loading (ETL) processes: development and implementation of automation solutions to extract data files.

1.2 LGG Advisors

In the ever-evolving business landscape, LGG Advisors emerges as a strategic partner, empowering organisations across industries to achieve sustainable growth and operational excellence. With a comprehensive suite of knowledge services and a flexible team augmentation model, LGG Advisors guides businesses through complex challenges and seizes emerging opportunities.

Their team of experts specialises in the core areas of Finance & Strategy, Risk & Compliance, Digital Transformation, Creative & Design, and Administrative & Support. By seamlessly integrating with clients' teams, they provide tailored solutions that align with their unique business goals and objectives. Their strategic approach fosters long-term value creation and sustainable growth.

With a global network of experts and a deep understanding of local markets, LGG Advisors caters to the specific needs of their clients worldwide. Their expertise transcends borders, ensuring businesses across geographies receive relevant and practical guidance to drive success.

During my internship, I was assigned to the Data Transformation Team responsible for handling all aspects of data, including collection, cleaning, analysis, and visualisation, to drive informed decision-making within the company and cater to our clients' requirements.

2. LITERATURE REVIEW

2.1 ETL Processes for Data Collection and Integration

Kimball & Ross (2013) describe ETL systems as an environment that encompasses a workspace, data structures, and a set of processes. These processes, as seen in Figure 1, include the extraction, transformation and loading of data.

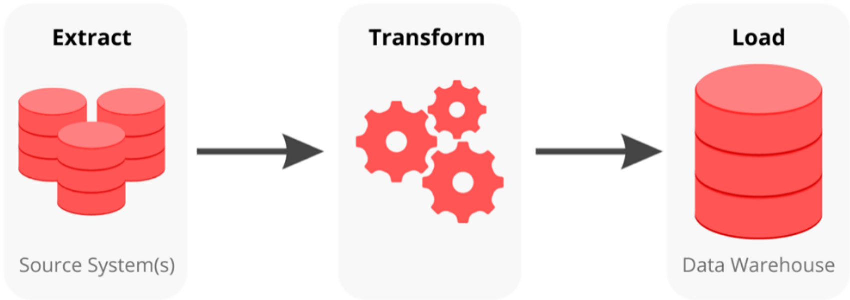

Figure 1: ETL Process. Source: The author.

The first step in integrating data into a DW environment is the extraction (Kimball & Ross, 2013) of data from various sources, which could include transactional systems, databases, spreadsheets, or other repositories (Dhanda & Sharma, 2016). During this step, it's essential to selectively extract relevant data while ensuring that the operational system's performance is not compromised (Dhanda & Sharma, 2016).

Once the data has been extracted from the source systems, it is transformed into a consistent format for analysis and reporting. This step may involve cleaning, combining and deduplicating data (Kimball & Ross, 2013). The transformation step can be the most challenging and resource-intensive aspect of ETL, as it requires complex rules and functions. The transformed data is then stored in a second staging area or directly loaded into the data warehouse (Dhanda & Sharma, 2016).

The final step in the ETL process is the loading stage which considers factors such as timing, system impact, and whether the loading is performed in batch or real-time mode. In this step, the transformed data is incorporated into the target system, usually a data warehouse (Dhanda & Sharma, 2016).

A data warehouse is an enterprise system used for the analysis and reporting of structured and semi-structured data from different sources, including point-of-sale transactions, marketing automation, customer relationship management, and more. It can store both current and historical data in one location and is intended to provide a long-term view of data over time, making it an important component of business intelligence (Google Cloud, 2023).

2.1.1 ETL Tools

The spectrum of ETL tools has expanded from traditional on-premises solutions like Informatica and Microsoft SSIS to cloud-based platforms such as AWS Glue and Azure Data Factory to accommodate diverse infrastructural preferences and scalability demands. These tools assist businesses in transferring data from several channels to a single destination (Sreemathy et al., 2021) and by minimising the need for "hand-coding" while developing a new data warehouse they save time and money (Mali & Bojewar, 2015).

The selection of an appropriate ETL tool depends on the organisation's specific needs and requirements. Gaikwad (2023) identifies four main types of ETL tools – enterprise ETL tools, open-source ETL tools, custom ETL tools, and cloud-based ETL services – and highlights that understanding the nuances between them is essential to make informed decisions tailored to each company's unique data management needs.

Enterprise ETL Tools are software programs built for large enterprises to manage a significant volume of data and are mainly used for data integration, warehousing, and business intelligence applications. Enterprise ETL systems frequently offer a distinct collection of features and capabilities that allow them to perform complicated data transformations and manipulation activities, plan and automate ETL operations, and manage large volumes of data sources and targets (Gaikwad, 2023). Examples of these tools are:

i. IBM InfoSphere DataStage: known for scalability and parallel processing, IBM DataStage offers a graphical interface for robust ETL processes. It supports complex transformations, diverse connectors, and scheduling for seamless data integration across varied environments (IBM, 2023).

ii. Informatica PowerCenter: serves as a comprehensive enterprise data integration platform, functioning as a unified solution for Cloud Data Integration, Data Migration, Data Quality, Data Replication and

Synchronization, Data Virtualization, and more. This tool comprises three primary components: Informatica PowerCenter Client Tools, Informatica PowerCenter Repository, and Informatica PowerCenter Server (Mali & Bojewar, 2015).

Open-Source ETL Tools are free software that may be customised to meet individual needs. Because the source code of open-source ETL solutions is publicly available, data scientists and analysts are able to easily evaluate, costumise, and improve the businesses' ETL processes (Gaikwad, 2023). Some examples are:

i. Pentaho Data Integration (PDI): formerly known as Kettle, PDI is an open-source tool with a graphical environment that is suitable for both small and big firms. The platform can be implemented on-premises and supports data aggregation and sorting from an array of data sources (Sreemathy et al., 2021).
ii. Talend: versatile tool used for integration across operational systems, migration, business intelligence, ETL, and data warehousing. Its standout feature is the ability to produce fully editable Java code, offering users customisation and flexibility in handling data integration tasks efficiently and precisely (J. Sreemathy et al., 2020).

Custom ETL Tools are tailored to satisfy the needs of a certain company. Though they give sufficient flexibility, they also need significant work because data pipelines must be built from the ground up. Organisations have to handle maintenance, produce documentation, and continuously test and check for development on their own. Custom ETL tools and pipelines are designed and built by organisations adopting scripting languages such as Python and SQL as well as technologies like Hadoop and Spark (Gaikwad, 2023).

Cloud-Based ETL Services enable enterprises to quickly and effectively do ETL operations in a cloud computing environment. Some ETL cloud services are extremely proprietary and can only be used inside the cloud vendor's framework, so companies cannot employ ETL cloud services on the platform of a separate cloud vendor (Gaikwad, 2023). Some of these services include:

i. AWS Glue: fully managed ETL service on AWS, offers serverless data integration, automating ETL processes for data lakes or warehouses. It supports multiple data formats, integrates well with other AWS services, and includes data cataloguing and metadata management features (AWS, 2022).

ii. Azure Data Factory: Microsoft's cloud-based service facilitates the creation, scheduling, and management of data pipelines. It supports various connectors, simplifies complex workflows through a visual interface, and integrates with Azure services for seamless data movement and transformation (Microsoft Azure, 2014).

To provide a concise and comparative overview of the diverse landscape of ETL tool types discussed above, Table I outlines the main purposes, characteristics, and examples of each tool.

TABLE I: ETL TOOLS PURPOSE, CHARACTERISTICS AND EXAMPLES. THE AUTHOR.

Type	***Purpose***	***Characteristics***	***Example***
Enterprise ETL Tools	Handle large volumes of data from diverse sources.	- Specialised for complex data transformations and manipulation; - Automation of ETL processes; - Integration capabilities.	- IBM DataStage; - Informatica PowerCenter.
Open-Source ETL Tools	Offer flexibility and cost-effectiveness.	- Publicly available source code for customisation; - Free software.	- Pentaho Data Integration (PDI); - Talend.
Custom ETL Tools	Provide maximum customisation to meet specific needs.	- Built to cater to unique company requirements; - Uses scripting languages and technologies.	Custom pipelines using Python, SQL, etc.
Cloud-Based ETL Services	Streamline ETL processes in cloud environments.	- Cloud-native and scalable solutions; - Vendor-specific limitations.	- AWS Glue; - Azure Data Factory.

2.1.2 Data Integrity and Quality in ETL Processes

Data integrity and quality are fundamental concepts in information analysis, crucial for ensuring that data is accurate and reliable but also consistent, relevant, and secure throughout its lifecycle. In today's data-driven world, where information serves as the cornerstone of decision-making and operations across industries, maintaining the integrity and quality of data is paramount, and even though these two concepts are often used interchangeably, they have distinct meanings and roles in ensuring data is trustworthy and useful (Przybycień, 2023).

Data integrity refers to the completeness, correctness, consistency, accessibility, and security of data, while data quality measures the amount of data integrity and, as a result, its dependability and appropriateness for its intended purpose (Przybycień, 2023). Data quality is, therefore, crucial for the success of any organisation that relies on data-driven decision-making.

ETL processes play a critical role in ensuring the integrity and consistency of data as it moves from various sources into a central repository. However, these processes can introduce data quality problems if they are not properly designed and implemented. Souibgui et al., 2019 give some examples of data quality problems in the different stages of the ETL process. Some of these common issues are:

i. **Duplicate issues**: when multiple records with the same or very similar information are present in a dataset, it can lead to inaccurate analysis and decision-making. An example of this issue would be a customer database containing multiple entries for the same individual due to variations in data entry, such as using different email addresses or misspellings of names.
ii. **Inconsistent data formats**: different data sources may have varying formats for representing the same information, making it difficult to combine and analyse the data effectively. For example, a date field in one source may be formatted as YYYY-MM-DD, while the same field in another source may be formatted as MM/DD/YYYY.
iii. **Missing or incomplete data**: when a required field in a record is empty, contains no data, or contains some but not all necessary information. A sales dataset might lack, for example, information like salesperson IDs or product

codes for certain transactions, making it challenging to trace sales or attribute them accurately.

iv. **Data integrity and accuracy issues**: when data violates the defined rules or constraints within the system and is incorrect or misleading. For example, a product price in the inventory system may be incorrect due to an error in data entry, which could lead to inaccurate financial reporting.

Data-driven decision-making is the cornerstone of successful organisations. However, if the data used for these decisions is inaccurate or unreliable, the resulting insights can be misleading and lead to poor decision-making that can have severe consequences. As seen from the abovementioned issues, poor data quality can lead to flawed conclusions, potentially cause financial losses and damage a company's reputation (Hibencarey, 2023). For instance, faulty sales forecasts based on inaccurate data can lead to overstocking or understocking, resulting in financial losses and customer dissatisfaction. Similarly, inaccurate financial reporting due to data quality issues can erode investor confidence and damage the company's reputation (Alooba, 2023).

Addressing these issues through effective ETL processes streamlines operations, reduces errors, and frees up resources to focus on more strategic initiatives. By ensuring data integrity and consistency, organisations can enhance operational efficiency, improve customer satisfaction, and gain a competitive edge in the marketplace (TestingXperts, 2023).

Souibgui et al. (2019) suggest two main approaches when dealing with data quality management: process-centred and data-centred. The latter can be divided into syntactic, semantic, and contextual oriented approaches.

The process-centred approach strategically manages data quality within business processes by incorporating user-defined quality characteristics. It involves integrating these characteristics into data processing, analysis, and decision-making workflows, recognising users' varied perceptions of quality. Specifically, tools like Bijoux, proposed by (Theodorou et al., 2014), and POIESIS are employed: Bijoux aids in generating diverse test scenarios based on user preferences, using an algorithm to explore different paths in the ETL process, while POIESIS introduces alternative flow patterns aligned with user-defined quality preferences. These tools provide users with insights into performance, reliability, deduplication, and null value filtering, enabling them to select the most suitable data quality measures tailored to their needs (Souibgui et al., 2019).

The data-centred approach, as mentioned before, enters its efforts directly on the data itself, aiming to ensure the quality of the data independent of specific processes and is proposed by Souibgui et al. (2019) to be divided into three approaches:

i. **Syntactic approach**: ensures data structure, format, and adherence to predefined rules and standards. It encompasses rectifying issues like data accuracy, completeness, consistency, and uniqueness and, for instance, addressing problems such as data duplication by employing algorithms and metrics to identify matching records.

ii. **Semantic approach**: delves into understanding the meaning and interpretation of data elements. It aims to guarantee the correct interpretation and usage of data across various systems. This involves tasks like semantic mappings to integrate disparate data sources and enriching data with semantic information to improve cleansing and accuracy.

iii. **Contextual approach**: acknowledges the context surrounding data creation, storage, and use. It involves understanding the environment in which data exists and applying measures to manage its quality in specific contexts or situations. This includes considering factors like usability, influence on data quality measures, and the significance of context in the ETL process.

Table II provides a comparison between the process-centred and data-centred approaches.

TABLE II: COMPARISON OF DATA QUALITY MANAGEMENT APPROACHES. THE AUTHOR.

Aspect	***Process-Centred Approach***	***Data-Centred Approach***
Focus	Manages data quality within business processes.	Concentrates on ensuring data quality independently of specific processes.
Primary Emphasis	User-defined quality characteristics in workflows.	Addresses data quality irrespective of processes.

Methodology	Integrates quality characteristics into workflows.	Modifies data directly to ensure quality standards and reliability.
Tools/Approach	Bijoux, POIESIS (tools for test scenario generation).	Syntactic, Semantic, Contextual approaches (focus on data structural, meaning, context).
Benefit	Enhances data quality aligned with user perceptions.	Ensures overall data reliability across systems and processes.

2.2 Power BI Dashboard Development

Power BI is a versatile business intelligence platform developed by Microsoft that empowers businesses to transform their data into actionable insights by combining Excel's "Power" add-ons (Kline, 2014) – Power Query, Power Pivot and Power View. It seamlessly integrates data from a wide range of sources, enabling users to analyse and visualise information intuitively. Power BI's capabilities span across various aspects of business intelligence, including data preparation, data visualisation, report generation (see the example given by Figure 2), and data sharing (Microsoft, n.d.-c).

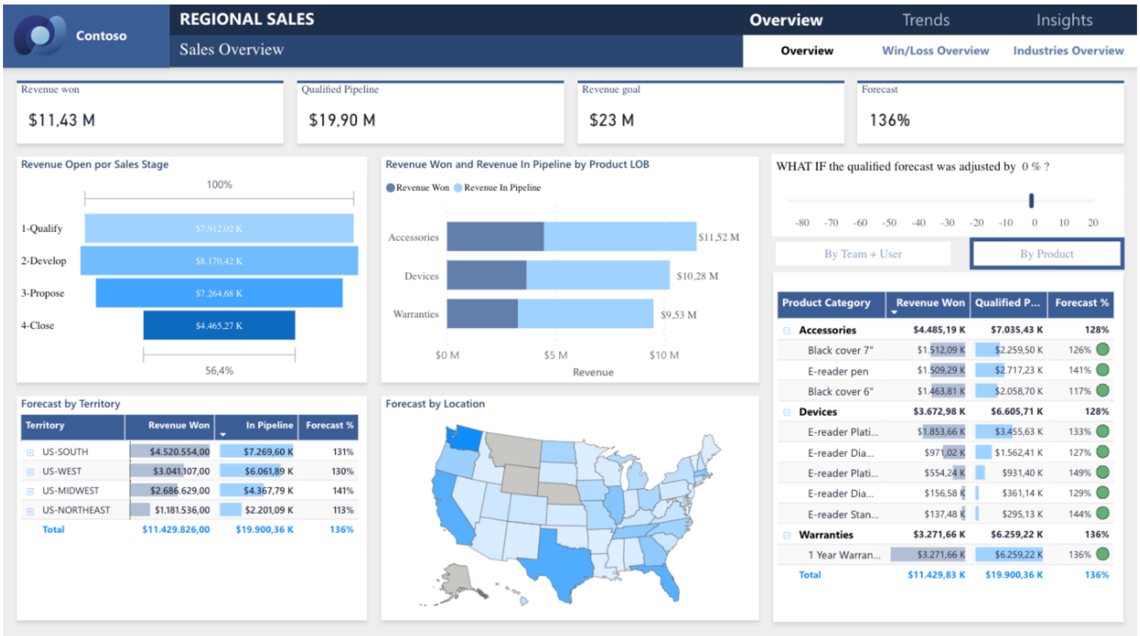

Figure 2: Example of a Power BI Report (Microsoft, 2023c).

Power BI extends and enhances Excel's capabilities to create a more sophisticated business intelligence platform. With Power Query, Power BI allows users to connect to diverse data sources, facilitating data transformation and preparation, while Power Pivot enables them to create intricate data models with relationships, calculated columns, and measures. The visualisation capabilities in Power BI surpass those of Power View by providing a versatile range of interactive charts, graphs, and advanced features. Its support for custom visuals, interactive dashboards, natural language querying, cloud-based collaboration, and integration with advanced analytics sets it apart as a more comprehensive and dynamic business intelligence platform. Additionally, Power BI introduces features like natural language querying and seamless integration with other Microsoft services, providing a comprehensive solution for data analysis (Becker & Gould, 2019).

Dashboard development is an essential aspect of Power BI; however, despite their widespread use in business, designing and executing effective dashboards can take time and effort, leading to failures in achieving their intended objectives (Kumar & Belwal, 2017). Therefore, there is a need for a systematic and strategic approach to dashboard development in Power BI to ensure that the created dashboards effectively support organisational objectives and facilitate informed decision-making.

At its core, Power BI is a comprehensive tool for gathering, organising, analysing, and presenting data. Its primary purpose is to equip businesses with the ability to make informed decisions based on data-driven insights. This encompasses a range of objectives such as gaining a comprehensive understanding of business operations (allows businesses to aggregate data from various sources, providing a full view of their operations), identifying trends and patterns (its powerful data visualisation capabilities enable businesses to identify trends, patterns, and correlations within their data, helping organisations understand customer behaviour, market trends, and operational inefficiencies), and making data-driven decisions (armed with insights from Power BI, businesses can make informed decisions that align with their strategic goals) (Microsoft Power BI, 2020).

In the broader context of business intelligence, Power BI incorporates various aspects of BI, as mentioned by Gestisoft (2023), including:

i. **Data Integration**: connects to various data sources – cloud-based services, on-premises databases, Excel spreadsheets, etc. – which enables businesses to centralise their data from multiple silos into a unified platform.
ii. **Data Preparation**: simplifies data preparation through data cleaning, transformation, and modelling techniques.
iii. **Data Visualisation**: offers various interactive data visualisation tools, enabling users to create charts, graphs, and dashboards that effectively communicate insights.
iv. **Report Generation**: facilitates the creation of interactive reports that can be shared with stakeholders across the organisation. These reports can be customised to suit specific needs and provide a centralised BI repository.
v. **Data Sharing**: enables data sharing within the organisation and beyond.

2.2.1 Data preparation in Power BI

Data preparation is a crucial step in the data analytics process that involves collecting, shaping, and organising data to be efficiently used in BI applications, such as Power BI. It helps clean, transform, and shape the data to ensure it is accurate, consistent, and ready for visualisation and insights (Stedman, 2022).

Power BI is a progression from the initial add-ins in Excel, including Power Query, Power Pivot, and Power View (Ferrari & Russo, 2016). While Power View was an early step in Microsoft's journey towards empowering users with self-service BI, Power BI represents the evolution of these tools, providing a more comprehensive and powerful solution for today's data-driven organisations.

Power Query is a powerful tool in Power BI that allows the cleaning, transformation, and shaping of data from various sources before it is loaded into the data model (Becker & Gould, 2019). It provides a user-friendly interface with many features, including data source connectivity, data transformation, formula language (M), applied steps, query folding, and custom functions (Microsoft, 2024b).

In the context of Power BI, data preparation is performed using Power Query, and it involves several key steps, including data collection, data cleaning and data transformation (Stedman, 2022):

i. **Data Collection**: gather data from various sources and import or load the data into Power BI's workspace.

ii. **Data Cleaning**: identifying and correcting errors, inconsistencies, and missing values in the data. Common techniques include:
    - Missing data: identifying and imputing missing values to prevent them from skewing the analysis.
    - Duplicate data: identifying and removing duplicate records to avoid counting the same data multiple times.
    - Irrelevant data/Outlier detection: identifying and addressing outliers that may distort the overall data distribution.
    - Inconsistent data: standardising data formats and naming conventions to enhance data readability and consistency.

iii. **Data Transformation**: involves manipulating and transforming data to make it suitable for visualisation and analysis. This step may involve:
    - Create Calculated Columns: extend the available data dimensions and provide additional filtering and analysis options.
    - Data Aggregation: summarising and grouping data effectively, such as calculating averages, sums, or counts.
    - Data Filtering: applying filters to focus the analysis on specific data points or segments of interest.

Power Pivot is a data modeling technology from Microsoft that allows you to create data models, analyse large amounts of data, and create sophisticated business intelligence solutions (Microsoft, n.d.-d). It is integrated into Power BI as a feature and has two essential components:

i. **Data Modeling**: organising data into a structured and efficient format that supports effective analysis. This involves creating a data model that reflects the relationships between different data sources. Key techniques include (Atlan, 2023):

- Data relationship definition: defining the relationships between tables with Power BI's model view using star schema to establish clear connections between data entities.
- Data normalisation: organising and structuring data to reduce redundancy and improve efficiency.
- Data denormalisation: involves duplicating data to improve query performance and reduce the number of joins required. However, it's important to keep in mind that this can lead to redundancy.

ii. **Data Analysis Expressions (DAX)**: formula language specifically designed for Power Pivot based on Excel's language. DAX functions are powerful tools for summarising, filtering, and manipulating data (Microsoft, n.d.-a). They can be used to calculate total sales, identify top-performing products, or generate custom metrics tailored to specific business needs. Below is an example of a DAX formula for a Total Sales measure using the SUM function (Ferrari & Russo, 2016).

$$Total\ Sales = SUM(Sales[SalesAmount])$$

2.2.2 Dashboard Design and Development Process

Creating a compelling Power BI dashboard involves an integrated approach that seamlessly blends design considerations with the development process. A successful dashboard not only presents data but does so in a way that is visually appealing and user-friendly and aligns with the organisation's objectives (AscenWork, 2023).

It's essential to start by clearly defining the dashboard's purpose, outlining specific business objectives, identifying key user pain points, understanding the audience, considering preferences for visualization types (Subotin, 2017), and choosing the appropriate dashboard type (Sindhu et al., 2020).

Dashboards can have different purposes. According to Bach et al. (2023) and Sindhu et al. (2020), there are three main types of dashboards based on these purposes: strategic (to support executive decision-making), tactical (to summarise department data), and operational (to provide information to front-line workers). Sindhu et al. (2020) categorise the different types of dashboards based on various criteria, such as purpose, business problems, users, goals, and visual features. The information is summarised in Table III.

TABLE III: DASHBOARD TYPES (Sindhu et al., 2020, pp. 2-5)

Type / ***Criteria***	***Strategic Dashboards***	***Operational Dashboards***	***Tactical Dashboards***
Purpose	Improves business procedures; Tracks KPIs; Monitors organisation's performance; Enhances planning for future scope.	Increases consistency of analysis; Monitors activities, individuals and groups; Improves speed and transparency of information.	Self-monitors management performance; Analyses trends.
Business Problem	Organisational KPIs and achievement of strategic goals.	Gaps in performance, time-sensitive data, and data awareness.	Deeper insights or trends, visibility into key processes.
Dashboard Users	Directors, executives, business analysts.	Managers and users involved in daily business operations.	Executives, data analysts, business analysts.
Dashboard Goals	Strategic goals, achieving KPI targets.	Tracking against goals, employee awareness, addressing issues.	Analytics goals, gaining visibility into key processes.
Visual Features	- Graphs: bar chart, pie chart, gauge chart; - Time horizon.	- Graphs: line chart, bar chart, pie chart, gauge chart; - Tables; - Filters.	- Graphs: bar chart, gauge chart; - Drill down feature; - Scenario analysis.

The design process relies on a holistic approach, which encompasses considering the user's requirements, best practices of data visualisation, and interactive elements that help to enhance user engagement. Many case studies and scholars, as Bach et al. (2023) and Pokhrel Jeevanand Awasthi (2021) mention, agree that dashboards should follow certain principles to be effective, Figure 3. These principles include:

i. Avoid overwhelming users and avoiding visual clutter;
ii. Use appropriate colours, chart types, and effective data labelling to help create an intuitive and easy-to-understand dashboard
iii. Carefully select a set of KPIs and avoid poor visual design;
iv. Align with existing workflows and avoid showing too much data;
v. Have both functional and visual features, meaning what the dashboard can do and how information is presented;
vi. Provide consistency and interaction attributes;
vii. Organise charts symmetrically, group them by attribute, and order them according to time.

Figure 3: Effective Dashboard Design Principles (Pokhrel Jeevanand Awasthi, 2021, p. 99).

In this context, it's also essential to consider the contributions of a pioneer in data visualisation, Edward Tufte. Tufte introduced several key principles that have become fundamental to effective data visualisation, like "data-ink ratio", "chart junk", "data density" and small multiples (Chen, 2017). The concept of "data-ink ratio" emphasises that the substantial portion of ink in a graphic should convey data information, adapting as data changes (Tufte, 2007), meaning that every element should serve a purpose in presenting data clearly and efficiently, and all non-essential elements (i.e., decorative features), should be minimised or eliminated. Tufte (2007) also highlighted the importance of avoiding "chart junk," which includes unnecessary visual elements that distract from the data, such as heavy grid lines, excessive text, and ornamental shading.

With “data density” and small multiples, Tufte (2007) focuses on maximising the amount of data shown through techniques like small multiples and sparklines, ensuring that graphics are clear, detailed, and efficient in conveying information.

The effective design of dashboards also relies on understanding and implementing design patterns. Bach et al. (2023) highlight eight groups of design patterns – data information, meta information, visual representation, page layout, screen space, structure, interaction and colour – that can be categorised into content patterns and composition patterns, as seen in Figure 4.

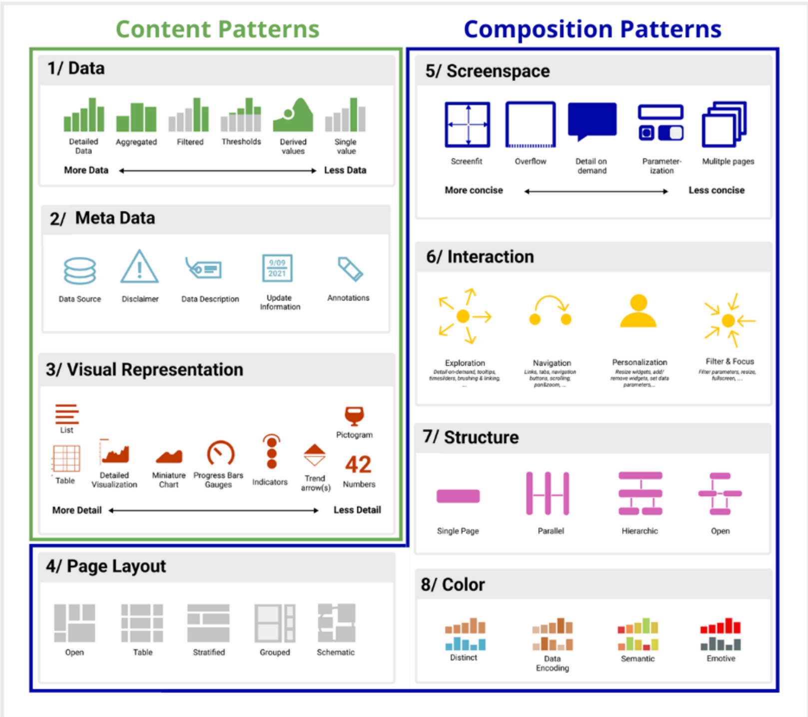

Figure 4: Design Patterns (Bach et al., 2023, p. 1).

A dashboard’s content comprises distinct dashboard elements, which are essential for data and display. There are three types of content-related design patterns: data information (identifying the types of information presented and the level of abstraction used), meta information (capturing additional information used to provide context and explanation, such as the current date), and data visualisation (standard solutions for presenting data information in dashboards) (Bach et al., 2023).

The way a dashboard is composed influences how its many content components are combined and presented. Dashboards show several information items, and the structure and style of a page are critical design factors. There are five aspects of composition: page layout (how widgets are placed up and arranged on a dashboard), screenspace (methods used to fit items on a single screen), structure (relation between multiple pages), interaction (interaction techniques seen in dashboards, which include data as the interface, user interface components, etc.), and colour (Bach et al., 2023).

Regarding the dashboard development process in Power BI, it involves several key considerations. Firstly, connecting to data sources. As mentioned before, Power BI allows users to connect to various sources such as Excel, SQL Server, or SharePoint (Microsoft, 2023d). It's essential to establish a solid connection and, if necessary, use the Power Query Editor to clean, transform, and shape the data (Microsoft, 2023b).

Once the data is loaded, the focus shifts to visualisation creation. Users can drag and drop fields onto the report canvas select visualisation types that suit their data, and customise the appearance and formatting to enhance clarity (Microsoft, 2023a).

Power BI offers an immense variety of visualisation options to cater to diverse data analysis needs, as summarised in Table IV. Users can choose from a range of charts, each designed to effectively convey specific types of information. Additionally, calculated columns and measures using Data Analysis Expressions (DAX) can be employed to derive meaningful insights from the dataset (Ferrari & Russo, 2016).

TABLE IV: SUMMARY OF POWER BI'S CHARTS (Ferrari & Russo, 2016, pp. 268 - 272)

Chart	***Description***
Line Chart	Display the trend of measures over time with a y-axis range that usually does not include zero.
Column Chart	Similar to a bar chart but vertically oriented.
Stacked Bar (Column) Chart	Compare different values of the same measure horizontally (vertically) or display different measures that are part of the same whole.
Scatter Chart	Show possible correlations between two measures.
Table	Display data in a textual form as a simple table with each attribute and measure in a single column.

Matrix	Extends the table, allowing grouping of measures by rows and columns.
Card	Displays a single numerical value of a measure textually on a coloured card.
Slicer	Used to filter one or more charts by selecting values of an attribute.

The ultimate goal of data visualisation is to provide insights and understanding that might not be immediately apparent from raw data alone (Kumar & Belwal, 2017). By employing effective visual elements, dashboards can help make sense of complex data and enable better decision-making.

As mentioned, visualisation design requires a careful selection of charts and graphs that align with the user's preferences. Therefore, choosing the appropriate chart type that best suits the presented data is crucial. Various chart types, such as bar charts, line graphs, and scatter plots, offer versatile options for visualising different data types. Gulbis (2016) shared a chart selection diagram created by Dr. Adrew Abela, Figure 5, to help choose the best chart when presenting data.

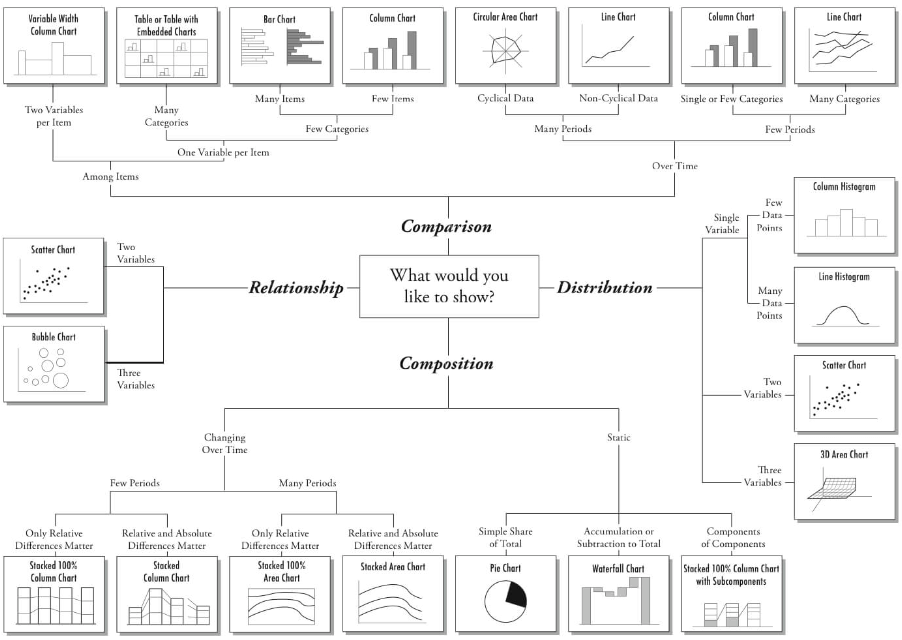

Figure 5: Chart Choice (Gulbis, 2016).

After creating an effective dashboard by following the design principles and patterns, the next step is to save and share it. Once satisfied with the dashboard, it can be saved in

Power BI Desktop and published to the Power BI Service, a platform for sharing and collaboration (Ferrari & Russo, 2016). Users can share dashboards with others, configure sharing options, collaborate in real-time, and schedule automatic data refreshes to ensure the dashboard always displays the latest data (Microsoft, 2024a).

3. METHODOLOGY

3.1 Internship Plan

The internship's objective was to develop comprehensive skills in data analytics. It focused on data acquisition, processing, and integration techniques within a database framework and improving the ability to construct dynamic and valuable dashboards.

Initially, the intern reproduced numerous reports the team was using for a customer, which served as a benchmarking exercise and an opportunity to become familiar with the team's methodology and procedures.

As the internship progressed, the intern became more involved in team activities. This included active participation in client calls and daily discussions with team members to gain insights into their respective roles and responsibilities. As part of this integration process, the intern was assigned increasing levels of responsibility, ultimately leading to the ownership of key projects.

The first significant project undertaken was the development of an Opportunity Report and Daily OPS Report for Client A, which involved comprehensive data analysis and dashboard creation tailored to the client's needs.

With the emergence of a new client, Client B, the intern was once again entrusted with a leadership role, spearheading the development and execution of the project.

3.2 ETL Process and Implementation

During the internship at LGG Advisors, we used a custom ETL process built combining different tools and technologies to create a solution that met the specific needs for data extraction, transformation, and storage. While a commercial ETL tool like Informatica or Talend might offer a more streamlined and integrated approach, a custom setup provides flexibility and control over the specific tools and steps involved in the ETL process. As described in Figure 6, this process used Power Automate to extract the files from the clients' servers, Python to transform the data – by performing data cleaning, manipulation, and preparation – and to load the transformed data into a structured MySQL database for further analysis using Power BI.

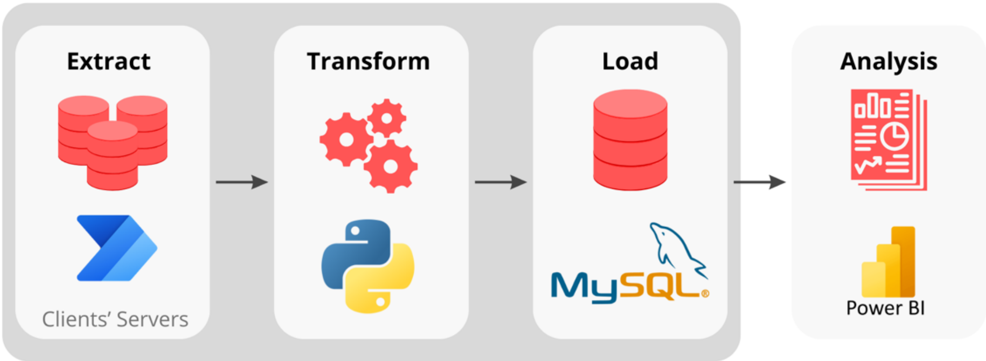

Figure 6: LGG Advisors' ETL Process. Source: The Author.

Here's a brief explanation of the different technologies employed in our custom ETL process:

i. **Power Automate**: cloud-based service by Microsoft that empowers businesses to automate repetitive tasks and streamline workflows encompassing a wide array of apps and services, simplifying task automation. Its drag-and-drop interface enables users to create flows without coding expertise. The help of pre-built connectors for popular applications, custom connectors for specialised tools, and trigger-based conditions enhance its versatility. (Microsoft, n.d.-b)

ii. **Pyhton**: as Python Software Foundation (2023) described, Python is an «interpreted, interactive, object-oriented programming language» recognised for its readability and simplicity. It supports a variety of programming styles besides object-oriented programming, including «procedural and functional programming» (Python Software Foundation, 2023). Python's versatility is evident in its widespread use across various domains, including web development, data science, artificial intelligence, and scientific computing.

iii. **MySQL**: open-source «relational database management system» (Oracle, n.d.) widely used for efficiently organising and managing large datasets. Following the principles of a relational database, MySQL uses tables with rows and columns to store and retrieve data, and it supports SQL for data manipulation. Known for its scalability, cross-platform compatibility, security features, and active community support, MySQL is a critical component in various web applications. Its versatility, performance optimisation, and integration with popular programming languages make it a popular choice for developers in creating dynamic and data-driven websites.

Our data extraction process revolved around using Power Automate to streamline our clients' daily and weekly data retrieval tasks. Power Automate securely logged into the different servers, navigated through the interfaces to select specific dates and configurations and then downloaded the required data in CSV or Excel formats. The extracted files were then stored in designated folders for subsequent processing by a Python-based pipeline meticulously developed by our team's Python specialists. This integration ensured seamless data transformation and loading, providing a streamlined and automated end-to-end solution. Subsequently, Power BI was employed to analyse the data and construct informative dashboards, completing the data lifecycle.

Toward the end of the internship, we started exploring the option of transferring our Power Automate flows to an AWS (Amazon Web Services) environment using Amazon EC2 (Elastic Compute Cloud) due to operational inefficiencies.

Amazon Web Services (AWS), as Wittig & Wittig (2023) mention, is a web services platform that provides computing, storage, and networking solutions at various abstraction levels. Abstraction levels in AWS represent different layers of complexity at which users can interact with the platform. They can choose lower levels for detailed control over specific components or higher levels for simplified management of abstracted services, offering flexibility based on their preferences and expertise (Re Ferre, 2018). Linking volumes to a virtual machine, for example, is considered a low level of abstraction, while storing and receiving data via a REST API is regarded as a high level of abstraction. It also offers services for hosting websites, running corporate applications, and mining massive volumes of data. Web services are available via the internet using standard web protocols and can be accessed by computers or humans via a UI. The most famous services supplied by AWS are EC2, which provides virtual computers, and S3, which offers storage space (Wittig & Wittig, 2023).

Amazon EC2 is a web service that allows users to rent virtual servers, known as instances, in the cloud (AWS, n.d.-a), which can easily be scaled up or down based on the user's computing needs. With EC2, users can run applications, host websites, and perform various computing tasks without investing in physical hardware. It provides flexibility, scalability, and the ability to configure instances with different specifications to meet specific requirements.

The need to manually trigger the daily and weekly flows during working hours and to upgrade Power Automate's subscription for concurrent execution proved to be time-consuming and pricey. In light of these challenges, migrating to an AWS environment, specifically leveraging Amazon EC2, gained traction as it promised a more scalable and cost-effective solution.

However, this approach proved to be unfeasible because the instance we initially selected, t3a.small, lacked the necessary power to handle the Power Automate workflows effectively. While T3 instances are optimized for applications with moderate CPU usage and occasional spikes (AWS, n.d.-b), they fell short of our requirements. Consequently, migrating Power Automate to this instance was impractical, and upgrading to a more capable instance would have incurred additional costs, undermining the cost-effectiveness we sought.

3.3 Transformations inside Power BI and Report Schemas

After processing the data with the built ETL, Power BI was used to analyse it. We started by creating a Power BI dataset that was connected to the MySQL database and this is where all measures and transformations were done. The necessary reports were then created based on this dataset using the Power BI datasets connection.

Sometimes, it was necessary to perform additional transformations within Power BI, such as renaming columns and tables, changing data types, and applying filters to remove duplicates. These steps ensured that any updates to the database did not cause data duplication issues in the dashboards.

The report schemas were all very similar. Each report included a stores view table containing information about all stores (store number, store name, market, state, director, etc.), a calendar table (created in Power BI as a new table using DAX), and various other tables (sales, labor, costs, etc.) connected to both the stores and calendar tables using one-to-many relationships.

4. WORK PERFORMED DURING THE INTERNSHIP

During the internship at LGG Advisors, I had the opportunity to develop a few dashboards for the company's clients. This development process began with a consultation with clients to understand their specific needs and expectations. This involved reviewing any existing reports or dashboards they had created, discussing their desired outcomes, and identifying the relevant data sources. The level of client clarity varied, with some clearly articulating their requirements while others required more guidance and exploration. To exemplify the work developed at LGG Advisors, I'll illustrate and briefly explain the challenges and the process involved in creating three specific reports.

One of the first reports I developed was an Opportunity Report for Client A, Figure 7.

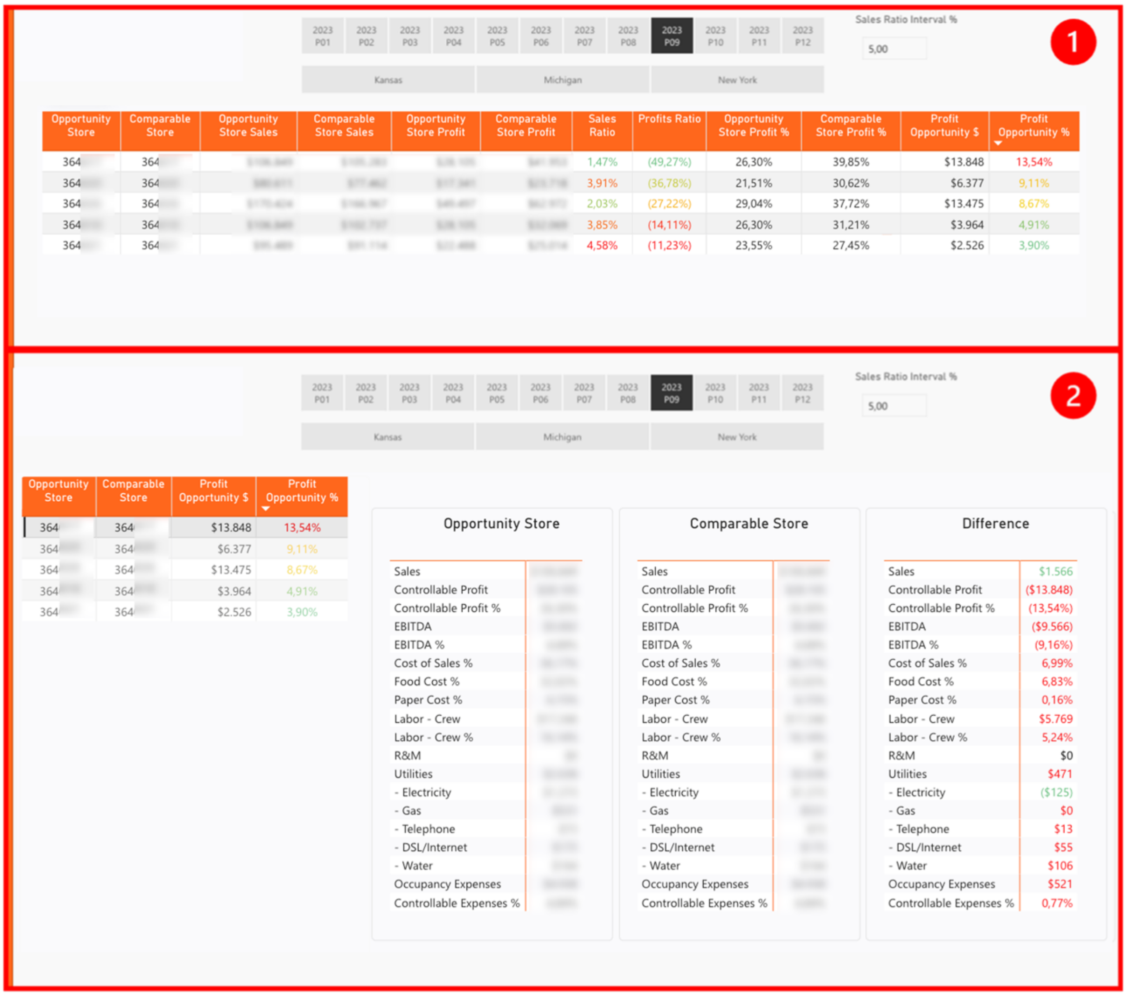

Figure 7: Opportunity Report

The Opportunity Report presents a comprehensive analysis of the organisation's performance metrics, offering a strategic lens to compare stores based on key performance indicators such as sales, profits, and cost of sales. Beyond the conventional examination of financial metrics, this report explored the complexities underlying variations among stores, particularly those with similar sales figures but divergent profit margins. By analysing these nuances, it aimed to discern the contributing factors, providing a unique opportunity to understand the intricacies that drive disparate outcomes within the organisation network.

The first big challenge we faced while creating this report revolved around comparing different stores based on their sales and profits since we just had one table with all this information in our data model. To address this issue effectively, we decided to split our original table into two separate tables. The first table, "Hierarchy", would hold all store-related information – such as store number, store name, store director, and state – while the second table, "KPIs", would capture sales, profits, utilities, labor costs, and other relevant data. These two tables were then connected using a one-to-many relationship based on store number and enabling both tables to filter each other, as seen in Figure 8.

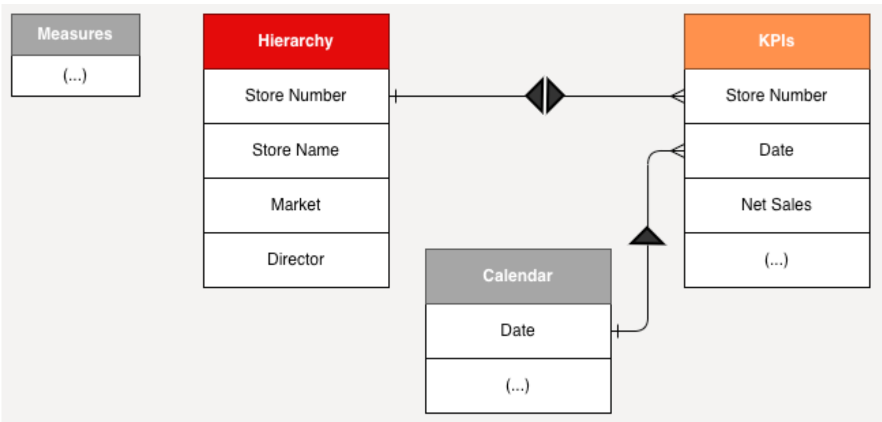

Figure 8: Opportunity Report ER Diagram. Source: The Author

In spite of our efforts, the initial split of the tables didn't fully resolve the issue. When attempting to display the store number from both the hierarchy table and the KPIs table in a visual table only presented stores that matched in both tables. To address this challenge, we implemented a solution by creating metrics for both Opportunity and Comparable stores. Let me illustrate this with a specific example using the sales metric:

$$OpportunityStoreSales = CALCULATE(SUM('KPIs'[TOTAL\ NET\ SALES]), 'KPIs'[Store\ number])$$

$$ComparableStoreSales = CALCULATE(SUM('KPIs'[TOTAL\ NET\ SALES]), 'Hierarchy'[Store])$$

Additionally, our aim was to identify a store as an "opportunity store" in comparison to a similar store, characterised by higher sales but lower profitability. To achieve this, we created the "sales and profit filter" metric, using it to refine the data presented in the table visual.

$$\begin{aligned}
&SalesProfitFilter = \\
&VAR\ Sales = [ComparableStoreSales] - [OpportunityStoreSales] \\
&VAR\ Profit = [ComparableStoreControllableProfit] - [OpportunityStoreControllableProfit] \\
&RETURN \\
&IF(Sales < 0\ \&\&\ Profit > 0, 1, 0)
\end{aligned}$$

Furthermore, to provide flexibility in selecting stores based on specific criteria, we introduced an additional filter called "sales ratio filter". This filter is anchored in the "sales ratio" metric, which computes the ratio between the comparable store sales and the opportunity store sales. The "sales ratio filter" with "sales ratio range", a numeric range parameter, enables the user to define a range within which the sales ratios should fall, allowing for a targeted display of stores meeting certain conditions.

$$\begin{aligned}
&SalesRatioFilter = \\
&VAR\ Result = 1 - ([ComparableStoreSales] / [OpportunityStoreSales]) \\
&VAR\ MinValue = ('Sales\ Ratio\ Range'[Sales\ Ratio\ Range\ Value] * -1)/100 \\
&VAR\ MaxValue = 'Sales\ Ratio\ Range'[Sales\ Ratio\ Range\ Value]/100 \\
&RETURN \\
&IF(Result >= MinValue\ \&\&\ Result <= MaxValue, 1, 0)
\end{aligned}$$

To enhance the interactive and user-friendly nature of the report, slicers were added to enable filtering of the data based on year, period, and store state, which allows users to easily delve into specific data subsets and explore it. In addition, conditional formatting was employed to highlight key metrics – sales ratio, profits ratio, opportunity %, and difference –, making it easier for users to identify significant values within the data. The combination of slicers and conditional formatting creates a comprehensive and informative report that caters to the needs of various users and facilitates data analysis.

For Client A, we also created a Daily OPS report, Figure 9, a more straightforward report that summarises the organisation's sales, transactions, and employee performance.

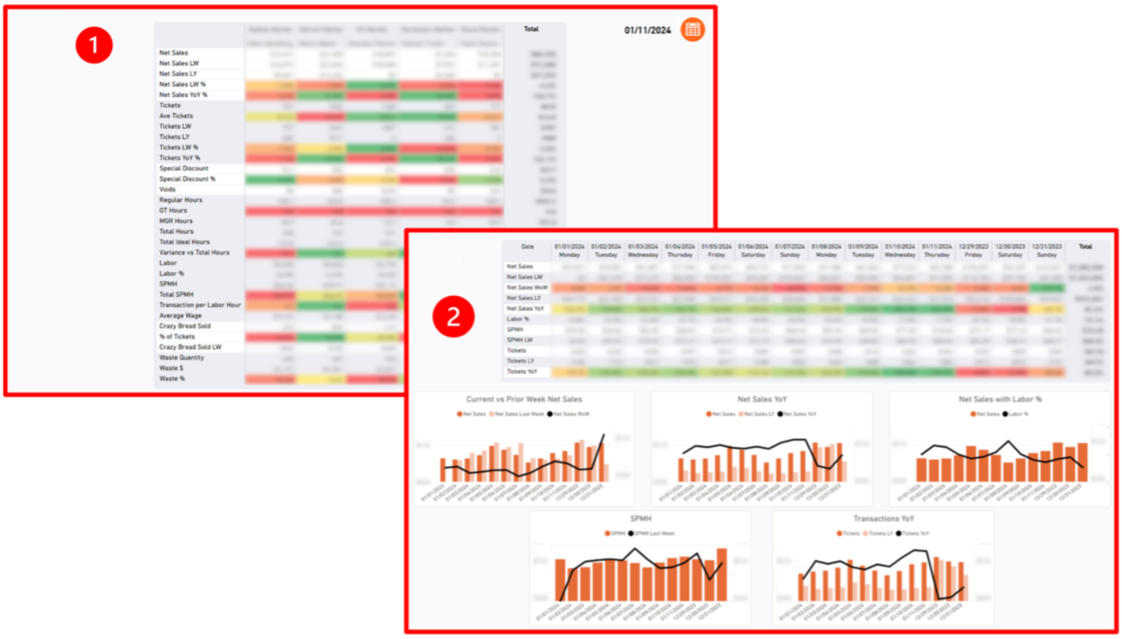

Figure 9: Daily OPS

Our approach involved leveraging an existing Excel report developed by the organisation's team, simplifying the creation process. Despite the foundation provided by the pre-existing report, specific enhancements were required to incorporate crucial historical context, like measures for comparisons for "LW" (last week) and "LY" (last year), as well as percentage values.

The report has seven pages, with the initial two pages, as shown in Figure 9, offering a comprehensive overview of the company's values and incorporating trend graphs to analyse specific metrics the client chose. The remaining five pages are explicitly tailored for store directors, facilitating a side-by-side comparison of values for their respective stores. This strategic segmentation ensures that the report caters to the diverse needs of both the organisation as a whole and individual store director, providing valuable insights for informed decision-making.

The most intricate report I developed during my internship was a combination of three reports – Profit and Loss (P&L), General Ledger (G&L) and Opportunity Report – into a single, user-friendly interface, empowering Client B to manage and analyse their financial performance across multiple dimensions effectively. A unique challenge drove the development of this report: the client's initial uncertainty about their desired result. This presented a design challenge, requiring the report's adaptation as the client's understanding of their data needs evolved.

The initial hurdle we faced was handling the data provided by the client – a sprawling and disorganised Excel file with over 80,000 (eighty thousand) rows, making the comprehension and analysis process significantly more complex. Despite this, we successfully delivered a report that seamlessly aligned with the client's envisioned outcome.

The integrated report comprises three key pages, illustrated in Appendix A – the Home page, Single Store page, and Opportunity Table page. The Home page covers information concerning all stores, allowing users to navigate among OPS Coaches, Area Coaches, and individual stores. Meanwhile, the Single Store page facilitates in-depth analysis of a specific store, and the Opportunity Table page, similar to Client A's Opportunity Report mentioned earlier, provides a comprehensive table for comparing, in this case, net revenue opportunity between stores.

Both the Home and Single Store pages are organised into three primary sections. The first section features the P&L, Figure 10, presenting values for each P&L line in both dollar amounts and percentages. These calculations were based on over 80 measures, as they resulted from intricate calculations of G&L lines. The second section houses 21 trend graphs for key P&L categories chosen by the client over the last 12 months. Each graph includes buttons for users to seamlessly change between dollar and percentage trends.

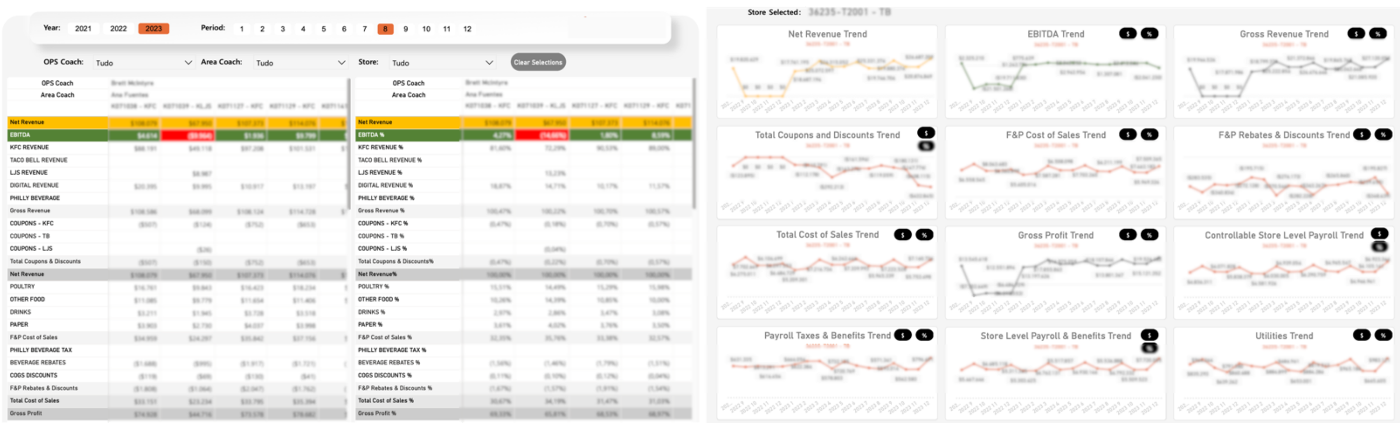

Figure 10: P&L Sections

The final section focuses on the G&L, Figure 11, incorporating a matrix of all G&L lines with a drill-down feature, enabling users to delve deeper into line details such as major categories, GL account descriptions, etc. When a specific value is selected, the line graph above the G&L matrix displays the trend for that item over the past 12 months.

Figure 11: G&L Section

The final page, the Opportunity Table, Figure 12, represents an enhanced version of the one found in Client A's Opportunity Report. In Client B's Opportunity Table dashboard, users can selectively choose the stores for comparison and adjust the range of net revenue opportunities they wish to explore. The tables at the bottom of the page allow users to compare dollar and percentage values for the main P&L lines, providing a comprehensive tool for strategic decision-making.

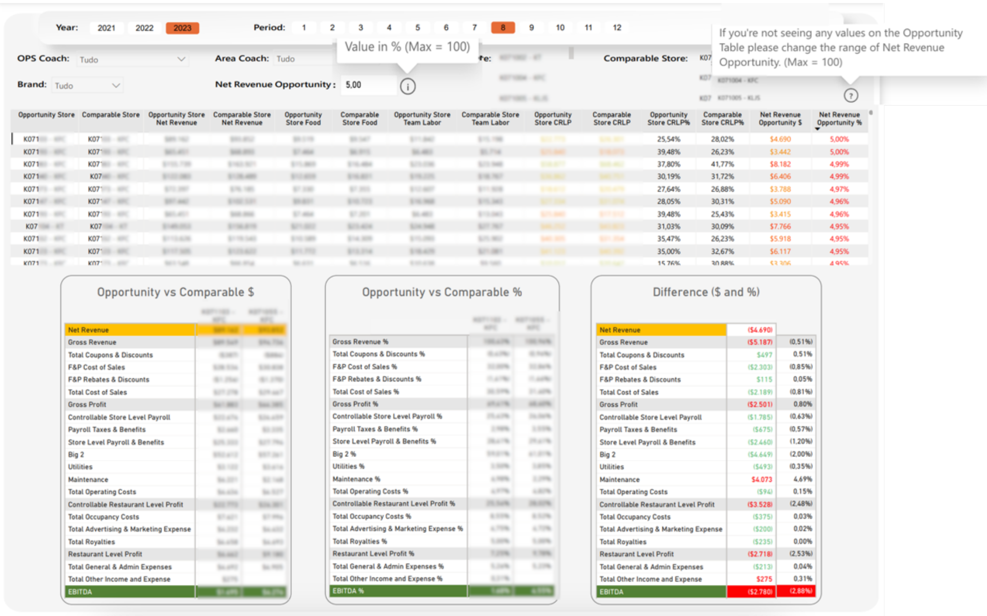

Figure 12: Opportunity Page

Finally, all pages feature navigation buttons – Home, Single Store, and Opportunity Table – and information buttons facilitate easy movement between pages and guide users to navigate and use the dashboards effectively. This navigation enhancement was achieved through a combination of pre-made Power BI buttons and the drill-through feature in Power BI.

After completing the reports and receiving positive feedback from clients, the final phase involved creating detailed documentation for each report. This documentation serves a crucial role in equipping users with the essential knowledge to interpret and leverage the information effectively in the decision-making process.

5. CONCLUSION

5.1 Key Findings

ETL and Power BI form an indispensable partnership in data-driven decision-making, seamlessly harmonising data acquisition, transformation, and visualisation to unlock hidden insights within our ever-expanding data landscape (Makara, 2023).

ETL, whether executed through Enterprise Tools, Open-Source Tools, Custom Tools, or Cloud-Based Services, extracts raw data from disparate sources, transforming it into a refined format optimized for analysis (Gaikwad, 2023). This thorough preparation ensures data accuracy, consistency, and relevance, laying the foundation for effective decision-making (TestingXperts, 2023).

In the realm of data visualisation and analysis, Power BI becomes the channel for translating processed data into meaningful insights through visually compelling dashboards and reports. This platform is built on theoretical concepts of data visualisation, including visual encoding and storytelling, to empower users to effortlessly interpret complex data and foster a culture of data-driven decision-making (Microsoft Power BI, 2018).

Through collaboration with LGG Advisors, the internship's objectives were met and exceeded, providing valuable insights into the practical aspects of data management and visualisation in a professional setting. The skills acquired in ETL processes and Power BI development will undoubtedly serve as a solid foundation for future accomplishments in data analytics.

In essence, this four-month internship has been instrumental in bridging the gap between theory and practice, equipping me with a robust skill set and a deeper understanding of the complexities involved in handling and presenting data effectively.

5.2 Contributions

This dissertation contributes to the existing literature on business intelligence by providing a comprehensive and in-depth understanding of the role of ETL processes and Power BI in transforming raw data into actionable insights for driving business growth and success.

In addition to the literature review section – which delves into the definition of ETL processes, explores various ETL tool types, and emphasises the importance of data integrity –, this report also provides insights into creating impactful dashboards that effectively communicate insights in Power BI.

The theoretical framework and application during the internship provide a unique and hands-on perspective that adds to the existing literature on BI, ETL processes and dashboard building in Power BI. The examples given by the developed ETL process and Power BI dashboards provide future researchers with concrete examples of how ETL processes and Power BI can transform raw data into actionable insights and how essential data visualisation and communication are in effective decision-making.

Furthermore, the concrete examples of Daily OPS report, Opportunity report and the combination of G&L and P&L reports act as templates for understanding the practical implementation of BI tools in real-world scenarios. They serve as a roadmap for organisations seeking to streamline their data management processes, enhance operational efficiency, and make data-driven decisions.

Overall, the ETL process and reports developed in this internship serve as a resource for future researchers looking to develop impactful BI reports using ETL processes and Power BI.

5.3 Future Work

This dissertation comprehensively analyses the partnership between ETL processes and Power BI, highlighting their pivotal role in data-driven decision-making. However, one aspect that warrants further exploration is the potential migration of Power Automate flows to an AWS environment using Amazon EC2.

Towards the end of the internship, preliminary discussions began regarding the feasibility of transferring the customised ETL process, which currently relies on Power Automate, to an AWS environment, specifically leveraging Amazon EC2. The existing ETL process, while effective, required manual triggering due to subscription limitations in Power Automate. Given the operational inefficiencies associated with manual triggering and the potential financial implications of subscription upgrades to automate the process, the exploration of migrating to an AWS environment gained traction.

However, several unanswered questions and considerations remain, necessitating further research and exploration in this domain.

Therefore, future research should focus on several key areas to assess the feasibility and benefits of migrating the customised ETL process from Power Automate to an AWS environment leveraging Amazon EC2. This includes evaluating the performance implications of running the process on EC2 instances and considering scalability options and performance benchmarks to handle varying data volumes and processing requirements efficiently. Additionally, it is important to conduct a comprehensive cost analysis to understand the financial implications, including upfront investments, ongoing operational costs, and potential cost savings compared to upgrading the Power Automate subscription. Investigating potential integration challenges between the existing ETL process and AWS services is crucial to ensure seamless migration and uninterrupted data flow. Finally, it is vital to address security and compliance considerations during migration. This includes implementing robust security measures and ensuring compliance with regulatory requirements to safeguard data integrity.

REFERENCES

Ali, S. M., Gupta, N., Nayak, G. K., & Lenka, R. K. (2016). Big data visualization: Tools and challenges. *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*, 656–660. https://doi.org/10.1109/IC3I.2016.7918044

Alooba. (2023, March 16). *The Importance of Data Quality in Analysis*. Https://Www.Alooba.Com/Articles/Importance-of-Data-Quality-in-Analysis/.

Amazon. (2022, October 18). *What is ETL (Extract Transform Load)?* Https://Aws.Amazon.Com/What-Is/Etl/?Nc1=h_ls.

AscenWork. (2023, October 9). *Creating Interactive Dashboards in Power BI: Tips and Best Practices*. Https://Www.Linkedin.Com/Pulse/Creating-Interactive-Dashboards-Power-Bi-Tips-Best-Practices-2f/.

Atlan. (2023, December 13). *What is Data Modeling? 5 Steps to Building a Robust Model*. Https://Atlan.Com/What-Is-Data-Modeling/.

AWS. (n.d.-a). *Amazon EC2*. Https://Aws.Amazon.Com/Ec2/?Nc1=h_ls.

AWS. (n.d.-b). *Amazon EC2 T3 Instances*. Https://Aws.Amazon.Com/Ec2/Instance-Types/T3/.

AWS. (2022, December 12). *AWS Glue*. Https://Aws.Amazon.Com/Glue/?Nc1=h_ls.

Bach, B., Freeman, E., Abdul-Rahman, A., Turkay, C., Khan, S., Fan, Y., & Chen, M. (2023). Dashboard Design Patterns. *IEEE Transactions on Visualization and Computer Graphics*, *29*(1), 342–352. https://doi.org/10.1109/TVCG.2022.3209448

Becker, L. T., & Gould, E. M. (2019). Microsoft Power BI: Extending Excel to Manipulate, Analyze, and Visualize Diverse Data. *Serials Review*, *45*(3), 184–188. https://doi.org/10.1080/00987913.2019.1644891

Chen, H. M. (2017). Information Visualization Principles, Techniques, and Software. In *Information Visualization*.

Dhanda, P., & Sharma, N. (2016). Extract Transform Load Data with ETL Tools . *International Journal of Advanced Research in Computer Science*, *7*(3).

Ferrari, A., & Russo, M. (2016). *Introducing Microsoft Power BI*. Microsoft Press.

Gaikwad, M. (2023, January 23). *Types of ETL Tools: The Complete Guide 101*. Https://Hevodata.Com/Learn/Types-of-Etl-Tools/#2.

Gestisoft. (2023, August 31). *What Are the Benefits of Power BI?* Https://Www.Gestisoft.Com/Blog/What-Are-the-Benefits-of-Power-Bi.

Google Cloud. (2023, September 15). *What is a Data Warehouse?* Https://Cloud.Google.Com/Learn/What-Is-a-Data-Warehouse.

Gulbis, J. (2016, March 1). *Data Visualization – How to Pick the Right Chart Type?* Https://Eazybi.Com/Blog/Data-Visualization-and-Chart-Types.

Hibencarey. (2023, July 20). *Data Quality: The Foundation of Informed Decision-Making*. Https://Medium.Com/@hibencarey/Data-Quality-the-Foundation-of-Informed-Decision-Making-Eafa195dd322.

IBM. (2020, August 26). *What is Data Modeling?* Https://Www.Ibm.Com/Topics/Data-Modeling.

IBM. (2023, November 29). *DataStage*. Https://Dataplatform.Cloud.Ibm.Com/Docs/Content/Dstage/Dsnav/Topics/Datastage.Html?Context=cpdaas.

IMD. (2024, January). *What is business analytics and why is it important?* Https://Www.Imd.Org/Reflections/What-Is-Business-Analytics-and-Why-Is-It-Important/.

J. Sreemathy, Infant Joseph V., S. Nisha, Chaaru Prabha I., & Gokula Priya R.M. (2020). Data Integration in ETL Using TALEND. *2020 6th*

International Conference on Advanced Computing and Communication Systems (ICACCS).

Kimball, R., & Ross, M. (2013). *The Data Warehouse Toolkit* (3rd ed.). John Wiley & Sons, Inc.

Kline, K. (2014, August 5). PowerBI: The Self-Service BI App That Can Change Your Life. *Https://Www.Dbta.Com/Columns/SQL-Server-Drill-Down/PowerBI-The-Self-Service-BI-App-That-Can-Change-Your-Life-98513.Aspx*.

Kumar, S. M., & Belwal, M. (2017). Performance dashboard: Cutting-edge business intelligence and data visualization. *2017 International Conference On Smart Technologies For Smart Nation (SmartTechCon)*, 1201–1207. https://doi.org/10.1109/SmartTechCon.2017.8358558

Makara, R. (2023, July 29). *Unleashing the Power of Data Transformation: A Beginner's Guide*. Https://Reconfigured.Io/Blog/Unleashing-Power-of-Data-Transformation-Beginners-Guide.

Mali, N., & Bojewar, S. (2015). A Survey of ETL Tools. *International Journal of Computer Techniques*, *2*(5), 20–27.

Microsoft. (n.d.-a). *Data Analysis Expressions (DAX) in Power Pivot*. Https://Support.Microsoft.Com/En-Us/Office/Data-Analysis-Expressions-Dax-in-Power-Pivot-Bab3fbe3-2385-485a-980b-5f64d3b0f730.

Microsoft. (n.d.-b). *Power Automate*. Https://Www.Microsoft.Com/En-Us/Power-Platform/Products/Power-Automate.

Microsoft. (n.d.-c). *Power BI - Data Visualization | Microsoft Power Platform*. Https://Www.Microsoft.Com/En-Us/Power-Platform/Products/Power-Bi.

Microsoft. (n.d.-d). *Power Pivot - Overview and Learning*. Https://Support.Microsoft.Com/En-Au/Office/Power-Pivot-Overview-and-Learning-F9001958-7901-4caa-Ad80-028a6d2432ed.

Microsoft. (2023a, October 11). *Work with Report view in Power BI Desktop*. Https://Learn.Microsoft.Com/En-Us/Power-Bi/Create-Reports/Desktop-Report-View.

Microsoft. (2023b, October 24). *Query overview in Power BI Desktop*. Https://Learn.Microsoft.Com/En-Us/Power-Bi/Transform-Model/Desktop-Query-Overview.

Microsoft. (2023c, November 10). *Regional Sales sample for Power BI: Take a tour*. Https://Learn.Microsoft.Com/En-Us/Power-Bi/Create-Reports/Sample-Regional-Sales.

Microsoft. (2023d, November 15). *Data sources in Power BI Desktop*. Https://Learn.Microsoft.Com/En-Us/Power-Bi/Connect-Data/Desktop-Data-Sources.

Microsoft. (2024a, January 1). *Ways to collaborate and share in Power BI*. Https://Learn.Microsoft.Com/En-Us/Power-Bi/Collaborate-Share/Service-How-to-Collaborate-Distribute-Dashboards-Reports.

Microsoft. (2024b, January 24). *What is Power Query?* Https://Learn.Microsoft.Com/En-Us/Power-Query/Power-Query-What-Is-Power-Query.

Microsoft Azure. (2014, October 27). *Azure Data Factory - Data Integration Service*. Https://Azure.Microsoft.Com/En-Us/Products/Data-Factory.

Microsoft Power BI. (2018, July 14). *What is data visualization?* Https://Powerbi.Microsoft.Com/En-Us/Data-Visualization/.

Microsoft Power BI. (2020, November 11). *What Is Business Intelligence?* Https://Powerbi.Microsoft.Com/En-Us/What-Is-Business-Intelligence/.

Oracle. (n.d.). *What is MySQL?* Https://Www.Oracle.Com/Mysql/What-Is-Mysql/.

Pokhrel Jeevan and Awasthi, A. (2021). Effectiveness of Dashboard and Intervention Design. In D. Sahin Muhittin and Ifenthaler (Ed.), *Visualizations and Dashboards for Learning Analytics* (pp. 93–116).

Springer International Publishing. https://doi.org/10.1007/978-3-030-81222-5_5

Przybycień, G. (2023, July 13). *Data integrity vs. data quality: Is there a difference?* Https://Www.Ibm.Com/Blog/Data-Integrity-vs-Data-Quality-Is-There-a-Difference/.

Python Software Foundation. (2023). *General Python FAQ* (3.12.1). https://docs.python.org/3/index.html.

Re Ferre, M. (2018, September 6). *Compute Abstractions on AWS: A Visual Story*. Https://Aws.Amazon.Com/Pt/Blogs/Architecture/Compute-Abstractions-on-Aws-a-Visual-Story/.

Sindhu, K., Jain, K. A., Prof.Sandhya, S., & Nasreen, Dr. A. (2020). *COMPARATIVE STUDY AND REVIEW OF PRINCIPLES FOR DASHBOARD DESIGN*. https://api.semanticscholar.org/CorpusID:235816131

Souibgui, M., Atigui, F., Zammali, S., Cherfi, S., & Yahia, S. Ben. (2019). Data quality in ETL process: A preliminary study. *Procedia Computer Science*, *159*, 676–687. https://doi.org/10.1016/J.PROCS.2019.09.223

Sreemathy, J., Brindha, R., Selva Nagalakshmi, M., Suvekha, N., Karthick Ragul, N., & Praveennandha, M. (2021). Overview of ETL Tools and Talend-Data Integration. *2021 7th International Conference on Advanced Computing and Communication Systems (ICACCS)*, *1*, 1650–1654. https://doi.org/10.1109/ICACCS51430.2021.9441984

Stedman, C. (2022, February). *What is data preparation? An in-depth guide to data prep*. Https://Www.Techtarget.Com/Searchbusinessanalytics/Definition/Data-Preparation.

Subotin, S. (2017, November 9). *Dashboard Design: Considerations and Best Practices*. Https://Www.Toptal.Com/Designers/Data-Visualization/Dashboard-Design-Best-Practices.

TestingXperts. (2023, September 4). *Data Quality Testing in ETL: Importance and Methods to Ensure Accuracy*. Https://Www.Testingxperts.Com/Blog/Data-Quality-Testing-in-Etl/Gb-En.

Theodorou, V., Nakuçi, E., Jovanovic, P., & Abelló, A. (2014, January). *Bijoux: Data Generator for Evaluating ETL Process Quality*. https://doi.org/10.1145/2666158.2666183

Tufte, E. R. (2007). *The Visual Display of Quantitive Information* (2nd ed.).

Wittig, A., & Wittig, M. (2023). *Amazon Web Services in Action, Third Edition: An In-depth Guide to AWS* (3rd ed.). Simon and Schuster.

APPENDIX A: CLIENT B'S REPORT