DATA WAREHOUSING REPORT

GROUP 7 MEMBERS	
NAME	ADMISSION NUMBER
DERRICK LUBANGA	169240
LINDAH KELIDA	169589
MONILADE OLADIRAN	169093
LORRAINE EYINDA	148454
CLAUDINE LINDA WA NCIKO	169375
DAVID NENE	169701
FRANCIS KABUTU	078232
SHARON TONUI	169194
JOSEPH RIDGE	166895
HENRY AYAH	169113
SIMALA LEONARD	121777

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF ACRONYMS

ERD	Entity Relational diagram
OLTP	Online Transactional Processing
OLAP	Online Analytical Processes
BI	Business Intelligence

CHAPTER 1: INTRODUCTION

In the domain of sports analysis, particularly within the context of an intense soccer season within a specific region, the challenges posed by vast amounts of data from diverse sources are undeniable. This project sets out to address these challenges by harnessing the capabilities of data warehousing and Online Analytical Processes (OLAP), offering an efficient and streamlined approach to dissecting the intricacies of soccer.

1.1 Background

Data Warehousing: Data warehousing is the foundational technology at the core of this project. It is a sophisticated process and technology integration that encompasses data collection, storage, and management, particularly in the context of extensive data originating from various sources within an organisation. Beyond data management, data warehousing serves as the cornerstone of our endeavour, creating a centralised and finely-tuned repository designed to facilitate reporting, analysis, and business intelligence (BI) endeavours.

1.2 Project Objectives

The primary objectives of this project are:

I. To design and develop a data warehouse using Pentaho.

II. To leverage data warehousing and OLAP techniques to derive insightful visualisations from a comprehensive soccer data warehouse.

CHAPTER 2: METHODOLOGY

The methodology is organised and will be executed according to specific activities and deliverables.

2.1 The approach

1. Download and install Pentaho software.
2. Create a data warehouse.
3. Search and retrieve data stored in various forms (database, CSV, Excel etc.). Herein, we used data stored in an SQL database.
4. Perform ETL on the data retrieved.
5. Store the data in our data warehouse *(Soccer_DWh data warehouse)*.
6. Perform the OLAP operations using the data warehouse.

2.2 Overview of Soccer Matches

An overview of soccer is that a match consists of two teams, a home team and away team. Home team is the team hosting the match, in that the match takes place at the home team's stadium. Away team is the team which is playing against the home team. A match in our database occurs only in a specific season and league of a country, although in real life a match may occur anytime either inside or outside an ongoing season and league of a country.

The objective of a match is to score goal against the enemy team, and should the scored goal(s) exceed the enemy's goal(s), then the team is declared as winner

A soccer team manager is responsible for preparing his/her team before a match, and preparation may be done by considering previous match statistics against the enemy team.

Seeing the trend of the team's performance in previous season(s) may also help the manager devise a plan for having better trends or maintain previous trends in this season or upcoming season(s).

2.3 Data Warehouse and Database Design

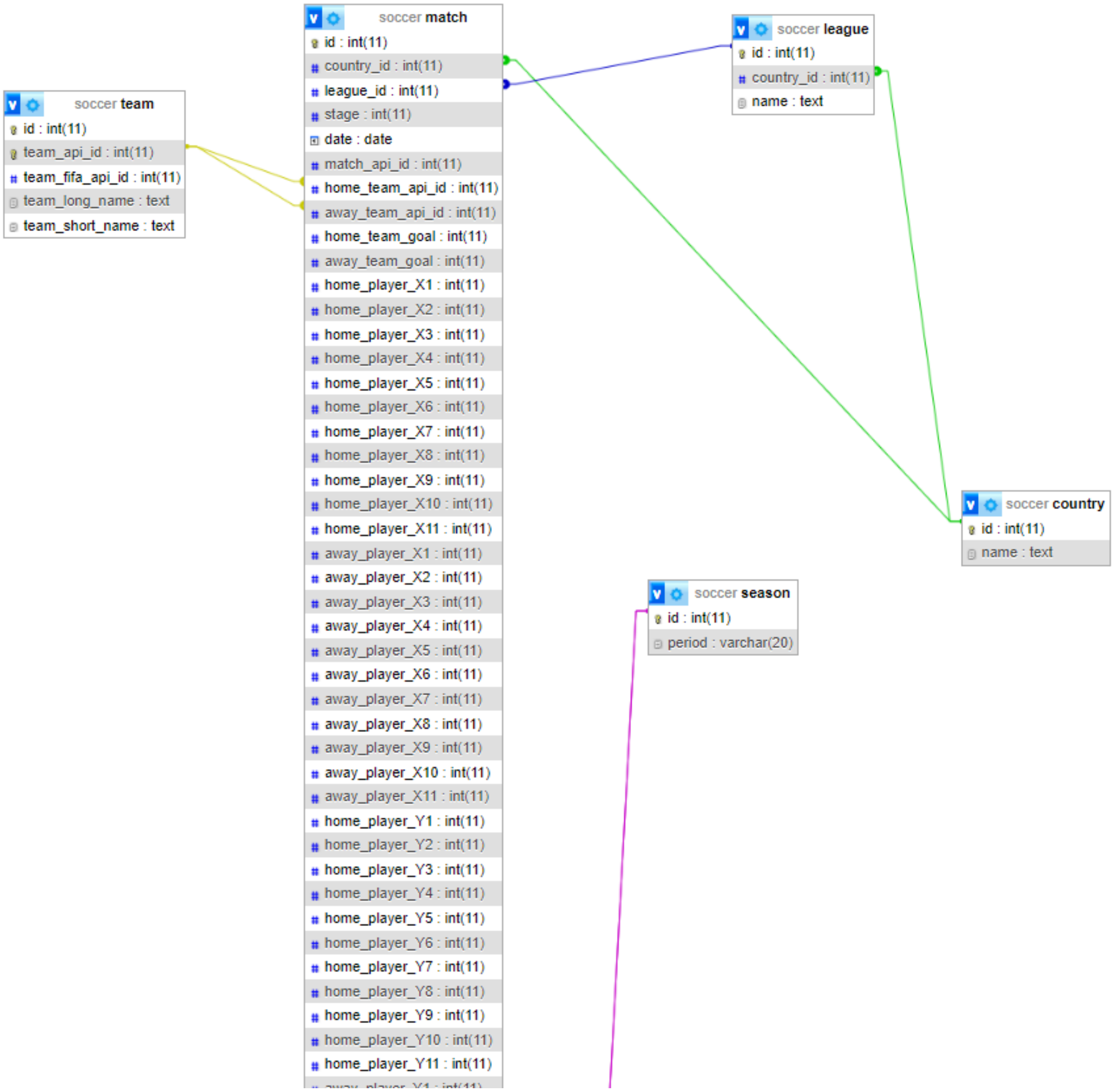


Figure 1: OLTP Database ERD Schema.

To start, as Kimball Lifecycle methodology steps state. We defined our project with the goal of delivering a dashboard of goal statistics for home team. With this project goal. We only use one project for this data warehouse so it can be inferred that our data mart is the data warehouse itself.

The next step is to identify our facts and dimensions. We investigated the OLTP database's data and figured that our dimensions would include month, season, team, and league dimensions. The league dimension will contain country names as well, since each league is located in a single country, but we would like to keep the opportunity for adding another league in the same country if possible. Our main fact on the fact table will be the number of goals scored (goal_for), goals conceded (goal_against), and we added another fact which is derived from the goals scored subtracted with goals conceded (total_goal). The data warehouse was implemented using a star schema.

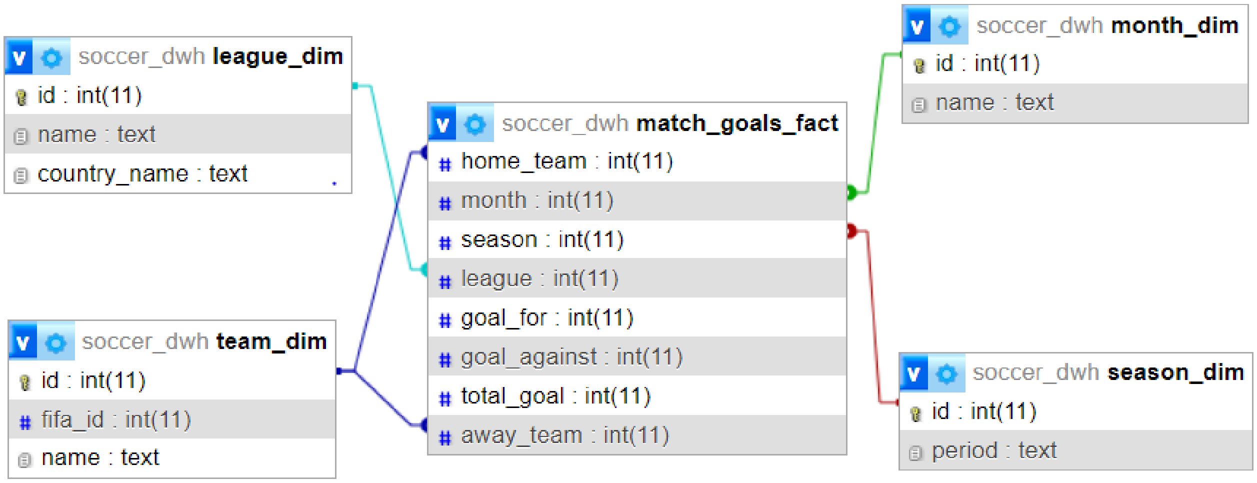


Figure 2: Data Warehouse Star Schema.

The fact table tracks goals scored, conceded, and the goal difference. The dimension tables included month, season, team, and league.

The data warehouse was hosted on a MySQL database.

2.4 ETL Using Pentaho Data Integration

We created five separate transformation files, four for our dimension tables and one for our fact table. By creating separate transformation files, we have a clear view of the transformation process for each dimension and fact. Although this separation requires more effort in running ETL compared to grouping all the ETL flows in one transformation file, this eases our ETL flow development

Pentaho Data Integration (PDI) was used to extract, transform, and load the data into the data warehouse.Separated transformation files were used for dimensions.

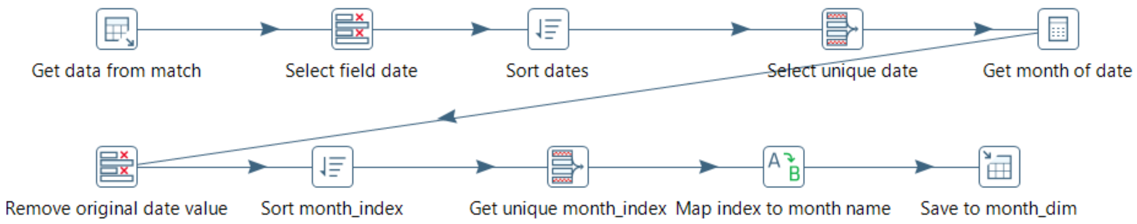


Figure 3: Month Dimension Transformation.

The above transformation shows how we extract unique month names only from the match dates in the match table in our OLTP database. We will not dive deep into how PDI works, and what each step does, but we will explain how we did the transformation. By fetching and selecting only the date data from the match table, we then select unique dates from those matches, as several matches occur on the same date. However, PDI throws a warning to sort the dates first prior to selecting unique values, therefore we fixed our ETL flow by adding sorting every time we would like to select unique values. Then, we get the month of date using PDI's calculator step. Unfortunately, we could not remove the original date in the calculator step, as we only want the month of date itself – which we assigned to a field/column with name month_index. Thus, we need to remove the original date value. Note that the month of date is in numerical value, such that October is represented as 10 as of this step. Then we select the unique month_index, and we map those indexes to month names by our own mapping definitions in the 'Map index to month name' step. Finally, we save the mapped values to our output table, which resides in our data warehouse with table name month_dim, representing month dimension in our data warehouse's fact table.

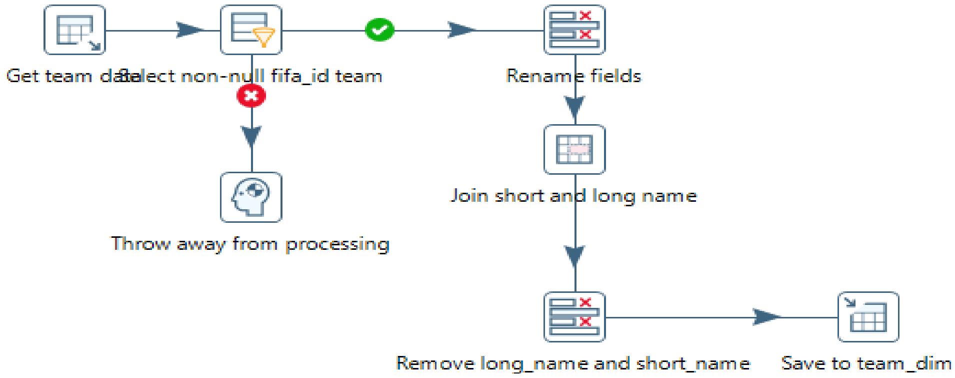

Get team data
Select non-null fifa_id team
Rename fields
Join short and long name
Throw away from processing
Remove long_name and short_name
Save to team_dim

Figure 4: Team Dimension Transformation

Our team dimension transformation is even more simple compared to league dimension transformation as it only selects all data, then filters those with null FIFA ID, which we take as an unregistered FIFA team. We then rename the fields for our next processing, which is transforming the data by joining the team's long name (example: Arsenal) with its short name (example: ARS) into the following format: <short_name> – <long_name>. Lastly, we save the transformed data to our team_dim table in the data warehouse.

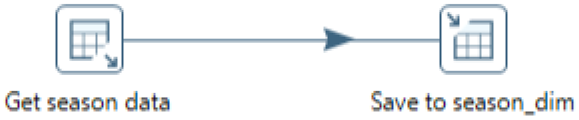


Figure 5: Season Dimensional Transformation

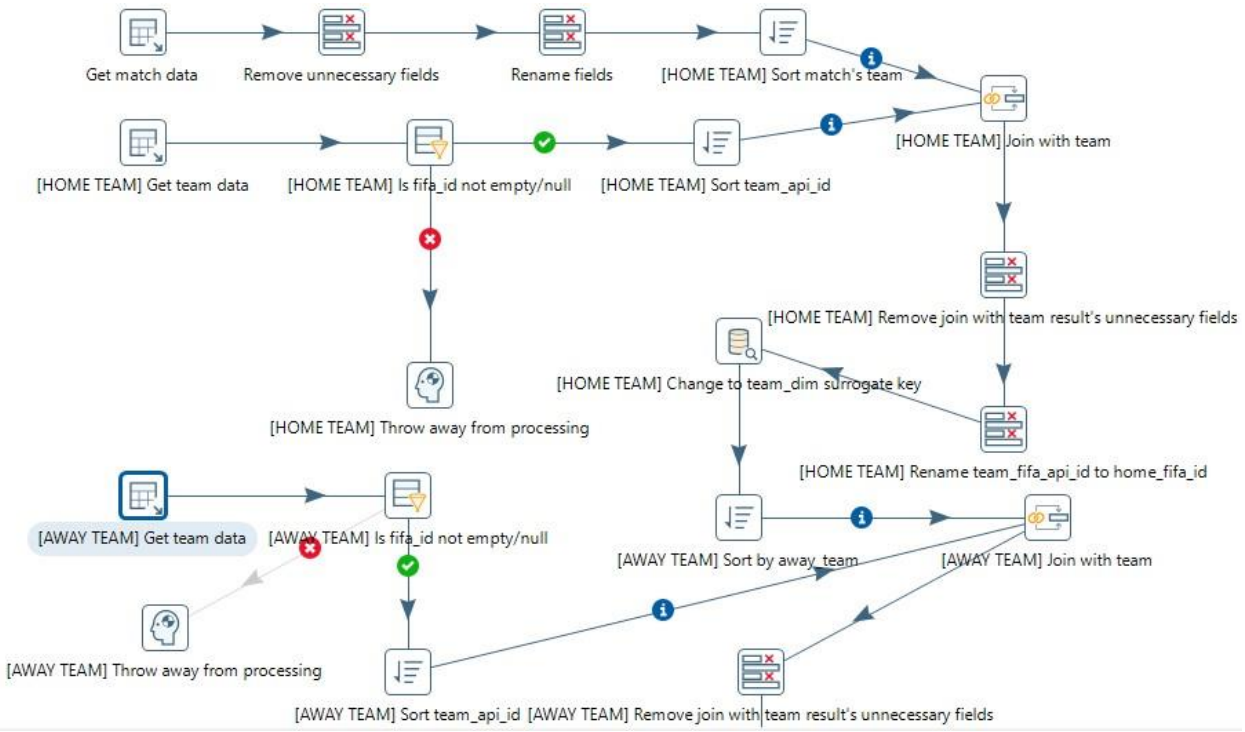

Get match data
Remove unnecessary fields
Rename fields
[HOME TEAM] Sort match's team
[HOME TEAM] Join with team
[HOME TEAM] Get team data
[HOME TEAM] Is fifa_id not empty/null
[HOME TEAM] Sort team_api_id
[HOME TEAM] Remove join with team result's unnecessary fields
[HOME TEAM] Change to team_dim surrogate key
[HOME TEAM] Throw away from processing
[HOME TEAM] Rename team_fifa_api_id to home_fifa_id
[AWAY TEAM] Get team data
[AWAY TEAM] Is fifa_id not empty/null
[AWAY TEAM] Sort by away_team
[AWAY TEAM] Join with team
[AWAY TEAM] Throw away from processing
[AWAY TEAM] Sort team_api_id
[AWAY TEAM] Remove join with team result's unnecessary fields

Figure 6: Fact table transformations

2.5 Testing and Deployment

The data warehouse was tested using a variety of methods, including unit testing, integration testing, and user acceptance testing. The data warehouse was deployed to a production environment once it has been successfully tested.

This methodology provides a high-level overview of the steps involved in creating a data warehouse for soccer match statistics. The specific steps may vary depending on the specific requirements of the project.

CHAPTER 3: ANALYSIS

In the ever-evolving landscape of sports management and analytics, data-driven decision-making has become a fundamental aspect of achieving success in the world of soccer. The ability to harness, process, and analyse vast volumes of data has transformed how soccer clubs, leagues, and organisations strategise, perform, and stay competitive. This analysis chapter delves into the core components of our data warehousing project, which leveraged powerful tools such as Pentaho, Qlik Sense, and OLAP (Online Analytical Processing) techniques to unearth valuable insights from a soccer dataset.

The integration of data warehousing and OLAP technology has not only enhanced our data management capabilities but has also empowered us to uncover critical insights that can drive performance improvements, inform strategic decisions, and offer a competitive edge in the soccer industry. Throughout this chapter, we will explore how we designed and implemented our data warehousing solution, the methodologies used for data extraction, transformation, and loading (ETL), and how Qlik Sense was employed to create interactive and intuitive dashboards for in-depth analysis.

Our analysis begins with the foundations of data warehousing, explaining the rationale behind its implementation, and the challenges we aimed to address. We will then delve into the architecture of our solution, highlighting the integration of Pentaho as our ETL tool and Qlik Sense as our primary analysis and visualisation platform.

Further, this chapter will showcase the various dimensions and measures we have defined within our data warehouse, demonstrating how OLAP cubes were constructed to facilitate multidimensional analysis. We will also explore the key performance indicators (KPIs) and metrics that have been devised to measure team performance, team dynamics, and even tactical strategies.

As we progress, we will provide concrete examples of the insights we have extracted from our data warehouse using Qlik Sense, illustrating the hierarchy of the decision-making processes within the soccer domain. These insights span team performance trends and team dynamics, showcasing the tangible benefits of our data-driven approach.

3.1 Dashboard Creation using Qlik Sense

This project was visualised using the Qlik Sense data visualisation tool. Qlik Sense is a data visualisation and business intelligence platform that is recommended for data warehousing due to its user-friendly nature, powerful querying capabilities, interactive dashboard creation, scalability, and collaborative features.

Qlik Sense allows manual table relationship customization for seamless data synchronisation within our warehouse. It efficiently loads data into the analysis tab, enabling various graph options like scatter plots, bar charts, KPIs, and more.

We've created four sheets: one for an overview and three for data representation based on time, location, and team. Each sheet illustrates the connection between fact and dimension tables, with dimensions like season, month, league, country, and team.

3.2 Key Features of Qlik Sense

Ease of Use: Qlik Sense is designed for non-technical users, making data warehousing accessible to a wider audience.

Powerful Querying: The platform enables complex data queries without requiring coding skills, facilitating comprehensive data exploration and analysis.

Interactive Dashboards: Qlik Sense empowers users to create interactive dashboards for intuitive data visualisation, aiding in the interpretation of data.

Scalability: It efficiently handles large datasets, ensuring its suitability for organisations dealing with substantial data volumes.

Collaboration: Qlik Sense supports collaborative data analysis projects, enabling multiple users to work together, share insights, and reach consensus.

Security: The platform offers robust security features to protect sensitive data, ensuring data integrity and compliance.

Administration: Qlik Sense's user-friendly administration tools simplify management and maintenance tasks.

Support: Users can access a range of support resources, including documentation, forums, and training, enhancing the overall user experience.

3.3 Use Cases

Interactive Dashboards: Qlik Sense can be employed to create interactive dashboards that provide business users with real-time insights, facilitating the monitoring of key performance indicators (KPIs) and trend identification.

Complex Queries: Complex data queries can uncover hidden insights within large datasets, making it an invaluable tool for data analysis beyond the capabilities of traditional reporting tools.

Collaborative Data Analysis: Multiple users can collaborate effectively on data analysis projects using Qlik Sense, enhancing decision-making processes and promoting data-driven practices.

Data Loading: MySQL ODBC connector was installed and used to connect to the data warehouse.

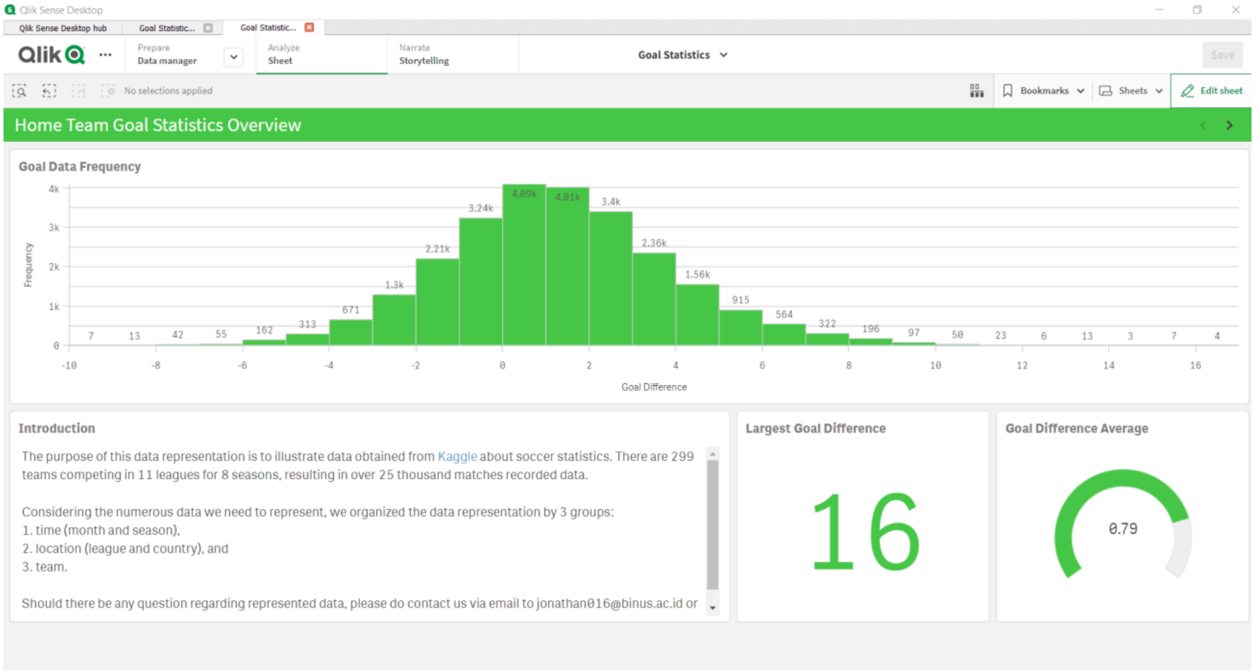
Qlik Sense automatically detects table relationships.

Dashboard Components:

Multiple analysis sheets were created for various aspects of goal statistics.

Customization and filtering capabilities of the graphs are discussed in the *"Visualisations"* section below.

3.4 Visualisations



Introduction

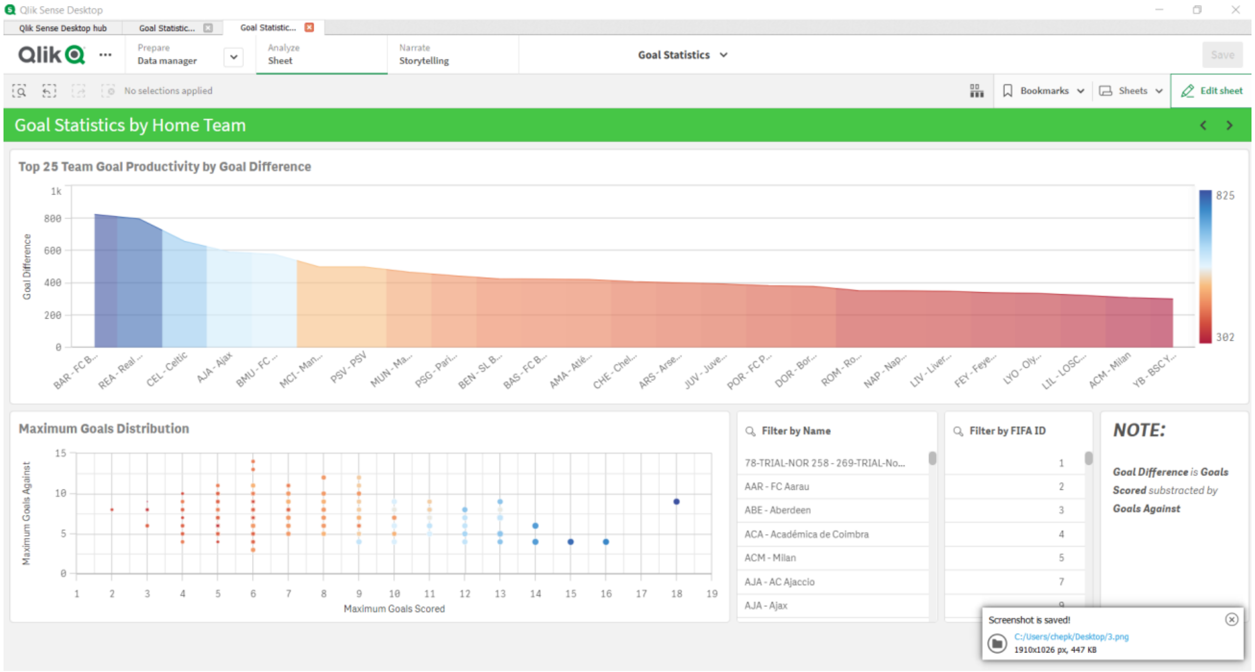
The purpose of this data representation is to illustrate data obtained from Kaggle about soccer statistics. There are 299 teams competing in 11 leagues for 8 seasons, resulting in over 25 thousand matches recorded data.

Considering the numerous data we need to represent, we organized the data representation by 3 groups:
1. time (month and season),
2. location (league and country), and
3. team.

Should there be any question regarding represented data, please do contact us via email to jonathan016@binus.ac.id or

Figure 7: Dashboard of Overview of Data Warehouse

A summary of the data warehouse's content, goal difference frequency, and some noteworthy KPIs that we want our end users to observe are shown in Figure 7. Additionally, we offer a text box with summaries and disclaimers about our dashboards.



Filter by Name	Filter by FIFA ID
78-TRIAL-NOR 258 - 269-TRIAL-No...	1
AAR - FC Aarau	2
ABE - Aberdeen	3
ACA - Académica de Coimbra	4
ACM - Milan	5
AJA - AC Ajaccio	7
AJA - Ajax	9

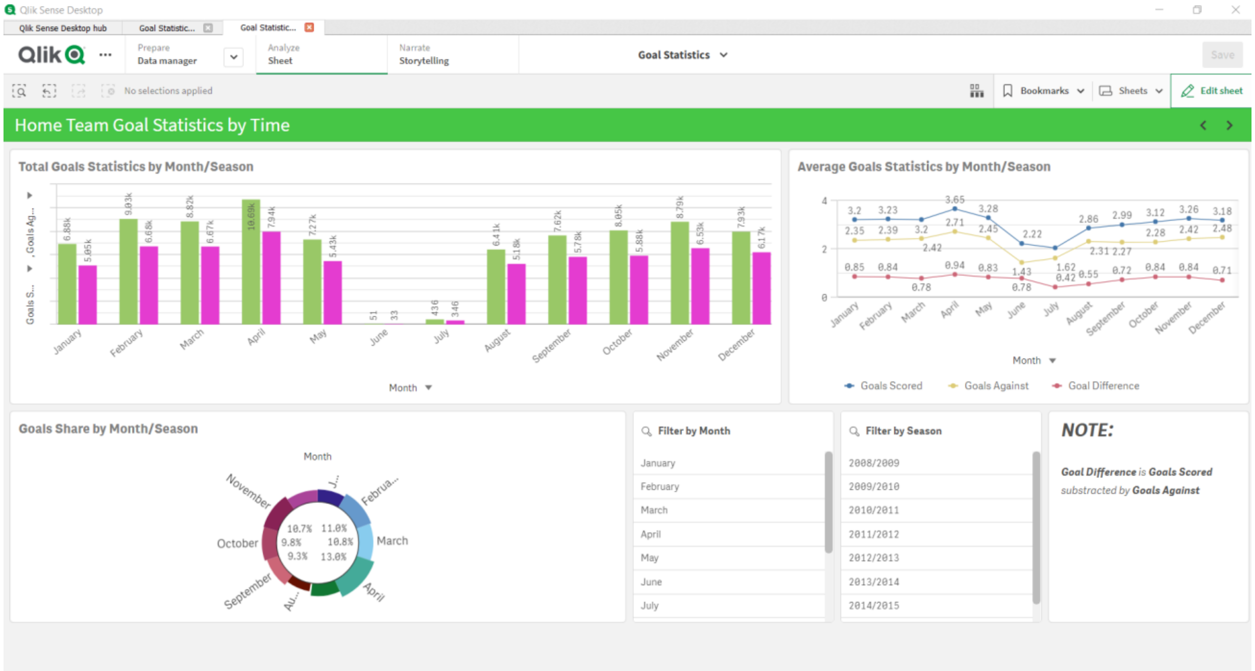
NOTE:

Goal Difference *is* ***Goals Scored*** *substracted by* ***Goals Against***

Screenshot is saved!
C:/Users/chepk/Desktop/3.png
1910x1026 px, 447 KB

Figure 8: Dashboard of Home Team Goal Statistics

Figure 8 displays our goal data dashboard broken down by the home team. We showed the top 25 teams' goal productivity along with a distribution graph and filtering options. Just by setting the maximum amount of data to be retrieved, the data for the top 25 teams could be collected. The filters offered include team name and FIFA ID.

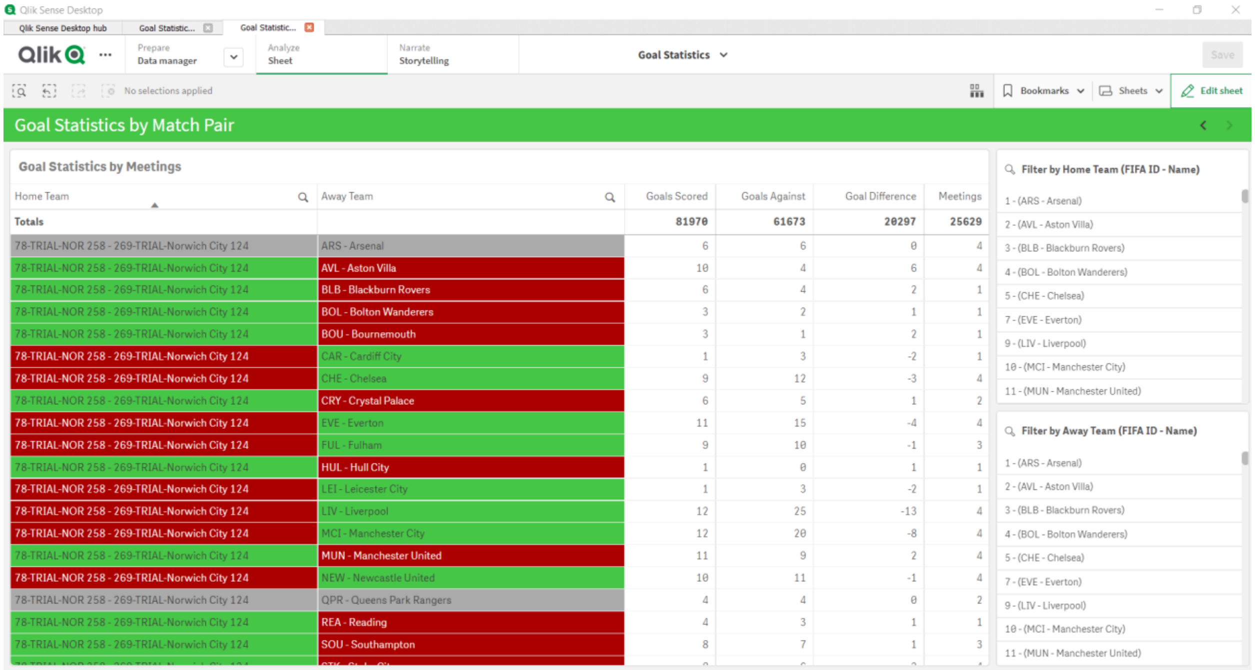


NOTE:

Goal Difference *is* ***Goals Scored*** *substracted by* ***Goals Against***

Figure 9: Dashboard of Time Dimension (Month and Season) of Data Warehouse

Given that both the month and the season dimensions fall under the same category, which is time, we provide a mixed-dimension representation of the data from our data warehouse. We offer filtering options, line charts, bar charts, and pie charts to display the goal statistics for each month or for the entire season. The graphs can be customised by choosing perspectives based on the month or season for each individual graph.



Goal Statistics by Match Pair

Goal Statistics by Meetings

Home Team	Away Team	Goals Scored	Goals Against	Goal Difference	Meetings
Totals		81970	61673	20297	25629
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	ARS - Arsenal	6	6	0	4
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	AVL - Aston Villa	10	4	6	4
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	BLB - Blackburn Rovers	6	4	2	1
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	BOL - Bolton Wanderers	3	2	1	1
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	BOU - Bournemouth	3	1	2	1
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	CAR - Cardiff City	1	3	-2	1
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	CHE - Chelsea	9	12	-3	4
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	CRY - Crystal Palace	6	5	1	2
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	EVE - Everton	11	15	-4	4
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	FUL - Fulham	9	10	-1	3
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	HUL - Hull City	1	0	1	1
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	LEI - Leicester City	1	3	-2	1
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	LIV - Liverpool	12	25	-13	4
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	MCI - Manchester City	12	20	-8	4
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	MUN - Manchester United	11	9	2	4
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	NEW - Newcastle United	10	11	-1	4
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	QPR - Queens Park Rangers	4	4	0	2
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	REA - Reading	4	3	1	1
78-TRIAL-NOR 258 - 269-TRIAL-Norwich City 124	SOU - Southampton	8	7	1	3

Filter by Home Team (FIFA ID - Name)

1 - (ARS - Arsenal)
2 - (AVL - Aston Villa)
3 - (BLB - Blackburn Rovers)
4 - (BOL - Bolton Wanderers)
5 - (CHE - Chelsea)
7 - (EVE - Everton)
9 - (LIV - Liverpool)
10 - (MCI - Manchester City)
11 - (MUN - Manchester United)

Filter by Away Team (FIFA ID - Name)

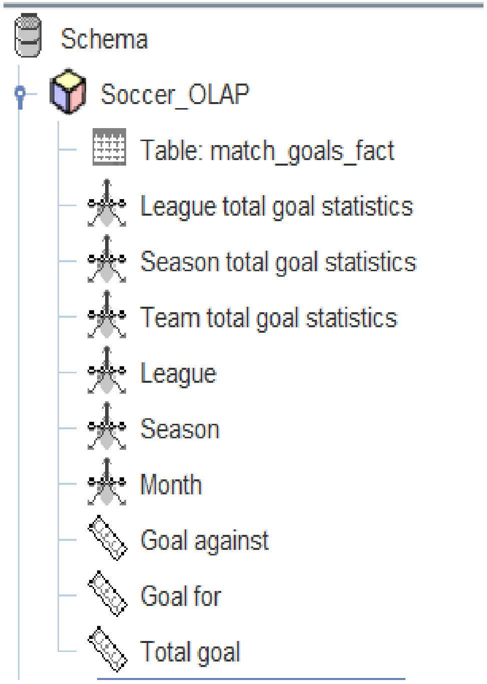
1 - (ARS - Arsenal)
2 - (AVL - Aston Villa)
3 - (BLB - Blackburn Rovers)
4 - (BOL - Bolton Wanderers)
5 - (CHE - Chelsea)
7 - (EVE - Everton)
9 - (LIV - Liverpool)
10 - (MCI - Manchester City)
11 - (MUN - Manchester United)

Figure 10: Dashboard of Match Pair Meetings Goal Statistics

Our final dashboard, Figure 10, represents meeting statistics broken down by match pairs. This implied that we had grouped the data in our data warehouse according to specific home and away team meetings. From the perspective of the home side, we had computed the overall goals scored, goals against, and goal differential. We had also used certain colours to embellish the table, with grey denoting a draw, green denoting a victory, and red denoting a defeat. Additionally, filtering options were provided. To create this dashboard, the team_dim data had to be loaded twice because Qlik Sense did not permit multiple relationships between the fact table and dimension table during a single load.

3.5 OLAP Operations

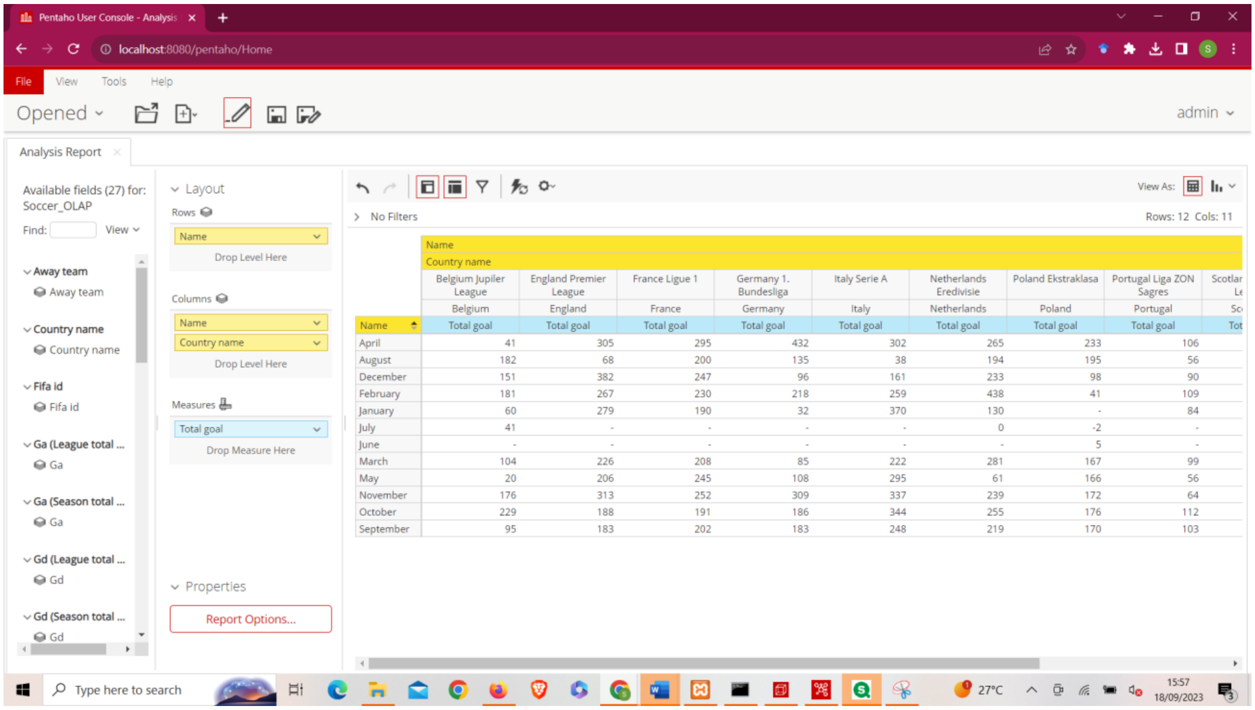
Pentaho Schema Workbench was used to design the cube. The schema was connected to the data warehouse to import the fact and dimension tables.
The cube's dimensions specify the hierarchies and levels within each dimension. The measures selected for this analysis were: '***goal against', 'goal for' and 'goal against'***



Schema
- Soccer_OLAP
  - Table: match_goals_fact
  - League total goal statistics
  - Season total goal statistics
  - Team total goal statistics
  - League
  - Season
  - Month
  - Goal against
  - Goal for
  - Total goal

Figure 11: Cube schema.

Once the cube design was complete, it was deployed to the Pentaho Analysis (Mondrian) server for multidimensional querying and analysis. This process involved generating XML files to describe the cube's structure and data sources



Rows: Name

Columns: Name, Country name

Measures: Total goal

No Filters

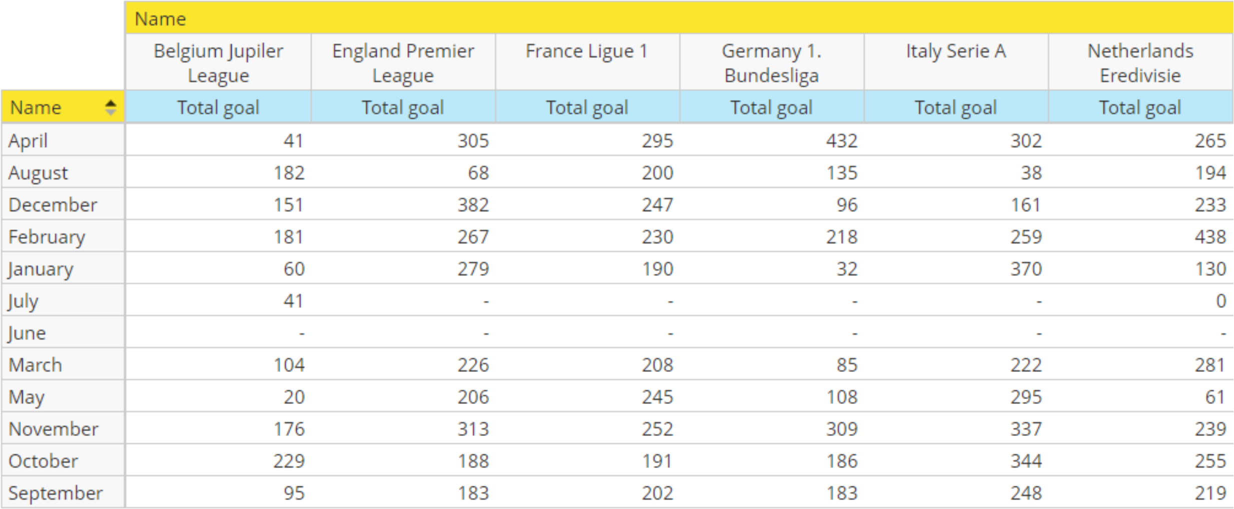
Rows: 12 Cols: 11

Name	Belgium Jupiler League	England Premier League	France Ligue 1	Germany 1. Bundesliga	Italy Serie A	Netherlands Eredivisie	Poland Ekstraklasa	Portugal Liga ZON Sagres	Scotlar Le
Country name	Belgium	England	France	Germany	Italy	Netherlands	Poland	Portugal	Sc
Name	Total goal	Total goal	Total goal	Total goal	Total goal	Total goal	Total goal	Total goal	Tot
April	41	305	295	432	302	265	233	106	
August	182	68	200	135	38	194	195	56	
December	151	382	247	96	161	233	98	90	
February	181	267	230	218	259	438	41	109	
January	60	279	190	32	370	130	-	84	
July	41	-	-	-	-	0	-2	-	
June	-	-	-	-	-	-	5	-	
March	104	226	208	85	222	281	167	99	
May	20	206	245	108	295	61	166	56	
November	176	313	252	309	337	239	172	64	
October	229	188	191	186	344	255	176	112	
September	95	183	202	183	248	219	170	103	

Figure 12: Three-dimension cube data, month, league and country

3.5.1 Slice

The data has been reduced to two dimensions by removing the country dimension.



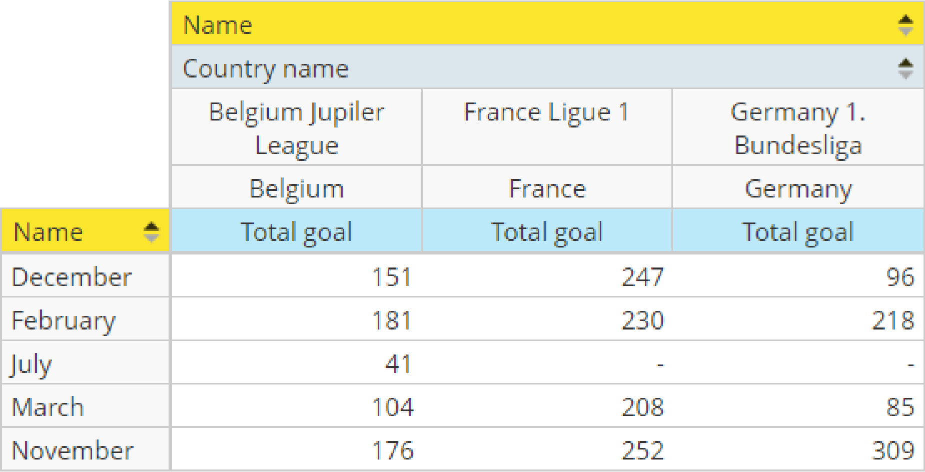
	Name					
	Belgium Jupiler League	England Premier League	France Ligue 1	Germany 1. Bundesliga	Italy Serie A	Netherlands Eredivisie
Name ⬍	Total goal	Total goal	Total goal	Total goal	Total goal	Total goal
April	41	305	295	432	302	265
August	182	68	200	135	38	194
December	151	382	247	96	161	233
February	181	267	230	218	259	438
January	60	279	190	32	370	130
July	41	-	-	-	-	0
June	-	-	-	-	-	-
March	104	226	208	85	222	281
May	20	206	245	108	295	61
November	176	313	252	309	337	239
October	229	188	191	186	344	255
September	95	183	202	183	248	219

Figure 13: Slice of original 3-dimensional cube

3.5.2 Dice

The figure below shows a sub-cube of the original 3-dimensional cube. The dice operation was based on the following criteria:

- Country = “Belgium”, “France”, “Germany”
- Month = “December”, “February”, “July”, “March”, “November”
- League = “Belgium Jupiler League”, “France Ligue 1”, “Germany 1. Bundesliga”



	Name ⬍		
	Country name ⬍		
	Belgium Jupiler League	France Ligue 1	Germany 1. Bundesliga
	Belgium	France	Germany
Name ⬍	Total goal	Total goal	Total goal
December	151	247	96
February	181	230	218
July	41	-	-
March	104	208	85
November	176	252	309

Figure 14: Dice- sub-cube of the original 3-dimensional cube

CONCLUSION AND RECOMMENDATIONS

In this report, we have presented the creation of a data warehouse for a soccer analytics project. We applied the Kimball Methodology, which is a popular approach for data warehousing. However, we omitted some steps due to the simplicity and limitations of our data. We have pointed out which parts do not follow the Kimball Methodology and how they should be done in a proper implementation.

We have also elaborated on our ETL process using Pentaho Data Integration, and the transformations that we performed. We have created a simple dashboard of five analysis sheets using Qlik Sense, and we have also generated reports using Pentaho Report Designer. These reports utilise graph representations and parameters to provide more detailed insights into the data.

Recommendations

Based on our experience in this project, we would recommend the following:

- Use the Kimball Methodology as a guide, but be flexible and adaptable to the specific needs of your project.
- Use an ETL tool like Pentaho Data Integration to automate the data loading and transformation process.
- Use a visualisation tool like Qlik Sense to create interactive dashboards and reports.

We hope that this report has been helpful in providing an overview of the data warehousing process. We believe that data warehousing is a valuable tool for businesses and organisations of all sizes. By providing a central repository for data, data warehouses can help organisations to make better decisions, improve efficiency, and gain a competitive advantage.

REFERENCES

Edited Book:
Wrembel, R. (Ed.). (2006). Data warehouses and OLAP: Concepts, architectures, and solutions. IGI Global.

Website Sources:

Pentaho Mondrian Workbench Documentation. (n.d.). Retrieved from https://mondrian.pentaho.com/documentation/workbench.php

Qlik Sense. (n.d.). Retrieved from https://www.qlik.com/us/products/qlik-sense

Hitachi Vantara. (2023). Pentaho Data Integration 9.4 Documentation. Retrieved from https://help.hitachivantara.com/Documentation/Pentaho/Data_Integration_and_Analytics/9.4/Products/Pentaho_Data_Integration

SQLServerCentral.com. (n.d.). The Kimball Approach. Retrieved from https://www.sqlservercentral.com/blogs/the-kimball-approach

Holistics. (n.d.). Kimball's Dimensional Data Modeling: A Comprehensive Guide. Retrieved from https://www.holistics.io/books/setup-analytics/kimball-s-dimensional-data-modeling/