UNIVERSITÀ DI PISA

LABORATORY OF DATA SCIENCE

Project Report

Ebrima Jallow (681743)

Academic Year 2024/2025

Contents

1 Introduction

This project involves the design and implementation of a data warehouse for tennis tournaments, based on real-world historical and structured data. The datasets include tournament-level, player-level, match-level, and country-level information, enabling analytical queries from different business perspectives.

The overarching objective is to develop a data mart that supports OLAP-style queries across various dimensions (time, geography, player attributes) and facts (e.g., number of matches, tournament stats, player performance). Through ETL processes and multidimensional modeling, the project prepares the data for efficient analytical querying and visualization.

2 Data Understanding

2.1 Tournament Data

The `sample_tourney.json` file provides information about tennis tournaments. Each record represents a tournament occurrence with attributes including:

- **tourney_id**: Unique identifier (e.g., "1968-303")
- **tourney_name**: Name of the tournament (e.g., "Buenos Aires")
- **surface**: Surface type (e.g., "Clay", "Grass", "Carpet", "Hard")
- **draw_size**: Number of players in the main draw
- **tourney_level**: Level of tournament (e.g., "A", "M", "G")
- **tourney_timestamp**: Unix timestamp representing tournament start date

Redundancy was observed in the dataset with multiple repeated entries per tournament. Data cleaning involved deduplication using a combination of `tourney_id` and `tourney_timestamp` to retain only unique tournament instances.

2.2 Country Data

The `countries.xml` file includes metadata about countries, such as:

- Country code and name
- ISO standard codes
- Geographic regions or continents

This file is used to map player nationalities and enable region-based analysis. The structure was flattened and normalized during ETL.

However, data quality issues were identified. For example, the country "Oman" was incorrectly labeled under the continent "Africa" instead of "Asia." Such inconsistencies were manually corrected.

Additionally, several country codes found in the `fact.csv` file were missing from the country dataset. Since the player dimension is extracted from the fact data, all missing countries had to be added to ensure consistency and avoid integrity errors during data import.

2.3 Fact Table Data

The `fact.csv` file contains match-level and player-level statistics, forming the core of the fact table. Notable columns include:

- Match outcomes and other metrics like profits etc.
- Player identifiers and tournament references

- Nationality fields, used for building geography (country) dimensions

The data underwent preprocessing to standardize formats, handle null values, and validate links to the tournaments and country datasets. The cleaned data is ready for analytical exploration via OLAP queries.

3 Data Cleaning Summary

The data cleaning process focused on preparing the datasets for integration into the data warehouse. Initial exploratory data analysis (EDA) identified several columns with over 70% missing values; these columns were deemed unsuitable for reliable analysis and were subsequently removed. For essential fields with missing data, such as player age, imputation strategies were employed. Specifically, the `match_id` was utilized as a proxy timeline to estimate missing ages, under the assumption that match IDs increment chronologically. This approach allowed for the estimation of player ages based on the temporal sequence of matches. Additionally, categorical variables with missing entries, like hand preference and surface type, were imputed using the mode within relevant groupings (e.g., by tournament or player). These cleaning steps ensured the datasets were consistent and ready for the ETL processes and subsequent analytical tasks.

4 Data Warehouse Schema

The warehouse follows a snowflake schema centered on a **Fact_Matches** table that connects to **Dim_Players** and **Dim_Tournaments**. Additional snowflaked dimensions such as **Dim_Countries** and **Dim_Time** provide hierarchical navigation for analysis.

- **Fact_Matches**: Contains match-level statistics (e.g., scores, rank points, match expenses).
- **Dim_Players**: Includes player details (e.g., name, hand preference, age).
- **Dim_Tournaments**: Contains metadata about tournaments (e.g., surface, level, location).
- **Dim_Countries**: Maps player nationalities to country names and continents.
- **Dim_Time**: Breaks down timestamps into year, month, day, and quarter for time-based analysis.

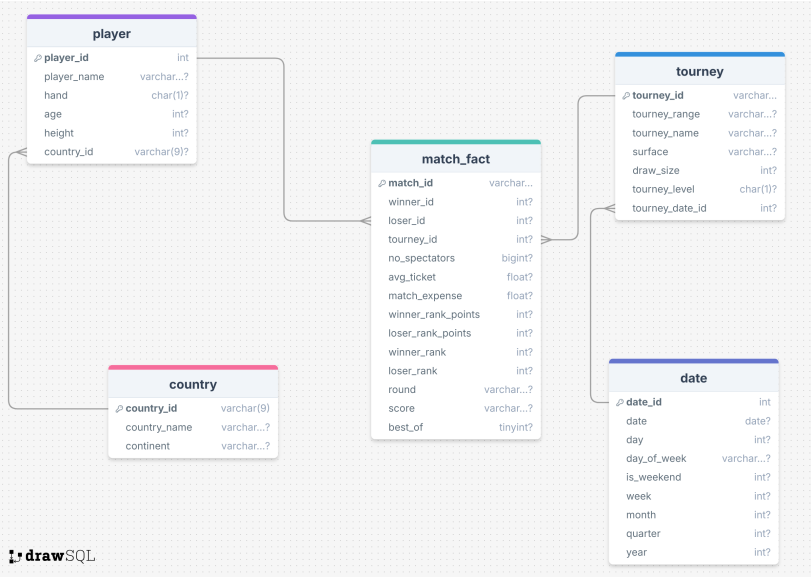

Figure 1: Snowflake Schema for Tennis Tournament Analytics

5 SQL Server Integration Services (SSIS)

The SSIS assignments began with the creation of additional tables as required in Assignment 6. For each original table, a duplicate was created using the naming convention `TABLENAME_SSIS`, with empty schemas.

A data flow within an SSIS project was developed to populate these tables with 20% of the data extracted in the preparation phase. A similar attempt was made using a Python script.

Eventually, full data loading into the data warehouse was done using the Python script, as it proved faster and more reliable. The Visual Studio environment occasionally froze or crashed, making it less dependable on the working machine.

5.1 Assignment 6: Nemesis

This query lists each player's nemesis—the opponent they lost the most matches to—along with the number of losses, grouped by year.

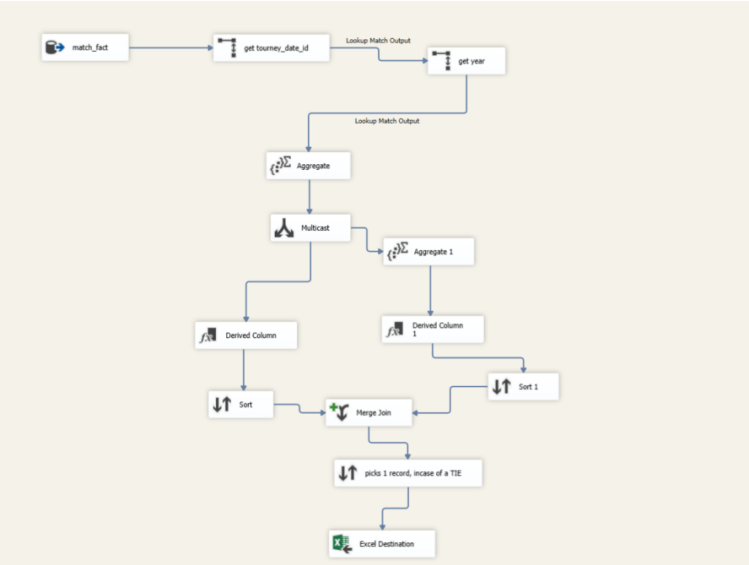

Figure 2: Yearly Nemesis per Player

5.2 Assignment 7: Outlier Matches

A match is considered an age-outlier if the age difference between the winner and loser exceeds 1.5 times the average age difference for matches in the same tournament. This query identifies the player who participated in the most outlier matches each year.

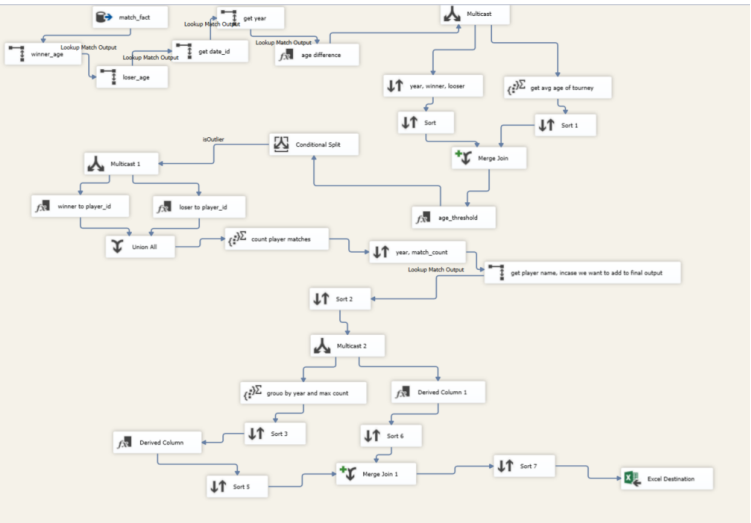

Figure 3: Players with Most Age-Outlier Matches by Year

6 Multidimensional Expressions (MDX)

6.1 Cube Design

The multidimensional cube was developed to support rich analytical operations over tennis match data, enabling advanced slicing, dicing, and aggregation across multiple business-relevant perspectives. Several hierarchies were implemented within the dimensions of time, geography, players, and tournaments, with a wide range of numeric measures to support in-depth analysis.

The cube includes the following:

- **Measures**: Match count, number of spectators, player rank points (winner and loser), profit, expenses, and other performance indicators.
- **Dimensions**: Player (both winner and loser), Tournament, Date, and Country (Geography)

6.2 Date Hierarchy

Two distinct hierarchies were defined:

The primary hierarchy follows a chronological structure: *Day* → *Month* → *Quarter* → *Week* → *Year*, enabling detailed time-series analysis and seasonal trend identification. A secondary hierarchy, focused on temporal behavior patterns, organizes dates by *Day of Week* and *Is Weekend*, supporting analyses such as weekday vs. weekend performance, attendance, or profitability. These hierarchies enrich the cube with versatile temporal perspectives crucial for understanding time-based dynamics in tennis matches and tournaments.

6.3 Country Hierarchy

To facilitate regional and continental breakdowns, a country (geography) dimension was constructed from the `countries.xml` file. It follows the structure: *Continent* → *Country*. Each player is linked to this hierarchy via their nationality, enabling MDX queries that summarize player outcomes or financial metrics by region.

6.4 Tournament Hierarchy

The tournament dimension was enriched with a composite hierarchy that integrates both structural and temporal elements. It begins with the *Tourney Level* (e.g., Grand Slam, ATP 1000), followed by the *Tourney*

Name, then *Surface Type*, and finally the complete temporal breakdown: *Date Id* → *Day* → *Month* → *Quarter* → *Year*. This unified hierarchy allows users to drill from high-level tournament categories all the way down to specific dates of play, while also analyzing surface-based performance variations within each level or tournament.

6.5 Player Hierarchy

To support geographical aggregation, a two-level hierarchy was established: *Continent* → *Country Name*. This structure, derived from the player's country of origin, enables region-based analysis of player performance and participation. The dimension design supports queries that compare metrics such as rank points, average age across countries and continents, offering a detailed view of global player distribution and trends in international competitiveness.

These hierarchies together create a flexible and powerful cube structure, enabling the project's MDX queries and supporting dashboard interactivity. They allow for diverse perspectives in understanding trends in player performance, tournament characteristics, and financial metrics across time and geography.

6.6 Assignment 9: Most Losing Player per Continent

This MDX query identifies the player who lost the most matches in each continent. The cube aggregates match outcomes based on player nationality, joined to the `Dim_Countries` table.

```
WITH
MEMBER [Measures].[Match Count By Continent] AS
(
  [Loser].[Player Id].CURRENTMEMBER,
  [Measures].[Match Fact Count]
)

SELECT
  {[Measures].[Match Count By Continent]} ON COLUMNS,

  GENERATE(
    [Country].[Continent].[Continent].MEMBERS,
    CROSSJOIN(
      { [Country].[Continent].CURRENTMEMBER },
      TOPCOUNT(
        [Loser].[Player Id].[Player Id].MEMBERS,
        1,
        ([Measures].[Match Fact Count], [Country].[Continent].CURRENTMEMBER)
      )
    )
  ) ON ROWS

FROM [Group ID Jallow DB CUBE]
```

6.7 Assignment 10: Number of Players per Tournament

This query calculates the number of unique players that participated in each tournament, providing insight into draw size and tournament engagement.

```
WITH
MEMBER [Measures].[Winner Players] AS
      DISTINCTCOUNT(
        NONEMPTY(
            [Winner].[Player Id].[Player Id].MEMBERS,
            [Measures].[Match Fact Count]
        )
    )

MEMBER [Measures].[Loser Players] AS
      DISTINCTCOUNT(
        NONEMPTY(
            [Loser].[Player Id].[Player Id].MEMBERS,
            [Measures].[Match Fact Count]
        )
    )

MEMBER [Measures].[Approx Total Players] AS
    [Measures].[Winner Players] + [Measures].[Loser Players]

SELECT
  {[Measures].[Approx Total Players]} ON COLUMNS,
  [Tourney].[Tourney Name].[Tourney Name].MEMBERS ON ROWS

FROM [Group ID Jallow DB CUBE]
```

6.8 Assignment 11: Profit Change by Quarter and Year

This query calculates the percentage change in profit between the same quarter across two consecutive years for each tournament.

```
WITH
-- Previous Year Profit (same quarter last year)
MEMBER [Measures].[Previous Year Profit] AS
    (
        [Measures].[Profit],
        ParallelPeriod(
            [Tourney].[Hierarchy].[Year],
            1,
            [Tourney].[Hierarchy].CurrentMember
        )
    )

-- % Change in Profit YoY
MEMBER [Measures].[YoY % Change] AS
    IIF(
        IsEmpty([Measures].[Previous Year Profit]) OR [Measures].[Previous Year Profit] = 0,
        NULL,
        ([Measures].[Profit] - [Measures].[Previous Year Profit])
        / [Measures].[Previous Year Profit]
    )

-- Label for Current Year-Quarter
MEMBER [Measures].[Current Year-Quarter] AS
    [Tourney].[Hierarchy].CurrentMember.Name

-- Label for Previous Year-Quarter
MEMBER [Measures].[Previous Year-Quarter] AS
    ParallelPeriod(
        [Tourney].[Hierarchy].[Year],
        1,
        [Tourney].[Hierarchy].CurrentMember
    ).Name

SELECT
  {
    [Measures].[Profit],
    [Measures].[Previous Year Profit],
    [Measures].[YoY % Change],
    [Measures].[Current Year-Quarter],
    [Measures].[Previous Year-Quarter]
  } ON COLUMNS,

  NON EMPTY
  CROSSJOIN(
    [Tourney].[Tourney Name].[Tourney Name].MEMBERS,
    [Tourney].[Hierarchy].[Quarter].MEMBERS
  ) ON ROWS

FROM [Group ID Jallow DB CUBE]
```

7 Interactive Dashboards with PowerBI

7.1 Dashboard 1: Global Distribution of Winner and looser Rank Point

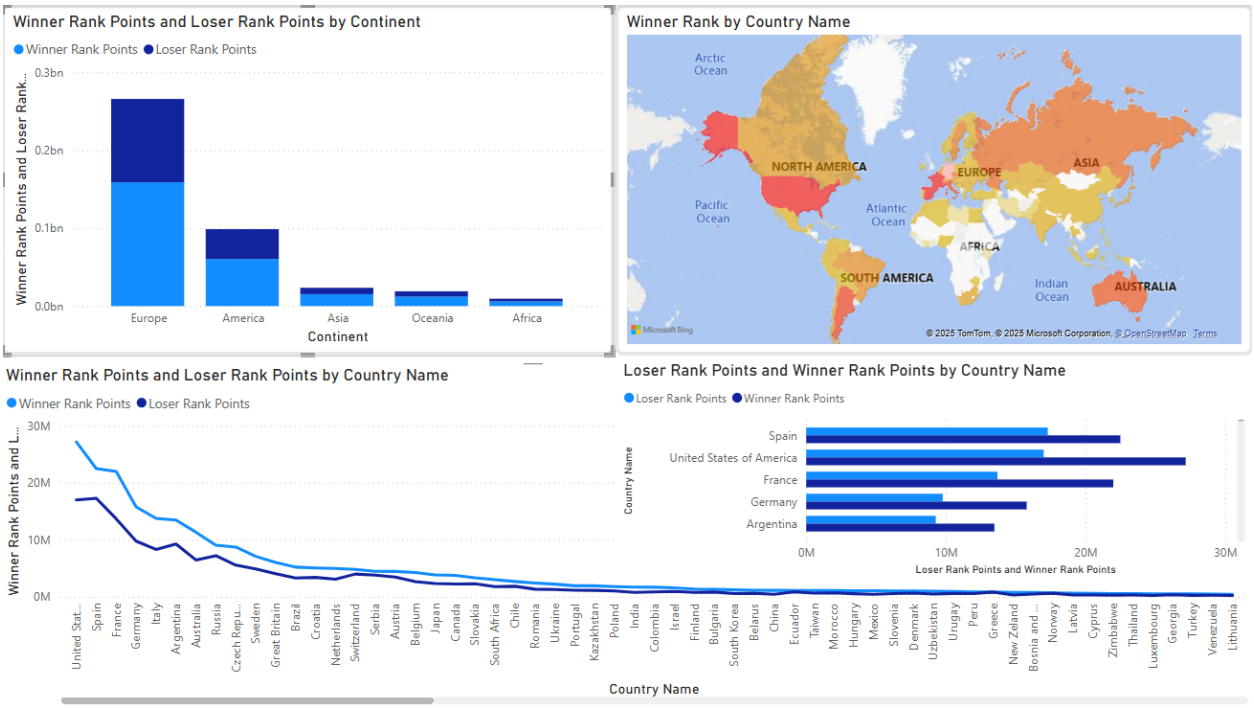

Figure 4: Geographical Distribution of Winner and looser Rank Point

This dashboard presents a geographical analysis of player performance, comparing winner and loser rank points across continents and countries. Europe leads with the highest cumulative rank points, followed by America and Asia. A world map highlights the spread of top-performing countries, while supporting charts offer a clear comparison between nations such as Spain, the United States, France, and Germany. The visuals provide insight into the global distribution of tennis competitiveness, emphasizing regional dominance and balance between winners and losers.

This dashboard is useful for identifying powerhouses in tennis performance and understanding geographic disparities in competitiveness.

7.2 Dashboard 2: Financial Analysis and Tournament Insights

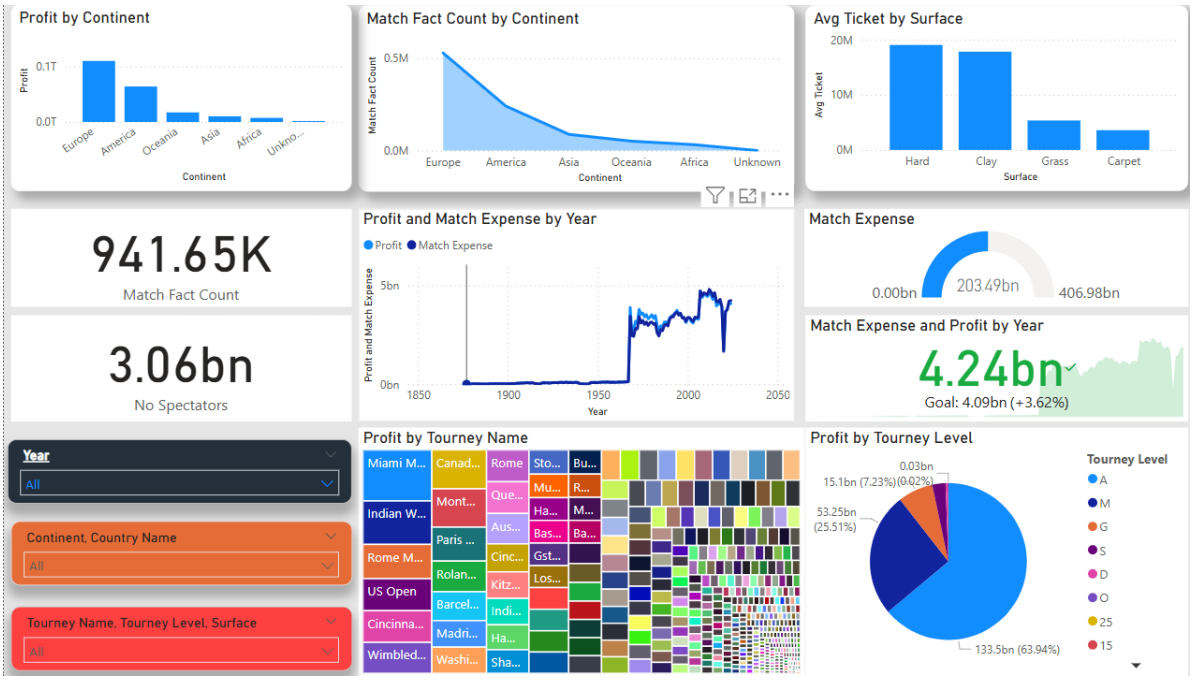

Figure 5: Match Profitability, Surface Analysis, and Tournament Trends

The second dashboard focuses on the financial aspects of tennis tournaments, including profit, match expenses, and ticket pricing across surfaces. It shows that hard and clay courts draw higher average ticket values, while Europe continues to dominate in both match count and profitability. Historical trends of profit and expenses are visualized, alongside KPIs that track overall revenue performance. Users can interactively filter by year, country/continent, and tournament characteristics. The dashboard also breaks down profit by individual tournaments and levels, revealing that higher-tier tournaments contribute the most to overall financial success.