**Business Intelligence**

Assignment 2

**“World Cup”**

***Nitzan Ziflinger***

***Noa Burg***

***Amit Nitzan***

**Part 1 – Define Data warehouse**

1. In a data warehouse, a star schema can have one fact table and a number of associated dimension tables in the center. It is referred to as a star schema because its structure resembles a star. The simplest type of Data Warehouse schema is the Star Schema data model.

Chart

Description automatically generated

We used a star diagram because:

• Dimensional hierarchies are stored in the dimensional table.

• It has a fact table in the center, surrounded by dimension tables.

• Only a single join creates the relationship between the fact table and any dimension tables in a star schema.

• Denormalized Data structure and query performance are also improved.

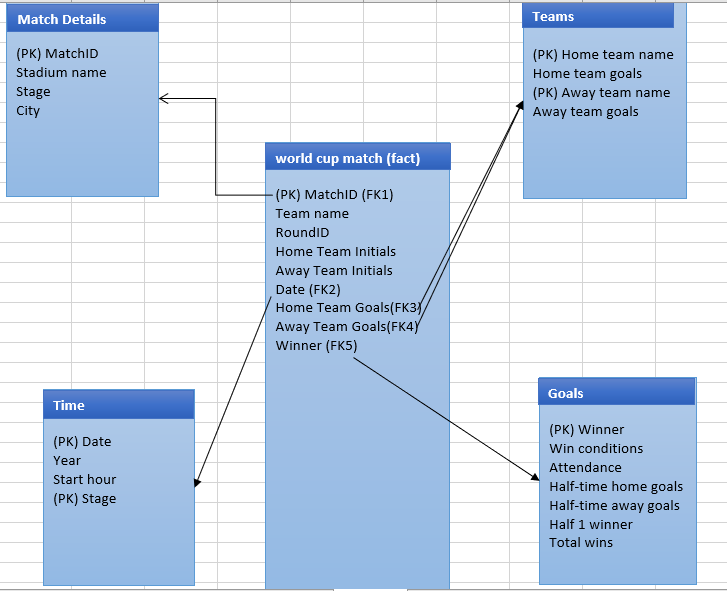
B.

A data warehouse is a decision support system that stores and processes historical data from across the organization, allowing it to be used for critical business analysis, reports, and dashboards.

Data warehouse systems store data from many sources. It typically includes structured online transaction processing (OLTP) data like ours include columns as: Year, Date, Start Hour, Stage, Stadium, City, Home Team Name, Home Team Goals, Away Team Goals, Away Team Name, Winner, Win conditions, Attendance, Half-time Home Goals, Half-time Away Goals, Half 1 Winner and total wins.

To improve query speed, the star schema splits the information from the fact table into denormalized dimension tables. To link fact tables to dimensions, only one join is required.

C.



D.

Star Schema databases are best suited for storing historical information. Star schemas, which are primarily read optimized, will provide good performance over large data sets.

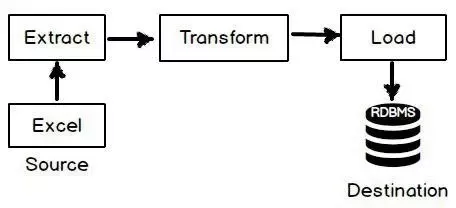
Star Schema helps in decision making because it increases performance, data slicing, and easier data comprehension.

Using our schema, a team coach can realize how much efforts the team needs to put in during a match in the two half of the world cup matches. In addition, this schema can help a team coach to understand which other competitor team can be more challenging and helps building new strategy on each stage at the world cup matches.

Furthermore, the coach will be able to choose and decide which players will take place in a specific game and when (first half, second half) according to the chances of the competitor team to win the match, which be presented in our analysis.

**Part 2 – Definition and implementation of ETL**

1. ETL is a process that stands for Extract, Transformation, and Loading. When we want to load data from any source to any destination, we must go through at least three steps. The first step is to extract data, then map and transform it before loading it to the destination.



**Extract:**

The goal of this "Extract" step is to collect the correct data, from the correct source, in the correct quantity.

**Transform**

**Filtering:**

We will filter the following columns, as they are not relevant to our analysis goals:

Attendance, Assistant 1, Assistant 2, Referee.

**Cleaning:**

For the win conditions column, we will clean the null values. The type of this column is a string, which means we can't replace the null values with median or mean, so the only option we can do here is to put the word "NULL" instead of the blank value.

**Sorting:**

We will sort by the A-Z of the winning team name. and then we will sort by year from Largest to Smallest.

Graphical user interface, application, table

Description automatically generated

**Load:**

The final step in the ETL process is data loading.

We will load the files to the data warehouse (attached).

|  |  |
| --- | --- |
| ETL Pipeline | |
| Reference Data | We reviewed the data. We will filter the following columns, as they are not relevant to our analysis goals:  Attendance, Assistant 1, Assistant 2, Referee. |
| Data Extraction | The data extracted into four different data set as per their types, such as teams, goals, time and match. |
| Data validation | We will do correction to resolve a few problems in the data. For example, we will make sure that there are no duplicate countries names (example: Czech Republic& Czechia) |
| Data Transformation | We will remove erroneous data, apply business rules, we will summarize the number of goals by the 1st, the 2nd and the whole game. |
| Stage | We will use MySQL as our staging DB. |
| Data Publishing | We will load the data to the target tables that were defined. |

Attached is a link to a Google Colab with the process that was done. [**https://colab.research.google.com/drive/14f0mujB9RTsQ131OH\_-E\_PvpDJKKGAkW?usp=sharing**](https://colab.research.google.com/drive/14f0mujB9RTsQ131OH_-E_PvpDJKKGAkW?usp=sharing)

**Below you can find a few (please check the attached code in order to see the full process) screenshots from the process:**

* **Replacing:**



* **Finding Outliers:**

