UNIVERSITY OF PISA

MASTER'S DEGREE IN
DATA SCIENCE & BUSINESS INFORMATICS

LABORATORY OF DATA SCIENCE
FIRST ASSIGNMENT

Build a Data Warehouse

MARCO CIOMPI [537856]
LUFTJAN SALIAJ [606507]
PASQUALE GORRASI [597817]

December 8, 2021

Contents

2 Part 2

2.1 Introduction

In Part 2 of the project we are required to solve some problems on the database we created in Part 1. The exercises has to be solved using Sequel Server Integration Services (SSIS) with computation on client side

2.2 Assignment 0 - *For every tournament, the players ordered by number of matches won.*

To obtain this list we retrieved data from the tables *Match* and *Player*, respectively the *tourney id, winner id* columns and the *player id* column. Then we merge the two databases ordering the id columns, and then we aggregate in order to have for every *tourney id* the list of players and for every *player id* the number of victories, retrieved by counting the *winner id*. Finally the list has been written on a csv file.

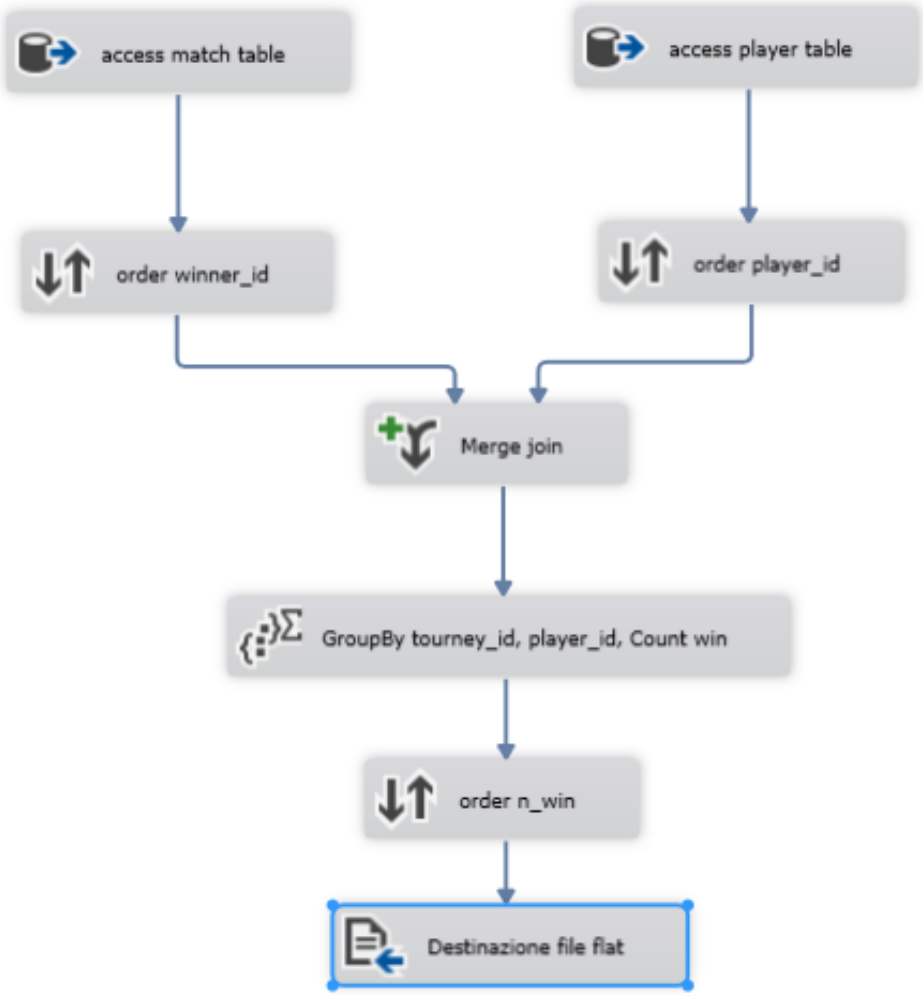

Figure 2.1: SSIS process to obtain the list.

2.3 Assignment 1 - *A tournament is said to be "worldwide" if no more than 30 percent of the participants come from the same continent. List all the worldwide tournaments.*

This assignment required an intricate procedure in order to avoid data loss and calculate the required percentage through the aggregation nodes.
The data were extracted from the table *Match*, the attributes *tourney id*, *loser id* and *winner id*. Then we appended the *country id* column from *Players* to tie the *continent* column from *Geography* using the lookup nodes. This step was done twice exploiting the multicast and union node in order not to lose data, (selecting just one between *loser id* and *winner id*), to start the process. Therefore we discarded the duplicate in *player id*.
Another multicast was required in order to aggregate by *tourney id* counting the total players on one side and all the distinct continent on the other, then we merged on *tourney id*.
Next we added a derivate column node to calculate the percentage of players coming from any continent and we kept just the higher percentage for any different tournament using an aggregation. This lead to a conditional subdivision fixing the percentage < 0.3 and the result is just one tourney considered "worldwide" by this logic.

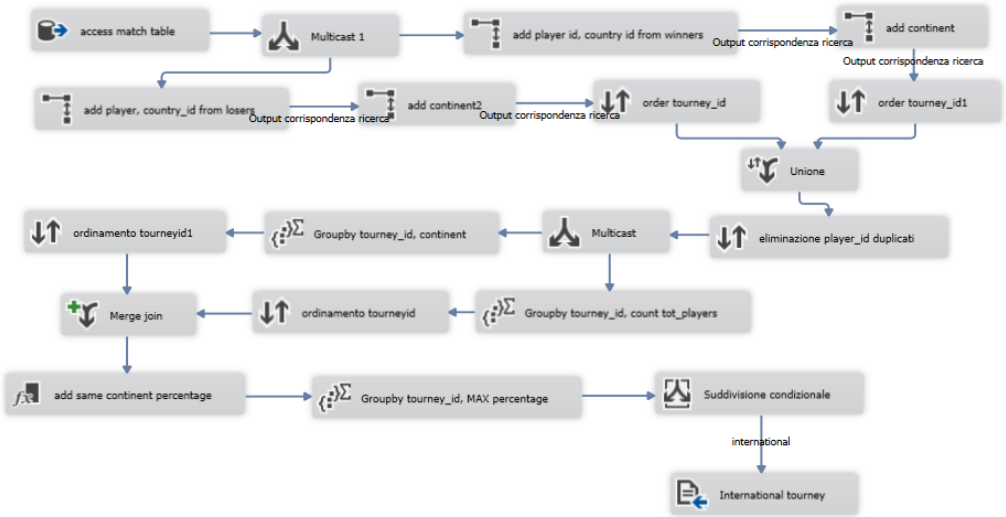

Figure 2.2: SSIS process to obtain the tournament.

2.4 Assignment 2 - *For each country, list all the players that won more matches than the average number of won matches for all players of the same country.*

For this query as in the previous one we gained data from the *Match* table casting for winners and losers and summing up the number of match won and lost by any player, merging at the end of the process. Then we retrieved *country id* from *Player* table.
Next we added a column with the derivate column node calculating the average of victories for any players, then with an aggregation node we calculated the winning average for any country. Finally with a conditional subdivision node we filtered just the players with a winning average above the average of their country.

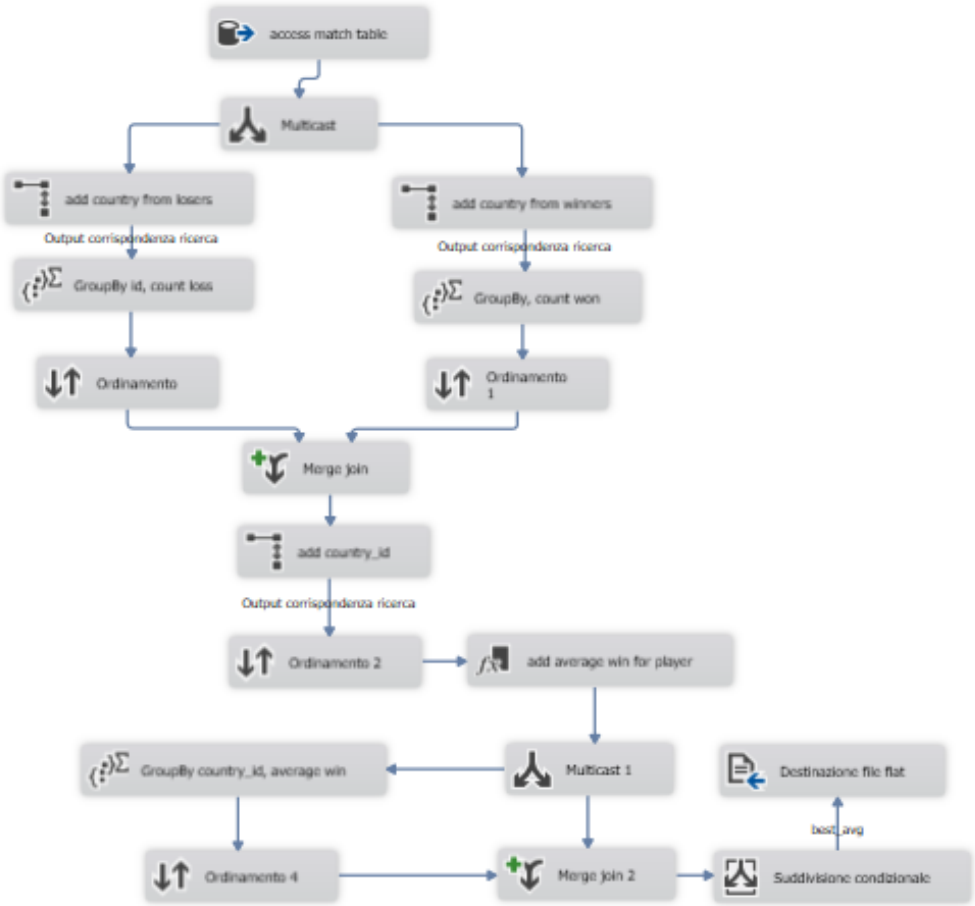

Figure 2.3: SSIS process to obtain the tournament.