LAB ASSIGNMENT:
ATP Men's Tour data

Bruna Kimura (up201902504)
Raphael Costa (up201902503)

1. Subject description. Assignment requirements.

For this activity, the dataset should comply some prerequisites:

1) The number of facts must be over 10000 (ten thousand) and must have, at least, one additive and one non-additive measure.
2) The dataset should have at least 4 dimensions, one of which temporal, and some of them are common to both kinds of facts

2. Planning: dimensional bus matrix, dimensions and facts dictionary

Dimensional bus matrix:

	Date	Player	Tournament	Court	Match
WRank	X	X	X		X
LRank	X	X	X		X
WPts	X	X	X	X	X
LPts	X	X	X	X	X
W1	X	X	X	X	X
L1	X	X	X	X	X
W2	X	X	X	X	X
L2	X	X	X	X	X
W3	X	X	X	X	X
L3	X	X	X	X	X
W4	X	X	X	X	X
L4	X	X	X	X	X
W5	X	X	X	X	X
L5	X	X	X	X	X
Wsets	X	X	X		X
Lsets	X	X	X		X
B365W	X	X	X		X
B365L	X	X	X		X
EXW	X	X	X		X
EXL	X	X	X		X

LBW	X	X	X		X
LBL	X	X	X		X
PSW	X	X	X		X
PSL	X	X	X		X
MaxW	X	X	X		X
MaxL	X	X	X		X
AvgW	X	X	X		X
AvgL	X	X	X		X

Dimensions and facts dictionary:

Dimensions	- ATP - Location - Tournament - Data - Series - Court - Surface - Round - Best Of - Winner - Loser - Comment
Facts	- WRank - LRank - WPts - LPts - W1 - L1 - W2 - L2 - W3 - L3 - W4 - L4 - W5 - L5 - Wsets - Lsets - B365W - B365L - EXW - EXL - LBW

	- LBL - PSW - PSL - MaxW - MaxL - AvgW - AvgL

3. MultiDimConceptual model

Star Scheme:

Court
- idCourt
- Court
- Surface

Tournament
- idTorneio
- Tournament
- Location
- ATP
- Serie

Result
- WRank
- LRank
- WPts
- LPts
- W1
- L1
- W2
- L2
- W3
- L3
- W4
- L4
- W5
- L5
- Wsets
- Lsets

Date
- idDate
- Day
- Month
- Year

Betting Odds
- B365W
- B365L
- EXW
- EXL
- LBW
- LBL
- PSW
- PSL
- MaxW
- MaxL
- AvgW
- AvgL

Match
- idMatch
- Round
- BestOf
- Comment
- Winner
- Loser
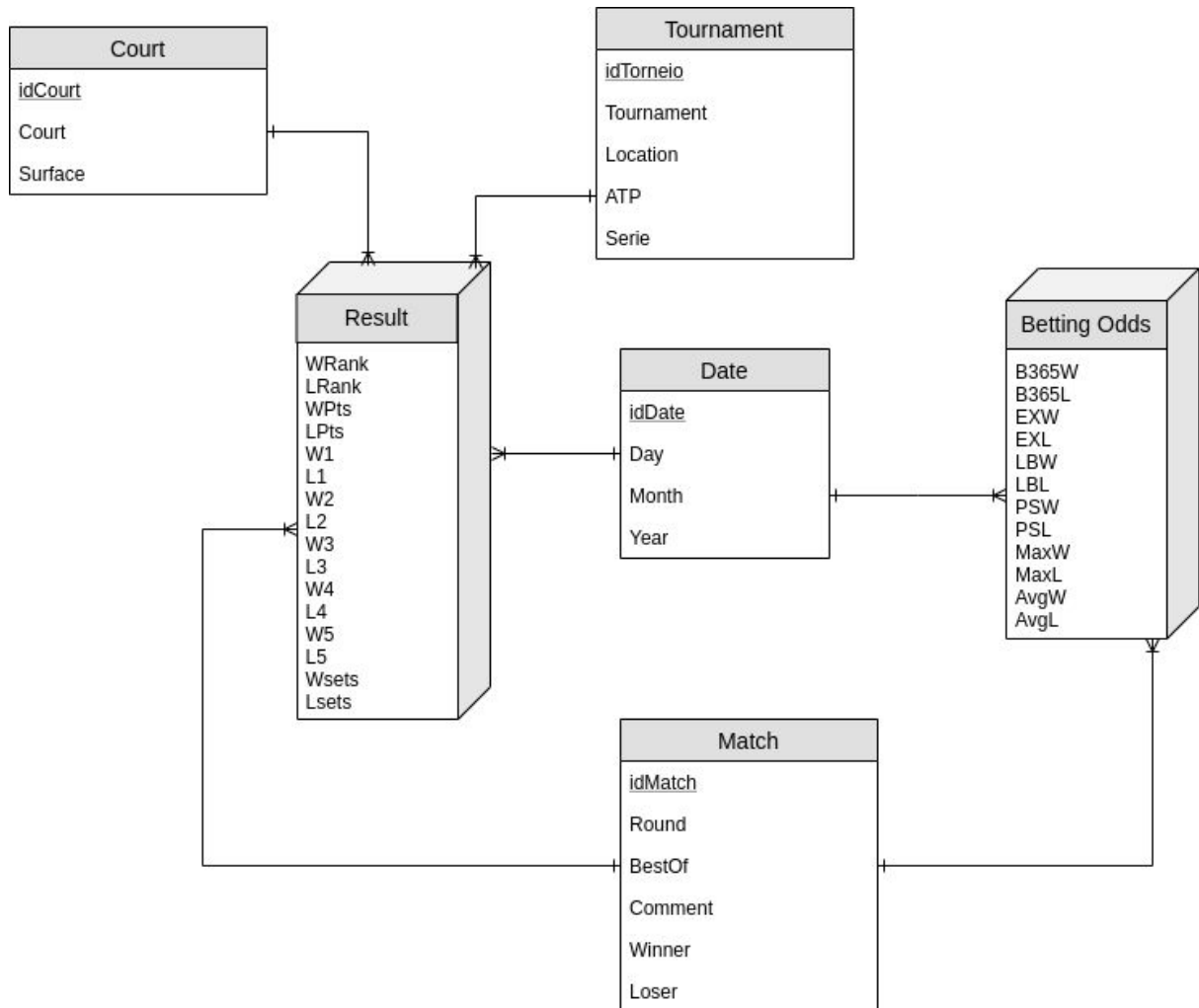
4. Dimensional data model (explained).

Column	Description	Category
ATP	Tournament number (men)	Dimension
Location	Venue of tournament	Dimension
Tournament	Name of tournament (including sponsor if relevant)	Dimension
Data	Date of match (note: prior to 2003 the date shown for all matches played in a single tournament is the start date)	Dimension
Series	Name of ATP tennis series (Grand Slam, Masters, International or International Gold)	Dimension
Court	Type of court (outdoors or indoors)	Dimension
Surface	Type of surface (clay, hard, carpet or grass)	Dimension
Round	Round of match	Dimension
Best of	Maximum number of sets playable in match	Dimension
Winner	Match winner	Dimension
Loser	Match loser	Dimension
Comment	Comment on the match (Completed, won through retirement of loser, or via Walkover)	Dimension
WRank	ATP Entry ranking of the match winner as of the start of the tournament	Fact
LRank	ATP Entry ranking of the match loser as of the start of the tournament	Fact
WPts	ATP Entry points of the match winner as of the start of the tournament	Fact
LPts	ATP Entry points of the match loser as of the start of the tournament	Fact

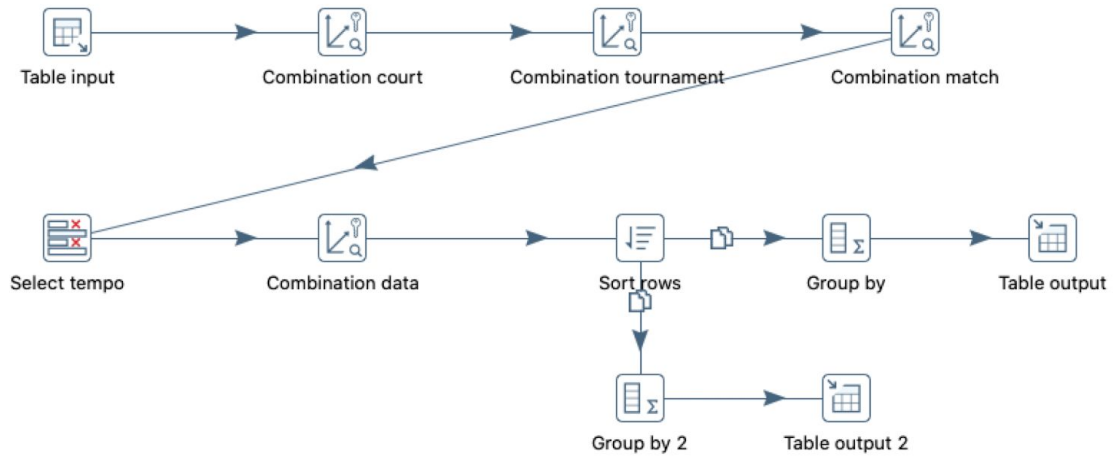|		
W1	Number of games won in 1st set by match winner	Fact
L1	Number of games won in 1st set by match loser	Fact
W2	Number of games won in 2nd set by match winner	Fact
L2	Number of games won in 2nd set by match loser	Fact
W3	Number of games won in 3rd set by match winner	Fact
L3	Number of games won in 3rd set by match loser	Fact
W4	Number of games won in 4rd set by match winner	Fact
L4	Number of games won in 4th set by match loser	Fact
W5	Number of games won in 5th set by match winner	Fact
L5	Number of games won in 5th set by match loser	Fact
Wsets	Number of sets won by match winner	Fact
Lsets	Number of sets won by match loser	Fact
B365W	Bet365 odds of match winner	Fact
B365L	Bet365 odds of match loser	Fact
EXW	Expekt odds of match winner	Fact
EXL	Expekt odds of match loser	Fact
LBW	Ladbrokes odds of match winner	Fact
LBL	Ladbrokes odds of match loser	Fact
PSW	Pinnacles Sports odds of match winner	Fact
PSL	Pinnacles Sports odds of match loser	Fact
MaxW	Maximum odds of match winner (as shown by Oddsportal.com)	Fact

MaxL	Maximum odds of match loser (as shown by Oddsportal.com)	Fact
AvgW	Average odds of match winner (as shown by Oddsportal.com)	Fact
AvgL	Average odds of match loser (as shown by Oddsportal.com)	Fact

5. Data sources selection

Tennis is one of the most played sport nowadays. Since the 12th century, humans have been playing this racket sport against one another, making Tennis one of the oldest sport known. The subject of this project consists on the ATP Men's Tour data. The ATP Tour is the where the global elite of tennis professional players compete against each other in a circuit of championships, organised by the Association of Tennis Professionals. The objective of it, of course, besides winning the tournaments is growing in the ATP Ranking. Depending on what position a player finishes a tournament, he wins a bunch of points in the ranking. Nowadays, the 1st player of this rank is Novak Djokovic, a Serbian player, followed by the spanish Rafael Nadal.

The dataset analysed was built using the information available in The Tennis-Data Betting, Results and Livescores Portal. It consists on the record of each match in all ATP Tournaments from 2015 to 2018, with the winner and loser infos, the place where the match was played, etc. Furthermore, there's also some bet odds info about that match. In other words, it's how much a site paid for a punter for making the correct bet on the winner of that match.

6. Transformations including incremental loadings

7. Jobs