Football Data Analysis

1. Player Valuations

   Dataset link:
   https://docs.google.com/spreadsheets/d/1X5XMxLCQ3Xz9yGl9gFLc-W_OmRLCjBI85LOaM0CsFHQ/edit?usp=sharing

Column name	datatype	
date	date	
datetime	date	
dateweek	date	
player_id	integer	
current_club_id	integer	
market_value	integer	
player_club_domestic_competition_id	string	

2. Club Games

   Dataset Link:
   https://docs.google.com/spreadsheets/d/14nkK7OxgltIb6IGqJJ5af6_qoTl3Uh-shfMn7l5xec0/edit?usp=sharing

Column name	string	
club_id	integer	
game_id	Integer	
own_goals	Integer	
own_position	Integer	
own_manager_name	string	
opponent_id	Integer	
opponent_goals	Integer	

opponent_position	Integer	
opponent_manager_name	string	
hosting	string	
is_win	integer	Generate a new column where it will say, win_status: win or loss

3. Competitions

Dataset Link: https://docs.google.com/spreadsheets/d/1yru26gCJzXFBRH7jcTxlctIaRmp2zCTj-mpqzMnyhR4/edit?usp=sharing

Column name	datatype	
competition_id	string	
pretty_name	string	
type	string	
sub_type	string	
country_id	integer	If country_id is less than zero then mark that record as bad record
country_name	string	
country_latitude	decimal	
country_longitude	decimal	
domestic_league_code	string	
name	string	
confederation	string	
url	string	From 'URL' column generates two more columns Baseurl and end_point_url Example: https://www.transfermarkt.co.uk/fc-reading/startseite/verein/1032 Base_url: https://www.transfermarkt.co.uk/ End_point_url:

		fc-reading/startseite/verein/1032

4. Players

Dataset Link:
https://docs.google.com/spreadsheets/d/14_H4c9tgjxuz2Ua-YY5zJRx2XoBg4Pp2oitPlGx2r8c/edit?usp=sharing

Column name	datatype	
player_id		
pretty_name		
club_id		
club_pretty_name		
current_club_id		
country_of_citizenship		
country_of_birth		
city_of_birth		
date_of_birth		
position		
sub_position		
name		
foot		
height_in_cm		Add one more column with the name “height_in_feet” convert “height_in_cm” value to feet
market_value_in_gbp		If values are not present then market it as zero ‘0’
highest_market_value_in_gbp		If values are not present then market it as zero ‘0’
agent_name		

contract_expiration_date		IF null then put date as future date 2999-12-31
domestic_competition_id		
club_name		
image_url		
last_season		
url		

5. Games

Dataset Link:
https://docs.google.com/spreadsheets/d/12fGXEu19YCgrh2vSDBYtCt9-oC0dcpafJ-UFLpcxVsE/edit?usp=sharing

Column name	string	
game_id		
competition_id		
competition_type		
season		
round		
date		
home_club_id		
away_club_id		
home_club_goals		
away_club_goals		
aggregate		
home_club_position		
away_club_position		
club_home_pretty_name		

club_away_pretty_name		
home_club_manager_name		
away_club_manager_name		
stadium		
attendance		
referee		
url		

6. club

Dataset link:
https://docs.google.com/spreadsheets/d/1omLwPpHYuf65BqwlrEGmUsgx8MKQniEy2K9Unf78aKE/edit?usp=sharing

Column name	string	Transformation logic
club_id		
name		Update '-' to '_' Replace other special character with ''(with nothing)
pretty_name		
domestic_competition_id		
total_market_value		
squad_size		
average_age		
foreigners_number		
foreigners_percentage		
national_team_players		
stadium_name		

stadium_seats		
net_transfer_record		Convert pounds to rupees Convert '+-0' to '-1'
coach_name		
url		From 'URL' column generates two more columns Baseurl and end_point_url Example: https://www.transfermarkt.co.uk/fc-reading/startseite/verein/1032 Base_url: https://www.transfermarkt.co.uk/ End_point_url: fc-reading/startseite/verein/1032

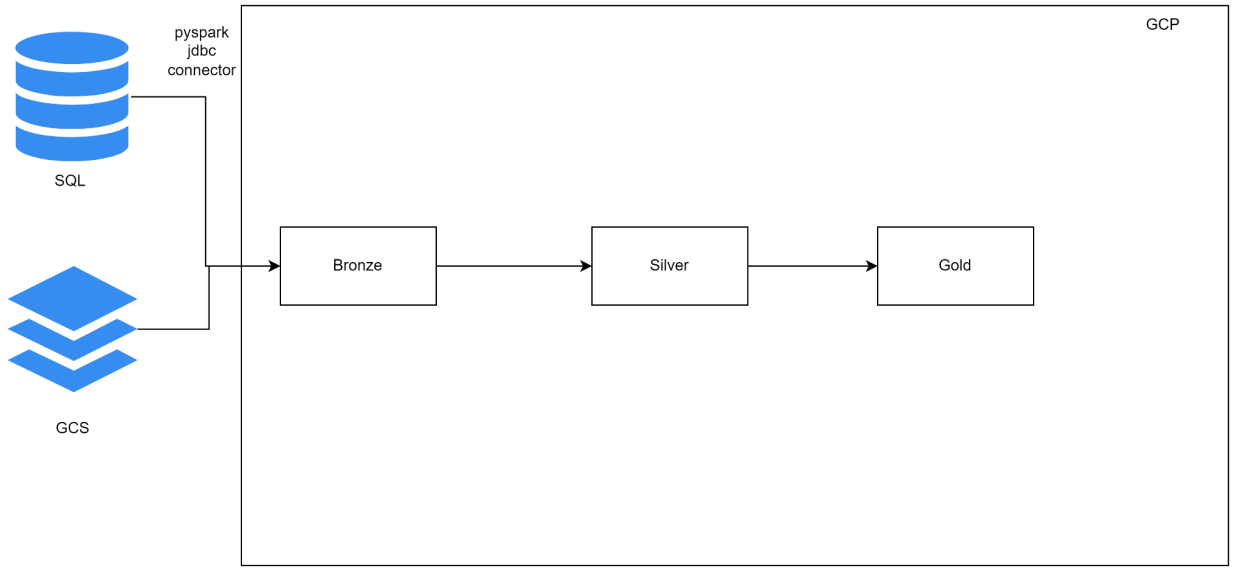
7. appearance

Column name		
appearance_id		
game_id		
player_id		
player_club_id		
date		
player_pretty_name		
competition_id		
yellow_cards		
red_cards		
goals		
assists		
minutes_played		

Load Type: Incremental-Append

Dataset name	Source type	Load Type	
Player valuations	GCS	full	
Club games	GCS	full	
competitions	Cloud SQL	full	
Players	Cloud SQL	full	
games	GCS	full	
club	Cloud SQL	full	
appearance	GCS	incremental	

High Level Data Flow



Dataset Location:
gs://football_analysis/<layer_name>/<dataset name>/<file_name>.parquet

- All dataset should follow parquet file format.

Gold layer: Transformation - Generate below tables in Bigquery
Dataset name: gold_db

1. Player club details - Combine player and club dataset
2. Details club games table -
3. Player market valuation - combine player and player valuation
4. Games stats for club - combine club with game dataset

Scenario Question to Do in Pyspark, DataFlow, and Bigquery

1. Find the average market value of all players for each club in a specific year. How has each player's market value changed compared to the average of their club over time?
2. Find the top10 players with the highest market value
3. Calculate winning percentage of club after their manager change
4. Compute the total market value of all clubs within each domestic competition and identify the competition with the highest total value.
5. Group players based on their age like 18-25, 26-30, 31-35,
    a. Calculate average goals per age group
    b. Calculate average assists per age group
    c. Minutes played per games per age group
6. Do clubs with more foreign players tend to win the game?
7. Calculate goal contribution rate for each player -
8. Calculate average market value for the club