ETL PROJECT: FOOTBALL DATA ANALYTICS

MISBAH BIN HOSSAIN

DATA ENGINEER STUDENT AT NACKADEMIN

ETL WORKFLOW OVERVIEW

The ETL workflow consists of:

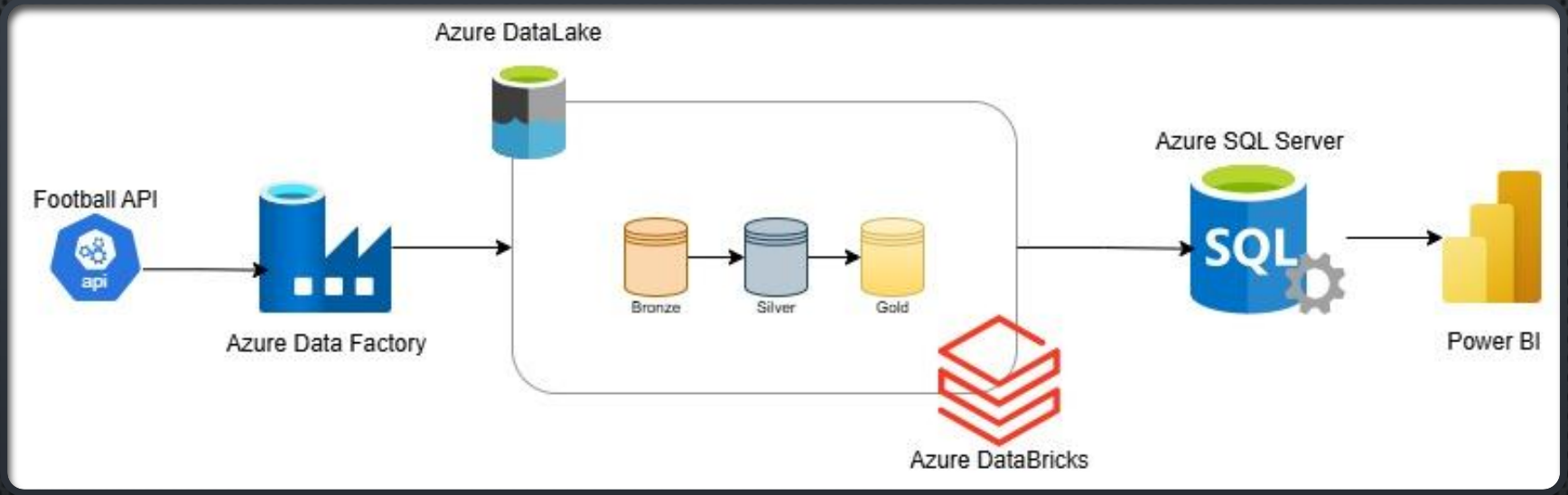
1. Data Extraction

2. Data Transformation

3. Data Loading

4. Data Visualization

Each phase is supported by different Azure and data engineering tools.



ETL WORKFLOW

FOOTBALL API

- Website - https://www.football-data.org/
- My Parameters:
  - Filter = **["matches","standings","teams","scorers"]**
  - Season= [2022","2023","2024"]
  - League = ["PL","SA","PD","BL1"]

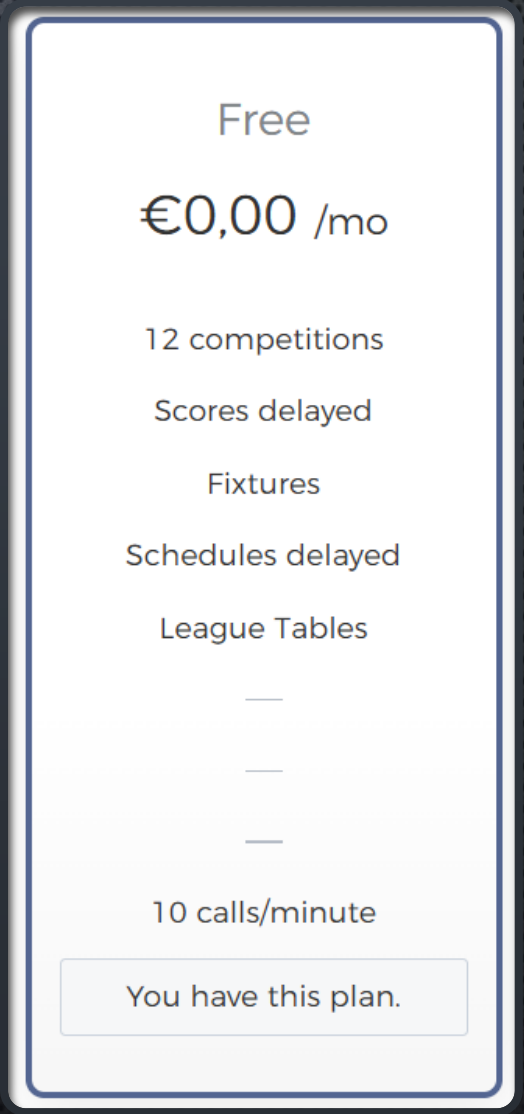
- Api call example:

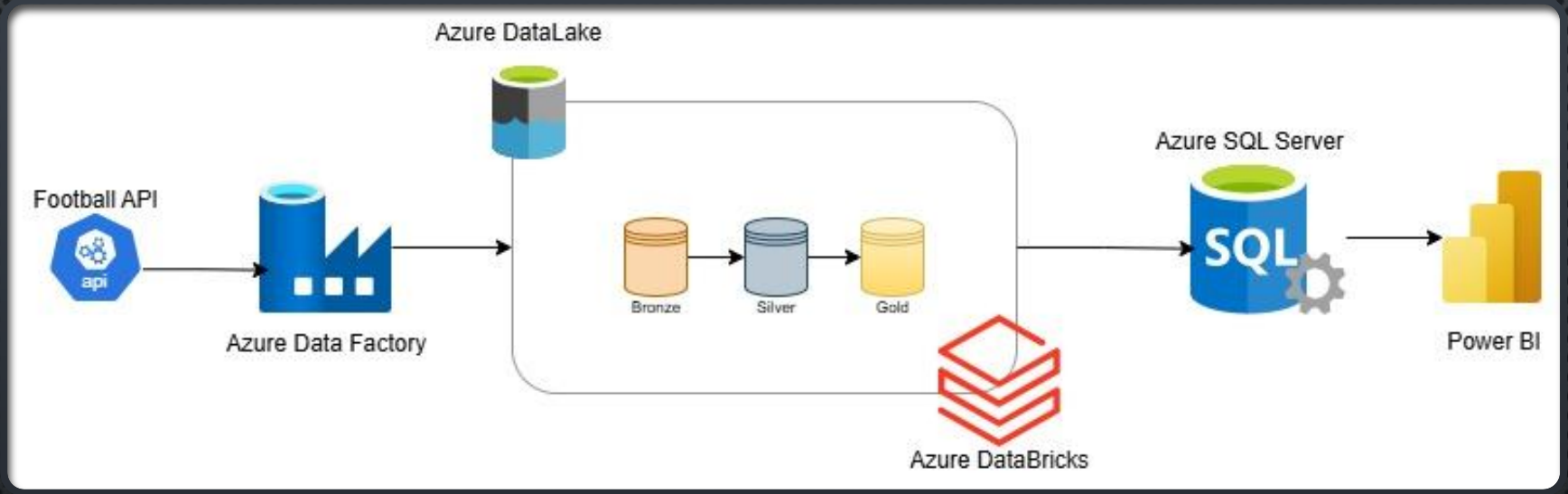
  https://api.football-data.org/v4/competitions/**PL/matches/?season=2024**

  https://api.football-data.org/v4/competitions/**PD/teams/?season=2023**

  https://api.football-data.org/v4/competitions/**SA/scorers/?season=2022**

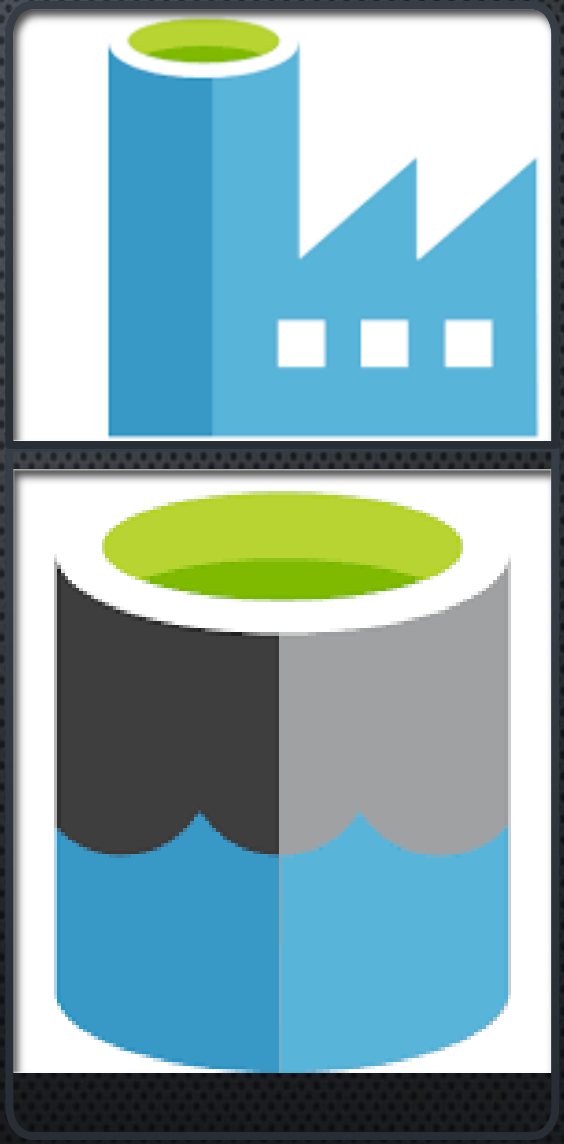
  https://api.football-data.org/v4/competitions/**BL1/standings/?season=2023**





ETL WORKFLOW

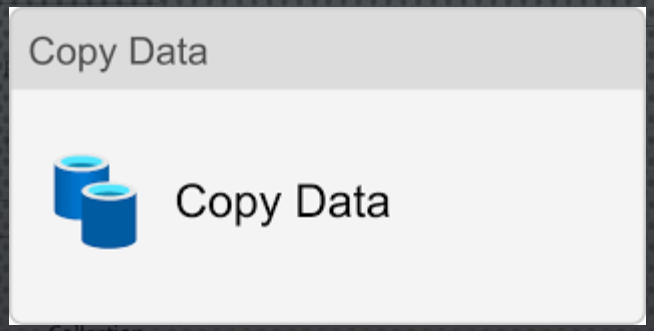
DATA EXTRACTION



Tools: Azure Data Factory, Azure Blob Storage

- - Dynamic pipelines fetch football data from API
- - Handles multiple parameters (season, league, filter types)
- - Data stored in Parquet format in Azure Blob Storage.

EXAMPLE OF COPY ACTIVITY



General | Source | Sink | Mapping | Settings | User properties

Source dataset *: Dynamic_import | Open | New | Preview data | Learn more

Dataset properties

Name	Value	Type
Filter	@item().filter	string
Season	@item().season	string
League	@item().league	string

General | Source | Sink | Mapping | Settings | User properties

Sink dataset *: Standing | Open | New | Learn more

Dataset properties

Name	Value	Type
Fil	@item().filter	string
Sea	@item().season	string
lea	@item().league	string

Copy behavior: Flatten hierarchy

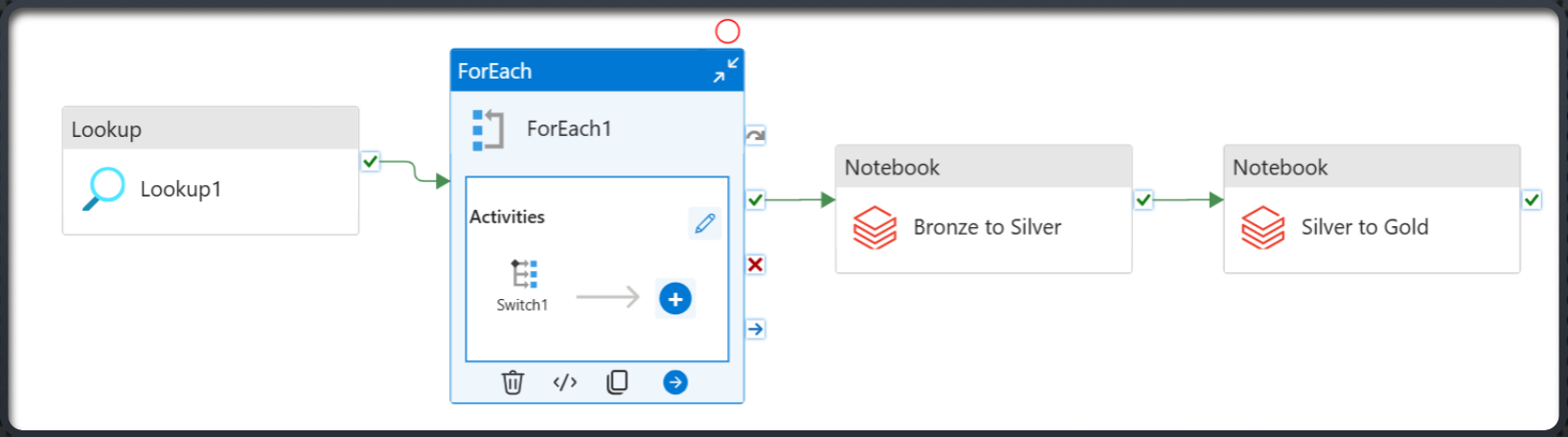
General | Source | Sink | Mapping | Settings | User properties

Import schemas | New mapping | Clear | Delete | Advanced editor

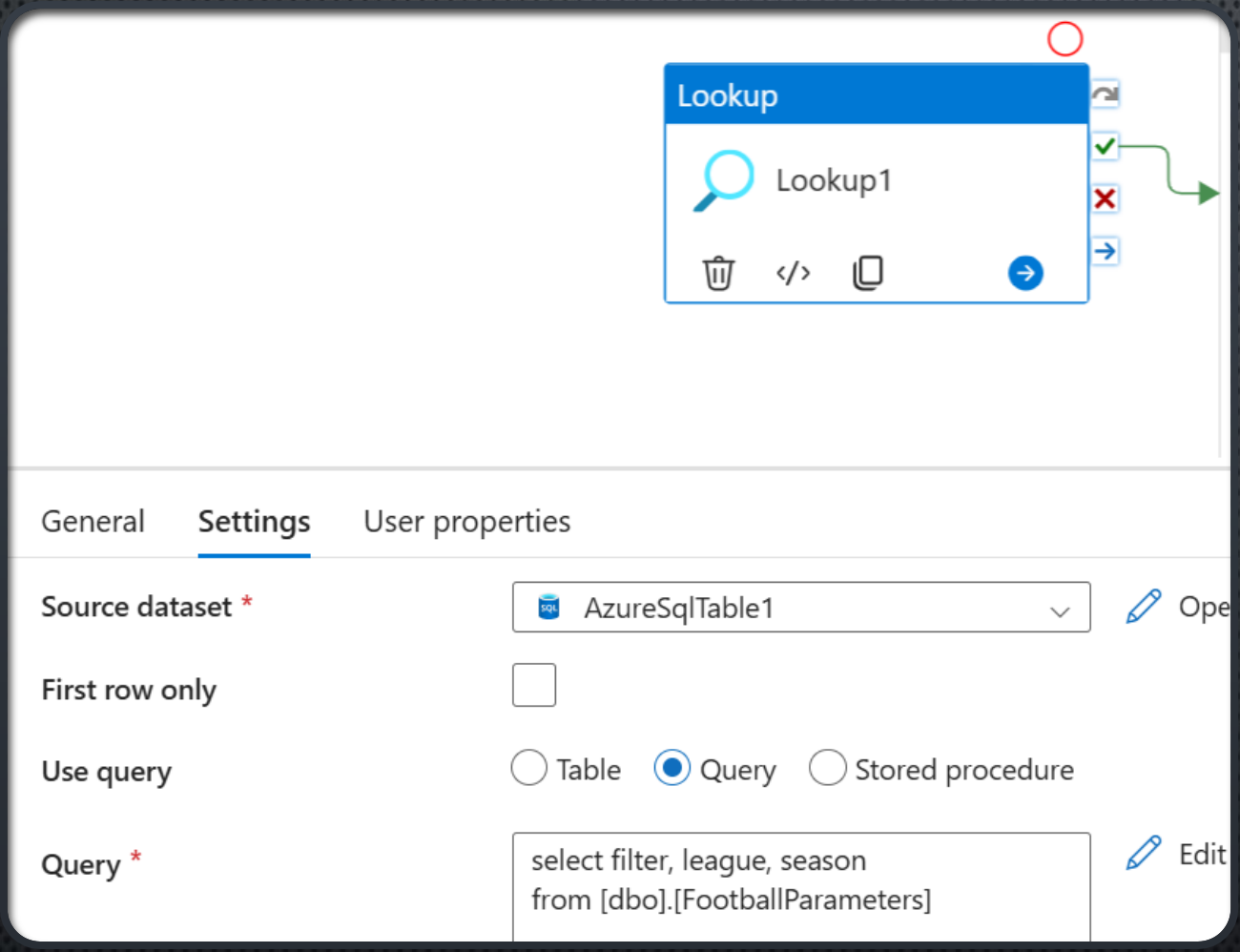
Collection reference: $['standings'][0]['table']

Map complex values to string

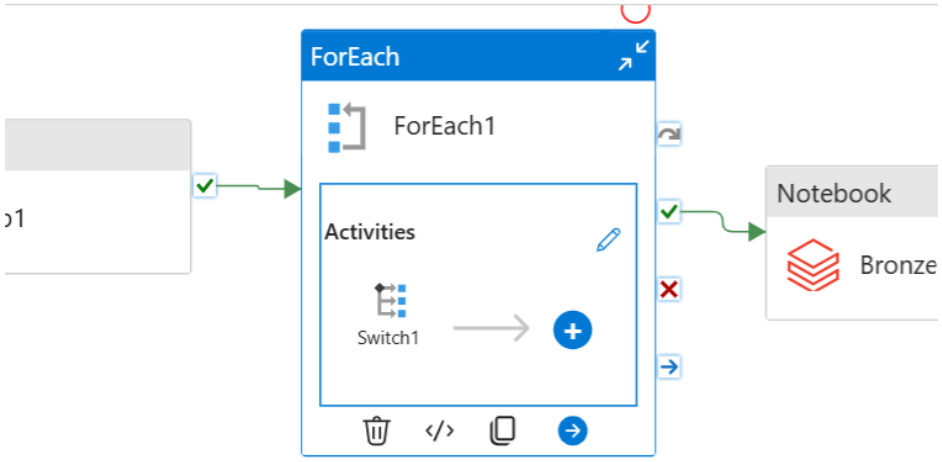
Name	Type	Collection reference	Column name	Type	Include
standings	[] array				
table	[] array	☑			



ADF PIPELINE



LOOKUP ACTIVITY



General | **Settings** | Activities (1) | User properties

Sequential ☑

Items @activity('Lookup1').output.value

FOREACH ACTIVITY

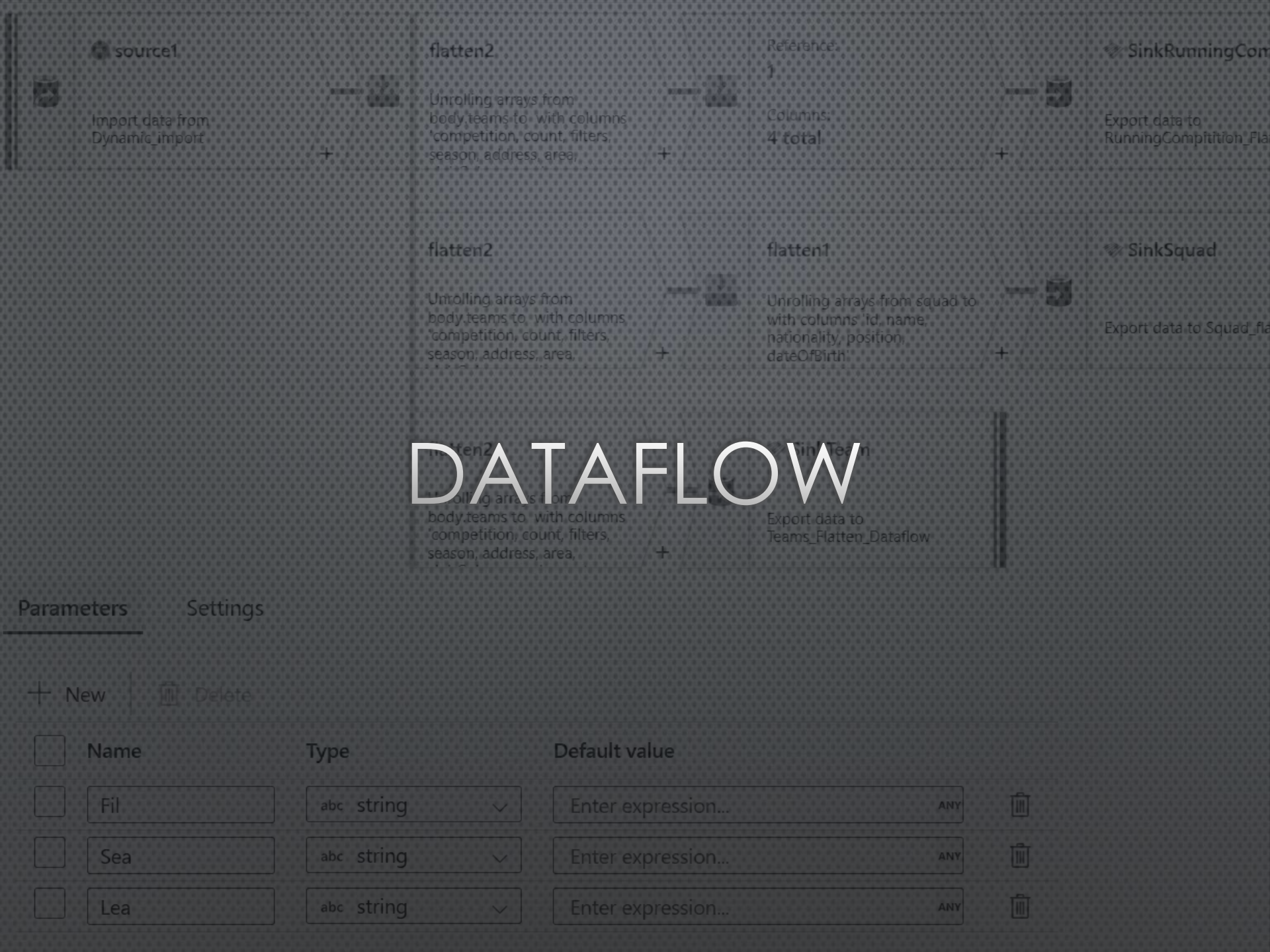
SWITCH ACTIVITY

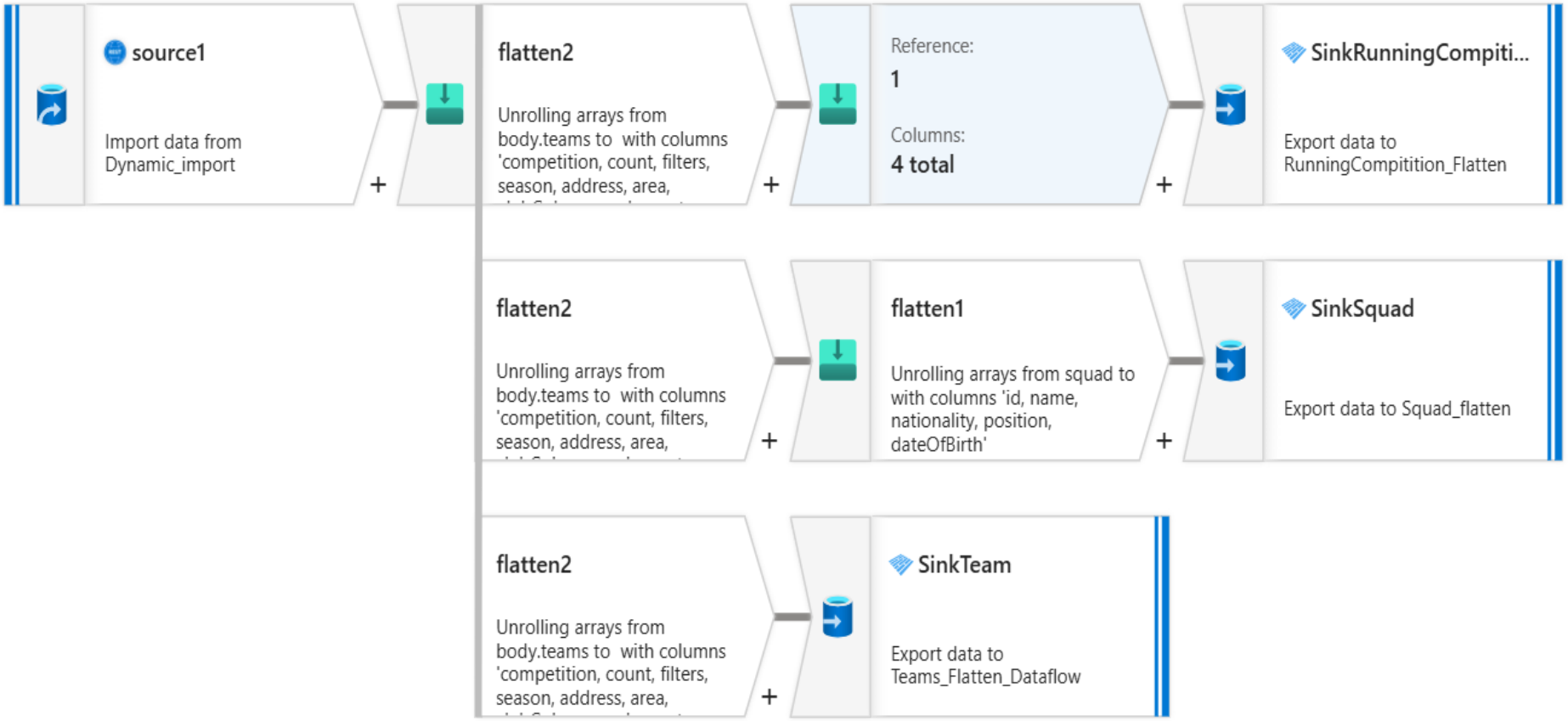
General | **Activities (8)** | User properties

Expression ⓘ @item().filter

+ Add case

Case ⓘ	Activity
Default	*No activities*
standings	Wait_Standing, Copy data -Stan… — 2 Activities
scorers	Wait_Scoreres, Copy data -Top … — 2 Activities
teams	Wait_Teams, Dataflow_Teams — 2 Activities
matches	Wait_Matches, Copy data -Mat… — 2 Activities

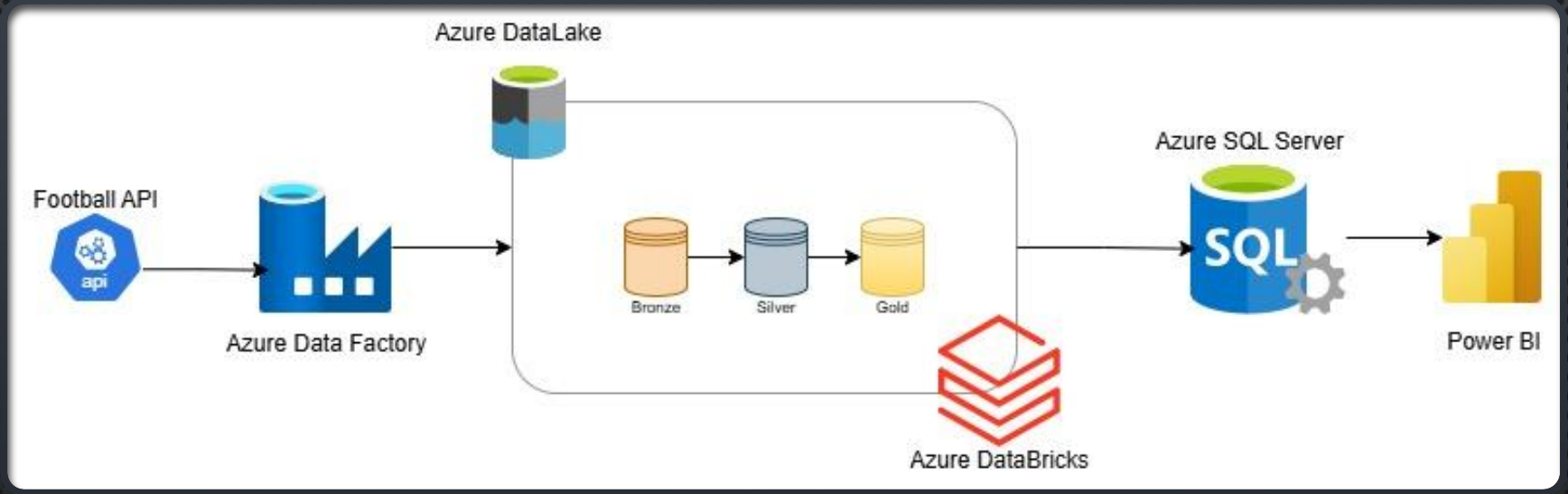

source1
Import data from Dynamic_import
flatten2
Unrolling arrays from body.teams to 'with columns 'competition, count, filters, season, address, area,
Referance
}
Columns:
4 total
SinkRunningCom
Export data to RunningCompetition_Fla
flatten2
Unrolling arrays from body.teams to 'with columns 'competition, count, filters, season, address, area,
flatten1
Unrolling arrays from squad to with columns 'id, name, nationality, position, dateOfBirth'
SinkSquad
Export data to Squad_fla
DATAFLOW
Unrolling arrays from body.teams to 'with columns 'competition, count, filters, season, address, area,
Export data to Teams_Flatten_Dataflow
Parameters
Settings
New
Delete
Name
Type
Default value
Fil
string
Enter expression...
ANY
Sea
string
Enter expression...
ANY
Lea
string
Enter expression...
ANY



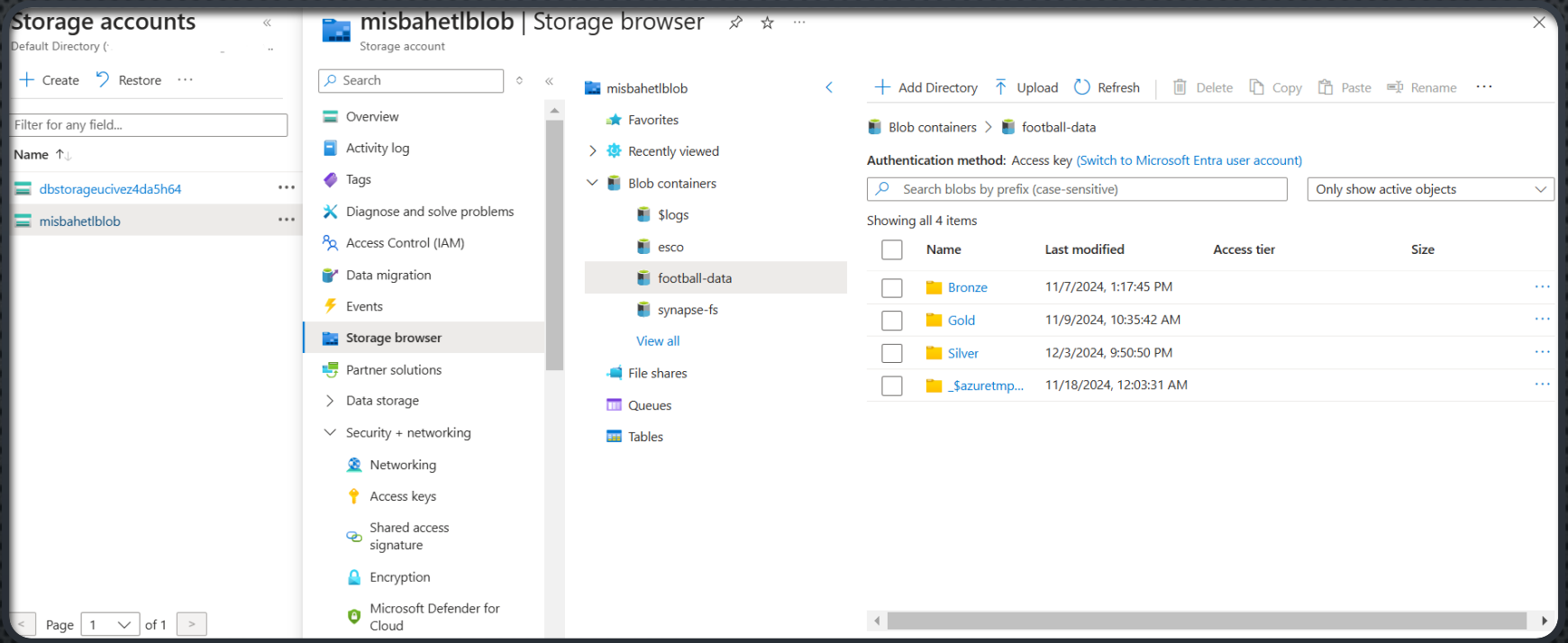
Parameters | Settings

+ New | Delete

Name	Type	Default value
Fil	abc string	Enter expression... ANY
Sea	abc string	Enter expression... ANY
Lea	abc string	Enter expression... ANY



ETL WORKFLOW

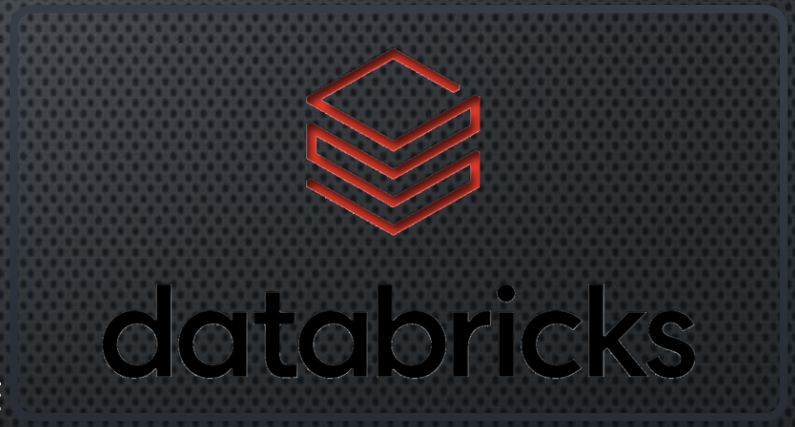

Storage accounts
Default Directory (
Create
Restore
Filter for any field...
Name
dbstorageucivez4da5h64
misbahetlblob
Page 1 of 1
misbahetlblob | Storage browser
Storage account
Search
Overview
Activity log
Tags
Diagnose and solve problems
Access Control (IAM)
Data migration
Events
Storage browser
Partner solutions
Data storage
Security + networking
Networking
Access keys
Shared access signature
Encryption
Microsoft Defender for Cloud
misbahetlblob
Favorites
Recently viewed
Blob containers
$logs
esco
football-data
synapse-fs
View all
File shares
Queues
Tables
Add Directory
Upload
Refresh
Delete
Copy
Paste
Rename
Blob containers > football-data
Authentication method: Access key (Switch to Microsoft Entra user account)
Search blobs by prefix (case-sensitive)
Only show active objects
Showing all 4 items
Name
Last modified
Access tier
Size
Bronze
11/7/2024, 1:17:45 PM
Gold
11/9/2024, 10:35:42 AM
Silver
12/3/2024, 9:50:50 PM
_$azuretmp...
11/18/2024, 12:03:31 AM

AZURE BLOB

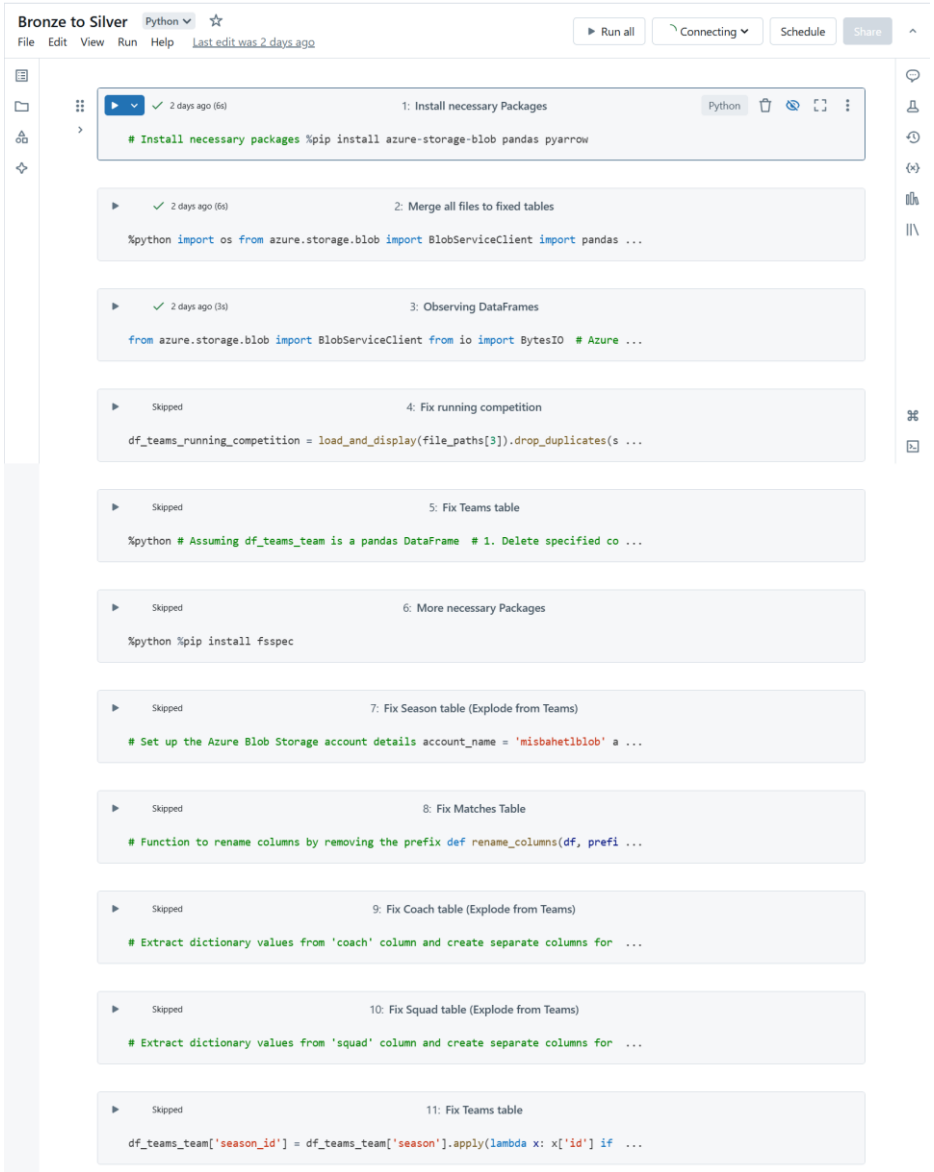
Data Transformation

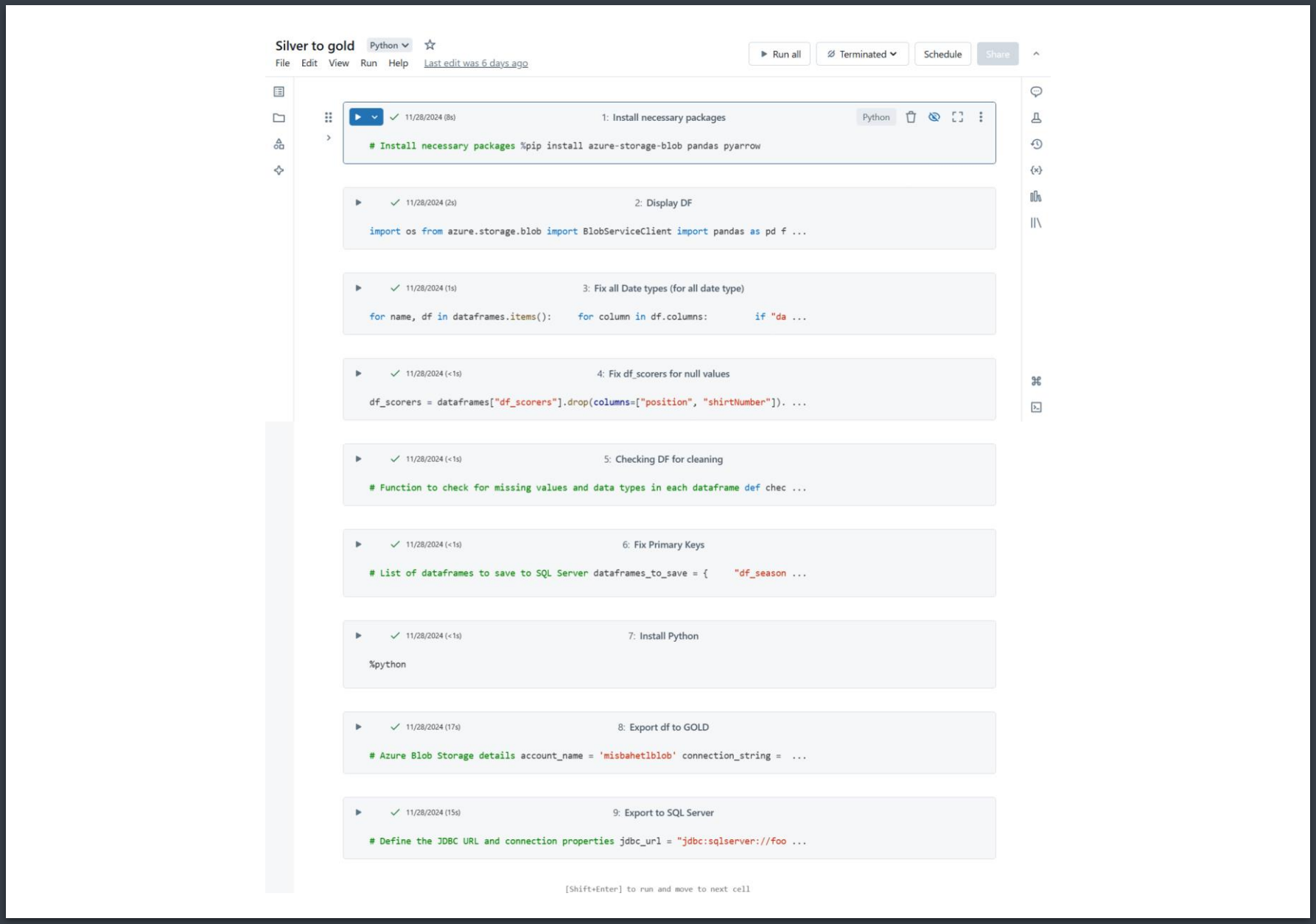
Tools: Azure Databricks (PySpark)

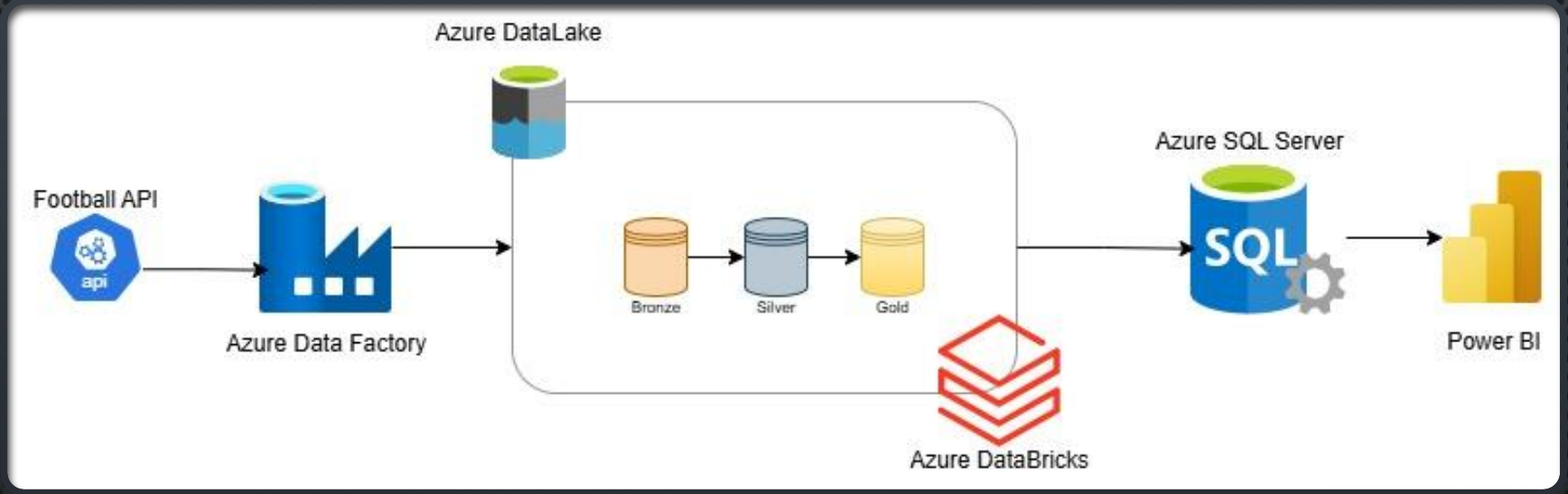
Process:
- **Bronze to Silver**: Clean and normalize raw data
- **Silver to Gold**: Aggregate and prepare data for insights



Outcome:
High-quality, analysis-ready datasets.


Bronze to Silver
Python
File Edit View Run Help Last edit was 2 days ago
Run all
Connecting
Schedule
Share
2 days ago (6s)
1: Install necessary Packages
Python
Install necessary packages %pip install azure-storage-blob pandas pyarrow
2 days ago (6s)
2: Merge all files to fixed tables
%python import os from azure.storage.blob import BlobServiceClient import pandas ...
2 days ago (3s)
3: Observing DataFrames
from azure.storage.blob import BlobServiceClient from io import BytesIO # Azure ...
Skipped
4: Fix running competition
df_teams_running_competition = load_and_display(file_paths[3]).drop_duplicates(s ...
Skipped
5: Fix Teams table
%python # Assuming df_teams_team is a pandas DataFrame # 1. Delete specified co ...
Skipped
6: More necessary Packages
%python %pip install fsspec
Skipped
7: Fix Season table (Explode from Teams)
Set up the Azure Blob Storage account details account_name = 'misbahetlblob' a ...
Skipped
8: Fix Matches Table
Function to rename columns by removing the prefix def rename_columns(df, prefi ...
Skipped
9: Fix Coach table (Explode from Teams)
Extract dictionary values from 'coach' column and create separate columns for ...
Skipped
10: Fix Squad table (Explode from Teams)
Extract dictionary values from 'squad' column and create separate columns for ...
Skipped
11: Fix Teams table
df_teams_team['season_id'] = df_teams_team['season'].apply(lambda x: x['id'] if ...


Silver to gold
Python
File Edit View Run Help
Last edit was 6 days ago
Run all
Terminated
Schedule
Share
11/28/2024 (8s)
1: Install necessary packages
Python
Install necessary packages %pip install azure-storage-blob pandas pyarrow
11/28/2024 (2s)
2: Display DF
import os from azure.storage.blob import BlobServiceClient import pandas as pd f ...
11/28/2024 (1s)
3: Fix all Date types (for all date type)
for name, df in dataframes.items(): for column in df.columns: if "da ...
11/28/2024 (<1s)
4: Fix df_scorers for null values
df_scorers = dataframes["df_scorers"].drop(columns=["position", "shirtNumber"]). ...
11/28/2024 (<1s)
5: Checking DF for cleaning
Function to check for missing values and data types in each dataframe def chec ...
11/28/2024 (<1s)
6: Fix Primary Keys
List of dataframes to save to SQL Server dataframes_to_save = { "df_season ...
11/28/2024 (<1s)
7: Install Python
%python
11/28/2024 (17s)
8: Export df to GOLD
Azure Blob Storage details account_name = 'misbahetlblob' connection_string = ...
11/28/2024 (15s)
9: Export to SQL Server
Define the JDBC URL and connection properties jdbc_url = "jdbc:sqlserver://foo ...
[Shift+Enter] to run and move to next cell

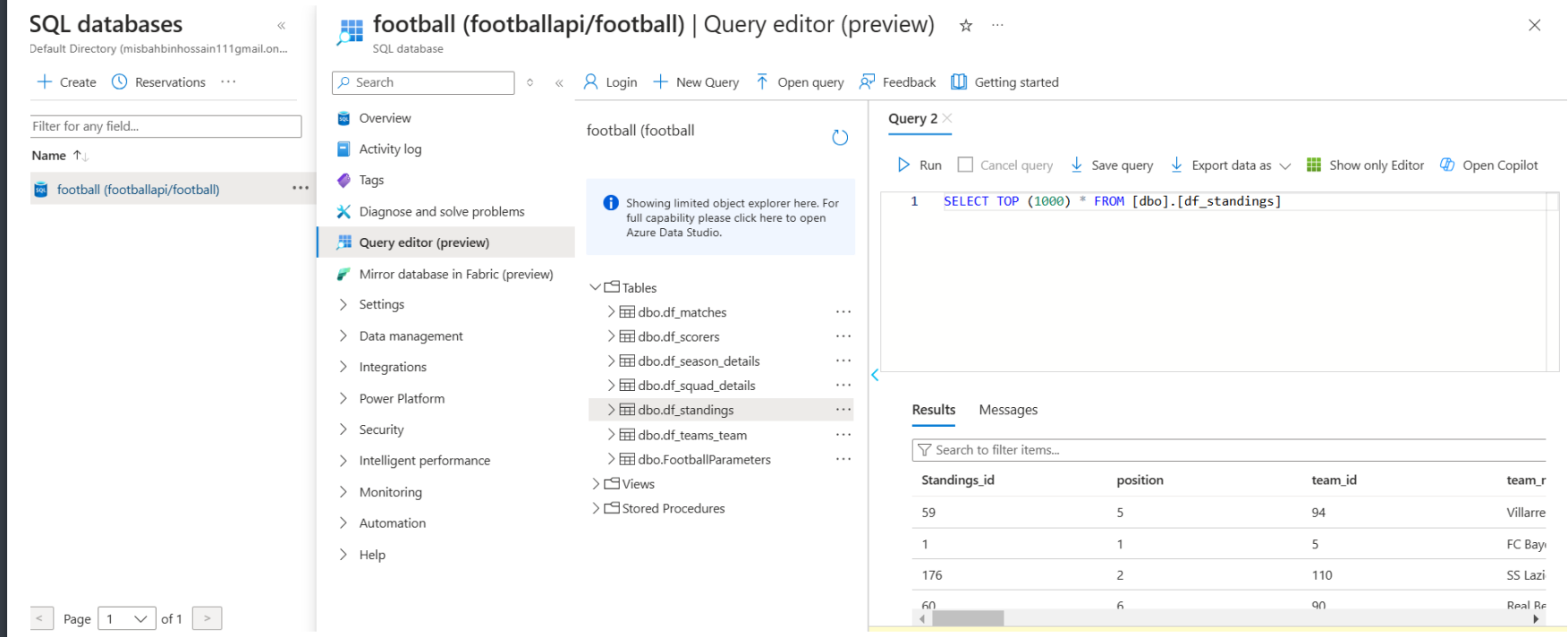


ETL WORKFLOW

DATA LOADING

TOOLS: AZURE DATABRICKS, SQL SERVER

- - TRANSFORMED DATA LOADED INTO SQL SERVER VIA JDBC
- - ENSURES EFFICIENT STORAGE AND ACCESS FOR ANALYSIS.


SQL databases
Default Directory (misbahbinhossain111gmail.on...
Create
Reservations
Filter for any field...
Name
football (footballapi/football)
Page 1 of 1
football (footballapi/football) | Query editor (preview)
SQL database
Search
Overview
Activity log
Tags
Diagnose and solve problems
Query editor (preview)
Mirror database in Fabric (preview)
Settings
Data management
Integrations
Power Platform
Security
Intelligent performance
Monitoring
Automation
Help
Login
New Query
Open query
Feedback
Getting started
football (football
Showing limited object explorer here. For full capability please click here to open Azure Data Studio.
Tables
dbo.df_matches
dbo.df_scorers
dbo.df_season_details
dbo.df_squad_details
dbo.df_standings
dbo.df_teams_team
dbo.FootballParameters
Views
Stored Procedures
Query 2
Run
Cancel query
Save query
Export data as
Show only Editor
Open Copilot
1 SELECT TOP (1000) * FROM [dbo].[df_standings]
Results
Messages
Search to filter items...
Standings_id
position
team_id
team_r
59
5
94
Villarre
1
1
5
FC Bay
176
2
110
SS Lazi
60
6
90
Real Be

DATA VISUALIZATION

Tool: Power BI

- - Team standings and rankings
- - Match results and summaries
- - Top scorer statistics
- - Squad-level details

Football Data Analysis

34	5	379	20
Max of points	Min of points	Total Goals	Total Teams

Competition Name

- [ ] Bundesliga
- [x] Premier League
- [ ] Primera Division
- [ ] Serie A

Season

- [ ] 2022
- [ ] 2023
- [x] 2024

14

Current Matchday

Position	Team Name	Points	PG	Won	Lost	Draw	GF	GA	GD	FormIcons
1	Liverpool FC	34	13	11	1	1	26	8	18	✓ ✓ ✓ ✓ □
2	Arsenal FC	25	13	7	2	4	26	14	12	✓ ✓ □ ✗ □
3	Chelsea FC	25	13	7	2	4	26	14	12	✓ ✓ □ □ ✓
4	Brighton & Hove Albion FC	23	13	6	2	5	22	17	5	□ □ □ □ □
5	Manchester City FC	23	13	7	4	2	22	19	3	✗ ✗ ✗ ✗ ✓
6	Nottingham Forest FC	22	13	6	3	4	16	13	3	✓ ✗ ✗ ✓ ✓
7	Tottenham Hotspur FC	20	13	6	5	2	28	14	14	□ □ □ □ □
8	Brentford FC	20	13	6	5	2	26	23	3	✓ □ ✓ ✗ ✓
9	Manchester United FC	19	13	5	4	4	17	13	4	✓ □ ✓ □ ✗
10	Fulham FC	19	13	5	4	4	18	18	0	□ □ □ □ □
11	Newcastle United FC	19	13	5	4	4	14	14	0	□ □ □ □ □
12	Aston Villa FC	19	13	5	4	4	19	22	-3	✗ □ ✗ ✗ □
13	AFC Bournemouth	18	13	5	5	3	20	19	1	✓ ✗ ✗ ✓ □
14	West Ham United FC	15	13	4	6	3	17	24	-7	✗ ✓ □ ✗ ✓
15	Everton FC	11	13	2	6	5	10	21	-11	✗ □ □ ✗ □
16	Leicester City FC	10	13	2	7	4	16	27	-11	✗ ✗ ✗ □ ✗

Football Data Analysis

Season
- 2022
- 2023
- 2024

Competition Name
- Bundesliga
- Premier League
- Primera Division
- Serie A

Nationality
- Select all
- Albania
- Algeria
- Andorra
- Angola

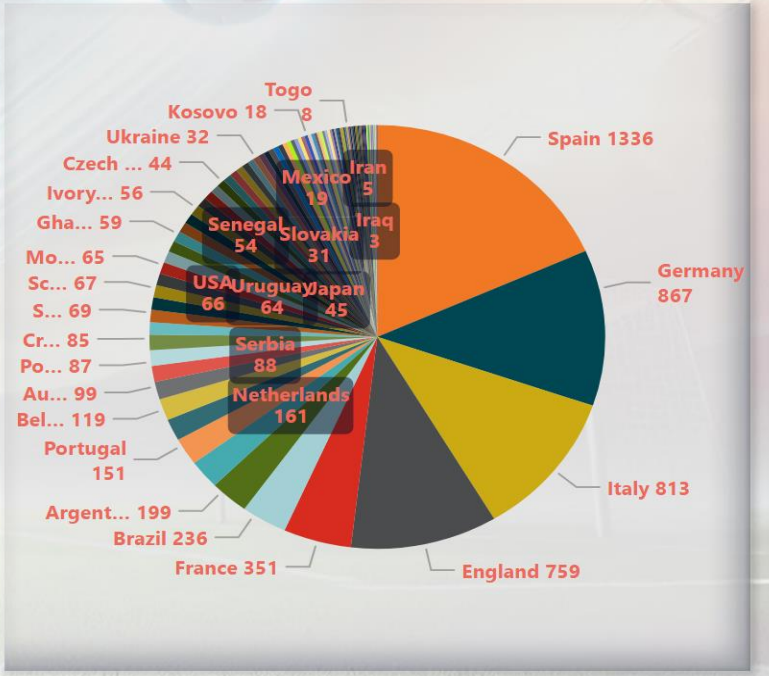
TeamName
- 1. FC Heidenheim 1846
- 1. FC Köln
- 1. FC Union Berlin
- 1. FSV Mainz 05
- AC Milan
- AC Monza
- ACF Fiorentina
- AFC Bournemouth
- Arsenal FC
- AS Roma
- Aston Villa FC
- Atalanta BC
- Athletic Club
- Bayer 04 Leverkusen
- Bologna FC 1909
- Borussia Dortmund
- Borussia Mönchengladba

name	Age	nationality	position	
Aarón	29	Spain	Goalkeeper	Ul
Aaron Ciammaglichella	19	Italy	Central Midfield	Tc
Aaron Cresswell	35	England	Left-Back	W
Aaron Hickey	22	Scotland	Right-Back	Br
Aaron Keto-Diyawa	21	England	Left-Back	W
Aaron Martín	27	Spain	Left-Back	1.
Aaron Martín	27	Spain	Left-Back	Ge
Aaron Ramsdale	26	England	Goalkeeper	Ar
Aaron Ramsdale	26	England	Goalkeeper	Sc
Aaron Ramsey	21	England	Attacking Midfield	Bu
Aaron Seydel	28	Germany	Offence	SV
Aaron Wan-Bissaka	27	England	Right-Back	M
Aaron Wan-Bissaka	27	England	Right-Back	W
Aaron Zehnter	20	Germany	Left Midfield	FC

26.01
Average of Age

112
No. of Nationality

1336
Max Nationality



Football Data Analysis

Teams

- Select all
- Brentford FC
- Brighton & Hove Albion FC

Match Day

- 1
- 2
- 3

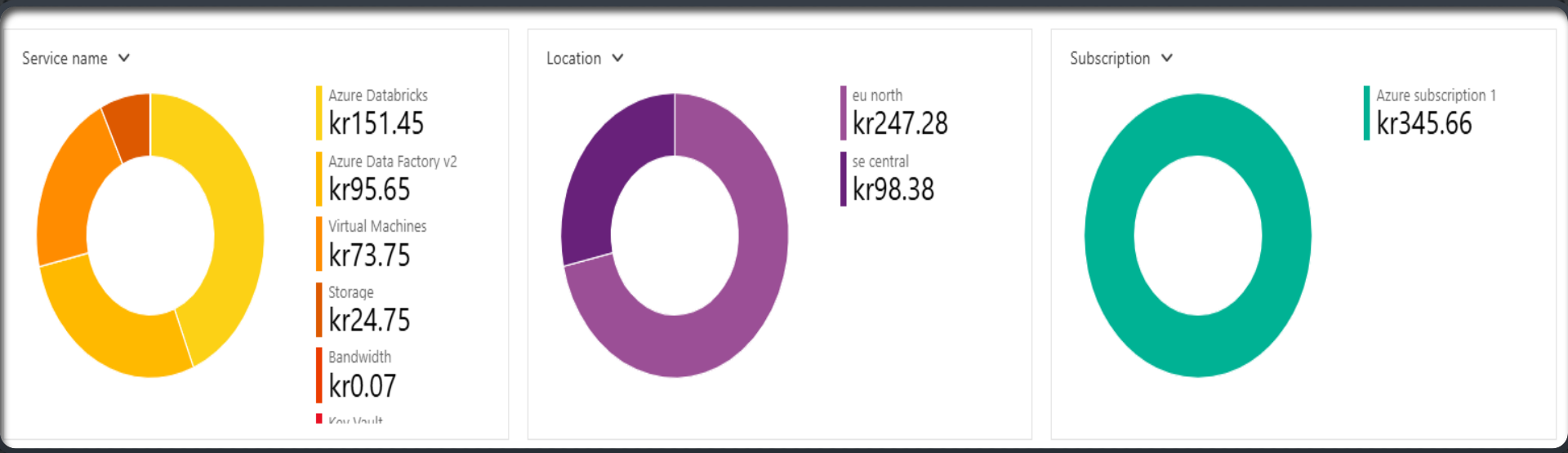
Home Team Name	Score Home	Game Time	Score Away	Away Team Name
Chelsea FC	3	12/1/2024 1:30:00 PM	0	Aston Villa FC
Liverpool FC	2	12/1/2024 4:00:00 PM	0	Manchester City FC
Manchester United FC	4	12/1/2024 1:30:00 PM	0	Everton FC
Nottingham Forest FC	1	11/30/2024 3:00:00 PM	0	Ipswich Town FC
Brentford FC	4	11/30/2024 3:00:00 PM	1	Leicester City FC
Brighton & Hove Albion FC	1	11/29/2024 8:00:00 PM	1	Southampton FC
Crystal Palace FC	1	11/30/2024 3:00:00 PM	1	Newcastle United FC
Tottenham Hotspur FC	1	12/1/2024 1:30:00 PM	1	Fulham FC
Wolverhampton Wanderers FC	2	11/30/2024 3:00:00 PM	4	AFC Bournemouth
West Ham United FC	2	11/30/2024 5:30:00 PM	5	Arsenal FC

Season

- 2022
- 2023
- 2024

Competition Name

- Bundesliga
- Premier League
- Primera Division
- Serie A



COST ANALYSIS

DEMO TIME

Conclusion: Applying Skills in the ETL Project

- **Database Design and SQL**: Applied advanced querying and modeling techniques for efficient data structuring in SQL Server.
- **Python Programming**: Used PySpark in Databricks for data cleaning and transformation.
- **DevOps**: Automated workflows and ensured reliability using Azure Data Factory and cloud tools.
- **Cloud Solutions**: Utilized Azure platforms for data storage, processing, and management.
- **ETL/ELT Expertise**: Designed a robust pipeline for API integration, data migration, and transformation.
- **System Understanding and Business Alignment**: Delivered actionable insights via a Power BI dashboard tailored to business needs.
- This project demonstrates my ability to integrate skills from the program into solving real-world data challenges.

THANK YOU FOR YOU ATTENTION

QUESTAINS ?