Scraping transfermarkt
a Data Pipeline Journey

Michele Tassoni 28-05-2021

Agenda

- What is transfermarkt
- Project motivation and description
- Project architecture
- Data stack
- Deployment
- Challenges
- Outlook

transfermarkt.com

- Website with historical footballing information.
- Popular for estimating player values alongside other football information like scores, results, transfer news, etc.
- More importantly for this project it provides information about football players like age, height, etc.
- It is in the to 25 most visited websites in Germany.

Project motivation

Passion for football

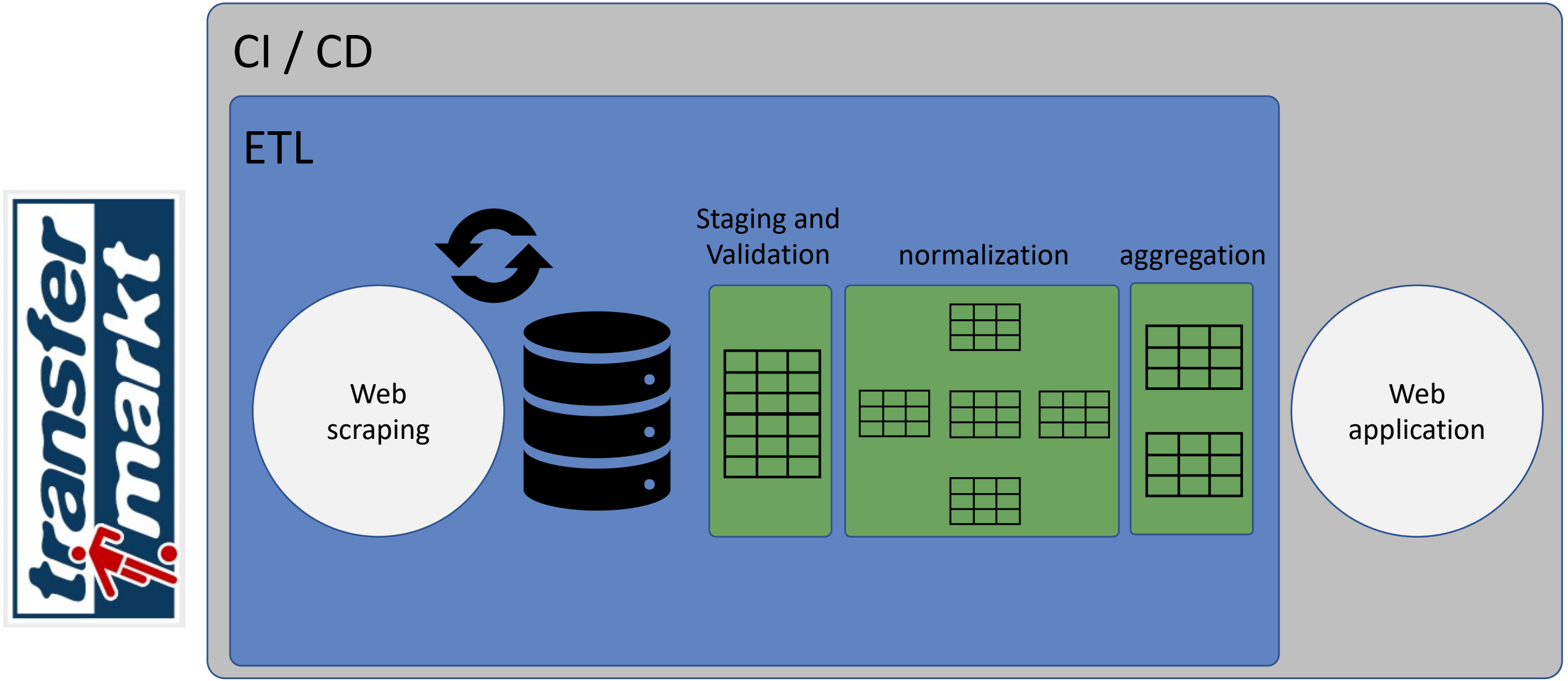
Learning about web scraping

Source of data that allows transformation and visualization

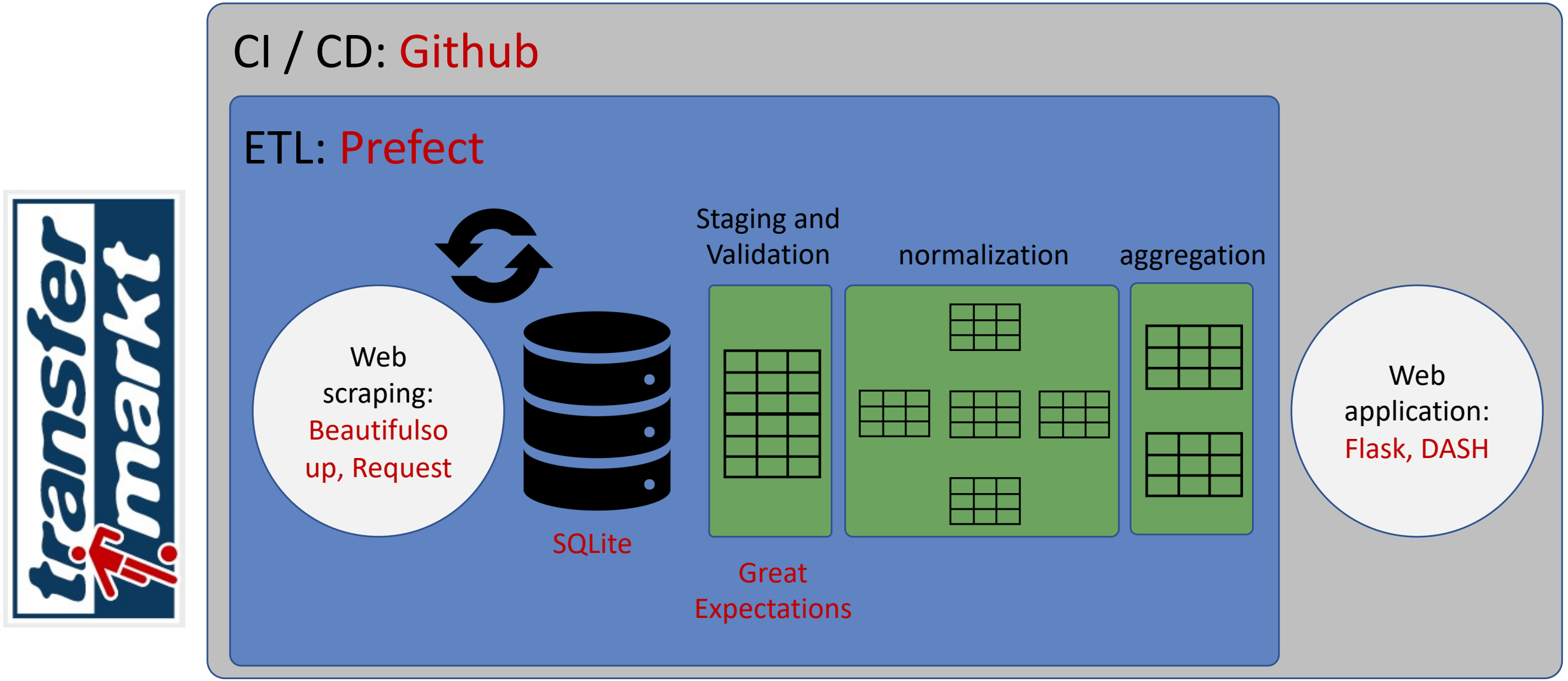
Project description

- Build a data product that:
  - Automatically collects the last 50 years of players data from the top five European football competitions
  - Stores the data into an SQL database
  - Use the data to build a dashboard that shows data aggregation results

Project architecture and data stack



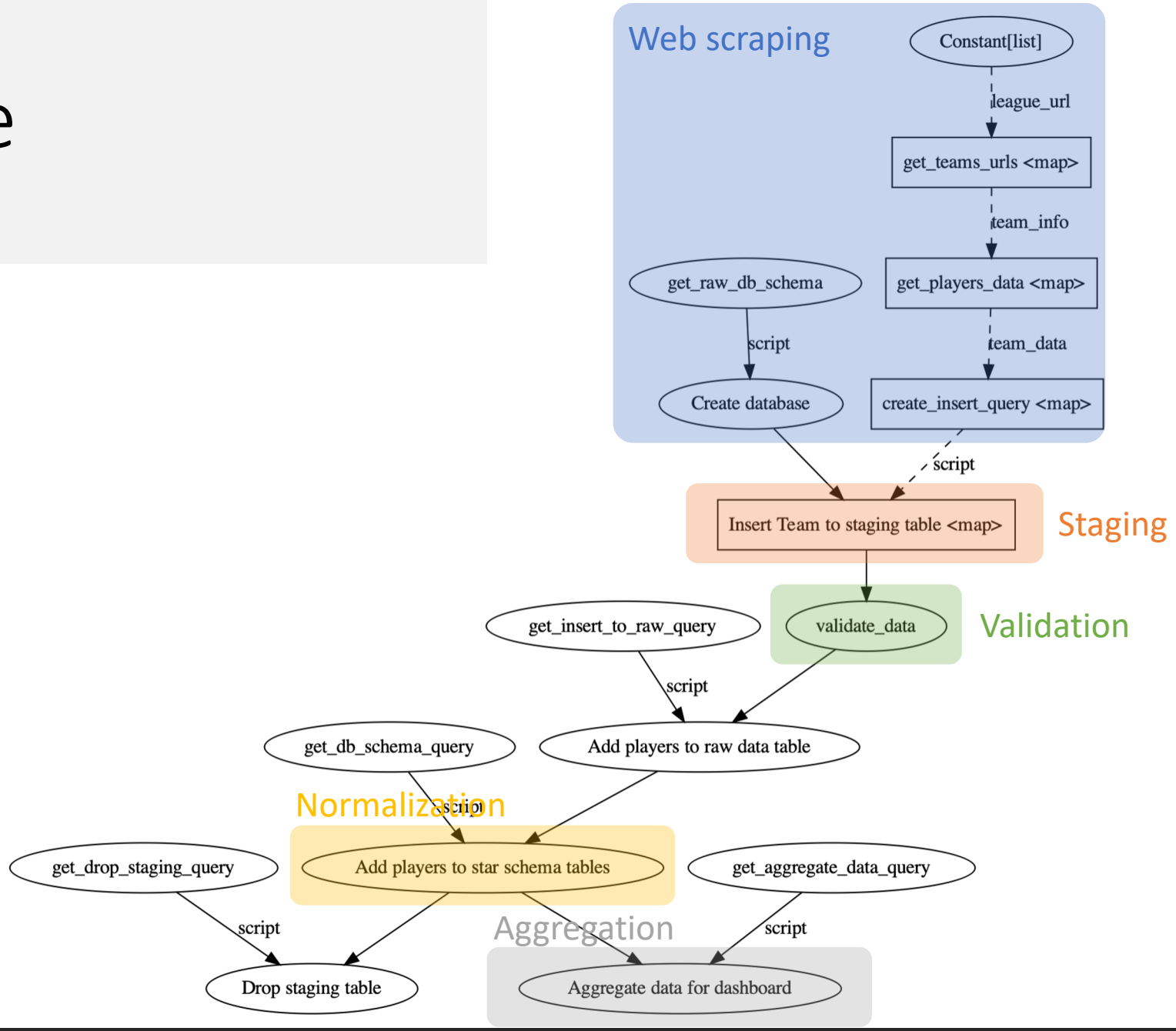
Project architecture and data stack



Prefect pipeline

It runs in two modes:

- “Populate database”
- “Update database” (scheduled)

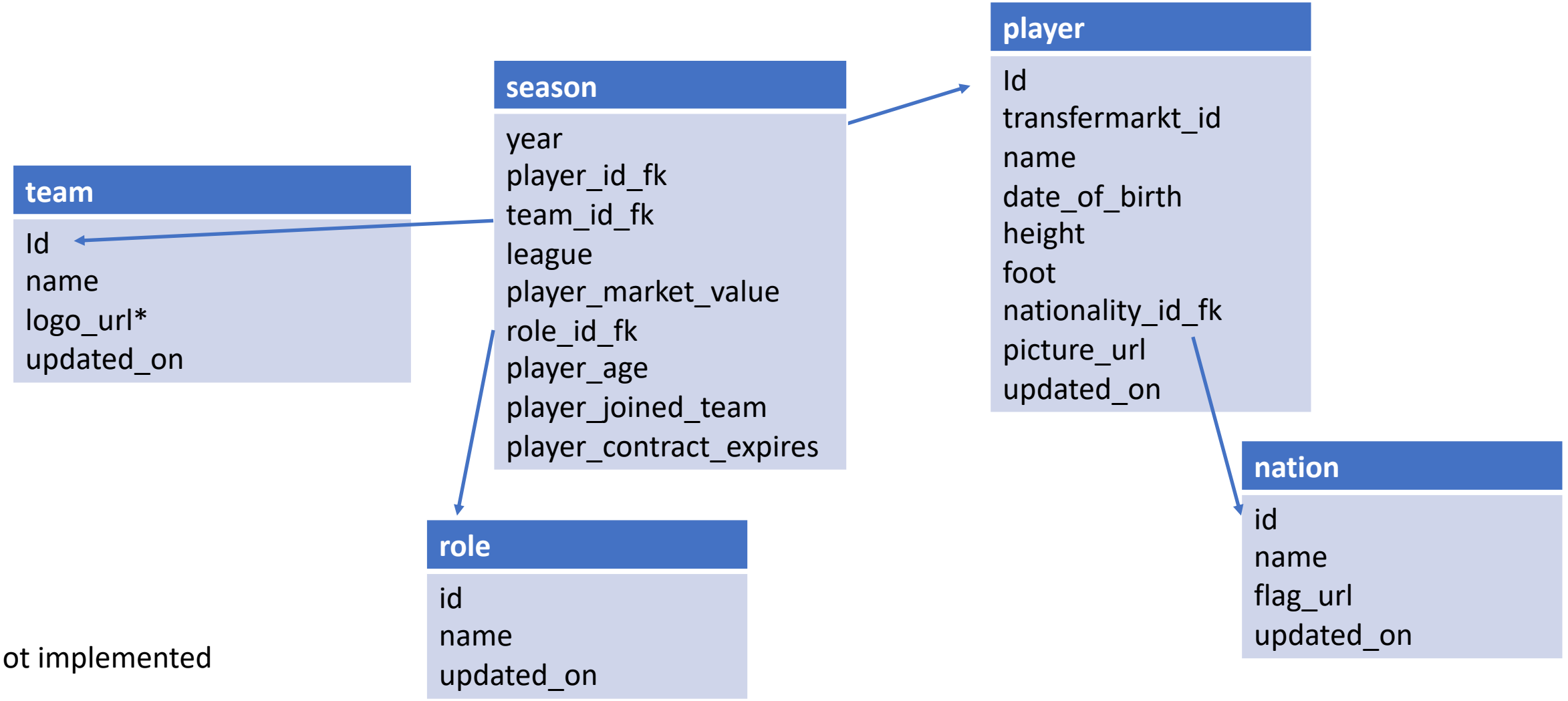


Raw data and data validation

Players
name
team
league
role
date_of_birth
age
height
foot
joined
contract_expires
market_value
nationality
nation_flag_url
player_picture_url
updated_on
season

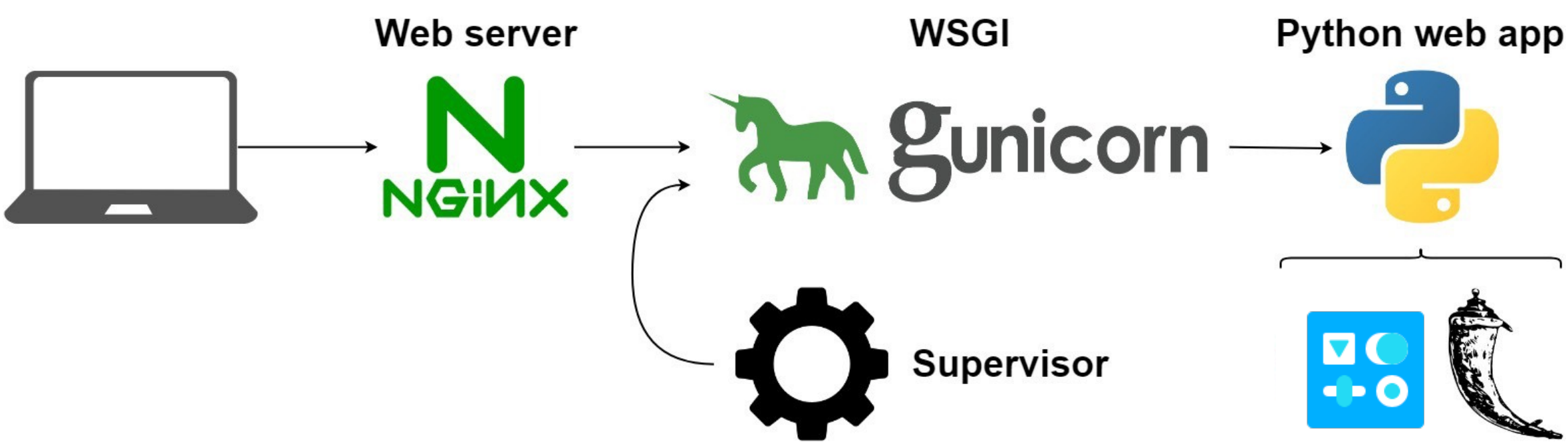
- Column names and types
- Column order
- Min and Max number of entries
- Min and Max values in each column
- Amount of missing values

Database schema

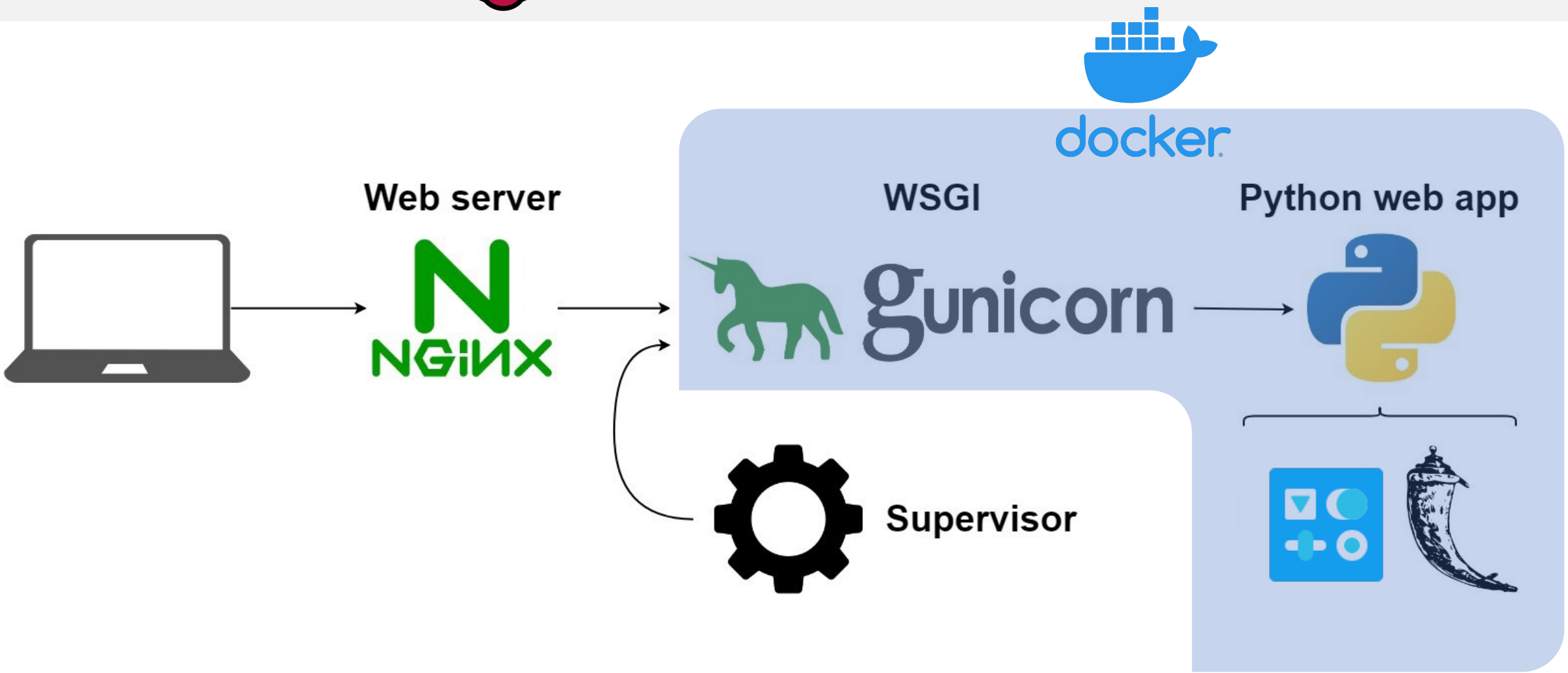


* = Not implemented

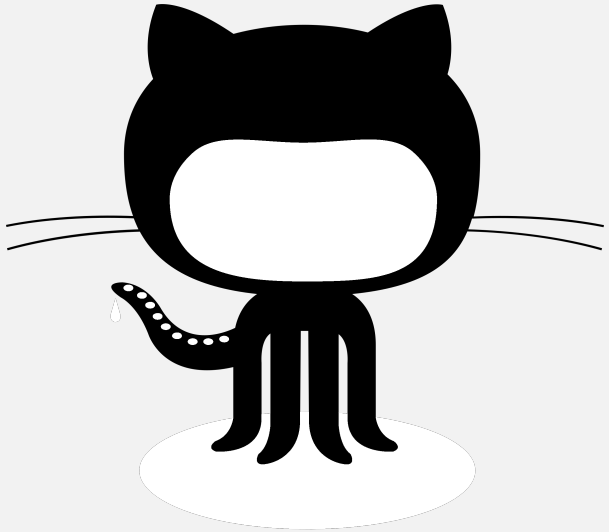
Deployment



Deployment

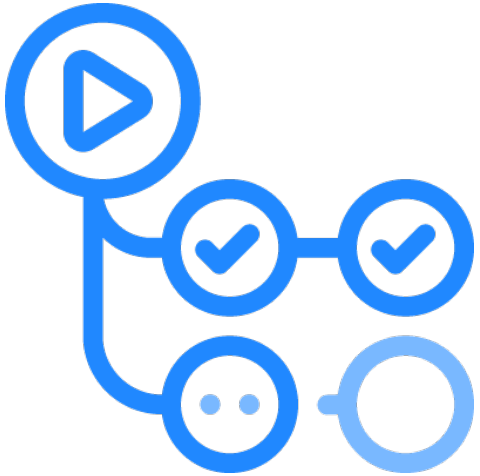

docker
Web server
WSGI
Python web app
gunicorn
NGINX
Supervisor

CI / CD



webhooks

The webhook is triggered when new changes are pushed to the main branch and directly deployed to the production server.
Gunicorn ensures server restart to apply the new version of the code.



A Github action is triggered to validate pool requests by running tests and linting.

Challenges

- Deployment on PythonAnywhere was the first choice
  - Over 500 Mb of virtual environment (Only Great Expectations takes >400 Mb)!

- Installing Virtual environment on Raspberry PI
  - Great expectations requires Scipy that is impossible to build on the Pi
  - Using PiWheels to install precompiled python packages

- Setting up a private Webserver

Summary

- How to develop an automated data pipeline that scrapes data from the internet to serve a web application for data visualization
- How to validate and normalize data for efficient database storage and querying
- How to setup a private Webserver
- How do define CI / CD tools using Github

Outlook

- Using the Prefect server to schedule scraping pipeline
- Add integration tests for the Prefect pipeline and the connection to transfermarkt
- Take care of the dashboard updates once the new data lands in the database
- Extract more information from the data and improve dashboard quality

Thank you for your attention

Visit my Github: https://github.com/mrvaita/footballers_value