

# Wykorzystanie bazy dokumentowej MongoDB

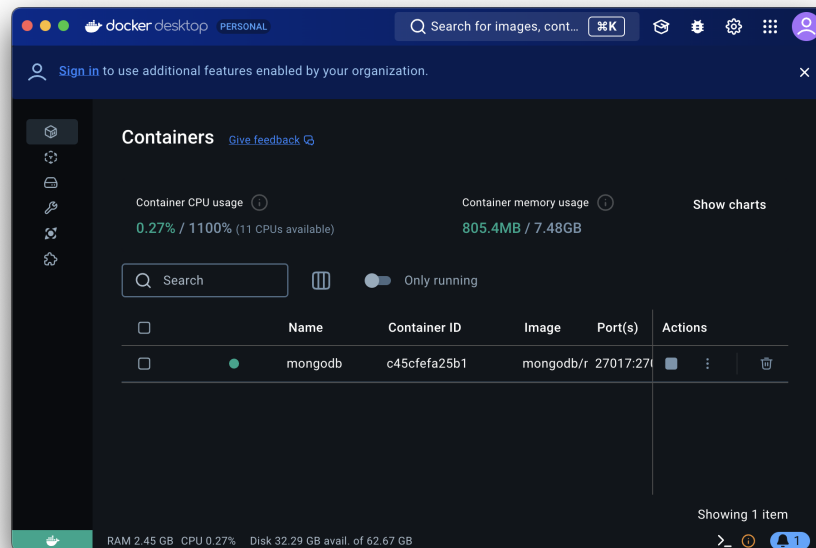
## Zbiór danych

<https://www.kaggle.com/datasets/davidcariboo/player-scores>

Zbiór danych dotyczący meczy piłkarskich. Każdy mecz ma wiele zdarzeń, co pozwala na zagnieżdżenie. Zdarzenia mają sporo cech. Zbiór jest dosyć spory (miliony rekordów).

## Instalacja

Baza została zainstalowana lokalnie na moim laptopie jako kontener Dockerowy.



## Upload danych do bazy

Napisałem program w Python, który umieszcza dane w bazie:

```
import os

import kagglehub
import pandas as pd
from pymongo import MongoClient
from tqdm import tqdm

path = kagglehub.dataset_download("davidcariboo/player-scores")
```

```

clubs = pd.read_csv(os.path.join(path, "clubs.csv"))
game_events = pd.read_csv(os.path.join(path, "game_events.csv"))
game_lineups = pd.read_csv(os.path.join(path, "game_lineups.csv"))
games = pd.read_csv(os.path.join(path, "games.csv"))
player_valuations = pd.read_csv(os.path.join(path, "player_valuations.csv"))
players = pd.read_csv(os.path.join(path, "players.csv"))

client = MongoClient("mongodb://admin:password@localhost:27017/")
assert client.server_info()["ok"] == 1

db = client["football_data"]
db.list_collection_names()

clubs_collection = db["clubs"]
clubs_collection.insert_many(clubs.to_dict(orient="records"))

players_collection = db["players"]
players["valuations"] = (
    player_valuations.groupby("player_id")
    .apply(lambda x: x.to_dict(orient="records"))
    .to_dict()
)
players_collection.insert_many(players.to_dict(orient="records"))

for game in tqdm(games.iterrows(), total=games.shape[0]):
    game_row = game[1]
    lineup = game_lineups[game_lineups["game_id"] == game_row["game_id"]]
    events = game_events[game_events["game_id"] == game_row["game_id"]]
    game_dict = game_row.to_dict()
    game_dict["lineup"] = lineup.to_dict(orient="records")
    game_dict["events"] = events.to_dict(orient="records")
    db["games"].insert_one(game_dict)

```

