BigData - project2

Platforma: Flink (DataStream API)

Zestaw 1 – Netflix-Prize-Data

Instrukcja uruchamiania projektu:

Przygotowania:

- pobierz dane dla zestawu 1 movie_titles.csv (https://www.cs.put.poznan.pl/kjankiewicz/bigdata/stream_project/movie_titles.csv) oraz netflix-prize-data (https://www.cs.put.poznan.pl/kjankiewicz/bigdata/stream_project/netflix-prize-data.zip) i umieść je w folderze `data`
- uruchom projekt w PyCharm
- stwórz nowy venv - wejdź w prawym dolnym rogu w `Python [nr wersji]`, następnie wybierz `add new interpreter` i potem `add local interpreter`, wybierz go i zainstaluj z manage packages `apache-flink==2.0.0` i `pymongo==4.13.0`
- uruchamiając skrypty wybieraj ten venv
- zainstaluj potrzebne biblioteki
- w celu zainstalowania wszystkich wymaganych pakietów możesz skorzystać z przygotowanego pliku `requirements.txt`
- podczas uruchamiania projektu korzystaj z kontenerów `FlinkAndFriends2025`, upewnij się, że wskazany parametr ma poprawną wartość w pliku `docker-compose.yml`:

```
KAFKA_ADVERTISED_LISTENERS: PLAINTEXT://localhost:9092
```

```
docker compose -p fandf up -d
```

- utwórz temat producenta w kontenerze Kafki:

```
docker exec -it fandf-kafka-1 /bin/bash
```

```
/opt/kafka/bin/kafka-topics.sh --create --bootstrap-server kafka:9092 \
 --replication-factor 1 --partitions 3 --topic netflix
```

- ustaw wartości poszczególnych parametrów skryptu `kafka-producer.py` zgodnie ze swoją konfiguracją:

```
CSV_FOLDER = 'data\\netflix-prize-data'
KAFKA_BOOTSTRAP_SERVERS = 'localhost:9092'
KAFKA_TOPIC = 'netflix'
```

- uruchom skrypt `kafka_producer.py` i obserwuj temat producenta, czy został zasilony

```
/opt/kafka/bin/kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic netflix --from-beginning
```

Spodziewany format wiadomości w temacie:

```
2000-01-29,7397,1428659,3
2000-01-29,7397,562156,3
2000-01-29,7397,417988,4
```

Wykrywanie anomalii:

- utwórz temat, do którego trafią wykryte anomalie:

```
/opt/kafka/bin/kafka-topics.sh --create --bootstrap-server kafka:9092 \
--replication-factor 1 --partitions 3 --topic netflix-anomalies
```

- pobierz potrzebne pliki `.jar` zgodnie z tutorialem do laboratoriów z Flinka, zwróć uwagę na plik `flink.properties` i dostosuj poszczególne propsy, w szczególności zwróć uwagę na: `static.file.path` oraz `pipeline.jars`
- uruchom skrypt `netflix_data_anomalies_detection.py` wraz z wybranymi parametrami D, L oraz O podanymi w run configuration IDE

- po chwili uruchom skrypt konsumenta z tematu odbiorczego kafki `kafka_consumer.py` i obserwuj wyniki, upewniając się wcześniej czy ma on poprawne parametry:

```
KAFKA_BOOTSTRAP_SERVERS = 'localhost:9092'
KAFKA_TOPIC = 'netflix-anomalies'
```

Przykładowa część wyniku dla parametrów `D:30 L:20 O:4.3`:

```
Nasłuchiwanie wiadomości z tematu: netflix-anomalies
Odebrano: {"window_start": "1999-12-09T01:00:00", "window_end": "2000-01-08T01:00:00", "title": "Apollo 13", "count": 20, "avg_rate"
Odebrano: {"window_start": "1999-12-10T01:00:00", "window_end": "2000-01-09T01:00:00", "title": "Die Hard", "count": 25, "avg_rate":
Odebrano: {"window_start": "1999-12-10T01:00:00", "window_end": "2000-01-09T01:00:00", "title": "The Matrix", "count": 32, "avg_rate
Odebrano: {"window_start": "1999-12-11T01:00:00", "window_end": "2000-01-10T01:00:00", "title": "October Sky", "count": 22, "avg_rat
Odebrano: {"window_start": "1999-12-11T01:00:00", "window_end": "2000-01-10T01:00:00", "title": "The Terminator", "count": 20, "avg_
```

ETL – obraz czasu rzeczywistego:

- utwórz kontener MongoDB:

```
docker run -d -p 27017:27017 --name mongodb mongo
```

- sprawdź parametry dotyczące MongoDB w pliku `flink.properties`
- uruchom skrypt `netflix_data_ETL_analysis.py` z wybranym parametrem delay (wartość A lub C) podanym w run configuration IDE

- po chwili uruchom skrypt `mongodb_reader.py` i sprawdź wyniki w kolekcji MongoDB

Przykładowy fragment wyniku dla trybu `A`:

```
{"_id": "683ad97fbffce212c887074d", "id": 4883, "title": "The Bodyguard", "month": "1999-12", "count_rate": 26, "sum_rate": 72.0, "u
{"_id": "683ad97fbffce212c887074f", "id": 4883, "title": "The Bodyguard", "month": "1999-12", "count_rate": 27, "sum_rate": 76.0, "u
{"_id": "683ad97fbffce212c8870751", "id": 4883, "title": "The Bodyguard", "month": "1999-12", "count_rate": 28, "sum_rate": 81.0, "u
```

Przykładowy fragment wyniku dla trybu `C`:

```
{"_id": "683ada53a0d9acc9c30e23f5", "id": 15940, "title": "Box of Moonlight", "month": "1999-12", "count_rate": 4, "sum_rate": 17.0,
{"_id": "683ada53a0d9acc9c30e23f7", "id": 9654, "title": "Very Bad Things", "month": "1999-12", "count_rate": 18, "sum_rate": 39.0, "
{"_id": "683ada53a0d9acc9c30e23f9", "id": 13787, "title": "Powder", "month": "1999-12", "count_rate": 29, "sum_rate": 97.0, "unique_
```

Restart środowiska:

- usuń tematy Kafki
- wykonaj skrypt `mongodb_restart.py`, aby usunąć kolekcję w MongoDB