

Projekat 1

Sistemi za obradu i analizu velike količine podataka

Izabrani skup podataka

- Oslo City Bikes https://oslobysykkel.no/en/open-data
- Sistem koji dozvoljava da se na određenim stanicama u
 Oslu iznajmi bicikla i da se ostavi na nekoj drugoj stanici
- Naplacuje se po trajanju vožnje
- Ima preko 200 stanica u Oslu na kojima može da se iznajmi bicikla
- Sistem koristi preko 100.000 ljudi u Oslu
- 2018. godine je zabeleženo skoro 3 miliona vožnji

Predmet analize

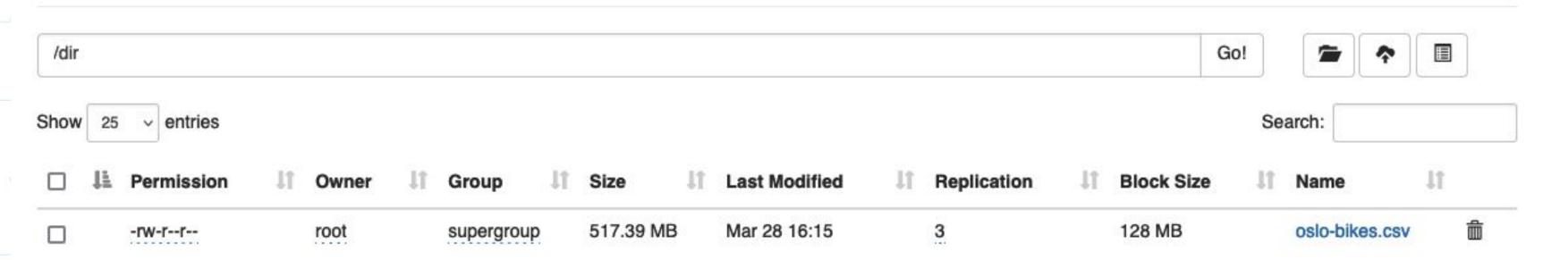
- Analizirani su podaci za celu 2022. godinu
- Fokus analize bio je:
 - Broj vožnji za izabranu polaznu stanicu i za izabrani vremenski period
 - Najpopularnije destinacije u zavisnosti za izabranu polaznu stanicu i izabrani vremenski period
- Statistički parametri za vremensko trajanje vožnji grupisani po polaznoj stanici i po danu u nedelji

Postavljanje podataka na hdfs

Podatke prvo prebacimo na namenode koristeći komande:

docker cp oslo-bikes.csv namenode:/data
 docker exec -it namenode bash
 hdfs dfs -mkdir /dir
hdfs dfs -put /data/oslo-bikes.csv /dir

Browse Directory



Prosleđivanje parametara

Aplikaciji je prilikom startovanja potrebno proslediti sledece parametre u .env fajlu:

```
START_ID=552
START_DATE='2022-01-01T00:00:00'
END_DATE='2022-01-31T00:00:00'
FILE_DATA='oslo-bikes.csv'
HDFS_DATA='hdfs://namenode:9000/dir/oslo-bikes.csv'
```

Broj vožnji za polaznu stanicu u nekom vremenskom periodu

```
data.filter(
data.start_station_id == station_id)
.filter(data.started_at >= from_date &
  data.started_at <= to_date)
.count()</pre>
```

Prvo filtriramo pdoatke tako da dobijemo samo one podatke gde je polazna stanica ista kao i zadata.

Zatim imamo filter sa AND uslovom gde nam je bitno da je vreme polaska u određenom vremenskom intervalu



Statistički podaci za dužinu vožnjr

```
data.groupBy("start_station_name", dayofweek("started_at")).agg(
  min("duration"), max("duration"), stddev("duration"), mean("duration")
)
```

Tražimo minimum, maksimum, standardnu devijaciju kao i prosek za trajanje vožnje a zatim grupišemo po polaznoj stanici i po danu u nedelju (od 1 do)



Performanse aplikacije

Zbog pokretanja klustera na racunaru sa M1 procesorom, posto ne postoji Native image za taj procesor, performanse su slabije nego sto se ocekuje

