



Upravljanje i analiza velikih skupova podataka

Projekat 1 - Apache Spark & HDFS

Luka Gavrilović 1823

Projekat 1

- Analiza kašnjenja letova korišćenjem Apache Spark-a:
 - Batch obrada velikog skupa podataka
 - Python API (PySpark) + Spark na Docker klasteru (BDE)
 - HDFS za skladištenje podataka
- Funkcionalnosti aplikacije:
 - Brojanje i filtriranje podataka po zadatim kriterijumima (npr. mesec, aviokompanija)
 - Statistička analiza atributa (min, max, avg, stddev) grupisanih po kolonama (npr. aviokompanija, aerodrom)

Skup podataka

- 2019 Airline Delays w/Weather and Airport Detail ($\approx 1.37\text{GB}$)
- https://www.kaggle.com/datasets/threnjen/2019-airline-delays-and-cancellations?select=full_data_flightdelay.csv
- Skup podataka sa detaljnim informacijama o avio-kompanijama, aerodromima i vremenskim uslovima

| | |
|--------------------------------|---|
| MONTH: | Month |
| DAY_OF_WEEK: | Day of Week |
| DEP_DEL15: | TARGET Binary of a departure delay over 15 minutes (1 is yes) |
| DISTANCE_GROUP: | Distance group to be flown by departing aircraft |
| DEP_BLOCK: | Departure block |
| SEGMENT_NUMBER: | The segment that this tail number is on for the day |
| CONCURRENT_FLIGHTS: | Concurrent flights leaving from the airport in the same departure block |
| NUMBER_OF_SEATS: | Number of seats on the aircraft |
| CARRIER_NAME: | Carrier |
| AIRPORT_FLIGHTS_MONTH: | Avg Airport Flights per Month |
| AIRLINE_FLIGHTS_MONTH: | Avg Airline Flights per Month |
| AIRLINE_AIRPORT_FLIGHTS_MONTH: | Avg Flights per month for Airline AND Airport |
| AVG_MONTHLY_PASS_AIRPORT: | Avg Passengers for the departing airport for the month |
| AVG_MONTHLY_PASS_AIRLINE: | Avg Passengers for airline for month |
| FLT_ATTENDANTS_PER_PASS: | Flight attendants per passenger for airline |
| GROUND_SERV_PER_PASS: | Ground service employees (service desk) per passenger for airline |
| PLANE_AGE: | Age of departing aircraft |
| DEPARTING_AIRPORT: | Departing Airport |
| LATITUDE: | Latitude of departing airport |
| LONGITUDE: | Longitude of departing airport |
| PREVIOUS_AIRPORT: | Previous airport that aircraft departed from |
| PRCP: | Inches of precipitation for day |
| SNOW: | Inches of snowfall for day |
| SNWD: | Inches of snow on ground for day |
| TMAX: | Max temperature for day |
| AWND: | Max wind speed for day |

Docker i klaster kontejnera

- Dockerfile za Spark aplikaciju
- docker-compose.yml klaster konfiguracija:
 - Spark master + 2 workera
 - Hadoop namenode & datanode
 - ResourceManager & NodeManager & HistoryServer
- Spark aplikacija pokreće se unutar mreže bde
- Pokretanje infrastrukture:
 - docker network create bde
 - pozicioniranje u folder gde se nalazi docker-compose.yml
 - docker-compose up --build -d

Docker i klaster kontejnera

| | | | | | | | |
|---|---|---|-----------------|--------------|--|-------------|------------------------------------|
| □ | ▼ | ● | project1 | - | - | - | - |
| □ | | ● | spark-worker-2 | dae14ad3b4de | bde2020/spark-worker:3.1.2-hadoop3.2 | 8072:8071 ↗ | |
| □ | | ● | spark-worker-1 | 5b5aab52b805 | bde2020/spark-worker:3.1.2-hadoop3.2 | 8071:8071 ↗ | |
| □ | | ● | datanode | 7165db46f3ea | bde2020/hadoop-datanode:2.0.0-hadoop | | |
| □ | | ● | nodemanager | de66c531ca10 | bde2020/hadoop-nodemanager:2.0.0-hac | | |
| □ | | ● | namenode | 8d5961068b8b | bde2020/hadoop-namenode:2.0.0-hadoo | 9000:9000 ↗ | Show all ports (2) |
| □ | | ● | resourcemanager | 96495fd86a2f | bde2020/hadoop-resourcemanager:2.0.0 | | |
| □ | | ● | spark-master | 1c86a8c37f15 | bde2020/spark-master:3.1.2-hadoop3.2 | 7077:7077 ↗ | Show all ports (2) |
| □ | | ● | historyserver | 76792c6f1606 | bde2020/hadoop-historyserver:2.0.0-had | | |

Postavljanje podataka na HDFS

- Pokretanje skripte:
 - Pozicioniranje u folder gde se nalazi skripta `hdfs-put-data.bat`
 - `./hdfs-put-data.bat`
- Nakon pokretanja, svi neophodni podaci će biti kopirani na HDFS

```
@echo off
docker cp flightDelays.py namenode:/data
docker cp utils.py namenode:/data
docker cp data/full_data_flightdelay.csv namenode:/data
docker exec -it namenode bash -c "hdfs dfs -mkdir /dir"
docker exec -it namenode bash -c "hdfs dfs -mkdir /dir/FlightDelays"
docker exec -it namenode bash -c "hdfs dfs -rm -r /dir/FlightDelays/flightDelays.py"
docker exec -it namenode bash -c "hdfs dfs -rm -r /dir/FlightDelays/utils.py"
docker exec -it namenode bash -c "hdfs dfs -rm -r /dir/full_data_flightdelay.csv"
docker exec -it namenode bash -c "hdfs dfs -put /data/flightDelays.py /dir/FlightDelays"
docker exec -it namenode bash -c "hdfs dfs -put /data/utils.py /dir/FlightDelays"
docker exec -it namenode bash -c "hdfs dfs -put /data/full_data_flightdelay.csv /dir"
```

```
C:\projects\big-data-projects\Project1>.\hdfs-put-data.bat
Successfully copied 6.66kB to namenode:/data
Successfully copied 5.63kB to namenode:/data
Successfully copied 1.37GB to namenode:/data
mkdir: `/dir': File exists
mkdir: `/dir/FlightDelays': File exists
Deleted /dir/FlightDelays/flightDelays.py
Deleted /dir/FlightDelays/utils.py
Deleted /dir/full_data_flightdelay.csv
```

Pregled podataka

- Hadoop namenode: <http://localhost:9870/>

Hadoop Overview Datanodes Datanode Volume Failures Snapshot Startup Progress Utilities ▾

Browse Directory

/dir Go!

Show 25 entries Search:

| <input type="checkbox"/> | Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name | |
|--------------------------|------------|-------|------------|---------|---------------|-------------|------------|--------------------------|--|
| <input type="checkbox"/> | drwxr-xr-x | root | supergroup | 0 B | Nov 18 17:19 | 0 | 0 B | FlightDelays | |
| <input type="checkbox"/> | -rw-r--r-- | root | supergroup | 1.27 GB | Nov 18 17:19 | 3 | 128 MB | full_data_flighdelay.csv | |

Showing 1 to 2 of 2 entries Previous 1 Next

Hadoop, 2019.



Browse Directory

/dir/FlightDelays Go!

Show 25 entries Search:

| <input type="checkbox"/> | Permission | Owner | Group | Size | Last Modified | Replication | Block Size | Name |
|--------------------------|------------|-------|------------|---------|---------------|-------------|------------|-----------------|
| <input type="checkbox"/> | -rw-r--r-- | root | supergroup | 4.85 KB | Nov 18 17:19 | 3 | 128 MB | flightDelays.py |
| <input type="checkbox"/> | -rw-r--r-- | root | supergroup | 3.85 KB | Nov 18 17:19 | 3 | 128 MB | utils.py |

Pokretanje aplikacije

1. Manuelno pokretanje

- docker run -it --network bde --env-file hadoop.env -p 4040:4040 --name spark bde2020/spark-base:3.1.2-hadoop3.2 bash
- /spark/bin/spark-submit --master spark://spark-master:7077 --py-files hdfs://namenode:9000/dir/FlightDelays/utils.py hdfs://namenode:9000/dir/FlightDelays/flightDelays.py --input hdfs://namenode:9000/dir/full_data_flightdelay.csv --month 7 --carrier "United Air Lines Inc." --stats_col PLANE_AGE --group_by_col CARRIER_NAME

Pokretanje aplikacije

2. Korišćenjem Spark Python template-a

- Pozicionirati se u folder koji sadrži `run.bat` skriptu
- Pokrenuti `run.bat` skriptu: `.\run.bat`

```
@echo off
docker build --rm -t bde/spark-app .
docker run --name flightDelays --net bde -p 4040:4040 bde/spark-app
```

```
FROM bde2020/spark-python-template:3.1.2-hadoop3.2

COPY flightDelays.py /app/
COPY utils.py /app/

ENV SPARK_MASTER_PORT 7077
ENV SPARK_APPLICATION_PYTHON_LOCATION /app/flightDelays.py
ENV SPARK_SUBMIT_ARGS "--executor-memory 3G --executor-cores 3 --py-files /app/utils.py"
ENV SPARK_APPLICATION_ARGS "--input hdfs://namenode:9000/dir/full_data_flightdelay.csv --month 7 --carrier 'United Air Lines Inc.' --stats_co
```

Aplikacija

- Argumenti aplikacije

```
def get_args():
    parser = argparse.ArgumentParser()
    parser.add_argument("--input", required=True, help="Path to CSV file (Local or HDFS)")
    parser.add_argument("--month", type=int, default=None)
    parser.add_argument("--carrier", type=str, default=None, help="Airline name")
    parser.add_argument("--stats_col", type=str, default="PLANE_AGE", help="Column to calculate statistics on")
    parser.add_argument("--group_by_col", type=str, default="CARRIER_NAME", help="Column to group by for statistics")
    return parser.parse_args()
```

Aplikacija

- Spark inicijalizacija i učitavanje dataset-a

```
def initialize(args):
    spark_session = SparkSession.builder \
        .appName(APP_NAME) \
        .getOrCreate()

    print("Current Spark master:", spark_session.sparkContext.master)

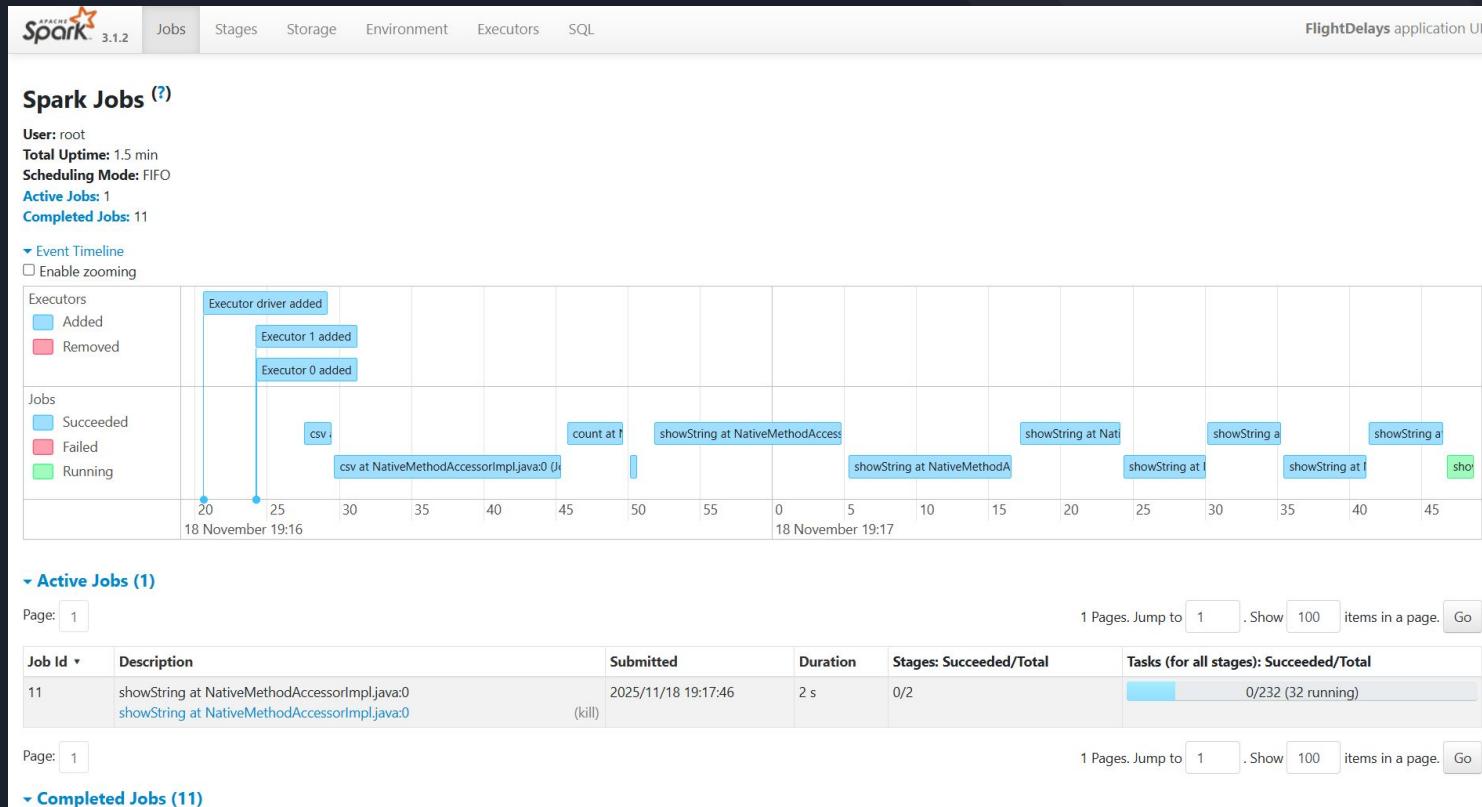
    spark_session.sparkContext.setLogLevel("ERROR")

    data_frame = spark_session.read.csv(args.input, header=True, inferSchema=True)

    return spark_session, data_frame
```

Praćenje rada Spark aplikacije

- <http://localhost:4040/>



Konfiguracije i promenljive okruženja

The screenshot shows the Apache Spark 3.1.2 Environment UI running on localhost:4040/environment/. The UI has tabs for Jobs, Stages, Storage, Environment (selected), Executors, and SQL. The Environment tab displays two sections: Runtime Information and Spark Properties.

Environment

Runtime Information

| Name | Value |
|---------------|-----------------------------------|
| Java Home | /usr/lib/jvm/java-1.8-openjdk/jre |
| Java Version | 1.8.0_275 (IcedTea) |
| Scala Version | version 2.12.10 |

Spark Properties

| Name | Value |
|------------------------------------|--|
| spark.app.id | app-20251118191620-0000 |
| spark.app.initial.file.urls | hdfs://namenode:9000/dir/FlightDelays/utils.py |
| spark.app.name | FlightDelays |
| spark.app.startTime | 1763493379022 |
| spark.driver.host | 45249099f53c |
| spark.driver.port | 36607 |
| spark.executor.id | driver |
| spark.files | hdfs://namenode:9000/dir/FlightDelays/utils.py |
| spark.master | spark://spark-master:7077 |
| spark.rdd.compress | True |
| spark.scheduler.mode | FIFO |
| spark.serializer.objectStreamReset | 100 |
| spark.submit.deployMode | client |
| spark.submit.pyFiles | /tmp/spark-a74c8288-81c1-4c09-82b0-cdc6e1db495b/utils.py |

Prikaz rezultata

- --input hdfs://namenode:9000/dir/full_data_flightdelay.csv
- --month 7
- --carrier "United Air Lines Inc."
- --stats_col PLANE_AGE
- --group_by_col CARRIER_NAME

| ----- Flights with delay ≥ 15 min per month ----- | |
|--|-------------------|
| MONTH | DelayedFlightsNum |
| 1 | 87682 |
| 2 | 98036 |
| 3 | 96589 |
| 4 | 98757 |
| 5 | 113530 |
| 6 | 135871 |
| 7 | 123238 |
| 8 | 119411 |
| 9 | 72834 |
| 10 | 90745 |
| 11 | 75576 |
| 12 | 115099 |

Prikaz rezultata

----- Flights with delay ≥15 min per airline -----

| CARRIER_NAME | DelayedFlightsNum |
|------------------------------|-------------------|
| Alaska Airlines Inc. | 39417 |
| Allegiant Air | 8072 |
| American Airlines Inc. | 181350 |
| American Eagle Airlines Inc. | 41153 |
| Atlantic Southeast Airlines | 23004 |
| Comair Inc. | 42687 |
| Delta Air Lines Inc. | 137361 |
| Endeavor Air Inc. | 35641 |
| Frontier Airlines Inc. | 31536 |
| Hawaiian Airlines Inc. | 6521 |
| JetBlue Airways | 68480 |
| Mesa Airlines Inc. | 34525 |
| Midwest Airline, Inc. | 49163 |
| SkyWest Airlines Inc. | 104124 |
| Southwest Airlines Co. | 271281 |
| Spirit Air Lines | 35585 |
| United Air Lines Inc. | 117468 |

Prikaz rezultata

```
----- Top 10 airports with highest number of delays ≥15 min -----
```

| DEPARTING_AIRPORT | DelayedFlightsNum |
|------------------------------|-------------------|
| Chicago O'Hare International | 74049 |
| Atlanta Municipal | 65892 |
| Dallas Fort Worth Regional | 65497 |
| Stapleton International | 55609 |
| Douglas Municipal | 44958 |
| Los Angeles International | 41061 |
| LaGuardia | 37766 |
| San Francisco International | 36856 |
| Houston Intercontinental | 33993 |
| Newark Liberty International | 33746 |


```
----- Flights with delay ≥15 min in month 7 -----
```

| MONTH | DelayedFlightsNum |
|-------|-------------------|
| 7 | 123238 |

Prikaz rezultata

```
----- Flights with delay ≥15 min in month 7 -----
```

| MONTH | DelayedFlightsNum |
|-------|-------------------|
| 7 | 123238 |

```
----- Flights with delay ≥15 min per airline in month 7 -----
```

| CARRIER_NAME | DelayedFlightsNum |
|------------------------------|-------------------|
| Alaska Airlines Inc. | 3490 |
| Allegiant Air | 1143 |
| American Airlines Inc. | 18424 |
| American Eagle Airlines Inc. | 4218 |
| Atlantic Southeast Airlines | 2180 |
| Comair Inc. | 4141 |
| Delta Air Lines Inc. | 16550 |
| Endeavor Air Inc. | 3336 |
| Frontier Airlines Inc. | 3629 |
| Hawaiian Airlines Inc. | 608 |
| JetBlue Airways | 6631 |
| Mesa Airlines Inc. | 3102 |
| Midwest Airline, Inc. | 4250 |
| SkyWest Airlines Inc. | 9803 |
| Southwest Airlines Co. | 24869 |
| Spirit Air Lines | 4209 |
| United Air Lines Inc. | 12655 |

Prikaz rezultata

| ----- Delayed flights for carrier United Air Lines Inc. per month ----- | | | | | |
|---|-------------------|--|--|--|--|
| MONTH | DelayedFlightsNum | | | | |
| 1 | 8057 | | | | |
| 2 | 7933 | | | | |
| 3 | 9552 | | | | |
| 4 | 9235 | | | | |
| 5 | 11210 | | | | |
| 6 | 13841 | | | | |
| 7 | 12655 | | | | |
| 8 | 11993 | | | | |
| 9 | 8352 | | | | |
| 10 | 8259 | | | | |
| 11 | 6789 | | | | |
| 12 | 9592 | | | | |

| ----- PLANE_AGE statistics by CARRIER_NAME ----- | | | | | |
|--|-------------------|---------------|---------------|-----------------|---------------|
| CARRIER_NAME | average_PLANE_AGE | min_PLANE_AGE | max_PLANE_AGE | stdev_PLANE_AGE | total_records |
| Southwest Airlines Co. | 12.12 | 0 | 22 | 5.76 | 1296329 |
| Delta Air Lines Inc. | 14.66 | 0 | 32 | 9.37 | 938346 |
| American Airlines Inc. | 11.23 | 0 | 26 | 6.76 | 903640 |
| United Air Lines Inc. | 15.23 | 0 | 30 | 6.91 | 601044 |
| SkyWest Airlines Inc. | 9.86 | 0 | 22 | 6.35 | 584204 |
| Midwest Airline, Inc. | 9.45 | 2 | 15 | 4.31 | 300154 |
| JetBlue Airways | 10.67 | 0 | 20 | 4.8 | 269596 |
| Alaska Airlines Inc. | 8.27 | 0 | 20 | 5.09 | 239337 |
| American Eagle Airlines Inc. | 10.97 | 0 | 21 | 6.47 | 228792 |
| Comair Inc. | 9.58 | 0 | 18 | 5.81 | 219324 |

only showing top 10 rows

Prikaz rezultata

```
----- Airport average monthly pass -----  
+-----+-----+-----+-----+  
| min(AVG_MONTHLY_PASS_AIRPORT) | max(AVG_MONTHLY_PASS_AIRPORT) | avg(AVG_MONTHLY_PASS_AIRPORT) | stddev_samp(AVG_MONTHLY_PASS_AIRPORT) |  
+-----+-----+-----+-----+  
| 70476 | 4365661 | 1588638.5315082518 | 1123847.2509635123 |  
+-----+-----+-----+-----+  
  
----- Effect of weather conditions on delays (month, avg TMAX, avg PRCP) -----  
  
+-----+-----+-----+  
| MONTH | AvgTMAX | AvgPRCP |  
+-----+-----+-----+  
| 1 | 49.41 | 0.15 |  
| 2 | 53.17 | 0.15 |  
| 3 | 59.56 | 0.11 |  
| 4 | 70.54 | 0.2 |  
| 5 | 76.25 | 0.19 |  
| 6 | 83.75 | 0.2 |  
| 7 | 89.2 | 0.14 |  
| 8 | 88.27 | 0.16 |  
| 9 | 84.73 | 0.12 |  
| 10 | 71.49 | 0.18 |  
| 11 | 58.95 | 0.09 |  
| 12 | 55.21 | 0.18 |  
+-----+-----+-----+
```

Prikaz rezultata

```
-- Average number of airline flights per airport --
```

| DEPARTING_AIRPORT | CARRIER_NAME | AvgMonthlyFlightsPerCarrier |
|--------------------------------|------------------------|-----------------------------|
| Spokane International | Southwest Airlines Co. | 115406.04 |
| Pensacola Regional | Southwest Airlines Co. | 115268.86 |
| Portland International Jetport | Southwest Airlines Co. | 115239.14 |
| Albany International | Southwest Airlines Co. | 114660.93 |
| Keahole | Southwest Airlines Co. | 112900.92 |
| Adams Field | Southwest Airlines Co. | 112896.61 |
| Kahului Airport | Southwest Airlines Co. | 112858.11 |
| Rochester Monroe County | Southwest Airlines Co. | 112824.96 |
| Honolulu International | Southwest Airlines Co. | 112728.41 |
| Long Beach Daugherty Field | Southwest Airlines Co. | 112458.5 |

```
only showing top 10 rows
```



```
-- Weather impact on delays --
```



```
--- Delayed flights ---
```

| avg_prcp_delayed | avg_snow_delayed | avg_temp_delayed | avg_wind_delayed |
|------------------|------------------|------------------|------------------|
| 0.16 | 0.06 | 71.13 | 8.72 |


```
--- On-time flights ---
```

| avg_prcp_on_time | avg_snow_on_time | avg_temp_on_time | avg_wind_on_time |
|------------------|------------------|------------------|------------------|
| 0.09 | 0.02 | 71.55 | 8.25 |



HVALA NA PAŽNJI!