Big Data - Projekat 1

Teodora Stamenković 1460

Dataset

Putanje za 442 taksi vozila u Portu u periodu od 01.07.2013 do 30.06.2014

- TRIP_ID
- CALL TYPE:
 - A poziv iz centrale
 - B direktno kontaktiranje taksija na stajalištu
 - o C ostalo
- ORIGIN_CALL identifikacija putnika ukoliko je CALL_TYPE='A'
- ORIGIN_STAND identifikacija taksi stajališta ukoliko je CALL_TYPE='B'
- TAXI_ID identifikacija taksi vozača
- TIMESTAMP vreme početka
- DAY_TYPE
 - o B praznik
 - C dan pre praznika
 - o A obican dan
- MISSING_DATA da li nedostaju GPS podaci
- POLYLINE lista parova (lat, lon) za svakih 15s putovanja

```
root
|-- TRIP_ID: long (nullable = true)
|-- CALL_TYPE: string (nullable = true)
|-- ORIGIN_CALL: integer (nullable = true)
|-- ORIGIN_STAND: integer (nullable = true)
|-- TAXI_ID: integer (nullable = true)
|-- TIMESTAMP: integer (nullable = true)
|-- DAY_TYPE: string (nullable = true)
|-- MISSING_DATA: boolean (nullable = true)
|-- POLYLINE: string (nullable = true)
```

Prikaz podataka

```
TRIP ID|CALL TYPE|ORIGIN CALL|ORIGIN STAND| TAXI ID| TIMESTAMP|DAY TYPE|MISSING DATA|
                                                                                                                POLYLINE
1372636858620000589
                                      nullI
                                                    null|20000589|1372636858|
                                                                                              false|[[-8.618643,41.14...|
                                      null
                                                                                              false|[[-8.639847,41.15...|
1372637303620000596
                             B
                                                       7 | 20000596 | 1372637303 |
|1372636951620000320|
                                      nulli
                                                    null|20000320|1372636951|
                                                                                              false|[[-8.612964,41.14...|
                                      null
                                                                                              false|[[-8.574678,41.15...|
1372636854620000520
                                                    null|20000520|1372636854|
|1372637091620000337|
                                      nullI
                                                    null|20000337|1372637091|
                                                                                              false|[[-8.645994,41.18...|
1372636965620000231
                                      null
                                                    null|20000231|1372636965|
                                                                                              false|[[-8.615502,41.14...|
                                      nulli
                                                    null|20000456|1372637210|
                                                                                              false|[[-8.57952,41.145...|
1372637210620000456
                                                                                    A
                                                                                              false|[[-8.617563,41.14...|
1372637299620000011
                                      null
                                                    null|20000011|1372637299|
1372637274620000403
                                      null
                                                    null|20000403|1372637274|
                                                                                              false|[[-8.611794,41.14...|
                             CI
                                      nulli
                                                                                              false|[[-8.615907,41.14...|
1372637905620000320
                                                    null|20000320|1372637905|
only showing top 10 rows
```

DataFrame Rows count : 1710670

Transformacija podataka

```
#drop missing values
dataset = dataset.filter(dataset["missing data"] == False)
dataset = dataset.withColumnRenamed("timestamp", "start time")
# check missing values
print(dataset.where(dataset["start time"].isNull()).count())
# data transformation
dataset = dataset.withColumn("coordinates", F.regexp replace("polyline", "[\[|\]]", ""))
dataset = dataset.withColumn("start lon", F.split(dataset["coordinates"], ",").getItem(0).cast('double'))
dataset = dataset.withColumn("start lat", F.split(dataset["coordinates"], ",").getItem(1).cast('double'))
dataset = dataset.withColumn("end lon", F.reverse(F.split(dataset["coordinates"], ",")).getItem(1).cast('double'))
dataset = dataset.withColumn("end lat", F.reverse(F.split(dataset["coordinates"], ",")).getItem(0).cast('double'))
dataset = dataset.withColumn("array of coordinates", F.split(dataset["coordinates"], ","))
dataset = dataset.withColumn("trip duration", F.size(F.col("array of coordinates")) * 7.5 )
dataset = dataset.withColumn("end time", dataset["start time"] + dataset["trip duration"])
dataset = dataset.drop("polyline")
dataset = dataset.drop("coordinates")
dataset = dataset.drop("array of coordinates")
dataset.show(10)
```

Prikaz podataka nakon transformacija

++			+	+		+
trip_id cal	l_type orig	jin_call ori	gin_stand taxi_id start_time day	_type mi	ssing_data start_lon start_lat end_lon end_lat trip_d.	uration end_time
<u> </u>				+		
1372636858620000589	c	null	null 20000589 1372636858	A	false -8.618643 41.141412 -8.630838 41.154489	345.0 1.372637203E9
1372637303620000596	B	null	7 20000596 1372637303	A	false -8.639847 41.159826 -8.66574 41.170671	285.0 1.372637588E9
1372636951620000320	c	null	null 20000320 1372636951	A	false -8.612964 41.140359 -8.61597 41.14053	975.0 1.372637926E9
1372636854620000520	c	null	null 20000520 1372636854	A	false -8.574678 41.151951 -8.607996 41.142915	645.0 1.372637499E9
1372637091620000337	C	null	null 20000337 1372637091	A	false -8.645994 41.18049 -8.687268 41.178087	435.0 1.372637526E9
1372636965620000231	c	null	null 20000231 1372636965	A	false -8.615502 41.140674 -8.578224 41.160717	390.0 1.372637355E9
1372637210620000456	C	null	null 20000456 1372637210	A	false -8.57952 41.145948 -8.603973 41.142816	540.0 1.37263775E9
1372637299620000011	c	null	null 20000011 1372637299	A	false -8.617563 41.146182 -8.6247 41.161554	510.0 1.372637809E9
1372637274620000403	c	null	null 20000403 1372637274	A	false -8.611794 41.140557 -8.589402 41.163309	570.0 1.372637844E9
1372637905620000320	c	null	null 20000320 1372637905	A	false -8.615907 41.140557 -8.604594 41.134158	285.0 1.37263819E9
++	+		+	+		+

only showing top 10 row

Filtriranje podataka

```
first_latitude = float(os.getenv('FIRST_LATITUDE'))
first_longitude = float(os.getenv('FIRST_LONGITUDE'))
second_latitude = float(os.getenv('SECOND_LATITUDE'))
second_longitude = float(os.getenv('SECOND_LONGITUDE'))
start_time = int(os.getenv('START_TIME'))
end_time = int(os.getenv('END_TIME'))
```

Podaci se filtriraju na osnovu ulaznih podataka (opseg koordinata i vremenski period)

```
print("Average trip duration grouped by call type")
dataset_trip_duration_by_call_type = dataset_filtered.groupBy("call_type").agg(F.avg("trip_duration"))
dataset_trip_duration_by_call_type.show()
```

Output:

```
1 Taxi stand with the widest range of trip duration is 59 (stddev = 1245.2054778031106)
```

Output:

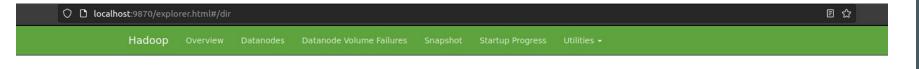
```
The longest trip was 0.9041666666666667 hours from taxi stand 15

The highest number of trips (19) started from taxi stand 15
```

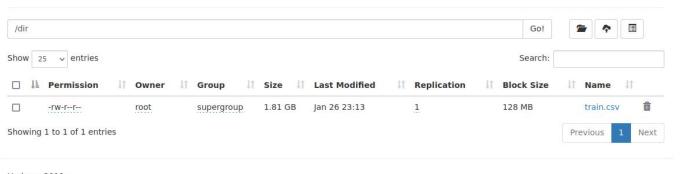
Output:

4 Taxi driver who spent the least time driving is 20000472 (0.75 minutes)

Čuvanje podataka na HDFS



Browse Directory

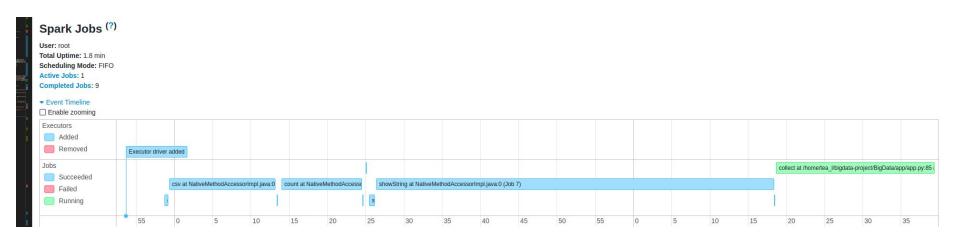


Hadoop, 2019.

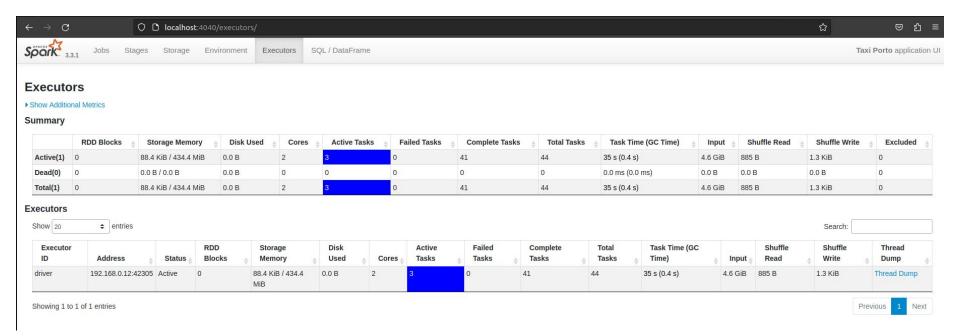
Lokalno izvršenje

```
input = "hdfs://localhost:9000/dir/train.csv"
spark = SparkSession.builder.appName(appName).master("local[2]").getOrCreate()
```

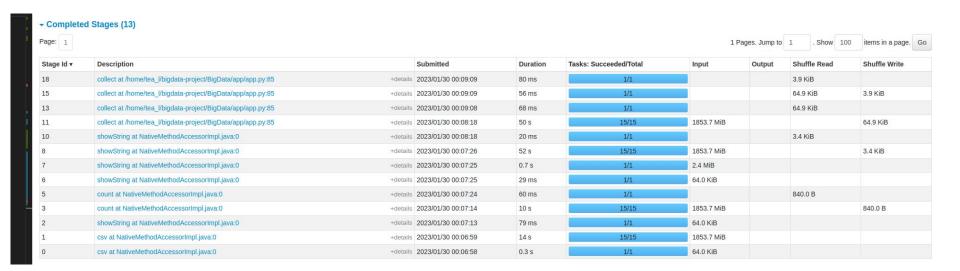
dataset = spark.read.option("inferSchema", True).option("header", True).csv(input)



Lokalno izvršenje - executors



Lokalno izvršenje - stages



Izvršenje na klasteru

```
input = "hdfs://namenode:9000/dir/train.csv"
spark = SparkSession.builder.appName(appName).master("spark://spark-master:7077").getOrCreate()
```

dataset = spark.read.option("inferSchema", True).option("header", True).csv(input)



Spark Master at spark://b115afcb22a3:7077

URL: spark://b115afcb22a3:7077

Alive Workers: 2

Cores in use: 8 Total, 8 Used

Memory in use: 13.5 GiB Total, 2.0 GiB Used

Resources in use:

Applications: 1 Running, 0 Completed

Drivers: 0 Running, 0 Completed

Status: ALIVE

→ Workers (2)

Worker Id	Address	State	Cores	Memory	Resources
worker-20230130072909-172.18.0.8-40165	172.18.0.8:40165	ALIVE	4 (4 Used)	6.8 GiB (1024.0 MiB Used)	
worker-20230130072909-172.18.0.9-39879	172.18.0.9:39879	ALIVE	4 (4 Used)	6.8 GiB (1024.0 MiB Used)	

▼ Running Applications (1)

Application ID	Name	Cores	Memory per Executor	Resources Per Executor	Submitted Time	User	State	Duration
app-20230130072940-0000	(kill) Taxi Porto	8	1024.0 MiB		2023/01/30 07:29:40	root	RUNNING	4.9 min

Izvršenje na klasteru



Stages

Storage

Environment

Executors

SQL

Taxi Porto application UI

Executors

▶ Show Additional Metrics

Summary

	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	♦ Input	Shuffle Read	Shuffle Write	Excluded
Active(3)	0	547.3 KiB / 1.1 GiB	0.0 B	8	6	0	1519	1525	46 min (2.7 min)	11.9 GiB	4.8 KiB	5.5 KiB	0
Dead(0)	0	0.0 B / 0.0 B	0.0 B	0	0	0	0	0	0.0 ms (0.0 ms)	0.0 B	0.0 B	0.0 B	0
Total(3)	0	547.3 KiB / 1.1 GiB	0.0 B	8	6	0	1519	1525	46 min (2.7 min)	11.9 GiB	4.8 KiB	5.5 KiB	0

Executors

Total Tasks Executor ID A Address Status **RDD Blocks** Storage Memory Disk Used Cores **Active Tasks** Failed Tasks Complete Tasks Task Time (GC Time) Input A Shuffle Read **Shuffle Write Thread Dump** 172.18.0.9:42115 Active 164.2 KiB / 366.3 MiB 0.0 B 763 767 24 min (1.5 min) 6.4 GiB 1.4 KiB 1.5 KiB Thread Dump stderr driver 0e07af938c64:35155 191.6 KiB / 366.3 MiB 0.0 B 0 0 0 0 0 0.0 ms (0.0 ms) 0.0 B 0.0 B 0.0 B Thread Dump 172.18.0.8:39993 191.6 KiB / 366.3 MiB 0.0 B 4 756 758 22 min (1.2 min) 5.5 GiB 3.4 KiB 4 KiB stdout Thread Dump Active stderr

Showing 1 to 3 of 3 entries

Show 20 \$ entries



Search:

