

Big Data - Projekat 1

Teodora Stamenković 1460

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
 - 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
 - B - praznik
 - C - dan pre praznika
 - 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	C	null	null	20000589	1372636858	A	false	[[-8.618643,41.14...
1372637303620000596	B	null	7	20000596	1372637303	A	false	[[-8.639847,41.15...
1372636951620000320	C	null	null	20000320	1372636951	A	false	[[-8.612964,41.14...
1372636854620000520	C	null	null	20000520	1372636854	A	false	[[-8.574678,41.15...
1372637091620000337	C	null	null	20000337	1372637091	A	false	[[-8.645994,41.18...
1372636965620000231	C	null	null	20000231	1372636965	A	false	[[-8.615502,41.14...
1372637210620000456	C	null	null	20000456	1372637210	A	false	[[-8.57952,41.145...
1372637299620000011	C	null	null	20000011	1372637299	A	false	[[-8.617563,41.14...
1372637274620000403	C	null	null	20000403	1372637274	A	false	[[-8.611794,41.14...
1372637905620000320	C	null	null	20000320	1372637905	A	false	[[-8.615907,41.14...

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	call_type	origin_call	origin_stand	taxi_id	start_time	day_type	missing_data	start_lon	start_lat	end_lon	end_lat	trip_duration	end_time
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 rows

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)

```
dataset_filtered = dataset.filter(((dataset["start_lat"] > first_latitude) & (dataset["start_lon"] > first_longitude)
& (dataset["start_lat"] < second_latitude) & (dataset["start_lon"] < second_longitude)
& (dataset["start_time"] > start_time) & (dataset["start_time"] < end_time))
| ((dataset["end_lat"] > first_latitude) & (dataset["end_lon"] > first_longitude)
& (dataset["end_lat"] < second_latitude) & (dataset["end_lon"] < second_longitude)
& (dataset["end_time"] > start_time) & (dataset["end_time"] < end_time)))
```

Analiza podataka

```
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()
```

Average trip duration grouped by call type

call_type	avg(trip_duration)
B	838.3928571428571
C	1038.75
A	843.75

Analiza podataka

```
trip_duration_stddev = dataset_filtered.groupBy("origin_stand").agg(F.stddev("trip_duration").alias("trip_duration_stddev"))
trip_duration_stddev = trip_duration_stddev.sort("trip_duration_stddev", ascending=False)
origin_stand = trip_duration_stddev.collect()[0]
result = "Taxi stand with the widest range of trip duration is " + str(origin_stand.asDict()["origin_stand"]) \
+ " (stddev = " + str(origin_stand.asDict()['trip_duration_stddev']) + ")\n"
```

Output:

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


Analiza podataka

```
dataset_max = dataset_filtered.filter(dataset_filtered["call_type"] == 'B').groupBy('origin_stand').agg(count('trip_id')\
    .alias('num_of_trips'), max('trip_duration').alias('max_trip_time'))
dataset_max_duration = dataset_max.sort("max_trip_time", ascending=False)
dataset_max_count = dataset_max.sort("num_of_trips", ascending=False)
max_duration = dataset_max_duration.collect()[0]
max_count = dataset_max_count.collect()[0]
result = "The longest trip was " + str(max_duration.asDict()['max_trip_time'] / 3600) + " hours from taxi stand " \
    + str(max_duration.asDict()['origin_stand']) + "\n"
result_count = "The highest number of trips (" + str(max_count.asDict()['num_of_trips']) + ") started from taxi stand " \
    + str(max_count.asDict()['origin_stand']) + "\n"
```

Output:

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

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

Analiza podataka

```
dataset_taxi = dataset_filtered.groupBy("taxi_id").agg(F.sum("trip_duration").alias('trip_duration_sum'))
dataset_taxi = dataset_taxi.filter(dataset_taxi["trip_duration_sum"] > 0).sort("trip_duration_sum", ascending=True)
taxi_driver = dataset_taxi.collect()[0]
result = 'Taxi driver who spent the least time driving is ' + str(taxi_driver.asDict()['taxi_id']) \
        + ' (' + str(taxi_driver.asDict()['trip_duration_sum'] / 60) + ' minutes)\n'
```

Output:

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

Čuvanje podataka na HDFS




localhost:9870/explorer.html#/dir

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"/>	-rw-r--r--	root	supergroup	1.81 GB	Jan 26 23:13	1	128 MB	train.csv 

Showing 1 to 1 of 1 entries

Previous

1

Next

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)
```

Spark Jobs (?)

User: root
Total Uptime: 1.8 min
Scheduling Mode: FIFO
Active Jobs: 1
Completed Jobs: 9

Event Timeline
☐ Enable zooming



Lokalno izvršenje - executors

←

→

↻

localhost:4040/executors/

☆

📌

☰

APACHE
Spark

3.3.1

Jobs

Stages

Storage

Environment

Executors

SQL / DataFrame

Taxi Porto application UI

Executors

[Show Additional Metrics](#)

Summary

Executors

Show entriesSearch:

Showing 1 to 1 of 1 entries

Previous **1** Next

Lokalno izvršenje - stages

Completed Stages (13)

Page: 1

1 Pages. Jump to 1. Show 100 items in a page. Go

Stage Id ▼	Description		Submitted	Duration	Tasks: Succeeded/Total	Input	Output	Shuffle Read	Shuffle Write
18	collect at /home/tea_1/bigdata-project/BigData/app/app.py:85	+details	2023/01/30 00:09:09	80 ms	1/1			3.9 KiB	
15	collect at /home/tea_1/bigdata-project/BigData/app/app.py:85	+details	2023/01/30 00:09:09	56 ms	1/1			64.9 KiB	3.9 KiB
13	collect at /home/tea_1/bigdata-project/BigData/app/app.py:85	+details	2023/01/30 00:09:08	68 ms	1/1			64.9 KiB	
11	collect at /home/tea_1/bigdata-project/BigData/app/app.py:85	+details	2023/01/30 00:08:18	50 s	15/15	1853.7 MiB			64.9 KiB
10	showString at NativeMethodAccessorImpl.java:0	+details	2023/01/30 00:08:18	20 ms	1/1			3.4 KiB	
8	showString at NativeMethodAccessorImpl.java:0	+details	2023/01/30 00:07:26	52 s	15/15	1853.7 MiB			3.4 KiB
7	showString at NativeMethodAccessorImpl.java:0	+details	2023/01/30 00:07:25	0.7 s	1/1	2.4 MiB			
6	showString at NativeMethodAccessorImpl.java:0	+details	2023/01/30 00:07:25	29 ms	1/1	64.0 KiB			
5	count at NativeMethodAccessorImpl.java:0	+details	2023/01/30 00:07:24	60 ms	1/1			840.0 B	
3	count at NativeMethodAccessorImpl.java:0	+details	2023/01/30 00:07:14	10 s	15/15	1853.7 MiB			840.0 B
2	showString at NativeMethodAccessorImpl.java:0	+details	2023/01/30 00:07:13	79 ms	1/1	64.0 KiB			
1	csv at NativeMethodAccessorImpl.java:0	+details	2023/01/30 00:06:59	14 s	15/15	1853.7 MiB			
0	csv at NativeMethodAccessorImpl.java:0	+details	2023/01/30 00:06:58	0.3 s	1/1	64.0 KiB			

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)
```



3.1.2

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



3.1.2

Jobs

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

Show 20 entries

Search:

Executor ID ▲	Address	Status	RDD Blocks	Storage Memory	Disk Used	Cores	Active Tasks	Failed Tasks	Complete Tasks	Total Tasks	Task Time (GC Time)	Input	Shuffle Read	Shuffle Write	Logs	Thread Dump
0	172.18.0.9:42115	Active	0	164.2 KiB / 366.3 MiB	0.0 B	4	4	0	763	767	24 min (1.5 min)	6.4 GiB	1.4 KiB	1.5 KiB	stdout stderr	Thread Dump
driver	0e07af938c64:35155	Active	0	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
1	172.18.0.8:39993	Active	0	191.6 KiB / 366.3 MiB	0.0 B	4	2	0	756	758	22 min (1.2 min)	5.5 GiB	3.4 KiB	4 KiB	stdout stderr	Thread Dump

Showing 1 to 3 of 3 entries

Previous 1 Next