

# Projekat 1 – Big data

Nikola Đorđević 1567

# Skup podataka - Brightkite check ins

Brightkite je nekada bio provajder društvenih mreža zasnovan na lokaciji, gde su korisnici delili svoje lokacije. Prikupljeno je 4 491 143 čekiranja korisnika, odnosno prijavljivanja na određenu lokaciju. Atributi koji su zastupljeni u skupu podataka su:

- User – id korisnika
- Check\_in\_time – vreme prijave na određenu lokaciju
- Longitude – geografska dužina lokacije
- Latitude – geografska širina lokacije
- Location\_id – id lokacije

# Primer slogova u skupu podataka

```
+-----+-----+-----+-----+
|user|check_in_time    |latitude |longitude  |location_id|
+-----+-----+-----+-----+
|0    |2010-10-17 03:48:53|39.747652|-104.99251 |88c46bf20db295831bd2d1718ad7e6f5|
|0    |2010-10-16 08:02:04|39.891383|-105.070814|7a0f88982aa015062b95e3b4843f9ca2|
|0    |2010-10-16 05:48:54|39.891077|-105.068532|dd7cd3d264c2d063832db506fba8bf79|
|0    |2010-10-14 20:25:51|39.750469|-104.999073|9848afcc62e500a01cf6fbf24b797732f8963683|
|0    |2010-10-14 02:21:47|39.752713|-104.996337|2ef143e12038c870038df53e0478cefc|
|0    |2010-10-14 01:31:51|39.752508|-104.996637|424eb3dd143292f9e013efa00486c907|
|0    |2010-10-13 22:05:43|39.7513   |-105.000121|d268093afe06bd7d37d91c4d436e0c40d217b20a|
|0    |2010-10-13 18:41:35|39.758974|-105.010853|6f5b96170b7744af3c7577fa35ed0b8f|
|0    |2010-10-13 05:57:23|39.827022|-105.143191|f6f52a75fd80e27e3770cd3a87054f27|
|0    |2010-10-12 21:56:49|39.749934|-105.000017|b3d356765cc8a4aa7ac5cd18caafd393|
+-----+-----+-----+-----+
```

# Filtriranje

- Iz skupa podataka izdajamo oblast koja se nalazi između 40. i 60. stepena geografske širine, i, između -104. i -120. stepena geografske dužine.
- Kao vremenski prozor koji posmaramo, uzimamao dane između 15.10.2010. i 31.10.2010. godine.

```
filtered_df = df.filter((df["latitude"] > latitude_lower) & (df["latitude"] < latitude_upper) & (df["longitude"] < longitude_upper) &  
                        (df["longitude"] > longitude_lower) & (df["check_in_time"] > date_lower) & (df["check_in_time"] < date_upper))
```

# Broj pojavljivanja korisnika u prethodno definisanoj oblasti i vremenskom intervalu

```
grouped_df = filtered_df.groupBy("user").count().alias("count")
result = grouped_df.select("user", "count")
result.show()
```

Podaci grupisani po korisnicima, za odredjenu oblast i odrjenji vremenski interval

```
+----+-----+
|user|count|
+----+-----+
|  15|    1|
|1092|    1|
|1863|   16|
|2879|    3|
|2937|    1|
|2878|    1|
+----+-----+
```

Ukupan broj prijavljivanja na određenu lokaciju koja pripada prethodno definisanoj oblasti, u definisanom vremenskog okviru

```
grouped_df = filtered_df.groupBy("location_id").count().alias("count")
result = grouped_df.select("location_id", "count")
result.show()
```

```
Podaci grupisani po lokacijama, za odredjenu oblast i odrjenji vremenski interval
+-----+
|          location_id|count|
+-----+
|9a404eee49b47d7df...|    1|
|5f9fd804781871335...|    1|
|eebadad0a22411ddb...|    1|
|7b062c7e83141c131...|    1|
|1ab85b9dff61c977b...|    1|
|d22e7db4cb988d813...|    2|
|5d2db576912de7550...|    1|
|f52bde44aa45c171a...|    1|
|f54c489c05b23dfcd...|    1|
|eeffd054a22411ddb...|    2|
|eeb9db9ea22411dd8...|    1|
|eeb46b50a22411dda...|    1|
```

# Kolona – time\_spent

Veštački je generisana nova kolona "time\_spent", koja sadrži informaciju o tome koliko je korisnik proveo vremena na nekoj lokaciji. Skup podaka nakon dodavanja te kolone je prikazan na slici.

user	check_in_time	latitude	longitude	location_id	time_spent
0	2010-10-17 03:48:53	39.747652	-104.99251	88c46bf20db295831...	124
0	2010-10-16 08:02:04	39.891383	-105.070814	7a0f88982aa015062...	99
0	2010-10-16 05:48:54	39.891077	-105.068532	dd7cd3d264c2d0638...	121
0	2010-10-14 20:25:51	39.750469	-104.999073	9848afcc62e500a01...	227
0	2010-10-14 02:21:47	39.752713	-104.996337	2ef143e12038c8700...	291
0	2010-10-14 01:31:51	39.752508	-104.996637	424eb3dd143292f9e...	165
0	2010-10-13 22:05:43	39.7513	-105.000121	d268093afe06bd7d3...	201
0	2010-10-13 18:41:35	39.758974	-105.010853	6f5b96170b7744af3...	78
0	2010-10-13 05:57:23	39.827022	-105.143191	f6f52a75fd80e27e3...	237
0	2010-10-12 21:56:49	39.749934	-105.000017	b3d356765cc8a4aa7...	178
0	2010-10-11 04:51:09	39.891077	-105.068532	6f3a2db56d4fa788f...	29
0	2010-10-09 07:45:25	39.891077	-105.068532	6f3a2db56d4fa788f...	148
0	2010-10-08 05:33:37	39.758974	-105.010853	11da318f0ea3c4a8f...	320

# Kolona – check\_in\_date

Dodata je nova kolona, "check\_in\_date", koja izvlači informaciju o datumu iz kolone "check\_in\_time", jer ona sadrži informaciju i o vremenu i o datumu. Ovo je urađeno da bismo u nastavku projekta mogli da grupišemo slogove po datumu.

```
# Dodavanje nove kolone koja ce da sadrzi samo informaciji o datumu  
df = df.withColumn("check_in_date", to_date(col("check_in_time"), "yyyy-MM-dd'T'HH:mm:ss'Z'"))
```



# Određivanje statističkih parametara za atribut time\_spent na određenoj lokaciji, određenog datuma

```
grouped_df = df.groupby("location_id", "check_in_date").agg(  
    min("time_spent").alias("min_time_spent"),  
    max("time_spent").alias("max_time_spent"),  
    avg("time_spent").alias("avg_time_spent"),  
    coalesce(stddev("time_spent"), lit(0.0)).alias("stddev_time_spent")  
)
```

Podaci grupisani po lokacijama i datumima:

location_id	check_in_date	min_time_spent	max_time_spent	avg_time_spent	stddev_time_spent
dd7cd3d264c2d0638...	2010-04-07	111	111	111.0	0.0
34d9fb11d6e4e875b...	2010-03-12	251	251	251.0	0.0
3c416dba5f9811deb...	2010-02-05	231	231	231.0	0.0
6db9d82229f1e0eb2...	2010-01-06	2	35	18.5	23.33452377915607
7648bcecaf9911deb...	2009-10-02	172	351	261.5	126.572113832392
7a22ad4e9e3f11ddb...	2009-08-16	1	188	94.5	132.22896808188437
2356b384f6c4c8095...	2010-06-24	325	325	325.0	0.0
901ec5410896ff121...	2010-04-23	23	23	23.0	0.0
8fde23d6245c11deb...	2010-01-21	276	345	319.2	27.79748190034486
ee81e0b8a22411dda...	2009-07-02	34	351	207.55555555555554	122.32856484802629
ee6b8534a22411dd9...	2009-04-21	51	325	188.8125	69.84575267449458

Određivanje statističkih parametara za atribut `time_spent` na određenoj lokaciji, određenog datuma, za svakog korisnika

```
grouped_df = df.groupby("user", "location_id", "check_in_date").agg(
    min("time_spent").alias("min_time_spent"),
    max("time_spent").alias("max_time_spent"),
    avg("time_spent").alias("avg_time_spent"),
    coalesce(stddev("time_spent"), lit(0.0)).alias("stddev_time_spent")
)
```

Podaci grupisani po lokacijama i datumima, za određenog korisnika:

user	location_id	check_in_date	min_time_spent	max_time_spent	avg_time_spent	stddev_time_spent
0 7a0f88982aa015062...		2010-08-07	8	8	8.0	0.0
0 f6f52a75fd80e27e3...		2010-06-30	53	328	190.5	194.45436482630058
0 5b1e75de2fd3acfc...		2009-06-18	338	338	338.0	0.0
0 ec71ac1659d211de9...		2009-06-15	245	273	259.0	19.79898987322333
0 0b8d9aa2489d4a8b0...		2009-05-27	90	90	90.0	0.0
1 c18334382591a7534...		2010-04-02	271	271	271.0	0.0
1 dbacc50b61ae8d7df...		2010-03-04	209	209	209.0	0.0
1 e4b85ebcf23911dda...		2009-10-19	333	333	333.0	0.0
1 ee8403d4a22411dd8...		2009-08-18	229	229	229.0	0.0
1 ee6b8534a22411dd9...		2009-05-06	6	6	6.0	0.0
1 ee81e0b8a22411dda...		2009-04-03	122	229	183.0	55.054518434003214
2 6e3768ee94fc11dea...		2010-06-16	325	325	325.0	0.0

# Najposećenija lokacija za svakog korisnika

```
+-----+
|user|max_visits|most_visited_location|
+-----+
|  0|      213| ba494e1ceff294a80...|
|  1|      297| c0b8fc511887b932c...|
|  2|      247| 3534836081ca2913b...|
|  3|      119| 3410de6ca0f4d8c77...|
|  4|      132| 9325dd207bed11dea...|
|  5|       92| ee81e0b8a22411dda...|
|  6|      110| dab6294e732d36c2d...|
|  7|      300| 74a3b18100c1bb81d...|
|  8|        4| 4080a0d997eaeb739...|
|  9|       23| 2d144622e8333f8de...|
| 10|       56| 737eafcefd83b2554...|
| 11|      441| af9bc69f55fa74cd3...|
| 12|      320| 81091f83e687d1bc0...|
| 13|      105| 48697cc9ceb4fe0fb...|
| 14|       63| d9988f09bff65eefb...|
| 15|      464| 50b5757e7bdc11de9...|
| 16|       39| 9e7eacc95b1ed4eec...|
| 17|      155| d1ab6766ae5f31894...|
| 18|       66| f99685b6643511deb...|
| 19|      210| 9c25dbbc444011deb...|
+-----+
```

# Datumi kada je svaka od lokacija bila najposećenija

```
+-----+-----+-----+
|      location_id|max_visits|most_busiest_date|
+-----+-----+-----+
|6d2b50de917111de9...|      1|    2009-08-25|
|d6de67abf64236233...|      4|    2009-12-05|
|2d766c02ea72f085d...|      3|    2009-01-12|
|fd8eacc4b1cb11dda...|      5|    2009-06-16|
|26a11bced05811ddb...|      3|    2009-06-02|
|3033e2d7a560b5a5c...|      2|    2008-12-27|
|11e6ea3656fe520b7...|      4|    2008-12-01|
|d1937763c05393367...|      2|    2008-11-16|
|5298aa3780e37d294...|      1|    2010-02-02|
|3e7f42009fcb11de8...|      1|    2009-09-26|
|8ee95c686be11deaf...|      1|    2009-08-12|
|2773ea36c7bfae09c...|      1|    2010-08-04|
|b6f6070ae12611ddb...|      2|    2009-07-25|
|c6ad07a0a83911dd8...|      2|    2008-11-01|
|700e4e2929211dda0...|      2|    2008-10-30|
|15325f65b6662f0b3...|      3|    2009-09-20|
|99774f916e92f1706...|      1|    2010-03-21|
|659ac30cabcd11dd9...|      1|    2008-11-15|
|4d6416d28b5511dd9...|      2|    2009-01-30|
|f79c95e5936a31191...|      1|    2009-12-09|
+-----+-----+-----+
```

# Prosečno vreme provedeno na svakoј od lokacija

```
+-----+-----+
|      location_id|      avg_time_spent|
+-----+-----+
|64b925364ac71005a...|      211.5|
|31c4ef9bbd5f8ebe2...|      219.7|
|e5acc4dbe743d713b...|       32.0|
|d9ba7fd062dd11de9...|      133.0|
|db3e5b3257cb11dea...|      306.0|
|8ec18f7210b811deb...|     164.75|
|3033e2d7a560b5a5c...|       85.4|
|ca54a3d67bc011dda...|      181.0|
|cc8795999b01637cf...|      309.0|
|7b9b133c85199c71d...|      345.0|
|f4491cf2fe7c88fb3...|      185.7|
|abc4c6f75f5453484...|174.57142857142858|
|d2c833d0d13d1c037...|      276.0|
|6ce8b78370a37d7b7...|      170.5|
|f79c95e5936a31191...|       90.0|
|6bdc8b8d41bcf11dea...|      198.0|
|4d9b5a6f08993325b...|      296.0|
|6d540229c0547a981...|208.33333333333334|
|fd8eacc4b1cb11dda...|170.77083333333334|
|cf86f32c17e211dea...|      324.0|
+-----+-----+
```

# Aplikacija u PyCharm-u, na lokalnoj mašini

The screenshot displays the PyCharm IDE interface. The main editor shows a Python script named `app.py` with the following code:

```
1 import sys
2 from pyspark.sql import SparkSession
3 from pyspark.sql.functions import *
4
5 spark = SparkSession.builder.appName("Demo").master("local[*]").getOrCreate()
6
7 #Učitavanje podataka
8 df = spark.read.option("inferSchema", True).option("header", True).csv(sys.argv[1])
```

The Run tool window at the bottom shows the execution results of the script. It displays a table with 20 rows of data, representing the top 20 rows of the CSV file. The data is as follows:

1 c18334382591a7534...	2010-04-02	271	271	271.0	0.0
1 dbacc50b61ae8d7df...	2010-03-04	209	209	209.0	0.0
1 e4b85ebcf23911dda...	2009-10-19	333	333	333.0	0.0
1 ee8403d4a22411dd8...	2009-08-18	229	229	229.0	0.0
1 ee6b8534a22411dd9...	2009-05-06	6	6	6.0	0.0
1 ee81e0b8a22411dda...	2009-04-03	122	229	183.0 55.054518434003214	
2 6e3768ee94fc11dea...	2010-06-16	325	325	325.0	0.0
2 06ccfd02e400b52bc...	2009-11-23	129	148	138.5 13.435028842544403	
2 261ab815f8853d072...	2009-10-25	120	251	185.5  92.63098833543772	
2 f3180c88fea911dd9...	2009-02-19	43	43	43.0	0.0
2 e194ee413805ba5a6...	2008-05-07	165	170	167.5 3.5355339059327378	
3 0b245362a81d22193...	2010-07-30	219	219	219.0	0.0
3 4af7820ab51a11dd8...	2010-05-18	23	107	65.0  59.39696961966999	
3 ee81ef22a22411ddb...	2010-03-14	28	305	166.5 195.86857838867365	
3 d2ecffd5c32390644...	2010-02-20	288	288	288.0	0.0

only showing top 20 rows

Process finished with exit code 0

The bottom status bar shows the file encoding as UTF-8, 4 spaces, and Python 3.9 (venv).

# Aplikacija na klasteru Spark Docker container-a (BDE)

Docker Desktop

Upgrade plan

Search

Ctrl+K

Sign in

Containers

Images

Volumes

Dev Environments

Extensions

Add Extensions

Containers

A container packages up code and its dependencies so the application runs quickly and reliably from one computing environment to another.

Only show running containers

	NAME	IMAGE	STATUS	PORT(S)	STARTED	ACTIONS
<input type="checkbox"/>	spark ff6800a5b445	bde2020/spark-base:3.1.2-hadoop3.2	Running	4040:4040	7 hours ago	
<input type="checkbox"/>	docker-spark	-	Running (8/8)			
<input type="checkbox"/>	spark-worker-2 3f42194a56d6	bde2020/spark-worker:3.1.2-hadoop3.2	Running	8072:8071	7 hours ago	
<input type="checkbox"/>	spark-worker-1 f7156e344aa2	bde2020/spark-worker:3.1.2-hadoop3.2	Running	8071:8071	7 hours ago	
<input type="checkbox"/>	resourcemanager 4537fd70ec85	bde2020/hadoop-resourcemanager:2.0.0-hadoop3.2	Running		7 hours ago	
<input type="checkbox"/>	historyserver b7ed07601fb0	bde2020/hadoop-historyserver:2.0.0-hadoop3.2.1-jav	Running		7 hours ago	
<input type="checkbox"/>	spark-master 381a106af556	bde2020/spark-master:3.1.2-hadoop3.2	Running	7077:7077 8070:8070	7 hours ago	
<input type="checkbox"/>	datanode 9eea97053df2	bde2020/hadoop-datanode:2.0.0-hadoop3.2.1-java8	Running		7 hours ago	
<input type="checkbox"/>	namenode dba155360717	bde2020/hadoop-namenode:2.0.0-hadoop3.2.1-java8	Running	9000:9000 9870:9870	7 hours ago	
<input type="checkbox"/>	nodemanager					

Showing 10 items

# Aplikacija na klasteru Spark Docker container-a (BDE)

```
bash-5.0# /spark/bin/spark-submit --master local[*] project_1.py hdfs://namenode:9000/dir_proj_1 40 60 -120 -104 2010-10-15 2010-10-31
```

```
bash-5.0# /spark/bin/spark-submit --master local[*] project_1.py data_proj_1/data_proj_1 40 60 -120 -104 2010-10-15 2010-10-31
```

```
23/01/26 22:03:32 INFO TaskSchedulerImpl: Killing all running tasks in stage 40: Stage finished
23/01/26 22:03:32 INFO DAGScheduler: Job 20 finished: showString at NativeMethodAccessorImpl.java:0, took 0.907657 s
23/01/26 22:03:32 INFO CodeGenerator: Code generated in 5.601293 ms
```

location_id	avg_time_spent
64b925364ac71005a...	103.5
31c4ef9bbd5f8ebe2...	163.65
e5acc4dbe743d713b...	173.0
d9ba7fd062dd11de9...	168.0
db3e5b3257cb11dea...	226.5
8ec18f7210b811deb...	170.75
3033e2d7a560b5a5c...	168.6
ca54a3d67bc011dda...	191.5
cc8795999b01637cf...	128.0
7b9b133c85199c71d...	164.0
f4491cf2fe7c88fb3...	193.0
abc4c6f75f5453484...	241.85714285714286
d2c833d0d13d1c037...	340.0
6ce8b78370a37d7b7...	148.0
f79c95e5936a31191...	99.5
6bdc8d41bcf11dea...	28.0
4d9b5a6f08993325b...	35.0
6d540229c0547a981...	226.33333333333334
fd8eacc4b1cb11dda...	172.55729166666666
cf86f32c17e211dea...	263.0

only showing top 20 rows



# Vreme potrebno za izvršenje aplikacije

- Izvršavanje aplikacije na lokalnoj mašini

```
Vreme potrebno za izvršenje aplikacije: 51.465081453323364
```

- Izvršavanje aplikacije u Docker-u

```
Vreme potrebno za izvršenje aplikacije: 75.35431814193726
```