Sistemi za obradu i analizu velike količine podataka - Prvi projekat

Nikola Božinović 1030

Podaci

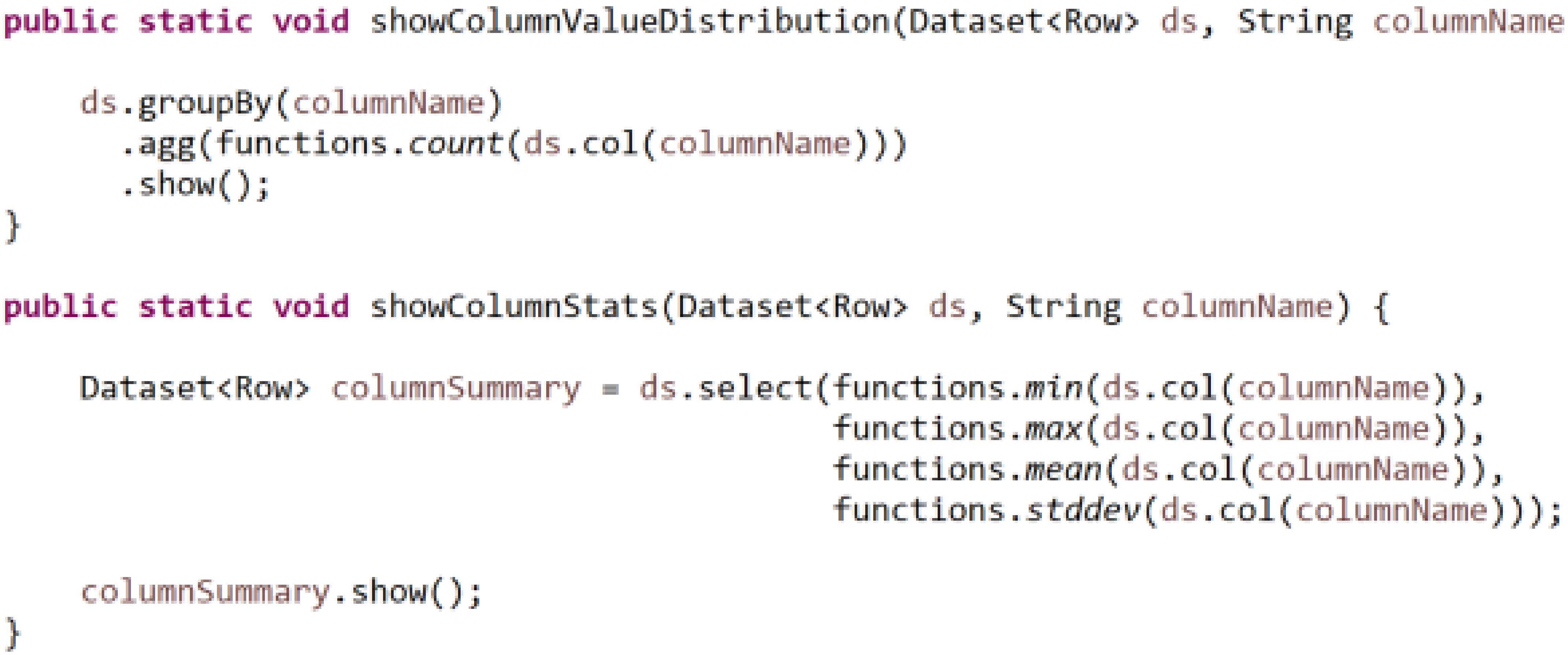
- **2014 New York City Taxi Trips**
- Dataset sadrži prostorno vremenske podatke koji se odnose na jedno putovanje taksijem u Njujorku u prva 2 meseca 2014. godine.
- Svi podaci se nalaze u jednom csv fajlu.

- **vendor_id**: A code indicating the TPEP provider that provided the record. Values are: 1= Creative Mobile Technologies, LLC and 2= VeriFone Inc.
- **pickup_datetime**: The date and time when the meter was engaged.
- **dropoff_datetime**: The date and time when the meter was disengaged.
- **passenger_count**: The number of passengers in the vehicle. This is a driver-entered value.
- **trip_distance**: The elapsed trip distance in miles reported by the taximeter.
- **pickup_longitude**: The longitude where the meter was engaged.
- **pickup_latitude**: The latitude where the meter was engaged.
- **rate_code**: The final rate code in effect at the end of the trip. Values are: 1= Standard rate, 2= JFK, 3= Newark, 4= Nassau or Westchester, 5= Negotiated fare and 6= Group ride.
- **store_and_fwd_flag**: This flag indicates whether the trip record was held in vehicle memory before sending to the vendor, aka "store and forward," because the vehicle did not have a connection to the server. Values are: Y= store and forward trip and N= not a store and forward trip.
- **dropoff_longitude** The longitude where the meter was disengaged.
- **dropoff_latitude**: The latitude where the meter was disengaged.
- **payment_type**: A numeric code signifying how the passenger paid for the trip. Values are: 1= Credit card, 2= Cash, 3= No charge, 4= Dispute, 5= Unknown, 6= Voided trip.
- **fare_amount**: The time-and-distance fare calculated by the meter.
- **surcharge**: Miscellaneous extras and surcharges.
- **mta_tax**: $0.50 MTA tax that is automatically triggered based on the metered rate in use.
- **tip_amount**: – This field is automatically populated for credit card tips. Cash tips are not included.
- **tolls_amount**: The total amount of all tolls paid in trip.
- **total_amount**: The total amount charged to passengers. Does not include cash tips.

Tehnologije

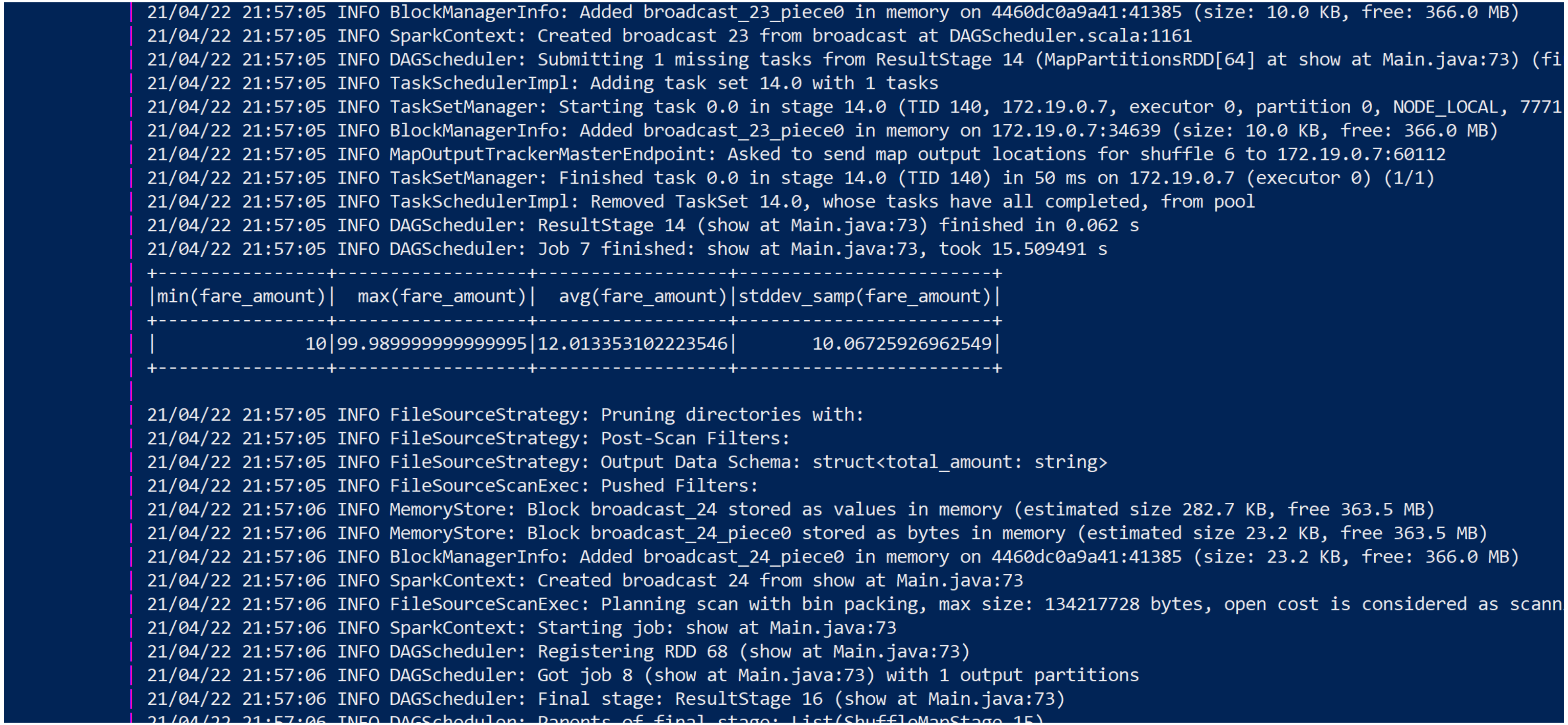
- Java 8
- Hadoop 2.7
- Spark 2.4.3
- Docker

Prikaz distribucije vrednosti i statistickih podataka jedne kolone



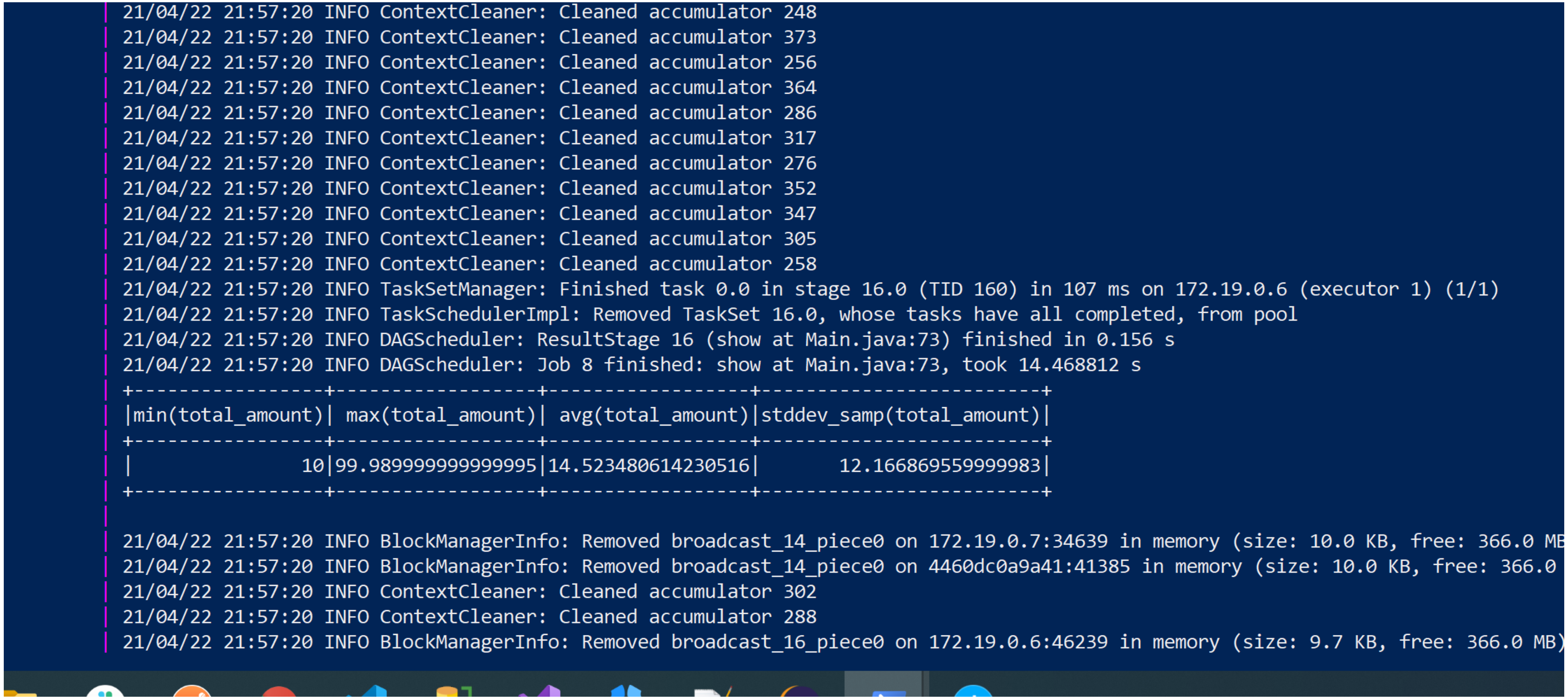
```
public static void showColumnValueDistribution(Dataset<Row> ds, String columnName) {

    ds.groupBy(columnName)
      .agg(functions.count(ds.col(columnName)))
      .show();
}

public static void showColumnStats(Dataset<Row> ds, String columnName) {

    Dataset<Row> columnSummary = ds.select(functions.min(ds.col(columnName)),
                                           functions.max(ds.col(columnName)),
                                           functions.mean(ds.col(columnName)),
                                           functions.stddev(ds.col(columnName)));

    columnSummary.show();
}
```

Fare amount statistika



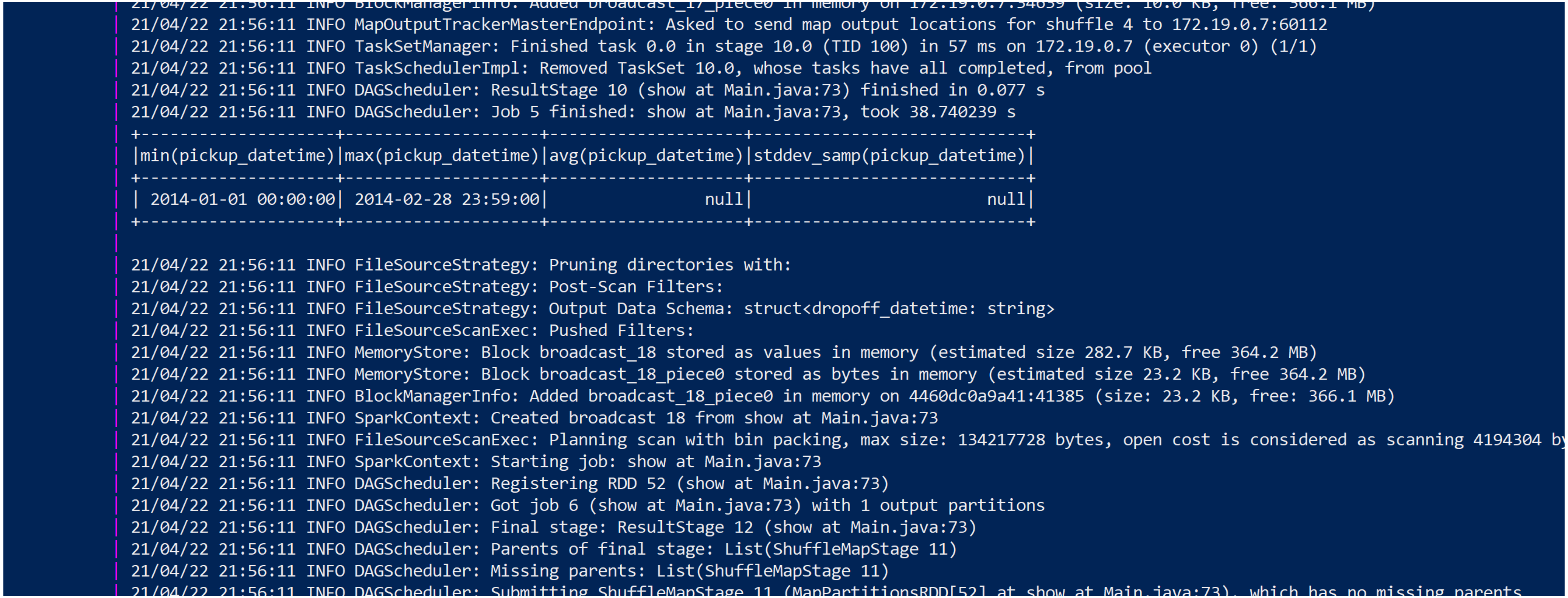
```
21/04/22 21:57:05 INFO BlockManagerInfo: Added broadcast_23_piece0 in memory on 4460dc0a9a41:41385 (size: 10.0 KB, free: 366.0 MB)
21/04/22 21:57:05 INFO SparkContext: Created broadcast 23 from broadcast at DAGScheduler.scala:1161
21/04/22 21:57:05 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 14 (MapPartitionsRDD[64] at show at Main.java:73) (fi
21/04/22 21:57:05 INFO TaskSchedulerImpl: Adding task set 14.0 with 1 tasks
21/04/22 21:57:05 INFO TaskSetManager: Starting task 0.0 in stage 14.0 (TID 140, 172.19.0.7, executor 0, partition 0, NODE_LOCAL, 7771
21/04/22 21:57:05 INFO BlockManagerInfo: Added broadcast_23_piece0 in memory on 172.19.0.7:34639 (size: 10.0 KB, free: 366.0 MB)
21/04/22 21:57:05 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 6 to 172.19.0.7:60112
21/04/22 21:57:05 INFO TaskSetManager: Finished task 0.0 in stage 14.0 (TID 140) in 50 ms on 172.19.0.7 (executor 0) (1/1)
21/04/22 21:57:05 INFO TaskSchedulerImpl: Removed TaskSet 14.0, whose tasks have all completed, from pool
21/04/22 21:57:05 INFO DAGScheduler: ResultStage 14 (show at Main.java:73) finished in 0.062 s
21/04/22 21:57:05 INFO DAGScheduler: Job 7 finished: show at Main.java:73, took 15.509491 s
+----------------+-----------------+------------------+------------------------+
|min(fare_amount)|  max(fare_amount)|   avg(fare_amount)|stddev_samp(fare_amount)|
+----------------+-----------------+------------------+------------------------+
|              10|99.98999999999995|12.013353102223546|       10.06725926962549|
+----------------+-----------------+------------------+------------------------+

21/04/22 21:57:05 INFO FileSourceStrategy: Pruning directories with:
21/04/22 21:57:05 INFO FileSourceStrategy: Post-Scan Filters:
21/04/22 21:57:05 INFO FileSourceStrategy: Output Data Schema: struct<total_amount: string>
21/04/22 21:57:05 INFO FileSourceScanExec: Pushed Filters:
21/04/22 21:57:06 INFO MemoryStore: Block broadcast_24 stored as values in memory (estimated size 282.7 KB, free 363.5 MB)
21/04/22 21:57:06 INFO MemoryStore: Block broadcast_24_piece0 stored as bytes in memory (estimated size 23.2 KB, free 363.5 MB)
21/04/22 21:57:06 INFO BlockManagerInfo: Added broadcast_24_piece0 in memory on 4460dc0a9a41:41385 (size: 23.2 KB, free: 366.0 MB)
21/04/22 21:57:06 INFO SparkContext: Created broadcast 24 from show at Main.java:73
21/04/22 21:57:06 INFO FileSourceScanExec: Planning scan with bin packing, max size: 134217728 bytes, open cost is considered as scann
21/04/22 21:57:06 INFO SparkContext: Starting job: show at Main.java:73
21/04/22 21:57:06 INFO DAGScheduler: Registering RDD 68 (show at Main.java:73)
21/04/22 21:57:06 INFO DAGScheduler: Got job 8 (show at Main.java:73) with 1 output partitions
21/04/22 21:57:06 INFO DAGScheduler: Final stage: ResultStage 16 (show at Main.java:73)
21/04/22 21:57:06 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 15)
```

Total Amount statistika



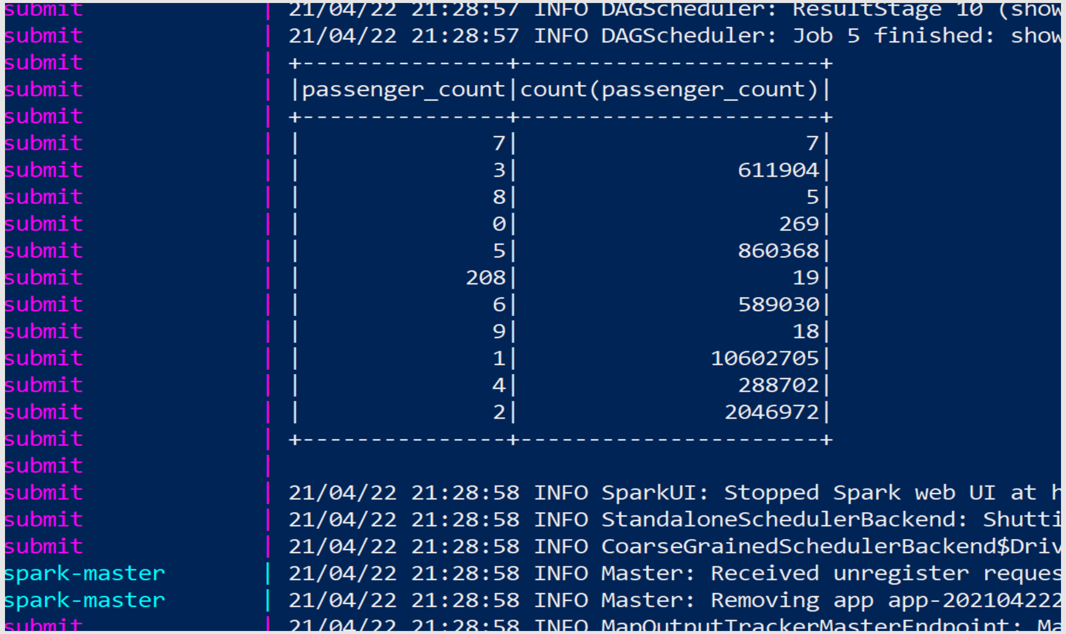
```
21/04/22 21:57:20 INFO ContextCleaner: Cleaned accumulator 248
21/04/22 21:57:20 INFO ContextCleaner: Cleaned accumulator 373
21/04/22 21:57:20 INFO ContextCleaner: Cleaned accumulator 256
21/04/22 21:57:20 INFO ContextCleaner: Cleaned accumulator 364
21/04/22 21:57:20 INFO ContextCleaner: Cleaned accumulator 286
21/04/22 21:57:20 INFO ContextCleaner: Cleaned accumulator 317
21/04/22 21:57:20 INFO ContextCleaner: Cleaned accumulator 276
21/04/22 21:57:20 INFO ContextCleaner: Cleaned accumulator 352
21/04/22 21:57:20 INFO ContextCleaner: Cleaned accumulator 347
21/04/22 21:57:20 INFO ContextCleaner: Cleaned accumulator 305
21/04/22 21:57:20 INFO ContextCleaner: Cleaned accumulator 258
21/04/22 21:57:20 INFO TaskSetManager: Finished task 0.0 in stage 16.0 (TID 160) in 107 ms on 172.19.0.6 (executor 1) (1/1)
21/04/22 21:57:20 INFO TaskSchedulerImpl: Removed TaskSet 16.0, whose tasks have all completed, from pool
21/04/22 21:57:20 INFO DAGScheduler: ResultStage 16 (show at Main.java:73) finished in 0.156 s
21/04/22 21:57:20 INFO DAGScheduler: Job 8 finished: show at Main.java:73, took 14.468812 s
+-----------------+------------------+------------------+-------------------------+
|min(total_amount)| max(total_amount)| avg(total_amount)|stddev_samp(total_amount)|
+-----------------+------------------+------------------+-------------------------+
|               10|99.989999999999995|14.523480614230516|       12.166869559999983|
+-----------------+------------------+------------------+-------------------------+


21/04/22 21:57:20 INFO BlockManagerInfo: Removed broadcast_14_piece0 on 172.19.0.7:34639 in memory (size: 10.0 KB, free: 366.0 MB
21/04/22 21:57:20 INFO BlockManagerInfo: Removed broadcast_14_piece0 on 4460dc0a9a41:41385 in memory (size: 10.0 KB, free: 366.0
21/04/22 21:57:20 INFO ContextCleaner: Cleaned accumulator 302
21/04/22 21:57:20 INFO ContextCleaner: Cleaned accumulator 288
21/04/22 21:57:20 INFO BlockManagerInfo: Removed broadcast_16_piece0 on 172.19.0.6:46239 in memory (size: 9.7 KB, free: 366.0 MB)
```

Pickup datetime statistika

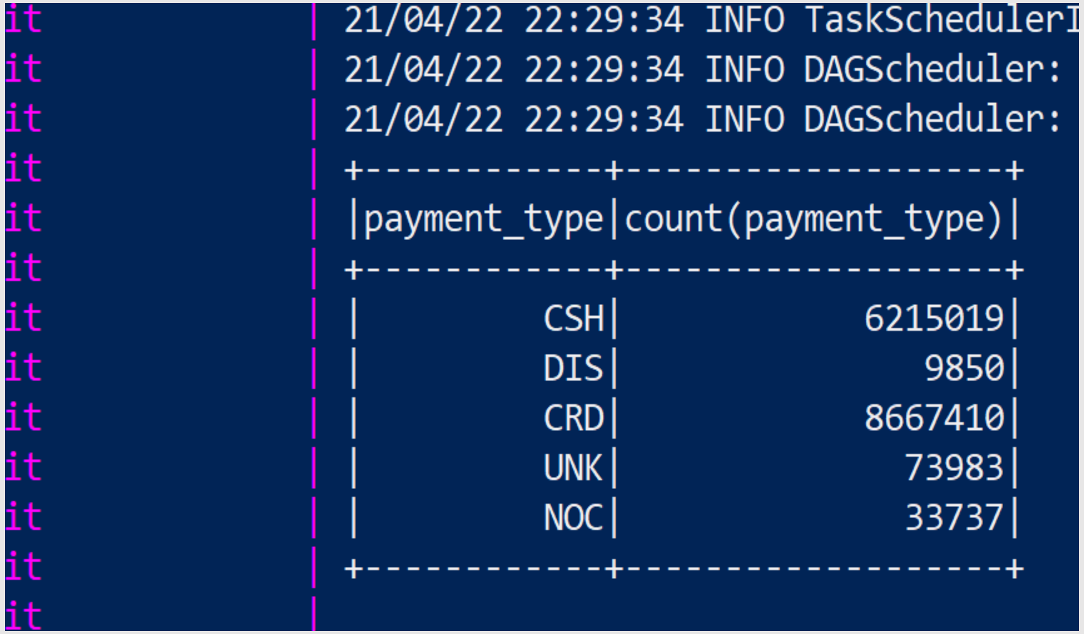


```
21/04/22 21:56:11 INFO BlockManagerInfo: Added broadcast_17_piece0 in memory on 172.19.0.7:34639 (size: 10.0 KB, free: 366.1 MB)
21/04/22 21:56:11 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 4 to 172.19.0.7:60112
21/04/22 21:56:11 INFO TaskSetManager: Finished task 0.0 in stage 10.0 (TID 100) in 57 ms on 172.19.0.7 (executor 0) (1/1)
21/04/22 21:56:11 INFO TaskSchedulerImpl: Removed TaskSet 10.0, whose tasks have all completed, from pool
21/04/22 21:56:11 INFO DAGScheduler: ResultStage 10 (show at Main.java:73) finished in 0.077 s
21/04/22 21:56:11 INFO DAGScheduler: Job 5 finished: show at Main.java:73, took 38.740239 s
+--------------------+--------------------+--------------------+----------------------------+
|min(pickup_datetime)|max(pickup_datetime)|avg(pickup_datetime)|stddev_samp(pickup_datetime)|
+--------------------+--------------------+--------------------+----------------------------+
| 2014-01-01 00:00:00| 2014-02-28 23:59:00|                null|                        null|
+--------------------+--------------------+--------------------+----------------------------+

21/04/22 21:56:11 INFO FileSourceStrategy: Pruning directories with:
21/04/22 21:56:11 INFO FileSourceStrategy: Post-Scan Filters:
21/04/22 21:56:11 INFO FileSourceStrategy: Output Data Schema: struct<dropoff_datetime: string>
21/04/22 21:56:11 INFO FileSourceScanExec: Pushed Filters:
21/04/22 21:56:11 INFO MemoryStore: Block broadcast_18 stored as values in memory (estimated size 282.7 KB, free 364.2 MB)
21/04/22 21:56:11 INFO MemoryStore: Block broadcast_18_piece0 stored as bytes in memory (estimated size 23.2 KB, free 364.2 MB)
21/04/22 21:56:11 INFO BlockManagerInfo: Added broadcast_18_piece0 in memory on 4460dc0a9a41:41385 (size: 23.2 KB, free: 366.1 MB)
21/04/22 21:56:11 INFO SparkContext: Created broadcast 18 from show at Main.java:73
21/04/22 21:56:11 INFO FileSourceScanExec: Planning scan with bin packing, max size: 134217728 bytes, open cost is considered as scanning 4194304 by
21/04/22 21:56:11 INFO SparkContext: Starting job: show at Main.java:73
21/04/22 21:56:11 INFO DAGScheduler: Registering RDD 52 (show at Main.java:73)
21/04/22 21:56:11 INFO DAGScheduler: Got job 6 (show at Main.java:73) with 1 output partitions
21/04/22 21:56:11 INFO DAGScheduler: Final stage: ResultStage 12 (show at Main.java:73)
21/04/22 21:56:11 INFO DAGScheduler: Parents of final stage: List(ShuffleMapStage 11)
21/04/22 21:56:11 INFO DAGScheduler: Missing parents: List(ShuffleMapStage 11)
21/04/22 21:56:11 INFO DAGScheduler: Submitting ShuffleMapStage 11 (MapPartitionsRDD[52] at show at Main.java:73), which has no missing parents
```

Raspodela po broju putnika i načinu plaćanja



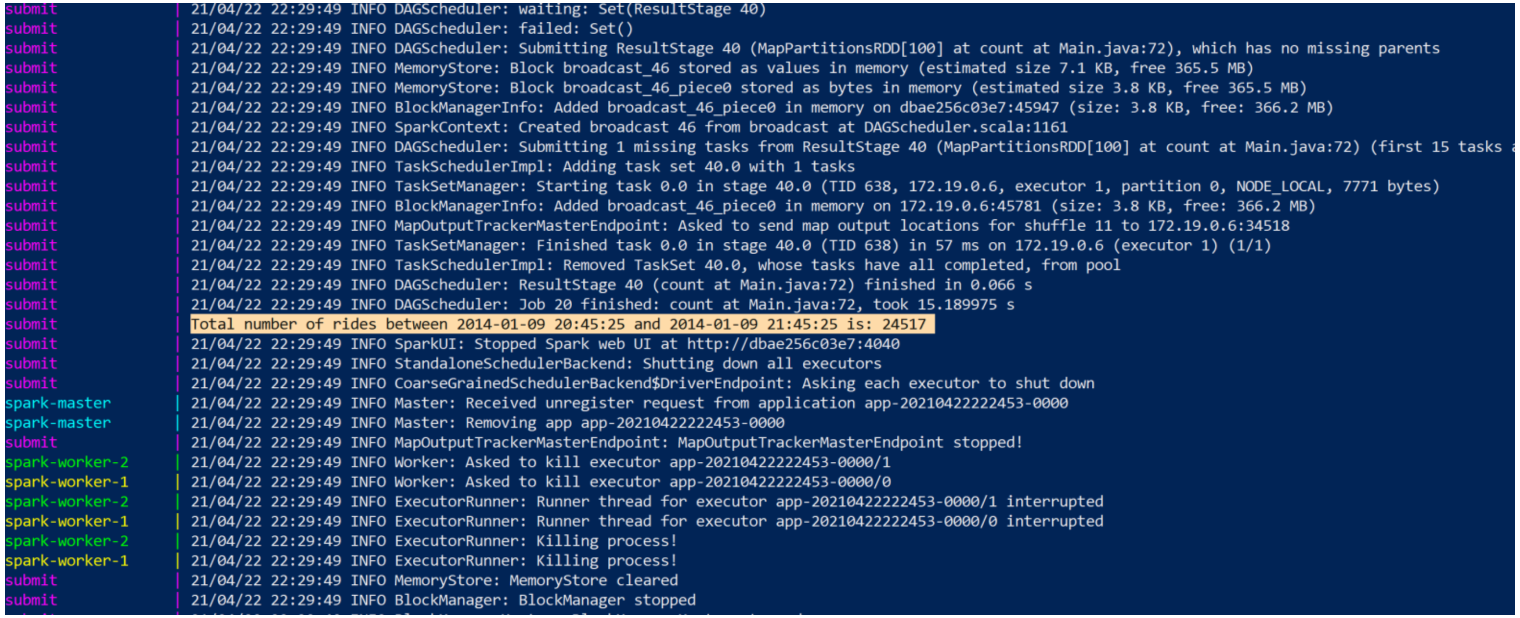
```
submit              | 21/04/22 21:28:57 INFO DAGScheduler: ResultStage 10 (show
submit              | 21/04/22 21:28:57 INFO DAGScheduler: Job 5 finished: show
submit              | +---------------+----------------------+
submit              | |passenger_count|count(passenger_count)|
submit              | +---------------+----------------------+
submit              | |              7|                     7|
submit              | |              3|                611904|
submit              | |              8|                     5|
submit              | |              0|                   269|
submit              | |              5|                860368|
submit              | |            208|                    19|
submit              | |              6|                589030|
submit              | |              9|                    18|
submit              | |              1|              10602705|
submit              | |              4|                288702|
submit              | |              2|               2046972|
submit              | +---------------+----------------------+
submit              |
submit              | 21/04/22 21:28:58 INFO SparkUI: Stopped Spark web UI at h
submit              | 21/04/22 21:28:58 INFO StandaloneSchedulerBackend: Shutti
submit              | 21/04/22 21:28:58 INFO CoarseGrainedSchedulerBackend$Driv
spark-master        | 21/04/22 21:28:58 INFO Master: Received unregister reques
spark-master        | 21/04/22 21:28:58 INFO Master: Removing app app-202104222
submit              | 21/04/22 21:28:58 INFO MapOutputTrackerMasterEndpoint: Ma
```



```
it             | 21/04/22 22:29:34 INFO TaskSchedulerI
it             | 21/04/22 22:29:34 INFO DAGScheduler:
it             | 21/04/22 22:29:34 INFO DAGScheduler:
it             | +------------+-------------------+
it             | |payment_type|count(payment_type)|
it             | +------------+-------------------+
it             | |         CSH|            6215019|
it             | |         DIS|               9850|
it             | |         CRD|            8667410|
it             | |         UNK|              73983|
it             | |         NOC|              33737|
it             | +------------+-------------------+
it             |
```

Broj vožnji u zadatom vremenskom periodu

```java
static void showNumberOfRidesInGivenTimePeriod(Dataset<Row> ds, String startDateTime, String endDateTime) {

    Long count = ds.filter(ds.col("pickup_datetime").geq(startDateTime)
                            .and(ds.col("dropoff_datetime").leq(endDateTime)))
                   .count();

    System.out.println(
            String.format("Total number of rides between %s and %s is: %d", startDateTime, endDateTime, count));
}
```

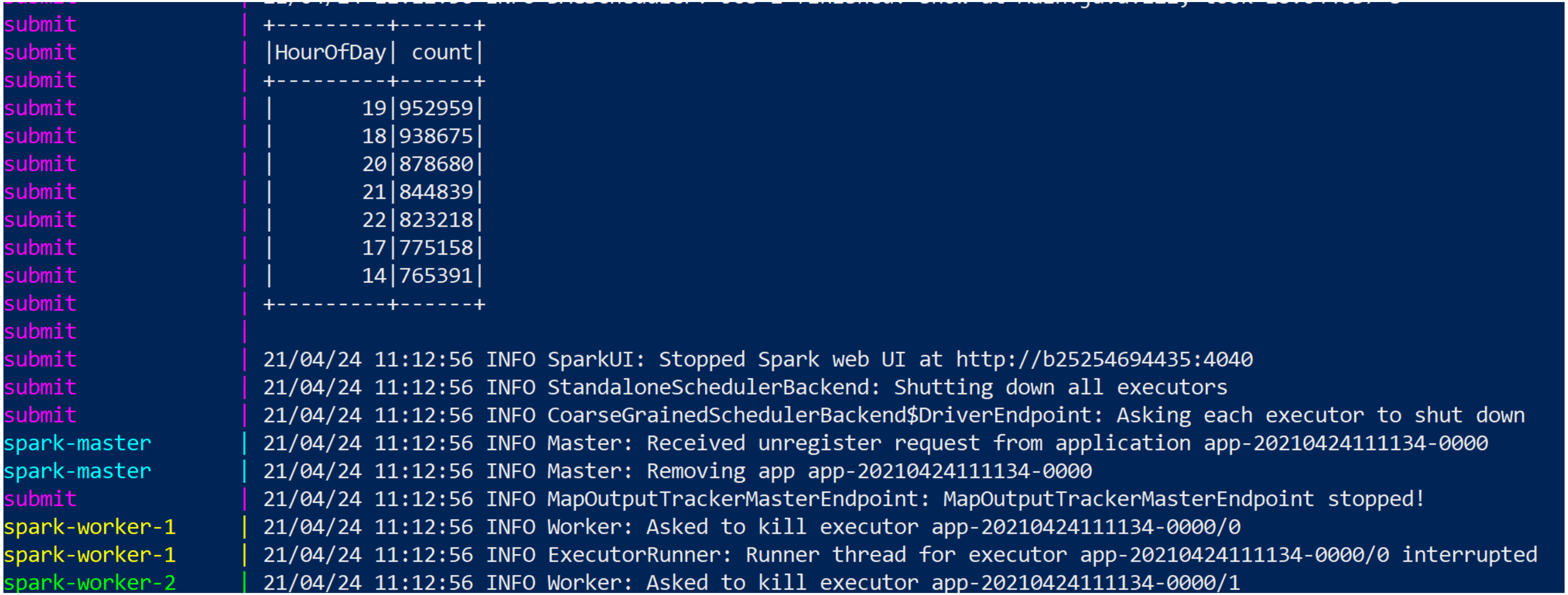


```
submit          | 21/04/22 22:29:49 INFO DAGScheduler: waiting: Set(ResultStage 40)
submit          | 21/04/22 22:29:49 INFO DAGScheduler: failed: Set()
submit          | 21/04/22 22:29:49 INFO DAGScheduler: Submitting ResultStage 40 (MapPartitionsRDD[100] at count at Main.java:72), which has no missing parents
submit          | 21/04/22 22:29:49 INFO MemoryStore: Block broadcast_46 stored as values in memory (estimated size 7.1 KB, free 365.5 MB)
submit          | 21/04/22 22:29:49 INFO MemoryStore: Block broadcast_46_piece0 stored as bytes in memory (estimated size 3.8 KB, free 365.5 MB)
submit          | 21/04/22 22:29:49 INFO BlockManagerInfo: Added broadcast_46_piece0 in memory on dbae256c03e7:45947 (size: 3.8 KB, free: 366.2 MB)
submit          | 21/04/22 22:29:49 INFO SparkContext: Created broadcast 46 from broadcast at DAGScheduler.scala:1161
submit          | 21/04/22 22:29:49 INFO DAGScheduler: Submitting 1 missing tasks from ResultStage 40 (MapPartitionsRDD[100] at count at Main.java:72) (first 15 tasks a
submit          | 21/04/22 22:29:49 INFO TaskSchedulerImpl: Adding task set 40.0 with 1 tasks
submit          | 21/04/22 22:29:49 INFO TaskSetManager: Starting task 0.0 in stage 40.0 (TID 638, 172.19.0.6, executor 1, partition 0, NODE_LOCAL, 7771 bytes)
submit          | 21/04/22 22:29:49 INFO BlockManagerInfo: Added broadcast_46_piece0 in memory on 172.19.0.6:45781 (size: 3.8 KB, free: 366.2 MB)
submit          | 21/04/22 22:29:49 INFO MapOutputTrackerMasterEndpoint: Asked to send map output locations for shuffle 11 to 172.19.0.6:34518
submit          | 21/04/22 22:29:49 INFO TaskSetManager: Finished task 0.0 in stage 40.0 (TID 638) in 57 ms on 172.19.0.6 (executor 1) (1/1)
submit          | 21/04/22 22:29:49 INFO TaskSchedulerImpl: Removed TaskSet 40.0, whose tasks have all completed, from pool
submit          | 21/04/22 22:29:49 INFO DAGScheduler: ResultStage 40 (count at Main.java:72) finished in 0.066 s
submit          | 21/04/22 22:29:49 INFO DAGScheduler: Job 20 finished: count at Main.java:72, took 15.189975 s
submit          | Total number of rides between 2014-01-09 20:45:25 and 2014-01-09 21:45:25 is: 24517
submit          | 21/04/22 22:29:49 INFO SparkUI: Stopped Spark web UI at http://dbae256c03e7:4040
submit          | 21/04/22 22:29:49 INFO StandaloneSchedulerBackend: Shutting down all executors
submit          | 21/04/22 22:29:49 INFO CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each executor to shut down
spark-master    | 21/04/22 22:29:49 INFO Master: Received unregister request from application app-20210422222453-0000
spark-master    | 21/04/22 22:29:49 INFO Master: Removing app app-20210422222453-0000
submit          | 21/04/22 22:29:49 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
spark-worker-2  | 21/04/22 22:29:49 INFO Worker: Asked to kill executor app-20210422222453-0000/1
spark-worker-1  | 21/04/22 22:29:49 INFO Worker: Asked to kill executor app-20210422222453-0000/0
spark-worker-2  | 21/04/22 22:29:49 INFO ExecutorRunner: Runner thread for executor app-20210422222453-0000/1 interrupted
spark-worker-1  | 21/04/22 22:29:49 INFO ExecutorRunner: Runner thread for executor app-20210422222453-0000/0 interrupted
spark-worker-2  | 21/04/22 22:29:49 INFO ExecutorRunner: Killing process!
spark-worker-1  | 21/04/22 22:29:49 INFO ExecutorRunner: Killing process!
submit          | 21/04/22 22:29:49 INFO MemoryStore: MemoryStore cleared
submit          | 21/04/22 22:29:49 INFO BlockManager: BlockManager stopped
```

Sati u danu u kojima je ima najviše vožnji

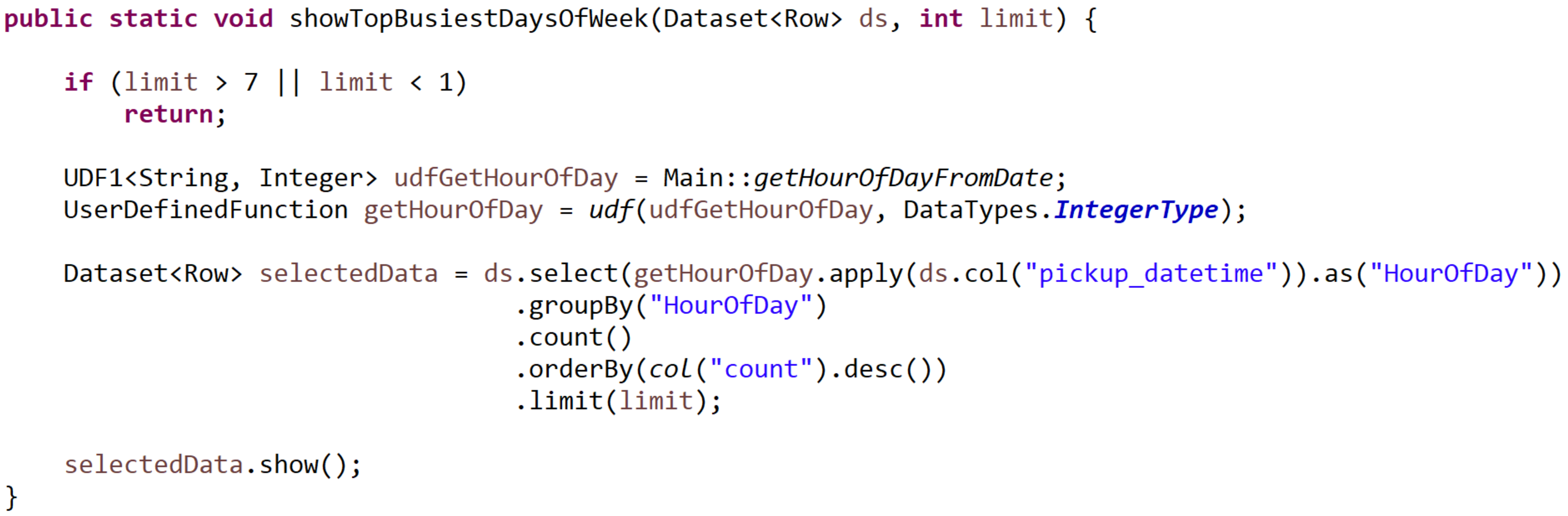
```java
public static void showTopBusiestHoursOfDay(Dataset<Row> ds, int limit) {

    if (limit > 24 || limit < 1)
        return;

    UDF1<String, Integer> udfGetDayOfWeak = Main::getDayOfWeekFromDate;
    UserDefinedFunction getDayOfWeak = udf(udfGetDayOfWeak, DataTypes.IntegerType);

    Dataset<Row> selectedData = ds.select(getDayOfWeak.apply(ds.col("pickup_datetime")).as("DayOfWeak"))
                                  .groupBy("DayOfWeak")
                                  .count()
                                  .orderBy(col("count").desc())
                                  .limit(limit);

    selectedData.show();
}
```

Prikaz 7 najprometnijih sati u danu



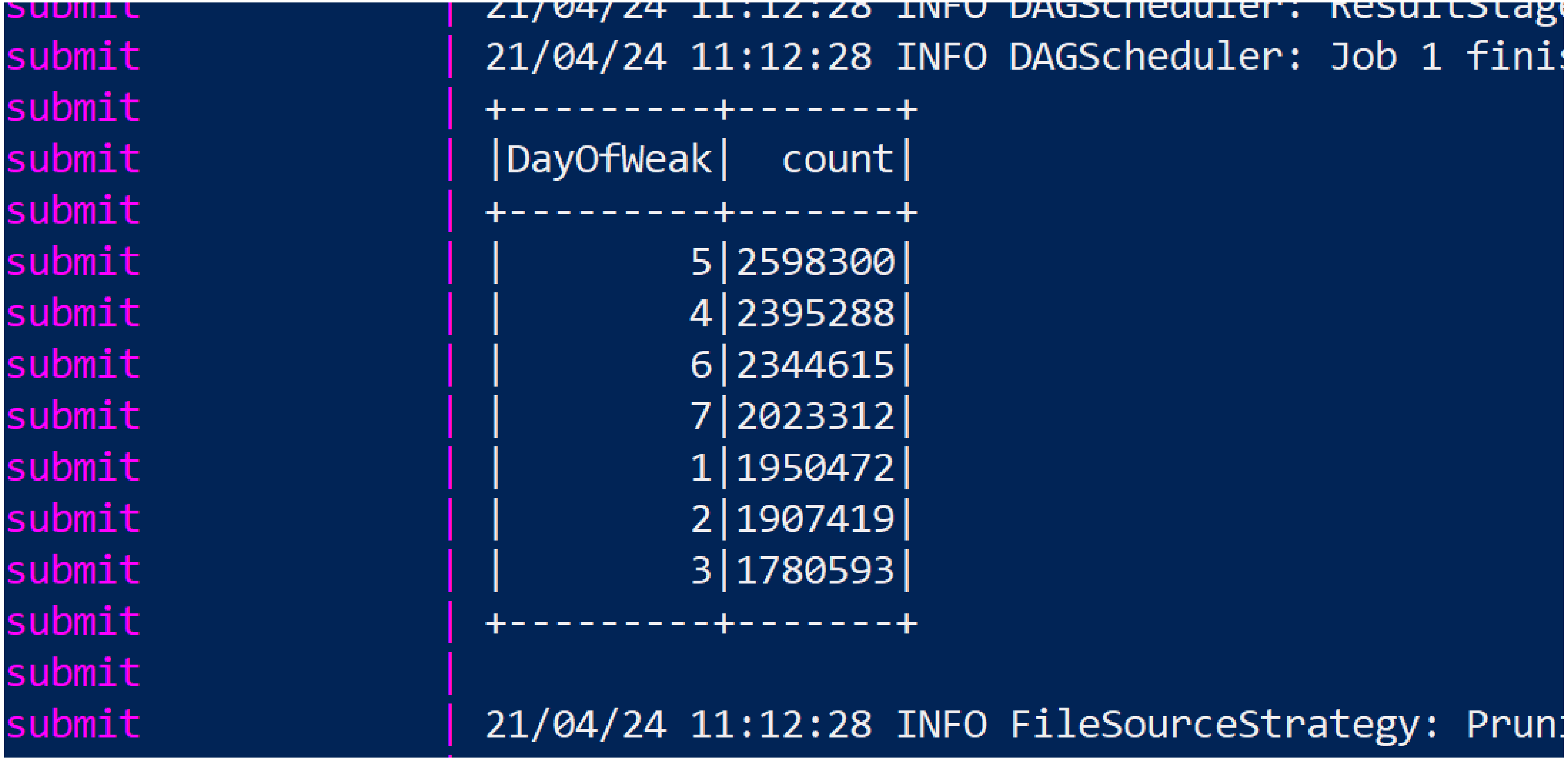
```
submit           | +---------+------+
submit           | |HourOfDay| count|
submit           | +---------+------+
submit           | |       19|952959|
submit           | |       18|938675|
submit           | |       20|878680|
submit           | |       21|844839|
submit           | |       22|823218|
submit           | |       17|775158|
submit           | |       14|765391|
submit           | +---------+------+
submit           |
submit           | 21/04/24 11:12:56 INFO SparkUI: Stopped Spark web UI at http://b25254694435:4040
submit           | 21/04/24 11:12:56 INFO StandaloneSchedulerBackend: Shutting down all executors
submit           | 21/04/24 11:12:56 INFO CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each executor to shut down
spark-master     | 21/04/24 11:12:56 INFO Master: Received unregister request from application app-20210424111134-0000
spark-master     | 21/04/24 11:12:56 INFO Master: Removing app app-20210424111134-0000
submit           | 21/04/24 11:12:56 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
spark-worker-1   | 21/04/24 11:12:56 INFO Worker: Asked to kill executor app-20210424111134-0000/0
spark-worker-1   | 21/04/24 11:12:56 INFO ExecutorRunner: Runner thread for executor app-20210424111134-0000/0 interrupted
spark-worker-2   | 21/04/24 11:12:56 INFO Worker: Asked to kill executor app-20210424111134-0000/1
```

Dani u nedelji u kojima ima najviše vožnji



```java
public static void showTopBusiestDaysOfWeek(Dataset<Row> ds, int limit) {

    if (limit > 7 || limit < 1)
        return;

    UDF1<String, Integer> udfGetHourOfDay = Main::getHourOfDayFromDate;
    UserDefinedFunction getHourOfDay = udf(udfGetHourOfDay, DataTypes.IntegerType);

    Dataset<Row> selectedData = ds.select(getHourOfDay.apply(ds.col("pickup_datetime")).as("HourOfDay"))
                                  .groupBy("HourOfDay")
                                  .count()
                                  .orderBy(col("count").desc())
                                  .limit(limit);

    selectedData.show();
}
```

Distribucija raspodele podataka na dane u nedelji

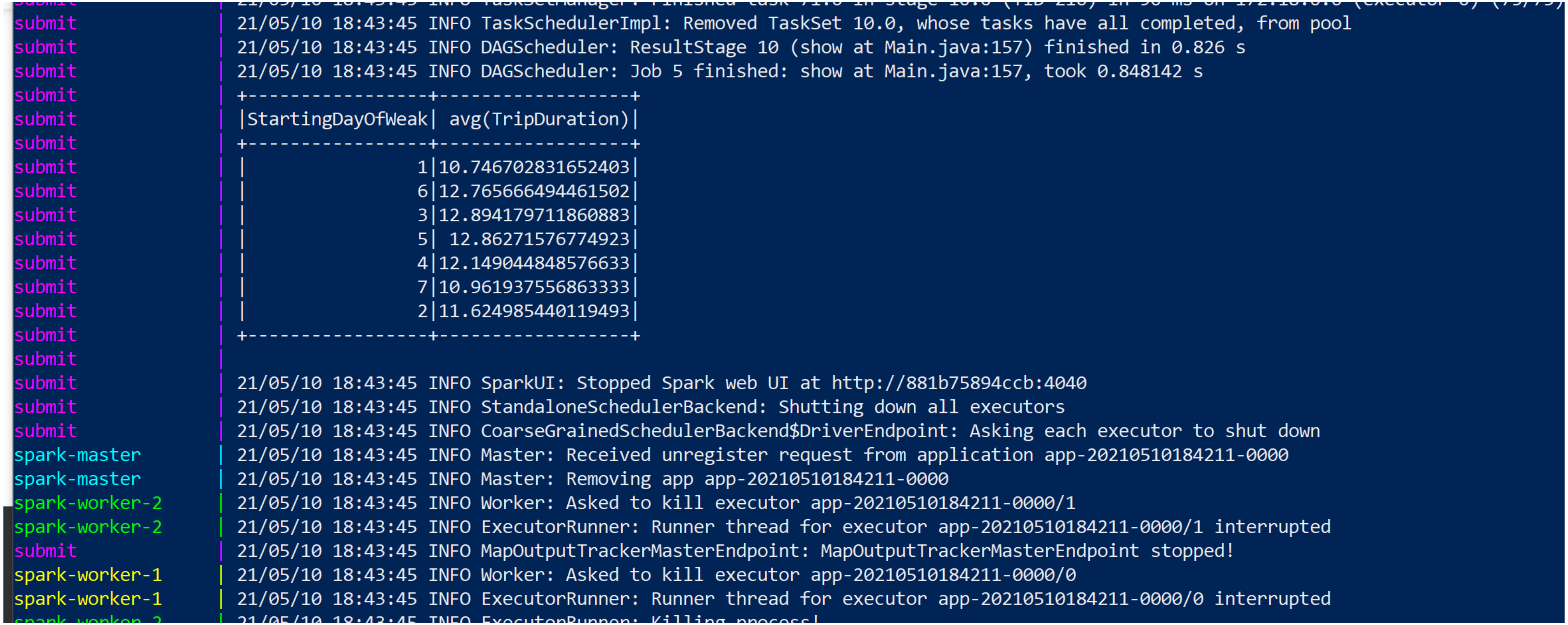


```
submit                  | 21/04/24 11:12:28 INFO DAGScheduler: ResultStage
submit                  | 21/04/24 11:12:28 INFO DAGScheduler: Job 1 finis
submit                  | +---------+-------+
submit                  | |DayOfWeak|  count|
submit                  | +---------+-------+
submit                  | |        5|2598300|
submit                  | |        4|2395288|
submit                  | |        6|2344615|
submit                  | |        7|2023312|
submit                  | |        1|1950472|
submit                  | |        2|1907419|
submit                  | |        3|1780593|
submit                  | +---------+-------+
submit                  |
submit                  | 21/04/24 11:12:28 INFO FileSourceStrategy: Pruni
```

Prikaz prosečnog trajanja vožnje po danima u nedelji

```
public static void showAverageTripDurationByDayInWeak(Dataset<Row> ds)
{
    UDF1<String, Integer> udfGetDayOfWeak = Main::getDayOfWeekFromDate;
    UserDefinedFunction getDayOfWeak = udf(udfGetDayOfWeak, DataTypes.IntegerType);


    UDF2<String, String, Long> udfGetTripDuration = Main::geTripDurationTime;
    UserDefinedFunction getTripDuration = udf(udfGetTripDuration, DataTypes.LongType);


    ds.select(getDayOfWeak.apply(ds.col("pickup_datetime")).as("StartingDayOfWeak"),
              getDayOfWeak.apply(ds.col("dropoff_datetime")).as("EndingDayOfWeak"),
              getTripDuration.apply(ds.col("pickup_datetime"), ds.col("dropoff_datetime")).as("TripDuration"))
    .filter(col("StartingDayOfWeak").equalTo(col("EndingDayOfWeak")))
    .groupBy("StartingDayOfWeak")
    .agg(avg("TripDuration"))
    .show();
}
```

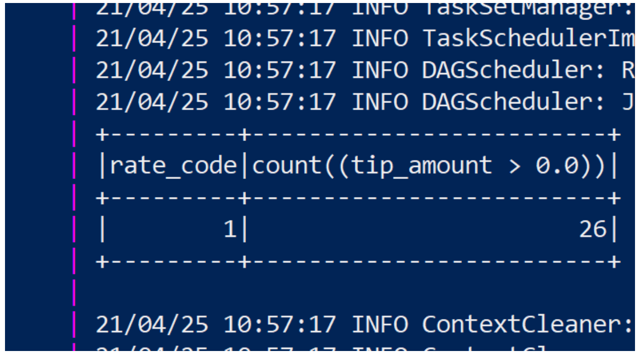
Prikaz prosečne dužine trajanja vožnje u minutima za dane u nedelji



```
submit              | 21/05/10 18:43:45 INFO TaskSetManager: Finished task 74.0 in stage 10.0 (TID 216) in 30 ms on 172.18.0.6 (executor 0) (75/75)
submit              | 21/05/10 18:43:45 INFO TaskSchedulerImpl: Removed TaskSet 10.0, whose tasks have all completed, from pool
submit              | 21/05/10 18:43:45 INFO DAGScheduler: ResultStage 10 (show at Main.java:157) finished in 0.826 s
submit              | 21/05/10 18:43:45 INFO DAGScheduler: Job 5 finished: show at Main.java:157, took 0.848142 s
submit              | +-----------------+------------------+
submit              | |StartingDayOfWeak| avg(TripDuration)|
submit              | +-----------------+------------------+
submit              | |                1|10.746702831652403|
submit              | |                6|12.765666494461502|
submit              | |                3|12.894179711860883|
submit              | |                5| 12.86271576774923|
submit              | |                4|12.149044848576633|
submit              | |                7|10.961937556863333|
submit              | |                2|11.624985440119493|
submit              | +-----------------+------------------+
submit              |
submit              | 21/05/10 18:43:45 INFO SparkUI: Stopped Spark web UI at http://881b75894ccb:4040
submit              | 21/05/10 18:43:45 INFO StandaloneSchedulerBackend: Shutting down all executors
submit              | 21/05/10 18:43:45 INFO CoarseGrainedSchedulerBackend$DriverEndpoint: Asking each executor to shut down
spark-master        | 21/05/10 18:43:45 INFO Master: Received unregister request from application app-20210510184211-0000
spark-master        | 21/05/10 18:43:45 INFO Master: Removing app app-20210510184211-0000
spark-worker-2      | 21/05/10 18:43:45 INFO Worker: Asked to kill executor app-20210510184211-0000/1
spark-worker-2      | 21/05/10 18:43:45 INFO ExecutorRunner: Runner thread for executor app-20210510184211-0000/1 interrupted
submit              | 21/05/10 18:43:45 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
spark-worker-1      | 21/05/10 18:43:45 INFO Worker: Asked to kill executor app-20210510184211-0000/0
spark-worker-1      | 21/05/10 18:43:45 INFO ExecutorRunner: Runner thread for executor app-20210510184211-0000/0 interrupted
spark-worker-2      | 21/05/10 18:43:45 INFO ExecutorRunner: Killing process!
```

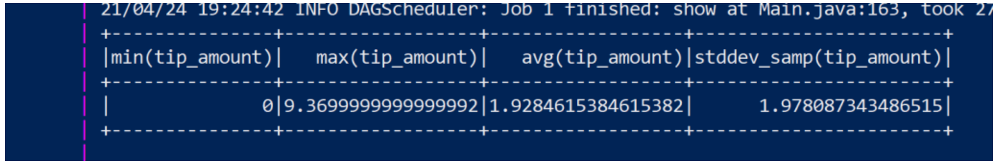
Prikaz raspodele tip amount-a po rate code-u i statistike tip amount-a na zadatoj lokaciji

```
public static void showTipAmountCountByRateCode(Dataset<Row> ds, double minLongitude, double maxLongitude,
    double minLatitude, double maxLatitude, String startTime, String endTime) {

  ds.filter(col("pickup_longitude").geq(minLongitude)
       .and(col("dropoff_longitude").leq(maxLongitude))
       .and(col("pickup_latitude").geq(minLatitude))
       .and(col("dropoff_latitude").leq(maxLatitude))
       .and(col("pickup_datetime").geq(lit(startTime)))
       .and(col("dropoff_datetime").leq(lit(endTime))))
   .groupBy(col("rate_code"))
   .agg(count(ds.col("tip_amount").gt(0.0)))
   .show();
}
```



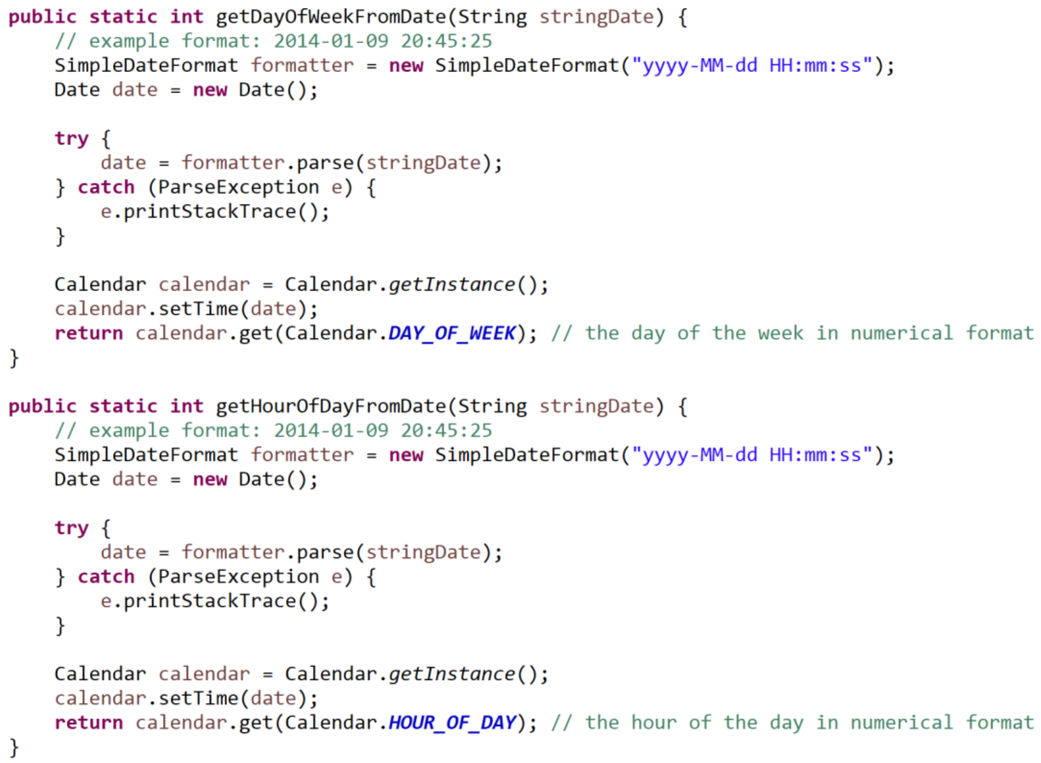
```
21/04/25 10:57:17 INFO TaskSetManager:
21/04/25 10:57:17 INFO TaskSchedulerIm
21/04/25 10:57:17 INFO DAGScheduler: R
21/04/25 10:57:17 INFO DAGScheduler: J
+---------+----------------------+
|rate_code|count((tip_amount > 0.0))|
+---------+----------------------+
|        1|                    26|
+---------+----------------------+

21/04/25 10:57:17 INFO ContextCleaner:
```

```
public static void showTipAmountStatsOnLocation(Dataset<Row> ds, double minLongitude, double maxLongitude,
    double minLatitude, double maxLatitude, String startTime, String endTime) {

  ds.filter(col("pickup_longitude").geq(minLongitude)
      .and(col("dropoff_longitude").leq(maxLongitude))
      .and(col("pickup_latitude").geq(minLatitude))
      .and(col("dropoff_latitude").leq(maxLatitude))
      .and(col("pickup_datetime").geq(functions.lit(startTime)))
      .and(col("dropoff_datetime").leq(functions.lit(endTime))))
      .select(functions.min("tip_amount"), functions.max("tip_amount"),
          functions.mean("tip_amount"), functions.stddev("tip_amount"))
  .show();

}
```

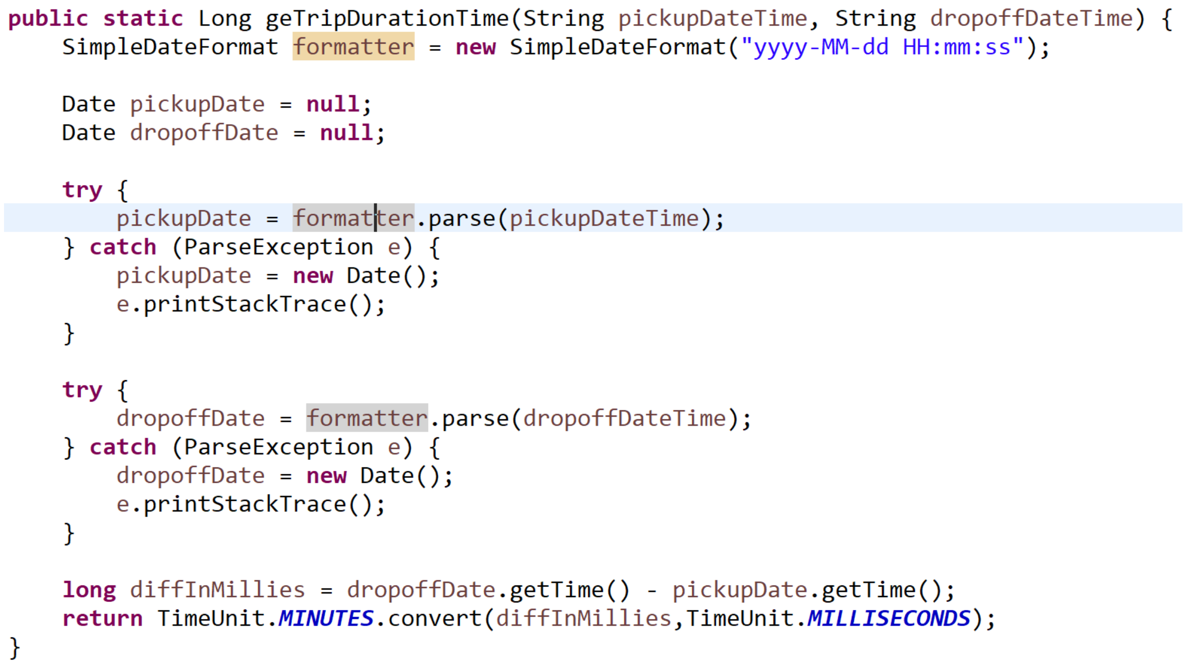


```
21/04/24 19:24:42 INFO DAGScheduler: Job 1 finished: show at Main.java:163, took 27
+---------------+-----------------+-----------------+---------------------+
|min(tip_amount)|  max(tip_amount)|  avg(tip_amount)|stddev_samp(tip_amount)|
+---------------+-----------------+-----------------+---------------------+
|              0|9.3699999999999992|1.9284615384615382|     1.978087343486515|
+---------------+-----------------+-----------------+---------------------+
```

Pomoćne funkcije



```java
public static int getDayOfWeekFromDate(String stringDate) {
    // example format: 2014-01-09 20:45:25
    SimpleDateFormat formatter = new SimpleDateFormat("yyyy-MM-dd HH:mm:ss");
    Date date = new Date();

    try {
        date = formatter.parse(stringDate);
    } catch (ParseException e) {
        e.printStackTrace();
    }

    Calendar calendar = Calendar.getInstance();
    calendar.setTime(date);
    return calendar.get(Calendar.DAY_OF_WEEK); // the day of the week in numerical format
}

public static int getHourOfDayFromDate(String stringDate) {
    // example format: 2014-01-09 20:45:25
    SimpleDateFormat formatter = new SimpleDateFormat("yyyy-MM-dd HH:mm:ss");
    Date date = new Date();

    try {
        date = formatter.parse(stringDate);
    } catch (ParseException e) {
        e.printStackTrace();
    }

    Calendar calendar = Calendar.getInstance();
    calendar.setTime(date);
    return calendar.get(Calendar.HOUR_OF_DAY); // the hour of the day in numerical format
}
```



```java
public static Long getTripDurationTime(String pickupDateTime, String dropoffDateTime) {
    SimpleDateFormat formatter = new SimpleDateFormat("yyyy-MM-dd HH:mm:ss");

    Date pickupDate = null;
    Date dropoffDate = null;

    try {
        pickupDate = formatter.parse(pickupDateTime);
    } catch (ParseException e) {
        pickupDate = new Date();
        e.printStackTrace();
    }

    try {
        dropoffDate = formatter.parse(dropoffDateTime);
    } catch (ParseException e) {
        dropoffDate = new Date();
        e.printStackTrace();
    }

    long diffInMillies = dropoffDate.getTime() - pickupDate.getTime();
    return TimeUnit.MINUTES.convert(diffInMillies,TimeUnit.MILLISECONDS);
}
```