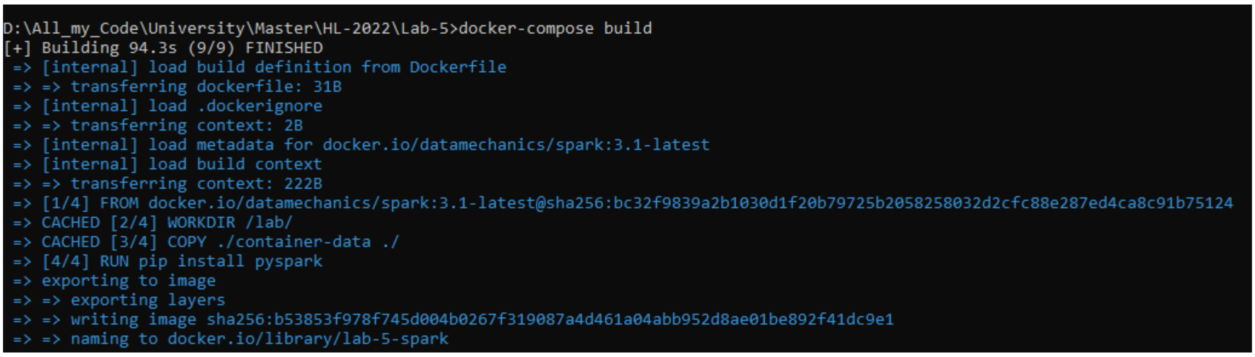
Step 1:

Data was generated using data-generation.py script. Data.txt file was generated based on London postcodes.csv

Step 2:

Build our image using docker-compose build

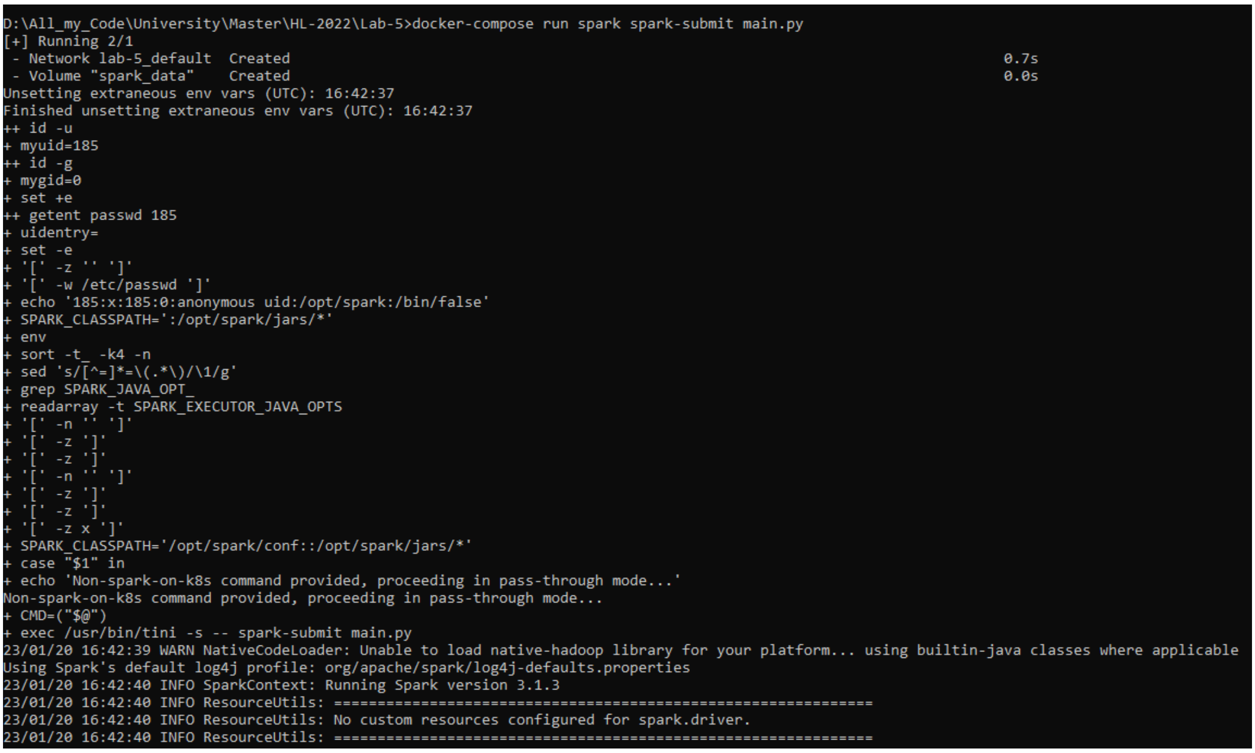


```
D:\All_my_Code\University\Master\HL-2022\Lab-5>docker-compose build
[+] Building 94.3s (9/9) FINISHED
 => [internal] load build definition from Dockerfile
 => => transferring dockerfile: 31B
 => [internal] load .dockerignore
 => => transferring context: 2B
 => [internal] load metadata for docker.io/datamechanics/spark:3.1-latest
 => [internal] load build context
 => => transferring context: 222B
 => [1/4] FROM docker.io/datamechanics/spark:3.1-latest@sha256:bc32f9839a2b1030d1f20b79725b2058258032d2cfc88e287ed4ca8c91b75124
 => CACHED [2/4] WORKDIR /lab/
 => CACHED [3/4] COPY ./container-data ./
 => [4/4] RUN pip install pyspark
 => exporting to image
 => => exporting layers
 => => writing image sha256:b53853f978f745d004b0267f319087a4d461a04abb952d8ae01be892f41dc9e1
 => => naming to docker.io/library/lab-5-spark
```

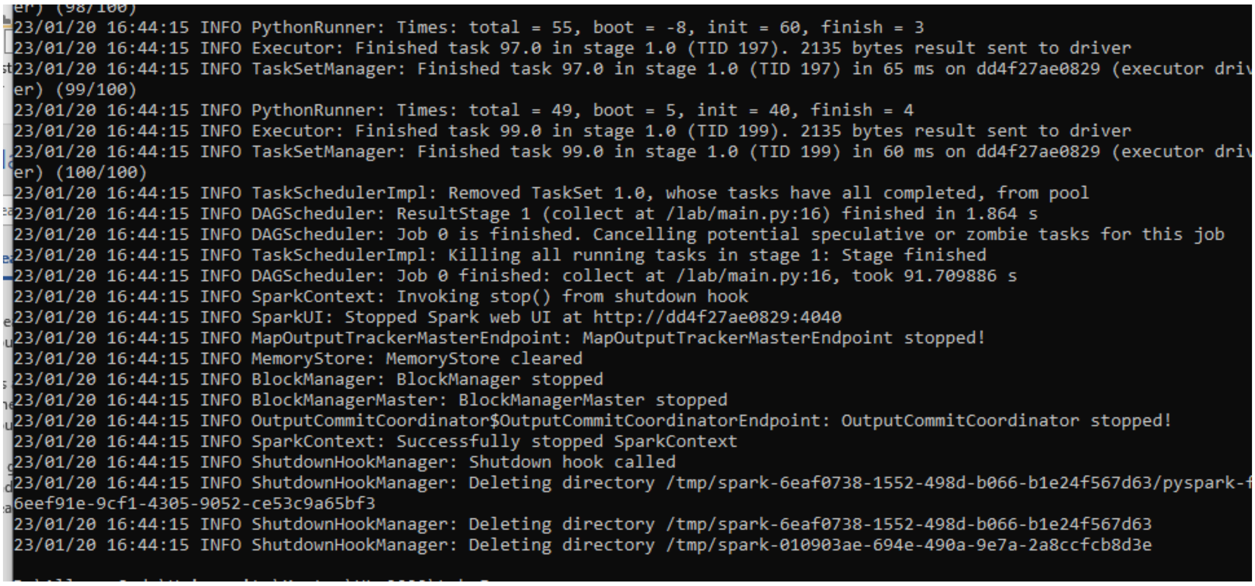
Step 3:

Execute main.py inside the container to get the needed output.

docker-compose run spark spark-submit main.py



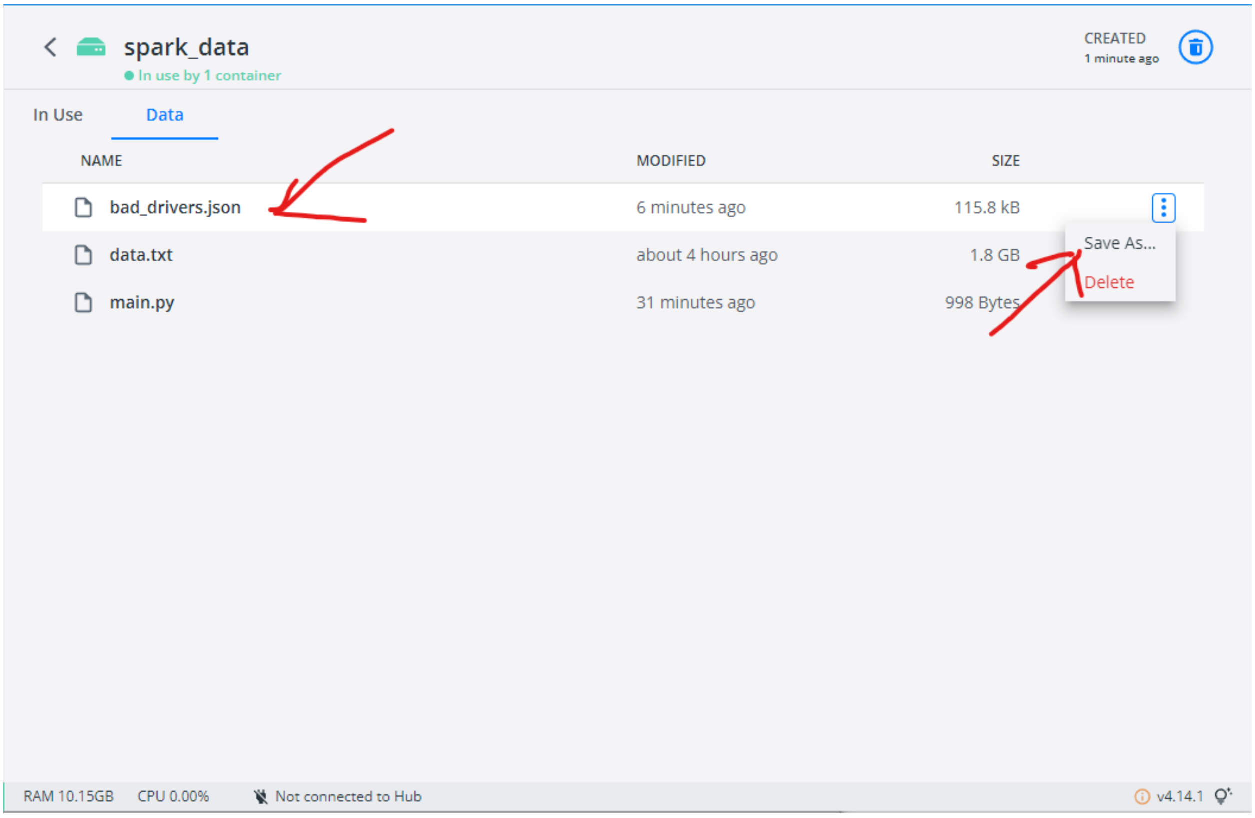
```
D:\All_my_Code\University\Master\HL-2022\Lab-5>docker-compose run spark spark-submit main.py
[+] Running 2/1
 - Network lab-5_default  Created                                                                                  0.7s
 - Volume "spark_data"    Created                                                                                  0.0s
Unsetting extraneous env vars (UTC): 16:42:37
Finished unsetting extraneous env vars (UTC): 16:42:37
++ id -u
+ myuid=185
++ id -g
+ mygid=0
+ set +e
++ getent passwd 185
+ uidentry=
+ set -e
+ '[' -z '' ']'
+ '[' -w /etc/passwd ']'
+ echo '185:x:185:0:anonymous uid:/opt/spark:/bin/false'
+ SPARK_CLASSPATH=':/opt/spark/jars/*'
+ env
+ sort -t_ -k4 -n
+ sed 's/[^=]*=\(.*\)/\1/g'
+ grep SPARK_JAVA_OPT_
+ readarray -t SPARK_EXECUTOR_JAVA_OPTS
+ '[' -n '' ']'
+ '[' -z ']'
+ '[' -z ']'
+ '[' -n '' ']'
+ '[' -z ']'
+ '[' -z ']'
+ '[' -z x ']'
+ SPARK_CLASSPATH='/opt/spark/conf::/opt/spark/jars/*'
+ case "$1" in
+ echo 'Non-spark-on-k8s command provided, proceeding in pass-through mode...'
Non-spark-on-k8s command provided, proceeding in pass-through mode...
+ CMD=("$@")
+ exec /usr/bin/tini -s -- spark-submit main.py
23/01/20 16:42:39 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
23/01/20 16:42:40 INFO SparkContext: Running Spark version 3.1.3
23/01/20 16:42:40 INFO ResourceUtils: ==============================================================
23/01/20 16:42:40 INFO ResourceUtils: No custom resources configured for spark.driver.
23/01/20 16:42:40 INFO ResourceUtils: ==============================================================
```



```
er) (98/100)
23/01/20 16:44:15 INFO PythonRunner: Times: total = 55, boot = -8, init = 60, finish = 3
23/01/20 16:44:15 INFO Executor: Finished task 97.0 in stage 1.0 (TID 197). 2135 bytes result sent to driver
23/01/20 16:44:15 INFO TaskSetManager: Finished task 97.0 in stage 1.0 (TID 197) in 65 ms on dd4f27ae0829 (executor driv
er) (99/100)
23/01/20 16:44:15 INFO PythonRunner: Times: total = 49, boot = 5, init = 40, finish = 4
23/01/20 16:44:15 INFO Executor: Finished task 99.0 in stage 1.0 (TID 199). 2135 bytes result sent to driver
23/01/20 16:44:15 INFO TaskSetManager: Finished task 99.0 in stage 1.0 (TID 199) in 60 ms on dd4f27ae0829 (executor driv
er) (100/100)
23/01/20 16:44:15 INFO TaskSchedulerImpl: Removed TaskSet 1.0, whose tasks have all completed, from pool
23/01/20 16:44:15 INFO DAGScheduler: ResultStage 1 (collect at /lab/main.py:16) finished in 1.864 s
23/01/20 16:44:15 INFO DAGScheduler: Job 0 is finished. Cancelling potential speculative or zombie tasks for this job
23/01/20 16:44:15 INFO TaskSchedulerImpl: Killing all running tasks in stage 1: Stage finished
23/01/20 16:44:15 INFO DAGScheduler: Job 0 finished: collect at /lab/main.py:16, took 91.709886 s
23/01/20 16:44:15 INFO SparkContext: Invoking stop() from shutdown hook
23/01/20 16:44:15 INFO SparkUI: Stopped Spark web UI at http://dd4f27ae0829:4040
23/01/20 16:44:15 INFO MapOutputTrackerMasterEndpoint: MapOutputTrackerMasterEndpoint stopped!
23/01/20 16:44:15 INFO MemoryStore: MemoryStore cleared
23/01/20 16:44:15 INFO BlockManager: BlockManager stopped
23/01/20 16:44:15 INFO BlockManagerMaster: BlockManagerMaster stopped
23/01/20 16:44:15 INFO OutputCommitCoordinator$OutputCommitCoordinatorEndpoint: OutputCommitCoordinator stopped!
23/01/20 16:44:15 INFO SparkContext: Successfully stopped SparkContext
23/01/20 16:44:15 INFO ShutdownHookManager: Shutdown hook called
23/01/20 16:44:15 INFO ShutdownHookManager: Deleting directory /tmp/spark-6eaf0738-1552-498d-b066-b1e24f567d63/pyspark-f
6eef91e-9cf1-4305-9052-ce53c9a65bf3
23/01/20 16:44:15 INFO ShutdownHookManager: Deleting directory /tmp/spark-6eaf0738-1552-498d-b066-b1e24f567d63
23/01/20 16:44:15 INFO ShutdownHookManager: Deleting directory /tmp/spark-010903ae-694e-490a-9e7a-2a8ccfcb8d3e
```

Step 4:

Get data from the container using Docker Desktop UI.



spark_data

● In use by 1 container

CREATED
1 minute ago

NAME	MODIFIED	SIZE
bad_drivers.json	6 minutes ago	115.8 kB
data.txt	about 4 hours ago	1.8 GB
main.py	31 minutes ago	998 Bytes

RAM 10.15GB CPU 0.00% Not connected to Hub v4.14.1