Report HW#3

Team Member

Pawel Urbanowicz	108015016
Martin Ledl	108012012
Tobias Kick	108998413

Github Repository

The Github repository for this project can be found here:
https://github.com/mledl/BDMA_HW/tree/master/HW3
Please check the repository for results under “/data/results/”, because the limited upload size restricted us from uploading the result files to cyberclassroom.

Responsibilities

Pawel Urbanowicz

- Configure the Production Spark Environment and run the final solution for benchmark.
- Take care of task 2 (Min-Hashing)
- Contribute to report.

Martin Ledl

- Setup project to work on HW#3.
- Take care of task 3(Locality-Sensitive Hashing)
- Contribute to report.

Tobias Kick

- Take care of task 1(shingling)
- Optimizing different aspects of data processing in all 3 tasks
- Contribute to report.

Environment Setup

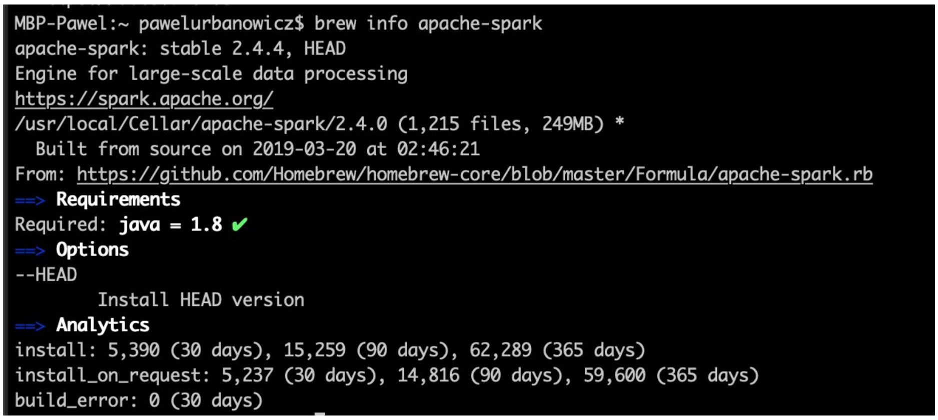
For local development we tested our code on a locally installed spark instance and for target stage we used Docker technology to wrap spark master instance, 2 spark workers instances and our Python script into separate containers. Software/Frameworks in use:

- Python 3.7 to write our code
- Spark version 2.4.4
- Hadoop version 3.2.1

- Docker engine version 18.09.2.

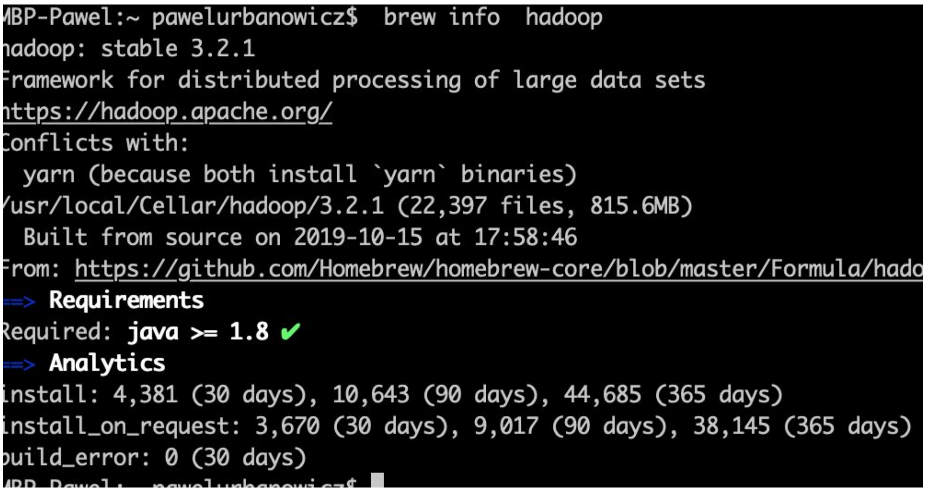
Environment setup for OSX

a. Install Spark
brew install apache-spark



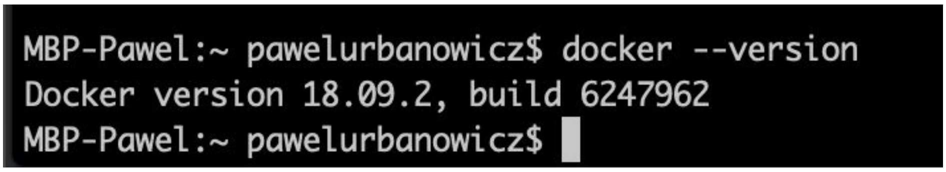
```
MBP-Pawel:~ pawelurbanowicz$ brew info apache-spark
apache-spark: stable 2.4.4, HEAD
Engine for large-scale data processing
https://spark.apache.org/
/usr/local/Cellar/apache-spark/2.4.0 (1,215 files, 249MB) *
  Built from source on 2019-03-20 at 02:46:21
From: https://github.com/Homebrew/homebrew-core/blob/master/Formula/apache-spark.rb
==> Requirements
Required: java = 1.8 ✔
==> Options
--HEAD
        Install HEAD version
==> Analytics
install: 5,390 (30 days), 15,259 (90 days), 62,289 (365 days)
install_on_request: 5,237 (30 days), 14,816 (90 days), 59,600 (365 days)
build_error: 0 (30 days)
```

b. Install Hadoop



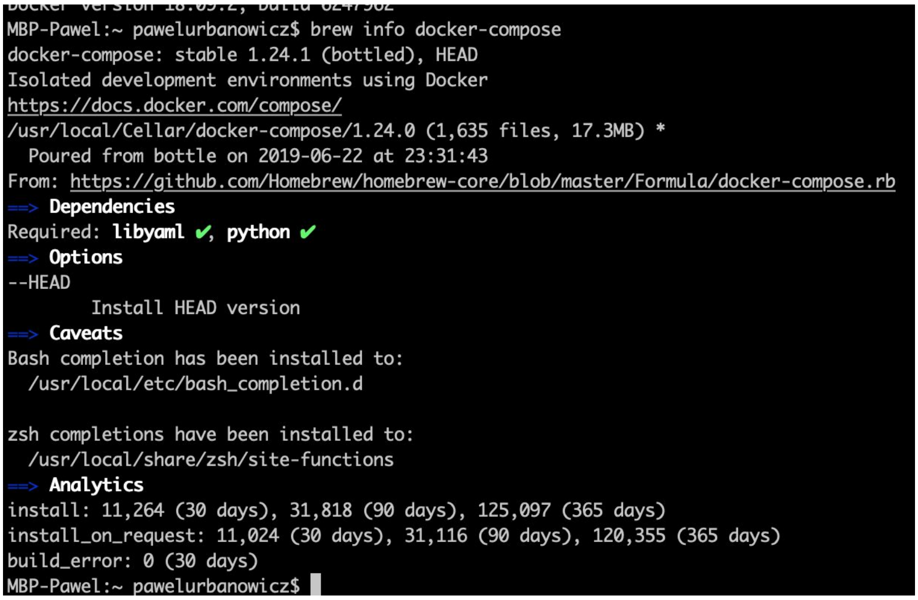
```
IBP-Pawel:~ pawelurbanowicz$  brew info  hadoop
nadoop: stable 3.2.1
ramework for distributed processing of large data sets
nttps://hadoop.apache.org/
onflicts with:
  yarn (because both install `yarn` binaries)
/usr/local/Cellar/hadoop/3.2.1 (22,397 files, 815.6MB)
  Built from source on 2019-10-15 at 17:58:46
rom: https://github.com/Homebrew/homebrew-core/blob/master/Formula/hado
==> Requirements
equired: java >= 1.8 ✔
==> Analytics
install: 4,381 (30 days), 10,643 (90 days), 44,685 (365 days)
install_on_request: 3,670 (30 days), 9,017 (90 days), 38,145 (365 days)
puild_error: 0 (30 days)
```

c. Install Docker Desktop for Mac
https://docs.docker.com/docker-for-mac/install/



```
MBP-Pawel:~ pawelurbanowicz$ docker --version
Docker version 18.09.2, build 6247962
MBP-Pawel:~ pawelurbanowicz$
```

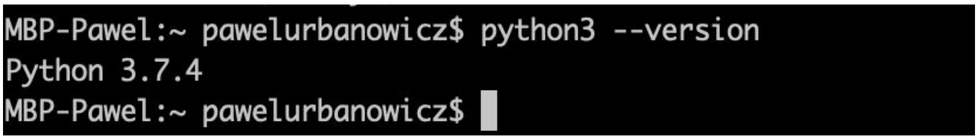
d. Install Docker Compose
   brew install docker-compose



```
Docker version 18.09.2, build 6247962
MBP-Pawel:~ pawelurbanowicz$ brew info docker-compose
docker-compose: stable 1.24.1 (bottled), HEAD
Isolated development environments using Docker
https://docs.docker.com/compose/
/usr/local/Cellar/docker-compose/1.24.0 (1,635 files, 17.3MB) *
  Poured from bottle on 2019-06-22 at 23:31:43
From: https://github.com/Homebrew/homebrew-core/blob/master/Formula/docker-compose.rb
==> Dependencies
Required: libyaml ✔, python ✔
==> Options
--HEAD
        Install HEAD version
==> Caveats
Bash completion has been installed to:
  /usr/local/etc/bash_completion.d

zsh completions have been installed to:
  /usr/local/share/zsh/site-functions
==> Analytics
install: 11,264 (30 days), 31,818 (90 days), 125,097 (365 days)
install_on_request: 11,024 (30 days), 31,116 (90 days), 120,355 (365 days)
build_error: 0 (30 days)
MBP-Pawel:~ pawelurbanowicz$
```

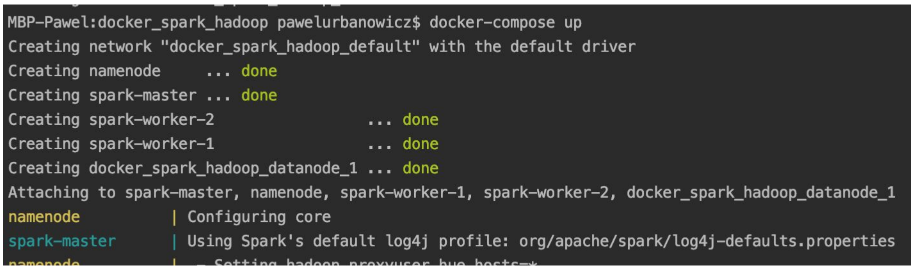
e. Install python 3
   brew install python



```
MBP-Pawel:~ pawelurbanowicz$ python3 --version
Python 3.7.4
MBP-Pawel:~ pawelurbanowicz$
```

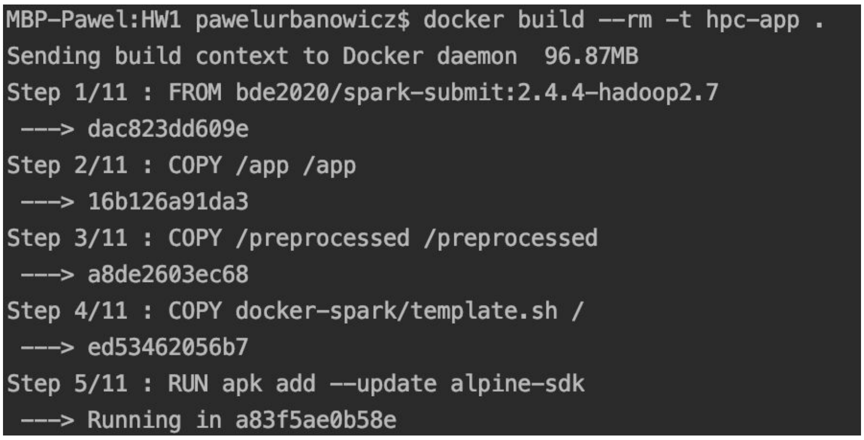
f. Clone code from repository
   git clone https://github.com/mledl/BDMA_HW

g. Open terminal and go to docker_spark_hadoop directory and run:
   docker-compose up



```
MBP-Pawel:docker_spark_hadoop pawelurbanowicz$ docker-compose up
Creating network "docker_spark_hadoop_default" with the default driver
Creating namenode     ... done
Creating spark-master ... done
Creating spark-worker-2                 ... done
Creating spark-worker-1                 ... done
Creating docker_spark_hadoop_datanode_1 ... done
Attaching to spark-master, namenode, spark-worker-1, spark-worker-2, docker_spark_hadoop_datanode_1
namenode          | Configuring core
spark-master      | Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
namenode          |  - Setting hadoop.proxyuser.hue.hosts=*
```

h. Go to HW2 directory to build image for python script
   docker build --rm -t app .

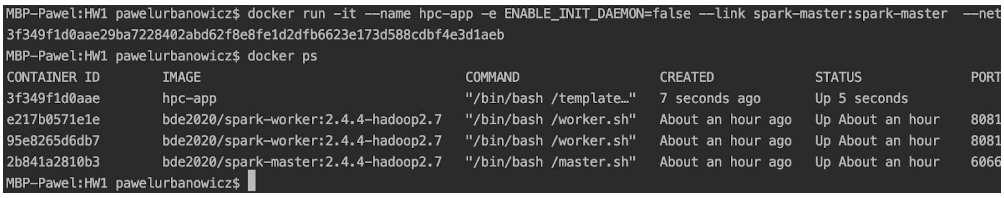


```
MBP-Pawel:HW1 pawelurbanowicz$ docker build --rm -t hpc-app .
Sending build context to Docker daemon  96.87MB
Step 1/11 : FROM bde2020/spark-submit:2.4.4-hadoop2.7
 ---> dac823dd609e
Step 2/11 : COPY /app /app
 ---> 16b126a91da3
Step 3/11 : COPY /preprocessed /preprocessed
 ---> a8de2603ec68
Step 4/11 : COPY docker-spark/template.sh /
 ---> ed53462056b7
Step 5/11 : RUN apk add --update alpine-sdk
 ---> Running in a83f5ae0b58e
```

It can take some time as some libraries must be built from sources

i. Add data to hadoop
docker cp data namenode:data
docker exec -it namenode bash
hadoop fs -put /data /data

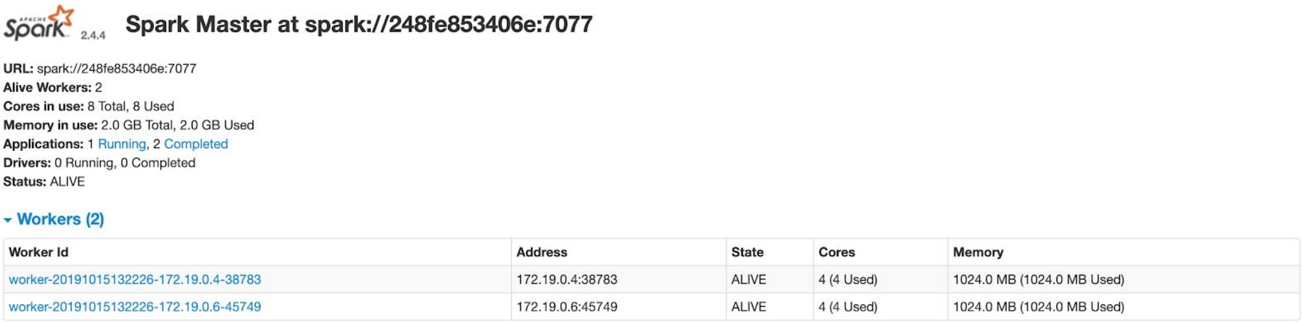
j. Run previously build image
docker run -it --name app -e ENABLE_INIT_DAEMON=false --link spark-master:spark-master --net docker_spark_hadoop_default -d app



```
MBP-Pawel:HW1 pawelurbanowicz$ docker run -it --name hpc-app -e ENABLE_INIT_DAEMON=false --link spark-master:spark-master  --net
3f349f1d0aae29ba7228402abd62f8e8fe1d2dfb6623e173d588cdbf4e3d1aeb
MBP-Pawel:HW1 pawelurbanowicz$ docker ps
CONTAINER ID        IMAGE                                COMMAND                  CREATED             STATUS              PORT
3f349f1d0aae        hpc-app                              "/bin/bash /template…"   7 seconds ago       Up 5 seconds
e217b0571e1e        bde2020/spark-worker:2.4.4-hadoop2.7   "/bin/bash /worker.sh"   About an hour ago   Up About an hour    8081
95e8265d6db7        bde2020/spark-worker:2.4.4-hadoop2.7   "/bin/bash /worker.sh"   About an hour ago   Up About an hour    8081
2b841a2810b3        bde2020/spark-master:2.4.4-hadoop2.7   "/bin/bash /master.sh"   About an hour ago   Up About an hour    6066
MBP-Pawel:HW1 pawelurbanowicz$ 
```

Result of setup:

http://localhost:8089/



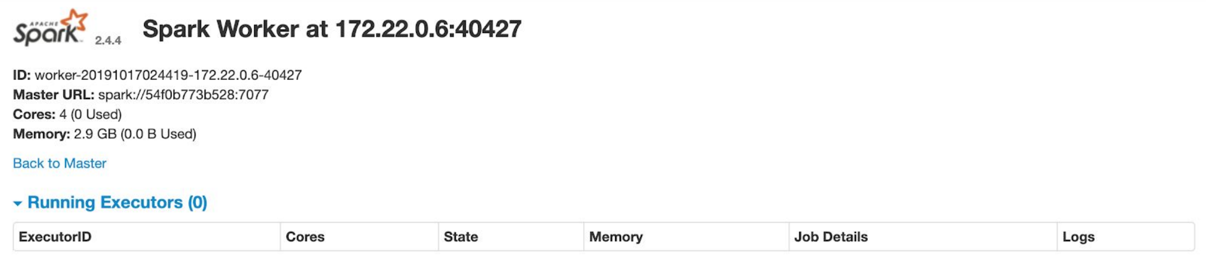
Spark Master at spark://248fe853406e:7077

URL: spark://248fe853406e:7077
Alive Workers: 2
Cores in use: 8 Total, 8 Used
Memory in use: 2.0 GB Total, 2.0 GB Used
Applications: 1 Running, 2 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

Worker Id	Address	State	Cores	Memory
worker-20191015132226-172.19.0.4-38783	172.19.0.4:38783	ALIVE	4 (4 Used)	1024.0 MB (1024.0 MB Used)
worker-20191015132226-172.19.0.6-45749	172.19.0.6:45749	ALIVE	4 (4 Used)	1024.0 MB (1024.0 MB Used)

http://localhost:8084/ and http://localhost:8085/



Spark Worker at 172.22.0.6:40427

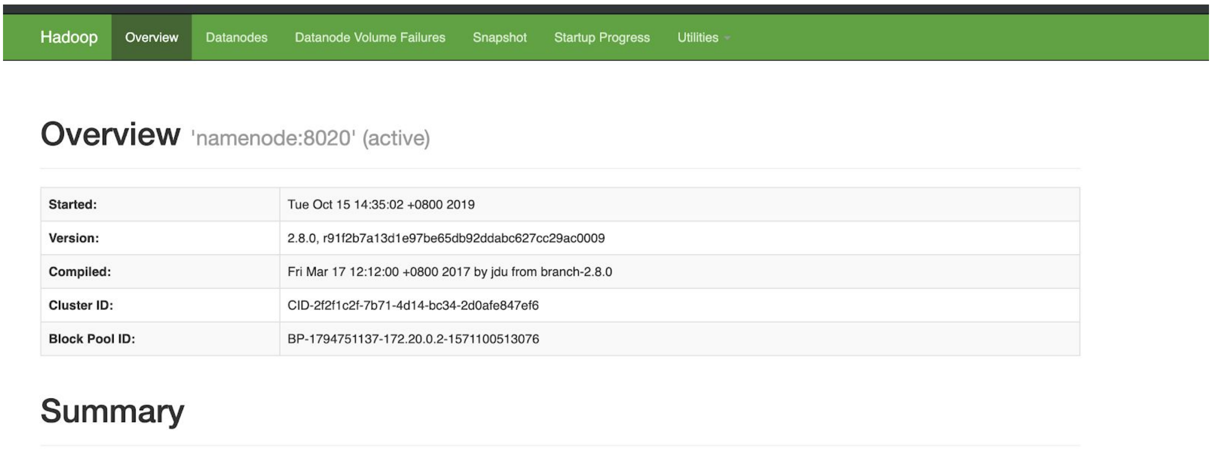
ID: worker-20191017024419-172.22.0.6-40427
Master URL: spark://54f0b773b528:7077
Cores: 4 (0 Used)
Memory: 2.9 GB (0.0 B Used)

Back to Master

Running Executors (0)

ExecutorID	Cores	State	Memory	Job Details	Logs

http://localhost:9870/

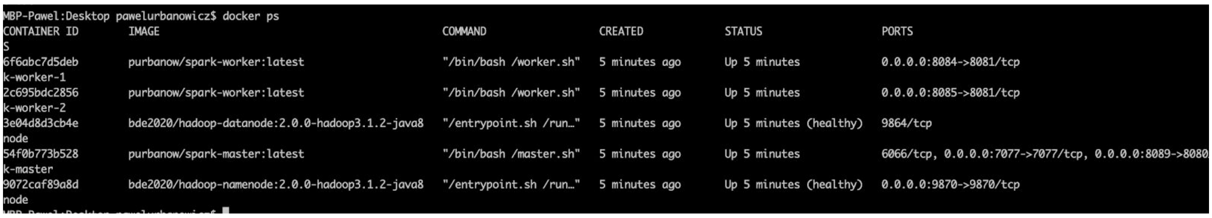


Overview 'namenode:8020' (active)

Started:	Tue Oct 15 14:35:02 +0800 2019
Version:	2.8.0, r91f2b7a13d1e97be65db92ddabc627cc29ac0009
Compiled:	Fri Mar 17 12:12:00 +0800 2017 by jdu from branch-2.8.0
Cluster ID:	CID-2f2f1c2f-7b71-4d14-bc34-2d0afe847ef6
Block Pool ID:	BP-1794751137-172.20.0.2-1571100513076

Summary

docker ps



```
MBP-Pawel:Desktop pawelurbanowicz$ docker ps
CONTAINER ID        IMAGE                                           COMMAND                  CREATED             STATUS                   PORTS
S
6f6abc7d5deb        purbanow/spark-worker:latest                    "/bin/bash /worker.sh"   5 minutes ago       Up 5 minutes             0.0.0.0:8084->8081/tcp
k-worker-1
2c695bdc2856        purbanow/spark-worker:latest                    "/bin/bash /worker.sh"   5 minutes ago       Up 5 minutes             0.0.0.0:8085->8081/tcp
k-worker-2
3e04d8d3cb4e        bde2020/hadoop-datanode:2.0.0-hadoop3.1.2-java8   "/entrypoint.sh /run…"   5 minutes ago       Up 5 minutes (healthy)   9864/tcp
node
54f0b773b528        purbanow/spark-master:latest                    "/bin/bash /master.sh"   5 minutes ago       Up 5 minutes             6066/tcp, 0.0.0.0:7077->7077/tcp, 0.0.0.0:8089->8080
k-master
9072caf89a8d        bde2020/hadoop-namenode:2.0.0-hadoop3.1.2-java8   "/entrypoint.sh /run…"   5 minutes ago       Up 5 minutes (healthy)   0.0.0.0:9870->9870/tcp
node
MBP-Pawel:Desktop pawelurbanowicz$
```

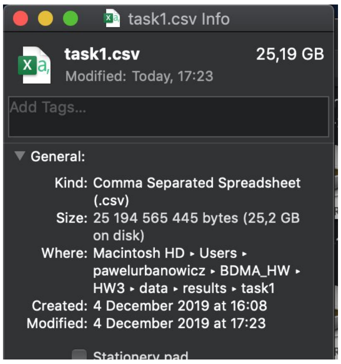
Data Preprocessing

- Replacing newlines by one whitespace
- Converting all text to lowercase
- removing <BODY> tag
- replacing ‘?’ and ‘$’ by one whitespace
- replacing multiline whitespaces with one

Output Format

Task 1: set representation: A MxN matrix: with rows as shingles and columns as documents (N=21,578)

25.12 GB file



task1.csv Info

task1.csv 25,19 GB

Modified: Today, 17:23

Add Tags...

General:

Kind: Comma Separated Spreadsheet (.csv)

Size: 25 194 565 445 bytes (25,2 GB on disk)

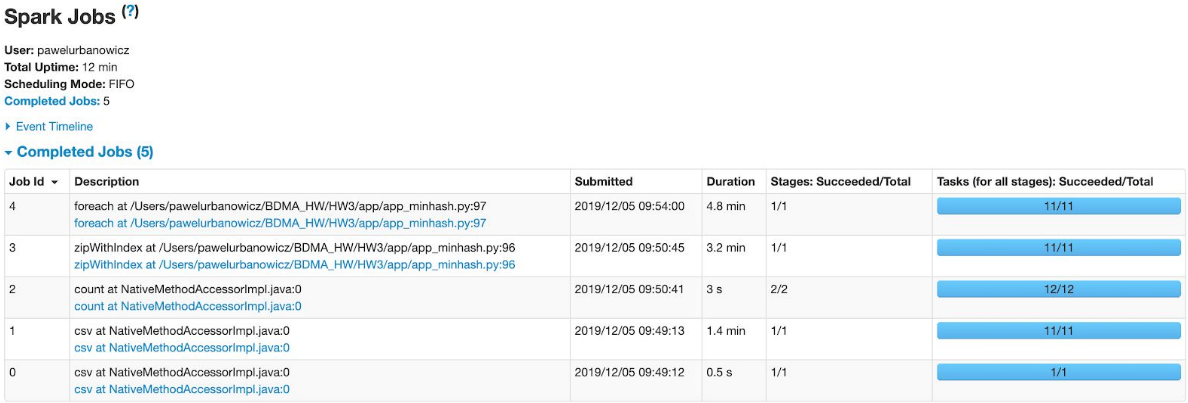
Where: Macintosh HD ▸ Users ▸ pawelurbanowicz ▸ BDMA_HW ▸ HW3 ▸ data ▸ results ▸ task1

Created: 4 December 2019 at 16:08

Modified: 4 December 2019 at 17:23

□ Stationery pad

Task 2: minhash signatures: The HxN signature matrix: with H as the number of hash functions, and N=21,578



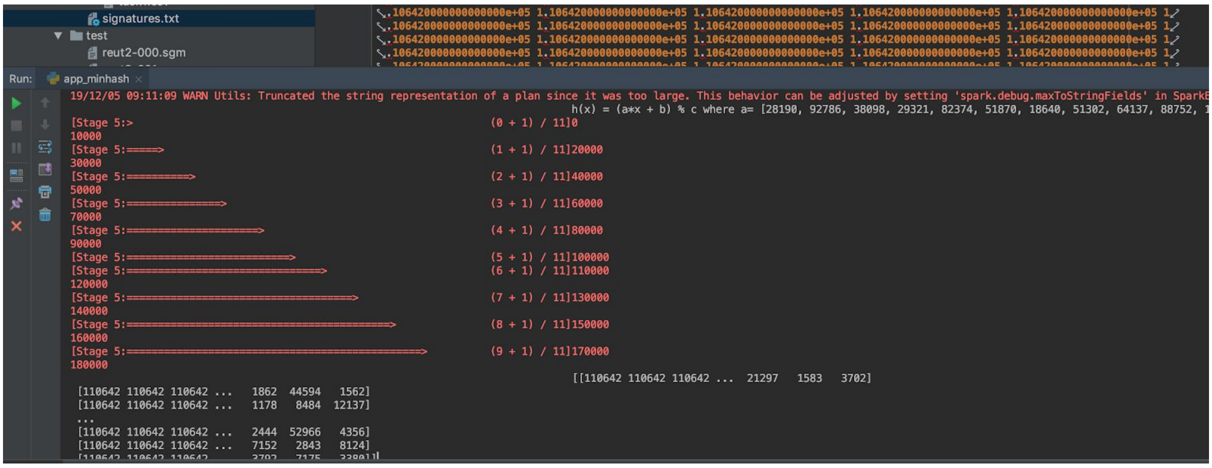
Spark Jobs (?)

User: pawelurbanowicz
Total Uptime: 12 min
Scheduling Mode: FIFO
Completed Jobs: 5

▸ Event Timeline

▾ Completed Jobs (5)

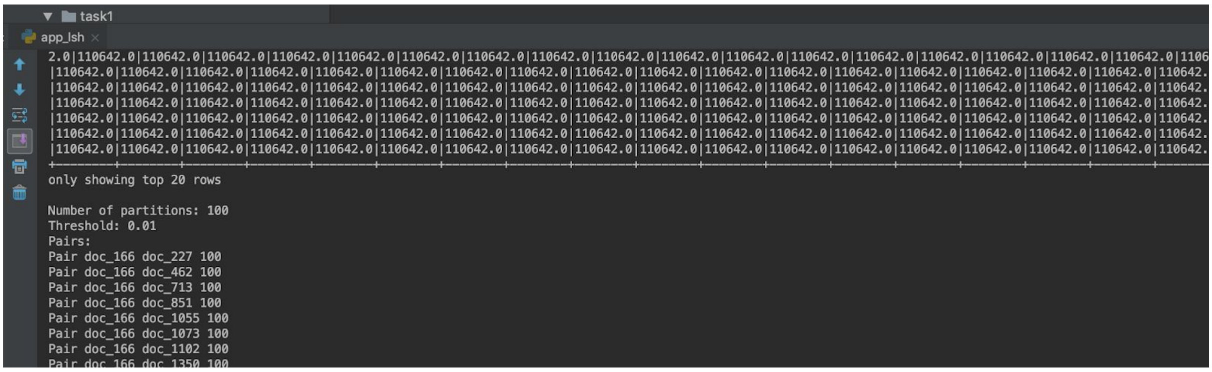
Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
4	foreach at /Users/pawelurbanowicz/BDMA_HW/HW3/app/app_minhash.py:97 foreach at /Users/pawelurbanowicz/BDMA_HW/HW3/app/app_minhash.py:97	2019/12/05 09:54:00	4.8 min	1/1	11/11
3	zipWithIndex at /Users/pawelurbanowicz/BDMA_HW/HW3/app/app_minhash.py:96 zipWithIndex at /Users/pawelurbanowicz/BDMA_HW/HW3/app/app_minhash.py:96	2019/12/05 09:50:45	3.2 min	1/1	11/11
2	count at NativeMethodAccessorImpl.java:0 count at NativeMethodAccessorImpl.java:0	2019/12/05 09:50:41	3 s	2/2	12/12
1	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2019/12/05 09:49:13	1.4 min	1/1	11/11
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2019/12/05 09:49:12	0.5 s	1/1	1/1



```
↘.106420000000000000e+05 1.106420000000000000e+05 1.106420000000000000e+05 1.106420000000000000e+05 1.106420000000000000e+05 1↙
↘.106420000000000000e+05 1.106420000000000000e+05 1.106420000000000000e+05 1.106420000000000000e+05 1.106420000000000000e+05 1↙
↘.106420000000000000e+05 1.106420000000000000e+05 1.106420000000000000e+05 1.106420000000000000e+05 1.106420000000000000e+05 1↙
↘.106420000000000000e+05 1.106420000000000000e+05 1.106420000000000000e+05 1.106420000000000000e+05 1.106420000000000000e+05 1↙
```

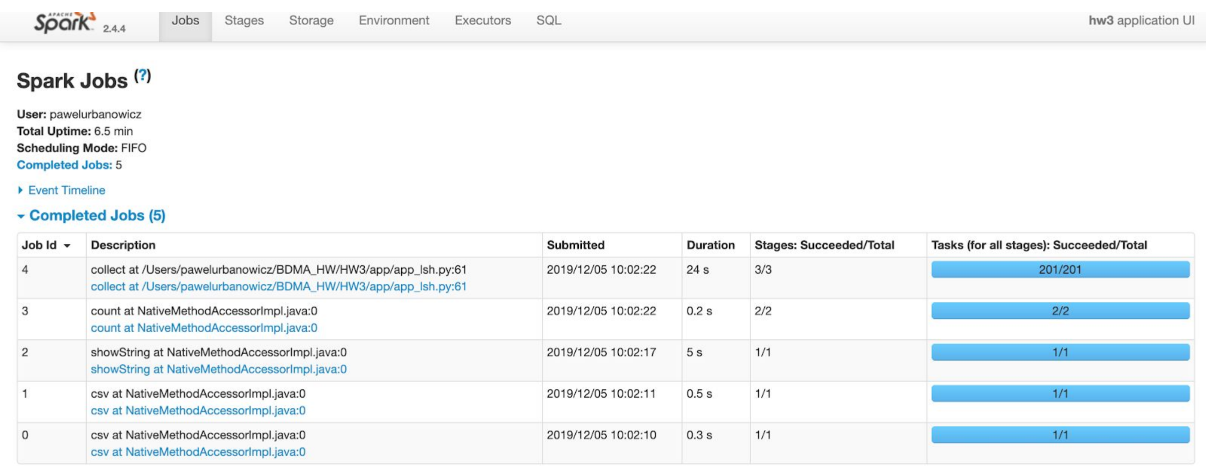
```
19/12/05 09:11:09 WARN Utils: Truncated the string representation of a plan since it was too large. This behavior can be adjusted by setting 'spark.debug.maxToStringFields' in SparkE
                                                                                  h(x) = (a*x + b) % c where a= [28190, 92786, 38098, 29321, 82374, 51870, 18640, 51302, 64137, 88752, 1
[Stage 5:>                                                          (0 + 1) / 11]0
10000
[Stage 5:=====>                                                     (1 + 1) / 11]20000
30000
[Stage 5:==========>                                                (2 + 1) / 11]40000
50000
[Stage 5:===============>                                           (3 + 1) / 11]60000
70000
[Stage 5:=====================>                                     (4 + 1) / 11]80000
90000
[Stage 5:==========================>                                (5 + 1) / 11]100000
[Stage 5:===============================>                           (6 + 1) / 11]110000
120000
[Stage 5:====================================>                      (7 + 1) / 11]130000
140000
[Stage 5:==========================================>                (8 + 1) / 11]150000
160000
[Stage 5:===============================================>           (9 + 1) / 11]170000
180000
                                                                                  [[110642 110642 110642 ...  21297   1583   3702]
 [110642 110642 110642 ...   1862  44594   1562]
 [110642 110642 110642 ...   1178   8484  12137]
 ...
 [110642 110642 110642 ...   2444  52966   4356]
 [110642 110642 110642 ...   7152   2843   8124]
 [110642 110642 110642 ...   3702   7175   3380]]
```

Task 3: candidate pairs



```
2.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|1106
|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.
|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.
|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.
|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.
|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.
|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.0|110642.
+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------+--------
only showing top 20 rows

Number of partitions: 100
Threshold: 0.01
Pairs:
Pair doc_166 doc_227 100
Pair doc_166 doc_462 100
Pair doc_166 doc_713 100
Pair doc_166 doc_851 100
Pair doc_166 doc_1055 100
Pair doc_166 doc_1073 100
Pair doc_166 doc_1102 100
Pair doc_166 doc_1350 100
```



Spark Jobs (?)

User: pawelurbanowicz
Total Uptime: 6.5 min
Scheduling Mode: FIFO
Completed Jobs: 5

▸ Event Timeline

▾ **Completed Jobs (5)**

Job Id ▾	Description	Submitted	Duration	Stages: Succeeded/Total	Tasks (for all stages): Succeeded/Total
4	collect at /Users/pawelurbanowicz/BDMA_HW/HW3/app/app_lsh.py:61 collect at /Users/pawelurbanowicz/BDMA_HW/HW3/app/app_lsh.py:61	2019/12/05 10:02:22	24 s	3/3	201/201
3	count at NativeMethodAccessorImpl.java:0 count at NativeMethodAccessorImpl.java:0	2019/12/05 10:02:22	0.2 s	2/2	2/2
2	showString at NativeMethodAccessorImpl.java:0 showString at NativeMethodAccessorImpl.java:0	2019/12/05 10:02:17	5 s	1/1	1/1
1	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2019/12/05 10:02:11	0.5 s	1/1	1/1
0	csv at NativeMethodAccessorImpl.java:0 csv at NativeMethodAccessorImpl.java:0	2019/12/05 10:02:10	0.3 s	1/1	1/1