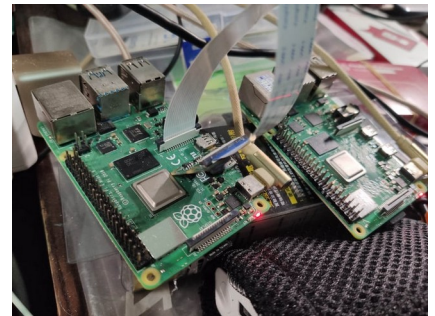


Big Data Mining Homework 1

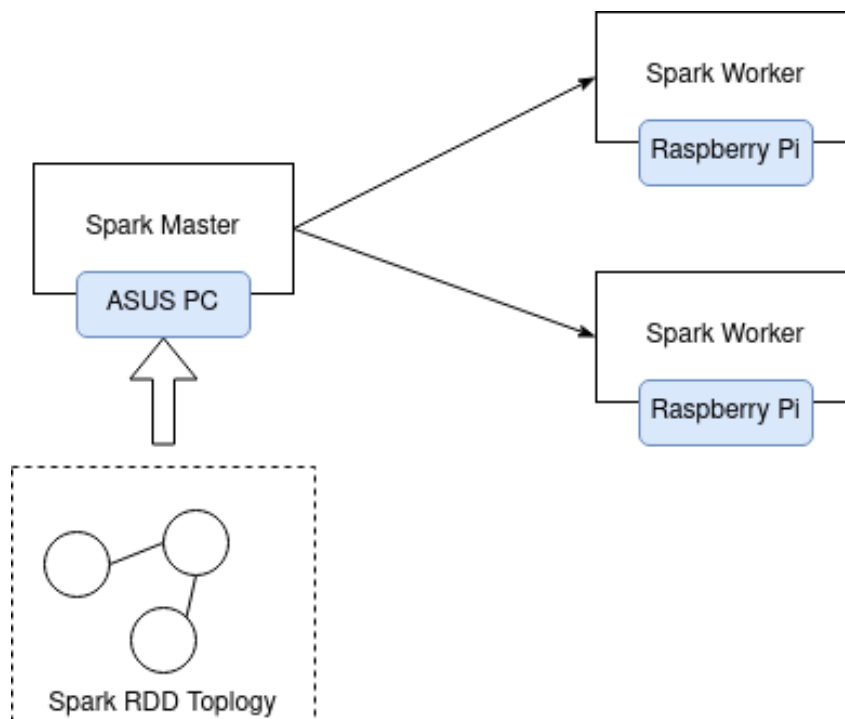
Spark Platform:

The platform consists of:

1. Raspberry Pi 4 Model B x2
 - OS: Linux Ubuntu 20.04 Server
 - CPU architecture: aarch64
 - RAM: 8GB
 - CPU: Broadcom BCM2711, Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz
 - Number of CPU: 4C (CPU) 1T (Thread Per CPU)
2. Asus-vivobook notebook
 - OS: Linux Ubuntu 20.04 LTS
 - CPU architecture: x86_64
 - RAM: 8GB
 - CPU: Intel(R) Core(TM) i3-8130U CPU @ 2.20GHz
 - Number of CPU: 4C (CPU) 2T (Thread Per CPU)



The simple architecture of spark cluster:



The generated output:

Log the RDD (Resilient Distributed Dataset) output

```
maskertim@SandiskUbuntu: ~/schoolwork/bdm2021/hw1
tracker-extract-files:1000
tracker-extract-files:125
voltage
maskertim@SandiskUbuntu:~/schoolwork/bdm2021/hw1$ ls /tmp/voltage/
_SUCCESS
maskertim@SandiskUbuntu:~/schoolwork/bdm2021/hw1$ cat /tmp/App.log
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<DOCTYPE log SYSTEM "logger.dtd">
<log>
<record>
<date>2021-10-23T05:46:20.795688Z</date>
<millis>1634967980795</millis>
<nanos>688000</nanos>
<sequence>0</sequence>
<logger>com.bdm.App</logger>
<level>INFO</level>
<class>com.bdm.App</class>
<method>main</method>
<thread>1</thread>
<message>global_active_power max:11.122</message>
</record>
<record>
<date>2021-10-23T05:46:24.325268Z</date>
<millis>1634967984325</millis>
<nanos>268000</nanos>
<sequence>1</sequence>
<logger>com.bdm.App</logger>
<level>INFO</level>
<class>com.bdm.App</class>
<method>main</method>
<thread>1</thread>
<message>global_active_power min:0.076</message>
</record>
<record>
<date>2021-10-23T05:46:26.43441Z</date>
<millis>1634967986434</millis>
<nanos>41000</nanos>
<sequence>2</sequence>
<logger>com.bdm.App</logger>
<level>INFO</level>
<class>com.bdm.App</class>
<method>main</method>
<thread>1</thread>
<message>global_active_power count:2049280</message>
</record>
<record>
<date>2021-10-23T05:46:31.857563Z</date>
<millis>1634967991857</millis>
<nanos>563000</nanos>
<sequence>3</sequence>
<logger>com.bdm.App</logger>
<level>INFO</level>
<class>com.bdm.App</class>
<method>main</method>
```

```
maskertim@SandiskUbuntu: ~/schoolwork/bdm2021/hw1
<thread>1</thread>
<message>global_intensity max:48.4</message>
</record>
<record>
<date>2021-10-23T05:48:24.257131Z</date>
<millis>1634968104257</millis>
<nanos>131000</nanos>
<sequence>16</sequence>
<logger>com.bdm.App</logger>
<level>INFO</level>
<class>com.bdm.App</class>
<method>main</method>
<thread>1</thread>
<message>global_intensity min:0.2</message>
</record>
<record>
<date>2021-10-23T05:48:26.875809Z</date>
<millis>1634968106875</millis>
<nanos>809000</nanos>
<sequence>17</sequence>
<logger>com.bdm.App</logger>
<level>INFO</level>
<class>com.bdm.App</class>
<method>main</method>
<thread>1</thread>
<message>global_intensity count:2049280</message>
</record>
<record>
<date>2021-10-23T05:48:32.585630Z</date>
<millis>1634968112585</millis>
<nanos>630000</nanos>
<sequence>18</sequence>
<logger>com.bdm.App</logger>
<level>INFO</level>
<class>com.bdm.App</class>
<method>main</method>
<thread>1</thread>
<message>global_intensity mean:4.627759310588324</message>
</record>
<record>
<date>2021-10-23T05:48:37.088871Z</date>
<millis>1634968117088</millis>
<nanos>871000</nanos>
<sequence>19</sequence>
<logger>com.bdm.App</logger>
<level>INFO</level>
<class>com.bdm.App</class>
<method>main</method>
<thread>1</thread>
<message>global_intensity std:4.444395175407247</message>
</record>
</log>
maskertim@SandiskUbuntu:~/schoolwork/bdm2021/hw1$
```

The completed data processed by spark cluster on Raspberry Pi

```
ubuntu@ubuntu:~$ jps
3008 Worker
3065 Jps
ubuntu@ubuntu:~$ ls /tmp
household_power_consumption.txt
hsperfdata_ubuntu
snap.lxd
spark-ubuntu-org.apache.spark.deploy.worker.Worker-1.pid
systemd-private-996c36f3ede847eaacae87c46d146678-fwupd.service-OgASVg
systemd-private-996c36f3ede847eaacae87c46d146678-systemd-logind.service-Y566li
systemd-private-996c36f3ede847eaacae87c46d146678-systemd-resolved.service-O3oyqh
systemd-private-996c36f3ede847eaacae87c46d146678-systemd-timesyncd.service-jhmh6f
ubuntu@ubuntu:~$ ls /tmp
global_active_power
global_intesity
global_reactive_power
household_power_consumption.txt
hsperfdata_ubuntu
snap.lxd
spark-cf951b75-848e-4d26-a4ed-b638c82f54f7
spark-ubuntu-org.apache.spark.deploy.worker.Worker-1.pid
systemd-private-996c36f3ede847eaacae87c46d146678-fwupd.service-OgASVg
systemd-private-996c36f3ede847eaacae87c46d146678-systemd-logind.service-Y566li
systemd-private-996c36f3ede847eaacae87c46d146678-systemd-resolved.service-O3oyqh
systemd-private-996c36f3ede847eaacae87c46d146678-systemd-timesyncd.service-jhmh6f
ubuntu@ubuntu:~$
```

```
0.6542880821812758
0.4721804373205265
0.33946381631482087
0.4413161177455915
0.1573642177455977
0.12649992588810418
0.259216553225336
0.3147723507468296
0.3518095490978253
0.3981060470340612
ubuntu@ubuntu:~$ ls /tmp
household_power_consumption.txt
hsperfdata_ubuntu
snap.lxd
spark-29be084-1b08-4b6f-85fa-87fc76a2bf05
spark-ubuntu-org.apache.spark.deploy.worker.Worker-1.pid
systemd-private-becf51db9d274e4aaabbf79dba8e61bb-fwupd.service-CJNK6g
systemd-private-becf51db9d274e4aaabbf79dba8e61bb-systemd-logind.service-fx5ACf
systemd-private-becf51db9d274e4aaabbf79dba8e61bb-systemd-resolved.service-v8gc4h
systemd-private-becf51db9d274e4aaabbf79dba8e61bb-systemd-timesyncd.service-qLGv5h
ubuntu@ubuntu:~$
```

Spark Web GUI and Logging of Spark Cluster

spark/pom.xml × tutorialkart.com × RDD (Spark 3.2) × AggregateFunc × Spark RDD Acti × Spark RDD Transf × Spark Master at: × 分享在Linux下 × + -

192.168.0.199:8080

Spark Master at spark://SandiskUbuntu:7077

URL: spark://SandiskUbuntu:7077
Alive Workers: 2
Cores in use: 8 Total, 0 Used
Memory in use: 2.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

| Worker Id | Address | State | Cores | Memory | Resources |
|---|---------------------|-------|------------|-------------------------|-----------|
| worker-20211023053328-192.168.0.201-45875 | 192.168.0.201:45875 | ALIVE | 4 (0 Used) | 1024.0 MiB (0.0 B Used) | |
| worker-20211023053342-192.168.0.200-42707 | 192.168.0.200:42707 | ALIVE | 4 (0 Used) | 1024.0 MiB (0.0 B Used) | |

Running Applications (0)

| Application ID | Name | Cores | Memory per Executor | Resources Per Executor | Submitted Time | User | State | Duration |
|----------------|------|-------|---------------------|------------------------|----------------|------|-------|----------|
|----------------|------|-------|---------------------|------------------------|----------------|------|-------|----------|

Completed Applications (1)

| Application ID | Name | Cores | Memory per Executor | Resources Per Executor | Submitted Time | User | State | Duration |
|-------------------------|-------------------|-------|---------------------|------------------------|---------------------|----------|----------|----------|
| app-20211023134559-0000 | power_consumption | 8 | 1024.0 MiB | | 2021/10/23 13:45:59 | maskerim | FINISHED | 3.1 min |

spark/pom.xml × tutorialkart.com × RDD (Spark 3.2) × AggregateFunc × Spark RDD Acti × Spark RDD Transf × Application: pow × 分享在Linux下 × + -

192.168.0.199:8080/app?appId=app-20211023134559-0000

Application: power_consumption

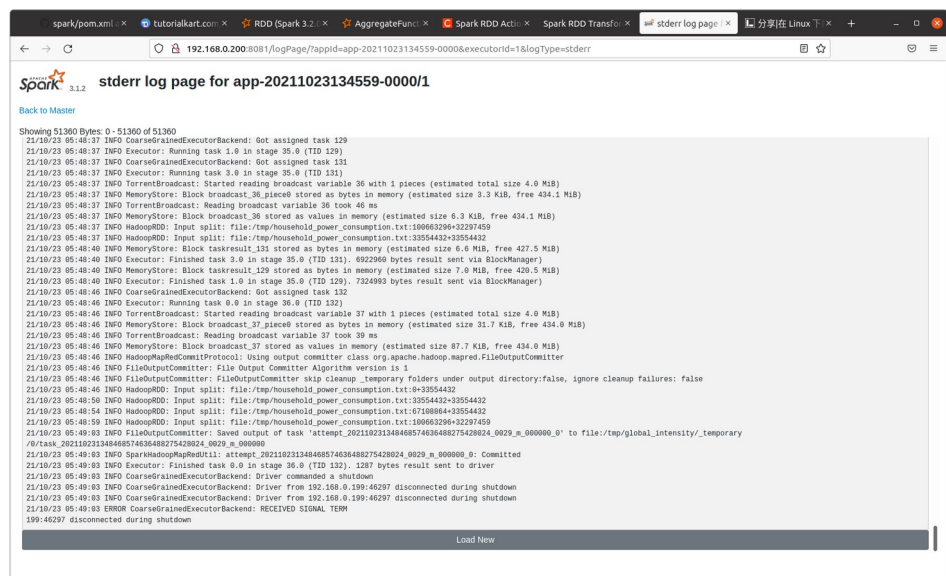
ID: app-20211023134559-0000
Name: power_consumption
User: maskerim
Cores: Unlimited (8 granted)
Executor Limit: Unlimited (2 granted)
Executor Memory: 1024.0 MiB
Executor Resources:
Submit Date: 2021/10/23 13:45:59
State: FINISHED

Executor Summary (2)

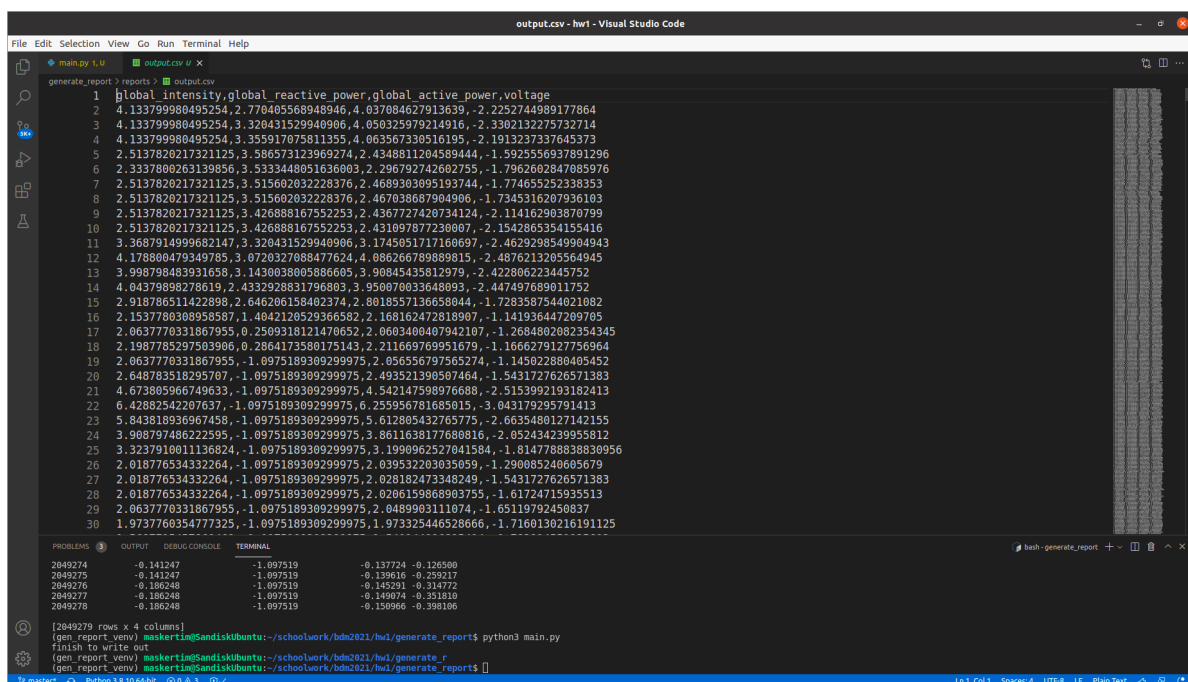
| ExecutorID | Worker | Cores | Memory | Resources | State | Logs |
|------------|--------|-------|--------|-----------|-------|------|
|------------|--------|-------|--------|-----------|-------|------|

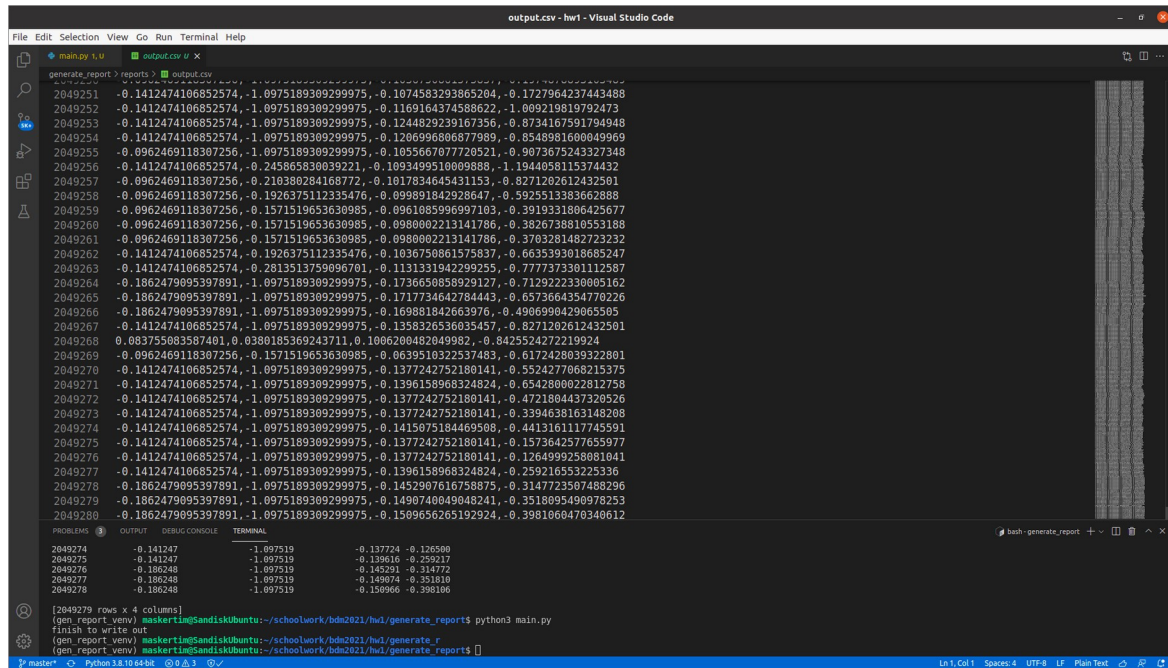
Removed Executors (2)

| ExecutorID | Worker | Cores | Memory | Resources | State | Logs |
|------------|---|-------|--------|-----------|--------|-------------------------------|
| 1 | worker-20211023053342-192.168.0.200-42707 | 4 | 1024 | | KILLED | stdout stderr |
| 0 | worker-20211023053328-192.168.0.201-45875 | 4 | 1024 | | KILLED | stdout stderr |



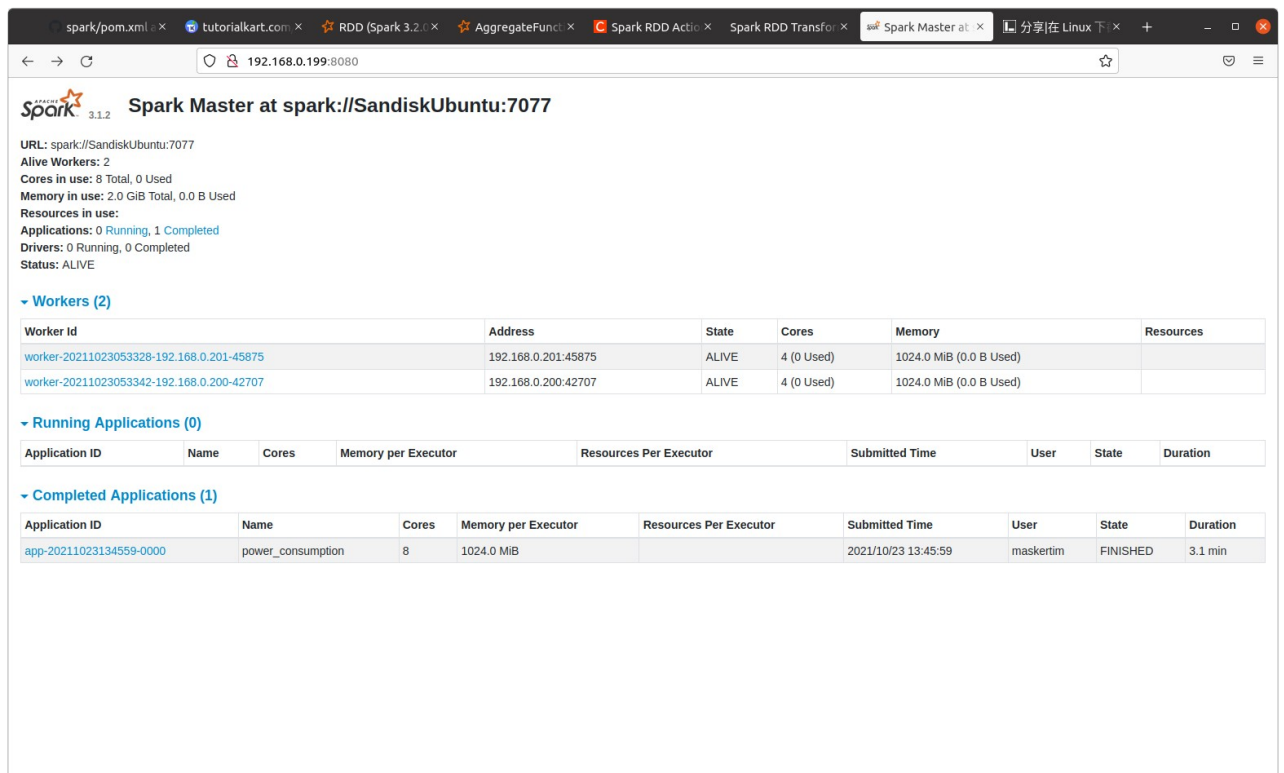
Report of min-max normalization (using formula of Z-Score)





What the workflow of implementation using spark?

1. Start spark master by “start-master.sh” command.
2. Start spark worker by “start-worker.sh [your spark url]” command to connect spark master.



spark/pom.xml × tutorialkart.com × RDD (Spark 3.2.0) × AggregateFunc × Spark RDD Actio × Spark RDD Transfo × Spark Master at spark://SandiskUbuntu:7077

URL: spark://SandiskUbuntu:7077
Alive Workers: 2
Cores in use: 8 Total, 0 Used
Memory in use: 2.0 GiB Total, 0.0 B Used
Resources in use:
Applications: 0 Running, 1 Completed
Drivers: 0 Running, 0 Completed
Status: ALIVE

Workers (2)

| Worker Id | Address | State | Cores | Memory | Resources |
|---|---------------------|-------|------------|-------------------------|-----------|
| worker-20211023053328-192.168.0.201-45875 | 192.168.0.201:45875 | ALIVE | 4 (0 Used) | 1024.0 MiB (0.0 B Used) | |
| worker-20211023053342-192.168.0.200-42707 | 192.168.0.200:42707 | ALIVE | 4 (0 Used) | 1024.0 MiB (0.0 B Used) | |

Running Applications (0)

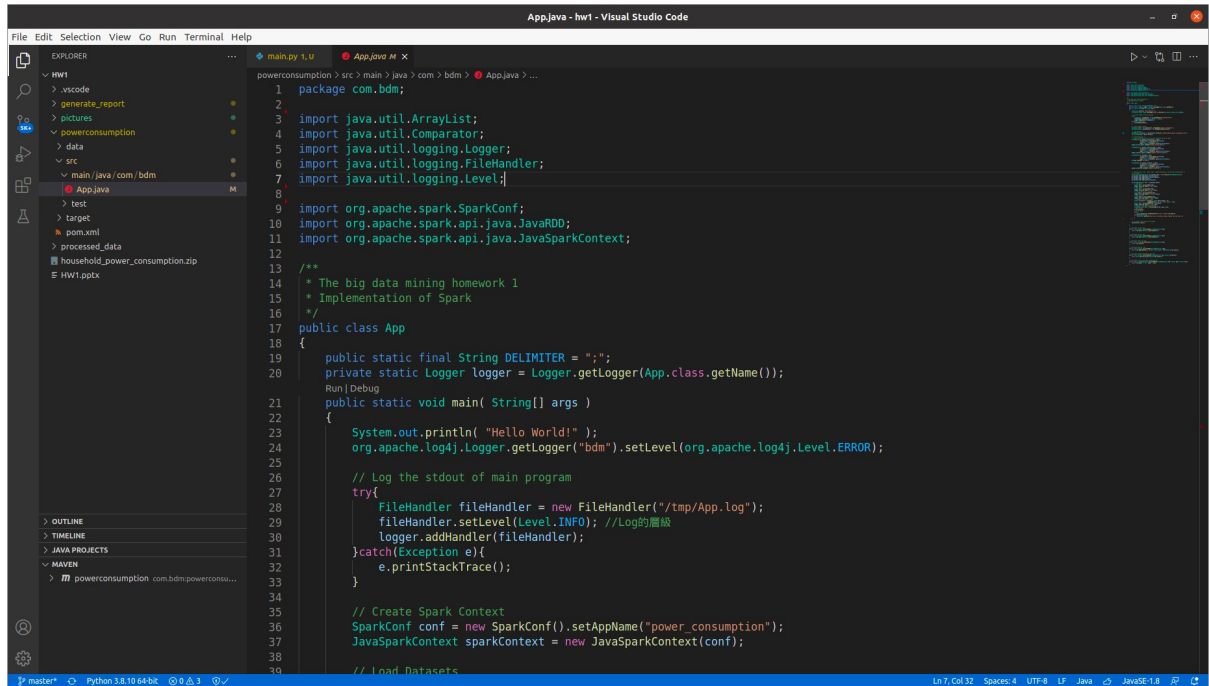
| Application ID | Name | Cores | Memory per Executor | Resources Per Executor | Submitted Time | User | State | Duration |
|----------------|------|-------|---------------------|------------------------|----------------|------|-------|----------|
|----------------|------|-------|---------------------|------------------------|----------------|------|-------|----------|

Completed Applications (1)

| Application ID | Name | Cores | Memory per Executor | Resources Per Executor | Submitted Time | User | State | Duration |
|-------------------------|-------------------|-------|---------------------|------------------------|---------------------|-----------|----------|----------|
| app-20211023134559-0000 | power_consumption | 8 | 1024.0 MiB | | 2021/10/23 13:45:59 | maskertim | FINISHED | 3.1 min |

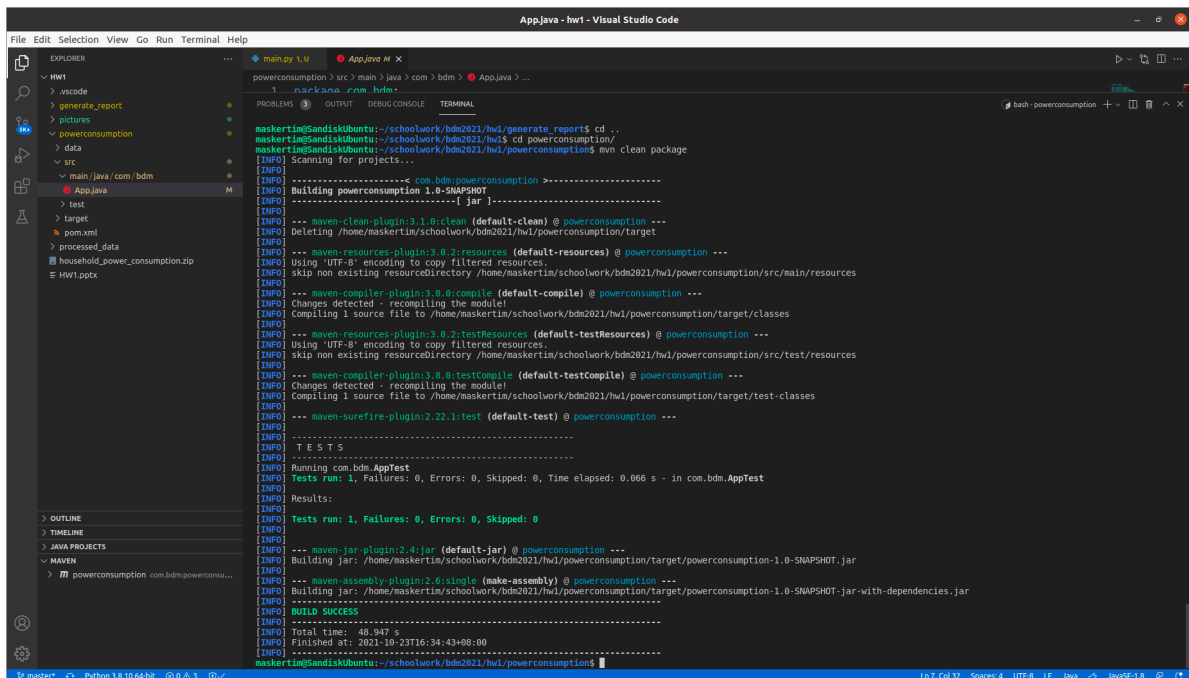
3. Writing the Spark API with any program (e.g., Java, R, Python, Scala) that spark provides, in this work, that uses Java Language.

4. If you finish to write a program, as follows:



```
1 package com.bdm;
2
3 import java.util.ArrayList;
4 import java.util.Comparator;
5 import java.util.logging.Logger;
6 import java.util.logging.FileHandler;
7 import java.util.logging.Level;
8
9 import org.apache.spark.SparkConf;
10 import org.apache.spark.api.java.JavaRDD;
11 import org.apache.spark.api.java.JavaSparkContext;
12
13 /**
14  * The big data mining homework 1
15  * Implementation of Spark
16  */
17 public class App
18 {
19     public static final String DELIMITER = ";";
20     private static Logger logger = Logger.getLogger(App.class.getName());
21
22     public static void main( String[] args )
23     {
24         System.out.println( "Hello World!" );
25         org.apache.log4j.Logger.getLogger("bdm").setLevel(org.apache.log4j.Level.ERROR);
26
27         // Log the stdout of main program
28         try{
29             FileHandler fileHandler = new FileHandler("/tmp/App.log");
30             fileHandler.setLevel(Level.INFO); //Log的级别
31             logger.addHandler(fileHandler);
32         }catch(Exception e){
33             e.printStackTrace();
34         }
35
36         // Create Spark Context
37         SparkConf conf = new SparkConf().setAppName("power_consumption");
38         JavaSparkContext sparkContext = new JavaSparkContext(conf);
39
40         // Load Datasets
```

5. Later using the maven (Java package management) to compile and package jar file. The command is “mvn clean package”. (Note that you may configure different settings in pom.xml, so this just a reference)



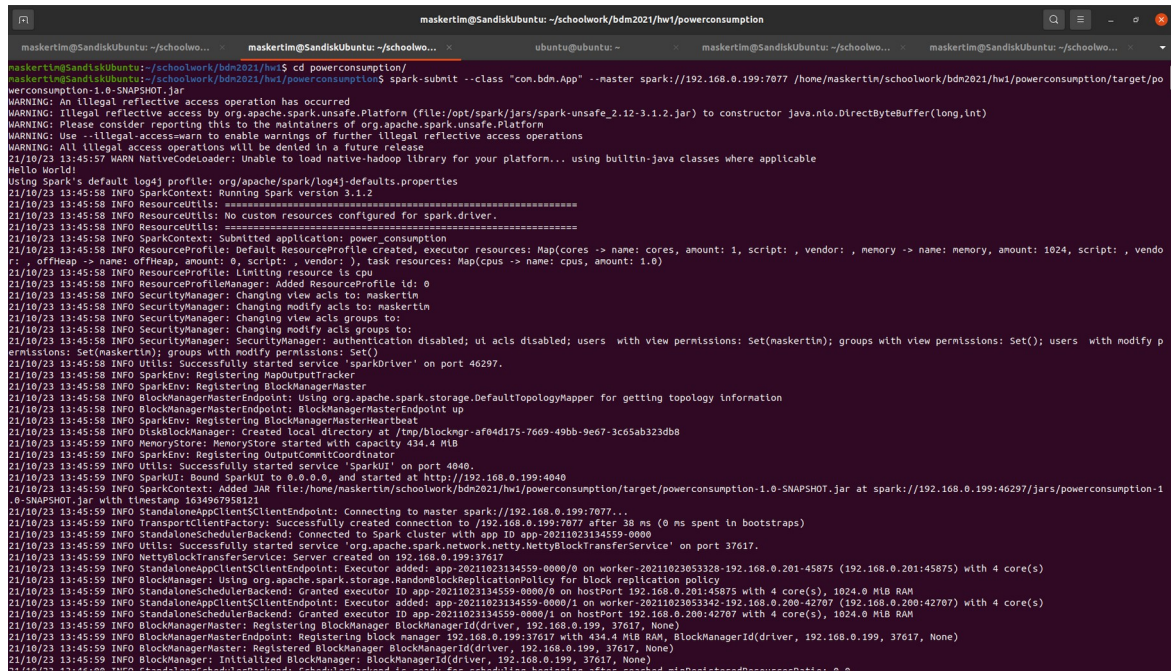
```
maskertingSandisk@buntu:~/schoolwork/bdm2021/hw1$ mvn clean package
[INFO] Scanning for projects...
[INFO]
[INFO] ----- com.bdm:powerconsumption -----
[INFO] Building powerconsumption 1.0-SNAPSHOT
[INFO]
[INFO] --- maven-clean-plugin:3.1.0:clean (default-clean) @ powerconsumption ---
[INFO] Deleting /home/maskerting/schoolwork/bdm2021/hw1/powerconsumption/target
[INFO]
[INFO] --- maven-resources-plugin:3.0.2:resources (default-resources) @ powerconsumption ---
[INFO] skip non existing resourceDirectory /home/maskerting/schoolwork/bdm2021/hw1/powerconsumption/src/main/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.8.0:compile (default-compile) @ powerconsumption ---
[INFO] Changes detected - recompiling the module!
[INFO] Compiling 1 source file to /home/maskerting/schoolwork/bdm2021/hw1/powerconsumption/target/classes
[INFO]
[INFO] --- maven-resources-plugin:3.0.2:testResources (default-testResources) @ powerconsumption ---
[INFO] skip non existing resourceDirectory /home/maskerting/schoolwork/bdm2021/hw1/powerconsumption/src/test/resources
[INFO]
[INFO] --- maven-compiler-plugin:3.8.0:testCompile (default-testCompile) @ powerconsumption ---
[INFO] Changes detected - recompiling the module!
[INFO] Compiling 1 source file to /home/maskerting/schoolwork/bdm2021/hw1/powerconsumption/target/test-classes
[INFO]
[INFO] --- maven-surefire-plugin:2.22.1:test (default-test) @ powerconsumption ---
[INFO]
[INFO] T E S T S
[INFO]
[INFO] Running com.bdm.AppTest
[INFO] Tests run: 1, Failures: 0, Errors: 0, Skipped: 0, Time elapsed: 0.066 s - in com.bdm.AppTest
[INFO]
[INFO] Results:
[INFO]
[INFO] Tests run: 1, Failures: 0, Errors: 0, Skipped: 0
[INFO]
[INFO] --- maven-jar-plugin:2.4:jar (default-jar) @ powerconsumption ---
[INFO] Building jar: /home/maskerting/schoolwork/bdm2021/hw1/powerconsumption/target/powerconsumption-1.0-SNAPSHOT.jar
[INFO]
[INFO] --- maven-assembly-plugin:2.6:single (make-assembly) @ powerconsumption ---
[INFO] Building jar: /home/maskerting/schoolwork/bdm2021/hw1/powerconsumption/target/powerconsumption-1.0-SNAPSHOT-jar-with-dependencies.jar
[INFO]
[INFO] BUILD SUCCESS
[INFO]
[INFO] Total time: 40.947 s
[INFO] Finished at: 2021-10-23T16:34:43+08:00
[INFO]
maskertingSandisk@buntu:~/schoolwork/bdm2021/hw1/powerconsumptions$
```

6. Finally, that can find a jar package in target directory, let just submit your jar file to spark cluster.

“spark-submit --class "com.bdm.App" --master spark://192.168.0.199:7077

/home/maskertim/schoolwork/bdm2021/hw1/powerconsumption/target/powerconsumption-1.0-SNAPSHOT.jar”.

As above that is my setting for spark cluster. Need to change some variable to apply different environment settings.



```
maskertim@SandiskUbuntu: ~/schoolwork/bdm2021/hw1/powerconsumption
maskertim@SandiskUbuntu: ~/schoolwork/bdm2021/hw1/powerconsumption$ spark-submit --class "com.bdm.App" --master spark://192.168.0.199:7077 /home/maskertim/schoolwork/bdm2021/hw1/powerconsumption/target/powerconsumption-1.0-SNAPSHOT.jar
WARNING: An illegal reflective access operation has occurred
WARNING: Illegal reflective access by org.apache.spark.unsafe.Platform (file:/opt/spark/jars/spark-unsafe_2.12-3.1.2.jar) to constructor java.nio.DirectByteBuffer(long,int)
WARNING: Please consider reporting this to the maintainers of org.apache.spark.unsafe.Platform
WARNING: Use --illegal-access=warn to enable warnings of further illegal reflective access operations
WARNING: All illegal access operations will be denied in a future release
21/10/23 13:45:57 WARN NativeCodeLoader: Unable to load native-heapoop library for your platform... using builtin-java classes where applicable
Hello World!
Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
21/10/23 13:45:58 INFO SparkContext: Running Spark version 3.1.2
21/10/23 13:45:58 INFO ResourceUtils: =====
21/10/23 13:45:58 INFO ResourceUtils: No custom resources configured for spark.driver.
21/10/23 13:45:58 INFO ResourceUtils: =====
21/10/23 13:45:58 INFO SparkContext: Submitted application: power_consumption
21/10/23 13:45:58 INFO ResourceProfile: Default ResourceProfile created, executor resources: Map(cores -> name: cores, amount: 1, script: , vendor: , memory -> name: memory, amount: 1024, script: , vendor: , offHeap -> name: offHeap, amount: 0, script: , vendor: ), task resources: Map(cpus -> name: cpus, amount: 1.0)
21/10/23 13:45:58 INFO ResourceProfileManager: Limiting resource is cpu
21/10/23 13:45:58 INFO ResourceProfileManager: Added ResourceProfile id: 0
21/10/23 13:45:58 INFO SecurityManager: Changing view acls to: maskertim
21/10/23 13:45:58 INFO SecurityManager: Changing modify acls to: maskertim
21/10/23 13:45:58 INFO SecurityManager: Changing view acls groups to:
21/10/23 13:45:58 INFO SecurityManager: Changing modify acls groups to:
21/10/23 13:45:58 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(maskertim); groups with view permissions: Set(); users with modify permissions: Set(maskertim); groups with modify permissions: Set()
21/10/23 13:45:58 INFO Utils: Successfully started service 'sparkDriver' on port 46297.
21/10/23 13:45:58 INFO SparkEnv: Registering MapOutputTracker
21/10/23 13:45:58 INFO SparkEnv: Registering BlockManagerMaster
21/10/23 13:45:58 INFO BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information
21/10/23 13:45:58 INFO BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
21/10/23 13:45:58 INFO SparkEnv: Registering BlockManagerMasterHeartbeat
21/10/23 13:45:58 INFO DiskBlockManager: Created local directory at /tmp/blockmgr-af04d175-7669-49bb-9e67-3c65ab323db8
21/10/23 13:45:59 INFO MemoryStore: MemoryStore started with capacity 434.4 MiB
21/10/23 13:45:59 INFO SparkEnv: Registering OutputCommitCoordinator
21/10/23 13:45:59 INFO Utils: Successfully started service 'SparkUI' on port 4040.
21/10/23 13:45:59 INFO SparkContext: Bound SparkUI to 0.0.0.0, and started at http://192.168.0.199:4040
21/10/23 13:45:59 INFO SparkContext: Added JAR file:/home/maskertim/schoolwork/bdm2021/hw1/powerconsumption/target/powerconsumption-1.0-SNAPSHOT.jar at spark://192.168.0.199:46297/jars/powerconsumption-1.0-SNAPSHOT.jar with timestamp 1634967958121
21/10/23 13:45:59 INFO StandaloneAppClient$ClientEndpoint: Connecting to master spark://192.168.0.199:7077...
21/10/23 13:45:59 INFO TransportClientFactory: Successfully created connection to /192.168.0.199:7077 after 38 ms (0 ms spent in bootstraps)
21/10/23 13:45:59 INFO StandaloneSchedulerBackend: Connected to Spark cluster with app ID app-20211023134559-0000
21/10/23 13:45:59 INFO NettyBlockTransferService: Server created on 192.168.0.199:37617
21/10/23 13:45:59 INFO StandaloneAppClient$ClientEndpoint: Executor added: app-20211023134559-0000/0 on worker-20211023053328-192.168.0.201-45875 (192.168.0.201:45875) with 4 core(s)
21/10/23 13:45:59 INFO BlockManager: Using org.apache.spark.storage.RandomBlockReplicationPolicy for block replication policy
21/10/23 13:45:59 INFO StandaloneSchedulerBackend: Granted executor ID app-20211023134559-0000/0 on hostPort 192.168.0.201:45875 with 4 core(s), 1024.0 MiB RAM
21/10/23 13:45:59 INFO StandaloneAppClient$ClientEndpoint: Executor added: app-20211023134559-0000/1 on worker-20211023053342-192.168.0.200-42707 (192.168.0.200:42707) with 4 core(s), 1024.0 MiB RAM
21/10/23 13:45:59 INFO StandaloneSchedulerBackend: Granted executor ID app-20211023134559-0000/1 on hostPort 192.168.0.200:42707 with 4 core(s), 1024.0 MiB RAM
21/10/23 13:45:59 INFO BlockManagerMaster: Registering block manager 192.168.0.199:37617 with 434.4 MiB RAM, BlockManagerId(driver, 192.168.0.199, 37617, None)
21/10/23 13:45:59 INFO BlockManagerMaster: Registered block manager 192.168.0.199, 37617, None)
21/10/23 13:45:59 INFO BlockManager: Initialized block manager: BlockManagerId(driver, 192.168.0.199, 37617, None)
```