# 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:
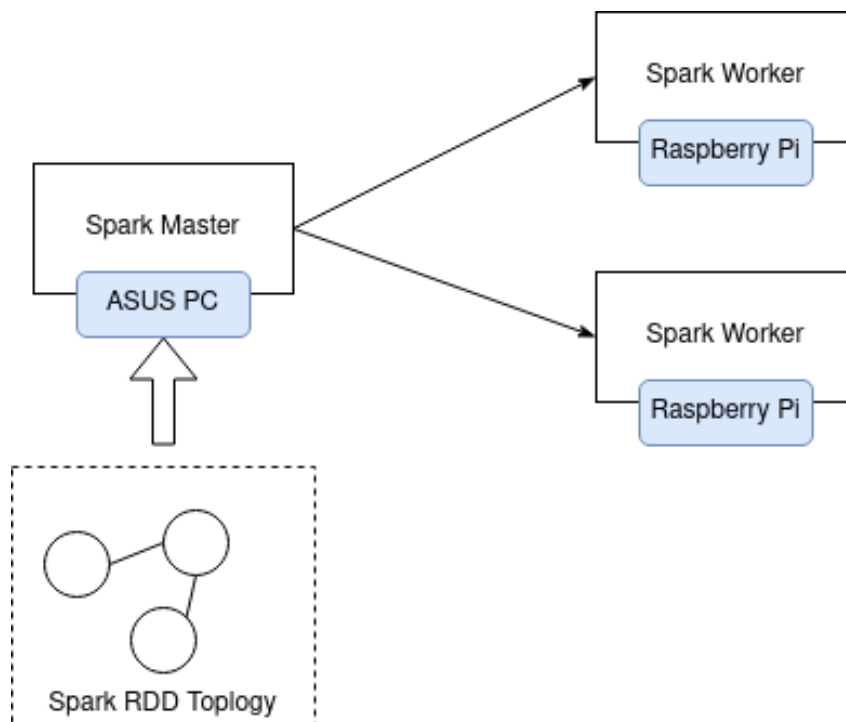
The generated output:

Log the RDD (Resilient Distributed Dataset) output

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>2021-10-23T05:46:26.434441Z</date>
  <millis>1634967986434</millis>
  <nanos>441000</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>

  <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$

Spark Web GUI and Logging of Spark Cluster

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

Application: power_consumption

ID: app-20211023134559-0000
Name: power_consumption
User: maskertim
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

Spark 3.1.2 **stderr log page for app-20211023134559-0000/1**

Back to Master

Showing 51360 Bytes: 0 - 51360 of 51360

Top of Log

```
Spark Executor Command: "/usr/lib/jvm/java-11-openjdk-arm64/bin/java" "-cp" "/opt/spark/conf/:/opt/spark/jars/*" "-Xmx1024M" "-Dspark.driver.port=46297"
"org.apache.spark.executor.CoarseGrainedExecutorBackend" "--driver-url" "spark://CoarseGrainedScheduler@192.168.0.199:46297" "--executor-id" "1" "--hostname" "192.168.0.200" "--cores" "4" "--app-id"
"app-20211023134559-0000" "--worker-url" "spark://Worker@192.168.0.200:42707"
========================================

Using Spark's default log4j profile: org/apache/spark/log4j-defaults.properties
21/10/23 05:46:03 INFO CoarseGrainedExecutorBackend: Started daemon with process name: 3211@ubuntu
21/10/23 05:46:03 INFO SignalUtils: Registering signal handler for TERM
21/10/23 05:46:03 INFO SignalUtils: Registering signal handler for HUP
21/10/23 05:46:03 INFO SignalUtils: Registering signal handler for INT
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 05:46:05 WARN NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
21/10/23 05:46:05 INFO SecurityManager: Changing view acls to: ubuntu,maskertim
21/10/23 05:46:05 INFO SecurityManager: Changing modify acls to: ubuntu,maskertim
21/10/23 05:46:05 INFO SecurityManager: Changing view acls groups to:
21/10/23 05:46:05 INFO SecurityManager: Changing modify acls groups to:
21/10/23 05:46:05 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(ubuntu, maskertim); groups with view permissions: Set(); users  with
modify permissions: Set(ubuntu, maskertim); groups with modify permissions: Set()
21/10/23 05:46:07 INFO TransportClientFactory: Successfully created connection to /192.168.0.199:46297 after 293 ms (0 ms spent in bootstraps)
21/10/23 05:46:07 INFO SecurityManager: Changing view acls to: ubuntu,maskertim
21/10/23 05:46:07 INFO SecurityManager: Changing modify acls to: ubuntu,maskertim
21/10/23 05:46:07 INFO SecurityManager: Changing view acls groups to:
21/10/23 05:46:07 INFO SecurityManager: Changing modify acls groups to:
21/10/23 05:46:07 INFO SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users  with view permissions: Set(ubuntu, maskertim); groups with view permissions: Set(); users  with
modify permissions: Set(ubuntu, maskertim); groups with modify permissions: Set()
21/10/23 05:46:07 INFO TransportClientFactory: Successfully created connection to /192.168.0.199:46297 after 16 ms (0 ms spent in bootstraps)
21/10/23 05:46:08 INFO DiskBlockManager: Created local directory at /tmp/spark-cf951b75-848e-4d26-a4ed-b638c82f54f7/executor-0471f6e3-263e-4fea-8f04-7036555185ab/blockmgr-ba8e079d-84cd-4ddc-a01c-5aeccd9163c7
21/10/23 05:46:08 INFO MemoryStore: MemoryStore started with capacity 434.4 MiB
21/10/23 05:46:09 INFO CoarseGrainedExecutorBackend: Connecting to driver: spark://CoarseGrainedScheduler@192.168.0.199:46297
21/10/23 05:46:09 INFO WorkerWatcher: Connecting to worker spark://Worker@192.168.0.200:42707
21/10/23 05:46:09 INFO TransportClientFactory: Successfully created connection to /192.168.0.200:42707 after 13 ms (0 ms spent in bootstraps)
21/10/23 05:46:09 INFO WorkerWatcher: Successfully connected to spark://Worker@192.168.0.200:42707
```

Spark 3.1.2 **stderr log page for app-20211023134559-0000/1**

Back to Master

Showing 51360 Bytes: 0 - 51360 of 51360

```
21/10/23 05:48:37 INFO CoarseGrainedExecutorBackend: Got assigned task 129
21/10/23 05:48:37 INFO Executor: Running task 1.0 in stage 35.0 (TID 129)
21/10/23 05:48:37 INFO CoarseGrainedExecutorBackend: Got assigned task 131
21/10/23 05:48:37 INFO Executor: Running task 3.0 in stage 35.0 (TID 131)
21/10/23 05:48:37 INFO TorrentBroadcast: Started reading broadcast variable 36 with 1 pieces (estimated total size 4.0 MiB)
21/10/23 05:48:37 INFO MemoryStore: Block broadcast_36_piece0 stored as bytes in memory (estimated size 3.3 KiB, free 434.1 MiB)
21/10/23 05:48:37 INFO TorrentBroadcast: Reading broadcast variable 36 took 46 ms
21/10/23 05:48:37 INFO MemoryStore: Block broadcast_36 stored as values in memory (estimated size 6.3 KiB, free 434.1 MiB)
21/10/23 05:48:37 INFO HadoopRDD: Input split: file:/tmp/household_power_consumption.txt:100663296+32297459
21/10/23 05:48:37 INFO HadoopRDD: Input split: file:/tmp/household_power_consumption.txt:33554432+33554432
21/10/23 05:48:40 INFO MemoryStore: Block taskresult_131 stored as bytes in memory (estimated size 6.6 MiB, free 427.5 MiB)
21/10/23 05:48:40 INFO Executor: Finished task 3.0 in stage 35.0 (TID 131). 6922960 bytes result sent via BlockManager)
21/10/23 05:48:40 INFO MemoryStore: Block taskresult_129 stored as bytes in memory (estimated size 7.0 MiB, free 420.5 MiB)
21/10/23 05:48:40 INFO Executor: Finished task 1.0 in stage 35.0 (TID 129). 7324993 bytes result sent via BlockManager)
21/10/23 05:48:46 INFO CoarseGrainedExecutorBackend: Got assigned task 132
21/10/23 05:48:46 INFO Executor: Running task 0.0 in stage 36.0 (TID 132)
21/10/23 05:48:46 INFO TorrentBroadcast: Started reading broadcast variable 37 with 1 pieces (estimated total size 4.0 MiB)
21/10/23 05:48:46 INFO MemoryStore: Block broadcast_37_piece0 stored as bytes in memory (estimated size 31.7 KiB, free 434.0 MiB)
21/10/23 05:48:46 INFO TorrentBroadcast: Reading broadcast variable 37 took 39 ms
21/10/23 05:48:46 INFO MemoryStore: Block broadcast_37 stored as values in memory (estimated size 87.7 KiB, free 434.0 MiB)
21/10/23 05:48:46 INFO HadoopMapRedCommitProtocol: Using output committer class org.apache.hadoop.mapred.FileOutputCommitter
21/10/23 05:48:46 INFO FileOutputCommitter: File Output Committer Algorithm version is 1
21/10/23 05:48:46 INFO FileOutputCommitter: FileOutputCommitter skip cleanup _temporary folders under output directory:false, ignore cleanup failures: false
21/10/23 05:48:46 INFO HadoopRDD: Input split: file:/tmp/household_power_consumption.txt:0+33554432
21/10/23 05:48:50 INFO HadoopRDD: Input split: file:/tmp/household_power_consumption.txt:33554432+33554432
21/10/23 05:48:54 INFO HadoopRDD: Input split: file:/tmp/household_power_consumption.txt:67108864+33554432
21/10/23 05:48:59 INFO HadoopRDD: Input split: file:/tmp/household_power_consumption.txt:100663296+32297459
21/10/23 05:49:03 INFO FileOutputCommitter: Saved output of task 'attempt_2021102313484685746364882754280024_0029_m_000000_0' to file:/tmp/global_intensity/_temporary
/0/task_2021102313484685746364882754280024_0029_m_000000
21/10/23 05:49:03 INFO SparkHadoopMapRedUtil: attempt_2021102313484685746364882754280024_0029_m_000000_0: Committed
21/10/23 05:49:03 INFO Executor: Finished task 0.0 in stage 36.0 (TID 132). 1287 bytes result sent to driver
21/10/23 05:49:03 INFO CoarseGrainedExecutorBackend: Driver commanded a shutdown
21/10/23 05:49:03 INFO CoarseGrainedExecutorBackend: Driver from 192.168.0.199:46297 disconnected during shutdown
21/10/23 05:49:03 INFO CoarseGrainedExecutorBackend: Driver from 192.168.0.199:46297 disconnected during shutdown
21/10/23 05:49:03 ERROR CoarseGrainedExecutorBackend: RECEIVED SIGNAL TERM
199:46297 disconnected during shutdown
```

Load New

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.

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:

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)

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.