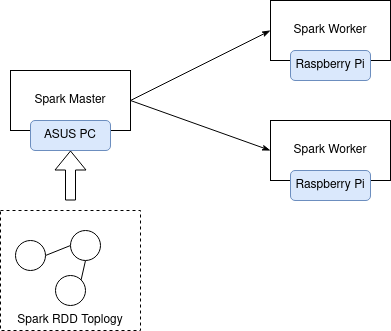
Big Data Mining Homework 1

# Spark Platform:

The platform consists of:

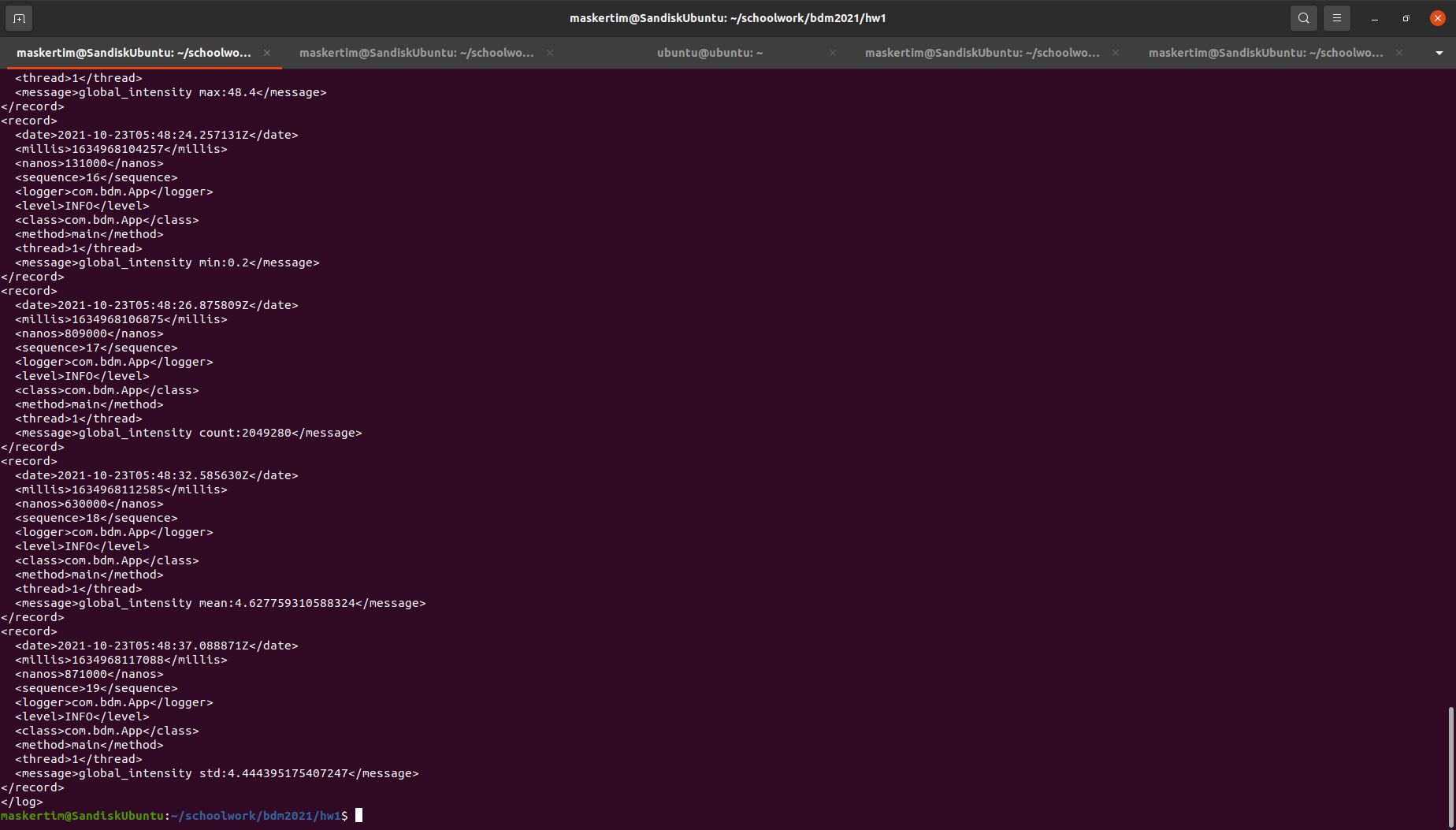
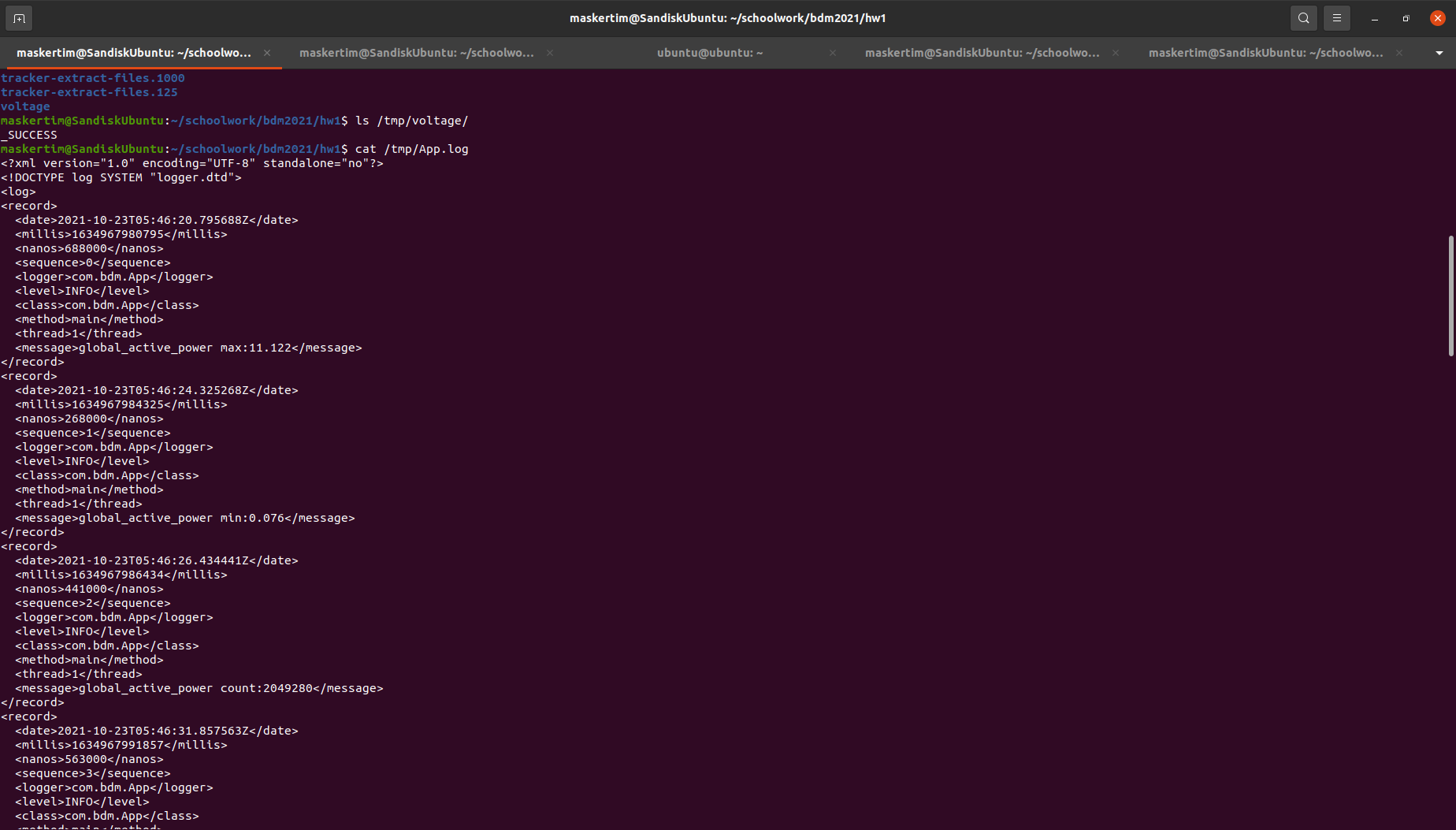
1. Raspberry Pi 4 Model B x2
   * OS: Linux Ubuntu 20.04 Server
   * CPU architecture: aarch64
   * RAM: 8GB
   * CPU: Broadcom BCM2711, Quad core Cortex-A72 (ARM v8) 64-bit SoC @ 1.5GHz
   * Number of CPU: 4C (CPU) 1T (Thread Per CPU)
2. Asus-vivobook notebook
   * OS: Linux Ubuntu 20.04 LTS
   * CPU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

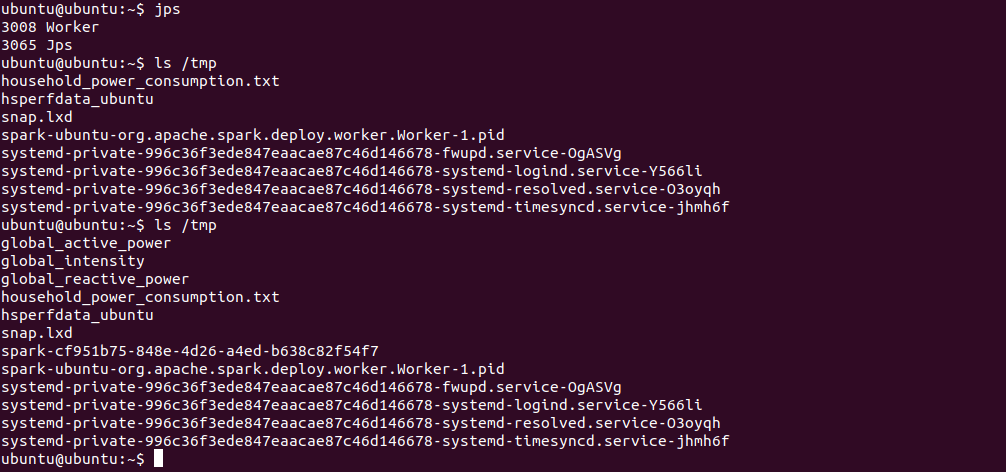


# The generated output:

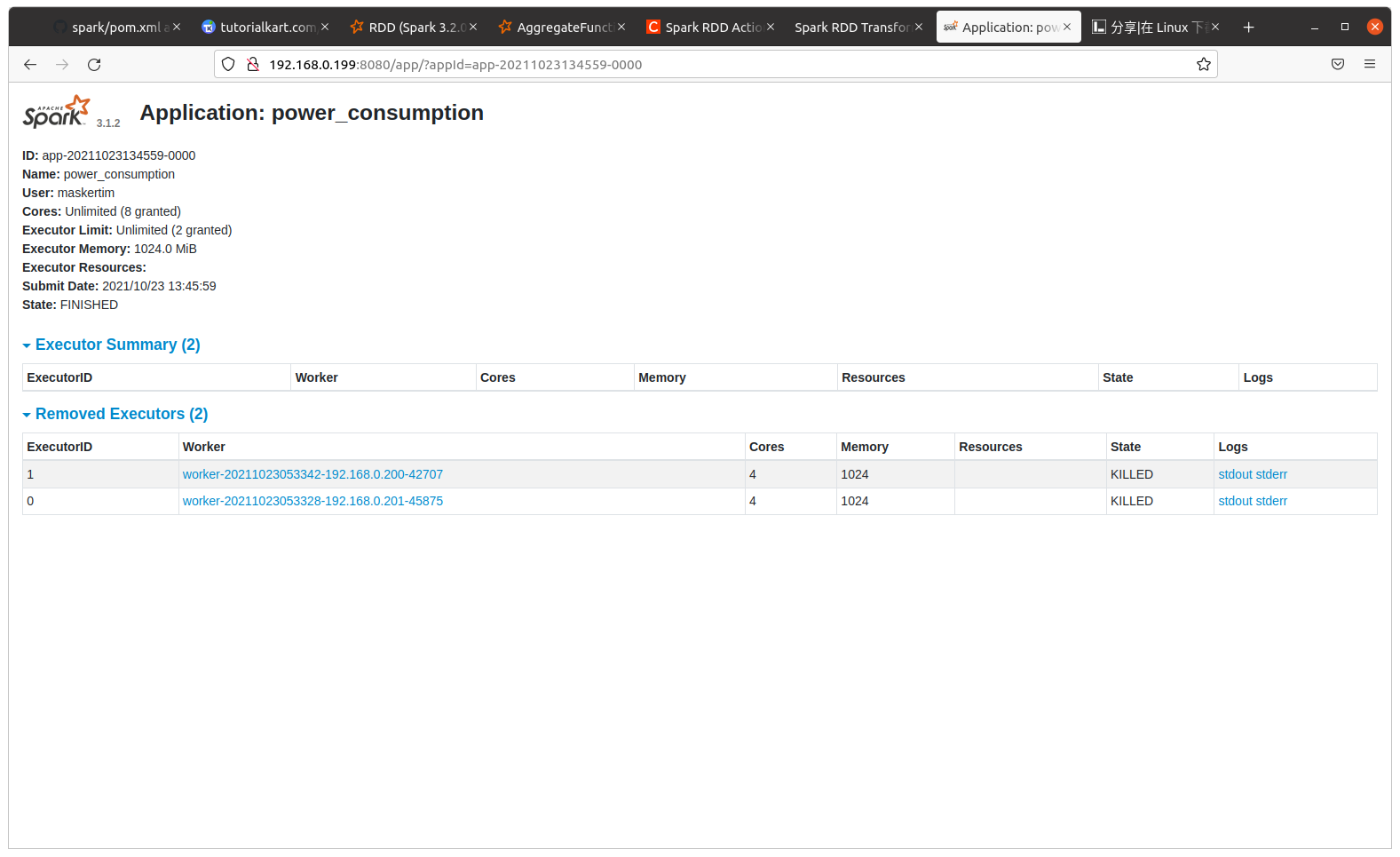
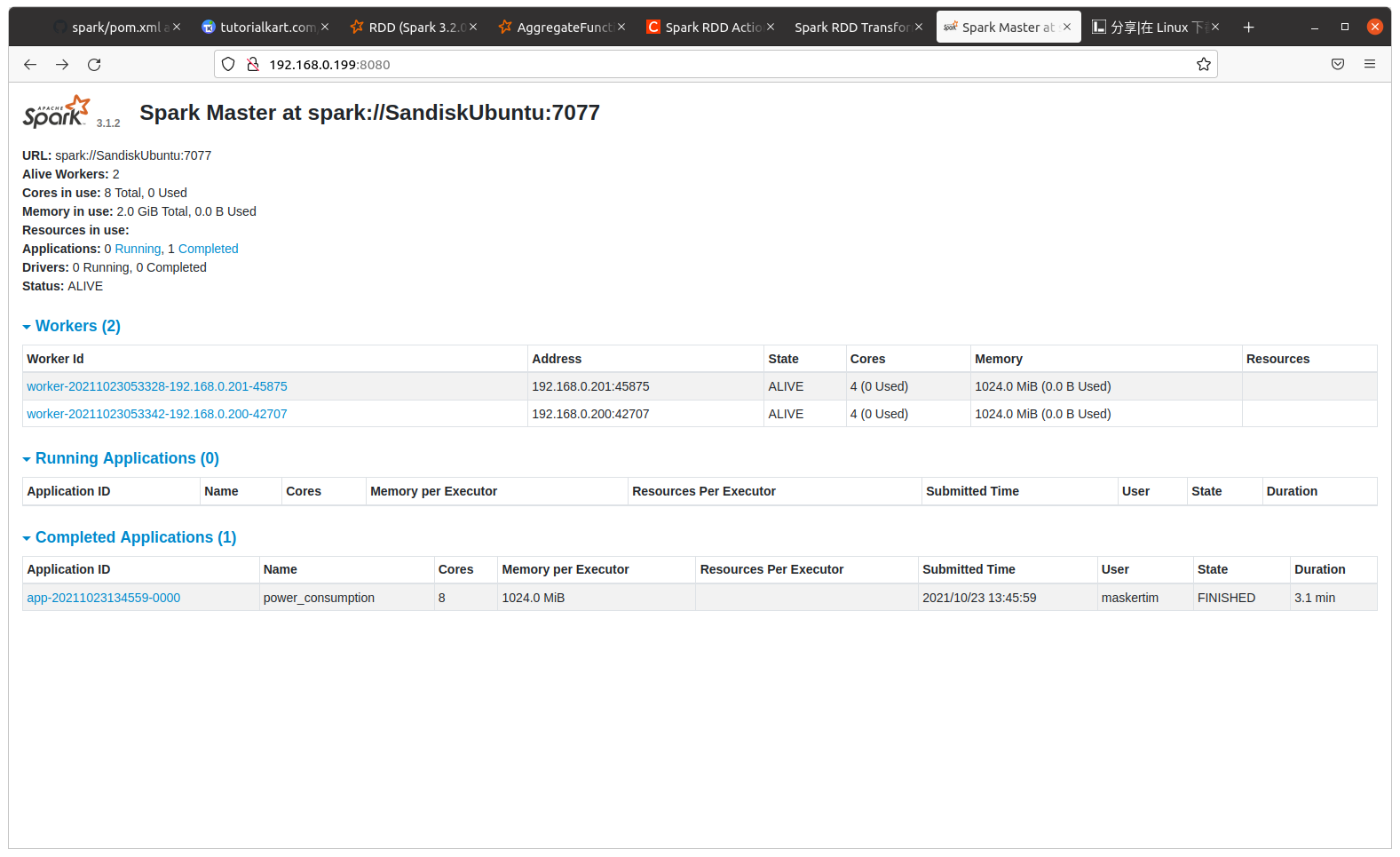
### Log the RDD (Resilient Distributed Dataset) output



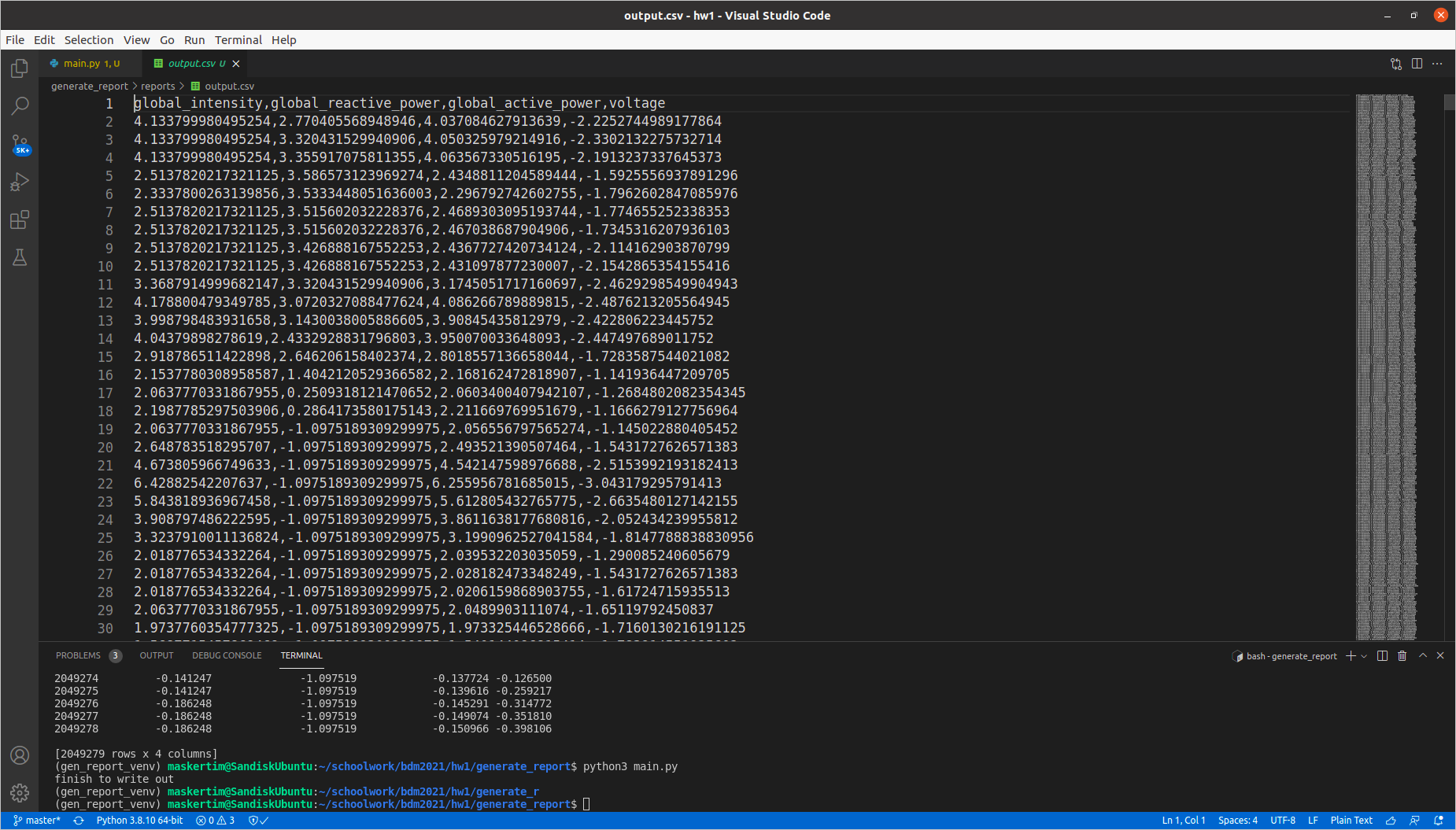
### The completed data processed by spark cluster on Raspberry Pi

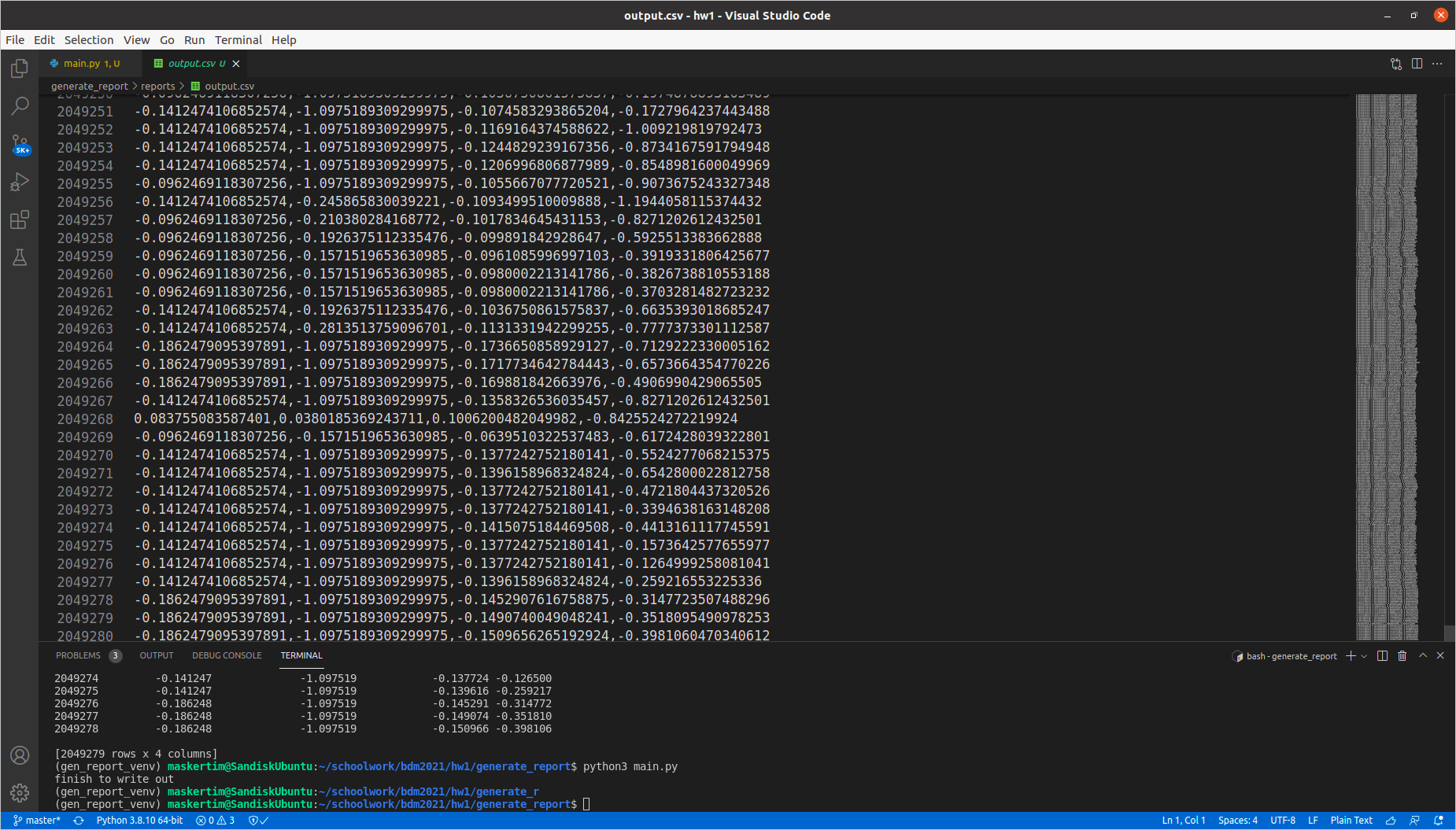


### Spark Web GUI and Logging of Spark Cluster

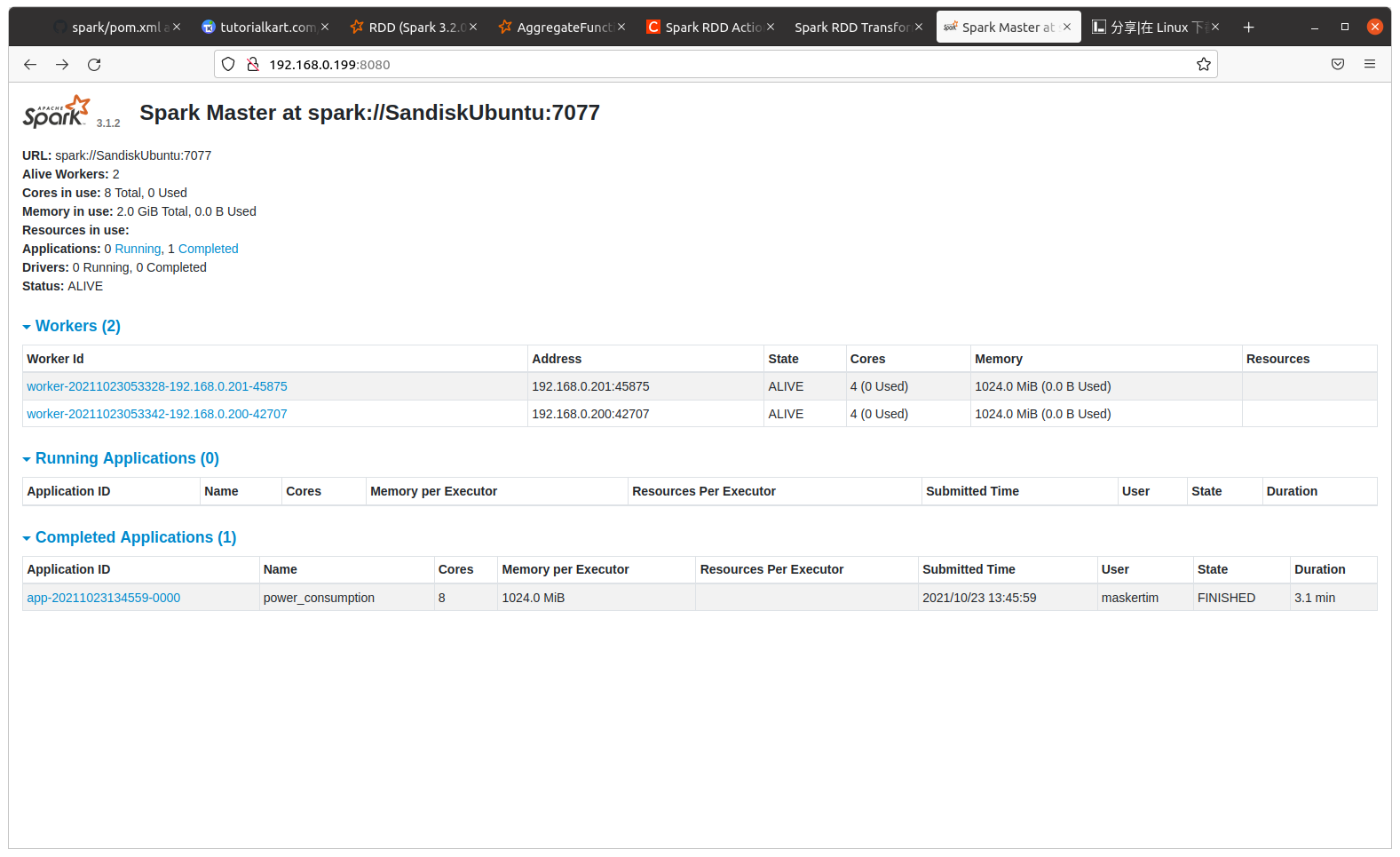
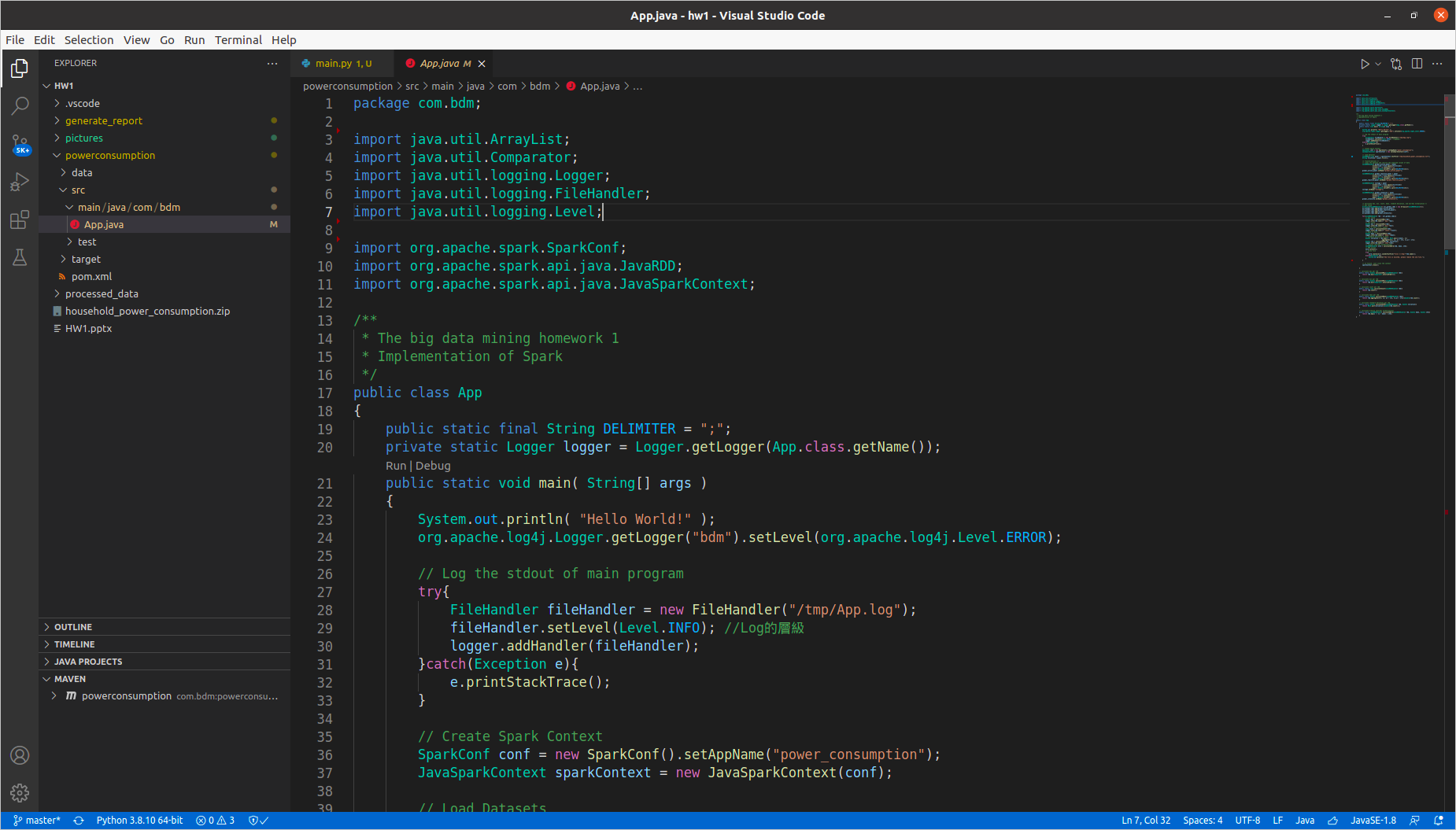
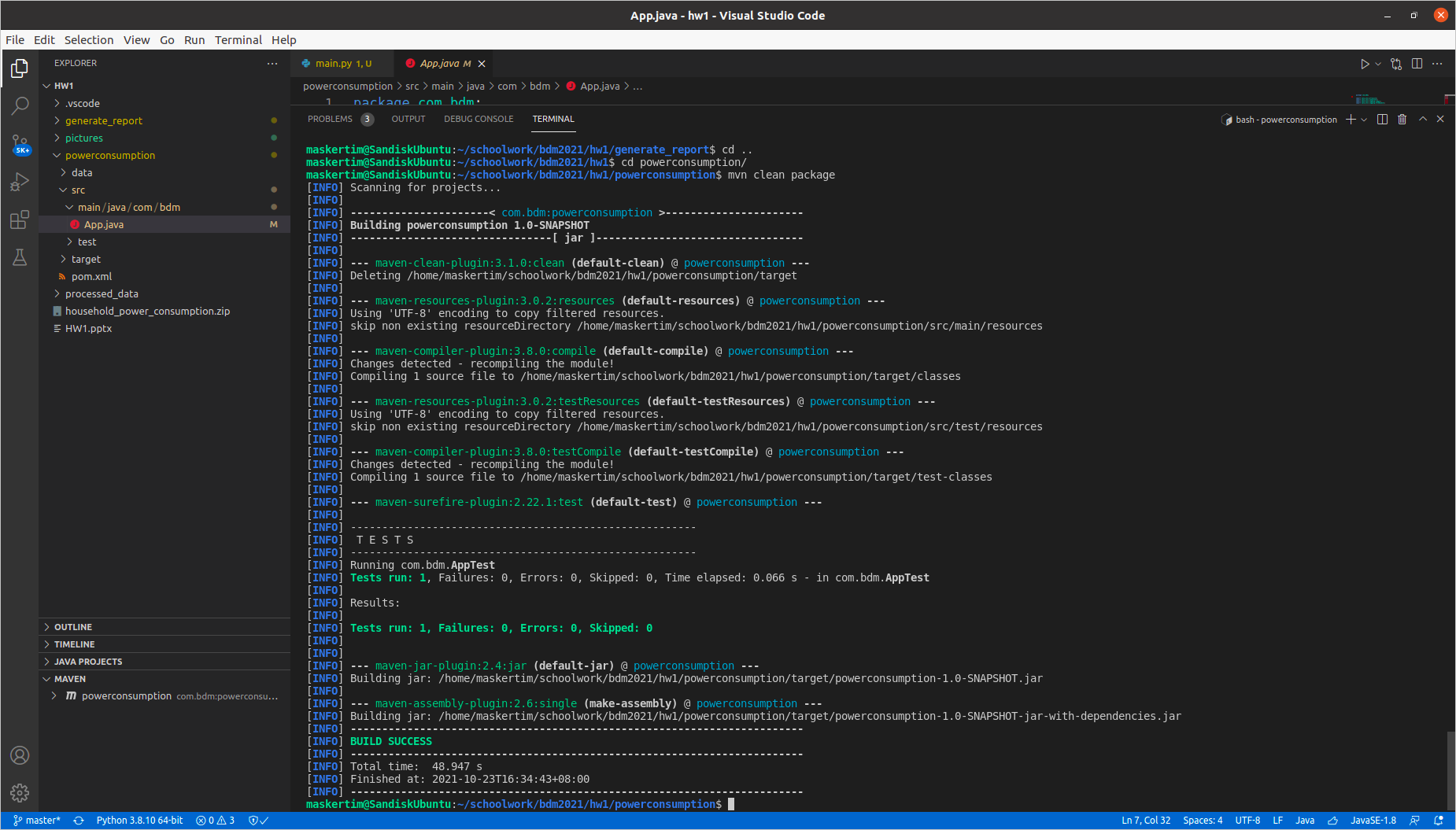


### Report of min-max normalization (using formula of Z-Score)





# 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.