TUGAS RESPONSI KELOMPOK

Logo

Description automatically generated

Disusun Oleh :

LUSI PRI HARTINI (205410059)

LALU ISARI AT TAMIMI (205410066)

MUHAMMAD DANU FIRDAUS (205410112)

PROGRAM STUDI INFORMATIKA

PROGRAM SARJANA

FAKULTAS TEKNOLOGI INFORMASI

UNIVERSITAS TEKNOLOGI DIGITAL INDONESIA

YOGYAKARTA

2023

Tugas :

Apache Spark

a. Buat infrastruktur spark pada mesin EC2 (boleh sesuai praktikum), untuk model

deploy bisa menggunakan container atau mesin AWS EC2. Konsep yang

digunakan adalah multinode

b. Boleh menggunakan script aplikasi hitung kata

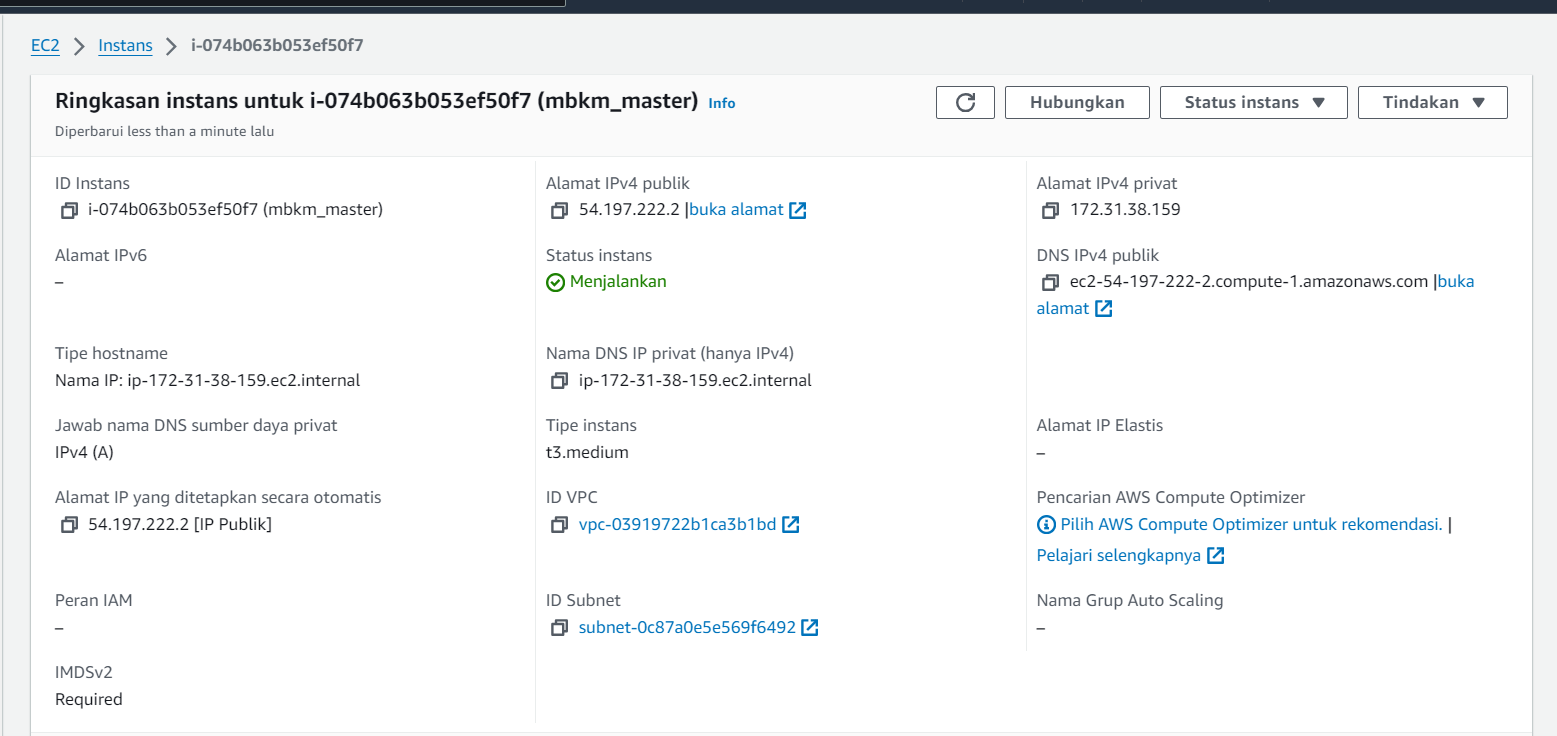
c. Melakukan proses eksperimen dengan dataset berukuran 25 MB, 50MB, 75MB,

100MB, 125MB, 150MB, 175MB, 200MB

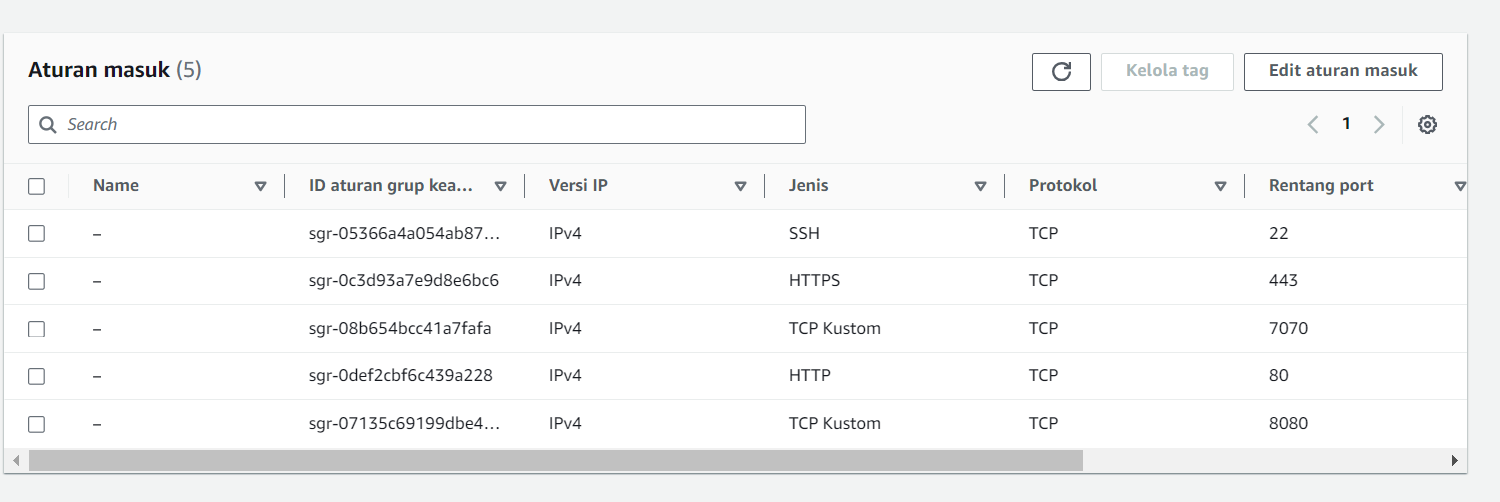
Langkah – langkah untuk mengerjakannya sebagai berikut :

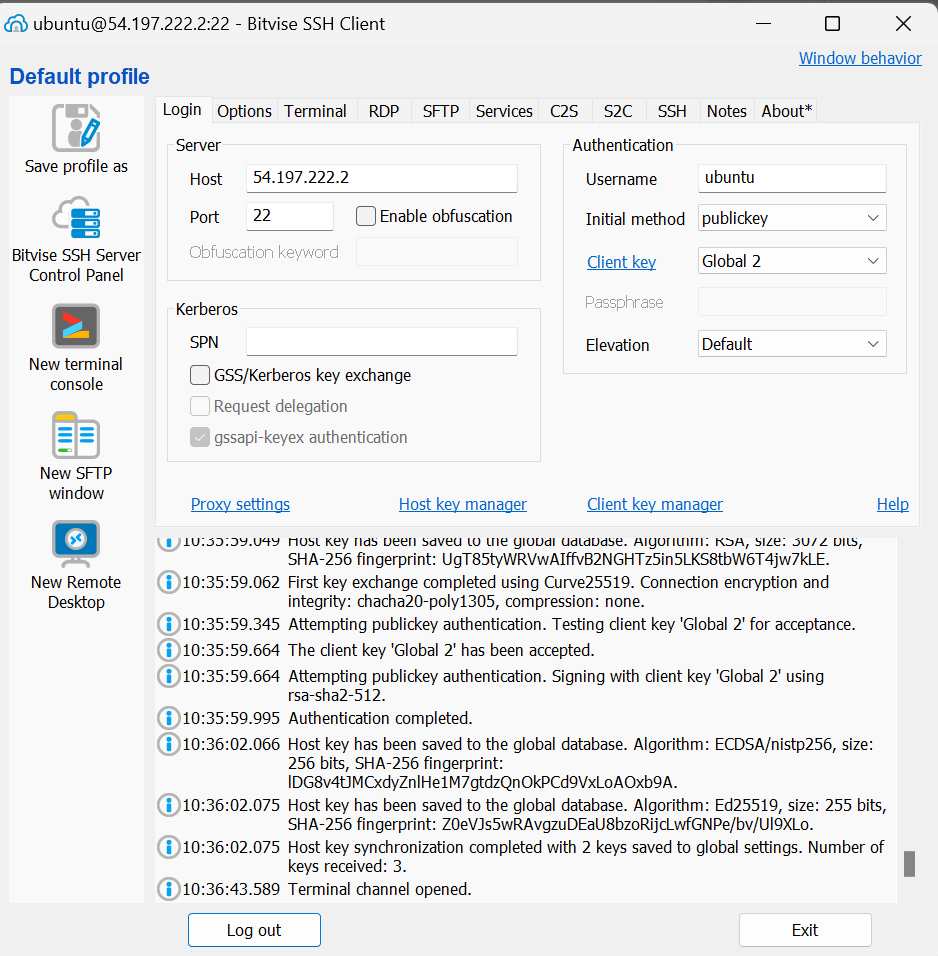
1. Menyiapkan mesin ec2 untuk master\_sparks

Masuk ke aws console, dan membuat ec2, berikut tampilan untuk spesifikasi ec2 yang baru dibuat



Setelah membuat mesin, menambahkan edit aturan masuk di group keamanan dan menambahkan port untuk 8080 sebagai server UI sparks dan 7070 sebagai nanti digunakan workers\_sparks karna ini multinode nanti akan membuat 1 mesin lagi untuk worker\_sparks nya.



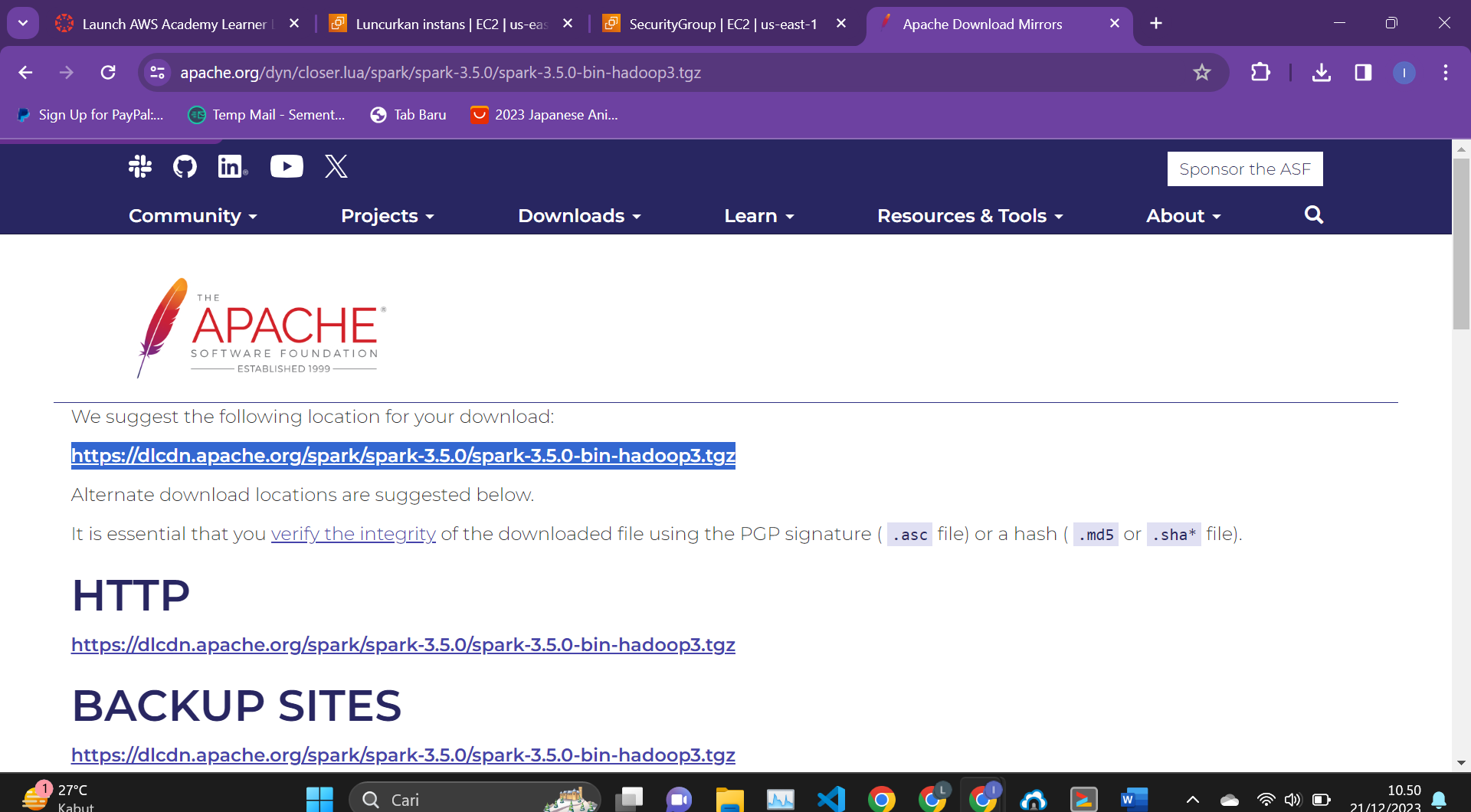
Pada praktikum ini menggunakan bitvise untuk memudahkan dalam upload file nya ke server, kalau menggunakan s3bucket amazon perlu membuat AMI saat proses pengaksesan file oleh skrip penghitung kata.  
  


1. Instalasi apache spark

Klik terminal dan jalankan sintaks :

* sudo apt update
* sudo apt upgrade,
* dan sudo apt install default-jdk

setelah proses instalasi jdk selesai, maka tinggal menginstall apache spark,   
buka website resmi apache spark



Setelah itu copy link download nya dan buka terminal ec2 yang terkoneksi di bitvise , kemudian masukkan sintaks” wget link download “



Setelah itu ekstrak hasil download nya :



1. Konfigurasi Lingkungan:

Tambahkan spark ke PATH

“

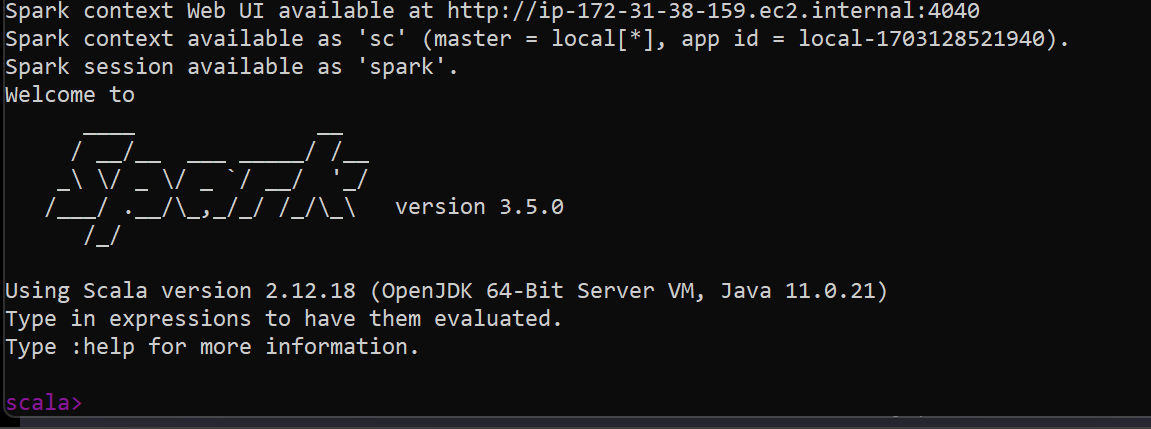
echo "export SPARK\_HOME=~/spark-3.5.0-bin-hadoop3" >> ~/.bashrc

echo "export PATH=$PATH:$SPARK\_HOME/bin" >> ~/.bashrc

source ~/.bashrc

“

Coba jalankan spark shell sebagai tanda instalasi berhasil :



1. Konfigurasi Spark Multinode

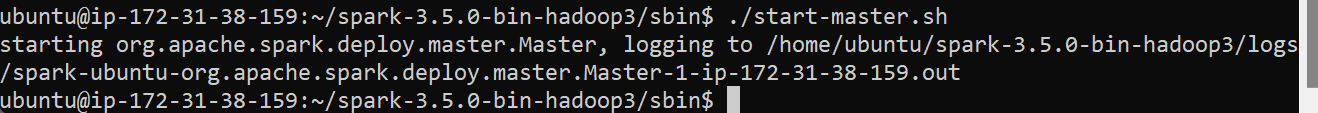
Pertama jalankan server untuk spark master yang dipilih ec2 yang tadi,

“

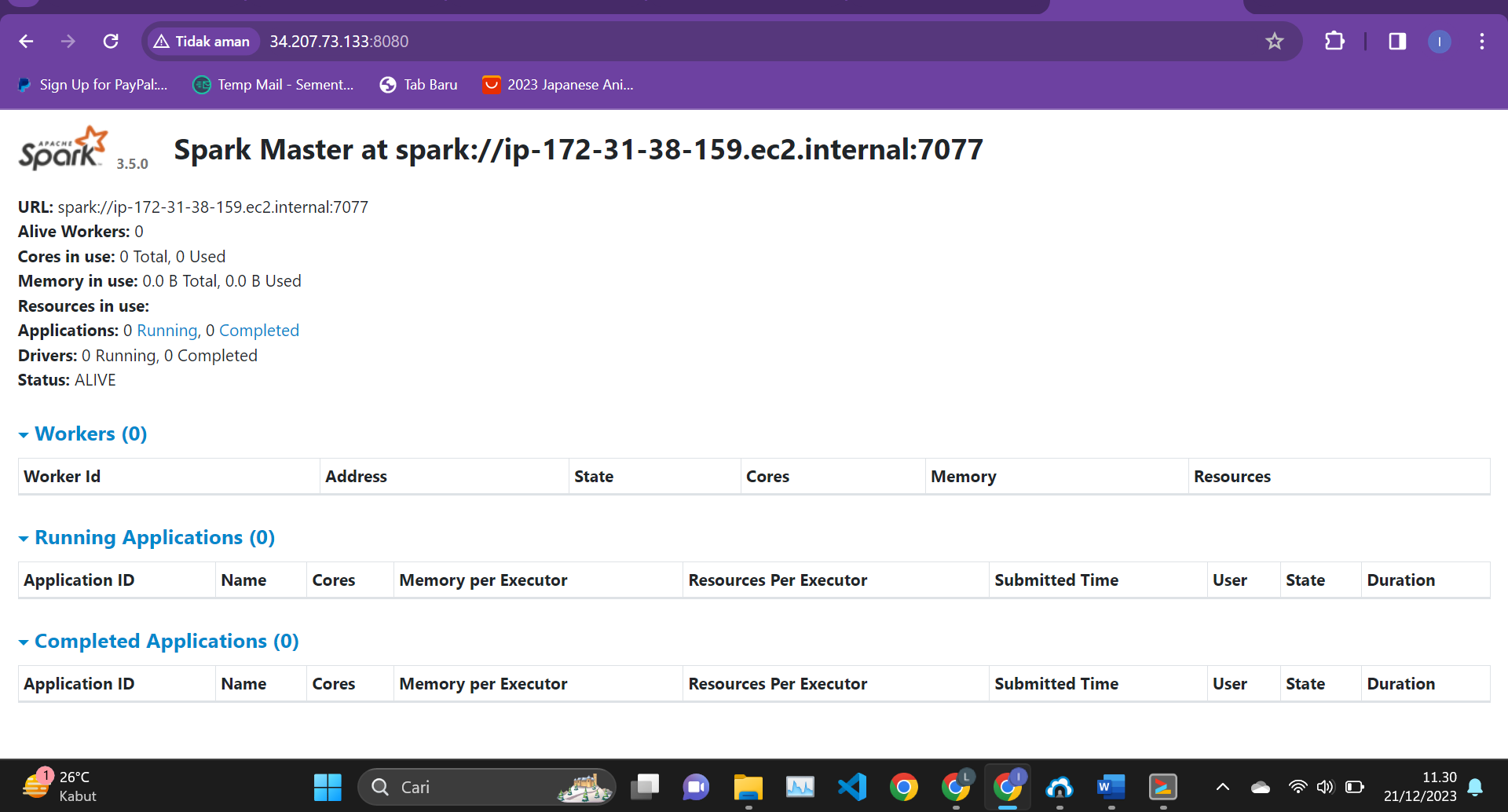
cd ~/spark-3.5.0-bin-hadoop3/sbin

./start-master.sh

“



Kemudian akses dan buka ip:8080 atau = http://34.207.73.133:8080/



Setelah itu buat lagi ec2 baru untuk workers spark, karena disini diminta untuk multinode , buat mesin ec2 yang sama aturan keamanan dan juga instalkan apache spark

Setelah terinstall, masuk ke path "

cd /path/to/spark/conf

“

Setelah itu masuk “

nano spark-env.sh

“

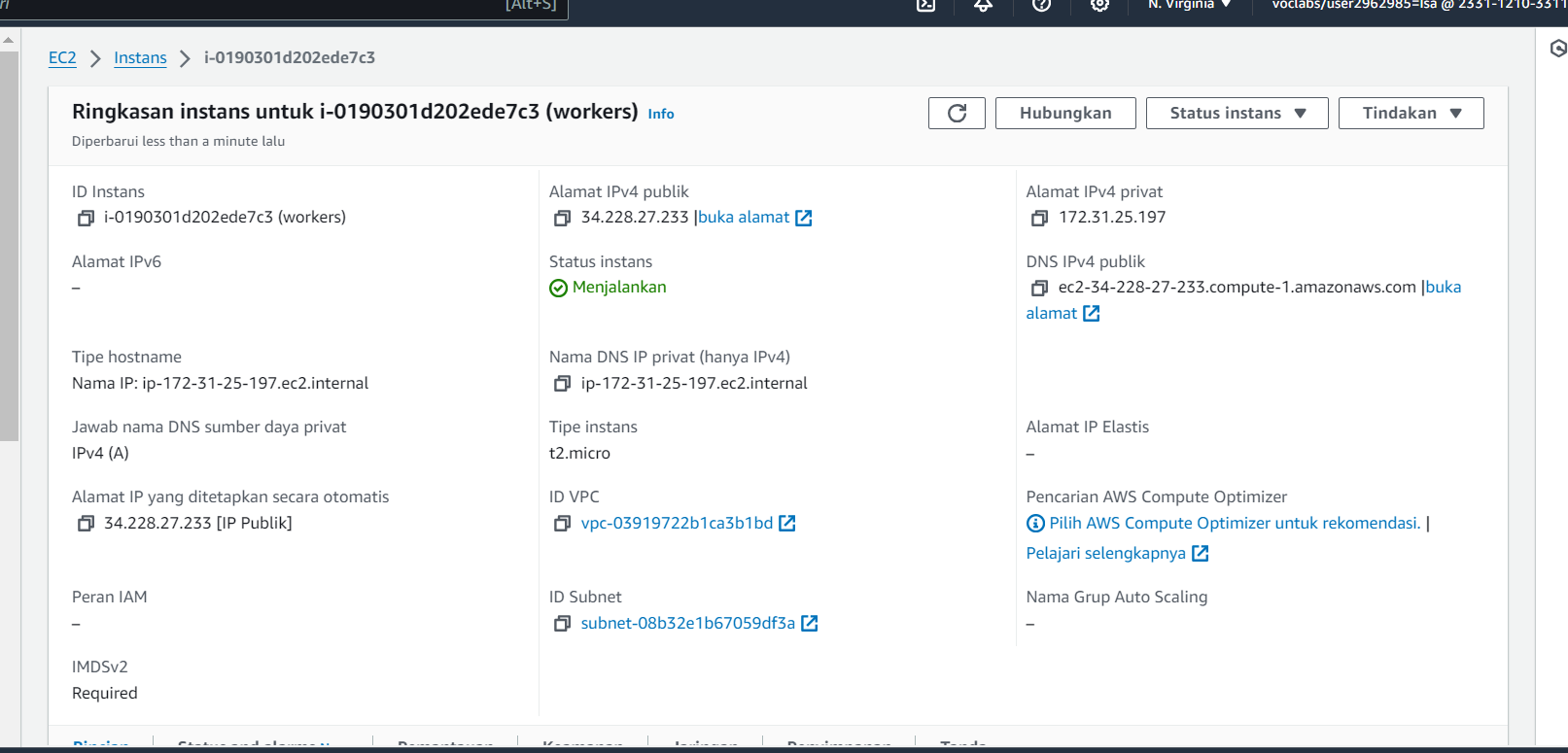
Untuk mensetting ip masternodes dan menyatakan ec2 ini sebagai workers,

“ SPARK\_MASTER\_HOST=ip-172-31-38-159.ec2.internal “

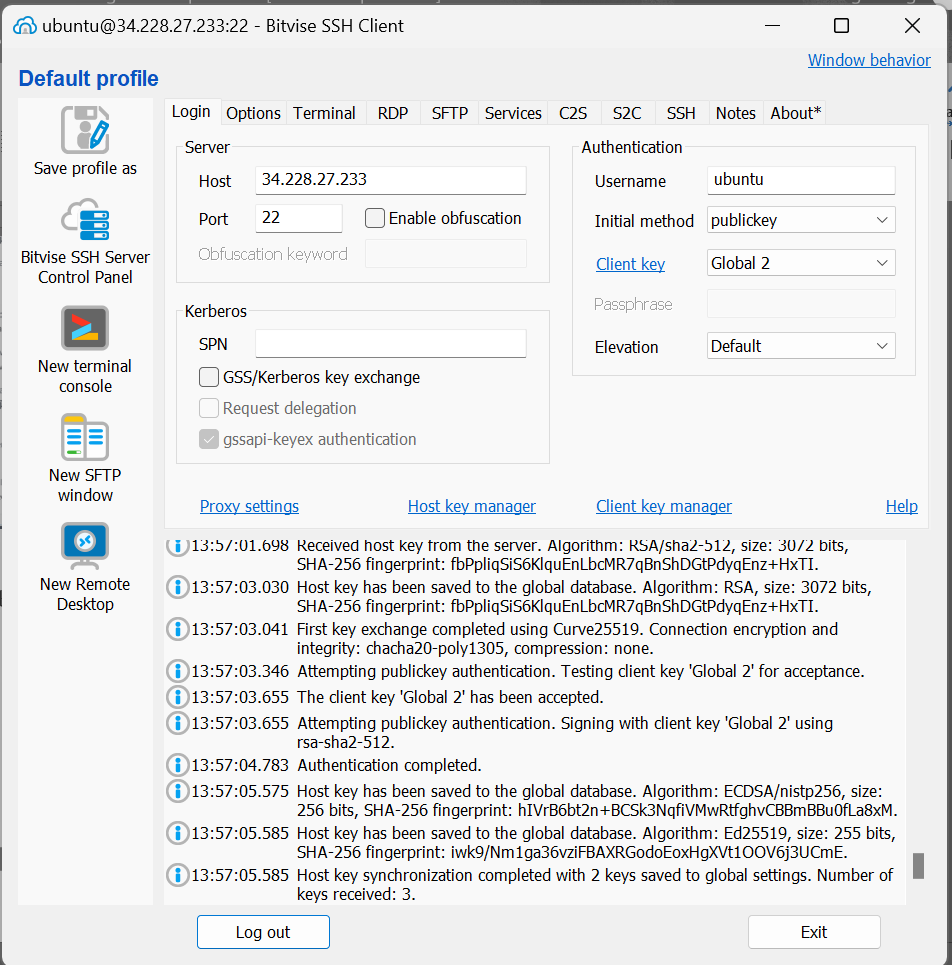
Setelah itu ctrl + o , ctrl + y, ctrl + x untuk menyimpan dan keluar,

Karena disini kami menggunakan bitvise tidak menggunakan bucket s3 amazon, karena itu kita perlu login juga di bitvise untuk ec2 nya ,

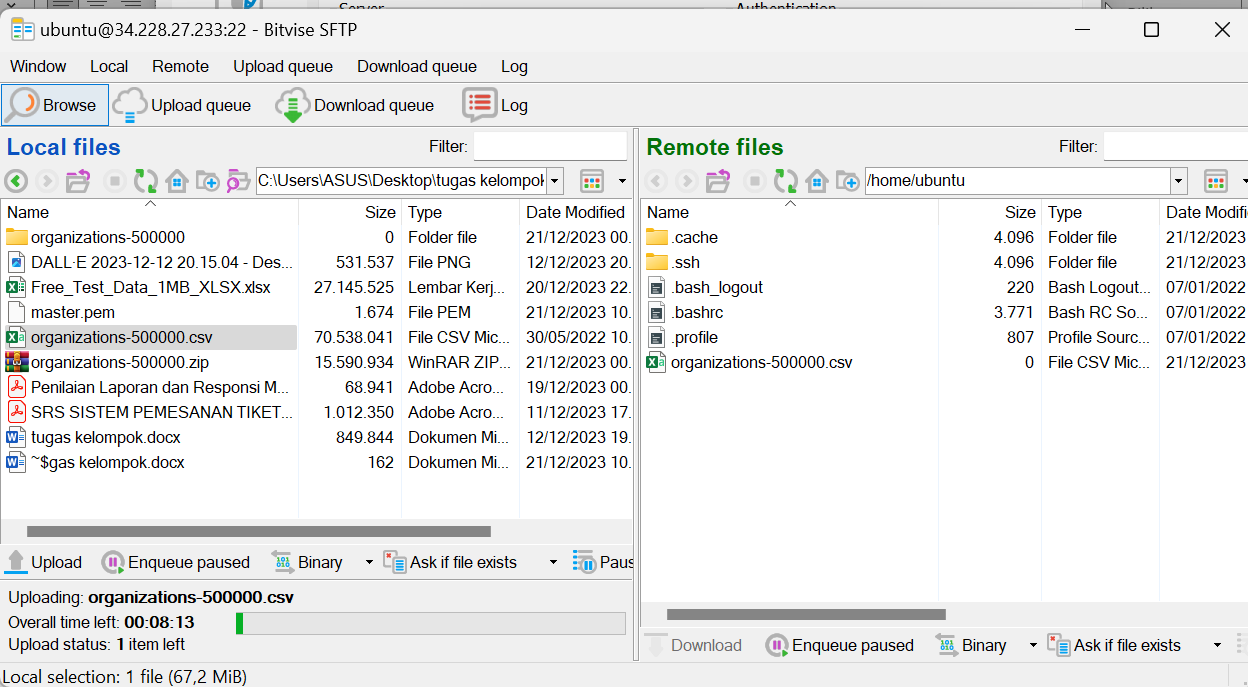
Berikut tampilan mesin ec2 workers :



Dan berikut proses koneksi nya di bitvise :



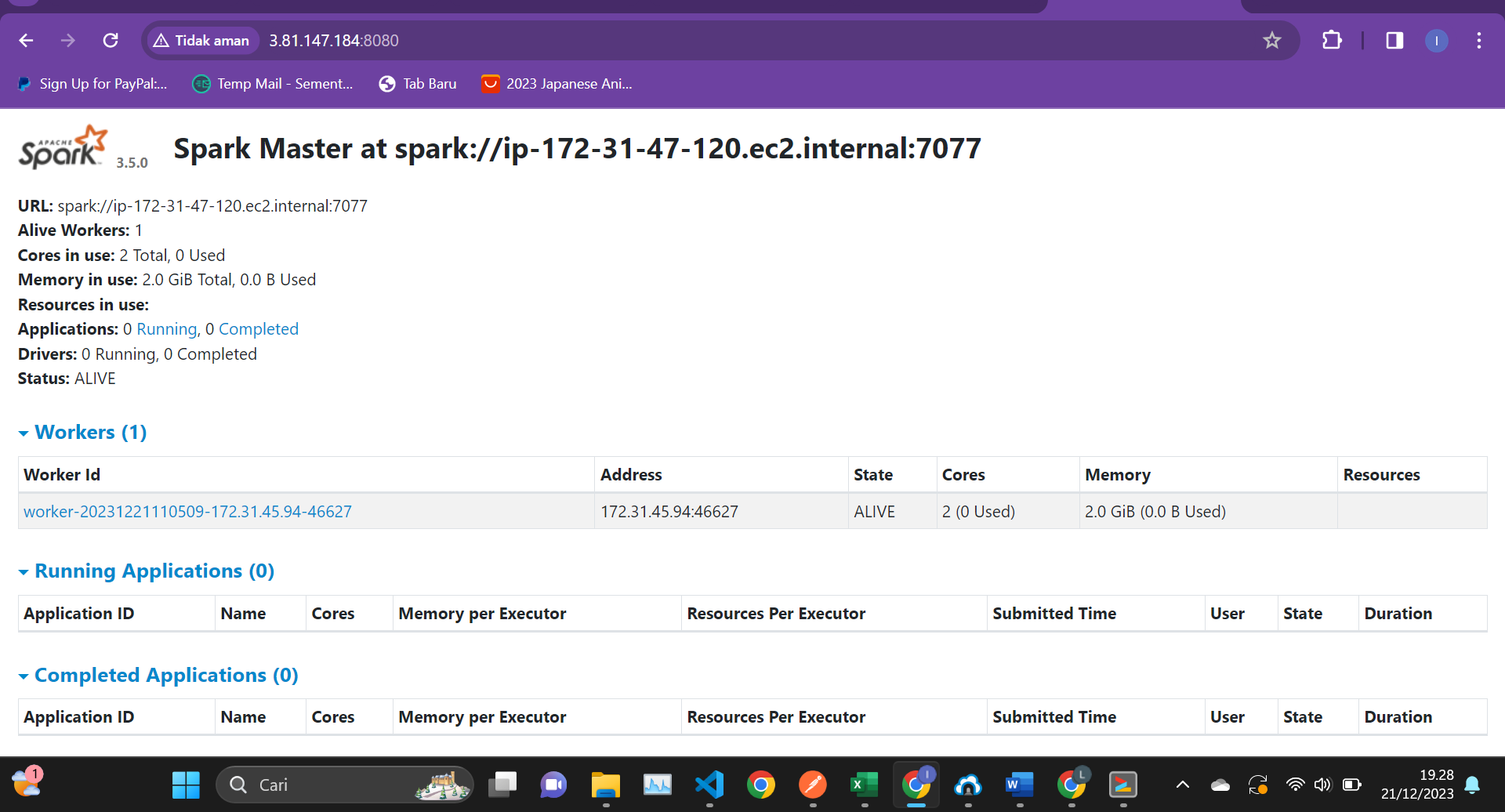
Setelah itu kita upload file csv ke kedua mesin, workers dan juga master spark nya : caranya dengan menekan tombol new SFTP window > drag and drop file csv :



Di ec2 workers jalankan sebagai settingan untuk workernya “ start-worker.sh spark://ip-172-31-47-120:7077 “

Nanti ip-172-31-47-120:7077 dengan alamat IP atau hostname master node yang sesuai.

Setelah menjalankan start-worker di directory tempat spark terinstall, maka setelah itu check UI spark master



Terlihat Alive Workers kita berhasil terhubung, ini berarti sudah multinode.

Setelah itu buat aplikasi atau script penghitung kata

Buat file sekaligus mengeditnya langsung dengan nano wordcount.py:

“ GNU nano 6.2 wordcount.py

from pyspark.sql import SparkSession

from pyspark.sql.functions import explode, split, col

# Inisialisasi Spark Session

spark = SparkSession.builder.appName("WordCount").getOrCreate()

# Membaca file CSV dan memilih kolom 'Description' saja

df = spark.read.option("header", "true").csv("/home/ubuntu/organizations-100.csv")

descriptions = df.select(split(col("Description"), " ").alias("words"))

# Memisahkan kata-kata dan menghitung frekuensinya

word\_counts = (descriptions.select(explode(col("words")).alias("word"))

.groupBy("word")

.count()

.sort("count", ascending=False))

# Menampilkan hasilnya

word\_counts.show()

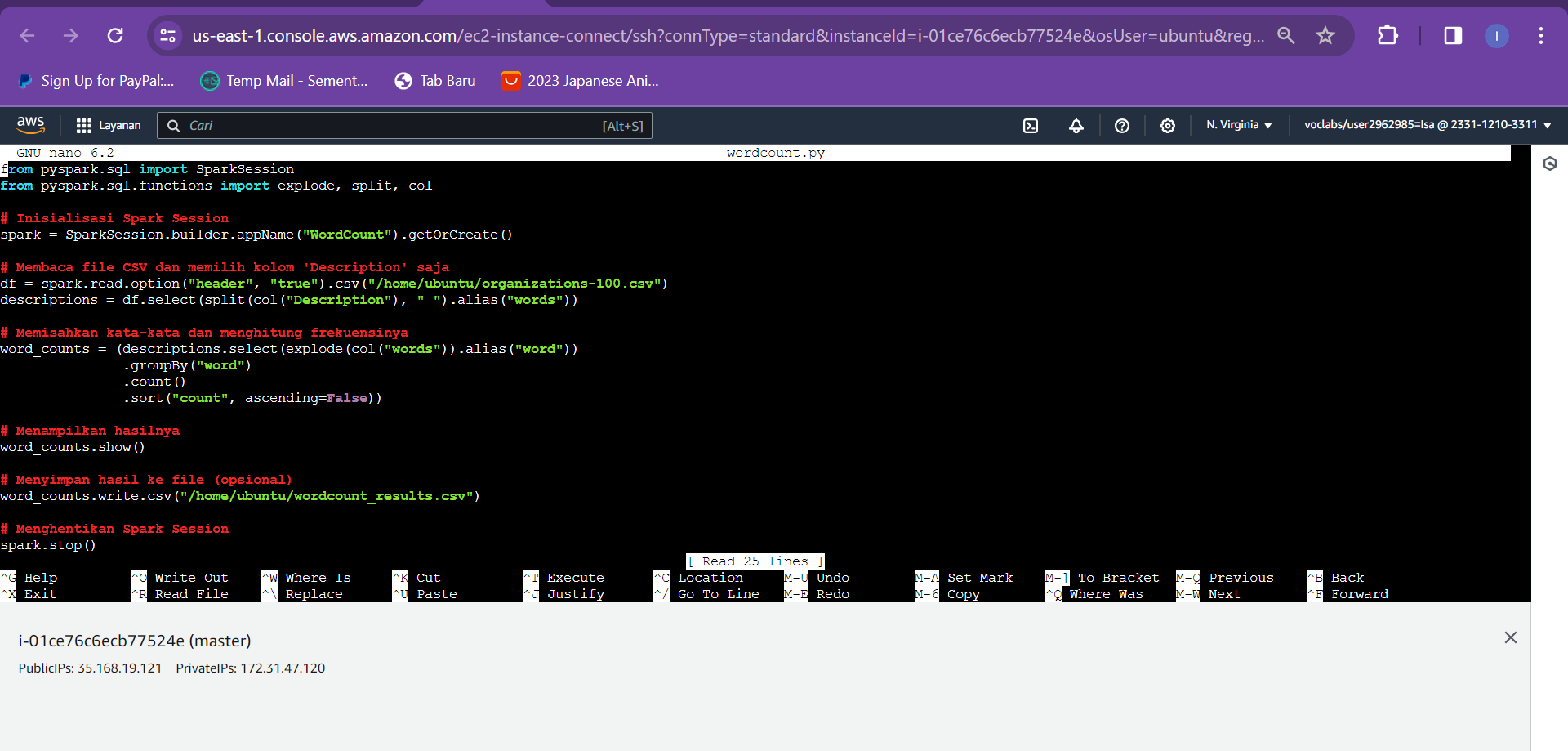
# Menyimpan hasil ke file (opsional)

word\_counts.write.csv("/home/ubuntu/wordcount\_results.csv")

# Menghentikan Spark Session

spark.stop()

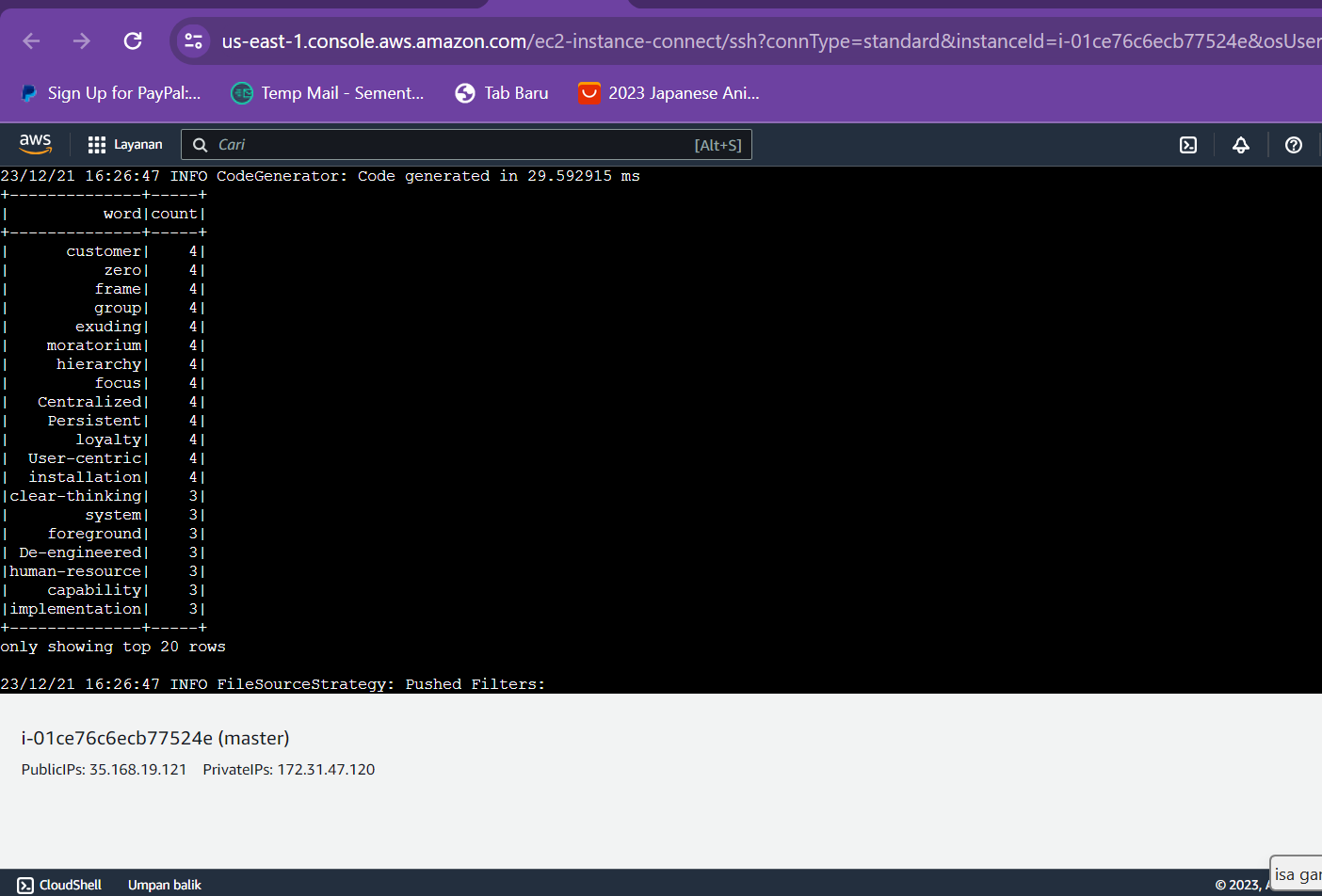
”



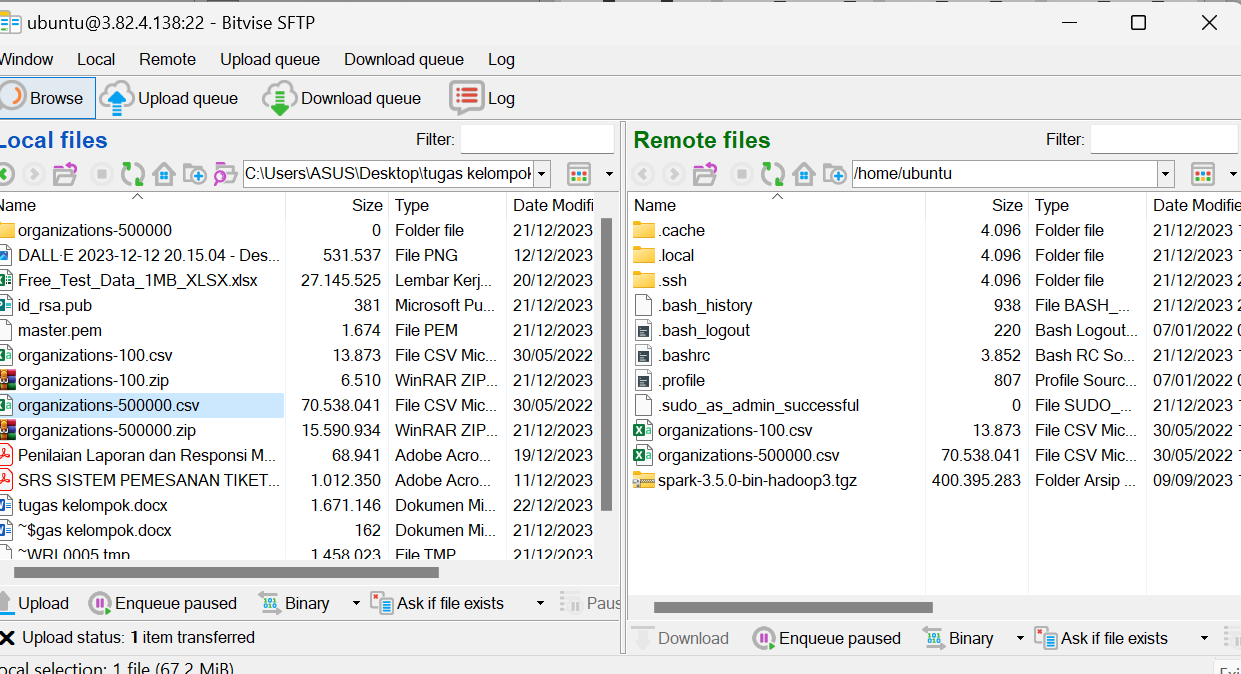
Untuk script penghitung katanya cukup di ec2 master spark saja, karna nanti workers akan menerima sebaran dari master.

Berikut hasilnya untuk file organizations-100.csv

Yang dimana dari file script penghitung kata yang akan menghitung kata di colom Description.

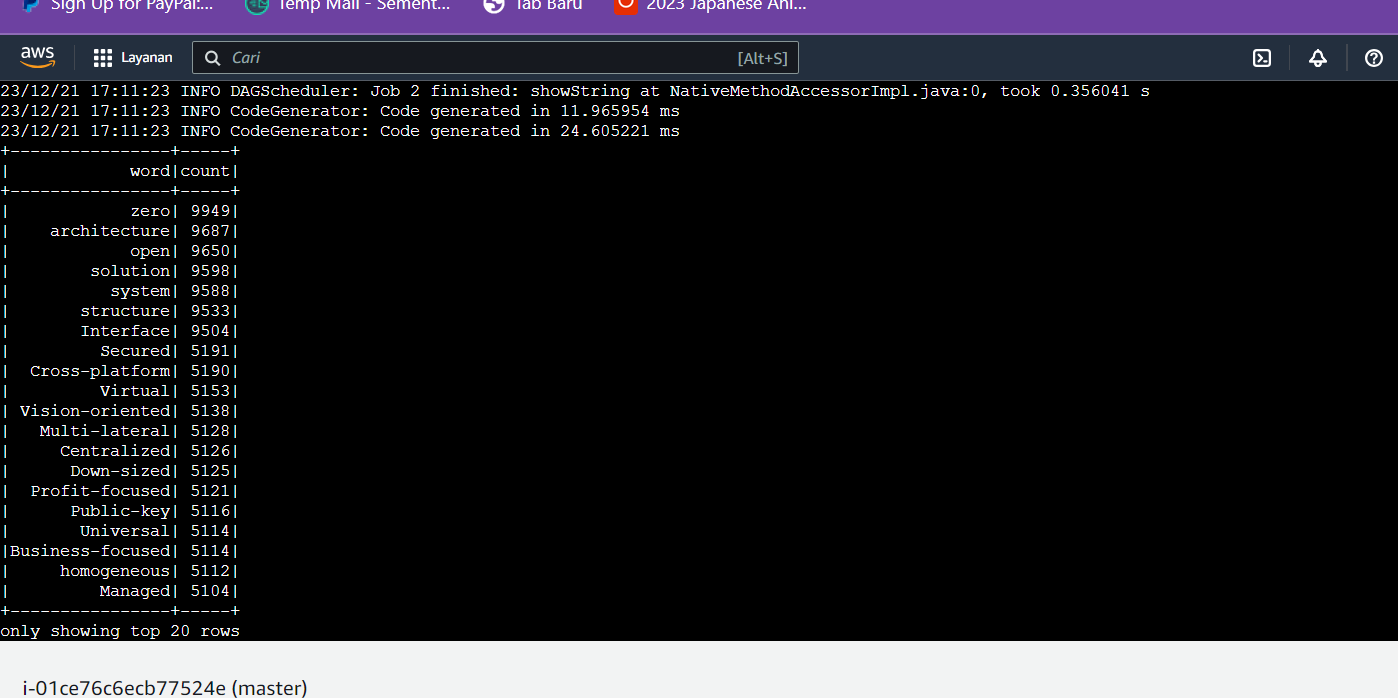


Kemudian kita coba lagi dengan file csv yang lebih besar,

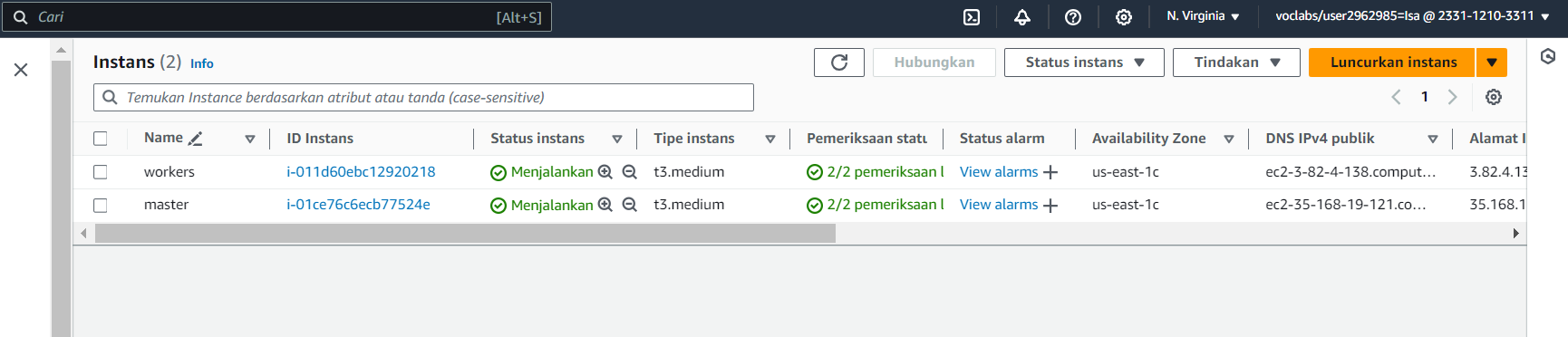


Bisa dilihat dari gambar diatas kami menambahkan file csv dengan size 76mb.

Dengan methode pencarian semua kata di nama kolom Descryption.



Berikut tampilan aws instance kami :



Dan berikut tampilan UI Spark dari Master :

Kode kunci :

start-master.sh => digunakan untuk menjalankan ec2 sebagai master sparks

start-worker.sh spark://ip-172-31-47-120:7077 => digunakan di ec2 workers untuk membuat ec2 itu terhubung dengan master spark sebagai worker

spark-submit wordcount.py => digunakan di master spark untuk menjalankan script di file wordcount.py