SCHOOL OF BIOTECHNOLOGY

Nile University

A HYBRID HIGH-PERFORMANCE COMPUTING AND BIG DATA FRAMEWORK FOR PARALLELIZED GENE-PROTEIN EXPRESSION ANALYSIS AND DISEASE CLASSIFICATION

CBIO312: HIGH PERFORMANCE COMPUTING

Name and ID:

Farah Ibrahim 221001140

Malak Atef 221000906

Zeina Ahmed 221000417

Under Supervision of: Dr. Mohamed EL-Sayeh

INTRODUCTION

- Project Goal: Build and compare traditional HPC and Big Data frameworks for analyzing gene/protein expression data.

Tools Used:

- VirtualBox + Ubuntu (for VM cluster)
- MPI + mpi4py (Task 1)
- Docker + Spark + PySpark (Task 2)
- scikit-learn, pandas, logistic regression

METHODOLOGY

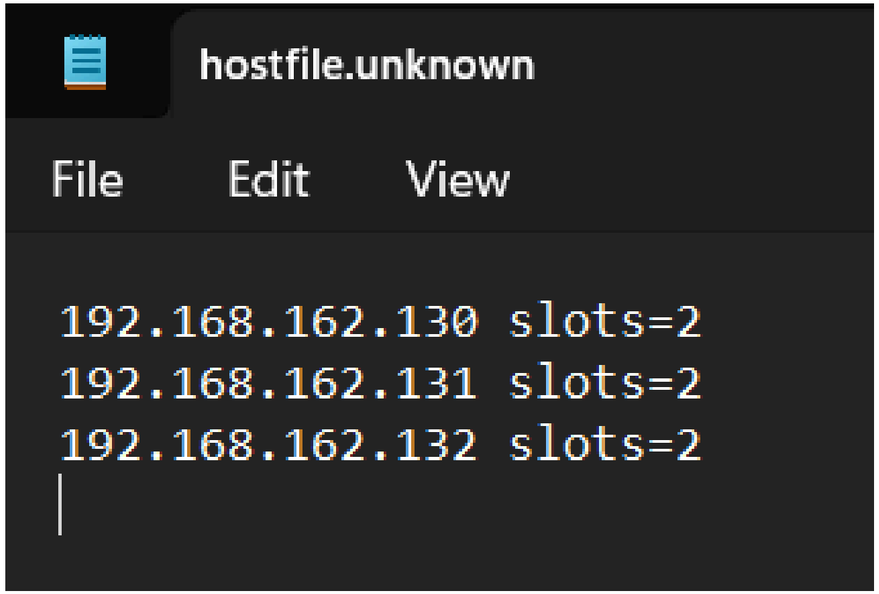
Task 1: Mini-HPC Cluster using MPI

- VM Setup:
  - 3 Ubuntu VMs (1 master, 2 workers)
  - NAT networking and hostnames mapped
- Passwordless SSH:
  - ssh-keygen and ssh-copy-id across nodes
  - Verified with test commands

METHODOLOGY & RESULTS

Task 1: Mini-HPC Cluster using MPI

- MPI Configuration:
  - Created hostfile
  - Installed openmpi, python3, mpi4py, scikit-learn
- Script: process_dataset_mpi.py
  - Part 1: Gene expression comparison (healthy vs diseased)
  - Part 2: Logistic regression on gene/protein features
  - Used mpirun -np 6 to run across nodes



METHODOLOGY & RESULTS

Task 2: Dockerized Spark Cluster

Docker Installation on All Nodes

- Master: docker swarm init
- Workers joined using docker swarm join

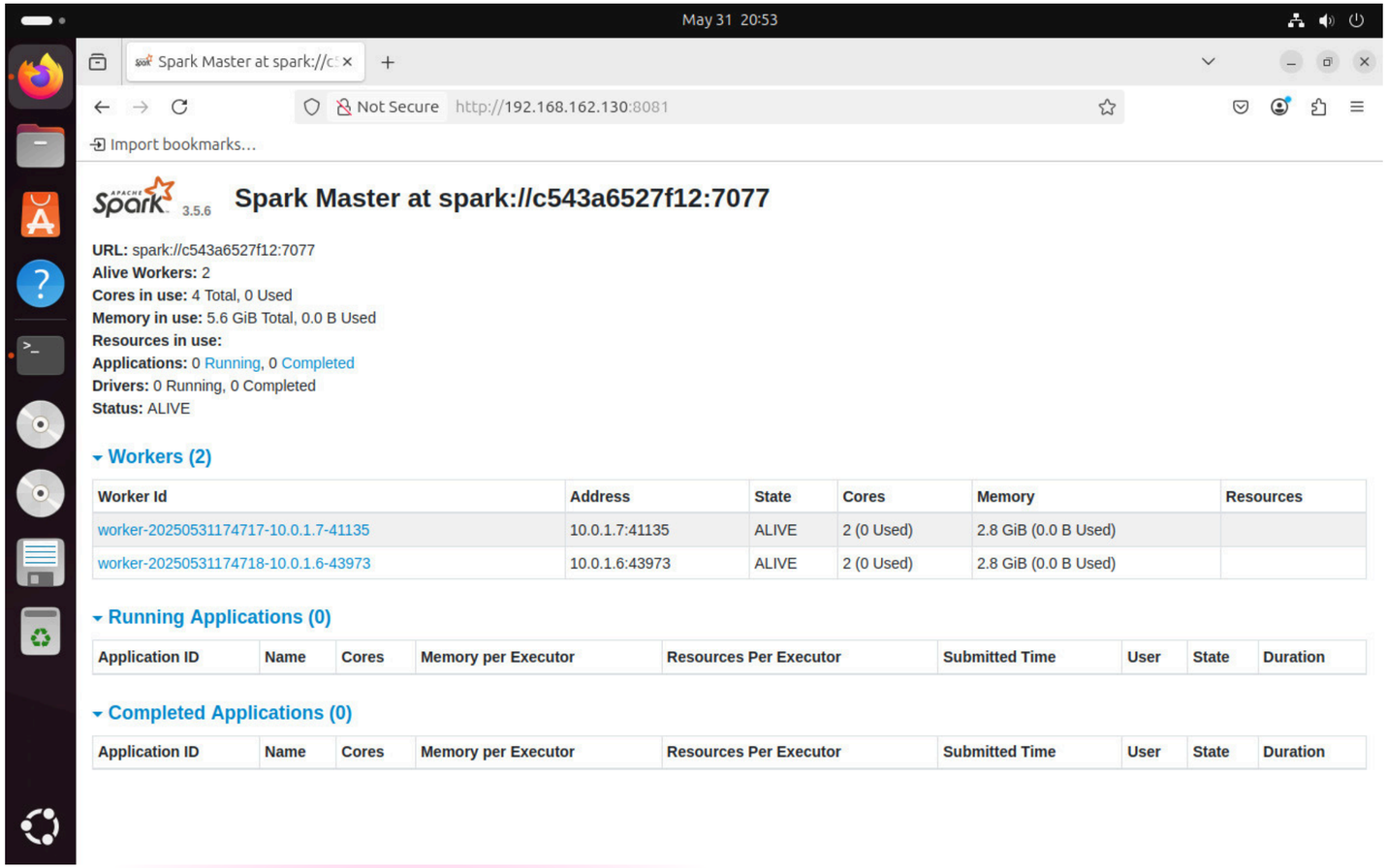
Swarm Initialization

Spark Stack Deployment

- Wrote spark-swarm.yml
- Deployed with docker stack deploy

METHODOLOGY & RESULTS

Task 2: Dockerized Spark Cluster



METHODOLOGY & RESULTS

Task 2: Dockerized Spark Cluster

ML Pipeline in PySpark

- Script: bio_classifier.py
- Read CSV → Feature engineering → Logistic regression
- Accuracy saved to result.txt **''Model Accuracy: 100.00%''**

ERRORS & CHALLENGES FACED

Networking Issues	Static IPs not persisting between reboots → fixed via netplan config
SSH Errors	Wrong permissions on .ssh folder → fixed with chmod 700 and 600
MPI Crashes	Hostfile misconfigured → corrected IPs and slot counts
Docker Permissions	Required newgrp docker to apply usermod

...

CONCLUSION

- We successfully deployed a hybrid computing architecture combining HPC and Spark.
- MPI was lightweight and faster to set up for smaller tasks.
- Spark offered better scalability and fault-tolerance.
- Both methods were effective at handling bioinformatics ML tasks.
- We gained real-world experience in systems setup, distributed ML, and debugging parallel computing environments.

THANK YOU