CS 496/522 Cloud Computing

Programming Assignment 1

Due Date: April 7, 2023

Make sure your answers are legible (submission in pdf); using Latex is encouraged. Succinct, but complete answers are encouraged. ***Please work alone, not in groups.***

---

1. We will give individual passwords to access the VMware cluster.
2. Log on to NMSU VPN to access the VMware ESXi servers. Open https://128.123.160.129 or https://128.123.160.133 based on what is assigned to you in the cluster configuration at the bottom of this document.
   If you have any issues accessing the cluster or permissions error with deploying a VM email gpanwar@nmsu.edu and sharad@nmsu.edu.
3. You deploy 4 nodes (1 light and 3 heavy) with specs:
   a. Light (Supervisor): 2 CPUs, 8 GB RAM, 120 GB HDD
   b. Heavy (Worker): 4 CPUs, 12 GB RAM, 120 GB HDD
   c. Name nodes as Lastname-nodeRole, for e.g., Panwar-Supervisor, Panwar-Worker1, Panwar-Worker2, and Panwar-Worker3.
4. Deploy Ubuntu-server (v.22) on all four nodes (ISO available in the Promise storage array).
5. Do the OS and system setup including any users you need to create and establish connectivity between the nodes.
   a. Setup the IP address configuration. Please refer to the cluster configuration at the bottom of this document to check what IPs you are assigned and stick to those IPs to avoid IP address conflicts with other students. All the devices will be in the 192.168.202.0/24 network.
   b. Setup firewall protection from all other networks outside of your 4-node cluster (This will be important, so you don't have issues with Hadoop nodes within the network belonging to other students). You are free to choose whatever firewall tool you are comfortable with and compatible with the OS, whether ufw, iptables, firewalld, etc.
   c. Make sure you keep notes on all commands you have used during setup and installation of packages so you can repeat it on the other nodes and in case you need to rebuild the system.
   d. Open Firewall access from 128.123.63.0/24, 128.123.64.0/24, and 10.253.0.0/16 over ssh to your master node.
   e. Create user named *admin* and give them sudo access. Add two ssh keys (one for Gaurav and another for Sharad) as authorized keys for the admin user:
      i. ssh-rsa
      AAAAB3NzaC1yc2EAAAADAQABAAABAQC1s0pOKp4QdVhZLXQU88g046kURJ9iB

EhhJu78Su6Emx0lGJxrSZnh74qsWZcnHUM2dEd8iAfjhAL/P2eO5EMBVOqn4x1g2r82ZIGpIMfE8lG42TXAJYb8xG7XzfVdceHThEuuEs+HAAxbLYKHXWWta+riAirTXf4JisDxU+D/N1Lyen4VyK9cuUZWQIEj97/OZkxXtVixtOeE/ToyEKjO4bplREEzBREMSsq6ZKhrP1oWHGUZ/0ntfs6C+8bYBCfH2RRLzzGYAK+J3UpLtl+UWLRp1ieM//7z32LQZ8Xyia7oSLxCXfY24ITNlJnubYHOSWGOGkLOM/AkfoWfDdSv gpanwar@nsol1

ii. ssh-rsa AAAAB3NzaC1yc2EAAAADAQABAAABgQC1s7PhiG8e7L4g6FNpWEub8VD2379IVglvFSmkpx5PFPR9lFT7W2fOPGXXcWvu0aZAKGwjcIQZlqGnMG9QHLsZ6LxtoJFSd+4/gm2DfigzuoHplA6e3oSC06j2B51Zj36vPMdpdDudc2FYBYfcQQeHZCxxoDDessw7SVnuLNhaMWIfWwxaZz5g5vmoAYoUTOqavtFNJAoxYKpkvaW7m+ZTiMk2hKBfAy03PCXvR9k3Xl22R+gZHlGKzZ9cFjOk2h/KIIoKwwXY7KCl0OySTGLptNpyvIsO+5UycdpLcnjFvJJNU84f5QVg0L39k2IYXzbbF6GqvnoTzG6x1f4bJcCUDsK0tNAuZX971C8aow/x+RubGh6IALXLZ6nUwrTnjiwogBnbsl/S6ytdk7ZpniT82Wd5qdmKqXdcZBOPHIujdkEgNrsbDLuXxEl72dMzqJpQgfbWuw8DI0TRaO4pKDlugvCilAYjw1FBL7TzDvWZXVBrICFjshpZvuh/BUTVk2U= theshrestha@theshrestha-desktop

f. Install and configure the Hadoop and Spark Frameworks.

6. Retrieve data available using scp from
    a. Host: 192.168.202.254
    b. Directory: /home/gpanwar/archive.zip
    c. Username: cc-user
    d. Password: dataneeded
    e. MD5sum for verification: 2d1c5b2988c94bd43793c6f0150a49d5 archive.zip
    f. Link to description of Dataset: https://www.kaggle.com/datasets/debashis74017/stock-market-data-nifty-50-stocks-1-min-data?select=1.+Data+description.txt

7. Write code to solve the following directives using both the Hadoop and the Spark frameworks.
    a. Find the maximum number of transactions in a day for all companies in a user supplied given time window.
    b. Maximum stock deviation is defined as (highest price for a day – lowest price for a day)/ lowest price for a day. Find the stock that had the highest maximum stock deviation in a day among all stocks and what was the corresponding value for a given time window.
    c. Find the maximum sell price in a day for a given company in the entire data set (No time window required).

NOTE: Your code should take in input parameters as mm/dd/yyyy for the start and end dates of the time window for 7.a and 7.b.

Make sure to cite your sources in your documentation wherever applicable. Do not copy paste code from other sources (no use of ChatGPT is allowed—code will be tested using ChatGPT).

Always log out of the VM in the VMWare console in the web browser. Please do not access anyone else's VMs within the same host.

Rubric:

1. Setting up VMs – 20 pts
2. Setting up Apache Hadoop and Spark frameworks – 20 pts
3. Performing the map-reduce and spark tasks – 30 pts
4. Clear and well-presented output – 10 pts
5. Documentation and README for running and grading the code. - 20 pts.

Cluster Configuration:

Public IP	Private IP	User	Device Role	ESXi Host
128.123.160.201	192.168.202.201	angelcam	supervisor	128.123.160.129
	192.168.202.202	angelcam	worker	
	192.168.202.203	angelcam	worker	
	192.168.202.204	angelcam	worker	
128.123.160.205	192.168.202.205	hchen	supervisor	128.123.160.129
	192.168.202.206	hchen	worker	
	192.168.202.207	hchen	worker	
	192.168.202.208	hchen	worker	
128.123.160.209	192.168.202.209	hunter30	supervisor	128.123.160.129
	192.168.202.210	hunter30	worker	
	192.168.202.211	hunter30	worker	
	192.168.202.212	hunter30	worker	
128.123.160.213	192.168.202.213	ftoaster	supervisor	128.123.160.129
	192.168.202.214	ftoaster	worker	
	192.168.202.215	ftoaster	worker	
	192.168.202.216	ftoaster	worker	
128.123.160.217	192.168.202.217	shanhead	supervisor	128.123.160.129
	192.168.202.218	shanhead	worker	
	192.168.202.219	shanhead	worker	
	192.168.202.220	shanhead	worker	
128.123.160.221	192.168.202.221	ccyr01	supervisor	128.123.160.133
	192.168.202.222	ccyr02	worker	
	192.168.202.223	ccyr03	worker	
	192.168.202.224	ccyr04	worker	
128.123.160.225	192.168.202.225	jiveyguy	supervisor	128.123.160.133
	192.168.202.226	jiveyguy	worker	

	192.168.202.227	jiveyguy	worker	
	192.168.202.228	jiveyguy	worker	
128.123.160.229	192.168.202.229	khanjari	supervisor	128.123.160.133
	192.168.202.230	khanjari	worker	
	192.168.202.231	khanjari	worker	
	192.168.202.232	khanjari	worker	
128.123.160.233	192.168.202.233	arajen97	supervisor	128.123.160.133
	192.168.202.234	arajen97	worker	
	192.168.202.235	arajen97	worker	
	192.168.202.236	arajen97	worker	
128.123.160.237	192.168.202.237	sosman	supervisor	128.123.160.133
	192.168.202.238	sosman	worker	
	192.168.202.239	sosman	worker	
	192.168.202.240	sosman	worker	