

NETW1005 S25 Project Description

Deadline is Tuesday 13th of May @11:59pm

Individual

- pg.1 - Overview and pre-requisites
- pg.2 & pg.3 - Tasks
- pg.4 - Restrictions and submission.

Overview

You are employed at a company which creates dashboard/reports on a daily basis. There are two main stages that occur each day to produce the dashboards:

- Ingestion.
- Filtering and reporting.

In the first stage, the company downloads csv files daily. While in the second stage, these csv files are filtered (cleaned for analyses) and analysed to produce the dashboards. On average, the filtering stage filters roughly 25% of the ingested data on a daily basis.

Problem Statement: Currently the company are facing issues regarding csv files storage whereas they are finding it difficult to scale their storage device since they have to manually increase the size every time they need more space. They are currently storing everything on a single partition. Extending this partition is often difficult and tiresome, having to unmount the partition and copy/move data to extend the partition.

Your job is to reorganize their storage system, making it more robust, extensible, automated and overall easier to manage and scale by performing the following:

1. Storage system design using logical volume managers.
2. Automation and auto-scaling.
3. SELinux employment.

Pre-requisites

- The system you are working on should have a 8 GB external storage disk (check the storage supplementary notes pdf posted on the cms to know how to add a disk on a virtual machine in Virtual Box)

Tasks

1. Storage system design

- (a) Separate ingestion and filtering stages, having them stored in different partitions or block devices and mount points. You don't have to copy or move data from the company's old storage system -start from empty mount points-
 - i. Create two directories under home directory, `~/ingestion` and `~/analyses`. You can safely assume that these directories are empty to start with -
- (b) Use the 8 GB disk to create an extensible storage system that is easy to manage and scale.
 - i. **IMPORTANT-** There are no specific requirements here except for one. The ingestion partition or block device has to be larger than analyses. That's it! How you decide to design the storage system, such as number of partitions, physical volumes, volume groups, overall size, etc. is ultimately YOUR decision and part of the objective of this project.

Again the only requirements For this Task are to separate the storage system for the two stages and have the ingestion partition overall size be larger than analyses and only use the 8 GB disk.

2. Automation and auto-scaling

- (a) You are required to create and schedule 3 scripts on a daily basis, first script is for ingestion, while the other two are for monitoring and auto scaling each block device/partition of ingestion and analyses.
 - i. **Script 1 -ingest.sh-** : This script simply checks the current date and subsequently downloads the respective csv file from a website and places it under `~/ingestion/data`. Make sure the directory is mounted and the storage system is set up before populating this directory. You can find the initial steps for the script in the section following the task descriptions below. What is required from you is to compress the csv file after downloading and not store the uncompressed version. The name of the compressed file should be the same in addition to its extension. Compression ratio is the most important aspect here and thus you choose the compression algorithm that would result in the smallest size.
 - ii. **Script 2 and 3 -lv1_monitor_extend.sh and lv1_monitor_extend.sh-** : Each script would monitor and scale the corresponding logical volume. The scripts should check if the overall consumption reached/exceeded 90%, then the logical volume should be extended by 10%. 10% of what? That is your decision.

- **For the sake of simplicity we will not address and schedule a script for filtering stage.**
- **Please note that the csv files are empty, use your knowledge of compression algorithms to choose the one that typically compresses the most.**
- **Your scripts must be error prone - that is, if there is no space left available, you should exit and MORE IMPORTANTLY if the script for monitoring `~/ingestion` is not able to extend the partition, YOU MUST AUTOMATICALLY CANCEL the ingestion script. In other words the ingestion script downloading should not run if the partition has reached 90% consumption and cant be extended. Again this must be done automatically (in the monitoring or ingestion script , or even a third separate script)**
- **Side note - A question you might be wondering is, why do I have to extend the logical volume when it reaches a certain limit rather than just giving it the full space of the block device from the beginning? Because you should take into consideration that the disk is not only used by ingestion and analyses and could also be possibly allocated to other directories/mount points.**

3. SELinux

- (a) Under /analyses directory there should be a directory called reports that will serve html reports over http server. You need to ensure that any file created under this directory has correct context type so that it could be served through http.

Ingestion script

```
#!/bin/bash
current_date=$(date +%d-%m-%Y)
# create dir under home dir if does not exist
mkdir -p ~/ingestion/data
wget -P ~/ingestion/data \
https://github.com/Badr-AL101/rh2-project-csvs/blob/main/$current_date.csv
```

Kindly note that the wget command takes the 2 arguments (directory and the URL) consecutively WITHOUT THE BACKSLASH IN THE MIDDLE. The backslash is used to indicate that the following content continues on the same line.

Restrictions

1. The ingestion script must be executed daily before noon (12 pm) to have the csv file ready for analyses.
2. The ingestion script should not run and be halted in case there is no additional space.
3. When exactly should these scripts run? Your decision. Just remember that the only restriction is that uploading script must be run before noon.
4. All directories and scripts should be named as stated.
5. The scripts should include comments that guide you through each step and clearly outline the flow of your ideas.

Submission

1. You need to submit your project as a report named after your info name_id_lab day_Project "Nawraz_Saeed_52-1234_Thursday_Project". At the start of the report ,please list your name, id and lab day.
2. Your report should include screenshot(s) of the following commands and their output :
 - (a) lsblk -fs
 - (b) cat /etc/fstab
 - (c) findmnt -verify
 - (d) pvdisk , vgdisplay , lvdisplay
 - (e) Screenshot of your cron table
 - (f) Screenshot of all your scripts
IMPORTANT -YOUR name and id must be included at the start of each script.
 - (g) semanage fcontext -l -C
 - (h) ls -Z /analyses/reports
 - (i) ls -al /ingestion/data
3. Feel free to include any additional screenshots/info you might feel relevant.
4. Please keep the report professional and organised.
5. Upload the pdf report to the following google form (only acceptable format is PDF).
Google form - <https://forms.gle/xrjKq6v2cR4LaqJQ9>