COMP0239 Coursework Challenge

Introduction

You have been employed by a company that wants to start analysing large datasets. It's your job to design and develop their new distributed computing architecture and implement a proof of concept by stress testing a data analysis task to test it out.

Coursework task

In this coursework you are required to build a distributed data analysis service across your cloud machines. You will also need to find an appropriately “sized” example analysis to benchmark and test your new data analysis service. You will have to implement this benchmark analysis as a pipeline. This analysis should implement/use an existing machine learning or statistical prediction method. You should not develop a new machine learning method. You must address the following features;

Part A: Example Analysis Features

1. Example analysis should be a pre-existing predictive method that can be expressed in the abstractions provided by you choose to install (i.e. a using MAP reduce in hadoop, using aa series of workflow steps, as a fully parallel MPI job, etc)
2. This can be a machine learning method or a more traditional predictive statistical method

Part B: Service Features

1. You should implement a distributed data analysis service across your 5 machines
2. Appropriate monitoring of the server hardware should be included
3. Appropriate monitoring of any running pipeline/data analysis should be included
4. Using your monitoring, you should appropriately benchmark and optimise the performance of your system
5. Once installed, tested and benchmarked, users of the service should be able to submit novel data to the service without having to write code themselves. That is, if there is an existing implemented pipeline (i.e the example analysis you have used as proof of concept) a user should be able to use this pipeline with new data without writing new pipeline code themselves.
6. Once complete, a user should have the ability to retrieve the outputs from their pipeline run.

Part C: Capacity Test

1. Your test dataset should be sufficiently large that it represents a meaningful test of the capacity of your distributed data analysis service
2. Such a capacity test should run your system at capacity for at least 24 hours.
3. You should seek to identify capacity improvements, such as change configuration options.

Resources

Machine ID	Number	Cores	RAM	HDD1	HDD2
Host	1	2	4GB	10GB	NA
Worker	4	4	32GB	50GB	(up to 200GB)
Total	5	18	108GB	210GB	800GB

Challenges/hints

1. Your distributed data analysis service can be implemented with any appropriate technology that can be scaled and can embody pipelines or parallelised data analysis task (celery, mpi, spark, hadoop, etc…). You are free to choose and implement what you feel is robust and appropriate for the class of analysis task you wish to address.
2. You should use the internet to find an appropriate, pre-existing predictive tool and appropriately input dataset that can be used to capacity test your system
3. You SHOULD NOT train and test your own predictive method, no marks will be awarded for doing this.
4. If the predictive pipeline is fast running and the input dataset have to be large enough to take at least 24 hours to process, or they can be long running and the input dataset can be smaller
5. The host instances are too small to run the calculations
6. Be mindful of load average and disk space on the machines while capacity testing them
7. You should only use these 5 machines for the coursework. There should not be a 6th CNC machine. Treat these 5 machines as a separate resource The Host machine in this set *is* the CNC machine

Deliverables

1. A short report (no more than 4,000 words) that explains the data analysis task you have chosen and explains how it has been implemented in the system you have installed. You should explain the capacity test that you have conducted and report statistics about it. Such as what is the maximum throughput of jobs that can be calculated per unit of time, how many concurrent jobs can be calculated, typically resource consumption of a given input, at what capacity does throughput start to reduce. You should discuss what changes could be made to improve capacity such as different configuration options. The report should include a link to a github repository
2. Your code repository must include instructions on how to use your code to install AND run your data analysis system. Instructions should assume that someone is starting with brand new host and n (n=5) machines that have nothing installed. The person marking your coursework should be able to check out your github to a fresh set of cloud machines, run your setup and analysis processes and produce the two required files on the host machine.

Tools and Datasets

There are many places online you might find complete machine learning tools that you could scale up. Scientific papers and assorted data repositories are a good place to start. You may wish to find an interesting dataset and then find a working analysis/prediction tool. Scientific papers often publish predictive methods and are accompanied by their datasets. You could choose an interesting method from another module you have taken this year. Here are some places you might also look

https://github.com/academic/awesome-datascience?tab=readme-ov-file#datasets
https://www.kaggle.com/
https://archive.ics.uci.edu/
https://scikit-learn.org/stable/
https://www.biorxiv.org/
https://arxiv.org/
https://github.com/facebookresearch/esm
https://github.com/agemagician/ProtTrans