CSC490 Assignment A2 - Data processing Pipelines and Infrastructure as code - 10% of Final Grade

Due: Feb 20, 10:59pm with a Github PR and on Quercus

The goal of this assignment is to start thinking about the datasets and infrastructure needed for your project.

1 Part One: Aspirational Datasets (10 marks)

Imagine you had a magic wand, what datasets would your team need to be successful in this project? Create an aspirational list of datasets with the data schemas you would like.

2 Part Two: Reality Check (15 marks)

Now review which datasets are actually available. Create a table of datasets that you could use for your project. Include public datasets, data you can scrape or data you can generate. For each dataset include a link to the source, a description of the data and commentary on why it is relevant to your project.

2.1 Dataset Sample Table

Relevant Item	Description	Commentary
Healthbench	Rubric based evaluation for Healthcare questions	HealthBench examples were created over the past year by a group of 262 physicians who have collectively practiced in 60 countries. These physicians are proficient in 49 languages and have training in 26 medical specialties.

3 Part Three: Data-processing pipelines(25 marks)

Design and implement a data processing pipeline for your project. This should include data ingestion, cleaning, transformation and data lake/warehouse design.

3.1 Key Requirements:

- Data schemas
- Pipeline diagrams with the technologies you are using (open source frameworks are helpful)
- When the pipelines will run and for which use cases
- Code for an initial version of this pipeline
- Include next steps for features you did not implement in the writeup

4 Part Four: Infrastructure as Code Implementation (30 marks)

Implement your pipeline infrastructure using code. You may choose any IaC framework, but your deployments should be straightforward and maintainable. For infrastructure components without existing modules:

- Create your own modules when possible (e.g., creating a Terraform module instead of using scripts)
- If using scripts, provide clear justification for this choice

4.1 Marking criteria:

- Code clarity and organization
- Infrastructure completeness
- Alignment with parts 1,2,3
- Implementation of multiple environments

5 Part Five: Disaster Recovery Demonstration (20 marks)

Disaster has struck! Record a screen capture of your team deleting your production environment and your IaaC to restore it. Marks will be given on the completeness of the deletion restoration and clarity of the process/code.

5.1 Demo should include:

- Data processing services or deployed applications
- Database systems and their data
- Configuration settings
- Access controls and security settings
- Verification of system functionality

6 Submission Instructions

Submit a pull request to the course Github repo with your assignment in a folder named a2 with your a2.pdf, include your team members names on the first page with student IDs. Also submit to quercus.