.

School of Computing, Engineering and Built Environment

Big Data Platforms

Level: M

Module Code: MMI227050

Assignment

Issue: 23rd March 2022

This coursework comprises 50% of the overall mark for the module.

Hand-in date: 2nd May 2022

An average student should be able to complete
this assignment in about 15 hours of work.

Attention is drawn to the university regulations on plagiarism. Whilst discussion of the coursework between individual students is encouraged, the actual work has to be undertaken individually. Collusion may result in a zero mark being recorded for the coursework for all concerned and may result in further action being taken.

DESIGN A CLOUD DATA PIPELINE TO PROCESS AN OPEN DATASET

There are many open datasets freely available relating to just about any area of interest or activity, including data about government, science, health, sport, music and so on. There are valuable insights to be gained from analysis of such data. Analyses can be performed using data pipelines which are built from platforms that ingest datasets and then transform and/or combine, store and query/visualise the data as required.

In this assignment you will consider the **conceptual design** of such a data pipeline, using platforms of the kinds that you have learned about in this module. The focus of the assignment is on the use of platforms capable of scaling to handle "big data", so **your design should be based on the use of distributed platforms which can be provisioned in the cloud**.

You will also be required to implement a **prototype** that implements and tests key data transformations and analytics within your design in a simplified form within a Databricks notebook.

Your assignment consists of two parts:

A. **Design report (30 marks)**

You should first research possible datasets and select the data that you want to use as the basis for your assignment. A list of resources to help you find suitable data will be made available on GCU Learn (see Reading & Links), but you may make use of suitable data from any source that you find. You may want to try to find data related to an area that you have a personal interest in and knowledge about.

IMPORTANT: before proceeding you MUST get approval from me for your choice of dataset. You must provide, through the relevant link in GCU Learn, the following information, and await approval:

- Name/nature of the dataset(s)
- URL(s) from which dataset(s) can be downloaded or of API documentation if it is a streaming source
- A brief description of the purpose for which you propose to use the data

If the dataset you choose is not considered to be suitable, or has been chosen already by another student, you may be asked to find an alternative. Each student should use a different dataset.

You should then proceed to devise and report on a high-level **design for a data pipeline** that could be used to perform your proposed analysis. Note the following:

- The pipeline should include stages as appropriate for: ingest, ETL, storage, analysis/visualisation.
- The pipeline should be designed for deployment on a single cloud service provider, and the platforms for each stage should be deployable or available as managed services on that provider's infrastructure.
- These may be services that are exclusive to your chosen cloud provider, or they may be third-party partner services that are available hosted by your chosen cloud provider. You will need to research the offerings that are available for your chosen provider.

This design should consider:

Overall concept (10 marks)

- The original format of the data (e.g. CSV, JSON) and illustration of the data schema.
- Any transformation to be applied to the data as ETL (Extract, Transform, Load)
- Potential analyses and/or visualisations to be performed. Given the focus of this module, I expect that analyses will be based on relatively simple filtering, projection and aggregation, rather than on ML (Machine Learning) algorithms, although there is no specific restriction on the analyses you can include.

Platforms (20 marks)

- The key components of the pipeline: for each component you should select a suitable cloud service (e.g. data store, file system, analytic engine)
- The overall design of your pipeline, illustrated with one or more diagrams
- The purpose of each component/service within your solution and the nature of the services, including any open-source platforms that the service is based on

You should base your choices on the module content and on additional research, and **you should justify your choices**. You should include appropriate references. Marks will be awarded on the basis of depth, completeness and relevance of the content within each of the above areas. Your report should be submitted in the form of a Word or PDF document.

Word limit for the report is 1500 ±10%

B. Prototype (20 marks)

You should implement a **prototype of your pipeline design using Apache Spark** within a Databricks notebook. The prototype will be a simplified working representation of your design concept. As a minimum it should demonstrate the following stages:

- Source data – storage and ingest
- ETL transformations - at least 2 distinct types of transformation
- Storage of data for analytics
- Query and visualisation – at least 2 distinct non-trivial queries with appropriate visualisation of results.

You should prepare your complete prototype in the form of a DataBricks notebook, and you should **make use of markdown cells to document your work**. The first markdown cell should contain a descriptive title for your prototype and your name and student number. It is suggested that you use Python as the programming language for your implementation.

Each processing stage of your pipeline should be represented by one or more executable notebook cells. Storage within your pipeline may be represented by file storage in the Databricks filesystem or by in-memory data structures. Your comments at each point should explain the purpose of the processing, where it fits into the overall data pipeline. It should be clear in your prototype where it is illustrating data being transferred from a storage platform to an analytic platform or vice versa.

Your prototype and documentation should be submitted in the form of a single notebook exported as PDF, including the output from executing the code in all the code cells. It should be possible for marking to view in the exported notebook the results of “running” the prototype.

Marks will be awarded for:
Relevance and appropriateness to solution described in Part A (5 marks)
Quality and completeness of implementation (10 marks)
Documentation within notebook (5 marks)

(End of assignment)