.

School of Computing, Engineering and Built Environment

Big Data Platforms

Level: M

Module Code: MMI227050

Coursework 2

Issue: April 2024
This coursework comprises 50% of the overall mark for the module.

Hand-in date: 3rd May 2024

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 AND PROTOTYPE A DATA PIPELINE TO PROCESS AN OPEN DATASET

Various open datasets are freely available, which span diverse areas including government, science, health, finances, sports, and music, among others. These datasets represent a rich source of insights waiting to be uncovered. To harness these insights, data pipelines can be employed. These pipelines are constructed using platforms that specialize in data ingestion, transformation, combination, storage, as well as querying and visualisation when needed. This approach enables in-depth analysis and interpretation of data, turning raw information into valuable, actionable knowledge.

In this assignment, you are tasked with designing and implementing a **prototype** data pipeline, complete with data transformations and analytics, within a **Databricks** notebook.

You are also required to consider scaling up the prototype using the types of platforms discussed in this module with a **conceptual design**. The focus of this conceptual design is on the utilization of platforms capable of scaling to handle "big data", so **your design should be based on the utilization of distributed platforms which can be deployed in the cloud**.

Therefore, this assignment consists of two parts.

Part A. Prototype within Databricks (60 marks)

You will be assigned to a dataset at random for this assignment.

You should design and implement ***a prototype data pipeline*** in ***Databricks***, aimed at storing, transferring, querying and visualising the assigned dataset. Your prototype must clearly demonstrate the following key stages of data processing:

- ***Source data management*** – This involves the initial storage and ingestion of source data, setting the foundation for subsequent transformations.
- ***ETL transformations*** – You are required to execute at least at least 2 distinct types of transformation, showcasing your ability to refine and optimize data for better analysis and insights.
- ***Storage of data for analytics*** - Post-transformation, the data should be stored in a format and structure that facilitates easy and efficient analysis.
- ***Query and visualisation*** –Demonstrate your ability to derive insights from the data through at least 2 distinct, non-trivial queries that utilize filtering, projection, and/or aggregation. Each query should be accompanied by an appropriate visualisation of results.

You are required to develop your complete prototype within a **Databricks notebook**, please ensure to **utilize markdown cells to document your process comprehensively**. The first markdown cell in your notebook should contain a descriptive title for your prototype, your name and your student number. It is recommended to use Python as the programming language for your implementation.

In your prototype, each processing stage should be clearly represented by one or more executable cells within the notebook. Data storage in your pipeline can be represented by file storage in the Databricks filesystem or by in-memory data structures. In the markdown cells, your comments at each point should explain the purpose of the processing, and 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 completed prototype and documentation (via markdown cells) should be ***submitted as a single PDF file***, generated from exporting your notebook. The PDF should include the outputs from executing the code in all the code cells. It should be possible for marking to view the results of "running" the prototype in the exported notebook.

The PDF file should be clearly labelled with "BDP CW2_A 2023-24" and contain the submitting student's name. e.g., *Harry Potter*'s report file should be named as:

"BDP CW2_A 2023-24_ H Potter"

Marks will be awarded for:

1. Quality and completeness of implementation (40 marks)
    a. Source data management
    b. ETL transformations
    c. Storage of data for analytics
    d. Query and visualisation

2. Documentation within notebook (20 marks)

Part B. Cloud-based Data Pipeline Design Report (40 marks)

[*Word limit for the report is 1200 ±10%*]

Based on your Databricks data pipeline implementation in part A, it's now time to upscale your design to a high-level **design for a cloud-based data pipeline**. This upscaled design should refine the purpose of using the assigned dataset, and the cloud-based data pipeline could be used to perform your proposed analysis. Consider the following points while designing the upscaled pipeline:

- Integrating appropriate stages for the cloud-based pipeline, such as ingestion, ETL, storage, and analysis/visualization, depending on your specific requirements.
- Designing the pipeline to be deployable on a single cloud service provider (e.g., AWS, GCP, Azure, etc.), utilizing either proprietary services or third-party solutions hosted by the cloud service provider.
- Conducting thorough research on the available services and tools provided by your chosen provider. These may be services that are exclusive to the cloud service provider, or they may be third-party partner services that are available and hosted by your chosen provider.

This cloud-based data pipeline design should consider:

1. Overall conceptual design (15 marks)

In advancing from your prototype in Part A, your conceptual design for a cloud-based data pipeline should **NOT** be confined to a single Databricks notebook and may incorporate additional datasets.

The conceptual design should follow this outline:

- Begin by describing the original format of the data (e.g. CSV, JSON), and provide a clear illustration of the data schema.
- Specify any necessary transformations to be applied to the data during ETL (Extract, Transform, Load), and justify each transformation. Although leveraging transformations from the prototype is acceptable, you should not be limited to those.
- Outline potential analyses and/or visualisations to be performed. Again, while you may use analyses and/or visualisations from the prototype, you are not limited to those. Given the focus of this module, it is expected that analyses will be based on relatively simple filtering, projection and aggregation, rather than on more complex ML (Machine Learning) algorithms, although there is no specific restriction on the analyses you can include.

2. Platforms (25 marks)

- For each key component in your upscaled pipeline, select a suitable cloud service (e.g. for data storage, file management, and analytics).
- Create diagrams to visualise the pipeline's architecture, showcasing the interconnectedness of components/services.
- Clearly explain the purpose of each component/service within your solution and the nature of the services, including any open-source platforms on which the service is based. This comprehensive approach ensures clarity in how each component contributes to the pipeline's functionality and the rationale behind service selection.

You should base your choices on the module content and additional research. **You should provide justification for your choices**. References should be included where appropriate. The report should be submitted in the form of a Word or PDF document, which should be clearly labelled with “BDP CW2_B 2023-24” and contain the submitting student’s name. e.g., *Harry Potter*’s report file should be named as:

“BDP CW2_B 2023-24_ H Potter”

Marks will be awarded on the basis of depth, completeness, and relevance of the content in each of the above areas.

(End of assignment)