# DESIGN AND PROTOTYPE A DATA PIPELINE TO PROCESS AN OPEN DATASET

A wide range of open datasets are available for free, covering various areas of interest and activity, such as government, science, health, sports, 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 design and implement a **prototype** data pipeline, complete with data transformations and analytics, within a **Databricks** notebook.

You will also be required to consider scaling the prototype using platforms of the kinds that you have learned about in this module with a **conceptual design**. The focus of this conceptual design 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**.

Your assignment consists of two parts.

### Part A. Prototype within Databricks (30 marks)

You will be assigned to a dataset randomly.

You should design and implement a prototype data pipeline to store, transfer, query and visualise the given dataset within Databricks. At minimum, the prototype should demonstrate the following processing 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 filtering, projection, and/or aggregation employed) 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 in your notebook should contain a descriptive title for your prototype, your name and your 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, 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 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:

1. Quality and completeness of implementation (20 marks)
   1. Source data (5 marks)
   2. ETL transformation (5 marks)
   3. Storage of data for analytics (5 marks)
   4. Query and visualization (5 marks)
2. Documentation within notebook (10 marks)

### Part B. Cloud data pipeline Design Report (20 marks)

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

Based on your Databricks data pipeline implementation, it's now time to upscale your design to a high-level **design for a cloud data pipeline**. This will enable you to refine the purpose for which you use the assigned data, and the cloud data pipeline could be used to perform your proposed analysis. Consider the following points while designing the upscaled pipeline:

* Include appropriate stages for the pipeline, such as ingest, ETL, storage, and analysis/visualization, depending on your specific requirements.
* Design the pipeline to be deployable on a single cloud service provider. 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 and hosted by your chosen cloud provider. You will need to research the offerings that are available for your chosen provider.

This design should consider:

1. Overall concept design (8 marks)

This concept design that you develop should be based on the prototype you implemented in part A, and it must extend beyond the prototype. It should **NOT** rely on a single DataBricks notebook.

The concept 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). While you may use transformations from the prototype, you are not limited to those, and all proposed transformations should be justified clearly.
  + 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.

1. Platforms (12 marks)
   * Choose a suitable cloud service for each key component of the upscaled pipeline (e.g. data store, file system, analytic engine).
   * Illustrate the overall design of the pipeline with one or more diagrams.
   * Clearly explain the purpose of each component/service in your solution and the nature of the services, including any open-source platforms on which the service is based.

You should base your choices on the module content and additional research, and **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. Marks will be awarded on the basis of depth, completeness, and relevance of the content in each of the above areas.

*(End of assignment)*