**Aim**

This coursework provides assessed practical experience in the use of technology to perform fundamental ETL processes within organisations using Python; alongside considerations of

Data and its management for modern organisations wishing to expand their operations and/or data volume. The assignment is split into two deliverables: Python ETL, A programming task for combining multiple data sources and pushing these to a central organisational data store; and the Scaling Up theoretical report, A report which poses a theoretical expansion scenario for a business. Please read the below sub-sections carefully, as these will detail the requirements for each component.

1. Task 1 - Python ETL: Programming Task

A SMB (small-to-medium business) has recently begun to utilise the data they obtain from their customers. Unfortunately, their business has multiple areas which all have customer data specific to that area, and this is fragmented within the organisation. E.g Credit Card data is only stored by the financial systems, employment within HR, etc. There is not a single cohesive record representing customers. The SMB is looking to unify these ahead of further data investigation, and to pool all this data together into a central datastore.

The data provided for this assessment is mock data representing a typical customer-facing business; these involve data such as names, banking credentials, family attributes, etc.

These data files are provided as a mixed modality in a variety of formats (CSV, JSON and XML). The work herein requires the processing of these data into a homogenous record, aligning the same customers from different sources together, which are then automatically entered into a Relational Database System using modern tools & libraries.

2. Raw data for this assignment is provided on canvas, containing a mix of .csv, .json, and .xml files. This data is synthetic, but derived from a realistic domain with data generated in accordance to 2016 UK Census data. You will need to apply your knowledge of data in order to correctly parse these and perform the task.

You are expected to read and extract data from these various formats, wrangle the data - solving inconsistencies if present - and bring data together into a singular format (See Figure1 as an example). These unified records are then to be mapped to a relational database using PonyORM Entities, with all unified records being entered into the database.