Module Title:	Programming for Data Analysis, Processing & Visualisation
Module Code:	B9DA100
Module Leader:	Clive Gargan / Darren Redmond / Abhishek Kaushik
Level:	9
Assessment Title	Python ETL Solution Design & Implementation
Assessment Number:	2 of 2
Restrictions on Time/Length	N/A
Individual/Group:	Individual
Assessment Weighting:	30%
Issue Date:	Week Beginning 11th January
Due Date:	29th January (6pm)
Feedback Date:	Hand in date plus 2 weeks
Mode of Submission	Moodle only

Learning Outcomes:

1. Critique data structures used to store and manage large datasets.
2. Write programs to process and manipulate large datasets in diverse platforms.
3. Apply practical skills in functional programming techniques to write code for a given dataset.
4. Appraise the extended ecosystem of tools and techniques related to Big Data.

Assessment Overview

The assignment focuses on the technical design and implementation of an ETL process using Python, Pandas and SQL Server.

Assessment Task

The file *FireBrigadeAndAmbulanceCallOuts.csv* is an extract generated from data.gov.ie (Ireland's Open Data Portal). This file is an activity log for Fire and Ambulance Annual Incidents between 2013 and 2019.

You are required to implement an ETL process to extract, transform and load the data.

(Note: you must use the file published to Moodle – do NOT regenerate the extract from data.gov.ie)

Data Dictionary

- Fields include date, area of incident (district ID) and response time data.
- The fields from MOB through to CD are generated by the vehicle (either by a button press or a voice message) and they reflect its changing status.
- TOC is the time the call is received in the control centre
- ORD is the time the vehicle is ordered, i.e., mobilised to the incident by a control operator.
- MOB is the time at which the vehicle is mobile to incident (the vehicle has started to move)
- IA is the time the vehicle is in attendance (the vehicle is stopped at the incident)
- LS leaving scene (the time the ambulance is leaving scene for hospital)
- AH the time at hospital (ambulance has arrived at hospital)
- MAV the time at which the vehicle is mobile and available (vehicle heading back to station)
- CD the time at which the vehicle is closing down (back at station, vehicle radio is being shut down)

ETL (Extract, Transform, Load) Process

Note: You must follow the steps below in the order they are presented.

You are required to implement an ETL process in Python to achieve the following:

- Load the CSV file (using Python).
- Output the total number of rows and columns.
- Output the number of non-null rows (by column).
- Output the number of null values (by column).
- Output the number of null values for all columns.
- Output the total number of call outs by Station Area.
- Output the total number of call outs by Date and Station Area.
- Output the total number of call outs by Station Area and Date where the description is either Fire Car or Fire Alarm.
- Replace any instance of "," (in any column) with an empty string.
- Replace any instance of "-" (in any column) with an empty string.
- Drop rows for the columns (AH, MAV, CD) where at least one row value is NULL.
- Drop any duplicate rows (except for the first occurrence).
- Output the minimum time difference between TOC and ORD.
- Using the resulting data set, post implementing the previous cleansing steps, load the data into a table in SQL Server. (Note: you must create the physical table in SQL Server to complete this task. Use the same column names as the columns in the CSV File.)

Technical Document

The accompanying Technical Document should cover (but not limited to):

1. Scope of the document.
2. Technical Design to include:
    a. ETL Architecture
    b. Pandas operations detailed for each requirement.
    c. Data Model
3. Testing.
4. Reflection on Learnings.
5. References.

Project Deliverables

The distribution of assessment marks will be as follows:

Deliverable	Breakdown of Marks	Submission Date
Python ETL	80%	29th January (6pm)
Technical Document (Should not be less than 300 words)	20%	29th January (9pm)

ASSESSMENT CRITERIA Criteria/Mark	< 40	40 - 49	50 - 59	60 – 69	70+
Python ETL (80%)					
Extract and Transformation Process: 60%	None or Incomplete Extraction Process.	Partially complete extraction process but no transformation.	Partially complete extraction process and partially complete transformation.	Complete extraction process and partially complete transformation.	Complete extraction process and complete transformation.
Loading Process 20%	None or incomplete loading process.	Partially complete loading process with poor exception handling.	Partially complete loading process with complete exception handling.	Complete loading process with complete exception handling but inconsistent naming conventions.	Complete loading process with complete exception handling and standardised naming conventions.
Technical Document (20%)					
Key Areas include: Overall Presentation, Description & Functionality and detailed Design.	Very Poor documentation. Description &functionality weak. Missing key parts (e.g. database schema). Poorly structured with spelling and syntax errors.	Poor documentation. Description and functionalities stated but lack clarity. Some key areas missing.	Adequate documentation with adequately stated details. Key areas are of reasonable standard.	Good documentation, all essential key areas covered. Description & functionalities clear.	Excellent documentation, Comprehensive design.

IMPORTANT NOTES:

1. **You must NOT use or copy parts of a solution from another student.**
2. **Extensions to assignment submission deadlines will be granted in exceptional circumstances only. The appropriate "Application for Extension" form must be used and supporting documentation (e.g. medical certificate) must be attached. Applications for extensions should be made directly to the Head of Year or Programme Leader in advance of the deadline date.**
3. **Late Submission Penalties: Immediately after the submission deadline for an item of continuous assessment, a penalty will be applied per day or part thereof. For the purposes of these penalties, a day is defined as any day of the week, including weekends and public holidays when the College may be closed. The minimum possible mark for late submission is 0%. The number of marks deducted depends on the lateness of the submission and will be deducted according to the following scale:**

   Where an assessment is submitted between 1 and 14 days late 2 marks per day are deducted

   An assessment submitted after the deadline but within 24 hours of the original deadline will attract the first day penalty, i.e. deduction of 2 marks

   Where an assessment is more than 14 days late it is annotated at the discretion of the lecturer but no marks can be awarded.