Configuration Manual

MSc Research Project
MSc in Data Analytics (MSCDAD_C)

Kranthi Kumar Yedla
Student ID: x23288892

School of Computing
National College of Ireland

Supervisor: Dr. David Hamill

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name: Kranthi Kumar Yedla

Student ID: X23288892

Programme: MSc in Data Analytics **Year:** 2025

Module: Research Practicum Part 2

Lecturer: Dr. David Hamill

Submission Due Date: 1st Sept 2025

Project Title: Integrating Autoencoders and GPU Acceleration for Proactive Anomaly Detection in ETL workflows

Word Count: 861 **Page Count:** 3

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature: Kranthi Kumar Yedla

Date: 1st Sept 2025

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

1. **System Requirements**

The experiments were performed on a 16 GB RAM and NVIDIA GeForce GTX 1650 Ti (4 GB VRAM) system running Windows 11. CUDA 12.1 and cuDNN have been installed to provide support to a GPU acceleration. To test on smaller-scale machines, the pipeline can be re-implemented on a CPU-only machine, although GPU is highly encouraged when the dataset size is up to 1 million records. Datasets, checkpoints, and log files will take at least 20 GB in free disk space.

2. Required Software and Tools

The project is implemented in Python 3.10 environment and the major packages are:

- PyTorch 2.2.2 (with CUDA support)
- Pandas 2.2.2, NumPy 1.26.4
- scikit-learn 1.5.1 (to preprocess utilities)
- Dagster 1.6+ (for pipeline orchestration)
- Matplotlib 3.9.2, Plotly 6.2.0 (for result visualization)
- boto3 1.34+ (for AWS S3 integration)

To recreate the environment: there is a requirements.txt file provided.
pip install -r requirements.txt3. Dataset Preparation
This project developed three artificial datasets to replicate the ETL conditions in the real world:

- Banking: 602,883 synthetic transaction records.
- Insurance: 999,830 synthetic policy and claim records.
- Sales: 200k synthetic records of retail transactions (as a mid-size benchmark).

All the datasets were produced in controlled forms and separated into two sections:

- Source (clean): skeleton data (synthetic) stored in source databases.
- Target (corrupted): aberrant injected (corrupted data, duplicates, missing rows, extreme values, and schema corruption) to resemble actual real ETL data quality problems.

The storage was in PostgreSQL (NeonDB) with a uniform aligning of the schema between the source and the target tables. This design enabled the anomaly detection system to be tested under controlled conditions and known ground-truth anomalies were used to assess the system.

3. Model Setup

A neural network with an autoencoder was written in PyTorch:

- Encoder: dense layers that reduce the input to the latent vector h.
- Decoder: Dense input reconstruction layers are symmetrical.
- Loss Function: Mean Squared Error (MSE).
- Optimizer: Adam, learning rate 0.001.
- Epochs: 100 (early convergence ~80 epochs).

Abnormalities were identified by computing reconstruction error per record and using dataset-specific threshold (e.g. 0.02867 in Banking, 0.000434 in Insurance).

4.Pipeline Orchestration (Dagster)

The ETL pipeline was orchestrated with Dagster:
1. Load clean data from source DB.
2. Credential anomalies -> populate target DB.
3. Rule-based checks and autoencoder anomaly detection (run validation).
4. Save logs (nulls.csv, duplicates.csv, missing.csv, autoencoder_anomalies.csv).
5. Upload logs to AWS S3 and email results to stakeholders.

5.Project Structure

- **jobs.py** → Dagster orchestration of ETL + validation.
- **[dataset]_anomaly_detection_to_s3.py** → validation + anomaly detection scripts.
- **validation_logs/** → CSV logs for each run.
- **metrics.txt / autoencoder_metrics.json** → performance logs.

6. Logs and Evaluation

Each run outputs:

- Rule logs: the number of nulls, duplicates, missing records.
- Autoencoder measurements: mean, standard deviation, minimum, maximum reconstruction error, anomalies, loss of training.
- Runtime: GPU validation times.

As an illustration, the outcome of Banking dataset: 6,029 anomalies identified (approximately 1 percent), 14.5s validation (GPU).

7. Troubleshooting:

• CUDA not found: NVIDIA drivers and CUDA toolkit must be installed. torch.cuda.is provided (CUDA is only supported with NVIDIA GPU).
• Database connection error: Verify connection string and SSL mode for NeonDB.
• S3 upload failure: Check AWS credentials in boto3 config.
• None of the anomalies identified: Adjust reconstruction threshold or validate anomaly injection step.

8. References

- PyTorch Foundation (2023). *PyTorch 2.1 Release Notes*. Available at: https://pytorch.org/blog/pytorch-2-1/
- Dagster Labs (2024). *Dagster Documentation*. Available at: https://docs.dagster.io/
- NVIDIA (2024). *CUDA Toolkit Documentation*. Available at: https://docs.nvidia.com/cuda/
- Chandola, V., Banerjee, A. & Kumar, V. (2009). *Anomaly Detection: A Survey*. ACM Computing Surveys, 41(3), 1–58.