Data Engineer Tech Test

**Introduction**

This package contains three folders designed to emulate an industry-standard Data Lake architecture. It includes a landing area for raw data, a staging area for cleansed data, and a transformed area for business-specific data manipulations.

The data you will be working with is in the landing folder. It comprises synthetic data that simulates customers of a bank, their accounts, and the transactions conducted from those accounts.

**Case Study**

As a Data Engineer at ABC Bank Ltd, your task is to clean and transform the data according to business requirements. *The deliverable must be a Python module* suitable for production environments. This module should encompass a comprehensive *pipeline capable of handling end-to-end processing*, designed for deployment by the DevOps team.

**Task 1**

The data has been successfully ingested into the landing area using Azure Data Factory. However, the data currently contains whitespace and newline characters that need to be cleaned. *Your task* is to cleanse the data and store the cleaned output in the staging area.

To ensure *data integrity and mitigate any potential data loss*, it is crucial to verify and ensure that the row and column counts between the landing and staging areas remain consistent throughout the cleansing process. This consistency check will help maintain the accuracy and completeness of the data as it moves from the landing area to the staging area.

**Task 2:**

A key stakeholder has requested specific data (as listed below) to be prepared to facilitate their analysis and predictions. They require transformations on the staging data that will empower analysts in their tasks.

The stakeholder intends to consume this data using Azure Synapse Analytics and emphasizes the use of *appropriate data types within Parquet files* to optimize performance. This ensures that the data is structured and formatted in a way that supports efficient querying and analysis within Synapse.

1. Aggregate transactions and transaction amounts categorized by account type.
2. Identification of customers with transactions exceeding £10,000 to detect high-risk or anomalous activity.
3. Calculation of cumulative account balances over time.
4. Calculation of monthly transaction trends.
5. Calculation of average transaction values per customer.
6. Identification of customer churn.

**Task 3**

Due to the expected increase in data volume to millions of records, it is imperative to evaluate the current efficiency of the pipeline. In response to stakeholder requirements, a *brief recommendation* is provided below for a phase 2 optimisation initiative aimed at enhancing pipeline performance.