ASSIGNMENT #1

Answers to Part 1 and 2
Submitted by Mahnoor Imran (25280082)

Part 1

(a) Data Heterogeneity

In this ELT pipeline, three different types of data sources were integrated within the Financial Technology domain I chose, specifically focusing on digital payments. Each source represents a different form of data structure, demonstrating real data heterogeneity.

The News API provides semi-structured data in JSON format. The response contains nested fields such as article metadata, source information, publication date and textual content. For example, each article includes attributes like title, author, publishedAt and a nested source object containing the publisher name. This structure preserves hierarchical relationships, making it semi-structured rather than purely tabular. The Kaggle PaySim dataset represents structured data in CSV format. It follows a fixed schema with clearly defined columns such as transaction type, transaction amount, account balances and fraud indicators. Each row represents a single financial transaction. This dataset is fully tabular and suitable for relational storage and statistical analysis. The Yahoo Finance dataset, retrieved using the yfinance library, provides structured time series data. It includes date indexed financial variables such as Open, Close, High, Low prices and trading volume for companies like PayPal, Visa and Mastercard. This dataset introduces a temporal dimension, where each observation is associated with a specific date.

Together, these sources demonstrate the pipeline's ability to handle semi-structured JSON data, structured transactional records and structured time series financial data, reflecting the diversity typically encountered in real data engineering systems.

(b) Extraction Challenges

Several technical and practical challenges were encountered during the extraction phase of the pipeline. One of the primary challenges was authentication management. Both the News API and Kaggle require secure API keys for access. These credentials had to be stored using environment variables rather than being hardcoded in the script, in order to follow best practices and meet assignment requirements.

Another challenge was handling API rate limits. The News API restricts the number of requests that can be made within a given time period. This means that pagination and request frequency must be carefully controlled to avoid request failures. In actual production systems, this would require retry logic and monitoring mechanisms.

Data format inconsistencies also presented challenges. The News API returned nested JSON data that required flattening before being converted into a DataFrame. In contrast, the Kaggle dataset was a flat CSV file with millions of rows, requiring careful handling to avoid memory issues in Colab. Additionally, the yfinance dataset sometimes returns multilevel column indexing when multiple tickers are downloaded simultaneously, which required restructuring before saving. These challenges reflect common EL pipeline complexities, including authentication management, rate limiting, data normalization and format harmonization across heterogeneous sources.

(c) Storage Justification

Storing the extracted data in multiple formats, specifically CSV and JSON, provides flexibility within a data engineering context. Each format serves a different purpose depending on downstream usage. CSV format is highly suitable for structured data analysis. It is humanreadable, lightweight and easily compatible with tools such as Pandas, Excel and business intelligence platforms. For the Kaggle transactional dataset and the Yahoo Finance timeseries data, CSV allows efficient loading and statistical analysis. JSON format, on the other hand, is better suited for preserving semi-structured or hierarchical data. It maintains nested relationships and metadata that may be lost in flat tabular storage. For example, the News API articles include nested source information that can be preserved more naturally in JSON format.

In practice, CSV would be chosen when the goal is numerical analysis or integration with relational databases. JSON would be preferred when working with APIs, NoSQL databases that require flexible schema design. By storing both formats, the pipeline ensures adaptability for different analytical and AI/ML workflows.

Part 2

(a) Cleaning Rationale

During the cleaning phase, duplicate records were removed to prevent inflated counts and biased analysis results. Missing values in the Kaggle dataset were minimal so they were replaced with zero where appropriate to maintain dataset integrity without significantly distorting distributions. For the Yahoo Finance dataset, missing values were handled using forward filling, as financial time series data often contains minor gaps due to non-trading days. Converting transaction types into categorical format improved memory efficiency and ensured appropriate handling during analysis. Date fields were standardized using datetime conversion to allow accurate time based visualization and sorting. These decisions were made to preserve data reliability while maintaining analytical usability.

(b) Visualization Insights

The visualizations reveal several meaningful patterns within the digital payments domain. The transaction type distribution shows that CASH_OUT and PAYMENT transactions dominate the dataset, indicating that withdrawal and payment activities are the most common behaviors in digital financial systems. The fraud analysis plot highlights that fraudulent transactions tend to cluster around specific transaction amounts, suggesting that fraud is not random but associated with particular financial patterns. The correlation heatmap further shows strong relationships between old and new account balances, which is expected in transaction based systems where balances are mathematically linked. Additionally, the stock price time series trend demonstrates a general upward movement for major digital payment companies, reflecting the growing adoption and expansion of fintech services globally. Together, these patterns illustrate how transaction behavior, fraud dynamics and market growth are interconnected within the digital payments ecosystem.

(c) Visualization Critique

The engineered datasets created through this ELT pipeline can support several high impact AI and machine learning applications within the digital payments ecosystem. The transactional dataset, particularly with fraud labels, can be used to train supervised learning models for fraud detection, where algorithms such as logistic regression, random forests or gradient boosting can identify suspicious transaction patterns in real time. The structured balance features and transaction amounts also make the data suitable for anomaly detection systems that flag unusual behavior even when explicit fraud labels are unavailable.

The time series stock market data enables forecasting applications, where models like ARIMA or other sequence-based neural networks can predict future price movements of digital payment companies. Such forecasting can support investment analysis, market risk assessment and fintech growth evaluation. Additionally, the news dataset introduces an opportunity for natural language processing tasks, such as sentiment analysis and topic modeling, which can be used to measure public perception of fintech companies and potentially link sentiment trends to stock performance.

By integrating structured financial transactions, semi-structured news data, and time series market data, the pipeline creates a foundation for multi-modal AI systems. These systems could combine behavioral, textual and financial signals to build better fraud detection models, credit risk scoring systems and strategic forecasting tools. Overall, the engineered datasets are not only suitable for exploratory analysis but are scalable for production level AI deployment in real-world financial technology environments.