**Practical AI CA1:**

**SP-Buy Platform**

**Name:** Shaun Kwo, Loh Yip Khai, Koh Zi Hao

**Class:** DAAA/FT/2B/03

**Group No.**: 1

**Project Overview**

The task is to develop a system to identify fraud activities on the SP-Buy platform, starting with the analysis of historical data and building an interactive dashboard for the business users. This project will be handled in two phases, and we are focusing on **Phase 1 (CA1)**.

**Phase 1 (CA1) Objective:**

The goal is to build a SQL database to ingest historical data, create a pipeline to analyze the data in Python, and deliver insights via an interactive dashboard for business users.

**User Stories (Including Stakeholder Impacts):**

1. **User Story 1 (Impact on Sellers)**  
   **As a** vendor on SP-Buy, **I need** a system that flags fraudulent orders, **so that** I can minimize revenue loss from fake refund requests and protect my business.

**Confirmation**

* Display flagged orders on a dashboard for vendor review.
* Show reasons for flagged status to help vendors assess order validity.
* Provide metrics on the effectiveness of fraud flagging for vendor reporting.

1. **User Story 2 (Impact on Customers)**  
   **As a** customer of SP-Buy, **I need** a fair and transparent fraud detection system, **so that** my legitimate transactions are not mistakenly flagged, and my shopping experience is not disrupted.

**Confirmation**

* Display transaction status (flagged/not flagged) and reasons if flagged.
* Show fraud detection criteria to assure transparency and build trust.

1. **User Story 3 (Impact on SP-Buy Platform)**  
   **As a** business executive at SP-Buy, **I need** to track fraud detection metrics via a real-time dashboard, **so that** I can make informed decisions to reduce fraud and improve the platform's reputation.

**Confirmation**

* Display real-time fraud detection metrics (e.g., number of flagged transactions, false positives, and fraud rates) on the dashboard.
* Allow filtering by time period, transaction type, and other relevant parameters.
* Provide trend analysis and data visualization for fraud metrics over time.

1. **User Story 4 (Fraud Analysis System)**  
   **As a** data analyst at SP-Buy, **I need** a system that analyzes historical transaction data to detect fraud patterns, **so that** I can provide actionable insights for reducing fraudulent activities on the platform.

**Confirmation**

* Display visualizations of fraud trends and correlations in the data.
* Highlight anomalies and emerging patterns to inform proactive fraud prevention strategies.

**Job Roles for Phase 1**

1. **Zi Hao (SQL Database Setup)**  
   **As a** SQL Database Engineer, **I need** to design and implement an SQL database to store transaction and user data, **so that** the data can be structured for efficient querying and fraud detection.
2. **Yip Khai (ETL Pipeline Creation)**  
   **As an** ETL Developer, **I need** to build a robust ETL pipeline between the SQL database and Python, **so that** data can be seamlessly extracted, transformed, and loaded for analysis.
3. **Shaun Kwo (Data Cleaning and Preprocessing)**  
   **As a** Data Scientist, **I need** to clean the data, **so that** I can ensure the quality and accuracy of the data used for fraud detection and analysis. **I also need** to preprocess the data so that there is unique primary (composite) key with foreign keys **so that** the data can be inserted into the database smoothly.
4. **TEAM MEMBER NAME(Exploratory Data Analysis)**  
   **As a** Data Analyst, **I need** to perform **EDA** on the transaction data, **so that** I can uncover patterns and trends in fraudulent activities and provide insights that can be visualized for further decision-making.
5. **Group (Dashboard Creation)**  
   **As** Data Visualization Specialists, we **can** create interactive dashboards to present fraud activity trends and KPIs, **so that** stakeholders can easily interpret the results and make decisions based on clear visual insights.

**Product Backlog**

The **Product Backlog** is a prioritized list of tasks the team must complete to implement the user stories. Here are the Product Backlog Items (PBIs):

|  |  |  |
| --- | --- | --- |
|  | Item | Effort Required (Man Hours) |
| 1 | **PBI\_01: Data Cleaning and Preprocessing**   * .strip() initial data set to remove additional spacing and \n. * Handle missing data. * Remove duplicated rows. * Check all **PRIMARY KEYs** are not duplicated. * Ensure all **FOREIGN KEYs** have a corresponding **PRIMARY KEY**. | 40 |
| 2 | **PBI\_02: Setup SQL Database**   * Define the database schema for historical data. * Create tables and relationships. * Populate the SQL database with historical data. * Write 3 complex SQL queries to analyze fraud patterns. | 16 |
| 3 | **PBI\_03: Build ETL Pipeline**   * Extract data from the SQL database using SQL queries. * Load data into Python using pandas and SQLAlchemy. | 4 |
| 4 | **PBI\_04: Exploratory Data Analysis (EDA)**   * Analyze fraud patterns (e.g., frequency of refunds, suspicious user behaviors). * Generate visualizations for fraud trends using Seaborn, Matplotlib. | 20 |
| 5 | . **PBI\_05: Interactive Dashboard Development**   * Deciding on which visualization software to use (Tableau or PowerBI) * Creating Cards with descriptive statistics * Create Pie chart and Donut Chart to display percentages or ratios * Create Line/ Area chart to visualize Time Series data * Create map for geospatial data | 30 |

**Sprint Planning**

Since this is a project that runs in **bi-weekly sprints**, we will divide the PBIs into smaller tasks and deliver incrementally after each sprint. Below is the Scrum Board plan for the first two sprints.

**Sprint 1: SQL Database and ETL Pipeline (Weeks 1-2)**

**Sprint Goal**: Setup SQL database and create ETL pipeline to load data into Python.

**1. Student bi-weekly performance summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Adm. No. | Name | No. of hours present | Progress1 | Remarks |
| 1. 2317933 | Shaun Kwo Rui Yu | 28 | A | * Excellent time organization and task delegation for sprints * Demonstrated strong understanding in SQL rules like Referential Integrity and Entity Integrity Rule aids with data preprocessing that enable smooth SQL bulk insert * Able to create functional Microsoft SQL database |
| 2. 2317454 | Loh Yip Khai | 18 | A | * Excellent attendance and engagement throughout the sprint. * Show good fundamental in using ETL Pipeline for Python |
| 3. 2317483 | Koh Zi Hao | 14 | A | * Excellent attendance and engagement throughout the sprint. * Demonstrated strong fundamentals of data engineering by completing 3 complex SQL Queries to gather insights * -Created ERD Diagram |

1 State whether: A=On Schedule B=Ahead Schedule for no. of days C=Behind Schedule for no. of days

**2. Weekly Scrum**

|  |  |
| --- | --- |
| Week No: 1-2 Date: 21/10/2024 | |
| Member Name 1: | **Shaun Kwo Rui Yu** |
| This Weeks’ Completed Deliverable | * Cleaned and preprocessed data. * Functional SQL database with transaction data. |
| Obstacles | * **Referential and Entity Integrity in SQL**: While working on SQL rules like Referential Integrity and Entity Integrity, I sometimes struggled to identify and resolve violations in large datasets. Ensuring data consistency during preprocessing and cleaning in Jupyter Notebook often required additional effort to debug errors before performing a bulk insert. * **Database Creation in Microsoft SQL**: When creating functional databases in Microsoft SQL, I occasionally faced challenges in designing a schema such as problems with connecting the order\_features and fraud\_labels tables together as I had to use a composite key of (country\_code, order\_id). |
| Member Name 2: | **Yip Khai** |
| This Weeks’ Completed Deliverable | * An ETL pipeline that loads data from SQL into Python. |
| Obstacles | * **Database Connection Error** Unable to connect to the database due to incorrect connection strings, driver issues, or firewall restrictions. * **Large Database**  Large SQL queries might take a long time to execute due to our data(training-data-customer-features\_v1.0.csv) having 2 million+ rows, or the database server may become overloaded. |
| Member Name 3: | **Zi Hao** |
| This Weeks’ Completed Deliverable | * 3 complex SQL queries providing fraud insights. |
| Obstacles | * Joins between large tables (e.g., order\_features and fraud\_labels) can be slow, especially when combined with aggregate functions. * Calculating averages or counts with multiple conditions (e.g., separate fraud and non-fraud cases) can make queries harder to debug and optimize. |

**Sprint 2: Data Analysis (Weeks 3-4)**

**Sprint Goal**: Perform Exploratory Data Analysis (EDA) and Data Visualization

**1.** **Student bi-weekly performance summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Adm. No. | Name | No. of hours present | Progress1 | Remarks |
| 1. 2317933 | Shaun Kwo Rui Yu | 6 | A | * Able to make use of Tutorial 4 or 5 for Big Data Management methods * Nice insights from fraud\_labels * Good work managing scrum document |
| 2. 2317454 | Loh Yip Khai | 7 | A | * EDA (Orders dataset) |
| 3. 2317483 | Koh Zi Hao | 7 | A | * Scrum Document * report * EDA (Customer dataset) |

1 State whether: A=On Schedule B=Ahead Schedule for no. of days C=Behind Schedule for no. of days

**2. Weekly Scrum**

|  |  |
| --- | --- |
| Week No: 3-4 Date: 4/11/2021 | |
| Member Name 1: | **Shaun** |
| This Weeks’ Completed Deliverable | **Big Data Management** **Convert CSV to HDF5 files**Advantage: Faster data loading and processingDisadvantage: Large storage space is needed as HDF5 takes up more data**Optimize data types for storage space and processing**INT64 is converted to INT16 and FLOAT16 to FLOAT64Advantage: Lesser memory used and faster loading and processingDisadvantage: Data is less accurate if data is cut off. **Exploratory Data Analysis for fraud\_labels dataset**   * Create pie chart to display the percentage of frauds for SP-Buy platform * Create sunburst chart for displaying percentage of fraud for each country * Create bar chart to display fraud ration which is   fraud orders/total orders |
| Obstacles | * I got some error when trying to convert the CSV files into saved HDF5 version as I did not install a python library that was required. * When I tried to create subplot of percentage by countries, it did not work so I ended up using plotly sunburst chart which looks interesting. |
| Member Name 2: | **Yip Khai** |
| This Weeks’ Completed Deliverable | **Data Analysis for Orders dataframe**   * Histogram for order\_value, num\_items\_ordered, refund value * Bar Chart of order\_value, num\_items\_ordered by region * Horizontal Bar Chart of Order Value by Payment Method Grouped by Country Code * Horizontal Bar Chart of Order Value by Payment Method Grouped by Country Code * Horizontal Bar Chart of num\_items\_ordered Value by Payment Method Grouped by Country Code * Horizontal Bar Chart of refund Value by Payment Method Grouped by Country Code * Bar Chart of Refund Value to Order Value Ratio by Country * Bar Chart of Refund to Order Ratio by Payment Method Grouped by Country Code |
| Obstacles | * **Formatting Histograms**  For Histograms, extreme values in order\_value, refund\_value, or num\_items\_ordered can skew the distribution, making the graph harder to interpret. To address this, I applied a log transformation to make the data more visually accessible. * **Formatting Axis Labels** When grouping by Payment Method and Country Code, I encountered issues with the payment method axis being cluttered and hard to read. To resolve this, I formatted the axis by rotating the labels and resizing the chart, ensuring all categories were clearly visible. |
| Member Name 3: | **Zi Hao** |
| This Weeks’ Completed Deliverable | * Conducting EDA (Exploratory Data Analysis) on Customer feature dataset * A pie chart for mobile\_verified and total\_payment\_last\_50days * A heatmap to find correlations for numerical variables * A interactive pie chart for Order Activity Level using nym\_orders\_last\_50days * Scatterplot and Correlation Coefficient between num of orders, num of refunds and num of cancellation, num of refunds * Comparative Bar Graph by region using num of orrders, num of cancelled orders, num of refund orders and total payment * A bar graph displaying the num of refund ratio for each region   (calculated by total number of refund orders/ total number of orders) |
| Obstacles | * Created multiple visual charts and graph during EDA phase but some was not insightful enough * Too many charts resulted in the decision to remove some but there are also some values in the removed charts |

**Sprint 3: Dashboard Refinement and Feedback (Weeks 5-6)**

**Sprint Goal**: Refine the dashboard based on user feedback and improve the data analysis models.

**1.** **Student bi-weekly performance summary**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Adm. No. | Name | No. of hours present | Progress1 | Remarks |
| 1. 2317933 | Shaun Kwo Rui Yu | 12 | A | * Scrum Document * Tableau Dashboard * Power Points Slides |
| 2. 2317454 | Loh Yip Khai | 12 | A | * Scrum Document * Tableau Dashboard * Power Point Slides |
| 3. 2317483 | Koh Zi Hao | 6 | A | * Scrum Document * Tableau Dashboard * Power Point Slides |

1 State whether: A=On Schedule B=Ahead Schedule for no. of days C=Behind Schedule for no. of days

**2.** **Weekly Scrum**

|  |  |
| --- | --- |
| Week No: 5-6 Date: 18/11/2021 | |
| Member Name 1: | **Whole Group (Shaun, Yip Khai, Zi Hao)** |
| This Weeks’ Completed Deliverable | -Transferred and created relevant charts and graphs into Tableau  1. Bar Chart   * No. Of Orders and Value (Fraud vs Not Fraud) * Refund Value Ratio   2. Pie Chart   * Fraud vs Not Fraud   3. Donut Chart   * % of Orders Refunded * % of Orders Cancelled   4. Map (Refund Ratio)  5. Area Chart (No. Of Order and Refund Orders)  6. Card   * Num Orders Last 50 Days * Total Payment Last 50 Days * Total No. Of Fraud |
| Obstacles | **Filter Problems:** There were problems with action filters for the Map (Refund Ratio), where we initially used Country Code for filtering but when filtered all charts except the Refund Value Ratio Chart were unable to be displayed.  **Inconveniences:** We initially created a second layer of pie charts, layered over the map, where we could use it as filter to display fraud vs not fraud by country. However, this was inconvenient and unpleasant as we could not display the statistics of the selected country  **KPI Chart Problem:** There were too many rows of data to load and display for our KPI Chart. Whenever we loaded up the dashboard with the KPI Chart, it would take at least 5 minutes to open it, causing much inconvenience. |

**Conclusion**

Using Agile, we have broken down the project into manageable user stories and product backlog items. The project will be developed over **bi-weekly sprints**, focusing first on building the SQL database and ETL pipeline, then on data analysis and dashboard creation. By the end of **Sprint 3**, the team will deliver a polished fraud detection system with an interactive dashboard for the company to reduce fraudulent activities on the platform.

This approach ensures that the team can adapt to changes, receive regular feedback, and deliver incremental value throughout the project.