Monash University

FIT5202 - Data processing for Big Data 2024 S2

Assignment 2B: Using Real-time Streaming Data and Machine Learning to Predict and Visualise eCommerce Fraud

Due: **(Week 11, Friday 23:55 11/Oct/2024)**
Worth: **15%** of the final marks

Background

In the dynamic world of eCommerce, **M**onash **F**ashion **C**orporation (**MFC**, an imaginary company) has established itself as a leading online retailer of fashion products, catering to a diverse and global customer base. With a wide range of products, from trendy apparel to stylish accessories, MFC has built a reputation for quality and customer satisfaction. Our platform leverages advanced technologies to provide a seamless shopping experience, ensuring customers can easily browse, select, and purchase their desired items. However, as our business has grown, so too have the challenges associated with maintaining the integrity and security of our transactions. Unfortunately, the rise of digital commerce has been accompanied by increased fraudulent activities, which pose significant risks to our financial stability and customers' trust. To address these challenges, we are committed to implementing cutting-edge solutions to detect and prevent fraudulent transactions in real-time.

Problem Statement & Project Objective (See A2 Part A)

We have already developed the machine learning models in part A of the assignment. **In part B, we will create a prototype streaming application to demonstrate the integration of machine learning models, Kafka, and Spark streaming. First, we predict potential fraud in** real time as customers browse the website. Then, we visualise the data to help the company make better business decisions.

Based on customers' browsing behaviours, we will focus on two aspects:

1) Predicting the potential fraudulent transaction to warn the company's operation team.
2) Predicting potential inventory requirements can help the company optimise its inventory and logistics.

<u>Required Datasets in Moodle:</u>

- All files from assignment 2 Part A(A2A).
- Your saved best model from A2A
- new_browsing_behaviour.csv: (New browsing behaviour data to simulate real-time streaming)
- new_transaction.csv: (New transactions linked to browsing behaviour)

What you need to achieve

The company requests a prototype application to ingest the new browsing behaviour data and predict potential fraud. To achieve this, you need to simulate the streaming data production using Kafka and then build a streaming application that ingests the data and integrates the machine learning model to predict potential fraud.

After the submission, a compulsory interview/demo will be arranged during Week 12 labs to demonstrate and discuss your prototype application.
The teaching team only marks submissions during the demo. Failure to attend this interview will result in 0 marks (for the whole part B).
(If you have an extension/special consideration, more demo sessions will be arranged after the SWOT.)

Architecture

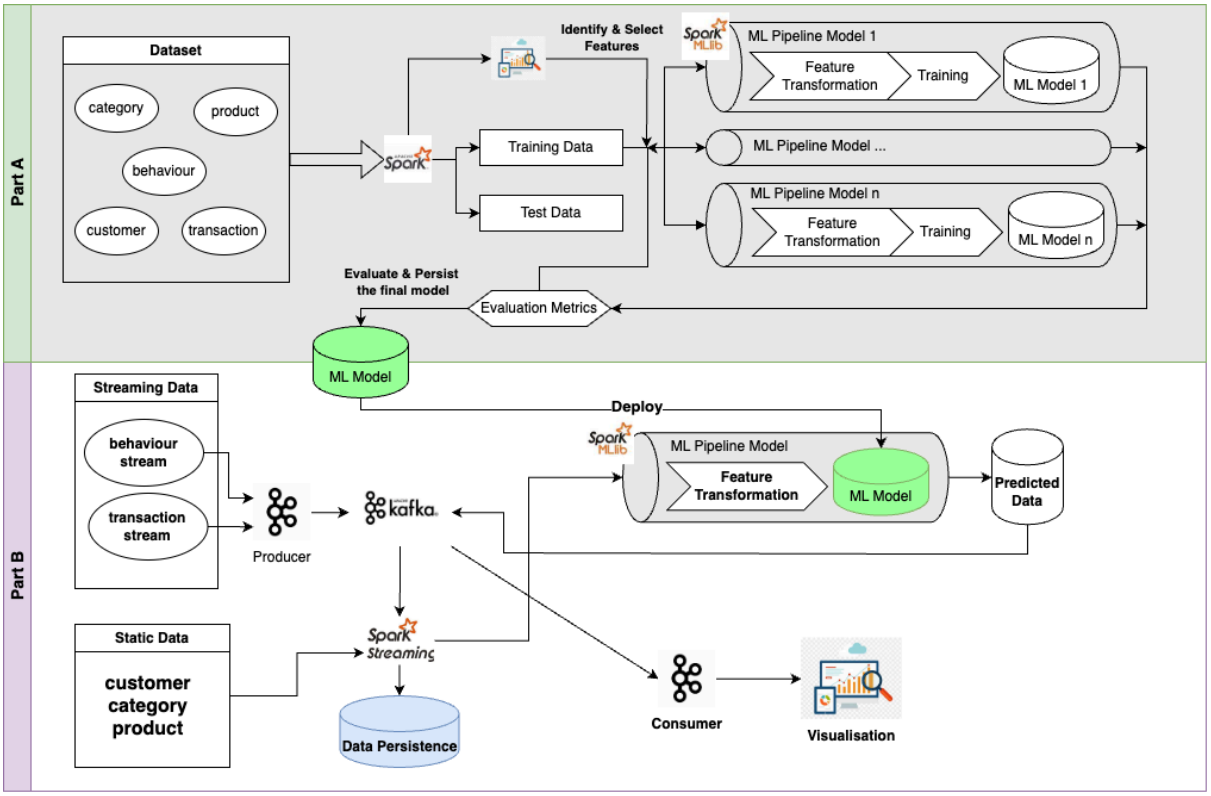

Fig 1: Overall Architecture for Assignment 2 (This assignment is Part B)

Part B of the assignment consists of creating data streams in Kafka and joining static data frames with streaming data to create features, using Spark streaming/MLLib to predict potential sales/inventory requirements, and visualising predicted data.
Overview:

1. In task 1, you will read and publish the data to the Kafka stream.
2. In task 2, you will process the data streams using Spark Structured Streaming and PySpark ML/DataFrame.
3. For task 3, you will read the data from the Kafka stream from part 2 and visualise it.

Please follow the steps to document the processes and write the codes in the Jupyter Notebook.

Getting Started

- Download the dataset from Moodle.
- Download three ipynb template files from Moodle
    - ***A2B-Task1_producer.ipynb*** file for streaming data production
    - ***A2B-Task2_spark_streaming.ipynb*** file for consuming and processing data using Spark Structured Streaming
    - ***A2B-Task3_consumer.ipynb*** file for consuming the data using Kafka and visualise

Part 1. Producing the data (10%)

In this task, we will implement Apache Kafka producers to simulate real-time data streaming. Spark is not allowed in this part since it's simulating a streaming data source.

1. Your program should send one batch of browsing behaviour data **every 5 seconds**. One batch consists of a **random 500-1000 rows** from the browsing behaviour dataset. The CSV shouldn't be loaded to memory at once to conserve memory (i.e. Read row as needed). Keep track of the **start and end** event_time. (You can assume the dataset is sorted by event_time.)
2. Add an integer column named '**ts**' for each row, a Unix timestamp in seconds since the epoch. Spead your batch out evenly for 5 seconds.
    a. For example, if you send a batch of 600 records at 2023-09-01 00:00:00 (ISO format: YYYY-MM-DD HH:MM:SS) -> (ts = 1693526400):
        - Record 1-120: ts = 1693526400
        - Record 121-240: ts = 1693526401
        - Record 241-360: ts = 1693526402
        - ….
3. Read the transactions between the start and end event_time in 1.1 every 5 seconds (the same frequency as browsing behaviour) and create a batch.
4. Send your two batches from 1.1 and 1.3 to Kafka topics with an appropriate name.

Note 1: In 1.1, "random 500-1000" means the number of rows is random, and the data file is still read sequentially.
Note 2: All the data except for the 'ts' column should be sent in the original String type without changing to any other type. This is because we are simulating a streaming access log and need to reduce the required processing at the source.

Save your code in ***Assignment-2B-Task1_producer.ipynb***.

Part 2. Streaming application using Spark Structured Streaming (50%)

In this task, we will implement Spark Structured Streaming to consume the data from task 1 and perform predictive analytics.

Important:

- **This task uses PySpark Structured Streaming with PySpark Dataframe APIs and PySpark ML.**
- **You also need your pipeline model from A2A to make predictions and persist the results.**
- **If you didn't complete A2A or failed to build an ML model, we still encourage you to attempt this part. You can use the "is_fraud" label to complete many tasks in this part; however, the maximum mark you will receive is capped at 80% (below HD).**

1. Write code to create a SparkSession, which 1) uses **four cores** with **a proper application name**; 2) use the Melbourne timezone; 3) ensure a checkpoint location has been set.
2. Similar to assignment 2A, write code to define the data schema for the data files, following the data types suggested in the metadata file. Load the static datasets (e.g. customer, product, category) into data frames. (You can use your code from 2A.)
3. Using the Kafka topics from the producer in Task 1, ingest the streaming data into Spark Streaming, assuming all data comes in the **String** format. Except for the 'ts' column, you shall receive it as an **Int** type.
4. Then, the streaming data format should be transformed into the proper formats following the metadata file schema, similar to assignment 2A. Perform the following tasks:
    a) For the 'ts' column, convert it to the timestamp format, we will use it as **event_ts**.
    b) If the data is late for more than 2 minutes, discard it.
5. Aggregate the streaming data frames and create features you used in your assignment 2A model. (note: customer ID has already been included in the stream.) Then, join the **static** data frames with the streaming data frame as our final data for prediction. Perform data type/column conversion according to your ML model and print out the Schema. (Again, you can reuse code from A2A).
6. The company is interested in the number of potential frauds as they happen and the products in customers' shopping carts (so that they can plan their stock level ahead.) Load your ML model, and use the model to predict/process each browsing session/transaction as follows:
    a) Every 10 seconds, show the total number of potential frauds (prediction = 1) in the **last 2 minutes,** and persist the raw data (see 7a).
    b) Every 30 seconds, find the top 20 products (order by quantity descending) in the last 30 seconds, show product ID, name and total quantity. We only need the non-fraud transactions (prediction=0) by extracting customer shopping cart details (sum of all items of ADD_TO_CART(ATC) events from browsing behaviour, you can also extract it from transactions).
7. Write a Parquet file and save the following data frames (tip: you may look at part 3 and think about what columns to save):
    a. Persist the **raw data from 6a** in **parquet format**. Every student may have different features/columns in their data frames depending on their model, at the bare minimum, we need some IDs to identify those frauds later on (transaction_id and/or session_id). After that, read the parquet file and show a few rows to verify it is saved correctly.

b. Persist the data from 6b in another parquet file.

8. Read the two parquet files from task 7 as **data streams** and send to Kafka topics with appropriate names
   (Note: **You shall read the parquet files as a streaming data frame and send messages to the Kafka topic when new data appears in the parquet file.)**

Save your code in ***Assignment-2B-Task2_spark_streaming*.ipynb**.

Part 3. Consuming data using Kafka and Visualise (20%)

In this task, we will implement an Apache Kafka consumer to consume the data from Part 2.

Important:

- **In this part, Kafka consumers are used to consume the streaming data published from task 2.8.**
- **This part doesn't require parallel processing, and please do not use Spark in this part. It's OK to use Pandas or any Python library to do simple calculations for the visualisation.**

1. (**Basic plot**) Plot a diagram with two subplots to show the following:
   a) Left subplot: Show a bar chart of potential fraud count every 10 seconds(visualise data from 7a);
   b) Right Subplot: Show a line chart of cumulative sales of top 20 products (non-fraud, sum of product*qty), update every 30 seconds. (visualise data from 7b)
2. (**Advanced plot**, open question) Be creative and create an advanced plot (not a line/bar chart).
   For example, plot a choropleth or bubble map to show where the fraudulent transaction happens the most.
   Choropleth: https://python-graph-gallery.com/choropleth-map/
   Bubble Map: https://python-graph-gallery.com/bubble-map/
   (hint: if you want to do a map plot, you may want to reduce the granularity of latitude and longitude by reducing the precision. See: Decimal degrees - Wikipedia)

Note: Both plots shall be real-time plots, which will be updated if new streaming data comes in from part 2. For the advanced plot, if you need additional data for the plots, you can add them in part 2.

Save your code in ***Assignment-2B-Task3_consumer.ipynb***.

Part 4: Demo and Interview (20%)

IMPORTANT: The interview is compulsory, and we only mark your A2B during the interview. No marks will be awarded if the interview is not attended(0 marks for the whole A2B, not just this section).

The demo/interview session details will be announced/arranged in Week 11. Please pay attention to the unit announcement email and Ed forum.

Each demo is roughly 10 minutes. You have 5-6 minutes to show your application; then, your marker will ask 3-4 questions to assess your understanding. Please come to your allocated demo session a few minutes early and ensure your laptop/application works correctly.

Demo/Interview is marked on a 5-level scale:

The demo is working, and the student has a competent understanding	20
Working demo, partial understanding	15
The demo is not working, partial understanding	10
The demo is not working, low understanding	5
No attendance or can't answer most of the questions	0

Assignment Marking Rubric

Detailed mark allocation is available in each task. For complex tasks and explanation questions, you will receive marks based on the quality of your work.

In your submission, the jupyter notebook file should contain the **code and its output**. It should follow *programming standards, readability of the code, and organisation of code*. Please find the PEP 8 -- Style Guide for Python Code for your reference. Here is the link: https://peps.python.org/pep-0008/ Penalty applies if your code is hard to understand with insufficient comments.

Submission

You should submit your final version of the assignment solution via Moodle. You must submit the following:

- A **zip** file named based on your authcate name (e.g. abcd1234). The zip file should contain
    - ***Assignment-2B-Task1_producer_authcate*.ipynb**
    - ***Assignment-2B-Task2_spark_streaming_authcate*.ipynb**
    - ***Assignment-2B-Task3_consumer_authcate.ipynb***

  The file in submission should be a ZIP file and *not any other kind of compressed folder (e.g. .rar, .7zip, .tar)*. Please do not include the data files in the ZIP file.
- The assignment submission should be uploaded and finalised by **(Week 11, Friday 23:55 11/Oct/2024).**

Other Information

Where to get help

You can ask questions about the assignment in the Assignments section in the Ed Forum, which is accessible on the unit's Moodle Forum page. This is the preferred venue for assignment clarification-type questions. You should check this forum regularly, as the responses of the teaching staff are "official" and can constitute amendments or additions to the assignment specification. Also, you can attend scheduled consultation sessions if the problem and the confusion are still unresolved.

Searching and learning on commercial websites/forums (e.g. Quora, Stack Overflow) is allowed. However, you should not post/ask assignment questions on those forums.

Plagiarism and collusion

Plagiarism and collusion are severe academic offences at Monash University. Students must not share their work with any other students. Students should consult the policy linked below for more information.

https://www.monash.edu/students/academic/policies/academic-integrity

See also the video linked on the Moodle page under the Assignment block.
Students involved in collusion or plagiarism will be subject to disciplinary penalties, which can include:

- The work not being assessed
- A zero grade for the unit
- Suspension from the University
- Exclusion from the University

Late submissions and Special Consideration

ALL Special Consideration, including within the semester, is now handled centrally. This means that students MUST submit an online Special Consideration form via Monash Connect. For more details, please refer to the **Unit Information** section in Moodle.

There is a **5% penalty per day including weekends** for a late submission. Also, the cut-off date is 7 days after the due date. No submission will be accepted (i.e. zero mark) after the cut-off date unless you have a special consideration.

Mark Release and Review

- Mark will be released within 10 business days after the submission deadline.
- Reviews and disputes regarding the mark will be accepted a maximum of 7 days after the release date (including weekends).

Generative AI Statement

As per the University's policy on the guidelines and practices pertaining to the usage of Generative AI:

AI & Generative AI tools may be used SELECTIVELY within this assessment.

Where used, AI must be used responsibly, clearly documented and appropriately acknowledged (see Learn HQ).

Any work submitted for a mark must:

1. Represent a sincere demonstration of your human efforts, skills and subject knowledge that you will be accountable for.
2. Adhere to the guidelines for AI use set for the assessment task.
3. Reflect the University's commitment to academic integrity and ethical behaviour.

Inappropriate AI use and/or AI use without acknowledgement will be

considered a breach of academic integrity.

The teaching team encourage students to apply their own critical thinking and reasoning skills when working on the assessments with assistant from GenAI. Generative AI tools may produce inaccurate content and this could have a negative impact on students' comprehension of big data topics.

Data source acknowledgement:

The dataset is a remix based on several real-world datasets. Transaction records are from real-world data, user name, age, dob, salary, etc. are randomly generated synthetic datasets.

We thank the authors/owners for sharing the original datasets.

1. eCommerce behavior data from multi category store | Kaggle
2. REES46
3. E-Commerce Data | Kaggle
4. Brazilian E-Commerce Public Dataset by Olist | Kaggle
5. Geoscape Geocoded National Address File (G-NAF) - Dataset - data.gov.au
6. Popular Baby Names - Dataset - data.sa.gov.au
7. Fashion Campus | Kaggle