REPORT

Section 1:

What are key information that can be derived from the plot?
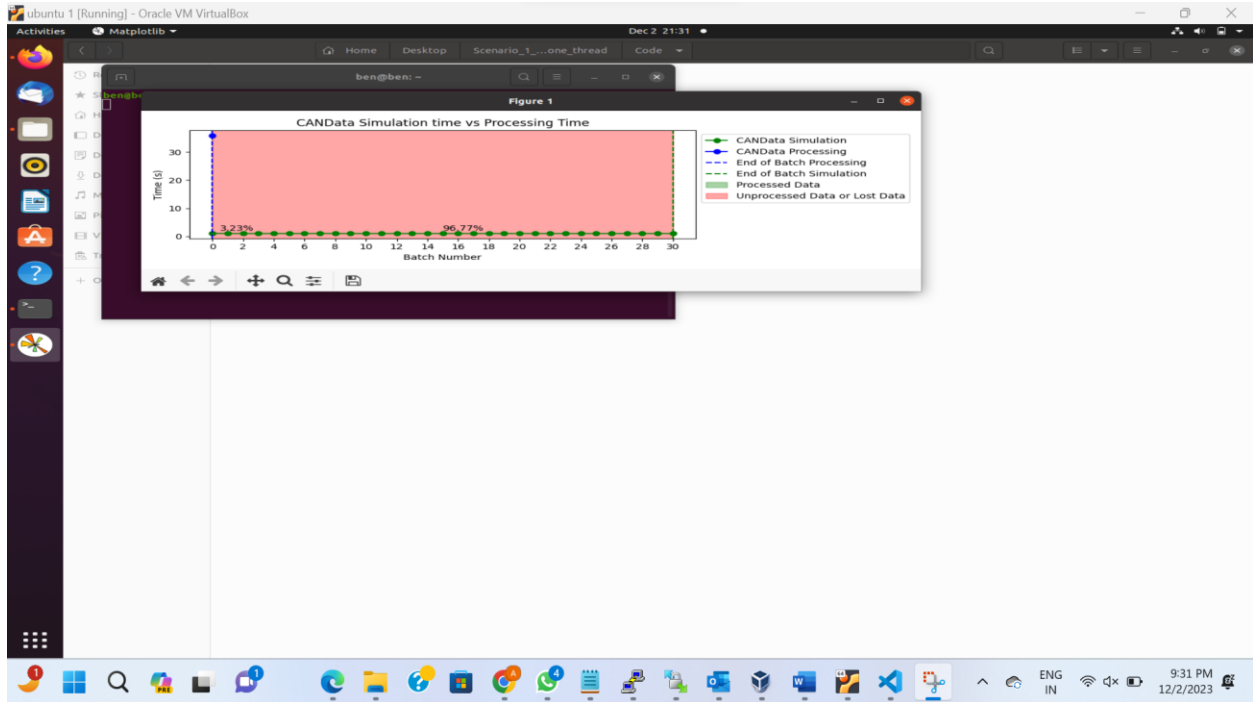

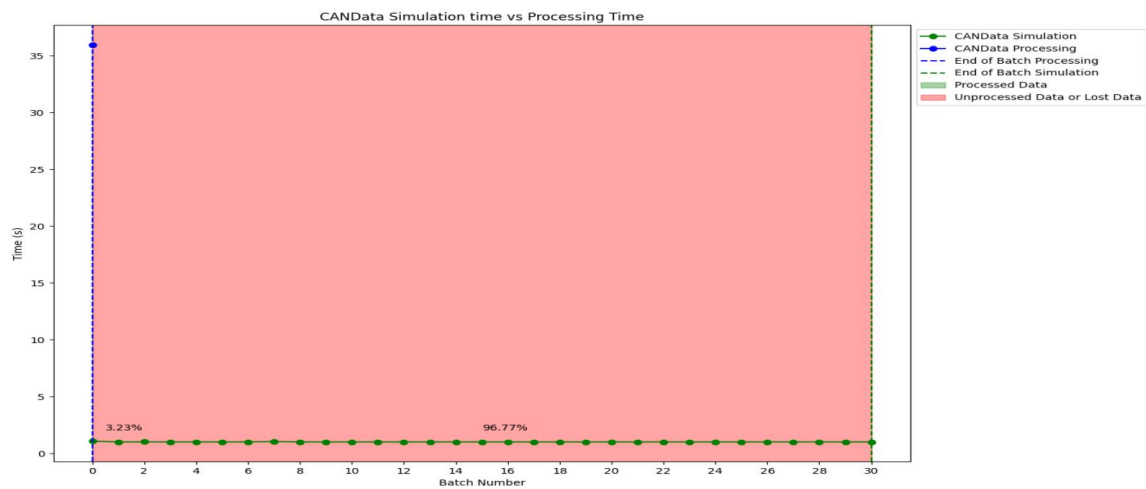Fig 1: CANData Simulation time vs Processing Time.

Fig 2: CANData Simulation time vs Processing Time.

- The graph represents the time taken for CANData simulation vs processing across different batches of data.
- The yaxis represents time in seconds, and the xaxis represents the batch number.
- **Only 3.23% of the CANData is processed.**
- **96.77% of the CANData is lost or unprocessed.**
- There is huge data loss and only 3.23% of the data is processed which is very less.
- The batch number ranges from 0 to 30, whereas the time ranges from 0 to 35.

Section 2:

1. Why do you think there is a data loss? Is it significant? Explain!

   a. **Splitting into Batches:** As we can see in the graph, there is significant loss of the process data(96.77 percentage) and only very less amount of the data(3.23%) is used. The data loss may happened while splitting the data into multiple batches.
   b. **Bottlenecks in the Pipeline:** The data transfer in the pipeline could be slow which might have caused data loss. While transferring the data into the pipeline the data may be discarded because of the pipeline storage or capacity is less.
2. What could be possible solutions to mitigate data loss?
   a. **Buffer Solution:** Adding a buffer solution like adding queue to manage data flow which ensures that data is stored to be processed. This ensures that data is not discarded by overloading the system.
   b. **Load Balancing:** Load balance the data into multiple nodes to prevent the data loss. By doing this we are reducing the load on the resources. This gives the resources time and storage to process the data without any loss.

Reflections:

1. The data needs to be monitored to identify the data loss. The graph plot helps us to identify how much data is lost and how much data is processed.
2. We also need to check the capacity of the system and the data size. If the data size is more or the processing of the data takes more space than the system storage the overflow may occur.
3. We also make sure the system which is processing is robust of handling such data.
4. Changing the parameters of the Joblib library may improve the prevention of the data.