**Capstone Project: Real-Time Retail Invoice Processing and Customer Loyalty Analytics**

**Project Overview:**

In this project, you will develop a **real-time data processing system** to track and analyze retail invoices and customer loyalty points. The system will ingest streaming data containing invoice details, store and customer information, and compute key metrics, such as total sales per store, the number of items sold, and customer loyalty points.

The project will be divided into three layers following the **Bronze, Silver, and Gold architecture** pattern to streamline data ingestion, transformation, and analytics.

**Objective:**

* Build a real-time streaming pipeline to process retail invoices.
* Track key business metrics, such as total sales, number of items sold, and most popular items.
* Implement a customer loyalty points system that awards points based on total purchases.

**1. Bronze Layer (Raw Data Ingestion)**

**Objective:**  
Ingest raw invoice data from a streaming source such as JSON or CSV files. The raw data will be minimally processed and stored for future transformations.

**Key Fields:**

* Invoice details: InvoiceNumber, CreatedTime, StoreID, PosID, CashierID
* Customer details: CustomerType, CustomerCardNo
* Transaction details: TotalAmount, NumberOfItems, PaymentMethod, TaxableAmount, CGST, SGST, CESS, DeliveryType, DeliveryAddress
* Invoice items: InvoiceLineItems (an array of items with ItemCode, ItemDescription, ItemPrice, ItemQty, and TotalValue)

**Data Ingestion Approach:**

* Ingest streaming invoice data from a folder where new files (JSON/CSV) are continuously added.
* Store the raw data in Parquet format in the Bronze layer for later cleaning and processing.

**2. Silver Layer (Data Cleaning and Transformation)**

**Objective:**  
Clean and transform the raw data by handling missing or invalid entries. Convert data types, compute derived columns, and normalize the schema for downstream analytics.

**Transformations:**

* Convert CreatedTime from a Unix timestamp (BIGINT) to a proper date and time format.
* Validate the TotalAmount and NumberOfItems fields to ensure they are non-null and positive.
* Parse the DeliveryAddress structure for consistency.
* Expand InvoiceLineItems to a flat structure for easier processing of individual item-level data.
* Calculate the total taxes (CGST + SGST + CESS) for each invoice.

**Output:**

* Write the clean and transformed data into the Silver layer, where it can be used for reporting and further aggregation.

**3. Gold Layer (Business Aggregation and Analytics)**

**Objective:**  
Use the clean data from the Silver layer to perform real-time analytics and generate insights into store performance and customer loyalty.

**Key Metrics:**

* **Store-Level Sales Analytics:**
  + Total sales (TotalAmount) per store (StoreID) and per cashier (CashierID).
  + Number of invoices processed per store.
  + Number of items sold per store and most popular items (ItemCode, ItemDescription).
* **Customer Loyalty Analytics:**
  + Calculate loyalty points for customers based on the TotalAmount spent (e.g., 1 point per $10 spent).
  + Track customer purchase behavior based on CustomerCardNo.
  + Identify high-value customers by filtering those with total points above a certain threshold.

**Output:**

* The aggregated results will be stored in the Gold layer, providing real-time insights into sales performance and customer behavior.
* These insights can be visualized through a dashboard (e.g., using Databricks SQL or another BI tool) for tracking store and customer KPIs.

**Real-Time Use Case Scenarios:**

1. **Store-Level Performance Monitoring:**  
   Store managers can use the real-time analytics system to monitor sales and operational performance across different locations. They can track which store has the highest sales, identify top-performing cashiers, and determine the most popular products sold.
2. **Customer Loyalty Tracking:**  
   The system will track customer purchases in real-time and award loyalty points accordingly. It can also identify high-value customers, enabling targeted marketing strategies such as personalized promotions or loyalty rewards for top customers.
3. **Inventory and Product Insights:**  
   The system can provide insights into product popularity, helping store managers make informed decisions regarding restocking and managing inventory based on customer demand.

**Technical Overview:**

* **Data Sources:**
  + Invoices from stores, streaming in real-time from JSON or CSV files.
* **Layers:**
  + **Bronze:** Raw data storage for invoice data.
  + **Silver:** Data cleaning and normalization.
  + **Gold:** Business metric aggregation and customer loyalty point tracking.
* **Tools and Technologies:**
  + **Spark Structured Streaming:** For real-time ingestion and processing.
  + **Parquet Format:** For efficient storage and querying.
  + **Databricks SQL/BI Tools:** To visualize and analyze the results in the Gold layer.