### **Subject**: **Sr. Java Backend Developer - Real-Time Streaming Engine Challenge**

### ***Introduction***

Congratulations on advancing to the **FINAL** round of interviews! You will join our organization upon passing this assessment. This document outlines a challenge designed to assess your skills in designing, architecting, and implementing a real-time streaming engine. Your solution should demonstrate your expertise in handling large volumes of data, ensuring data quality, implementing real-time processing, and optimizing for performance and scalability.

### ***Problem Statement***

You are required to design and implement a real-time streaming pipeline using streaming platforms/technologies (KafkaStreams/Apache Flink) that processes millions of records with minimal latency. This engine will handle data ingestion, processing, and aggregation from multiple sources to generate real-time analytics and insights.

### ***Data Sources***

The application will ingest data from the following sources:

1. **GPS Real-Time Transactions** (real time 4 sources via Kafka)
   * **Schema**: CustomerID, CarID, OfficeID, AgentID, TRXN Timestamp, CarDrivingStatus (Stopped/Idle/Moving), CurrentLongitude, CurrentLatitude, CurrentArea, KM
   * **Note:** these are 4 source systems (SRC01, SRC02, SRC03, SRC04) generating data with identical schema
2. **Sales Transactions (realtime API, sales system will submit the sales transaction to the Engine via API calls)**
   * **Schema**: AgentID, OfficeID, CarID, CustomerID, Amount (USD)
3. **Customer Data** (**realtime API, our engine pulls the new customer details via API calls to the CRM Oracle DB**)
   * **Schema**: CustomerID, Mobile No, Name, Gender, Age, Nationality, PassportNo, ID No, Home Address, Lease Start Date, Lease Period
4. **Car Data** (realtime file update)
   * **Schema**: CarID, Car Make, Car Model, Plate No, Registration Date, Registration Expiry Date
5. **Office Branch Data** (batch file update)
   * **Schema**: OfficeID, Mobile No, Area, Office No, Working Hours
6. **Sales Agent Data** (batch file update)
   * **Schema**: AgentID, Mobile No, Name, Gender, Age, Nationality, OfficeID

Below are sample data and structures for the above sources (dummy data only to put you in context):

A screenshot of a computer

Description automatically generated

#### **High Level Basic Requirements**

1. Process incoming data streams in real-time.
2. Apply data quality checks to clean the data.
3. Build real-time aggregates for analytics purposes (below are basic sample aggregations, feel free to add more based on the data):
   * Count of sales per month and year.
   * Volume of sales per month and year.
   * Count of Car Driving Status per area (Idle, Moving, Stopped).
4. Enrich transactional data by joining with reference data.
5. Filter data based on city name (e.g., "Dubai").
6. Anonymize mobile numbers using SHA256.
7. **Data Transfer:** 
   * Transfer output filtered data file via sFTP in batches of 2000 records each file.
   * Sync the output data in realtime through API to an external system.
   * Output Data Structure: BatchID, TRXN Timestamp, Car Driving Status, Current Longitude, Current Latitude, Current Area, KM, Amount (USD), Customer Hashed Mobile No, Name, Gender, Age, Nationality, PassportNo, ID No, Home Address, Lease Start Date, Lease Period, Care Make, Car Model, Plate No, Registration Date, Registration Expiry Date, Office Mobile No, Area, Office No, Working Hours, Agent Mobile No, Name, Gender, Age Band, Nationality
   * Age Band [<18, 19-30, 31-50, 51+]
8. Store and maintain both anonymized and non-anonymized copies of the data.
9. Implement security measures to protect the data and access to the system and its DBs.

### ***Deliverables:***

1. **Detailed Solution Design Document (MUST BE DELIVERED)**: End to end detailed design document that encompasses architecture**,** design with workflows and diagrams, technologies involved, frameworks, tools, integrations, best practices, security, operational management, automation, etc. Design the overall system and its components following the best practices and your experience distinguishing yourself among other candidates. Making sure it is an enterprise-wide application.
2. **Implementation (MUST BE DELIVERED)**: Write code for the below flow and demonstrate it in your IDE:
   1. **Source:** Mimic data producers to generate dummy data based on the schemas above.
   2. **Ingestion:** Read the records (transactional and references), reject the bad records based on the schemas.
   3. **Enrichment:** Join the data together (transactional and references)
   4. **Anonymization:** Apply SHA256 on customer mobile number
   5. **Filtration:** Filter out the records for City “Dubai”
   6. **Output:** Generate the final output files based on 500 record per file
3. **Optional Deliverables (PLUS, NICE TO HAVE)**:
   1. Design an alternative solution using any other streaming platform or methodology
   2. Implement and demonstrate your solution in a Docker environment.
   3. Build dashboards to show insights and analytics (e.g., sales aggregates).

### ***Presentation***

During the interview, you will present your solution and code from IDE. Ensure you cover:

1. **Architecture and Design**: Walk through your design document following all details and hints mentioned before.
2. **Challenges and Solutions**: Highlight all possible challenges and how you will prevent them and resolve them.
3. **Tools & Software**: Describe the tools, technologies and software required in development, code management, DBs, monitoring, etc. highlighting the strength of the tool and why it is needed and function.
4. **Hardware**: Estimate the required hardware needed to run this engine
5. **Guidelines:** Based on your experience and best practices, detail and describe the tactics and guidelines that need to be followed in the development, data security, testing, UAT, deployment and on production environment and postproduction and maintenance. E.g. design patterns, usage of data structures and algorithms, CI/CD, etc.
6. **Plan**: Create an end-to-end delivery plan with timelines, milestones, resources required, why needed, engagement period of each resource, etc.
7. **Code Demonstration**: Show your code (mandatory part: as requested before, PLUS part: if you managed to implement/code extra) and run through it.
8. **Extra**: In the context of this assessment, please feel free to add and enrich and tune as much as you can to distinguish yourself and show your experience and strength
9. **Apply all your experiences and best practices and right technologies.**
10. **Q&A**: Be prepared to answer questions about your design choices and implementation.

We are excited to see you demonstrate your maximum strength and skills in this assessment. **This is a crucial step in our selection process, and we encourage you to showcase your full potential and do your best to rank number one among the other candidates.**

Your performance in this final interview will determine your place in our organization. We believe in your abilities and look forward to potentially welcoming you to our team.

Once you receive this document, in case you need further clarifications or explanations, feel free to immediately send all your enquiries at one time by email and we will get clarified. This is applicable only on the same day you receive this document.

***Good Luck & Best regards,***