Data Engineering

ETL Process and SQL Analysis

Task Description:

You have been provided with a CSV file representing Call Detail Records (CDRs). This file contains information about subscriber's telecom traffic including voice, data, and SMS events.

No	Column Name	Description
1	**timestamp**	Event Timestamp
2	**caller_msisdn**	Caller's phone Number
3	**callee_msisdn**	Called phone Number
4	**event_type**	Type of the CDR event (sms, voice, data)
5	**caller_city**	City where caller initiated the event
6	**callee_city**	City of the callee number
7	**duration**	call duration (voice only)
8	**volume**	Used Data Volume (Data Only)
9	**cost**	Cost charged for the event
10	**is_roaming**	If the CDR related to roaming Event or not

Objectives:

1. ETL Process:

- Write a Python/Bash,… script - any other convenient script is accepted - to extract the CSV files, transform the data by handling any missing or inconsistent values, and load the cleaned data into a PostgreSQL database.
    - **Missing Data**: caller_msisdn column cannot be null
    - **Missing Data**: In voice records (voice event_type), duration column and in data records (data event_type), volume column cannot be null
    - **Inconsistent Data**: event_type should be from the provided list (sms, voice, data)
    - **Inconsistent Data**: caller and callee number column should be number

Note: Records with inconsistent and missing data (on the aforementioned columns) should be removed and not loaded to database.

2. SQL Analysis:

- Determine the top 10 cities with the highest total voice call durations.
- Identify the top 10 cities generating the most revenue (cost) for all types of events separately and overall, for last week **(2025-06-01 – 2025-06-07)**.
- List the top 10 subscribers (caller_msisdn) by total event cost across all types.
- List the top 10 Roamer subscribers by data volume usage.

Database connection information:

Database Server: Postgresql

Server: 192.168.5.111

Port: 5432

Database name: subscriber_traffic

Database user: tester

User password: tester@321

Traffic File Path: /home/tester/subscriber_traffic.csv

Available Ubuntu Desktop:

Remote IP (RDP): 192.168.5.111

Username: tester

Password: tester@321

Available Software on this remote: VScode, Datagrip