Batch Data Processing for Rental Marketplace Analytics

Project Overview: A rental marketplace platform (similar to Airbnb) requires an end-to-end data pipeline to enable analytical reporting on rental listings and user interactions. The platform stores application data in an AWS Aurora MySQL database, and the goal is to build a data warehouse using Amazon Redshift for business intelligence and reporting.

You are required to implement a batch data processing pipeline that:

- Extracts rental listing data from AWS Aurora MySQL and loads it into Amazon S3 as an intermediate storage layer.
- Ingests data from S3 into Amazon Redshift for structured processing.
- Implements a multi-layer architecture in Redshift with Raw, Curated, and Presentation layers.
- Uses AWS Glue for extraction, transformation, and loading (ETL).
- Orchestrates the workflow using AWS Step Functions to ensure efficient execution.

Key Business Metrics to Derive:

Once the data is available in Redshift's Presentation Layer, generate the following insights:

Rental Performance Metrics:

- **Average Listing Price:** Daily mean price of all available rentals.
- **Occupancy Rate:** Daily percentage of available rental days booked by users.
- **Most Popular Locations:** Daily areas with the highest number of bookings.
- **Top Performing Listings:** Daily properties generating the highest revenue.

User Engagement Metrics:

- **Total Bookings per User:** Number of rentals booked by each user per day.
- **Average Booking Duration:** Mean length of stay across all bookings per day.
- **Repeat Customer Rate:** Percentage of users who book more than once **within the same day**

Evaluation Criteria

- Efficient ETL implementation using AWS Glue & Step Functions.
- Correct data validation and transformation logic.
- Optimized Redshift schema for analytical queries.
- Well-documented setup and troubleshooting guide.