Capstone Project: Customer360 Data Integration & Reporting with AWS Glue

Objective

Build a complete end-to-end ETL pipeline using AWS Glue to create a Customer360 dataset by integrating customer details from CSV files in S3, transaction data from a PostgreSQL RDS, and enriching it with geolocation metadata. The final output should be written to S3 in Parquet format and partitioned for analytics consumption.

Scenario

A retail company collects customer and transaction data from multiple sources:
- Customer Profiles are uploaded daily into an S3 bucket as CSV files.
- Transactions are stored in a PostgreSQL RDS instance.
- Geolocation metadata is provided in a JSON file for customer addresses.

You are tasked to:
1. Clean and deduplicate customer data.
2. Join customer and transaction datasets.
3. Enrich data with geolocation metadata.
4. Output to partitioned Parquet files in another S3 bucket for analytics.

Datasets Provided

Dataset Name	Description	File link
customers.csv	Customer master data (S3 Source)	https://drive.google.com/file/d/1zDVSODnm0Epliqdru1efa-ZxuuBLauF6/view?usp=drive_link
transactions.sql	SQL dump of transaction data for PostgreSQL RDS	https://drive.google.com/file/d/1OhH4Y66Tv-NjDO7yqzeKOcLAFQyo9pzu/view?usp=drive_link
geolocation.json	Metadata about ZIP codes and locations	https://drive.google.com/file/d/1AOegG1VFJOhYszFNTZKIxpD9IbKjIRMT/view?usp=drive_link

Tasks to Perform

Phase 1: Crawler & Catalog

- Create a crawler for customers.csv in S3 and geolocation.json.
- Use JDBC connection to catalog PostgreSQL transactions table.

Phase 2: ETL Job 1 – Customer Data Cleansing

- Remove duplicate customers using customer_id.
- Drop rows with null email or zip_code.
- Write cleaned customer data to S3 in Parquet format (intermediate path).

Phase 3: ETL Job 2 – Join & Enrichment

- Read cleaned customer data, transaction data (JDBC), and geolocation metadata (JSON).
- Join all 3 sources.
- Enrich customer info with city/state info from geolocation metadata.
- Calculate total_transaction_amount and transaction_count.

Phase 4: Final Output

- Write the final Customer360 data to S3 in Parquet format.
- Partition by state and year.

Phase 5: Logging & Monitoring

- Add custom log statements and error handling.

Phase 6: Workflow & Trigger

- Create a Glue workflow that:
  - Runs crawler → cleansing job → enrichment job → sends success status.

Success Criteria

- Final output is partitioned by state and year.
- CloudWatch logs capture job start/end and errors.
- IAM roles use least privilege policy.
- Final job handles schema evolution if customer file changes.