AWS S3, Glue, Athena, and Power BI for Vehicle Insurance Claim Fraud Detection and Analysis

📌 1. Objective

To build a cloud-powered data analysis pipeline that identifies fraudulent vehicle insurance claims and uncovers patterns across customer demographics and vehicle characteristics.

📃 2. Dataset

Source: [Kaggle - Vehicle Claim Fraud Detection]
Attributes Include:

- AgeOfPolicyHolder, VehiclePrice, PolicyType, Make, AgeOfVehicle
- MaritalStatus, Sex, FraudFound_P (Target: 1 = Fraudulent)

☁ 3. Cloud Infrastructure Setup (AWS)

a. S3 Bucket

- Created an S3 bucket.
- Uploaded dataset to a /dataset/ folder.

b. AWS Glue

- Created an IAM role with read/write access to S3.
- Created a crawler to infer schema and populate the AWS Glue Data Catalog.

c. AWS Athena

- Queried the structured dataset using SQL in Athena.
- Performed aggregations on fraud data grouped by policy types, age, vehicle price, and gender.

🔗 4. Power BI Integration

- **ODBC Connection** established between Power BI and Athena.

- Used AWS access key credentials and region-specific configuration.
- Imported query outputs into Power BI for live dashboarding.

📊 5. Power BI Dashboard Highlights

✅ Key Metrics

- **Total Claims**: **15,000**
- **Fraudulent Claims**: **923**
- **Fraud Rate**: **5.99%**

✅ Fraud vs Non-Fraud Distribution

- Non-Fraudulent: **~94%**
- Fraudulent: **~6%**

✅ Fraud Patterns

- **By Vehicle Age**: Higher fraud rates in older vehicles (6–7+ years).
- **By Vehicle Price**: Vehicles costing **more than $69,000** had noticeably more fraud.
- **By Policy Type**:
  - *Sport - Collision*: 13.79%
  - *Utility - All Perils*: 12.06%
  - *Sedan - All Perils*: 10.06%

✅ By Car Make

- High fraud rates for:
  - *Mercedes*, *Acura*, *Saturn*, *Saab*, and *Ford*

✅ Demographics-Based Filtering

- **Gender**, **Marital Status**, and **Age of Policy Holder** were used as filters to further slice the data.

📈 6. Visualizations Used

- **Stacked Bar Charts**: Fraud rate by policy type and vehicle age.
- **Donut Chart**: Overall fraud vs non-fraud distribution.

- **KPI Cards**: Claims volume, fraud volume, and fraud rate.
- **Filters/Slicers**: By sex, marital status, vehicle make, and more.

✅ 7. Findings (from Power BI Dashboard)

Based on attached dashboard data:

- Fraud is **concentrated among older vehicles and higher-value cars**.
- Certain **policy types and car makes** are fraud-prone.
- Fraud distribution shows clear **demographic and policy-based patterns**, which can be used for **rule-based risk assessment** or **predictive modeling**.

🚀 8. Future Enhancements

- Train a machine learning model (e.g., Random Forest or XGBoost) using Glue ML/ SageMaker.
- Automate fraud alerts with AWS Lambda + SNS.
- Add real-time dashboards using QuickSight or embed Power BI into an application.