University of New Haven

TAGLIATELA COLLEGE OF ENGINEERING

POWER ON

AWS RIDE ANALYTICS LAKEHOUSE

S3 + GLUE (PARQUET) + ATHENA + QUICKSIGHT

TEAM 2

TEAM

Faryal Bahawi

Documentation Lead

Aadil Shakya

Data Engineer

Jatin Nabhoya

Cloud Architect

Project Details:

- RidePulse: AWS Ride Analytics Platform
- Project Objective
- Tools and Technologies :
- Expected Outcome

Business scenario overview

Problem Statement:

- Uber collects millions of ride records daily; pickup time, location, fare, and distance.
- The data is massive and messy, making it difficult to manage and analyze in real time.
- The opportunity is to build a cloud-based system that can handle this scale and help find key insights like peak hours and busy locations.

Solution Requirements:

- Use Amazon S3 for secure, scalable storage of raw trip data.
- Use AWS Glue to clean/transform data and write Parquet files back to S3.
- Register tables in the Glue Data Catalog and query with Amazon Athena.
- Visualize insights in Amazon QuickSight.

Solution overview(High Level Description)

- A modern data engineering pipeline on AWS Cloud designed to process NYC Taxi (Uber-like) datasets.
- The pipeline ingests raw CSV data, performs schema discovery and transformation using Mage.ai and AWS Glue, stores optimized data in S3 (Parquet format)
- Enables analytics through Athena and Amazon QuickSight.

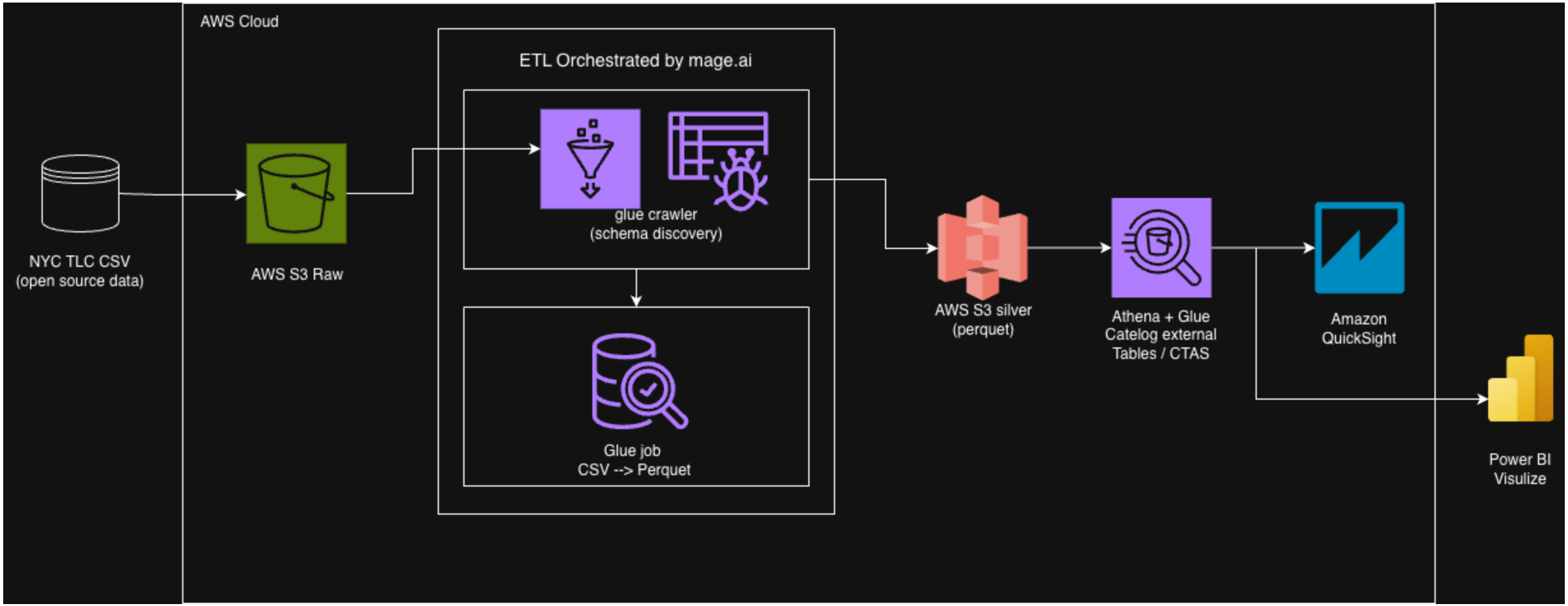
Solution overview(Design Consideration)

- Amazon S3 (Raw & Silver Layers): Raw CSV data stored in S3
- Mage.ai Orchestration: Manages the ETL flow
- AWS Glue: Automates data catalog creation and conversion from CSV to Parquet.
- Amazon Athena: Provides serverless querying on Parquet data using the Glue Data Catalog.
- Amazon QuickSight / Power BI: Used to visualize trip data, trends, and KPIs for business insights.

Solution overview(Use Cases)

- Monitor trip trends such as pickup/drop-off times, locations, and distances.
- Analyze fare amounts, tips, and payment patterns for performance metrics.
- Demonstrate an end-to-end AWS ETL pipeline integrating data ingestion, transformation, and visualization.
- Serve as a reusable framework for modern cloud-based analytics projects

Architecture diagram of the solution



Dataset overview

- Data source : https://www.nyc.gov/site/tlc/about/tlc-trip-record-data.page
- **TLC Trip Record Data (official landing page)** – monthly Yellow/Green taxi trip files with pickup/dropoff times, locations, fares, tips, distance, payment type, etc.
- **Taxi Zone Lookup (CSV)** – maps LocationID → Borough, Zone, Service Zone (used to label PULocationID/DOLocationID).
- Taxi Zone Maps and Lookup Tables

- **Taxi Zone Lookup Table** (CSV)
- **Taxi Zone Shapefile** (PARQUET)
- **Taxi Zone Map – Bronx** (JPG)
- **Taxi Zone Map – Brooklyn** (JPG)
- **Taxi Zone Map – Manhattan** (JPG)
- **Taxi Zone Map – Queens** (JPG)
- **Taxi Zone Map – Staten Island**

Measurable Outcomes

- **Business impact:** A single source of truth (GOLD tables) powers weekly Ops & Finance reporting—clear decisions on **peak hours, hotspot zones, revenue & tip mix**.
- **Performance & cost:** Converting CSV→**Parquet with partitions** cuts scanned data by **~80%+** and keeps dashboards fast (**p95 < 10s**).
- **Reliability & governance: Data-quality rules**, Glue Catalog schemas, and **Mage.ai** orchestration (with alerts) deliver consistent, on-time refreshes.