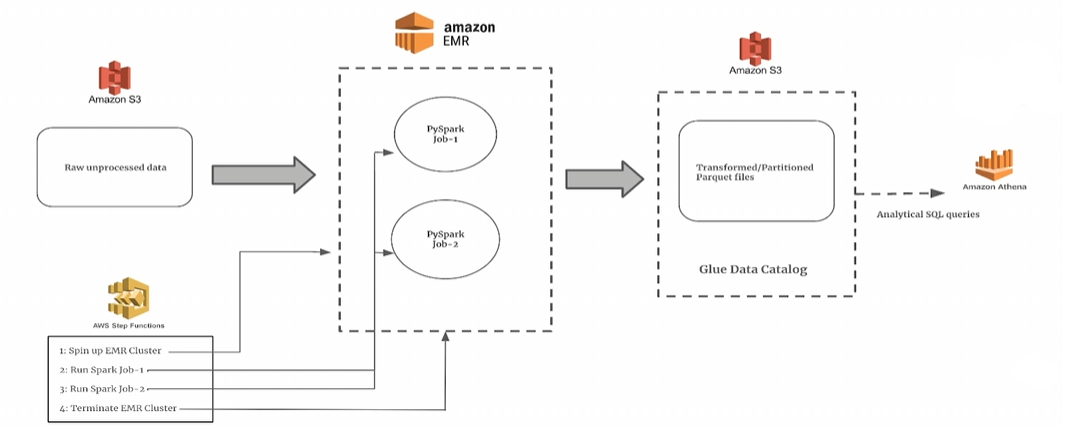
Big Data Processing of Rental Vehicles using EMR, S3 Data Lake, and Athena

# Introduction

Welcome to the structured documentation for Big Data Processing of Rental Vehicles. This project involves leveraging AWS EMR, S3, and Athena for large-scale data processing. The aim is to establish a data lake and process vehicle rental data efficiently.

# Architecture Overview

Below is the architecture diagram used for this solution:



# Technical Stack

The project utilizes the following AWS services:  
- \*\*Amazon S3\*\*: Storage layer for raw and processed data.  
- \*\*AWS EMR\*\*: Managed big data cluster for executing PySpark jobs.  
- \*\*AWS Athena\*\*: Query service to run analytical SQL queries on S3 data.  
- \*\*AWS Step Functions\*\*: Workflow automation for managing EMR clusters and jobs.

# Process Workflow

The end-to-end workflow consists of the following steps:  
1. \*\*Raw Data Storage\*\*: Data is stored in Amazon S3 in CSV format.  
2. \*\*EMR Execution\*\*: PySpark jobs process and transform data.  
3. \*\*Data Lake Formation\*\*: Processed data is stored in optimized Parquet format.  
4. \*\*Glue Data Catalog\*\*: Crawlers infer schema and store metadata.  
5. \*\*Athena Queries\*\*: SQL queries run on the processed data for analytics.

# Conclusion

This project effectively demonstrates the power of AWS services in handling large-scale data processing. Using EMR and Athena significantly reduces query processing time and cost.