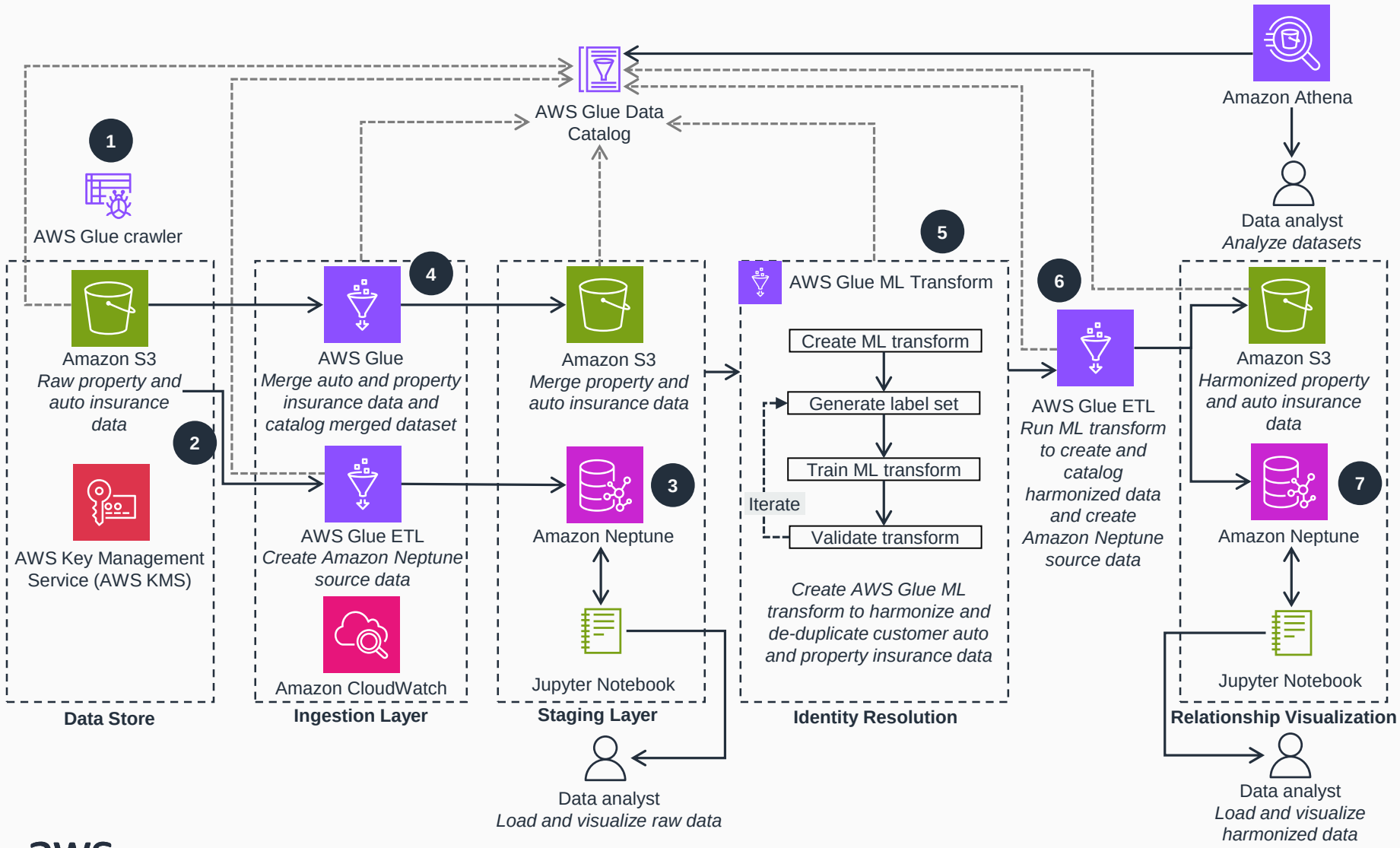
Guidance for Identifying and Resolving Duplicate Customer Records on AWS

This architecture diagram shows how to harmonize data using AWS Glue and AWS LakeFormation FindMatches ML.



1. Using an **AWS Glue crawler**, catalog the raw property and auto insurance data as tables in **AWS Glue Data Catalog**.
2. Using an **AWS Glue** extract, transform, and load (ETL) job, transform raw insurance data into a CSV format that **Amazon Neptune Bulk Loader** can accept.
3. When the data is in a CSV format, use an **Amazon SageMaker** Jupyter notebook to run a PySpark script to load the raw data into **Neptune**, and visualize it in a Jupyter notebook.
4. Run an **AWS Glue** ETL job to merge the raw property and auto insurance data into one dataset, and catalog the merged dataset. This dataset will have duplicates. No relations are built between the auto and property insurance data at this point.
5. Create and train an **AWS Glue ML transform** job to harmonize the merged data, remove duplicates, and build relations between the related data.
6. Run the **AWS Glue ML transform** job. The job also catalogs the harmonized data in the **Data Catalog** and transforms the harmonized insurance data into a CSV format that **Neptune Bulk Loader** can accept.
7. When the data is in a CSV format, use a Jupyter notebook to run a PySpark script and load the harmonized data into **Neptune**. Visualize the data in a Jupyter notebook.

Reviewed for technical accuracy February 1, 2024

AWS Reference Architecture