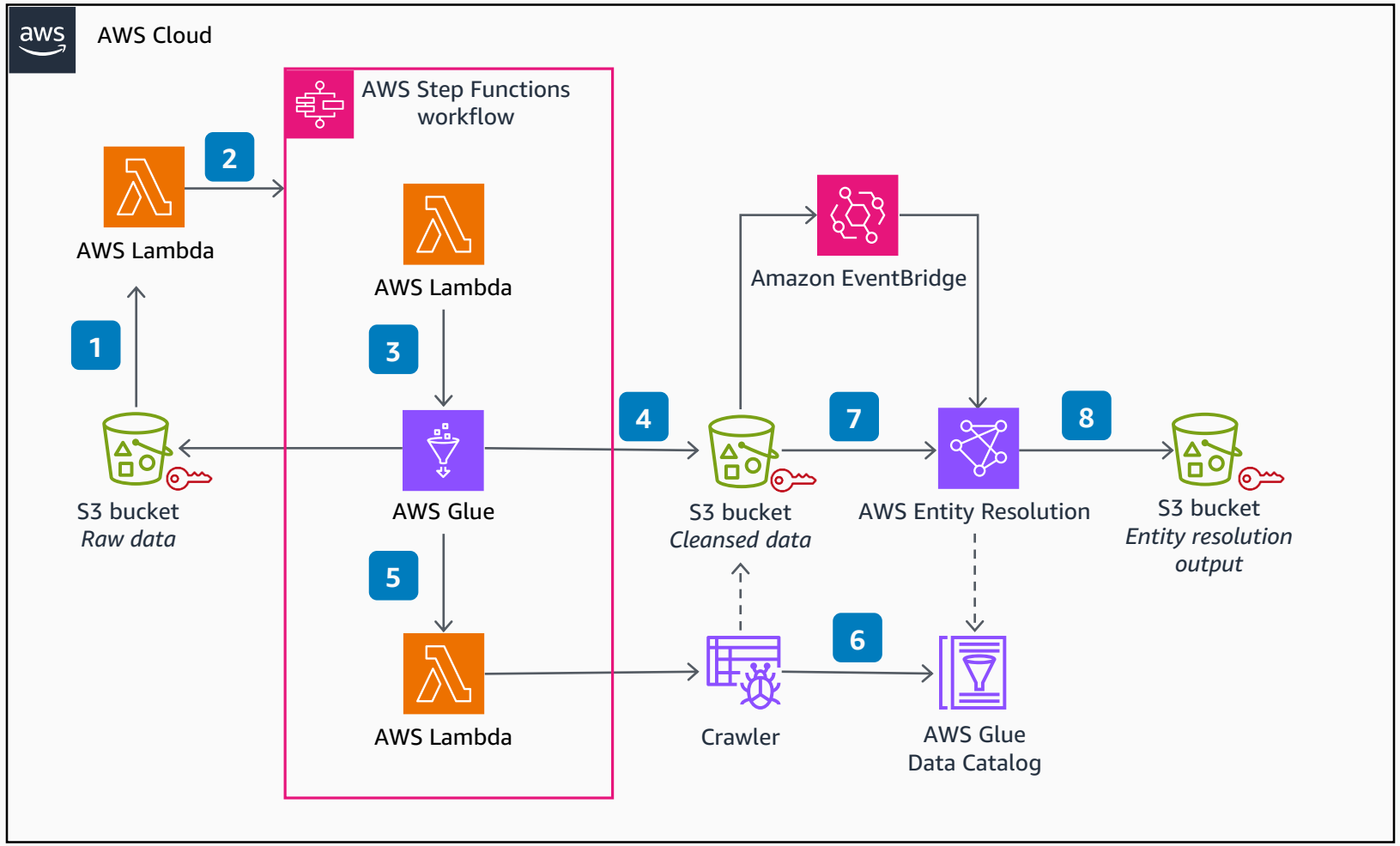
Guidance for Customizing Normalization Library for AWS Entity Resolution

This architecture shows how stored data moves through an AWS Step Functions workflow to be prepared for entity resolution.



1. The **Amazon Simple Storage Service (Amazon S3)** bucket hosts the input data source. **Server-side encryption with Amazon S3 managed keys (SSE-S3)** encrypt data in the bucket. As new data arrives in the bucket, it initiates a notification.
2. **Amazon S3** sends the notification to the **AWS Lambda** function, which then initiates an **AWS Step Functions** workflow.
3. **Lambda** within the **Step Function** initiates an **AWS Glue** job to pass information related to the new data files.
4. The **AWS Glue** job executes the normalization rules on the newly arrived data and writes the output to an **S3** bucket, encrypted using **SSE-S3**.
5. Once the **AWS Glue** job succeeds, **Lambda** invokes a crawler to crawl the recently written files to the cleansed **S3** bucket.
6. The **AWS Glue Crawler** crawls the newly written files in the **S3** bucket and updates the **AWS Glue Data Catalog** table with updated information related to schema and partition, if applicable.
7. As new data gets written into the **S3** bucket with cleansed data, **Amazon EventBridge** notifies **AWS Entity Resolution** that there is new data available for entity resolution.
8. **AWS Entity Resolution** performs the entity resolution on the data and writes the corresponding output to an **S3** bucket, encrypted using SSE-S3.

Reviewed for technical accuracy July 10, 2023

AWS Reference Architecture