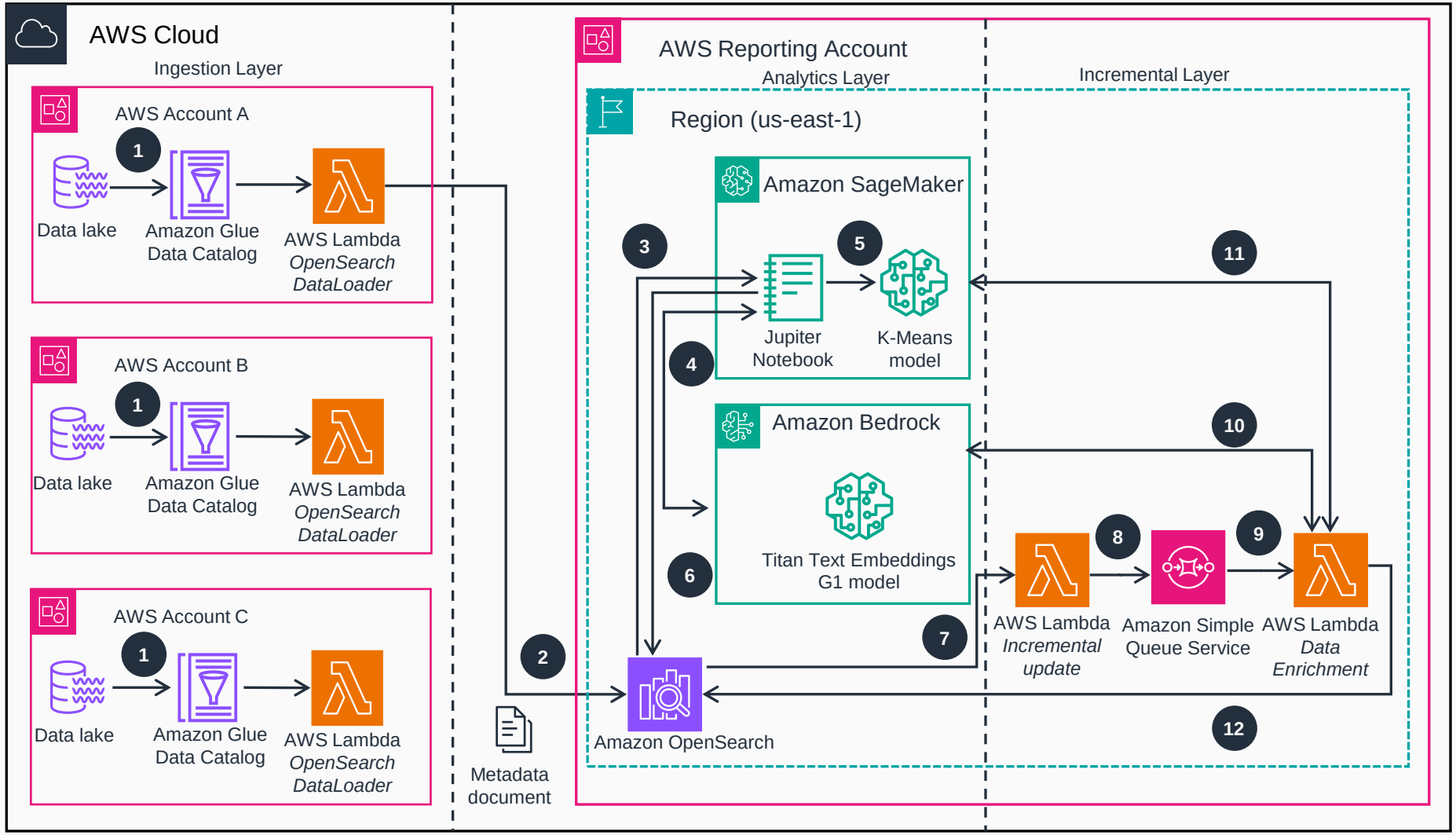
Guidance for Deduplicating Syndicated Data on AWS

This architecture diagram shows how to obtain an aggregated view of similar tables across multiple AWS accounts within an AWS Organization.



1. Each customer account that contributes data (Account A,B, and C) must have at least one table in the AWS Glue Data Catalog using an **AWS Glue** crawler or manual processes.
2. A scheduled **AWS Lambda** function, the *DataLoader*, will execute in external accounts at a fixed time every day. It extracts table information from the Data Catalog, transforms it to a metadata document, and sends it to **Amazon OpenSearch** in the reporting account located in the us-east-1 AWS Region.
3. Raw data collected from all accounts and Regions is read from **OpenSearch**.
4. Vector embeddings for each table are generated using the **Amazon Bedrock** Titan Text Embeddings G1 model based on the column names.
5. A K-Means algorithm model will be trained using **Amazon SageMaker,** and then deployed to the Amazon SageMaker Serverless Inference endpoint.
6. Each table will be labeled using the deployed K-Means model, and the vector-enriched dataset will be stored back in **OpenSearch**.
7. A **Lambda** function responsible for incremental updates will execute daily to query the raw metadata documents and identify the changes (delta) since the last data enrichment.
8. The delta identified in the raw metadata documents will be sent to **Amazon Simple Queue Service** (Amazon SQS).
9. The **Amazon SQS** service will trigger the *Data Enrichment* **Lambda** function.
10. Vector embeddings will be generated for each table based on the column names using the Titan Text Embeddings G1 model.
11. Each table is labeled using the deployed K-Means model.
12. The vector-enriched dataset will be stored back in **OpenSearch**.

Reviewed for technical accuracy July 30, 2024

AWS Reference Architecture