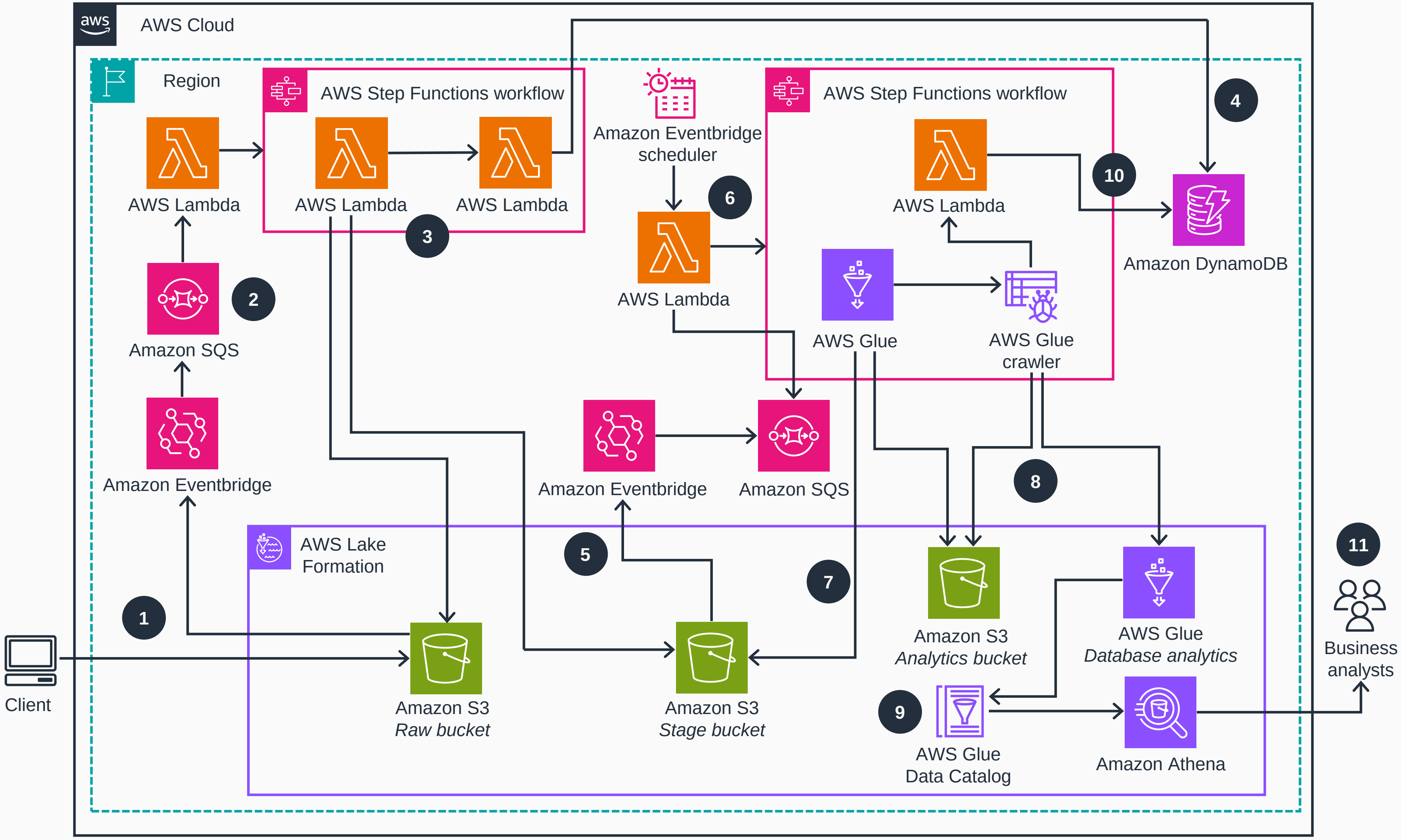
Guidance for Data Lakes on AWS

This architecture diagram shows how to build a data lake on AWS in addition to demonstrating how to process, store, and consume data using serverless AWS analytics services.



1 The data administrator uploads JSON files in the **Amazon Simple Storage Service (Amazon S3)** raw bucket. Object creation in **Amazon S3** triggers an event in **Amazon EventBridge**.

2 **EventBridge** has a rule that sends a message in **Amazon Simple Queue Service (Amazon SQS)**, which invokes an **AWS Lambda** function.

3 The **Lambda** function triggers the **AWS Step Functions** workflow, in which another **Lambda** function reads files from the **S3** raw bucket and performs transformation. It also writes the new set of JSON files in the **S3** stage bucket.

4 A **Lambda** function updates the **Amazon DynamoDB** table with the **Step Functions** job status.

5 Once the files are created in the **S3** stage bucket, it triggers an event in **EventBridge**, which has a rule that sends a message in **Amazon SQS** with created file details.

6 The **Eventbridge** scheduler runs at certain intervals and invokes a **Lambda** function that retrieves messages from **Amazon SQS** and starts another **Step Functions** workflow.

7 **AWS Glue** extract, transform, load (ETL) reads the data from the **AWS Glue** database stage, then converts the files from JSON to Parquet format.

8 **AWS Glue** ETL writes the Parquet files in the **S3** analytics bucket. **AWS Glue** crawler crawls the Parquet files in the same bucket and then creates analytics tables in **AWS Glue** database analytics.

9 All the staging and analytics catalogs are maintained in the **AWS Glue** Data Catalog.

10 A **Lambda** function updates the **DynamoDB** table with the **Step Functions** job status.

11 Business analysts use **Amazon Athena** to query the **AWS Glue** database analytics.

Reviewed for technical accuracy December 26, 2024

AWS Reference Architecture