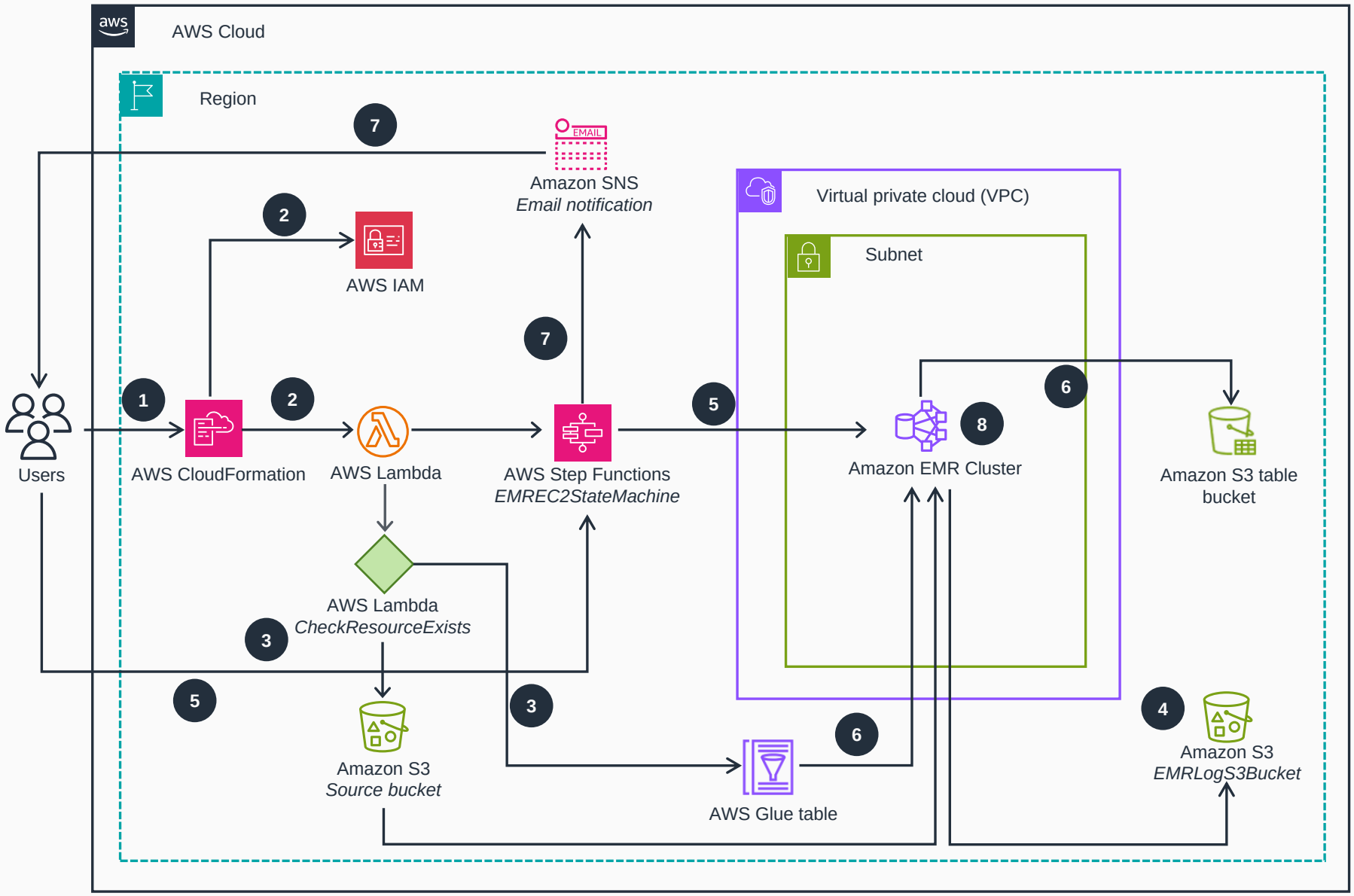
Guidance for Migrating Tabular Data from Amazon S3 to S3 Tables

This architecture diagram illustrates the migration of Amazon S3 Iceberg tables using a Create Table As Select (CTAS) approach. CTAS enables the creation of new tables based on the data from existing tables.



1. The user deploys this solution using **AWS CloudFormation** by creating a stack through the **AWS Management Console**.
2. **CloudFormation** deploys resources including **AWS Lambda**, **AWS Identity and Access Management** (IAM), custom resources, **AWS Step Functions**, and a PySpark Script.
3. The *CheckResourceExists* **Lambda** function checks for the existence of a source **Amazon Simple Storage Service** (Amazon S3) bucket and the **AWS Glue** table for migration.
4. The *EMRLogS3Bucket* **Amazon S3** bucket is created by **CloudFormation** to store the **Amazon EMR** cluster logs, as well as the PySpark script for the Apache Spark on **Amazon EMR** jobs.
5. The *EMREC2StateMachine* **Step Functions** task is manually invoked by the user to orchestrate the creation of an **Amazon EMR** cluster and the execution of an Apache Spark job.
6. The Apache Spark jobs running on the **Amazon EMR** cluster use the Create Table As Select (CTAS) functionality to migrate data from the source **AWS Glue** table and source **Amazon S3** bucket to the target **Amazon S3 table** bucket.
7. Upon completion of the migration workflow, the *EMREC2StateMachine* **Step Functions** task sends a notification email to the user by **Amazon Simple Notification Service** (Amazon SNS).
8. The **Amazon EMR** cluster is terminated by the *EMREC2StateMachine* **Step Functions** task.