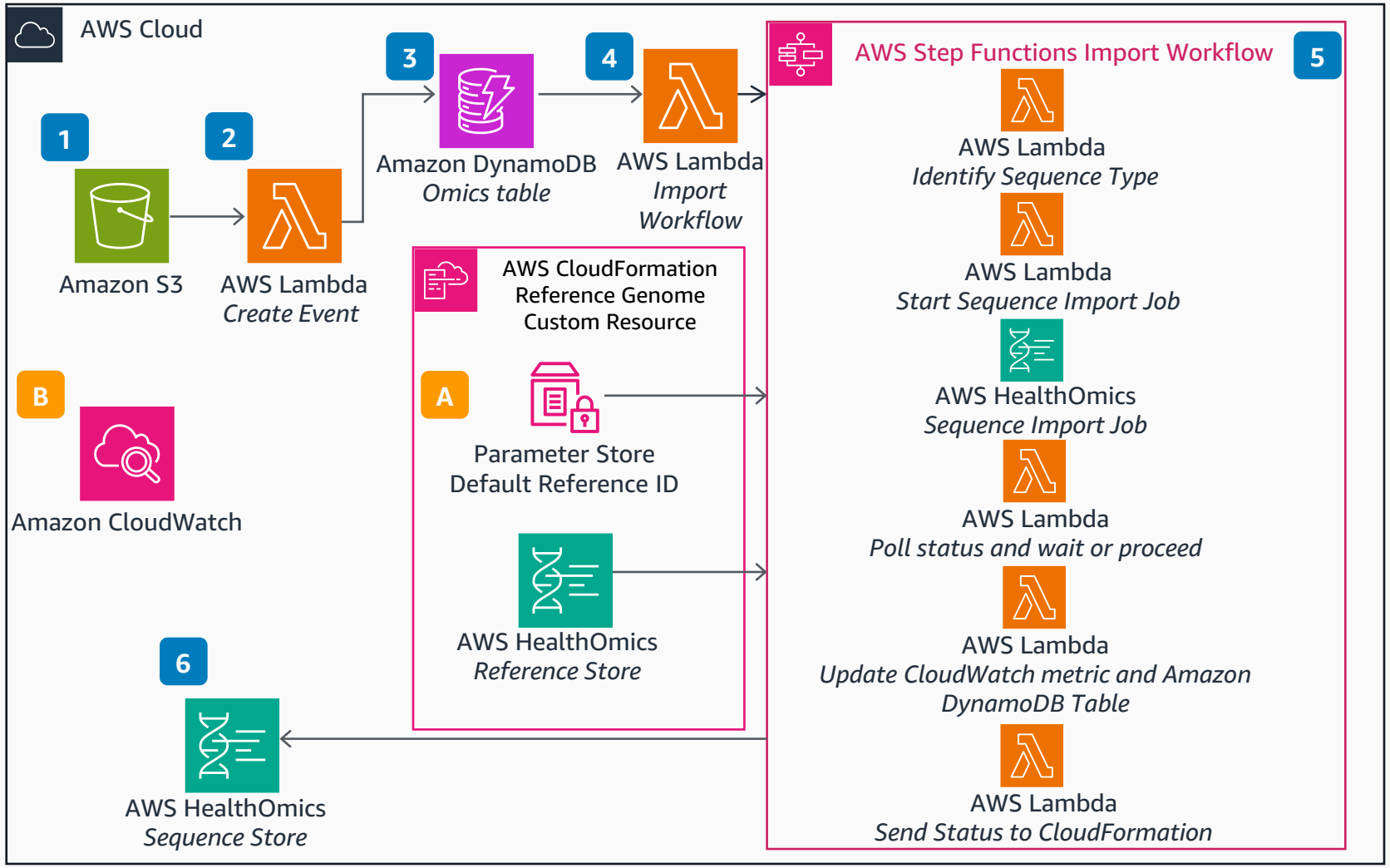
Guidance for Migration & Storage of Sequence Data with AWS HealthOmics

This architecture diagram shows how to automatically import omics sequence data from Amazon S3 into AWS HealthOmics Storage. It uses AWS Lambda for compute, Amazon DynamoDB for metadata management, and AWS Step Functions to orchestrate.



AWS Reference Architecture

1. If you have already followed the directions in How to move and store your genomics sequencing data with AWS DataSync, you will have a pre-existing **Amazon Simple Storage Service** (Amazon S3) bucket. If you do not have an **Amazon S3** bucket, you can create one using either the **AWS Management Console** or AWS Command Line Interface (AWS CLI).

2. The **Amazon S3** Object Created Event invokes an **AWS Lambda** function to create a record in the **Amazon DynamoDB** table.

3. Creation of a record in the Auto Load Omics Table creates an item in a **DynamoDB** stream.

4. The **DynamoDB** stream event invokes **Lambda**, which starts the sequence import workflow.

5. **AWS Step Functions** workflow using multiple **Lambda** functions and native **Step Functions** tasks is initiated to import data. Detailed workflow is located in the code repository.

6. The original sequence is loaded into **AWS HealthOmics** Storage.

A. **Custom Resource:** The sequence import requires a reference genome in the **HealthOmics** Reference store. This Guidance uses an additional **AWS Cloud Development Kit** (AWS CDK) construct that creates a reference and adds the acquirer reference number (ARN) for that reference as a parameter in AWS Systems Manager Parameter Store.

B. **Custom Metric:** Success or failure of the **HealthOmics** import job is recorded as a custom metric in **Amazon CloudWatch**. This allows detailed monitoring of imported statistics.

Reviewed for technical accuracy October 16, 2023