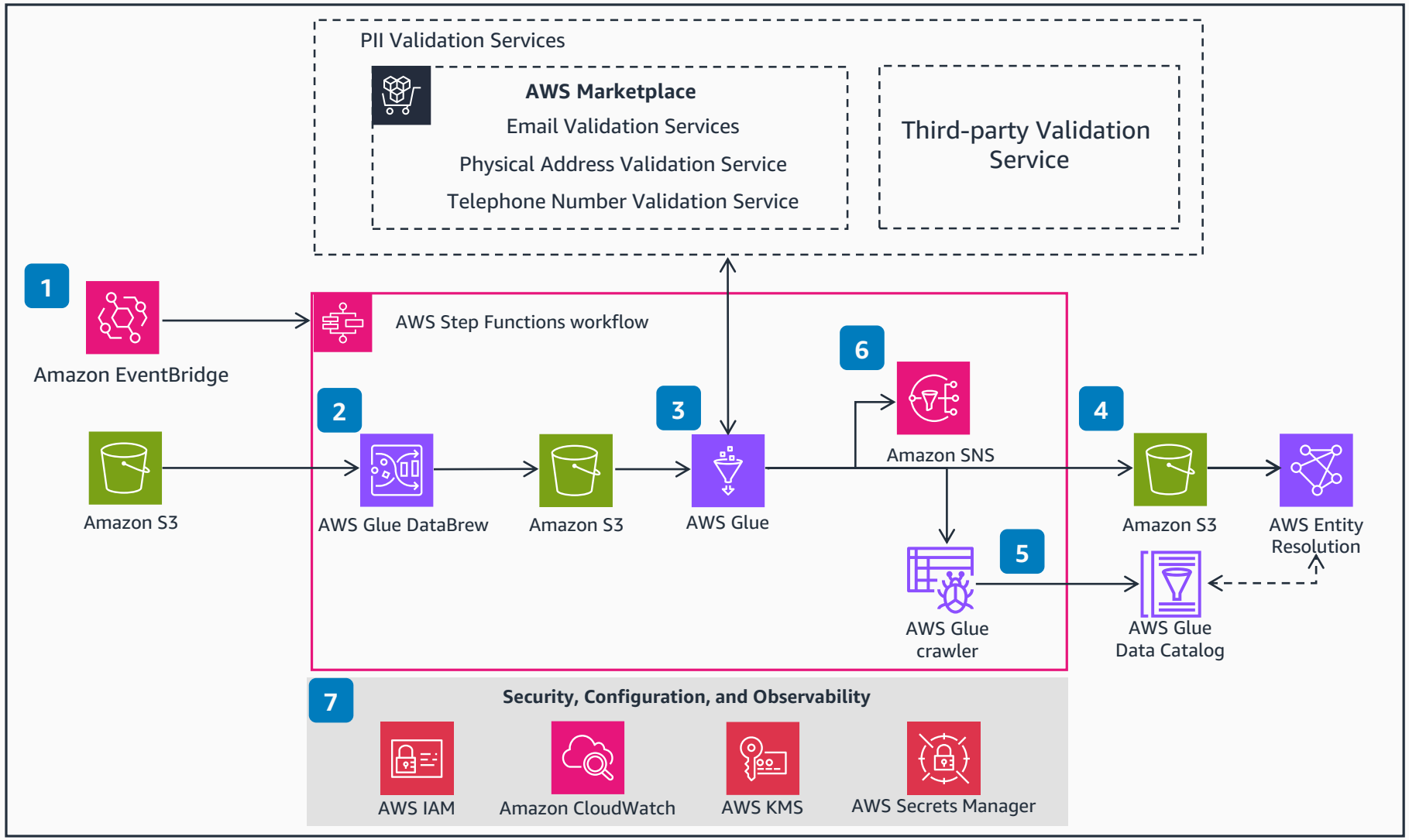
Guidance for Preparing and Validating Records for AWS Entity Resolution

This Guidance provides an overview of the architecture and services required to prepare and validate data, including physical address, phone, and email for AWS Entity Resolution.



1. Create a rule in **Amazon EventBridge** to schedule the data processing in **AWS Step Functions**. The **Step Functions** state machine includes data cleaning and validation steps.

2. Use **AWS Glue DataBrew** recipe to transform the data from the source **Amazon Simple Storage Service** (Amazon S3) location. Use this step to normalize data, which will give better outcomes with data validation API.

3. Use **AWS Glue** to read the output of the **DataBrew** job, enabling the invocation of the respective personally identifiable information (PII) entity validation services in small batches.

4. **AWS Glue** writes the validated data to the target curated **Amazon S3** bucket for **AWS Entity Resolution** to consume.

5. An **AWS Glue** crawler job is initiated to "refresh" the table definition or metadata of the curated **Amazon S3** storage location, and stores it in the AWS Glue Data Catalog.

6. An event is published to **Amazon Simple Notification Service** (Amazon SNS) to inform the user that the new curated data files are now available for consumption.

7. **Security, Configuration, and Observability**
This Guidance uses the following AWS services to promote security and access control:
   - **AWS Identity and Access Management** (IAM): Least-privilege access to specific resources and operations.
   - **AWS Key Management Service** (AWS KMS): Provides encryption for data at rest and data in transit, using Pretty Good Privacy (PGP) encryption of data files.
   - **AWS Secrets Manager**: Provides hashing keys for PII data.
   - **Amazon CloudWatch**: Monitors logs and metrics across all services used in this Guidance.

Reviewed for technical accuracy July 10, 2023

© 2023, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS Reference Architecture