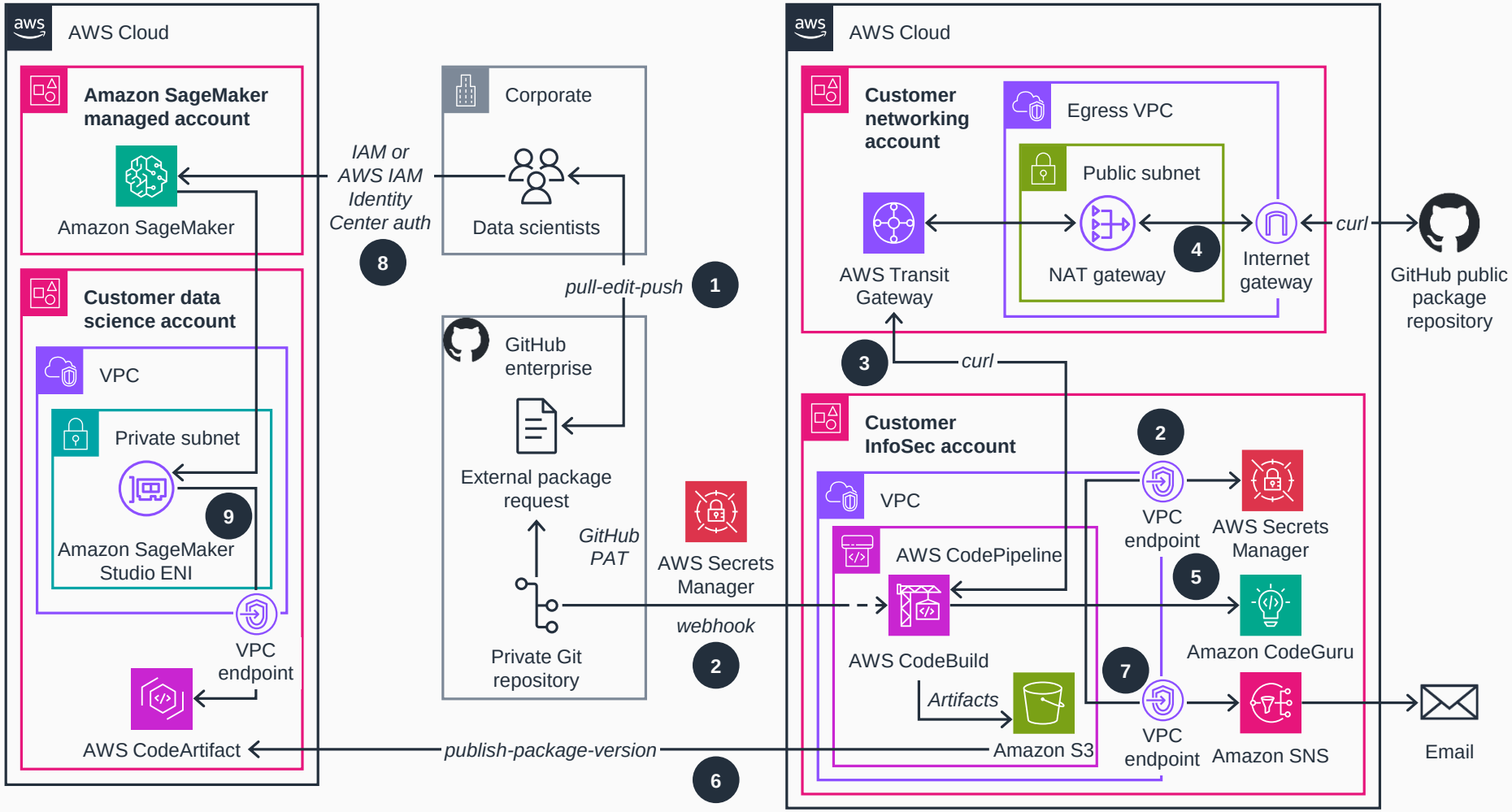
Guidance for Secure Access to External Package Repositories on AWS

This architecture diagram helps streamline the process for data scientists to access external package repositories while maintaining InfoSec compliance.



1. Your data scientist pulls the external package dependency request file from a GitHub Enterprise private repository, appends the external package repository names and ZIP URLs, and then pushes the updated request file into the private repository.
2. The request file check-in invokes an **AWS CodePipeline** orchestration, which is secured by a personal access token (PAT) in **AWS Secrets Manager** and accessed using an **Amazon Virtual Private Cloud** (Amazon VPC) endpoint.
3. The **CodePipeline** build stage includes an **AWS CodeBuild** project that parses the request file and downloads the external package repositories. The external packages are stored as a build-stage output artifacts in **Amazon Simple Storage Service (Amazon S3)**.
4. Centralized inbound and outbound internet traffic occurs through a NAT gateway attached to the output virtual private cloud (VPC) in your networking account.
5. **Amazon CodeGuru** Security performs security scans on the downloaded external package repositories.
6. If the security scans return lower than medium severity, the build stage creates a new private **AWS CodeArtifact** package version asset in your data science account.
7. **Amazon Simple Notification Service (Amazon SNS)** emails the results, positive or negative, to your requesting data scientist.
8. Your data scientist authenticates to the **Amazon SageMaker** Studio domain through the **AWS Identity and Access Management** (IAM) or **AWS IAM Identity Center** mode. A **SageMaker** Studio notebook installs the InfoSec-validated external packages using the corresponding private **CodeArtifact** package version assets (for example, aws codeartifact get-package-version-asset).
9. A **SageMaker** Studio elastic network interface (ENI) deployed in the VPC that you manage uses the VPC endpoint for private network access to **CodeArtifact**.

Reviewed for technical accuracy May 21, 2024

AWS Reference Architecture