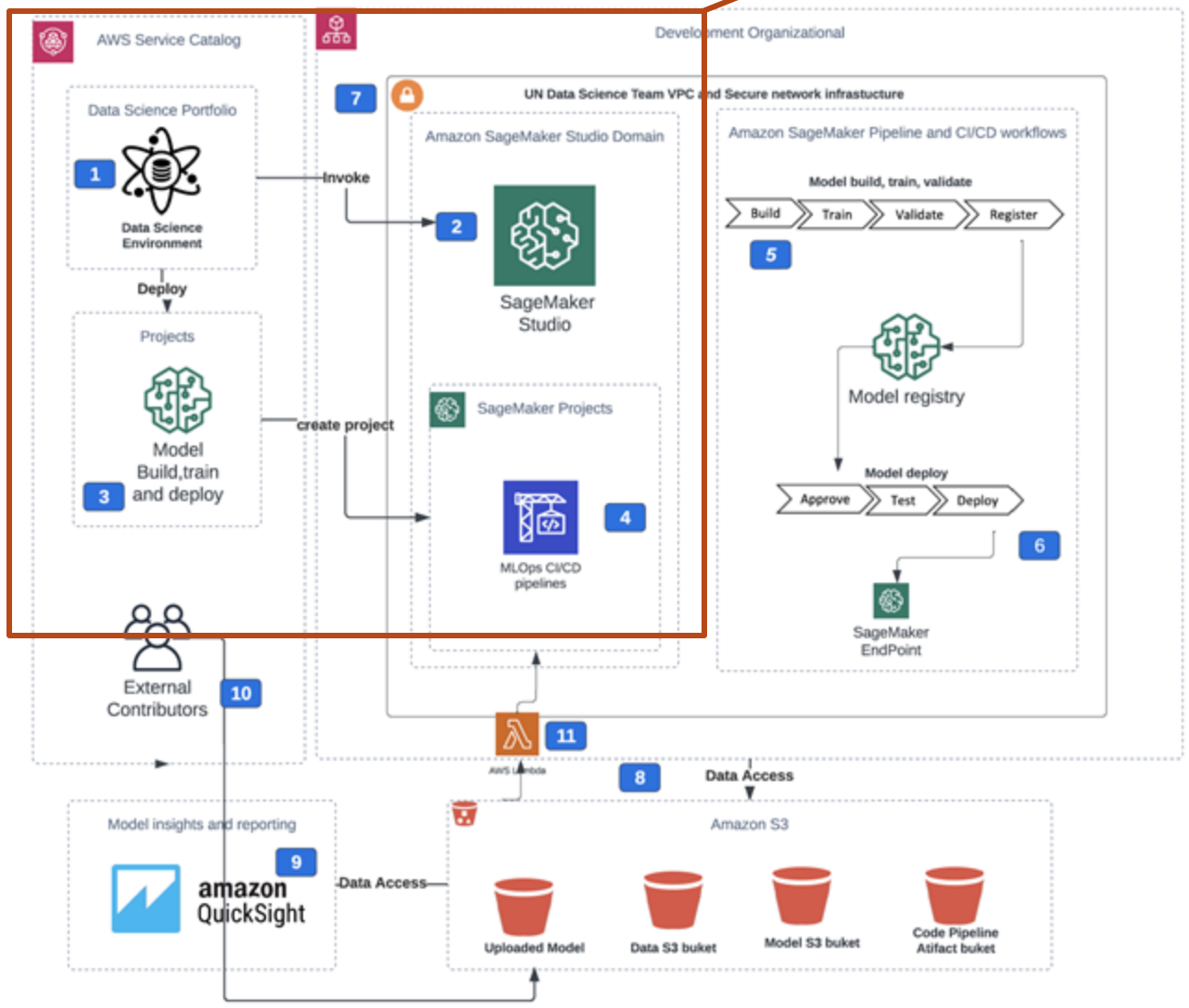
Project Resilience Architecture

Prem Krishnamurthy
Senior AI/ML Architect at Vanguard Groups

Architecture



Component 1: AWS Service Catalog

The end-to-end deployment of the data science environment is provided as a self-provisioned AWS Service Catalog offering. One of the primary benefits of using the AWS Service Catalog for self provisioning is that authorized users can configure and deploy available goods and AWS resources without requiring full privileges or access to AWS services. All AWS Service Catalog products are deployed under a service role with a defined set of permissions that are independent of the user's permissions.

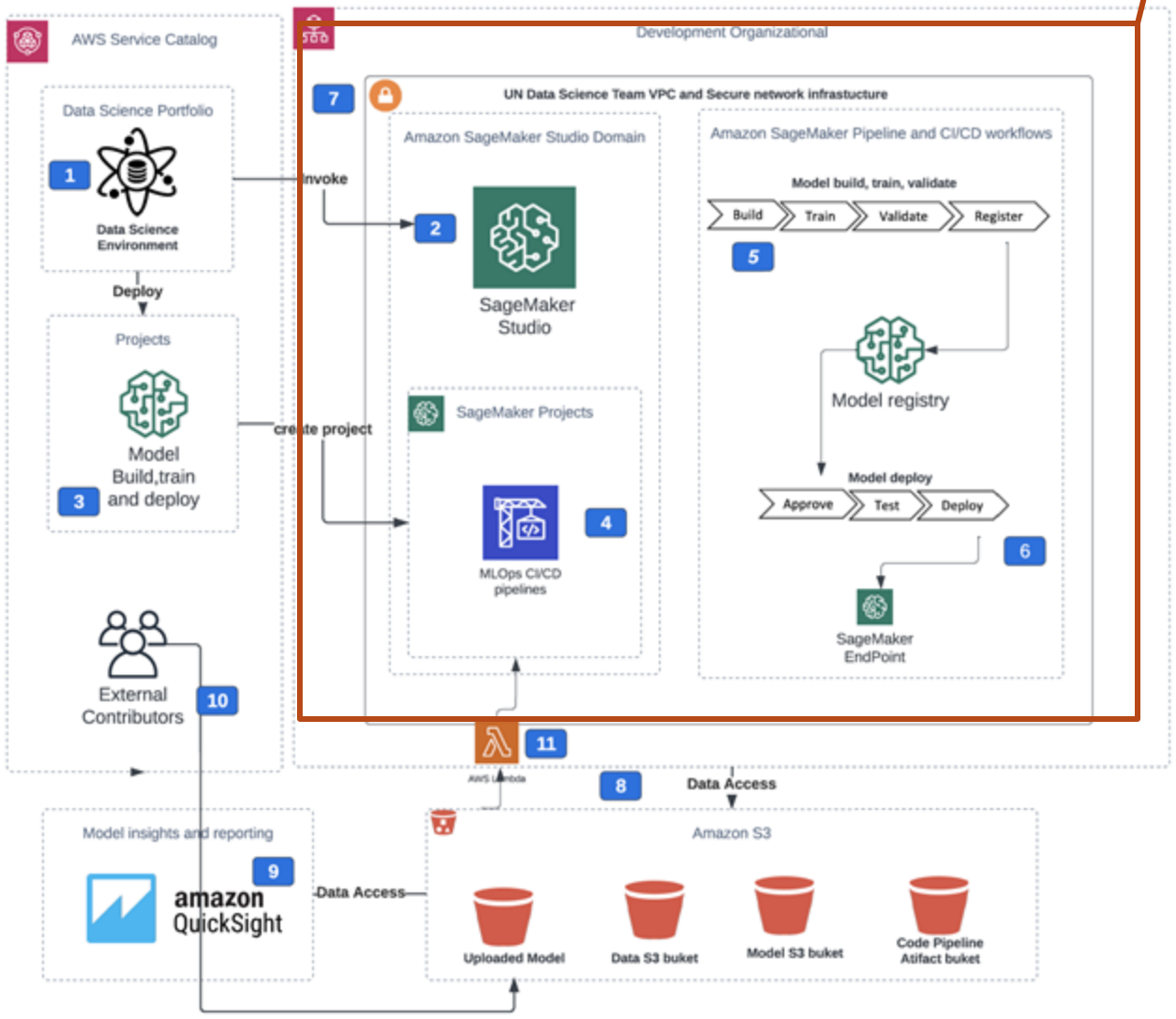
Component 2: Studio domain

The AWS ServiceCatalog product Data Science Environment provides a Studio domain. A Studio domain includes an approved user list, configuration information, an Amazon Elastic File System (Amazon EFS) volume, and S3 connections. The Amazon EFS volume or S3 stores user data, such as notes, resources, and artifacts.

Components 3 and 4: SageMaker MLOps project templates

The solution provides customized SageMaker MLOps project template versions. Each MLOps template provides an automated model development and deployment workflow based on continuous integration and delivery (CI/CD). The provided templates are configured for the secure deployment of multi-account models and are completely integrated into the data science environment. The Studio project templates are made available via the AWS Service Catalog. The templates include the seed code repository with Studio notebooks, which offers a secure configuration for SageMaker workloads including processing, training jobs, and pipelines.

Architecture



Components 5 and 6: CI/CD workflows

The MLOps projects employ Pipelines and AWS CodePipeline, AWS CodeCommit, and AWS CodeBuild to implement CI/CD. Additionally, SageMaker project templates allow a CI/CD workflow.

Pipelines is responsible for orchestrating workflows throughout each step of the ML process and automating tasks, such as data loading, data transformation, training, tuning and validation, and deployment. Each model is monitored by SageMaker Model Registry, which contains model metadata, including training and validation metrics and data lineage, as well as model versions and the model's approval status.

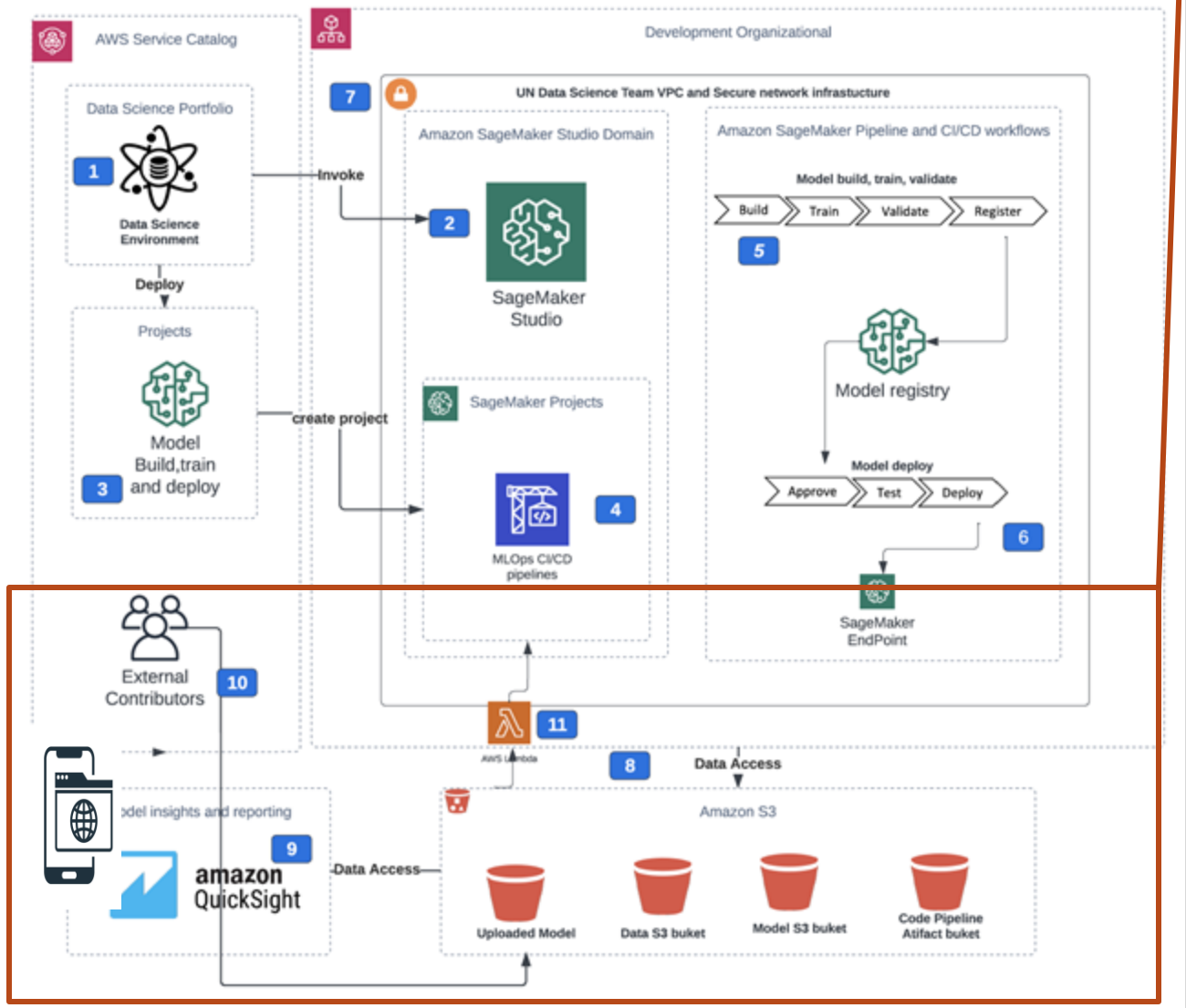
CodePipeline distributes the model to the accounts with staging and production environments specified as targets. During infrastructure building, CloudFormation templates pre-create the necessary resources. This solution facilitates the secure deployment of multi-account models via AWS Organizations

Component 7: Secure infrastructure

Studio is deployed within a dedicated VPC. Each elastic network interface utilized by a SageMaker domain or workload is generated within a private, dedicated subnet and is associated with the defined security groups. Using an optional NAT gateway, the data science environment VPC can be configured with internet access.

You can also operate this VPC offline, without inbound or outbound internet access. AWS PrivateLink is used for all access to AWS's public services. Traffic between your VPC and AWS services is not accessible to the public internet and never leaves the Amazon network.

Architecture



Component 8: Data Lake security

All data in the data science environment, which is stored in Using customer-managed CMKs, all data stored in Amazon Simple Storage Service (Amazon S3) buckets and Amazon Elastic Block Storage (Amazon EBS) and EFS volumes is encrypted at rest. Transport Layer Security (TLS) version 1.2 is used to protect all data transfers between platform components, API calls, and inter-container communication.

The combination of the S3 bucket and user policies and S3 VPC endpoint policy governs data access from Studio notebooks or any SageMaker workload to the environment's S3 buckets.

Component 9: Amazon QuickSight and Web App

Amazon QuickSight and WebApp enables all members of your organization to comprehend your data by asking questions in natural language, navigating via interactive dashboards, or automatically searching for patterns and outliers using machine learning.

Component 10: Bring your Own Model

The pre-trained machine learning model is uploaded your model to the "Upload Amazon S3 bucket", compiles the model for the SageMaker pipeline, and packages your model so that it can be deployed as the SageMaker interference.