
AWS
re:Invent

ARC312

Architectural best practices for machine learning applications

Mohsen Ansari
Solutions Builder
AWS

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Outline

Taking applications to production

Taking intelligent applications to production

Well-architecting an intelligent application

Building intelligent applications on AWS



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Taking applications to production



Write

Write source code for your software



Build

Build / compile the software



Test

Test to make sure it works as expected



Deploy

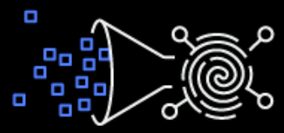
Deploy the software to servers



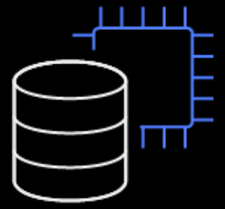
Use

Use the software in production

Taking intelligent applications to production



Data ingestion



Data preparation



Feature engineering



Model training



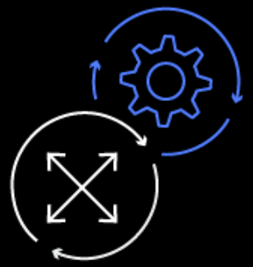
Model evaluation & tuning



Model deployment



inference



Model monitoring / re-training

The AWS ML stack

AI SERVICES

VISION	SPEECH		TEXT			SEARCH	CHATBOTS	PERSONALIZATION	FORECASTING	FRAUD	DEVELOPMENT	CONTACT CENTERS
Amazon Rekognition	Amazon Polly	Amazon Transcribe *+Medical*	Amazon Comprehend *+Medical*	Amazon Translate	Amazon Textract	Amazon Kendra	Amazon Lex	Amazon Personalize	Amazon Forecast	Amazon Fraud Detector	Amazon CodeGuru	Contact Lens *For Amazon Connect*

ML SERVICES

Amazon SageMaker

Ground Truth

AWS Marketplace for ML

SageMaker Studio IDE

Built-in algorithms	Notebooks	Experiments	Processing	Model training and tuning	Debugger	Autopilot	Model hosting	Model monitor

Neo

Augmented AI

ML FRAMEWORKS & INFRASTRUCTURE

TensorFlow, mxnet, PYTORCH, GLUON, Keras, DeepGraphLibrary

Deep Learning AMIs & Containers

GPUs & CPUs

Elastic Inference

Inferentia

FPGA

TensorFlow, the TensorFlow logo and any related marks are trademarks of Google Inc.

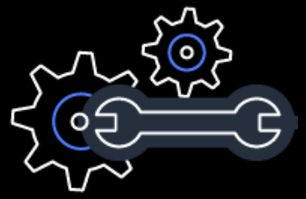
Well-architecting an intelligent application



AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Well-Architected Framework



Operational excellence



Security



Reliability



Performance efficiency



Cost optimization

https://aws.amazon.com/architecture/well-architected

ML Lens for Well-Architected Framework

Well-Architected pillar	Key Best Practices for ML
Cost Optimization	Use managed services to reduce cost
Operational Excellence	Version ML inputs and artifacts Automate machine learning deployment pipelines
Reliability	Train once and deploy across environments
Security	Restrict access to ML services
Performance Efficiency	Continuously monitor and measure system performance

https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens/welcome.html

Building intelligent applications on AWS



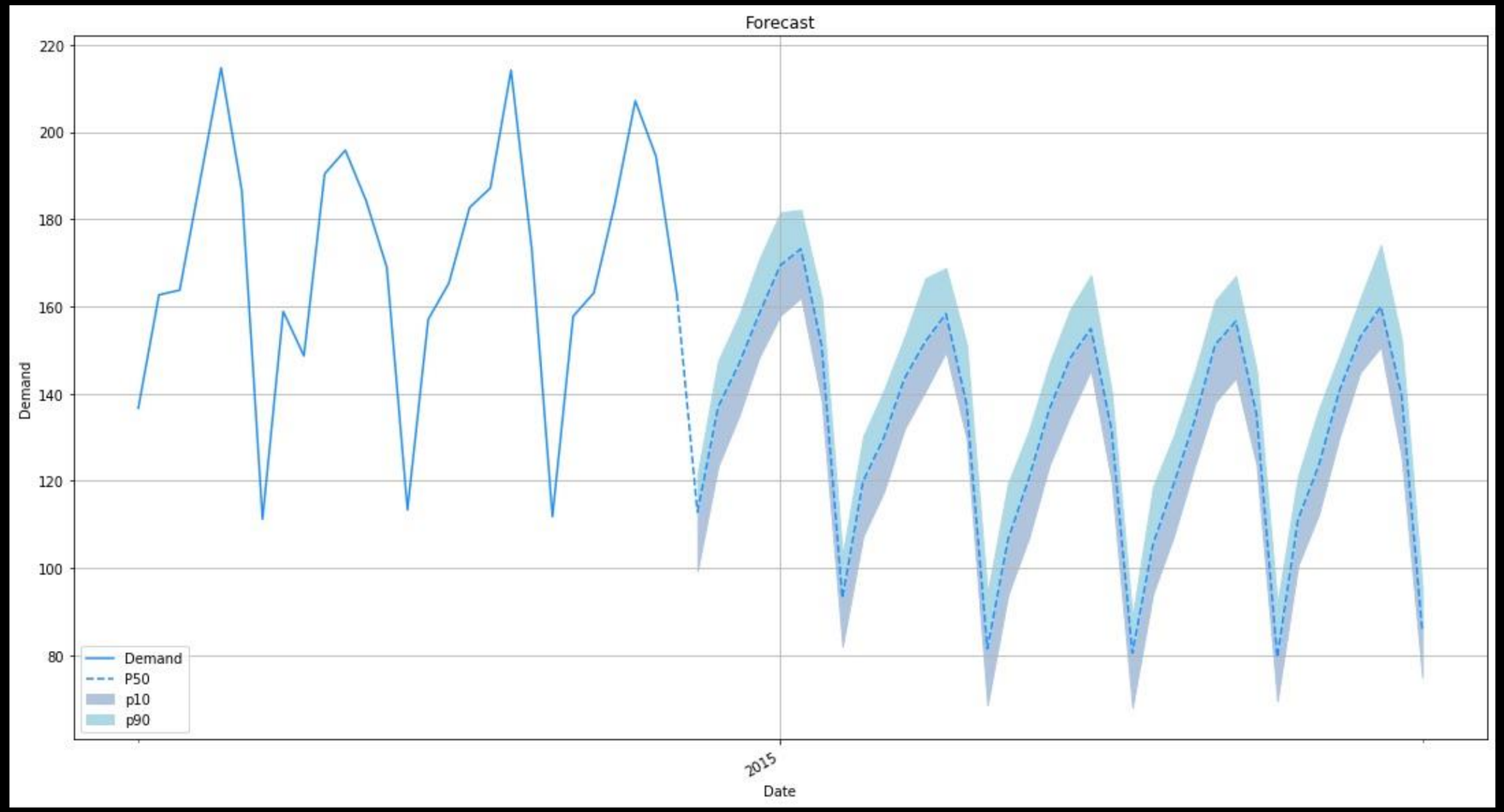
AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Intelligent applications with AWS AI services

Improving forecast accuracy with machine learning

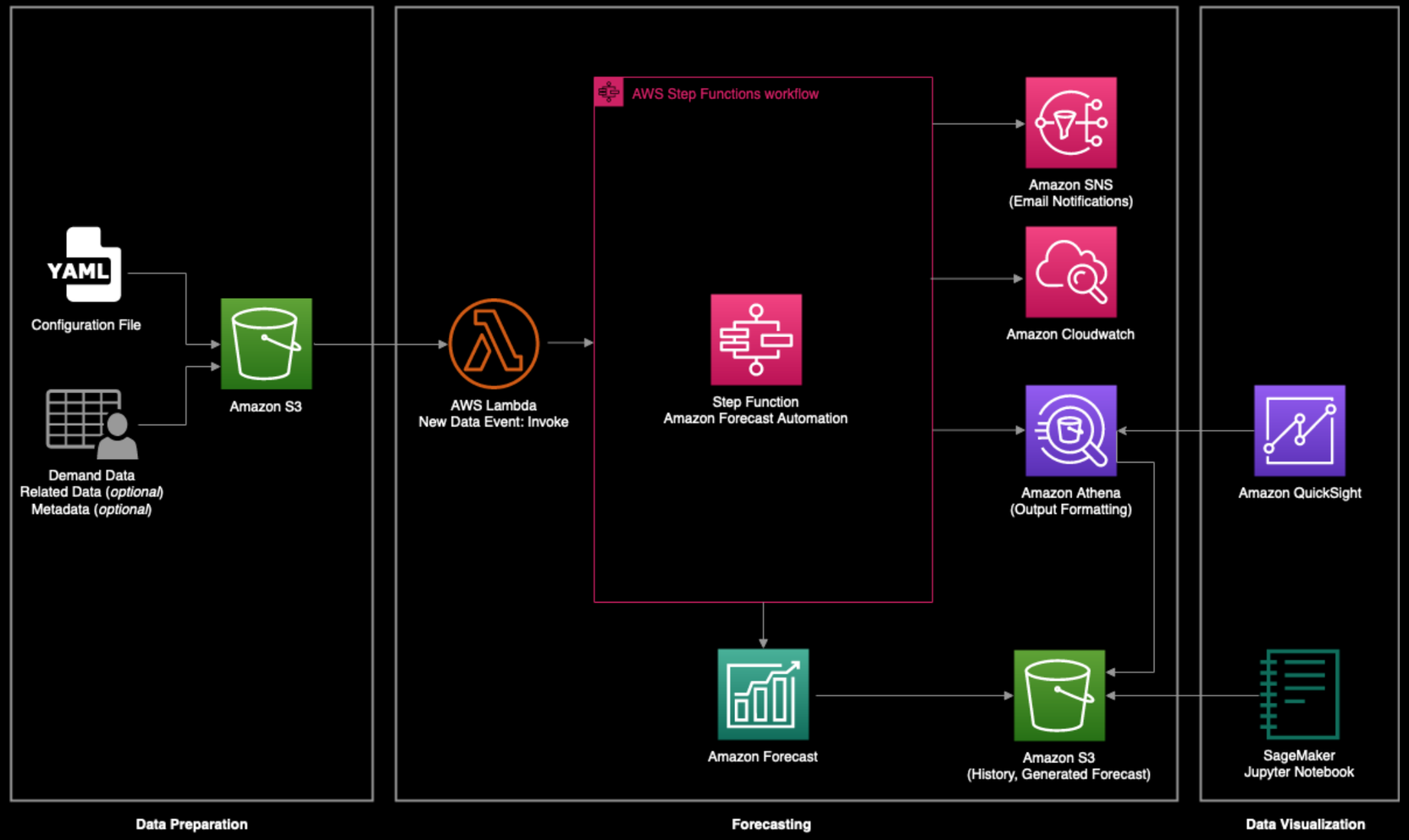
App8



https://aws.amazon.com/solutions/implementations/improving-forecast-accuracy-with-machine-learning/

Intelligent applications with AWS AI services

Improving forecast accuracy with machine learning



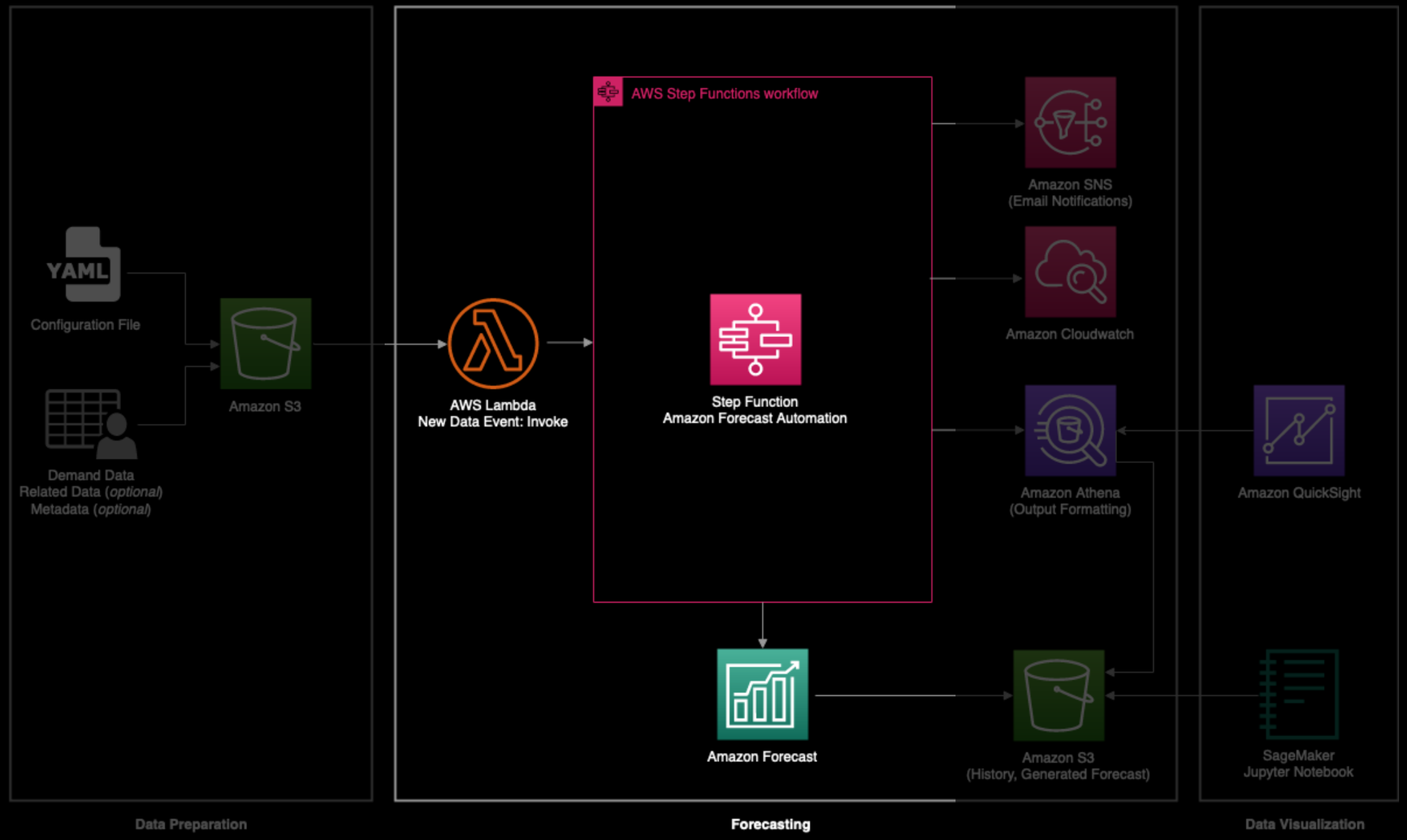
https://aws.amazon.com/solutions/implementations/improving-forecast-accuracy-with-machine-learning

ML Lens for Well-Architected Framework

Well-Architected Pillar	Key Best Practices for ML
Cost Optimization	Use managed services to reduce cost
Operational Excellence	Version ML inputs and artifacts Automate machine learning deployment pipelines
Reliability	Train once and deploy across environments
Security	Restrict access to ML services
Performance Efficiency	Continuously monitor and measure system performance

https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens

Cost optimization using managed services



https://aws.amazon.com/solutions/implementations/improving-forecast-accuracy-with-machine-learning

Cost optimization

- Use managed services to reduce cost of ownership
- Experiment with small datasets
- Right-size training and model hosting instances
- Account for inference architecture based on consumption patterns
- Define overall ROI and opportunity cost

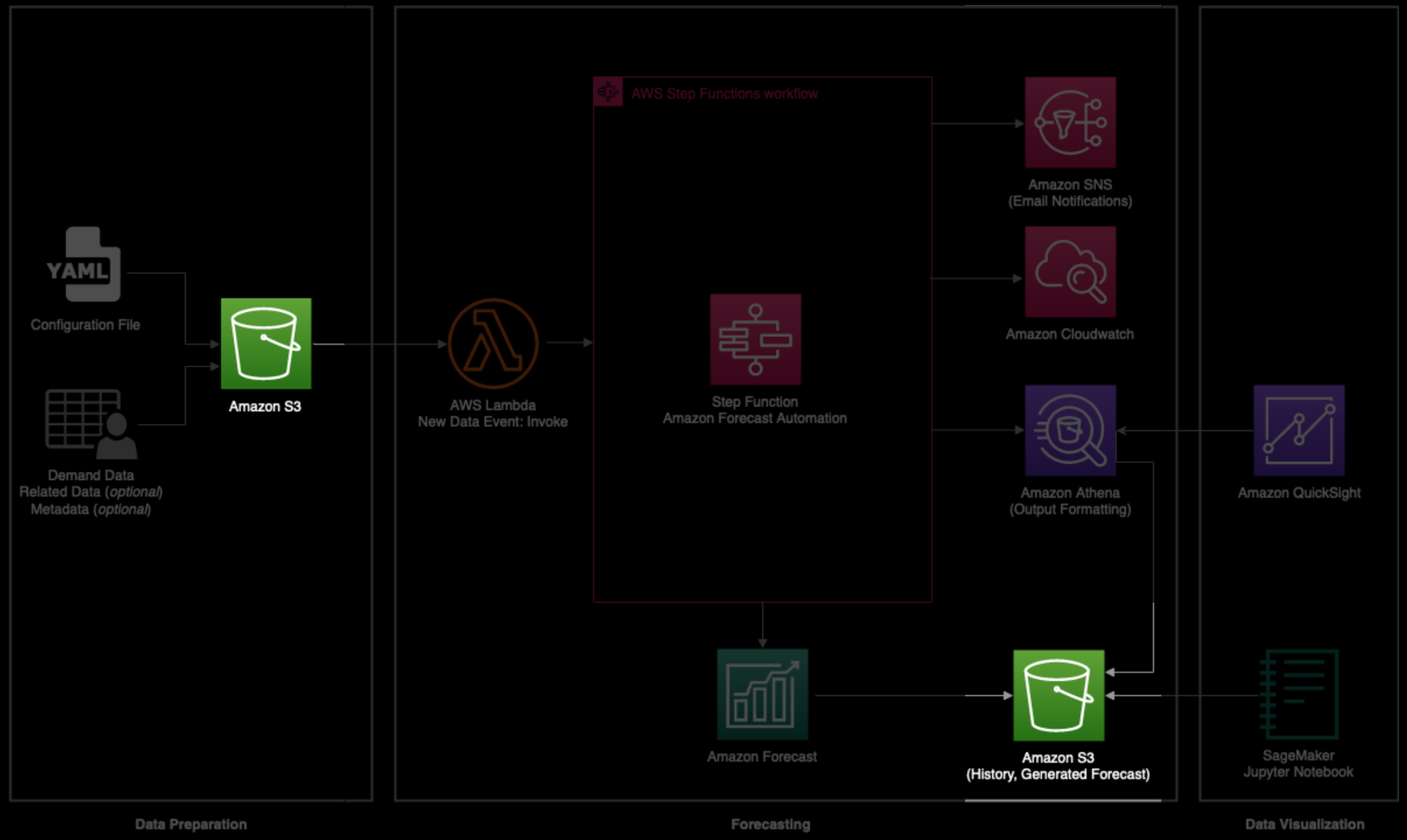


ML Lens for Well-Architected Framework

Well-Architected Pillar	Key Best Practices for ML
Cost Optimization	Use managed services to reduce cost
Operational Excellence	Version ML inputs and artifacts Automate machine learning deployment pipelines
Reliability	Train once and deploy across environments
Security	Restrict access to ML services
Performance Efficiency	Continuously monitor and measure system performance

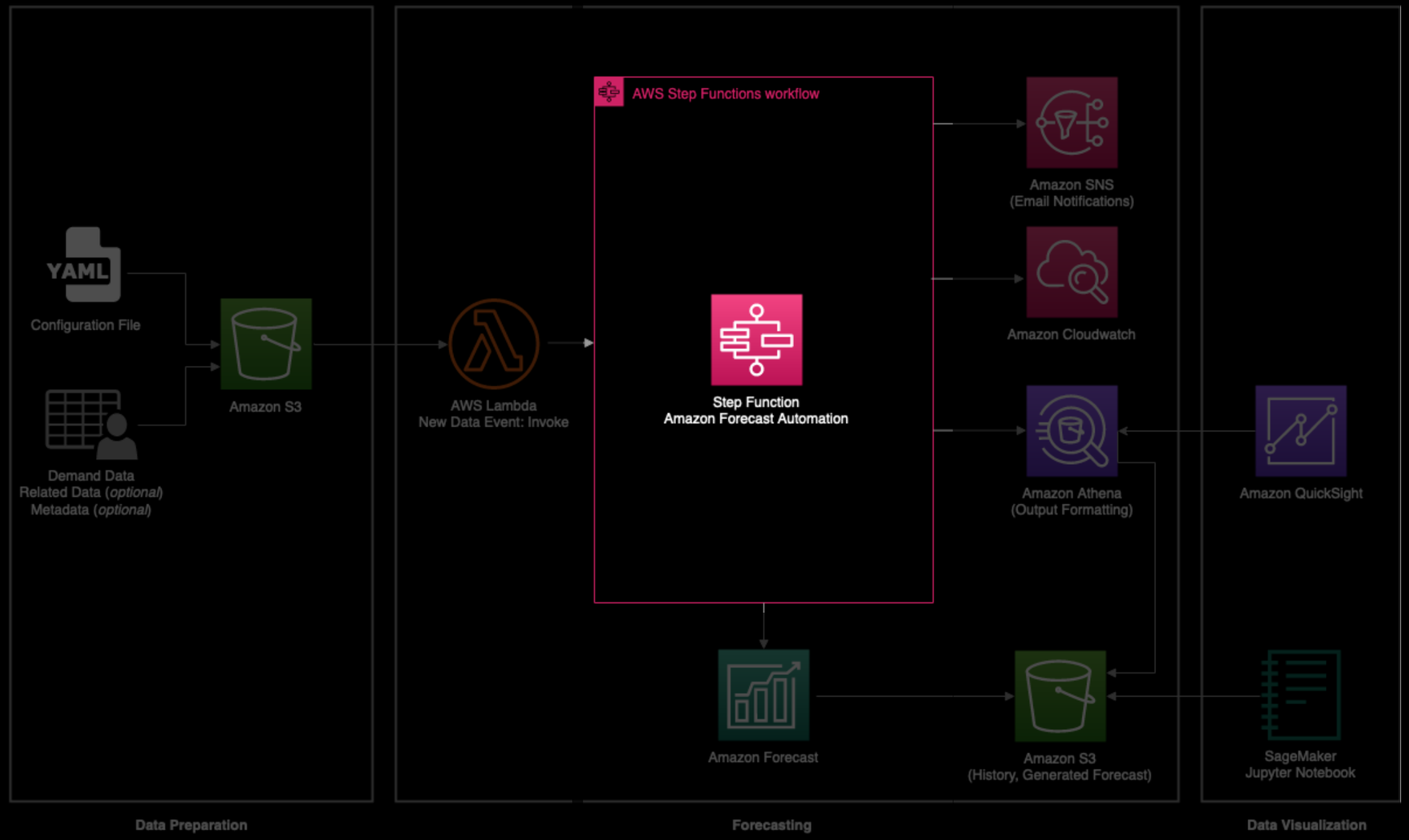
https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens

Versioning ML inputs and artifacts



https://aws.amazon.com/solutions/implementations/improving-forecast-accuracy-with-machine-learning

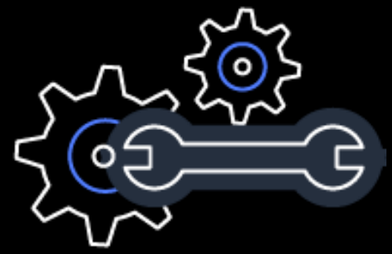
Automating ML deployment pipelines



https://aws.amazon.com/solutions/implementations/improving-forecast-accuracy-with-machine-learning

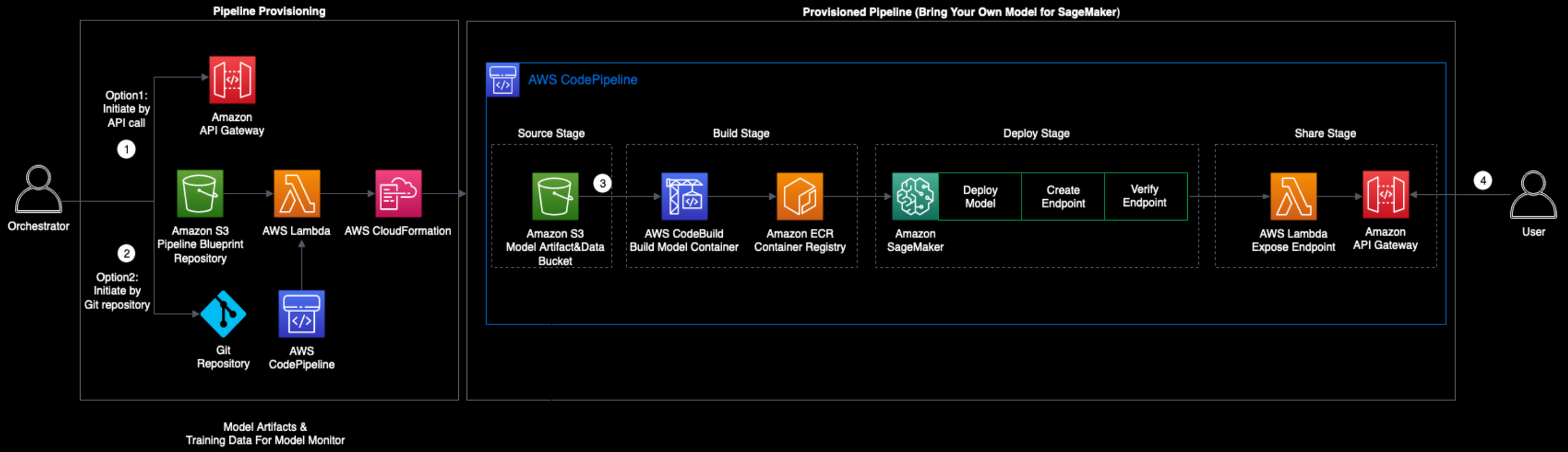
Operational excellence

- Continuously monitor and measure ML workloads
- Establish a model retraining strategy
- Version machine learning inputs and artifacts
- Automate machine learning deployment pipelines



Intelligent applications with AWS ML services

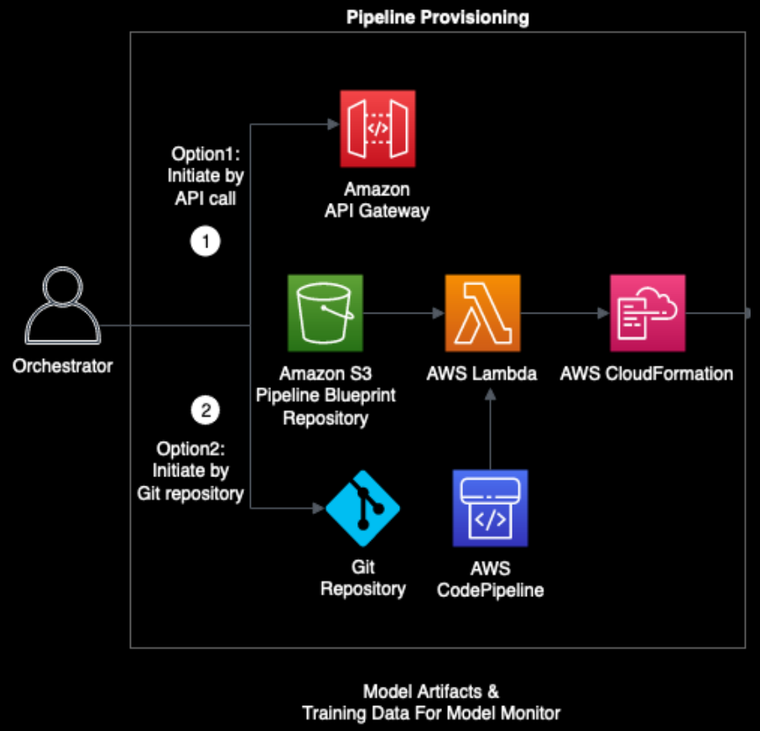
AWS MLOps Framework



https://aws.amazon.com/solutions/implementations/aws-mlops-framework

AWS MLOps Framework

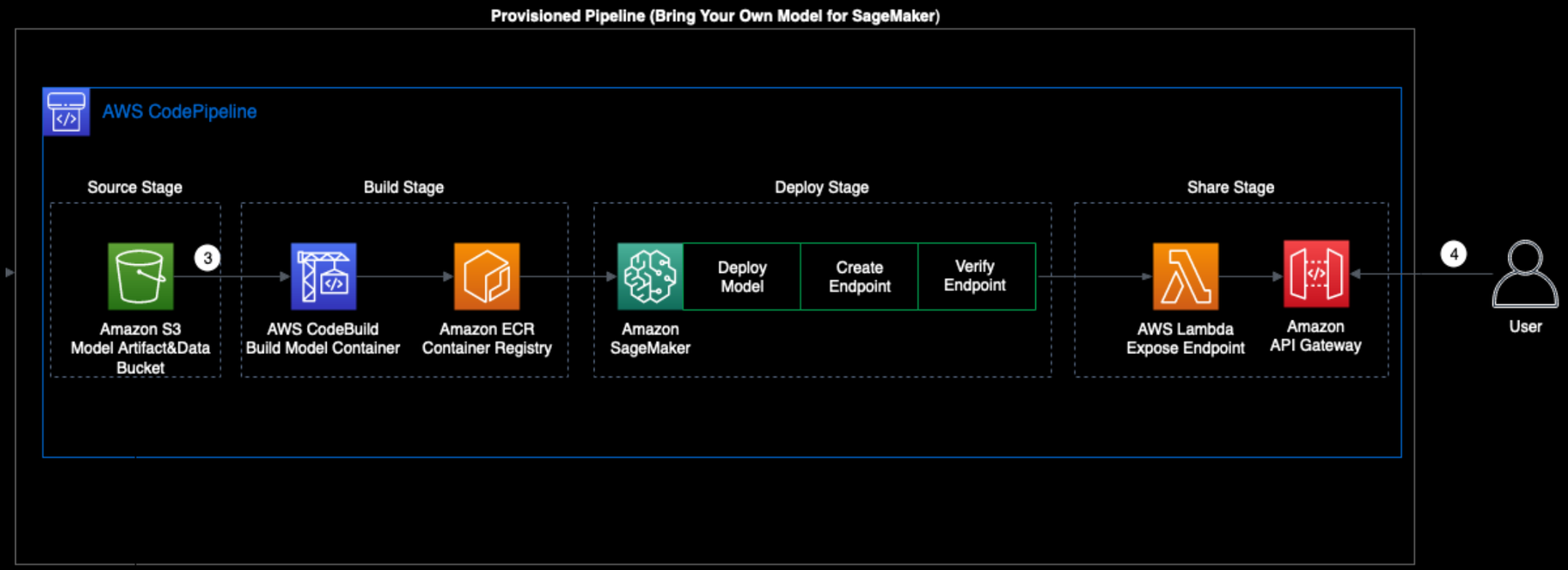
Pipeline provisioner



https://aws.amazon.com/solutions/implementations/aws-mlops-framework

AWS MLOps Framework

Pipeline for ML model deployment



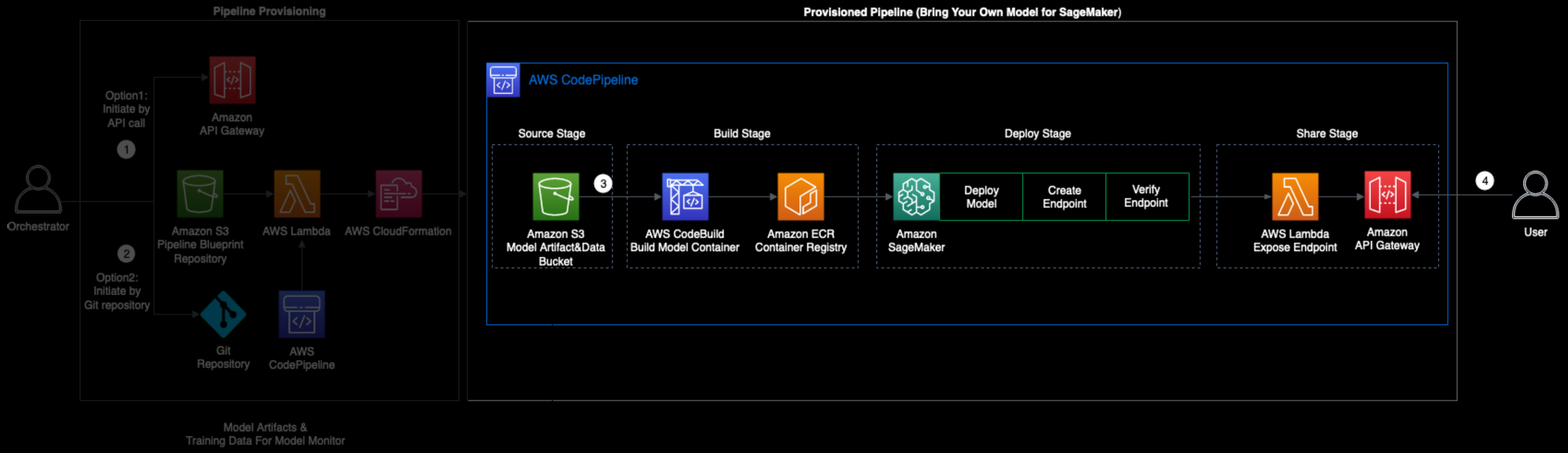
https://aws.amazon.com/solutions/implementations/aws-mlops-framework

ML Lens for Well-Architected Framework

Well-Architected Pillar	Key Best Practices for ML
Cost Optimization	Use managed services to reduce cost
Operational Excellence	Version ML inputs and artifacts Automate machine learning deployment pipelines
Reliability	Train once and deploy across environments
Security	Restrict access to ML services
Performance Efficiency	Continuously monitor and measure system performance

https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens

Train once and deploy across environments



https://aws.amazon.com/solutions/implementations/aws-mlops-framework

Reliability

- Manage changes to model inputs through automation
- Train once and deploy across environment



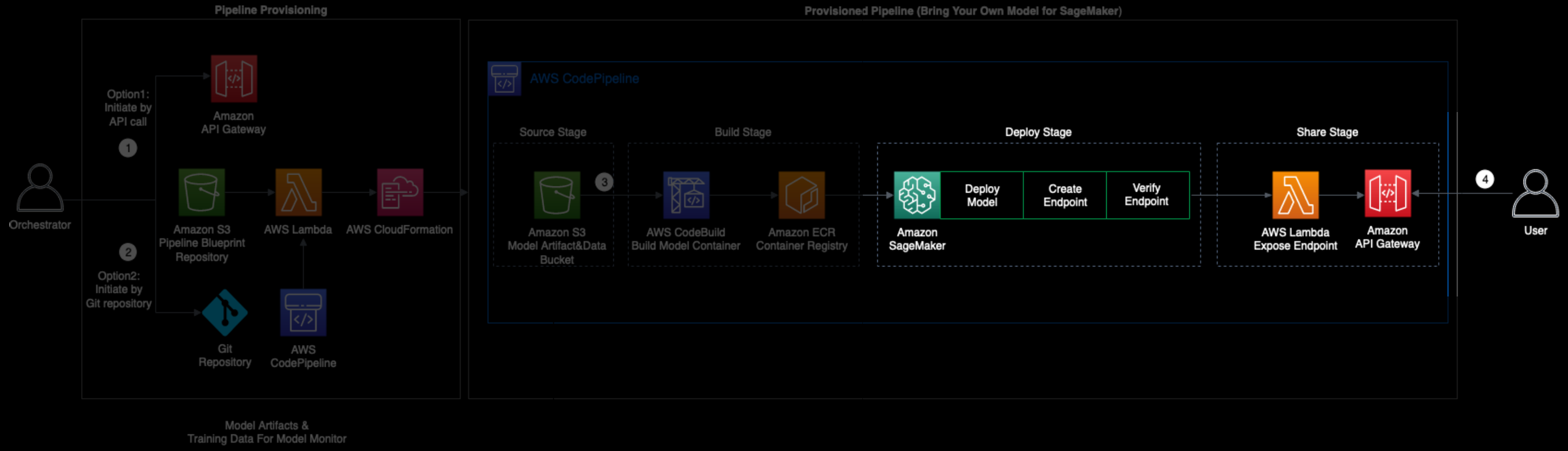
ML Lens for Well-Architected Framework

Well-Architected Pillar	Key Best Practices for ML
Cost Optimization	Use managed services to reduce cost
Operational Excellence	Version ML inputs and artifacts Automate machine learning deployment pipelines
Reliability	Train once and deploy across environments
Security	Restrict access to ML services
Performance Efficiency	Continuously monitor and measure system performance

https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens

Intelligent applications with AWS ML services

AWS MLOps Framework



https://aws.amazon.com/solutions/implementations/aws-mlops-framework

Security

- Restrict access to ML systems
- Ensure data governance
- Enforce data lineage
- Enforce regulatory compliance

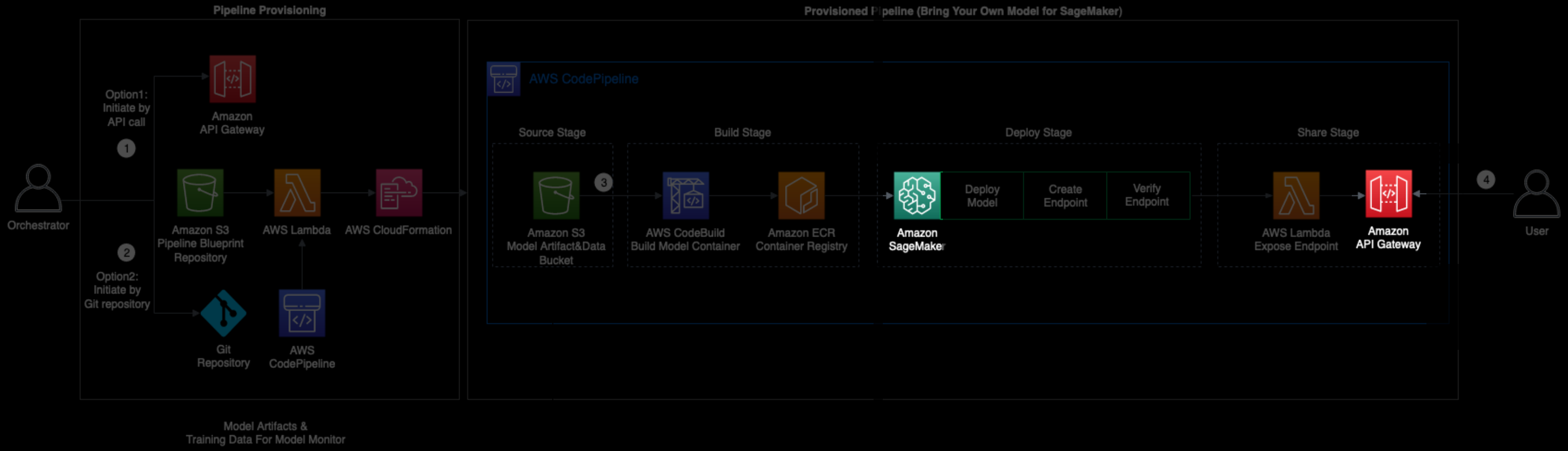


ML Lens for Well-Architected Framework

Well-Architected Pillar	Key Best Practices for ML
Cost Optimization	Use managed services to reduce cost
Operational Excellence	Version ML inputs and artifacts Automate machine learning deployment pipelines
Reliability	Train once and deploy across environments
Security	Restrict access to ML services
Performance Efficiency	Continuously monitor and measure system performance

https://docs.aws.amazon.com/wellarchitected/latest/machine-learning-lens

Continuously monitor system performance



https://aws.amazon.com/solutions/implementations/aws-mlops-framework

Performance efficiency

- Optimize compute for your ML workloads
- Define latency and network bandwidth performance requirements
- Continuously monitor and measure system performance



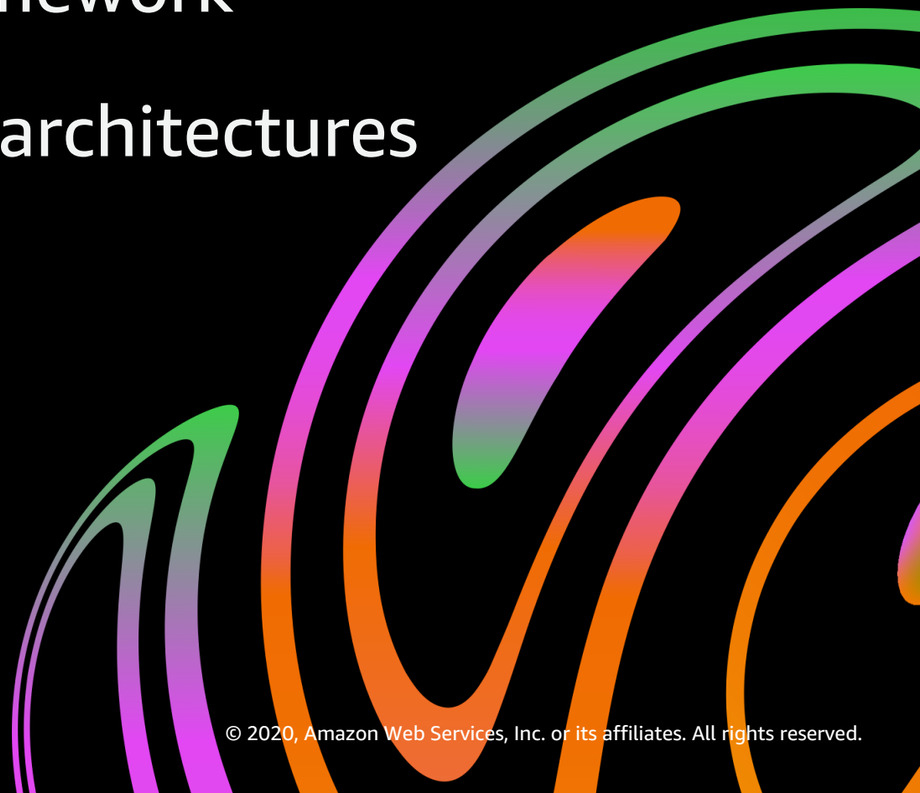
Takeaways

Building applications vs. building intelligent application

AI / ML services on AWS

Machine Learning Lens for Well-Architected Framework

Apply Well-Architected best practices in AI / ML architectures



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Architecture resources

AWS Well-Architected Framework

The official best practices for Architecting in the AWS Cloud

https://aws.amazon.com/architecture/well-architected

AWS Well-Architected Labs

Hands-on labs to help you learn, measure, and build using architectural best practices

https://wellarchitectedlabs.com/

AWS Architecture Center

Official AWS repository for all Architecture resources

https://aws.amazon.com/architecture

AWS Solutions Library

Vetted reference implementations and Well-Architected patterns

https://aws.amazon.com/solutions/



© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Thank you!

Mohsen Ansari

Solution Builder
AWS

Linkedin.com/in/mohsenansari

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Please complete the session survey

AWS re:Invent

© 2020, Amazon Web Services, Inc. or its affiliates. All rights reserved.