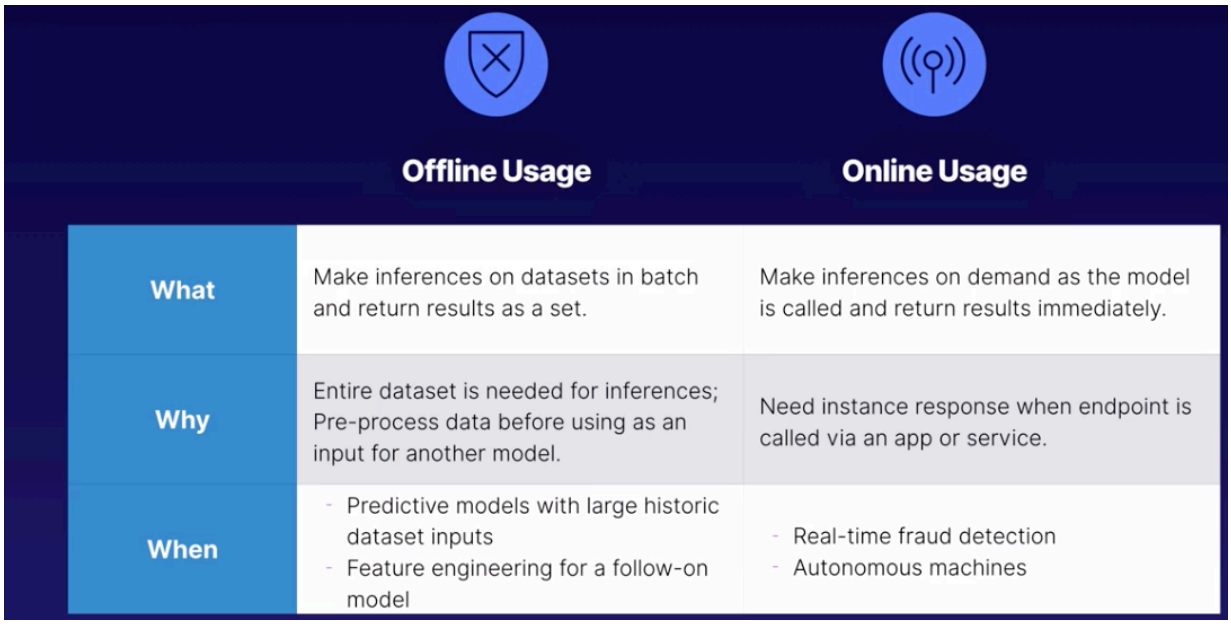
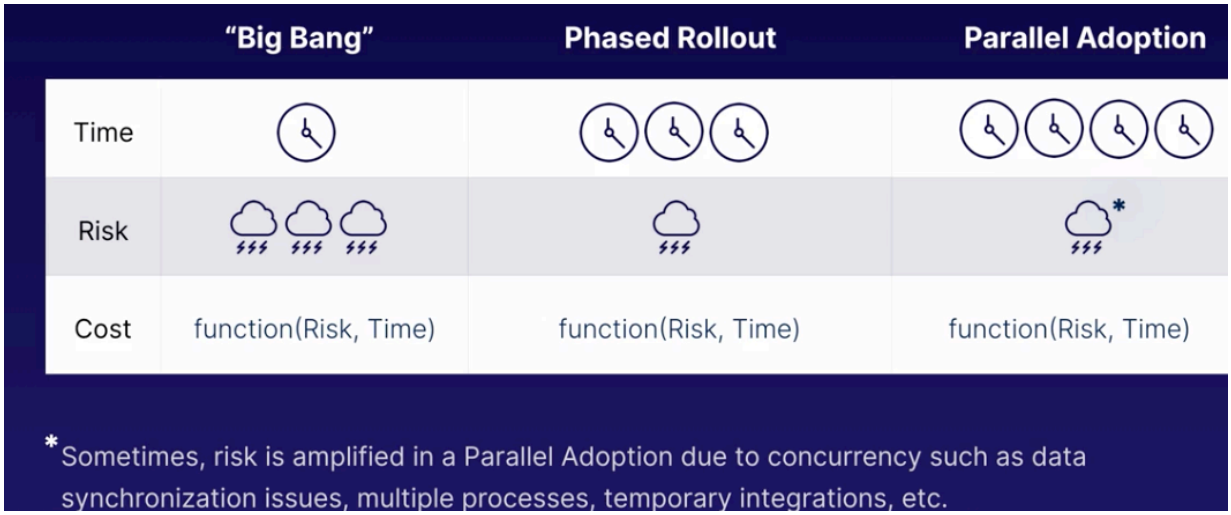
Cloud Guru - 8 - Implementation and Operations

2 types of usage



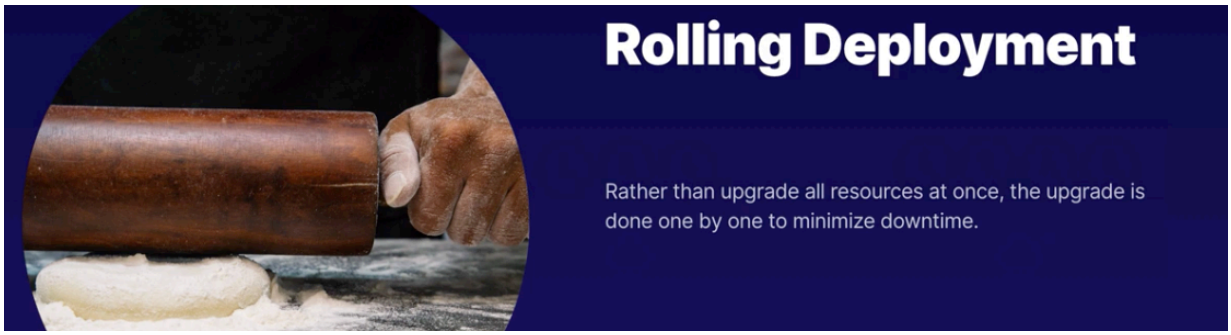
	Offline Usage	Online Usage
What	Make inferences on datasets in batch and return results as a set.	Make inferences on demand as the model is called and return results immediately.
Why	Entire dataset is needed for inferences; Pre-process data before using as an input for another model.	Need instance response when endpoint is called via an app or service.
When	- Predictive models with large historic dataset inputs - Feature engineering for a follow-on model	- Real-time fraud detection - Autonomous machines

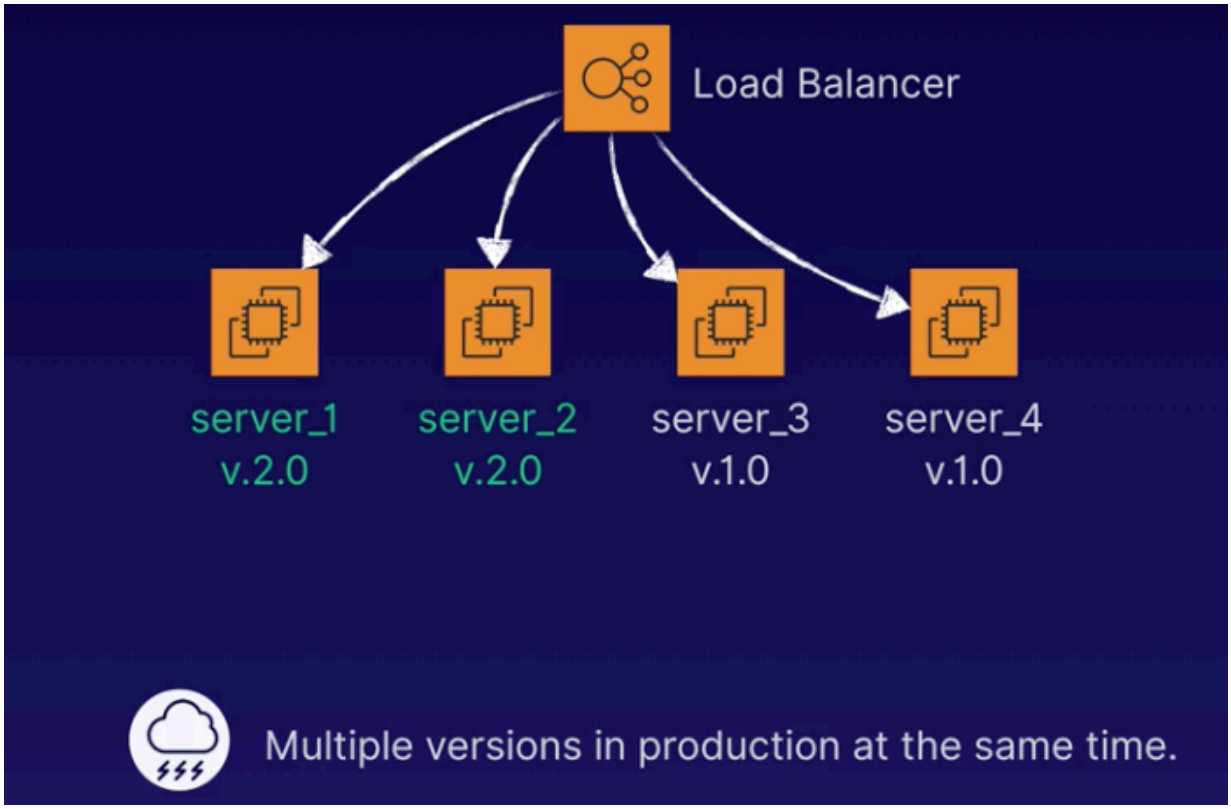
Type of deployments:



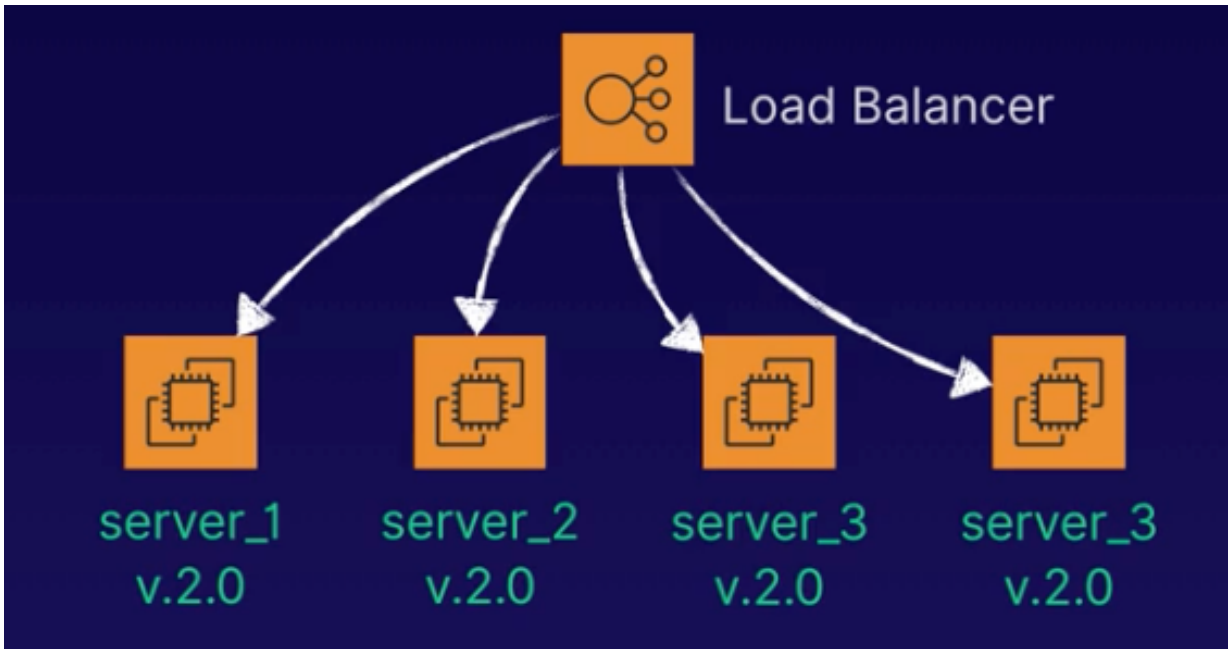
	"Big Bang"	Phased Rollout	Parallel Adoption
Time			
Risk			*
Cost	function(Risk, Time)	function(Risk, Time)	function(Risk, Time)

*Sometimes, risk is amplified in a Parallel Adoption due to concurrency such as data synchronization issues, multiple processes, temporary integrations, etc.

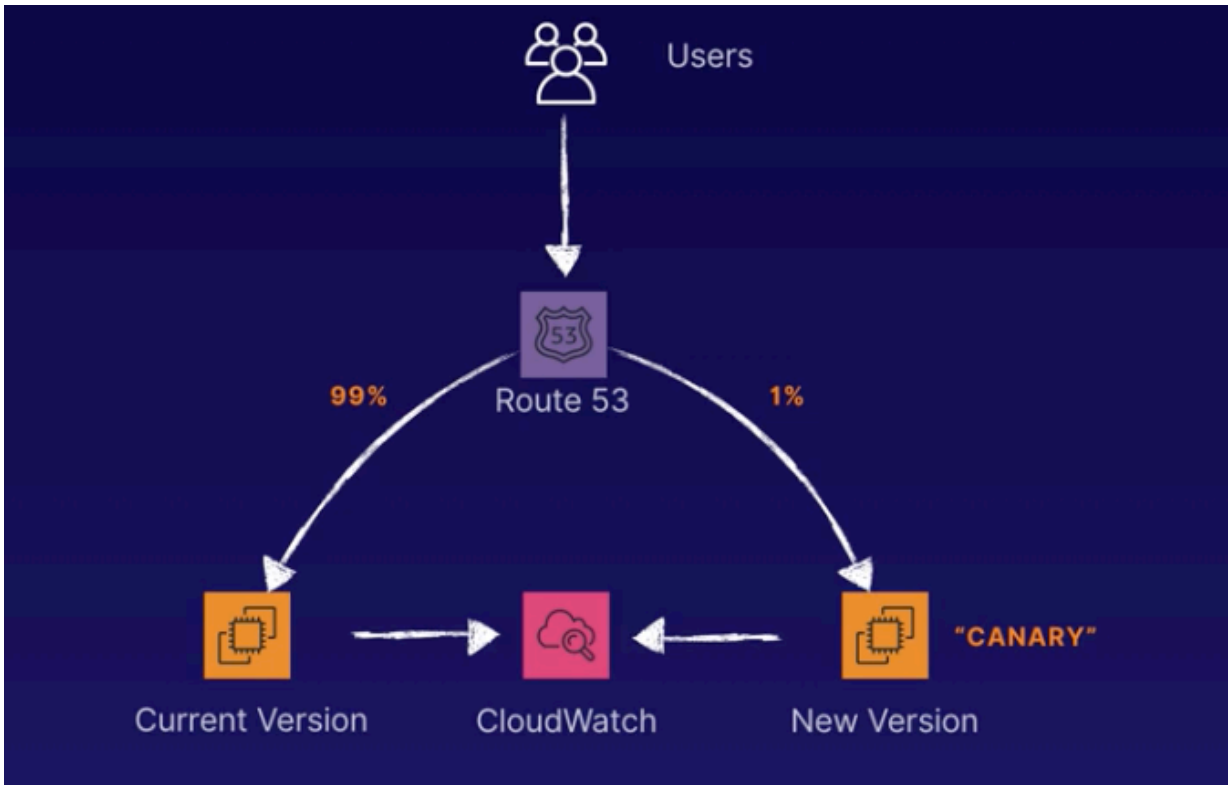

Rolling Deployment
Rather than upgrade all resources at once, the upgrade is done one by one to minimize downtime.


Load Balancer
server_1
v.2.0
server_2
v.2.0
server_3
v.1.0
server_4
v.1.0
Multiple versions in production at the same time.

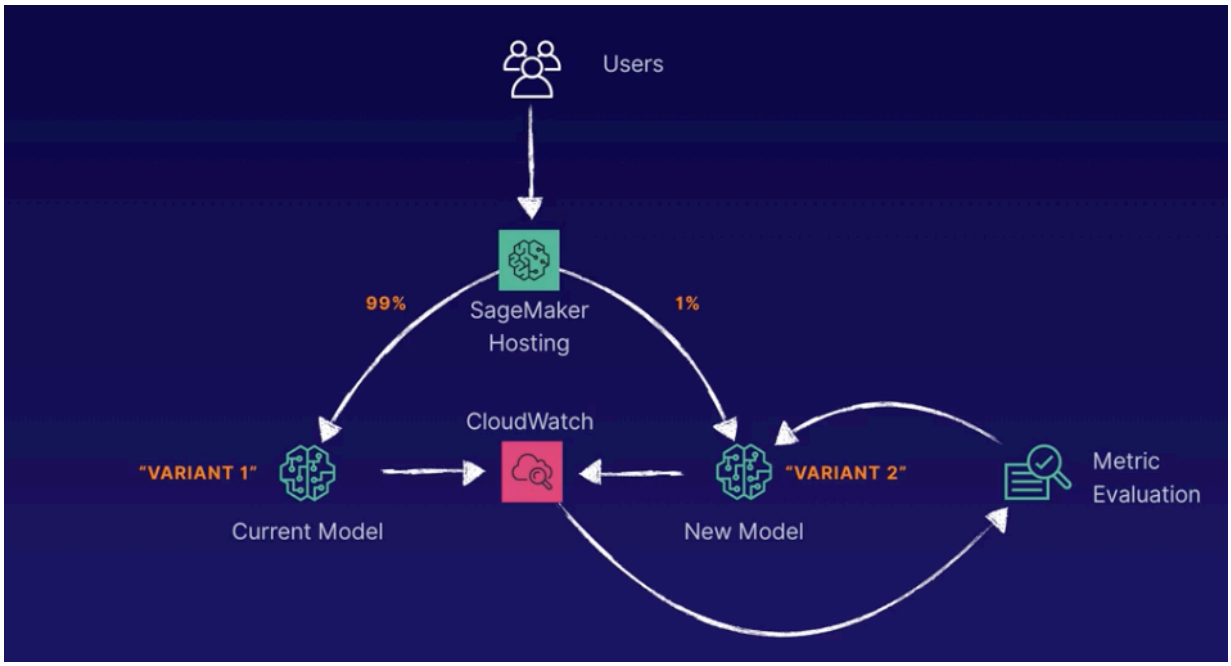
Eventually:


Load Balancer
server_1
v.2.0
server_2
v.2.0
server_3
v.2.0
server_3
v.2.0

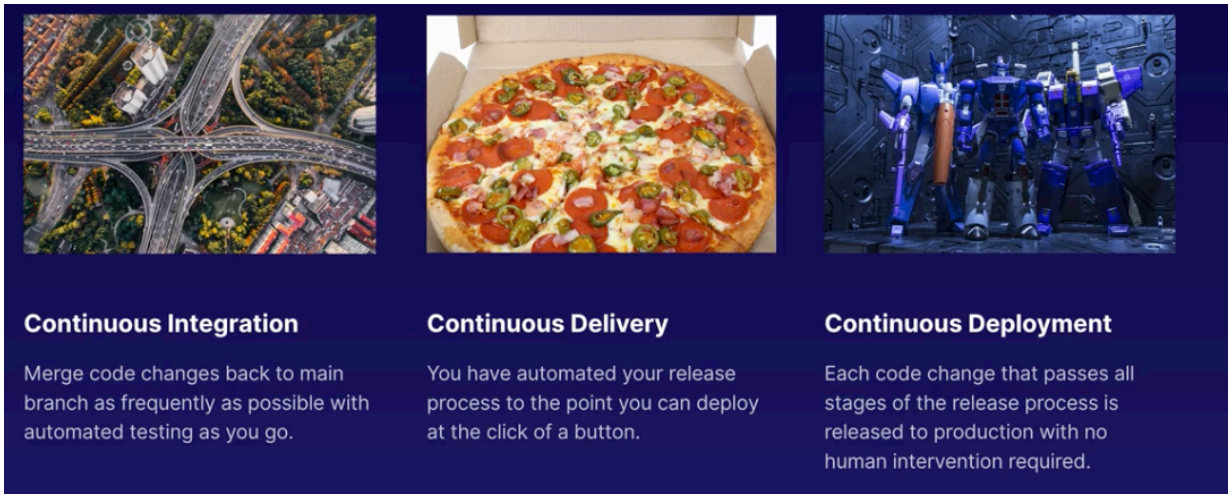

Canary Deployment
Deploy a new version into production so it sees a small portion of the total traffic and evaluate the performance and behavior.


Users
53
Route 53
99%
1%
"CANARY"
Current Version
CloudWatch
New Version

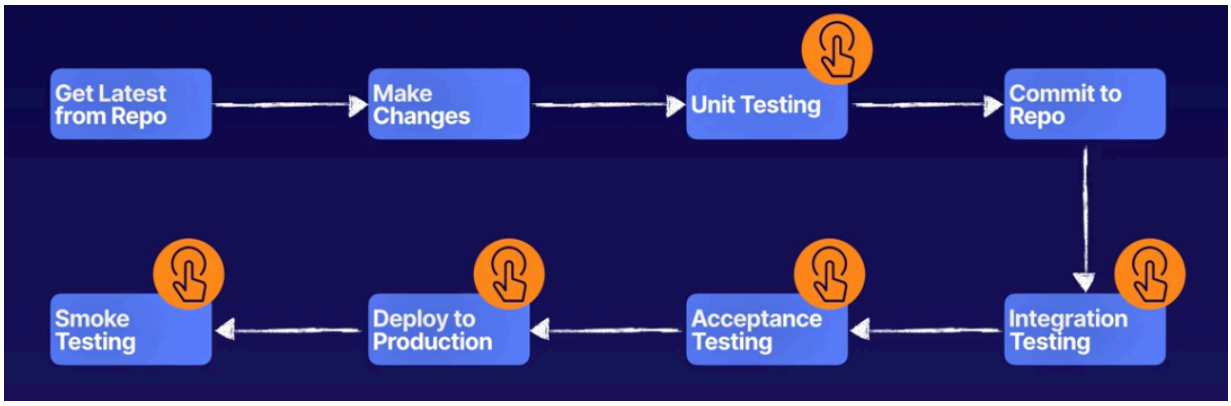

B
A/B Testing
Deploy a new version into production and configure a set amount of new inbound traffic to use the new (B) version, recording follow-on data about the outcome of those to used the new version.
Once enough data is collected, make a decision on whether to fully deploy the new version or make adjustments.



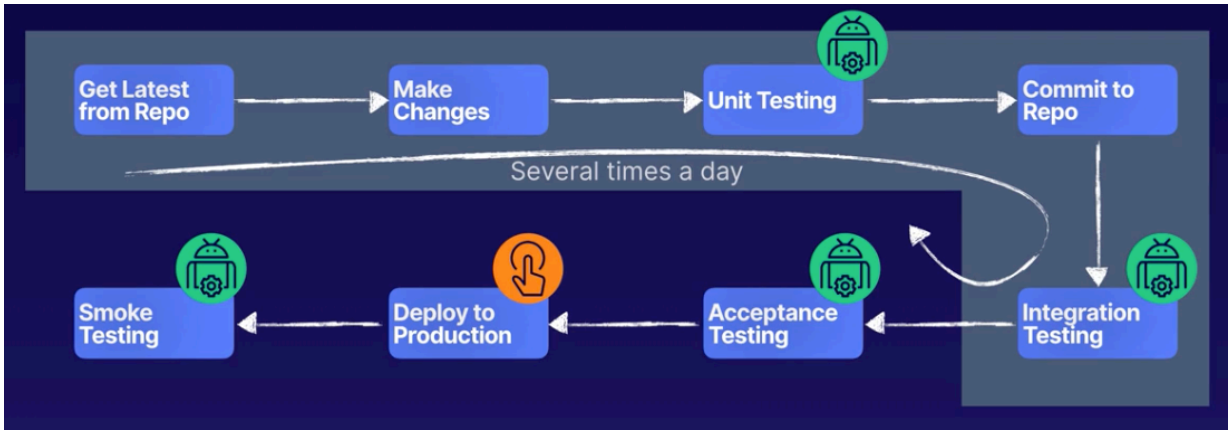
CI/CD pipeline



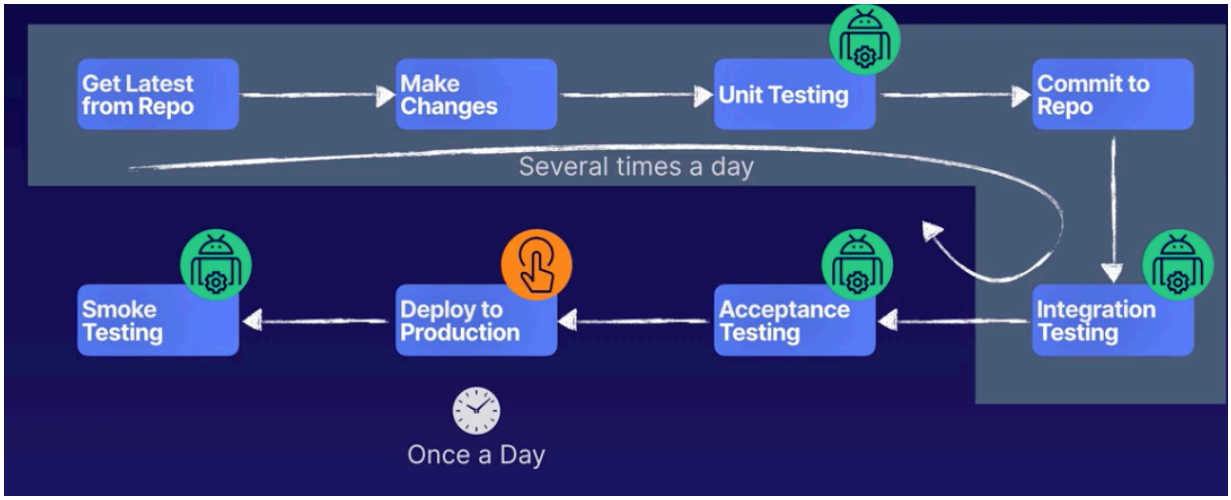
Typical Development Lifecycle:
Involves a human for different gates



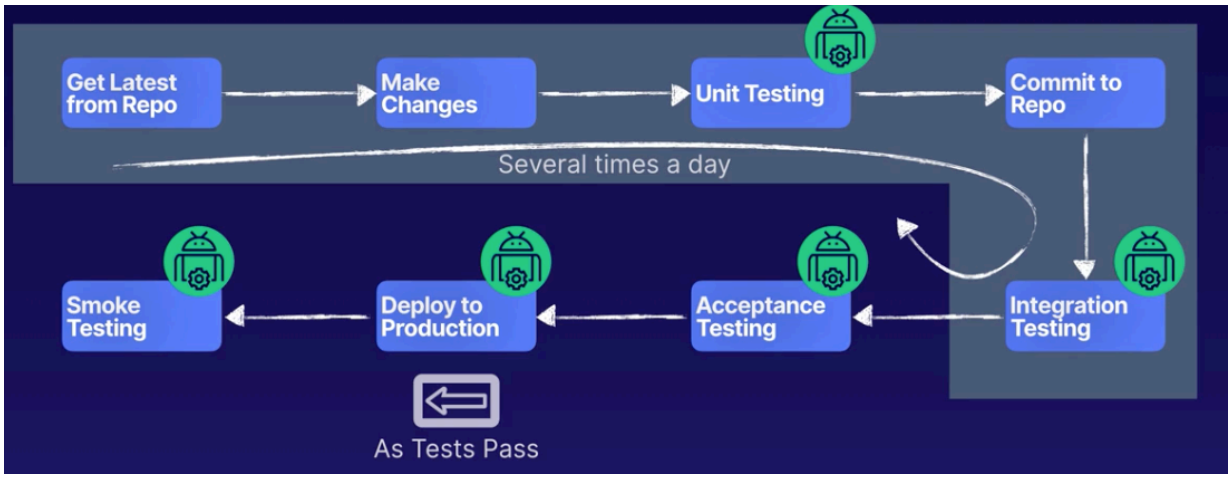
Continuous integration:
With each check-in, integration tests to make sure nothing breaks



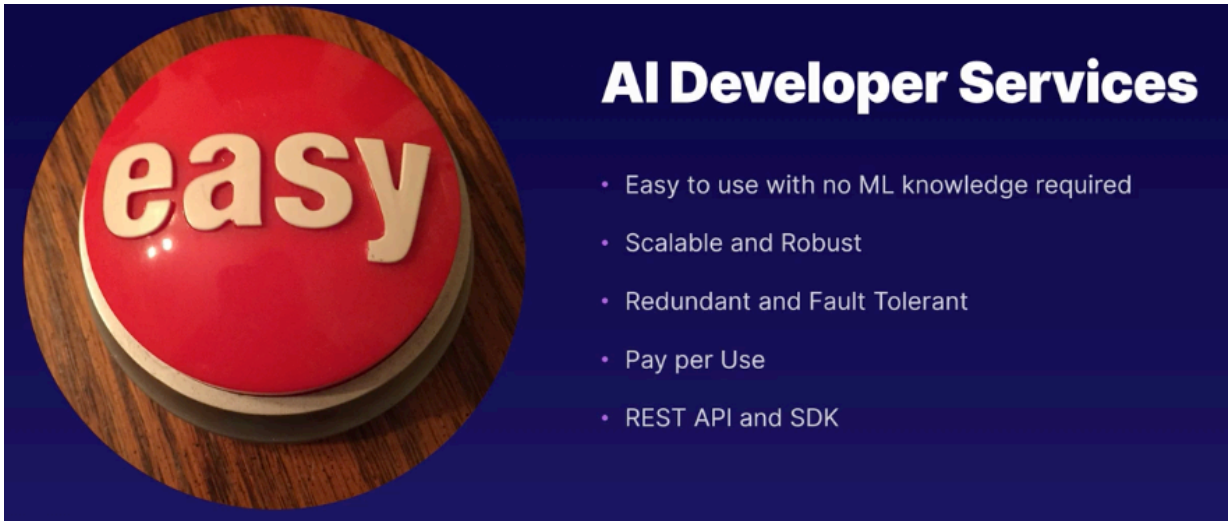
Continuous delivery:
Same process, but can release once per day



Continuous Deployment:



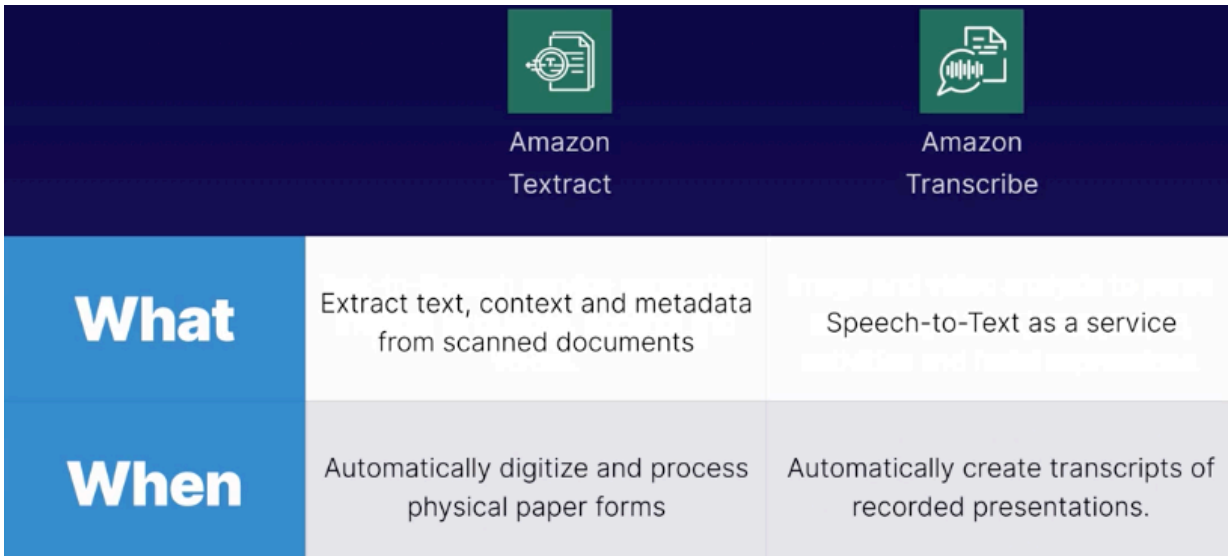
AI Developer services



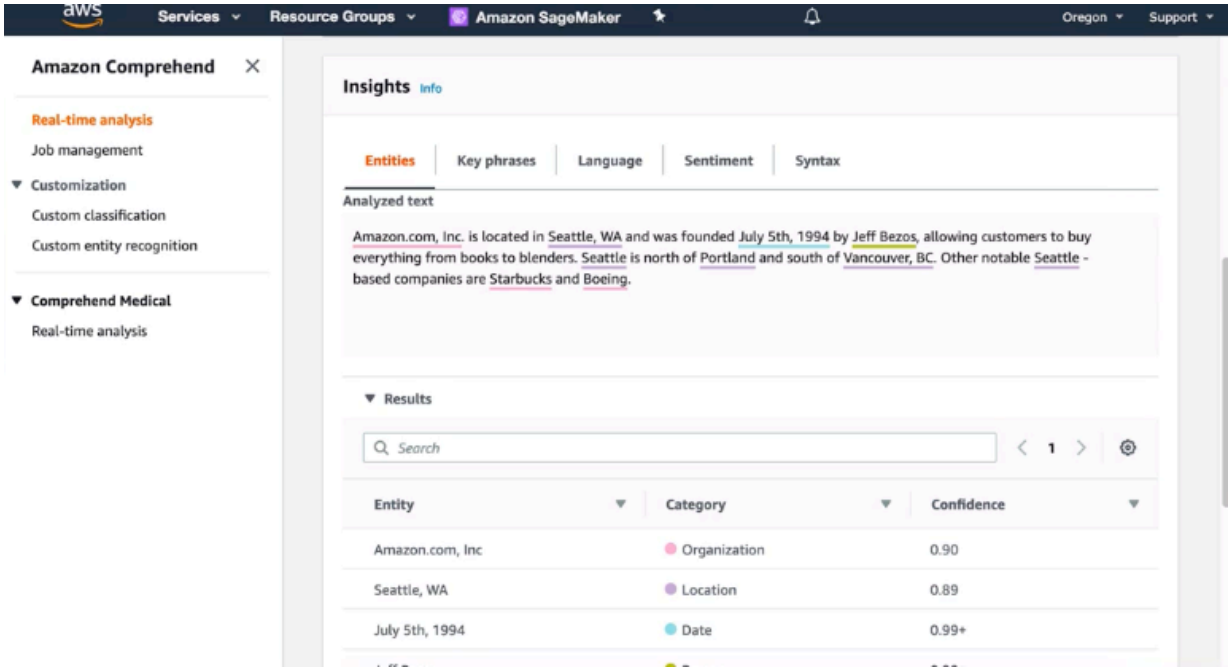
	Amazon Comprehend	Amazon Forecast
What	Natural Language Processing (NLP) service that finds insight and relationships within text.	Combines time-series data with other variables to deliver highly accurate forecasts.
When	Sentiment analysis of social media posts	Forecast seasonal demand for a specific color of shirt.

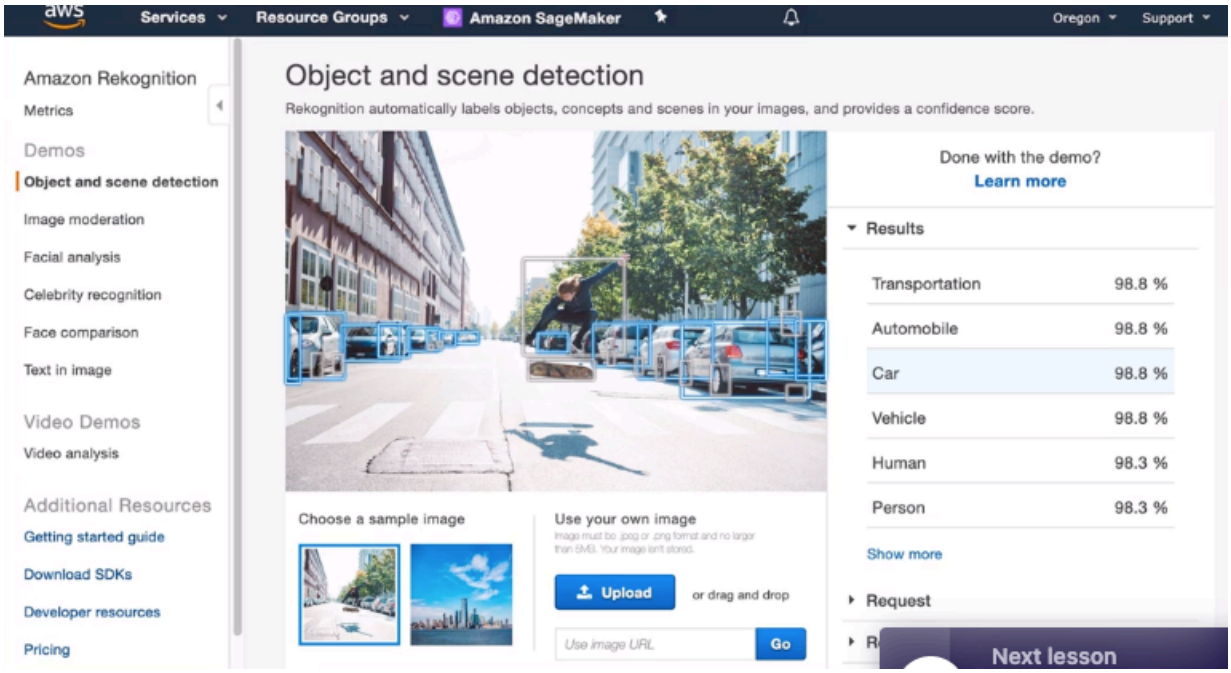
	Amazon Lex	Amazon Personalize
What	Build conversational interfaces that can understand the intent and context of natural speech.	Recommendation engine as a service based on demographic and behavioral data.
When	Create a customer service chatbot to automatically handle routine requests.	Provide potential upsell products at checkout during a web transaction.

	Amazon Polly	Amazon Rekognition
What	Text-to-Speech service supporting multiple languages, accents and voices.	Image and video analysis to parse and recognize objects, people, activities and facial expressions.
When	Provide dynamically generated personalized voice response for inbound callers.	Provide an additional form of employee authentication though facial recognition as they scan an access badge.



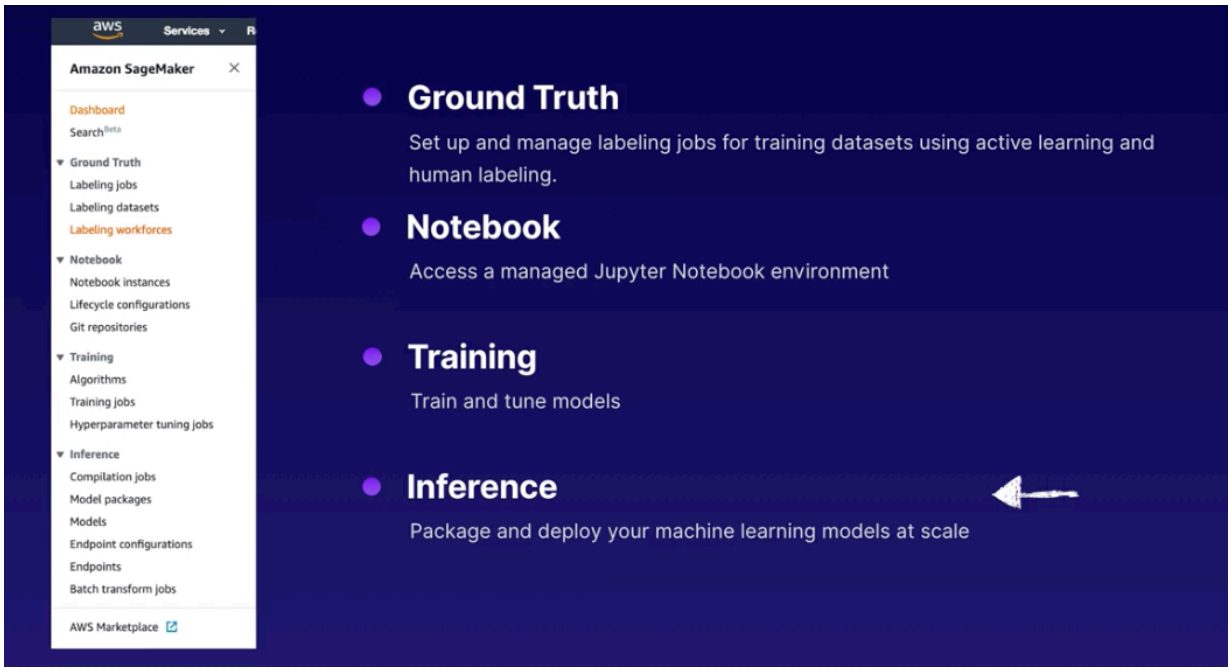
Console experience
√ç



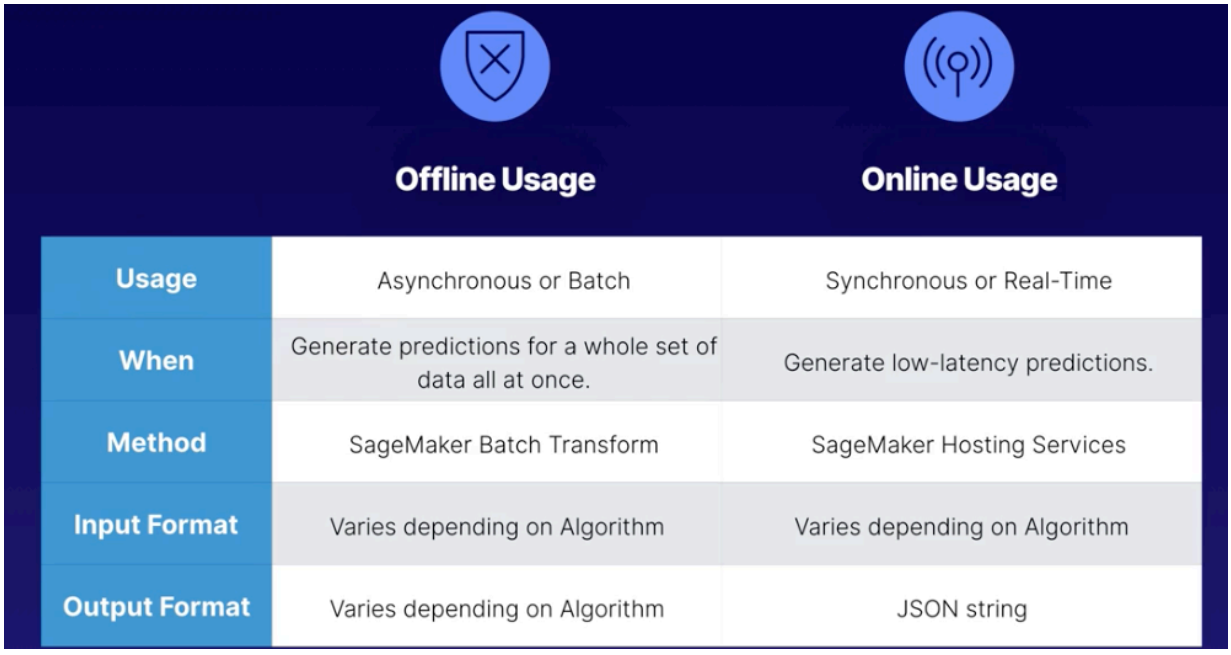


Amazon SageMaker Deployments

Inferences - deployment

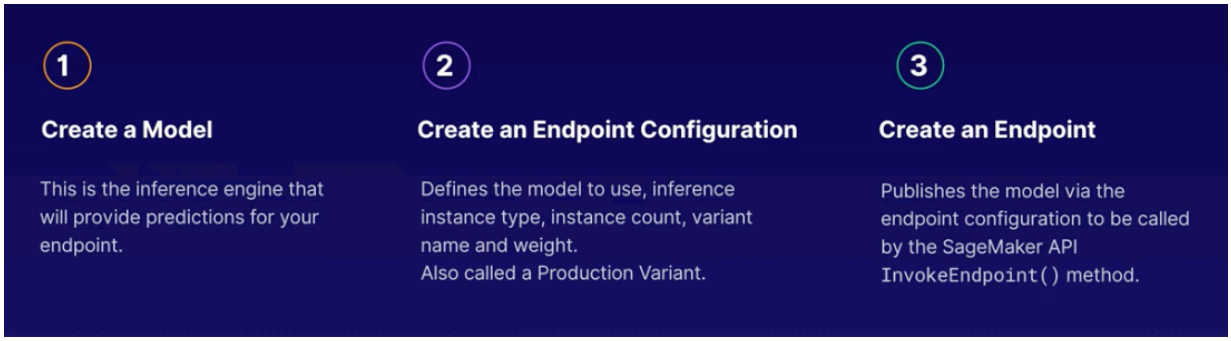


2 types of deployment: Offline vs Online

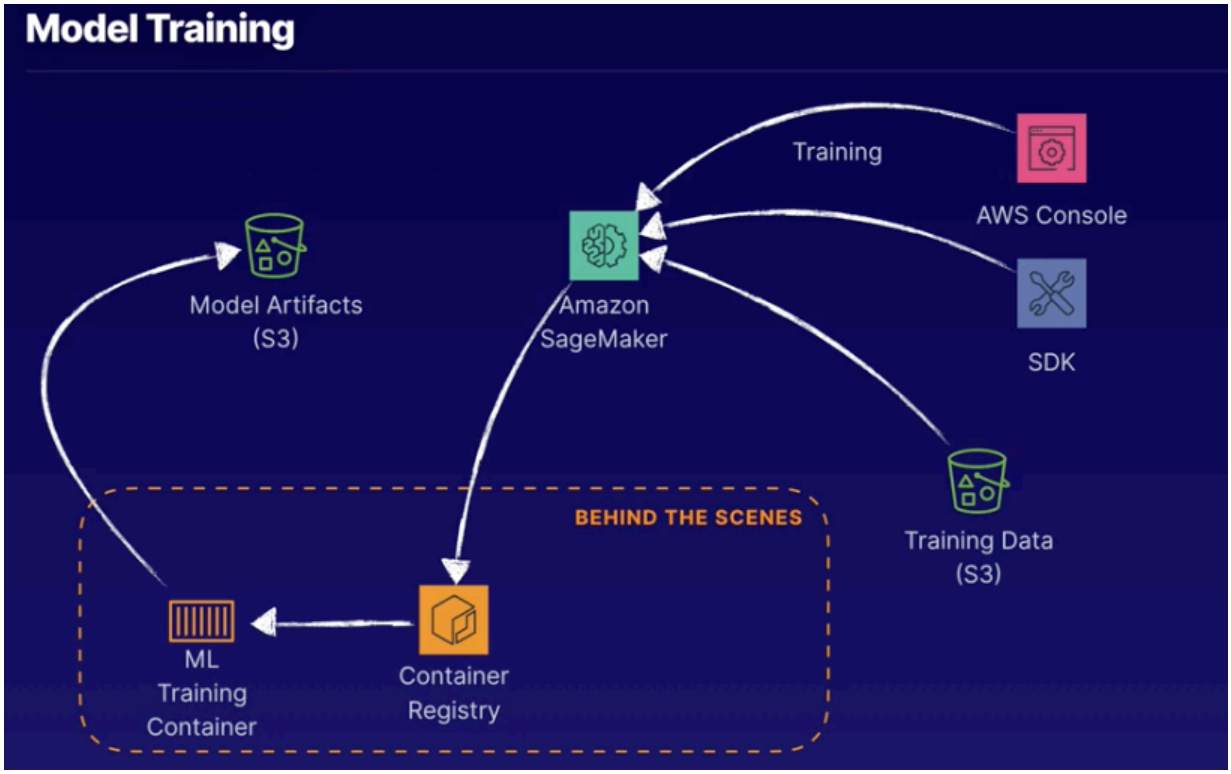


	Offline Usage	Online Usage
Usage	Asynchronous or Batch	Synchronous or Real-Time
When	Generate predictions for a whole set of data all at once.	Generate low-latency predictions.
Method	SageMaker Batch Transform	SageMaker Hosting Services
Input Format	Varies depending on Algorithm	Varies depending on Algorithm
Output Format	Varies depending on Algorithm	JSON string

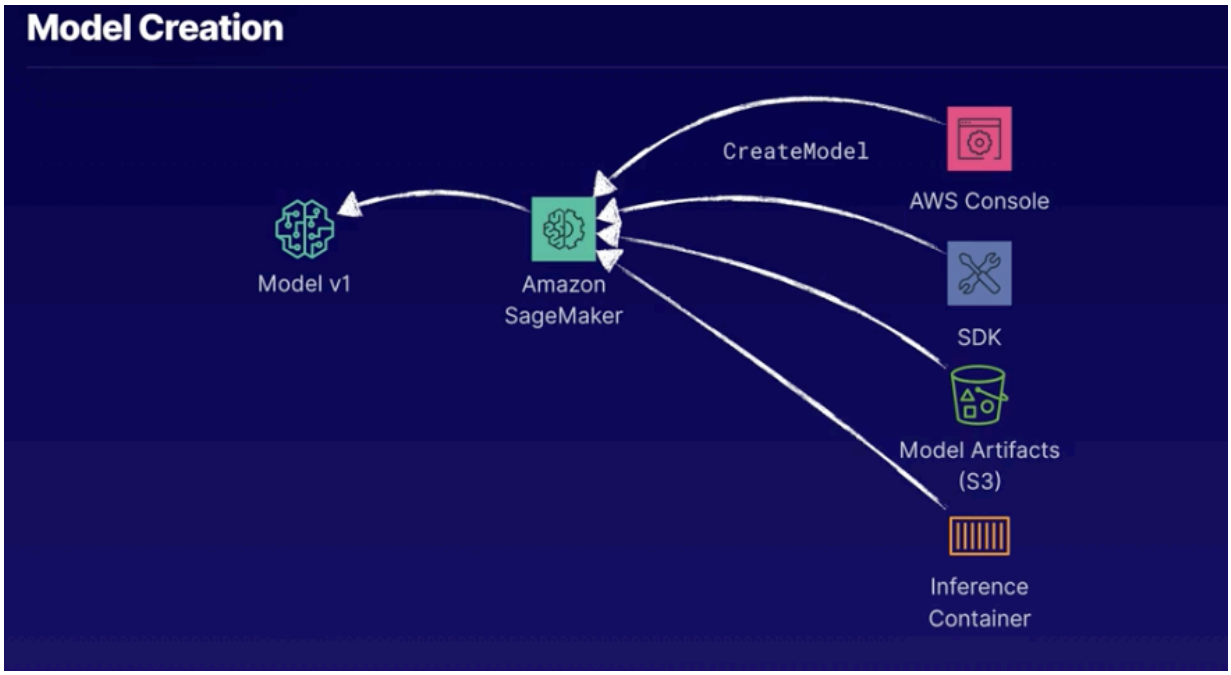
Steps:



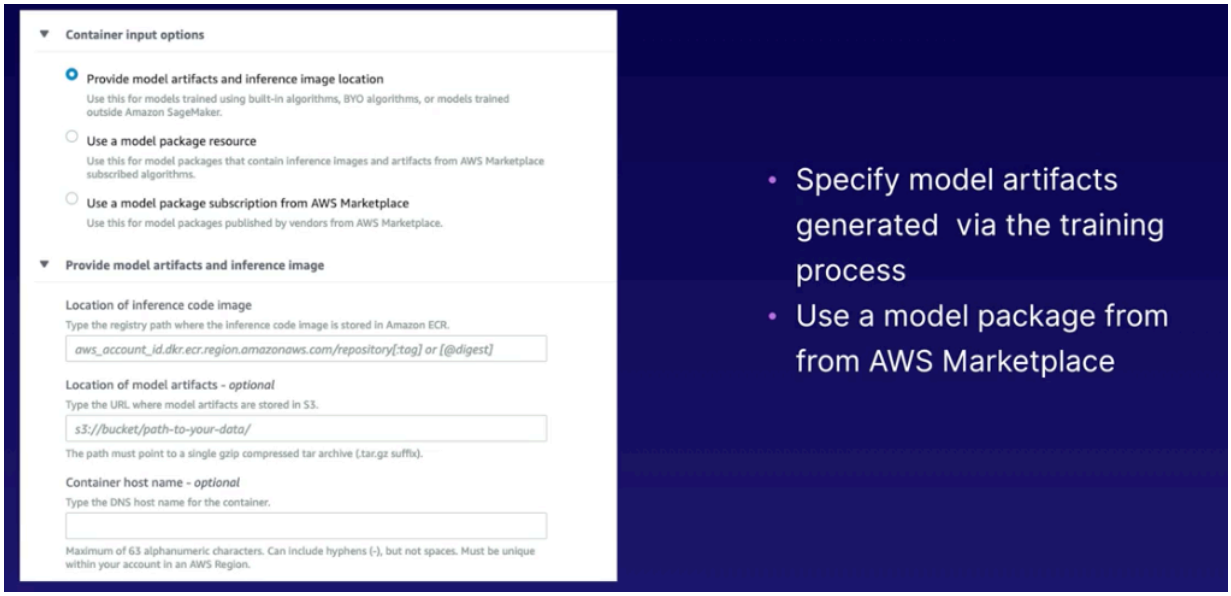
Model Training



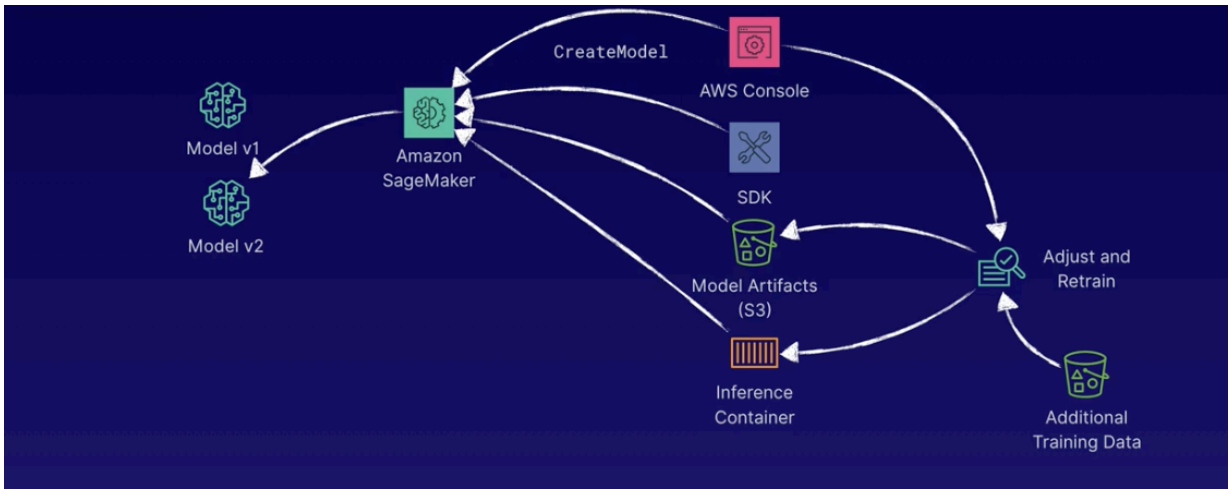
Model Creation (once training is finished)

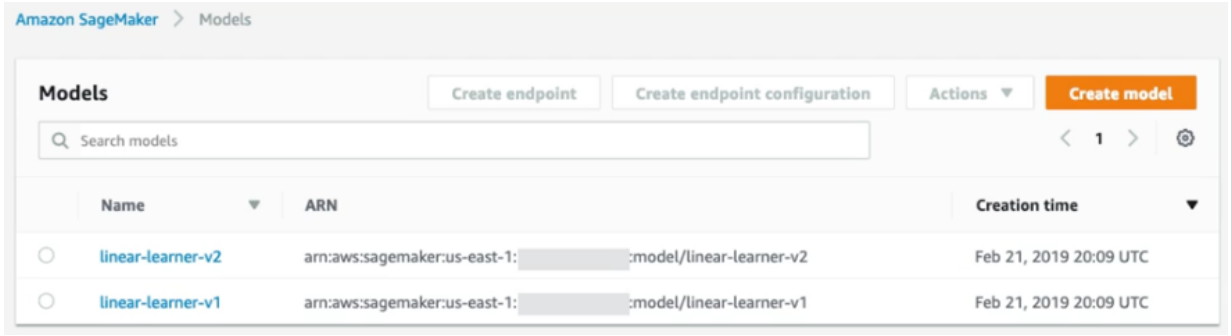


Console

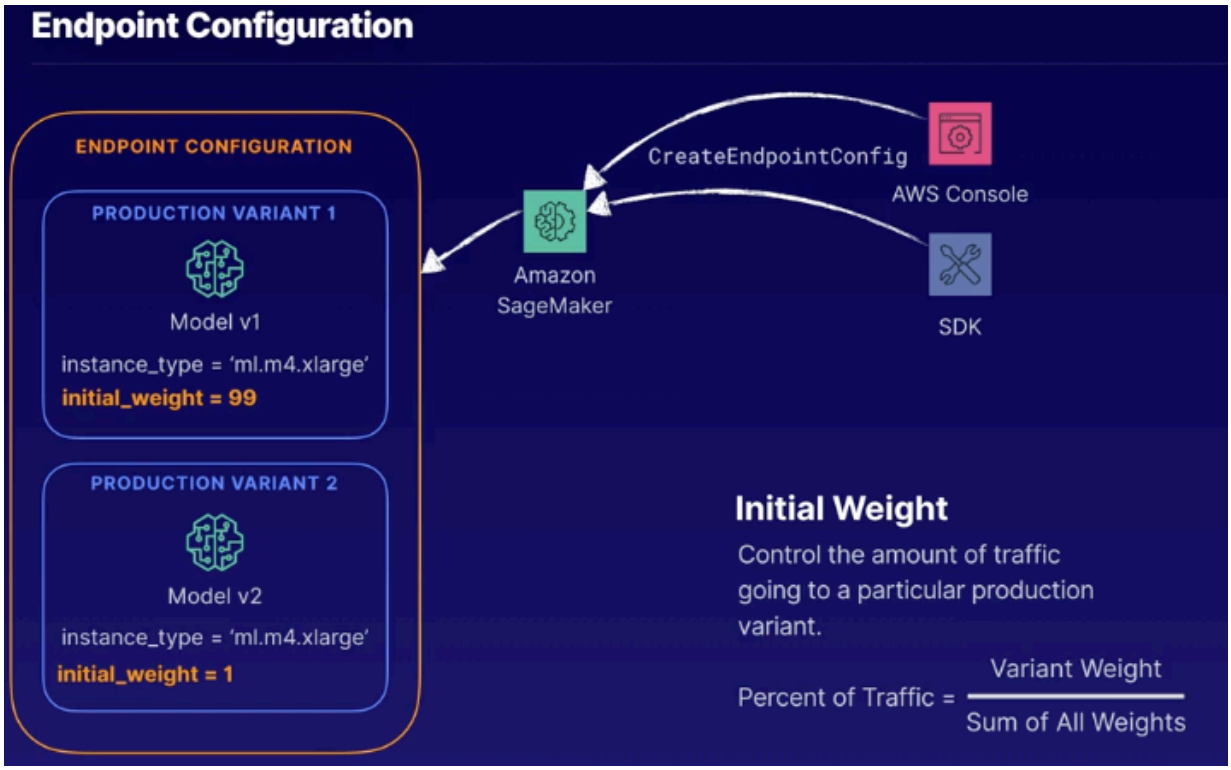


After model has been on prod for a while, we can RE-train it, and create a new version of that model





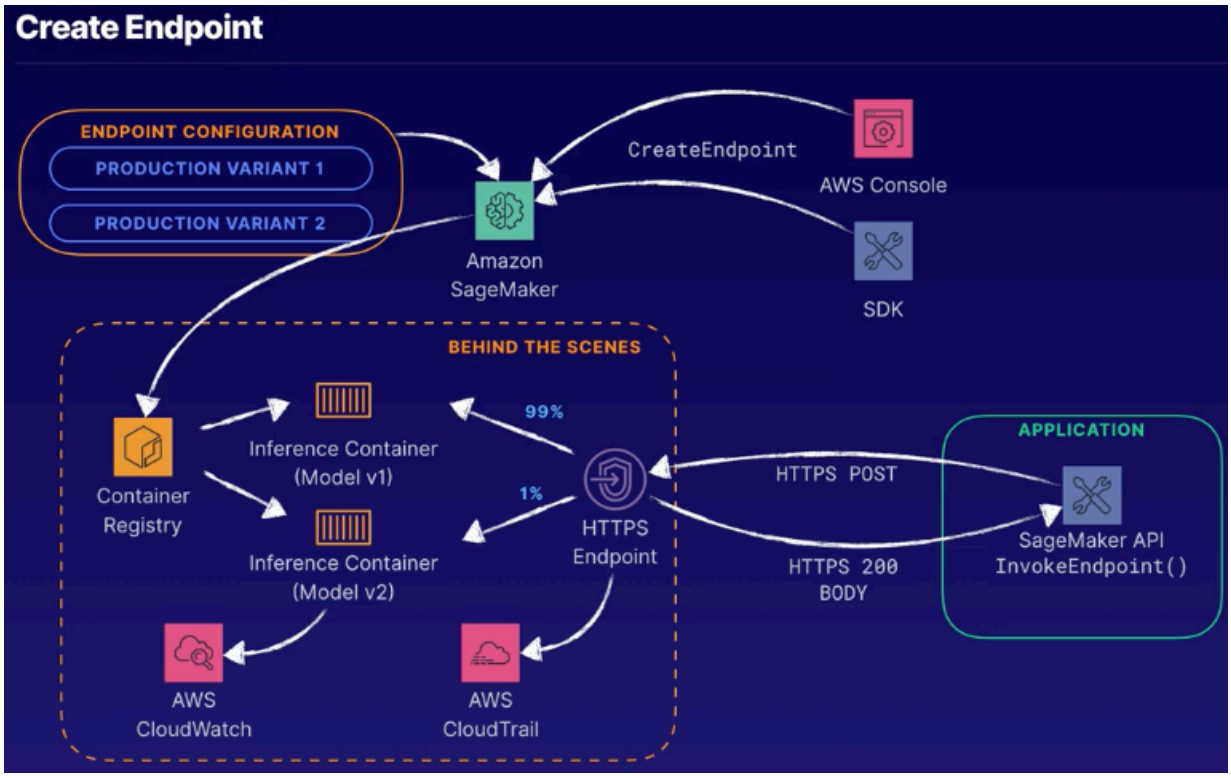
Endpoint configuration - to enable one or multiple production variant(s) of a model
We can define the wright to control what version gets how much traffic



Notes for Create endpoint:

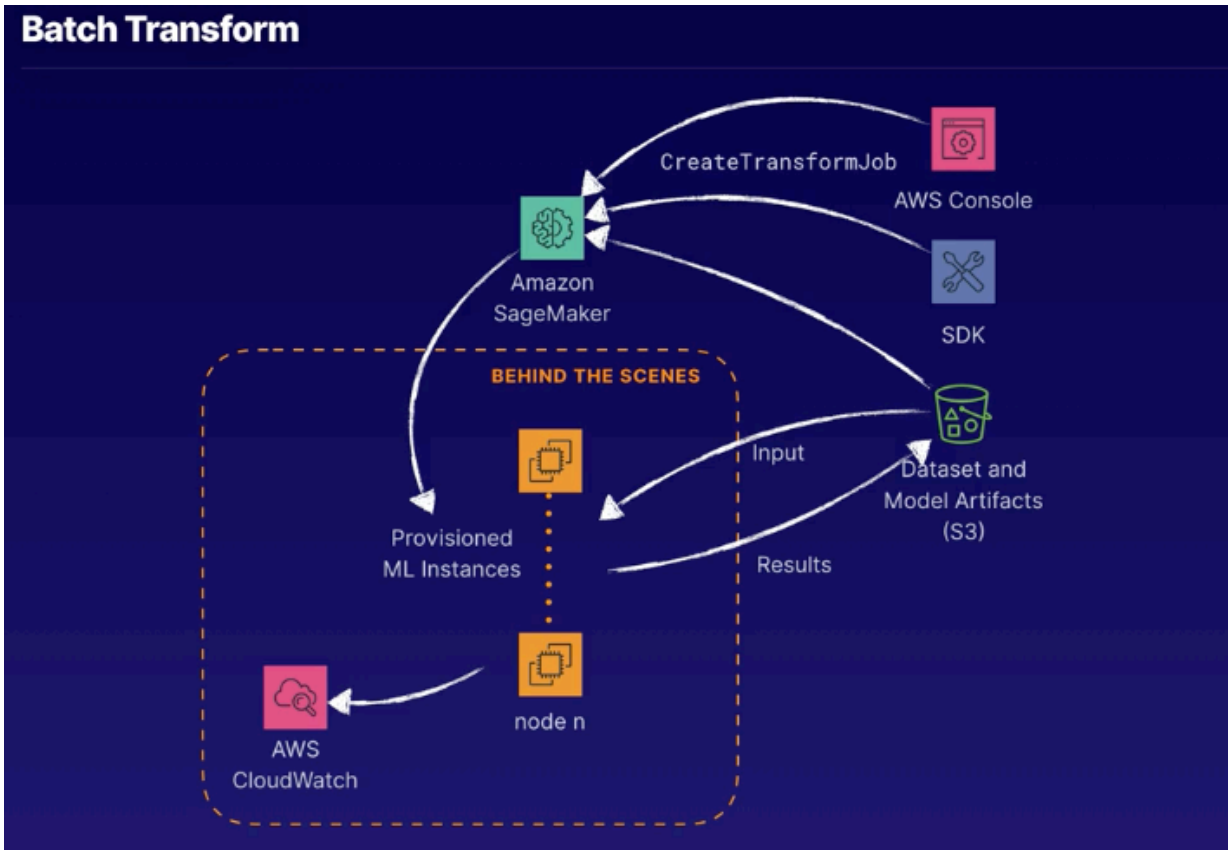
- **end point** traffic is logged into **CloudDtrail**
- **Container** logs is logged into **CloudWatch**

InvokeEndpoint() allows us to get the inferences

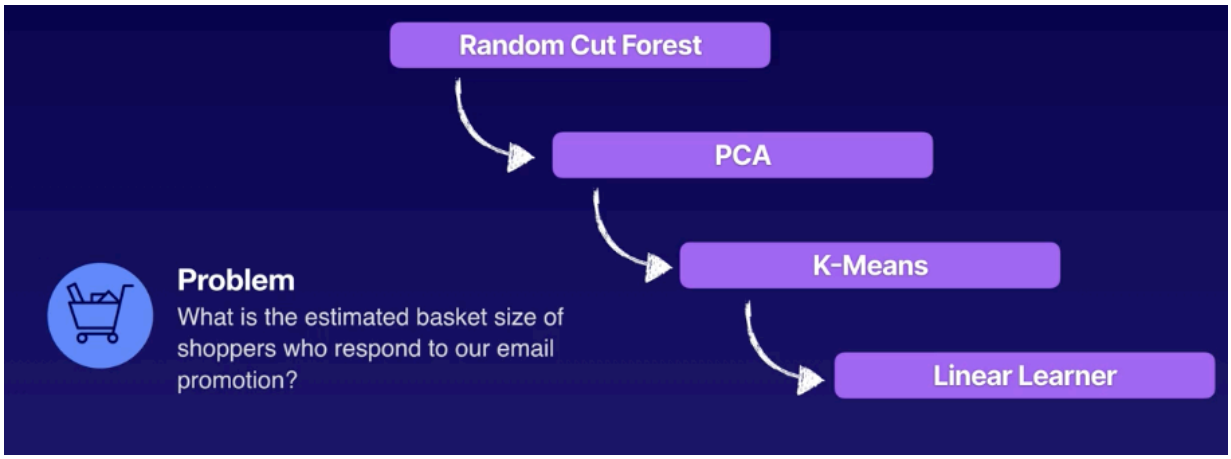


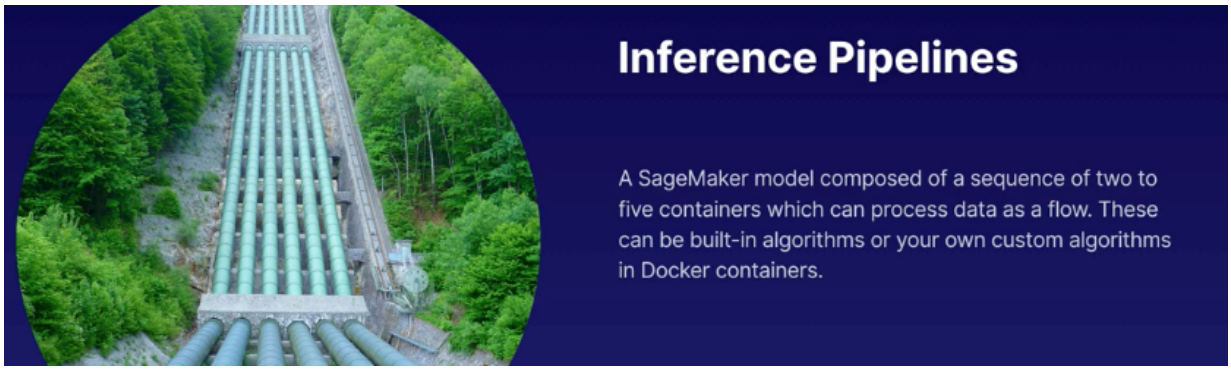
Batch Transform starts in a similar manner
SM provisions batch resources. Once inference jobs are complete, results get dropped into S3

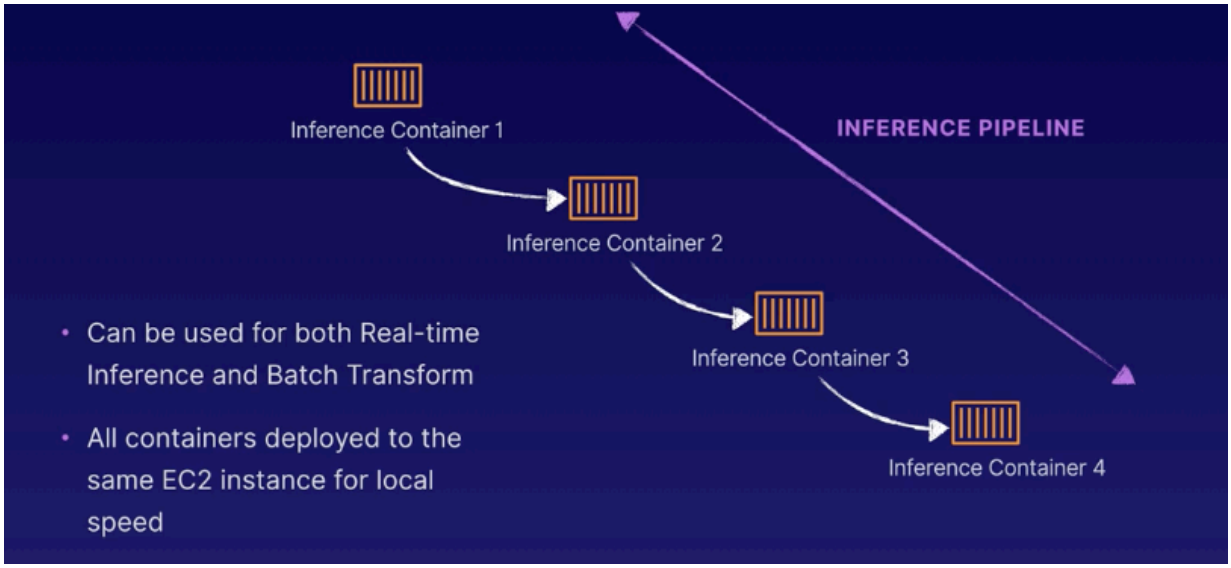
This is common to evaluate the whole dataset, like in Forecasting



Inference pipelines



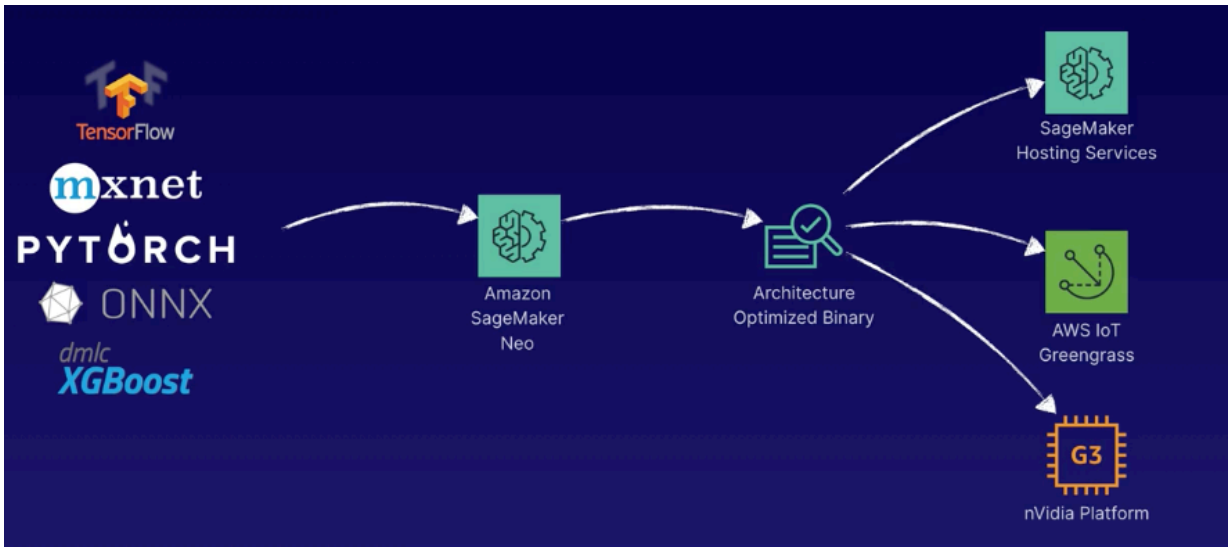




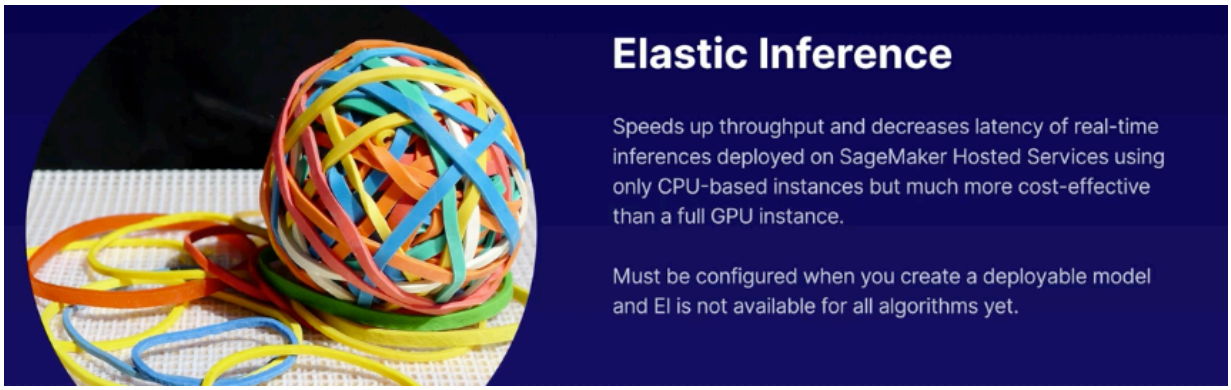
SageMaker Neo

Neo is a compiler and runtime component that allows us to optimize models for specific architectures => raspberry pie, IoT device,...

If we used a model with MXnet, we can use Neo to optimize this model to run on an IoT device or Nvidia cud platform without changing anything about the model

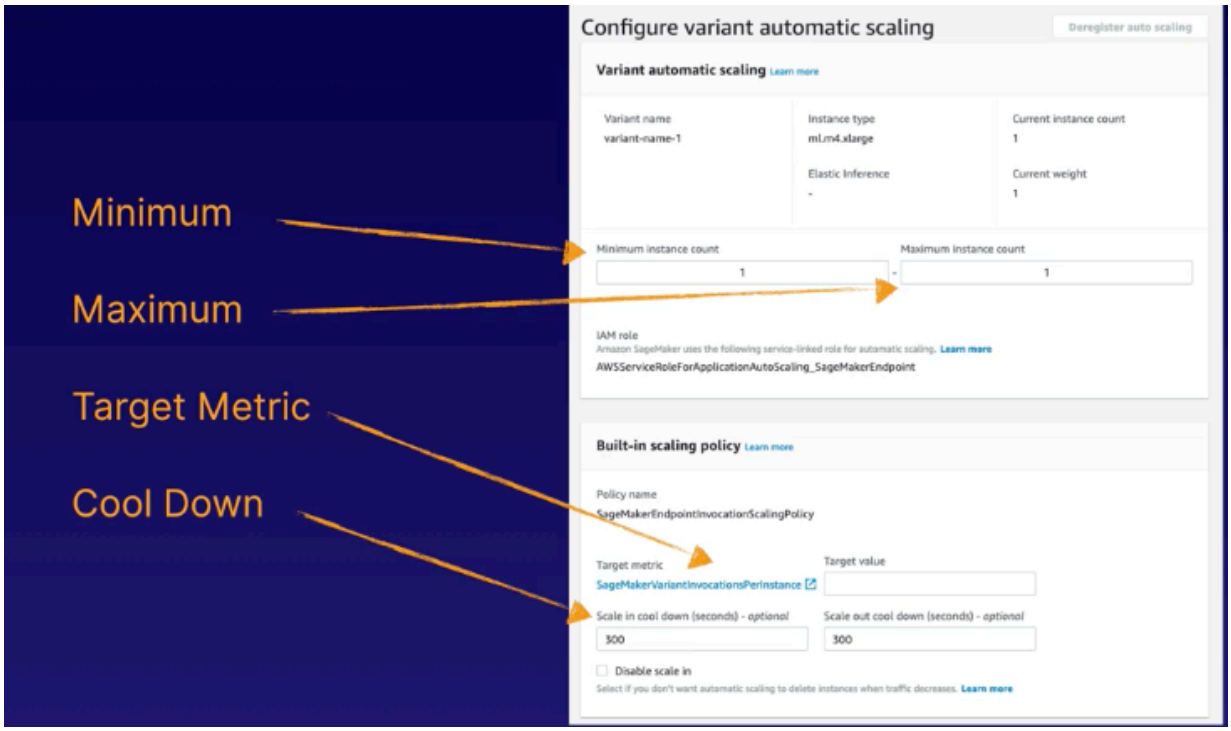


Elastic Inference = Turbo boos

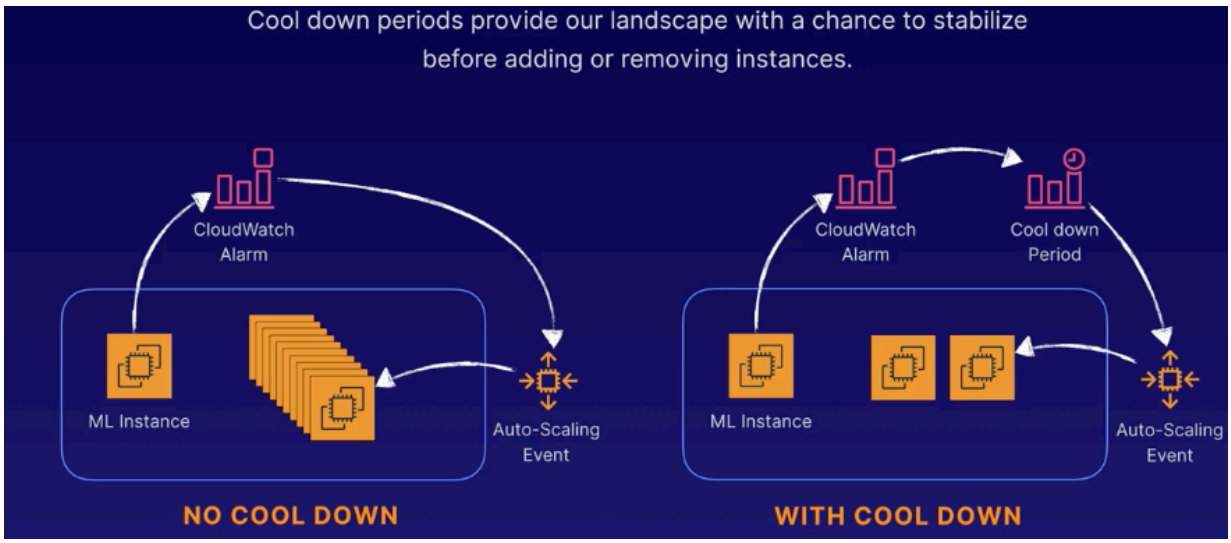


SageMaker supports **Automatic Scaling**

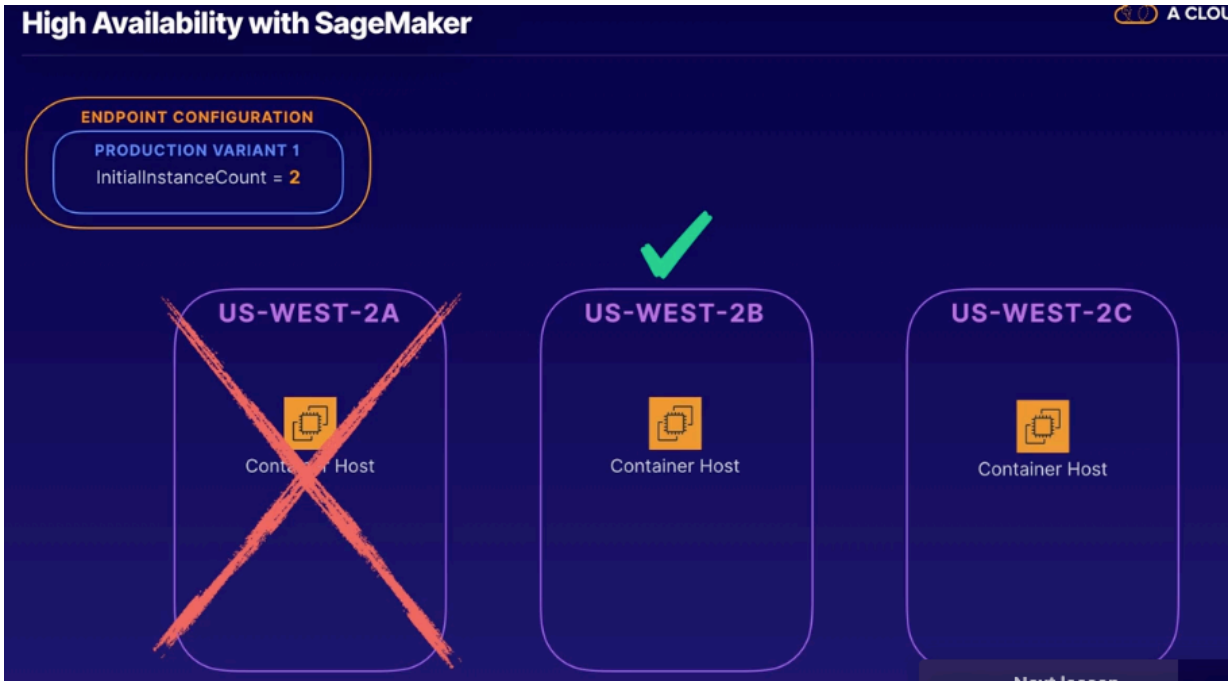
We define a Target Metric we want SM to watch



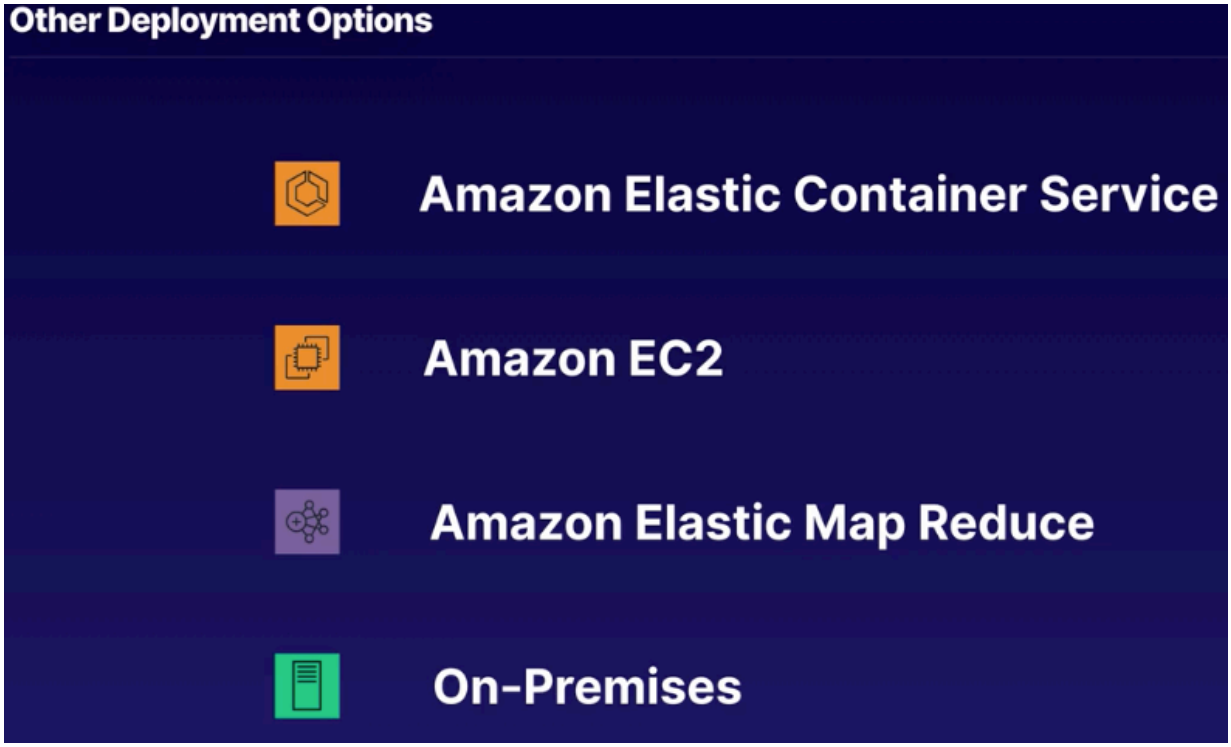
Cool Down period - to smooth scale in and scale up (buffer for newly introduced instances before removing them or adding more)

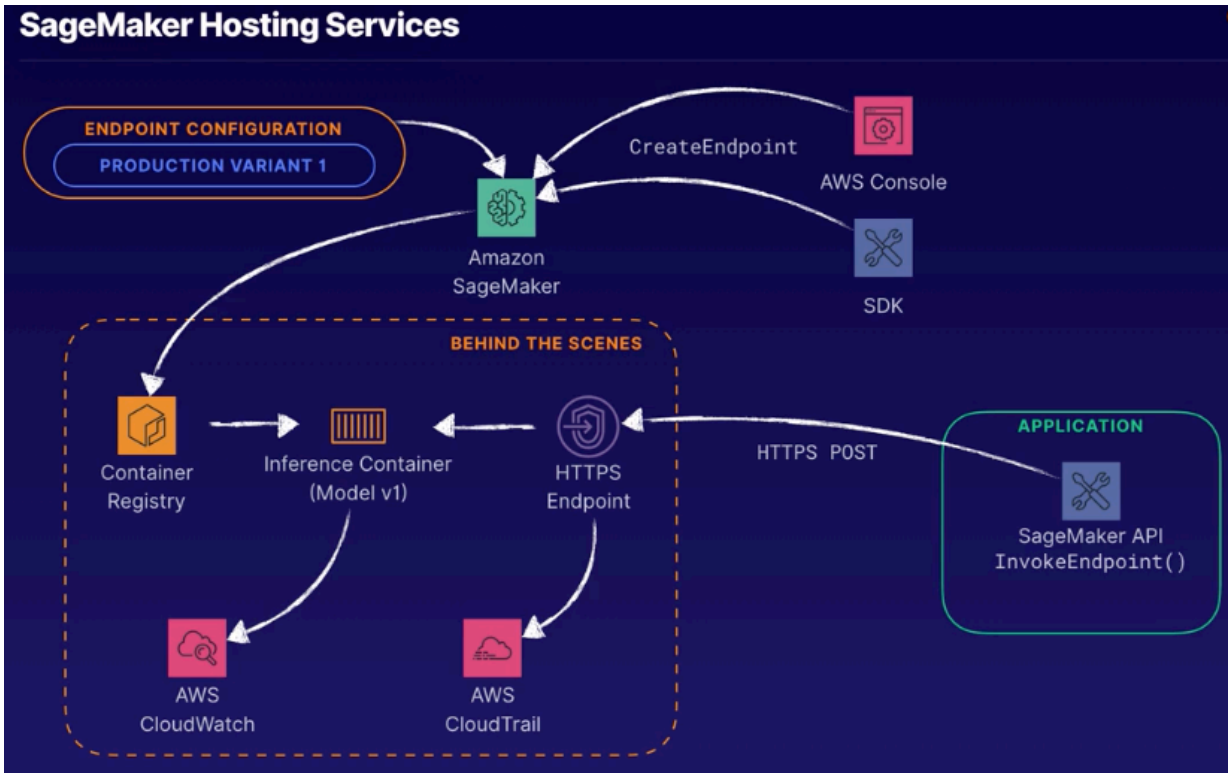


SM recommends we set up multiple instances for high availability

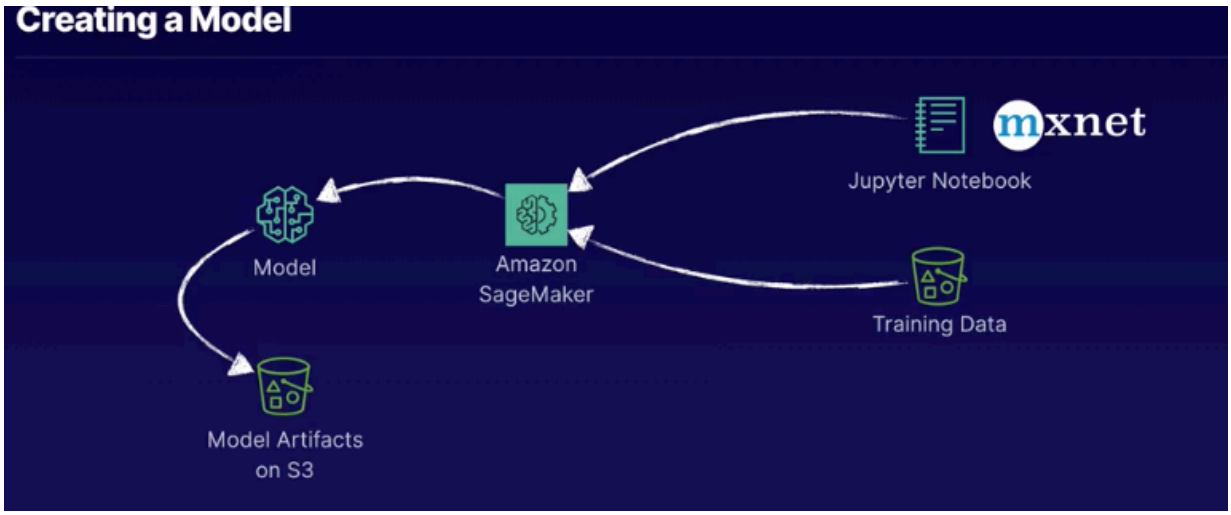


Other ML Deployment options

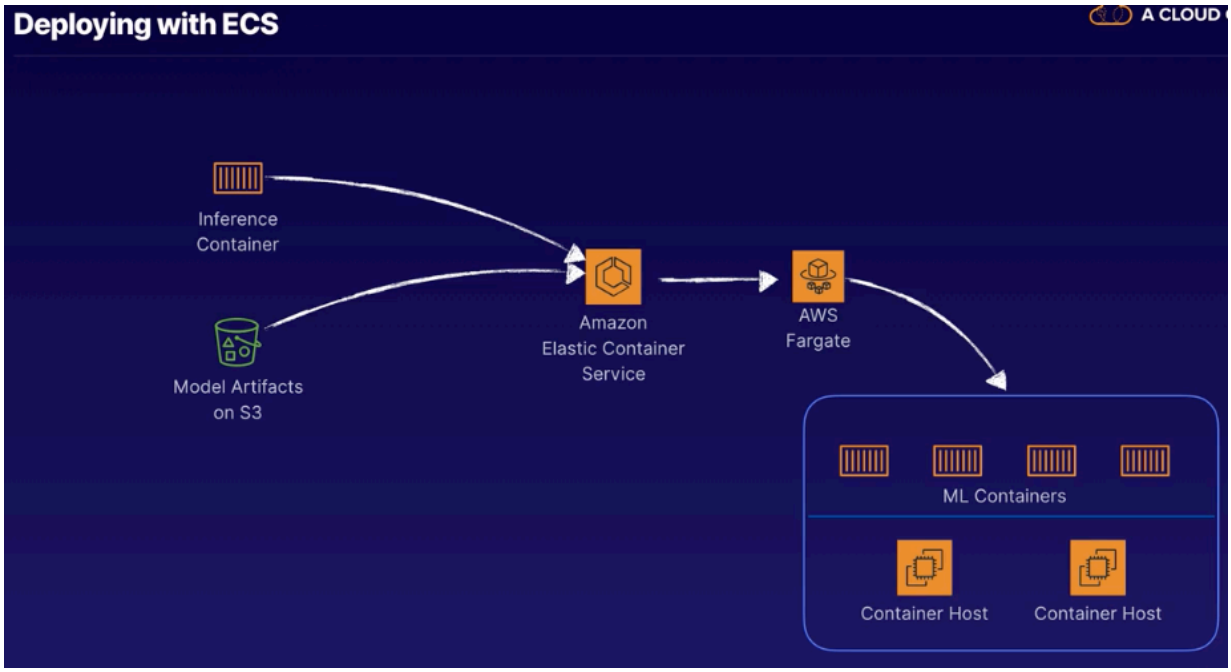




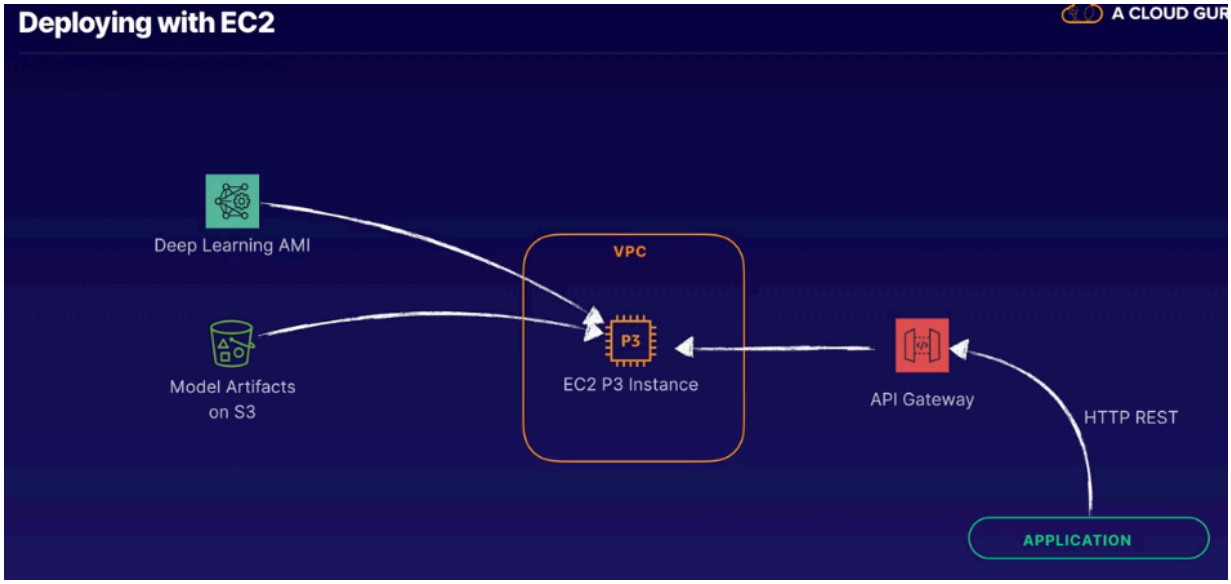
To deploy to other places, it stats the same way:



ECS deployment:

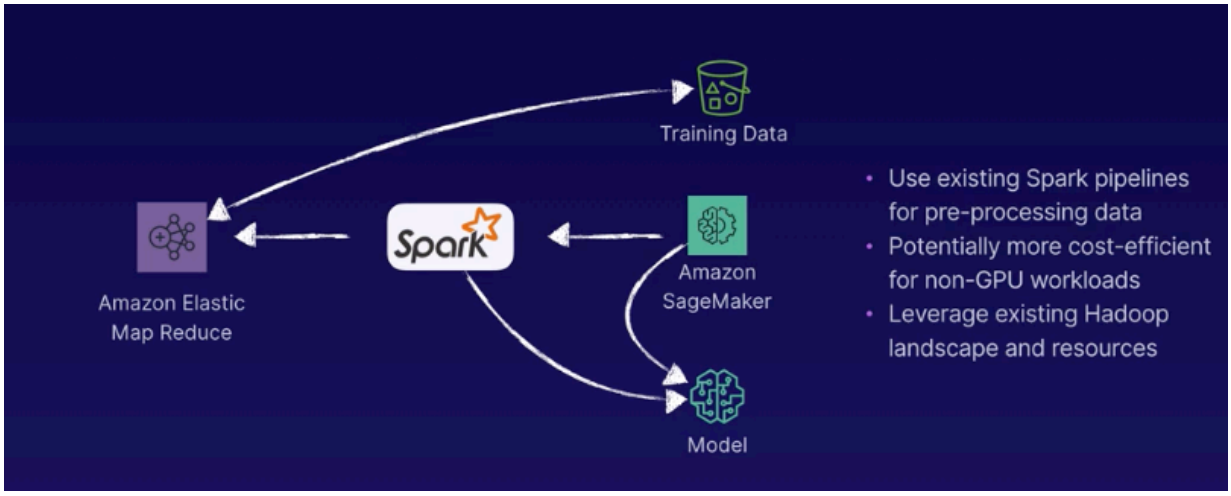


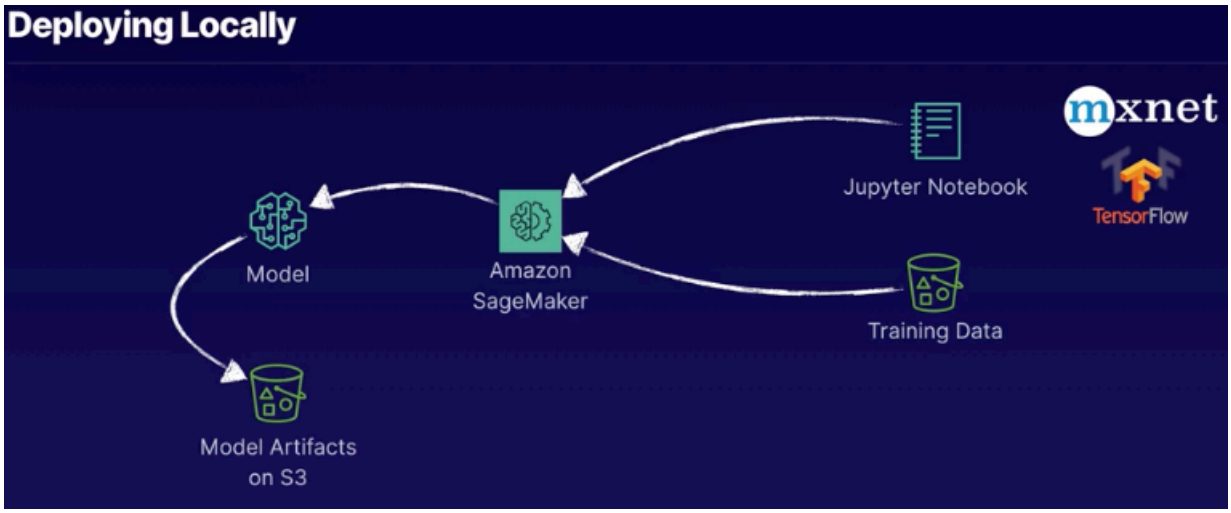
EC2 deployment:



Deploying toe EMR

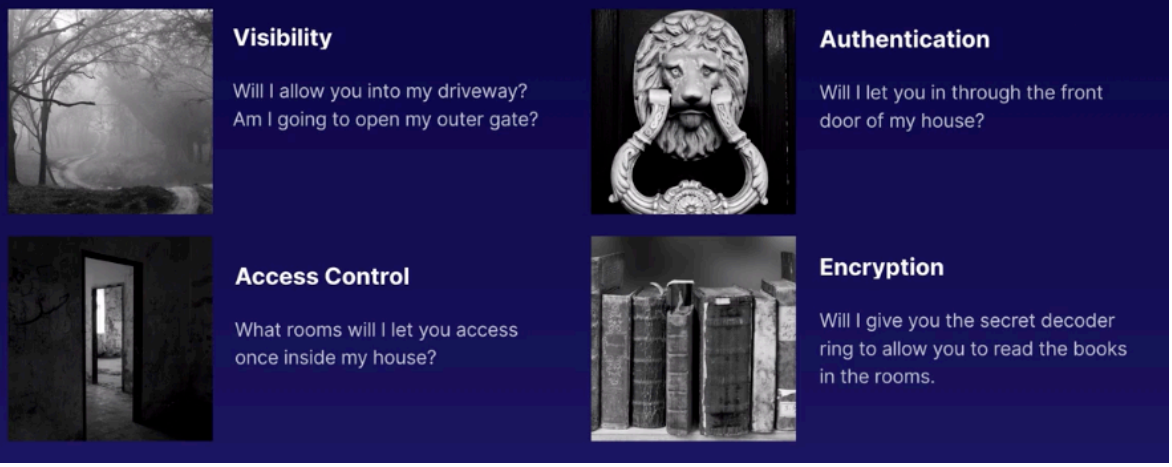

Deploying with EMR (Spark)
Model
Model Artifacts
on S3
Spark
spark-submit
--packages com.amazonaws:aws-java-sdk:1.11.238 \
--deploy-mode cluster \
--conf spark.driver.userClassPathFirst=true \
--conf spark.executor.userClassPathFirst=true \
--jars SageMakerSparkApplicationJar.jar,...
...
Amazon Elastic
Map Reduce


Training Data
Amazon Elastic
Map Reduce
Spark
Amazon
SageMaker
Model
• Use existing Spark pipelines for pre-processing data
• Potentially more cost-efficient for non-GPU workloads
• Leverage existing Hadoop landscape and resources

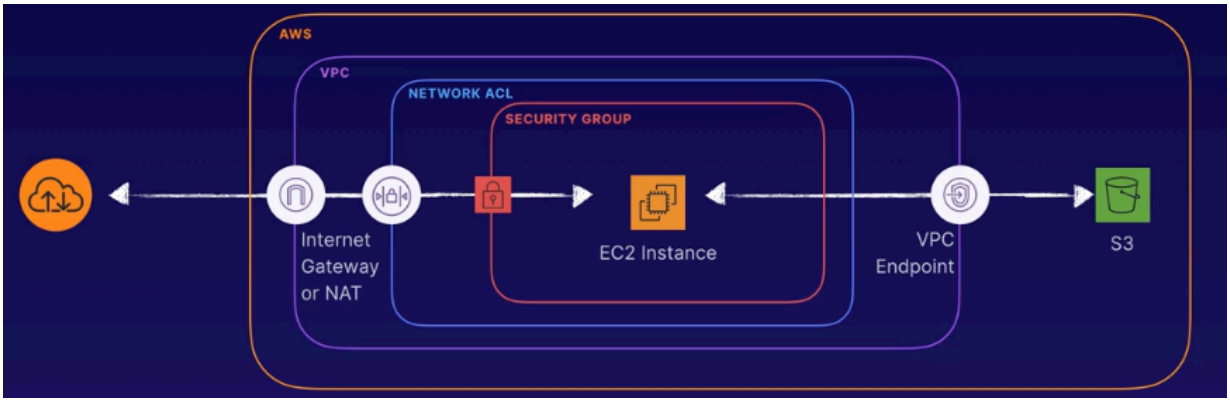

Deploying Locally
Jupyter Notebook
mxnet
TensorFlow
Model
Amazon
SageMaker
Training Data
Model Artifacts
on S3

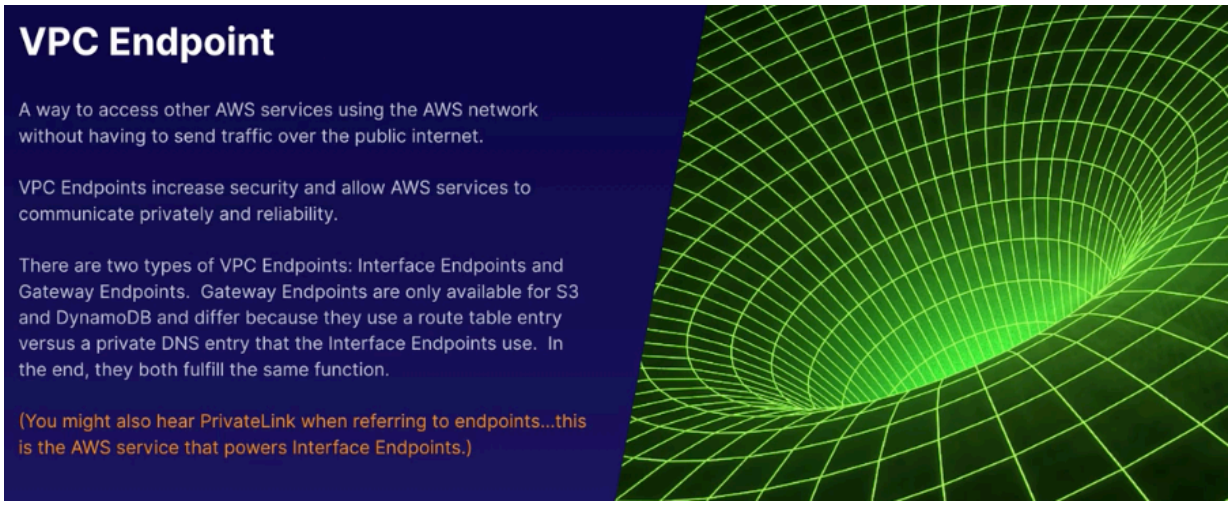
Security

Using scary house as an analogy to security

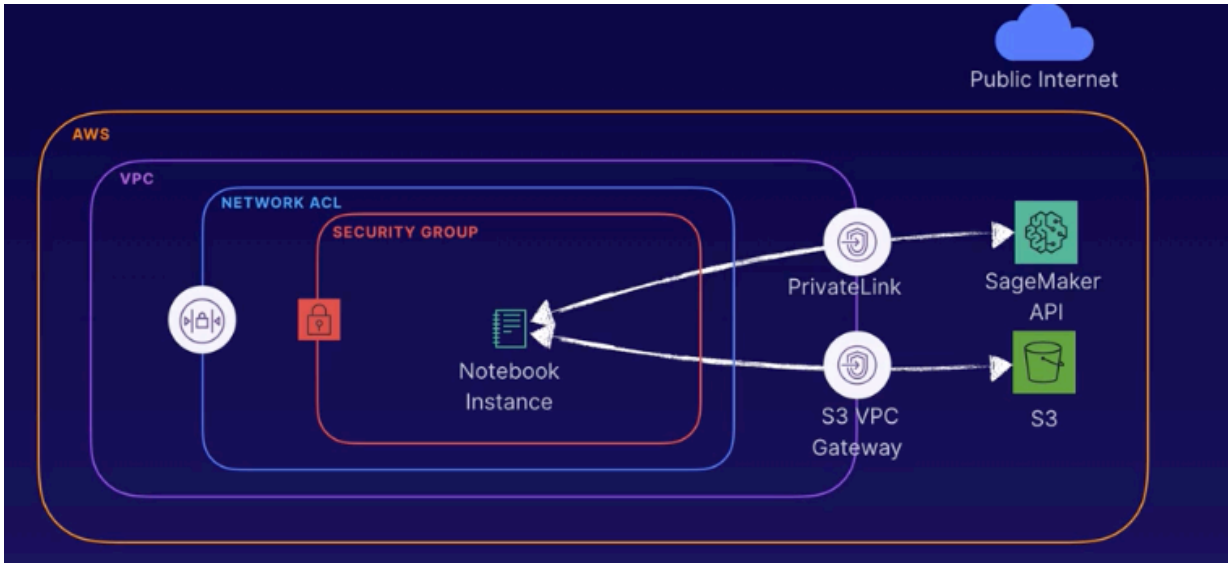

Visibility
Will I allow you into my driveway?
Am I going to open my outer gate?
Authentication
Will I let you in through the front door of my house?
Access Control
What rooms will I let you access once inside my house?
Encryption
Will I give you the secret decoder ring to allow you to read the books in the rooms.


Visibility
Will I allow you into my driveway?
Am I going to open my outer gate?
VPC Endpoints, Network Access Control Lists, Security Groups
Authentication
Will I let you in through the front door of my house?
Identity and Access Management
Access Control
What rooms will I let you access once inside my house?
Identity and Access Management
Encryption
Will I give you the secret decoder ring to allow you to read the books in the rooms.
Key Management Service

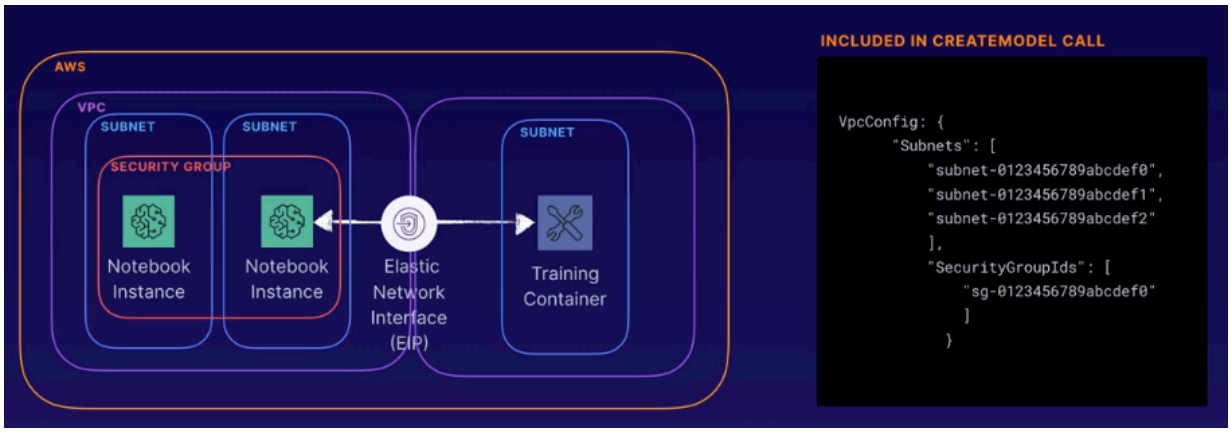

AWS
VPC
NETWORK ACL
SECURITY GROUP
Internet Gateway or NAT
EC2 Instance
VPC Endpoint
S3



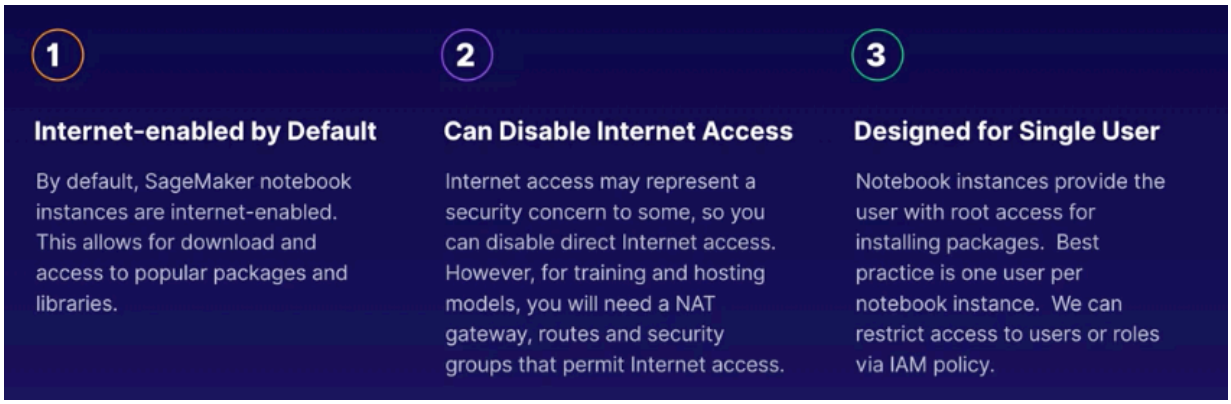
Securing ML resources, not going over Public internet



When we create a model for the first time, we can tell SM which subsets and security groups we want the training job to use
=> use EIP - Elastic Network Interface



Notebook instances



2 types of IAM policies:
- Identity-based
- Resource-based

	Identity-based Policy	Resource-based Policy
What	A permissions policy attached to IAM identities.	A permissions policy attached to a resource.
Why	Allow or deny access to a role or user.	Allow or deny access at the resource.
When	Only allow user “mary” to access the notebook instance “Mary_Notebook”.	Restrict R/W access to an S3 bucket by specific AWS accounts.

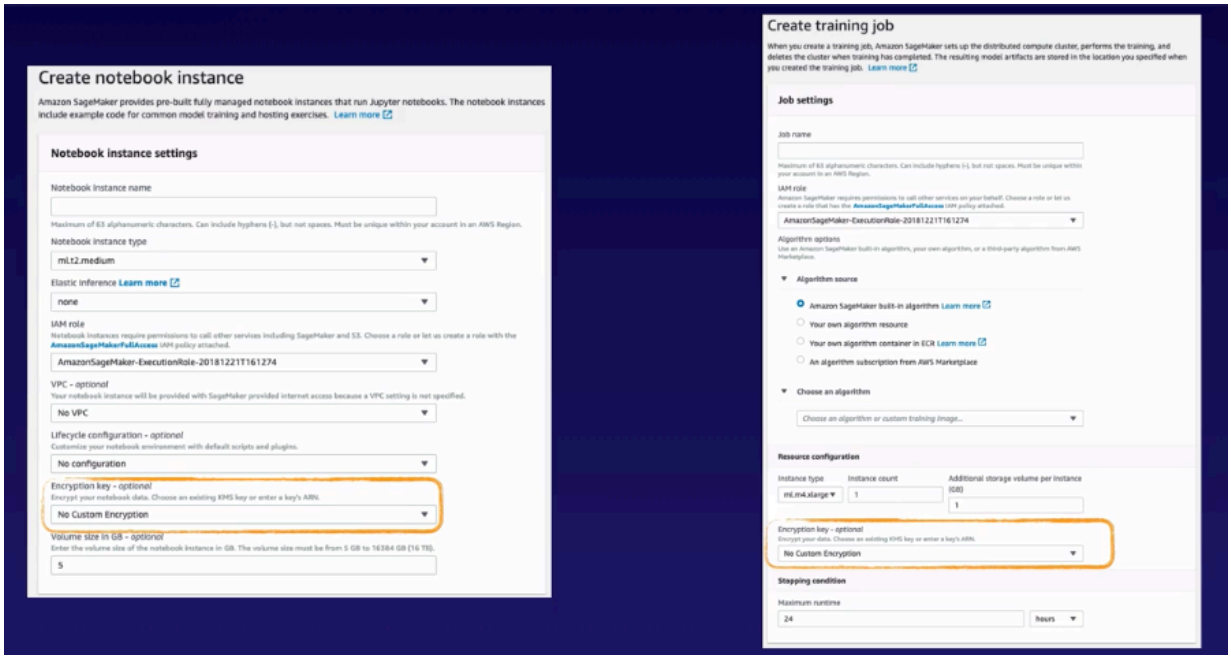
Example of IAM permission policy that we can attach to an IAM user

PassRole action => user of this policy has the ability to hand of their role to SM service

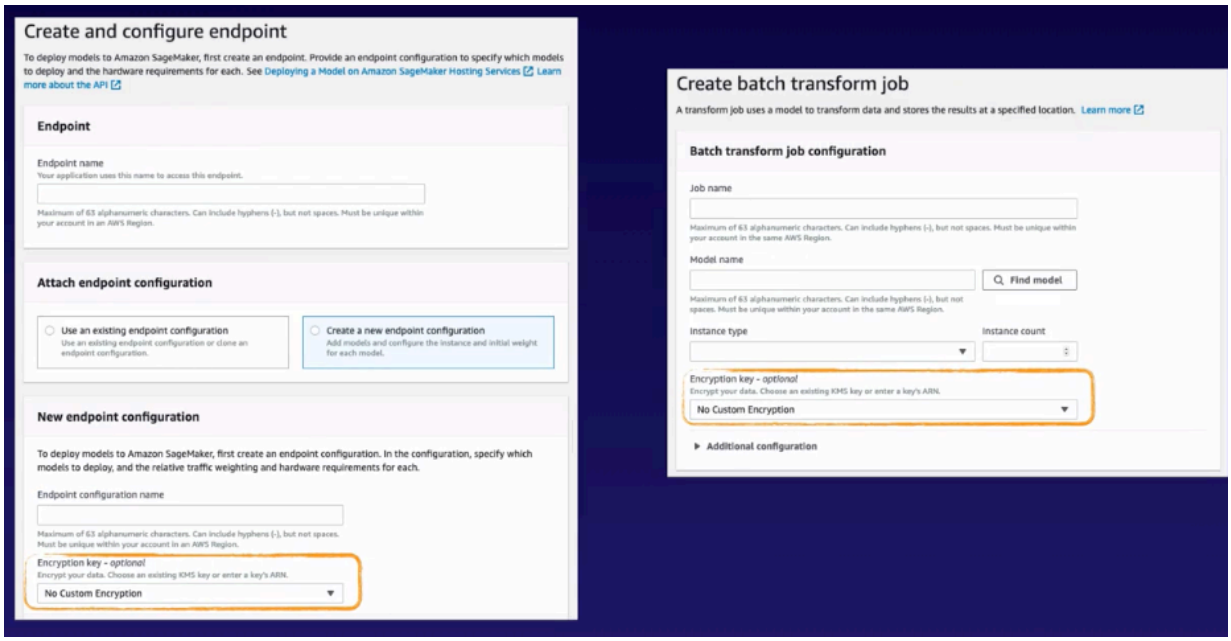
2 types of encryption:
- at rest
- in transit

	What	Generic Example	AWS Example
Encryption At Rest	Store data in an encrypted format	Use GPG to encrypt a file.	Configure S3 bucket to use S3-KMS encryption.
Encryption in Transit	Encrypt the data stream carrying the data	Using the TLS to encrypt an HTTP connection (HTTPS).	Use Certificate Manager with CloudFront to provide TLS support for a custom domain.

SM offers options to encrypt for notebooks and training jobs

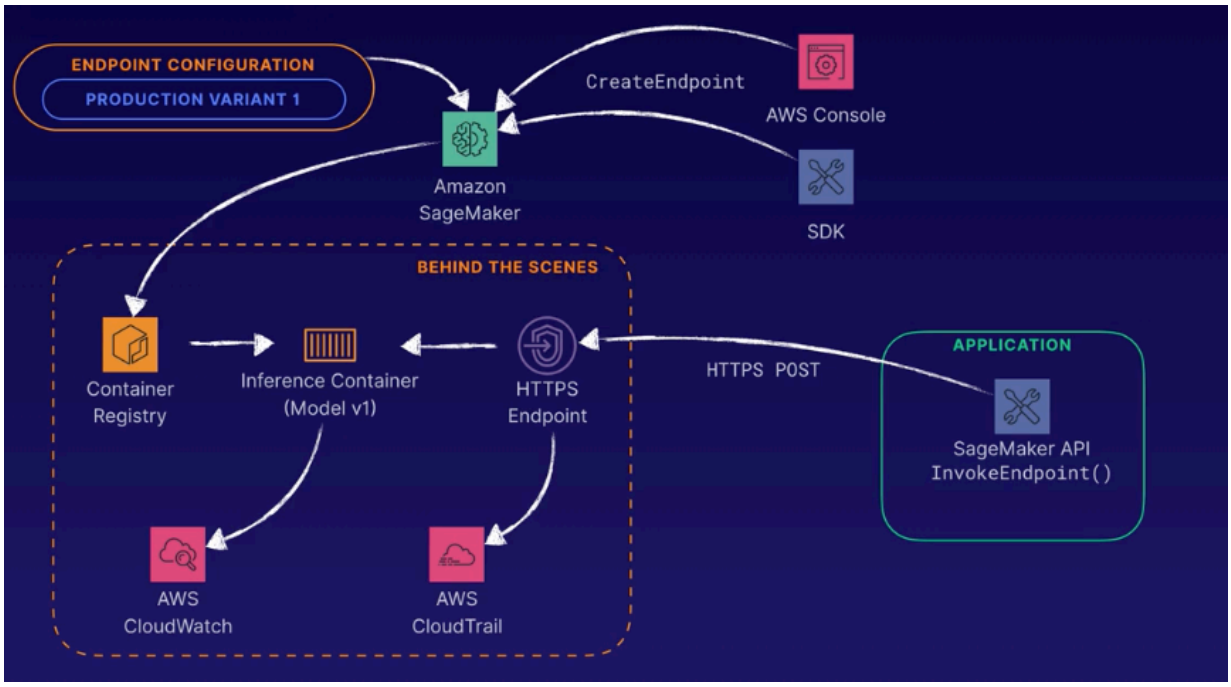


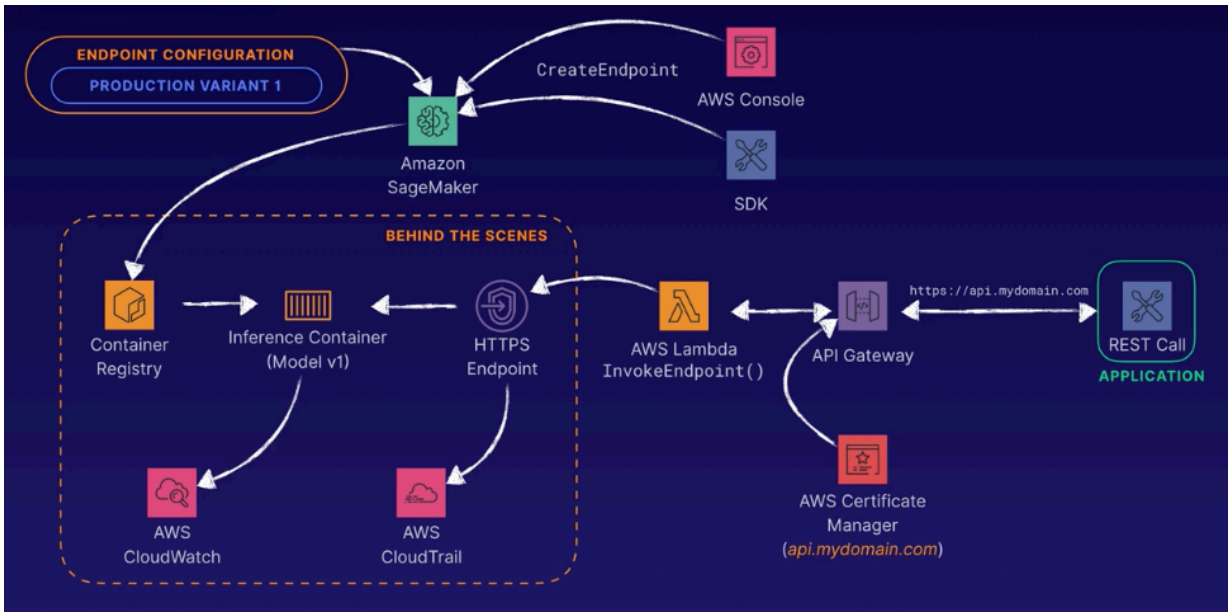
Same with end points or batch transform



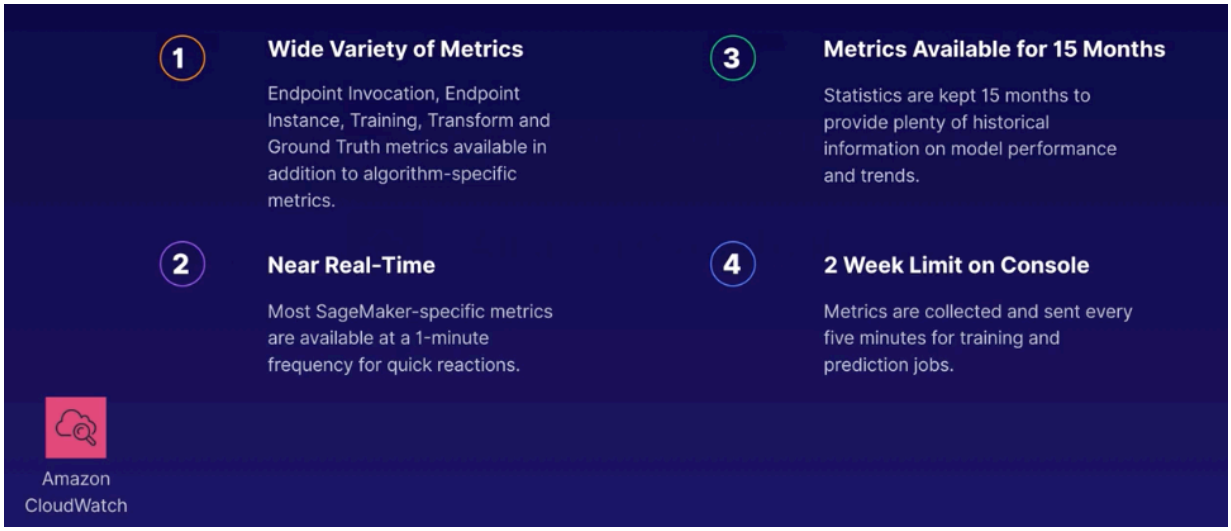
All these examples are encryption AT REST

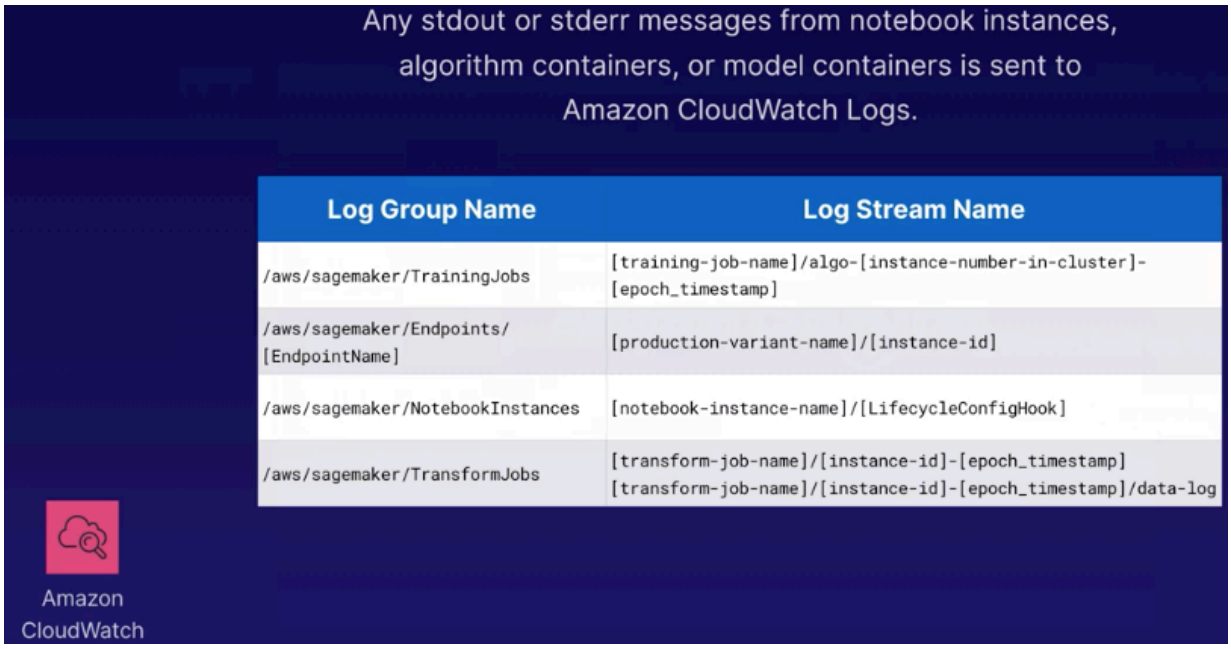
For encryption in TRANSIT: using httpS



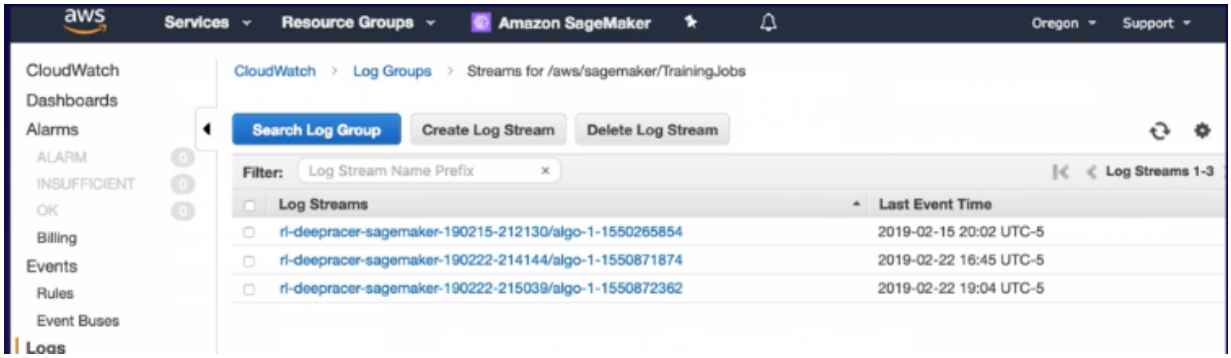


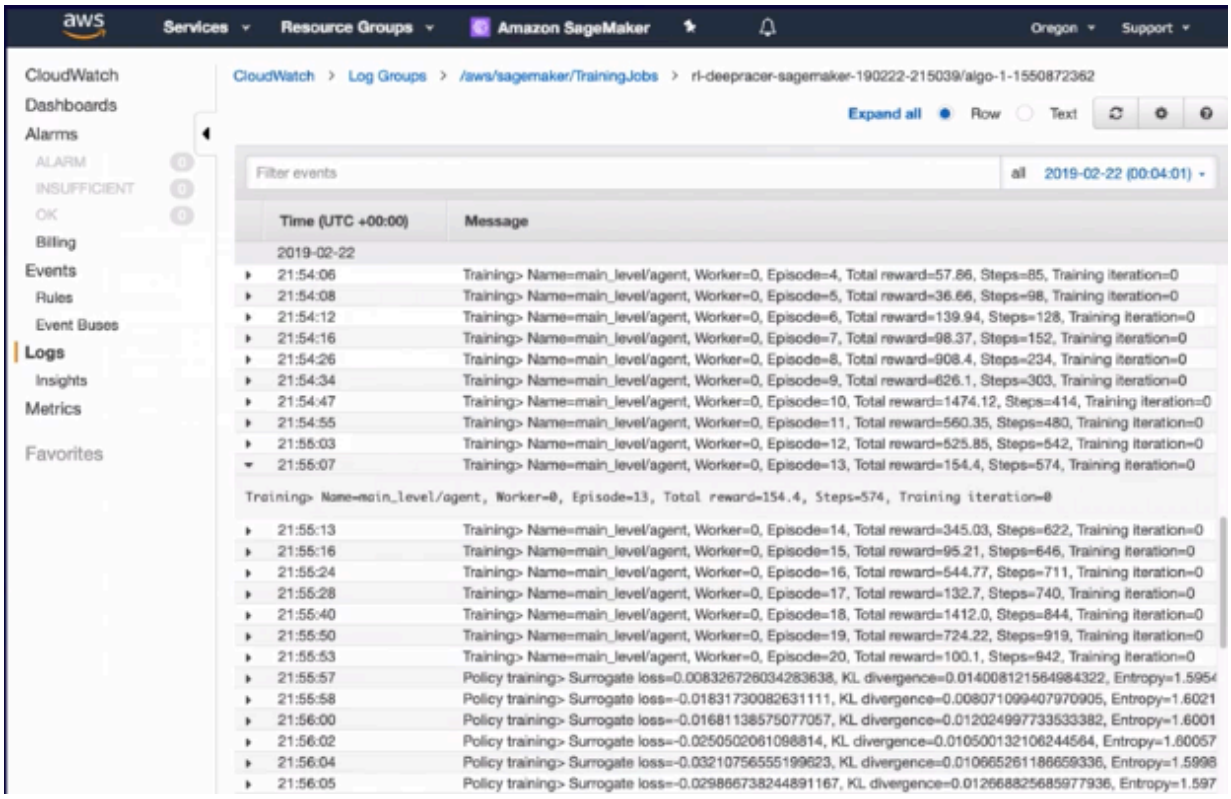
Monitor and Evaluate



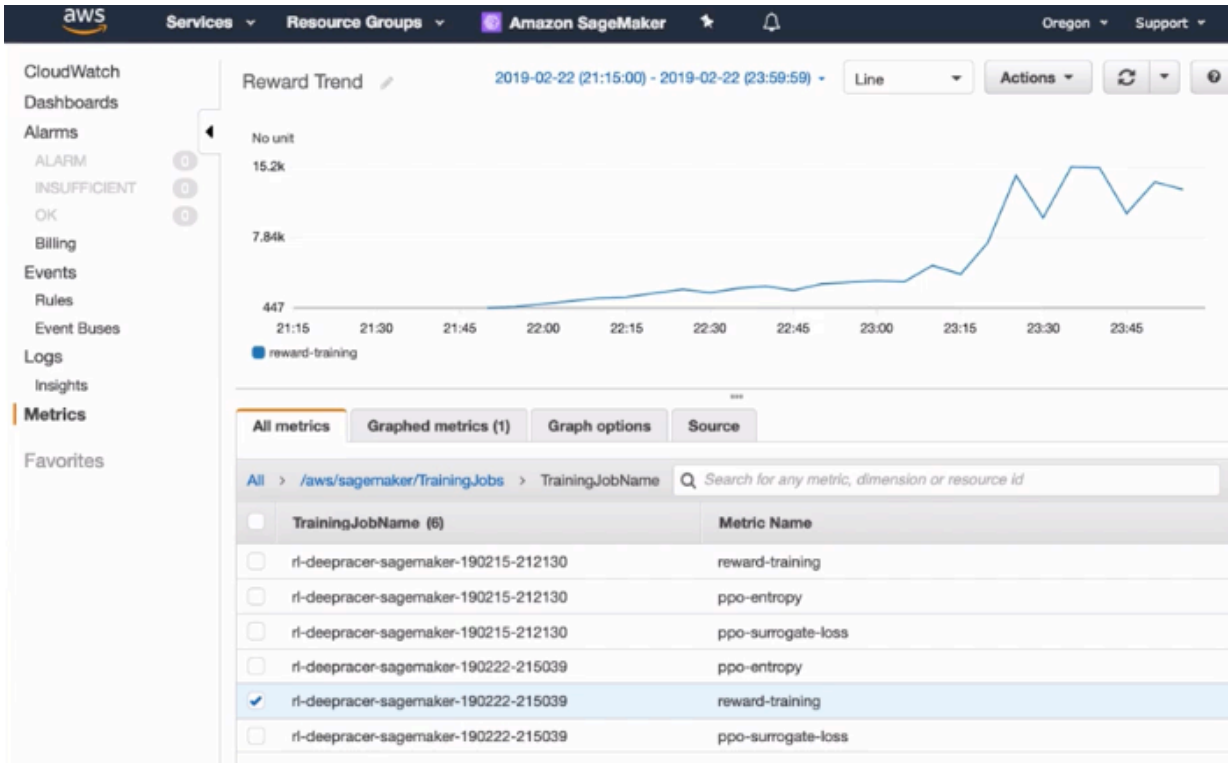


Example with deepracer

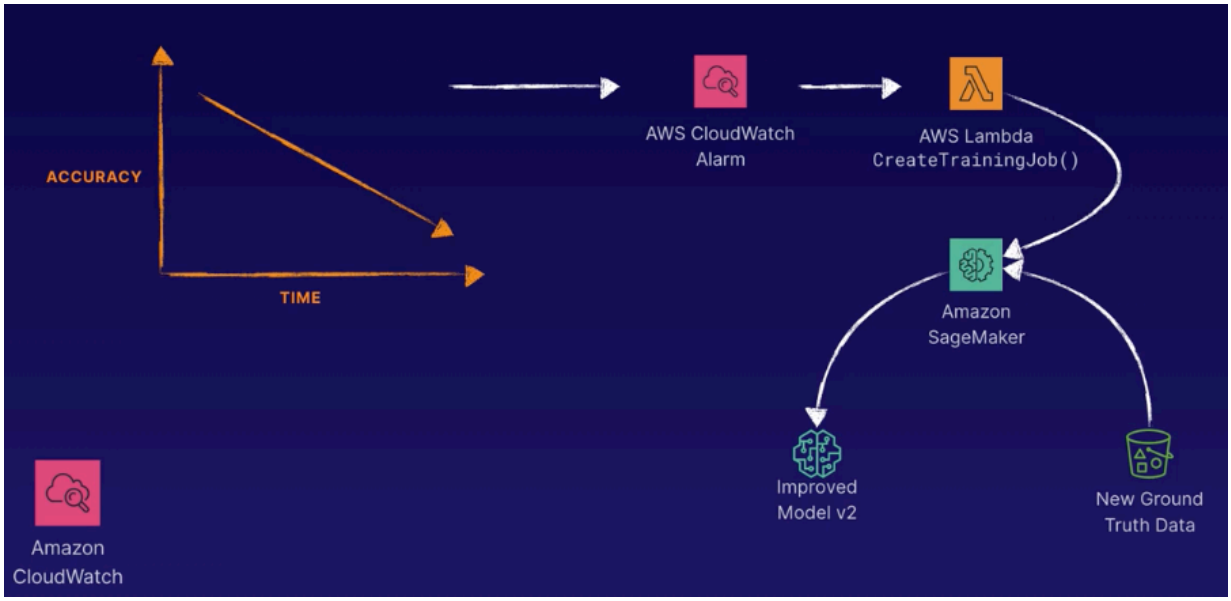




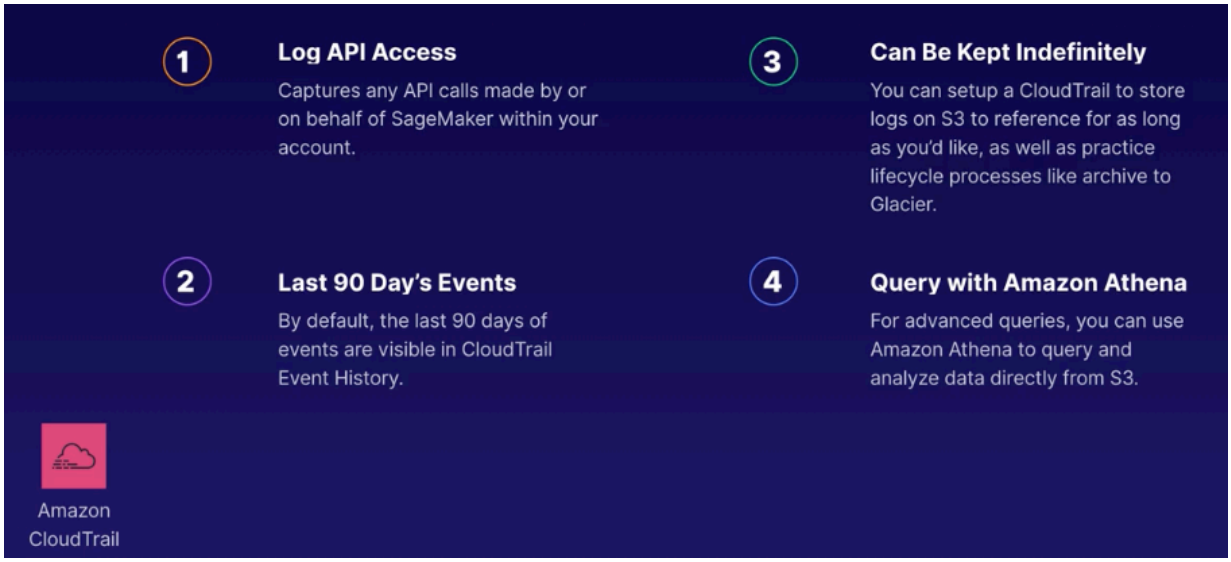
Can also view some metrics in graph format - reward function below



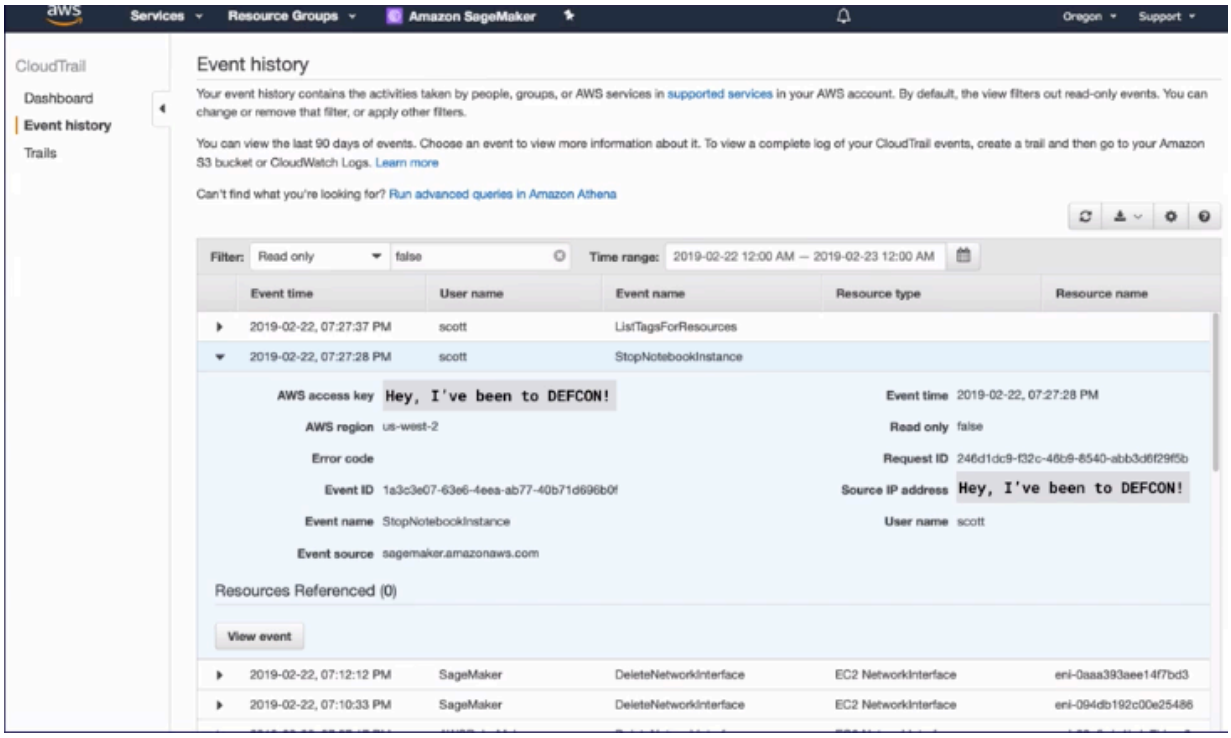
Re-Training
Over time, our model may be less accurate
We can use Cloudwatch alarms to trigger lambda which can initiate another training job

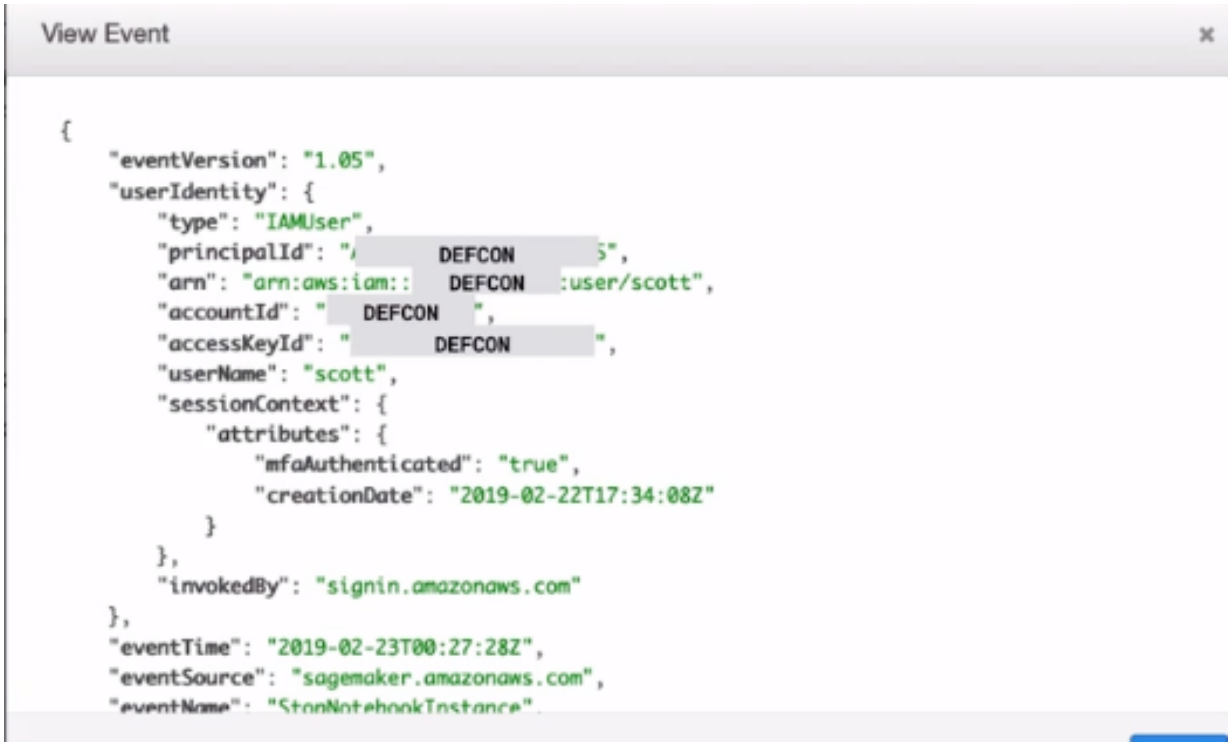


Amazon CloudTrail
Captures API access in/out of the account



Example of Cludtrail entry - stopping a notebook instance


Services
Resource Groups
Amazon SageMaker
Oregon
Support
CloudTrail
Dashboard
Event history
Trails
Event history
Your event history contains the activities taken by people, groups, or AWS services in supported services in your AWS account. By default, the view filters out read-only events. You can change or remove that filter, or apply other filters.
You can view the last 90 days of events. Choose an event to view more information about it. To view a complete log of your CloudTrail events, create a trail and then go to your Amazon S3 bucket or CloudWatch Logs. Learn more
Can't find what you're looking for? Run advanced queries in Amazon Athena
Filter: Read only false
Time range: 2019-02-22 12:00 AM — 2019-02-23 12:00 AM
Event time
User name
Event name
Resource type
Resource name
2019-02-22, 07:27:37 PM scott ListTagsForResources
2019-02-22, 07:27:28 PM scott StopNotebookInstance
AWS access key Hey, I've been to DEFCON!
AWS region us-west-2
Error code
Event ID 1a3c3e07-63e6-4eea-ab77-40b71d696b0f
Event name StopNotebookInstance
Event source sagemaker.amazonaws.com
Event time 2019-02-22, 07:27:28 PM
Read only false
Request ID 246d1dc9-f32c-46b9-8540-abb3d6f29f5b
Source IP address Hey, I've been to DEFCON!
User name scott
Resources Referenced (0)
View event
2019-02-22, 07:12:12 PM SageMaker DeleteNetworkInterface EC2 NetworkInterface eni-0aaa393aee14f7bd3
2019-02-22, 07:10:33 PM SageMaker DeleteNetworkInterface EC2 NetworkInterface eni-094db192c00e25486


View Event
{
"eventVersion": "1.05",
"userIdentity": {
"type": "IAMUser",
"principalId": " DEFCON ",
"arn": "arn:aws:iam:: DEFCON :user/scott",
"accountId": " DEFCON ",
"accessKeyId": " DEFCON ",
"userName": "scott",
"sessionContext": {
"attributes": {
"mfaAuthenticated": "true",
"creationDate": "2019-02-22T17:34:08Z"
}
},
"invokedBy": "signin.amazonaws.com"
},
"eventTime": "2019-02-23T00:27:28Z",
"eventSource": "sagemaker.amazonaws.com",

Exam tips

Concepts

- Understand the difference between offline and online usage for models and when you might use each.
- Have a high-level understanding of the types of deployments and respective pros and cons.
- Understand A/B Testing and how it might be used to introduce a newly updated model.
- Understand the concepts behind continuous integration, continuous delivery and continuous deployment.

AI Developer Services

- The AI developer services provided by AWS are scalable, fault-tolerant and ready to use.
- Know each service, its purpose and in what use-cases they might be used.
- Experiment with each service via the Console or CLI to get a feel for what it does and how it works.

Amazon SageMaker Deployments

- Know the three main steps in creating a deployment using SageMaker Hosting Services.
- Understand ProductionVariants and how they can be used to introduce new evolutions of your models.
- Know how to calculate the percentage of traffic given weights for each production variant.
- Understand the purpose and limitations of Interface Pipelines, SageMaker Neo, Elastic Inference, Auto-Scaling and Cool Down
- Recall best practice suggestions for high availability with SageMaker.

Other ML Deployment Options

- Know the four other options for ML deployment aside from SageMaker Hosting Services and when you might chose them instead of SageMaker.

Security

- Be familiar with using VPCs, NACLs, SGs, IAM and KMS to secure Amazon ML and SageMaker resources.
- Understand a VPC Endpoint and why it increases security.
- Know that Notebook Instances are Internet-enabled by default but this can be disabled given certain conditions and introducing limitations.

Monitor and Evaluate

- Understand the difference between CloudWatch and CloudTrail.
- Know that CloudWatch has a limited storage while CloudTrail can have unlimited storage if you log to an S3 bucket.
- Understand how you might use metrics to trigger events like re-training.