Patronus

Capsule8 ML Platform

Phase 1 Overview

June 2020

CAPSULE8

Why do we need our own ML platform ?

- Increase speed of iteration of ML projects
- All teams (outside of data science) should have access to data products and the ability to conduct quick experiments with sensor data
- A launch pad for eventual deployment of ML strategies for cross node behavioral analytics

CAPSULE8

Project Objective

- Provide a fast, repeatable and scalable ML platform
- Provides a repeatable pipeline for rapidly developing and testing ML models in internal C8 clusters
- Sets up the stage for designing/ extending this platform for deployment in production clusters in client environment

CAPSULE8

What is Sagemaker ?

- Fully managed AWS ML service
- Designed for speed and iteration
- Higher level API for working with pre-optimized ML frameworks, like MXNet, Tensorflow, Scikit-learn and XGBoost
- **Lower level API** that allows running custom jobs
- Any library or API we can fit in a docker image can be used with sagemaker

CAPSULE8

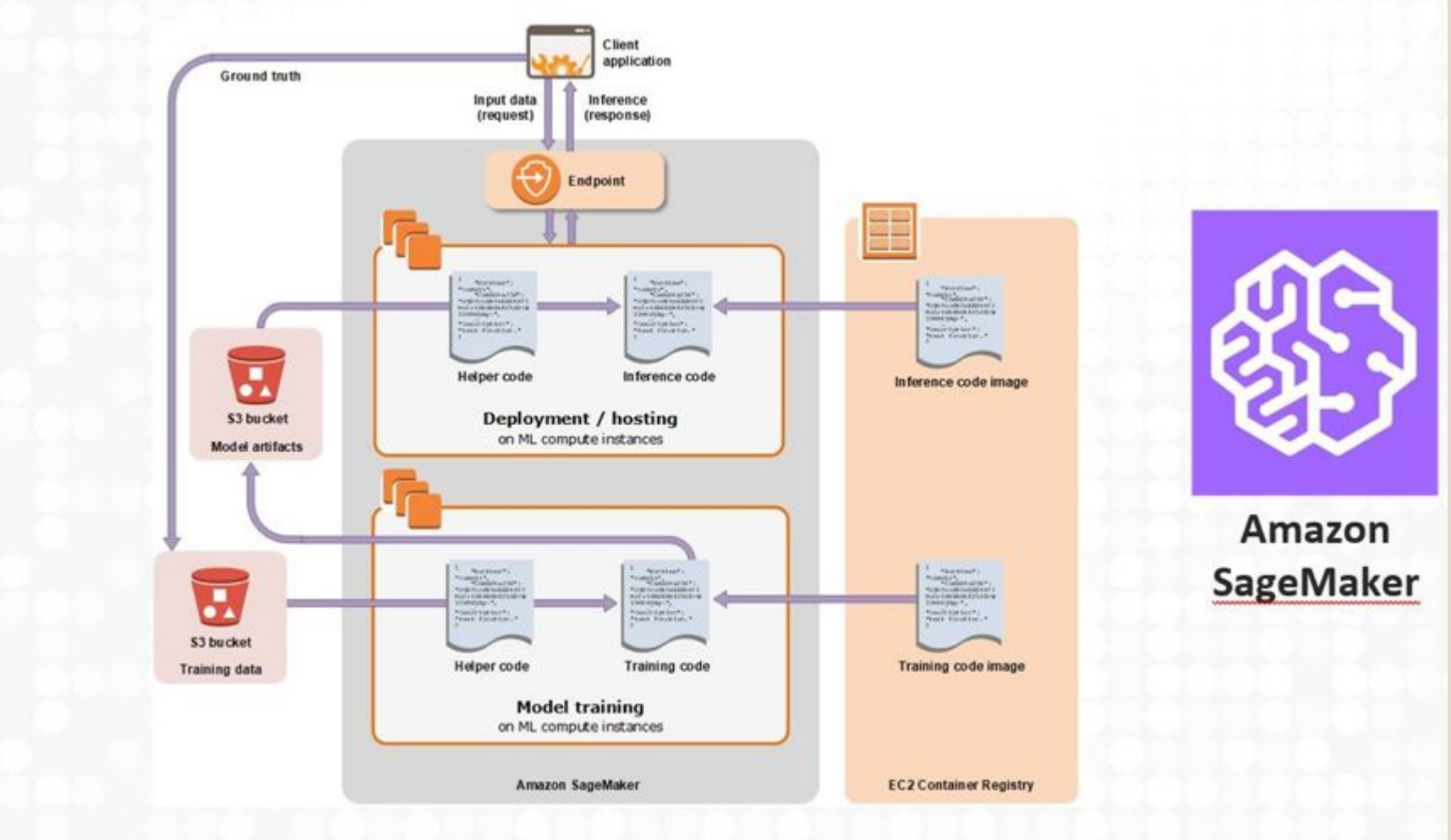
Sagemaker Architecture

AWS Sagemaker builds on top of other AWS services:

- S3 (blob storage)
- ECR (Docker registry)
- EC2 (compute)

CAPSULE8

Sagemaker Architecture



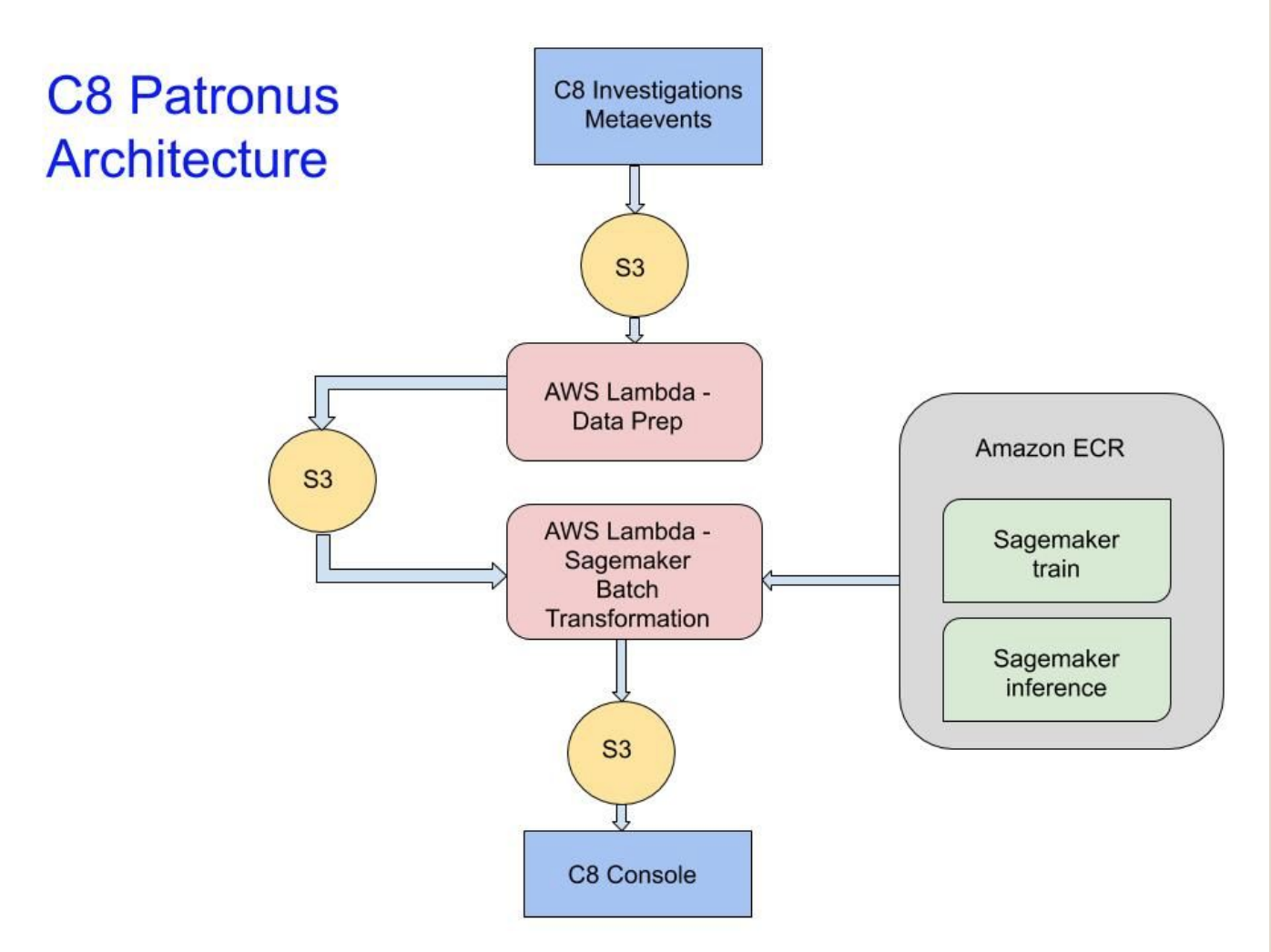
CAPSULE8

Sagemaker Workflow

1. Wrapping model training/ serving pipeline in a sagemaker-compatible docker image
2. Uploading that image to Amazon ECR
3. **Using sagemaker python api** to schedule the train/inference pipeline as a job on EC2 cluster
4. Saving the model artifacts to S3 bucket
5. Consuming the trained estimator as a web service or a batch job

CAPSULE8

Patronus Architecture



CAPSULE8

How Patronus works

For deploying a model from C8/patronus to a C8 test cluster:

1. Metaevents from C8 Investigations stored in S3 is the data source

2. Create a Sagemaker instance and integrate it with C8/Patronus repo, so all the training and inference code is made available in sagemaker instance

3. The model training and inference scripts are packaged in a Sagemaker compatible docker image and pushed to ECR

4. Model training is done by calling the Sagemaker estimator function and training on a baseline non malicious dataset from the test cluster

CAPSULE8

How Patronus works (cont.)

- Lambda function 1 that gets triggered everytime a new parquet file is added to this S3 bucket, and does the data munging functions, converting the raw data to model ingestible format. It pushes the cleaned data as CSV to another S3 bucket

- Lambda function 2 gets triggered for every new clean dataset added in the second S3 bucket, uses the pretrained model from the stored model artifacts, and performs batch inference, sending alerts to another S3 bucket

- Lambda function 3 gets triggered everytime a new ML alert is seen and converts them into C8 Console compatible json and pushes them to the Console alerts S3 bucket

CAPSULE8

Patronus in production

- This project should serve as a starting point to think about deploying capsule8 models in production.
- The architecture of Patronus is designed to be as close to online in-prod as possible, as in the continuous development/ deployment of ML models in Capsule8 internal clusters
- Currently this is designed to work in Capsule8's test production env, meaning deployed on our test clusters.
- How would we ship this to a client env is more of a product question. But this should provide a jump start to thinking along those lines.

CAPSULE8

Where can production models live ?

- ML Marketplace
- Capsule8-Console
- Capsule8-Sensor
- Capsule8-X

CAPSULE8