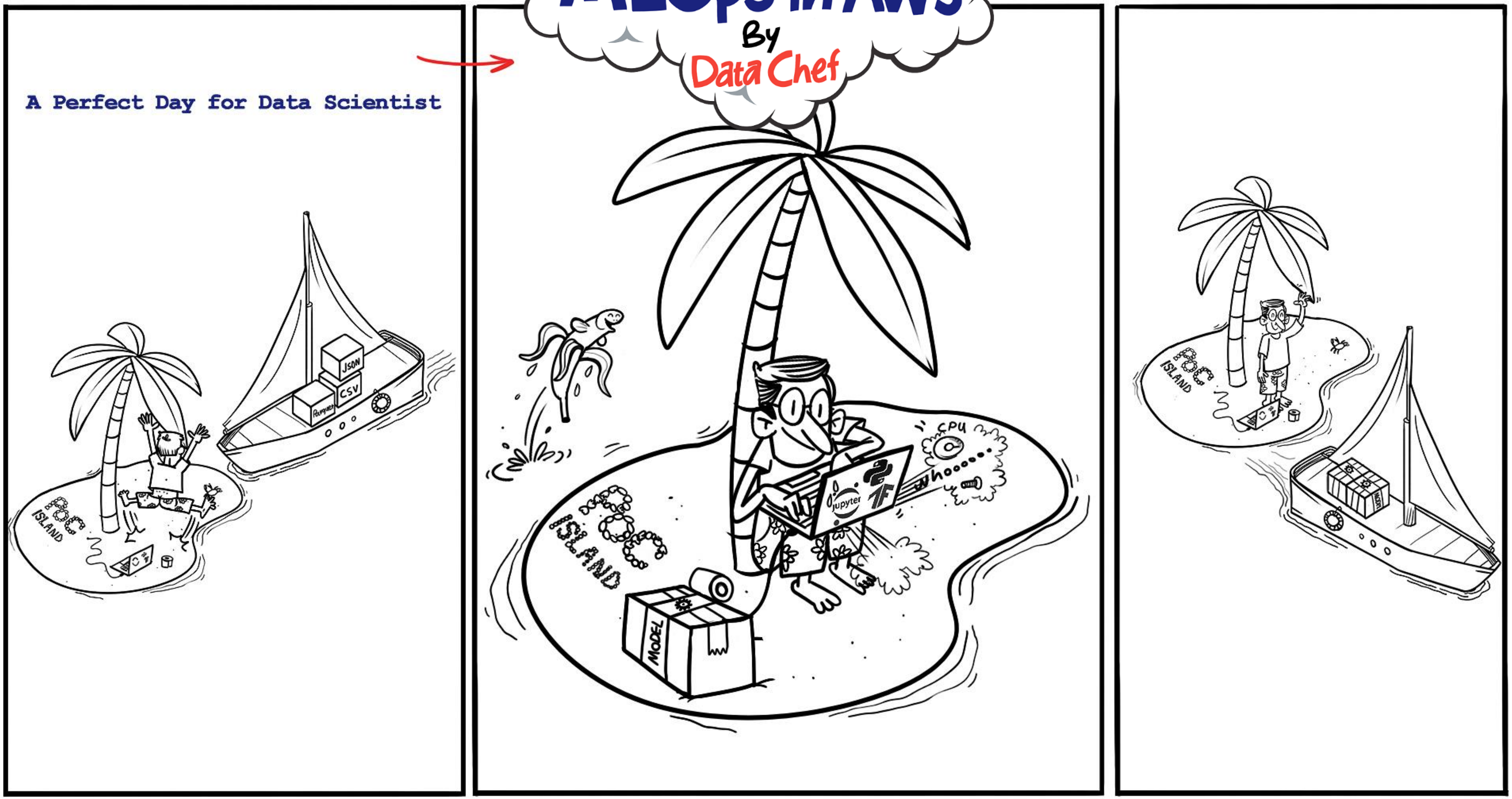
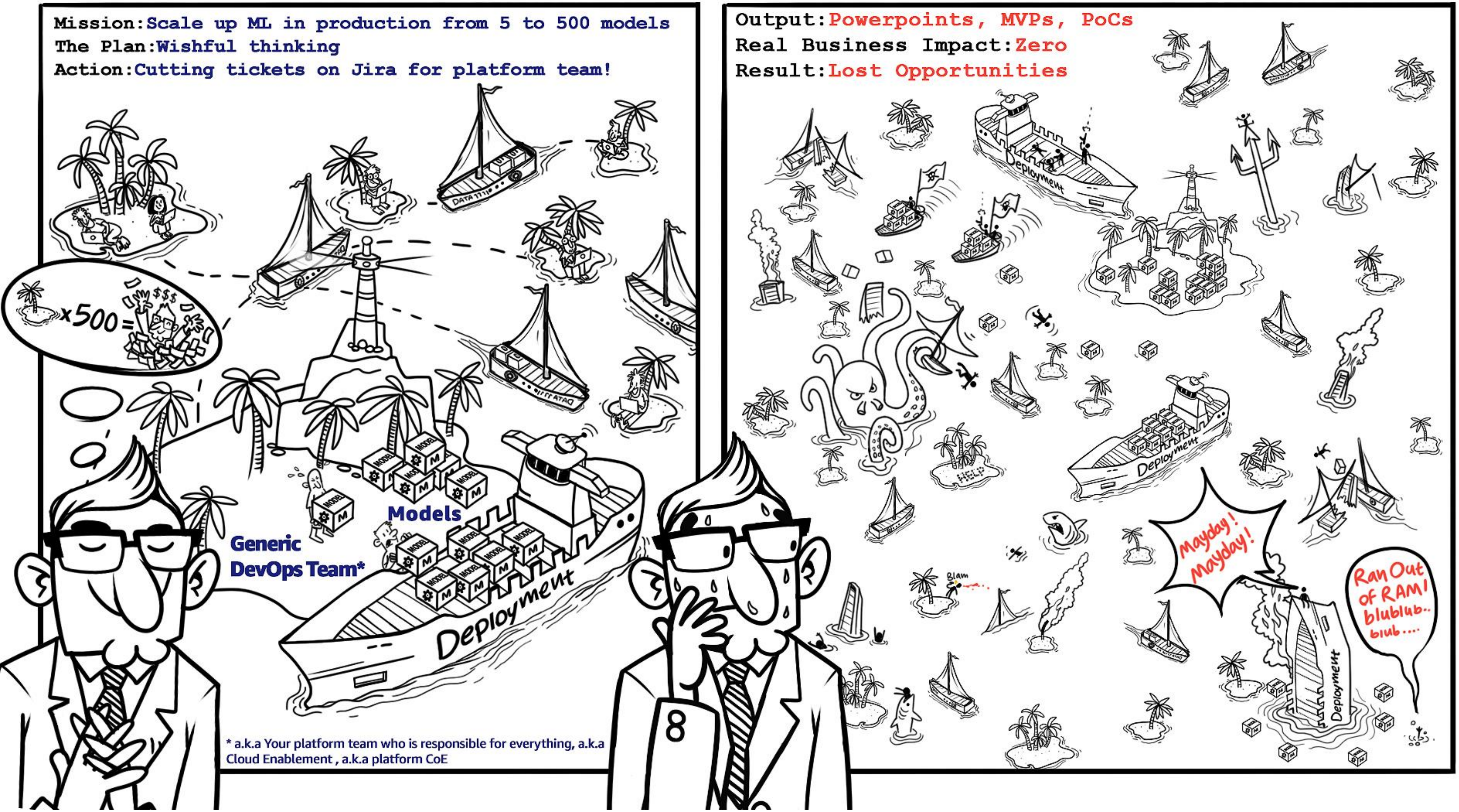
MLOps in AWS

By Data Chef

A Perfect Day for Data Scientist





Mission: Scale up ML in production from 5 to 500 models

The Plan: Wishful thinking

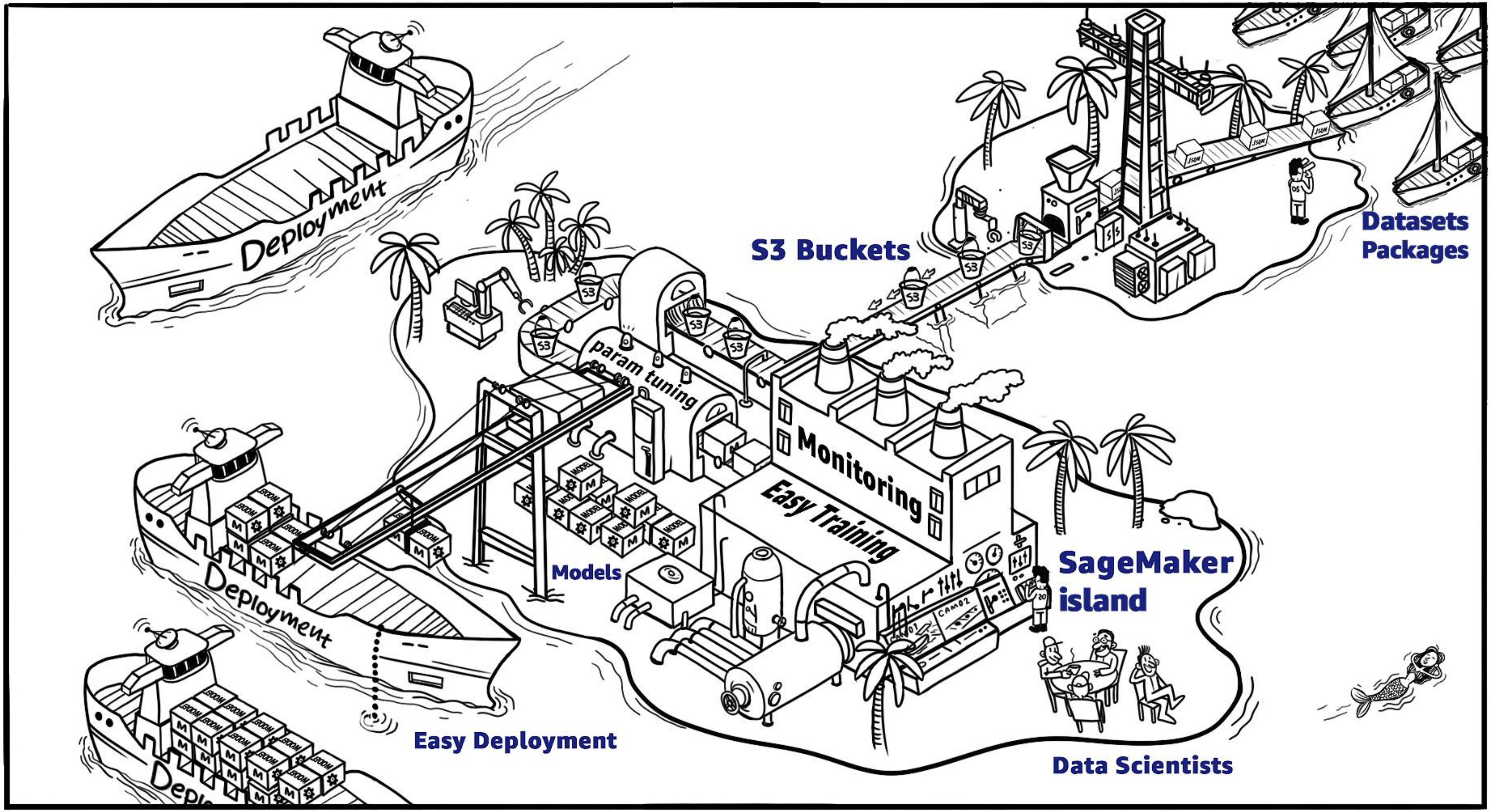
Action: Cutting tickets on Jira for platform team!

* a.k.a Your platform team who is responsible for everything, a.k.a Cloud Enablement , a.k.a platform CoE

Output: Powerpoints, MVPs, PoCs

Real Business Impact: Zero

Result: Lost Opportunities


Deployment
S3 Buckets
Datasets
Packages
param tuning
Monitoring
Easy Training
Models
SageMaker
island
Deployment
Easy Deployment
Data Scientists
S3
JSON
MODEL
M
CAM02
Depl

Introduction

Nowadays, everybody knows about AI and its potential role in the success of a business. But most of those who have already tried to involve AI in their business have not had a pleasant experience. Few projects actually make it to an actual product. And in rare cases they do, problems arise in deployment and scaling. Isolated workflows make collaboration really difficult.

What is MLOps?

Short for Machine Learning Operations, MLOps seek to minimize the difficulties most companies are struggling with within Machine Learning Pipelines. That is to say, MLOps works to facilitate the communication between Data Scientists and Operation Professionals and automate the process as much as possible. In short, MLOps makes machine learning pleasant and painless!

MLOps in Action

MLOps in AWS can resolve many issues most companies deal with in terms of coordination, scaling, costs, reliability, and security. Here we investigate the role of MLOps in each section.

Data Engineering

Instead of having local copies of the data, which comes with security issues, data can be streamed and organized in S3 buckets, a durable, reliable, and secure storage service offered by Amazon. Other ML services can easily and securely utilize S3. Plus, resolving the need for transfer speeds up the process.

Collaborative Development

Data Scientists can cooperate on developing instead of isolated work on different models. They can monitor the progress and walk towards a unified goal.

On-demand Infrastructure

While companies have to procure the hardware needed for model development in traditional ways, AWS offers on-demand compute power, effectively reducing costs, without the managers having to worry about over/under-provisioning resources. For each step, one is only charged for the amount of time they are using AWS, and once the job finishes automatically, there would be no extra expense.

AI Made Easy

With Amazon SageMaker, data scientists have access to many optimized algorithms and frameworks for most of the problems a business might face. This omits the need for experimenting with different models, handling dependencies, and hardware optimization.

Plus, with the Hyperparameter Tuning service, any model can automatically be optimized for the task at hand in a time-efficient manner.

Easy Deployment

Once a model is trained in SageMaker, artifacts are automatically stored in S3, ready to be used. The deployment in an endpoint comes with a mere click for real-time inference or batch transforms. Scaling up is amazingly easy. With the Amazon Autoscaling feature, you can scale out your deployment when the demand goes high and automatically scale down when demands are lower, thus optimizing the cost.

Integration

A model endpoint can be integrated with other services, for example, via an API gateway, based on the task at hand.

Automation - CI/CD

Once your workflow is complete, you can save the template in a Cloud Formation script and apply the same workflow, many times in different places. AWS will take care of provisioning resources, securing the infrastructure, and ensuring continuous development and integration. **Pleasant and painless!**