Automating the creation Machine Learning infrastructure

Roman Golovnya
19 September 2020

About me

Roman Golovnya (https://www.linkedin.com/in/romangolovnya) graduated with postgrad degree in Cloud Computing NCI. Currently, he works in ResMed as AWS Data Engineer. He develops scalable data solutions using python, Apache Spark and AWS services. In a free time, he organises the events for this meetup group, plays tennis and participate in Kaggle competitions.

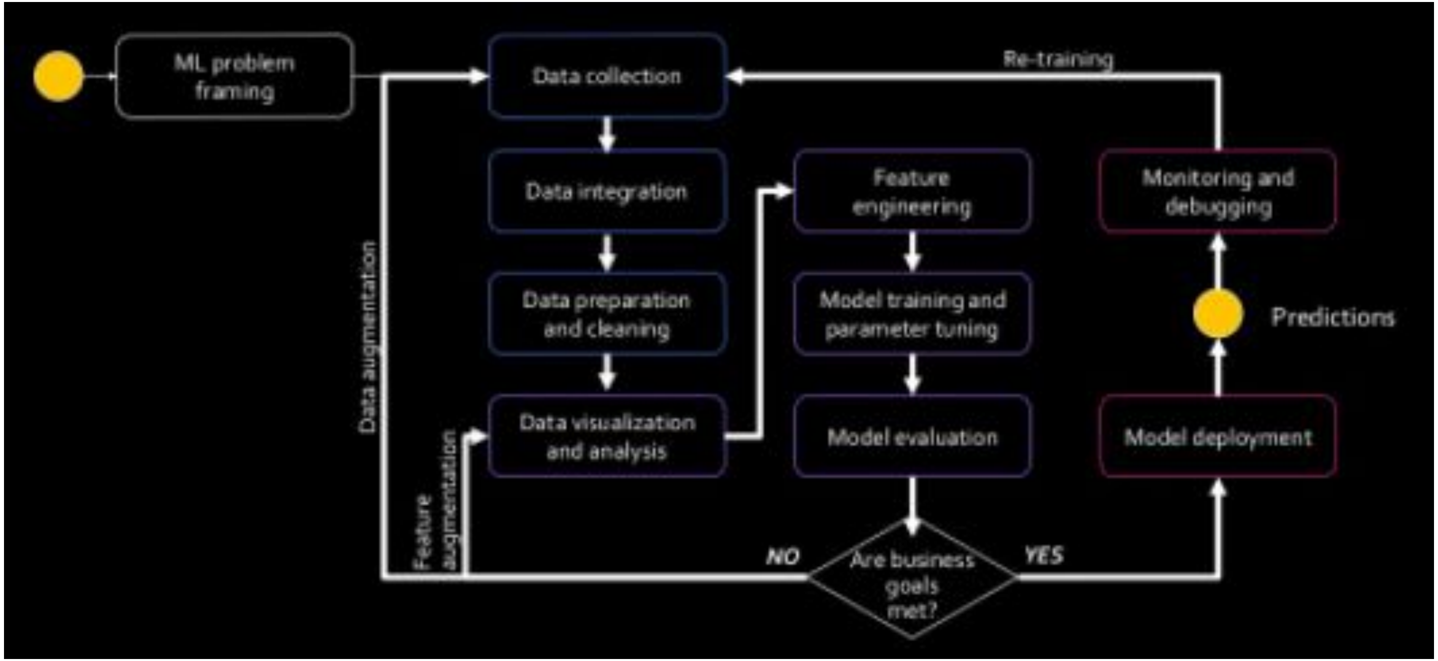
Agenda

- Data Science projects complexity
- AWS SageMaker
- Infra as a code AWS cloudformation, Terraform
- ML with containers, kubernetes AWS EKS , Kubeflow

- Simple ML project - POC
- *All relevant technology installed
- Data locally, relevantly small data
- Simple model
- No production ready
- No security concern
- A few people works on this project
- Cost: nearly zero
- Risk: low
- Technology: jupyter notebook/lab, RStudio, Sklearn,can be containerised

- Complex ML project
- *All relevant technology installed
- Data in the cloud, large datasets, multiple datasets
- Complex model, DL, NLP
- production ready
- Require scalability
- security concern
- Bigger team work on the project
- Cost: pay per usage but save time
- Risk: high cost if left working
- All services should be shut down after use!!!
- Technology: SageMaker, Kubeflow, containers, EKS , **

ML project cycle



ref Julien SIMON AWS Global Evangelist ML/AI

ML platforms

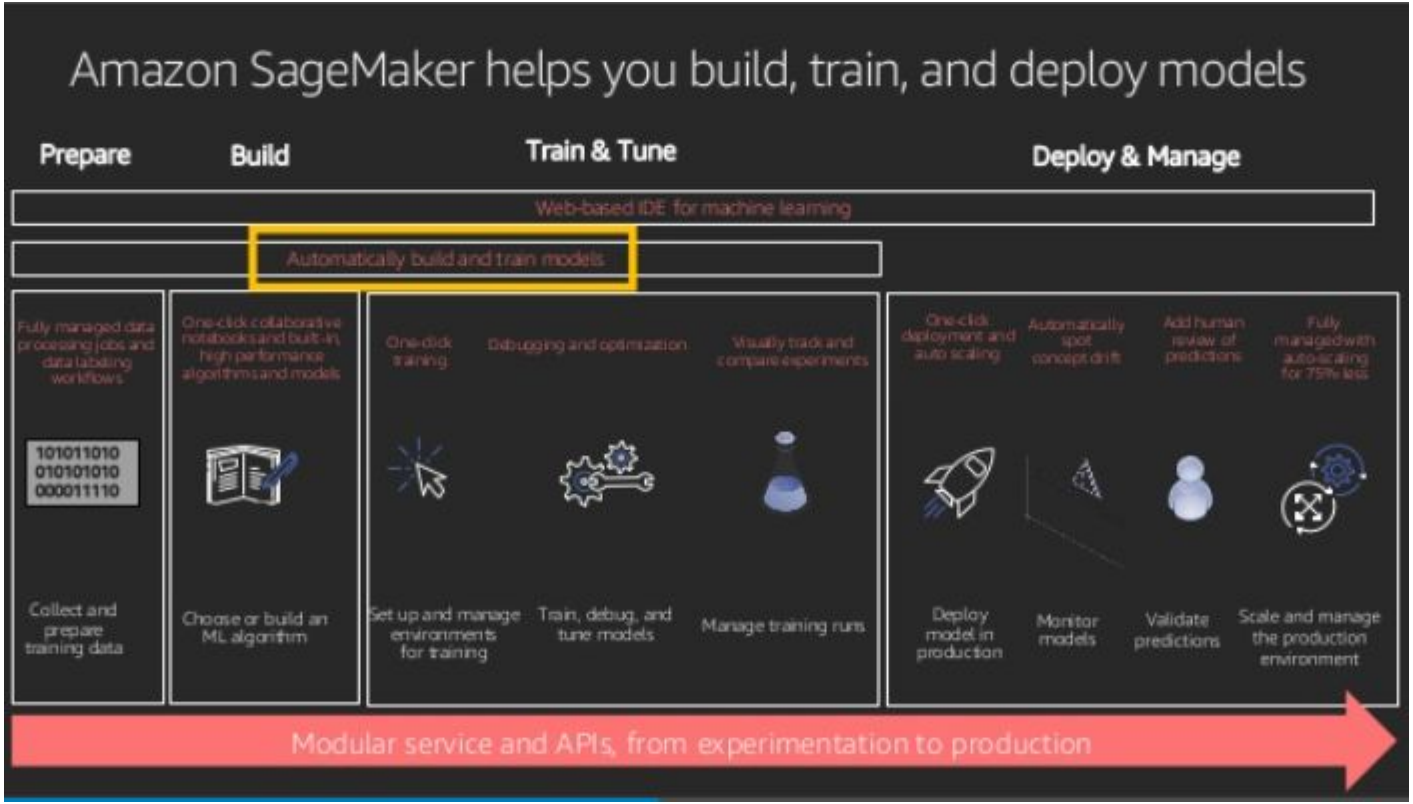
- AWS SageMaker *
- Azure ML Studio*
- H2O open source/*
- KubeFlow
- MLFlow
- Databricks *
- Michelangelo (Uber)
-
- *paid

ML open source platforms

- **Kubeflow** is a free and open-source machine learning platform designed to enable using machine learning pipelines to orchestrate complicated workflows running on Kubernetes. Kubeflow was based on Google's internal method to deploy TensorFlow models to Kubernetes called TensorFlow Extended.
- https://www.kubeflow.org/

- **MLflow** by Databricks - an open source platform for managing the end-to-end machine learning lifecycle
- An open source platform for the machine learning lifecycle
- https://mlflow.org/

AWS SageMaker https://aws.amazon.com/sagemaker/



Amazon SageMaker helps you build, train, and deploy models

Prepare | Build | Train & Tune | Deploy & Manage

Web-based IDE for machine learning

Automatically build and train models

Fully managed data processing jobs and data labeling workflows

Collect and prepare training data

One-click collaborative notebooks and built-in, high performance algorithms and models

Choose or build an ML algorithm

One-click training

Set up and manage environments for training

Debugging and optimization

Train, debug, and tune models

Visually track and compare experiments

Manage training runs

One-click deployment and auto scaling

Deploy model in production

Automatically spot concept drift

Monitor models

Add human review of predictions

Validate predictions

Fully managed with auto-scaling for 75% less

Scale and manage the production environment

Modular service and APIs, from experimentation to production

Amazon SageMaker is a fully-managed platform that enables data scientists to quickly and easily build, train, and deploy machine learning models at any scale.

AWS SageMaker https://aws.amazon.com/sagemaker/

Train machine learning models

Organize, track, and evaluate training runs using Amazon SageMaker Experiments

Analyze, detect, and alert problems for machine learning using Amazon SageMaker Debugger

Deploy machine learning models - one-click deployment,

Amazon SageMaker Model Monitor - keep models accurate over time using

Integrate with Kubernetes for orchestration and management

Use Kubeflow Pipelines for job orchestration and scheduling

Use Amazon Elastic Inference save money

SageMaker endpoint allows you to make real-time inferences via a REST API

Sagemaker AutoML automatically choose, train & optimise model

SageMaker Studio docker images not working yet.

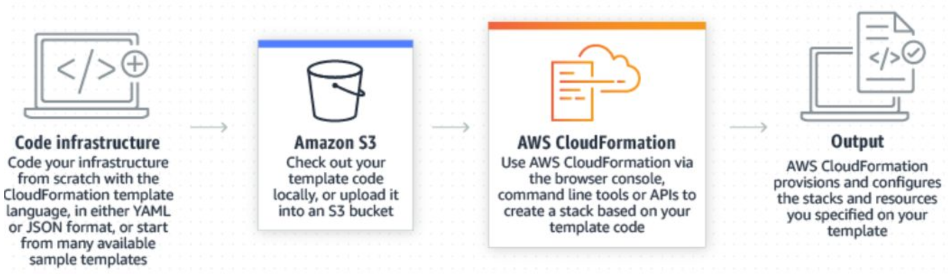
Amazon SageMaker supports the deep learning frameworks: TensorFlow, PyTorch, Apache MXNet, Chainer, Keras, Gluon, Horovod, Scikit-learn, and Deep Graph Library.

Automating creation the infrastructure

- Use infrastructure as a code
- Containerised ML - DS toolsets - installation & update
- Containerised ML-DL-NLP, EKS, Kubeflow - code for scalability

AWS cloudformation

- AWS CloudFormation provides a common language for you to **model and provision** AWS and third party application resources in your cloud environment. AWS CloudFormation allows you to use programming languages or a simple text file **to model and provision, in an automated and secure manner, all the resources needed for your applications across all regions and accounts.** This gives you a single source of truth for your AWS and third party resources.


Code infrastructure
Code your infrastructure from scratch with the CloudFormation template language, in either YAML or JSON format, or start from many available sample templates
Amazon S3
Check out your template code locally, or upload it into an S3 bucket
AWS CloudFormation
Use AWS CloudFormation via the browser console, command line tools or APIs to create a stack based on your template code
Output
AWS CloudFormation provisions and configures the stacks and resources you specified on your template

https://aws.amazon.com/cloudformation/ ref

https://github.com/aws/aws-cdk AWS Cloud Development Kit (AWS CDK)

https://github.com/awslabs/aws-cloudformation-templates

Practical application - AWS cloudformation, SageMaker & AWS services

https://github.com/awslabs/fraud-detection-using-machine-learning

Hashicorp Terraform

- Terraform is an open-source infrastructure as code software tool created by HashiCorp. It enables users to **define and provision** a datacenter infrastructure using a high-level configuration language known as Hashicorp Configuration Language (HCL) terraform

- Automate infrastructure provision
- Write declarative configuration file
- Consistent and repeatable workflows
- Reproducible and reusable infrastructure
- Versioning infrastructure with shared code

https://github.com/hashicorp/terraform

https://www.terraform.io/

Practical application

https://medium.com/@yuyasugano/machine-learning-infrastructure-terraforming-sagemaker-part-1-e5e22e368248

Kubeflow

- Kubeflow is a free and open-source machine learning platform designed to enable using machine

  learning pipelines to orchestrate complicated workflows running on Kubernetes. Kubeflow was based on Google's internal method to deploy TensorFlow models to Kubernetes called TensorFlow Extended.
- https://www.kubeflow.org/
- https://aws.amazon.com/blogs/opensource/enterprise-ready-kubeflow-securing-and-scaling-ai-and-machine-learning-pipelines-with-aws/
- https://www.eksworkshop.com/
-

ML serverless, orchestration

- Lambda
- Step functions
- AWS Step Functions allows you to build resilient workflows using AWS services such as Amazon DynamoDB, AWS Lambda, and Amazon SageMaker.
- Use the **Step Functions Data Science SDK** with **Amazon SageMaker Processing** to create and visualize **end-to-end machine learning workflows.** Workflows can be built in Python and visualized within Jupyter Notebooks. Data scientists can build and iterate on their machine learning pipelines and then write out a CloudFormation template that can be used by engineering teams to take the workflow into production, supporting the MLOps use-case.

  Practical

  https://medium.com/@elesin.olalekan/automating-machine-learning-workflows-pt2-sagemaker-processing-sagemaker-and-aws-step-functions-5e86314121b5

Conclusion:

- Think about ML projects in production from day one
- Use infra as code: cloudformation or terraform
- Containerize ML apps and run them in cluster if needed
- Serverless ML workflows

Resources:

- https://www.amazon.com/Learn-Amazon-SageMaker-developers-scientists-ebook/dp/B08FMWJXGN Julien Simon AWS
- https://github.com/awslabs/amazon-sagemaker-examples
- https://www.oreilly.com/library/view/data-science-on/9781492079385/ Chris Fregly, Antje Barth 2021
- https://github.com/data-science-on-aws/workshop/
- https://github.com/antonbabenko/terraform-best-practices-workshop
- https://aws.amazon.com/blogs/opensource/enterprise-ready-kubeflow-securing-and-scaling-ai-and-machine-learning-pipelines-with-aws/
- https://www.kubeflow.org/
- https://www.eksworkshop.com/

Thank you!

Any Questions?