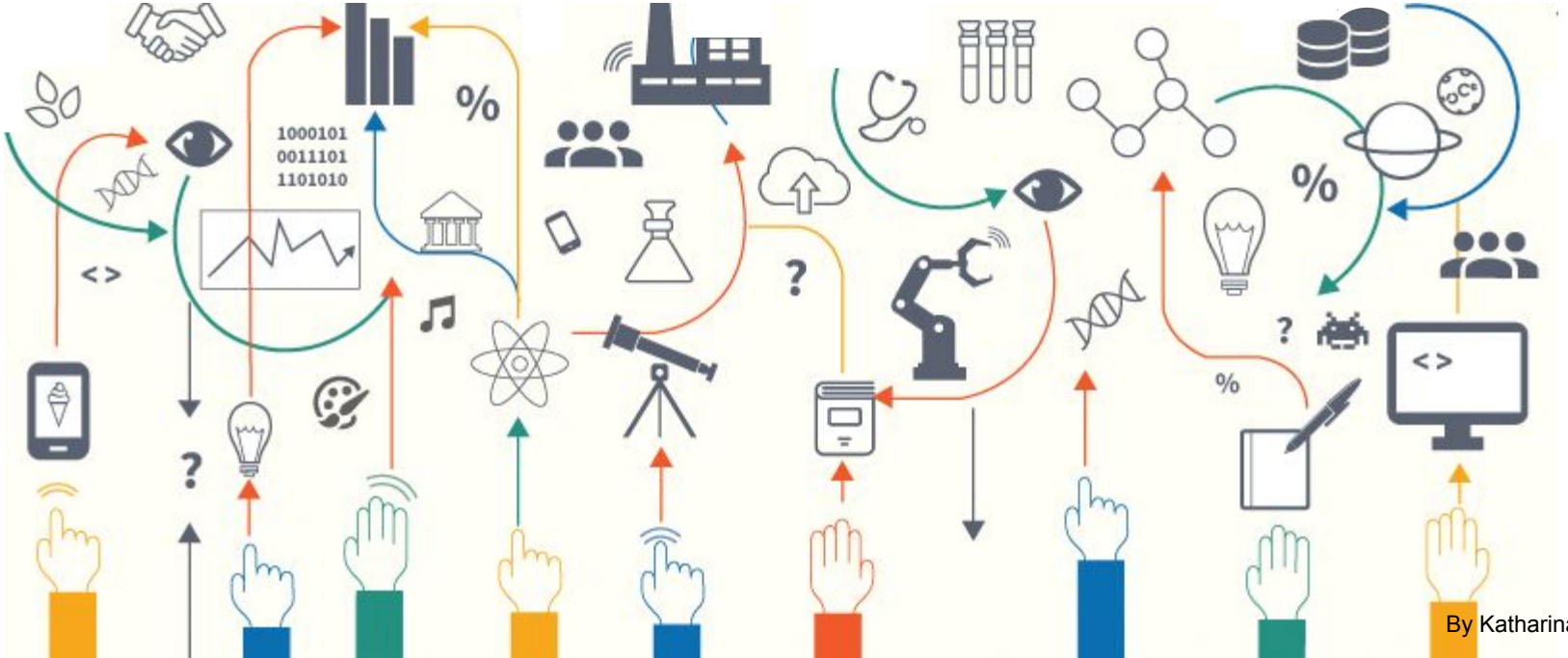
Data Science in the Cloud

With 3 approaches



By Katharina Hofschen

Motivation

Data science and ML often require a **lot of compute resources** that developers' PCs might not have.

Jupyter notebooks on Kaggle or Colab are either **limited or not highly available**, so lets run jupyter servers on the cloud and work there:

1. Approach: Manually create EC2 and install conda & jupyter lab
2. Approach: Automate and expand setup of 1st approach
3. Approach: Use Sagemaker instance notebooks or studio

The AWS ML Stack

Broadest and most complete set of machine learning capabilities

AI SERVICES

HEALTH AI
- NEW Amazon HealthLake
- Amazon Transcribe Medical
- Amazon Comprehend Medical

INDUSTRIAL AI
- NEW AWS Panorama + Appliance
- NEW Amazon Monitron
- NEW Amazon Lookout for Equipment
- NEW Amazon Lookout for Vision

ANOMALY DETECTION
- NEW Amazon Lookout for Metrics

CODE AND DEVOPS
- NEW Amazon DevOps Guru
- Amazon CodeGuru

VISION
- Amazon Rekognition

SPEECH
- Amazon Polly
- Amazon Transcribe +Medical

TEXT
- Amazon Comprehend +Medical
- Amazon Translate
- Amazon Textract

SEARCH
- Amazon Kendra

CHATBOTS
- Amazon Lex

PERSONALIZATION
- Amazon Personalize

FORECASTING
- Amazon Forecast

FRAUD
- Amazon Fraud Detector

CONTACT CENTERS
- Contact Lens
- Voice ID For Amazon Connect

ML SERVICES

Amazon SageMaker

Label data

SAGEMAKER STUDIO IDE

NEW Aggregate & prepare data | NEW Store & share features | Auto ML | Spark/R | NEW Detect bias | Visualize in notebooks | Pick algorithm | Train models | Tune parameters | NEW Debug & profile | Deploy in production | Manage & monitor | NEW CI/CD

Human review

NEW: SageMaker JumpStart

NEW: Model management for edge devices

FRAMEWORKS & INFRASTRUCTURE

TensorFlow | mxnet | PyTorch | GLUON | Keras | learn | DeepGraphLibrary

Deep Learning AMIs & Containers | GPUs & CPUs | Elastic Inference | Trainium | Inferentia | FPGA

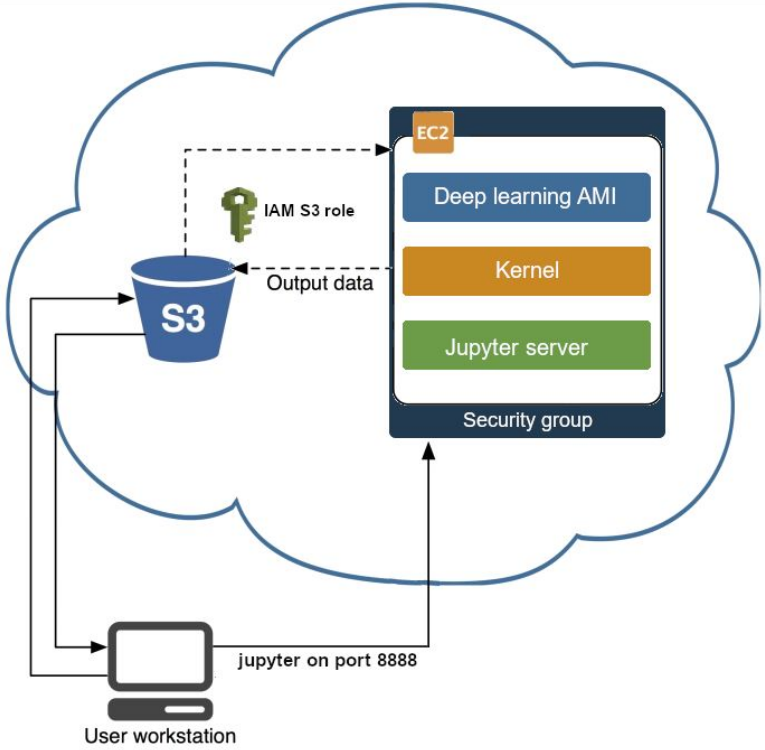
© 2020, Amazon Web Services, Inc. or its Affiliates.

aws

Steps to run a jupyter notebook on an EC2

1. Expose the port 8888 in security group (limit to your IP)
2. Get and install conda
3. Install all required packages
4. Configure the jupyter configuration (add encrypted password)
5. Initialize conda
6. Restart bash and activate conda environment
7. Start jupyter notebook
8. Access jupyter notebook on public IP:8888 of EC2

Automation with Terraform



Main.tf

- Specify all infrastructure needed
- Auto Scaling group
- AMI
- Security group
- IAM S3 role and policy

Variables.tf

- Secret keys, instance types etc.

Install.sh

- Script to set python env and start jupyter


Live Demo
TERRAFORM


WORKS ON MY MACHINE
38
OPS PROBLEM NOW!

Amazon SageMaker

Prepare →

SageMaker Ground Truth
Label training data for machine learning

SageMaker Data Wrangler NEW
Aggregate and prepare data for machine learning

SageMaker Processing
Built-in Python, BYO R/Spark

SageMaker Feature Store NEW
Store, update, retrieve, and share features

SageMaker Clarify NEW
Detect bias and understand model predictions

Build →

SageMaker Studio Notebooks
Jupyter notebooks with elastic compute and sharing

Built-in and Bring-your-own Algorithms
Dozens of optimized algorithms or bring your own

Local Mode
Test and prototype on your local machine

SageMaker Autopilot
Automatically create machine learning models with full visibility

SageMaker JumpStart NEW
Pre-built solutions for common use cases

Train & tune →

One-click Training
Distributed infrastructure management

SageMaker Experiments
Capture, organize, and compare every step

Automatic Model Tuning
Hyperparameter optimization

Distributed Training Libraries NEW
Training for large datasets and models

SageMaker Debugger NEW
Debug and profile training runs

Managed Spot Training
Reduce training cost by 90%

Deploy & manage →

One-click Deployment
Fully managed, ultra low latency, high throughput

Kubernetes & Kubeflow Integration
Simplify Kubernetes-based machine learning

Multi-Model Endpoints
Reduce cost by hosting multiple models per instance

SageMaker Model Monitor
Maintain accuracy of deployed models

SageMaker Edge Manager NEW
Manage and monitor models on edge devices

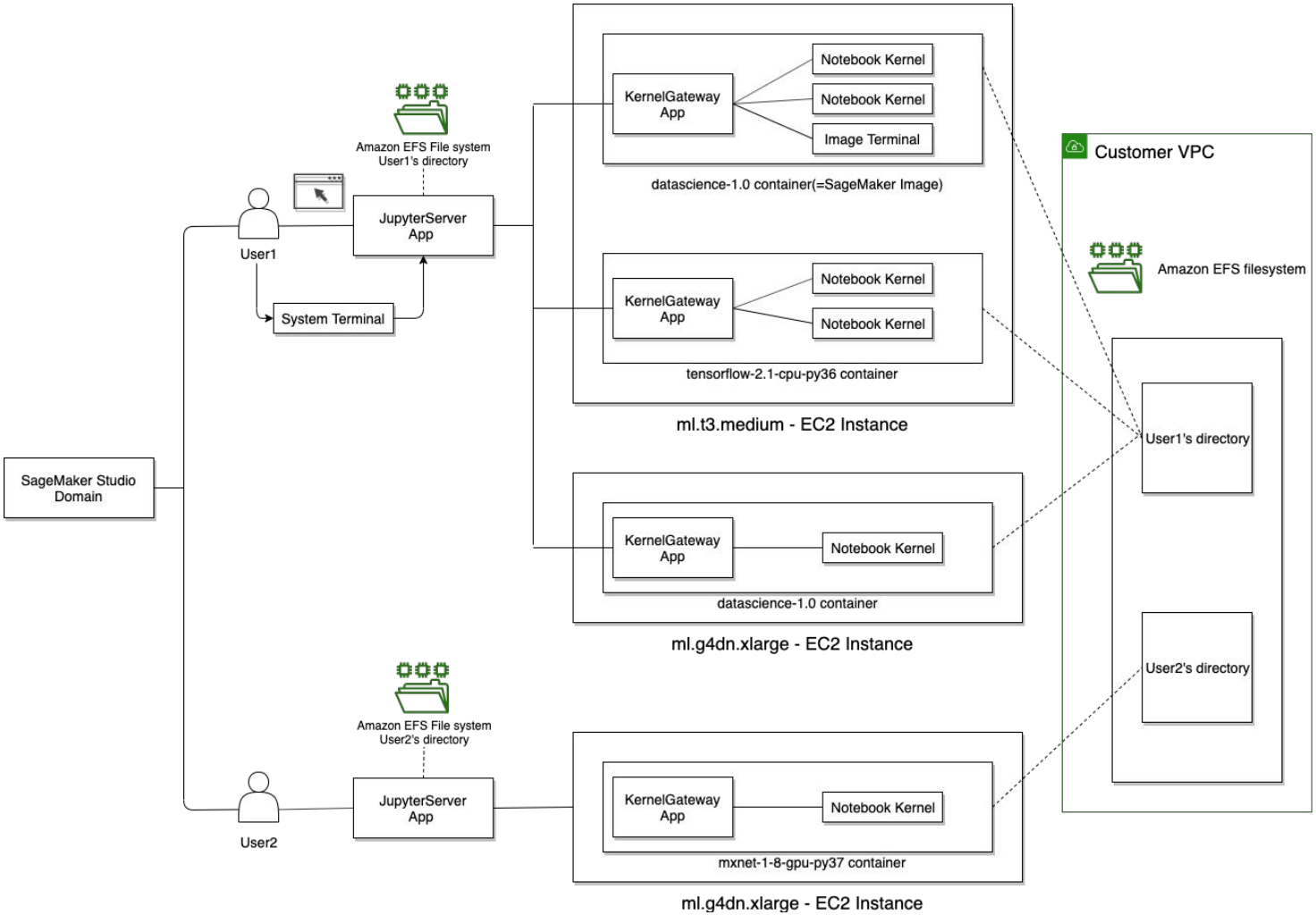
SageMaker Pipelines NEW
Workflow orchestration and automation

SageMaker Studio

Integrated development environment (IDE) for ML

Live Demo

Amazon SageMaker Studio


Customer VPC
Amazon EFS filesystem
User1's directory
User2's directory
SageMaker Studio Domain
Amazon EFS File system User1's directory
User1
JupyterServer App
System Terminal
KernelGateway App
Notebook Kernel
Notebook Kernel
Image Terminal
datascience-1.0 container(=SageMaker Image)
KernelGateway App
Notebook Kernel
Notebook Kernel
tensorflow-2.1-cpu-py36 container
ml.t3.medium - EC2 Instance
KernelGateway App
Notebook Kernel
datascience-1.0 container
ml.g4dn.xlarge - EC2 Instance
Amazon EFS File system User2's directory
User2
JupyterServer App
KernelGateway App
Notebook Kernel
mxnet-1-8-gpu-py37 container
ml.g4dn.xlarge - EC2 Instance

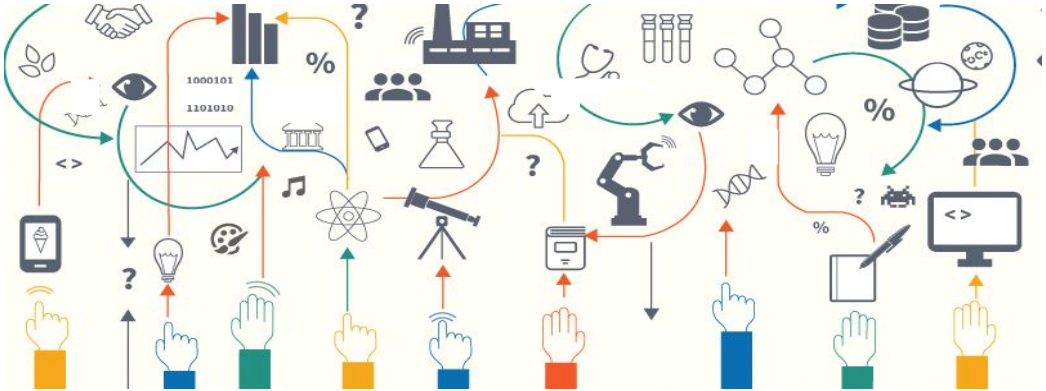
Conclusions & Future work

Sagemaker currently offers many tools for data science IDEs:

- Many capabilities to get anyone started with ML
- Perfect from separating data scientists from cloud computing
- Good tool for implementing MLOps methods

In the future:

- Work on my automation script to ensure data and model persistence across sessions
- Emulate the Studio setup to improve collaboration



Thank you! Any questions?