MLaaS MOD Israel

Miki Kenneth
Director, Software Engineering, Office of the CTO

Menny Tsarfati,
Solution Architect, Manager, IGC

“Without data, you’re just another person with an opinion.”
— W. Edwards Deming
Menny Tsarfati
Solution Architect, Manager, IGC

Miki Kenneth
Director, Software Engineering, Office of the CTO

“Learning is not compulsory... neither is survival.”
— W. Edwards Deming
Agenda

- Market trends and challenges
- Data as the Next/Current value Evolution/Revolution
- MOD Use case
  - Why
  - System background
  - Solution
  - 5 Steps of improvement
- The outcomes
- Lesson learned
- Next step
- Call for action
- QA

“In God we trust; all others bring data.”
— W. Edwards Deming
Companies Are Failing in Their Efforts to Become Data-Driven

by Randy Bean and Thomas H. Davenport
FEBRUARY 05, 2019

https://hbr.org/2019/02/companies-are-failing-in-their-efforts-to-become-data-driven
Here are some of the alarming results from the survey

72% of survey participants report that they have yet to forge a data culture
69% report that they have not created a data-driven organization
53% state that they are not yet treating data as a business asset
52% admit that they are not competing on data and analytics.

Further, the percentage of firms identifying themselves as being data-driven has declined in each of the past 3 years – from 37.1% in 2017 to 32.4% in 2018 to 31.0% this year.
Sense of urgency

92% of survey respondents reported that the pace of their big data and AI investments is accelerating; 88% report a greater urgency to invest in big data and AI; 75% cite a fear of disruption as a motivating factor for big data/AI investment. In addition, 55% of companies reported that their investments in big data and AI now exceed $50MM, up from 40% just last year.

The world is changing very fast. Big will not beat small anymore. It will be the fast beating the slow. —Rupert Murdoch
Motivation

• Big Data – Data is doubled every year
• Unstructured data
• Data Engineering services
• Data Science services
• New business needs New Marketplace
Data life cycle

1. **Generation or capture**: In this phase, data comes into an organization, usually through data entry, acquisition from an external source or signal reception, such as transmitted sensor data.

2. **Maintenance**: In this phase, data is processed prior to its use. The data may be subjected to processes such as integration, scrubbing and extract-transform-load (ETL).

3. **Active use**: In this phase, data is used to support the organization’s objectives and operations.

4. **Publication**: In this phase, data isn’t necessarily made available to the broader public but is just sent outside the organization. Publication may or may not be part of the life cycle for a particular unit of data.

5. **Archiving**: In this phase, data is removed from all active production environments. It is no longer processed, used or published but is stored in case it is needed again in the future.

6. **Purging**: In this phase, every copy of data is deleted. Typically, this is performed on data that is already archived.

Source: [https://whatis.techtarget.com](https://whatis.techtarget.com)
MoD Cloud

- Modern platform – Greenfield
- Full Red Hat stack – Openstack, Openshift, Ansible ….
- Multi hybrid cloud
- “Amazon like” Multi tenancy
- On Premise disconnected environment
- EaaS
- Self Service
- Multi site
- Edge computing
- Security
MOD - The Why

- Cost
- Time to market
- Data market-place
- Research reproduction
- Analytics accuracy
- Re-use of models developed
- Platform utilization

Make it simple, Make it available, Fast deliveries. Fast fail faster recovery
Common data science pipeline

tensorflow, keras, pytorch or scikit-learn
Experiment reproduction

- Huge number of evaluations and results
- What is the best experiment?
- Collaboration
  - Model reuse
Hyper parameter optimization

- Container based
- Full resources utilization
- Smart optimization algorithm
- Time to value: months -> days
- Better performance and accuracy
Model deployment

- The relation between data scientist and containers
- Container based, automatic deployment of common ML frameworks
- Model A/B testing out of the box
- Version control
Research reproduction

Today’s, Machine learning researches challenge is to keep track of previous experiments. For example, if you want to build a decision tree model for a specific problem, you will obviously try different hyper-parameters, pre-processing functions and other mathematical operations, without keeping the results of these experiments. We solved this problem by developing a tool inside our Jupyter notebooks, which saves automatically every relevant metric of the experiment, including model hyper-parameters, pre-processing functions and even the code itself.
Re-use of developed models and their publication

Self-service for the full life cycle
The application enables the usage of Openshift projects in the self-service portal and re-use or the basis for a new project with changing parameters
The Journey
Current State - Journey

Month #1
Current State - Journey

DI380 gen9 x3 each with 256GB ram
4x800GB SSD’s in raid10 - 100% sequential write
~2GB/s
Network – Mellanox Tesla V100
Current State - Journey
TESLA V100 tensor core gpu

5,120 CUDA cores
640 NEW Tensor cores
7.8 FP64 TFLOPS | 15.7 FP32 TFLOPS
| 125 Tensor TFLOPS
20MB SM RF | 16MB Cache
32 GB HBM2 @ 900GB/s | 300GB/s NVLink
Tensor core automatic mixed precision

3x Speedup With Just Two Lines of Code

Tools and Libraries
Maintain Network Accuracy

Training Speedup Over 3x

Inference Speedup Over 4x
Next step
OUTCOMES
Costs

The amount of information, Variety, Speed of streaming and analysis require expensive computers with graphics processors and in some case dedicated servers such as Nvidia DGX. Today there aren’t many options for customization the research workspace, and usually a data scientist go for the biggest environment and pays more than he actually need to. We developed, based on JupyterHub tool, a platform which can allocate resources dynamically, where each workspace (A jupyter notebook), has its own resources and is actually a container (That can be customized!) on Openshift.

Developed a kubernetes device plugin that enables splitting a single GPU into multiple “vGPUs" with time scheduling.
Utilization of the platform and the speed of execution of the analysis

Not only a data scientist can now select his workspace and allocate specific resources, he can also stop this workspace and create a new one in minutes, instead of an exhausting process which includes killing a server, creating a new one, installing relevant packages again and again. We growth from 4 users to 60 in 3 month.
Research reproduction

The challenge issue in today’s Machine learning researches is to keep track of previous experiments. For example, to build a decision tree model for a specific problem, you will obviously try different hyper-parameters, pre-processing functions and other mathematical operations, without keeping the results of these experiments.

We solved this problem by developing a tool inside our Jupyter notebooks, which saves automatically every relevant metric of the experiment, including model hyper-parameters, preprocessing functions and even the code itself.
Hyper Parameters Optimization

Common challenge is adapting a machine learning model to the data. Every model has dozens of hyper-parameters and tuning them to the best result is a hard and exhausting job. In order to solve this, we developed Hyper-parameters optimization framework, which takes as input the search space and in response starts a “master” node which generates different parameters choices and multiple “workers” which evaluating them, all hosted on Openshift. The master and workers are communicating through a RabbitMQ which is also, hosted on Openshift. We then display the best composition of parameters and their results to the user, in order to create the final model. The search itself is done using various techniques including Bayesian search (TPE), random search and a genetic algorithm.
Model Deployment

There is a lack of knowledge among data scientists when it comes to deploying their model to production. To solve this, we developed a simple interface which takes as input the S3 address of the exported model and a preprocessing function, and then turns the model (ANY model written in tensorflow, keras, pytorch or scikit-learn) into a deployment hosted on Openshift, replicated, scalable, production-ready application ready for prediction requests.
You can build it yourself … or … better talk with the communities

MLFlow - https://www.mlflow.org/

Hyperopt - https://github.com/hyperopt/hyperopt

Sigopt - https://github.com/sigopt

KubeFlow - https://www.kubeflow.org/

His-tagged
Automated machine learning (AutoML) is the process of automating the end-to-end process of applying machine learning to real-world problems. In a typical machine learning application, practitioners must apply the appropriate data pre-processing, feature engineering, feature extraction, and feature selection methods that make the dataset amenable for machine learning. Following those preprocessing steps, practitioners must then perform algorithm selection and hyperparameter optimization to maximize the predictive performance of their final machine learning model. As many of these steps are often beyond the abilities of non-experts, AutoML was proposed as an artificial intelligence-based solution to the ever-growing challenge of applying machine learning. Automating the end-to-end process of applying machine learning offers the advantages of producing simpler solutions, faster creation of those solutions, and models that often outperform models that were designed by hand.

Demo time  .....
Marvin Minsky: "The science of making machines do things that would require intelligence if done by men."
Outcomes

- Simplicity
- Higher utilization of Environment
- Higher utilization of Talents (Data scientist)
- Average data science process from 3 month to 3 days
- Accuracy – average improvement of 20%
- Data science - No need for “Techy” skills
- Platform of choice new Data culture
Lesson Learned

• Data scientist are looking for the data and models – Platform should be transparent – Make it easy to consume and experiment
• Reuse and Reproduce as a values
• In a very short time It will become a platform and a “Normal service” in your private cloud.
• Ready for any
• Collaborate with the Academy – Be part of thier experiments and … use the Talents and new Algorithms
• Full life cycle of AutoML
• Usage Growth with simplicity and accessibility
• Open, Open, Open

We are just at the beginning
What Next and Call to action

- Follow and contribute to Kubeflow
- @Kubecon Barcelona – Openshift track – MOD Demo
- Edge – promoting solutions to the edge
- Security and Privacy – No real roles
- Data Validation / Certification
- Shared project with Academy
- Share Reference Architecture
Q&A
THANK YOU

Miki Kenneth – mkenneth@redhat.com
Menny Tsarfati – menny@redhat.com