EasyBuild for Kubernetes-Based Data Science Platforms

7th EasyBuild User Meeting

2022-01-26

Guillaume Moutier

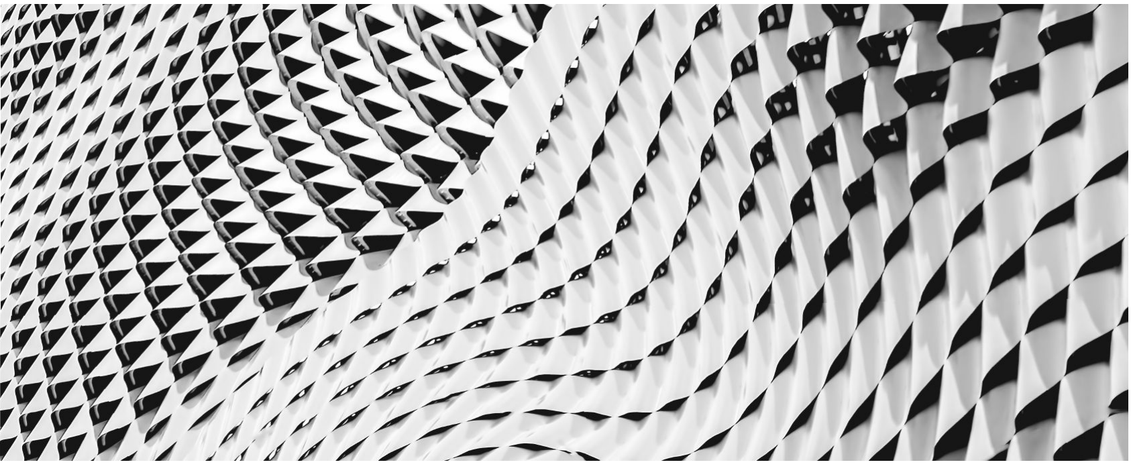
Sr Principal Data Engineering Architect

Red Hat

What we'll discuss today

- Introduction and Context
- The problem
- The solution
- Demo
- What's next?

Introduction and Context



"Don't adventures ever have an end? I suppose not. Someone else always has to carry on the story."
Bilbo Baggins

Red Hat

Who am I?

Former head of IT Architecture

at Laval University in Québec

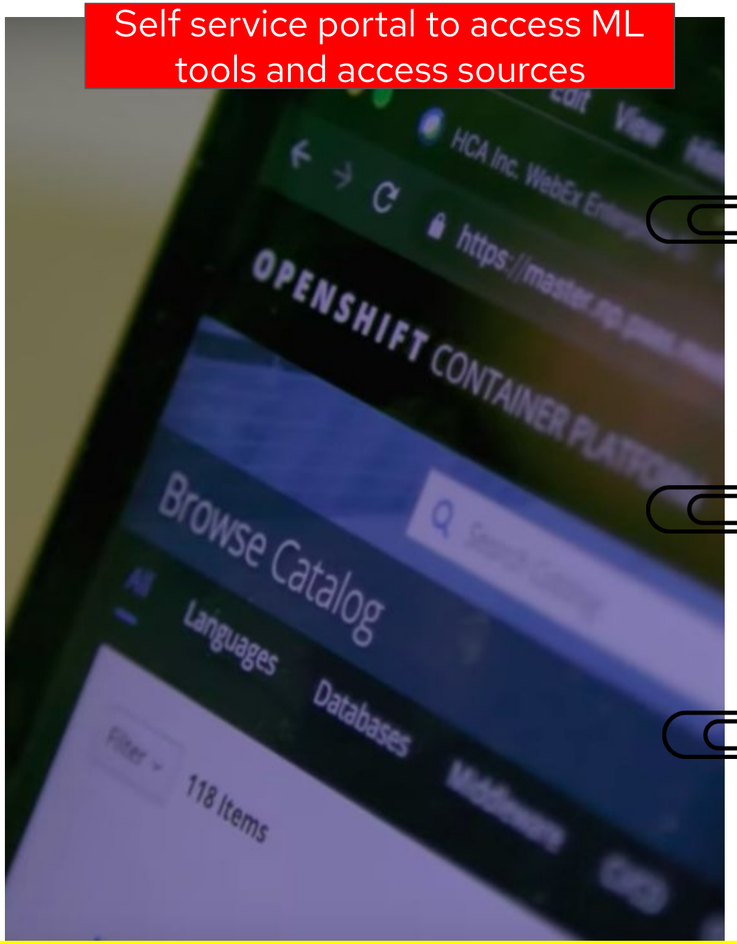
Now a Data Engineering Architect with the

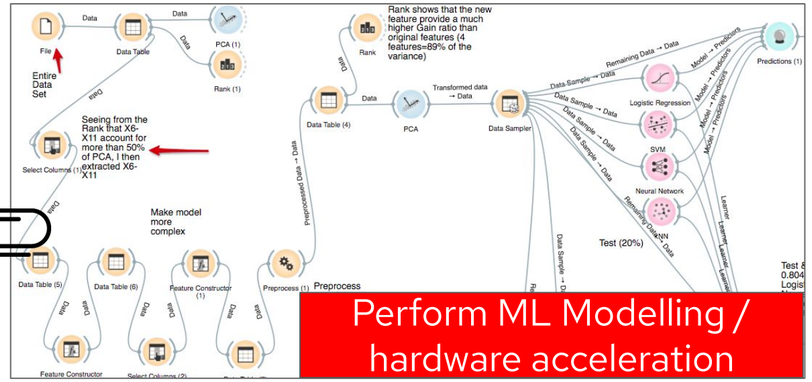
Red Hat OpenShift Data Science team

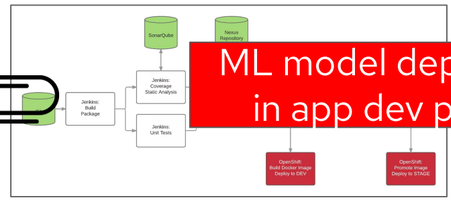
Red Hat
OpenShift
Data Science

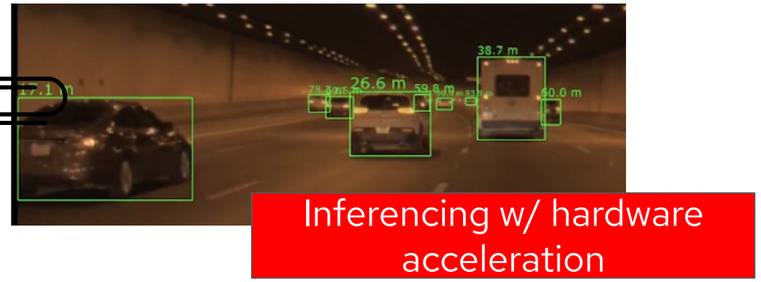
What does a Data Scientist care about?

As a Data Scientist, I want a "self-service cloud like" experience for my Machine Learning projects, where I can access a rich set of modelling tools, data, and computational resources, share and collaborate with colleagues, and deliver my work into production with speed, agility and repeatability to drive business value!







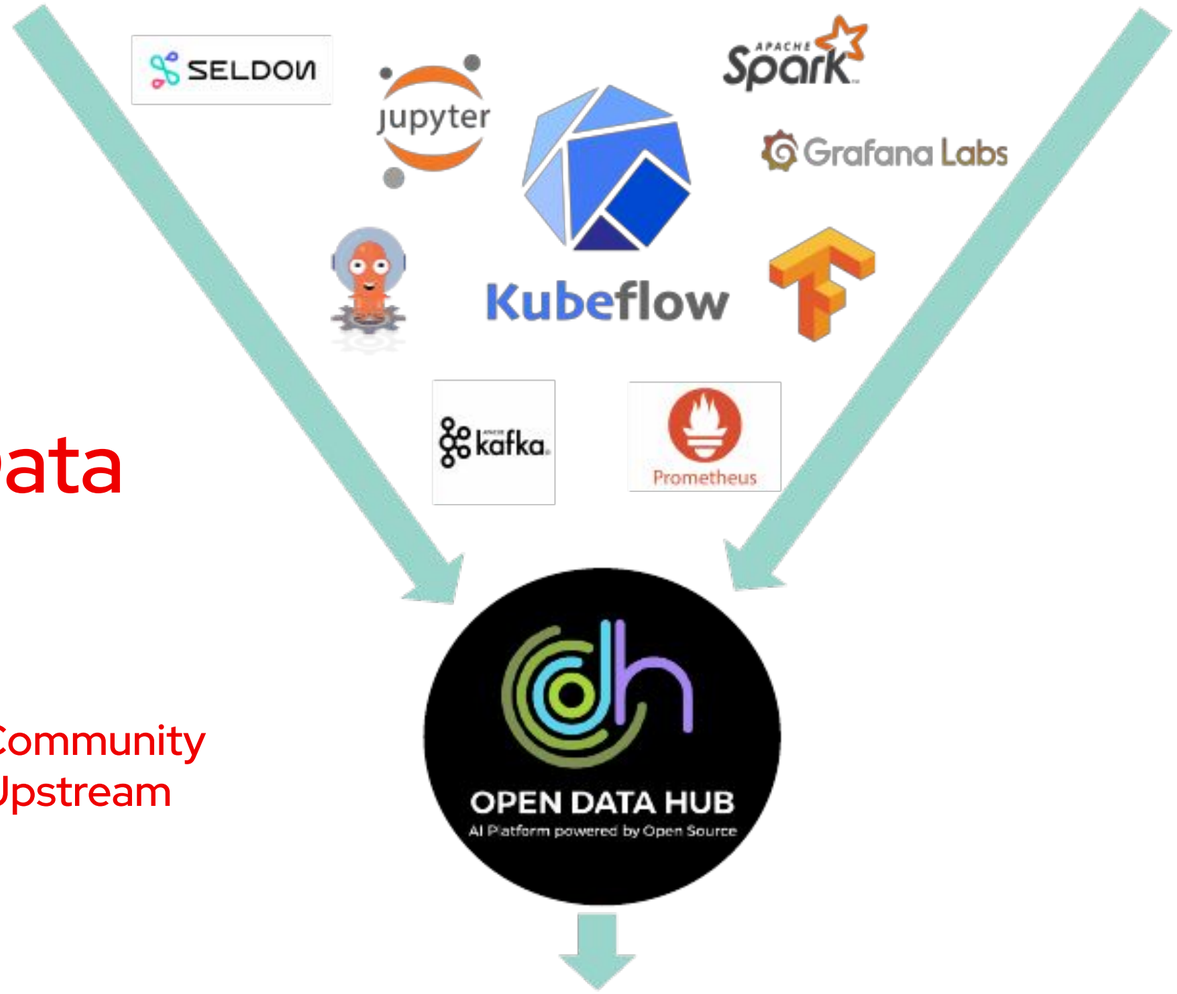


Data Scientists care less about infrastructure platform unless it integrates with their ML tooling, and provides them the agility, flexibility, portability, & scalability.

Red Hat

Open Data Hub

Open Source Community
Downstream/Upstream



Red Hat Application Services

Red Hat OpenShift

Red Hat OpenShift Data Foundation

Goals

- Provide an end-to-end AI/ML platform on OpenShift
- One stop easy operator deployment for the platform on OCP
- Provide Tools for each stage in the AI/ML platform and for all AI/ML user personas optimized for OpenShift
- Provide monitoring tools for model and services used by DevOps
- Provide development tools for Data Scientists
- Provide ETL tools used by Data Engineers
- AI/ML pipelines and long processing tasks.

Red Hat

Red Hat OpenShift Data Science

Addressing AI/ML experimentation and integration use cases on a managed platform

Cloud Service

Available on Red Hat OpenShift Dedicated (AWS) and Red Hat OpenShift Service on AWS

Increased capabilities/collaboration

Combines Red Hat components, open source software, and ISV certified software available on Red Hat Marketplace

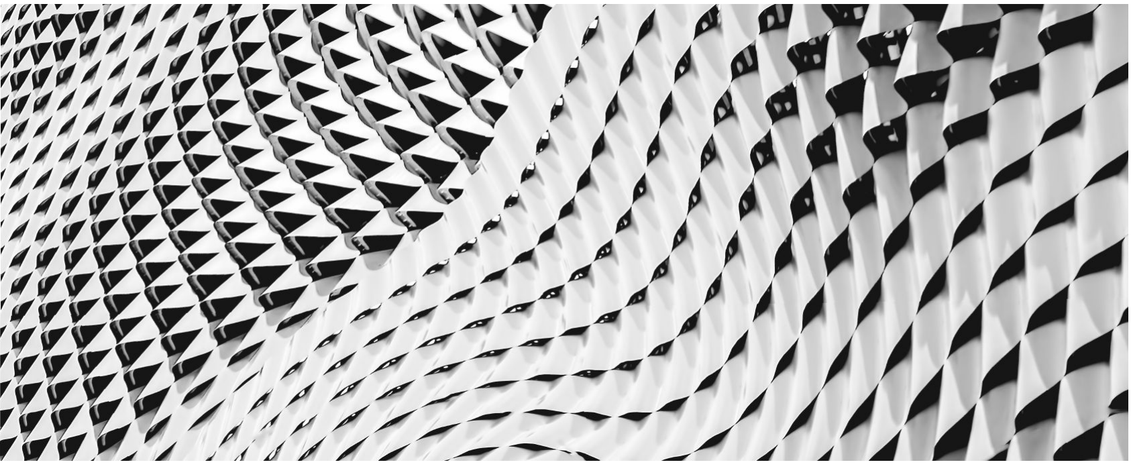
Core data science workflow

Provides data scientists and intelligent application developers the ability to build, train, and deploy ML models

Rapid experimentation use cases

Model outputs are hosted on the Red Hat OpenShift managed service or exported for integration into an intelligent application

The problem



"I have a bad feeling about this..."

Luke Skywalker

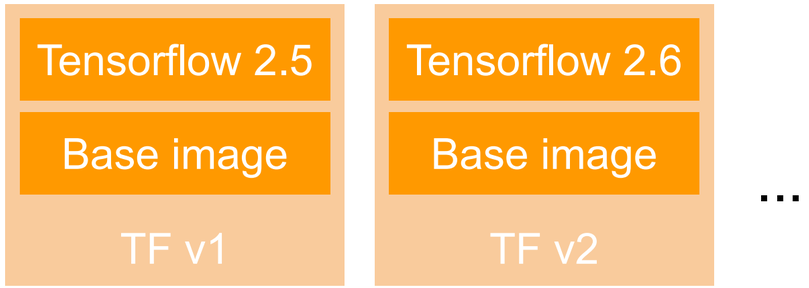
Red Hat

The on-demand notebook example

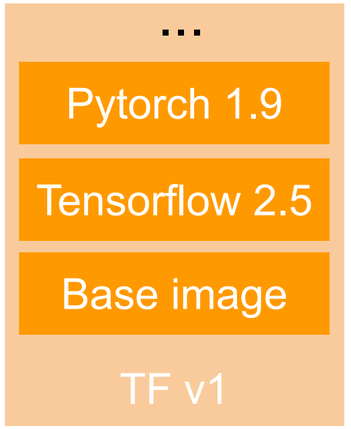
The "base" container image includes:

- Python at a specific version
- Some useful libraries and applications
- Jupyter with pre-defined extensions

Users want more, what to do?

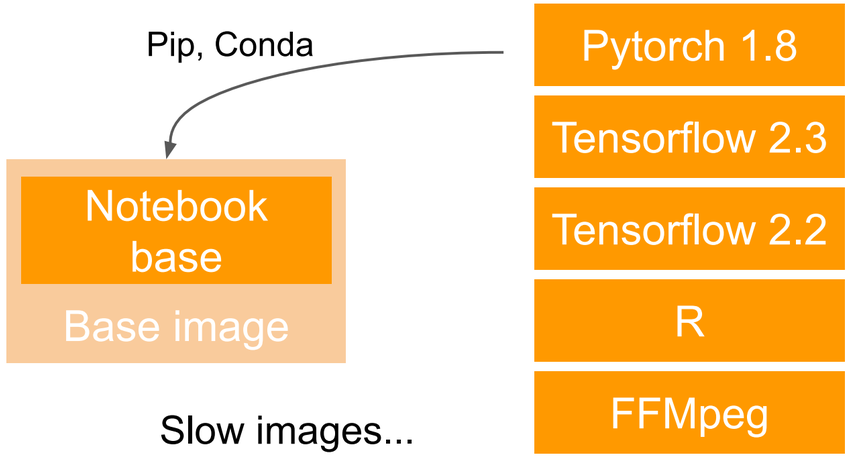


Multiple versions => Multiple images



Overweight images...

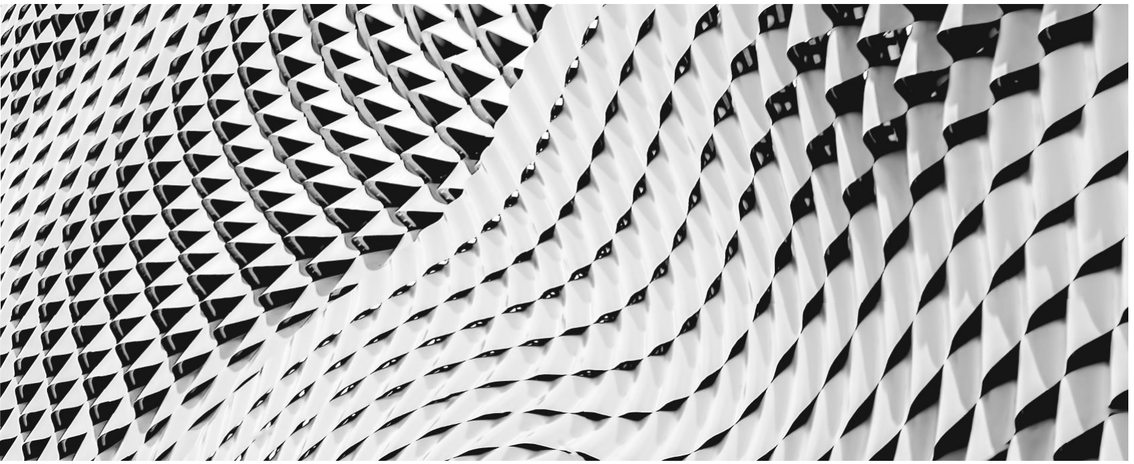
Multiple tools => Multiple images



Slow images...

They want more, what to do?

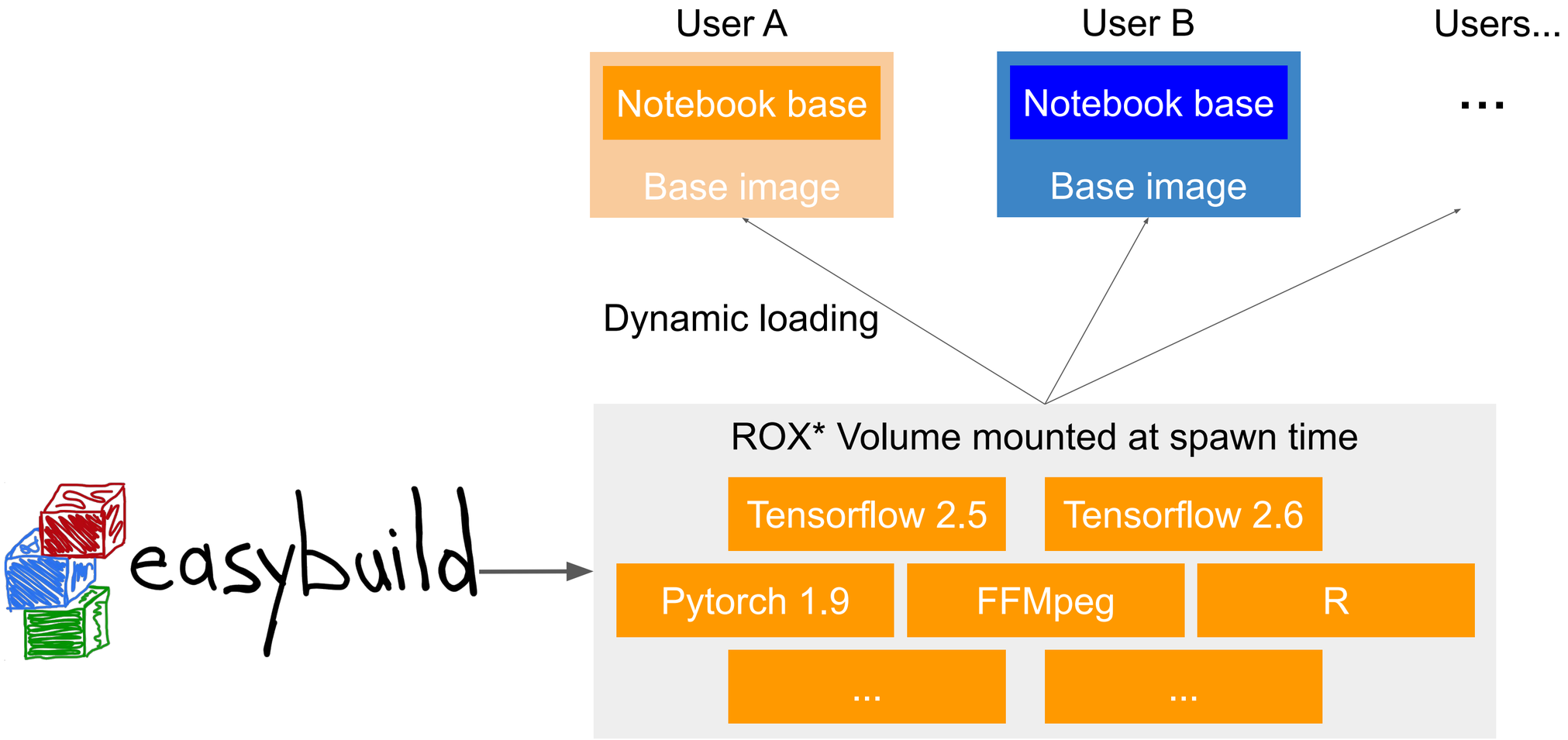
The solution



"You Did Good Son.
I'm Proud Of You."
Admiral Anderson

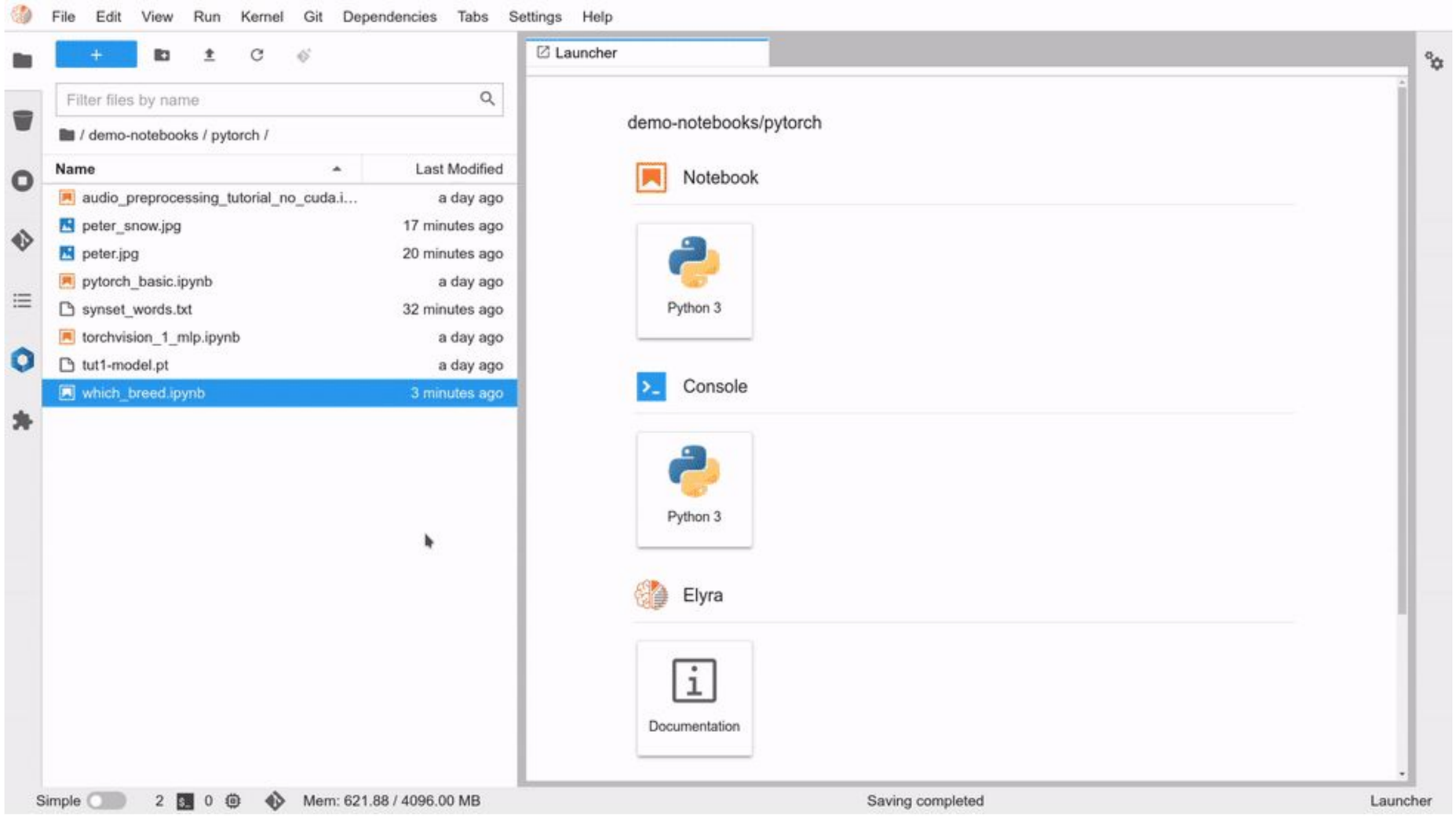
Red Hat

Enter the environment modules!



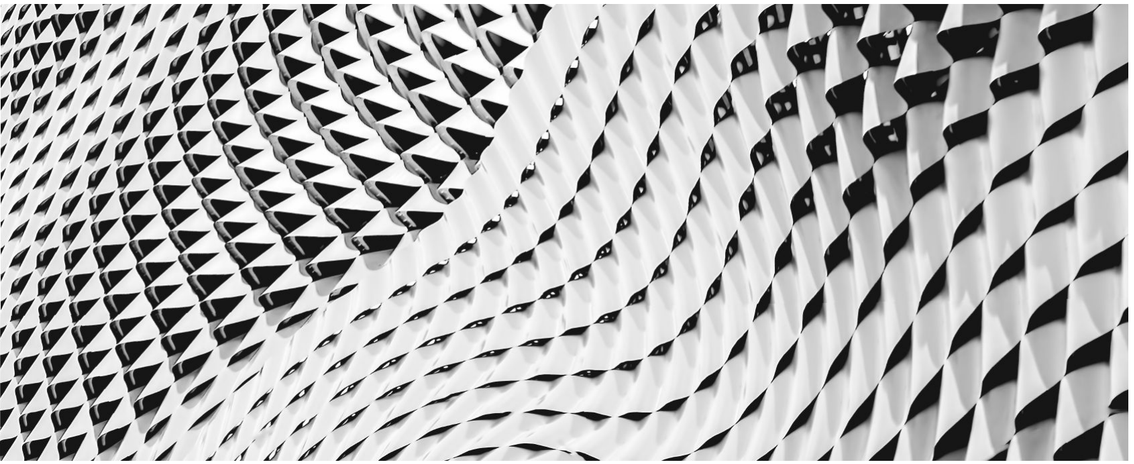
*ROX: Read-Only Many, a volume that can be mounted simultaneously into many containers

And a miraculous Jupyter Extension!



Red Hat

Demo

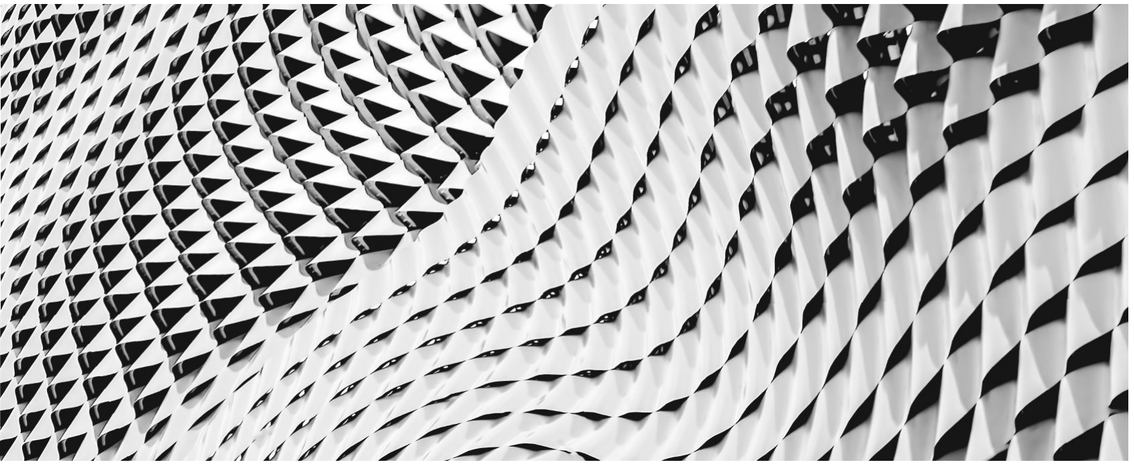


"Beam me up Scotty."

Captain James T. Kirk

Red Hat

What's Next?



"We're going to need a bigger boat."
Chief Brody

Red Hat

What's Next: Performance

- From preliminary tests, some packages provide better performance than their "pip" counterparts.

  E.g. Image recognition training with Tensorflow 2.6.0 => **17.6% faster**

- Concordant with other "optimized" versions of tools.

- Will require further tests and benchmarking.

Great opportunity for optimized environments, non x86 environments,...

What's Next: Modules lifecycle

- Self-build modules: EasyBuild documentation 💯, and tons of sweat...
- Module repo: synch mechanism, self-serve UI to select what to sync.
- Container image creation to directly package the required modules (self-sustained images).

Great opportunity to expand the community

References

- Open Data Hub project: http://opendatahub.io/
- Red Hat OpenShift Data Science:

  https://www.redhat.com/en/technologies/cloud-computing/openshift/openshift-data-science
- Easybuild+Lmod+Open Data Hub Project: https://github.com/guimou/odh-highlander

Thank you

Red Hat is the world's leading provider of enterprise open source software solutions. Award-winning support, training, and consulting services make Red Hat a trusted adviser to the Fortune 500.


linkedin.com/company/red-hat


youtube.com/user/RedHatVideos


facebook.com/redhatinc


twitter.com/RedHat

Red Hat