Flexible Tools for Data Science Education

Damien Eversmann
Chief Architect
Education

Will McGrath
Product Marketing Manager
Cloud Data Services Business Unit

EDUCAUSE

Red Hat

What we will discuss today

- History of Data Science
- Problem Statement
- Elements of a good solution
- How did Red Hat get involved with Machine Learning?
- And why Educational institutions?
- The Future

What is Data Science?

Data science is an interdisciplinary field that uses scientific methods, processes, algorithms and systems to extract knowledge and insights from noisy, structured and unstructured data, and apply knowledge from data across a broad range of application domains. Data science is related to data mining, machine learning and big data.

Source:
https://en.wikipedia.org/wiki/Data_science

Where Does It Come From

An Abbreviated History

1962 — "Data Analysis"
American statistician, John Tukey, uses the term to define a field close to data science.

1985 — Isn't it just Statistics
Statistician Chien-Fu Jeff Wu uses "data science" as a replacement term for "statistics".

1997 — It's better than Statistics
Wu again advocates "data science" over "statistics" as the latter just implies "accounting".

2001 — Statistics Applied
Computer Scientist William S. Cleveland proposes "data science" as a term for applied statistics.

2008 — Data Scientist
The title begins to be common in the private sector initially at LinkedIn and Facebook.

Today — Fact or Fiction
Some say it is a buzzword like "Big Data", but companies and enterprises everywhere won't make major moves without it.

Source:
https://en.wikipedia.org/wiki/Data_science
https://www2.isye.gatech.edu/~jeffwu/presentations/datascience.pdf
https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

Intelligent apps are just one of the ways we see Data Science in action

An intelligent app is an app where...

Part of the code was **written by a human**

Part of the code is a **model** created from **data and training**

Examples of intelligent applications

- **Recommendation engines**
  Netflix, Amazon, etc..
- **Virtual assistant**
  Siri, Alexa, etc...
- **Detecting fraudulent activity**
  Money laundering, spam, hacking, insurance
- **Quantifying risks and making smart decisions**
  Insurance, loans
- **Pattern detection**
  Images, videos: how many cars, humans, etc. ?
- **Analyze specialized data**
  Seismic data for oil and gas
- **Teach AI to play video games**
  AI opponents
- **Text analysis**
  Summarization, accuracy, offensive, plagiarism detection
- **Medical**
  Tumour detection
- **Customer retention**
  Predict who's about to leave

Poll Question: Do you teach or work in the data science field?

A. Yes, a significant amount of time
B. Only a little
C. Not yet, but looking at doing so in the future
D. No, mainly interested in learning more

The Problem Defined

Overhead on campus...

"Half my students use the 'laptop-ate-my-homework' excuse to ask for deadline extensions"

Gayatri, Professor

"I waste the first 3 hours of EVERY course helping set up the student's environments."

Sarah, Teaching Assistant

"So of course my laptop decided to blue screen of death an hour before the deadline. I panicked, and now every computer in the house has Pytorch on it."

Ashesh, Undergraduate Student

"Every year, I need to do more with less."

Diane, Chief Technology Officer

"I wish I had something else than Excel to teach Linear Regressions"

Pao-lu, Professor

"Hello IT. Have you tried turning off and on again?"

Roy, Faculty IT

"I've had to rename 'office hours' to 'tech support hours'"

Igor, Adjunct Professor

Pain points for educational institutions

- **High variability of needs** across faculties, teachers, and students
- **High peaks of activity** during live class and the last hour before deadline
- **High variability and low reliability** of student-owned devices
- Different classes can have **different or contradicting software requirements**
- **Waste of time, resources, and talent** on zero-value activities
- **Budgets, resources, and skills** are split between central IT and faculty IT
- **Rapidly evolving needs** based on subjects being taught
  - Data science was not taught 10 years ago
  - MLOps is not taught…yet

What Red Hat tells its commercial companies translates into opportunity for educational institutions

Talent shortage

Lack of key skills makes it difficult to find and retain talent

Lack of self-service access to AI/ML tools & infrastructure

Slows data scientists and developers from doing their job

Complexity to operationalize AI projects

Slow, manual, siloed operations slow AI lifecycle execution

Educating for the future isn't just next gen data scientists

Data Scientists

Data Engineers

Developers

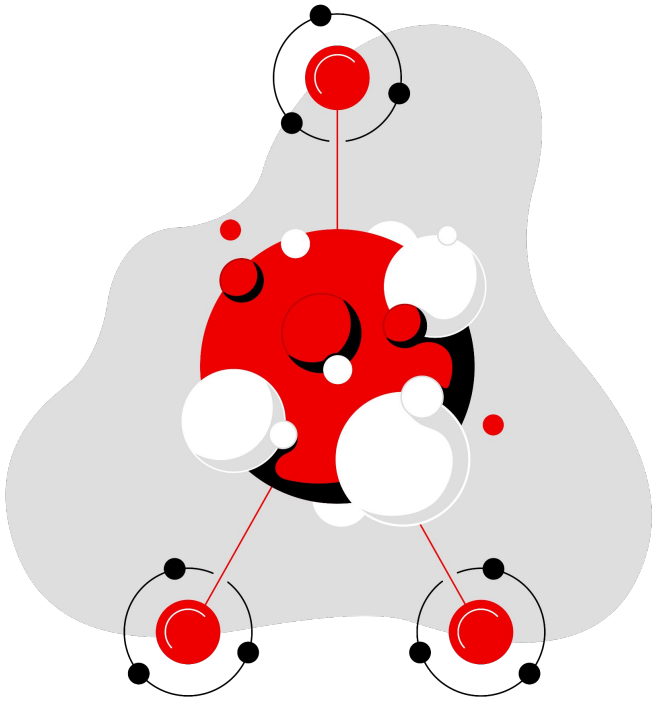
Architects

MLOps Engineers

Elements of a good solution

Open Source

Some of us think this is a no-brainer



- Open Source is secure and stable
  - More eyes on the code means fewer flaws and exploits
- Open Source is nimble and feature rich
  - Anyone can make changes and participation is encouraged
- Many/Most of the common Data Science tools are either entirely Open Source or are based in Open Source
- Most of all, Higher Education is the original Open Source
  - From Unix in the 70s and GNU in the 80s to Linux and beyond in the 90s, it came from EDU

Flexible and Customizable

Not “One Size Fits None”

- There are endless tools for Data Science, Artificial Intelligence and Machine Learning.
- The configurations of those tools increase the permutations exponentially.
- A good solution should provide an opinionated selection but endless possibilities.

Cloud-Friendly

Available to All

- Not all schools have multi-million dollar data centers.
  - They should still be able to teach and sponsor research in Data Science fields
- Not all students have high-end, multi-GPU personal computers
  - They should still be able to learn and do Data Science

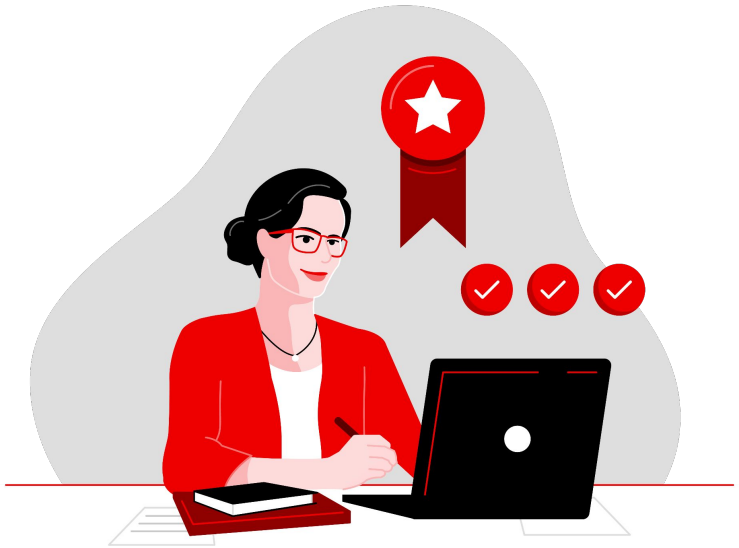
Rapidly Updating

Keeping up with the Joneses

- The pace of AI/ML research is break neck.
- The tools that support it appear and change and evolve along with it.
- A good solution should be able to do the same.

Easy

... as the push of a button



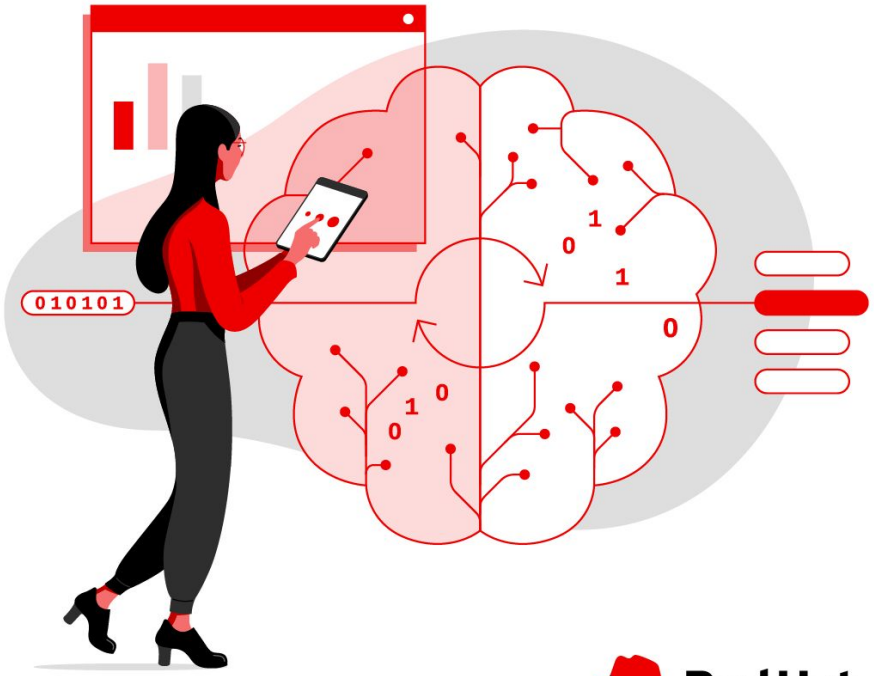
- These days, everyone is being asked to do more with less.
- This is especially true with Higher Education...
- ... And even more so in IT and Data Science.
- A good solution needs to be simple to install, configure and manage.

Poll Question: What elements of a data science instructional solution resonate the most with you?

A. Open source
B. Flexible and customizable
C. Cloud friendly
D. Rapidly updating
E. Easy
F. All of above
G. Other _____

How did Red Hat get involved with Machine Learning?

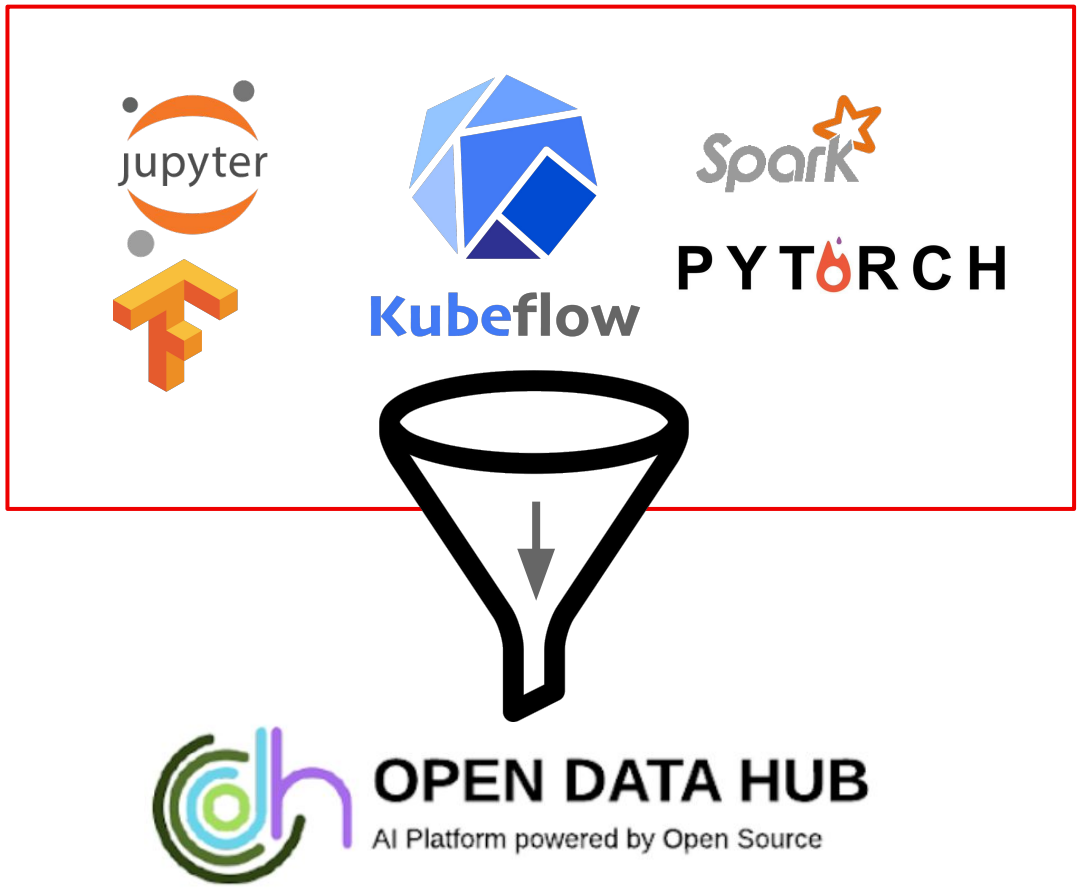
Open Data Hub: the origin story



- **Began as CI/CD engineering project for build insights**
  Terminate batch jobs early
- **Expanded to an all open source blueprint of AI technologies**
  Customer demand from a number of visitors to exec briefing center
- **Built boutique AI consulting service**
  Mainly about helping customers provide AI infra on Kubernetes
- **Introduced commercial version based on subset of components**
  Continued customer ask: can Red Hat support the open source components?

What is Open Data Hub?

100% open source-based ML architecture blueprint built for Kubernetes



Learn more: https://opendatahub.io

EDUCAUSE

Red Hat

Based on Open Data Hub and Operate First

Upstream code enhanced with operational excellence

Open Data Hub

Community driven upstream meta-project demonstrating AI/ML platform on Red Hat OpenShift comprised of open source projects

Operate First

Subset Open Data Hub operated at scale for community and university audiences to infuse operational excellence

Red Hat OpenShift Data Science

Subset of Operate First delivered as a cloud service on Red Hat OpenShift Managed on Amazon Web Services with optional ISV offerings

Open Data Hub: https://opendatahub.io
Operate First: https://openinfralabs.org/ and https://www.operate-first.cloud/

What is 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

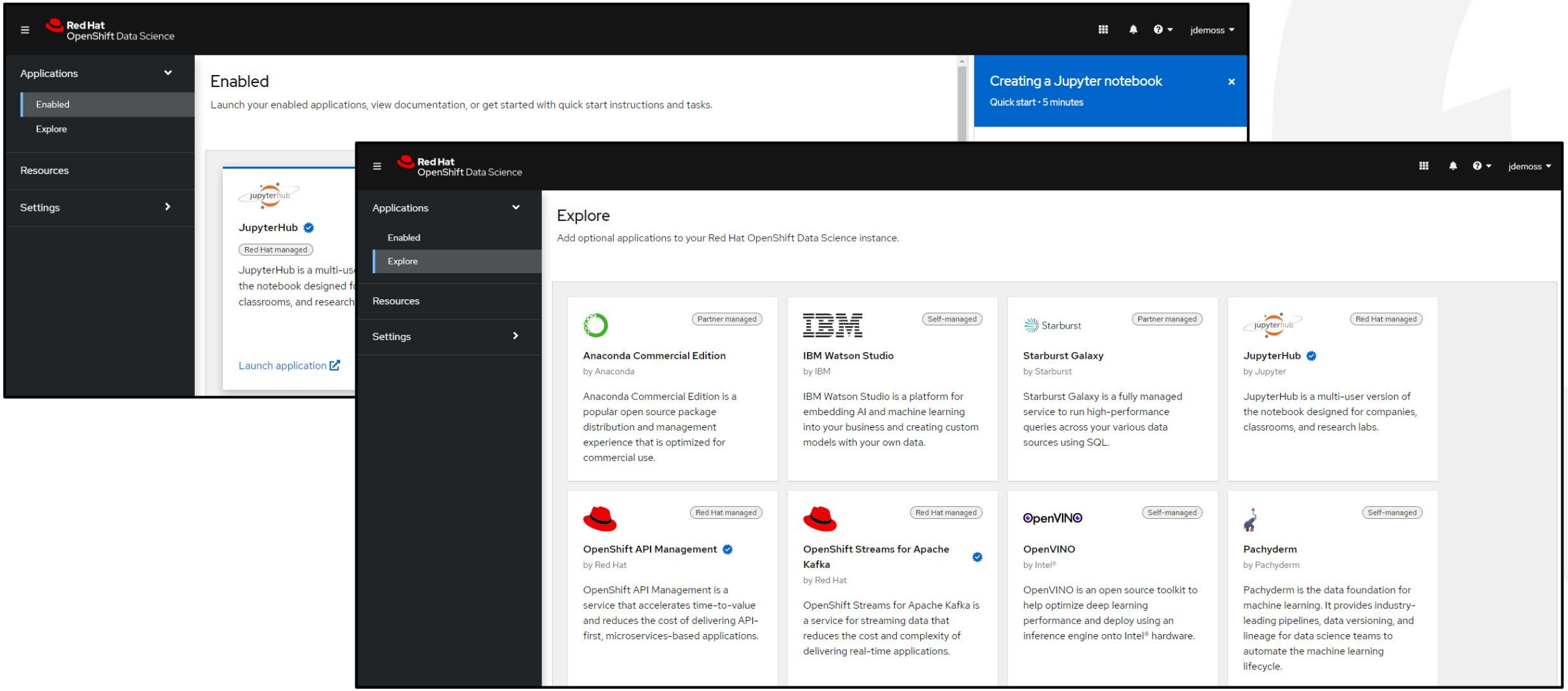
Core data science workflow

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

Dashboard user interface



EDUCAUSE Red Hat

Jupyter Spawner - including GPUs

Start a notebook server

Select options for your notebook server.

Notebook image

- Minimal Python
  Python v3.8
- PyTorch
  Python v3.8, PyTorch v1.8, CUDA
- Standard Data Science
  Python v3.8
- TensorFlow
  Python v3.8, TensorFlow v2.7, CUDA v11.4
- CUDA
  Python v3.8, CUDA v11.4

Deployment size

Container size

Small

Number of GPUs

0

Number of GPUs

1

- 0
- 1 ✓

Add more variables

Number of GPUs

0

- 0 ✓
- 1
- 2
- 3
- 4

Note: Only way to get GPUs for ROSA in next few months

Red Hat

Standard Jupyter Notebook Interface

File Edit View Run Kernel Git Tabs Settings Help

Filter files by name

/ test /

Name	Last Modified
my.NoteBook.ipynb	seconds ago
test.py	5 hours ago

my.NoteBook.ipynb

Code

```
[2]: a = "Hello World"
     print(a)

     Hello World

[ ]:
```

1000970000@:~

```
top - 20:45:28 up 15 days, 14:17,  0 users,  loa
Tasks:   6 total,   1 running,   5 sleeping,   0
%Cpu(s):  1.7 us,  1.2 sy,  0.0 ni, 96.6 id,  0.
MiB Mem : 255122.8 total, 160018.0 free,  21253.
MiB Swap:      0.0 total,      0.0 free,      0.

    PID USER      PR  NI    VIRT    RES    SHR S
      1 1000970+  20   0   17004   3120   2884 S
    109 1000970+  20   0  579868 107876  15628 S
    172 1000970+  20   0   35332  22188   3460 S
    398 1000970+  20   0  657748  48544  13628 S
    411 1000970+  20   0   35200  21752   3264 S
    446 1000970+  20   0   52052   3792   3372 R
```

Simple 2 1 Python 3 | Idle Saving completed Mode: Edit Ln 1, Col 1 my.NoteBook.ipynb

Why Educational Institutions?

BOSTON UNIVERSITY

- **Implemented interactive lecture and lab environment** for computer scientists and engineers based on Red Hat OpenShift Data Science
- **Currently over 300 users** including over 100 concurrent
- **Integrates with the Boston University online textbook material,** also authored using the Red Hat OpenShift Data Science
- **Fast time to solution:** cloud services environment enabled BU to configure and deploy in December for classes that started in January
- **Lowers cost:** auto-scales based on demand; enables bursty interactive use cases at optimized cost

Jonathan Appavoo (BU professor) : Steps toward open source education webinar

Red Hat OpenShift Data Science

Red Hat® OpenShift® Data Science provides a computing environment for students, faculty, and researchers that is:

Simple

Managed

Supported

Scalable

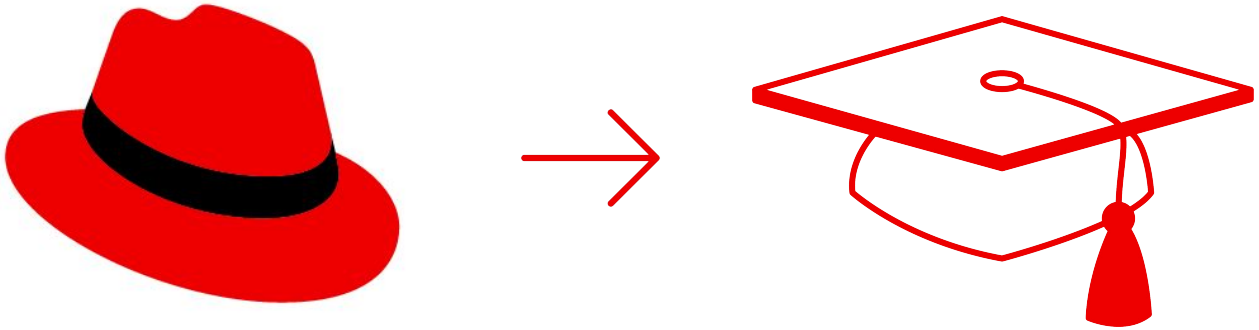
Benefits for faculty and research

- **Spend more time teaching,** less time debugging laptops
- **Default notebook images:** what the industry uses for data science
- **(BYO) custom notebook images:** computer science, stats, economics, psychology, etc...
- **Access** to technology partner software, if required
- **Consistent,** reliable, and fair environments for all
- **Not just for class:** research projects, publications, etc...

Benefits to students

- Browser-based environment
- Zero-install, any device
- Environment is available 24/7
- If it works from your device, it will work from:
  - Teacher's device
  - Friend's device
- Available during class and for assignments

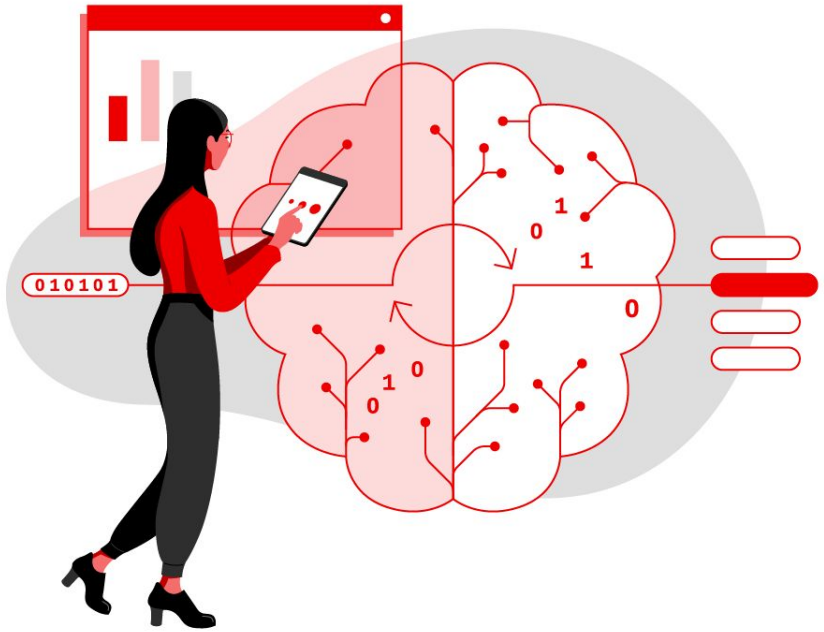
Ready your students ready for their careers



This hat can lead to another

The Future

What to expect in coming months



- On prem version - Beta in Sept
- GCP & Azure support
- Better MLOps capabilities through Kubeflow
- Running OpenShift Data Science in same cluster with Open Data Hub components

Red Hat OpenShift Data Science

Learn more ▸

Try it ▸

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

EDUCAUSE

Red Hat