Lessons in transitioning a data science team to JupyterHub

JupyterHub Workshop 07/22

DATASCIENCE

Alec Aivazis & Mike Polino

What is this talk about?

1. What we use Jupyter for and why **- 2 min**
2. Why we decided to move to JupyterHub **- 1 min**
3. Five challenges we’ve dealt with moving our team to JH **- 5 min**

What does DataScience do?

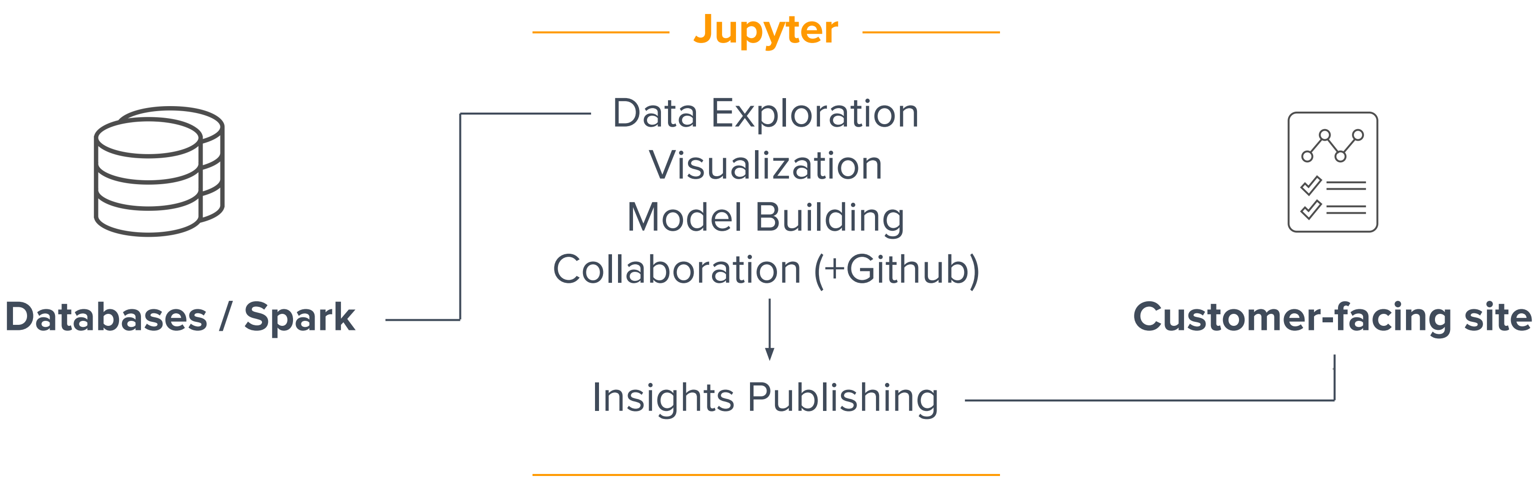
DataScience offers **consulting solutions** and **tools** that let companies get business value out of their data.

In particular, we do a lot of prediction based on large and complex datasets.

We use:

- Python a lot
- Spark
- Almost anything else for building tools

How have we used Jupyter?



Databases / Spark

Jupyter

Data Exploration
Visualization
Model Building
Collaboration (+Github)

Insights Publishing

Customer-facing site

Our custom Jupyter plugins/tools include credential management, data caching, and nbconvert workflow for publishing to customer-facing website.

Icons: database by Abhishek Rana via Noun Project, page by Alfredo Hernandez

What didn’t go well?

- VM drift, difficult debugging, reduced control for engineering team
- Discomfort in git (especially for diff’ing json docs)
- Limited by RAM on our laptops

We’re moving to an AWS-hosted JupyterHub, which brings us closer to our data and Spark clusters, gives control back to engineering team, lets us build better solutions for our needs.

This transition didn’t come without challenges. Here are five interesting things we experienced:

Challenge 1

Sharing notebooks across Docker containers without mounting volumes or using git

Solution: our restful content manager (open source, link below)

https://github.com/datascienceinc/RestfulContentManager

Challenge 2

```python
import os

while True:
    os.fork()
```

Solution: App Armor and compusec policies

Challenge 3

Users want control over their notebook server resources (RAM / cores)

Solution: Mesos and DockerSpawner, re-size accessible from the UI

Challenge 4

Analysts want to collaborate quickly, don’t want to “check out” code

Solution: locking mechanism and websockets

Challenge 5

Power users want to install notebook-specific Python packages

Solution: Python virtualenv per notebook

Thanks!

Alec (engineer): alec@datascience.com

Mike (product manager): mpolino@datascience.com