DATA SCIENCE IN THE OPEN CLOUD EXCHANGE MODEL

Using the Massachusetts Open Cloud Open Data Hub to perform Data Science experiments

Steven Huels - Red Hat AI Center of Excellence
Vaclav Pavlin - Red Hat AI Center of Excellence

January 26, 2019
THE PROPRIETARY PUBLIC CLOUD
The Good, the Bad, and the Evil of ‘The New Mainframe’

Benefits:
- Elasticity
- Integration
- Security
- OPEX vs CAPEX
- Service abstraction & time to value
- Operational excellence made accessible
- Built on OSS

The Cost:
- Vertical lock-in
- Data Gravity
- One way route: uses OSS but limited contribution back
- Life cycle dependency
- Black-box-services
- Reproducibility?

The Public Cloud has in effect re-invented the Mainframe: black-box services with a controlled ecosystem on leased hardware.
WHY NOT JUST USE A PRIVATE CLOUD

Challenges of a private cloud

- Operational complexity of deploying and managing a complex stack
- Lack of diversity in provided services
- Poor user experience
- Costly to support in terms of training and operations
WHAT ABOUT THE OPENSTACK CLOUDS

Vexxhost, OVH, CityCloud, Internap, Conoha, Fuga, Cloudwatt, Catalyst, Dreamhost, Elastx, Enter Cloud Suite, Open Telekom Cloud, Zetta, Ultimum, Limestone, Auro

PROS
Distribution of deployments exceeds the number of supported regions in any public cloud offering

CONS
None have the scale to attract the diversity of services of the public cloud
Divergent services, differing base images and flavors
No mandate or funds to address these issues
IS A DIFFERENT MODEL POSSIBLE?

An “Open Cloud eXchange (OCX)”
OPEN CLOUD EXCHANGE

What is it?

An alternative cloud model...

where many stakeholders, rather than just a single provider, participate in implementing and operating the cloud

multi-sided marketplace in which participants cooperate and compete with each other

users can freely choose which set of services to use from among any number of competing services and solutions
OPEN CLOUD EXCHANGE

Use cases

- Respond to occasional large scale demands
- Open source providers and other participants can stand up services
- Hardware and software vendors provide access to new technology and continuously deliver updates
- Organizations can jointly fund and operate a shared environment
SO ... THIS REALLY ISN’T THAT CRAZY

Let’s recap

- Current clouds are incredibly expensive...
- Much of industry locked out of current clouds...
- Lots of great open source software...
- Lots of great niche markets; markets important to us...
- Lots of users concerned by vendor lock in...
- This does not need to be AWS scale to be worth it...

"Past a certain scale; little advantage to economy of scale" — John Goodhue
THE MASSACHUSETTS COLLABORATORS

NetApp
EnterpriseDB
EMC2
Mellanox Technologies
Intel
Lenovo
UMASS
VERITAS
BROCADE
MIT
UMASS
TWO SIGMA
CISCO
redhat
BOSTON UNIVERSITY

CanonicaL
MathWorks
Juniper Networks
Riverbed
Oak Ridge National Laboratory
PLEXXi
Dell
U.S. Air Force
DataDirect Networks
Northeastern
MASSACHUSETTS OPEN CLOUD

What is it?

90,000 sq ft and growing with 10s of thousands of users

ELASTIC HARDWARE
- 250 cores
- 1.5 TB RAM

KAIZEN
- 2000 cores
- 20 TB RAM

STORAGE
- Ceph
- 2+ PB

RED HAT OPENSHIFT
Container Platform
THE OPEN DATA HUB IN THE MOC
Collaborate on a Data & AI platform for the Hybrid Cloud

- Meta-Project to integrate Open Source projects into a practical service oriented solution.
- Foster collaboration between communities, vendors, user-enterprises, operators, and academics following Open Source best practices.
- Focus on integration, service abstraction, and continuity to free users from owning the full complexity of the underlying stack.
- Ensure reproducibility by using Open Source.
THE OPEN DATA HUB IN THE MOC

Collaborate on a Data & AI platform for the Hybrid Cloud

- Offer flexible entry points (e.g. storage tier, app platform tier, workflow tier, pre-defined transparent-box services).

- Enable related projects to pick-and choose services, co-locate at the right layers.

- Operate the core platform as part of the MOC, provide easy access to Research and Open Source communities with full transparency and modularity.
EARLY ADOPTERS OF ODH IN MOC

Enabling researchers and business to analyze their data

- BestBees - [https://bestbees.com/](https://bestbees.com/)
  - New England Agriculture & Food Security - NASA DEVELOP Summer 2018 @ GSFC ([video](https://bestbees.com/))
- Citation Attribution Network

+ A lot of interest
# COMMON PATTERNS

Data-Centric Application Platform for the Hybrid Cloud

<table>
<thead>
<tr>
<th>SELF SERVICE UI / CLI &amp; SERVICE CATALOG</th>
<th>API &amp; SERVICE ROUTING</th>
<th>COMMON SERVICES</th>
</tr>
</thead>
<tbody>
<tr>
<td>MANAGEMENT CONSOLE / AIOPS (PROMETHEUS</td>
<td>PRE-DEFINED AI LIBRARY</td>
<td>DATA SCIENCE TOOLCHAIN</td>
</tr>
<tr>
<td>LINEAGE (CODE AND DATA)</td>
<td>PRIVATE MICROSERVICES</td>
<td></td>
</tr>
<tr>
<td>LANGUAGE RUNTIMES</td>
<td>ANALYTICS &amp; AI PROCESSING TOOLCHAINS</td>
<td></td>
</tr>
<tr>
<td>DEVOPS ALM (CI/CD, CODE &amp; DATA)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

SECURE HYBRID-CLOUD DATA-CENTRIC APP PLATFORM (KUBERNETES, LINUX, S3, KAFKA)
DATA SCIENCE IN THE OPEN DATA HUB

Early adopter data science experimentation workflow

Ceph + JupyterHub + Apache Spark + TensorFlow
DEMO

Two personas

Operations

Data Scientist
DEMO
OPEN DATA HUB IN CZECH REPUBLIC

Local universities are interested in working with Open Data Hub

- We are currently working with
  - Masaryk University in Brno and CERIT Scientific Cloud
  - AI Center at Czech Technical University in Prague

- Project examples
  - Dynamic Pricing in Open Cloud Model
  - AI for Dependency Recommendation
  - AIOps on Open Data Hub
# LESSONS LEARNT AND WHAT’S NEXT

What we learnt by running Open Data Hub in MOC and what we are doing next

## LESSONS LEARNT

<table>
<thead>
<tr>
<th>LESSONS LEARNT</th>
<th>WHAT’S NEXT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Capacity planning is hard for Data Science</td>
<td>Add components running internally to the Gitlab repo</td>
</tr>
<tr>
<td>OpenShift v3 is hard to operate at scale</td>
<td>Integrate AI/ML model serving</td>
</tr>
<tr>
<td>Aligning priorities and setting up communication &amp; process around operating a</td>
<td>Integrate AIOps &amp; AI Library</td>
</tr>
<tr>
<td>big data clusters is hard</td>
<td>Initiate the Open Cloud Marketplace</td>
</tr>
</tbody>
</table>
TRY IT YOURSELF

Interested in being an early adopter?

We are looking for collaborators on the Open Data Hub

- [OpenDataHub.io](#)
- Mailing list: [contributors@lists.opendatahub.io](mailto:contributors@lists.opendatahub.io)

We are looking for Early Adopters to exercise the MOC system

- **Contact**
  - Steven Huels ([shuels@redhat.com](mailto:shuels@redhat.com))
  - Sherard Griffin ([shgriffi@redhat.com](mailto:shgriffi@redhat.com))
  - Vašek Pavlín ([vasek@redhat.com](mailto:vasek@redhat.com))
THANK YOU