Data Lake as a Service

It's a lake.. but for data.. presented as a service

Openlab Summer Students Lightning Talks 2021

Muhammad Aditya Hilmy <aditya.hilmy@cern.ch>

06 / 09 / 2021
A bit of context

• We will have HL-LHC and other experiments coming online.
• Data volume expected to increase by >10x.
• We need to think about how to store and manage the data.
• The Data Lake is a place where experiments can ‘dump’ their data.
• …and scientists can ‘fish’ data from.
• The challenge: making sure the scientists can ‘fish’ easily.
Making data fishing easier

• The Data Lake has a lot of moving parts.
• The goal of the service is to hide the complexities of the Data Lake from the scientists.
• This way, scientists can focus their time on doing science-y stuff.
Full demo video: https://youtu.be/AxzAsXTEaxw
A humble beginning

• Started as an idea presented on CS3 2020 by the Rucio team [1]
• Developed “Rucio JupyterLab Extension” as a part of Google Summer of Code 2020 [2,3]
• Deployed the extension as “Data Lake as a Service” as a part of Openlab Online (😊) Summer Students Programme 2021

[1] https://indico.cern.ch/event/854707/contributions/3680520/
## The Data Lake as a Service

<table>
<thead>
<tr>
<th>Userspace</th>
<th>JupyterHub</th>
</tr>
</thead>
<tbody>
<tr>
<td>Userspace Mount</td>
<td>OpenID Connect Auth</td>
</tr>
<tr>
<td>Scratch FUSE Mount</td>
<td>OIDC Token Exchange</td>
</tr>
<tr>
<td>Data Staging FUSE Mount</td>
<td>Environment Selection</td>
</tr>
<tr>
<td>OIDC Token Autorenewal</td>
<td></td>
</tr>
<tr>
<td>OIDC Token Injection</td>
<td></td>
</tr>
<tr>
<td>Rucio JupyterLab Extension</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Cron</th>
<th>Storage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Scratch Cleanup</td>
<td>CephFS</td>
</tr>
<tr>
<td>EOS Subject Mapping</td>
<td>EOS</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Infrastructure</th>
</tr>
</thead>
<tbody>
<tr>
<td>Openstack</td>
</tr>
<tr>
<td>Kubernetes</td>
</tr>
<tr>
<td>FluxCD</td>
</tr>
<tr>
<td>GitLab CI</td>
</tr>
<tr>
<td>Docker</td>
</tr>
<tr>
<td>Helm</td>
</tr>
</tbody>
</table>

5 Muhammad Aditya Hilmy
A bit of early mishaps
I broke production system on my first week :) 

25-06-2021 11:31:22 - Spyridon Trigazis
Hello,

It was Muhammad Aditya Hilmy the deleted both pools.
2021-06-22 11:58:40.630 INFO neutron.wsgi [req-ed93c560-efe3-4070-a94f-e0366f218844 muhilmy 586f1e21-4512-4480-890b-756cf4b7facb - default default] 137.138.121.172,10.100.1.0 "DELETE /v2.0/lbaas/pools/bf36789b-a9b5-4dc9-aec8-c3ca41e13abf HTTP/1.1" status: 204 len: 149 time: 1.0106199

2021-06-22 11:58:11.909 13 INFO neutron.wsgi [req-df990818-9eb1-4e21-916a-cd028e229016 muhilmy 586f1e21-4512-4480-890b-756cf4b7facb - default default] 137.138.121.172,10.100.1.0 "DELETE /v2.0/lbaas/pools/58603cd6-cd6d-468b-84a5-333490415ee0 HTTP/1.1" status: 204 len: 149 time: 0.9526761

I will check with my colleagues if we can manually recreate the pools, but heat operations will fail since the pool IDs will be new.

Cheers,
Spyros
Use Cases

Data discovery and access

Data analysis
  Use the files directly in the JupyterLab interface to create plots.

Data preparation and processing
  Use the service to preprocess data in the Data Lake, and once done, upload it back to Rucio.

Data preservation
  Use the service to produce data and reupload them to the Data Lake

Submitting jobs to external service (remote computing)
  Users can use the convenience of the extension to browse data in Rucio and access the file PFN directly from the notebook code.
Data Lake as a Service

- Desktop Data Lake-as-a-Service
- Location-aware Remote Configuration
- More Kernel Support
- Remote Configuration from Notebook File
- Token Support for Direct Download
- File Upload Functionality
- Integration with SWAN
- Integration with ESCAPE ESAP
- Latency Hiding Layer
Acknowledgements

Here are the people who’ve helped me throughout the journey:

Riccardo Di Maria (IT-SC-RD)
Xavier Espinal (IT-SC-RD)
Alba Vendrell Moya (IT-SC-RD)
Rizart Dona (IT-SC-RD)
David Smith (IT-SC-RD)
Cristian Contescu (IT-ST-PDS)
Elvin Alin Sindrilaru (IT-ST-PDS)
Andreas Joachim Peters (IT-ST-PDS)

Mihai Patrascoiu (IT-ST-PDS)
Enrico Bocchi (IT-ST-GSS)
Diogo Castro (IT-ST-GSS)
Mario Lassnig (EP-ADP-CO)
Martin Barisits (EP-ADP-CO)
“Thank you”

- Albert Einstein, at some point in time

Attributions:
The SpongeBob GIF is a copyright of Viacom International
The stick figure is taken from XKCD (Randall Munroe)
This picture of Albert Einstein is taken from Wikimedia Commons
QUESTIONS?

mhilmy@hey.com

LinkedIn: Muhammad Aditya Hilmy
GitHub: didithilmy