Jupyter Notebooks as scientific gateways to access cloud computing and distributed storage

Fernando Aguilar
IFCA (CSIC-UC)
aguilarf@ifca.unican.es
XDC

• The eXtreme DataCloud is a software development and integration project

• Develops scalable technologies for federating storage resources and managing data in highly distributed computing environments
  • Focus efficient, policy driven and Quality of Service based DM

• The targeted platforms are the current and next generation e-Infrastructures deployed in Europe
  • European Open Science Cloud (EOSC)
  • The e-infrastructures used by the represented communities

• Addresses the EINFRA-21-2017 (b)-2: “Computing e-infrastructure with extreme large datasets”
  • Deal with heterogeneous datasets
  • Bring to TRL8 and include in a unified service catalogue services and prototype at least at TRL6
## XDC Consortium

<table>
<thead>
<tr>
<th>ID</th>
<th>Partner</th>
<th>Country</th>
<th>Represented Community</th>
<th>Tools and system</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>INFN (Lead)</td>
<td>IT</td>
<td>HEP/WLCG</td>
<td>INDIGO-Orchestrator</td>
</tr>
<tr>
<td>2</td>
<td>DESY</td>
<td>DE</td>
<td>Research with Photons (XFEL)</td>
<td>dCache</td>
</tr>
<tr>
<td>3</td>
<td>CERN</td>
<td>CH</td>
<td>HEP/WLCG</td>
<td>EOS, DYNAFED, FTS, RUCIO</td>
</tr>
<tr>
<td>4</td>
<td>AGH</td>
<td>PL</td>
<td></td>
<td>ONEDATA</td>
</tr>
<tr>
<td>5</td>
<td>ECRIN [ERIC]</td>
<td></td>
<td>Medical data</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>UC</td>
<td>ES</td>
<td>Lifewatch</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>CNRS</td>
<td>FR</td>
<td>Astro [CTA and LSST]</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>EGI.eu</td>
<td>NL</td>
<td>EGI communities</td>
<td></td>
</tr>
</tbody>
</table>

- 8 partners, 7 countries
- 6 research communities represented + EGI
- XDC Total Budget: 3.07Meuros
What is Jupyter?

- Notebooks are documents produced by Jupyter Notebook App.
- Contains both source code and enriched text.
- Understood by humans, actionable by machines: scripts, data analytics, etc.
- Useful for teaching, user applications.
- Different kernels/programming languages.
Jupyter: Main Menu
Jupyter Hub

- JupyterHub brings the power of notebooks to groups of users.
- It gives users access to computational environments and resources without burdening the users with installation and maintenance tasks.
- Users with their own workspace.
- Features:
  - **Customizable** - JupyterHub can be used to serve a variety of environments. It supports dozens of kernels with the Jupyter server.
  - **Flexible** - Authentication is pluggable, supporting a number of authentication protocols (such as OAuth and GitHub).
  - **Scalable** - JupyterHub is container-friendly.
  - **Portable** - JupyterHub is entirely open-source.
Docker + AAI

- JupyterHub adopts DockerSpawner to deploy the user workspace.
- A Docker image can be configured with any software/package required by the user.
- It can be configured to perform any kind of actions when the user logs in, logs out, etc.
- The Authentication and Authorization is compatible with multiple social IDs (Github, Google, etc.).
- It can also be configured with AAI standards, like OpenID Connect.
- Information about the user (username, tokens) can be sent to the docker container deployed as workspace for the user.
OpenID-Connect - INDIGO IAM

Flexible authentication support (SAML, X.509, OpenID Connect, username/password, ...)

Account linking

Registration service for moderated and automatic user enrollment

AUP enforcement support

Mobile-friendly organization management tools

Easy integration in off-the-shelf components thanks to OpenID Connect/OAuth

Slide: Andrea Ceccanti
Scientific Gateway

- Integrating different components using the same AAI, a new complete environment can be deployed, including all the required components for researchers.
- Data gathering, configuration, programming, visualization: JupyterHub
- Computing needs: PaaS Orchestrator (jobs submission)
- Storage needs: Onedata. Cloud storage.
- INDIGO IAM and OIDC standard are the “glue” to integrate the different elements.
INDIGO - PaaS Orchestrator

Kind of “batch system” to send “jobs” to Cloud Computing resources.

Collects high-level deployment requests and translate them into action to coordinate resources interacting with the underlying cloud infrastructures.

Allows to implement workflows with different steps (data ingestion, data processing, etc.).

New features are developed in XDC project.
Onedata

Distributed storage space to store not only data, but also customized metadata.

Organized in Spaces (user, can be shared), providers and zones.

One “OneZone” federates multiple providers, that can be geographically distributed.

Data can be access via web, but is POSIX-compliant (directly mounted).

New features are developed in XDC project.
Use Case Example

Data Ingestion

PaaS Orchestrator
To be Integrated Docker Ready

AAI

Welcome to eXtreme-DataCloud

SAT

Repo

Meteo

GUI

Sign in

Water Quality Model

PaaS Orchestrator
Integrated Docker Ready

Onedata

Water Quality Model

AAI
Example

Sign in with XDC IAM

Welcome to eXtreme-DataCloud

Username

Password

Sign in

Files Running Clusters

Select items to perform actions on them.

- datasets
  - XDC.ipynb
    - Running 6 hours ago 7.51 kB
  - README.md
    - 20 days ago 49 B
  - test.sh
    - 20 days ago 1.06 kB
  - XDC_nb.py
    - 5 days ago 43 kB
Example
Example

```python
In [1]: %run -i XDC_nb.py
%matplotlib notebook
menu
```

---

**Data Ingestion**

- **Onedata to**: MDAuNWxvY2F0dW9udGh5Xp
- **Reservoirs/...**: CDP  Sanabria
- **Initial Date**: 09/01/2019
- **End Date**: 09/08/2019
- **Actions**: cloud_mask, cloud_coverage, list_files, download_sat_data

---

**Searching models**

```sql
Job List
---
Job ID: 11e0d492-8cb0-a123-8e9e-024269aad334 | Creation time: 2019-09-11T07:05+0000 | Status: CREATE_FAILED
Job ID: 11e0d087-93b8-b23-8e9e-024269aad334 | Creation time: 2019-09-06T09:20+0000 | Status: CREATE_COMPLETE
Job ID: 11e0d07a-9f65-8a02-b99e-024269aad334 | Creation time: 2019-09-06T09:09+0000 | Status: CREATE_FAILED
Job ID: 11e0d0a3-8b39-b09a-024268aad334 | Creation time: 2019-09-10T02:54+0000 | Status: CREATE_COMPLETE
Job ID: 11e0d01f-6b34-8e09e-024269aad334 | Creation time: 2019-09-05T13:31+0000 | Status: CREATE_COMPLETE
```

---

**Requesting access token**

- **Status code**: 200
  ```json
  "cloudProviderName": "provider-RECAS-BARI",
  "createdBy": {
    "issuer": "https://iam.extreme-datacloud.eu/",
    "subject": "e4b6b9f-0e94-40c2-b76d-bf2335f582684"
  }
  ```

---
Example
Thanks!

Fernando Aguilar
aguilarf@ifca.unican.es