Rucio - Exascale Data Management

Ruturaj Gujar
Project Details

- Organization: Rucio
- Mentors:
  - Martin Barisits
  - Mario Lassnig
  - Thomas Beerman
  - Cedric Serfon
- Work Report
About Rucio

- It is a large scale data management system.
- It provides the functionality to organize, manage and access a large amount of scientific data (in the order of petabytes).
- It also provides monitoring and data analytics.
Project Tasks

1. Add Username/Password authentication type to Rucio WebUI.
3. Create a dataset following mechanism.
Username/Password Authentication

- Rucio only supported x509 certificate authentication.
- Need was to add a new username/password authentication for the clients.
- Not all clients have userpass credentials.
- So based on the option provided in the config file, the corresponding auth type is invoked.
Username/Password Authentication

Login to Rucio

Username
Enter username here...

Password
Enter password here...

Login
Single Sign-On

- Some users already have a CERN account.
- So instead of asking them to create a new rucio account, provide an option to login with CERN SSO.
- This would reduce the need to login to Rucio separately if the user is already logged in to CERN SSO.
- SAML was used to make requests to the CERN SSO.
Dataset Following Mechanism

**Problem**
- Many operations can affect a dataset or its rules and can have effects on the end users.
  - **File loss:** The files that are lost or corrupted are removed from the dataset which change the content of the dataset.
  - **Dataset erased:** The dataset was erased because it was buggy or superseded by a new version.
  - **Lifetime expiration:** All datasets have a finite lifetime. If the lifetime of a dataset expires, users can ask for an extension of the lifetime.
- All these operations are run without notifying the users.
- So users really interested in these events might be at a disadvantage.
Dataset Following Mechanism

Solution

- Create a following mechanism for the datasets.
- Users can select a particular dataset and start following it.
- As and when any event occurs, entries are made into the database with the Account IDs of the users following that dataset.
- A daemon will then summarise all the events affecting a dataset and send an auto-generated report to each user following the dataset.
Things Learnt

- Working with large codebase
- Coding conventions to be followed
- Way to communicate with the mentors
- Improved docker skills
- Worked with webpy and SAML
- Created REST API
- Followed test driven development
Thank You!