Science & Technology Facilities Council

UK Research and Innovation

EGI Dataset Accounting and the WLCG

Adrian Coveney, Greg Corbett

2018-07-10

CHEP 2018

Outline

1. The APEL Team
2. Current usage Accounting
3. EGI-Engage and EOSC-hub
4. Dataset Definition
5. Dataset Usage Accounting
6. EGI-Engage Prototype
7. Transferring the Data
8. Prototype Development
9. Conclusion
10. Contacts

The APEL Team

- We develop and run the centralized usage accounting system on behalf of the WLCG and EGI infrastructures
- We work closely with the GOCDB team (configuration management data base also operated for WLCG and EGI)
- Both based at STFC RAL

Current Usage Accounting

- APEL provides grid job, cloud VM, and storage space usage accounting
- Collects records from WLCG and partner infrastructures into a central Repository
- Aggregates the records by various groupings – date, user, VO, number of processors, etc.
- Sends the processed data to the Accounting Portal for display

EGI-Engage and EOSC-hub

- EGI-Engage
  - EU H2020 project to expand offered capabilities, by engaging with large infrastructures, long-tail of science, industry/SMEs
  - One objective was a prototype open data platform
- EOSC-hub
  - EU H2020 project to create integration and management systems for the future European Open Science Cloud
  - Brings together EGI, EUDAT, INDIGO-DataCloud

Dataset Definition

- A logical set of files which may exist in several places at once but to which it is possible to assign some form of persistent unique identifier
- Treated separately to storage space accounting which accounts for disk allocation and usage without concern over how it is used
- For accounting, it is assumed that this unique identifier is available
- File-based accounting vs. uniquely identified dataset accounting
  - In the area of data accounting, WLCG is mainly concerned with the optimisation of storage space by minimising the storage of data that is infrequently read

Dataset Usage Accounting

- Aim to enable:
  - infrastructure managers to make replication decisions, i.e. moving data closer to where it is used to make more efficient use of the infrastructure
  - scientists to assess dataset impact by measuring how many times datasets are reused and thus how popular they are

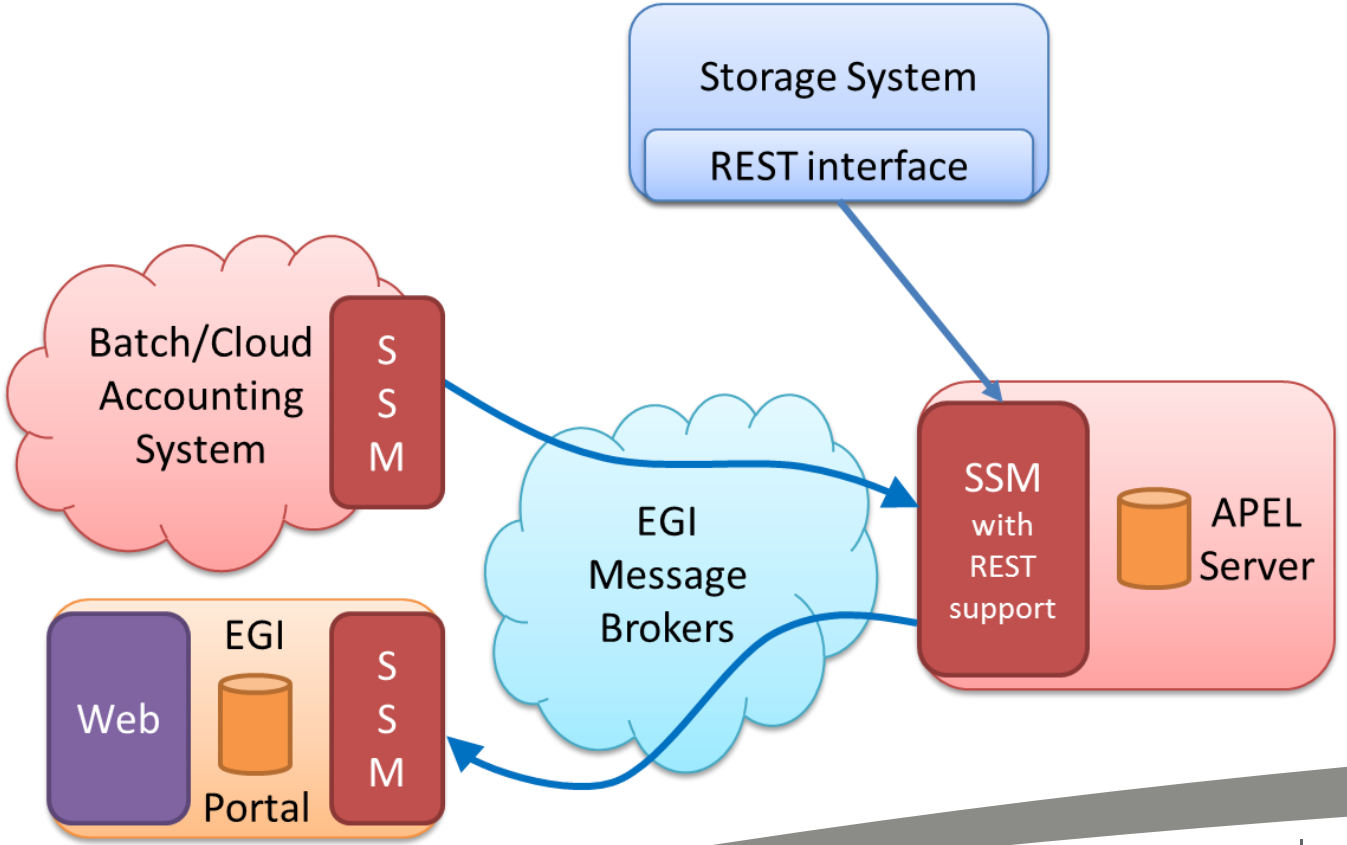
- Do this by collecting metrics such as number of access events for each PID

EGI-Engage Prototype

- Performed survey to see what metrics people considered important
- Used Onedata as the source of test data
  - a global data management system, providing access to distributed storage resources
- Focused on getting basic metrics out of Onedata and integrating that with the APEL accounting system
- New record format designed as extension to the Open Grid Forum Usage Record version 2 to enable dataset usage to be collected

Transferring the Data

- For current production services, records sent via the EGI Message Brokers using SSM (developed by APEL)
- Implemented a new method for obtaining accounting records by querying the Onedata REST API
- Modified SSM can be set to pull records from this interface for later loading into the repository


Storage System
REST interface
Batch/Cloud Accounting System
S S M
EGI Message Brokers
SSM with REST support
APEL Server
Web
EGI
Portal
S S M

Prototype Development

- At the time, metrics from Onedata API were limited to resource provider metrics but we were still able to iterate the design

- Record changes over the course of development:
  - Split access events into read and write
  - Added support for different types of data set ID (non-DOI unique IDs, such as PIDs)
  - Added a way to identify systems

Conclusion

- Basic dataset accounting prototype developed that introduced a new method for retrieving metrics to the APEL software system
- Further development needed as part of EOSC-hub as Onedata matures
- Additionally, need to investigate integration with EUDAT sources of dataset usage metrics (as part of overarching EOSC-hub objective)
- Aim to provide a unified view on datasets across these different infrastructures

Contacts

- WLCG Accounting Task Force
  - wlcg-accounting@cern.ch
- APEL Team
  - apel-admins@stfc.ac.uk

Thank you