Contribution ID: **19** Type: **not specified**

DataNet: A flexible metadata overlay over file resources

Monday 17 November 2014 15:55 (20 minutes)

Managing and sharing data stored in files results in a challenge due to data amounts produced by various scientific experiments [1]. While solutions such as Globus Online [2] focus on file transfer and synchronization, in this work we propose an additional layer of metadata over file resources which helps to categorize and structure the data, as well as to make it efficient in integration with web-based research gateways. A basic concept of the proposed solution [3] is a data model consisting of entities built from primitive types such as numbers, texts and also from files and relationships among different entities. This allows for building complex data structure definitions and mix metadata and file data into a single model tailored for a given scientific field. A data model becomes actionable after being deployed as a data repository which is done automatically by the proposed framework by using one of the available PaaS (platform-as-a-service) platforms and is exposed to the world as a REST service, which can be accessed from any computing site or a personal computer through the HTTP protocol. Data stored in such a repository can be shared by using various access policies (e.g. user-based or group-based) and can be managed from a wide range of applications. The repository is a self-contained application which can be scaled to improve transfer throughput and can integrate many underlying file storage technologies (currently it supports the GridFTP protocol). The generated REST interface allows data querying and file transfers directly from user web browsers without going through additional servers (this is possible thanks to using the CORS mechanism which is now supported by all major web browsers including mobiles).

Using a PaaS platform as a deployment base for the repository gives an advantage of extending it with different metadata storage backends which can be more suitable for handling metadata schema of certain data models while keeping the source model unchanged. The framework supports it by a a plugin system for different storage backends. Such flexible approach allows to adapt the platform to specific requirements without rewriting everything from scratch.

Using a single web endpoint for a repository gives the impression of using a cloud-based service to end users and other services (user credential delegation is also supported) while reusing existing storage facilities maintained in computing centers.

Acknowledgements. This research has been supported by the European Union within the European Regional Development Fund program POIG.02.03.00-00-096/10 as part of the PL-Grid Plus project.

References

[1] Witt, S.D., Sinclair, R., Sansum, A., Wilson, M.: Managing large data volumes from scientific facilities. ERCIM News 2012(89) (2012)

[2] Foster, I.: Globus Online: Accelerating and Democratizing Science through Cloud-Based Services, Internet Computing, IEEE , vol. 15, no. 3, pp. 70,73, May-June 2011

[3] Harężlak, D., Kasztelnik, M., Pawlik, M., Wilk, B., and Bubak, M.: A Lightweight Method of Metadata and Data Management with DataNet, eScience on Distributed Computing Infrastructure, Eds. Bubak, M., Kitowski, J., Wiatr, K., Springer International Publishing, Lecture Notes in Computer Science, vol. 8500, 2014, pp. 164-177

Authors: Mr WILK, Bartosz (ACC CYFRONET AGH); HARĘŻLAK, Daniel (A); Mr PAWLIK, Maciej (ACC CYFRONET AGH); Mr KASZTELNIK, Marek (ACC CYFRONET AGH); BUBAK, Marian (A)

Presenter: HARĘŻLAK, Daniel (A)

Session Classification: Technology and research

Track Classification: Technology and research