Contribution ID: **75** Type: **not specified**

Towards smart file-based data stores

Monday 30 January 2017 17:30 (20 minutes)

Nowadays, due to the data deluge and the need for the high availability of data, online file-based data stores have gained an unprecedented role in facilitating data storage, backup and sharing [1]. Up to date, the role of these file storage systems has been, largely, passive i.e. they host files and serve files to clients upon request. The simplistic approach of these file data stores means that they are easily deployed and integrate into other applications but may have limitations when hosted files are part of larger distributed data-oriented computations. First, data locality plays a crucial role on the performance of a data-oriented application, file servers may be too far from the computation or may have unreliable network between computation and data which will introduce bottlenecks and overhead in the running application. Second, larger computations tend to produce many intermediate result files which can easily inundate a file data store either from capacity or network limitations.
Our proposed approach tackles these two points by proposing a hybrid data- compute store where data stores can have a limited role in computing thus bringing together computation and data; this is extension of the concept presented in[2] and [3]. The main concept of our solution is that, in many scientific applications, data and computation are tightly coupled thus it makes sense to store the functions alongside the data in a unified database. One simple example is transcoding of images e.g. two same image files with different resolution. By capturing this information at the data store as part of the file metadata we can introduce some optimization routines. Instead of storing multiple images at different resolutions we can store one raw image and a set of transcoder functions that get called by the database when a particular image with a resolution is requested. The implication of mixing functions and data together means that datastores can prioritize on storage space by, safely, removing data which can be regenerated from the stored functions. As one can imagine this concept can be extended to larger computations such as work where one file is subsequently transformed into many other files which are all linked together through workflow functions.

References

[1] S. Koulouzis, A. Belloum, M. Bubak, P. Lamata, D. Nolte, D. Vasyunin, C. de Laat, Distributed Data Management Service for VPH Applications, IEEE Internet Computing 20 (2), 34-4, 2016

[2] R. Cushing, M. Bubak, A. Belloum, C. de Laat, Beyond scientific workflows: Networked open processes, IEEE 9th International Conference on eScience, 357-364, 2013

[3] R Cushing, A Belloum, M Bubak, C de Laat, Towards a data processing plane: An automata-based distributed dynamic data processing model, Future Generation Computer Systems 59, 21-32, 2016

Author: CUSHING, Reginald (University of Amsterdam)

Co-authors: BELLOUM, Adam (University of Amsterdam); BUBAK, Marian (AGH Krakow); Dr KOULOUZIS, Spiros (University of Amsterdam)

Presenters: CUSHING, Reginald (University of Amsterdam); Dr KOULOUZIS, Spiros (University of Amsterdam)

Session Classification: Technology