Contribution ID: **43**

Type: **Presentation**

Supporting Big Data Processing via Science Gateways

Wednesday, 11 November 2015 17:20 (20 minutes)

With the rapid increase of data volumes in scientific computations, the importance of utilising parallel and distributed computing paradigms in data processing is becoming more and more important. Hadoop is an open source implementation of the MapReduce framework supporting processing large datasets in parallel and on multiple nodes in a reliable and fault-tolerant manner. Scientific workflow systems and science gateways are high level environments to facilitate the development, orchestration and execution of complex experiments from a user-friendly graphical user interface. Integrating MapReduce/Hadoop with such workflow systems and science gateways enables scientists to conduct complex data intensive experiments utilising the power of the MapReduce paradigm from the convenience provided by science gateway frameworks. This presentation describes an approach to integrate MapReduce/Hadoop with scientific workflows and science gateways.

As workflow management systems typically allow a node to execute a job on a compute infrastructure, the task of integration can be translated into the problem of running the MapReduce job in a workflow node. The input and output files of the MapReduce job have to be mapped into the inputs and outputs of a workflow node. Besides executing the MapReduce job, the necessary execution environment (the Hadoop cluster) should also be transparently set up before and destroyed after execution. These operations should also be carried out from the workflow without further user intervention. Therefore, the concept of infrastructure aware workflow is utilised where first the necessary execution environment is created dynamically in the cloud, followed by the execution of workflow tasks, and finally breaking down of the infrastructure releasing resources.

As implementation environment for the above concept, the WS-PGRADE/gUSE science gateway framework and its workflow solution has been utilized. However, the solution is generic and can also be applied to other grid or cloud based workflow systems. Two different approaches have been implemented and compared: the Single Node Method where the above described process is implemented as a single workflow node, and the Three Node Method where the steps of creating the Hadoop cluster, executing the MapReduce jobs, and destroying the Hadoop execution environment are separated. While the Single Node Method is efficient when embedding a single MapReduce experiment into a workflow, the Three Job Method allows more flexibility for workflow developers and results in better performance in case of multiple MapReduce experiments that can share the same Hadoop cluster. Both approaches support multiple storage solutions for input and output data, including local files on the science gateway, and also cloud-based storage systems such as Swift object storage and Amazon S3. These storage types can be freely mixed and matched when defining input and output data sources/destinations of the workflow.

The current implementation supports OpenStack based clouds with a more generic solution including OpenNebula and generic EGI Federated Cloud support on its way.

The presentation will describe the implementation concept and environment, will provide benchmarking experiments regarding the efficiency of the implemented approaches, and demonstrate how the solution can be utilised by scientific user communities.

Summary

This work is partially funded by the CloudSME Cloud-
Based Simulation platform for Manufacturing and Engineering
Project No. 608886 (FP7-2013-NMP-ICT-FOF).

Primary author: KISS, Tamas (University of Westminster, London, UK)

Co-authors: TERSTYANSZKY, Gabor (University of Westminster); Dr CASTELLLI, Giuliano (University of Westminster); KACSUK, Peter (MTA SZTAKI); Mr GUGNANI, Shashank (BITS-Pilani K.K. Birla Goa Campus, Goa, India)

Presenter: KISS, Tamas (University of Westminster, London, UK)

Session Classification: Exploiting the EGI Federated clouds - Paas & SaaS workshop