Contribution ID: **19** Type: **Presentation (15' + 5' for questions)**

Optimizing Cloud Resource Allocation in Containerized Bioinformatics Workflows Using the Cloud Monitoring Kit (CMK)

Tuesday, 29 October 2024 17:10 (20 minutes)

Workflow languages have become indispensable for defining reproducible and scalable data analysis pipelines across various scientific domains, including bioinformatics, medical imaging, astronomy, high-energy physics, and machine learning. In recent years, languages such as the Common Workflow Language (CWL), Workflow Description Language (WDL), and Nextflow have gained significant traction alongside established solutions like Snakemake and Galaxy workflows.

Despite these advancements, resource allocation and monitoring in cloud environments remain significant challenges. Scientific tools often utilize assigned resources irregularly, leading to inefficiencies. Each analytical task specifies its required resources—such as CPUs, memory, and disk space—but selecting appropriate values is critical to ensure sufficient resources without over-provisioning.

To address these issues, the Cloud Monitoring Kit (CMK) was designed as a flexible, event-driven architecture, to generate uniform, aggregated metrics from containerized workflow tasks originating from different workflow management systems in a cloud environment. CMK offers essential insights through intuitive dashboards that display individual and aggregated metrics relevant to job performance. Developers can leverage CMK to monitor resource consumption and adjust system configurations during development or tool integration, enhancing efficiency and performance. Operations staff benefit from continuous performance monitoring and troubleshooting capabilities, crucial for maintaining system reliability. Scientists gain a robust analytical platform to scrutinize data, facilitating informed decisions regarding system configurations. The adaptability of CMK makes it particularly valuable where precise resource management and systematic optimization are essential.

In this contribution, we discuss our experiences implementing CMK in an industrial AWS cloud environment for processing bioinformatics data. We summarize the lessons learned during this process, highlighting the benefits and limitations of using CMK in a real-world setting. Furthermore, we explore how the data collected by CMK can be utilized to improve and optimize resource usage by informing task resource assignments. By closing the loop between monitoring and resource allocation, it is possible to assign informed values to task resources, reducing inefficiencies caused by over-provisioning or underutilization. The implementation of the CMK architecture for AWS Batch is available at https://github.com/biobam/cmk

GM and RN would like to thank Grant PID2020-113126RB-I00 funded by MICIU/AEI/10.13039/501100011033.

Primary authors: NICA, Robert (Universitat Politècnica de València); Dr GÖTZ, Stefan (BioBam Bioinformatics S.L.); MOLTÓ, Germán (Universitat Politècnica de València)

Presenter: NICA, Robert (Universitat Politècnica de València)

Session Classification: IBERGRID

Track Classification: R&D for computing services, networking, and data-driven science