Contribution ID: **8** Type: **Demonstration**

Notebook-as-a-VRE (NaaVRE): Scaling Jupyter Notebooks

Thursday, 21 October 2021 13:00 (10 minutes)

The study of many scientific problems concerning environmental challenges sets significant computational demands, such as large data volumes, advanced modeling techniques, and distributed computing facilitates. To conduct such investigations, a researcher often has to reuse virtual assets, e.g., observational data or images, AI models, operational workflows, and infrastructure services from different parties, for building computational experiments. Jupyter in such a scenario, allows researchers to effectively implement their experimental logic using scripting languages to document and share experiments with their necessary inputs and parameters.
In the case of ecological and biodiversity scientists, AI and statistical models are used to analyze a host of different observations, such as specimen records, citizen science observations, eDNA, and remote sensing imaging. For example, high-resolution Light Detection and Ranging (LiDAR) datasets are widely used to monitor changes in an ecosystem structure, and to predict that the distribution of species across space and time. Most often, Jupyter notebooks are used in this analysis,
as they allow researchers to effectively implement an experimental logic using languages such as Python. However, Jupyter faces challenges of utilizing remote infrastructure:
Difficulty to find and reuse a notebook at the cell level. The lack of metadata hampers the discovery of useful fragments of code (namely Cells).
Lackingflexibility and portability to reuse a notebook as part of a workflow. The tightly coupled functions of a notebook, i.e. library dependencies make the reusability and portability of the code fragments difficult.
Difficulty to scale the notebook to remote infrastructures. Current notebook environments, like Jupyter Hub, use pre-configured infrastructures. When processing huge data volumes or computationally complex tasks, dynamically allocated cloud resources are needed for parallelizing distributed computing tasks.
To tackle those challenges, we propose a VRE solution that can be embedded into Jupyter as an extension that enables exporting individual cells as docker containers that can be composed as workflows. Our solution is composed of the following components:
Containerizer: It tracks the user's interactions with the notebook and in real-time updates the code and metadata based on the cell's modified content. When a user wishes to publish their cell, the containerizer builds a docker image off-premise with its metadata using our infrastructure automator called Software Defined Infrastructure Automator (SDIA).
Experiment manager: This component is responsible for loading and editing the cells'metadata catalog as well as for the submission and monitoring of workflows on the provisioned infrastructure.
Software-Defined Infrastructure Automator (SDIA). SDIA automates the planning, provisioning, monitoring, and adaptation of applications and their infrastructure on multi-cloud provider cloud offerings.

About the speakers:

Spiros Koulouzis is a researcher at the University of Amsterdam. His research interests include scientific workflows, as well as distributed and parallel systems. Spiros has a Ph.D. in computer science from the University of Amsterdam.

Zhiming Zhao is currently a senior researcher in the group of System and Network Engineering (SNE) at University of Amsterdam (UvA). He obtained his bachelor and master degrees in Computer Science from Nanjing Normal University (NJNU) and East China Normal University (ECNU) in 1993 and 1996 in China respectively. He obtained his Ph.D. in Computer Science from University of Amsterdam (UvA) in 2004. He has strong research interest in advanced computing and network technologies, time critical and data intensive

systems, Cloud computing, scientific workflows and software agents. He coordinates research and development activities in the EU H2020 project SWITCH (Software Workbench for interactive time critical and highly self-adaptive cloud applications), and in the "Data for Science"theme in the EU H2020 environmental science cluster project ENVRIPlus. He also leads the research tasks of research sustainability in the EU H2020 VRE4EIC project, and of semantic linking in the EU FP7 ENVRI project.

By submitting my abstract, I agree that my personal data is being stored in accordance to conference Privacy Policy

Most suitable track

Delivering services and solutions

Primary authors: KOULOUZIS, Spiros (University of Amsterdam); ZHAO, Zhiming (EGI.eu); Mr BIANCHI, Riccardo (Multiscale Networked Systems, University of Amsterdam 2LifeWatch ERIC, vLab&Innovation Center); Dr SIAMAK , Farshidi (Multiscale Networked Systems, University of Amsterdam); Ms RUYUE , Xin (Multiscale Networked Systems, University of Amsterdam); Ms WANG, Yuandou (Multiscale Networked Systems, University of Amsterdam); Mr NA , Li (Multiscale Networked Systems, University of Amsterdam); SHI, Yifang (2LifeWatch ERIC, vLab&Innovation Center 3Institute for Biodiversity and Ecosystem Dynamics (IBED)); TIMMERMANS, Joris (LifeWatch ERIC, vLab&Innovation Center 3Institute for Biodiversity and Ecosystem Dynamics (IBED)); KISSLING, W. Daniel (LifeWatch ERIC, vLab&Innovation Center 3Institute for Biodiversity and Ecosystem Dynamics (IBED))

Presenters: KOULOUZIS, Spiros (University of Amsterdam); ZHAO, Zhiming (EGI.eu)

Session Classification: Demonstration