Contribution ID: **112** Type: **Presentation**

Just-in-time data ingestion for distributed metagenomics assembly in multiple clouds with OneData

Wednesday, January 29, 2020 4:20 PM (20 minutes)

- Accessing large amounts of data in the cloud poses several problems:
  - Many bioinformatics applications require POSIX access, which does not scale well. Re-writing the application is not always an option.
  - Data sitting in the cloud costs money, whether it's being used or not.
- An ideal solution in many cases would be to provide federated data access to data stored on-premises, with caching in the cloud to reduce latency and use of network bandwidth.
- We have been using OneData to provide federated, secure, scalable access from the cloud to our on-premises storage. OneData offers tuneable caching, block-level access, and many other features that make it very attractive for our use-cases
- I present the tests we have done with OneData, including performance measurements from our first production use-case, distributed assembly of marine metagenomes in multiple clouds. We discuss our plans for rolling it out as a production service in the coming year.

Primary author: TIWARI, C.D. (EMBL-EBI)

Presenter: TIWARI, C.D. (EMBL-EBI)

Session Classification: Fabric and platforms for Global Science

Track Classification: User Voice: Novel Applications, Data Science Environments & Open Data