Lessons Learned from Petabyte Scale Biomedical Data Commons and Clouds

Robert Grossman
University of Chicago
Open Cloud Consortium

May 19, 2015
XLDB 2015
Data commons co-locate data, storage and computing infrastructure, and commonly used tools for analyzing and sharing data to create a resource for the research community.

The image above is the web portal for the NCI Genomic Data Commons. The University of Chicago is collaborating with the National Cancer Institute (NCI) to develop the NCI Genomic Data Commons (GDC). The GDC is being developed and operated with NCI funding through a subcontract from Leidos Biomedical Research, Inc. at the Frederick National Laboratory for Cancer Research.
Future State: virtual comprehensive cancer center.

Data commons containing genomic and clinical data.

This is a visualization of the time required to download data, process it, and upload it into the GDC (horizontal axis) vs the size of the data (vertical axis, ranging from GB ($10^9$) to TB ($10^{12}$)).
Object-based storage with S3 compatible interface

Scalable lightweight workflow

Community data products, including data harmonization

Database services

Data submission portal

APIs for data submission

APIs for data access

Devops supporting virtualized environments & containers (OpenStack VMs, Docker containers, scheduling)

Long running middleware services (Digital ID services, metadata services, high perf. transport)

Data access portal

Co-located “pay for compute”
At What Scale?

• Data centers are sometimes divided into “pods,” which can be built out as needed.
• A reasonable scale for a data commons is one of these pods (“cyberpod”).
• Let’s use the term “datapod” for the data & analytic cyber infrastructure that scales to a cyberpod.
• Think of as the scale out of a database to a cyberpod.
We are developing an open source software stack for data commons that scales to cyberpods.
Questions?

For more information:
rgrossman.com
@bobgrossman
opensciencedatacloud.org
cdis.uchicago.edu