Cloud Dataverse

Dataverse: Leonid Andreev, Gustavo Durand, Pete Meyer

MOC: Jeremy Freudberg, Anuj Thakur, Ata Turk

Dataverse team and MOC team

Mercè Crosas; Orran Krieger; Piyanai Saowarattitada
Massachusetts Open Cloud (MOC) & BDaaS

MOC: shared public cloud, multi-landlord model
- MIT, Harvard, NU, BU, Umass
- Lenovo, Intel, Brocade, TwoSigma, Cisco, USAF, Mass

BDaaS: spins up on-demand Big Data environment
- Exploit: Dataverse, Sahara, Ceph; prefetches data into local cluster
The Dataverse Project

• **Dataverse** is an open-source software platform for building data repositories
• It provides an incentive to share data
• Gives **credit** through data citation and control over data access
• Builds a **community** to:
  – define new standards and best practices
  – foster new research in data sharing
Dataverse around the world

Installed in 20 repositories worldwide;
Hosting dataverses from > 500 institutions.
Harvard Dataverse Repository

- 63,000 datasets with 360,000 files; 12 new datasets published per day
- 1.9 Million downloads; 1,500 downloads per day
- 13,000 registered users

Dataverse Community

- 300 members in the community list
- 35 GitHub core contributors
- Annual Community meeting with 200 participants
- semi-monthly community calls

http://dataverse.harvard.edu

http://dataverse.org
20 Installations
1,888 Dataverses
64,490 Datasets
1,969,761 Downloads

dataverse.org
Local copy of Datasets

Finland

Iceland

Dataverse installation

WHAT IF?
Local copy of Datasets

Iceland

Dataverse installation

Finland

Remember the MOC?
What need to happen?
Introducing... Cloud Dataverse

Local copy of Datasets
https://github.com/IQSS/dataverse/pull/3239/commits

Dataverse installation
https://github.com/IQSS/dataverse

OpenStack Sahara:
- Hadoop
- Spark
- Pig
- etc…
Summary

• Reduce computational time with **dataset locality**
• Leverage existing **flexibility** of **OpenStack design**; Swift, Sahara, OpenStack Core services
• Could use simpler orchestration around Sahara; too many options could get confusing very fast …
• **Simple** concept with tremendous usage and we are making it happen!

[dataverse.org] [Info.massopencloud.org]