compute canada

A Service Orientation to High Performance & Distributed Computing

September 22, 2015

Canada's ARC Platform Today & Tomorrow

Consolidation & Renewal

Distributed Across Canada

50 Systems

27 Data Centres

200,000 cores, 2 Pflops, 20 PB

200 Experts

~10,000 Users

Consolidation & Concentration by 2017

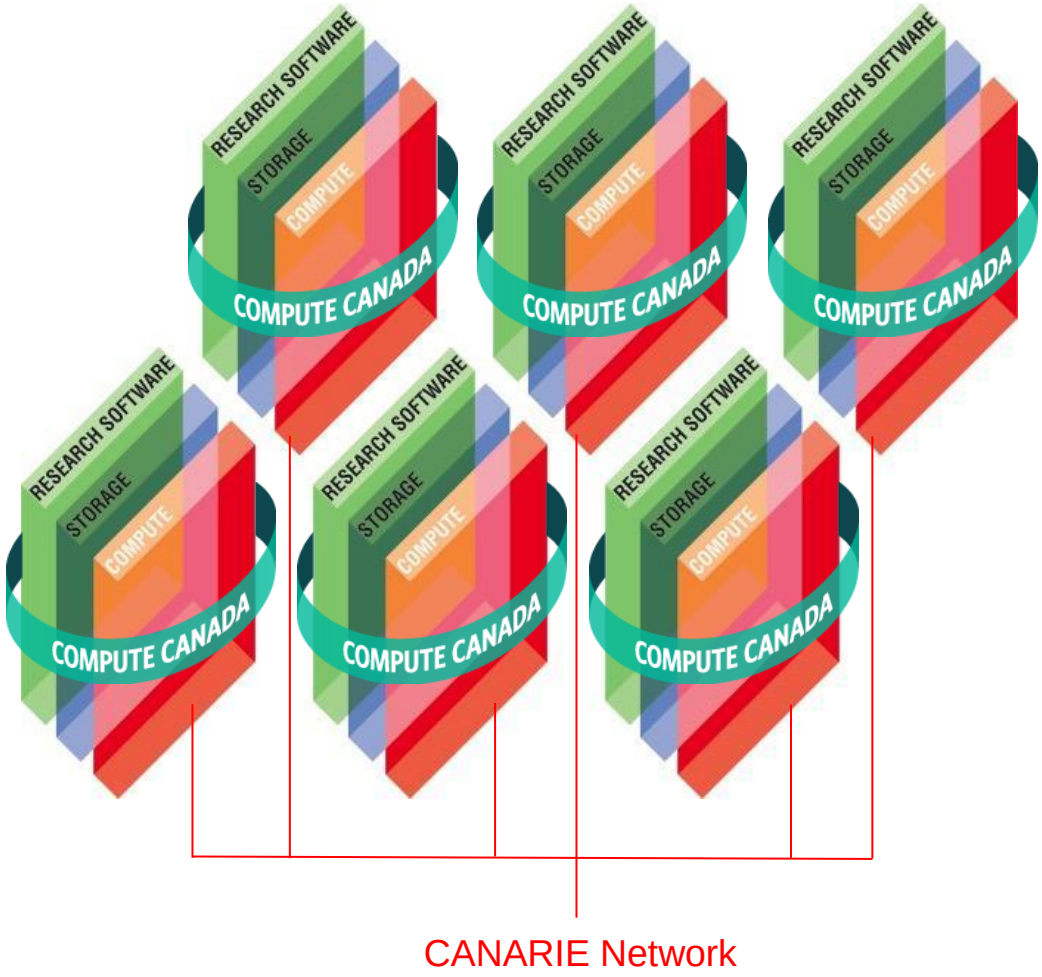
18 Systems

13 Data Centres

240,000 cores, 13.4 Pflops, 80 PB

200 Experts

~20,000 Users



compute canada | calcul canada

A Few Achievements

200 experts accelerating results for more than 8,500 researchers including close to 3,000 faculty members

More than 3,700 peer-reviewed publications, 40 patents, 23 inventions, and 7 companies*

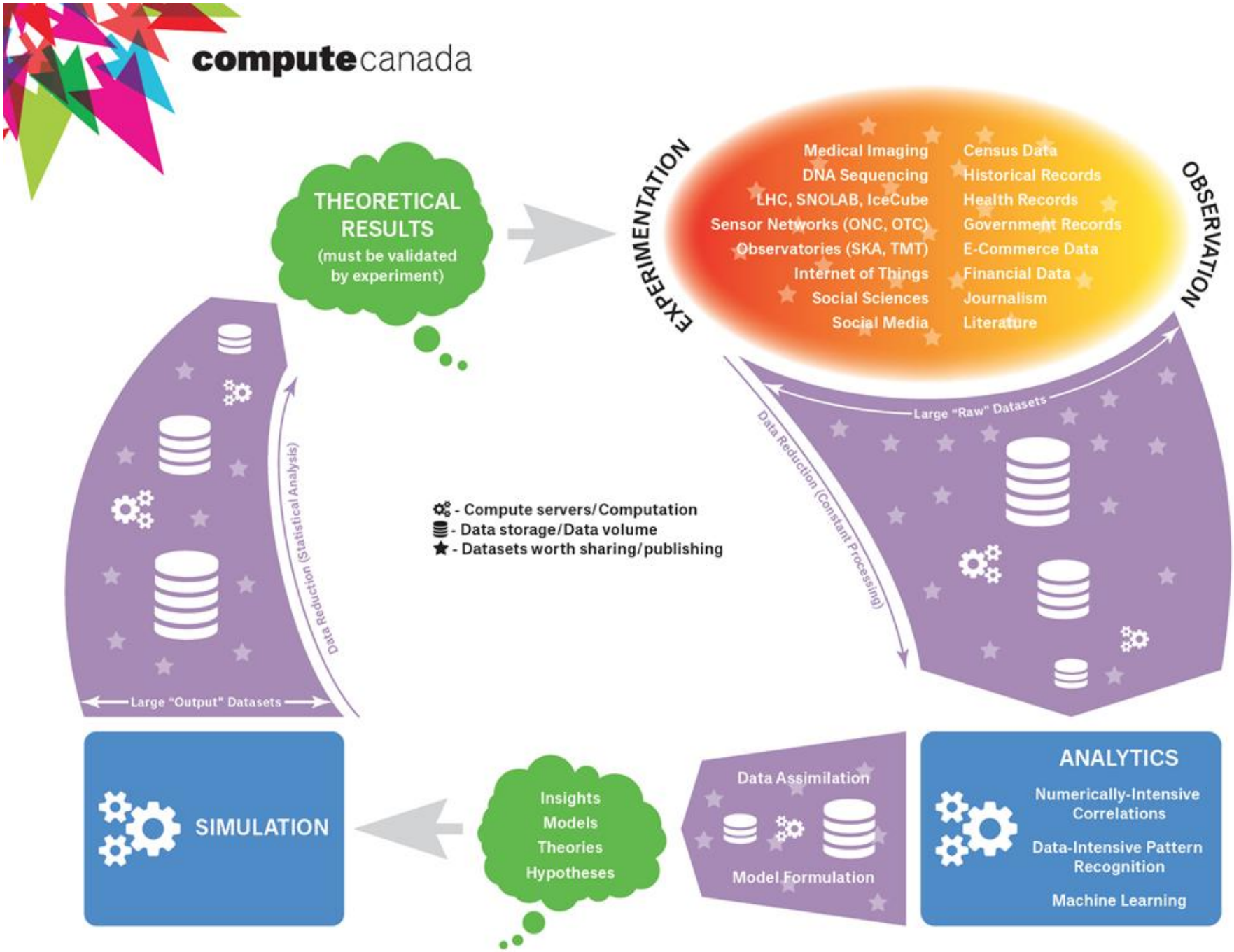
*since 2012

Storing and managing over 15 petabytes of active research data

Delivering 54,000 hours of training to more than 11,000 researchers

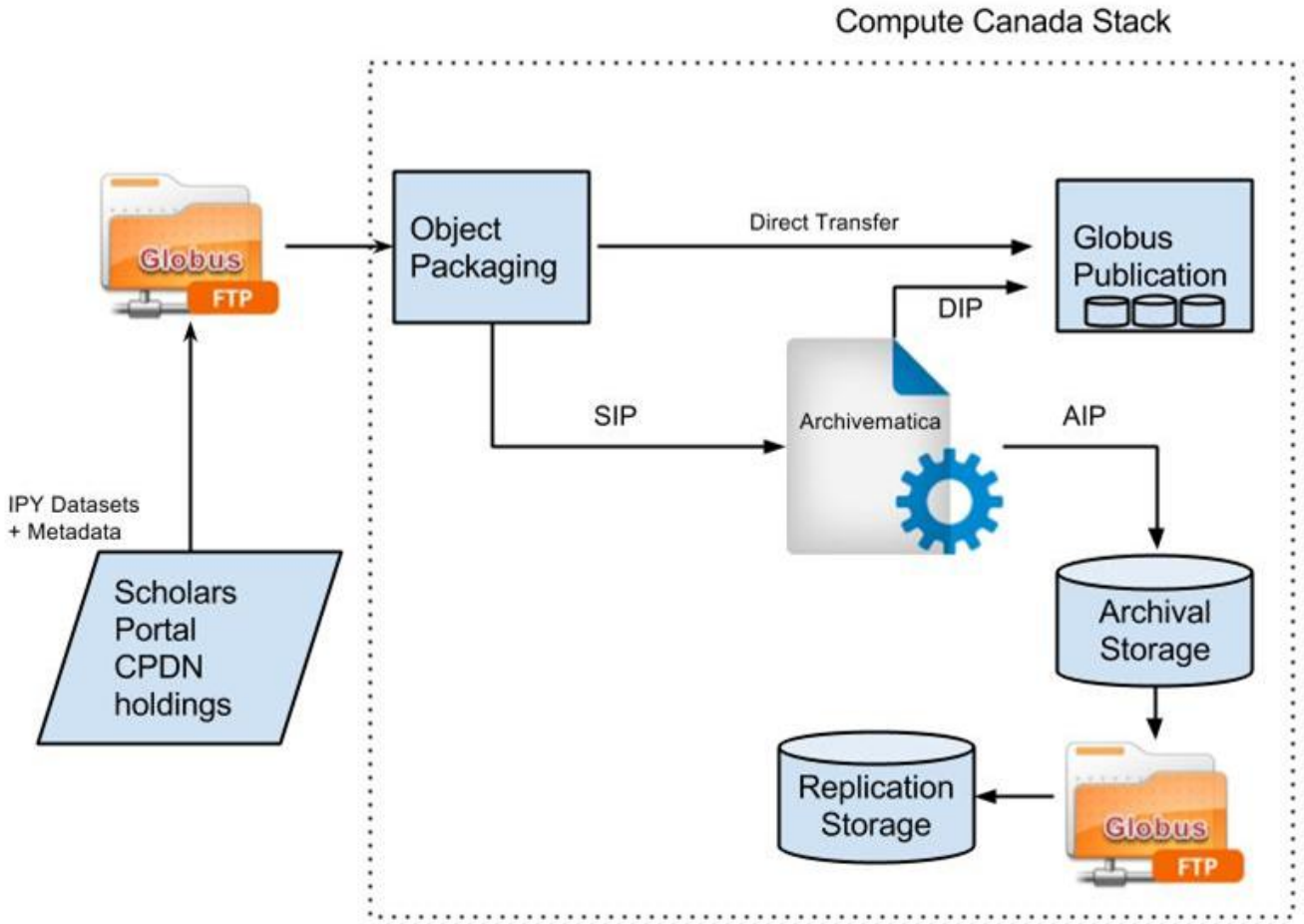
Serving users at more than 70 Canadian universities

compute canada | calcul canada


computecanada
THEORETICAL RESULTS
(must be validated by experiment)
EXPERIMENTATION
OBSERVATION
Medical Imaging
DNA Sequencing
LHC, SNOLAB, IceCube
Sensor Networks (ONC, OTC)
Observatories (SKA, TMT)
Internet of Things
Social Sciences
Social Media
Census Data
Historical Records
Health Records
Government Records
E-Commerce Data
Financial Data
Journalism
Literature
Large "Raw" Datasets
Data Reduction (Constant Processing)
Data Reduction (Statistical Analysis)
Large "Output" Datasets
- Compute servers/Computation
- Data storage/Data volume
- Datasets worth sharing/publishing
SIMULATION
Insights
Models
Theories
Hypotheses
Data Assimilation
Model Formulation
ANALYTICS
Numerically-Intensive Correlations
Data-Intensive Pattern Recognition
Machine Learning

CC-RDC-PORTAGE Federated RDM Pilot: Canadian Polar Data Network

CPDN is the domain repository for the Canadian International Polar Year (IPY) and Northern research data



Conclusions

- Project demonstrated relatively complete path for migration and preservation of existing data collections
  - **Globus File Transfer** service to move files from existing IPY collection to final repository location
  - **Archivematica** service to generate dissemination (DIPs) and preservation (AIPs) products
  - Automated ingestion of normalized data+metadata into **Globus Publishing** for data access and discovery
  - Automated replication of preservation products (AIPs) across Compute Canada storage sites using **Globus File Transfer**

- The scalability of Globus Transfer and Globus Publishing is promising
- Still considerable work to be done to “scale up” the pipeline
- These results may serve as a foundation for RDM solution(s) which will serve a very wide range of use-cases across many disciplines

compute canada | calcul canada

Feature Wishlist

HTTP file access

- anonymous (no Globus account) download

API for Globus Publishing

- customizability of user interface is a driver

Self-service form configurator

Ingestion from existing collections with existing metadata

- bypass manual data entry

Ingestion in-place without transferring data

Expansion of download features to get multiple datasets at a time

Data Management Framework

a PID system (like unique DOI)
ID system for actors (like ORCID)
Registry system for Trusted Repositories
Metadata system
Schema Registry System
Registry System for Semantic Categories, Vocabularies, etc.
Registry System for Data Types
Registry System for Practical policies
Prefabricated PP Modules

Distributed Authentication System
Authorization Record Registry System
Systems to aggregate and Harvest metadata
Workflow Engine and Environment
Conversion Tool Registry
Analytics Component Registry
Repository API
Repository System
Certification and Trusted Repositories
Training Modules

compute canada

Merci Beaucoup!

September 22, 2015