RDM: Active Research Data Lifecycle

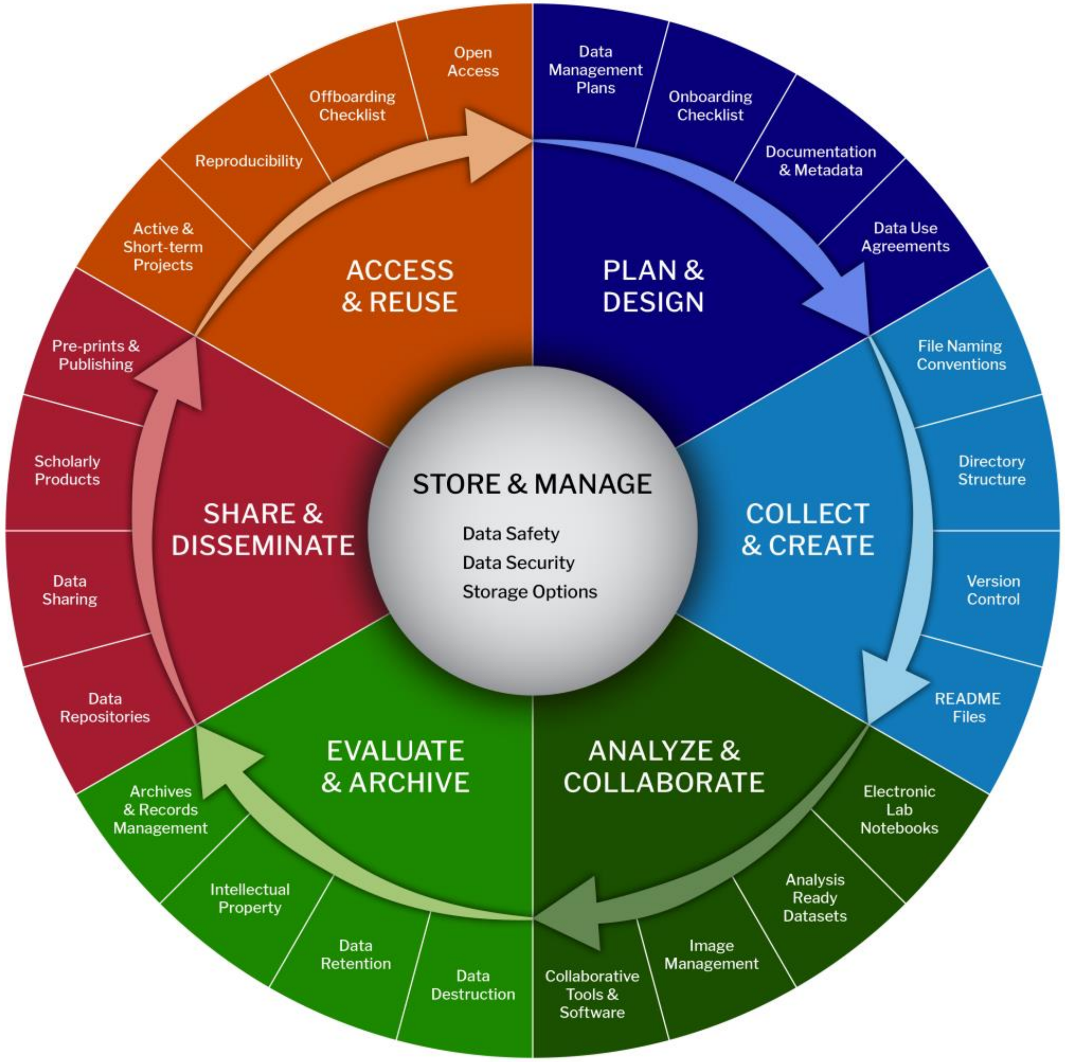
Abdel Yousif, PhD,
Director, Research Computing Services (RCS), IT
abdel.yousif@ucalgary.ca

September 15, 2022

UNIVERSITY OF CALGARY

Research Data Lifecycle

- Active Research Data phases
  - Collect and Create
  - Analyze and Collaborate
  - Access & Reuse: Reproducibility
- Active research data includes
  - Machine & human generated data
  - Metadata: involves ETL and sometime ML
  - Synthetic data (artificially generated): AI
  - IoT and sensor (passive) data
- Data security levels
  - Level 1 & 2: public / internal
  - Level 3 & 4: Private / Sensitive
  - https://www.ucalgary.ca/legal-services/sites/default/files/teams/1/Standards-Legal-Information-Security-Classification-Standard.pdf



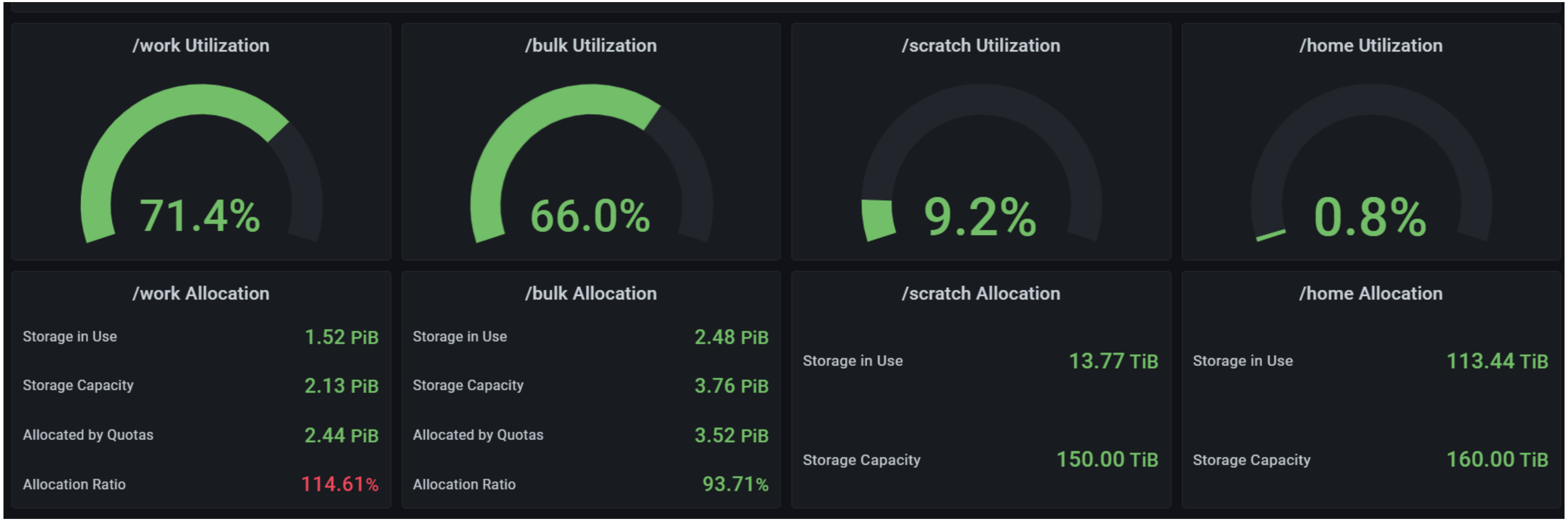
Biomedical Data Lifecycle

UNIVERSITY OF CALGARY

Collect & Create

- Create an organizational workflow for the data, detailing any methods, procedures, and/or protocols used for data collection
  - ✓ Important step to help define you research data destination in IT
- Seek ethics approval for sensitive data based projects
- Pick the right storage system for your data
  - ✓ Secure Compute Data Storage (SCDS) for sensitive data ($$)
  - ✓ Computationally complex workflow: storage for data analysis using HPC
    - ✓ Examples: Bioinformatics and Computational Fluid Dynamics (CFD) data
    - ✓ $$$
  - ✓ Virtual Machines data back up ($$)
  - ✓ AcademicFS: two geo-replicated copies of level 1 & 2 research data ($$)

A Snap Shot of High Speed Storage (HPC System)



/work Utilization

71.4%

/bulk Utilization

66.0%

/scratch Utilization

9.2%

/home Utilization

0.8%

/work Allocation

Storage in Use	1.52 PiB
Storage Capacity	2.13 PiB
Allocated by Quotas	2.44 PiB
Allocation Ratio	114.61%

/bulk Allocation

Storage in Use	2.48 PiB
Storage Capacity	3.76 PiB
Allocated by Quotas	3.52 PiB
Allocation Ratio	93.71%

/scratch Allocation

Storage in Use	13.77 TiB
Storage Capacity	150.00 TiB

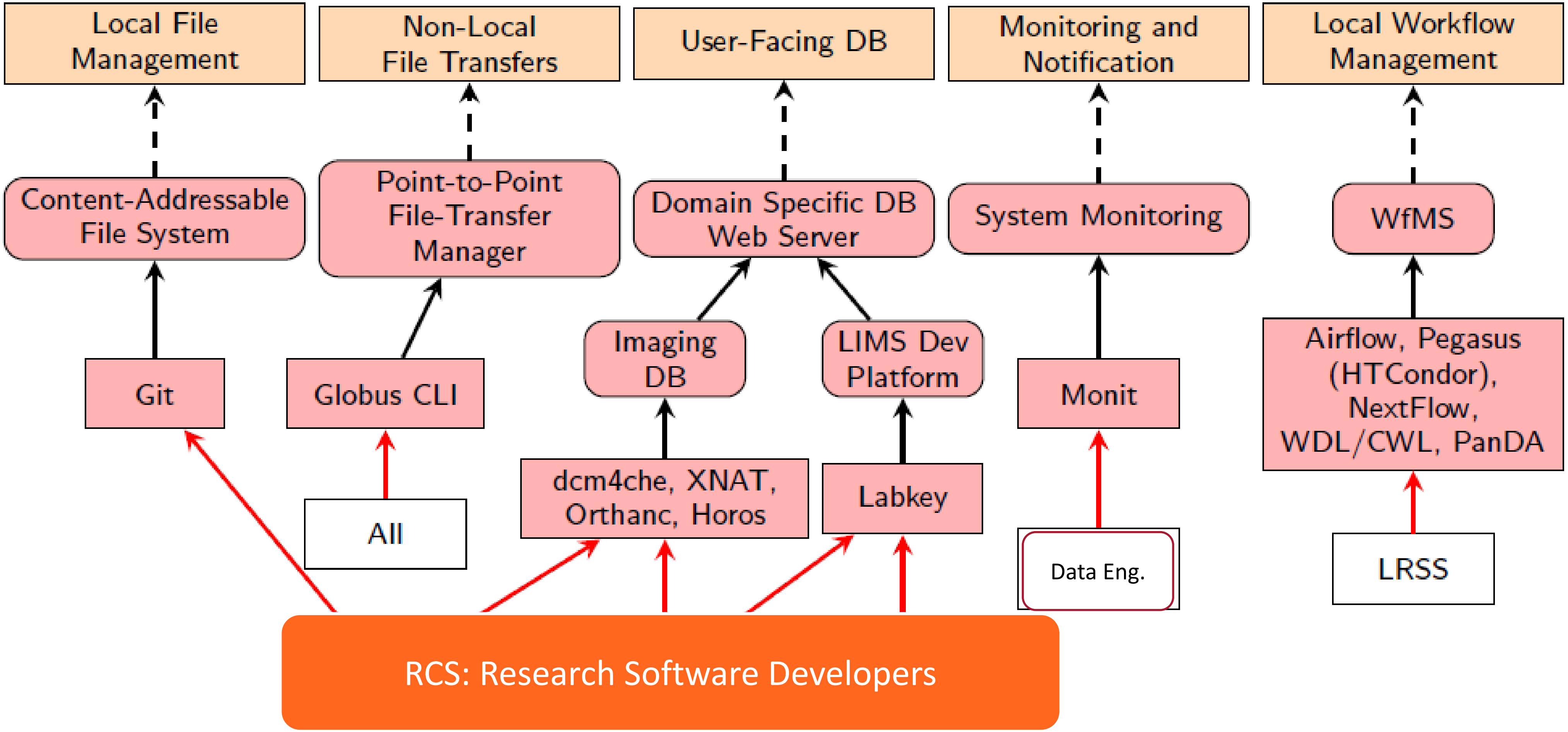
/home Allocation

Storage in Use	113.44 TiB
Storage Capacity	160.00 TiB

- On average, ARC storage investment is $0.4M annually
- RDM requires a new investment in all storage capacities due to warranty expiry
  - ❑ 2023/2024: $0.5M
  - ❑ 2024/2025: $1.8M
- A combination of VPR/VPS investment + small cost recovery percentage (< 20%) from researchers

UNIVERSITY OF CALGARY

Analyze & Collaborate: A Data Pipeline Point of View*



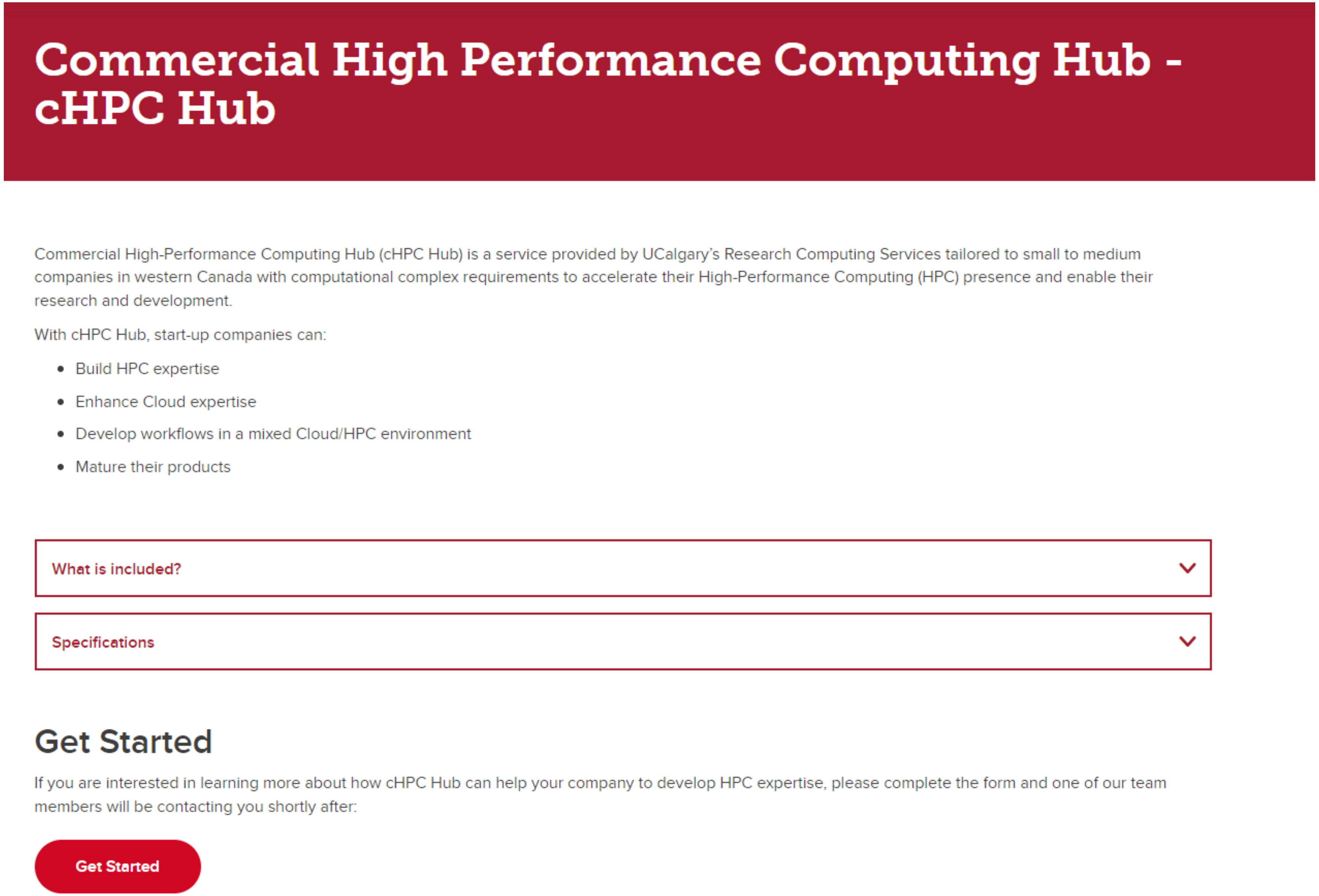
Local File Management
Non-Local File Transfers
User-Facing DB
Monitoring and Notification
Local Workflow Management
Content-Addressable File System
Point-to-Point File-Transfer Manager
Domain Specific DB Web Server
System Monitoring
WfMS
Git
Globus CLI
Imaging DB
LIMS Dev Platform
Monit
Airflow, Pegasus (HTCondor), NextFlow, WDL/CWL, PanDA
All
dcm4che, XNAT, Orthanc, Horos
Labkey
Data Eng.
LRSS
RCS: Research Software Developers

* This is work in progress based on a CANARIE grant to RCS

Access & Reuse: Reproducibility

- Mostly done manually through data lineage of README files
  - Gets extremely complex with complex data pipelines
  - Researchers lose opportunities to expand research outcomes
- Reproducibility must be addressed through the intersection of RDM, Research Software (RS) and Advanced Research Computing (ARC)
  - How can a researcher reproduce results without mimicking the same ARC infrastructure and version of software used for processing?
  - Researchers will still face trade offs between available storage for large data sets on different “result” points on a data pipeline
- RCS is pursuing an automated approach to research reproducibility
  - This is RCS focus over the next two years

The Lifecycle Beyond Research: Commercialization



Commercial High Performance Computing Hub - cHPC Hub

Commercial High-Performance Computing Hub (cHPC Hub) is a service provided by UCalgary’s Research Computing Services tailored to small to medium companies in western Canada with computational complex requirements to accelerate their High-Performance Computing (HPC) presence and enable their research and development.

With cHPC Hub, start-up companies can:

- Build HPC expertise
- Enhance Cloud expertise
- Develop workflows in a mixed Cloud/HPC environment
- Mature their products

What is included?

Specifications

Get Started

If you are interested in learning more about how cHPC Hub can help your company to develop HPC expertise, please complete the form and one of our team members will be contacting you shortly after:

Get Started

Research Computing Services (RCS) Resources

- Research computing website: https://it.ucalgary.ca/research-computing-services
- For storage requests: https://it.ucalgary.ca/research-computing-services/our-resources/storage-solutions
- For HPC/Self Managed VM requests: email support@hpc.ucalgary.ca
- For RCS commercial services: https://it.ucalgary.ca/research-computing-services/CHPC
- For research software support: support@hpc.ucalgary.ca
- For data pipelines design and deployment services: ian.percel@ucalgary.ca

Thank you for your time!

Contact: abdel.yousif@ucalgary.ca

UNIVERSITY OF CALGARY