Breakout Session 3: Track B

ATLAS-D2K - Exploring Cloud Optimization

Dr. Hongsuda Tangmunarunkit
Supervising Computer Scientist, University of Southern California

ATLAS-D2K - Exploring Cloud Optimization

Exploring AWS native solutions to improve data portal efficiency – fault tolerance, scalability, cost

Hongsuda Tangmunarunkit, Laura Pearlman
Carl Kesselman, Todd Valerius

ATLAS-D2K Center

GenitoUrinary Development Molecular Anatomy Project
Create a high resolution molecular anatomy of gene expression for the developing organs of the GU tract

(Re)Building a Kidney
Create approaches for the isolation, expansion, and differentiation of appropriate kidney cell types and their integration into complex structures that replicate human kidney function.

The Analysis, Technology, Leadership, Administration, and Science - Data to Knowledge (ATLAS-D2K) Center

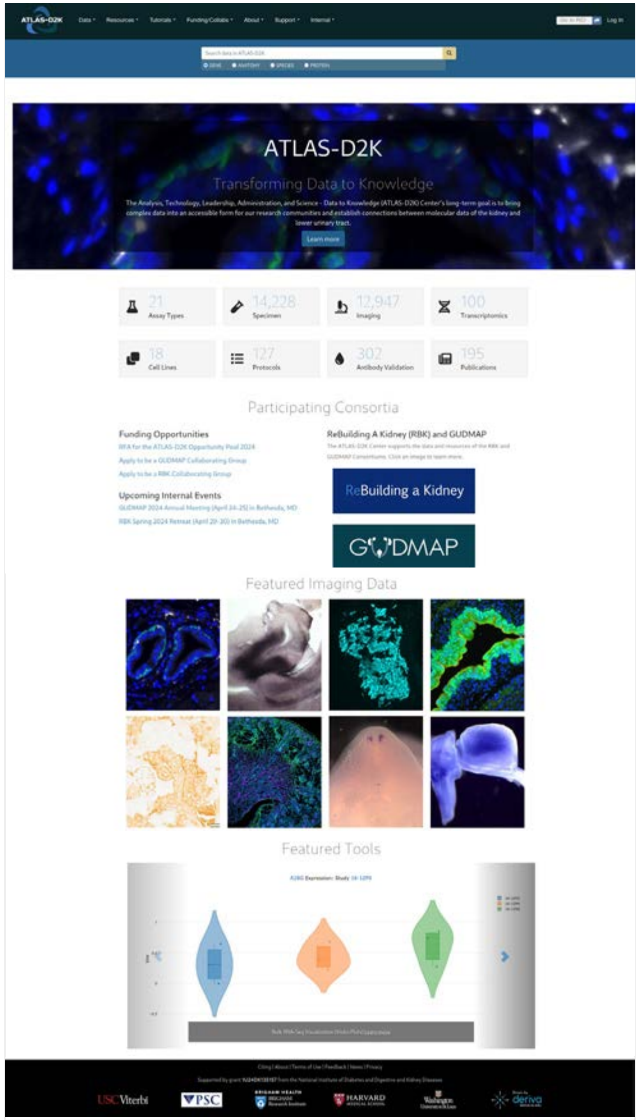
Goal: bring complex data into an accessible form for our research communities and establish connections between molecular data of the kidney and lower urinary tract.

Role: partner with consortium members to create open high-quality reusable data assets and tools

Data Asset Role:
- A research asset to consortium and community
  - Whole is greater than the sum of its parts
- Transparency and reproducibility of scientific data
  - Follow *F.A.I.R.* data principles: Findable, Accessible, Interoperable, Reusable
  - Data ”modeled”, curated, and published openly

Data *Usability* Role:
- Put metadata to work
- Make sharing with attribution easy
- Make tools that enable direct use of the data
- Visualization tools for data interaction
- Re-analysis and QC of existing data

www.atlas-d2k.org



ATLAS-D2K Infrastructure

ATLAS-D2K Data Portal
www.atlas-d2k.org
(www.gudmap.org, www.rebuildingakidney.org)

Users:
- Consortium members
- Broader communities
- Public

Consortia Members
- GUDMAP
- RBK

Collaborating Consortia
- KPMP
- HuBMAP
- FaceBase

Public

External Entities
- GEO
- GitHub
- DataCite (DOI)
- dbGAP

Ingest/Export

Data Management System (DMS)
unified and integrated repository deriva

Website
web content

Data (raw, processed)
- 2D/3D microscopy & imaging: immunohistochemistry, histology, *in situ* hybridization, micro-CT, nano-CT
- Omics: transcriptomics, epigenomics, metabolomics, proteomics
- Annotated gene expression
- Imaging mass cytometry
- Transgenic cell-lines & mouse strains

Resources
- Reagents: antibodies, antibody validations, primers
- Protocols
- Publications
- Chemical compounds
- Instructional videos
- Videos
- Multi-modal data collections
- Ontologies

Tools
- mRNA-Seq analysis
- Single-cell visualization
- Image display
- Image annotation

Dissemination & Outreach
- Center information
- Consortium information
- Opportunities (e.g. pilot projects)
- Tutorials, documentation, training
- Announcements, news, blog posts
- Shortcuts to highlighted data

Data types: 21 | # Species: 3 (mouse, human, dog) | # Labs: 35
Specimen: 14K+ | # Cell-lines: 18 | # Users (12 mo): 14K
Imaging files: 42K+ | # Genes with data: 43K+ | # Files: 366K+ (13.4TB)
Transcriptomics studies: 100 | # Anatomy with data: 739 | # Download (12 mo): 28K+

Key Capabilities

Data Discovery & Access
- Online search & browse tools
- Direct access through persistent identifiers (RecordID and DOI)
- Data export and download
- APIs (ReST, Python, R, javascript)

Data Processing
- Images and videos (visualization and annotations)
- Sequencing bioinformatics & visualization (mRNA-Seq, scRNA-Seq)
- DOI management

FAIR Data Catalog & Store
- Metadata design (i.e. data model)
- Persistent and citable identifiers
- Open metadata & data access
- Ontologies and controlled terms, metadata and file standards
- Versioned data objects and point-in-time metadata snapshots

Data Curation & Publication
- Curation Process
- Online data curation tools
- Client and CLI tools
- Online image annotation tool
- Collections
- Data citation (DOI)

Data Visualization
- Interactive 2D image & annotation viewer
- Interactive 3D image viewer with surface and ortho-slice views
- Interactive single-cell and mRNA-Seq expression visualization
- Scored expression & array visualization

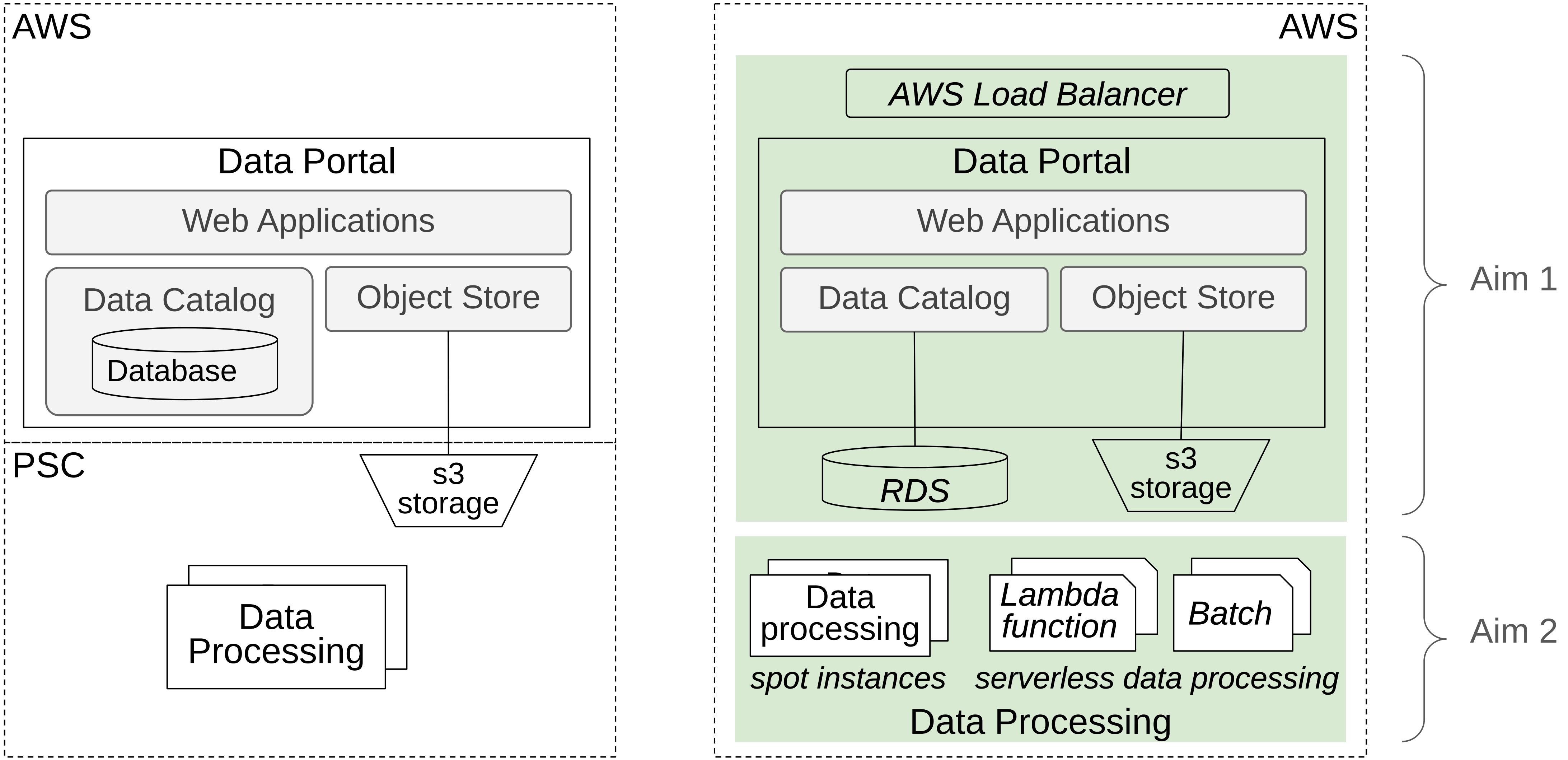
Specific Aims

1. Evaluate AWS native solutions to increase ATLAS-D2K Data Portal fault-tolerance, minimize system downtime, and lower operation/maintenance efforts
    1. Exploring AWS Relational Database Service (RDS) for reliable (highly available) managed database service.
    2. Evaluate AWS Elastic Load Balancing (ELB) to reduce system downtime and hence improve application availability with respect to system upgrade and server failure.

1. Evaluate AWS cost-effective approaches for running data processing tasks
    1. Evaluate spot instances for running data processing tasks.
    2. Evaluate AWS serverless data processing architecture (e.g., Lambda functions, AWS batch) for different data processing classes.

Processing classes	Representative tasks	CPU	Memory	Disk space	Expected execution time
small	2D image processing: small files (<= 200MB, 98% images)	2 cores	10 GB	2 GB	1-15 mins
moderate	2D image processing: files 200 MB to 20 GB (maximum)	2 cores	32 GB	100 GB	15-80 mins
data intensive	3D Image processing (130 MB - 6.5 GB)	4 cores	16 GB	500 GB	5-20 minutes
memory intensive	2D image processing: high-resolution multi-channel images (100K x 86K avg dimension)	4 cores	36-512 GB	100 GB	1-3 hours
computing-intensive	mRNA-Seq analysis (10GB - 110 GB)	16 cores	32 GB	1 TB	3-4 hours

Design Architecture



A. Hybrid architecture (parent proposal)

B. Cloud optimization (supplemental)

Expected Outcomes

1. Evaluate AWS RDS and ELB (compared to a baseline system with Postgres on EC2 and no load-balancing)
    1. Higher fault-tolerance ⇧
    2. Lower system downtime ⇩
    3. Lower operation/maintenance efforts ⇩
    4. Higher Cost ⇧

1. Evaluate AWS spot instances and serverless architecture for running data processing tasks (compared to general-purpose EC2 systems)
    1. Lower cost ⇩
    2. Higher scalability ⇧
    3. Increase deployment & billing complexity ⇧

Acknowledgement

ATLAS-D2K

- DERIVA Team: Karl Czajkowski, Josh Chudy, Mike D’Arcy, Robert Schuler, Aref Shafaeibejestan, Serban Voinea, Cris Williams

Supported by grant **1U24DK135157** from the National Institute of Diabetes and Digestive and Kidney Diseases

USC Viterbi PSC BRIGHAM HEALTH BRIGHAM Research Institute HARVARD MEDICAL SCHOOL Washington University in St. Louis Driven by deriva deriva.isi.edu