Bringing visibility to food security data results:
harvests of PRAGMA and RDA

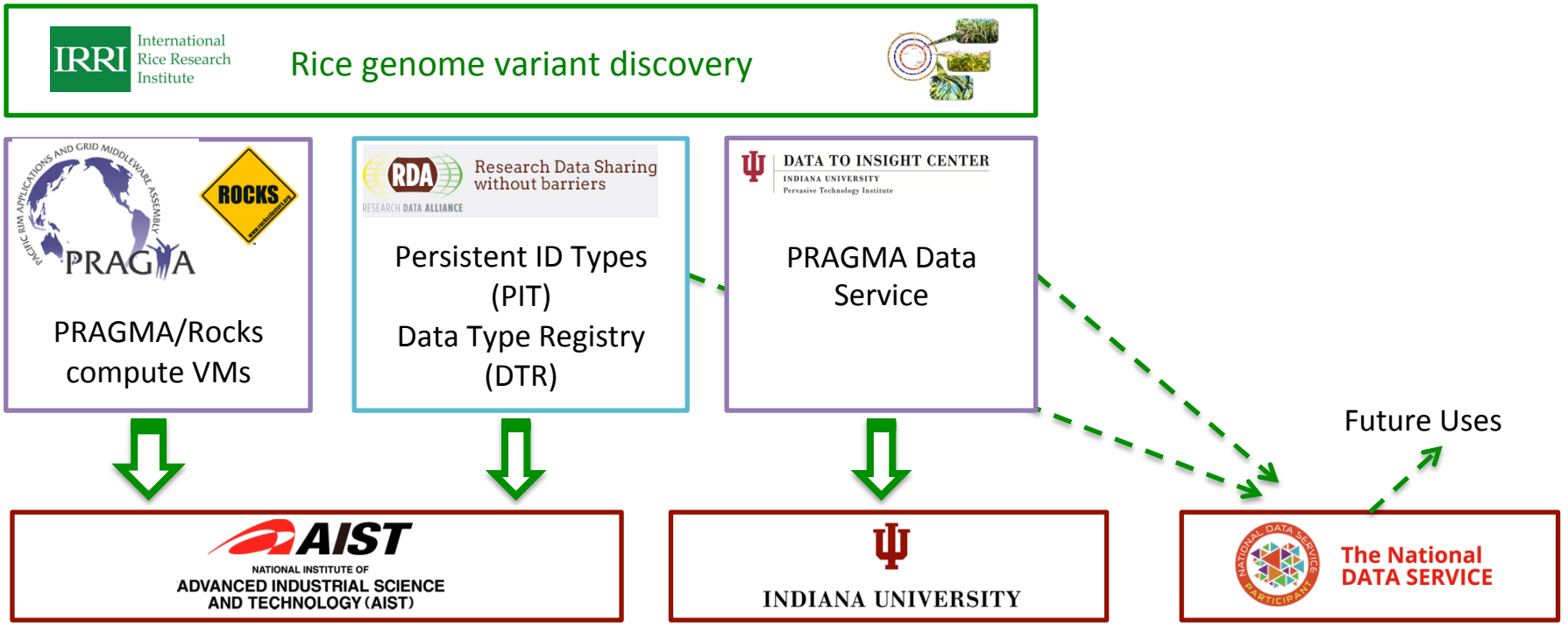
AIST-IU-IRRI

Presenter: Quan (Gabriel) Zhou
Venice Juanillas, Ramil Mauleon, Jason Haga, Beth Plale

9/8/16

Bringing visibility to food security data results: harvests of PRAGMA and RDA

Co-PIs: Beth Plale, Indiana University, USA; Jason Haga, AIST, Japan



Launch the use of two RDA products in Asia by utilizing the PRAGMA community and tools to work with a new rice genome group in the Philippines and implement software services at AIST (Japan) using the outputs of the PID Information Types and Data Type Registries Working Groups.

Goals:

- Seek an agreeable PID attribute type profile to harvest data objects from varied scientific domains for wider adoption;
- Implement RDA PIT and DTR recommendations to support data citation of rice genomes data objects;
- Developed Software will be installed additionally at the National Data Service to stimulate adoption in the US.

Research Motivation

- Use RDA PIT/DTR model to improve sharing and interoperability of scientific data objects by embedding minimum metadata in persistent data identifier
- Provide a framework with both repository and PID service to provide long-term access and findability to heterogeneous data objects across scientific boundaries
- Propose a methodology to automatically harvest data objects from scientific workflows and improve reproducibility of workflow execution

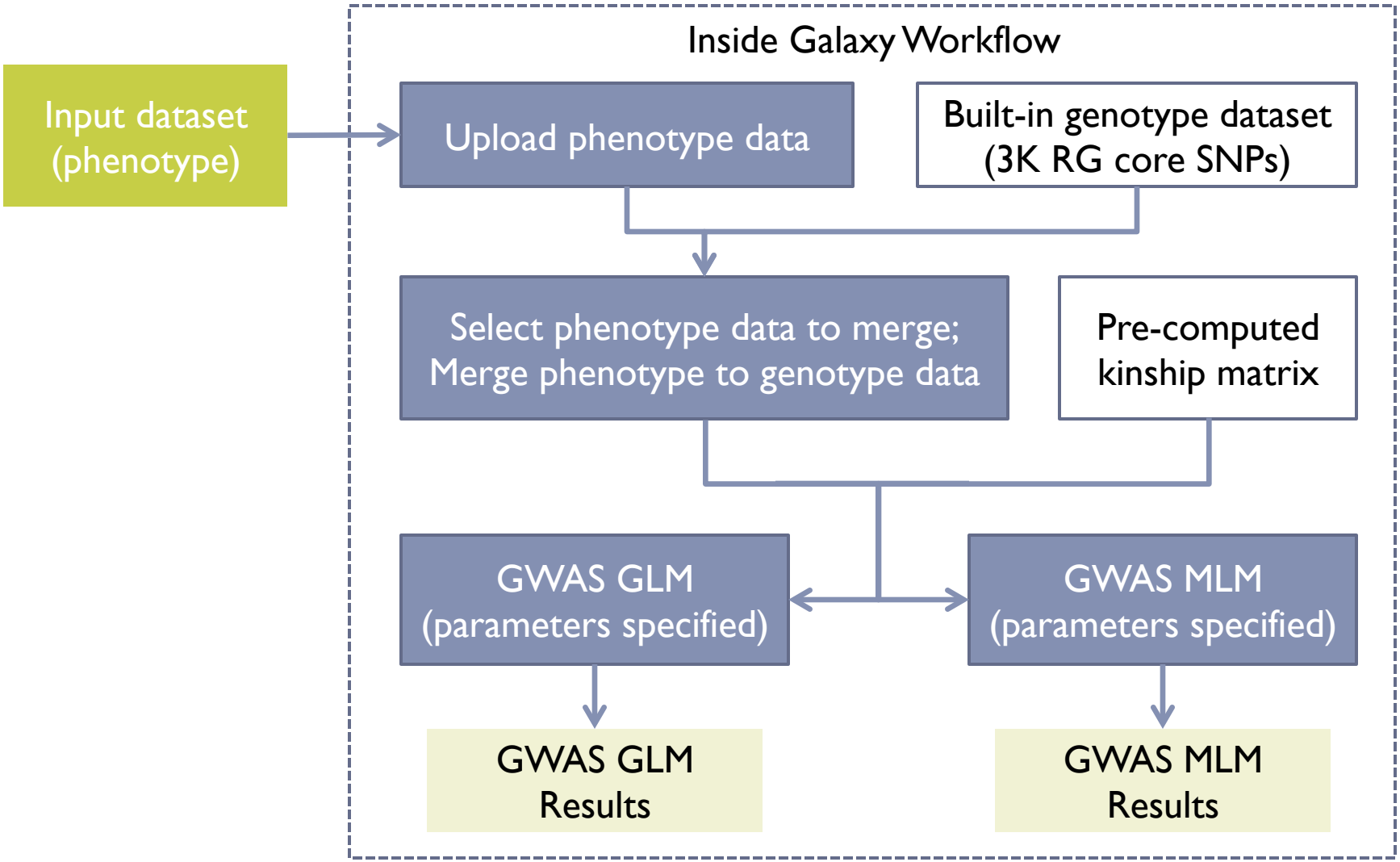
Purpose of IRRI TASSEL in Galaxy

- Enable collaborators to do their own GWAS analysis of their own phenotyping data for 3KRG using common analysis framework

- Enable IRRI to harvest phenotyping data and results of analysis with 3KRG from the collaborators to train stronger rice genomics model

- To collect relevant information from rice genomes workflows and help future users to validate the results by reproducing the workflow during their decision-making

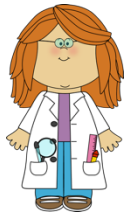
Typical use cases



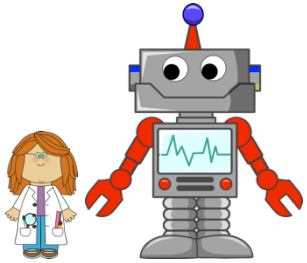
IRRI's Need from Data Domain

- Install TASSEL data handling and compute steps (blue boxes) as Galaxy Tools
- Define Galaxy workflows to execute each use case
  - Enable input of user-adjustable parameters of analysis
- Run the workflow selected, version the analysis done
  - Workflow id
  - Input file used
  - Parameters supplied to tool in each step
  - Output file
- Provide a means to share analysis results back to IRRI

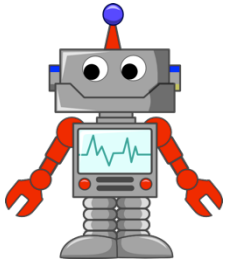
General Events Timeline of Demo



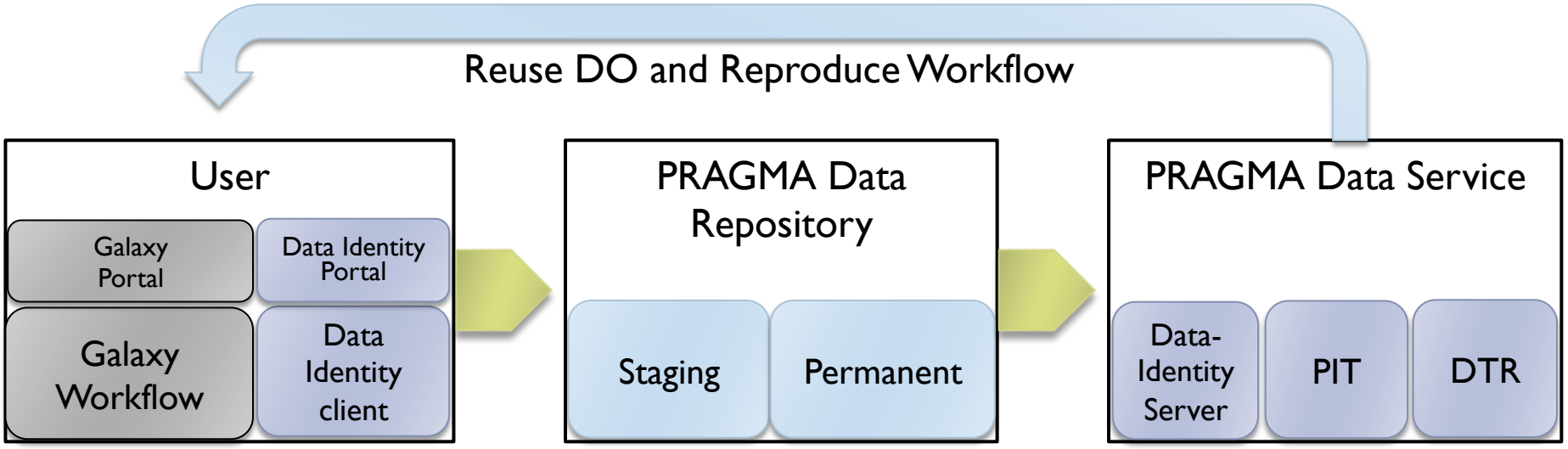
1. End-user



2. Repository Service



3. PID Service



User experimental DO

DO goes in repository database

DO gets assigned persistent identifier and landing page

Demo – Reproducing Rice Genomes Workflow

Deployment Diagram

PRAGMA Data Identity Service

Data Identity Server

PIT Service

DTR Service

PRAGMA Data Repository

Repository Rest API

MongoDB DataBase

AIST Data Identity Cluster, Tsukuba, Japan

Data Identity GUI

Galaxy Workflow Server

Tassel 5 Workflow

Data Identity Tassel Client

Rice Genomics Workflow VM SDSC, CA, US

Handle.net Service

Handle V8 Instance

PID Service Instance CNRI, Virginia, US

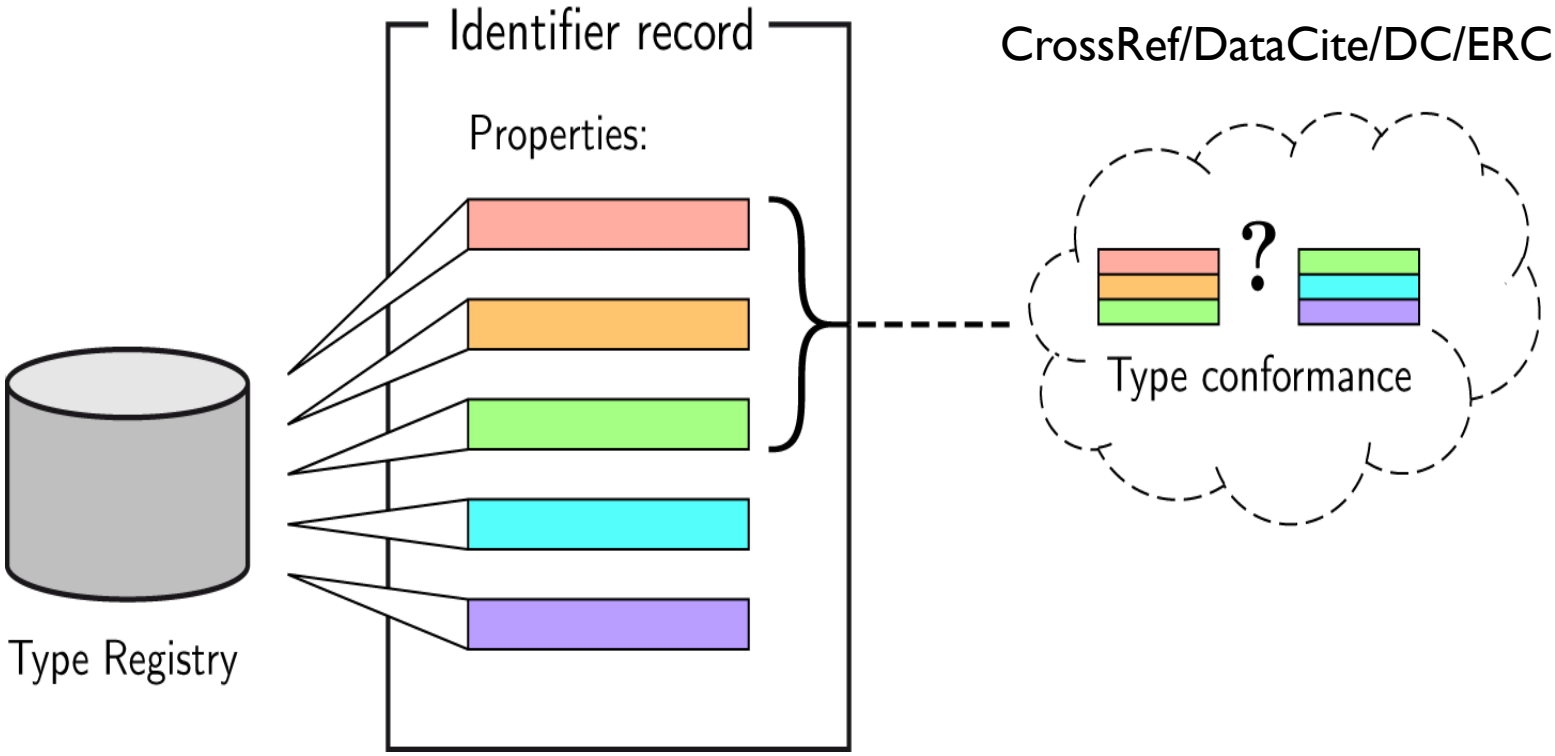
New architectural components

- Handle service : handle.net V8 instance
  - What: Assign and resolve handle PIDs for individual data objects
  - Where: Provided by CNRI server https://38.100.130.12 with prefix 11723
- PRAGMA Data Repository: Host scientific DOs across domains
  - What: Used to host and collect heterogeneous scientific DOs with associated metadata from varied disciplines
  - Where: Deployed at AIST data identity cluster with 3 MongoDB replications
- PRAGMA Data Identity Service
  - What: Service to interact with PRAGMA data repo and PIT/DTR service to collect scientific DOs and register for PID with minimum metadata
  - Where: Server deployed at AIST data identity cluster; Client and GUI hosted at rocks-53 Biolinux server at SDSC.
- RDA PIT/DTR Service : PRAGMAPIT-ext server V0.2 and CNRI Cordra Server V1.0.7
  - What: Used to host profile/type definition for PID metadata and data objects for interoperability and sharing across service providers
  - Where: Deployed at AIST data identity cluster

Component I RDA PIT/DTR outcomes

- Different service providers such as DOI, ARC, Handle can provide identifier metadata which can use different terminologies;
- PIT WG develops an API enables consensus on some essential types;
- In this model, every PID record consists of a number of properties. (A minimum metadata associated with PID) Every property bears a PID and its essential elements are a name, a range and a value. Only the PID and the value are stored in PID records, while the name and range are available from the registered property definition in the data type registry.

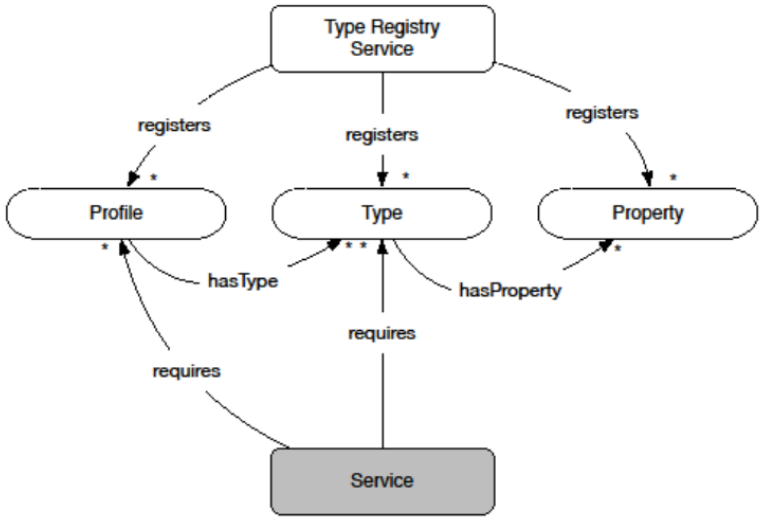
PID Information Type Model



PRAGMA PIT extension

We extend RDA PIT service V0.1.0 with improved compatibility and APIs to support query on Profile level.

- Features:
  - Compatible with latest Data Type Registry version (Cordra 1.0.7)
  - Support query on Profile level
  - Github Code Base: https://github.com/Data-to-Insight-Center/RDA-PRAGMA-Data-Service/tree/master/pragmapit-ext



size
checksum
version
part-of
has-parts
replica locations
...

Climate sciences

size
checksum
timestamps
version
predecessor
successor
...

Material sciences

. . .

Core profile

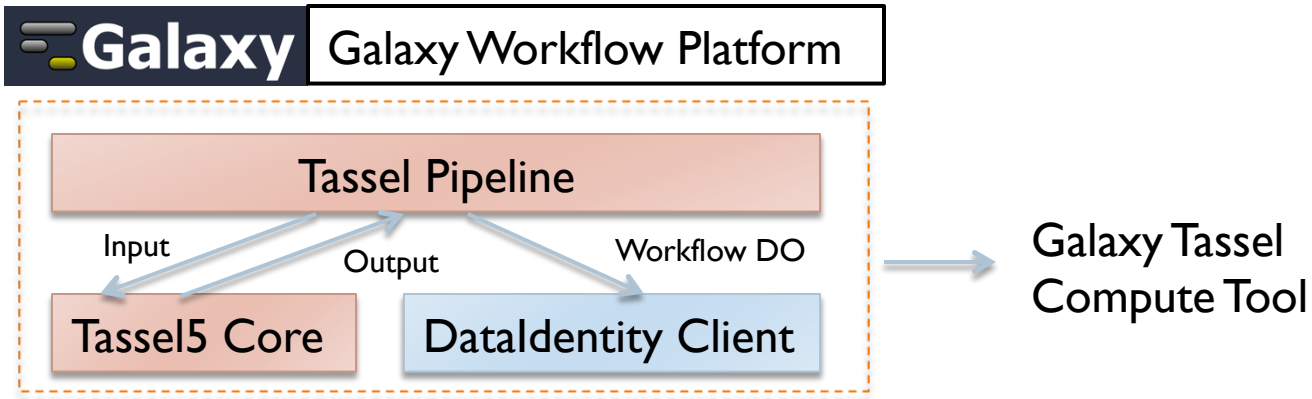
Component II - PRAGMA Data Identity Service

- Brings persistent IDs and registration of data objects generated by scientific analysis that is carried out from scientific experiments such as workflows. The data service leverages two recent recommendations from the Research Data Alliance (RDA, https://rd-alliance.org/): Persistent Identifier Information Type (PIT) and Data Type Registry (DTR).

- API resources
  - Create DO PID with PID metadata profile (PIT model is applied)
  - Resolve DO PID with metadata profile as human readable format
  - Get/Set resource links (landing page, metadata URL)
  - Get Data Type Definition using community profile PID (interaction with PIT service)
  - Get full inter-identifier links

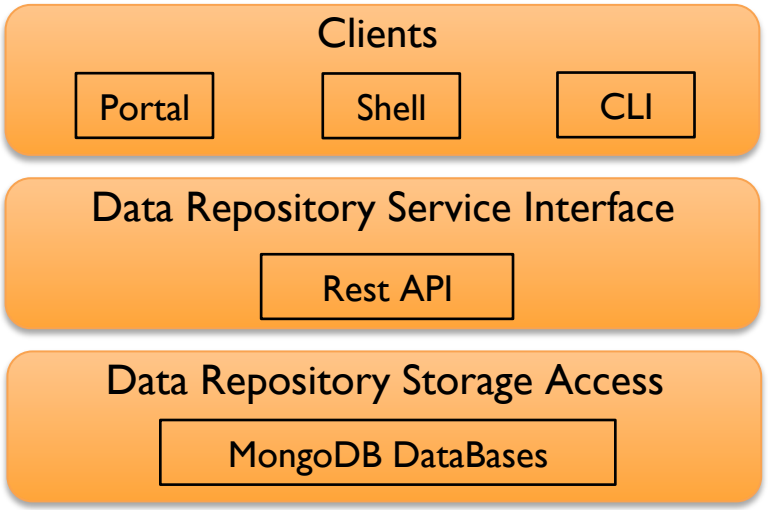
Component II – Data Identity Client for Galaxy

- Data Identity Client added into Galaxy Tassel5 Workflow to harvest workflow data objects
  - Minimum instrumentation - Interact with Tassel5 pipeline script without touching Tassel core code base
  - User transparency - Automatically harvest DOs when workflow is executed from Galaxy engine
  - Plug & play model – With minor updates to client this framework can be used to harvest DO from applications across domains



Component III PRAGMA Data Repository

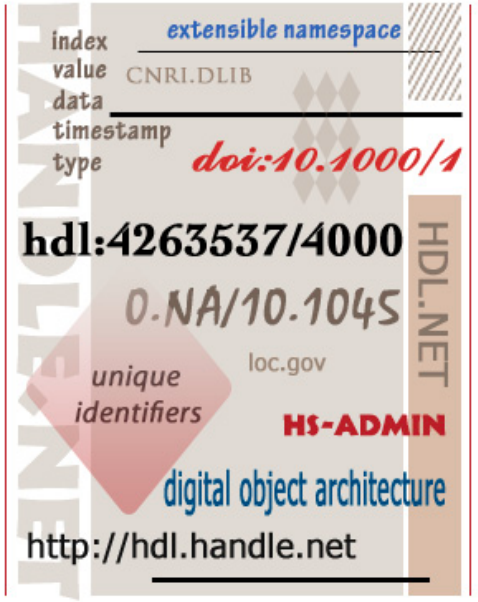
- Implemented with MongoDB, which provides persistent feature that distributes the database among different machines while maintain replicas in other machines.
- A single framework to store both metadata and data and offer users the possibility to decide the information they want to have as data objects metdata.
- 2 separate databases: Staging DB and permanent Repo DB
  - CRUD operation support for DOs in staging DB
  - DOs in permanent Repo DB only support READ; UPDATE and DELETE not allowed



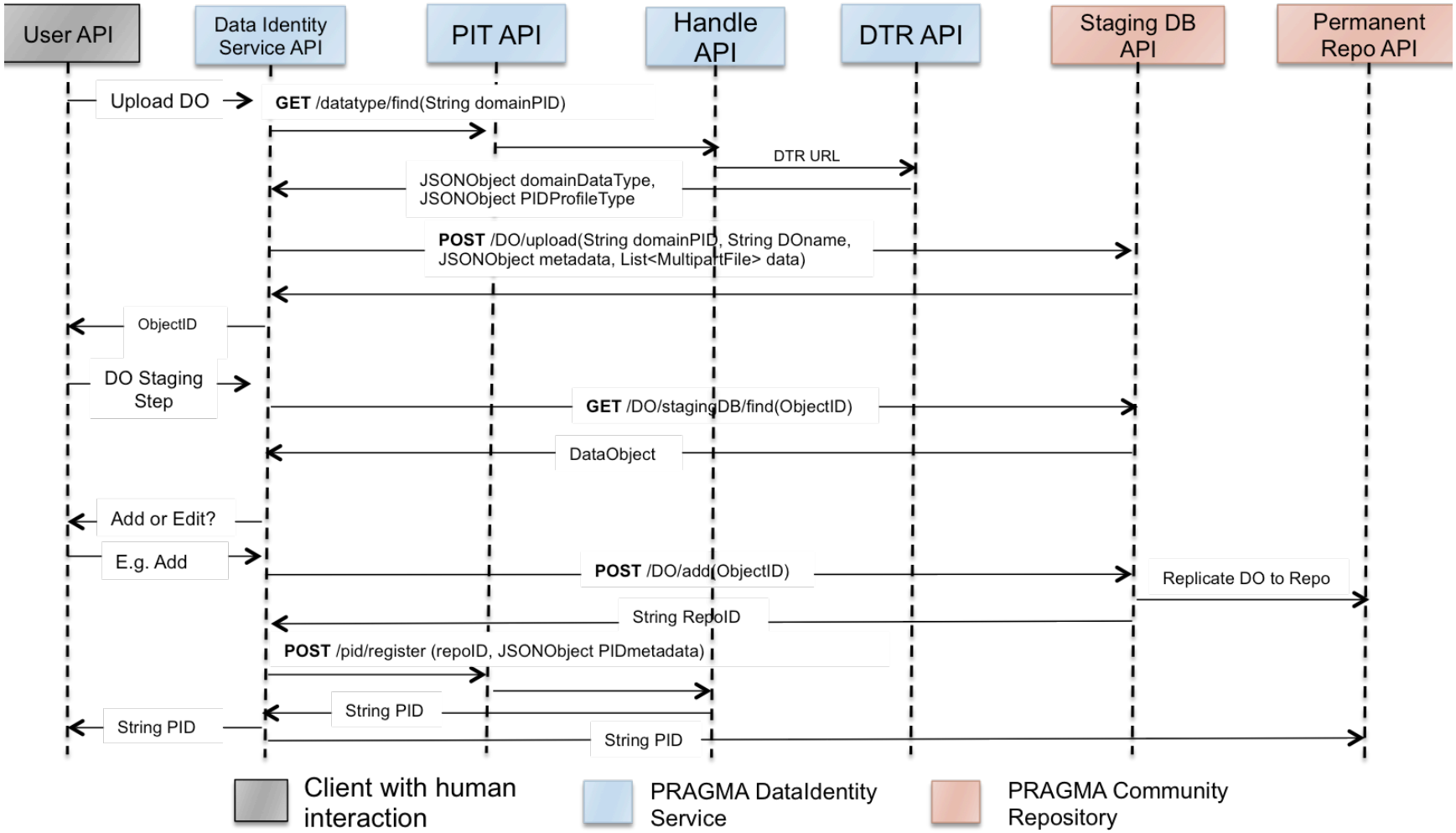
Component IV - Handle Service @ CNRI

- CNRI hosted a handle server V8 instance for our evaluation;

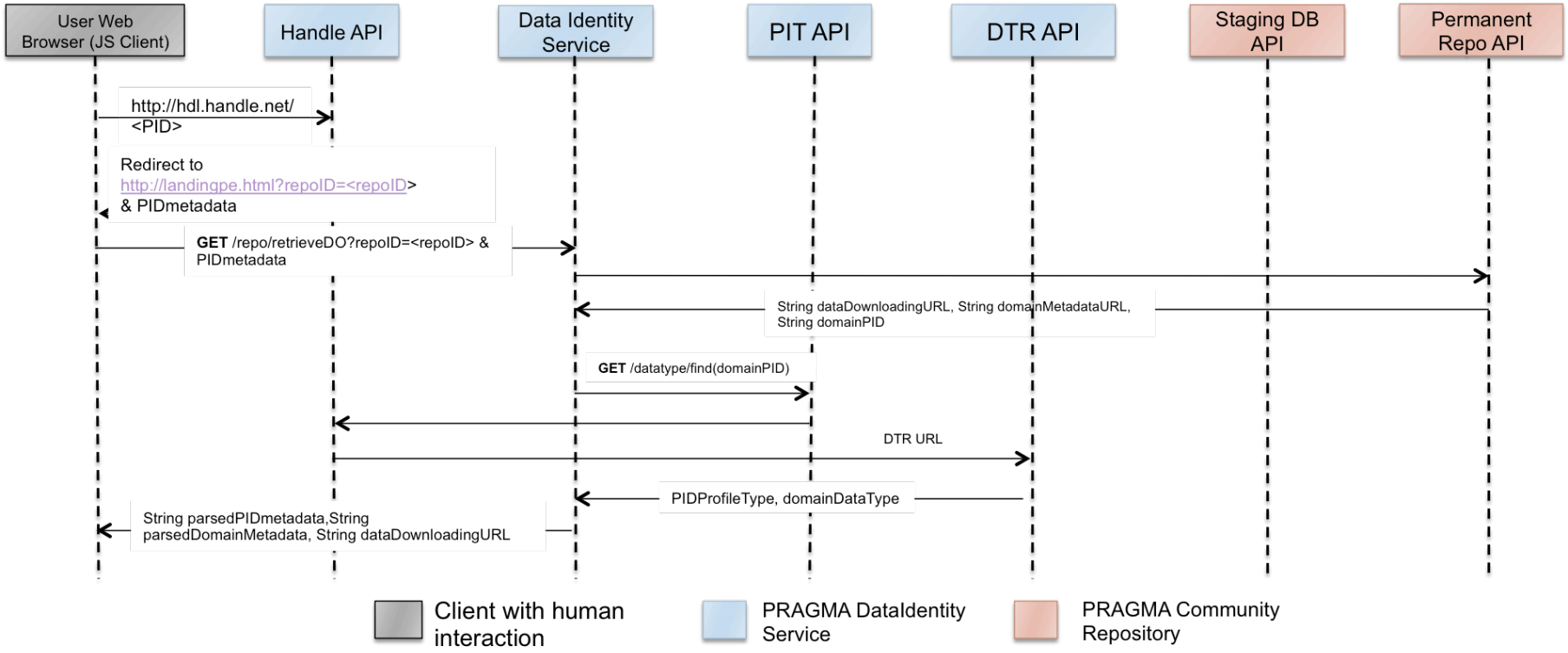
- Handle instance configurations:
  - https://38.100.130.12:8000/
  - Handle prefix: 11723



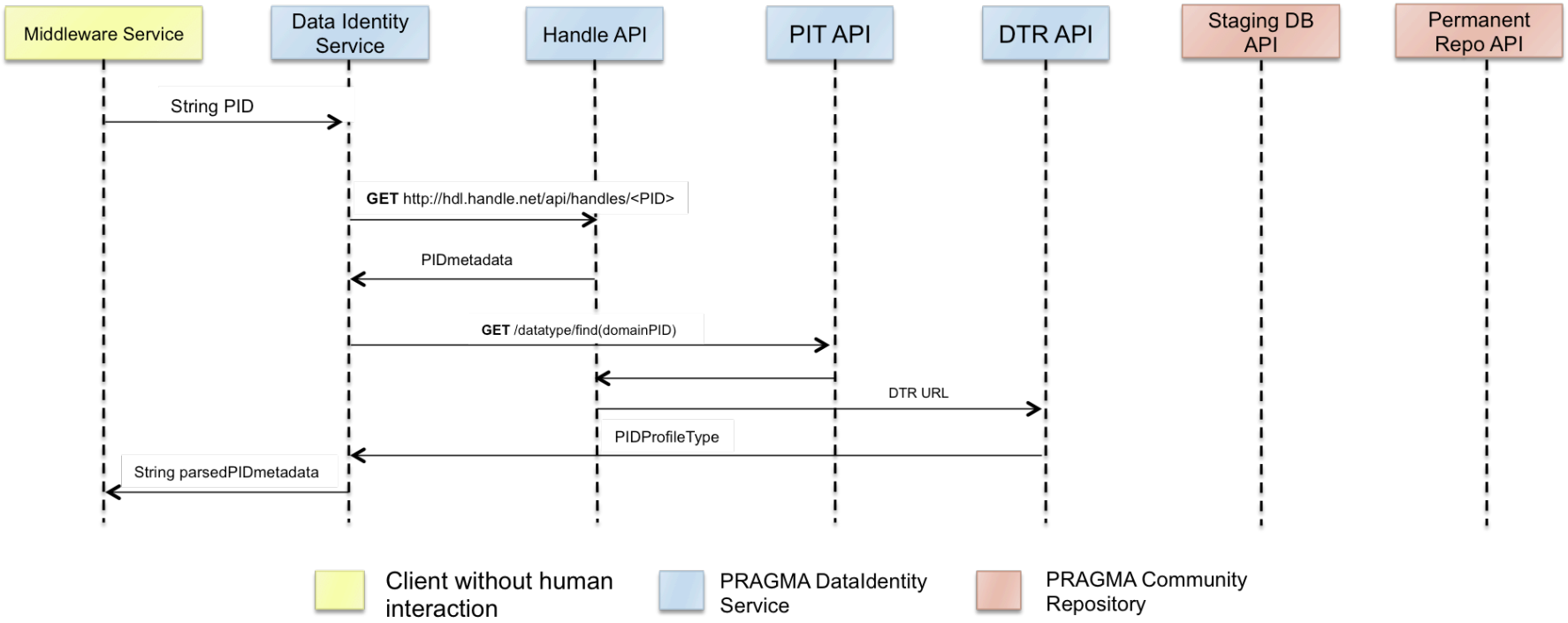
DO Upload Timeline Diagram



DO Retrieval Timeline Diagram



Middleware Service Timeline Diagram



Infrastructure and Resources

Data Identity GUI

Galaxy Workflow Server

- Tassel 5 Workflow
- Data Identity Tassel Client

Rice Genomics Workflow VM
SDSC, CA, US

PRAGMA Data Identity Service

- Data Identity Server
- PIT Service
- DTR Service

PRAGMA Data Repository

- Repository Rest API
- MongoDB DataBase

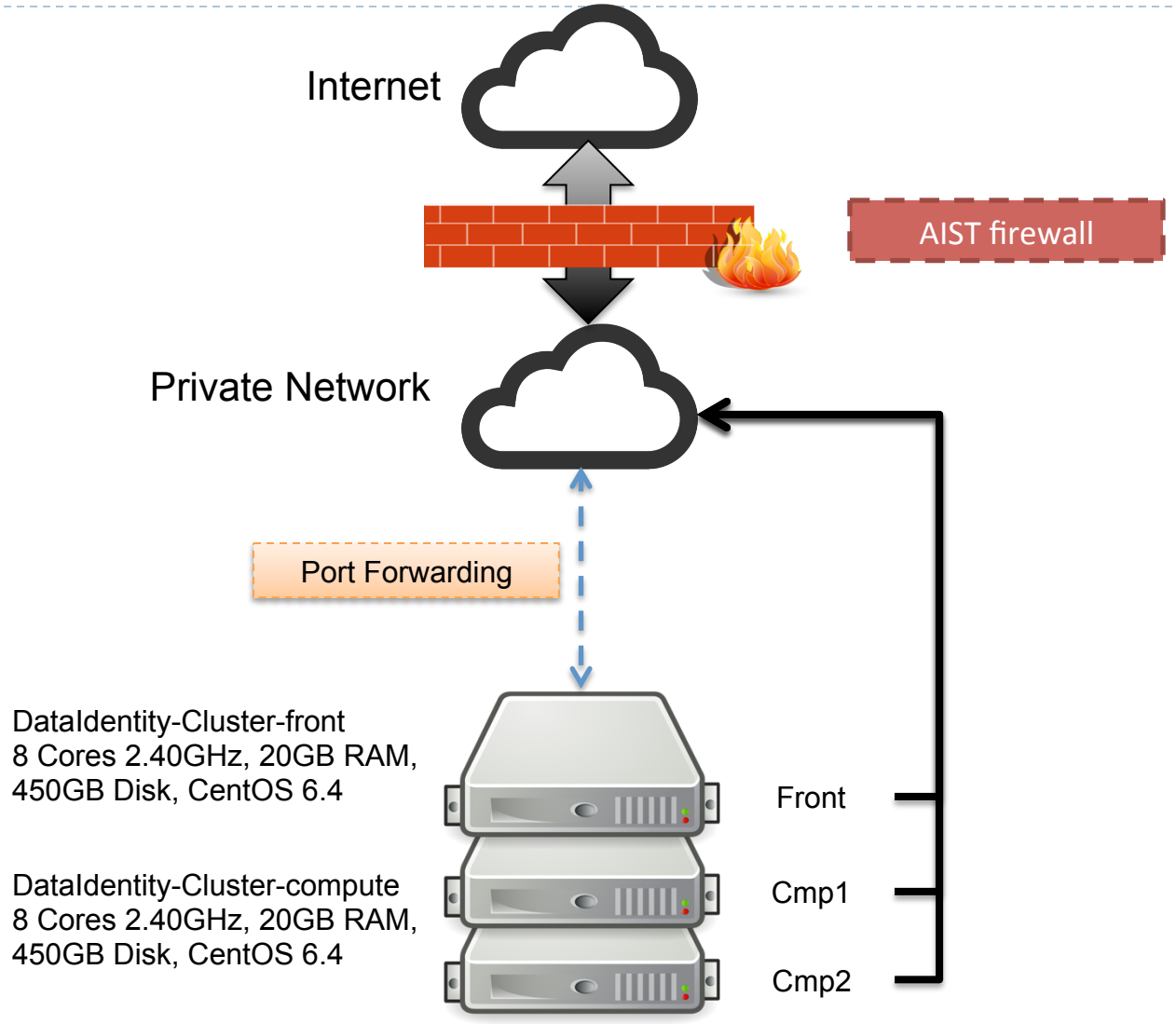
AIST Data Identity Cluster,
Tsukuba, Japan

Handle.net Service

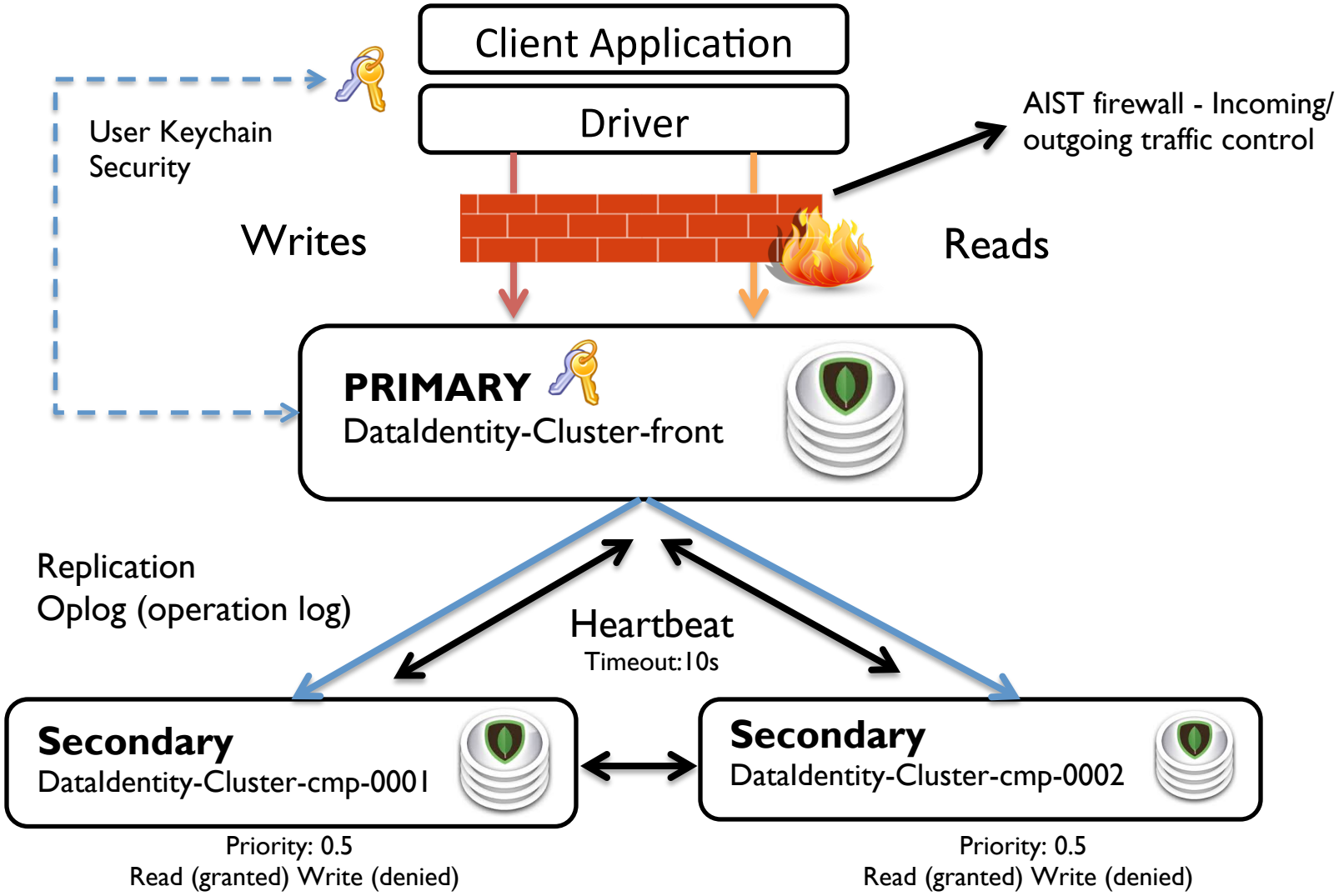
- Handle V8 Instance

PID Service Instance
CNRI, Virginia, US

AIST Data Identity Cluster Setup



Persistent Data Host in AIST



Success to Date

- The PRAGMA Data Identity Services is a user transparent means of harvesting DOs from applications and assignment of PIDs to scientific outcomes.
  - Reviewed and informed by core members of the rice genomics team
  - Usability study with members of IRRI rice genomics community planned for occur Fall 2016
  - Software is stable. Early version of the Open software available on Github.
  - Built with default PID information types and metadata

Future Work

- Data identity services and PRAGMA repository can be used in other applications running on and off the PRAGMA Cloud testbed.

- Next steps:
  - User interface and hardening over Fall 2016 for more robust operation
  - Refine metadata types based on user group study feedback
  - Extend data server (mongoDB) with basic preservation capabilities
  - Install services at US National Data Service (NDS)
  - Serve as basis for US testbed
  - Evaluate provenance capture

More Information

- For more information, please visit the following URLs:
  - Open Code Base & API Documents

https://github.com/Data-to-Insight-Center/RDA-PRAGMA-Data-Service
https://github.com/Data-to-Insight-Center/PRAGMA-Data-Repository

  - RDA PID Information Types Working Group

https://rd-alliance.org/groups/pid-information-types-wg.html

  - RDA Data Type Registries WG:

https://rd-alliance.org/groups/data-type-registries-wg.html

  - CNRI Handle.Net Registry

https://www.handle.net/

  - IRRI Tassel 5 Rice Genomes Workflow

http://rocks-53.sdsc.edu:8080/

References

[1] Bradbury PJ, et al. 2007. TASSEL: Software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633-2635.

[2] Mauleon, R. et al., 2012. IRRI GALAXY: bioinformatics for rice. ISCB-Asia/SCCG, 2012, Shenzhen, China

[3] Li J.Y., Wang J., and Zeigler R.S., 2014. The 3,000 rice genomes project: new opportunities and challenges for future rice research. GigaScience, 2014. 3: 8. DOI: 10.1186/2047-217X-3-8

[4] Weigel, T. et al., 2013. A Framework for Extended Persistent Identification of Scientific Assets. Data Science Journal, 12, pp.10–22. DOI: http://doi.org/10.2481/dsj.12-036

Acknowledgement

- This project funded in part by Research Data Alliance/US through grant from MacArthur Foundation; by NSF grant # OCI-#1234983, and by funding from AIST ICT International Team.

- With special thanks to CNRI for hosting handle V8 server for evaluation RDA PIT/DTR tool. We thank Tobias Weigel from RDA for all the instructions and discussions about RDA output. We also thank Ramil Mauleon and Venice Juanillas from IRRI for tremendous help from Rice Genomes domain.

The End

Thanks!