Breakout Session 3: Track A

ImmPort - NIH STRIDES: Facilitating Access of Immunological Data in ImmPort for Analyses

Mr. Srinivas Chepuri
Lead Enterprise Architect, ImmPort - NIAID/NIH

ImmPort - NIH STRIDES

Facilitating access of immunological data in ImmPort for analyses

Agenda

- Motivation
  - Facilitating access of immunological data in ImmPort for analyses
- Achievements
  - Data Sharing
    - ImmPort Shared Data in S3 Buckets
  - Data Analysis
    - ImmPort Galaxy
  - Interoperability
    - GA4GH DRS API
- Lessons Learned
- Challenges
- FY2024 plans
  - Facilitating understanding of shared disease mechanisms leveraging interoperability standards across dbGaP/SRA, ImmPort, and Kids First DRC

Motivation

Goal: Facilitating access of immunological data in ImmPort for analyses

1. FY22: Analysis of ImmPort flow cytometry data within Galaxy platform on the cloud
    a. Cloud-enabling ImmPort flow cytometry data within NIH AWS STRIDES environment for easier data access from within the ImmPort Galaxy platform
    b. Eliminate the download and upload steps for ImmPort shared flow cytometry files to ImmPort Galaxy workspaces
2. FY23: NIAID Ecosystem use case
    a. Use STRIDES for hosting the private raw data files for the Asthma and Allergic Diseases Cooperative Research Centers (AADCRC) program in the cloud and for serving them via GA4GH DRS API to analytical platforms like SevenBridges.
3. FY24: Host shared data in the cloud for standardized data access via GA4GH DRS API.
    a. S3 Urls for streaming and downloading the files

FY2022: ImmPort Data Cloud Accessibility for analysis within Galaxy

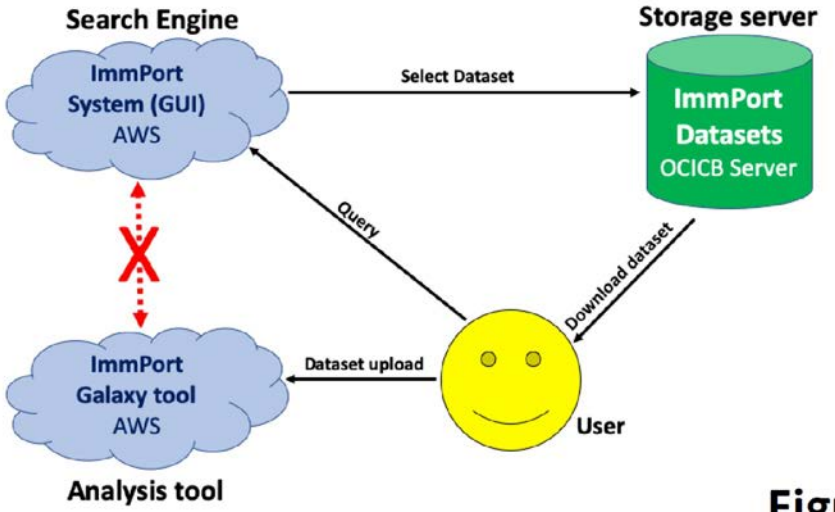


Figure 1

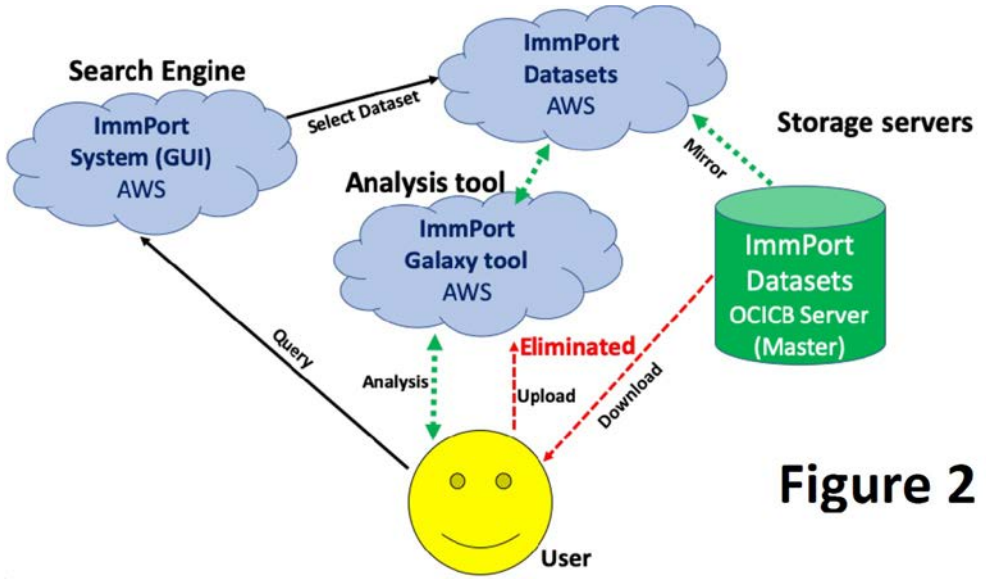
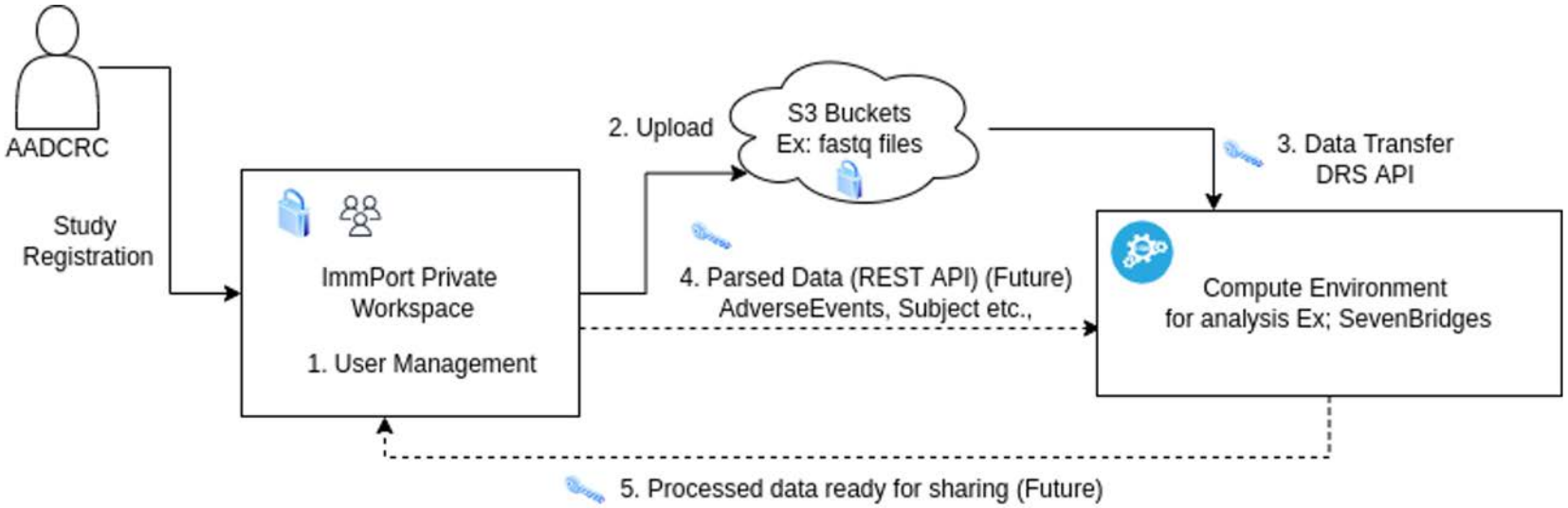


Figure 2

- Host Shared data on AWS STRIDES
  - S3 Bucket
- DR37 released on 12/15/2020
  - 4.6TB
  - OCICB to AWS
- Deploy ImmPort Galaxy on AWS STRIDES
  - EC2 instance

- Mount S3 bucket as a file system within Galaxy
  - Preload shared data library
  - No Egress costs, shared Data available locally for analysis
- Eliminated the need to download and upload data
- Develop analysis workflows in ImmPort Galaxy
  - Ex: FLOCK, FlowSOM, FlowAI, MetaCyto

FY2023: NIAID Data Ecosystem use case



- Hosted the private files (FASTQ) on AWS STRIDES Cloud
- Developed GA4GH DRS API providing standardized set of data access methods regardless of where it's stored and how it's managed.
  - Logical DRS Object ID for each file
  - ImmPort OAuth2 Integration
  - Two Access methods
    - S3 Signed and Streaming Urls
- NIAID Data Ecosystem
  - SevenBridges/Velsera
  - Connect ImmPort
    - OAuth2 Workflow
  - Download/Stream ImmPort data using DRS API
  - Develop analysis pipeline
- Processed data from the compute environment is uploaded back into ImmPort for public sharing

Lessons Learned

1. Ease of use
   - ImmPort team has extensive AWS engineering experience
   - Seamless migration from the existing ImmPort AWS environment to NIH STRIDES AWS environment
     i. Provisioning EC2 instances with the images from the ImmPort AWS account
   - Scalability - CPU, GPU
2. Migration of data from the on-prem file system to S3 bucket
   - AWS Configuration and Security Credentials
   - Opening of the TCP 443 port on the NIH firewall
   - Installation of AWS S3 Sync command line interface
     i. 4.6TB transferred in minutes
3. Mounting the S3 bucket as a local file system using S3FS-FUSE for the Galaxy software
4. Complexity
   - Replicating the ATO security controls
   - Load balancing configuration to use the *.immport.org domain for STRIDES environment

Challenges

1. Egress Costs
    - Egress costs if data is accessed outside AWS region
2. Alternative approaches for avoiding egress costs
    - Mount S3 buckets as local file systems on the analysis servers
        i. Galaxy, JupyterHub, NextFlow
    - AWS Open Data Sponsorship Program
        i. Covers the costs of storage and data transfer for a period of two years, can be renewed
3. Review any data location constraints
    - OCICB vs Cloud for Private & Controlled data sets
4. Additional licence/agreement requirements for cloud

FY2024 Plans

1. An NCPI project: Facilitating understanding of shared disease mechanisms leveraging interoperability standards across dbGaP/SRA, ImmPort, and Kids First DRC
    - Shared data on Cloud
        i. S3 Buckets in NIH STRIDES
    - GA4GH DRS API
        i. Standardized data access
        ii. Download and Streaming access methods
    - HAPI FHIR server
        i. HL7 Fast Healthcare Interoperability Resources (FHIR) for metadata and phenotypic/clinical variable representation.
    - Integration with cloud based compute workspaces like CAVATICA