iRODS at Bristol Myers Squibb

Status and Prospects. Leveraging iRODS for scientific applications in Amazon AWS Cloud

Mohammad Shaikh | Oleg Moiseyenko

Scientific Cloud Computing

iRODS UGM, Jun 9-11, 2020
**R&D: Delivering Innovative Medicines to Patients**

12 new medicines for Patients since 2011

R&D Investment in 2018: $5.1 BILLION on a non-GAAP basis*

5 PERCENT Increase over 2017.

Data as of January, 2019

*This non-GAAP amount excludes significant upfront and milestone payments for business development transactions and other specified R&D items. A reconciliation of GAAP to non-GAAP measures can be found on our website at [www.bms.com](http://www.bms.com). The GAAP amount is $6.3B.

~5,700 R&D Colleagues Worldwide
It’s all about data, Big Data!

**Scientific data sets**
- NGS data
- Proteomics
- Flow Cytometry
- Imaging data
- High-Throughput screening
- Mass spectrometry
- Databases

**Data governance**
- 25 years of retention
- Backups

**Major data sources**
- Raw data from labs
- Scratch space
- Results data
- External collaborations
- Public & government agencies
- R&D

![Exponential growth](chart.png)

- (Tens of PB’s)

From GB’s to PB’s scale
Lab data challenges

Data accessibility and sharing
- Silos between teams (organizational resistance)
- Generating insights in a timely manner, visualization and sharing

Networking, storage & computing power
- Efficient data exchanges, storage and processing

Replicating results
- Testing, validating, retesting,…

Data mining
- Lack of good metadata annotation

Data standards & compliancy
- Different formats, data integration and validation

Data insights are only as good as the data that drives them.
1. Instruments writes raw data into local scratch space
2. Raw data pushed to S3 by Storage Gateway/DataSync or via AWS CLI S3 commands
3. iRODS system scans S3 buckets regularly
4. Applications request data via iRODS metadata catalog
iRODS base architecture

- Client asks for data
- Data requests go to iRODS server
- Server looks up information in iCAT
- iCAT tells which iRODS server has data
- Data is retrieved from its physical location

Local data stores

- East 1
- East 2
- West 1
- West 2

UNIFIED NAMESPACE

- Metadata Catalog (iCAT)
- iRODS Server
- iRODS Rule Engine
- AWS
- S3 Bucket 1
- S3 Bucket 2
- S3 Bucket 3

BMS Scientific Instruments

BMS Scientists

MetaLnx browser

iQuery

API calls
iRODS for Computational Genomics

iRODS resources on cloud specs
- Consumers: m4.2xlarge (8vCPU/32GB)
- Provider: m4.10xlarge (40vCPU/160GB)
- Workers: c4.4xlarge (16vCPU/30GB)
- Redis server: r4.8xlarge (32vCPU/244GB)
- Metalnx: m4.large (4vCPU/16GB)
- Database: db.m4.4xlarge (16vCPU/64GB)
iRODS in NGS data processing pipeline

- NGS Labs Data
- AWS Direct Connect
- AWS CLI
- AWS Storage Gateway
- S3 Raw Data Bucket
- BMS NGS360
- AWS Batch
- Sequence alignment
- S3 Data Bucket
- Project Registry API
- S3 Result Bucket
- Virtual Private Cloud
- Seven Bridges
- S3 bucket A
- S3 bucket D
- S3 “drop” buckets
- AWS Lambda
- Gene expression database
- NGS QC Analysis
- Applications
- Scientists
- Vendor A
- Vendor B
- Vendor C
- Vendor D
- Collaborations (clinical data)

- iRODS
- Bristol Myers Squibb
- R&D, Informatics & Predictive Sciences
- NOT FOR PROMOTIONAL USE
iRODS in Discovery Imaging Platform

Scientific Instruments

Local storage layer

Images on local server (NFS)

Images on local server (NFS)

Images on local server (NFS)

Image Metadata database

AWS Direct Connect 10 Gb/s

AWS Snowball

On-premises

Storage Gateway Hardware appliance

BMS AWS Cloud

S3 bucket for transformed images

S3 bucket 1

S3 bucket 2

S3 bucket 3

S3 bucket A

S3 bucket B

S3 bucket N

S3 bucket N+1

S3 object store

Image transformation

Collaborator’s Cloud

Image transformation

Transformation

Scientists

Images on local server (NFS)

Images on local server (NFS)

Images on local server (NFS)

Images on local server (NFS)

Images on local server (NFS)

Local storage layer

Scientific Instruments

Local storage layer

Images on local server (NFS)

Images on local server (NFS)

Images on local server (NFS)

Image Metadata database

AWS Direct Connect 10 Gb/s

AWS Snowball

On-premises

Storage Gateway Hardware appliance

BMS AWS Cloud

S3 bucket for transformed images

S3 bucket 1

S3 bucket 2

S3 bucket 3

S3 bucket A

S3 bucket B

S3 bucket N

S3 bucket N+1

S3 object store

Image transformation

Collaborator’s Cloud

Image transformation

Transformation

Scientists

Images on local server (NFS)

Images on local server (NFS)

Images on local server (NFS)

Images on local server (NFS)

Images on local server (NFS)
### iRODS & Data Lake Integration

<table>
<thead>
<tr>
<th>Feature</th>
<th>iRODS – System of Records</th>
<th>Data Lake – System of Insights</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lab data move to cloud</td>
<td>✓</td>
<td>X</td>
</tr>
<tr>
<td>Technical meta data</td>
<td>✓ ✓</td>
<td>X</td>
</tr>
<tr>
<td>Business meta data</td>
<td>✓ ✓ ✓</td>
<td>X</td>
</tr>
<tr>
<td>Data acquisition</td>
<td>✓ ✓ ✓</td>
<td>X</td>
</tr>
<tr>
<td>Analytics</td>
<td>✓ ✓ ✓ ✓</td>
<td>X</td>
</tr>
<tr>
<td>File Management</td>
<td>✓ ✓ ✓ ✓</td>
<td>X</td>
</tr>
<tr>
<td>Operational analytics</td>
<td>✓ ✓ ✓ ✓</td>
<td>X</td>
</tr>
<tr>
<td>Domain specific</td>
<td>✓ ✓ ✓ ✓</td>
<td>X</td>
</tr>
<tr>
<td>Insights, Cross-functional</td>
<td>✓ ✓ ✓ ✓</td>
<td>X</td>
</tr>
<tr>
<td>File Management</td>
<td>✓ ✓ ✓ ✓</td>
<td>X</td>
</tr>
<tr>
<td>External Workflows</td>
<td>✓ ✓ ✓ ✓</td>
<td>X</td>
</tr>
</tbody>
</table>

**Legend:**
- ✓: Preferred platform
- X: Capability not existing on the platform
Roadmap to iRODS

- **Nov 2017**: Initial assessment, Pilot SoW
- **Feb 2018**: Production SoW
- **Mar-Aug 2018**: iRODS Pilot
- **Sep 2018**: iRODS Production deployment for Computational Genomics
- **Sep’18 - Jul’19**: NFS / S3 data syncs
- **Aug 2018**: AWS infrastructure setup for iRODS (1st iRODS Production)
- **Aug 2019**: 2nd iRODS Production environment in cloud
- **Dec 2018**: Dec 2018
- **Dec 2019**: iRODS Consortium membership
- **Nov 2020**: iRODS for ECL Labs Project
- **2021**: iRODS for Discovery Imaging Platform
- **2022**: We’re here today!

Towards Data Farm
Towards iRODS Data Farm

Global Search Index on top of iRODS Metadata catalog

- East - West Coasts Data Federation
  - East Coast Data Federation
    - East Zone 1
    - East Zone 2
    - East Zone 3
  - West Coast Data Federation
    - West Zone 1
    - West Zone 2

- Region 1, East
- Region 2, East
- Region 3, East
- Region 4, West
- Region 5, West

- Data Lake
- Data analytics
- Data providers
- Scientific groups
- Applications
Processing Data at Scale

Using iRODS for managing petabytes of data in hundreds of millions of files on distributed storage resources spread across the country.

- Number of S3 buckets: 200+
- Number of objects in S3: 800+ millions
- Size of dataset: 10+ PB
- Processing rate (regular data ingest): 5 millions objects per hour
iRODS data ingest – standard approach

Challenges

- iRODS catalog is always behind
- Negative space / Deleted files
Near real time data ingest – AWS Lambda function

Data Labs → Amazon S3 bucket → Amazon SNS → Amazon SQS → Amazon Lambda → iRODS Metadata catalog

Gene expression database

NGS QC Analysis

Applications
Updating iRODS Catalog with multiple S3 events

- Data Lab 1
- S3 bucket 1
- Amazon SNS 1
- Amazon SQS
- Amazon Lambda
- Amazon Systems Manager
- iRODS Metadata catalog
- Amazon CloudWatch

- Data Lab 2
- S3 bucket 2
- Amazon SNS 2

- Data Lab n
- S3 bucket n
- Amazon SNS n

- Amazon SNS 1
- Amazon SNS n

- Data Lab 1
- Data Lab 2
- Data Lab n

- Gene expression database
- NGS QC Analysis
- Applications

R&D, Informatics & Predictive Sciences
iRODS S3 Client AWS Lambda Function

This AWS Lambda function updates an iRODS Catalog with events that occur in one or more S3 buckets. Files created, renamed, or deleted in S3 appear quickly in iRODS.

- iRODS is assumed to have its associated S3 Storage Resource(s) configured with **HOST_MODE = cacheless_attached**
- If SQS is involved, it is assumed to be configured with **batch_size = 1**
- Handler: irods_client_aws_lambda_s3.lambda_handler
- Runtime: Python 3.7
- Environment Variables:
  - IRODS_COLLECTION_PREFIX : /tempZone/home/rods/lambda
  - IRODS_ENVIRONMENT_SSM_PARAMETER_NAME : irods_default_environment
  - IRODS_MULTIBUCKET_SUFFIX : _s3
iRODS S3 Client AWS Lambda Function

This AWS Lambda function updates an iRODS Catalog with events that occur in one or more S3 buckets. Files created, renamed, or deleted in S3 appear quickly in iRODS.

- You must configure your Lambda to trigger on all **ObjectCreated** and **ObjectRemoved** events for a connected S3 bucket.
- The connection information is stored in the **AWS Systems Manager --> Parameter Store** as a JSON object string.
- SSL Support
- This Lambda function can be configured to receive events from multiple sources at the same time.
- GitHub repository: [https://github.com/irods/irods_client_aws_lambda_s3](https://github.com/irods/irods_client_aws_lambda_s3)
- Release 1.0 date: May 12, 2020
Thank you

Acknowledgements

• BMS Cloud team
• iRODS support team
• Consortium members
• Conference papers
• Open source community

Mohammad Shaikh | Director | Scientific Computing Services | Cloud Computing & DevOps
100 Nassau Park Blvd, #300, Princeton, NJ 08540
Phone: 609.419.6352
Email: mohammad.shaikh@bms.com

Oleg Moiseyenko | Associate Director | Scientific Computing Services | Cloud Computing & DevOps
100 Nassau Park Blvd, #300, Princeton, NJ 08540
Phone: 609.419.6330
Email: oleg.moiseyenko@bms.com