Bristol-Myers Squibb iRODS Journey

Employing iRODS to manage petabytes of genomics data on cloud

Oleg Moiseyenko
Sr. Scientific Cloud Engineer
Bristol-Myers Squibb Company
If you’re going to fight the battle of your life, you’ve got to stay positive – in the midst of any storm, there’s always something to be grateful for.

Carol Willis
Renal cell carcinoma patient, benefiting from an *Opdivo-Yervoy* combination

Latest update - April 2019
Our Mission

To **discover, develop and deliver** innovative **medicines** that help patients prevail over serious diseases.
**Bristol-Myers Squibb Delivering in 2018**

**DELIVERING by the NUMBERS**

$22.6 BILLION in Revenue

9% Revenue Growth VS. 2017

**PRODUCT REVENUES**

$ BILLIONS

- **$6.7**
- **$6.4**
- **$2.7**
- **$3.5**
- **$2.0**
- **$1.3**
- **Other***

*Includes Empliciti, Baraclude, Sustiva, Reyataz, Hepatitis C franchise and Other Brands*
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.

~5,700 R&D Colleagues Worldwide

*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. The GAAP amount is $6.3B.

Data as of January, 2019
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**
(Tens of PB’s)

From GB’s to PB’s scale
Key considerations for data management system

BMS acceptance criteria

- Cloud integration
- Petabyte scalable
- CLI interface
- Rich API
- **Metadata driven**
  - NFS – S3 connectivity
  - User’s access management
- **Security**
  - Low price tag
  - Low administrative efforts
- **Established presence in life science & healthcare**
  - Support
Cloud advantages

• S3 object store
  • Unlimited size
  • Data protection: 99.999999999% durability
  • Build-in data distribution & replication
  • Easy integration with other cloud micro services
• No hardware / storage technology lock-in
• Cloud elasticity: vertical & horizontal
• Backups (versioning, snapshots, lifecycle rules)
• PaaS platform for database technologies
• High data security
• Low cost
iRODS deployment on cloud

**NetApp**

- Corporate data center
- Local server (NFS)
- Local server (NFS)
- Local server (NFS)

**AWS**

- Internet gateway
- VPC
- EC2
- iRODS Catalog Consumer
- iRODS Catalog Provider
- iRODS Ingest Worker
- iRODS Redis Server
- iRODS Metalnx Server
- Primary
- Standby
- iRODS RDS Database (PostgreSQL)

**AWS Direct Connect 10 Gb/s**

**Enterprise Data Lake**

- Genomics DataHub

**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)

**Virtual private cloud**

**Availability Zone 1**

- iRODS RDS Database (PostgreSQL)

**Availability Zone 2**

- iRODS RDS Database (PostgreSQL)

**Data replication**

- S3 bucket A
- S3 bucket B
- S3 bucket N
- S3 bucket N+1

**S3 object store**
iRODS use cases

Data management
- Moving data from labs to cloud
- Managing various scientific datasets
- Providing access to clinical data sets

NFS/S3 data sync
- Sync S3 object store with on-prem data stores (NFS)
- Confirm no deltas left
- Provide logs for audits
- Unmount local storage

ML based data enrichment
- ML and deep learning algorithms classify image data
- iRODS catalog is updated with tags with classification information

Data Lake integration
- Integrate iRODS meta data catalog with Clinical data lake
- Enterprise Data lake ingestion tools use iSQL to read iRODS meta data catalog
Our journey to BMS iRODS Data Farm

Initial assessment, Pilot SoW  
Feb 2018

iRODS Pilot  
Mar-Aug 2018

AWS infrastructure setup for iRODS  
Sep 2018

Production SoW  
Nov 2017

iRODS Production deployment for Computational Genomics  
Sep 2018

NFS / S3 data syncs  
Nov’18 – May’19

AWS infrastructure setup for iRODS  
Nov’18 – May’19

BMS iRODS Consortium membership  
Aug 2019

2nd iRODS Production environment in cloud  
Aug 2019

We’re here today!

Towards BMS Data Farm  
2020

Genomics DataHub  
2020

NOT FOR PRODUCT PROMOTIONAL USE
iRODS: Pros & Cons

**Pros**

- Easy to deploy
- Metadata driven
- Flexible rule engine
- Same names for logical/physical file paths
- Established presence in life science
- Rich API
- Data virtualization
- Flexible & PB-scalable system
- High data retention requirements (10-25 years)
- ACL's and permissions support
- Secure data sharing
- Workflows automation & data replication

**Cons**

- Higher complexity level
- Requires advance development
- No mechanism to enforce good metadata system (“garbage in, garbage out”)
- No user-friendly front-end interface
Challenges

- MD5 checksums
- Scanning speed: every million files on S3 takes about two hours to scan on the NFS side
- Data replication speed
- Non-readable characters in file names
- Permission issues
- Redis cache issue (once)
- Verification upon data sync process completion
BMS Wishlist

- NFSRODS integration
- Minio iRODS Gateway
- Better LDAP/AD integration
- Metadata templates
- iRODS catalog structure specs
- Advance SQL support
- Push notifications instead of polling
- Database performance optimization
- User-friendly front end
- Improved documentation
- AWS EC2 spot instances for workers
BMS – iRODS: Next steps

- Capture, manage, apply metadata to data collections
- Deliver continuous data scans for S3 store
- Unify access to metadata; metadata enrichment
- Unify the governance approach for iRODS
- Advance development: rules, policies, etc.
- Genomics DataHub (gateway to BMS data lake)
- LDAP integration for user’s authentication
- Dashboarding/system health (iRODS audit plugin)
- Towards BMS Data Farm (zones federation)
Acknowledgements

**BMS iRODS Core Team**
- Mohammad Shaikh
- Isaac Neuhaus
- Carlos Rios
- Mark Russo
- Oleg Moiseyenko

**iRODS Consortium Team**
- Jason Coposky
- Terrell Russell

**BMS iRODS Cross Team**
- Dan Huston
- Valerie Williams
- Eric Sison
- Dmitry Khavich
- Paul O’Malley
- Gopal Prakriya

- Sponsor, Business Partner: Ajay Shah