iRODS at Bristol Myers Squibb

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

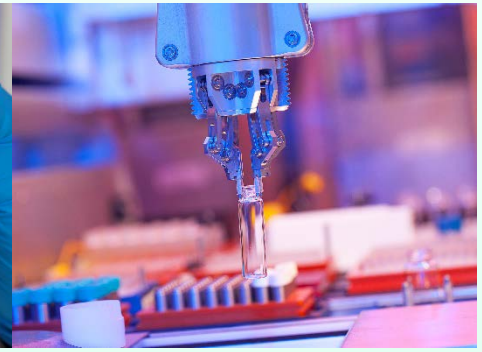
Mohammad Shaikh | Oleg Moiseyenko

Scientific Cloud Computing

iRODS UGM, Jun 9-11, 2020

R&D: Delivering Innovative Medicines to Patients



40 compounds in development

12 new medicines for Patients since 2011

~5,700 R&D Colleagues Worldwide

R&D Investment IN 2018

$5.1 BILLION on a non-GAAP basis*

5 PERCENT Increase over 2017.

*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

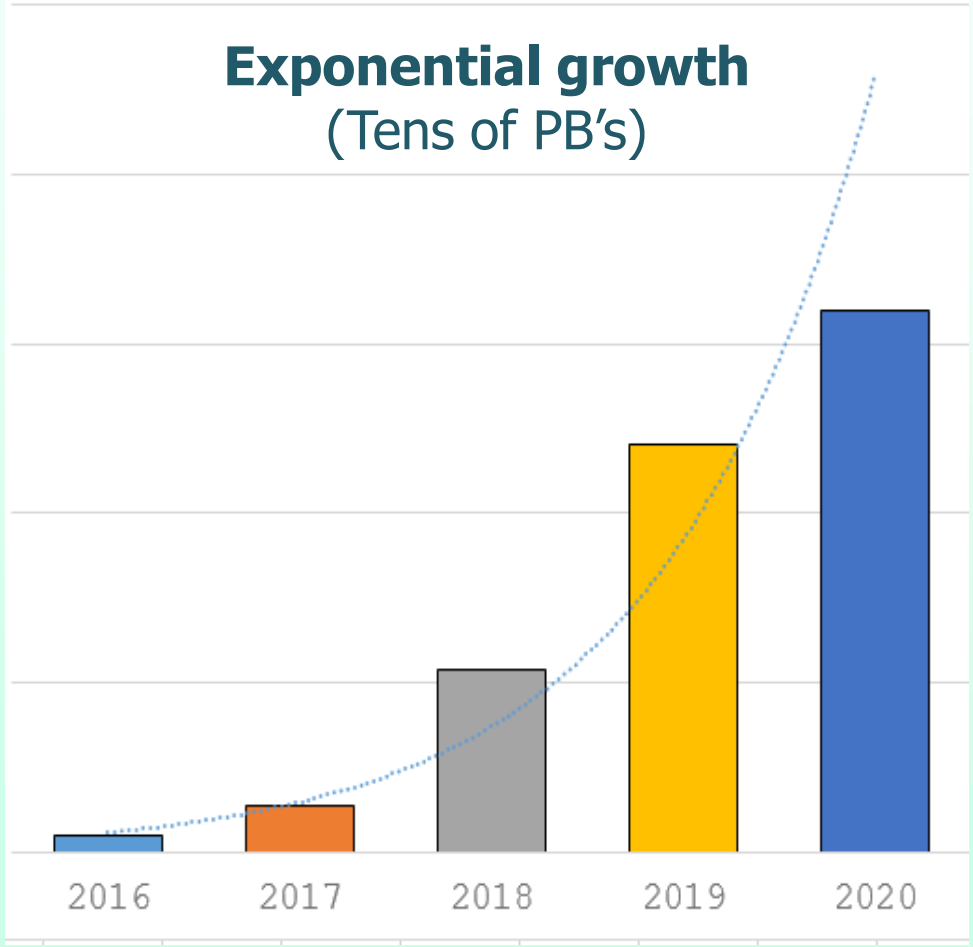
It's all about data, Big Data!

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

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,…

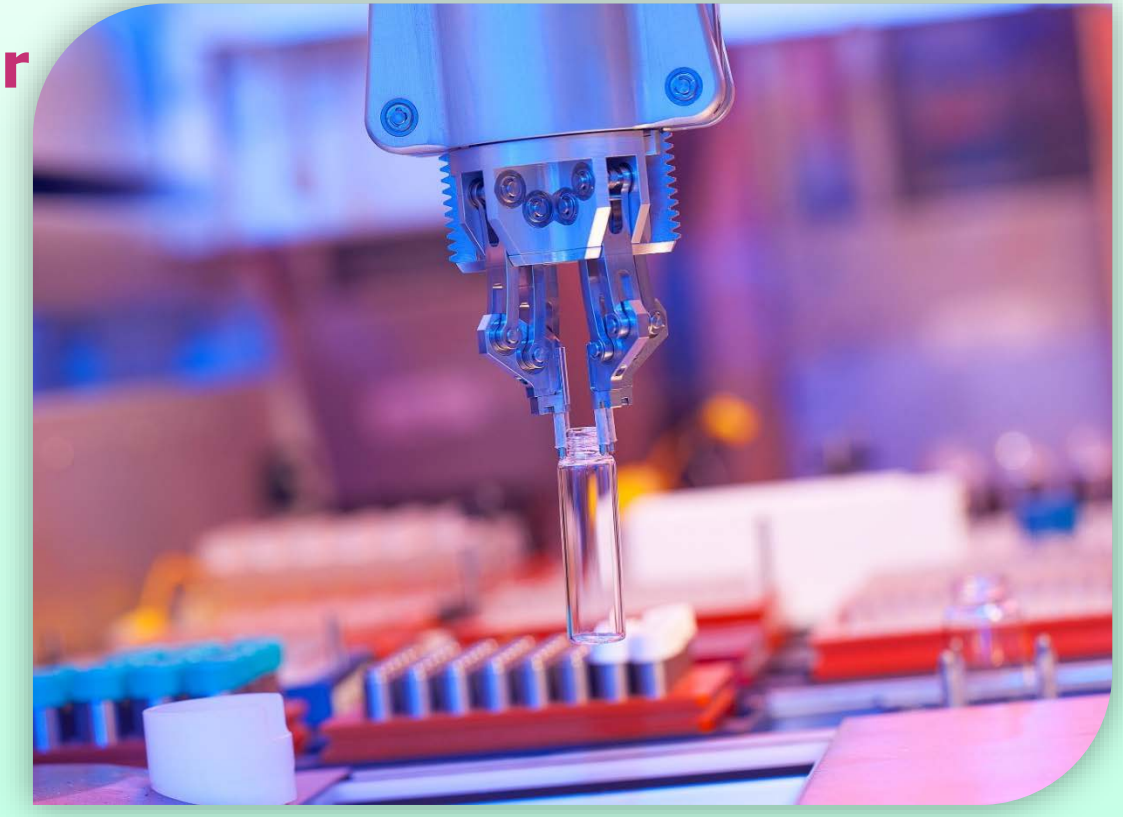
Data mining

- Lack of good metadata annotation

Data standards & compliancy

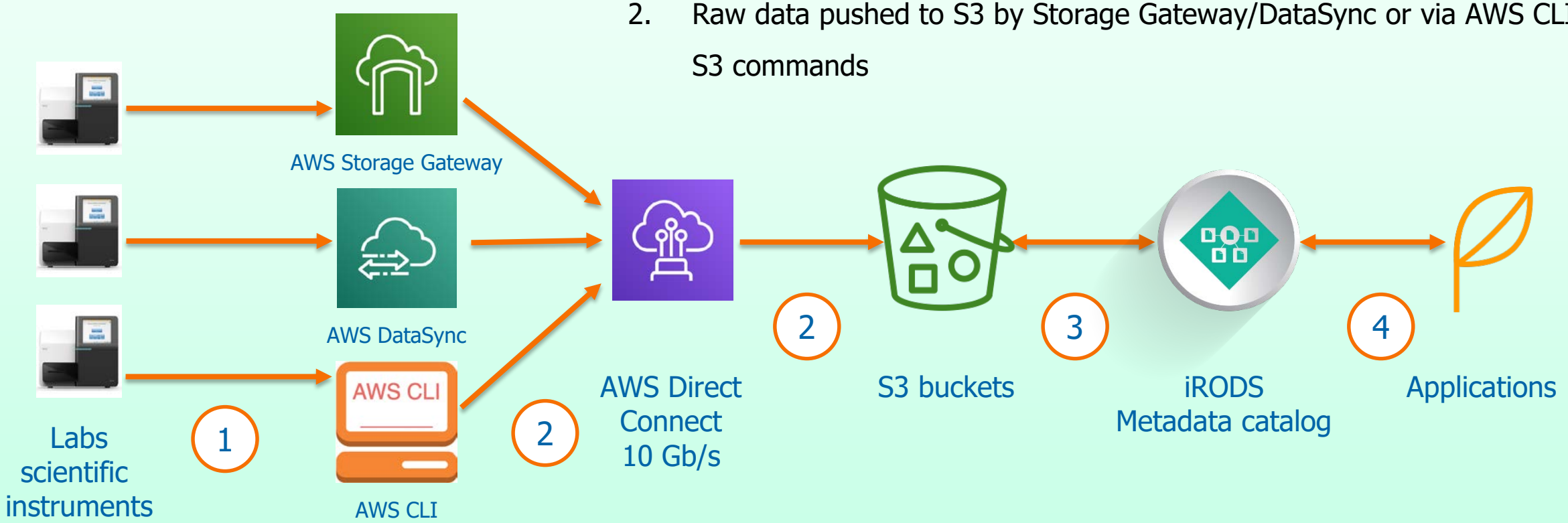
- Different formats, data integration and validation

Data insights are only as good as the data that drives them



Typical data flow diagram

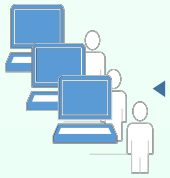
1. Instruments writes raw data into local scratch space
2. Raw data pushed to S3 by Storage Gateway/DataSync or via AWS CLI S3 commands



3. iRODS system scans S3 buckets regularly
4. Applications request data via iRODS metadata catalog

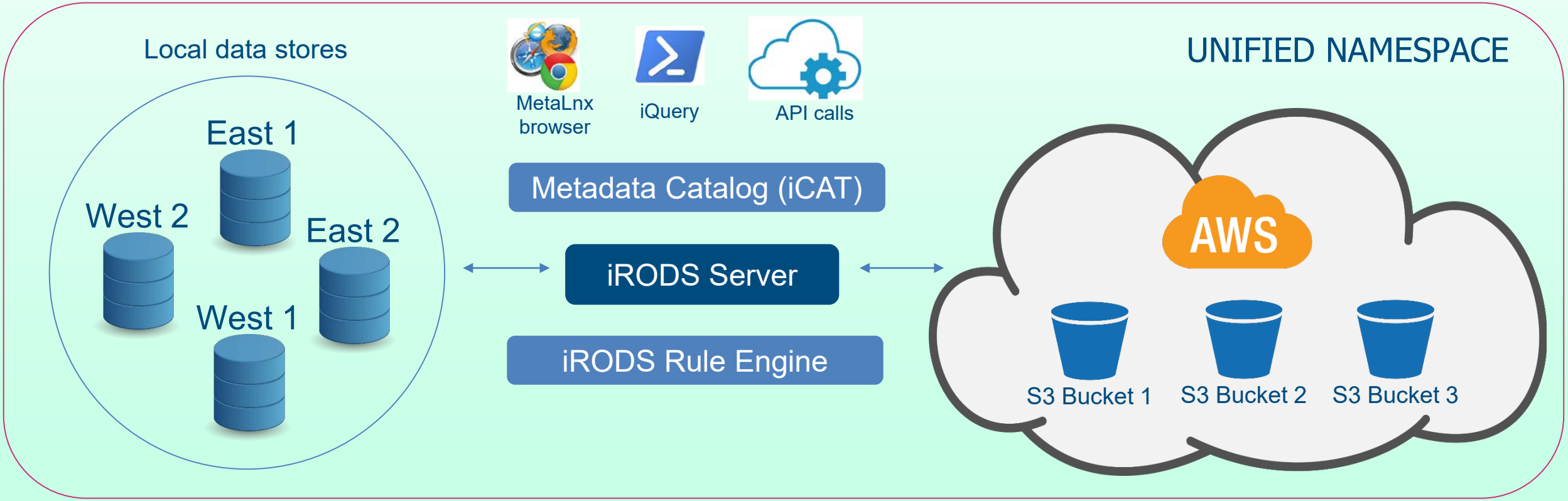
iRODS base architecture

BMS Scientific Instruments

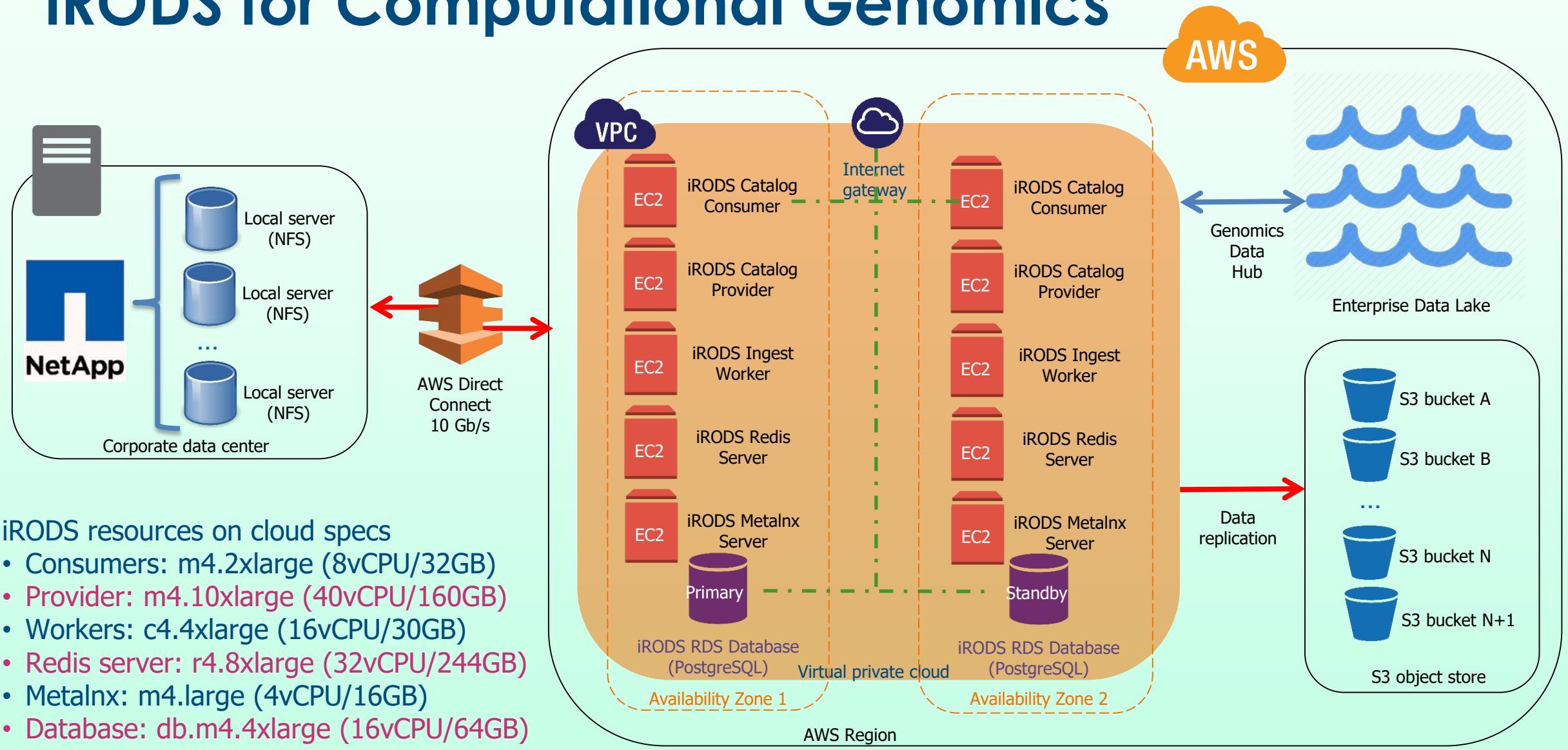


BMS Scientists

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



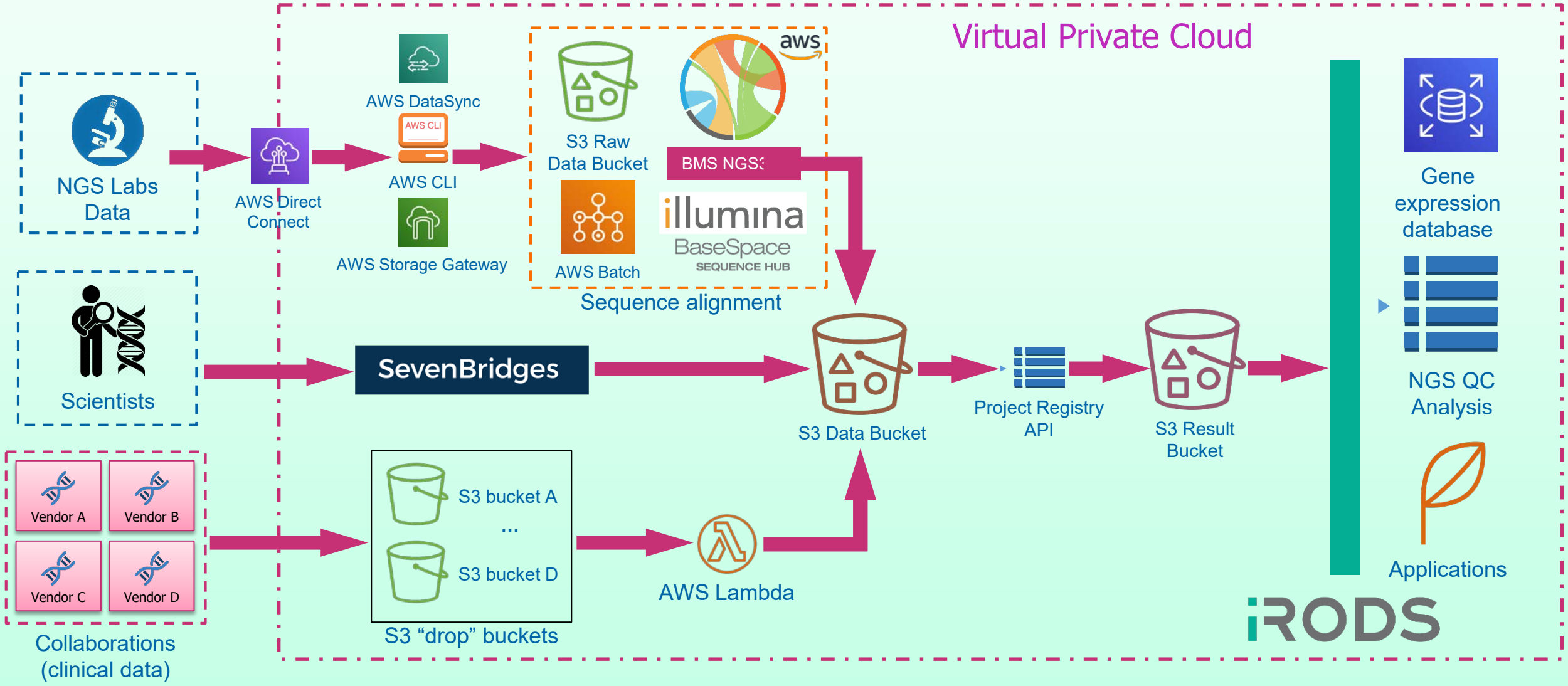
iRODS for Computational Genomics



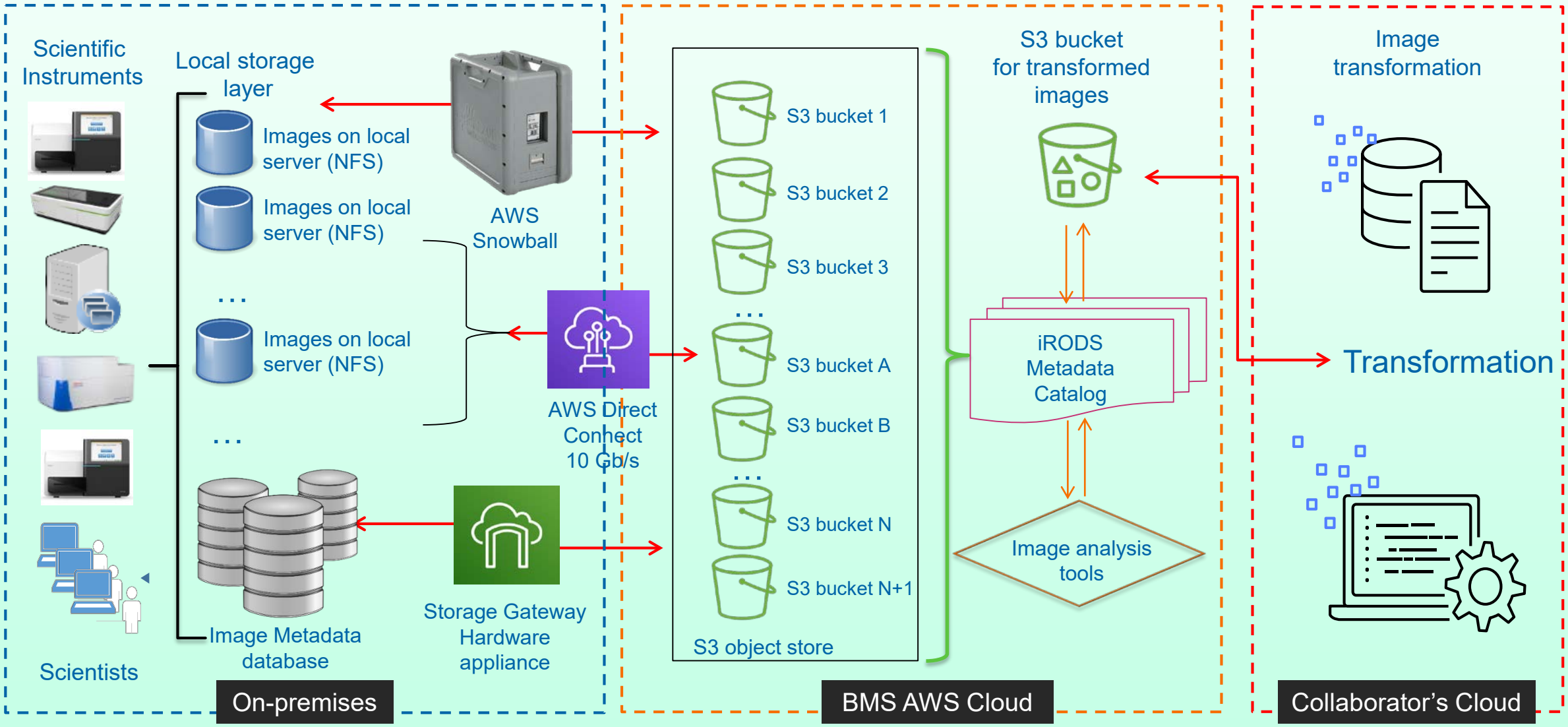
iRODS resources on cloud specs

- Consumers: m4.2xlarge (8vCPU/32GB)
- Provider: m4.10xlarge (40vCPU/160GB)
- Workers: c4.4xlarge (16vCPU/30GB)
- Redis server: r4.8xlarge (32vCPU/244GB)
- MetaInx: m4.large (4vCPU/16GB)
- Database: db.m4.4xlarge (16vCPU/64GB)

iRODS in NGS data processing pipeline



iRODS in Discovery Imaging Platform



Flow Cytometry Data Flows

Raw

fcs files

Storage

NuGenesis

FileCatcher

Exp. Design

Bio Signals Analytics

Workbook for Biologics registration

Chemistry workbook

LIMS

Analysis

FCS Express

FlowJo

Cytobank

Tracking

Shared Drives

Spotfire

Aggregation

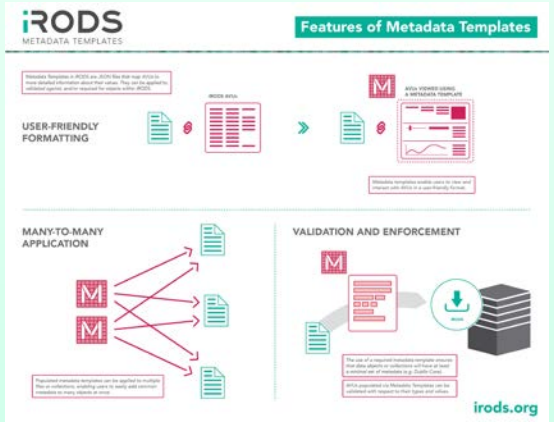
Analytics & Data Lakes

iRODS
S3

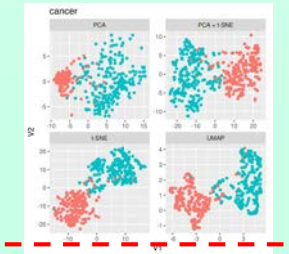
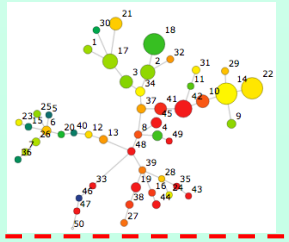
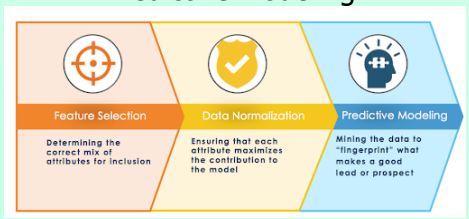
Slide credit:
Goce Bogdanoski

Flow Cytometry – Digital Intelligence / ML

2
User
Instrument Data Generation
Flow Cytometry Data Hub
UI
AWS DataSync or SmartSync
3
Data Storage
FC Database
6
4
Data Standardization guidelines
1
Metadata Guidelines
File Nomenclature
Existing Source Integration
Data Dictionaries
User Access to HPC on the Cloud
Amazon EC2
5¹
Automated Gating (AaaS)
- AltraBio
- Cytapex Bioinformatics
- Astrolabe

FileSelector App
5²
5³
File Management, Tracking & Auditing


- Automated Ingest
- Storage Tiering
- Indexing
- Compliance
- Auditing
- Publishing
- Provenance
- Integrity

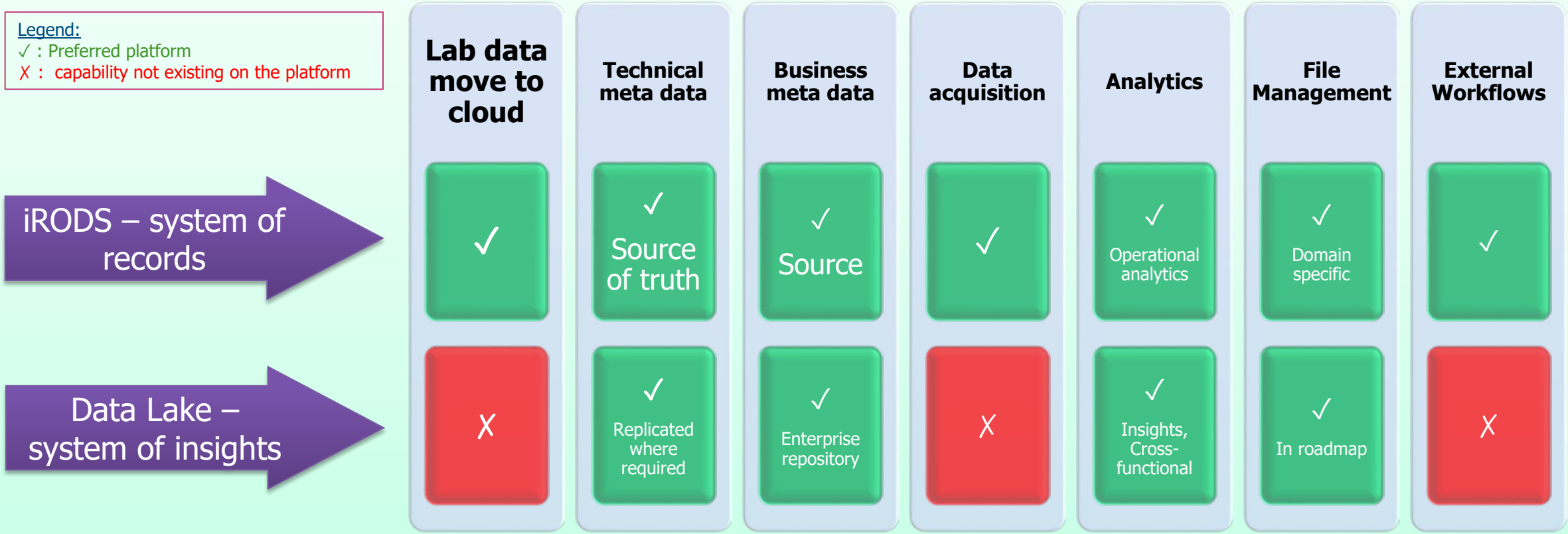
Unsupervised Analysis Pipeline
Dimensionality Reduction

Clustering

Data Pre-processing & Clean-up
Predictive Modeling


- t-SNE
- FlowSOM
- Citrus
- Spotfire
- Disqover
- Signals

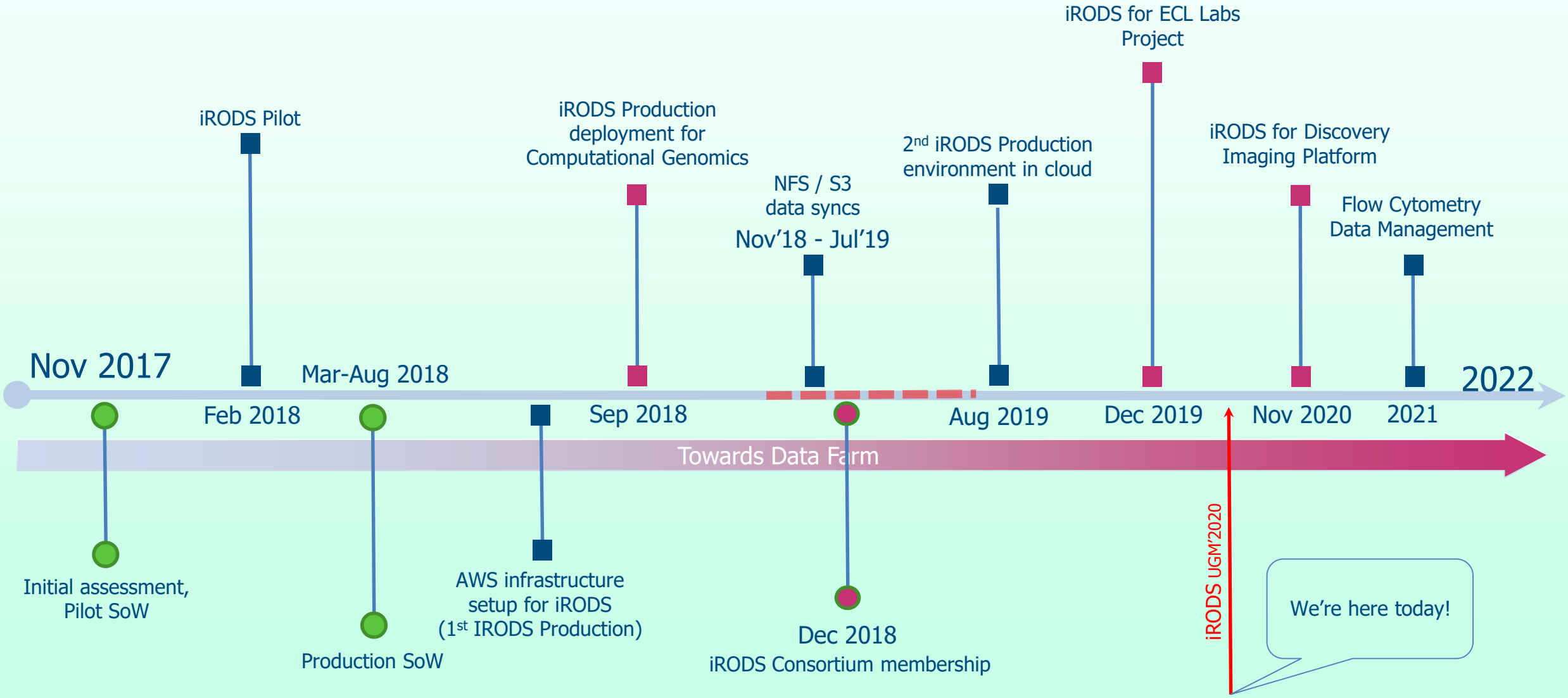
Slide credit: Goce Bogdanoski

iRODS & Data Lake Integration

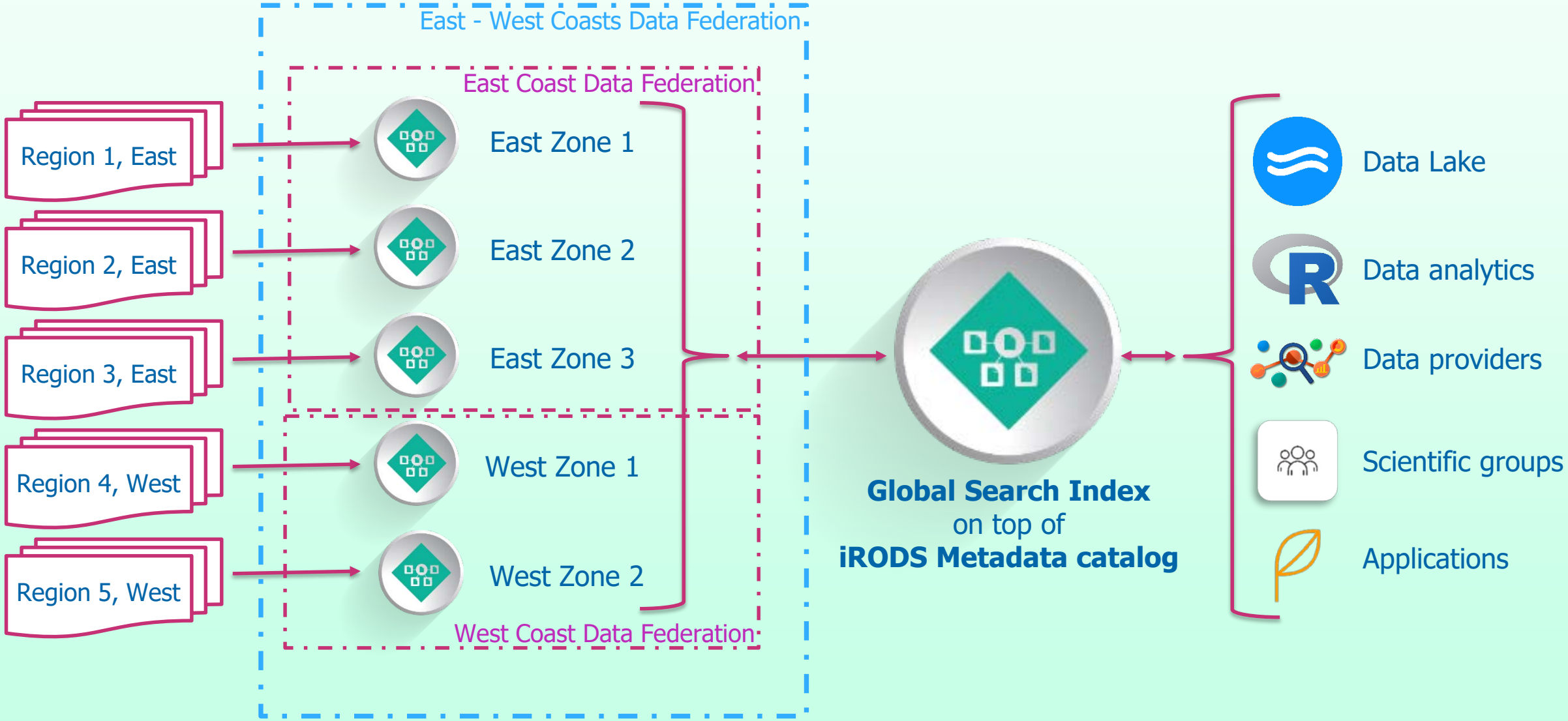
Legend:
✓ : Preferred platform
X : capability not existing on the platform



Roadmap to iRODS



Towards iRODS Data Farm

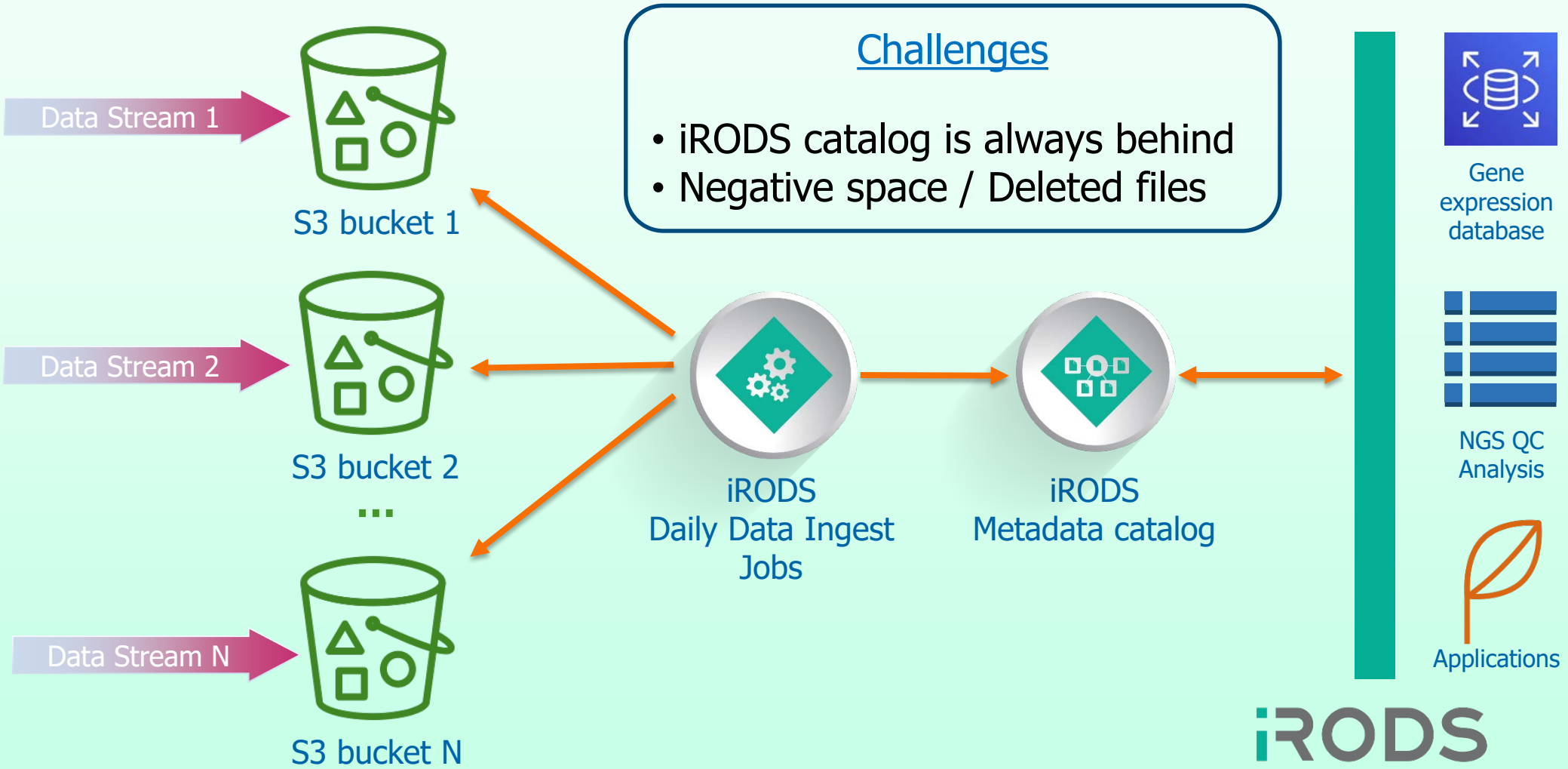


Processing Data at Scale

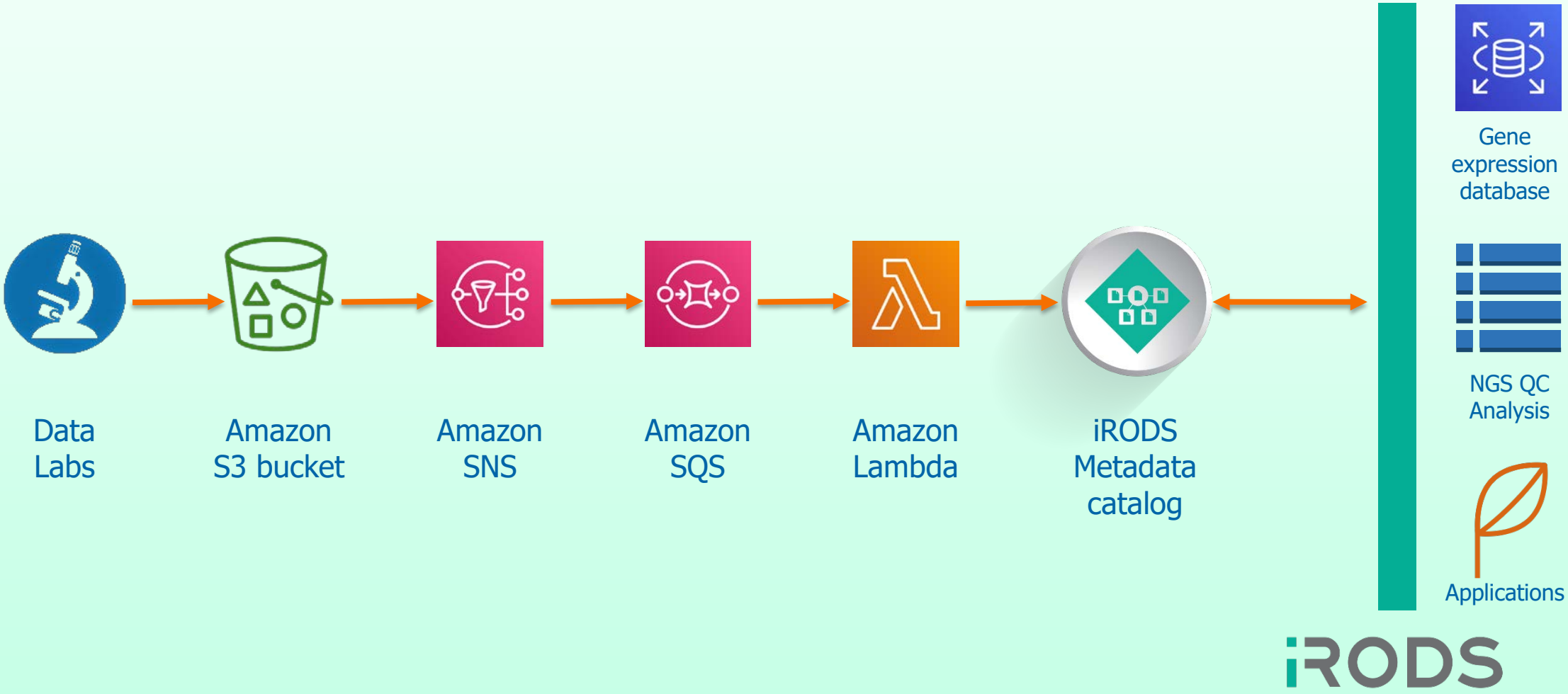
Using iRODS for managing petabytes of data in hundreds of millions of files on distributed storage resources spread across the country.

- 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**

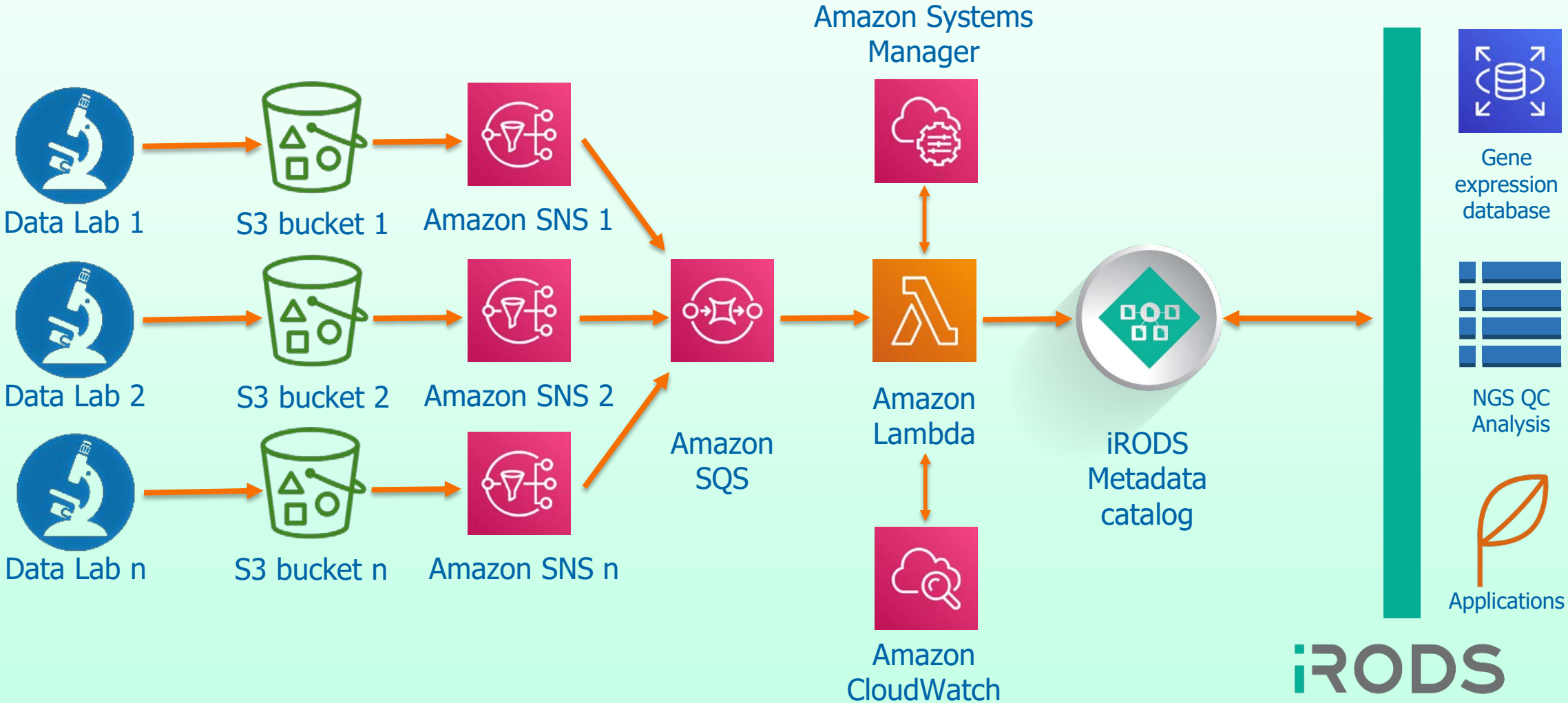
iRODS data ingest – standard approach



Near real time data ingest – AWS Lambda function



Updating iRODS Catalog with multiple S3 events



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
- 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

Bristol Myers Squibb™