iRODS Client: AWS Lambda Function for S3 1.0

Terrell Russell, Ph.D.
@terrellrussell
Chief Technologist, iRODS Consortium

June 9-12, 2020
iRODS User Group Meeting 2020
Virtual Event

Design Goals

- Play nicely with the universe of tools that already know how to write to S3 directly
- Allow those updates within the S3 namespace to smoothly flow into the iRODS Catalog
- Trigger automated data management due to crossing the policy boundary

iRODS Client: AWS Lambda Function for S3 1.0

Considerations

- Lambda can run Python code
- iRODS provides a python client library

Success would be...

- near-real-time, asynchronous, catalog updates for creates/moves/deletes

iRODS Client: AWS Lambda Function for S3 1.0

Files created, renamed, or deleted in S3 appear quickly in iRODS.

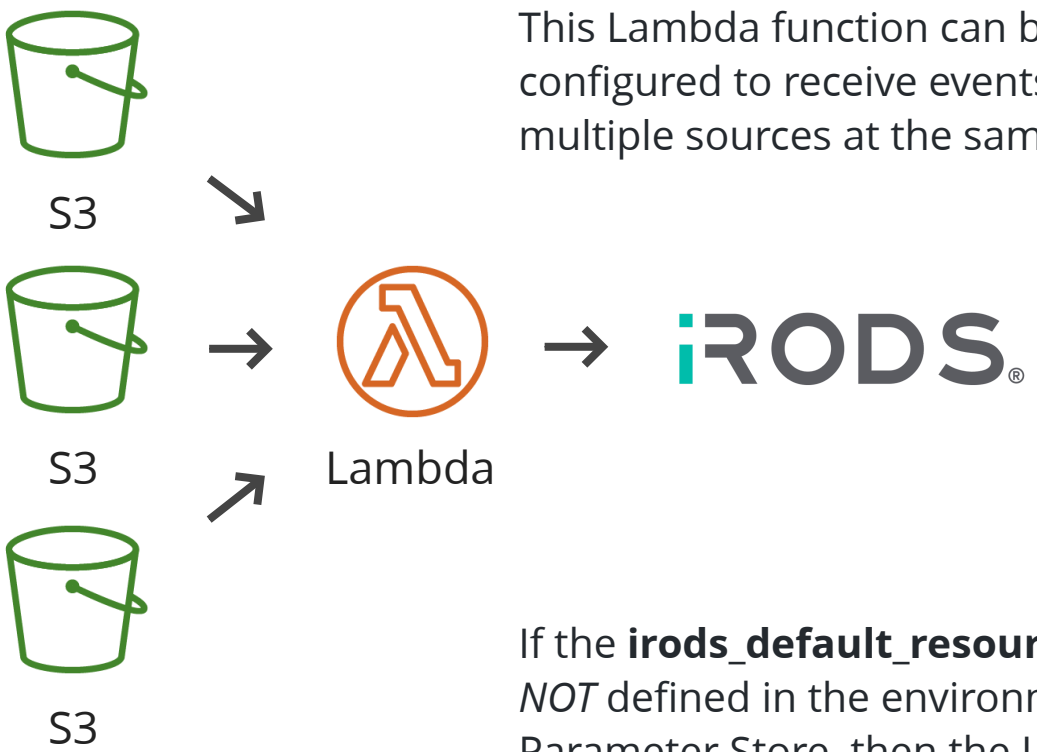
iRODS is assumed to have its associated S3 Storage Resource(s) configured with **HOST_MODE=cacheless_attached**.

You must configure your Lambda to trigger on all **ObjectCreated** and **ObjectRemoved** events for a connected S3 bucket.

The iRODS connection information is stored in the **AWS Systems Manager > Parameter Store** as a JSON object string.

SSL to iRODS is supported by placing a certificate in a relative path within the Lambda package.

iRODS Client: AWS Lambda Function for S3 1.0



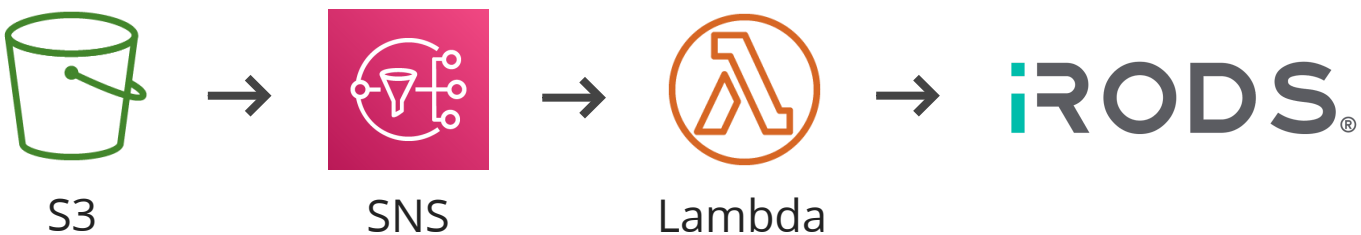
This Lambda function can be configured to receive events from multiple sources at the same time.

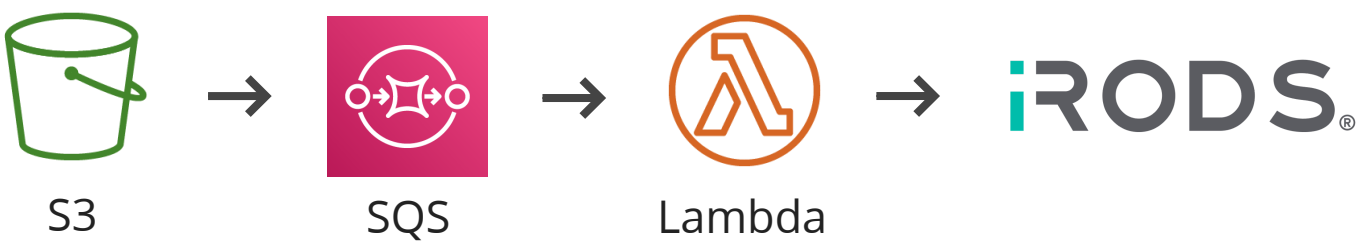
If the **irods_default_resource** is *NOT* defined in the environment in the Parameter Store, then the Lambda function will derive the name of a target iRODS Resource.

By default, the Lambda function will append **_s3** to the incoming bucket name.

iRODS Client: AWS Lambda Function for S3 1.0

The following AWS configurations are supported at this time:





iRODS Client: AWS Lambda Function for S3 1.0

Limitations

- S3 is decoupled from the Lambda. A **rename** is actually a **create** and a **delete** message. To iRODS, this becomes a new data object. This means any metadata AVUs associated with the now-deleted data object is lost. Could be remedied with full checksum comparison. Other ideas welcome.

- SQS configuration is limited to **batch_size = 1**. Operating on more than one message at a time would reduce the cost of running this Lambda at AWS. Unclear how to signal partial success at this time.

Questions?

https://github.com/irods/irods_client_aws_lambda_s3

Thank You!

Pre-release testing environment provided by Bristol Myers Squibb.