
AWS
re:Invent
NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

ANT308

Seamless data sharing using Amazon Redshift

BP Yau

Sr Product Manager – Amazon Redshift
Amazon Web Services

Asser Moustafa

Sr Analytics Specialist Solutions Architect
Amazon Web Services

Agenda

Introduction

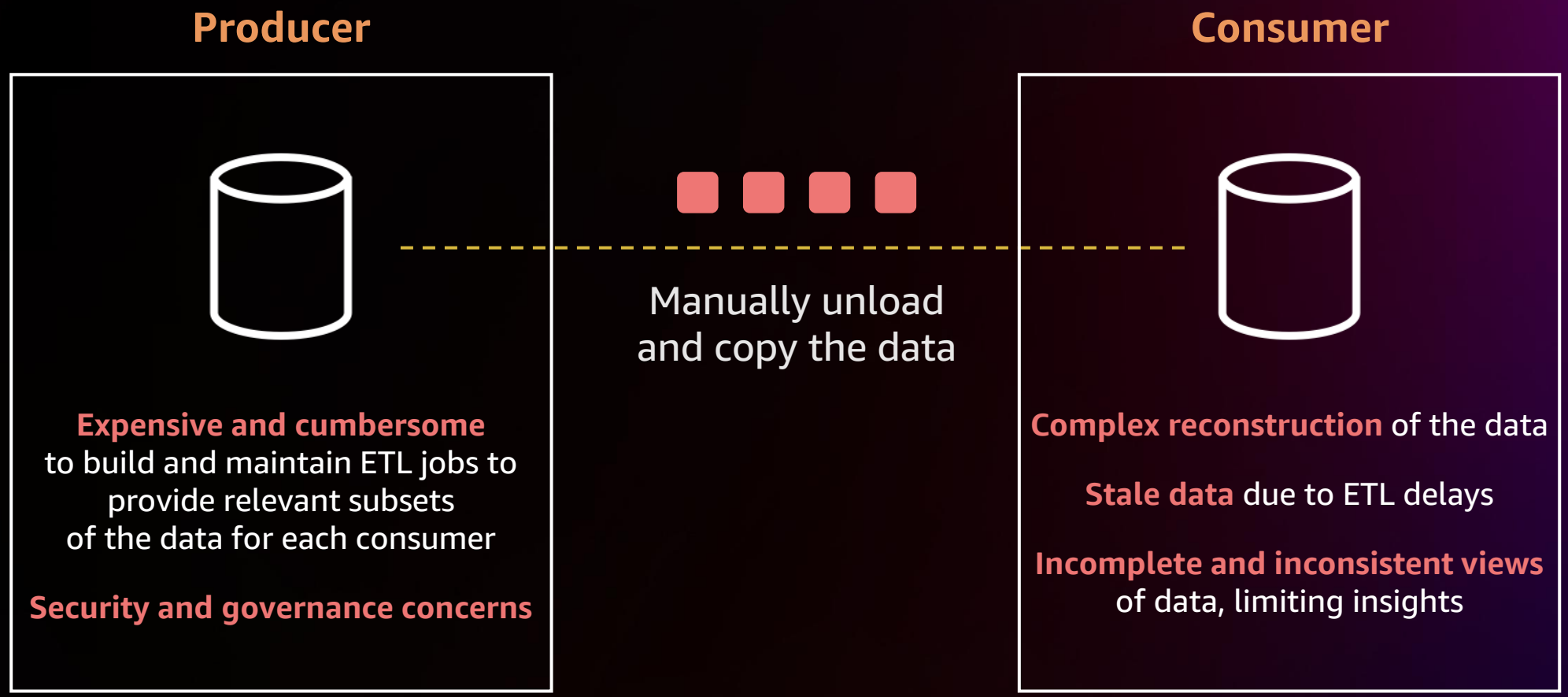
Demo

Workshop

Best practices and considerations

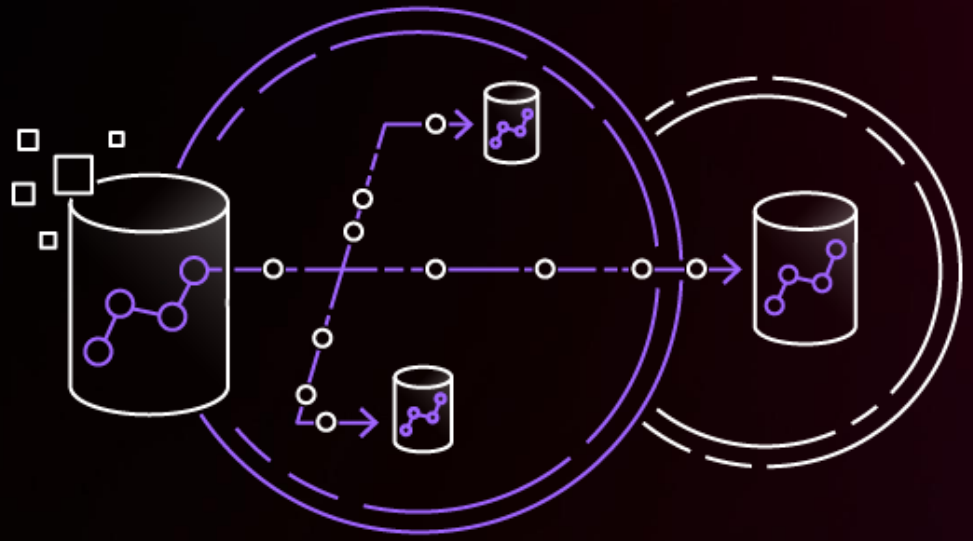
Introduction

Sharing data in organizations is complex and offers inconsistent views to users



Amazon Redshift data sharing

A **secure** and **easy** way to share **live** data across Amazon Redshift clusters **within the same or different AWS accounts and regions**



Data sharing builds on Amazon Redshift managed storage

HIGH PERFORMANCE DATA ACCESS WHILE PRESERVING WORKLOAD ISOLATION

Producer pays for Amazon Redshift managed storage and consumers pay for consumer cluster

Workloads accessing shared data are isolated from each other and the producer

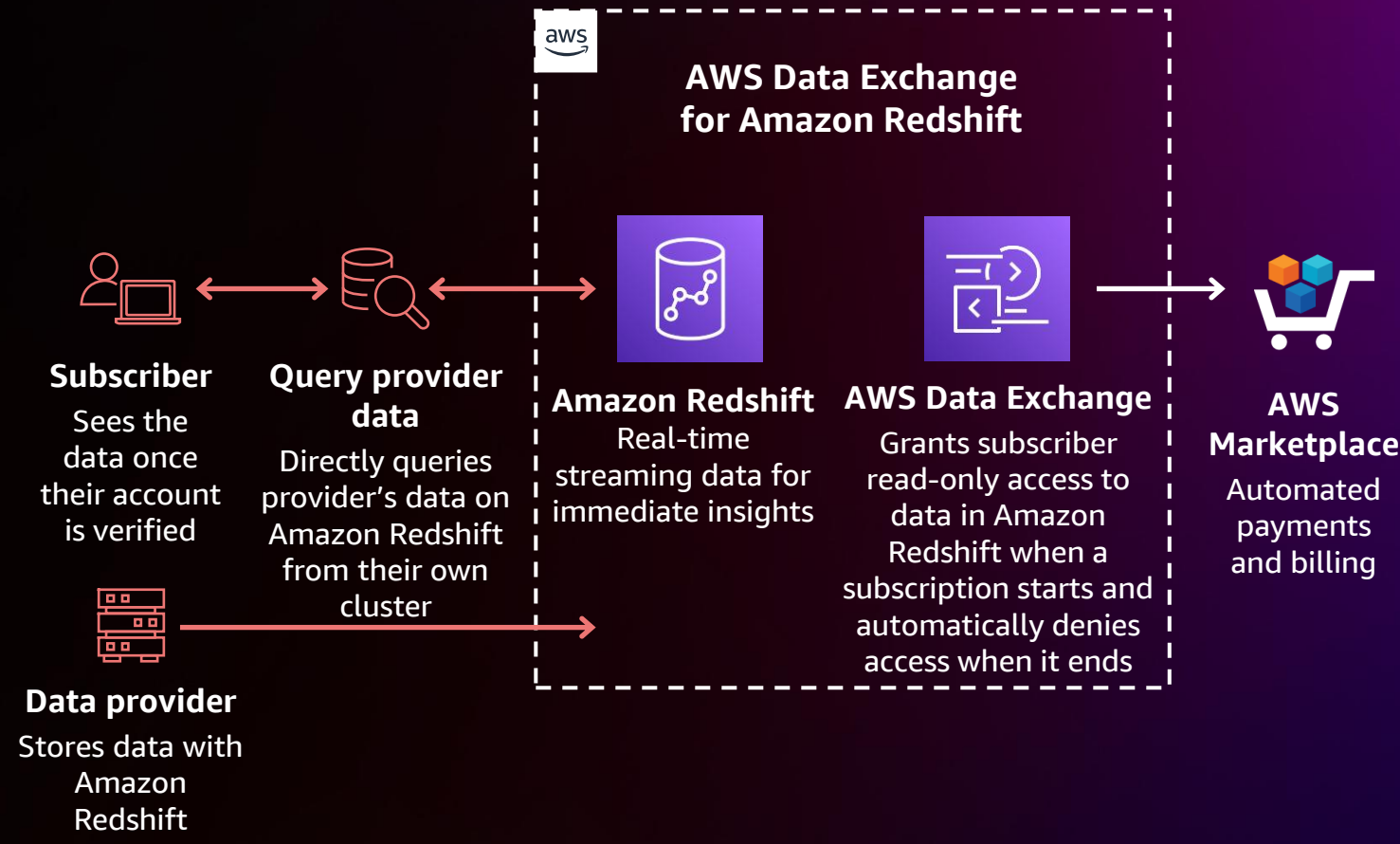
Find, subscribe to, and query third-party data

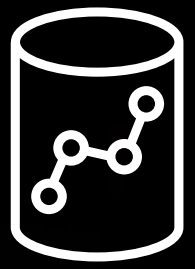
Data providers can package their Amazon Redshift data as data products, with pricing terms and conditions

Subscribers can search and subscribe to products

Get automated “live access” to the data in your Amazon Redshift cluster or data lake

Providers update their data directly at the source, without having to ETL or generate a new file





Amazon Redshift Serverless

Get insights from data in seconds without having to manage data warehouse infrastructure

Get insights from data more easily

Deliver consistently high performance

Optimize costs

YOU
focus on insights

Automatic scaling

Compute provisioning

Automated patching

Automatic failover

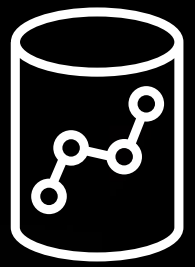
Advanced monitoring

Backup and recovery

Routine maintenance

Security and industry compliance

AWS
takes care of the rest



Amazon Redshift Serverless

Common use cases

Load and get started with querying

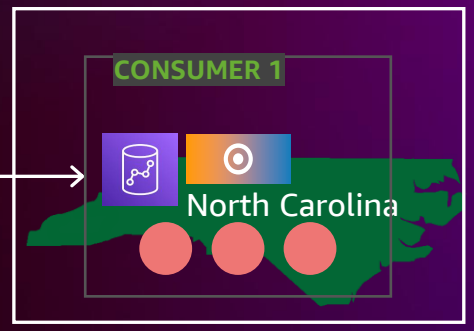
Variable and spiky workloads

Periodic workloads

Data sharing: Analytics as a service

HEALTHCARE ANALYTICS – IN-PATIENT BED OCCUPANCY DATA BY STATE

Amazon Redshift Serverless



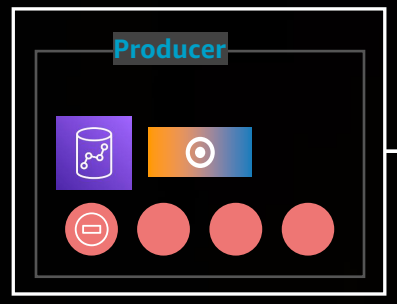
Consumer Account 2
State = “NC”

- Associate consumer Serverless endpoint
- Query data

Subscribers (Tenants)

state	name	namespace	Account
US	primary	Primary Cluster	
NC	tenant1	Consumer 1	AWSAccount1
VA	tenant2	Consumer 2	AWSAccount2

Healthcare Analytics Provider
Amazon Redshift Serverless



Producer Account 1

View: V_InpatientBedOccupancy

Create Datashare: Healthshare

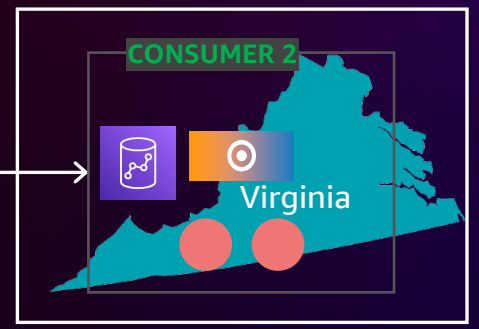
Add View:
V_InpatientBedOccupancy

- Grant usage to Consumer Account
- Approve access to consumer

InpatientBedOccupancy

state	collectiondate	Total_Inpatient_Beds
NC	11/21/20	4543
NC	5/5/20	43534
NC	3/8/20	2433
VA	1/23/20	3453
VA	11/3/20	23435
VA	10/1/20	98234

Amazon Redshift Serverless



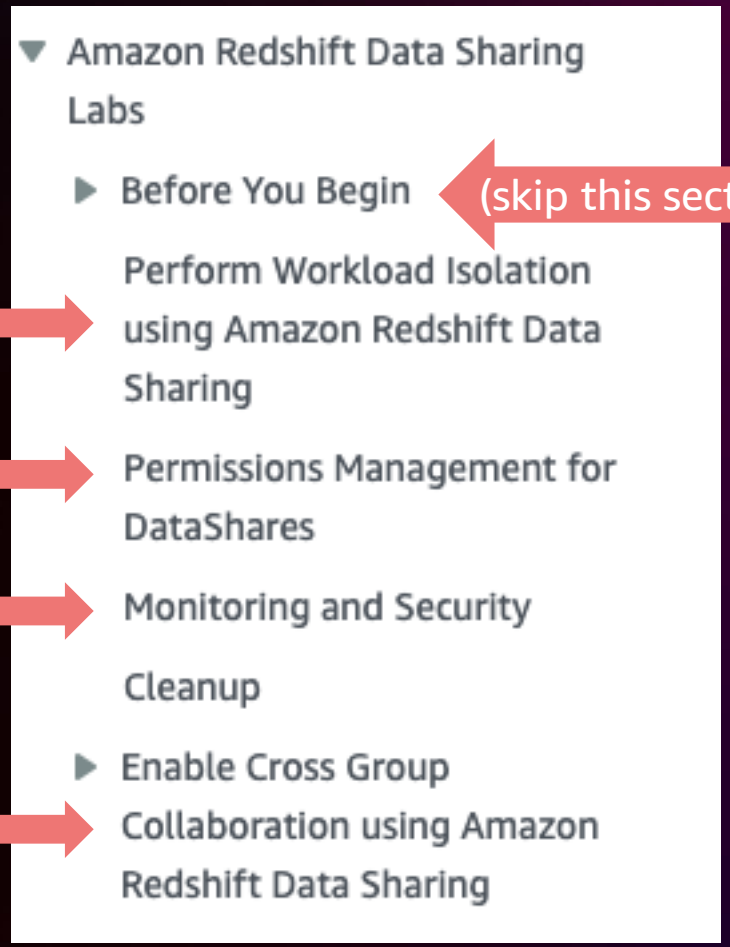
Consumer Account 3
State = “VA”

Workshop

Seamless data sharing using Amazon Redshift

Workshop goals

- Perform workload isolation for ETL and BI workloads
- Permissions management for data shares
- Monitoring and security of data shares
- [Optional] Enable cross group/account collaboration using Amazon Redshift data sharing



Getting started with this workshop

Pre-provisioned

As a participant, you will have access to an AWS account with any optional pre-provisioned infrastructure and IAM policies needed to complete this workshop

Temporary

The AWS account will be available only for the duration of this workshop; you will lose access to the account thereafter

Region-focused

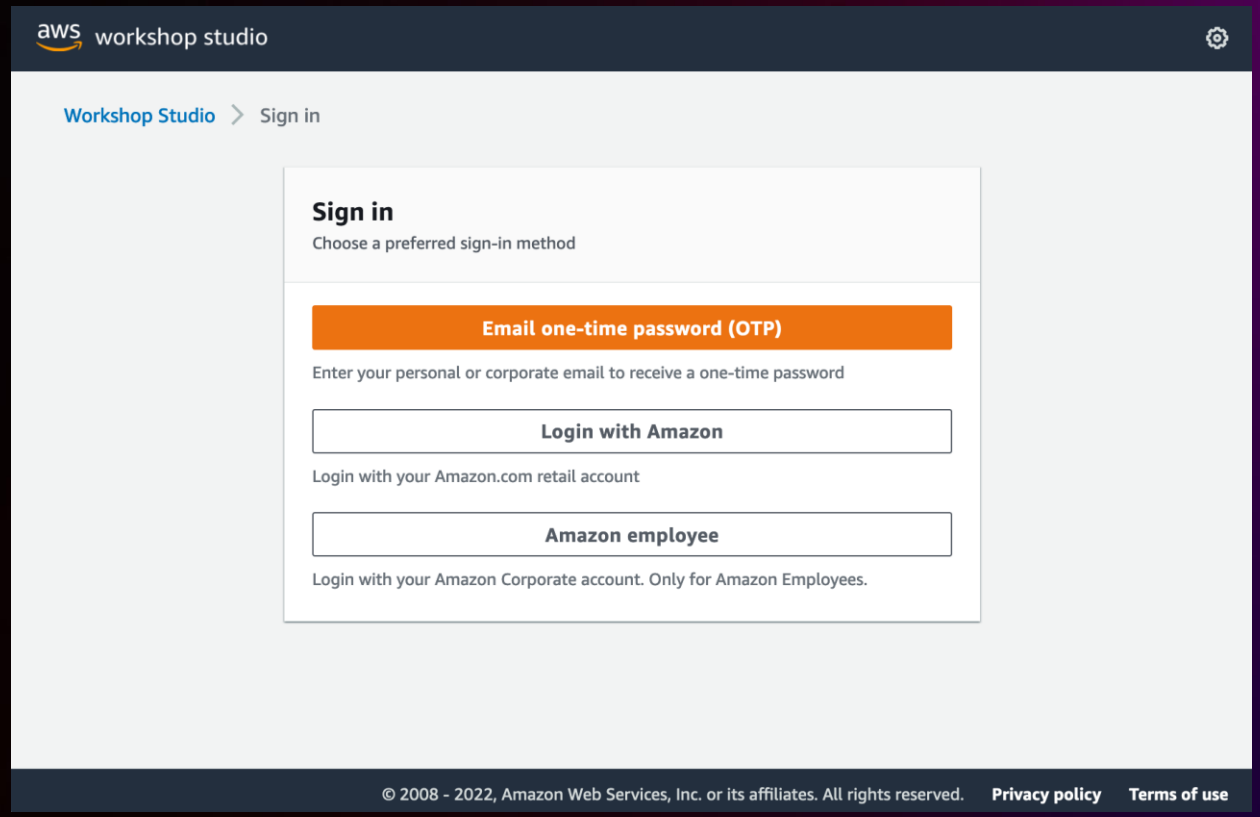
The optional pre-provisioned infrastructure will be deployed to a specific region; check your workshop content to determine whether other regions will be used

Terms and conditions

Be sure to review the terms and conditions of the event; do not upload any personal or confidential information in the account

Step 1: Sign in using your preferred method

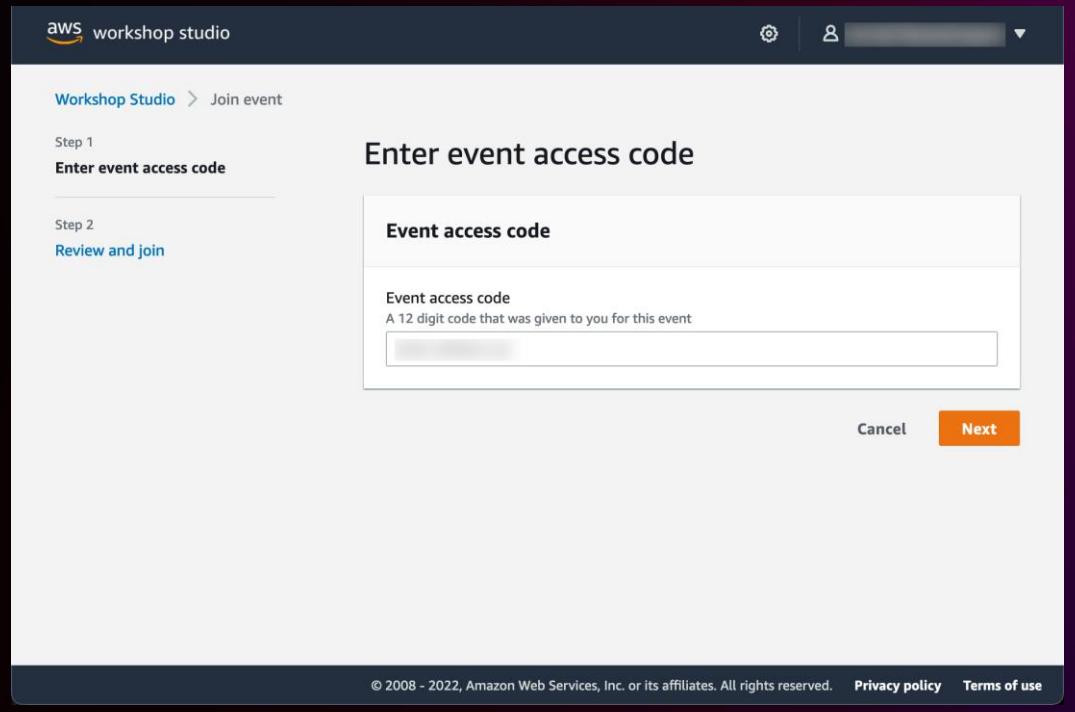
https://catalog.workshops.aws/join



Step 2: Enter event access code

ENTER THE 12-CHARACTER EVENT ACCESS CODE (IF YOU WERE GIVEN A ONE-CLICK JOIN LINK, YOU CAN SKIP THIS STEP)

https://catalog.workshops.aws/join

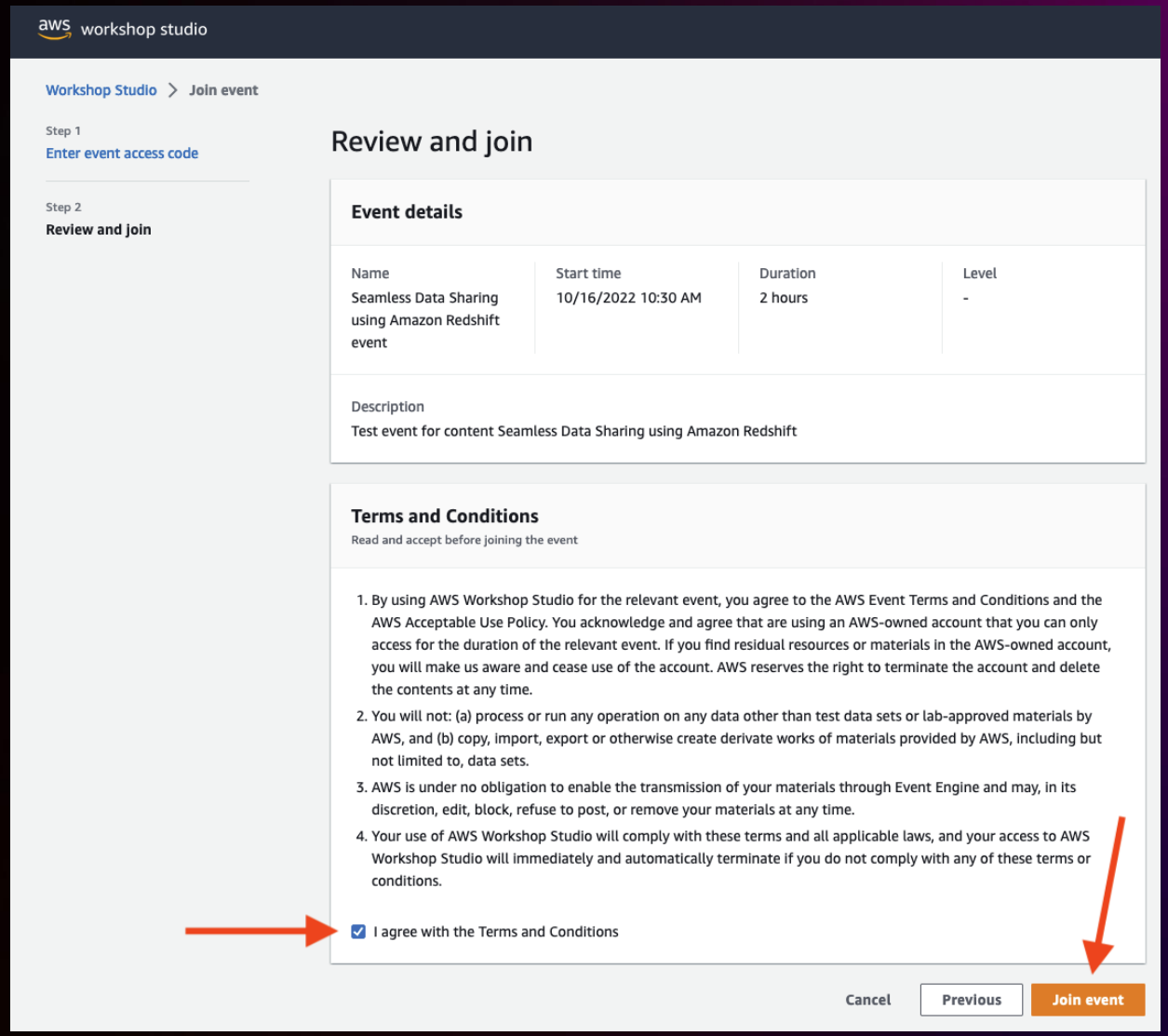


d29e-01211d-2c

Step 3: Review terms and join event

Event Access Code: d29e-01211d-2c

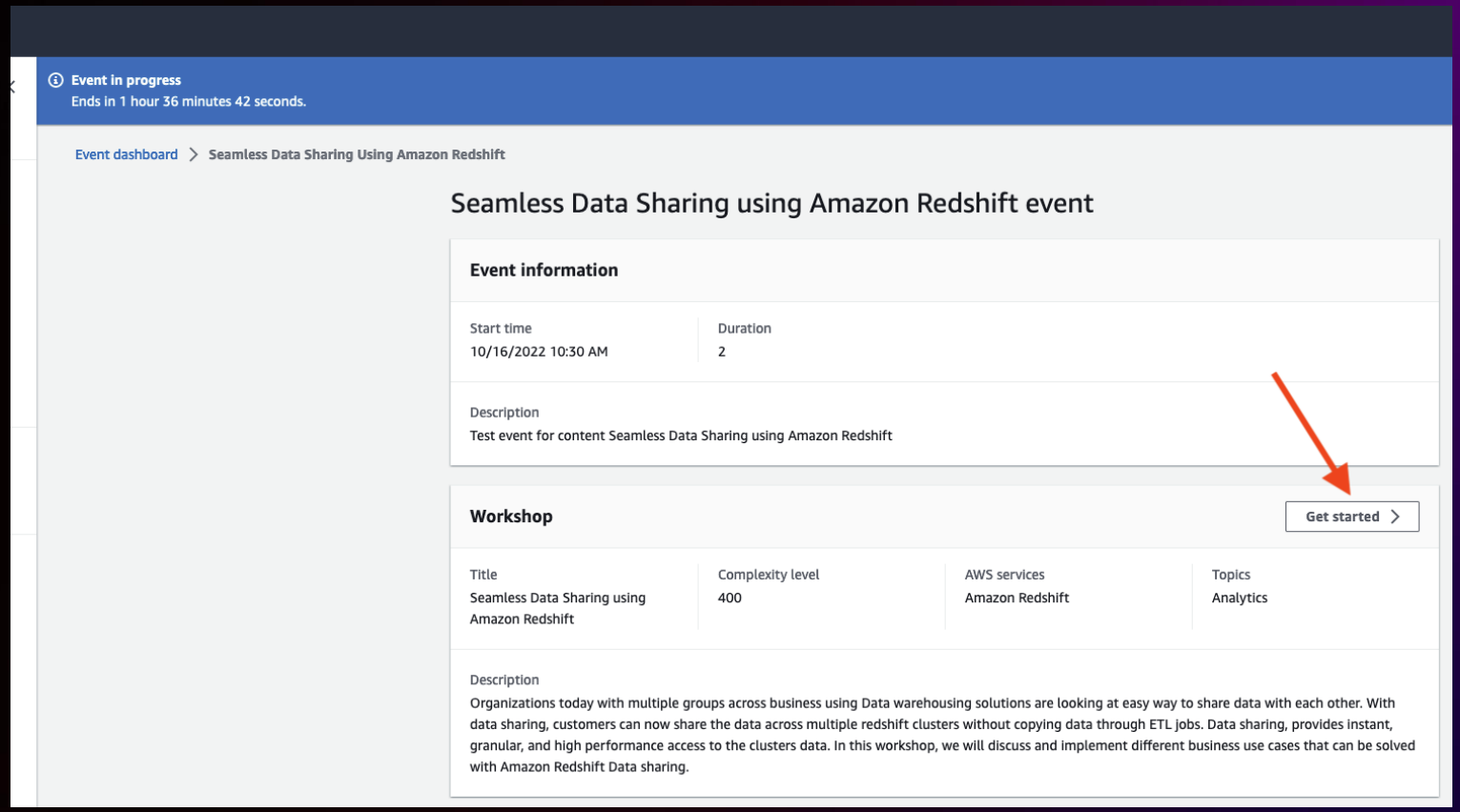
https://catalog.workshops.aws/join



Step 4: Get started with the workshop

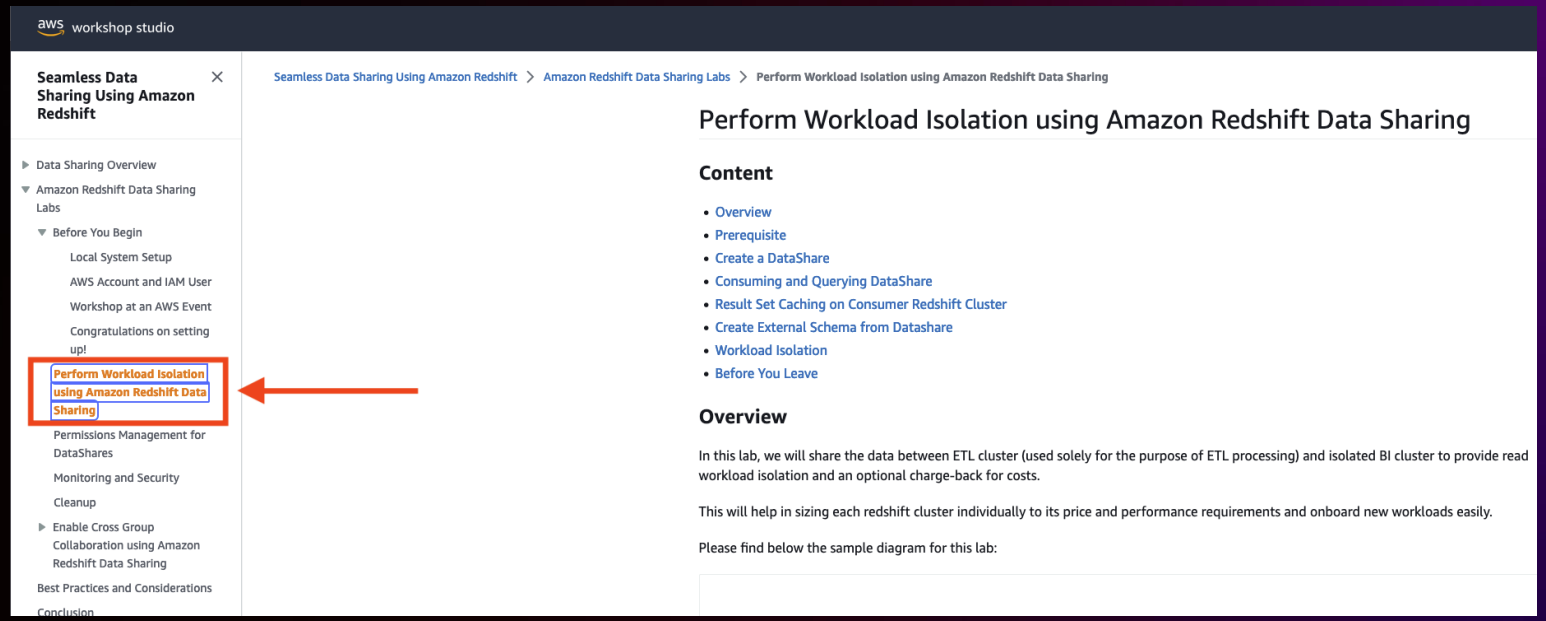
Event Access Code: d29e-01211d-2c

https://catalog.workshops.aws/join



Step 5: Access the workshop

Event Access Code: d29e-01211d-2c

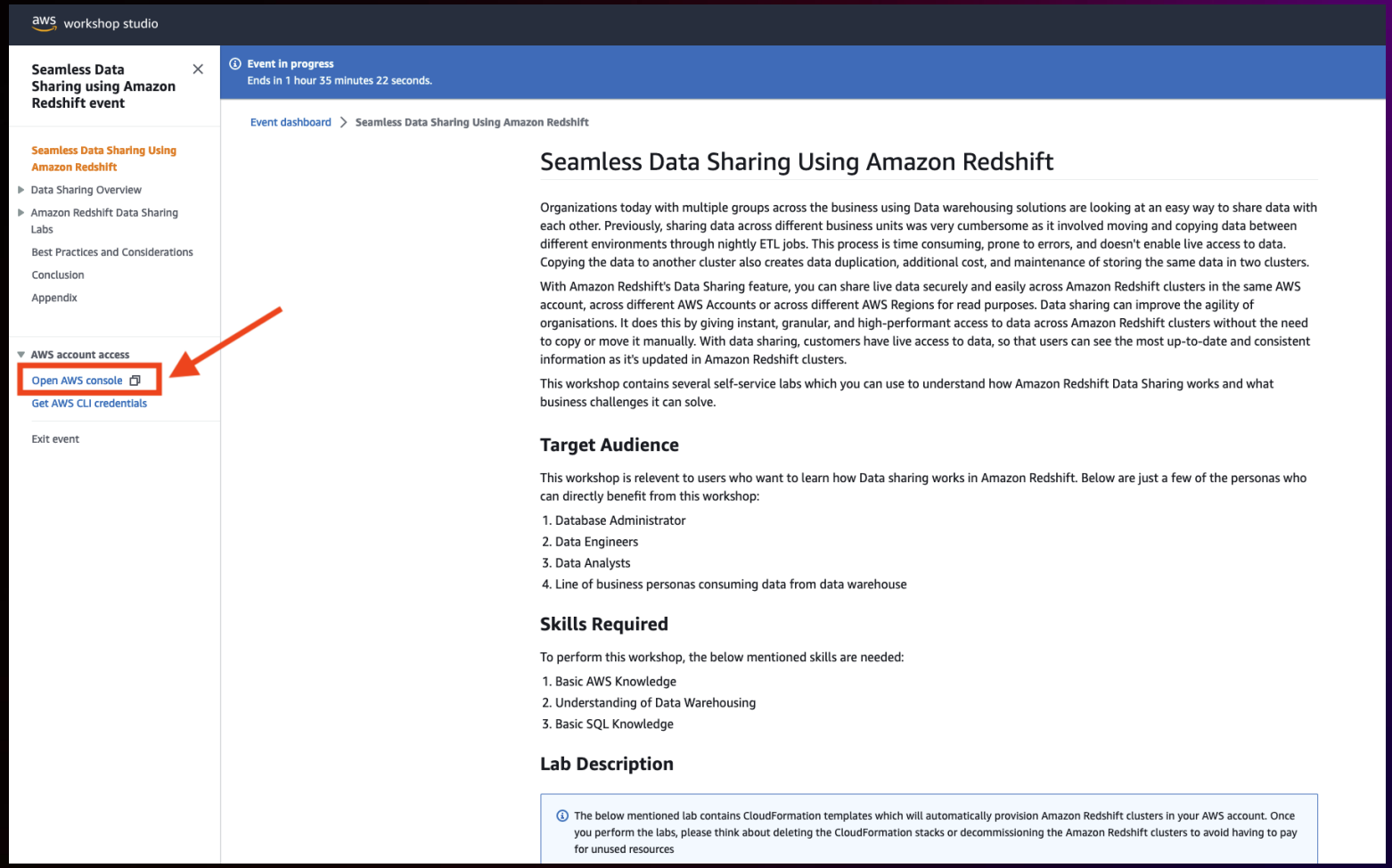


https://catalog.workshops.aws/join

Step 6: Access AWS account

Event Access Code: d29e-01211d-2c

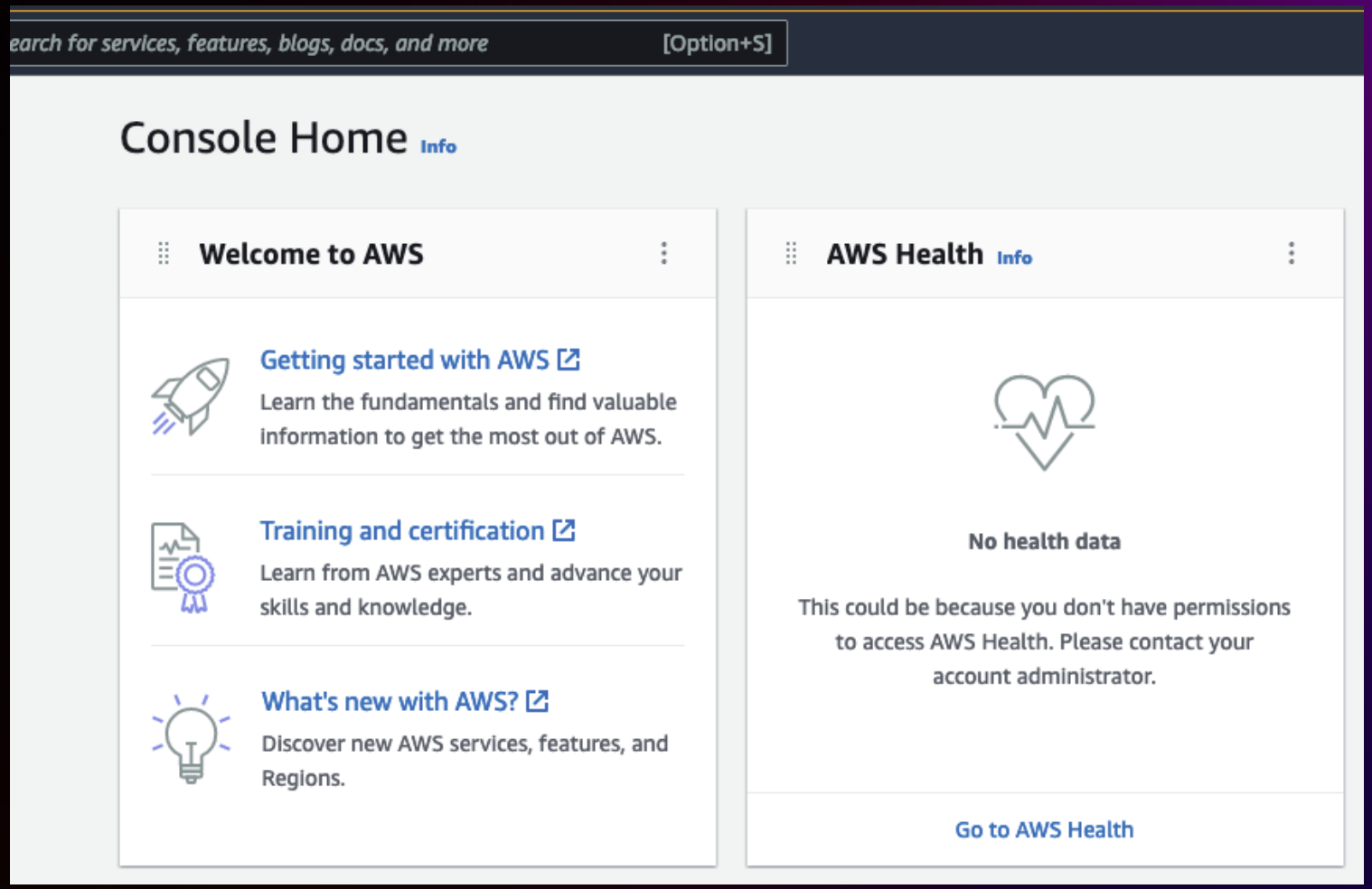
https://catalog.workshops.aws/join



Step 7: Access AWS console

Event Access Code: d29e-01211d-2c

https://catalog.workshops.aws/join



Best practices and considerations

Best practices/considerations

PERFORMANCE

Materialized view ownership

Producer

- Centralized management of MVs

Consumer

- Customized view on consumer based on its use case

Cross-region data sharing

- Data transfer cost on consumer – if cache, no data transfer
- Performance will vary compared to in-region due to network throughput

Table maintenance

If running vacuum manually, use vacuum re-cluster wherever possible (does not run merge and full delete), especially in use cases where:

- Large objects on producer where frequent maintenance is needed as a result of frequent DMLs
- Benefits the data sharing workloads by reducing the block metadata sync times

Best practices/considerations

SECURITY

INCLUDENEW datashare property

- When set to true, producer will automatically add future objects to an existing datashare, which helps reduce datashare objects management, but disabling it provides greater control over datashare objects

Fine-grained access control on consumer

- LBV or MV on consumer on shared objects
- Central access control with AWS Lake Formation– **NEW**

Audit data share usage and changes

- SVL_DATASHARE_CHANGE_LOG – Records the activity and usage of datashares on the consumer cluster (https://docs.aws.amazon.com/redshift/latest/dg/r_SVL_DATASHARE_CHANGE_LOG.html)
- SVL_DATASHARE_USAGE_CONSUMER – Records the activity and usage of datashares on the consumer cluster (https://docs.aws.amazon.com/redshift/latest/dg/r_SVL_DATASHARE_USAGE_CONSUMER.html)
- SVL_DATASHARE_USAGE_PRODUCER – Records the activity and usage of datashares on the producer cluster (https://docs.aws.amazon.com/redshift/latest/dg/r_SVL_DATASHARE_USAGE_PRODUCER.html)

Best practices/considerations

PRODUCER–CONSUMER DEPENDENCIES

Workload isolation, encryption, read consistency

- Queries on the consumer cluster will have no impact in terms of performance or activity on the producer cluster
- Data sharing works seamlessly with homogenous encryption configurations
- All the queries involving shared objects on the consumer cluster follow read-committed transaction consistency while checking for visible data for that transaction

Metadata access

Use restrictive filtering while querying the system views for metadata (such as *SVV_ALL_COLUMNS*, *SVV_ALL_SCHEMAS*, *SVV_ALL_TABLES*), for example, instead of the query

```
select * from svv_all_tables;
```

which will try to collect metadata for all the shared and local objects, making it very heavy in terms of metadata scans, especially for shared objects, use the following query to achieve a similar result

```
SELECT table_name, column_name, data_type FROM svv_all_tables WHERE table_name = < tablename > AND schema_name = < schemaname > AND database_name = < databasename > ORDER BY ordinal_position;
```

You can also use the *SVV_DATASHARE** system views to exclusively see shared object-related information

Best practices/considerations

COMPLIMENTARY FEATURES

Real-time analytics and data sharing with streaming data

- Streaming ingestion with KDS on producer
- Data exposed through MV and can be shared to consumer for real-time analytics

Concurrency scaling

- Concurrency scaling is supported for data sharing queries

Redshift Serverless integration

- The new serverless platform provides out-of-box data sharing support for both provisioned/serverless producers/consumers

AWS Data Exchange (ADX)

- Monetize dataset
- Data as service
- Public dataset without ETL

A few other notable complimentary features

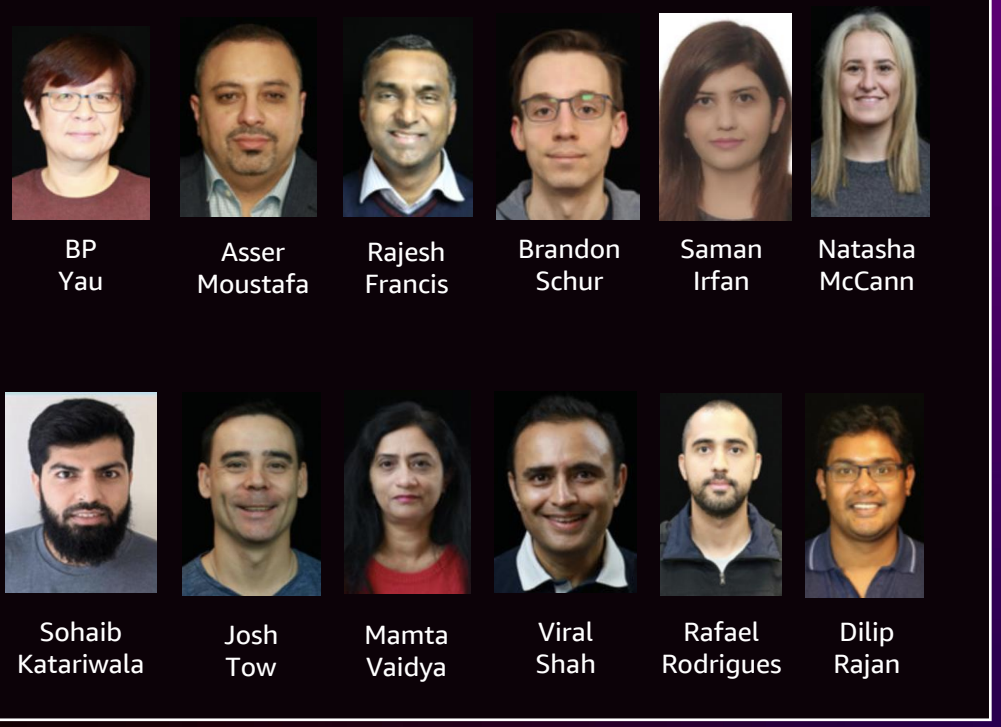
- Spectrum
- Federated query

Thank you!

BP Yau
boonyau@amazon.com

Asser Moustafa
aserm@amazon.com

Thank you from the entire workshop team for attending the workshop



Please complete the session survey in the **mobile app**