

# AWS re:Invent

NOV. 28 – DEC. 2, 2022 | LAS VEGAS, NV

# 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



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

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

## Producer



**Expensive and cumbersome** to build and maintain ETL jobs to provide relevant subsets of the data for each consumer

**Security and governance concerns**



Manually unload and copy the data

## Consumer



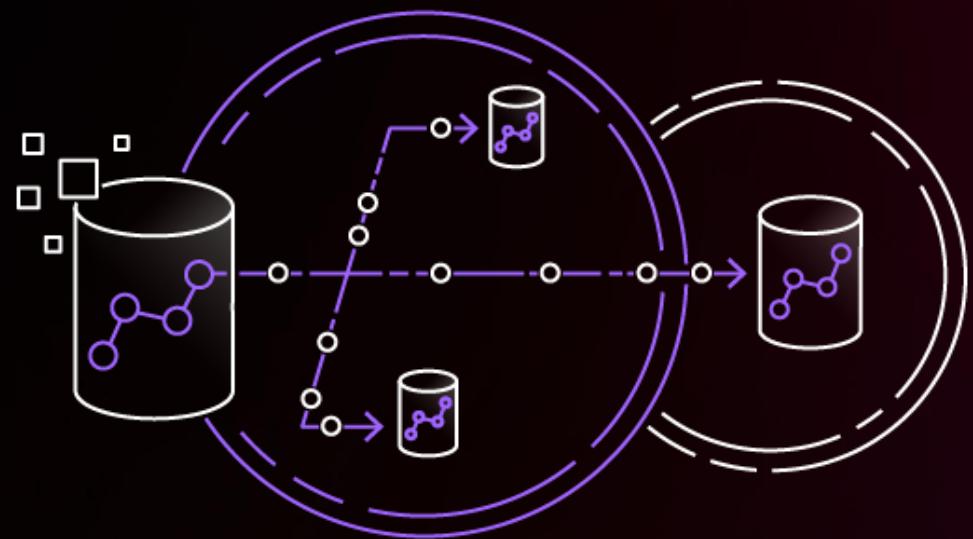
**Complex reconstruction** of the data

**Stale data** due to ETL delays

**Incomplete and inconsistent views** of data, limiting insights

# 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

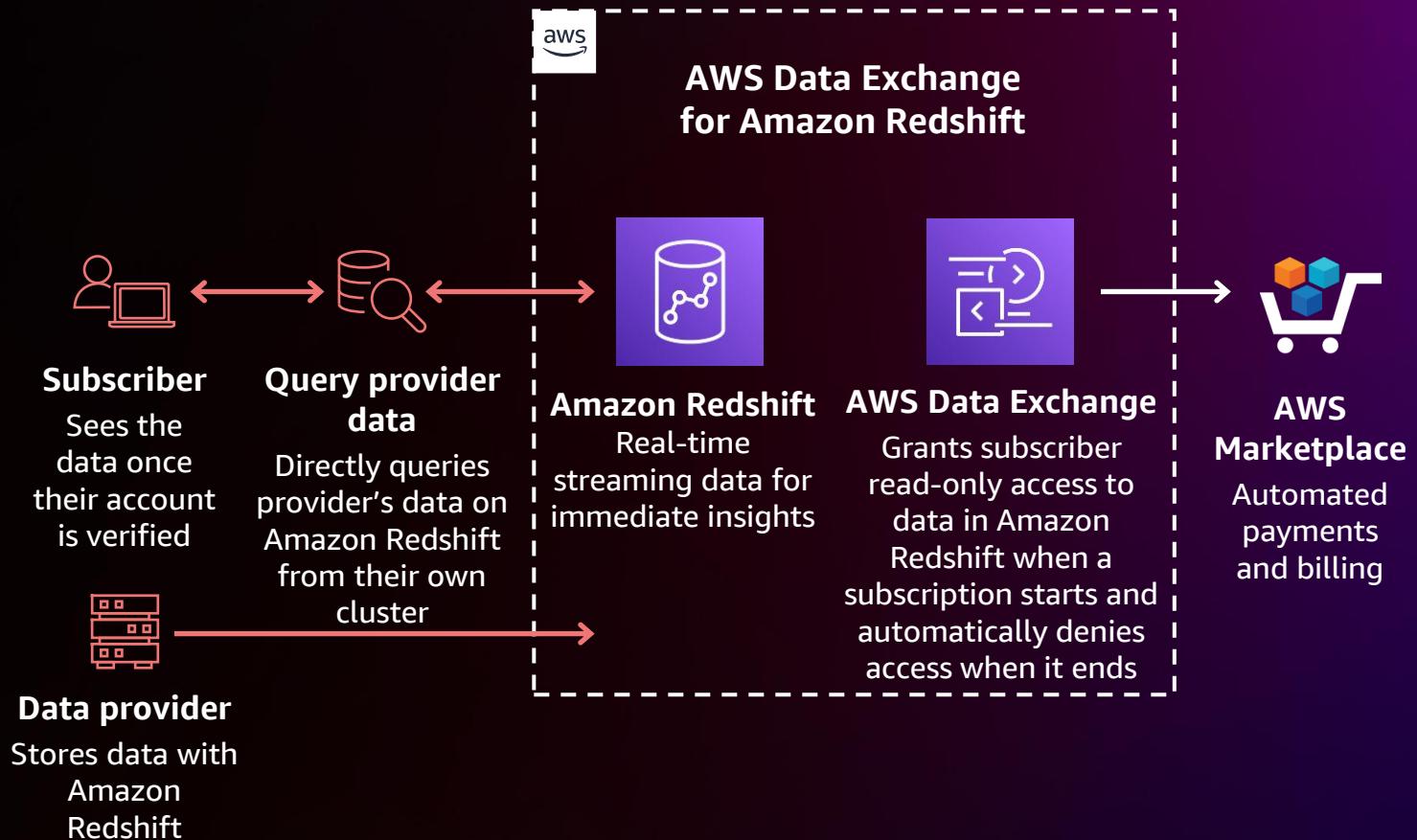
# Find, subscribe to, and query third-party data

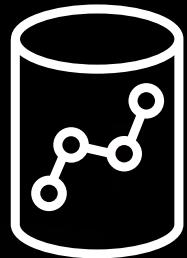
Data providers can package their Amazon Redshift data as data products, with pricing terms and conditions

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

Automatic scaling

Compute provisioning

Automated patching

Automatic failover

Advanced monitoring

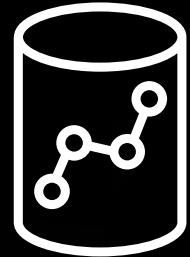
Backup and recovery

Routine maintenance

Security and industry compliance

**YOU**  
focus on  
insights

**AWS**  
takes care  
of the rest



# Amazon Redshift Serverless

Common use cases



Load and get started  
with querying



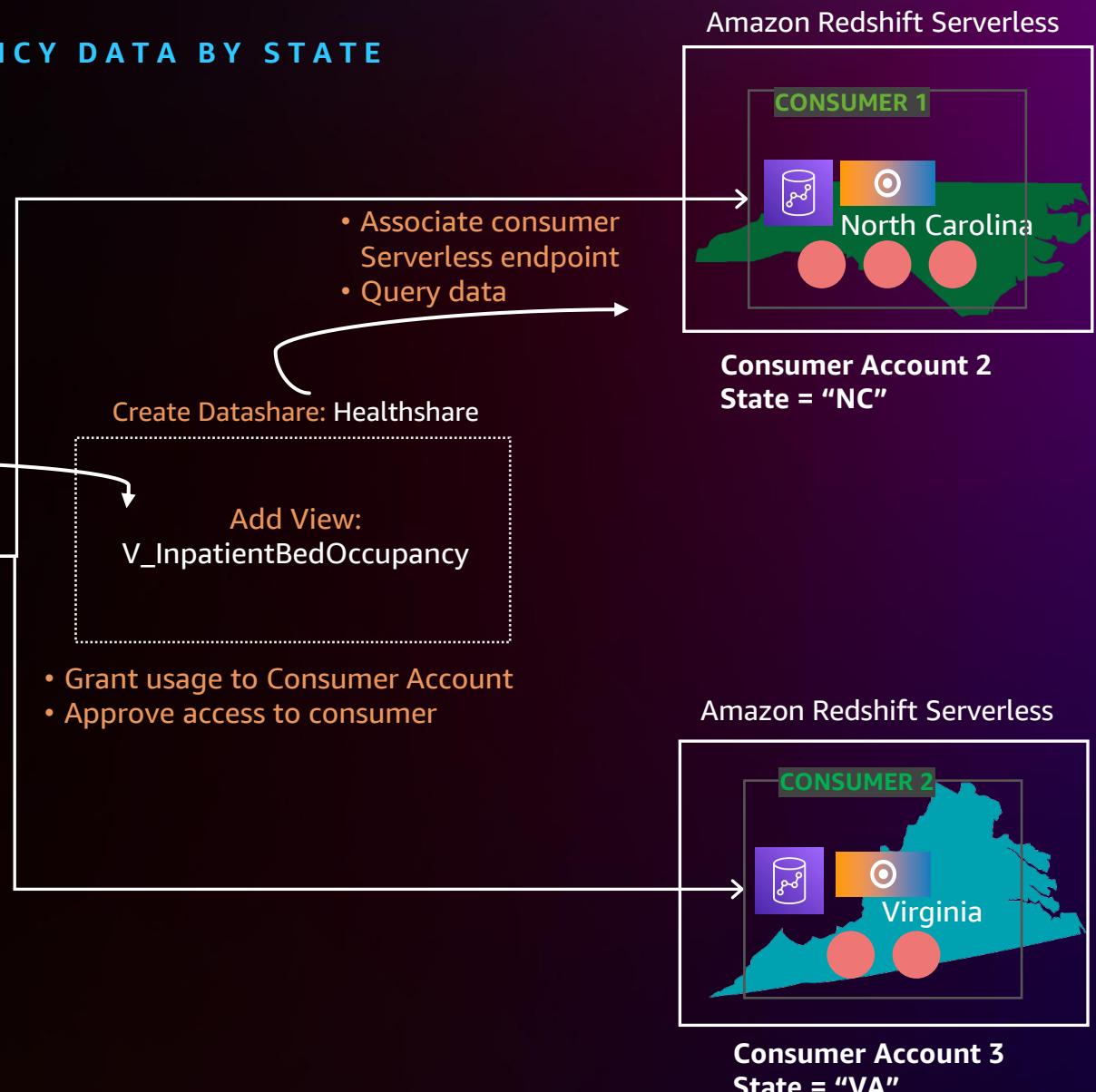
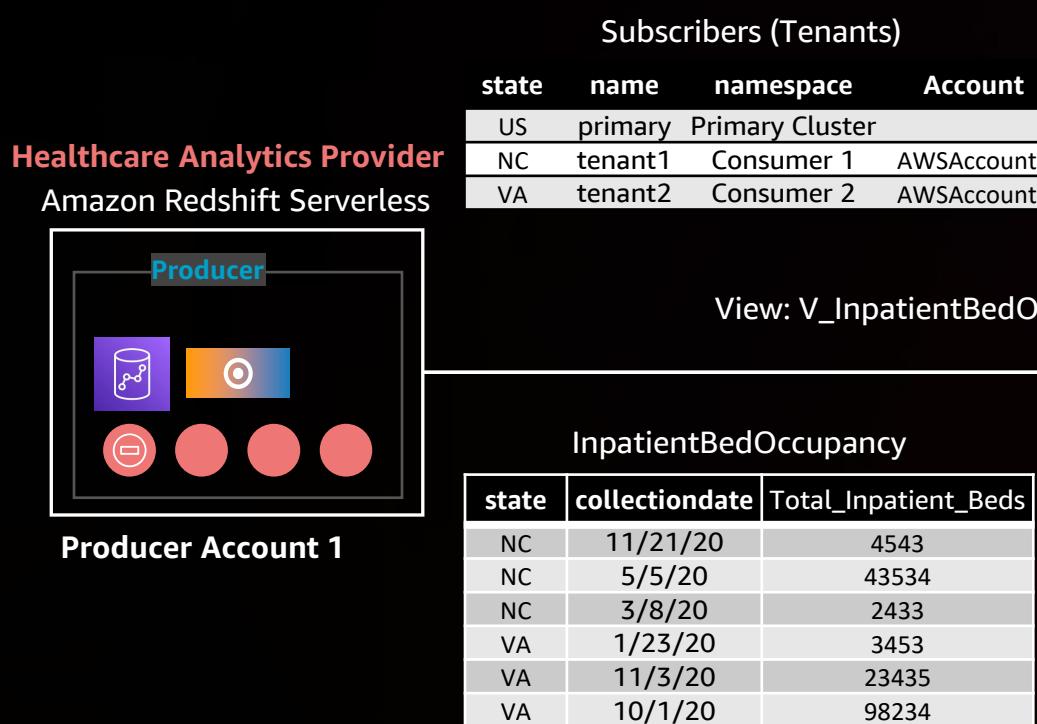
Variable and spiky  
workloads



Periodic workloads

# Data sharing: Analytics as a service

## HEALTHCARE ANALYTICS – IN-PATIENT BED OCCUPANCY DATA BY STATE



# Workshop



© 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.

# 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

▼ Amazon Redshift Data Sharing

Labs

► Before You Begin (skip this section for re:Invent)

► Perform Workload Isolation using Amazon Redshift Data Sharing

► Permissions Management for DataShares

► Monitoring and Security Cleanup

► Enable Cross Group 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>

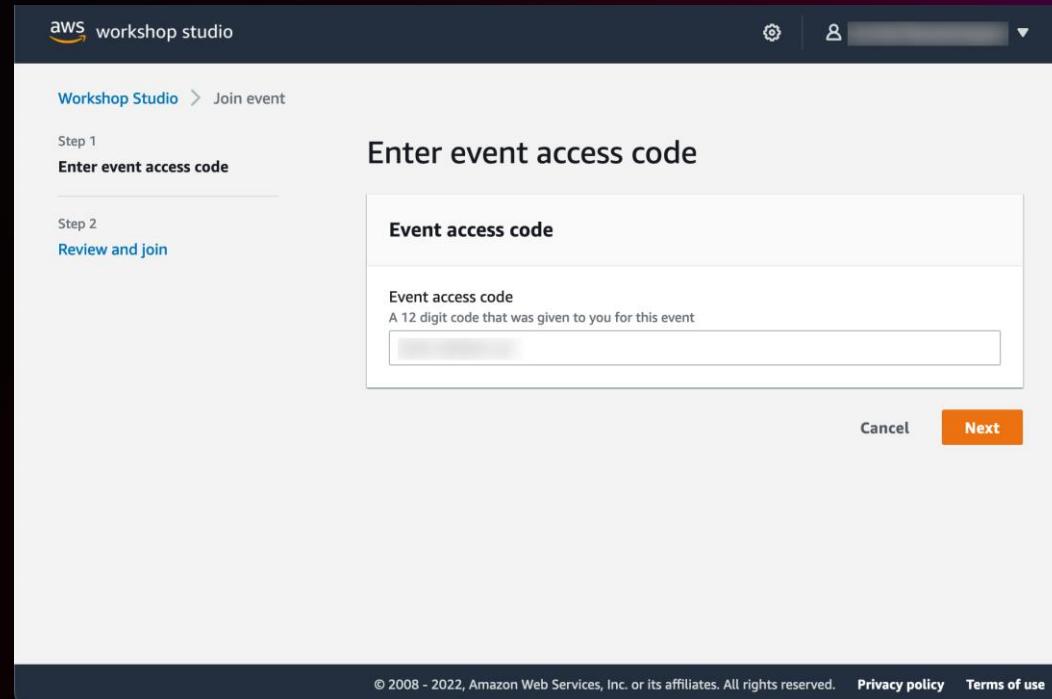
A screenshot of the AWS Workshop Studio sign-in interface. The top navigation bar includes the 'aws workshop studio' logo and a gear icon for settings. The main title 'Workshop Studio &gt; Sign in' is displayed. The sign-in section is titled 'Sign in' with the sub-instruction 'Choose a preferred sign-in method'. Three options are listed: 'Email one-time password (OTP)' (highlighted with an orange background), 'Login with Amazon' (with a light gray background), and 'Amazon employee' (with a light gray background). Each option has a corresponding descriptive text below it. At the bottom of the page, a footer bar contains the text '© 2008 - 2022, Amazon Web Services, Inc. or its affiliates. All rights reserved.' and links for 'Privacy policy' and 'Terms of use'.

# 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>

aws workshop studio

Workshop Studio > Join event

Step 1  
Enter event access code

Step 2  
Review and join

### Review and join

#### Event details

Name	Start time	Duration	Level
Seamless Data Sharing using Amazon Redshift event	10/16/2022 10:30 AM	2 hours	-

Description  
Test event for content Seamless Data Sharing using Amazon Redshift

#### Terms and Conditions

Read and accept before joining the event

1. By using AWS Workshop Studio for the relevant event, you agree to the AWS Event Terms and Conditions and the AWS Acceptable Use Policy. You acknowledge and agree that are using an AWS-owned account that you can only access for the duration of the relevant event. If you find residual resources or materials in the AWS-owned account, you will make us aware and cease use of the account. AWS reserves the right to terminate the account and delete the contents at any time.

2. You will not: (a) process or run any operation on any data other than test data sets or lab-approved materials by AWS, and (b) copy, import, export or otherwise create derivative works of materials provided by AWS, including but not limited to, data sets.

3. AWS is under no obligation to enable the transmission of your materials through Event Engine and may, in its discretion, edit, block, refuse to post, or remove your materials at any time.

4. Your use of AWS Workshop Studio will comply with these terms and all applicable laws, and your access to AWS Workshop Studio will immediately and automatically terminate if you do not comply with any of these terms or conditions.

I agree with the Terms and Conditions

Cancel Previous **Join event**

# Step 4: Get started with the workshop

Event Access Code: d29e-01211d-2c



<https://catalog.workshops.aws/join>

Event in progress  
Ends in 1 hour 36 minutes 42 seconds.

Event dashboard > Seamless Data Sharing Using Amazon Redshift

### Seamless Data Sharing using Amazon Redshift event

**Event information**

Start time	10/16/2022 10:30 AM	Duration	2
------------	---------------------	----------	---

Description  
Test event for content Seamless Data Sharing using Amazon Redshift

**Workshop**

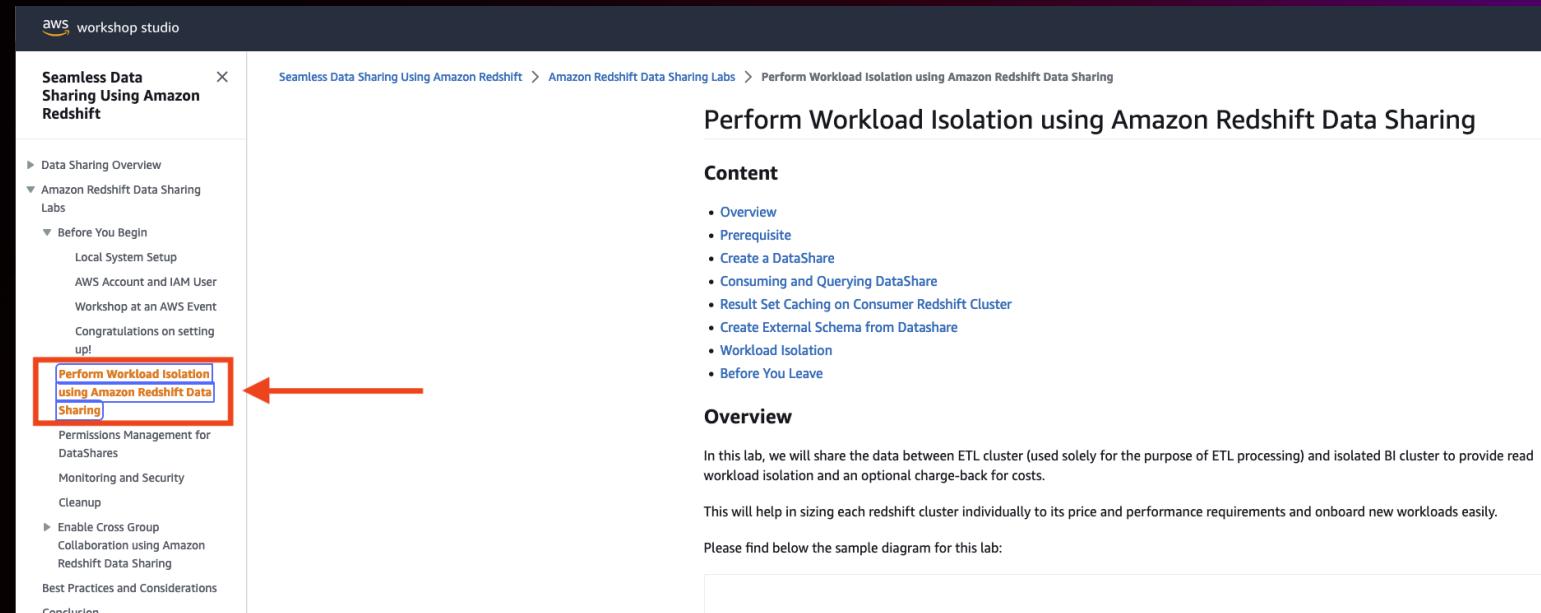
Title	Complexity level	AWS services	Topics
Seamless Data Sharing using Amazon Redshift	400	Amazon Redshift	Analytics

Description  
Organizations today with multiple groups across business using Data warehousing solutions are looking at easy way to share data with each other. With data sharing, customers can now share the data across multiple redshift clusters without copying data through ETL jobs. Data sharing, provides instant, granular, and high performance access to the clusters data. In this workshop, we will discuss and implement different business use cases that can be solved with Amazon Redshift Data sharing.

**Get started >**

# Step 5: Access the workshop

Event Access Code: d29e-01211d-2c



The screenshot shows the AWS Workshop Studio interface. The left sidebar has a tree view with the following structure:

- Seamless Data Sharing Using Amazon Redshift
- ▶ Data Sharing Overview
- ▼ Amazon Redshift Data Sharing Labs
  - ▼ Before You Begin
    - Local System Setup
    - AWS Account and IAM User
    - Workshop at an AWS Event
    - Congratulations on setting up!
  - Perform Workload Isolation using Amazon Redshift Data Sharing** (this item is highlighted with a red box and has a red arrow pointing to it from the left)
  - Permissions Management for DataShares
  - Monitoring and Security
  - Cleanup
- ▶ Enable Cross Group Collaboration using Amazon Redshift Data Sharing
- Best Practices and Considerations
- Conclusion

The main content area shows the title "Perform Workload Isolation using Amazon Redshift Data Sharing". Below it is a "Content" section with a bulleted list of topics:

- Overview
- Prerequisite
- Create a DataShare
- Consuming and Querying DataShare
- Result Set Caching on Consumer Redshift Cluster
- Create External Schema from Datasource
- Workload Isolation
- Before You Leave

Below the content is an "Overview" section with the following text:

In this lab, we will share the data between ETL cluster (used solely for the purpose of ETL processing) and isolated BI cluster to provide read workload isolation and an optional charge-back for costs.

This will help in sizing each redshift cluster individually to its price and performance requirements and onboard new workloads easily.

Please find below the sample diagram for this lab:

<https://catalog.workshops.aws/join>

# Step 6: Access AWS account

Event Access Code: d29e-01211d-2c



<https://catalog.workshops.aws/join>

aws workshop studio

Seamless Data Sharing using Amazon Redshift event

Event in progress  
Ends in 1 hour 35 minutes 22 seconds.

Event dashboard > Seamless Data Sharing Using Amazon Redshift

### Seamless Data Sharing Using Amazon Redshift

Organizations today with multiple groups across the business using Data warehousing solutions are looking at an easy way to share data with each other. Previously, sharing data across different business units was very cumbersome as it involved moving and copying data between different environments through nightly ETL jobs. This process is time consuming, prone to errors, and doesn't enable live access to data. Copying the data to another cluster also creates data duplication, additional cost, and maintenance of storing the same data in two clusters. With Amazon Redshift's Data Sharing feature, you can share live data securely and easily across Amazon Redshift clusters in the same AWS account, across different AWS Accounts or across different AWS Regions for read purposes. Data sharing can improve the agility of organisations. It does this by giving instant, granular, and high-performant access to data across Amazon Redshift clusters without the need to copy or move it manually. With data sharing, customers have live access to data, so that users can see the most up-to-date and consistent information as it's updated in Amazon Redshift clusters.

This workshop contains several self-service labs which you can use to understand how Amazon Redshift Data Sharing works and what business challenges it can solve.

#### Target Audience

This workshop is relevant to users who want to learn how Data sharing works in Amazon Redshift. Below are just a few of the personas who can directly benefit from this workshop:

1. Database Administrator
2. Data Engineers
3. Data Analysts
4. Line of business personas consuming data from data warehouse

#### Skills Required

To perform this workshop, the below mentioned skills are needed:

1. Basic AWS Knowledge
2. Understanding of Data Warehousing
3. Basic SQL Knowledge

#### Lab Description

The below mentioned lab contains CloudFormation templates which will automatically provision Amazon Redshift clusters in your AWS account. Once you perform the labs, please think about deleting the CloudFormation stacks or decommissioning the Amazon Redshift clusters to avoid having to pay for unused resources

AWS account access

Open AWS console

Get AWS CLI credentials

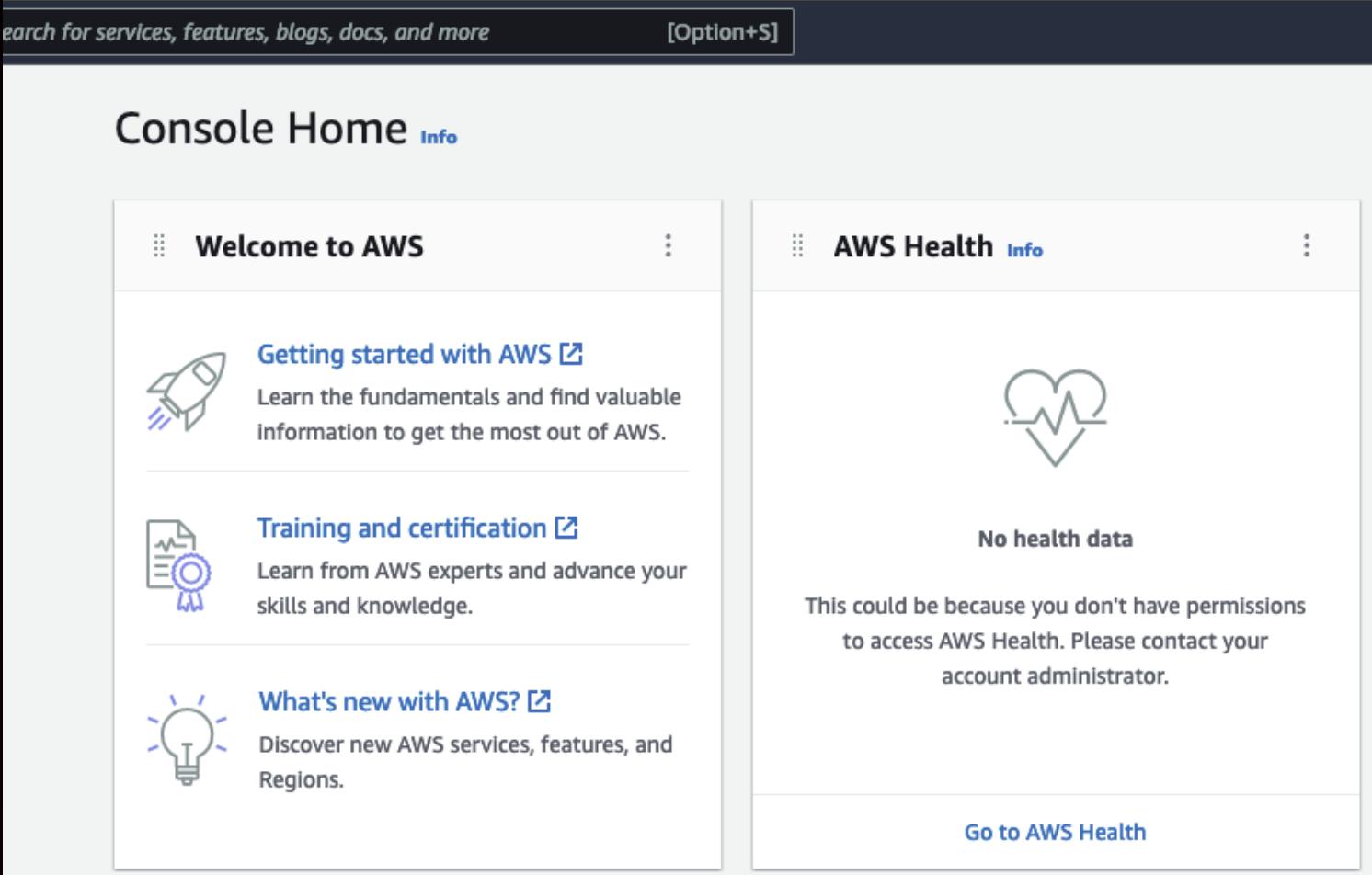
Exit event

# Step 7: Access AWS console

Event Access Code: d29e-01211d-2c



<https://catalog.workshops.aws/join>



The screenshot shows the AWS Console Home page. At the top, there is a search bar with the placeholder text "Search for services, features, blogs, docs, and more" and a keyboard shortcut "[Option+S]". Below the search bar, the page title "Console Home" is displayed with an "Info" link. The main content area is divided into three sections: "Welcome to AWS", "AWS Health", and "AWS Marketplace".

- Welcome to AWS**
  - Getting started with AWS**  [Info](#)  
Learn the fundamentals and find valuable information to get the most out of AWS.
  - Training and certification**  [Info](#)  
Learn from AWS experts and advance your skills and knowledge.
  - What's new with AWS?**  [Info](#)  
Discover new AWS services, features, and Regions.
- AWS Health**  [Info](#)
  - No health data**
  - This could be because you don't have permissions to access AWS Health. Please contact your account administrator.
  - [Go to AWS Health](#)
- AWS Marketplace**  [Info](#)
  - No products**
  - This could be because you don't have permissions to access AWS Marketplace. Please contact your account administrator.
  - [Go to AWS Marketplace](#)

# 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](https://docs.aws.amazon.com/redshift/latest/dg/r_SVL_DATASHARE_CHANGE_LOG.html) – 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](https://docs.aws.amazon.com/redshift/latest/dg/r_SVL_DATASHARE_CHANGE_LOG.html))
- [SVL\\_DATASHARE\\_USAGE\\_CONSUMER](https://docs.aws.amazon.com/redshift/latest/dg/r_SVL_DATASHARE_USAGE_CONSUMER.html) – 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](https://docs.aws.amazon.com/redshift/latest/dg/r_SVL_DATASHARE_USAGE_CONSUMER.html))
- [SVL\\_DATASHARE\\_USAGE\\_PRODUCER](https://docs.aws.amazon.com/redshift/latest/dg/r_SVL_DATASHARE_USAGE_PRODUCER.html) – 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](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](mailto:boonyau@amazon.com)

Asser Moustafa

[aserm@amazon.com](mailto:aserm@amazon.com)

Thank you from the entire workshop team for attending the workshop



BP  
Yau

Asser  
Moustafa

Rajesh  
Francis

Brandon  
Schur

Saman  
Irfan

Natasha  
McCann



Sohaib  
Katariwala

Josh  
Tow

Mamta  
Vaidya

Viral  
Shah

Rafael  
Rodrigues

Dilip  
Rajan



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