AWS
re:Invent

Rapidly evaluate AWS analytics solutions with Amazon Redshift

Zach Christopherson

Database Engineer – Redshift
Amazon Web Services

Brandon Schur

Database Engineer – Redshift
Amazon Web Services

AWS re:Invent

aws

Workshop agenda

- Five 5-minute micro-presentations

- Five 10-to-15-minute hands-on labs

- Introduce toolkit capabilities

- Explore JupyterLab environment

- Demonstrate modern Amazon Redshift

- Recap and next steps

Workshop staff

Zach

Brandon

Joe

Juan

Faraz

Junpei

Cristina

Feature evaluation and adoption challenges

- Comprehensive product evaluation is time consuming

- Cutting corners often results in poor data points on key criteria

- Forgoing relevant features impacts optimal price/performance

- Knowing which features will benefit you, and how to best implement them

Features

Effort/time

Lab environment
orientation

Benchmarking
managed datasets

Configure intelligent
workload management

Data lake
integration

Benchmarking highly
concurrent workloads

Amazon Redshift workshop toolkit

- Simplifies and accelerates onboarding new workloads, migrations, feature and product evaluations

- Provides AWS-approved architectures, configurations, content, and workload patterns

- Publicly available assets and service-generated content based on customer engagements
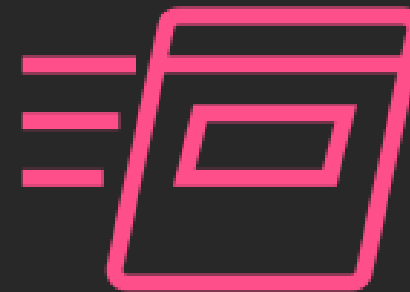
Architectures
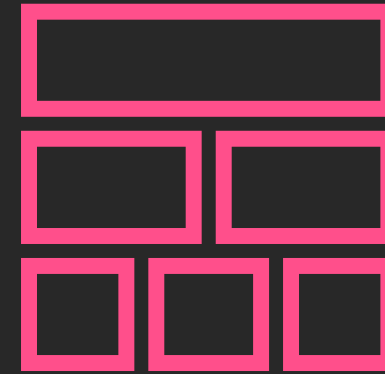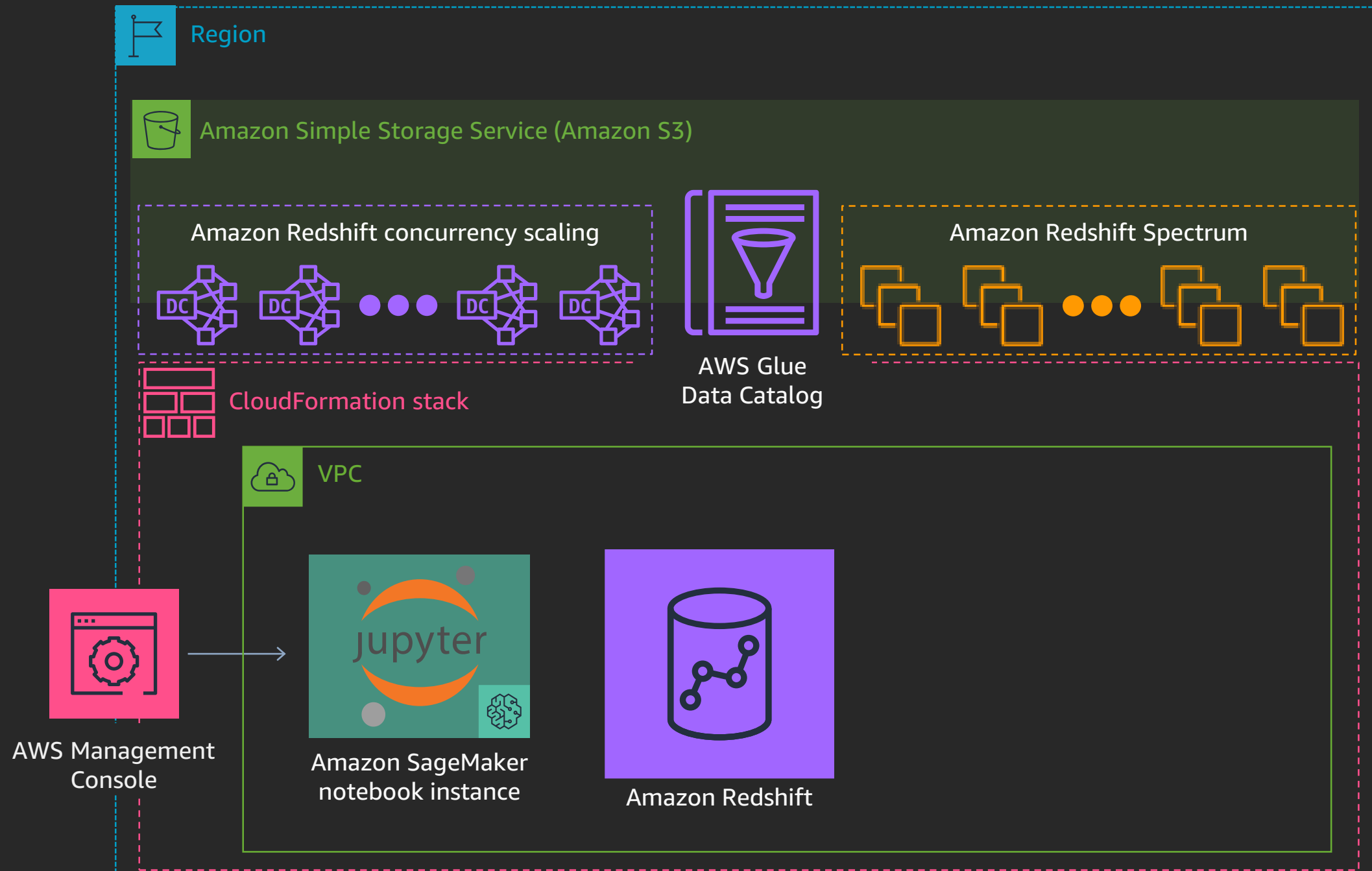
Datasets

Evaluation tools

Demonstrations

Architecture and lab environment

Lab architecture

AWS Event Engine

https://dashboard.eventengine.run

The AWS Event Engine was created to help AWS field teams run Workshops, GameDays, Bootcamps, Immersion Days, and other events that require hands-on access to AWS accounts.

Event Engine Login

ARC101-R2 Introduction to AWS

Team Hash: a1b2c3d4e5f6

Go to: https://dashboard.eventengine.run
Please contact your event operator if you have any questions!

AWS
re:Invent

https://dashboard.eventengine.run

https://dashboard.eventengine.run/dashboard

Demonstrations

Confidence **Ease of use** **Effort/Time**

- Notebook-style presentation

- Widget-based actions (click to start)

- Real-time visualizations

- Comprehensive documentation

- Feature guides

- Common workloads and use cases
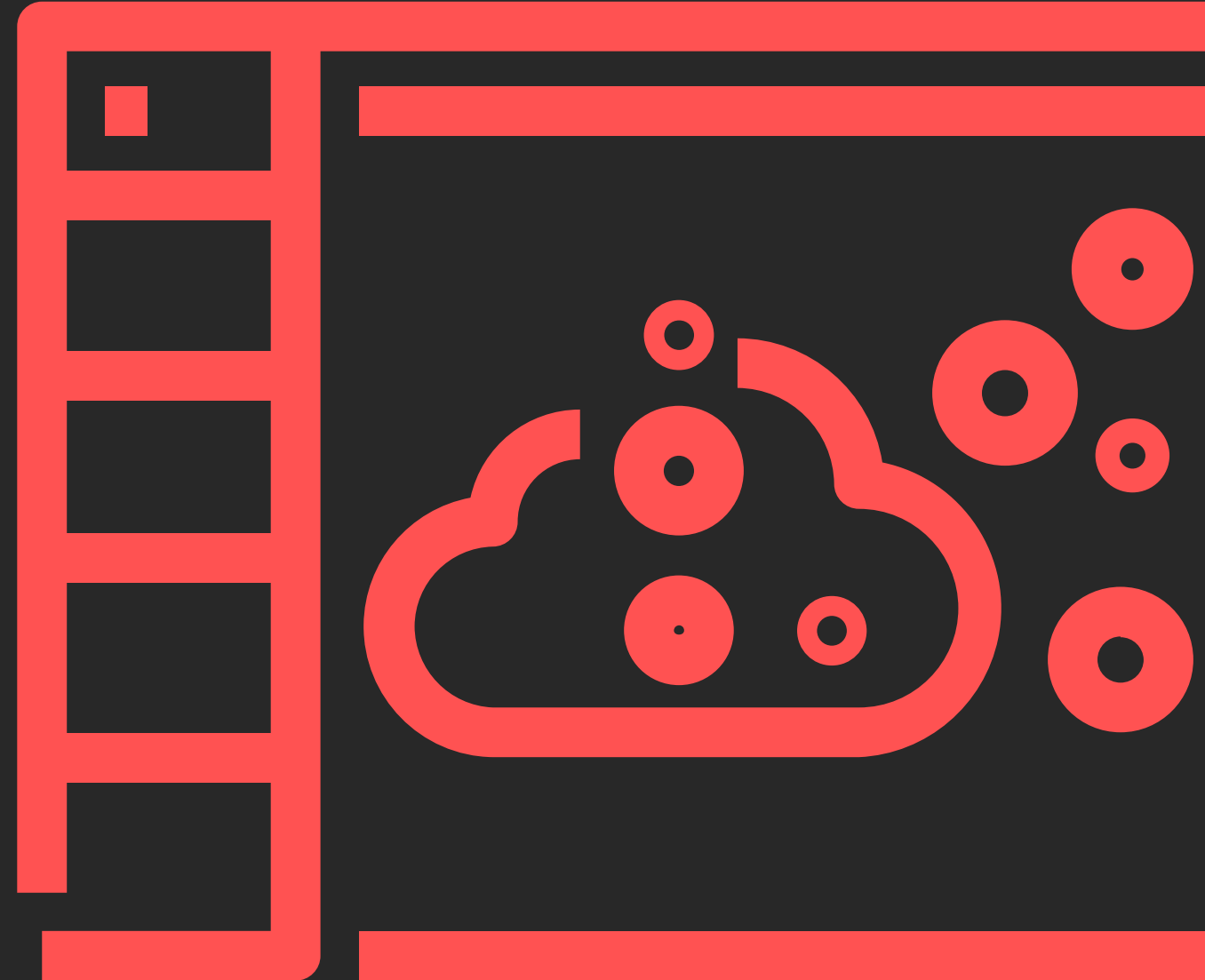
Lab #1 – Lab environment orientation

- 10 minutes

- Login at https://dashboard.eventengine.run

- Follow Event Engine README instructions

- Access Amazon SageMaker hosted JupyterLab

- Complete Lab1.ipynb

Lab #1: Lab environment orientation

Recap Lab #1 – Lab environment onboarding

- Lab guide tips for success

- Toolkit configuration details

- Interactive spatial analysis widgets

- Demonstrations as Jupyter notebooks

- Toolkit metadata and assets

AIM413 – Deep dive on Project Jupyter
Thursday, Dec 5, 1:45 PM - 2:45 PM – MGM, Level 3, Premier Ballroom 313

Benchmarking managed datasets

Datasets

Solution TCO   **Confidence**   **Ease of use**   **Effort/time**

- Native Amazon Redshift and Amazon Redshift Spectrum ready

- Preloaded or one-click load

- Industry standard (TPC-DS[†], TPC-H[†])

- Optimized and OOB table DDL

- Scale factors (1 GB → 1 PB)

- Open datasets:

  - Amazon product reviews

  - Wikipedia clickstream

[†]Datasets and workloads are inspired by the TPC standard

Evaluation tools

Benchmark utilities
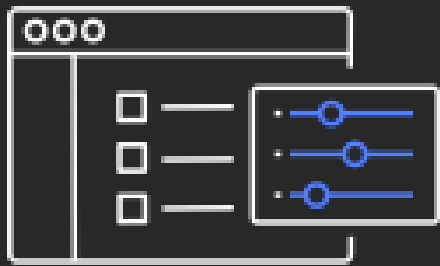
- Configurable clients
- Pre-/post-run validation
- Serialized and scheduled execution
- Dependency management
- Cluster modifications

Interpreted results

- Real-time feedback
- Proven summarizations
- Automatic analysis
- Persisted raw and summary
- Amazon SNS and Amazon CloudWatch Events

Solution TCO

Confidence

Ease of use

Effort/time

Lab #2 – Benchmarking managed datasets

- 15 minutes

- Locate Lab2.ipynb

- Perform TPC-DS benchmark tasks

- Compare real-time and final results

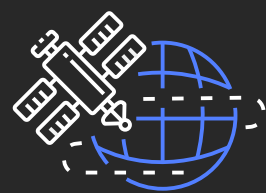# Lab #2: Benchmarking managed datasets

aws

Recap Lab #2 – Benchmarking managed datasets

- Robust evaluation summaries

- Community assets (perf matrices)

- Benchmark client as a building block

- Where was the time spent?

Workload management

Automatic workload management

Dynamic memory allocation

ML classifier

Adaptive concurrency

Improves throughput
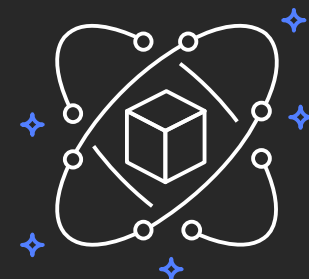
Query prioritization

Importance-based execution

Dashboarding →

Reporting →

Ad hoc/
exploratory →

	Standard WLM	Automatic WLM + Priorities
Dashboarding	80 QpH	224 QpH
Reporting	80 QpH	28 QpH
Exploratory	80 QpH	7 QpH

Lab #3 – Intelligent workload management

- 15 minutes

- Locate Lab3.ipynb

- Perform benchmark tasks and configuration changes

- Compare to previous runs, real-time, and final results

Lab #3: Enhanced workload management

Recap Lab #3 – Intelligent workload management
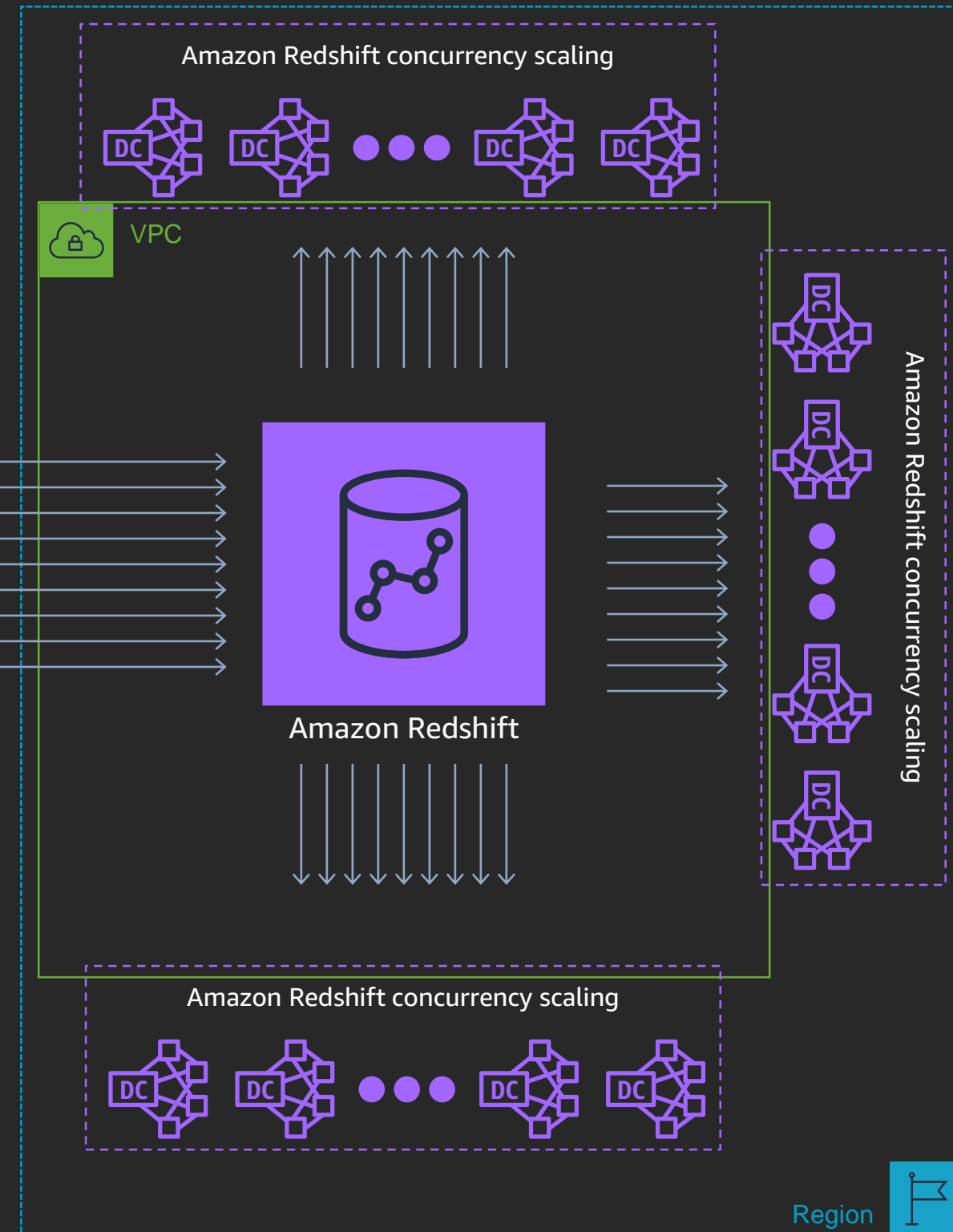
- Throughput and latency changes

- Identifying where the time went

- How can we tell if this is best for you?

Highly concurrent workloads

Elastic solutions for elastic workload

- Handles unpredictable and non-uniform workloads

- Eliminates the need to over-provision for peak activity

- Free credits accumulate at one hour per day

- Per-second billing only when queries are active

Consistent throughput at concurrency

Improving Scalability

Throughput (Queries per Hour) vs Concurrent Users

- 12,414 QpH @ 210 CUs
- 354 QpH @ 5 CUs
- 206 QpH @ 5 CUs
- Cloud DW 3TB DC2.8XL

Legend:
- Jan'19
- Apr'19
- Aug'19
- Nov'19
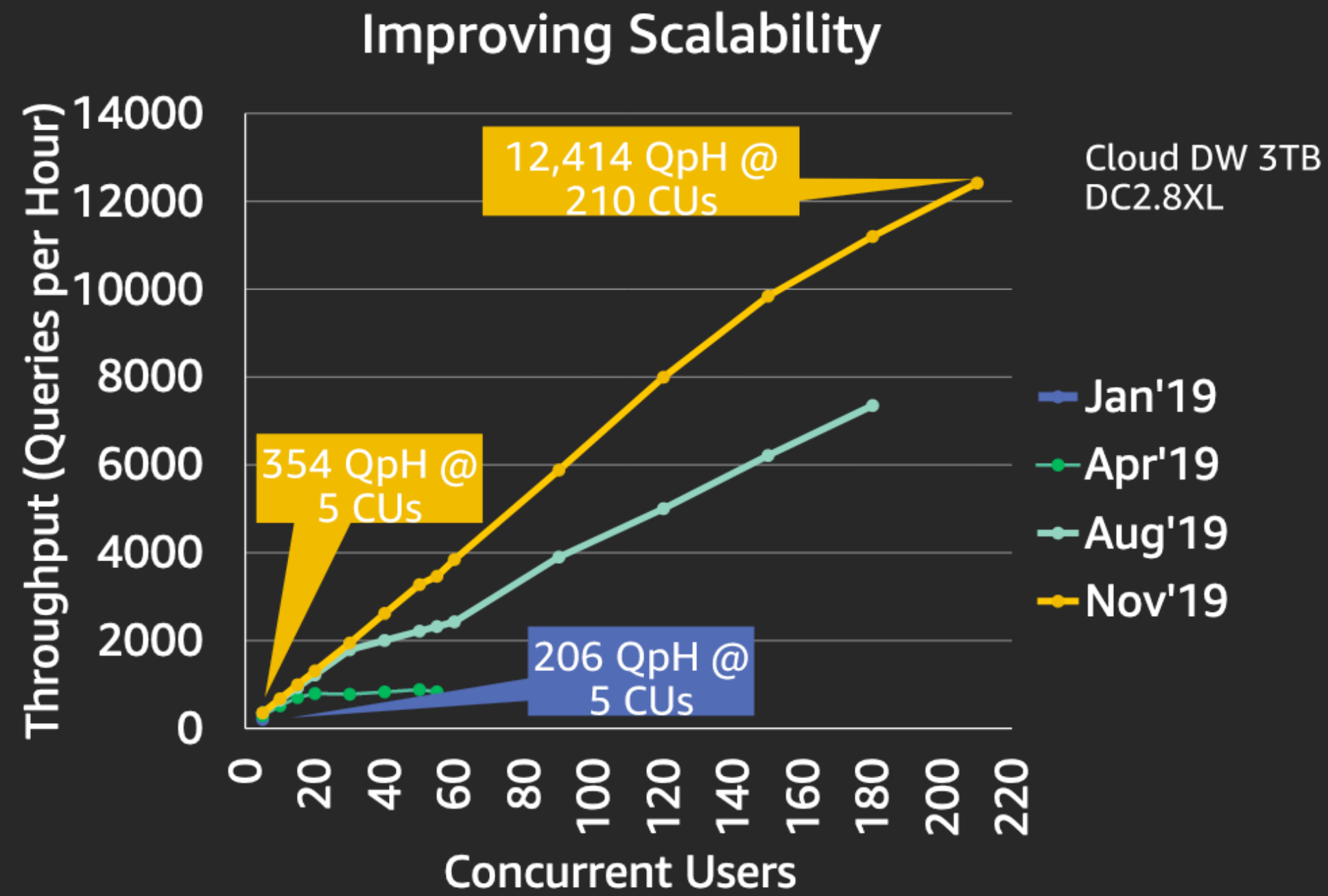
Lab #4 – Highly concurrent workloads

- 15 minutes

- Locate Lab4.ipynb

- Perform benchmark tasks with configuration changes

- Compare Lab#3

Lab #4: Highly concurrent workloads

Recap Lab #4 – Concurrency scaling

- System throughput and query latency

- Staying in the free tier

- Controlling cost and usage patterns

Accessing your data lake

AWS re:Invent

Data lake access with Amazon Redshift Spectrum

- High bandwidth

  - Parallelism (many small straws)

- Optimized data scans

  - Avoids unnecessary work

  - Columnar formats and compression

  - Computation push-down

- Nested data access for flexibility and JOIN elimination

- Common open file formats
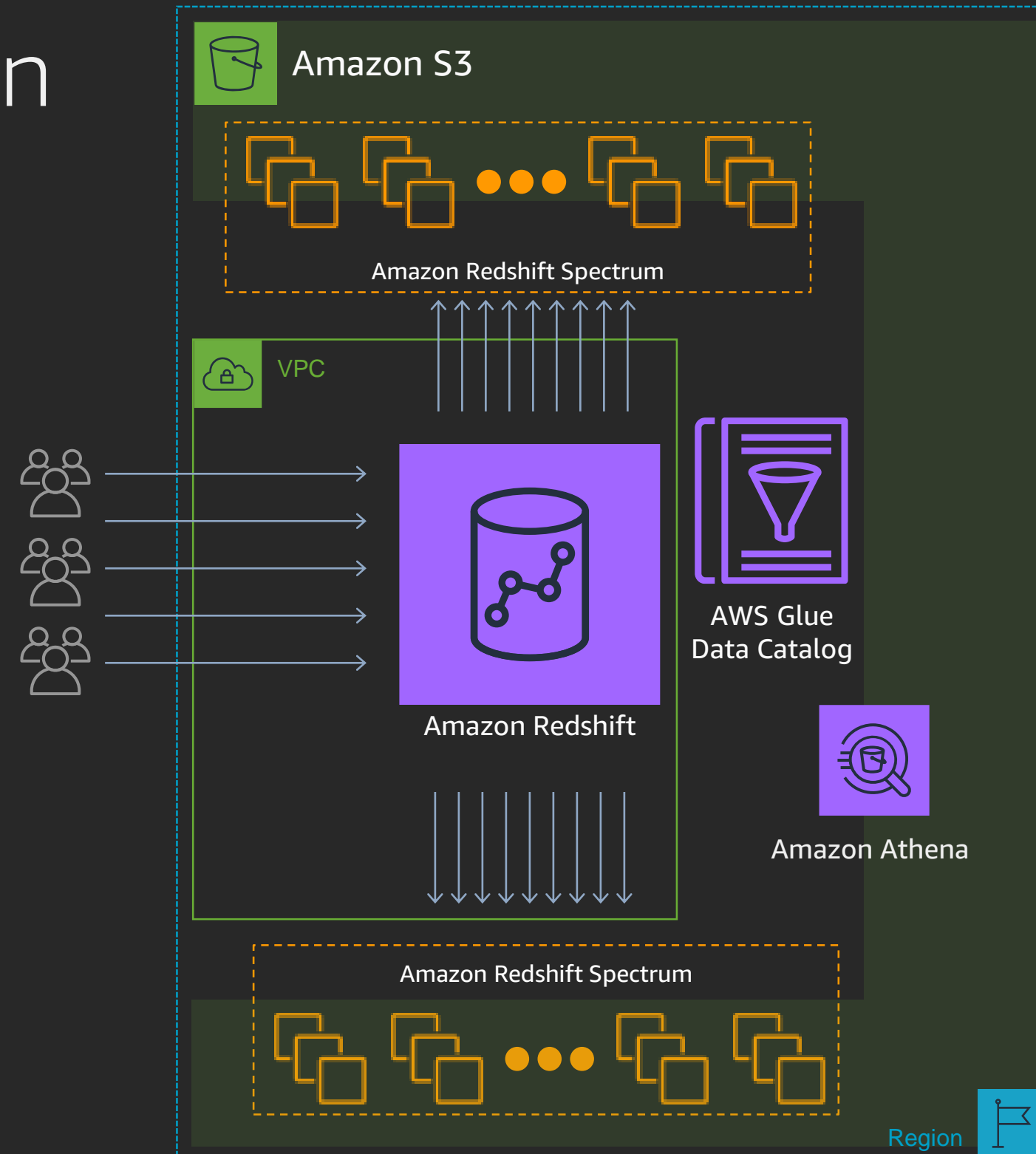
Lab #5 – Accessing your data lake

- 20 minutes

- Complete Lab5.ipynb

- Nested data access

- Usage patterns

- Additional query engines

Lab #5: Accessing your data lake

Recap Lab #5 – Accessing your data lake

- File format and partition comparisons

- Downsizing while maintaining performance

- Simple bootstrap for your data lake

- Concurrency scaling Amazon Redshift Spectrum queries

Next steps

aws

Next steps

Feedback welcome

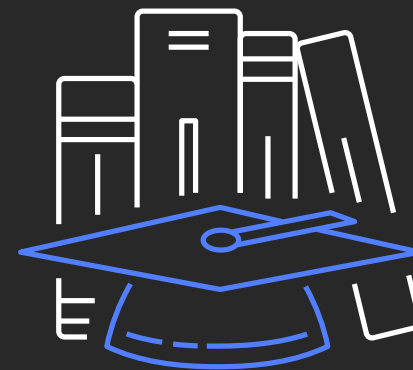- Submit feature requests and issues
- We're here to help

Make it your own

- Bring your own datasets and workloads
- Test at realistic scales for your requirements
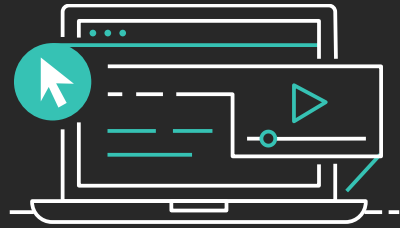- Fork and connect to other database engines

Stay updated

- Watch for new feature guides
- New managed datasets and workloads
- Performance resources routinely updated

Learn big data with AWS Training and Certification

Resources created by the experts at AWS to help you build and validate data analytics skills

New free digital course, Data Analytics Fundamentals, introduces Amazon S3, Amazon Kinesis, Amazon EMR, AWS Glue, and Amazon Redshift

Classroom offerings, including Big Data on AWS, feature AWS expert instructors and hands-on labs

Validate expertise with the **AWS Certified Big Data - Specialty** exam or the new **AWS Certified Data Analytics - Specialty** beta exam

Visit aws.amazon.com/training/paths-specialty/

aws training and certification

Thank you!

Zach Christopherson

chriz@amazon.com

Brandon Schur

bschur@amazon.com

AWS re:Invent

aws

Please complete the session survey in the mobile app.