Replicate, Analyze, and Visualize Datasets Using AWS Database Migration Service and Serverless Big Data Technologies

Agenda

- Workshop intro and outcomes
- Use-case and proposed solution
- Lab setup and environment
- AWS Service Introductions - Database Migration Services, Amazon Aurora Serverless, AWS Glue, Amazon Athena, Amazon QuickSight
- Hands-on Lab

Workshop Details

Requirements and expectations

- Attendees use temporary AWS accounts to run the lab with IAM admin permissions
- Basic knowledge of AWS services (Amazon Aurora, Amazon Simple Storage Service (S3), AWS DMS, AWS SCT, Amazon Athena, Amazon QuickSight, AWS CloudFormation, AWS Glue)
- Comfortable working on the AWS console and configure AWS services
- Working knowledge of relational databases (SQL Server)

Expected outcomes

- Use AWS serverless technologies to replicate, analyze, and visualize data from relational and semi-structured datasets to gain insight and drive business outcomes
- Leverage AWS Schema Conversion Tool (SCT) and AWS Database Migration Service (DMS) to convert MS SQL server database to Amazon Aurora Serverless MySQL
- Use AWS Glue to build data catalog & transform data
- Use Amazon Athena to run ad-hoc SQL queries and interactively analyze data stored on Amazon S3 data store
- Use Amazon QuickSight to visualize data and drive additional insights (Optional)

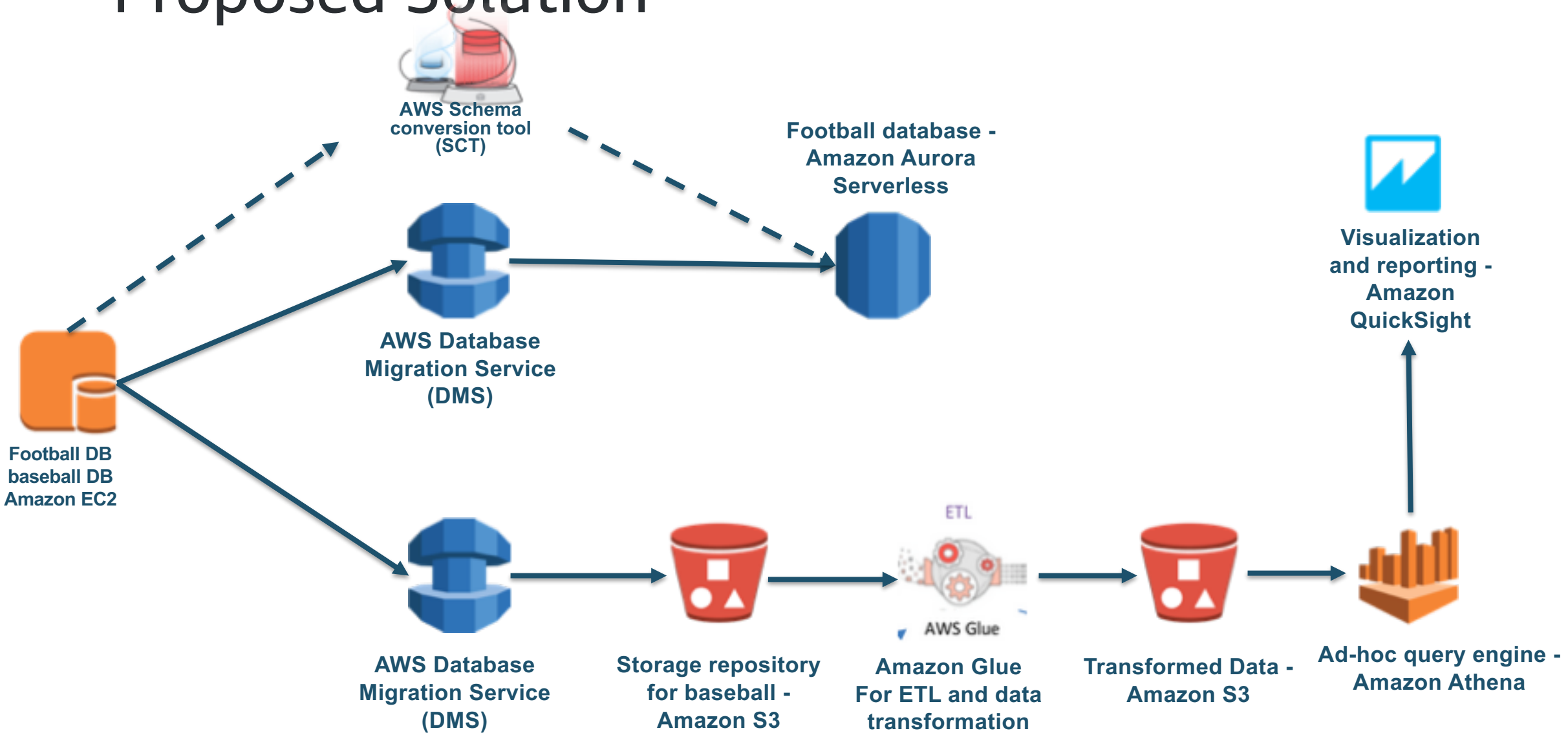
aws

Use-Case

- The datasets - football and baseball data stored on a single source SQL server database
- Need to fan-out source data into independent target data stores for separation and query patterns based on seasons
- Football is the current season, data to be replicated to target serverless relational database on AWS for transactional and complex queries
- Baseball season has ended, data replicated to target data store on Amazon S3 to transform data using AWS Glue and run ad-hoc queries using Amazon Athena
- Need for data analysis and visualization on target football dataset using Amazon QuickSight

aws

Proposed Solution



Solution Steps

1. Launch CloudFormation template to setup Amazon Virtual Private Cloud (VPC), source and target database environment
2. Use AWS Schema Conversion Tool (SCT) to convert schema from source SQL server to target Amazon Aurora Serverless MySQL for football DB
3. Use AWS Database Migration Service (DMS) to replicate source SQL server data to target Amazon Aurora Serverless MySQL for football DB
4. Use AWS Database Migration Service (DMS) to replicate source SQL server data to target Amazon S3 for baseball DB
5. Use AWS Glue to build data catalog and transform data to Parquet
6. Setup and configure Amazon Athena to run ad-hoc queries on baseball data stored on Amazon S3
7. Setup and configure Amazon QuickSight to build visualization for football data stored on Amazon Aurora Serverless MySQL

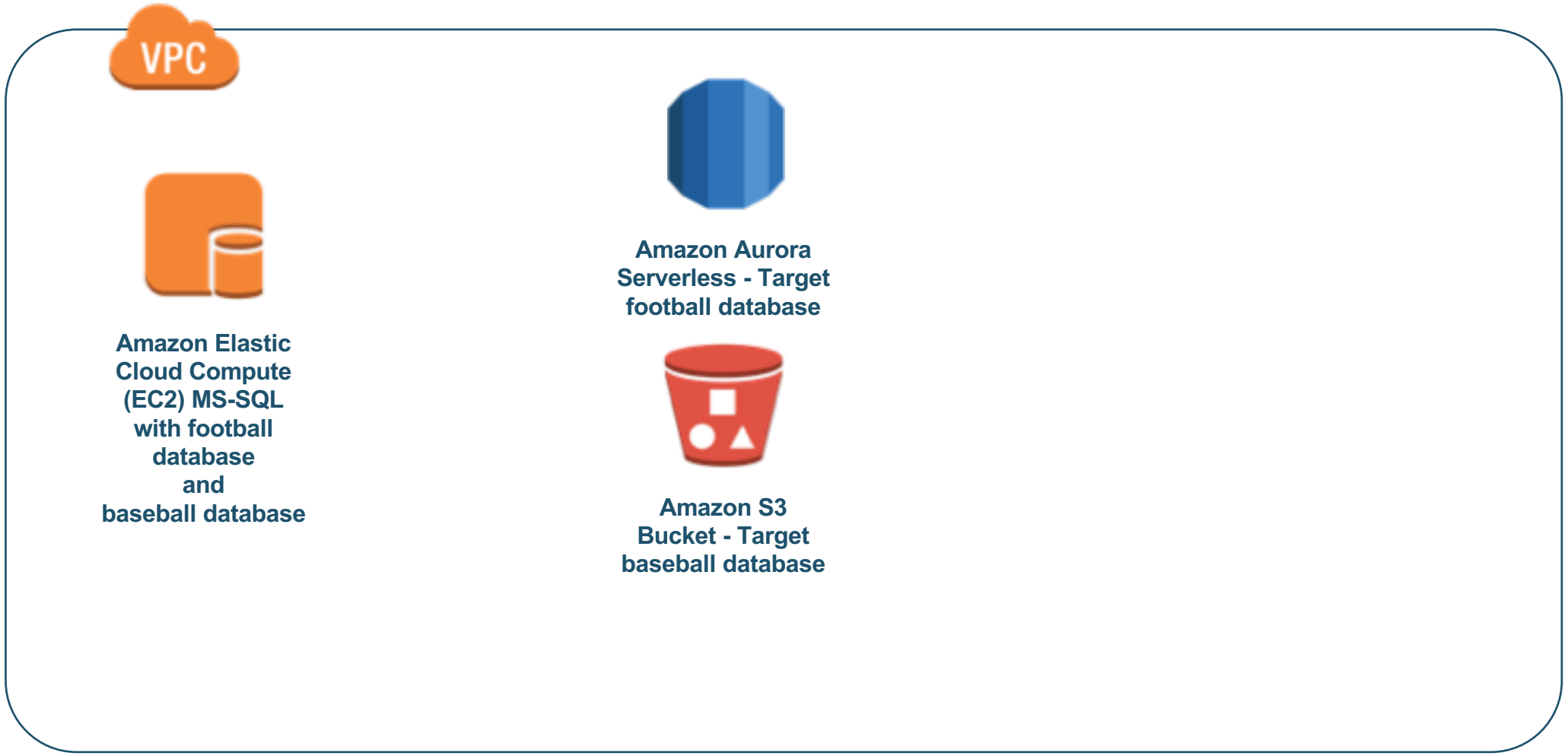
aws

Hands on Lab Overview

- AWS Region for the lab: **eu-west-1 (Ireland)** (this is enabled by default)
- Follow instructions in each of the exercises
    - Exercise 1 - Region verification, launch CloudFormation stack, verify VPC, source and target database, Amazon S3 bucket
    - Exercise 2 - Source SQL server configuration, AWS SCT schema conversion steps
    - Exercise 3 - AWS DMS configuration and data migration to Amazon S3 and Amazon Aurora Serverless
    - Exercise 4 – AWS Glue to build data catalog and transform data to Parquet
    - Exercise 5 - Amazon Athena configuration and queries
    - Exercise 6 - Amazon QuickSight configuration and visualization

- GitHub Repo contains:
    - CloudFormation template
    - ConfigureSQLServer.sql—source SQL Server configuration script
    - AthenaQueries—queries for Athena text file
    - Lab guide and exercises

AWS CloudFormation launch environment



AWS service overview

AWS DMS and SCT

AWS Database Migration Service (DMS) easily and securely migrate and/or replicate your databases *and* data warehouses to AWS

AWS Schema Conversion Tool (SCT) convert your commercial database and data warehouse schemas to open-source engines or AWS-native services, such as Amazon Aurora and Amazon Redshift

Migrated over 30,000 databases. And counting …

aws

Serverless Computing + Relational Databases

What if you could have a relational database that:

- Starts up on demand, shuts down when not in use?
- Seamlessly scales with no instances to manage?
- Is billed per second for the database capacity you use?
- Is compatible with existing database applications?

Introducing Amazon Aurora Serverless:

On-demand, auto-scaling database for applications with variable workloads

aws

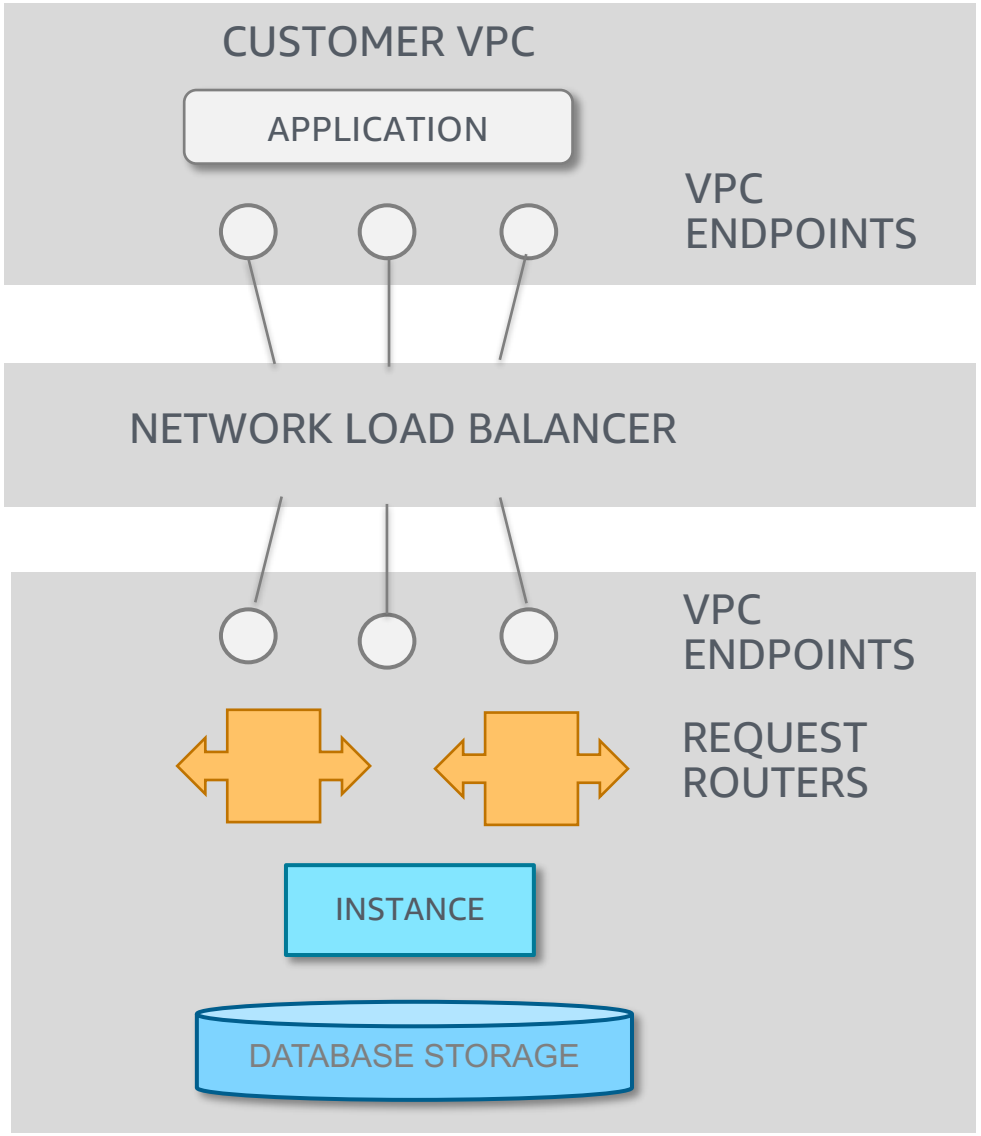
Aurora Serverless Use Cases

- Infrequently used applications (e.g. low-volume blog site)
- Applications with variable load - peaks of activity that are hard to predict (e.g. news site)
- Development or test databases not needed on nights or weekends
- Multi-tenant SaaS applications

Aurora Serverless Provisioning

When you provision a database, Aurora Serverless:

- Creates an Aurora storage volume
- Creates endpoint in your VPC for the application connection
- Configures network load balancer behind endpoint
- Configures request routers (multi-tenant) to route database traffic
- Provisions initial instance capacity



AWS Glue automates the undifferentiated heavy lifting of ETL

Discover — Automatically discover and categorize your data making it immediately searchable and queryable across data sources

Develop — Generate code to clean, enrich, and reliably move data between various data sources; you can also use their favorite tools to build ETL jobs

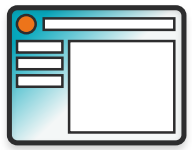
Deploy — Run your jobs on a serverless, fully managed, scale-out environment. No compute resources to provision or manage.

aws

AWS Glue: Components

Data Catalog

- Hive Metastore compatible with enhanced functionality
- Crawlers automatically extracts metadata and creates tables
- Integrated with Amazon Athena, Amazon Redshift Spectrum



Job Authoring

- Auto-generates ETL code
- Build on open frameworks – Python and Spark
- Developer-centric – editing, debugging, sharing

Job Execution

- Run jobs on a serverless Spark platform
- Provides flexible scheduling
- Handles dependency resolution, monitoring and alerting

aws

Amazon Athena

Amazon Athena is an interactive query service that makes it easy to analyze data directly from Amazon S3 using Standard SQL

Query data directly from Amazon S3

- No loading of data
- Query data in its raw format
  - Athena supports multiple data formats
  - Text, CSV, TSV, JSON, weblogs, AWS service logs
  - Convert to an optimized form like **ORC** or **Parquet** for the **performance and cost**
- No ETL required
- Stream data directly from Amazon S3
- Take advantage of Amazon S3 durability and availability

aws

Amazon QuickSight is a business analytics service that lets business users quickly and easily visualize, explore, and share insights from their data.

aws

Deep integration with AWS data sources

Amazon QuickSight is **deeply integrated** with AWS data sources like Amazon Redshift, Amazon RDS, Amazon S3, Athena and others, as well as third-party sources like Excel, Salesforce, as well as on-premises databases.

Please fill out the CSAT survey:

http://bit.ly/2miAb2S