ABD313

AWS re:INVENT

Building an End-to-End Serverless Data Analytics Solution on AWS

Gowri Balasubramanian, AWS Solution Architect
Karthik Kumar Odapally, AWS Solution Architect
Rajeev Srinivasan, AWS Solution Architect
Rudy Chetty, AWS Solution Architect

November 27, 2017

AWS re:Invent

aws

Agenda

- Presentation
  - Service Introduction
  - Reference Architecture
  - Query Performance Best Practices
  - Workshop Overview

- Hands on Workshop
  - Lab1: Serverless Analysis of Data in Amazon Simple Storage (Amazon S3) using Amazon Athena
  - Lab2: Visualization Using Amazon QuickSight
  - Lab3: Serverless ETL and Data Discovery Using Amazon Glue [Optional]
  - Lab4: Analysis of Data in Amazon S3 Using Amazon Redshift Spectrum [Take Home]

SERVICE INTRODUCTION

AMAZON ATHENA

AMAZON QUICKSIGHT

AWS GLUE

Amazon Athena

Start Querying Instantly

Serverless. No ETL.

Open. Powerful. Standard.

Built on Presto. Runs standard SQL.

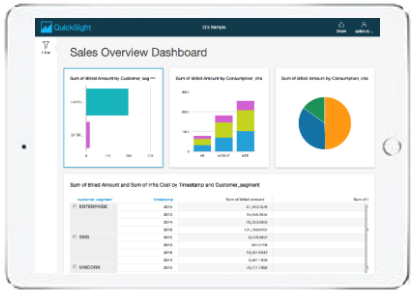
Pay per Query

Only pay for data scanned.

Amazon QuickSight

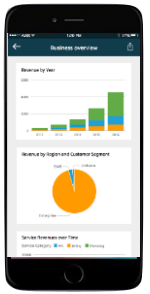
Analyses

Analyses are visual explorations of your data. Multiple users can collaborate on analyses with the ability to modify and change them in any way.



Dashboards

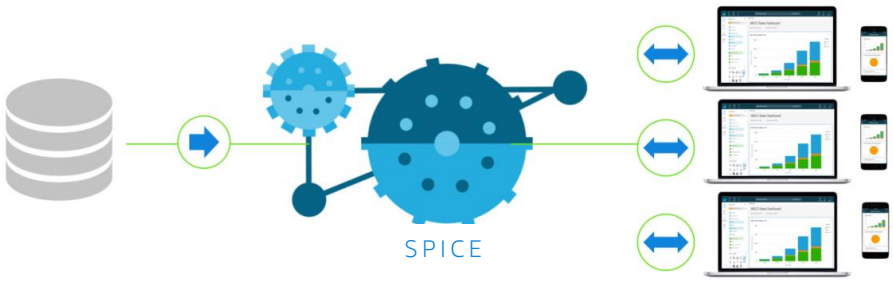
You can share your analyses as read only dashboards. Viewers can interact with and filter the visualizations without modifying them.

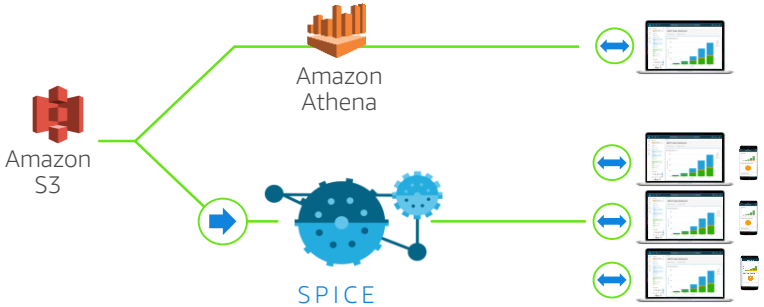


Storyboards

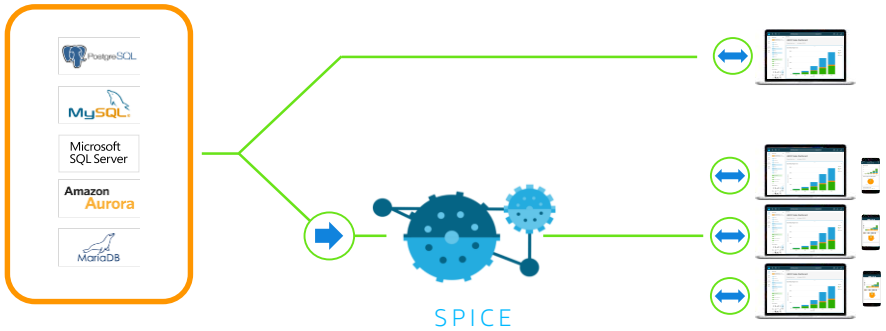
Let you combine visualizations into a guided tour that you can share with other users.

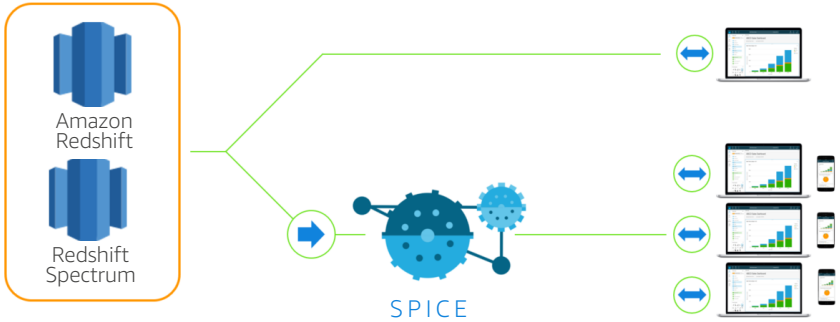
Amazon QuickSight—Data Sources



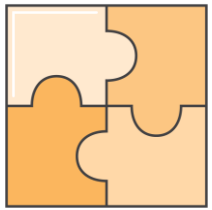


Amazon QuickSight—Data Sources

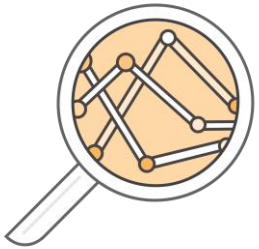




AWS Glue



Integrated
Data Catalog



Automated
Data Discovery

Code
Generation

Developer
Endpoints

Flexible
Job Scheduler

AWS Glue—Components

Data Catalog

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

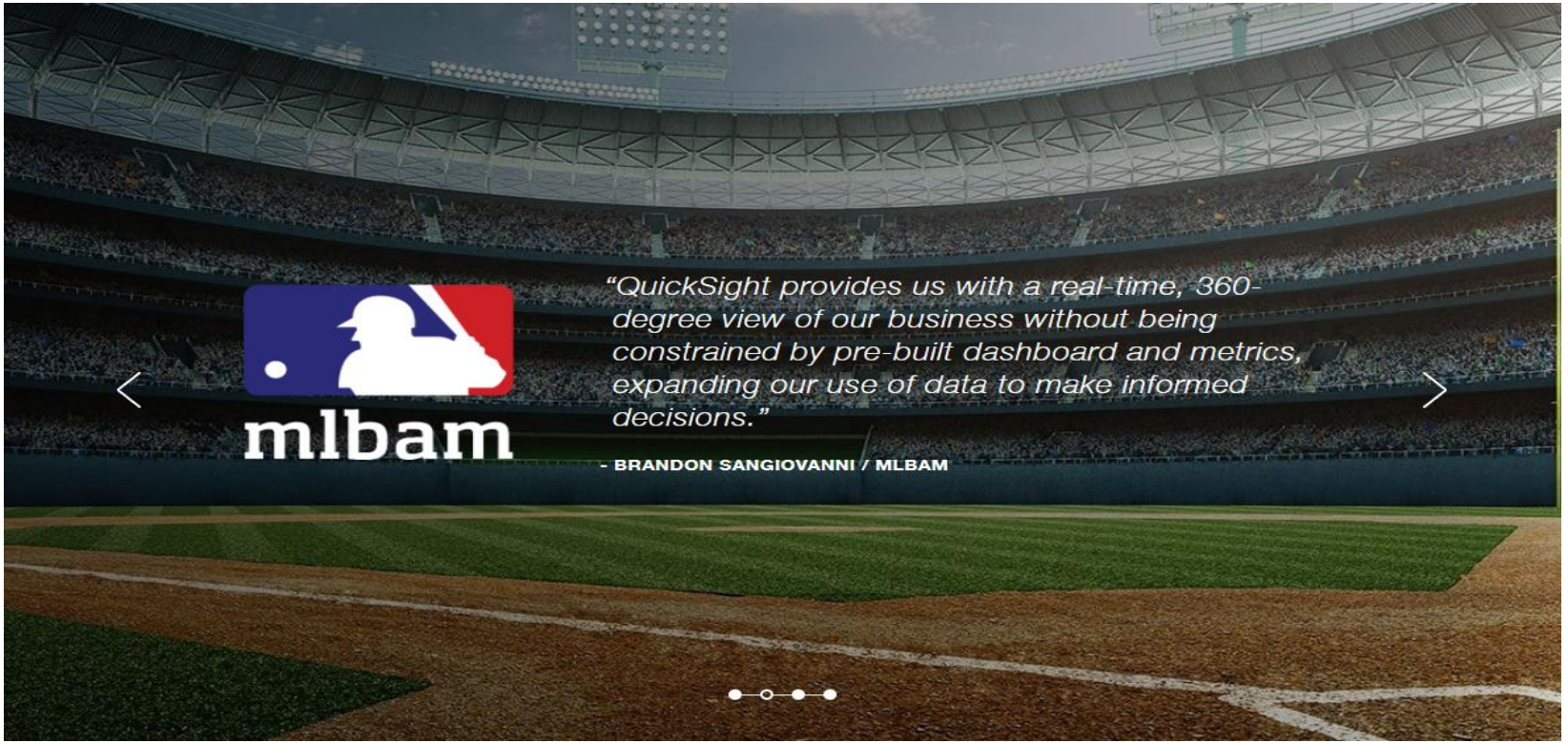
Job Authoring

- 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

Customer Reference—Amazon QuickSight



Customer Reference—Amazon Athena

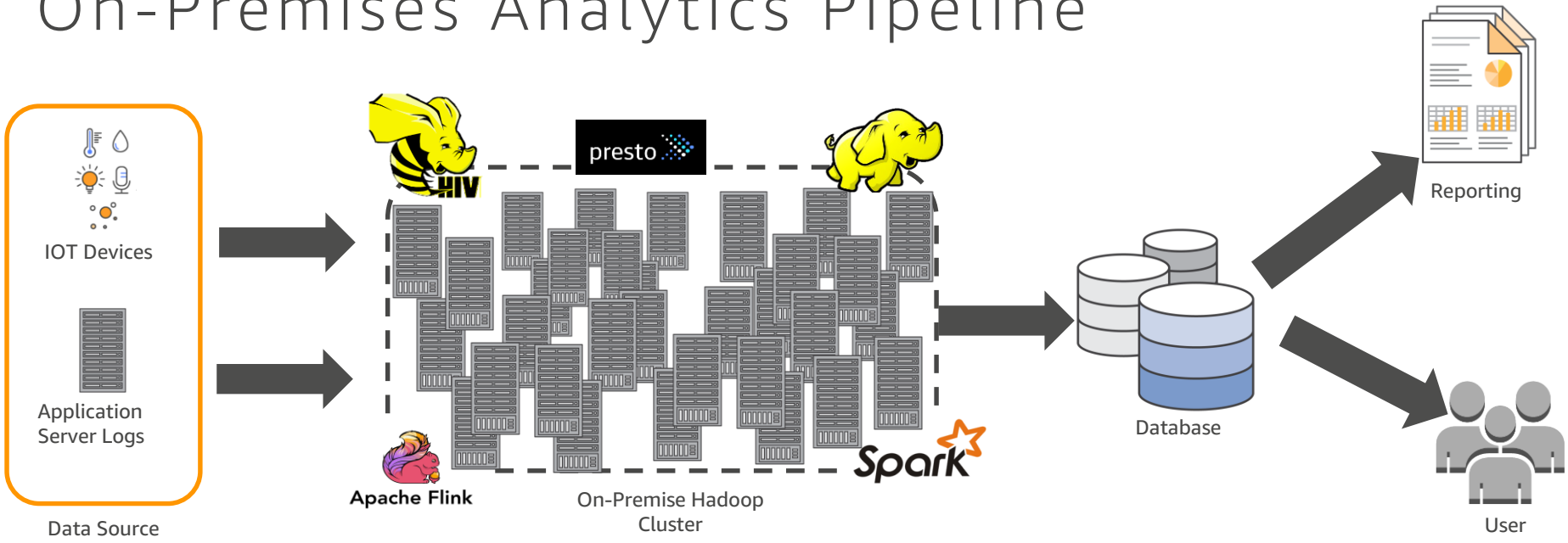
One of the big attractions of Amazon Athena is that it's serverless and purely consumption based. We only pay when we're actually querying the data, and we don't have to keep a cluster running all the time.

—Matt Chesler,
Director of DevOps, Movable Ink

Movable Ink

Reference Architecture

On-Premises Analytics Pipeline



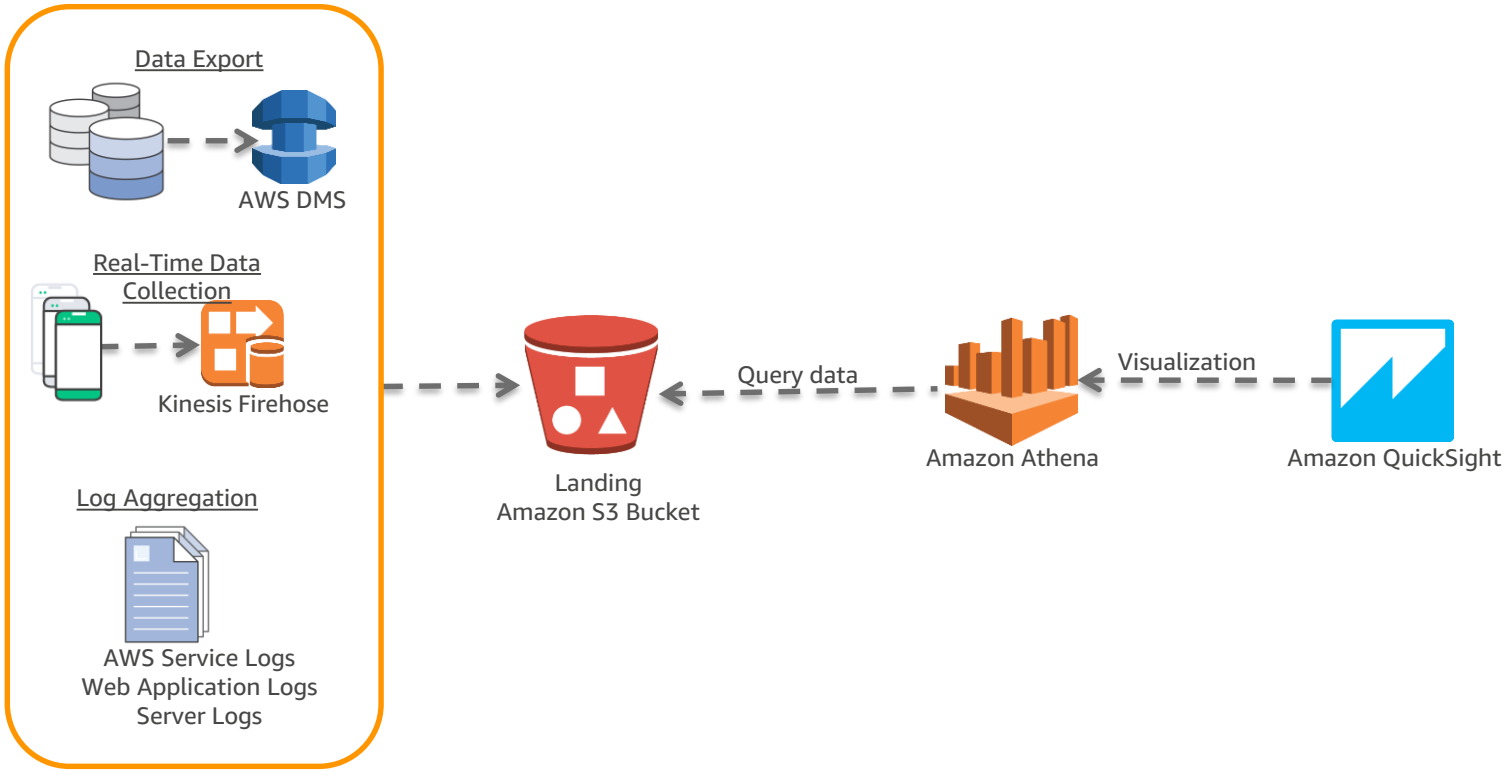

Static : Not Scalable


Outages Impact

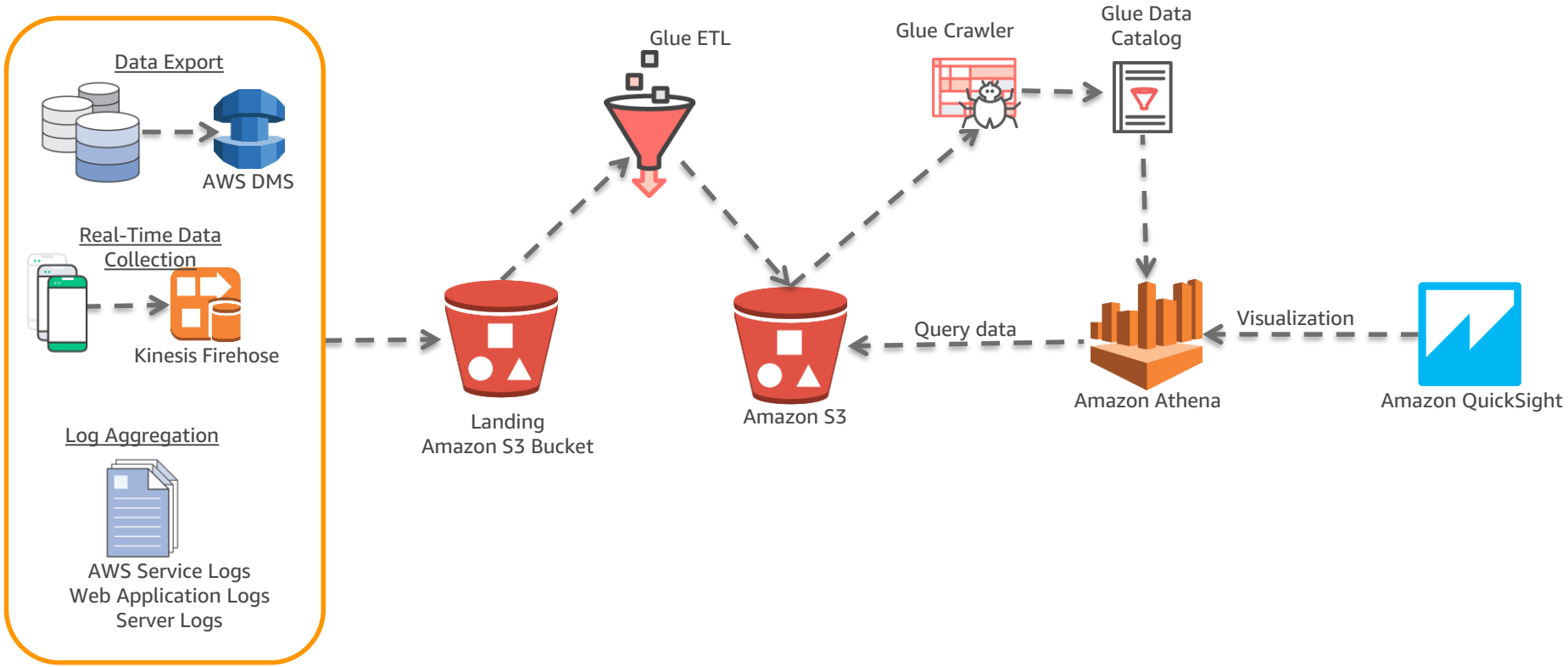

Storage Compute


Always On

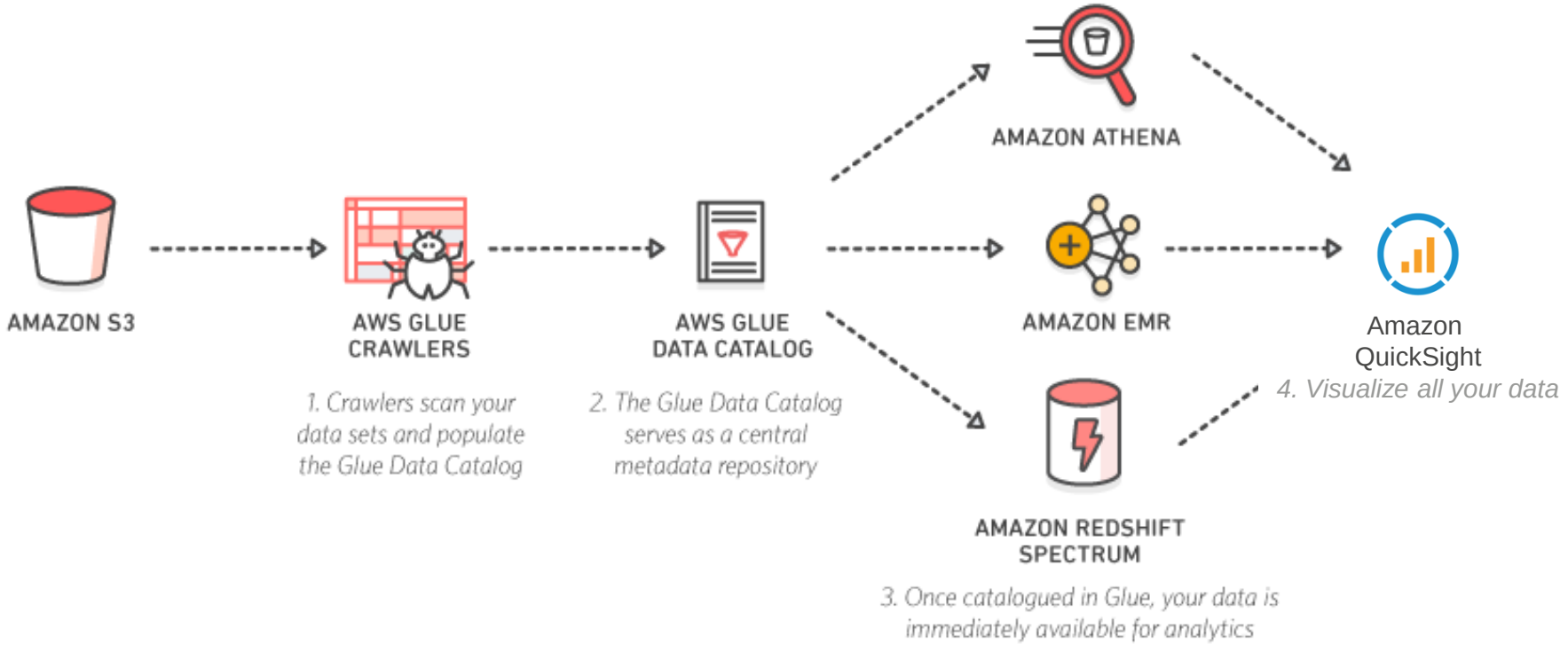
Reference Architecture

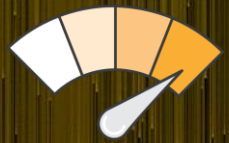


Reference Architecture—ETL



Reference Architecture





Query Performance

Best Practices—Storage

✓**Partition** your data

✓Optimize **columnar data** store generation

SELECT count(*) as count FROM **taxi_rides_csv**
(Run time: 20.06 seconds, Data scanned: 207.54GB, Row Count: 1,310,911,060)

SELECT count(*) as count FROM **taxi_rides_parquet**
(Run time: 5.76 seconds, Data scanned: 0KB, Row Count: 2,870,781,820)

✓**Compress and split** files

✓Optimize **file size**

Best Practices—Query

- ✓ Optimize **ORDER BY**

  SELECT * FROM nytaxirides WHERE year = 2011 AND month = 5 AND type = 'yellow' ORDER BY ratecode
  (Run time: 3 minutes 6 seconds)

  SELECT * FROM nytaxirides WHERE year = 2011 AND month = 5 AND type = 'yellow' ORDER BY ratecode **LIMIT 1000**
  (Run time: 3.01 seconds)

- ✓ Optimize **joins**

- ✓ Optimize **GROUP BY**

- ✓ Optimize the **LIKE operator**

Best Practices—Query

✓ Use **approximate functions**

SELECT **count(distinct tpep_pickup_datetime)** FROM nytaxidata
(Run time: 30.82 seconds)

SELECT **approx_distinct(tpep_pickup_datetime)** FROM nytaxidata
(Run time: 25.21 seconds)

✓ Only **include the columns** that you need

SELECT * FROM nytaxirides WHERE year = 2011 AND type = 'yellow' AND month = 5
(Run time: 2 minutes 59 seconds, Data scanned: 382.88MB)

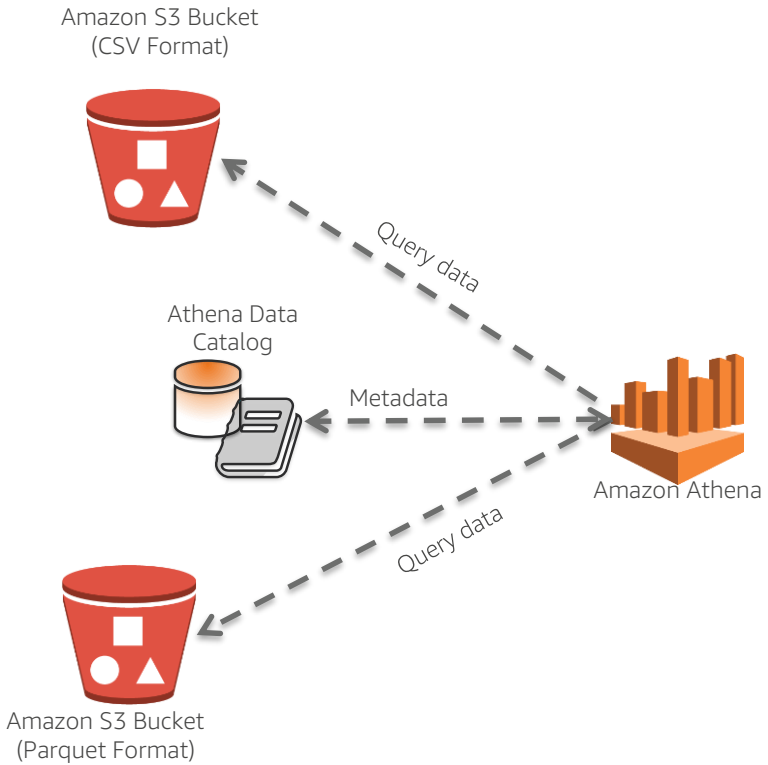
SELECT **vendorid, ratecode, passenger_count** FROM nytaxirides WHERE year = 2011 AND type = 'yellow' AND month = 5
(Run time: 38.79 seconds, Data scanned: 10.06MB)

Hands-On Workshop

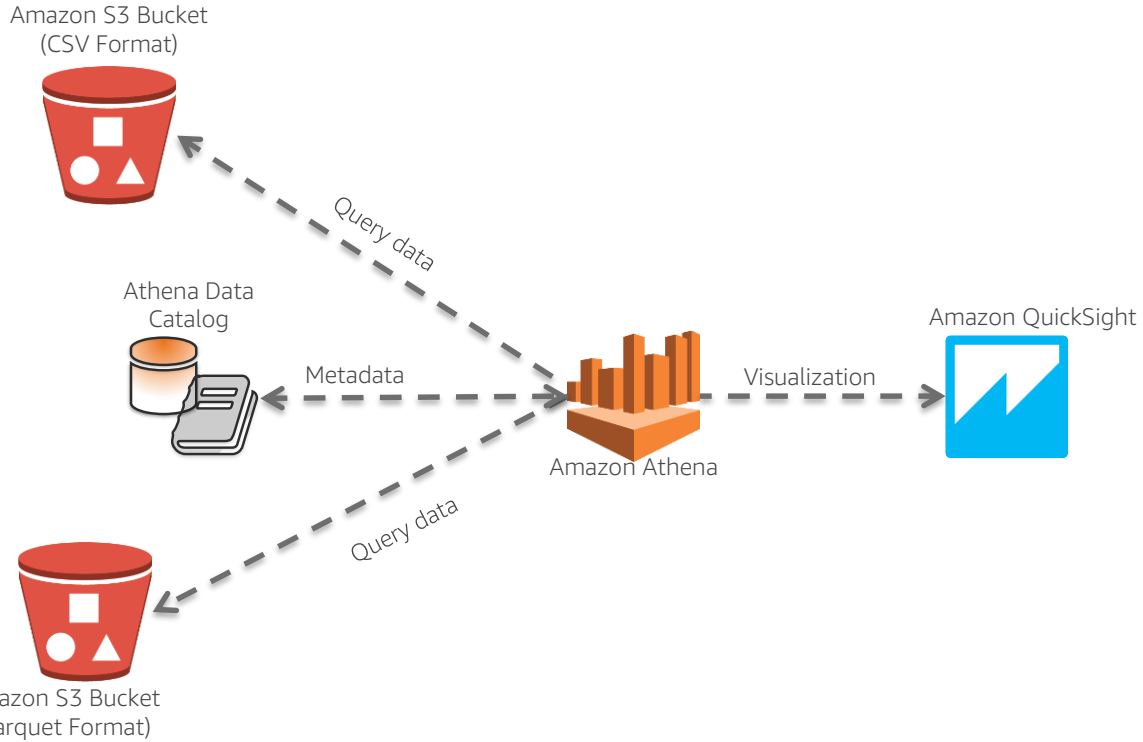
AWS re:Invent

aws

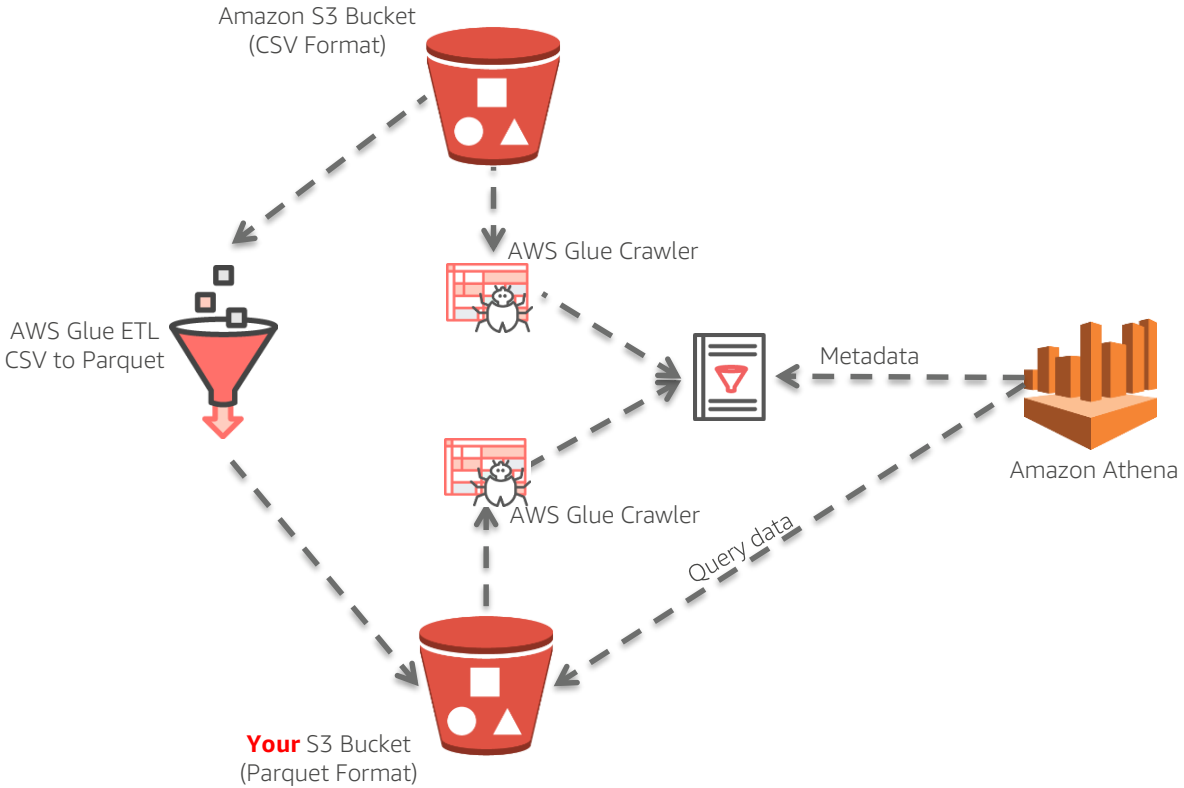
Lab 1: Serverless Analysis Using Amazon Athena



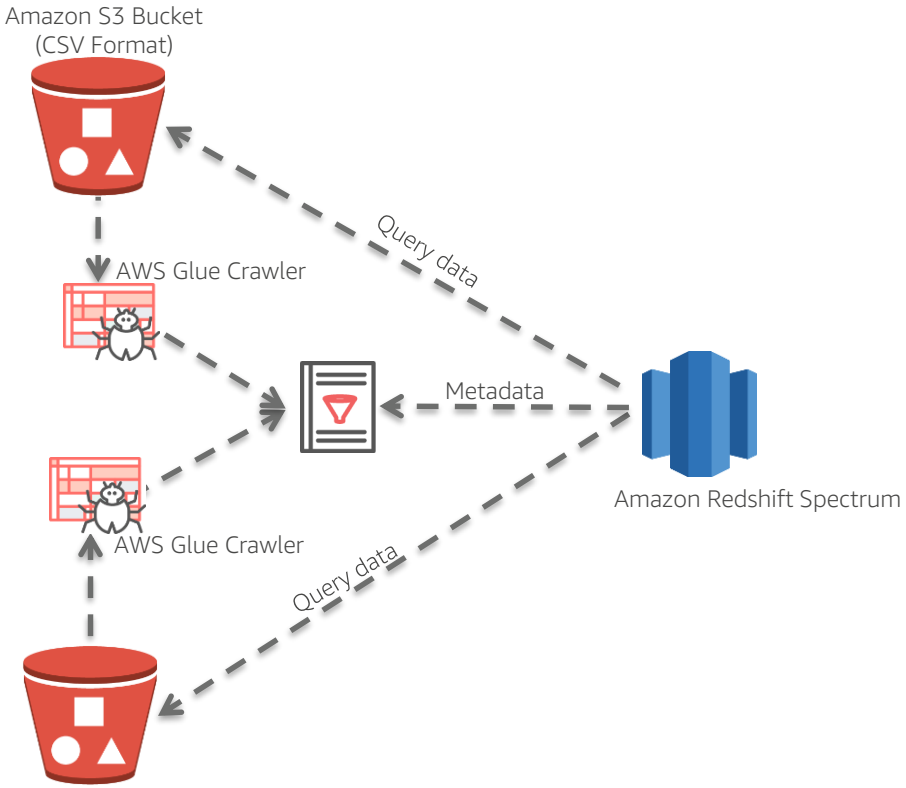
Lab 2: Visualization Using Amazon QuickSight



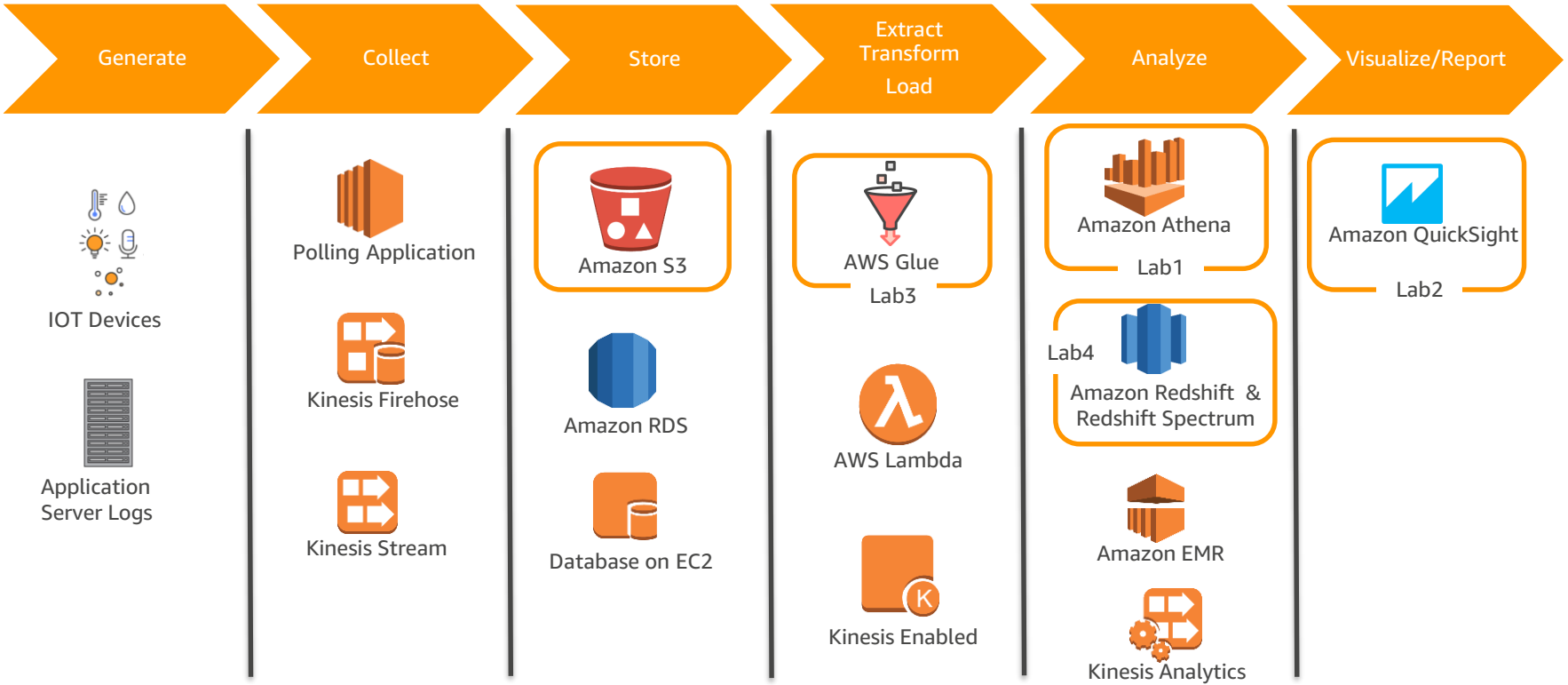
Lab 3: ELT and Data Discovery Using AWS Glue



Lab 4: Analysis Using Redshift Spectrum

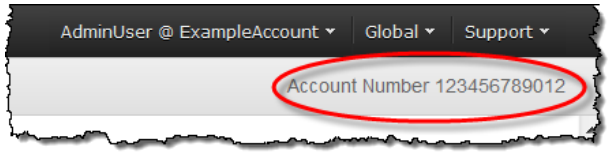


Analysis & Visualization Pipeline on AWS



Workshop

- Please collect the **credit coupon**. You can apply this coupon towards completing the labs in this workshop.
- Create an **AWS Account**, if you don't have one. Please **do not** use your production account for the labs.
- Provide your **AWS Account ID** for whitelisting to any of the AWS personnel who are staffing the workshop. Choose **Support** on the navigation bar on the upper right, and then choose **Support Center**. Your currently signed-in account ID appears in the upper-right corner below the **Support** menu.



- Navigate to the following web link for workshop lab instruction

http://bit.ly/2jgx6vd

- Choose Oregon region for the labs.

Thank you!

Please complete your survey

aws