AWS re:Invent

STG315

Query in Place with AWS

John Mallory
Amazon Web Services
Business Development Manager

Ridge XU
Amazon Web Services
Solutions Architect

Anson Shen
Amazon Web Services
Solutions Architect

AWS re:Invent

aws

Agenda

Amazon Simple Storage Service (Amazon S3) Select & Amazon Glacier Select

Amazon Athena and AWS Glue

Amazon Redshift Spectrum

Amazon EMR

Breakout repeats

Tuesday, 11/27/18

Query in Place

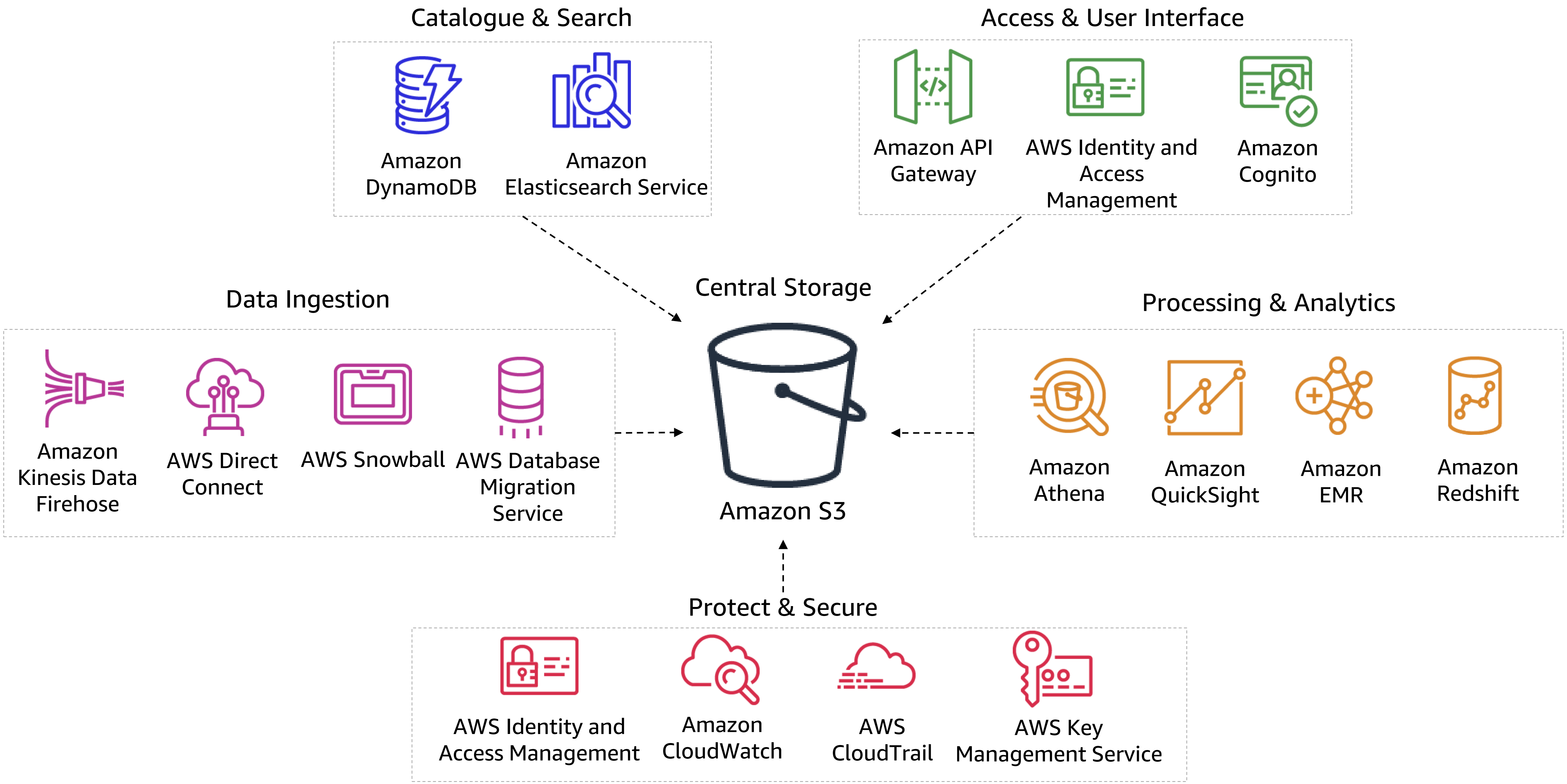
10:00 a.m. - 11:00 a.m. | Grand Ballroom D, Table 2, T1

Tuesday, 11/28/18

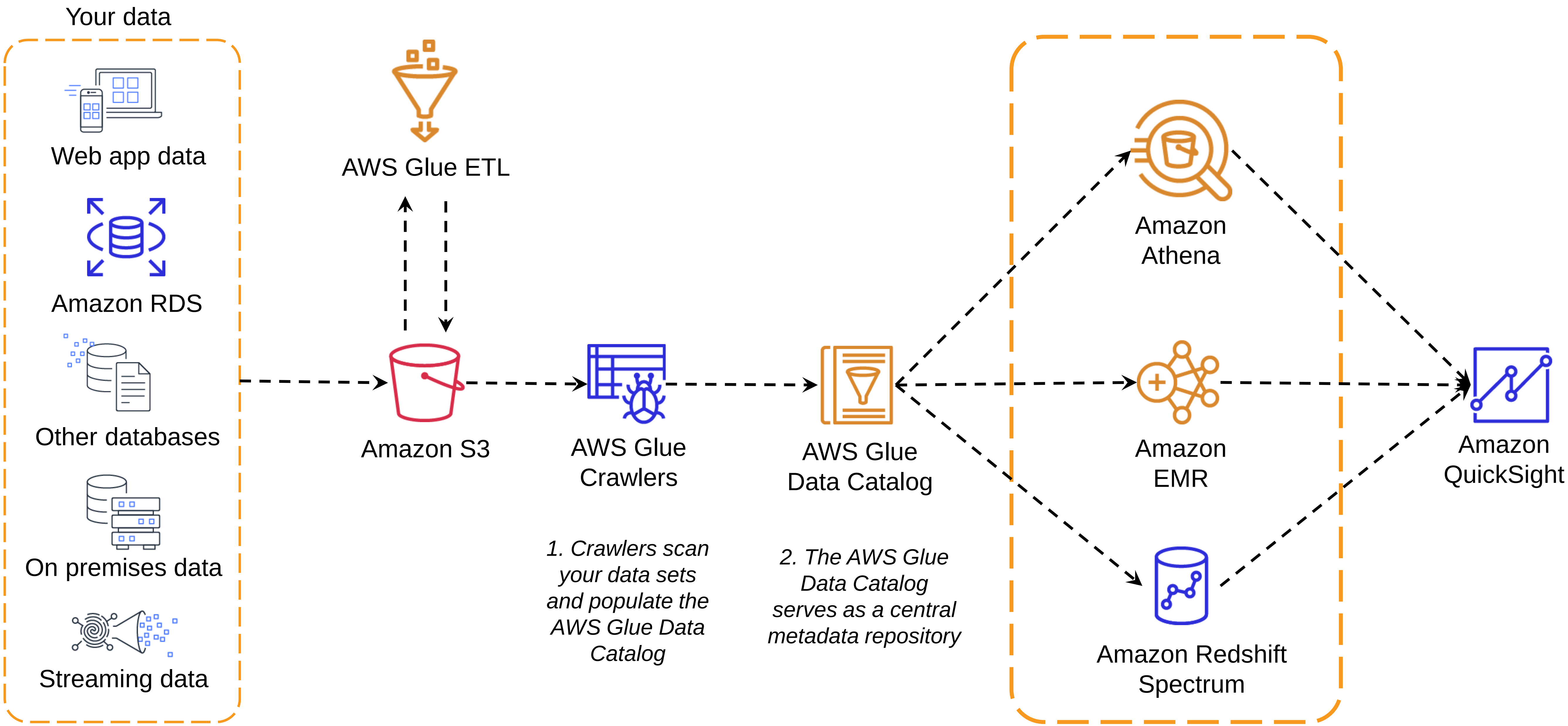
Query in Place

11:30 a.m. - 12:30 p.m. | Grand Ballroom D, Table 5, T1

Build a data lake on AWS



Data lake on Amazon S3



1. Crawlers scan your data sets and populate the AWS Glue Data Catalog

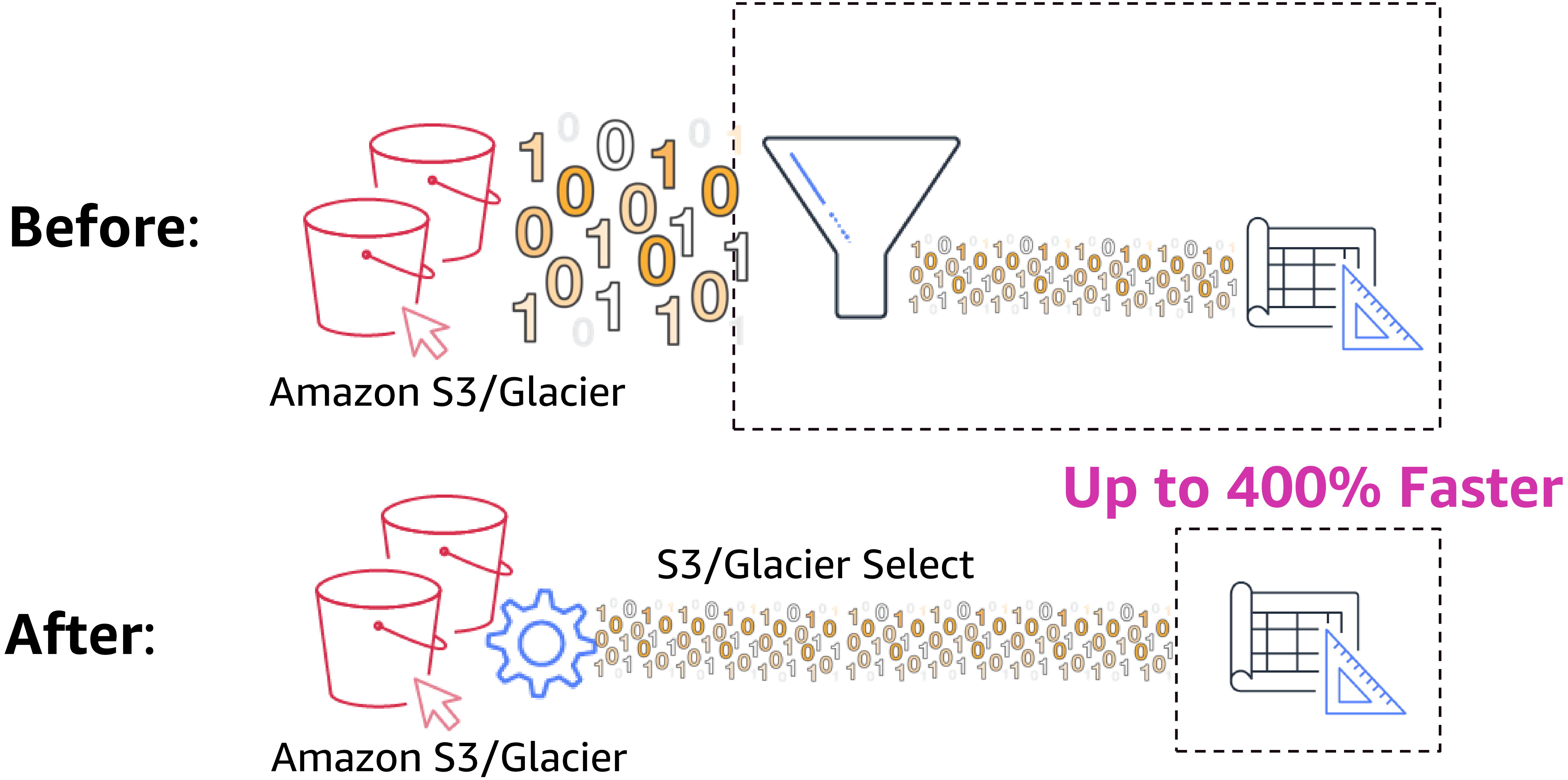
2. The AWS Glue Data Catalog serves as a central metadata repository

Introducing

Amazon S3 Select and Amazon Glacier Select

Select subset of data from an object based on an SQL expression

Amazon S3/Glacier Select: Accelerating big data



Before:

Amazon S3/Glacier

Up to 400% Faster

After:

S3/Glacier Select

Amazon S3/Glacier

Up to 80% Cheaper

Amazon S3 Select: Serverless MapReduce

Before

200 seconds and 11.2 cents

```
# Download and process all keys
for key in src_keys:
 response =
s3_client.get_object(Bucket=src_bucket
, Key=key)
 contents = response['Body'].read()
 for line in contents.split('\n')[:-
1]:
   line_count +=1
   try:
     data = line.split(',')
      srcIp = data[0][:8]
    ….
```

After

95 seconds and costs 2.8 cents

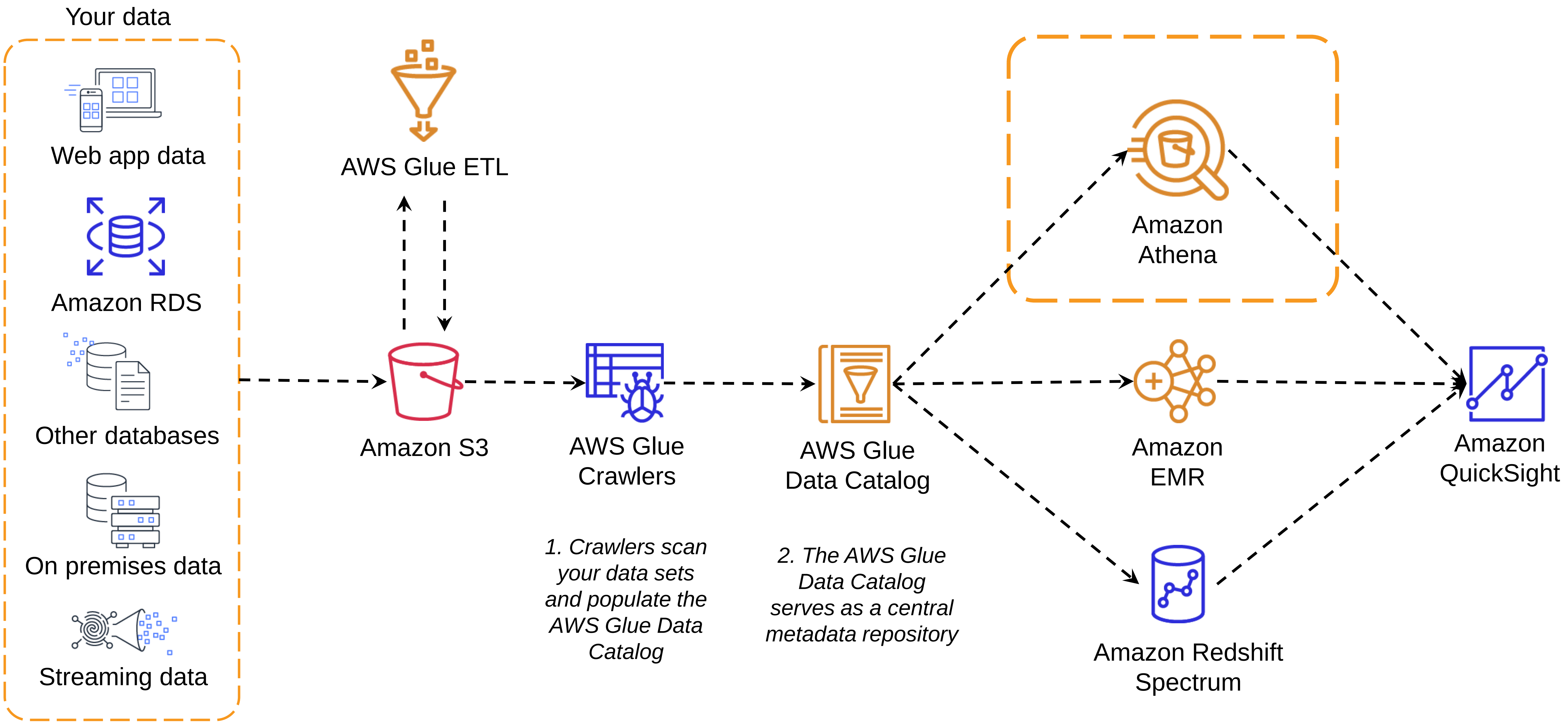
```
# Select IP Address and Keys
for key in src_keys:
 response =
s3_client.select_object_content
   (Bucket=src_bucket, Key=key,
expression =
    SELECT SUBSTR(obj._1, 1, 8),
obj._2 FROM s3object as obj)
 contents = response['Body'].read()
 for line in contents:
   line_count +=1
   try:
     ….
```

2X Faster at 1/5 of the cost

Let's build!

AWS re:Invent

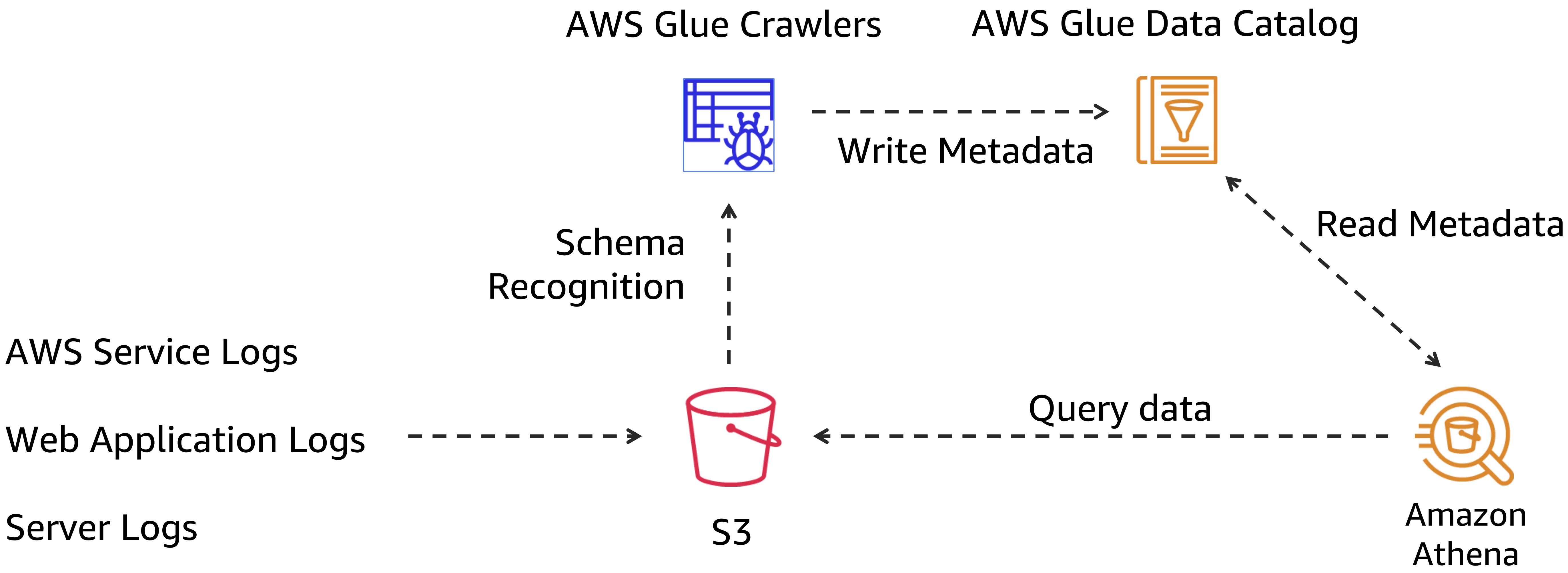
Data lake on Amazon S3



1. Crawlers scan your data sets and populate the AWS Glue Data Catalog

2. The AWS Glue Data Catalog serves as a central metadata repository

Amazon Athena querying logs with AWS Glue Catalog



AWS Glue Crawlers

AWS Glue Data Catalog

Write Metadata

Schema Recognition

Read Metadata

AWS Service Logs

Web Application Logs

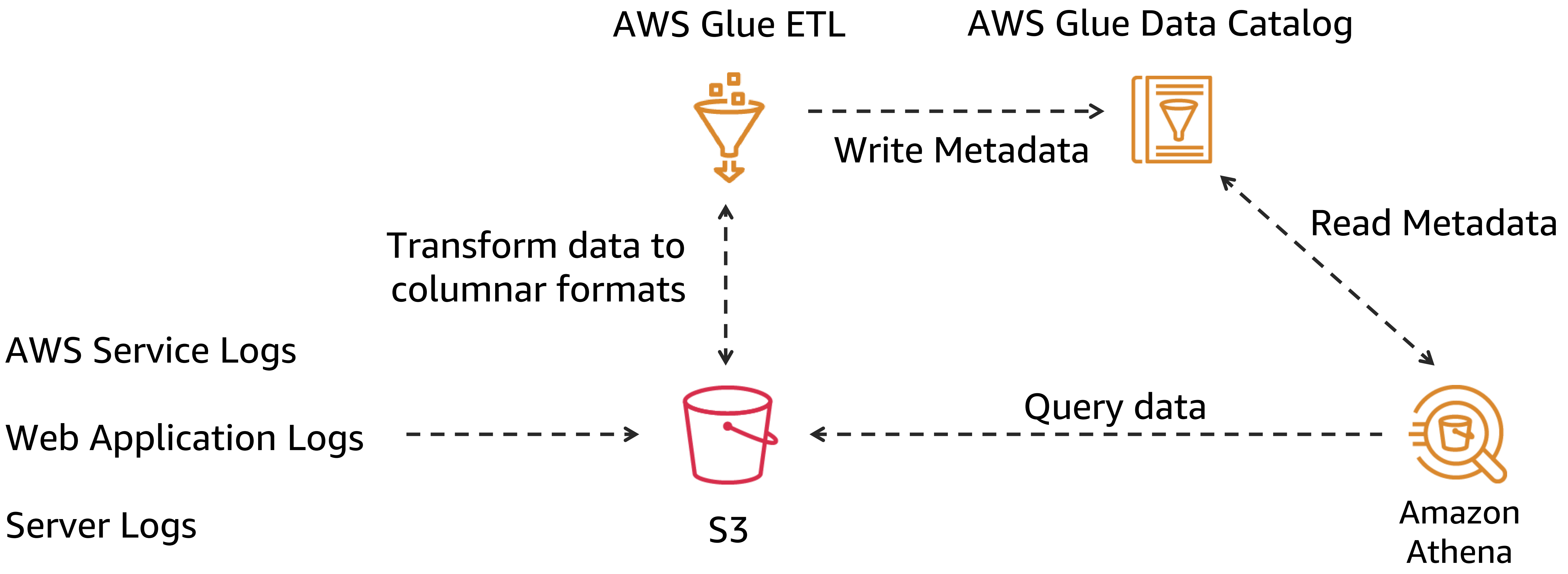
Server Logs

Query data

S3

Amazon Athena

Amazon Athena querying logs with AWS Glue Catalog


AWS Glue ETL

AWS Glue Data Catalog

Write Metadata

Transform data to columnar formats

Read Metadata

AWS Service Logs

Web Application Logs

Server Logs

S3

Query data

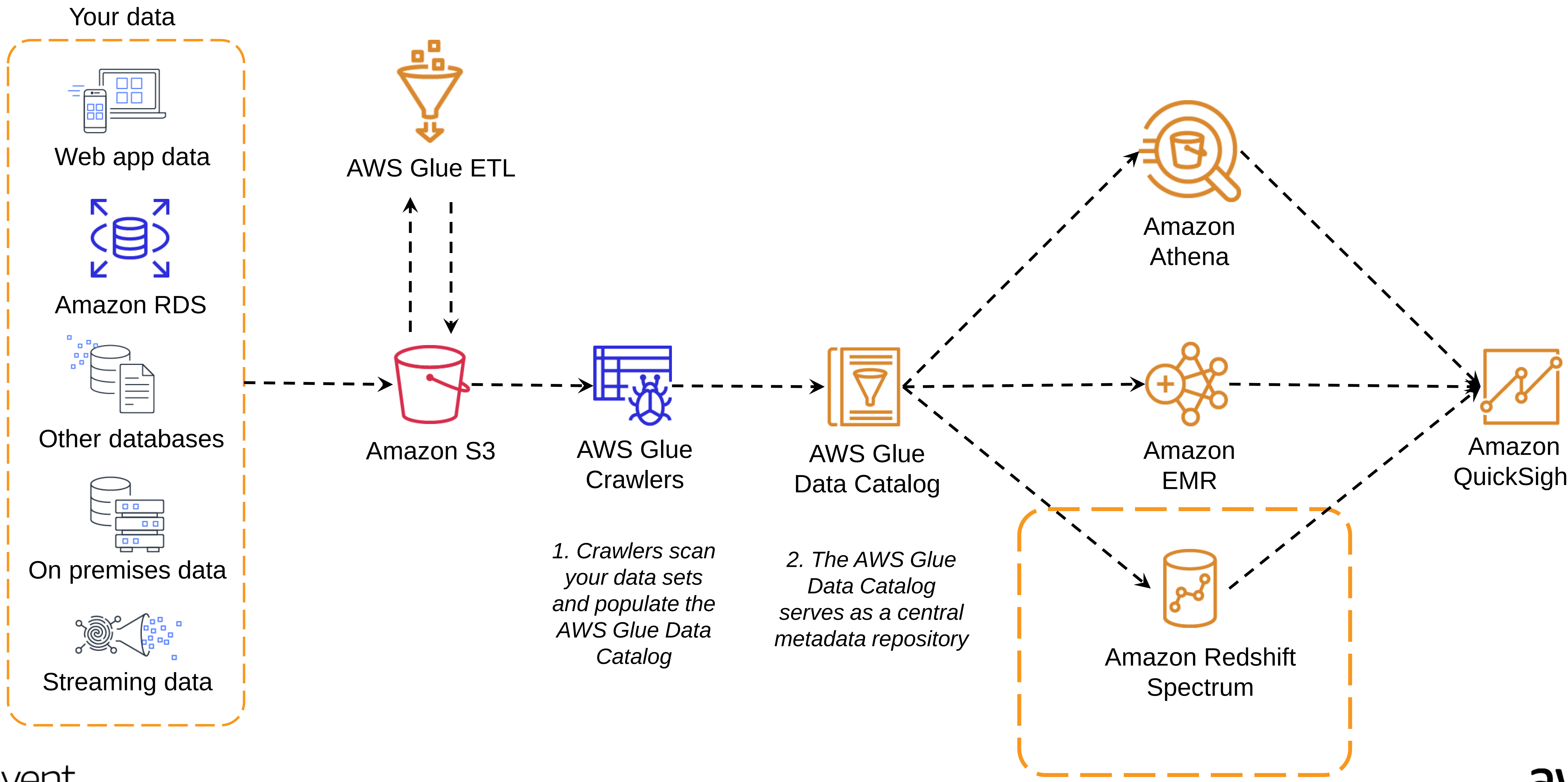
Amazon Athena

Let’s build!

AWS re:Invent

aws

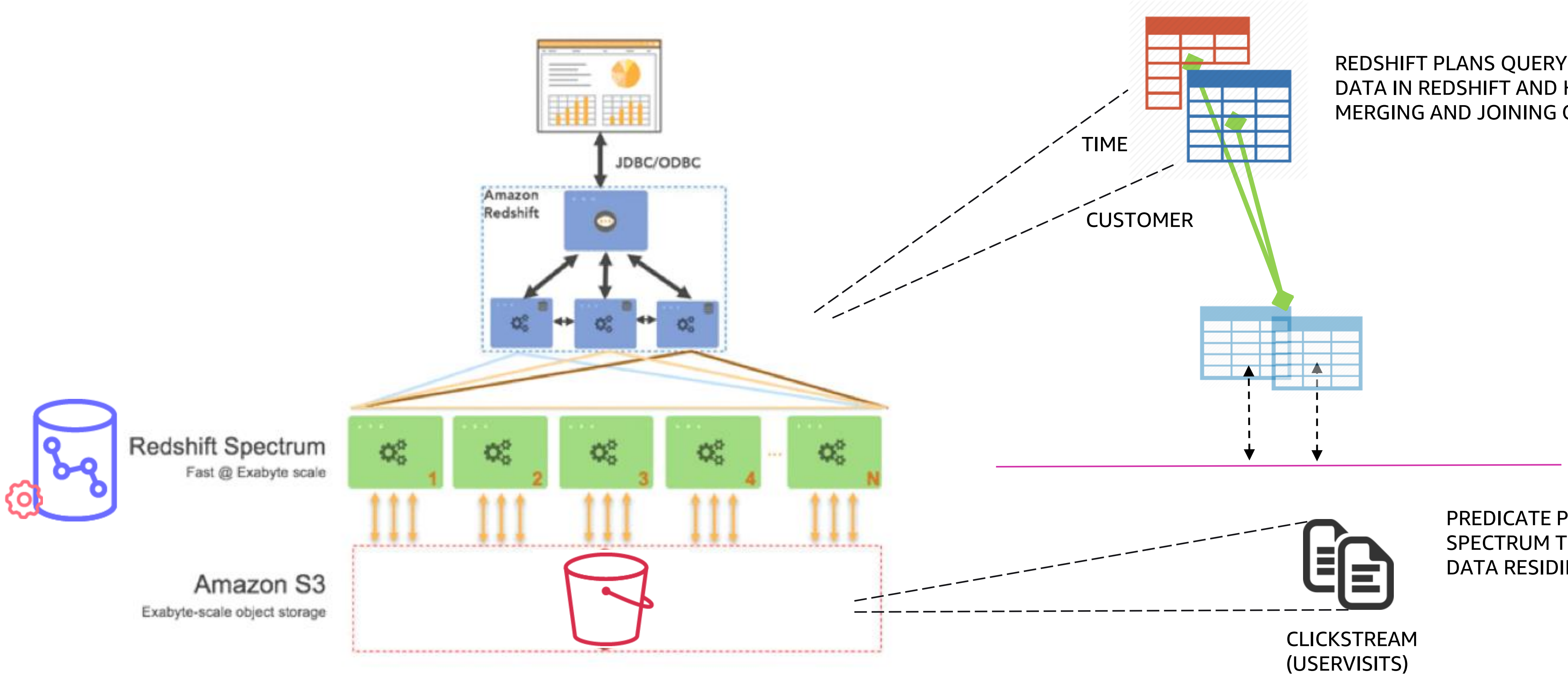
Data lake on Amazon S3



1. Crawlers scan your data sets and populate the AWS Glue Data Catalog

2. The AWS Glue Data Catalog serves as a central metadata repository

Amazon Redshift Spectrum



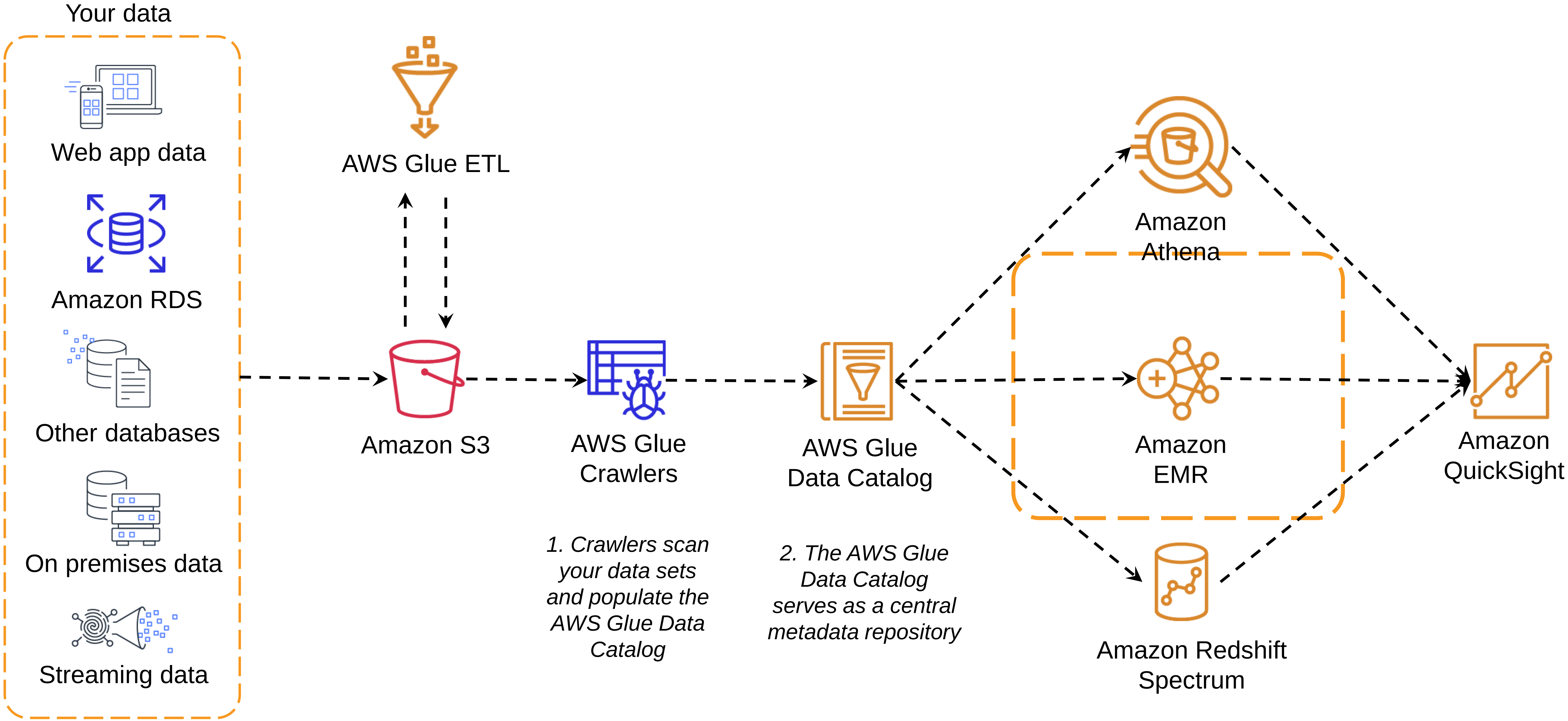
REDSHIFT PLANS QUERY, PROCESSES DATA IN REDSHIFT AND HANDLES MERGING AND JOINING OF DATASETS

PREDICATE PUSHDOWN TO SPECTRUM TO PROCESS DATA RESIDING IN S3

Let’s build!

AWS re:Invent

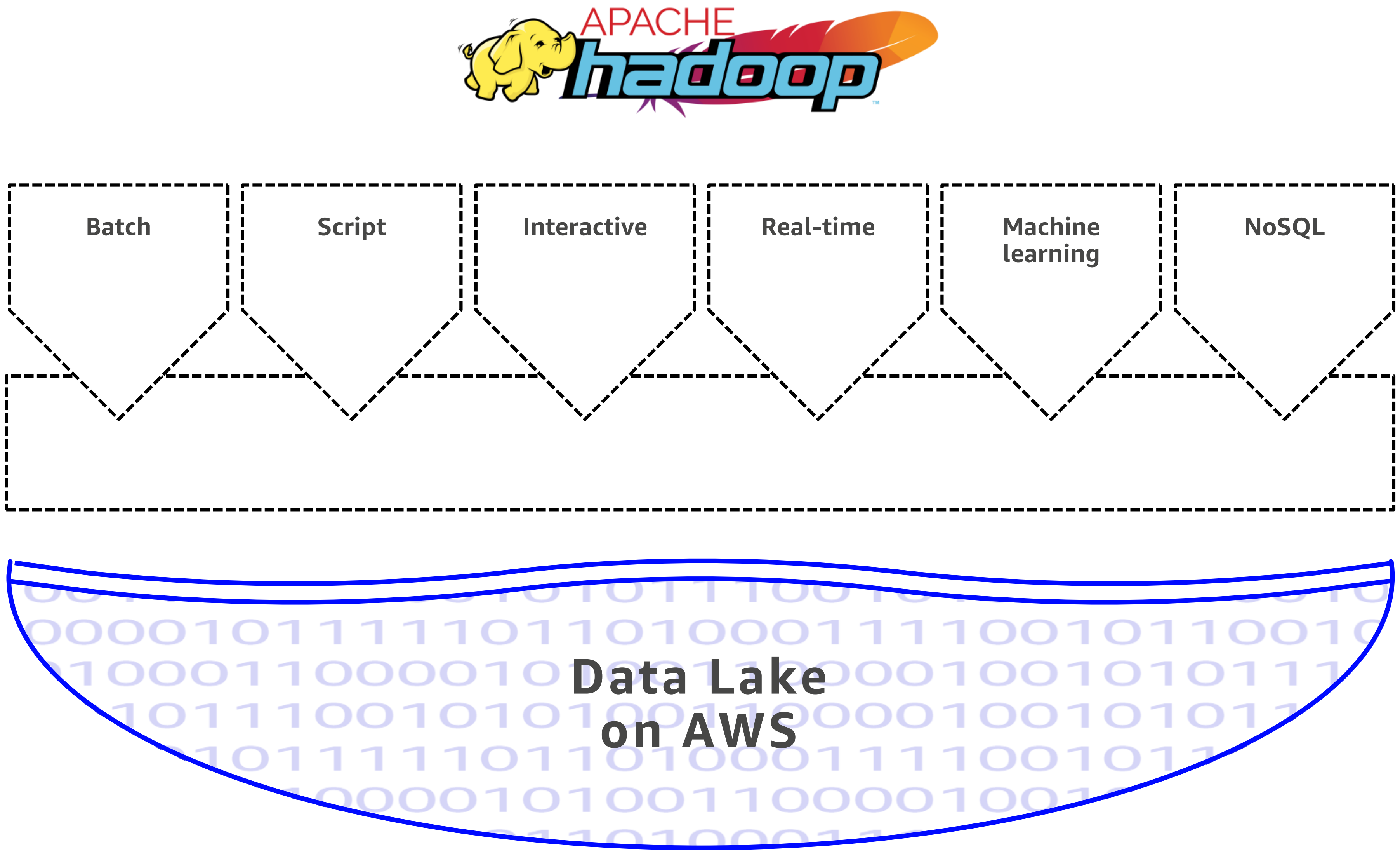
Data lake on Amazon S3



1. Crawlers scan your data sets and populate the AWS Glue Data Catalog

2. The AWS Glue Data Catalog serves as a central metadata repository

Agility - Hadoop/Spark analytics



- Distributed processing
- Diverse analytics
  - Batch/Script (Hive/Pig)
  - Interactive (Spark, Presto)
  - Real-time (Spark)
  - Machine learning (Spark)
  - NoSQL (HBase)
- For many use cases
  - Log and clickstream analysis
  - Machine learning
  - Real-time analytics
  - Large-scale analytics
  - Genomics
  - ETL

Let’s build!

Thank you!

Ridge XU
ridgexu@amazon.com

Anson Shen
ansons@amazon.com

!

Please complete the session survey in the mobile app.