AWS re:Invent

STG402

Querying data in place with Amazon S3 and analytics tools

John Mallory

Storage Business Development

Amazon Web Services

AWS re:Invent

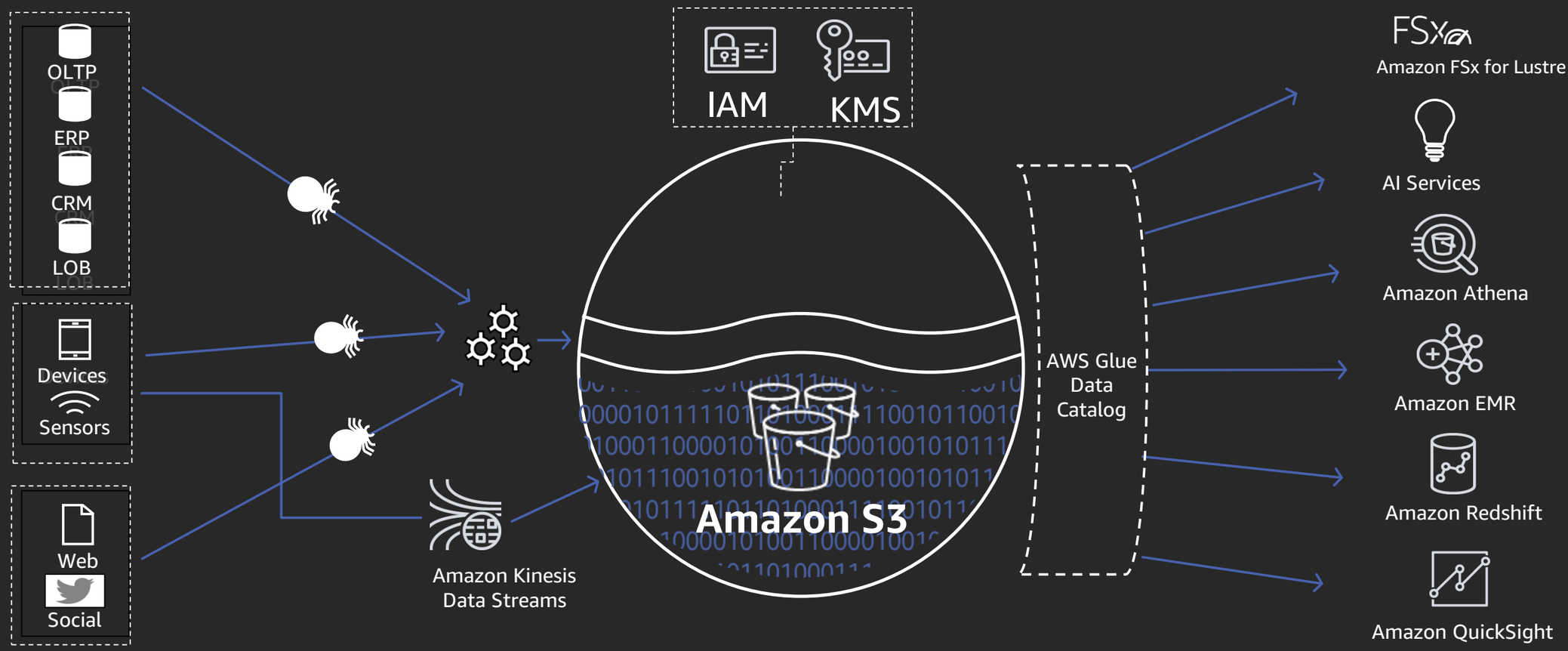
aws

How it works: Data lakes and analytics on AWS

Build data lakes quickly
- Identify, crawl, and catalog sources
- Ingest and clean data
- Transform into optimal formats

Simplify security management
- Enforce encryption
- Define access policies
- Implement audit login

Enable self-service and combined analytics
- Analysts discover all data available for analysis from a single data catalog
- Use multiple analytics tools over the same data



Focus: Reduced Time to Business Outcomes

Processing and querying in place

User-Defined Functions

AWS Lambda

Bring your own functions & code

Execute without provisioning servers

Fully Managed Process & Query

AWS Glue

Amazon Athena

Amazon Redshift

Amazon SageMaker

Catalog, transform & query data in Amazon S3

No physical instances to manage

Focus on Agility and Extracting Data Value

Process and query data in place on Amazon S3

Amazon Athena

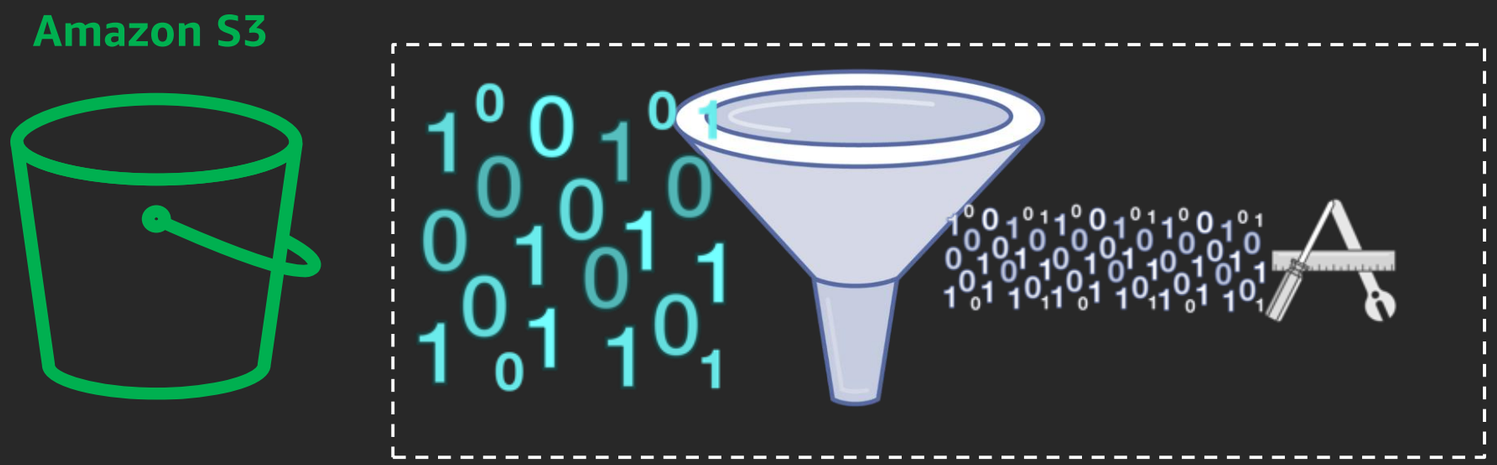
Amazon Redshift Spectrum

Amazon SageMaker

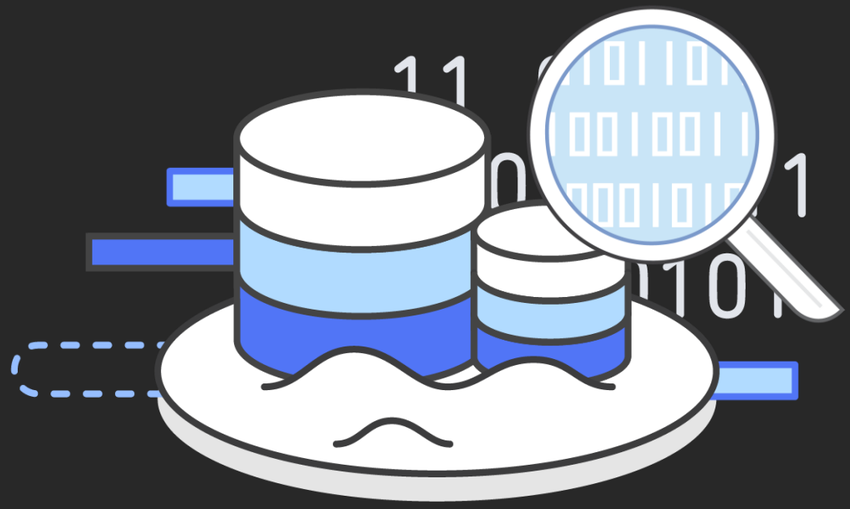
AWS Glue

Today: All of these tools…

retrieve a lot of data they don’t need from Amazon S3 and then scan and filter objects



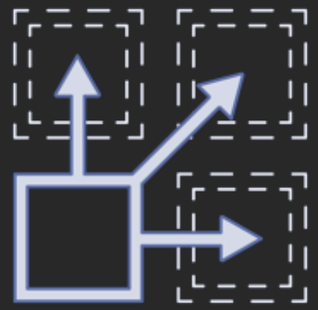
Amazon S3 Select changes the equation



Select a subset of an object’s data with a SQL expression

Amazon S3 scans and filters the object and returns matching results

Amazon S3 Select



Easy to use

Standard SQL expression

Integrated

Works just like a GET request

Open

AWS SDK, adopted for many AWS services and by many APN Partners

Amazon S3 Select usage

```
SELECT s.country, s.city from S3Object s where s.city = 'Seattle'
```

- Operates within the Amazon S3 system
- SQL Statement operates on a per-object basis
- Returns SQL query filtered results
- Supports CSV, JSON, and Parquet formats
- Integrated with Spark, Hive, and Presto on Amazon EMR
- Scan Range Selects: Up to 10x performance boost for large objects **NEW!**

```
SELECT * FROM s3object s WHERE s.Category = 'ASSAULT'
and s.PdDistrict = 'MISSION' and s.\"Date\" BETWEEN '2010-12-31' AND '2012-01-01'
```

Amazon S3 Select-Serverless Applications

Amazon S3

AWS Lambda

Amazon S3 Select

Amazon SNS

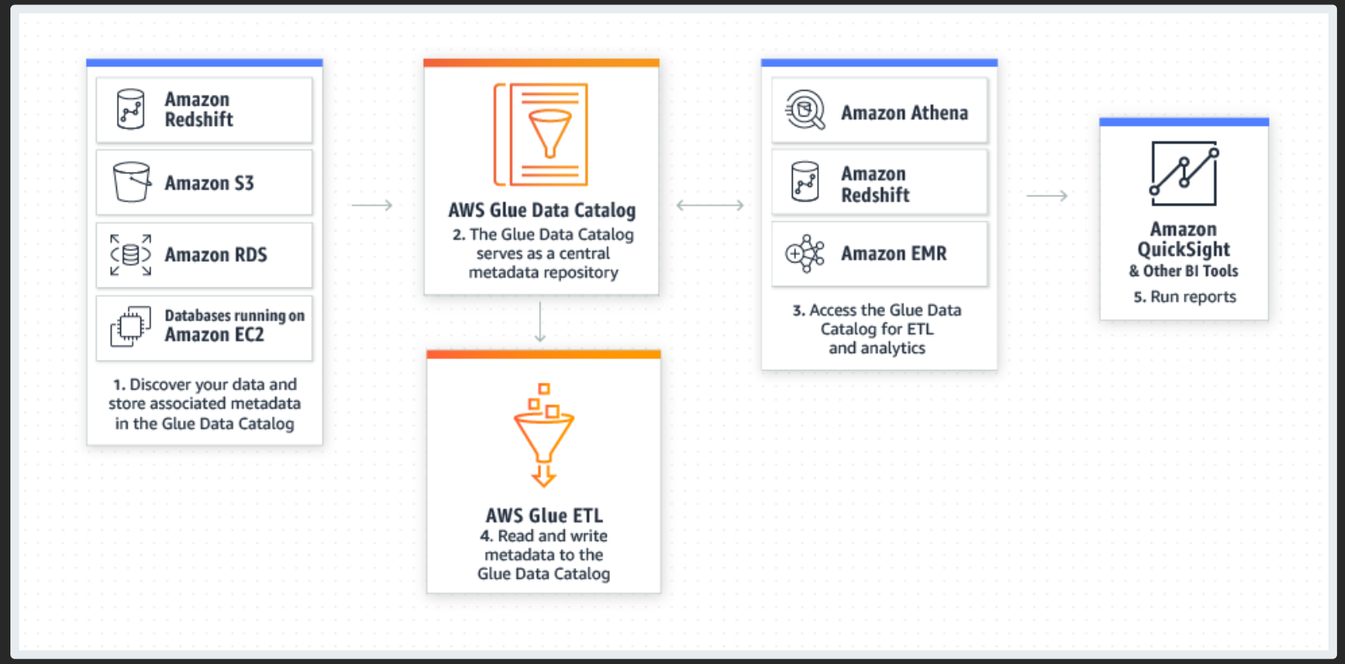
Set up a catalog, ETL, and data prep
with AWS Glue

Serverless provisioning, configuration, and scaling to run your ETL jobs on Apache Spark

Pay only for the resources used for jobs

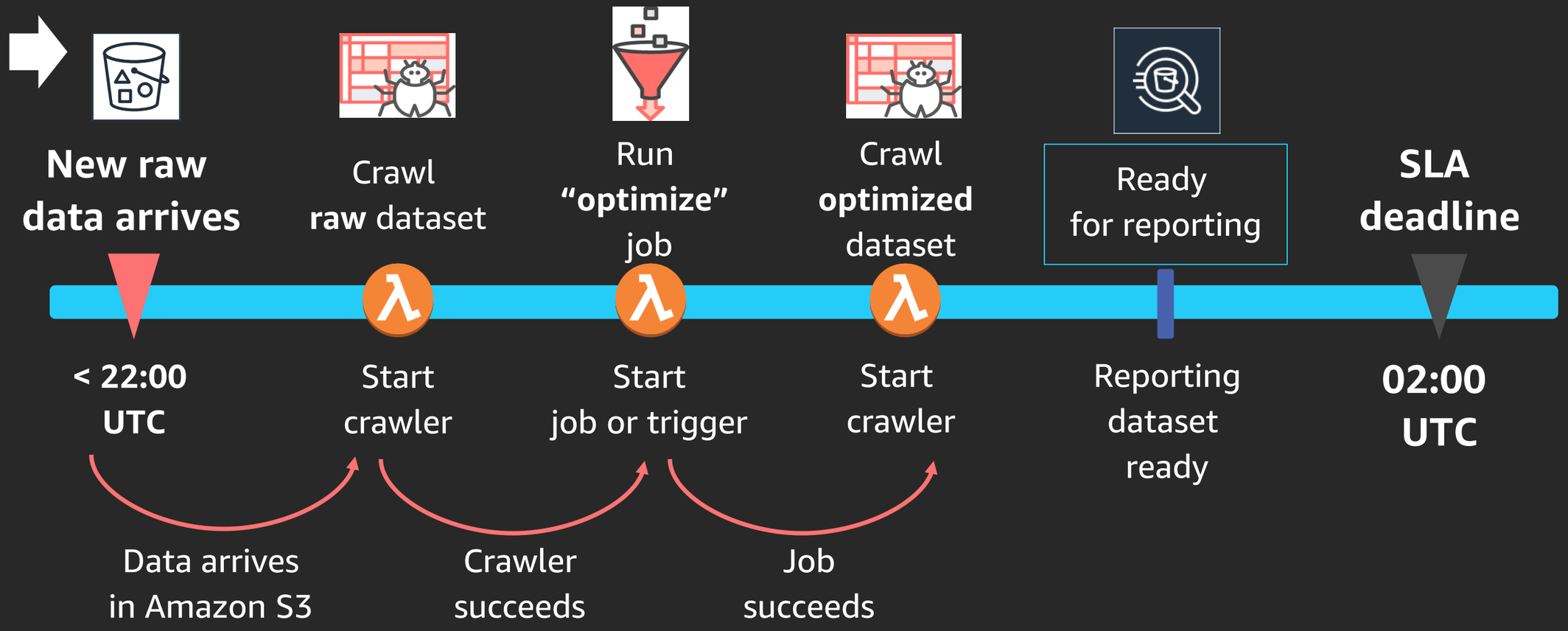
Crawl your data sources, identify data formats, and suggest schemas and transformations

Automates the effort in building, maintaining, and running ETL jobs



Event-driven batch ingest pipeline

Let **Amazon CloudWatch Events** and **AWS Lambda** drive the pipeline



Amazon Athena: Interactive Analysis

Interactive query service to analyze data in Amazon S3 using standard SQL

No infrastructure to set up or manage and no data to load

Query instantly

Zero setup cost; just point to Amazon S3 and start querying

Pay per query

Pay only for queries run; save 30–90% on per-query costs through compression

Open

ANSI SQL interface, JDBC/ODBC drivers, multiple formats, compression types, and complex joins and data types

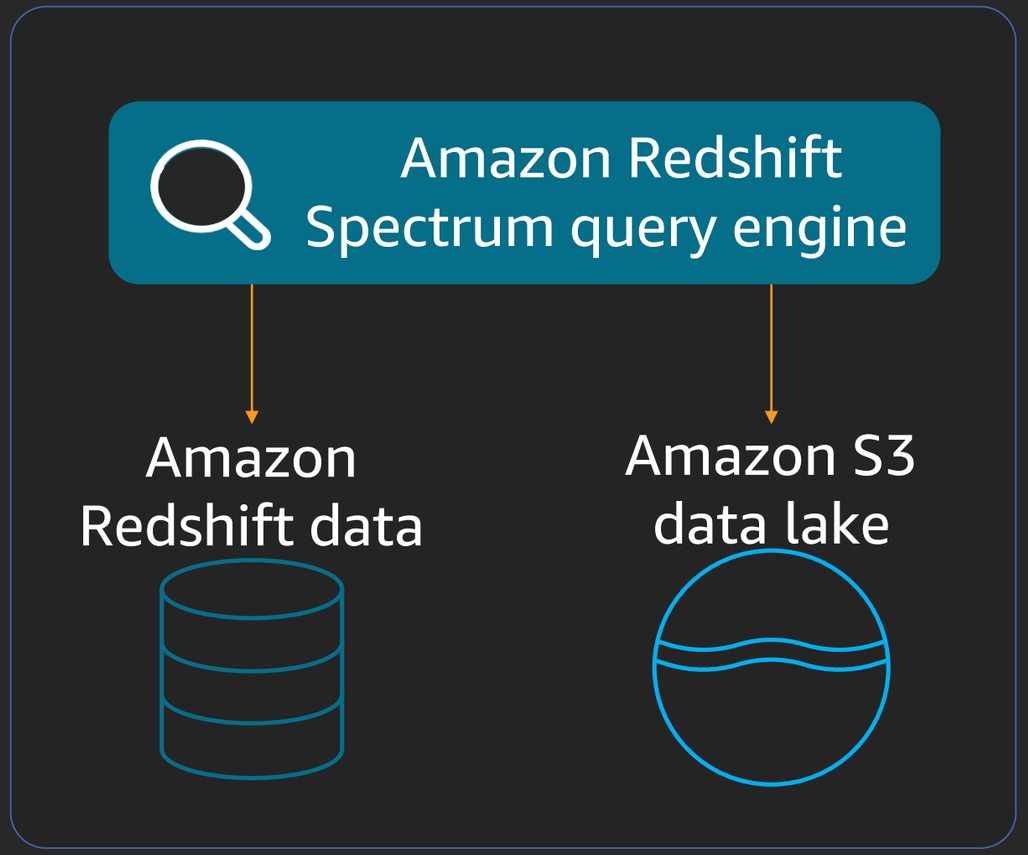
Easy

Serverless: Zero infrastructure, zero administration

Integrated with Amazon QuickSight

Amazon Redshift Spectrum: Data lake analytics

Query across your Amazon Redshift data warehouse and your Amazon S3 data lake



Run Amazon Redshift SQL queries against S3

Scale compute and storage separately

Fast query performance

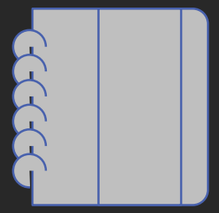
Unlimited concurrency

CSV, ORC, Grok, Avro & Parquet data formats

On demand, pay per query based on data scanned

Amazon SageMaker

The quickest and easiest way to get ML models from idea to production

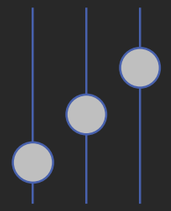


Pre-built notebooks for common problems

Built-in, high-performance algorithms

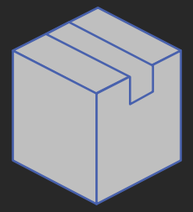
BUILD

One-click training

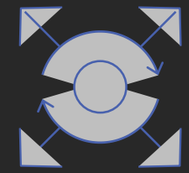


Automatic Model Tuning

TRAIN



One-click deployment



Fully managed hosting with auto-scaling

DEPLOY

Use Amazon Athena to filter S3 inventory reports

This query selects bucket, object key, and version id for unencrypted objects

```
select s._1, s._2, s._3 from s3object s where s._6 = 'NOT-SSE'
```

Example results:

batchoperationsdemo,0100059%7Ethumb.jpg,lsrtIxksLu0R0ZkYPL.LhgD5caTYn6vu

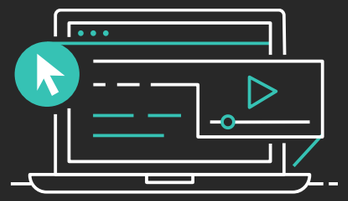
batchoperationsdemo,0100074%7Ethumb.jpg,sd2M60g6Fdazoi6D5kNARIE7KzUibmHR

batchoperationsdemo,0100075%7Ethumb.jpg,TLYESLnl1mXD5c4BwiOIinqFrktddkoL

Q&A

Learn storage with AWS Training and Certification

Resources created by the experts at AWS to help you build cloud storage skills



45+ free digital courses cover topics related to cloud storage, including:

- Amazon S3
- AWS Storage Gateway
- Amazon S3 Glacier
- Amazon Elastic File System (Amazon EFS)
- Amazon Elastic Block Store (Amazon EBS)

Classroom offerings, like Architecting on AWS, feature AWS expert instructors and hands-on activities

Visit aws.amazon.com/training/path-storage/

aws training and certification

Thank you!

AWS re:Invent

aws

Please complete the session survey in the mobile app.