FSV303

AWS re:INVENT

Queryable Archive + Data Lake

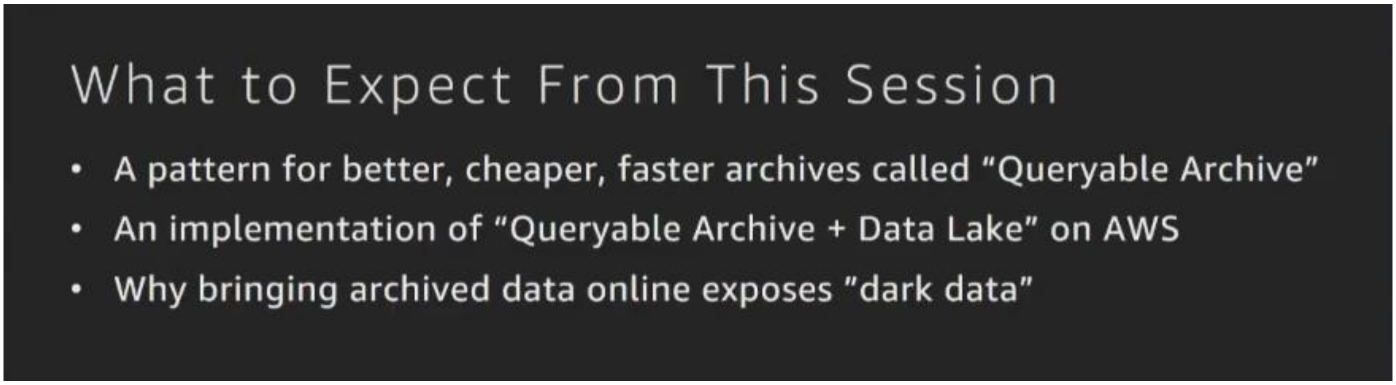
FSV303 - Building Queryable Archives and Data Lakes for Financial Services

George Smith, Global Financial Services Solutions Architect at AWS

November 27, 2017

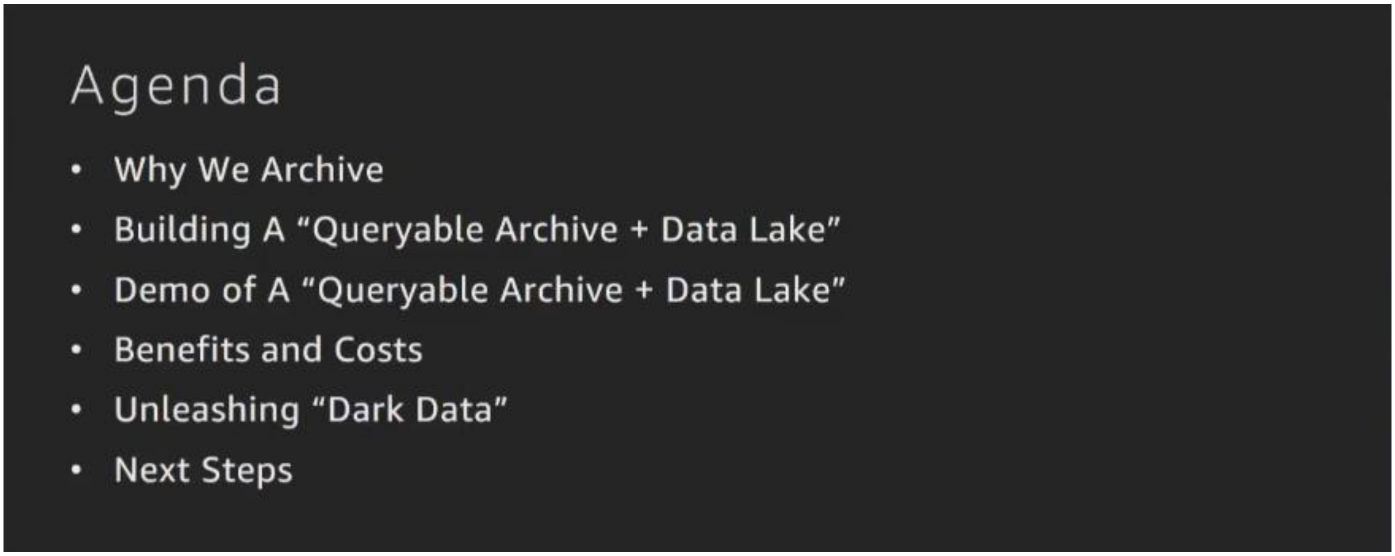
AWS re:Invent
aws

Financial institutions today must manage multiple data types from a wide variety of sources. Among these various data types, archive data presents a particular challenge: it is invisible to much of the organization and not easily leveraged by the lines of business for analytics, insight, and product innovation. Faced with massive volumes of archive data, Financial Services organizations are finding that delivering insights in a timely manner requires a data storage and analytics solution with more agility and flexibility than traditional data management systems can provide. In this session, we will discuss a design pattern that (1) brings this data into a highly available, lower-cost queryable archive within AWS than you currently have and (2) migrates that data to a data lake that the entire organization can use to extract insight and drive innovation. We will walk through a strategy that addresses the following topics: storing archive data in compressed, cost-effective, and readily available formats; creating lifecycle policies to archive older data sets and make them easily accessible; fully utilizing the features of object storage to enrich the data lake; and applying AWS analytics tools to gather business insights.



What to Expect From This Session

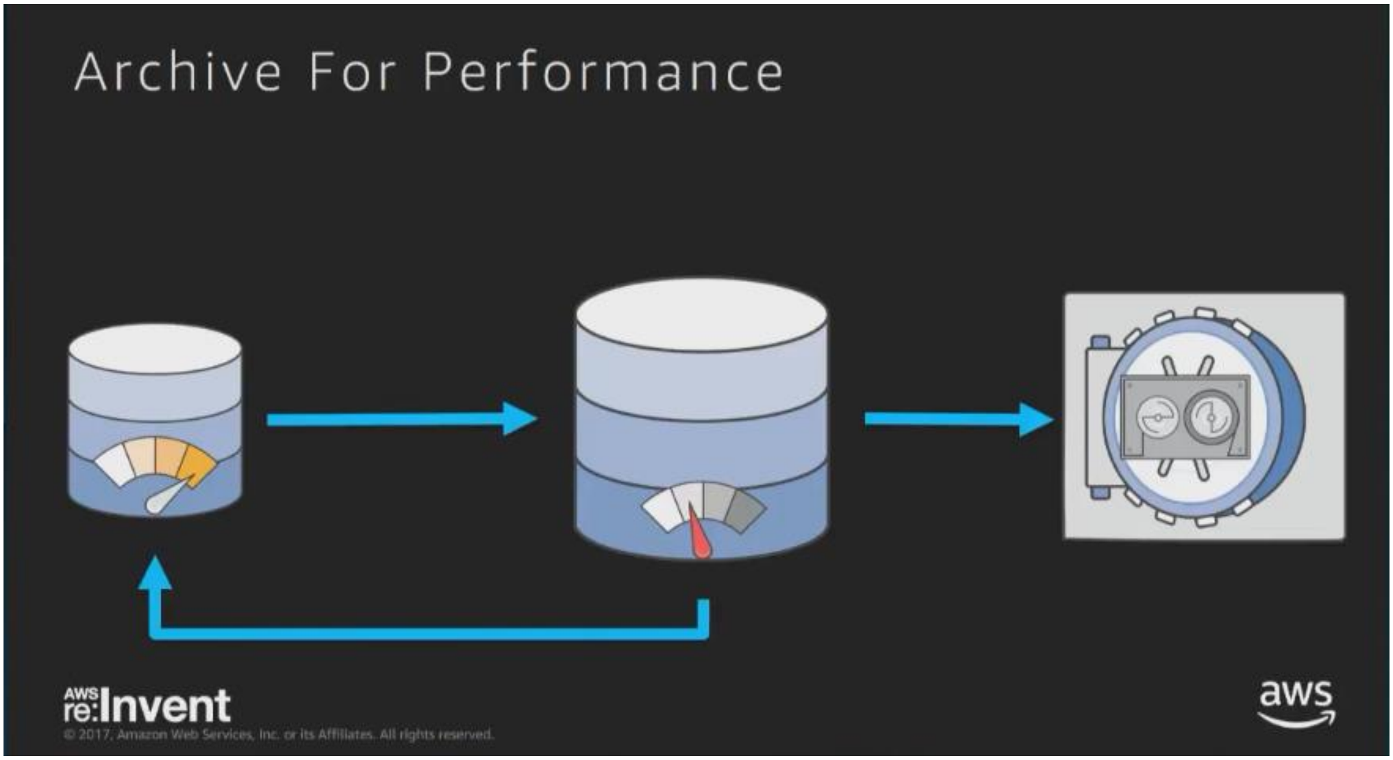
- A pattern for better, cheaper, faster archives called “Queryable Archive”
- An implementation of “Queryable Archive + Data Lake” on AWS
- Why bringing archived data online exposes ”dark data”



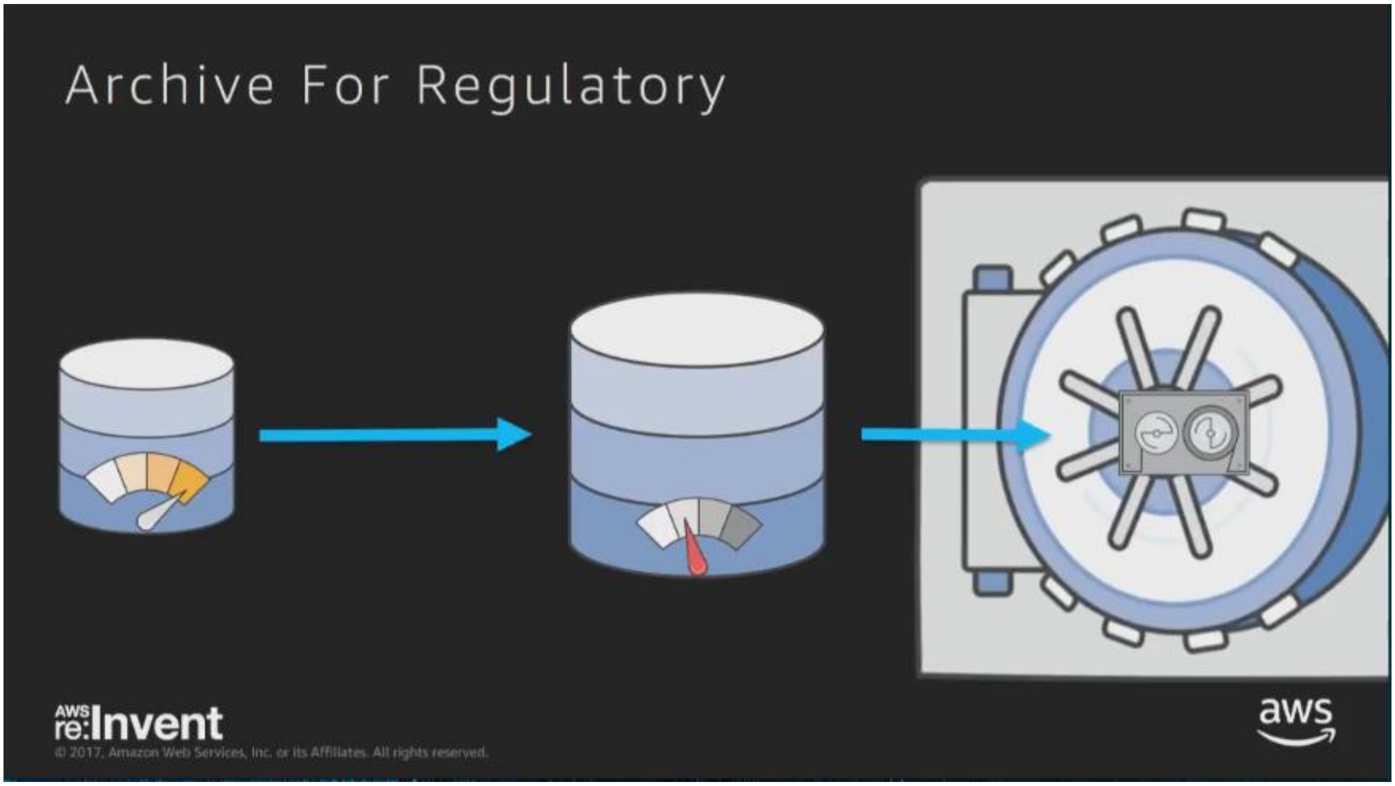
Agenda

- Why We Archive
- Building A “Queryable Archive + Data Lake”
- Demo of A “Queryable Archive + Data Lake”
- Benefits and Costs
- Unleashing “Dark Data”
- Next Steps

Why We Archive

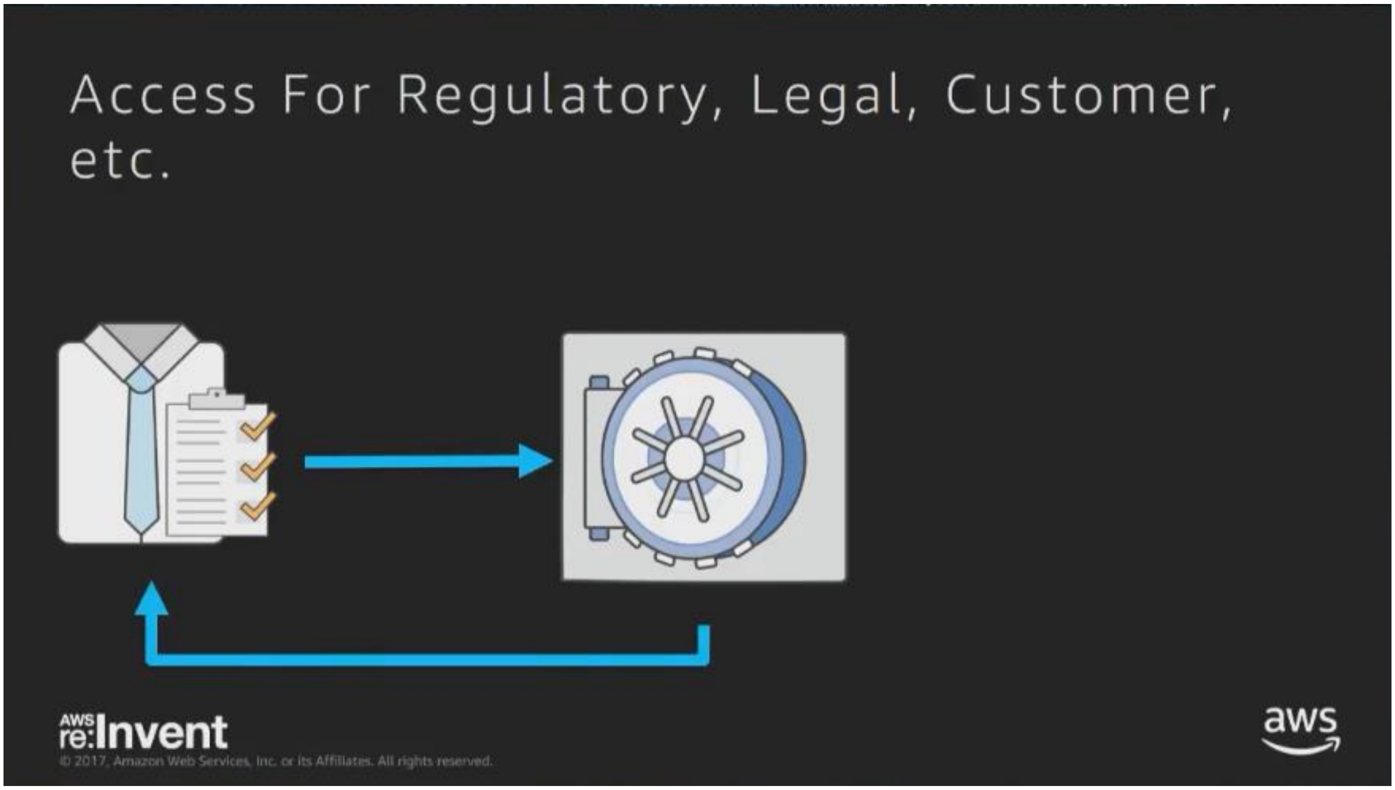


Archive For Performance

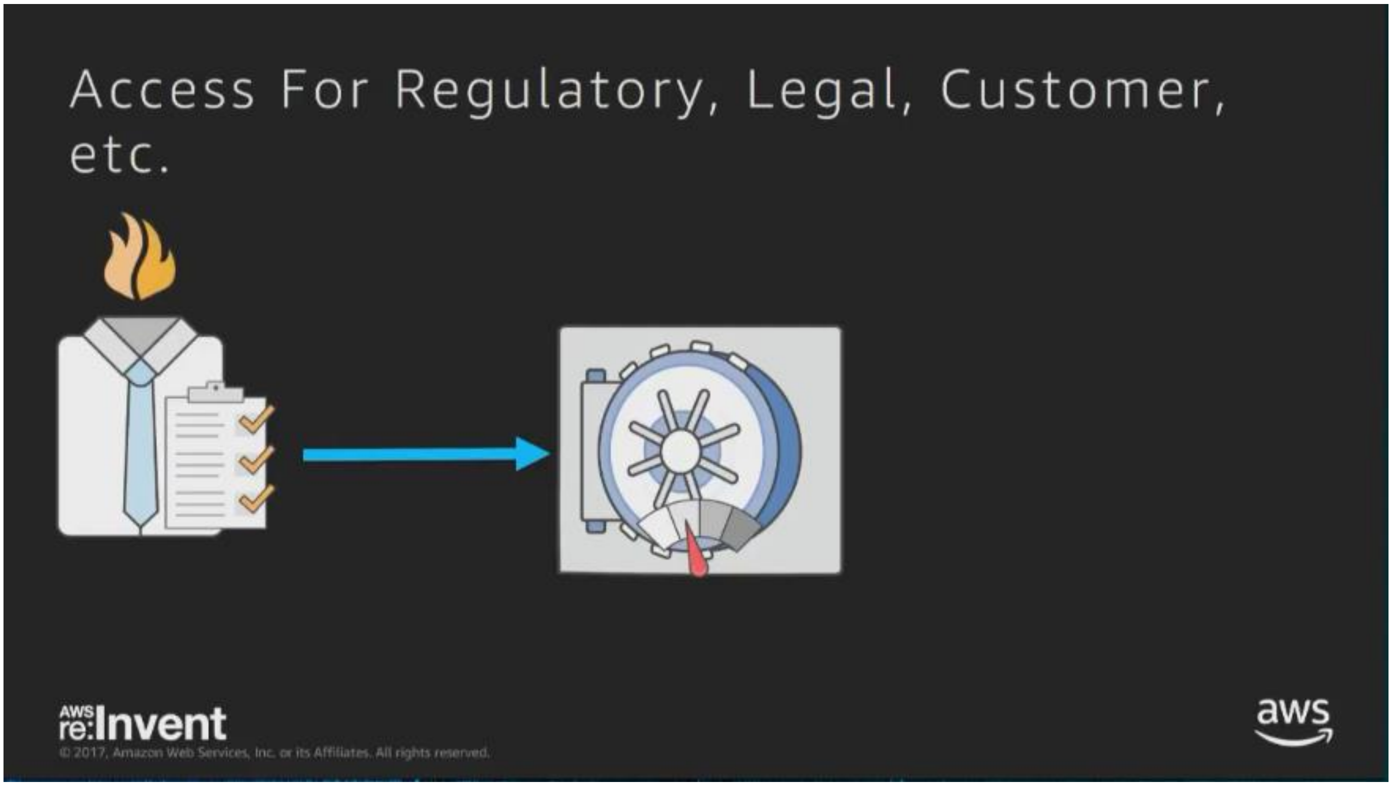


Archive For Regulatory

We now have way more storage attached to compute which is not what we intended during design phase

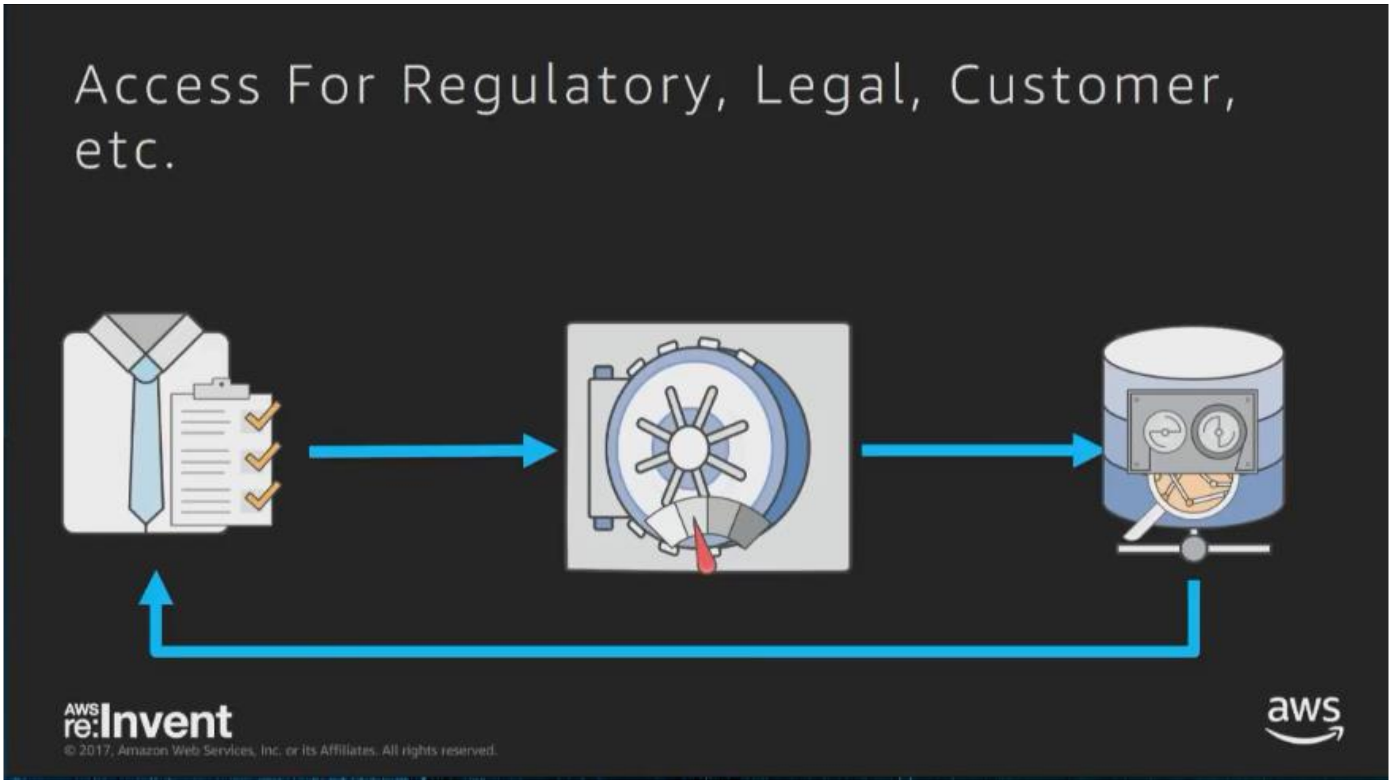


Access For Regulatory, Legal, Customer, etc.



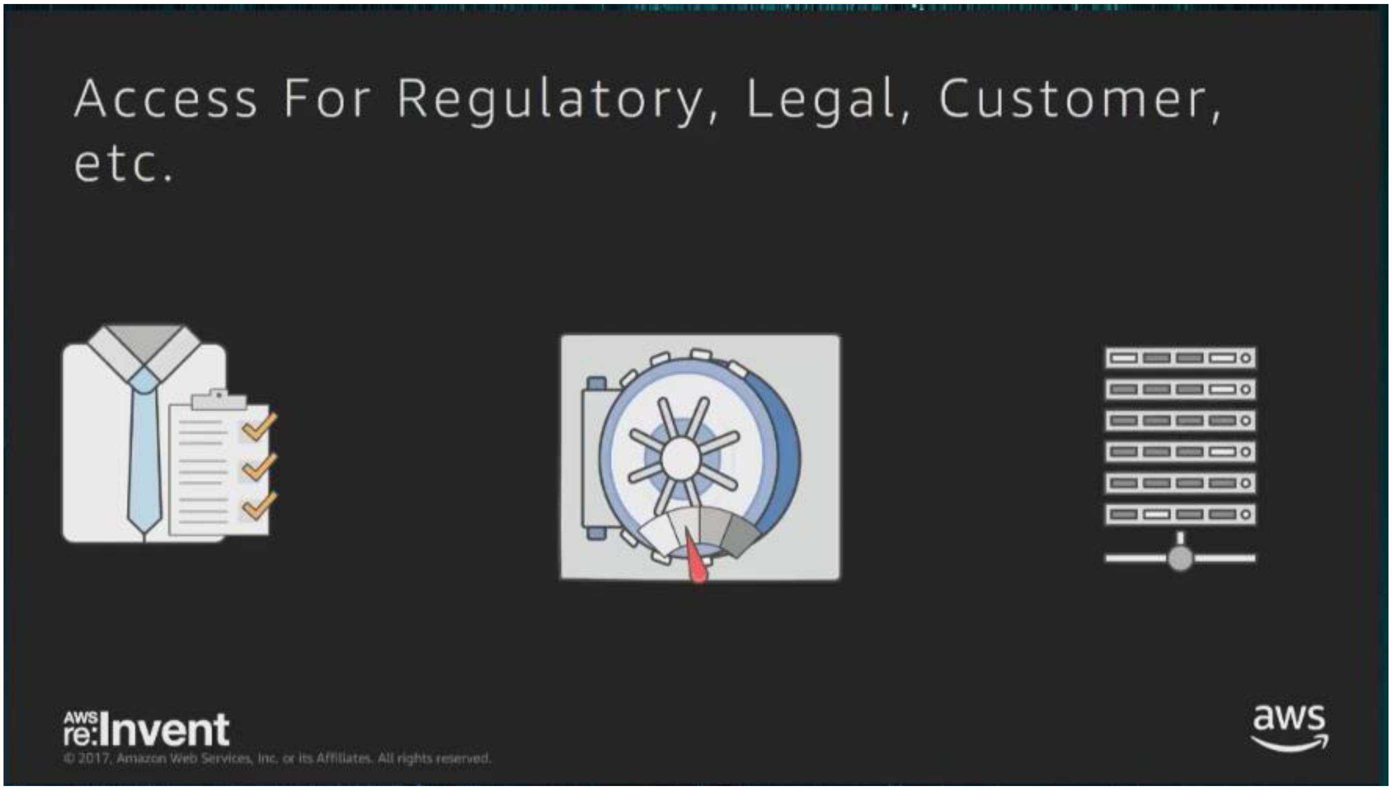
Access For Regulatory, Legal, Customer, etc.

we now have an end user who needs access to data and can’t get it



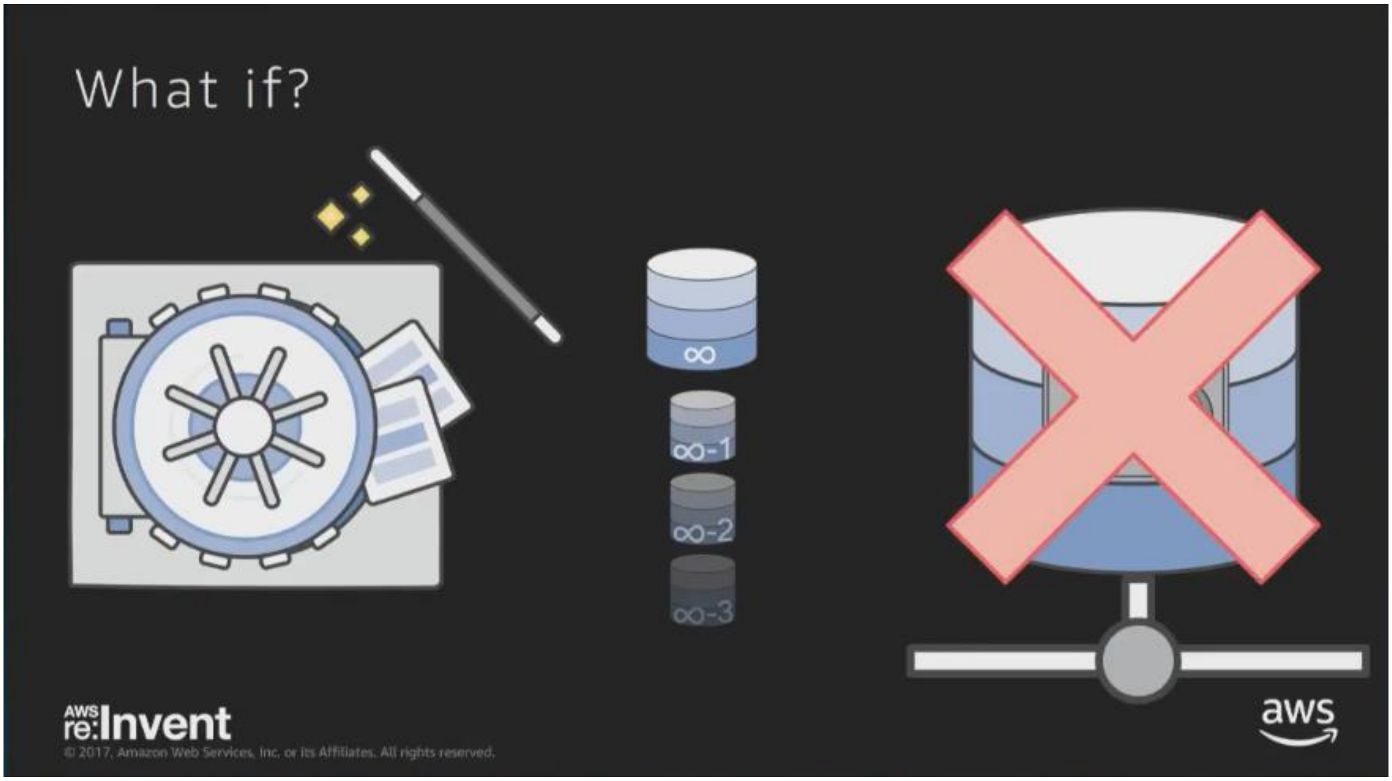
Access For Regulatory, Legal, Customer, etc.

We have to set up spare database and copy the archives into it, then get the date out for the customer



Access For Regulatory, Legal, Customer, etc.

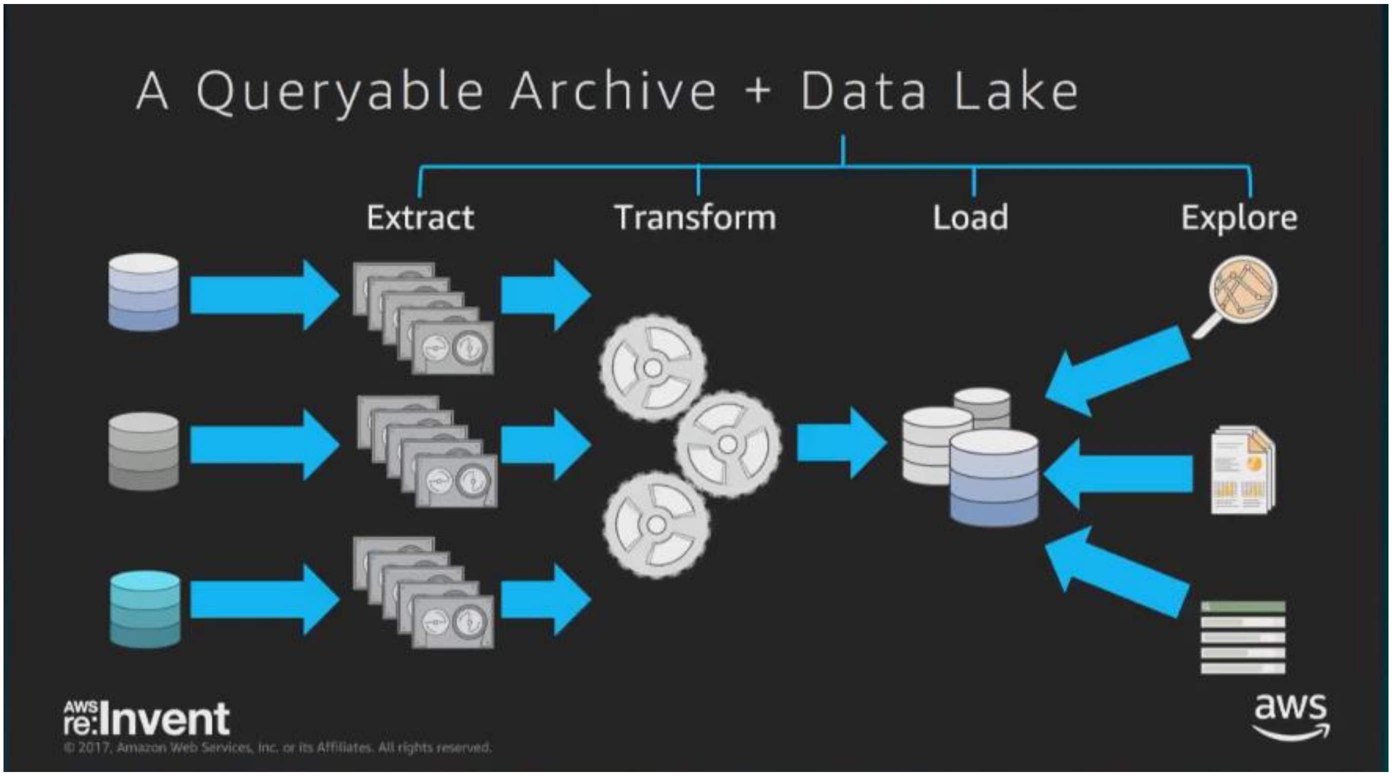
We will then tear everything down again



What if?

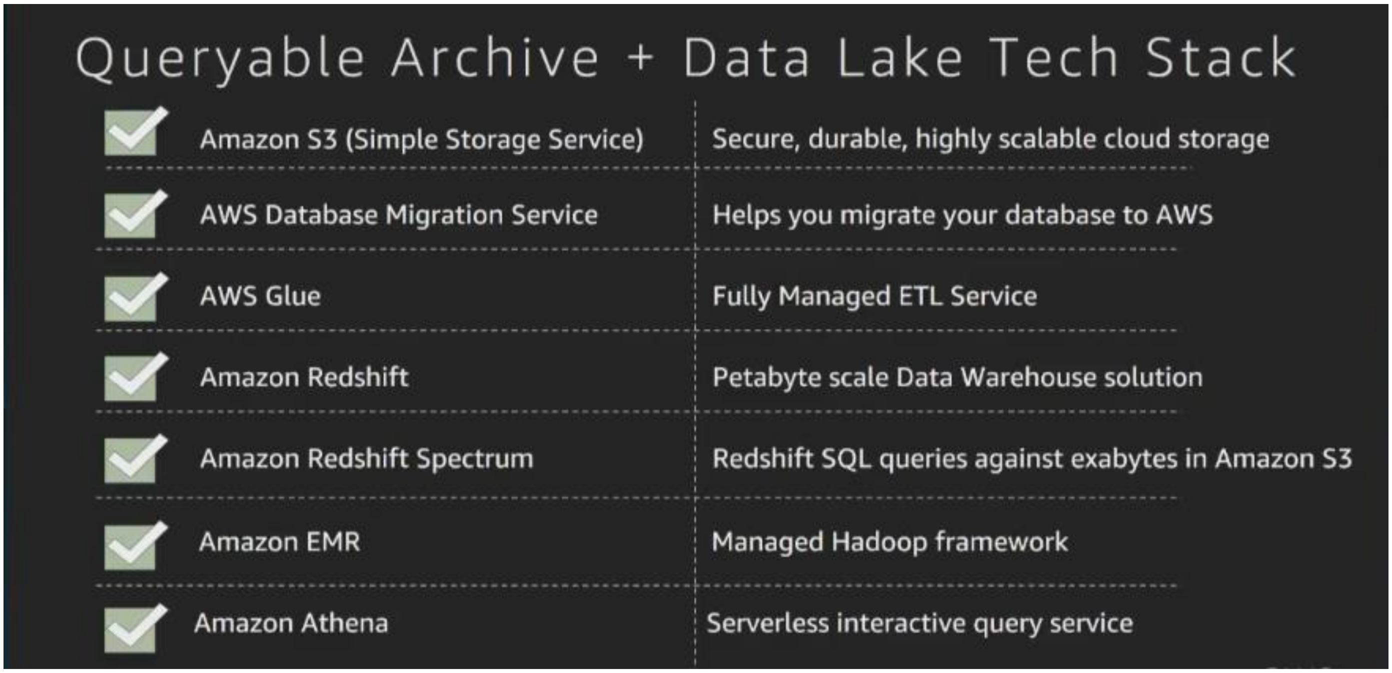
Building A "Queryable Archive + Data Lake"*

* (Without breaking the bank!)



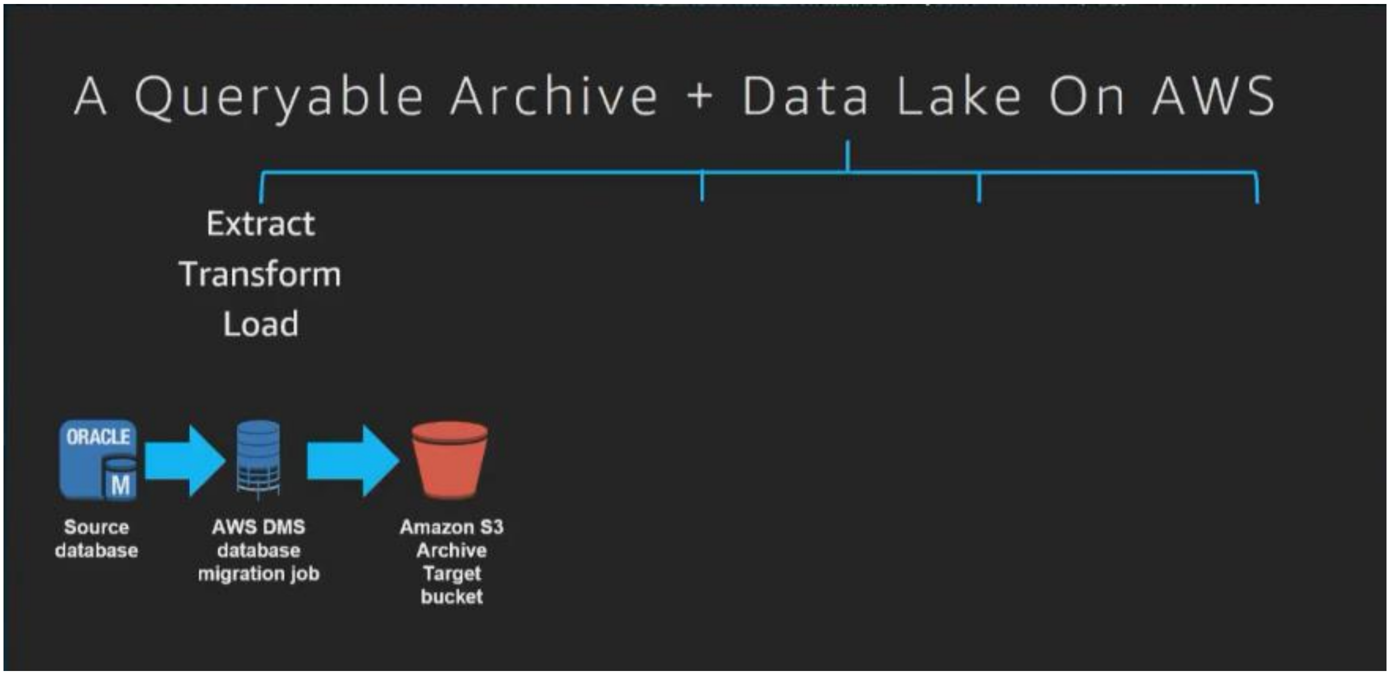
A Queryable Archive + Data Lake

We start out with all our historical financial datasets that are sitting in our databases in data centers today. Once the data is loaded into the specific AWS database services, we have a variety of tools to start exploring that data



Queryable Archive + Data Lake Tech Stack

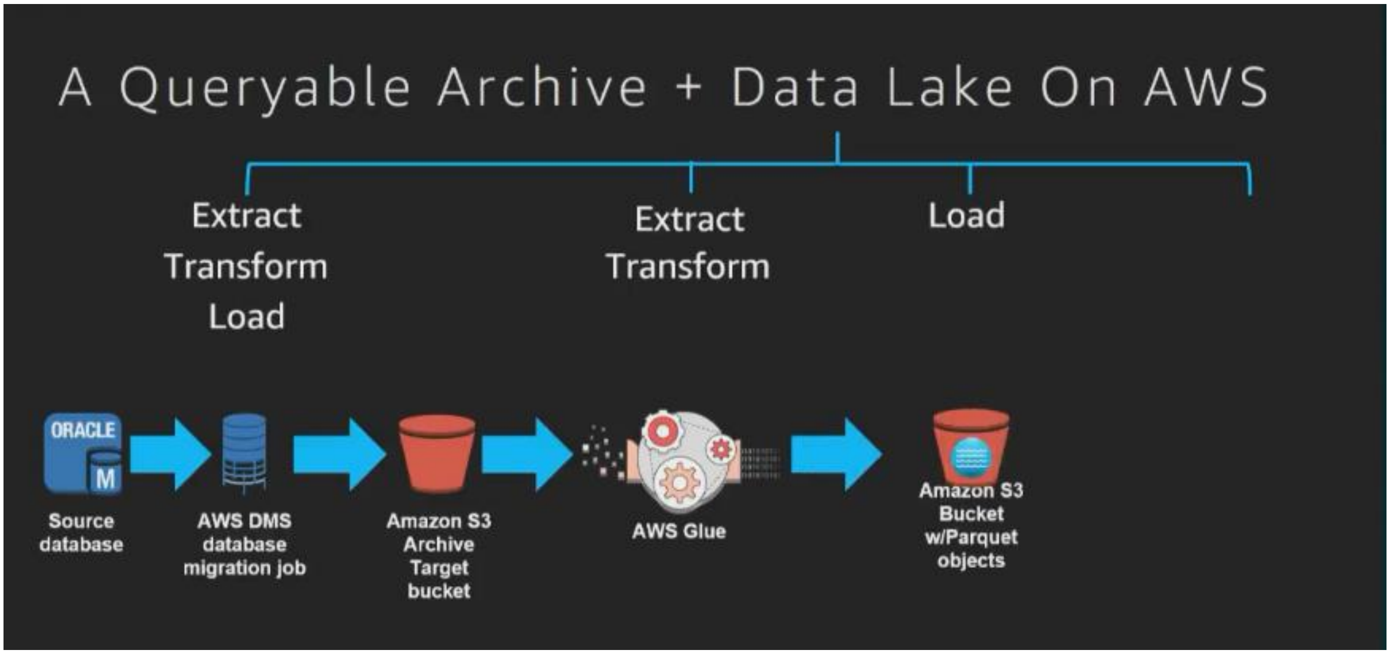
✓	Amazon S3 (Simple Storage Service)	Secure, durable, highly scalable cloud storage
✓	AWS Database Migration Service	Helps you migrate your database to AWS
✓	AWS Glue	Fully Managed ETL Service
✓	Amazon Redshift	Petabyte scale Data Warehouse solution
✓	Amazon Redshift Spectrum	Redshift SQL queries against exabytes in Amazon S3
✓	Amazon EMR	Managed Hadoop framework
✓	Amazon Athena	Serverless interactive query service



A Queryable Archive + Data Lake On AWS

Extract
Transform
Load
ORACLE
M
Source database
AWS DMS database migration job
Amazon S3 Archive Target bucket

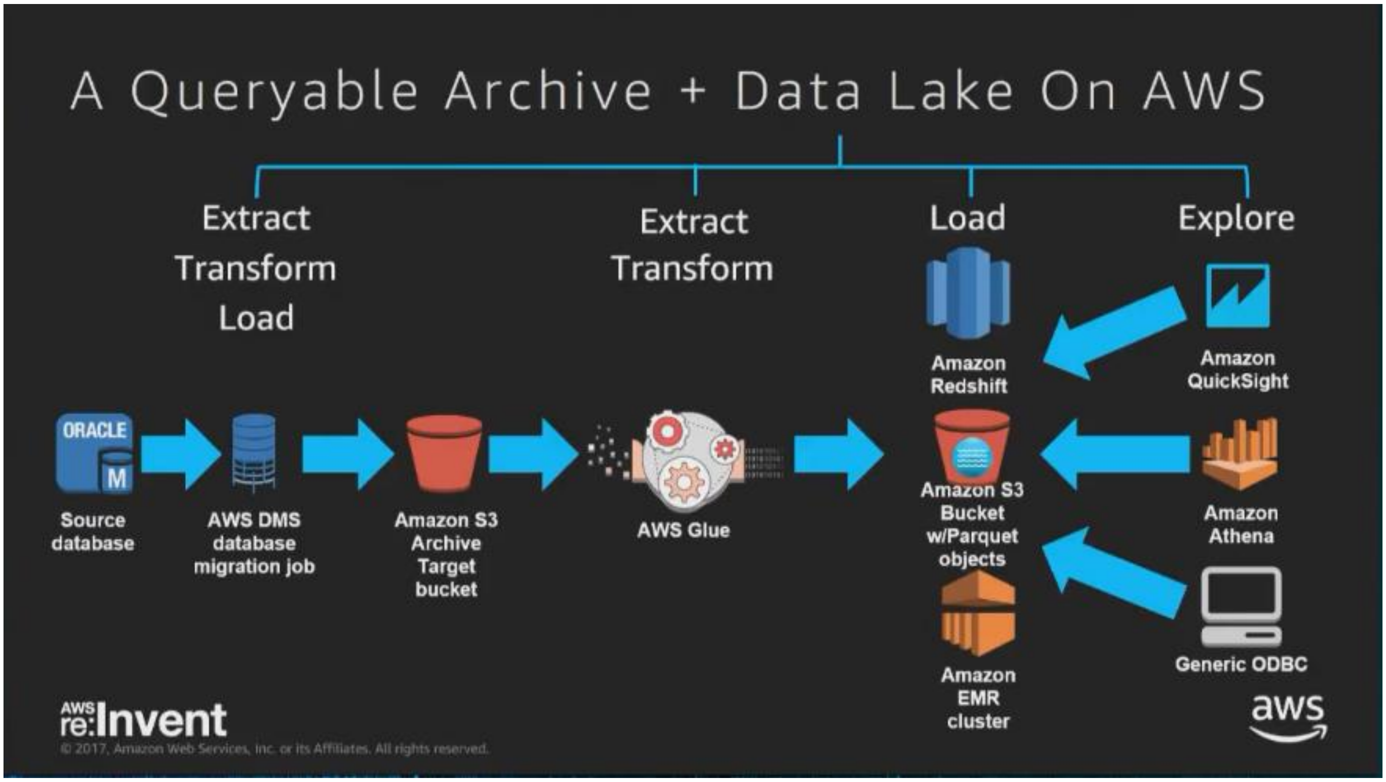
We use the AWS Data Migration Service and now have the data in S3 in a compressed CSV format.



A Queryable Archive + Data Lake On AWS

Extract
Transform
Load
Extract
Transform
Load
ORACLE
M
Source database
AWS DMS database migration job
Amazon S3 Archive Target bucket
AWS Glue
Amazon S3 Bucket w/Parquet objects

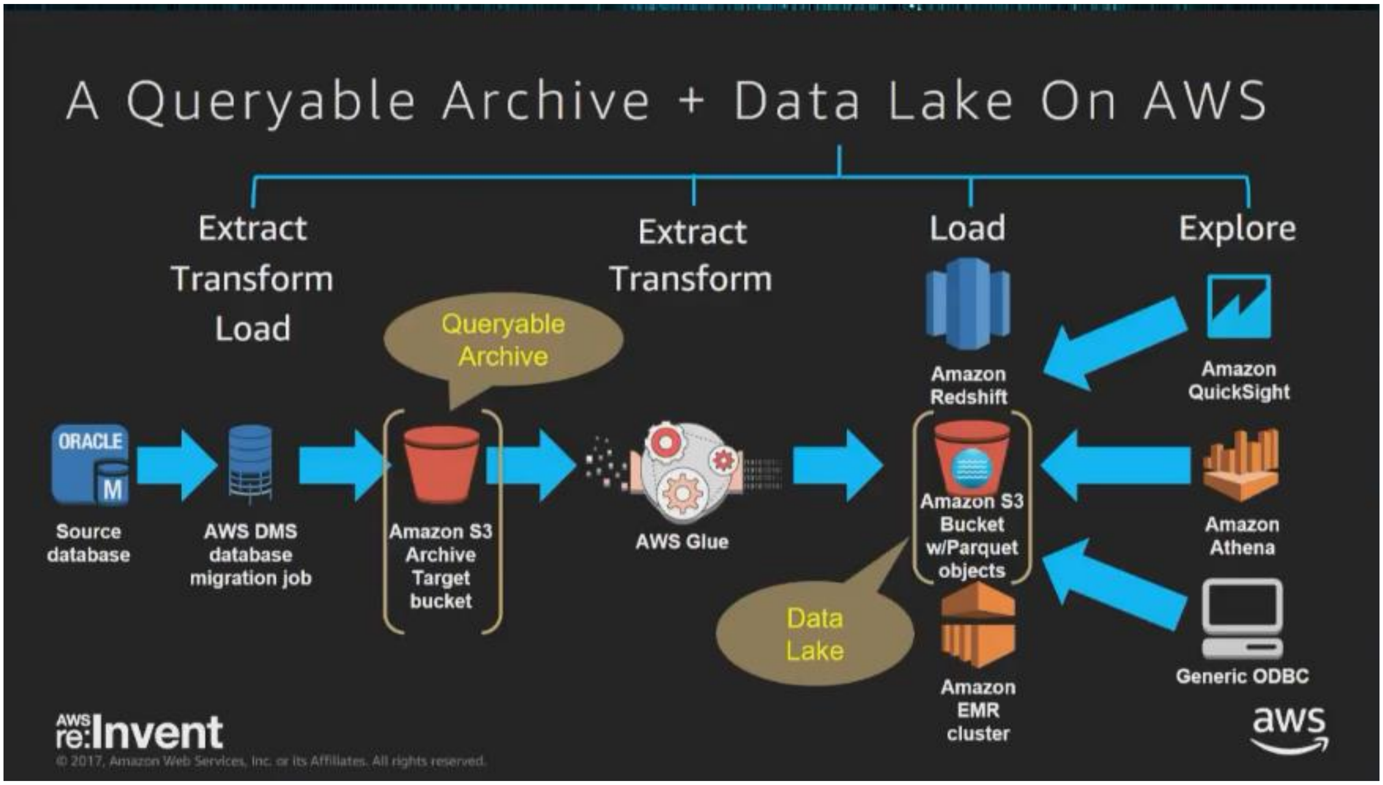
We now again use AWS Glue to transform the data into our enterprise data format and store the data in S3 in the Parquet object format



A Queryable Archive + Data Lake On AWS

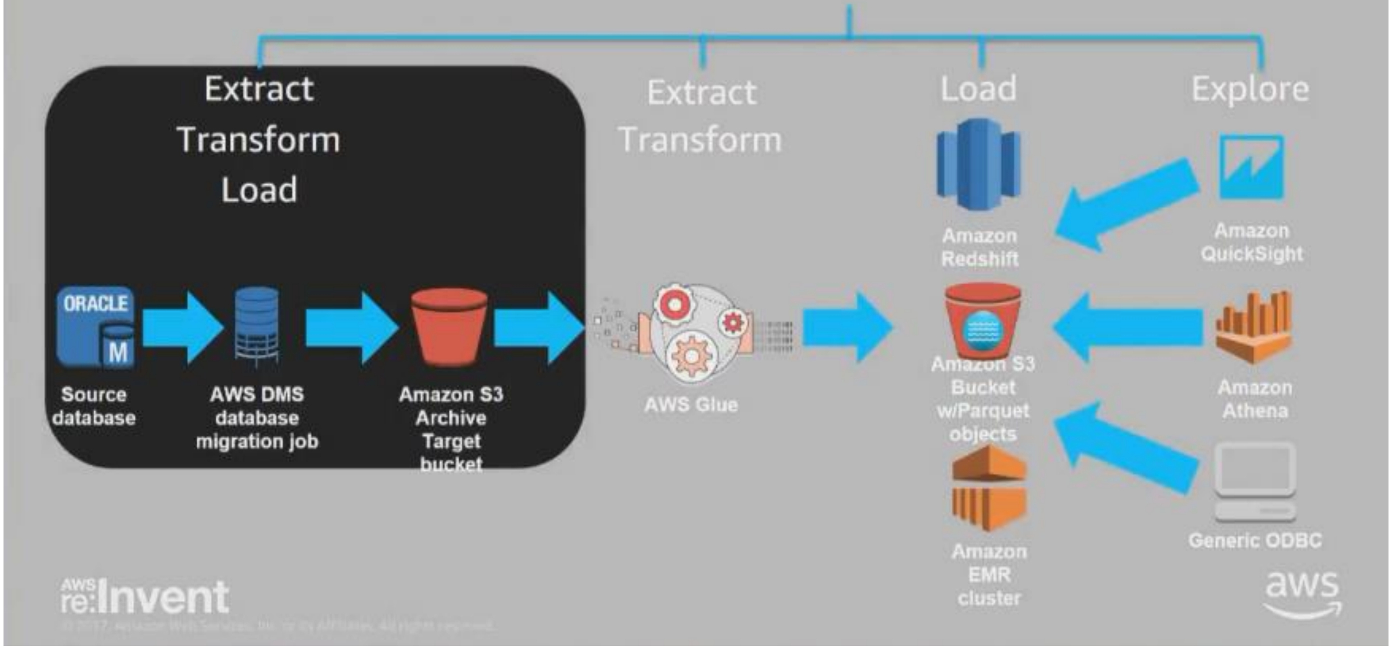
We can now access and explore the data

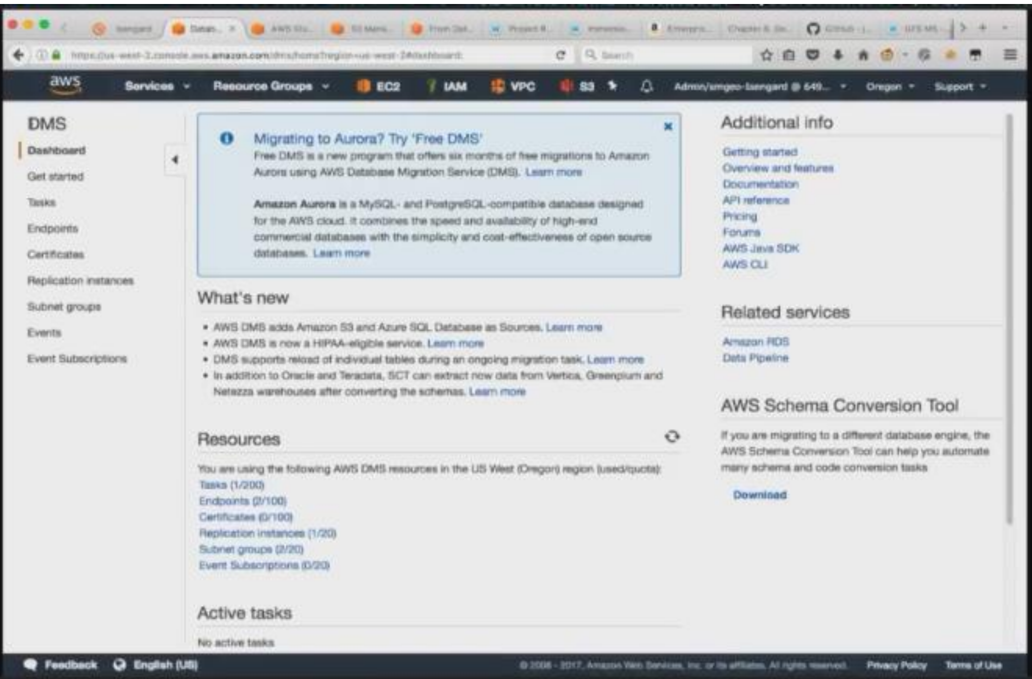
Demo of a "Queryable Archive + Data Lake"



A Queryable Archive + Data Lake On AWS

A Queryable Archive + Data Lake On AWS





DMS

Migrating to Aurora? Try 'Free DMS'

Free DMS is a new program that offers six months of free migrations to Amazon Aurora using AWS Database Migration Service (DMS). Learn more

Amazon Aurora is a MySQL- and PostgreSQL-compatible database designed for the AWS cloud. It combines the speed and availability of high-end commercial databases with the simplicity and cost-effectiveness of open source databases. Learn more

What's new

- AWS DMS adds Amazon S3 and Azure SQL Database as Sources. Learn more
- AWS DMS is now a HIPAA-eligible service. Learn more
- DMS supports reload of individual tables during an ongoing migration task. Learn more
- In addition to Oracle and Teradata, SCT can extract now data from Vertica, Greenplum and Netezza warehouses after converting the schemas. Learn more

Resources

You are using the following AWS DMS resources in the US West (Oregon) region (used/quota):

Tasks (1/200)
Endpoints (2/100)
Certificates (0/100)
Replication instances (1/20)
Subnet groups (2/20)
Event Subscriptions (0/20)

Active tasks

No active tasks

Additional info

Getting started
Overview and features
Documentation
API reference
Pricing
Forums
AWS Java SDK
AWS CLI

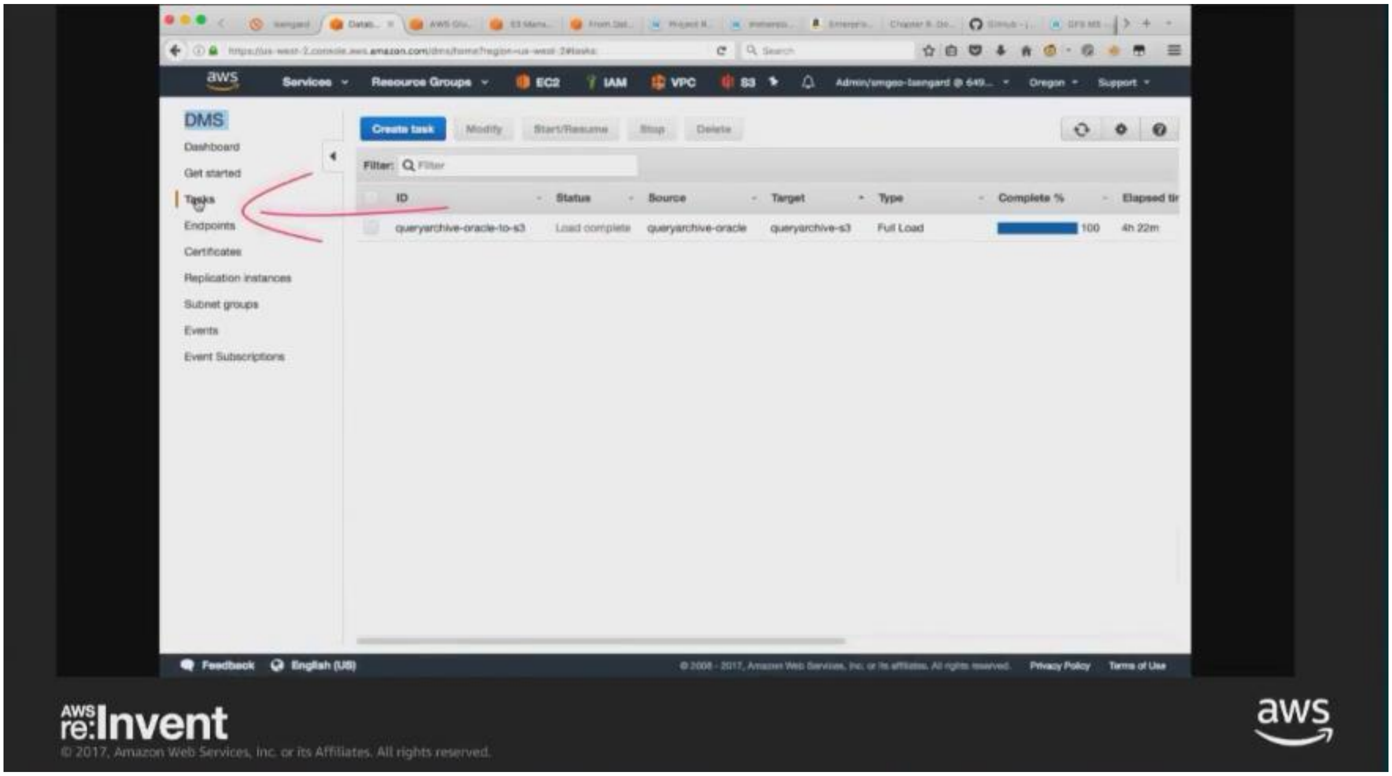
Related services

Amazon RDS
Data Pipeline

AWS Schema Conversion Tool

If you are migrating to a different database engine, the AWS Schema Conversion Tool can help you automate many schema and code conversion tasks

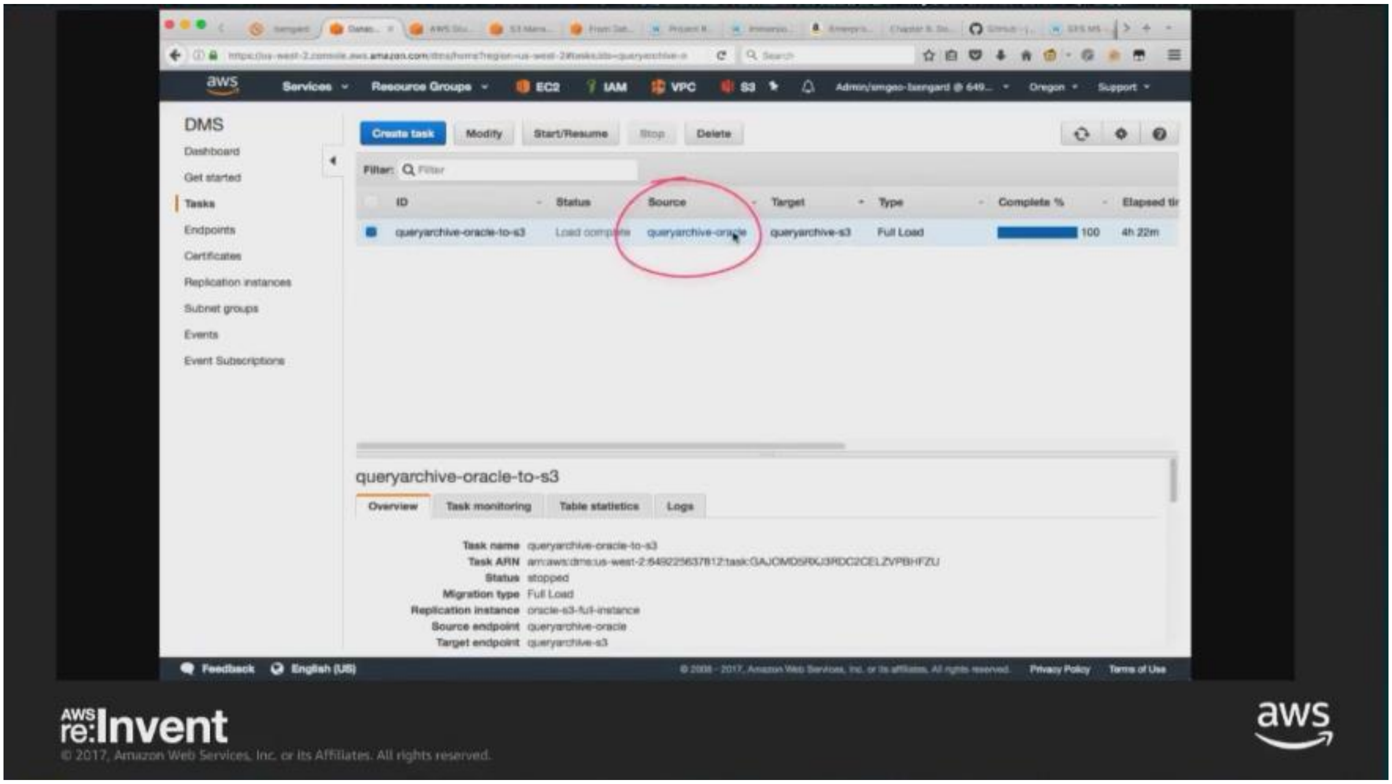
Download



	ID	Status	Source	Target	Type	Complete %	Elapsed tir
□	queryarchive-oracle-to-s3	Load complete	queryarchive-oracle	queryarchive-s3	Full Load	100	4h 22m

© 2008 - 2017, Amazon Web Services, Inc. or its affiliates. All rights reserved.

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



	ID	Status	Source	Target	Type	Complete %	Elapsed tir
■	queryarchive-oracle-to-s3	Load complete	queryarchive-oracle	queryarchive-s3	Full Load	100	4h 22m

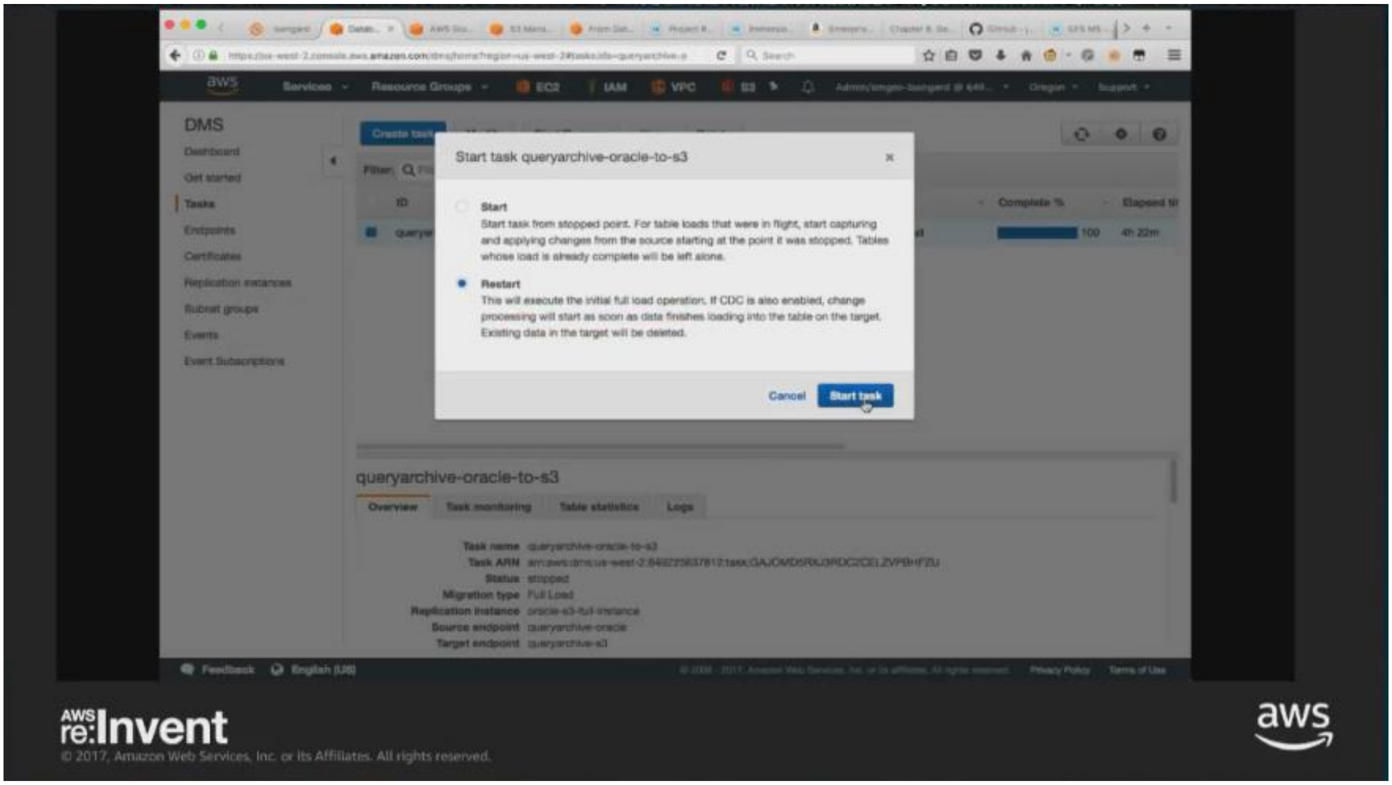
queryarchive-oracle-to-s3

Task name queryarchive-oracle-to-s3
Task ARN arn:aws:dms:us-west-2:649225637812:task:GAJOMDSRXJ3RDC2CELZVPBHFZU
Status stopped
Migration type Full Load
Replication instance oracle-s3-full-instance
Source endpoint queryarchive-oracle
Target endpoint queryarchive-s3

© 2008 - 2017, Amazon Web Services, Inc. or its affiliates. All rights reserved.

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

A Task is a batch job that moves data from one point to another



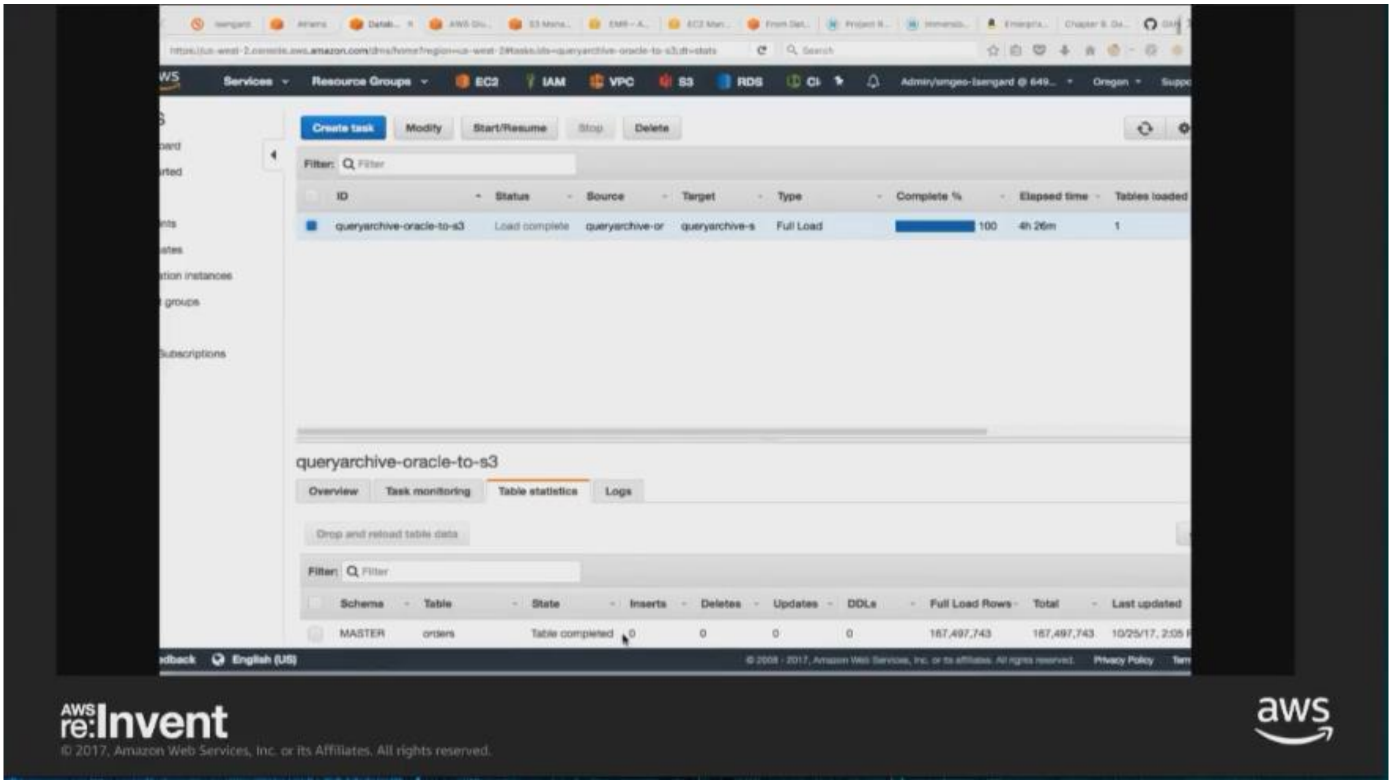
Start task queryarchive-oracle-to-s3

○ **Start**
Start task from stopped point. For table loads that were in flight, start capturing and applying changes from the source starting at the point it was stopped. Tables whose load is already complete will be left alone.

● **Restart**
This will execute the initial full load operation. If CDC is also enabled, change processing will start as soon as data finishes loading into the table on the target. Existing data in the target will be deleted.

queryarchive-oracle-to-s3

Task name	queryarchive-oracle-to-s3
Task ARN	arn:aws:dms:us-west-2:649225637812:task:GAJOMD5RIU3RDC2CELZVPBHFZU
Status	stopped
Migration type	Full Load
Replication instance	oracle-s3-full-instance
Source endpoint	queryarchive-oracle
Target endpoint	queryarchive-s3

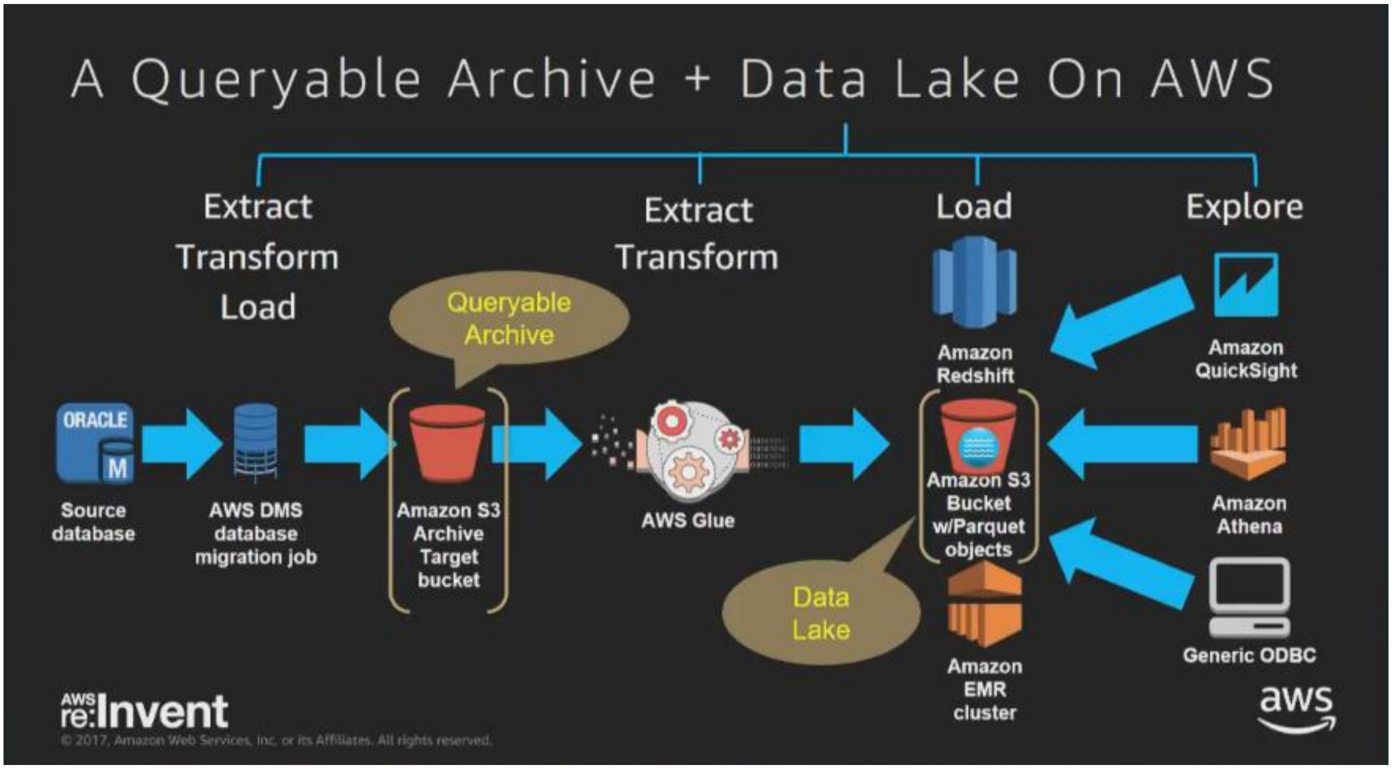


ID	Status	Source	Target	Type	Complete %	Elapsed time	Tables loaded
queryarchive-oracle-to-s3	Load complete	queryarchive-or	queryarchive-s	Full Load	100	4h 26m	1

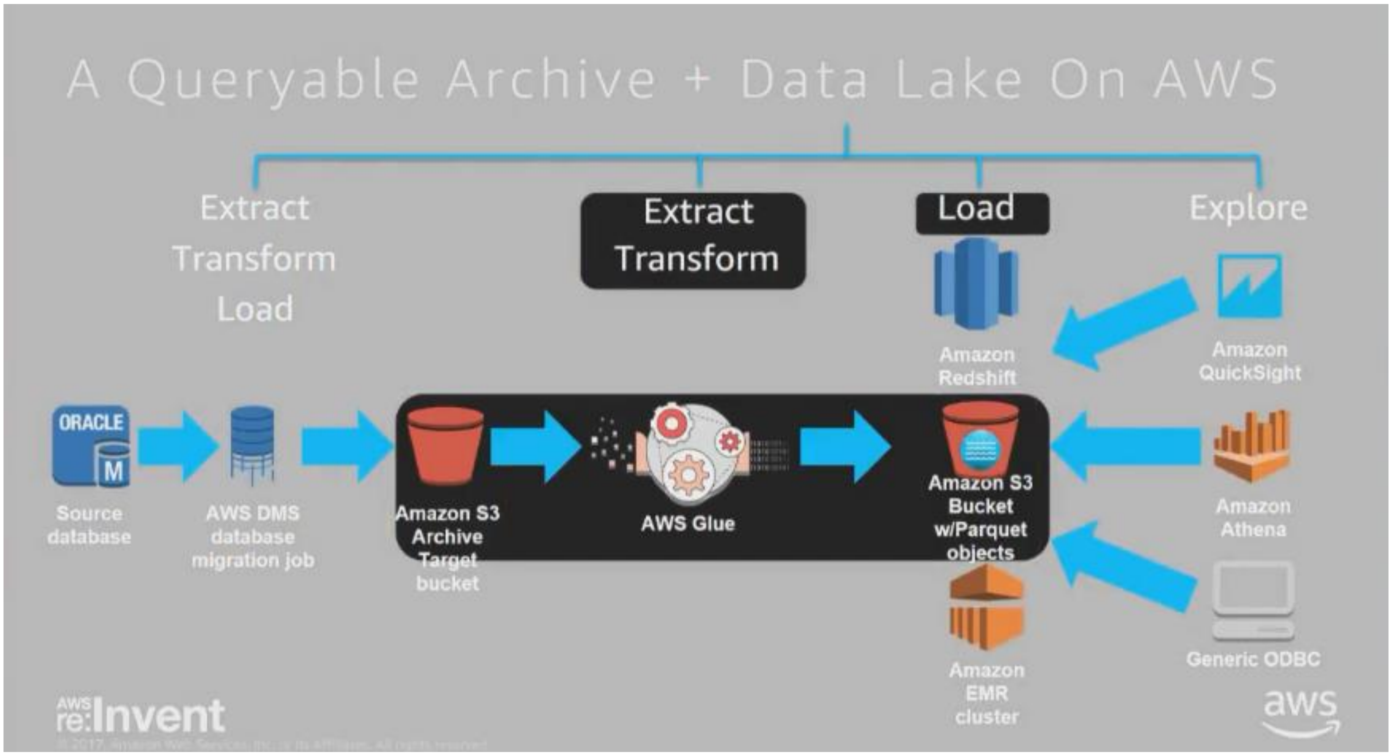
queryarchive-oracle-to-s3

Schema	Table	State	Inserts	Deletes	Updates	DDLs	Full Load Rows	Total	Last updated
MASTER	orders	Table completed	0	0	0	0	167,497,743	167,497,743	10/25/17, 2:05

The job is done 4h 26m later and we have copied 167,497,743 rows from the Oracle database to S3 bucket in compressed CSV format. The dataset contains orders and order executions by a trading desk.

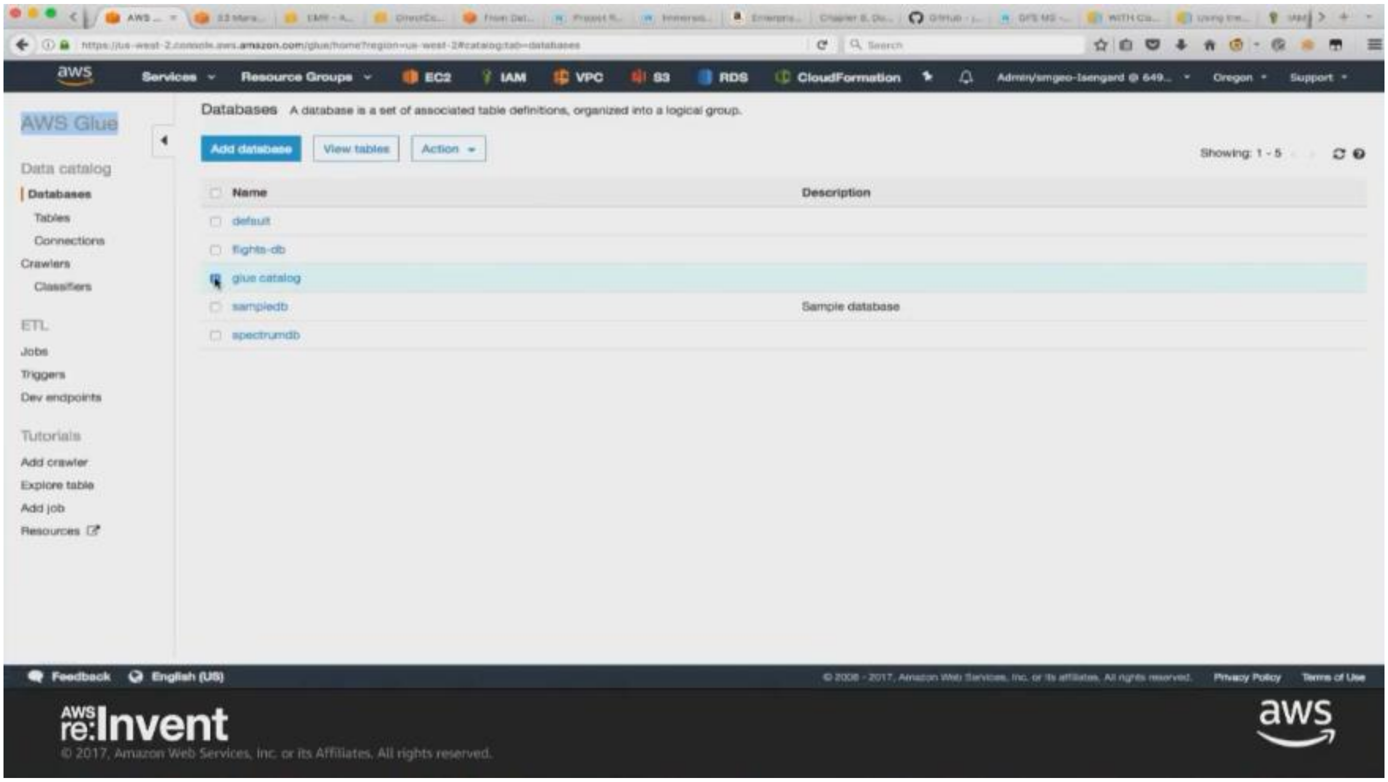


A Queryable Archive + Data Lake On AWS



A Queryable Archive + Data Lake On AWS

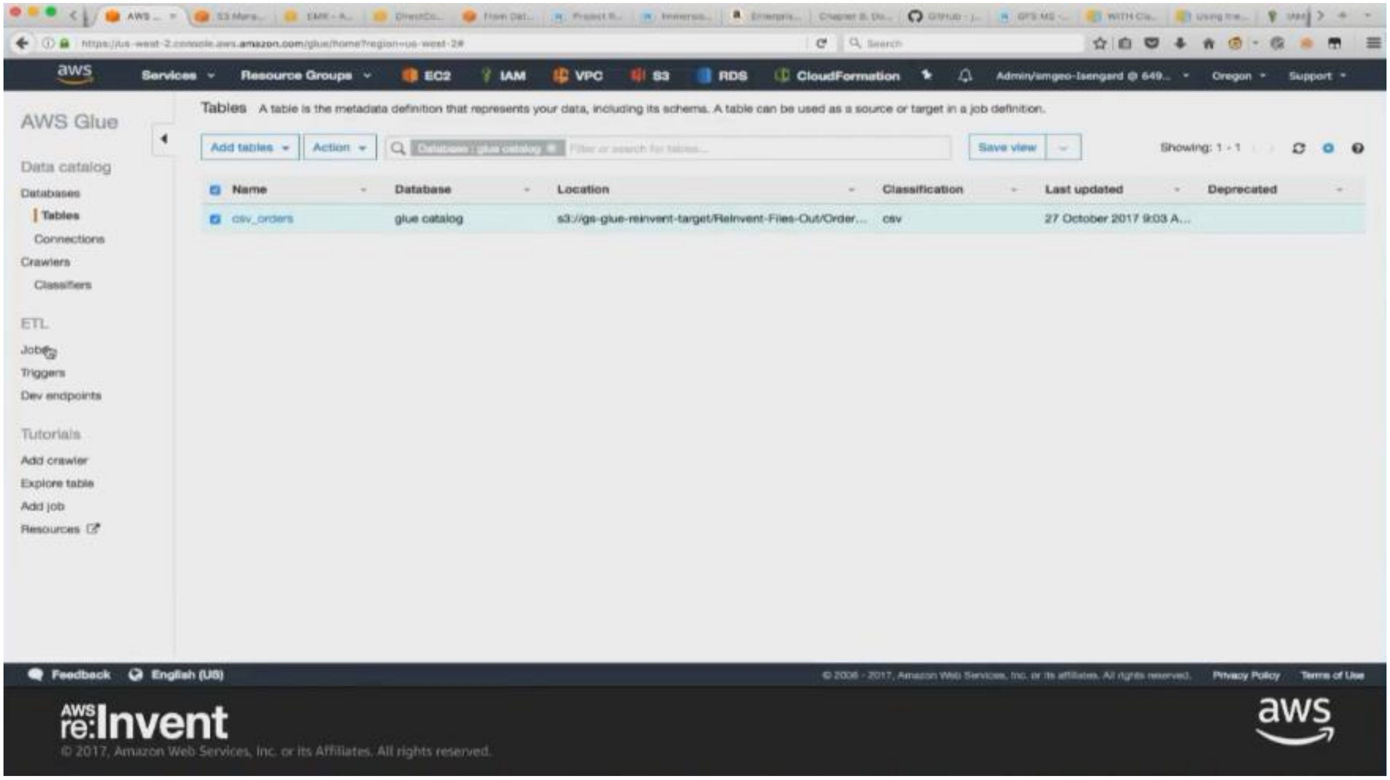
We now want to transform the compressed CSV data into the enterprise data lake format for the data model in Parquet objects. You will be transforming and enriching the data along the way



Databases A database is a set of associated table definitions, organized into a logical group.

☐ Name	Description
☐ default	
☐ flights-db	
☑ glue catalog	
☐ sampledb	Sample database
☐ spectrumdb	

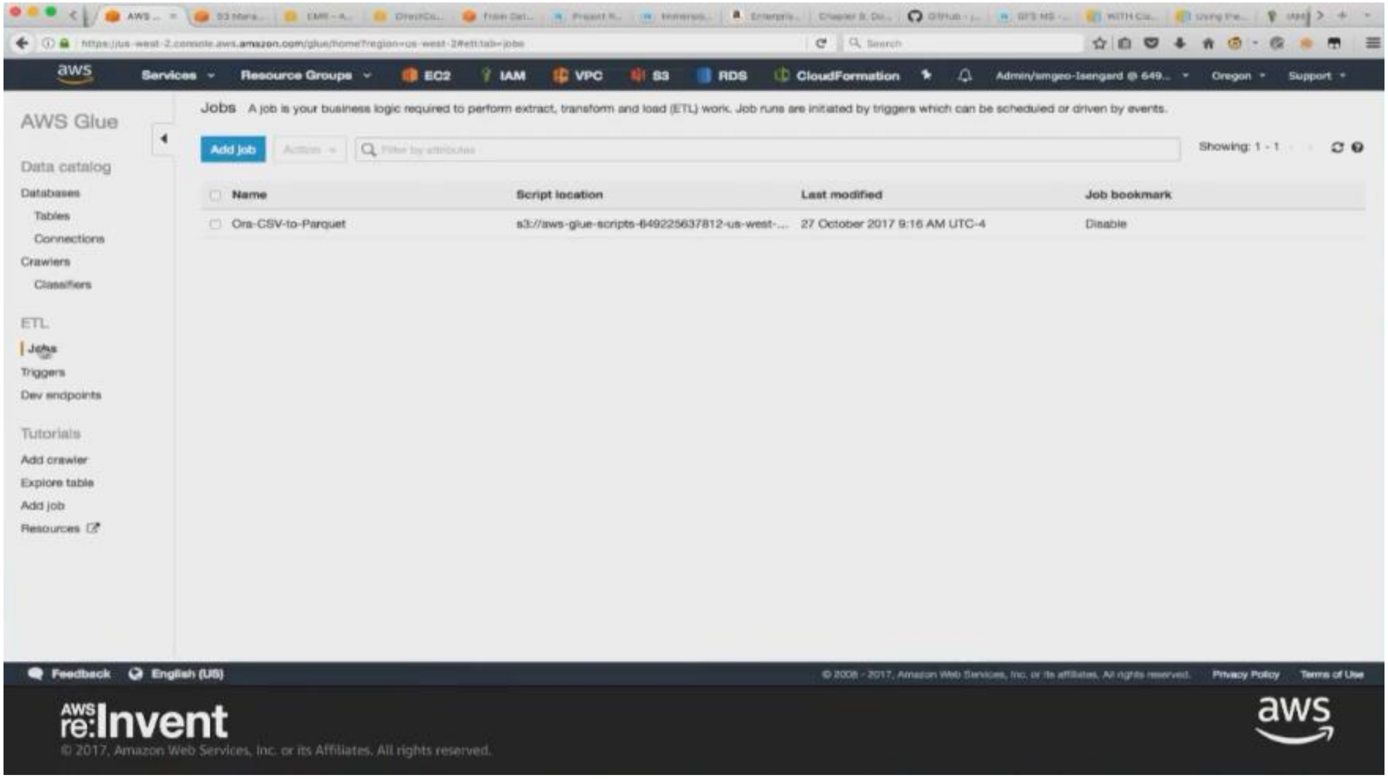
We are going to be using AWS Glue for this batch job,



Tables A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

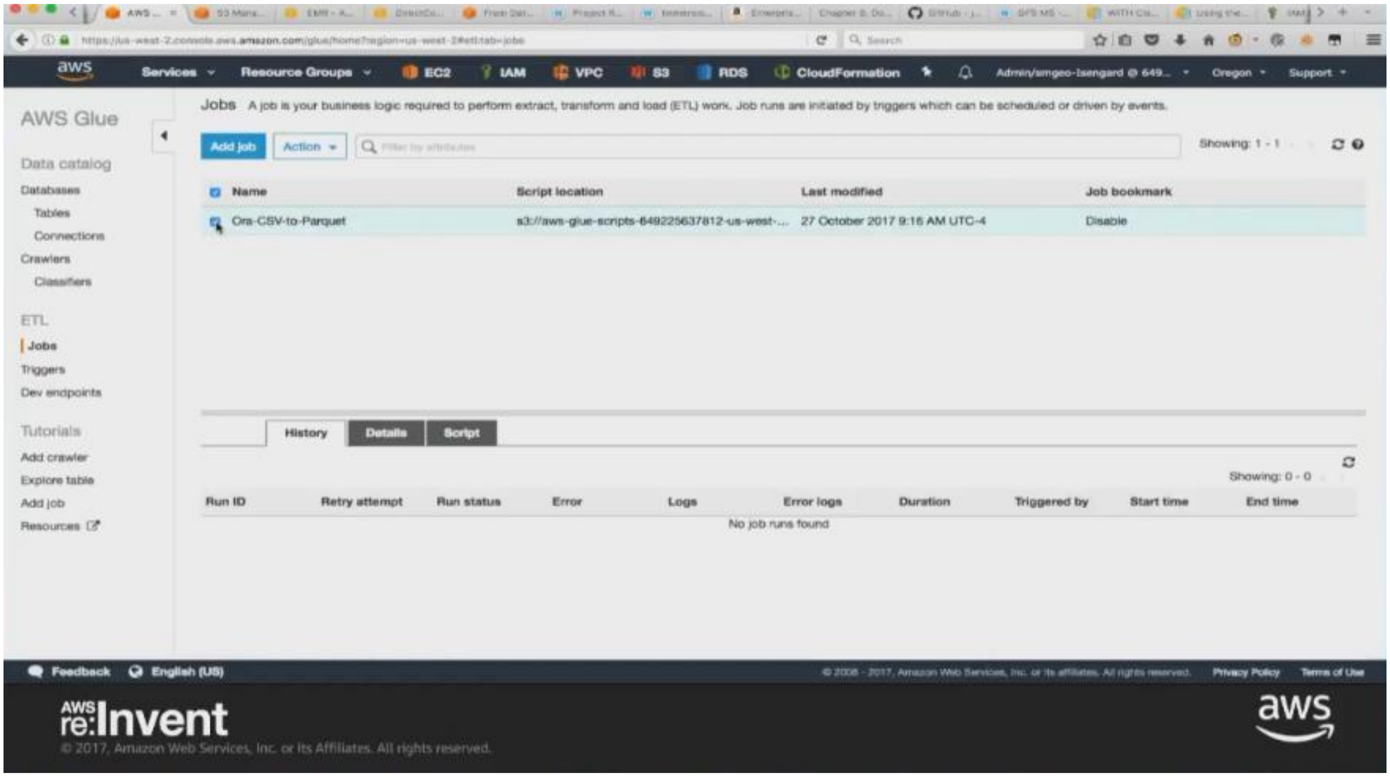
☑ Name	Database	Location	Classification	Last updated	Deprecated
☑ csv_orders	glue catalog	s3://gs-glue-reinvent-target/Reinvent-Files-Out/Order...	csv	27 October 2017 9:03 A...	

Within the glue_catalog database, we have the table containing our compressed CSV files



Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are initiated by triggers which can be scheduled or driven by events.

	Name	Script location	Last modified	Job bookmark
□	Ora-CSV-to-Parquet	s3://aws-glue-scripts-649225637812-us-west-...	27 October 2017 9:16 AM UTC-4	Disable

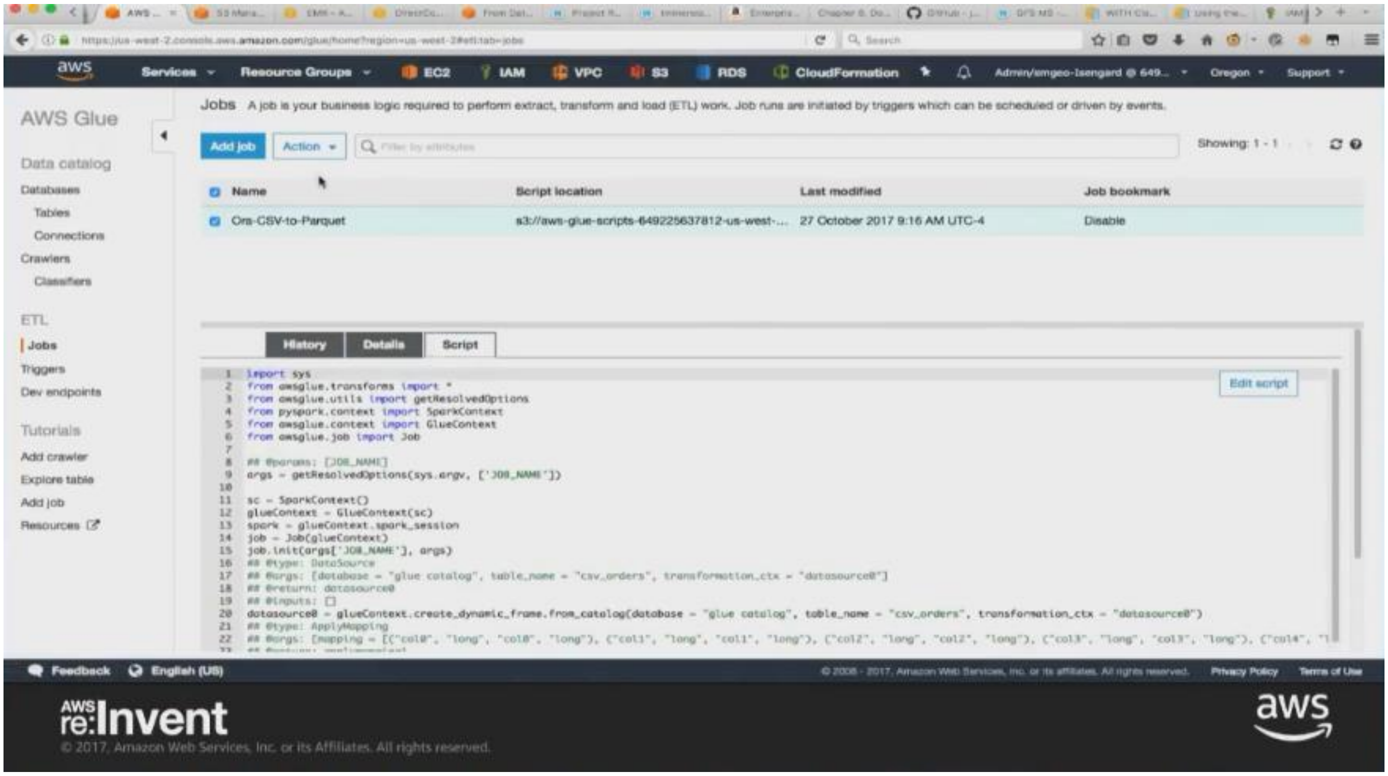


Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are initiated by triggers which can be scheduled or driven by events.

☑	Name	Script location	Last modified	Job bookmark
☑	Ora-CSV-to-Parquet	s3://aws-glue-scripts-649225637812-us-west-...	27 October 2017 9:16 AM UTC-4	Disable

Run ID	Retry attempt	Run status	Error	Logs	Error logs	Duration	Triggered by	Start time	End time

No job runs found

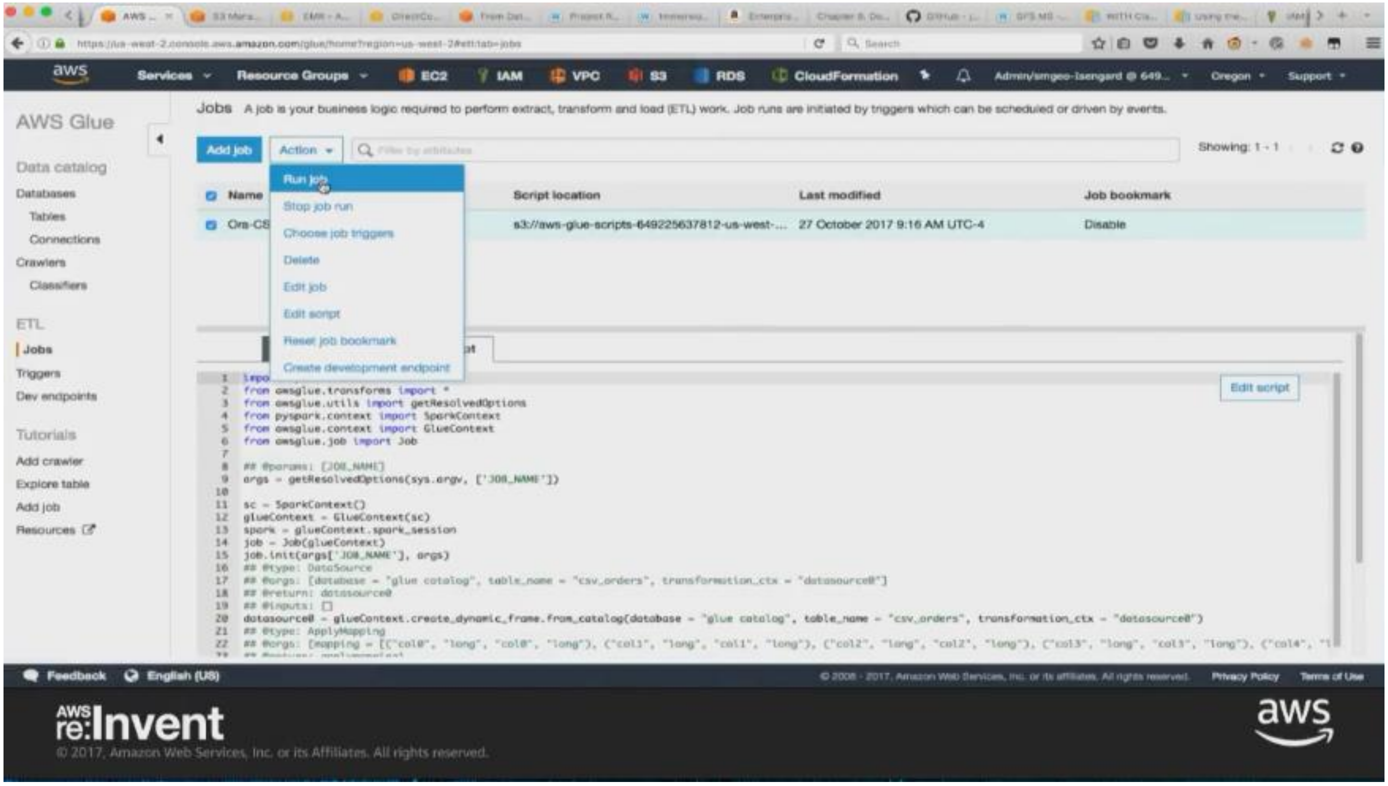


Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are initiated by triggers which can be scheduled or driven by events.

☑ Name	Script location	Last modified	Job bookmark
☑ Ora-CSV-to-Parquet	s3://aws-glue-scripts-649225637812-us-west-...	27 October 2017 9:16 AM UTC-4	Disable

```
import sys
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

## @params: [JOB_NAME]
args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)
## @type: DataSource
## @args: [database = "glue catalog", table_name = "csv_orders", transformation_ctx = "datasource0"]
## @return: datasource0
## @inputs: []
datasource0 = glueContext.create_dynamic_frame.from_catalog(database = "glue catalog", table_name = "csv_orders", transformation_ctx = "datasource0")
## @type: ApplyMapping
## @args: [mapping = [("col0", "long", "col0", "long"), ("col1", "long", "col1", "long"), ("col2", "long", "col2", "long"), ("col3", "long", "col3", "long"), ("col4", "l
```

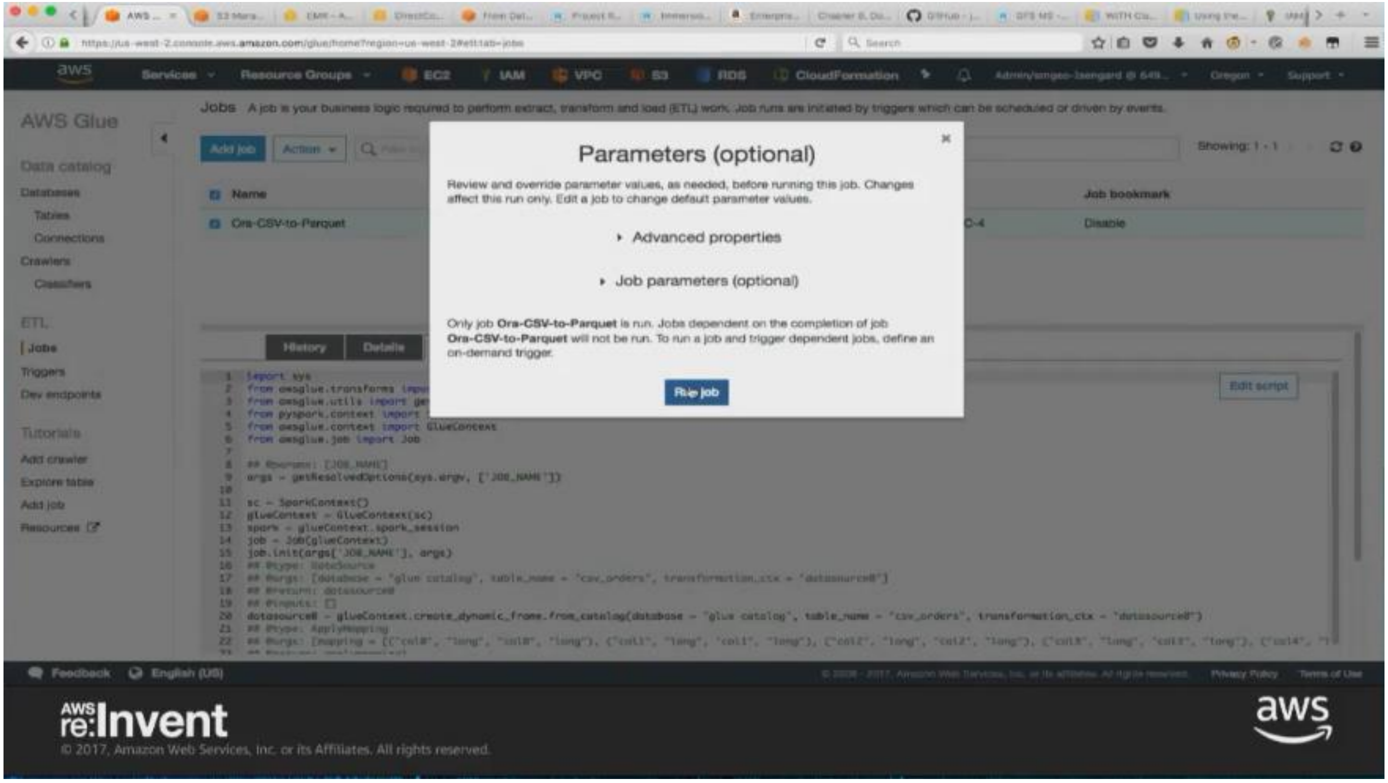
AWS Glue generate python code for the CSV_to_Parquet job for us



Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are initiated by triggers which can be scheduled or driven by events.

☑ Name	Script location	Last modified	Job bookmark
☑ Ora-CS	s3://aws-glue-scripts-649225637812-us-west-...	27 October 2017 9:16 AM UTC-4	Disable

```
impo
from awsglue.transforms import *
from awsglue.utils import getResolvedOptions
from pyspark.context import SparkContext
from awsglue.context import GlueContext
from awsglue.job import Job

## @params: [JOB_NAME]
args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)
## @type: DataSource
## @args: [database = "glue catalog", table_name = "csv_orders", transformation_ctx = "datasource0"]
## @return: datasource0
## @inputs: []
datasource0 = glueContext.create_dynamic_frame.from_catalog(database = "glue catalog", table_name = "csv_orders", transformation_ctx = "datasource0")
## @type: ApplyMapping
## @args: [mapping = [("col0", "long", "col0", "long"), ("col1", "long", "col1", "long"), ("col2", "long", "col2", "long"), ("col3", "long", "col3", "long"), ("col4", "l
```



Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are initiated by triggers which can be scheduled or driven by events.

Name		Job bookmark
Ora-CSV-to-Parquet	C-4	Disable

Parameters (optional)

Review and override parameter values, as needed, before running this job. Changes affect this run only. Edit a job to change default parameter values.

▸ Advanced properties

▸ Job parameters (optional)

Only job **Ora-CSV-to-Parquet** is run. Jobs dependent on the completion of job **Ora-CSV-to-Parquet** will not be run. To run a job and trigger dependent jobs, define an on-demand trigger.

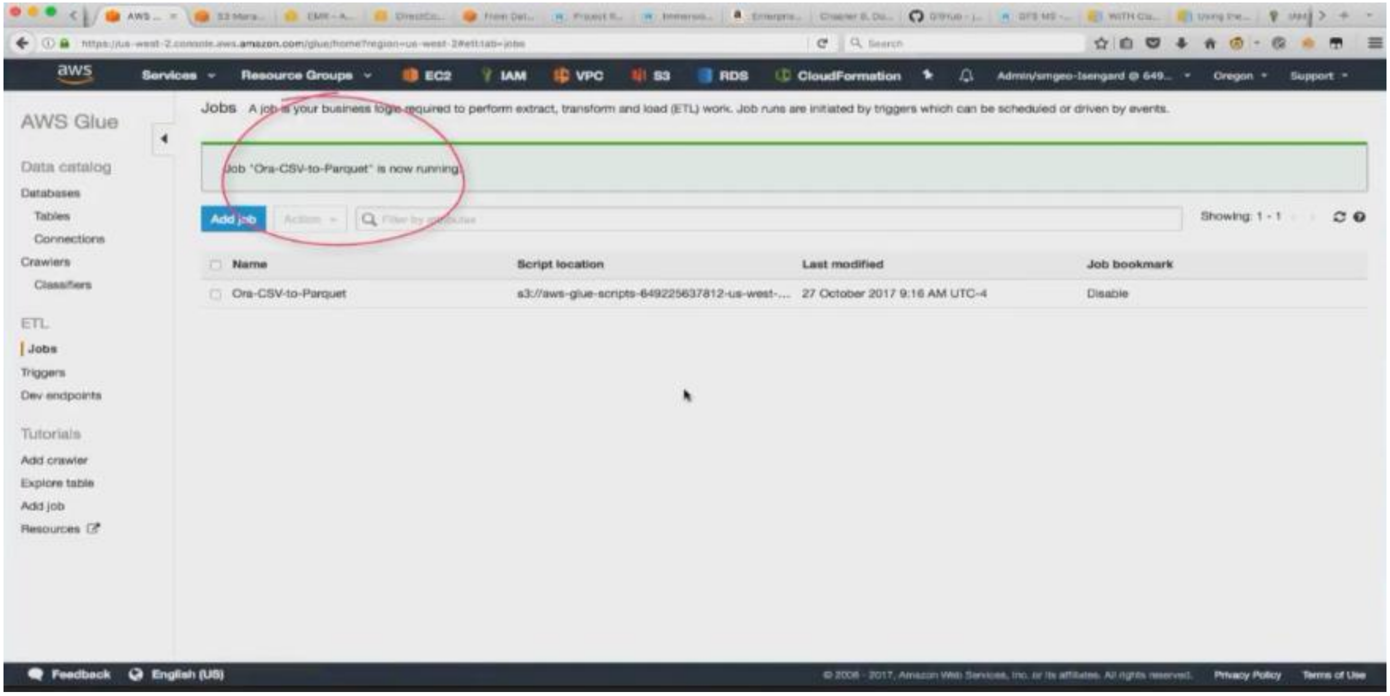
Run job

```
import sys
from awsglue.transforms impo
from awsglue.utils import ge
from pyspark.context import
from awsglue.context import GlueContext
from awsglue.job import Job

## @params: [JOB_NAME]
args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()
glueContext = GlueContext(sc)
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'], args)
## @type: DataSource
## @args: [database = "glue catalog", table_name = "csv_orders", transformation_ctx = "datasource0"]
## @return: datasource0
## @inputs: []
datasource0 = glueContext.create_dynamic_frame.from_catalog(database = "glue catalog", table_name = "csv_orders", transformation_ctx = "datasource0")
## @type: ApplyMapping
## @args: [mapping = [("col0", "long", "col0", "long"), ("col1", "long", "col1", "long"), ("col2", "long", "col2", "long"), ("col3", "long", "col3", "long"), ("col4", "l
```

© 2008 - 2017, Amazon Web Services, Inc. or its affiliates. All rights reserved.

AWS re:Invent

© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.

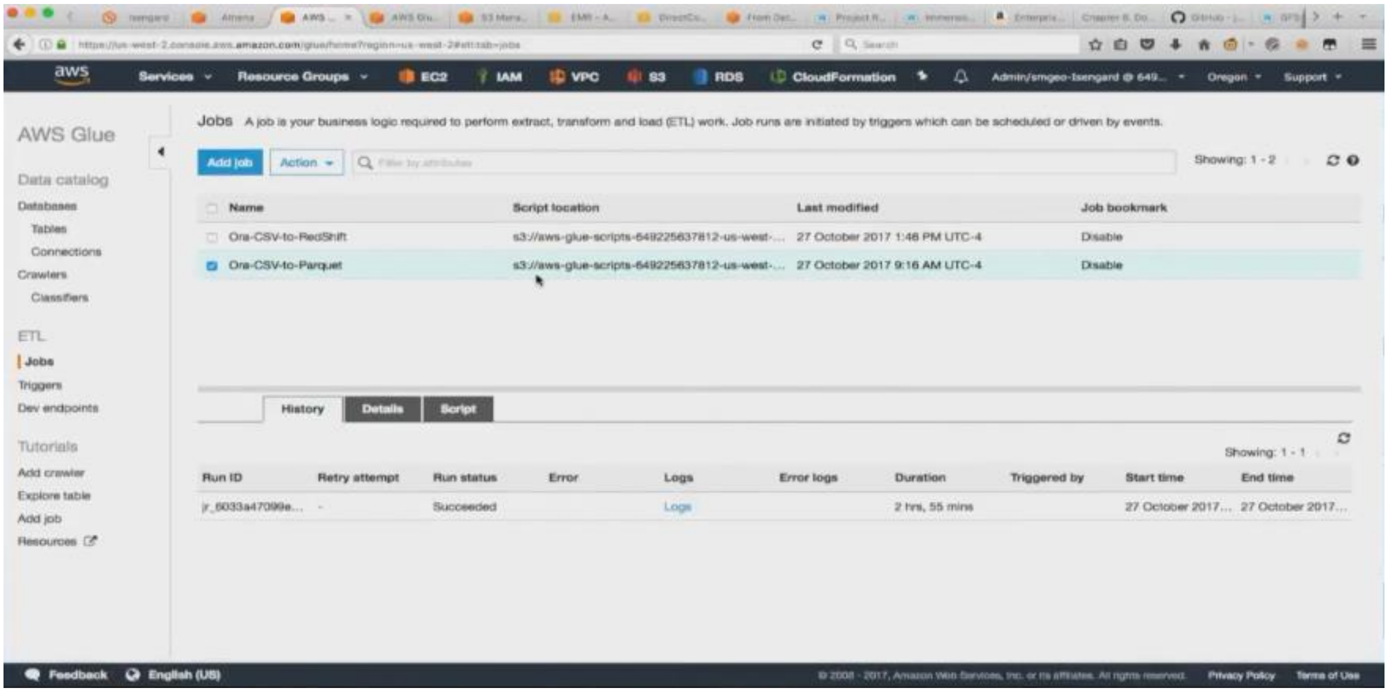


Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are initiated by triggers which can be scheduled or driven by events.

Job "Ora-CSV-to-Parquet" is now running

Name	Script location	Last modified	Job bookmark
Ora-CSV-to-Parquet	s3://aws-glue-scripts-649225637812-us-west-...	27 October 2017 9:16 AM UTC-4	Disable

© 2008 - 2017, Amazon Web Services, Inc. or its affiliates. All rights reserved.



Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are initiated by triggers which can be scheduled or driven by events.

Showing: 1 - 2

	Name	Script location	Last modified	Job bookmark
□	Ora-CSV-to-RedShift	s3://aws-glue-scripts-649225637812-us-west-...	27 October 2017 1:46 PM UTC-4	Disable
☑	Ora-CSV-to-Parquet	s3://aws-glue-scripts-649225637812-us-west-...	27 October 2017 9:16 AM UTC-4	Disable

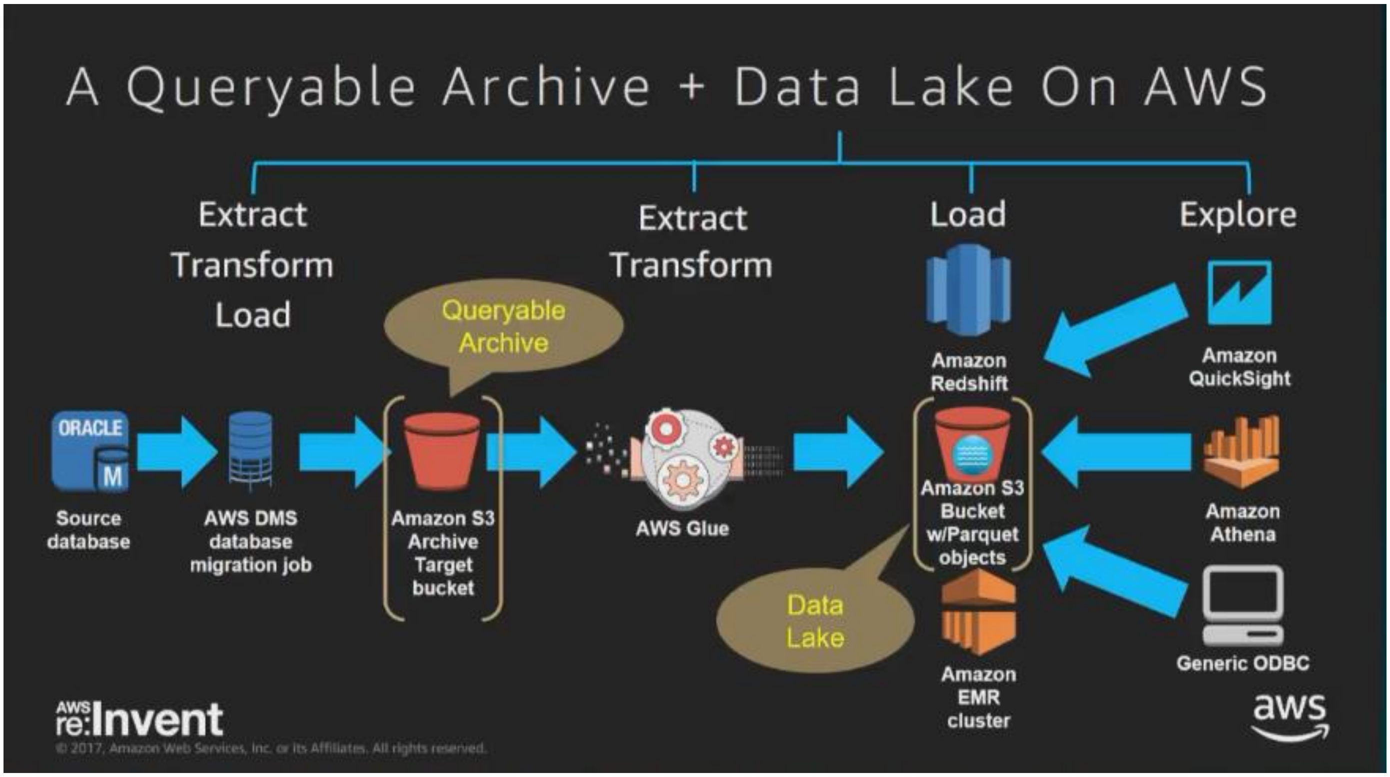
History | Details | Script

Showing: 1 - 1

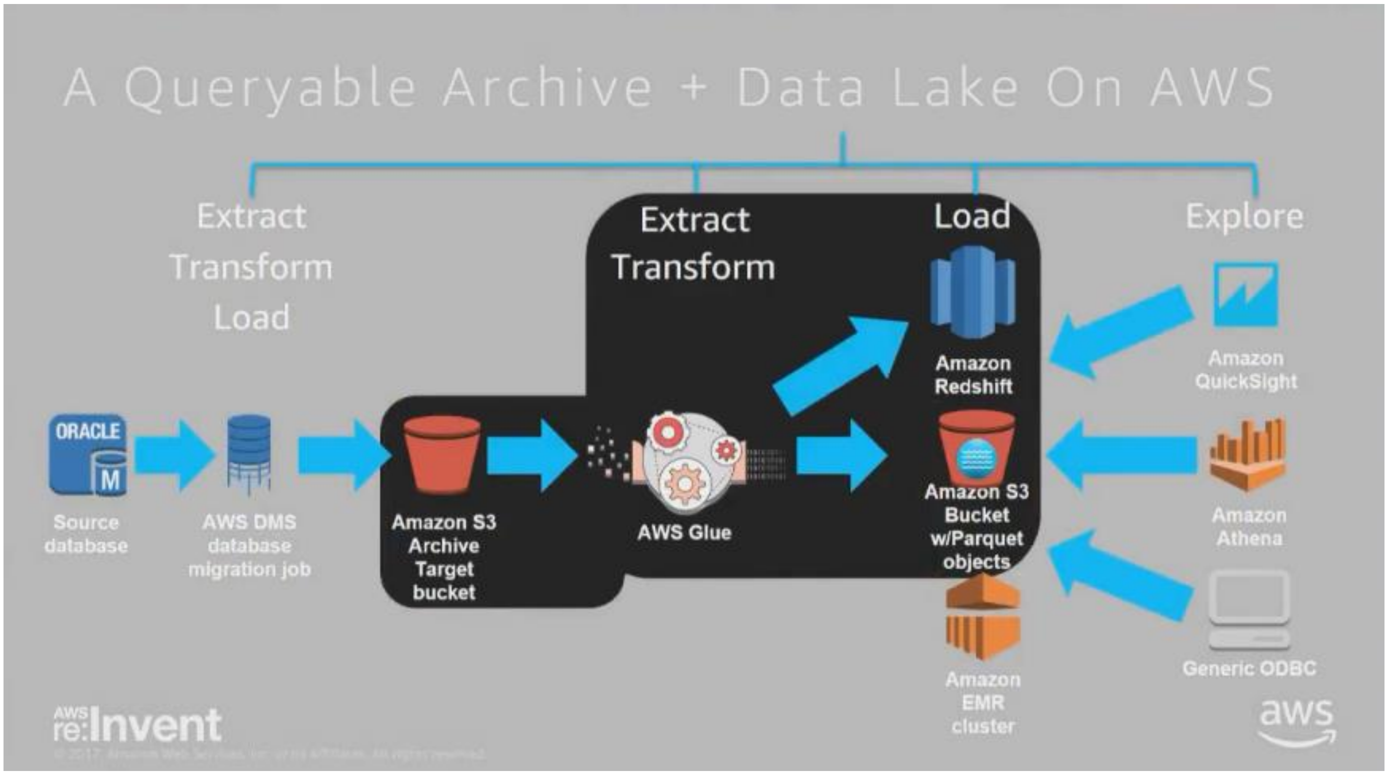
Run ID	Retry attempt	Run status	Error	Logs	Error logs	Duration	Triggered by	Start time	End time
jr_6033a47099e...	-	Succeeded		Logs		2 hrs, 55 mins		27 October 2017...	27 October 2017...

© 2008 - 2017, Amazon Web Services, Inc. or its affiliates. All rights reserved. Privacy Policy Terms of Use

The job finishes after 2h 55mins and we have loaded all 167,497,743 rows into our data lake.

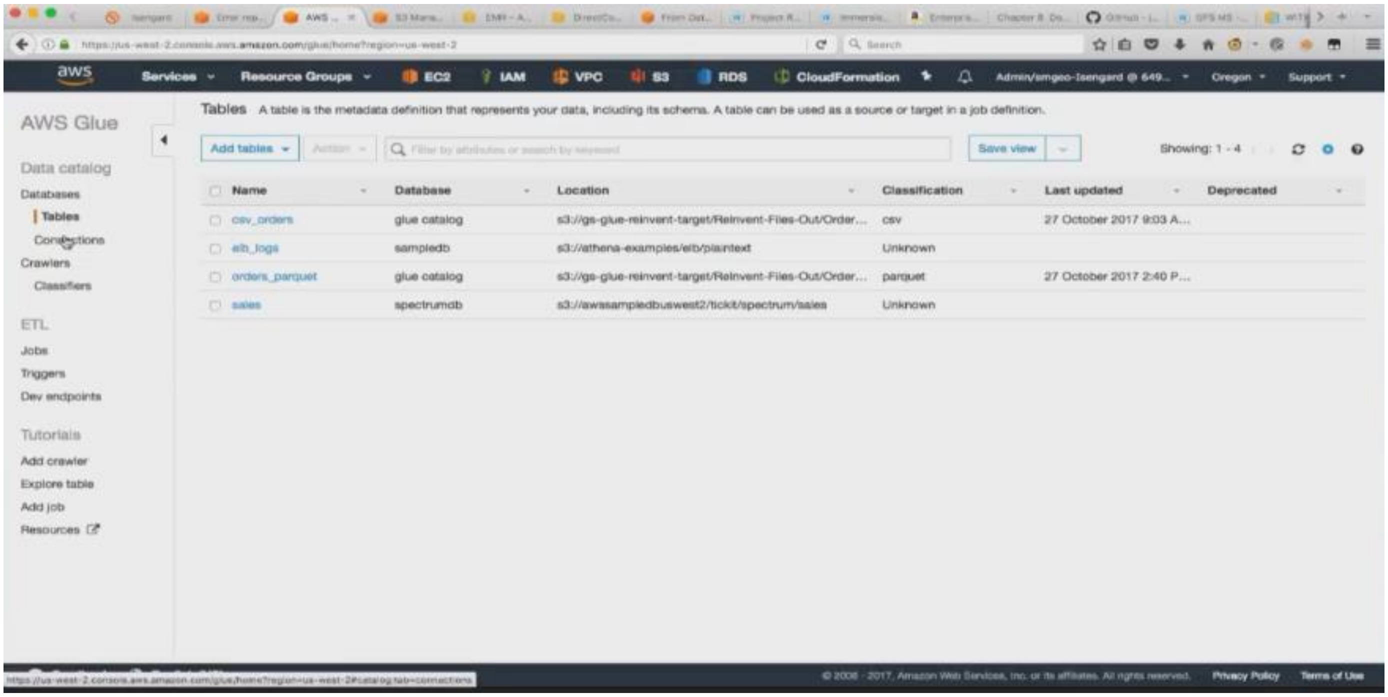


© 2017, Amazon Web Services, Inc. or its Affiliates. All rights reserved.



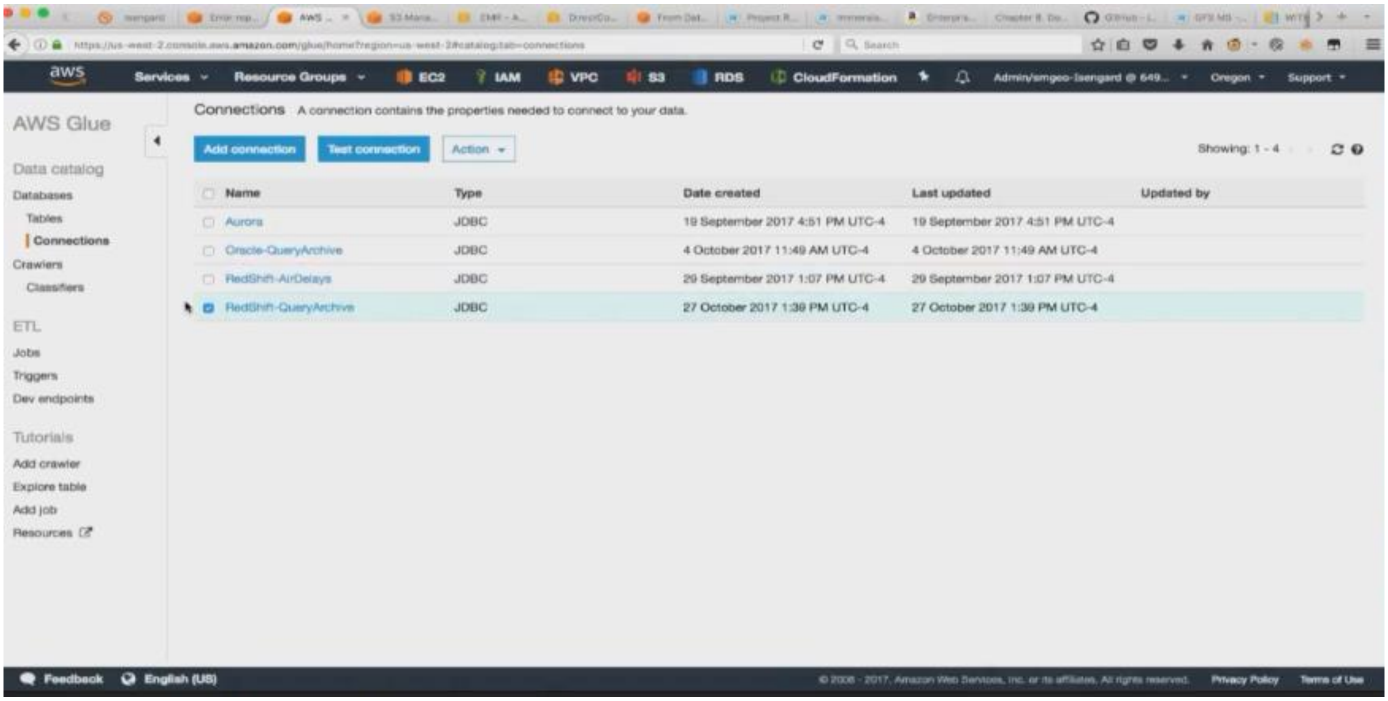
A Queryable Archive + Data Lake On AWS

Now we have data in 2 locations.



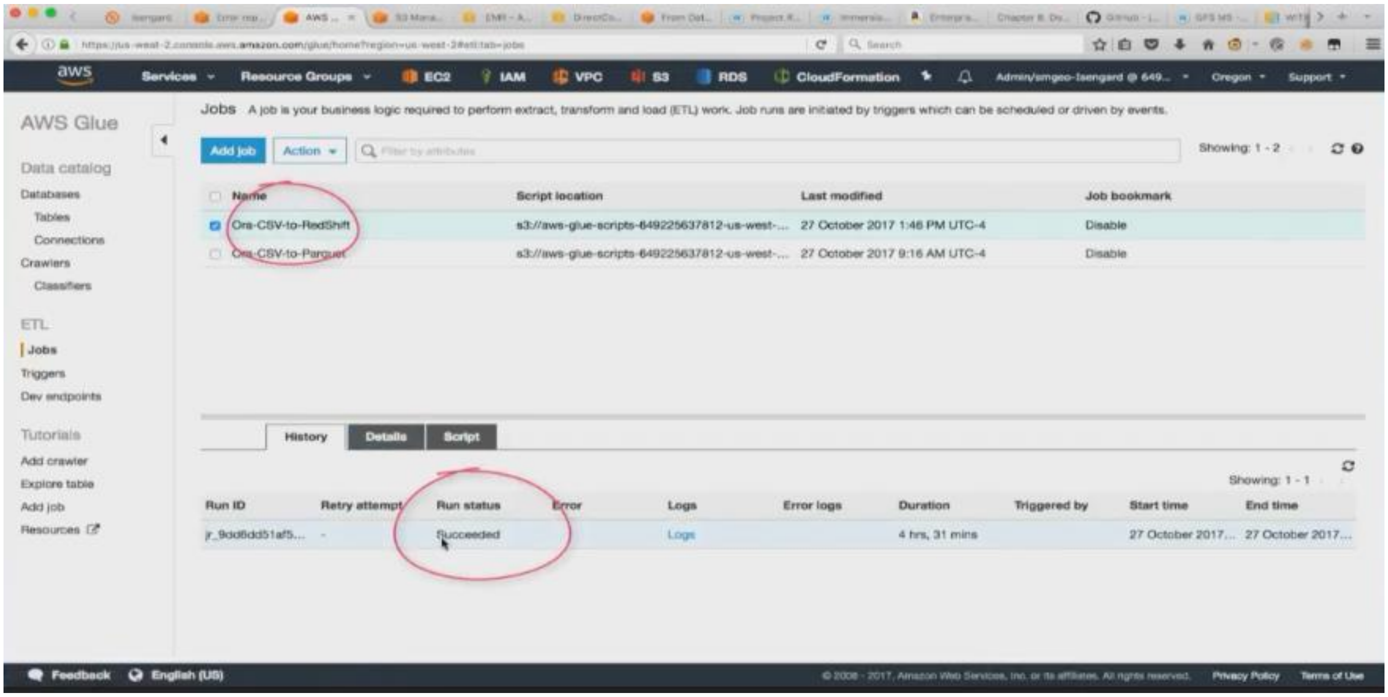
Tables A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Name	Database	Location	Classification	Last updated	Deprecated
csv_orders	glue catalog	s3://gs-glue-reinvent-target/Reinvent-Files-Out/Order...	csv	27 October 2017 9:03 A...	
elb_logs	sampledb	s3://athena-examples/elb/plaintext	Unknown		
orders_parquet	glue catalog	s3://gs-glue-reinvent-target/Reinvent-Files-Out/Order...	parquet	27 October 2017 2:40 P...	
sales	spectrumdb	s3://awssampledbuswest2/tickit/spectrum/sales	Unknown		



Connections A connection contains the properties needed to connect to your data.

	Name	Type	Date created	Last updated	Updated by
□	Aurora	JDBC	19 September 2017 4:51 PM UTC-4	19 September 2017 4:51 PM UTC-4	
□	Oracle-QueryArchive	JDBC	4 October 2017 11:49 AM UTC-4	4 October 2017 11:49 AM UTC-4	
□	RedShift-AirDelays	JDBC	29 September 2017 1:07 PM UTC-4	29 September 2017 1:07 PM UTC-4	
☑	RedShift-QueryArchive	JDBC	27 October 2017 1:39 PM UTC-4	27 October 2017 1:39 PM UTC-4	

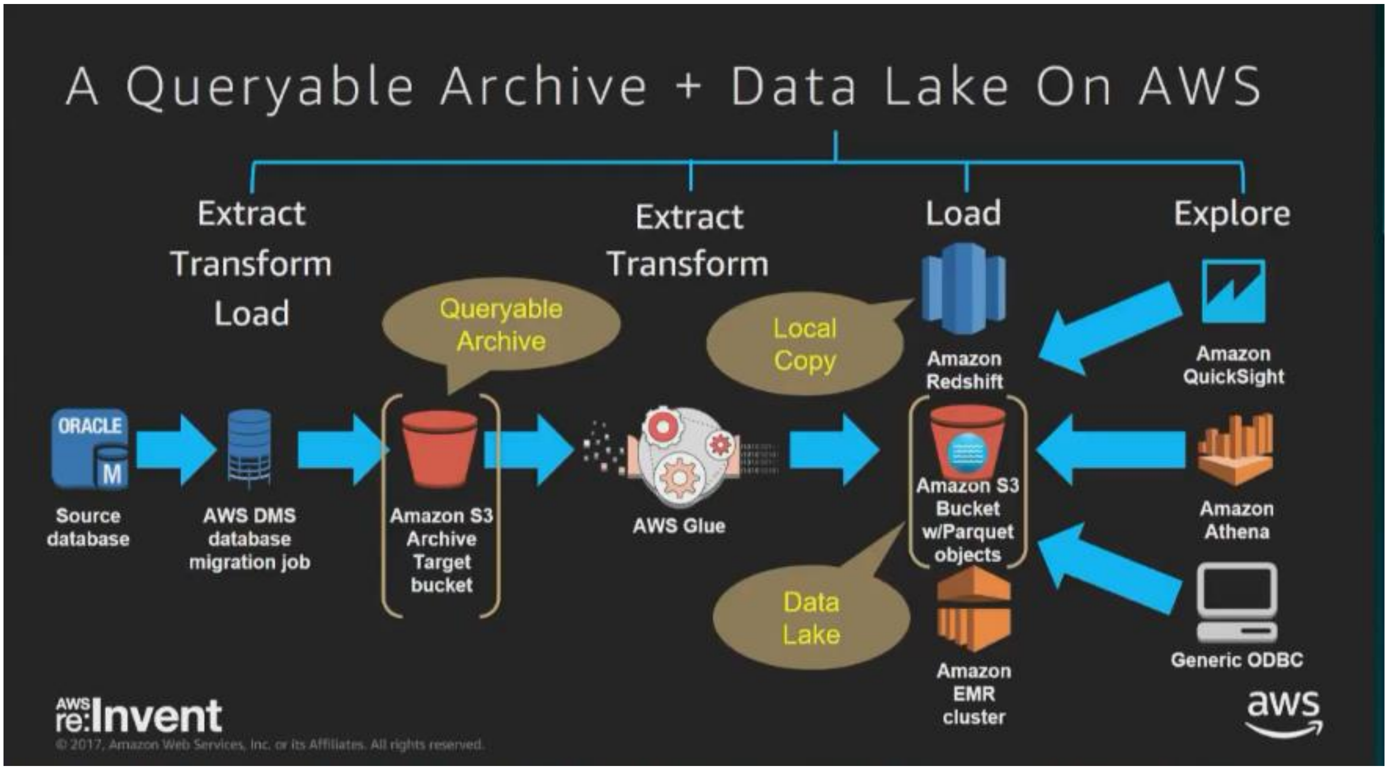


Jobs A job is your business logic required to perform extract, transform and load (ETL) work. Job runs are initiated by triggers which can be scheduled or driven by events.

	Name	Script location	Last modified	Job bookmark
☑	Ora-CSV-to-RedShift	s3://aws-glue-scripts-649225637812-us-west-...	27 October 2017 1:46 PM UTC-4	Disable
□	Ora-CSV-to-Parquet	s3://aws-glue-scripts-649225637812-us-west-...	27 October 2017 9:16 AM UTC-4	Disable

Run ID	Retry attempt	Run status	Error	Logs	Error logs	Duration	Triggered by	Start time	End time
jr_9dd6dd51af5...	-	Succeeded		Logs		4 hrs, 31 mins		27 October 2017...	27 October 2017...

We also load data into Redshift



A Queryable Archive + Data Lake On AWS

Extract
Transform
Load

Extract
Transform

Load

Explore

Queryable Archive

Local Copy

Data Lake

Source database

AWS DMS database migration job

Amazon S3 Archive Target bucket

AWS Glue

Amazon Redshift

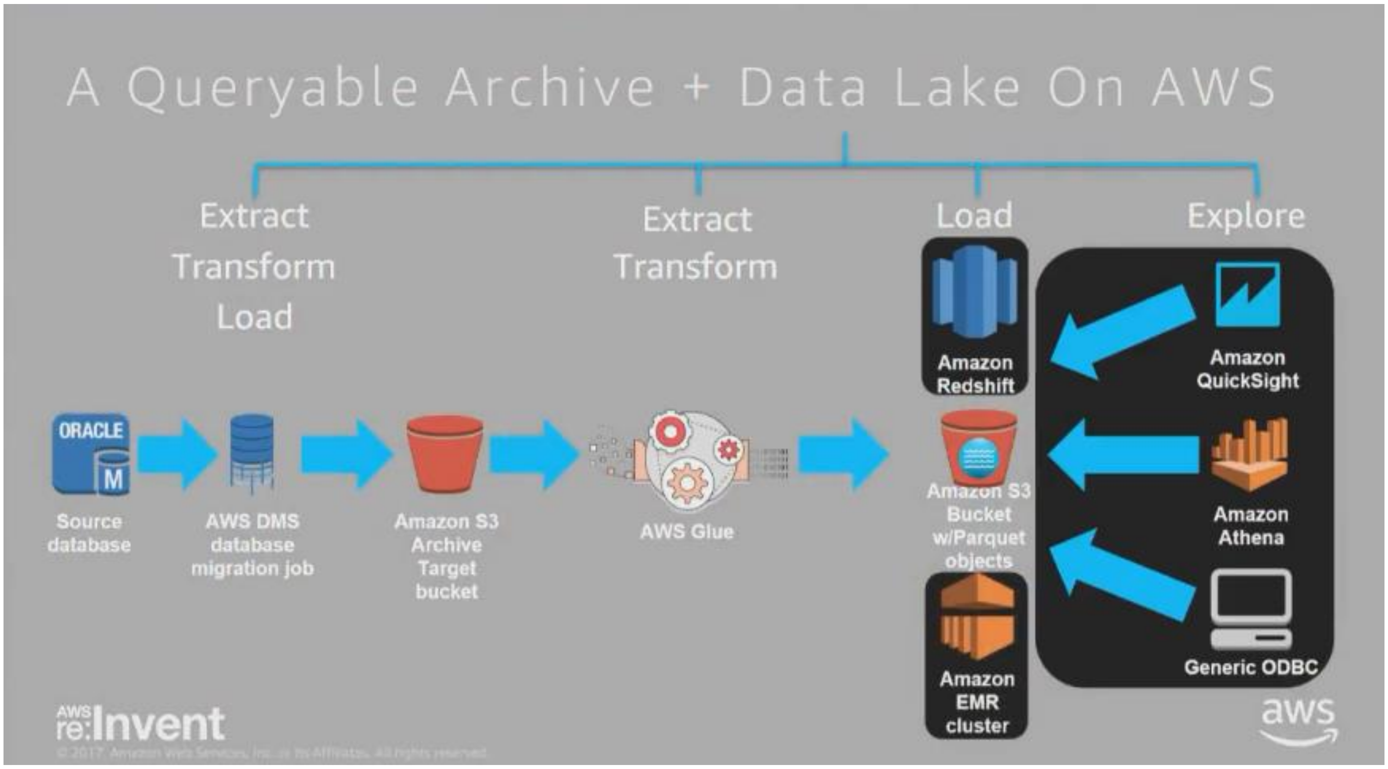
Amazon S3 Bucket w/Parquet objects

Amazon EMR cluster

Amazon QuickSight

Amazon Athena

Generic ODBC



A Queryable Archive + Data Lake On AWS

Extract
Transform
Load

Extract
Transform

Load

Explore

Source database

AWS DMS database migration job

Amazon S3 Archive Target bucket

AWS Glue

Amazon Redshift

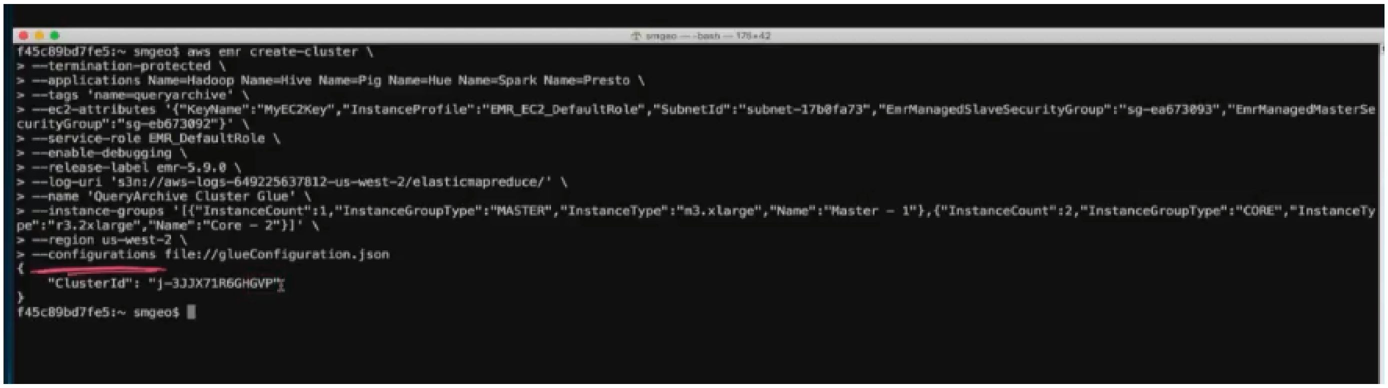
Amazon S3 Bucket w/Parquet objects

Amazon EMR cluster

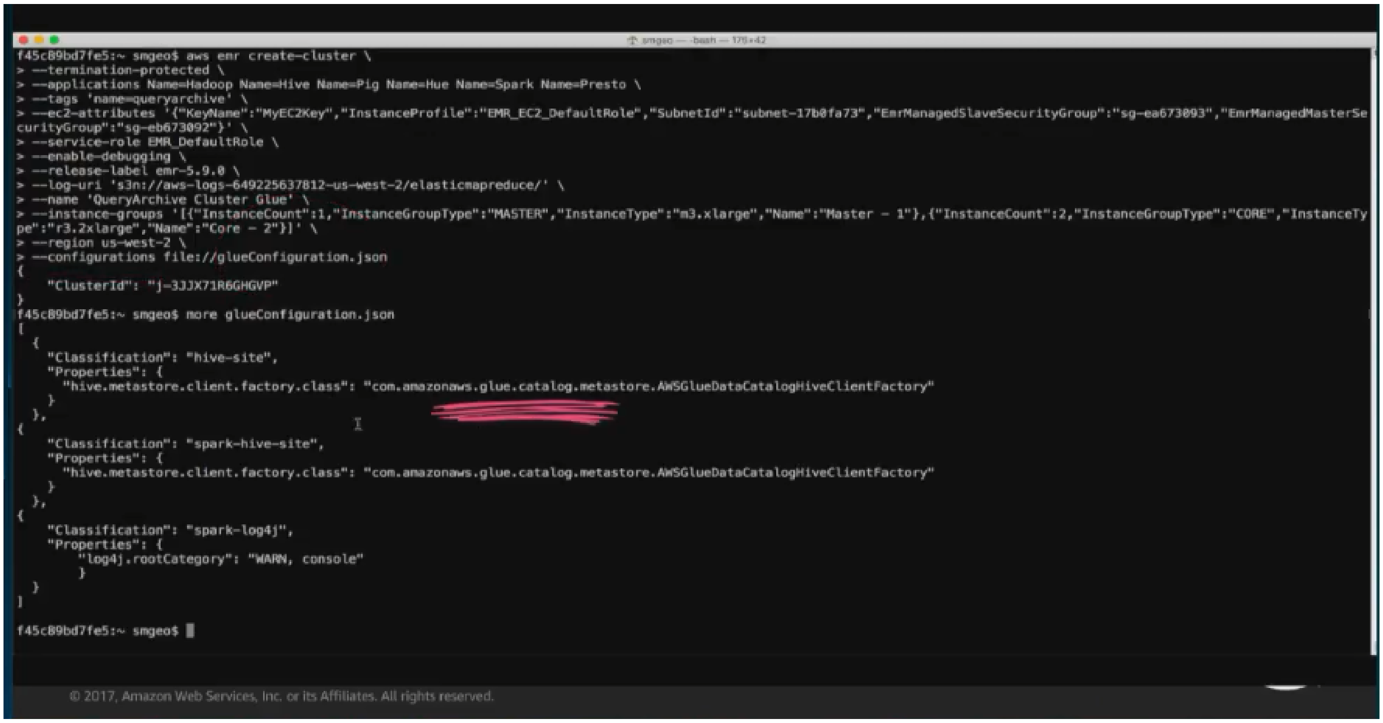
Amazon QuickSight

Amazon Athena

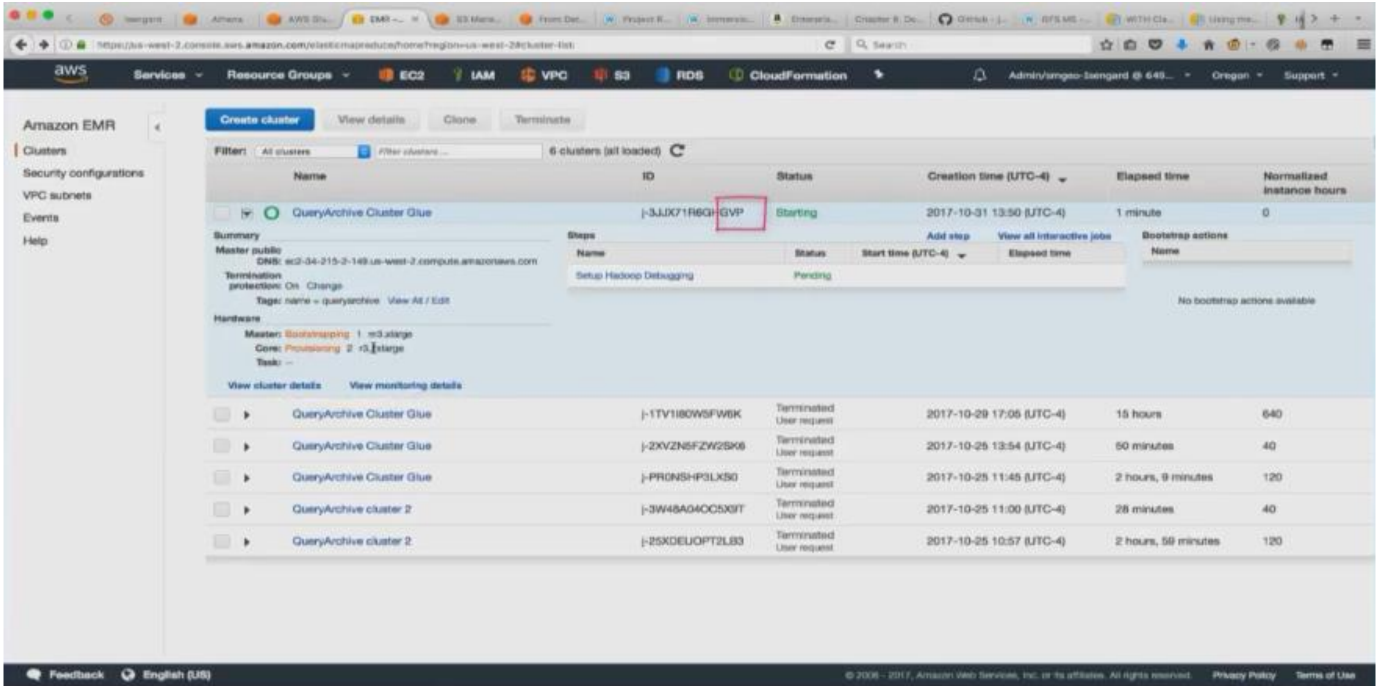
Generic ODBC



```
f45c89bd7fe5:~ smgeo$ aws emr create-cluster \
> --termination-protected \
> --applications Name=Hadoop Name=Hive Name=Pig Name=Hue Name=Spark Name=Presto \
> --tags 'name=queryarchive' \
> --ec2-attributes '{"KeyName":"MyEC2Key","InstanceProfile":"EMR_EC2_DefaultRole","SubnetId":"subnet-17b0fa73","EmrManagedSlaveSecurityGroup":"sg-ea673093","EmrManagedMasterSe
curityGroup":"sg-eb673092"}' \
> --service-role EMR_DefaultRole \
> --enable-debugging \
> --release-label emr-5.9.0 \
> --log-uri 's3n://aws-logs-649225637812-us-west-2/elasticmapreduce/' \
> --name 'QueryArchive Cluster Glue' \
> --instance-groups '[{"InstanceCount":1,"InstanceGroupType":"MASTER","InstanceType":"m3.xlarge","Name":"Master - 1"},{"InstanceCount":2,"InstanceGroupType":"CORE","InstanceTy
pe":"r3.2xlarge","Name":"Core - 2"}]' \
> --region us-west-2 \
> --configurations file://glueConfiguration.json
{
    "ClusterId": "j-3JJX71R6GHGVP"
}
f45c89bd7fe5:~ smgeo$ 
```



```
f45c89bd7fe5:~ smgeo$ aws emr create-cluster \
> --termination-protected \
> --applications Name=Hadoop Name=Hive Name=Pig Name=Hue Name=Spark Name=Presto \
> --tags 'name=queryarchive' \
> --ec2-attributes '{"KeyName":"MyEC2Key","InstanceProfile":"EMR_EC2_DefaultRole","SubnetId":"subnet-17b0fa73","EmrManagedSlaveSecurityGroup":"sg-ea673093","EmrManagedMasterSe
curityGroup":"sg-eb673092"}' \
> --service-role EMR_DefaultRole \
> --enable-debugging \
> --release-label emr-5.9.0 \
> --log-uri 's3n://aws-logs-649225637812-us-west-2/elasticmapreduce/' \
> --name 'QueryArchive Cluster Glue' \
> --instance-groups '[{"InstanceCount":1,"InstanceGroupType":"MASTER","InstanceType":"m3.xlarge","Name":"Master - 1"},{"InstanceCount":2,"InstanceGroupType":"CORE","InstanceTy
pe":"r3.2xlarge","Name":"Core - 2"}]' \
> --region us-west-2 \
> --configurations file://glueConfiguration.json
{
    "ClusterId": "j-3JJX71R6GHGVP"
}
f45c89bd7fe5:~ smgeo$ more glueConfiguration.json
[
  {
    "Classification": "hive-site",
    "Properties": {
      "hive.metastore.client.factory.class": "com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory"
    }
  },
{
    "Classification": "spark-hive-site",
    "Properties": {
      "hive.metastore.client.factory.class": "com.amazonaws.glue.catalog.metastore.AWSGlueDataCatalogHiveClientFactory"
    }
  },
{
    "Classification": "spark-log4j",
    "Properties": {
        "log4j.rootCategory": "WARN, console"
        }
  }
]

f45c89bd7fe5:~ smgeo$ 
```



6 clusters (all loaded)

	Name	ID	Status	Creation time (UTC-4)	Elapsed time	Normalized instance hours
	QueryArchive Cluster Glue	j-3JJX71R6GHGVP	Starting	2017-10-31 13:50 (UTC-4)	1 minute	0
	QueryArchive Cluster Glue	j-1TV1I80W5FW6K	Terminated User request	2017-10-29 17:05 (UTC-4)	15 hours	640
	QueryArchive Cluster Glue	j-2XVZN5FZW2SX6	Terminated User request	2017-10-25 13:54 (UTC-4)	50 minutes	40
	QueryArchive Cluster Glue	j-PR0NSHP3LXS0	Terminated User request	2017-10-25 11:45 (UTC-4)	2 hours, 9 minutes	120
	QueryArchive cluster 2	j-3W48A04OC5X9T	Terminated User request	2017-10-25 11:00 (UTC-4)	28 minutes	40
	QueryArchive cluster 2	j-25XDEUOPT2L83	Terminated User request	2017-10-25 10:57 (UTC-4)	2 hours, 59 minutes	120

Summary
Master public DNS: ec2-34-215-2-149.us-west-2.compute.amazonaws.com
Termination protection: On Change
Tags: name = queryarchive View All / Edit

Hardware
Master: Bootstrapping 1 m3.xlarge
Core: Provisioning 2 r3.2xlarge
Task: --

View cluster details View monitoring details

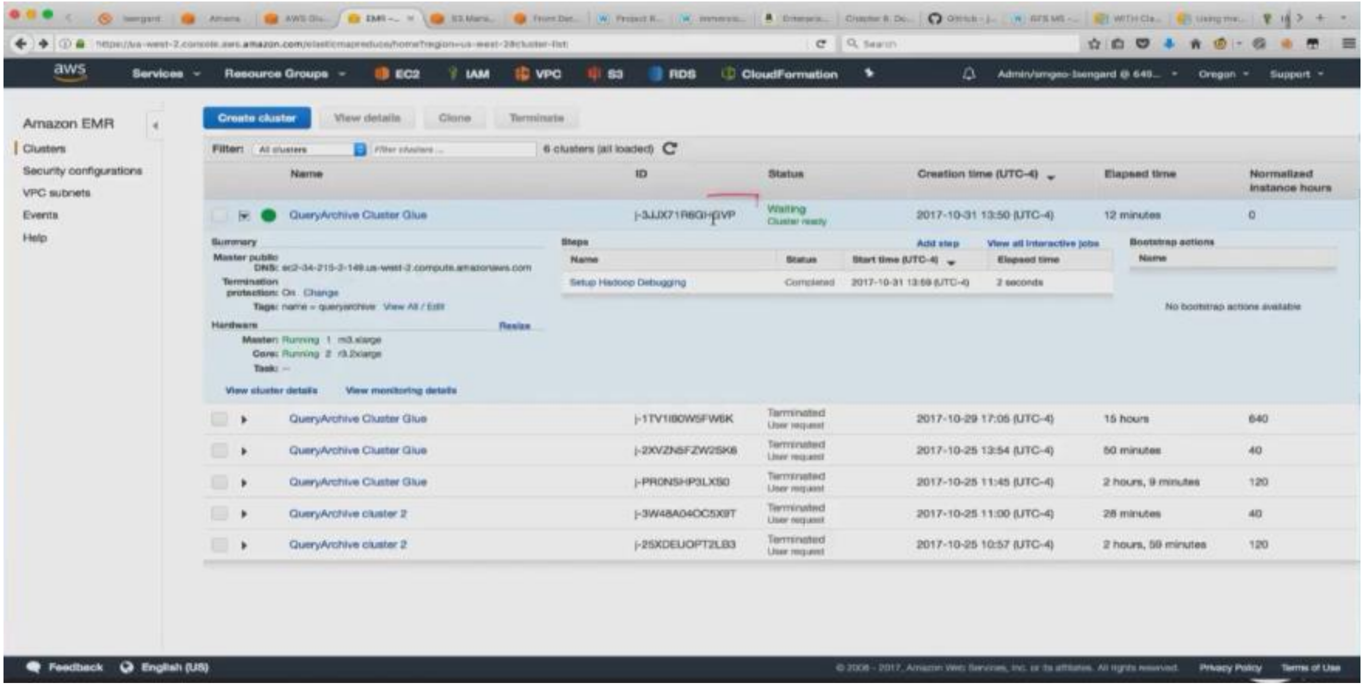
Steps Add step View all interactive jobs

Name	Status	Start time (UTC-4)	Elapsed time
Setup Hadoop Debugging	Pending		

Bootstrap actions

Name

No bootstrap actions available



Filter: All clusters | Filter clusters ... | 6 clusters (all loaded)

Name	ID	Status	Creation time (UTC-4)	Elapsed time	Normalized instance hours
QueryArchive Cluster Glue	j-3JJX71R6GHGVP	Waiting Cluster ready	2017-10-31 13:50 (UTC-4)	12 minutes	0
QueryArchive Cluster Glue	j-1TV1I80W5FW6K	Terminated User request	2017-10-29 17:05 (UTC-4)	15 hours	640
QueryArchive Cluster Glue	j-2XVZNSFZW2SK6	Terminated User request	2017-10-25 13:54 (UTC-4)	50 minutes	40
QueryArchive Cluster Glue	j-PR0NSHP3LXS0	Terminated User request	2017-10-25 11:45 (UTC-4)	2 hours, 9 minutes	120
QueryArchive cluster 2	j-3W48A04OC5X9T	Terminated User request	2017-10-25 11:00 (UTC-4)	28 minutes	40
QueryArchive cluster 2	j-25XDEUOPT2LB3	Terminated User request	2017-10-25 10:57 (UTC-4)	2 hours, 59 minutes	120

Summary

Master public DNS: ec2-34-215-2-149.us-west-2.compute.amazonaws.com

Termination protection: On Change

Tags: name = queryarchive View All / Edit

Hardware Resize

Master: Running 1 m3.xlarge

Core: Running 2 r3.2xlarge

Task: --

View cluster details View monitoring details

Steps Add step View all interactive jobs

Name	Status	Start time (UTC-4)	Elapsed time
Setup Hadoop Debugging	Completed	2017-10-31 13:59 (UTC-4)	2 seconds

Bootstrap actions

Name

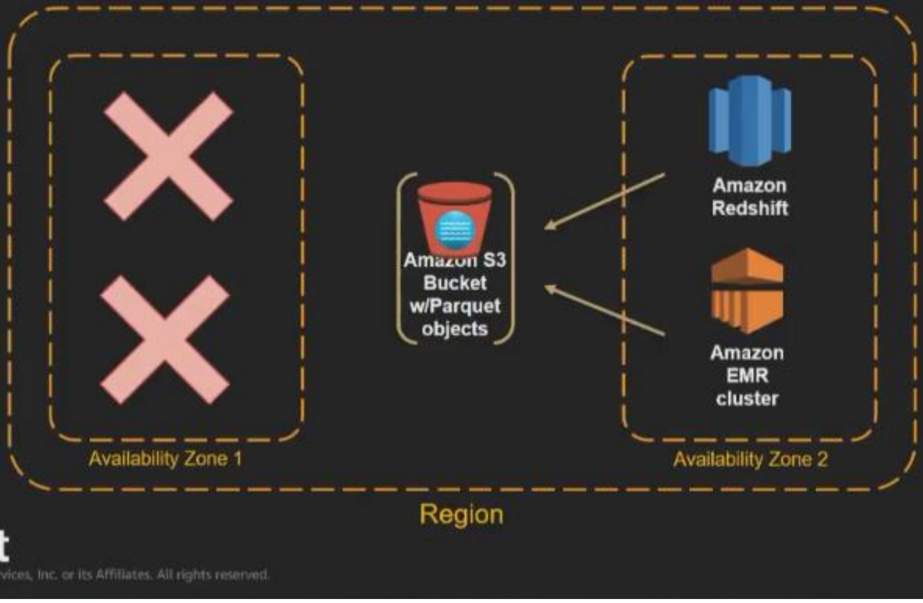
No bootstrap actions available

© 2008 - 2017, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Benefits

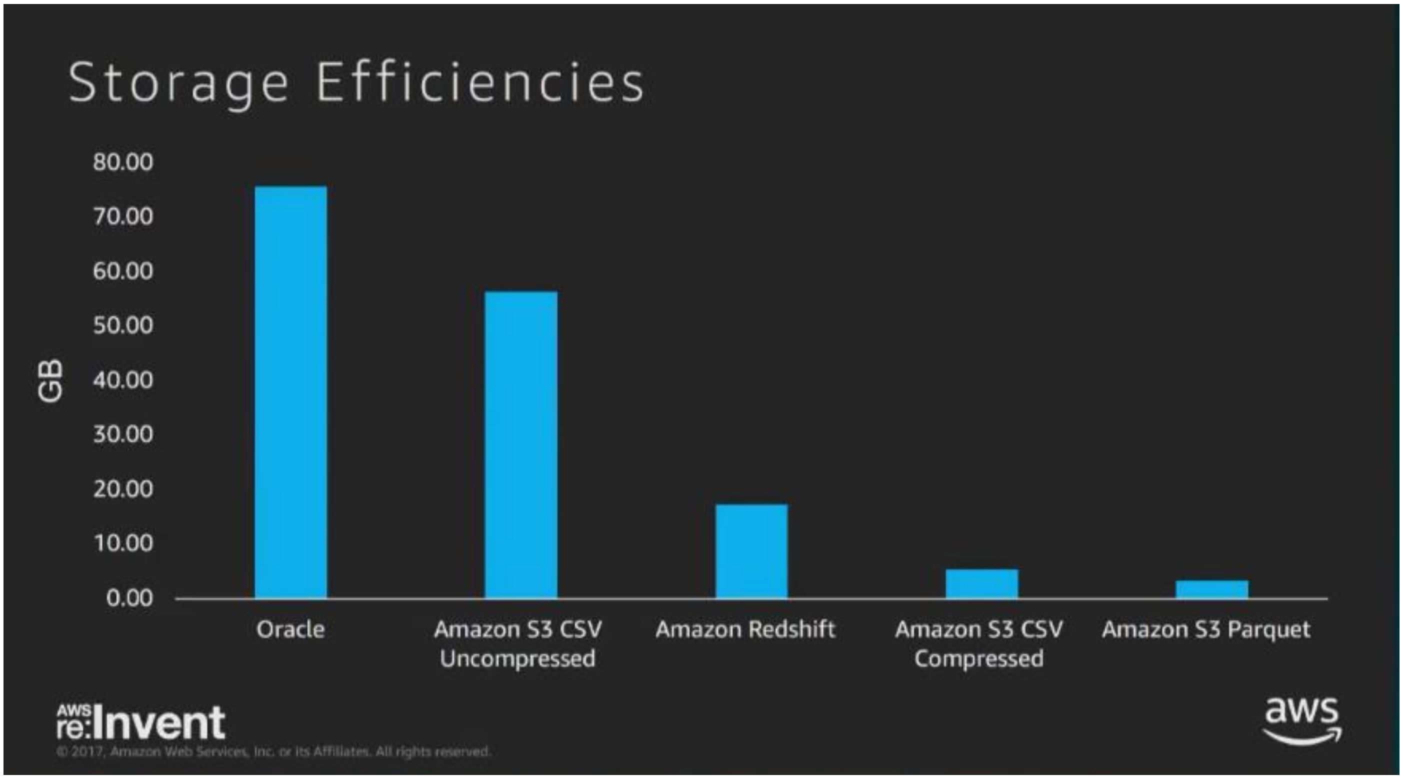
- Single golden source of truth
- Right sizing compute for workload
- Increased availability

A Queryable Archive + Data Lake On AWS

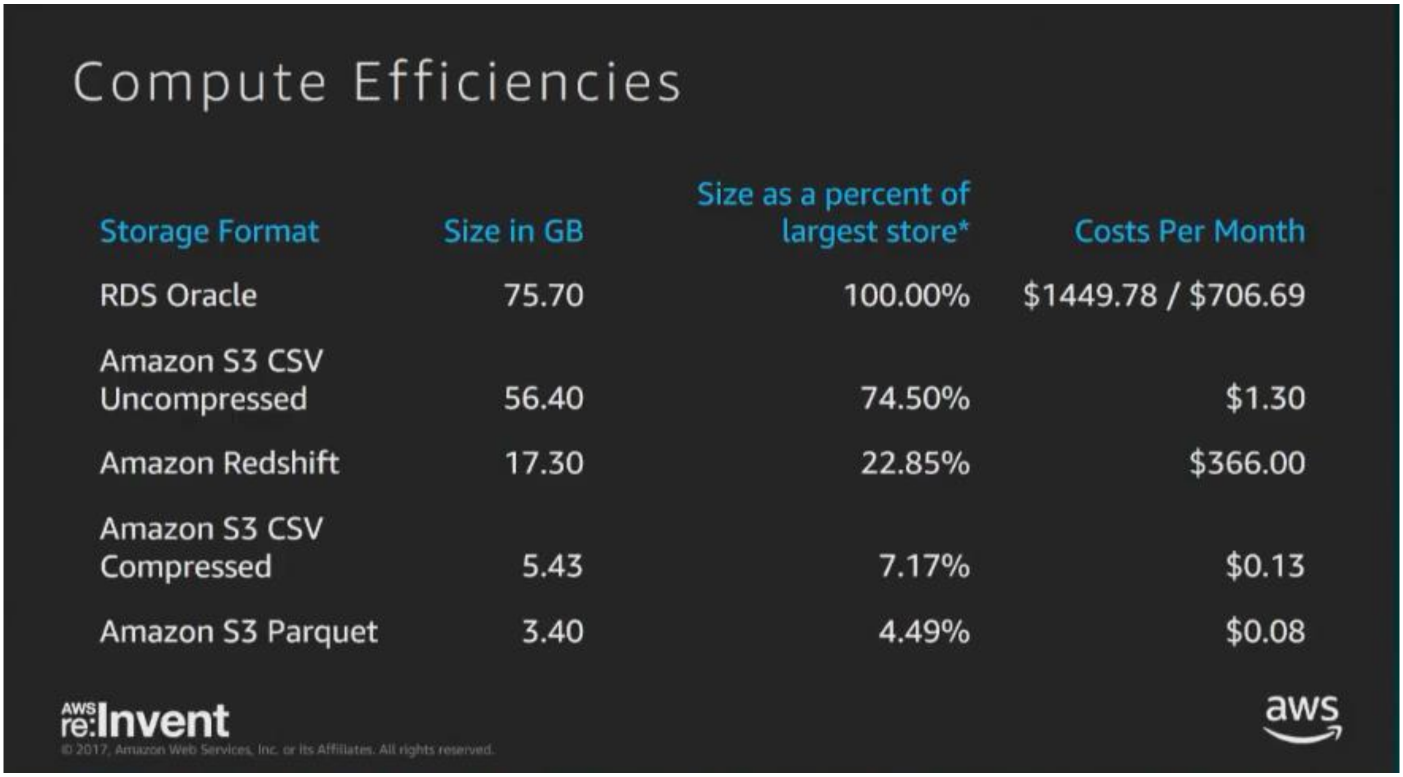


AWS re:Invent

aws

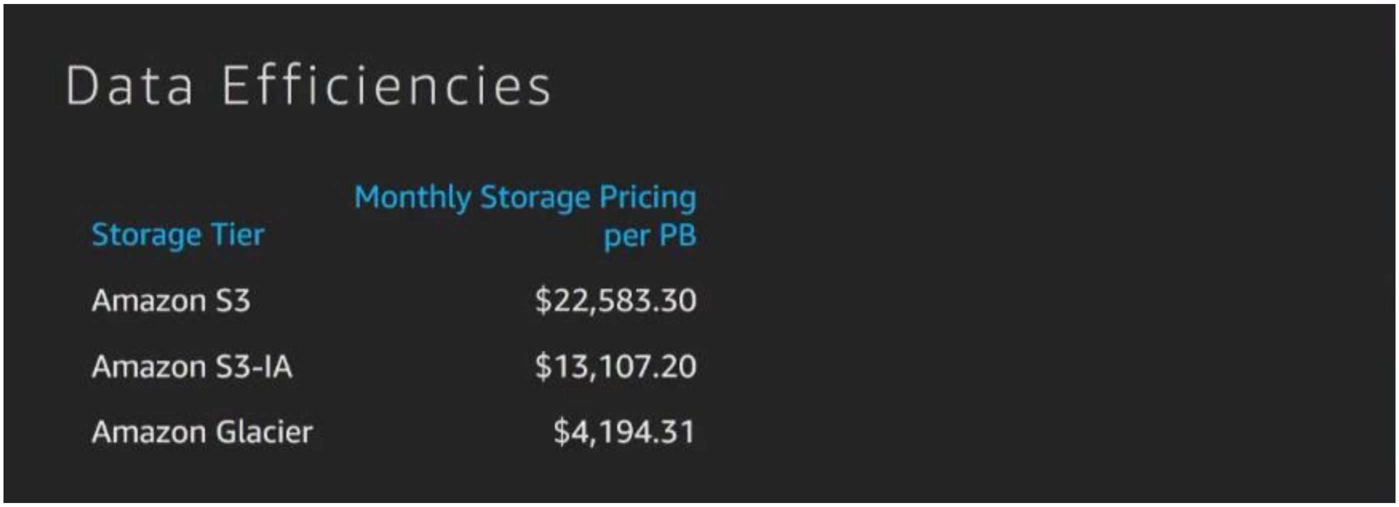


Storage Efficiencies



Compute Efficiencies

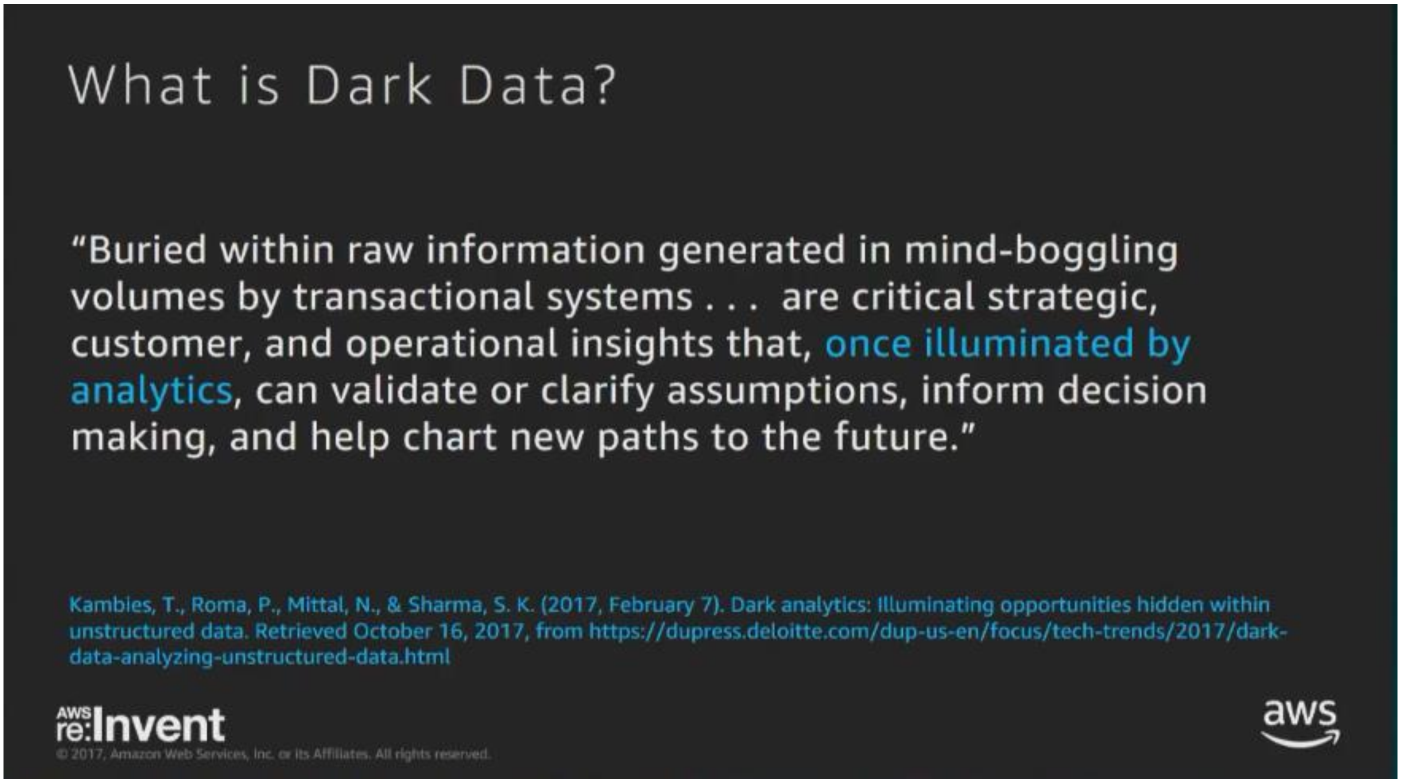
Storage Format	Size in GB	Size as a percent of largest store*	Costs Per Month
RDS Oracle	75.70	100.00%	$1449.78 / $706.69
Amazon S3 CSV Uncompressed	56.40	74.50%	$1.30
Amazon Redshift	17.30	22.85%	$366.00
Amazon S3 CSV Compressed	5.43	7.17%	$0.13
Amazon S3 Parquet	3.40	4.49%	$0.08



Data Efficiencies

Storage Tier	Monthly Storage Pricing per PB
Amazon S3	$22,583.30
Amazon S3-IA	$13,107.20
Amazon Glacier	$4,194.31

Unleashing "Dark Data"



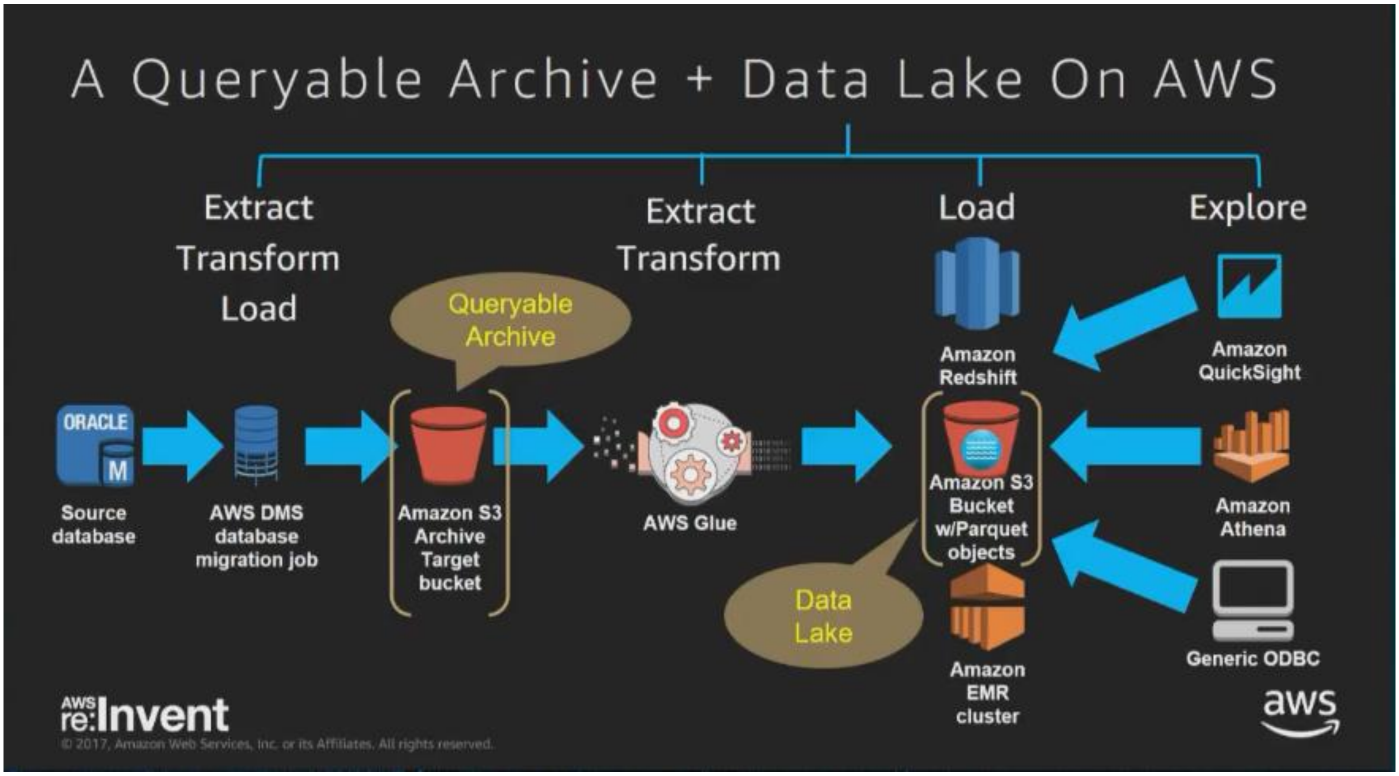
What is Dark Data?

"Buried within raw information generated in mind-boggling volumes by transactional systems . . . are critical strategic, customer, and operational insights that, once illuminated by analytics, can validate or clarify assumptions, inform decision making, and help chart new paths to the future."

Kambies, T., Roma, P., Mittal, N., & Sharma, S. K. (2017, February 7). Dark analytics: Illuminating opportunities hidden within unstructured data. Retrieved October 16, 2017, from https://dupress.deloitte.com/dup-us-en/focus/tech-trends/2017/dark-data-analyzing-unstructured-data.html

AWS re:Invent

aws



A Queryable Archive + Data Lake On AWS

Extract
Transform
Load

Extract
Transform

Load

Explore

Queryable Archive

Source database

AWS DMS database migration job

Amazon S3 Archive Target bucket

AWS Glue

Amazon Redshift

Amazon S3 Bucket w/Parquet objects

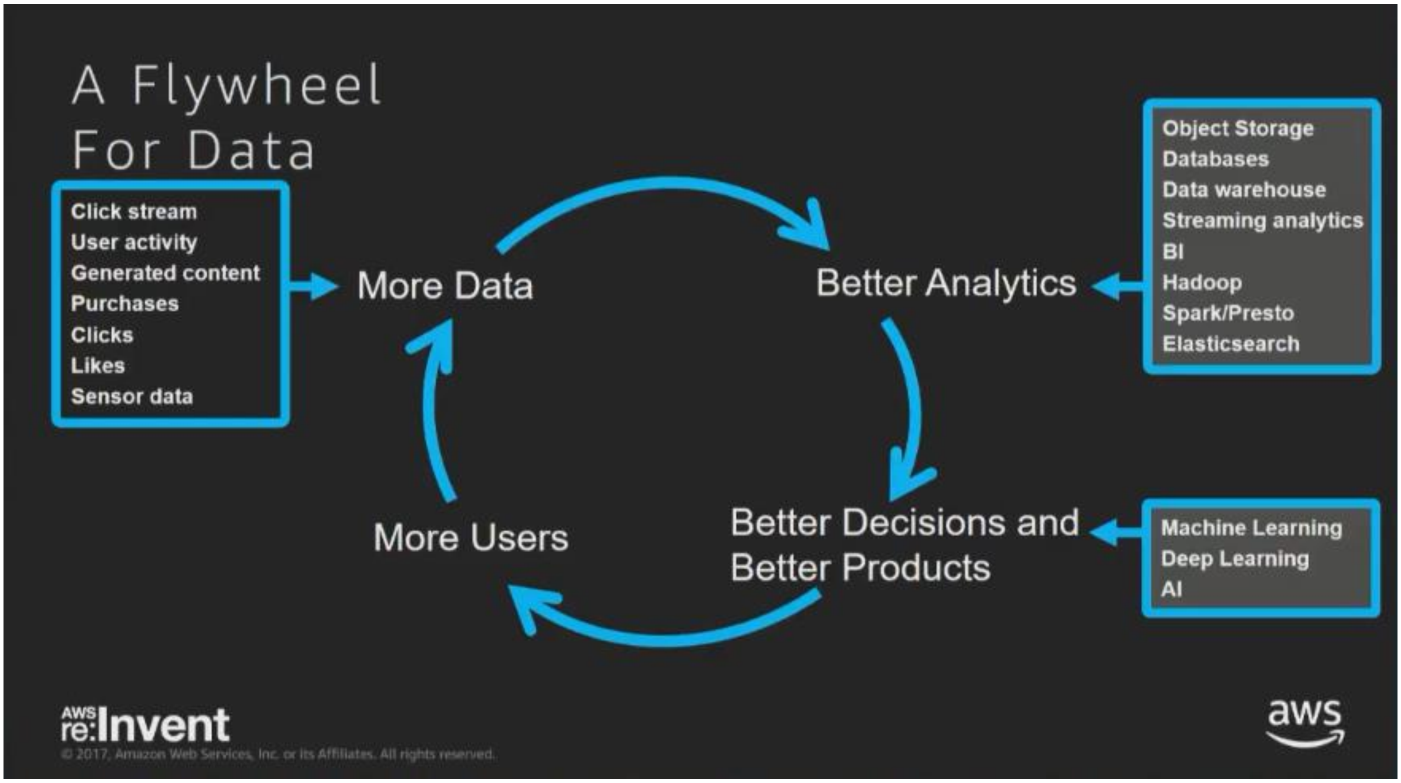
Amazon EMR cluster

Data Lake

Amazon QuickSight

Amazon Athena

Generic ODBC



A Flywheel For Data

Click stream
User activity
Generated content
Purchases
Clicks
Likes
Sensor data

More Data

Better Analytics

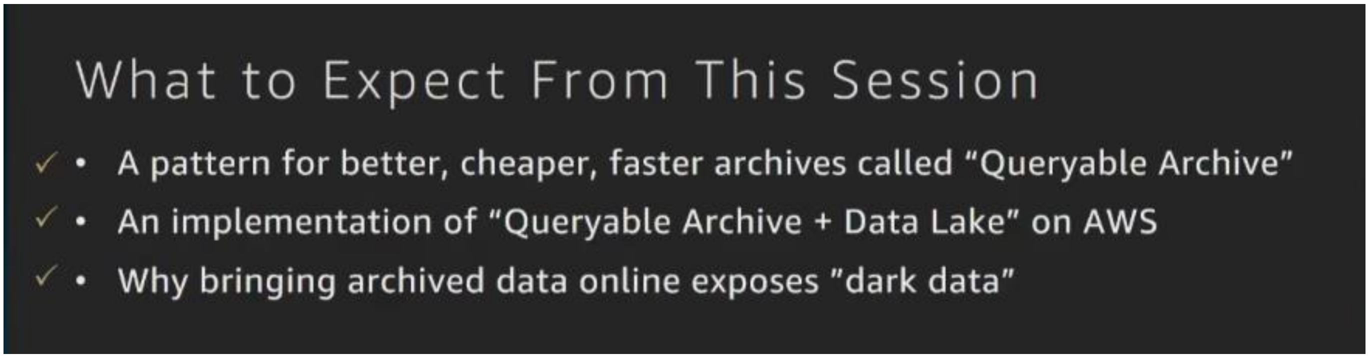
Better Decisions and Better Products

More Users

Object Storage
Databases
Data warehouse
Streaming analytics
BI
Hadoop
Spark/Presto
Elasticsearch

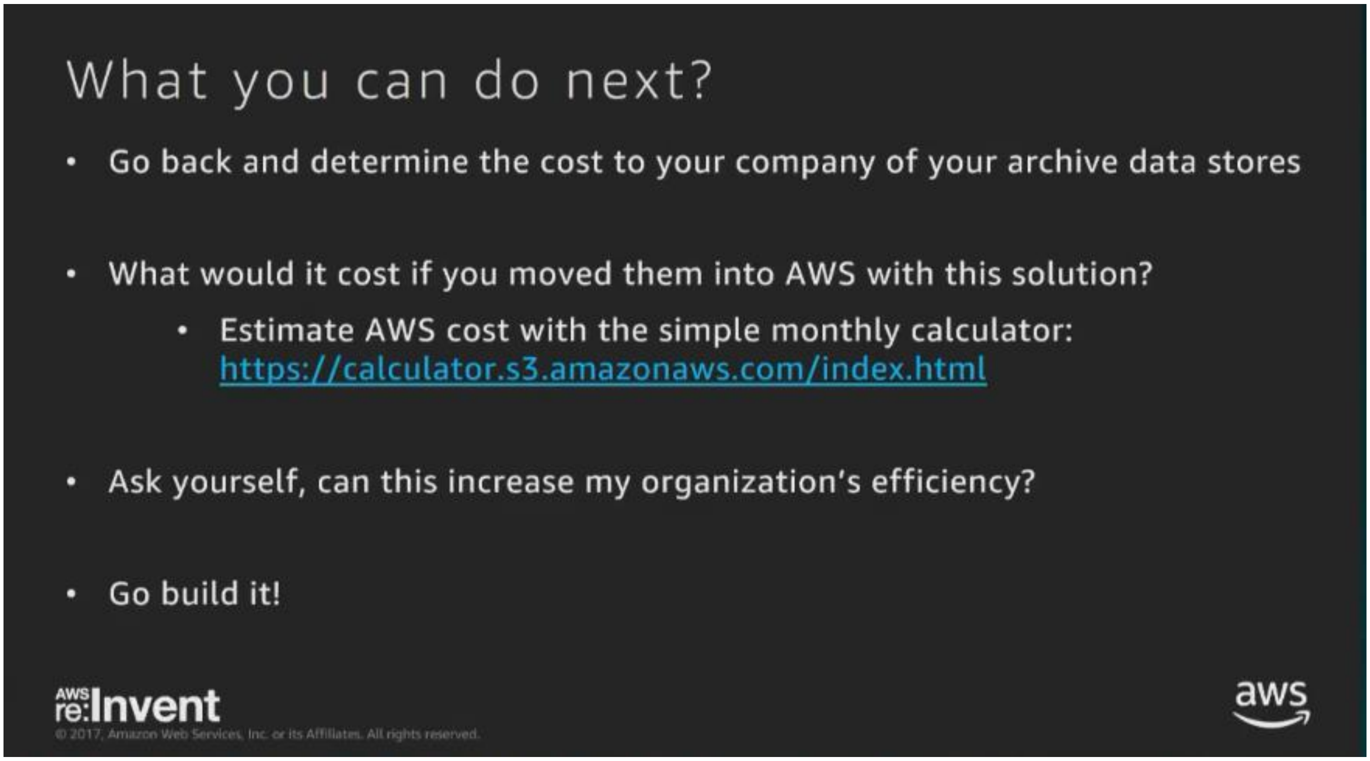
Machine Learning
Deep Learning
AI

Next Steps



What to Expect From This Session

- ✓ A pattern for better, cheaper, faster archives called “Queryable Archive”
- ✓ An implementation of “Queryable Archive + Data Lake” on AWS
- ✓ Why bringing archived data online exposes ”dark data”



What you can do next?

- Go back and determine the cost to your company of your archive data stores
- What would it cost if you moved them into AWS with this solution?
  - Estimate AWS cost with the simple monthly calculator: https://calculator.s3.amazonaws.com/index.html
- Ask yourself, can this increase my organization’s efficiency?
- Go build it!

Thank you

George Smith, Global Financial Services Solutions Architect at AWS
smgeo@amazon.com