AWS re:Invent

ANT406-R

# Build a single query to analyze data across Amazon Redshift and Amazon S3

**Jenny Chen**

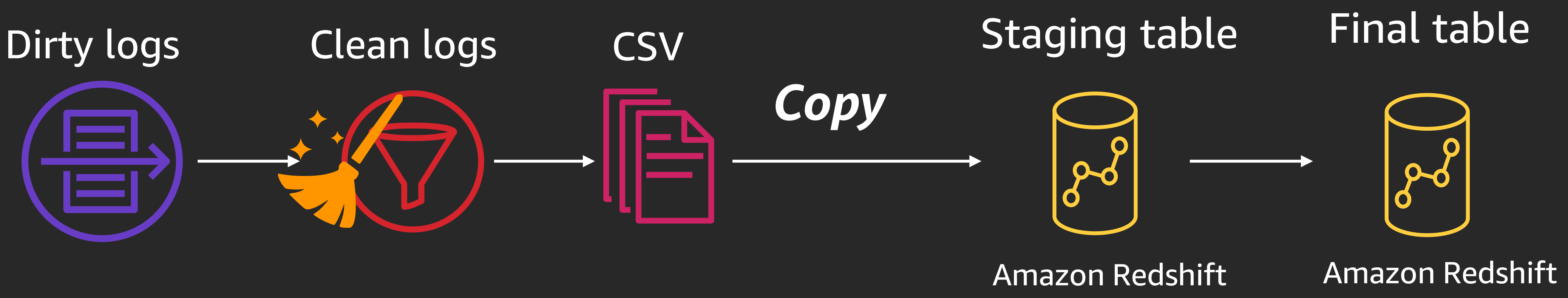
Database Engineer – Amazon Redshift

Amazon Web Services

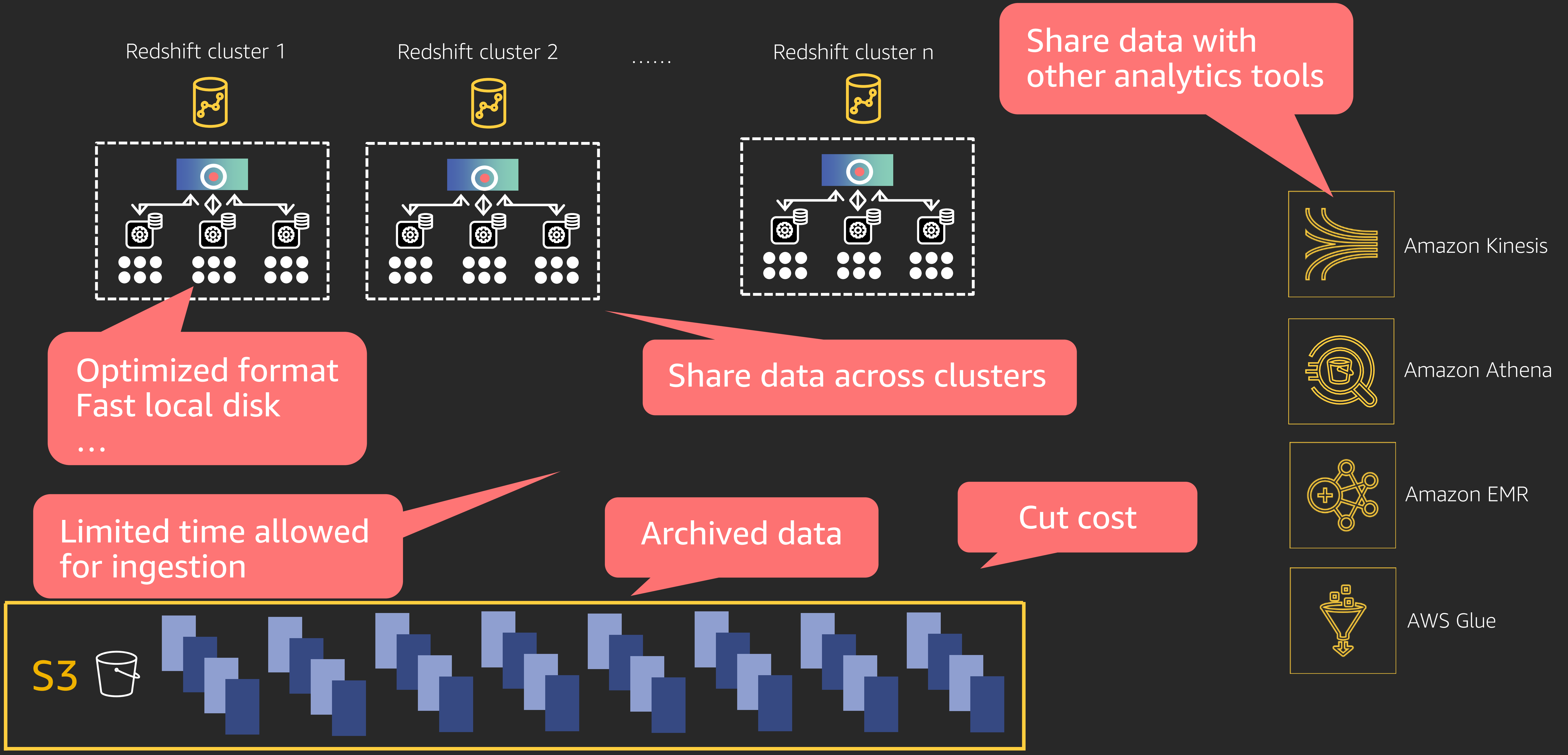
# Why would you query data across Amazon Redshift & Amazon S3?

# Classic ingestion

Clean and transform before and after copy into Amazon Redshift



# Why to have data in Amazon Redshift and Amazon S3



Share data with other analytics tools

Optimized format
Fast local disk
…

Share data across clusters

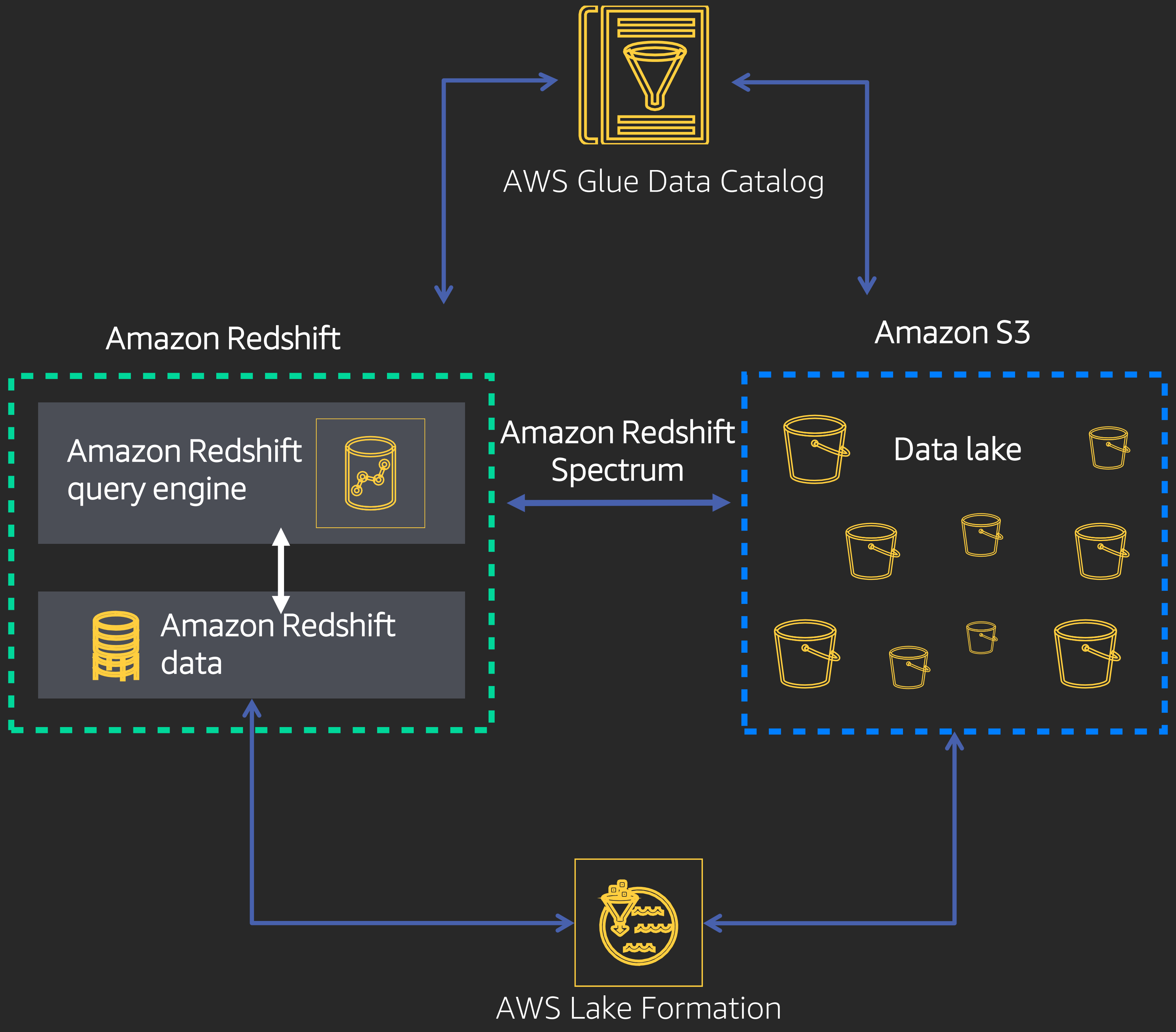
Limited time allowed for ingestion

Archived data

Cut cost

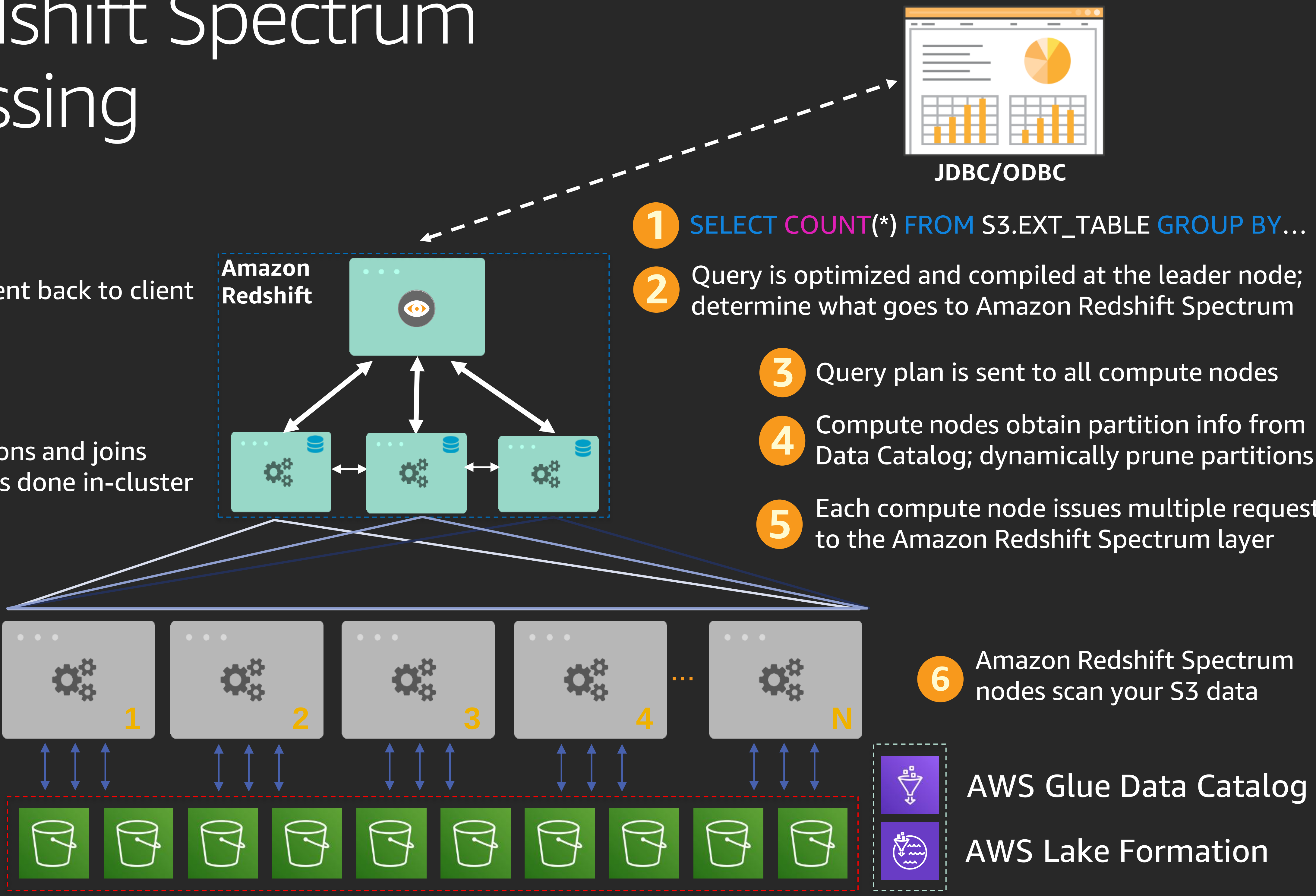
# With data in Amazon Redshift and Amazon S3, how to query both?

# What you need to build



- Amazon Redshift cluster
  - Amazon Redshift Spectrum
- External catalog
- Data in Amazon S3

# Amazon Redshift Spectrum query processing



1 SELECT COUNT(*) FROM S3.EXT_TABLE GROUP BY…

2 Query is optimized and compiled at the leader node; determine what goes to Amazon Redshift Spectrum

3 Query plan is sent to all compute nodes

4 Compute nodes obtain partition info from Data Catalog; dynamically prune partitions

5 Each compute node issues multiple requests to the Amazon Redshift Spectrum layer

6 Amazon Redshift Spectrum nodes scan your S3 data

7 Amazon Redshift Spectrum filters, joins, and aggregates

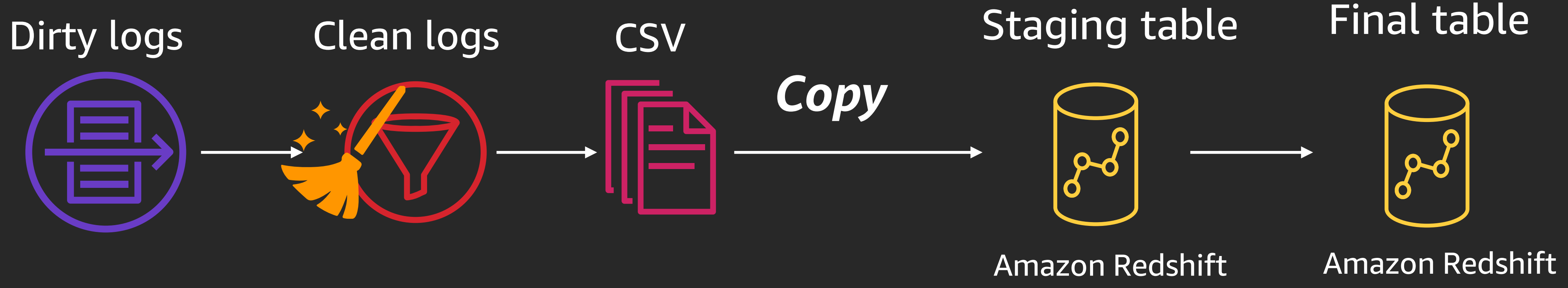
8 Final aggregations and joins with local tables done in-cluster

9 Result is sent back to client

# Simplified ingestion

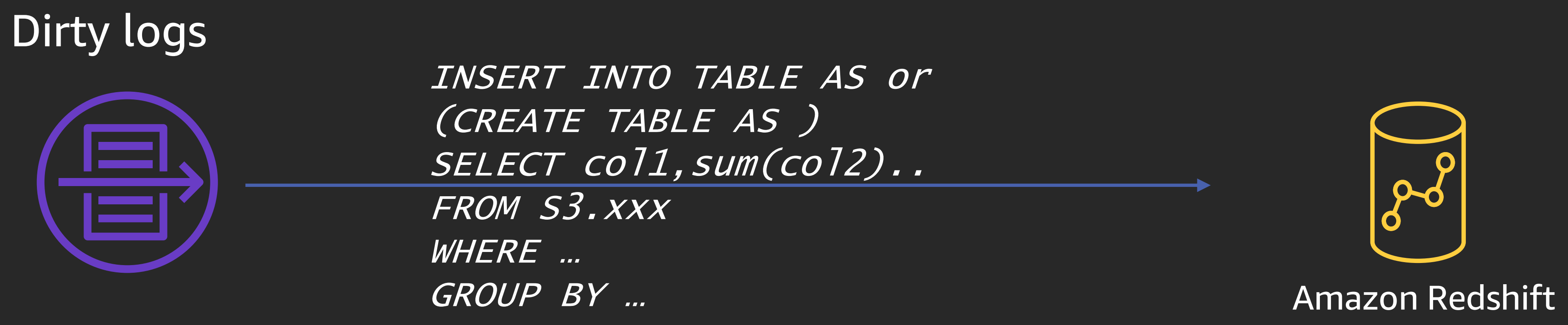
## Clean and transform before and after copy into Amazon Redshift

### Before



Dirty logs → Clean logs → CSV → *Copy* → Staging table (Amazon Redshift) → Final table (Amazon Redshift)

### After



Dirty logs → Amazon Redshift

```
INSERT INTO TABLE AS or
(CREATE TABLE AS )
SELECT col1,sum(col2)..
FROM S3.xxx
WHERE …
GROUP BY …
```

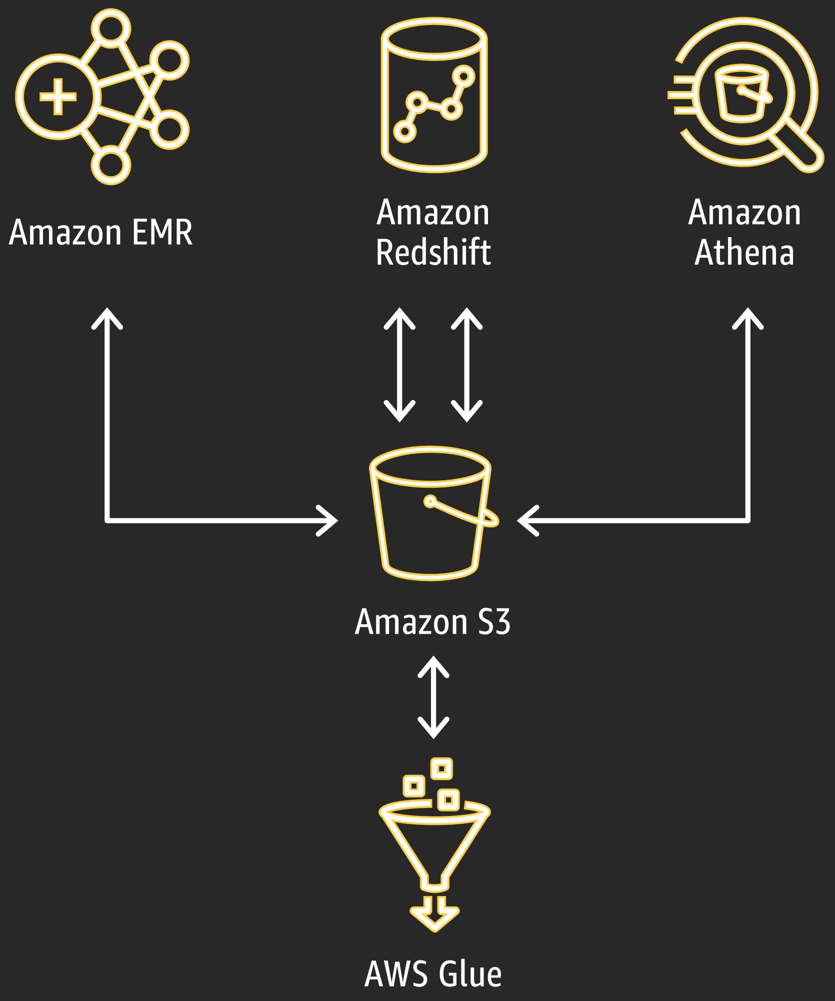
# Demos

# Demos

- Demo 1: Query Amazon Redshift audit logs, unload data to Parquet
- Demo 2: Ways to conjunct data in Amazon Redshift and data in Amazon S3
- Demo 3: Integration with Lake Formation
- Demo 4: Query AWS CloudTrail logs (nested JSON)

# Demo 1: Unload Amazon Redshift audit logs as Parquet to S3 with built-in auto partition

Amazon Redshift now supports exporting data to Amazon S3 in Parquet format. This makes **sharing data across the data lake easier and faster, without conversion.**



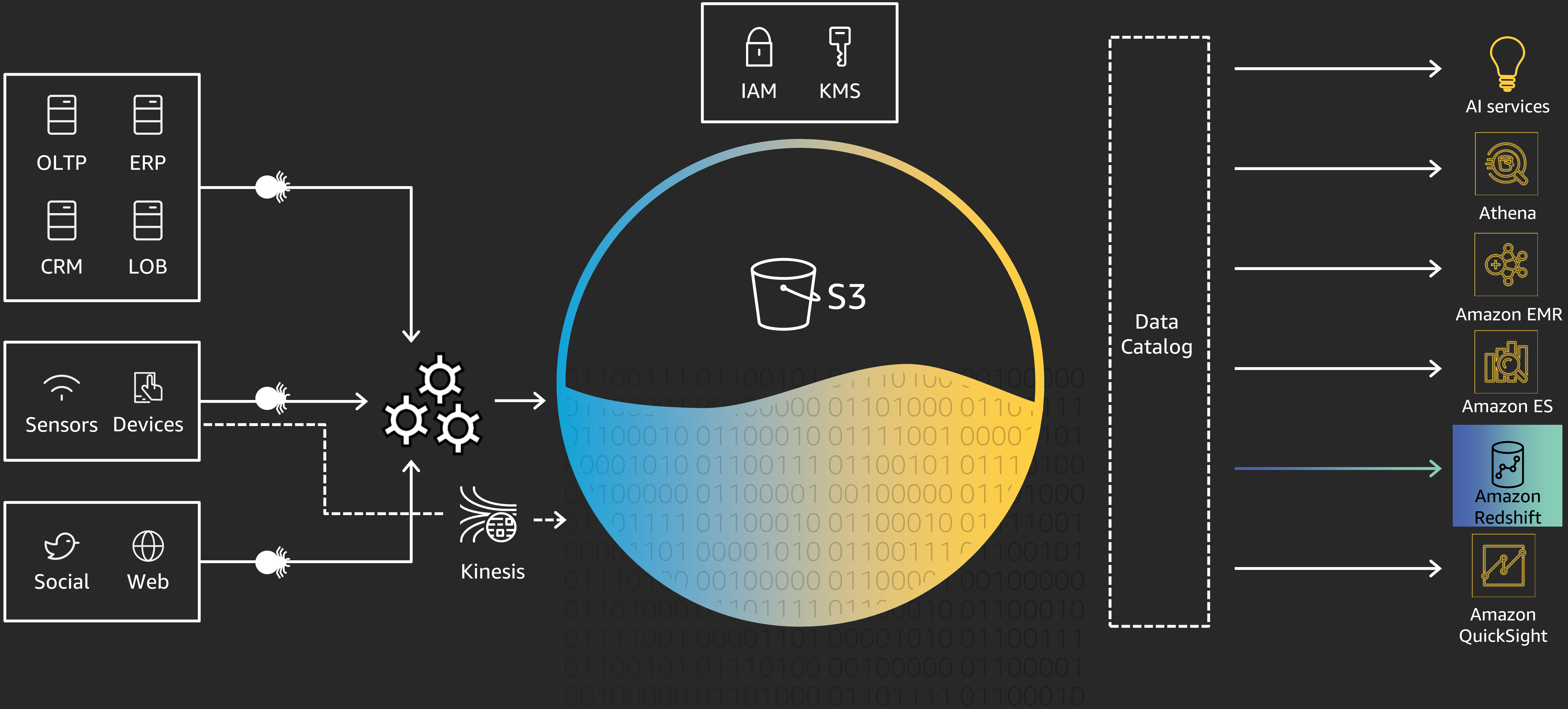
**Parquet is an open data format** supported by Amazon EMR, Amazon Athena, and Amazon Redshift

```
UNLOAD
('select * from lineitem')
TO
's3://mybucket/unload/lineitem/'
FORMAT as PARQUET
PARTITION BY (cdate);
```

# Demo 2: Conjunction data in Amazon Redshift and S3

- Join between small local dimension table and large external fact table
- Using UNION ALL between cold and hot data
- Using late-binding view as unified interface

# Demo 3: Integration with Lake Formation



# Demo 4: Query CloudTrail logs (nested JSON)

Analyze nested and semi-structured data in Amazon S3 with Amazon Redshift Spectrum
Allows easy ETL (extract, transform, and load) of nested data into Amazon Redshift using CTAs
Support for open file formats: Parquet, ORC, JSON, Ion, and Avro
Support for struct, map, and array
Uses dot notation to extend your existing SQL

Example: Find click frequency for links on “/home”:

```
s3data.clickStream: <<
{ “session_time”: “20171013 14:05:00”,
  “clicks”: [ {“page”: “/home”, “referrer”: “”},
              {“page”: “/products”, “referrer”:
“/home”} ]
},
{ “session_time”: “20171013 14:06:00”,
  “clicks”: [ {“page”: “/contact”, “referrer”:
“/home”} ]
} >>
```

```
SELECT c.page,
            COUNT(*) AS count
FROM s3data.clickStream s,
          s.clicks c
WHERE s.session_time > ‘2017-10-01
00:00:00’
      AND c.referrer = “/home”
GROUP BY c.page;
```

# Q&A

# Learn big data with AWS Training and Certification

Resources created by the experts at AWS to help you build and validate data analytics skills

New free digital course, Data Analytics Fundamentals, introduces Amazon S3, Amazon Kinesis, Amazon EMR, AWS Glue, and Amazon Redshift

Classroom offerings, including Big Data on AWS, feature AWS expert instructors and hands-on labs

Validate expertise with the **AWS Certified Big Data - Specialty** exam or the new **AWS Certified Data Analytics - Specialty** beta exam

Visit aws.amazon.com/training/paths-specialty/

# Thank you!

**Jenny Chen**
chenjuan@amazon.com

Please complete the session survey in the mobile app.

AWS re:Invent

aws