AWS re:Invent

ANT314-R

Extend analytics beyond your data warehouse

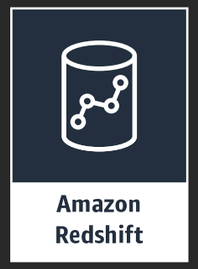
Ippokratis Pandis
Senior Principal Engineer
Amazon Web Services

Eric Kamm
Senior Engineer
Annalect

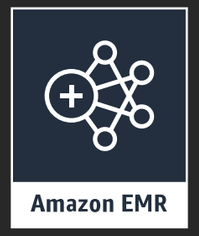
© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

What is analytics?

- Analysis based on data collected by transactional/customer-facing operations
- Everything is quantifiable and measurable
- Facilitate data-driven decisions and operations

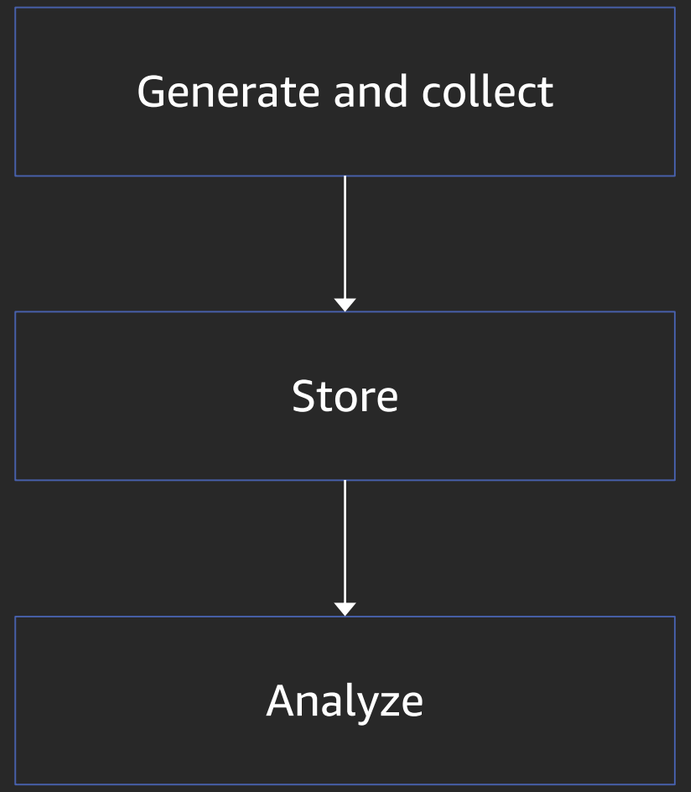








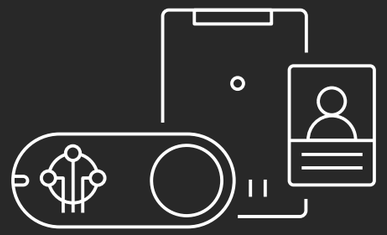


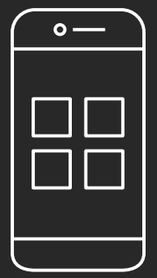


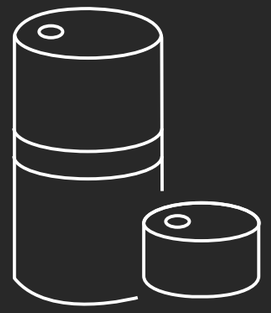
It has never been easier to collect data

New Report

First Nam	Last Nam	Action ID	Application	Click Date	Clicks	First Paramet	Second Parameter	Third Parameter
[illegible]	[illegible]	/AnswerView	End-User p	02/11/2010 12:25 PM	Gs6V	2688	Creating a report for specific user transactions	No Value
[illegible]	[illegible]	/AnswerView	End-User p	02/11/2010 12:25 PM	Gs6V	2688	Creating a report for specific user transactions	No Value
[illegible]	[illegible]	/AnswerList/Sear	End-User p	02/11/2010 12:25 PM	Gs6V	clicktrack	No Value	No Value
[illegible]	[illegible]	/AnswerList/Sear	End-User p	02/11/2010 12:25 PM	Gs6V	clicktrack	No Value	No Value
[illegible]	[illegible]	/social/ssoError	End-User p	02/11/2010 12:22 PM	Gs6V	No Value	No Value	No Value
[illegible]	[illegible]	/error	End-User p	02/11/2010 12:22 PM	Gs6V	No Value	No Value	No Value







APACHE

NGINX

Generate and collect

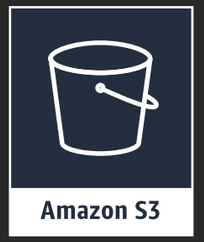
Store

Analyze

It has never been easier and cheaper to store data

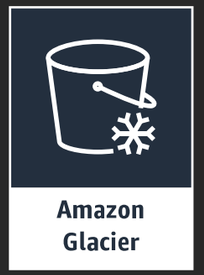
Amazon Simple Storage Service (Amazon S3)

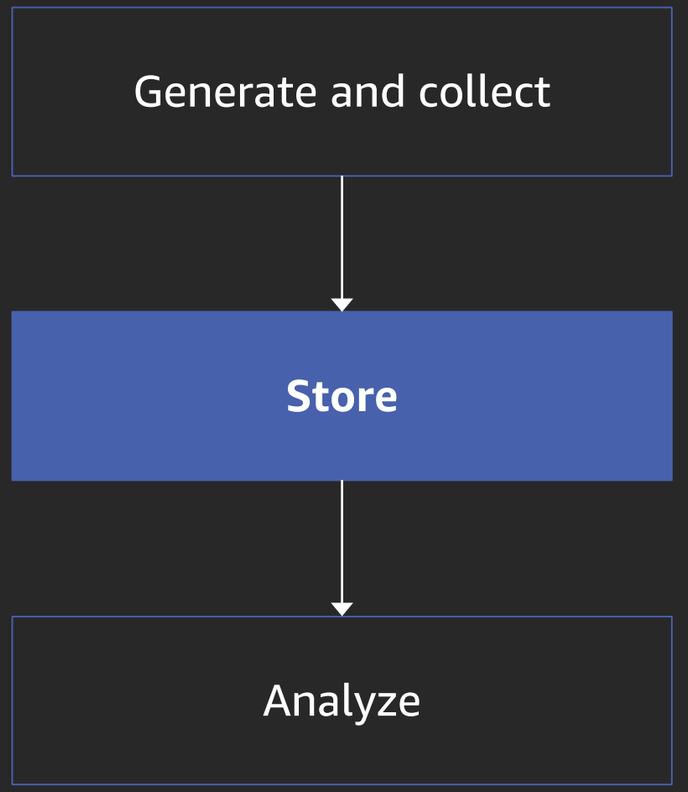
- S3: $23.4 per TB/mo
- S3 IA: $12.8 per TB/mo



Amazon Simple Storage Service Glacier

- Glacier: $4.1 per TB/mo
- Glacier Deep Archive: $1 per TB/mo





The tyranny of “OR” in Analytics

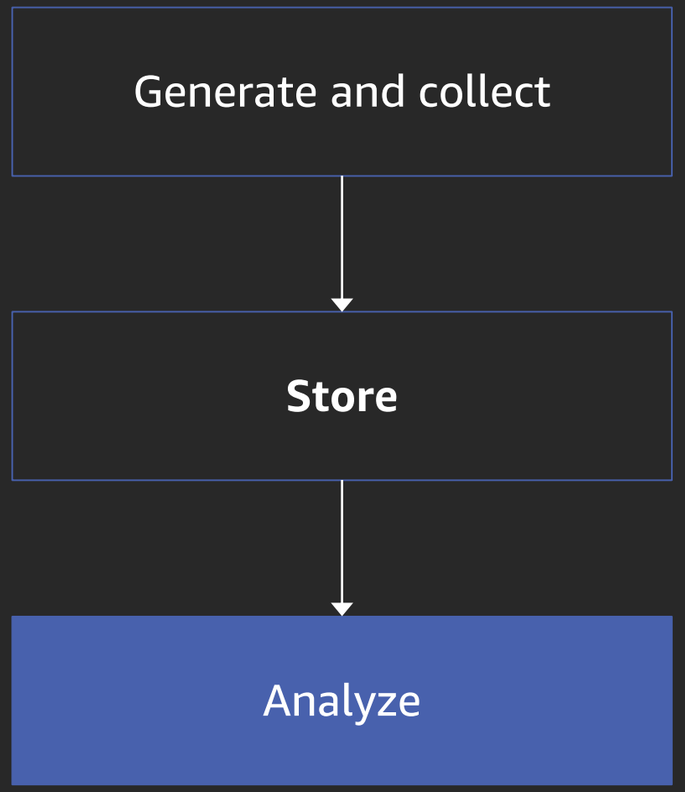
Parallel data warehouse

- Parallel database
- Own storage manager & transactionality
- Proprietary columnar data format

versus

Big data framework

- Open-source, large-scale distributed query engines
- Operate on data stored on external storage (S3, HDFS)
- Loose control over the stored data



Escaping the tyranny of “OR” with Amazon Redshift

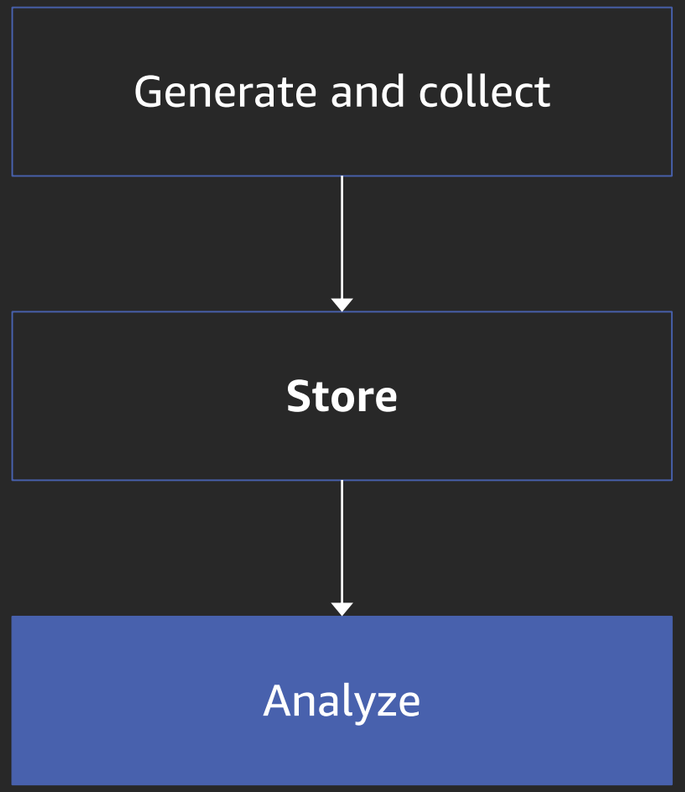
Top-of-the-line performance for any

Data scale

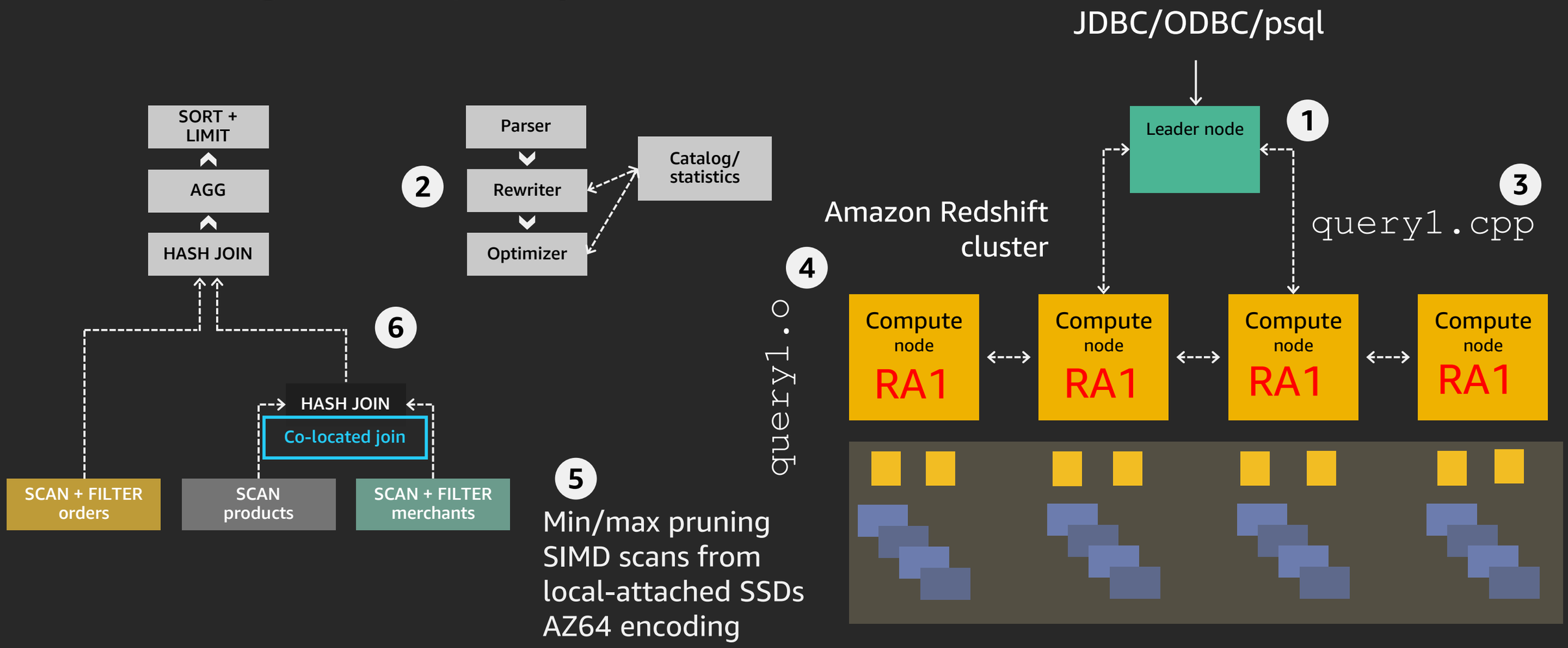
- a/ Local attached data
- b/ Very large data in Amazon S3

Concurrency

- a/ Capacity of a single cluster
- b/ Capacity of multiple clusters



Executing this query in Amazon Redshift



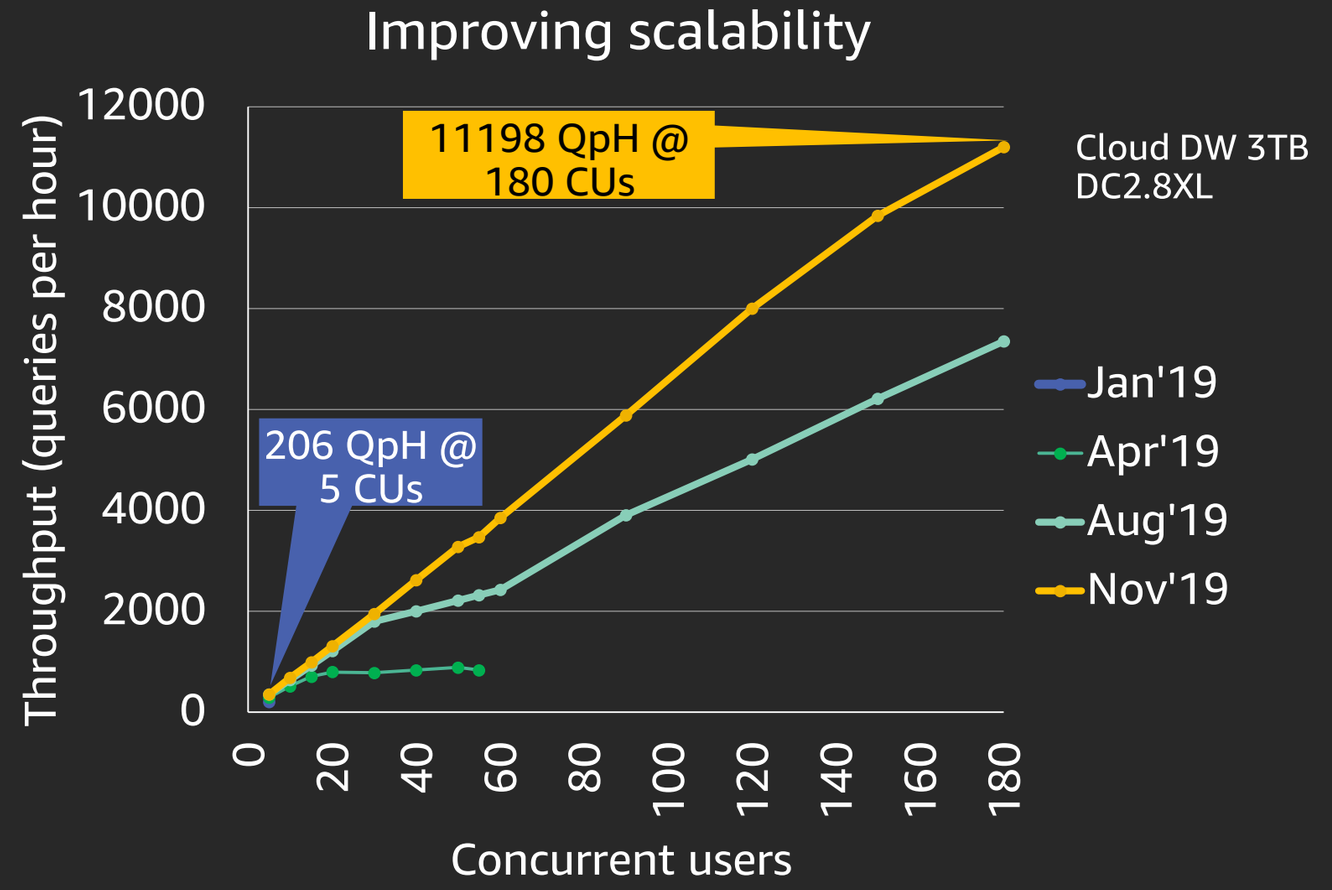
Top-of-the-line data warehousing performance

Continuous performance and scalability improvements

Maintain performance and scalability leadership through continuous telemetry and benchmarking

Within 2019

- Concurrency scaling
  - Added support for Amazon Redshift Spectrum, Unloads, CTEs, etc.
- AZ64 encoding
- Global compiled code cache
- AutoWLM & query priorities
- Auto Analyze & Auto Vacuum
- DistKey & SortKey advisor
- Runtime filters (bloom filters)
- Materialized views
- Unload to Parquet



>50x improvement in concurrency in 2019

Data lake: The new information hub

- A **centralized, secure repository** that enables you to **govern, discover, share,** and **analyze structured and unstructured data** at any scale

Typical steps of building a data lake

Ingestion and cleaning



Data engineer

Security

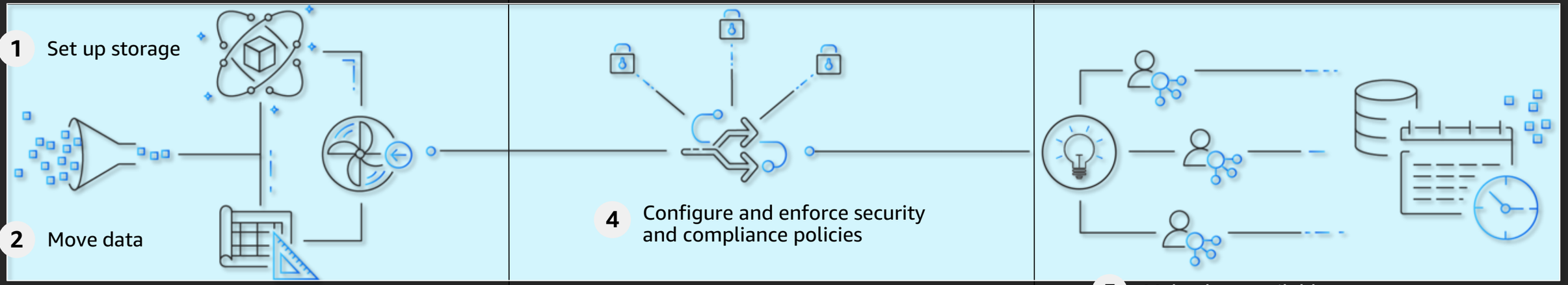


Data security officer

Analytics and ML



Data analyst



3 Cleanse, prep, and catalog data

5 Make data available for analytics

AWS Lake Formation

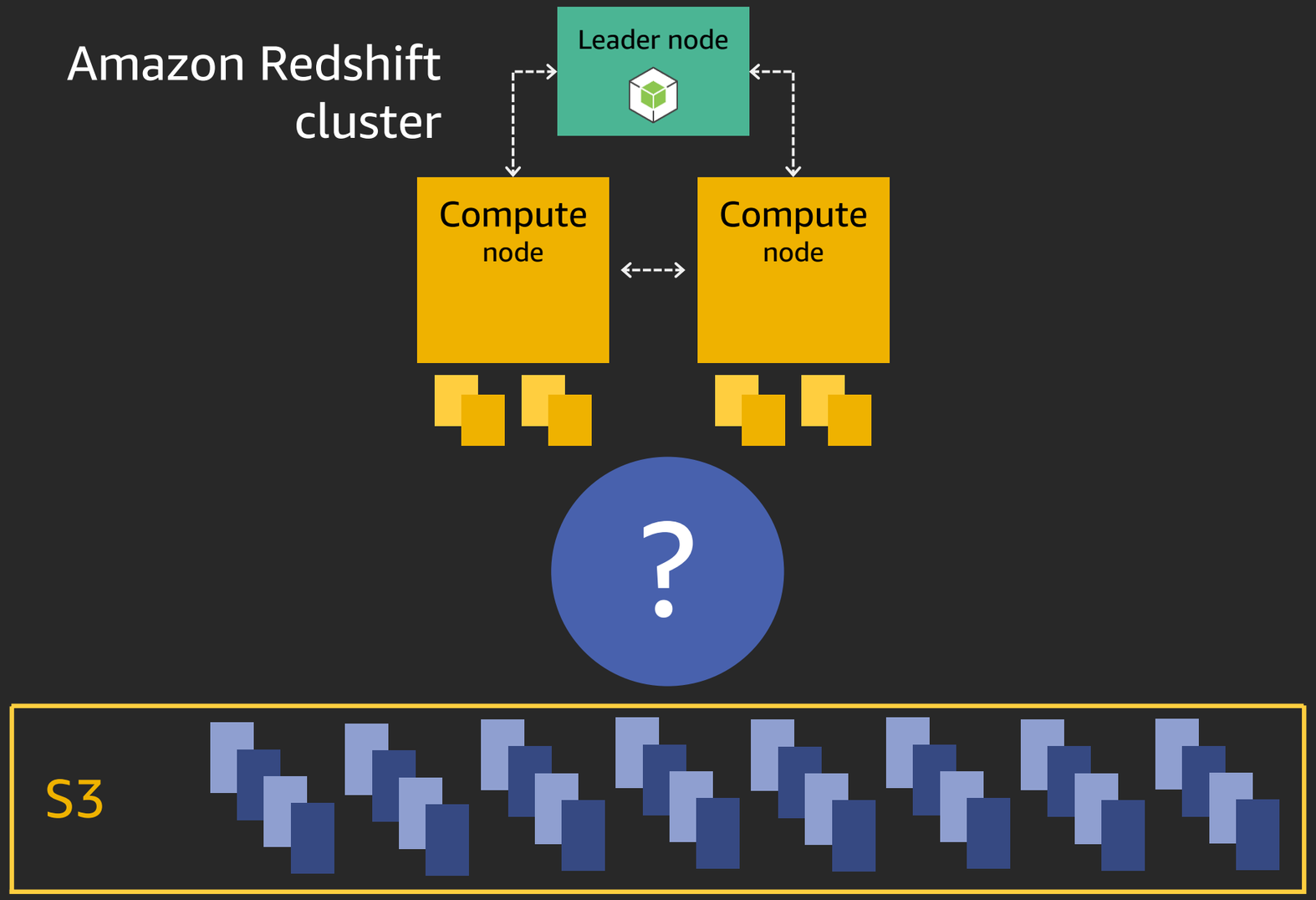
Lake Formation lets you

build secure data lakes in **days**

Amazon Redshift and AWS Lake Formation

- Easy integration
- Column-level security with Lake Formation

Querying big data from Amazon S3



Fast, large-scale queries out of Amazon S3

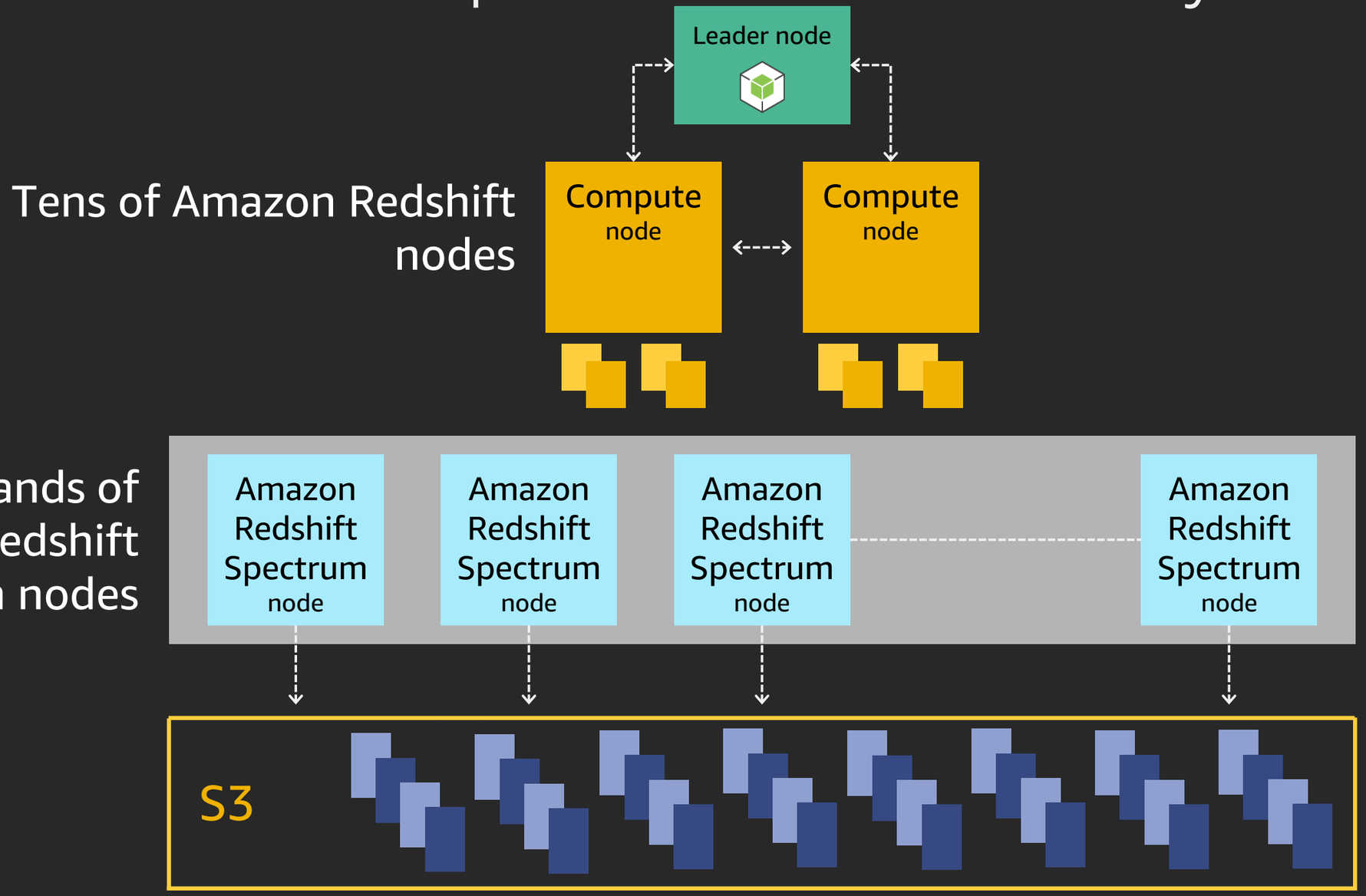
1. **High bandwidth**
   - Parallelism (many small straws)

2. **Minimize the amount of data to read from Amazon S3**
   - Avoid doing unnecessary work
   - Columnar formats and compression

3. **Reduce the amount of data sent back to Amazon Redshift**
   - Computation push-down

4. **Exploit repeating nature of queries**
   - Smart and incremental caching of intermediate results

Amazon Redshift Spectrum execution layer



Partitioning pruning – Static and dynamic

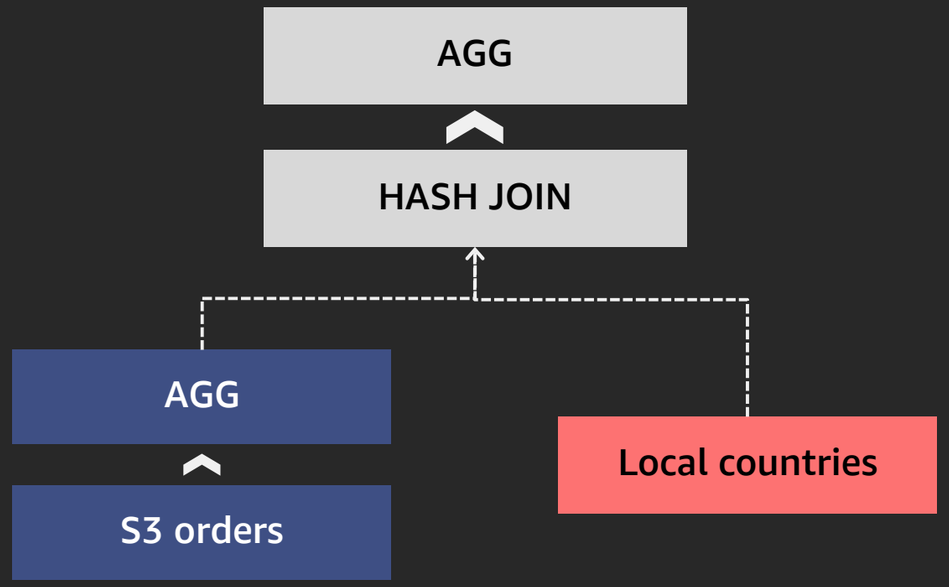
Partitioning scheme:
s3://Orders/ [YYYY] / [MM] / [DD] / [Country] /

20	365	195
1	1	195
1	1	12

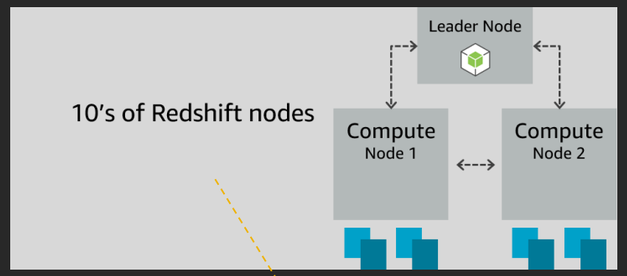
Partitions to process

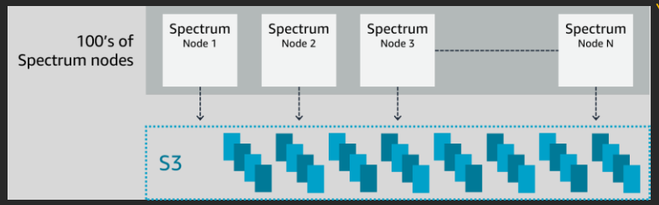
1,423,500
195
12

```
SELECT country, count(*)
FROM s3.Orders o, local.Countries c
WHERE o.CountryID = c.ID                          DYNAMIC
AND c.Continent = 'South America'
AND o.YYYY = 2017 AND o.MM = 12                   STATIC
AND o.DD = 24
GROUP BY 1;
```



Executing this query in Amazon Redshift Spectrum





SORT + LIMIT

AGG

HASH JOIN

HASH JOIN

SCAN orders

Tens of billions

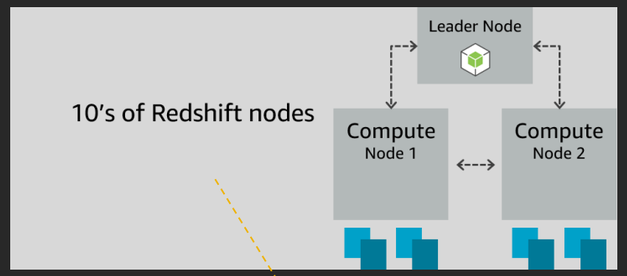
SCAN products

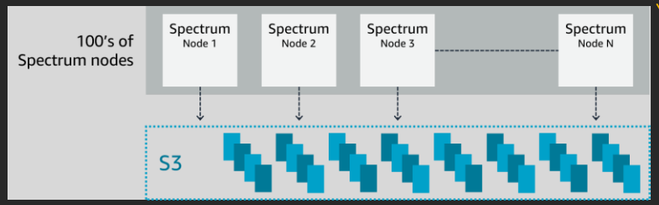
Millions

SCAN + FILTER merchants

Hundreds of thousands

Executing this query in Amazon Redshift Spectrum





SORT + LIMIT

AGG

HASH JOIN

TELEMETRY

Up to order of magnitude reduction

AGG

SCAN orders

Tens of billions

HASH JOIN

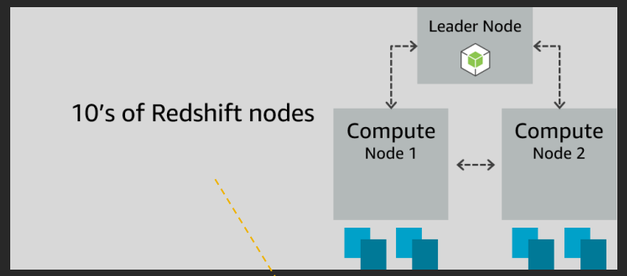
SCAN products

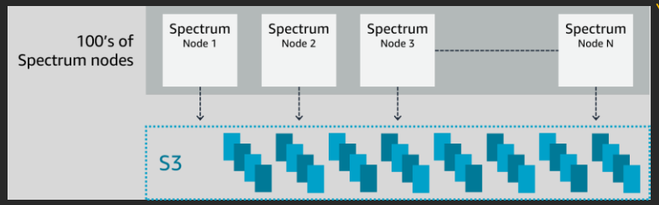
Millions

SCAN + FILTER merchants

Hundreds of thousands

Amazon Redshift Spectrum Request Accelerator





SORT + LIMIT

AGG

HASH JOIN

TELEMETRY

"Smart" caching layer

AGG

SCAN orders

Tens of billions

HASH JOIN

SCAN products

Millions

SCAN + FILTER merchants

Hundreds of thousands

Significant acceleration of repeating queries; incremental

Using Amazon Redshift with Redshift Spectrum for elastic ELT

“Smart” transformation requiring complex SQL transformations

Phase one

- Ingest using simple file-based COPY-command, or
- Smart pre-filtering, aggregation, and cleansing with Amazon Redshift Spectrum

Phase two

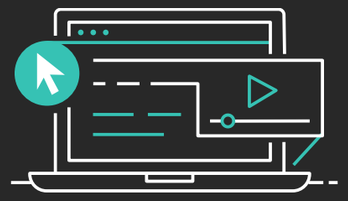
- Execute of business logic and other transformations using SQL UDFs, Stored Procedures, etc.

Phase three

- Retain data in Amazon Redshift optimized for local high-performance query execution, or
- UNLOAD to Parquet files for data lake–ready contribution

Learn analytics 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, Amazon Redshift, Amazon Athena, and AWS Lake Formation



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/

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws training and certification

Annalect

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Agenda

Our story of analytics on AWS

What we set out to deliver

How we do it

Tips and optimizations

The future

Our story of analytics on AWS

AWS re:Invent

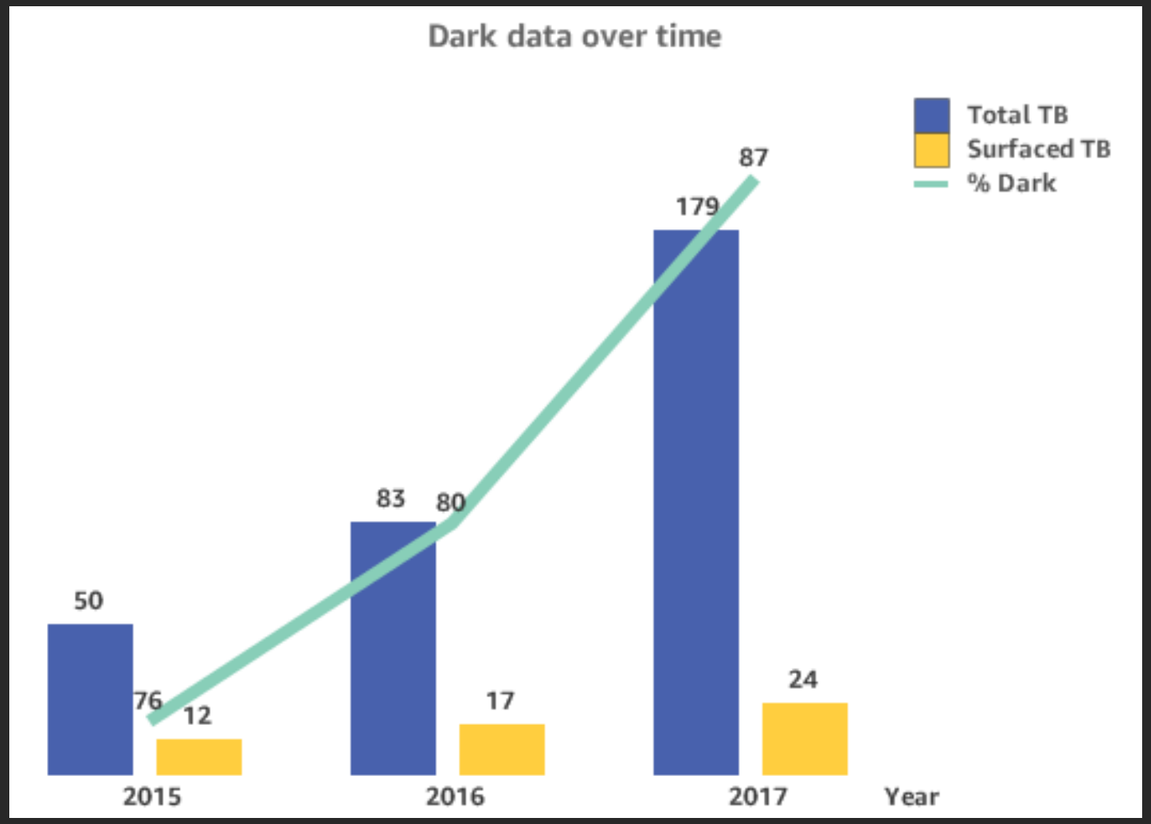
© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws

Our Amazon Redshift usage over the years

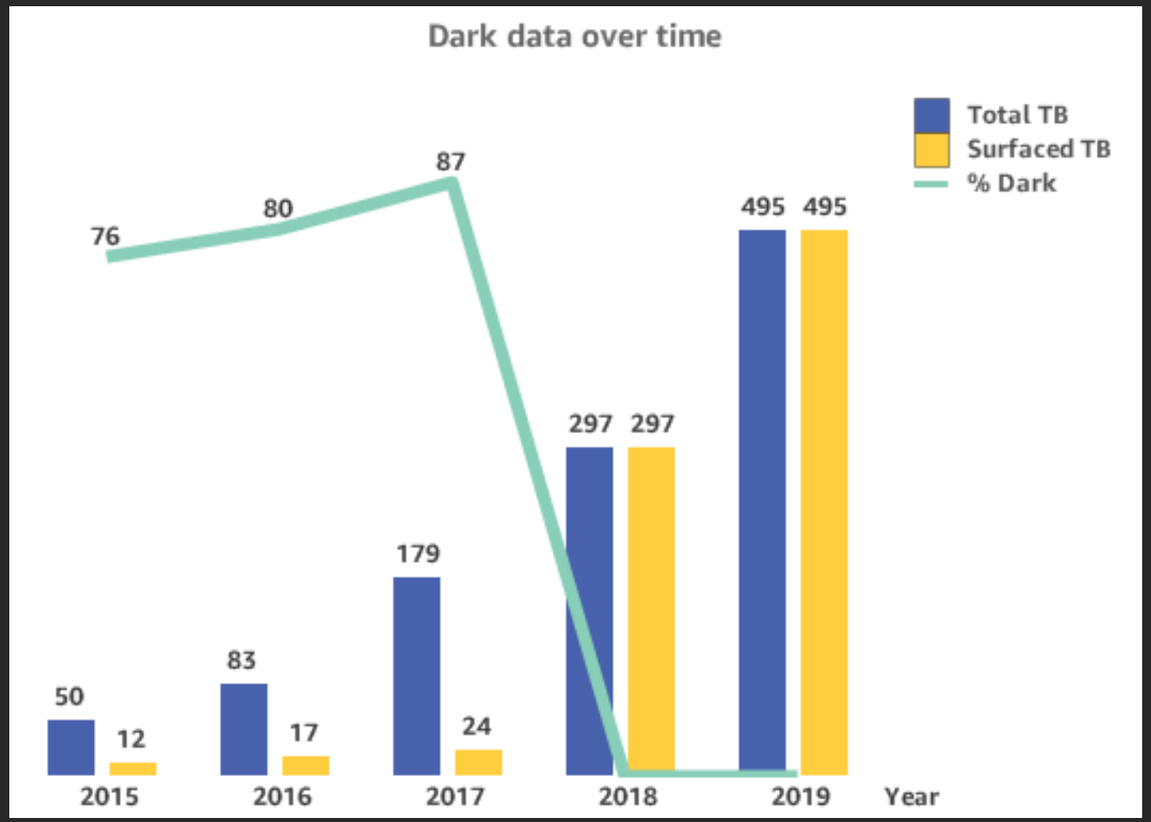
2015–2017

- Data for scores of clients + vendors
- 15–20 smaller dense storage clusters; all gone now
- Only highest-priority clients; limited data time frame



Today

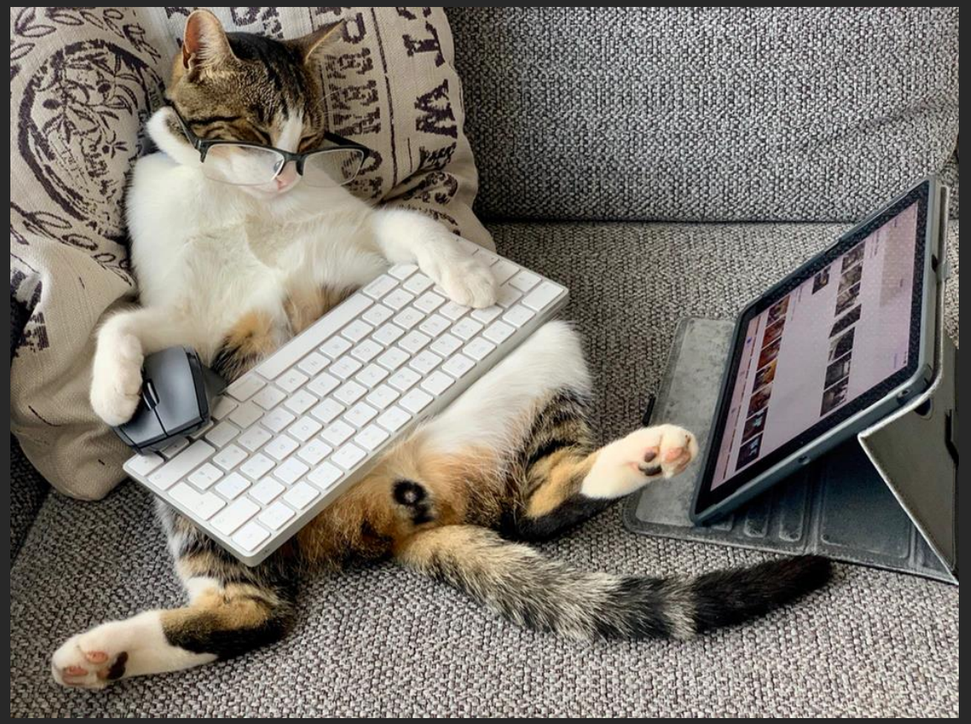
- 2017–present (the Amazon Redshift Spectrum era)
  - Data for scores of clients + vendors
  - 4–5 smaller dense compute clusters
  - All clients; much longer data time frame
- Some stats
  - Raw data size: 500 TB
  - Prepared data size: 350 TB
  - Total events: 3 trillion



Data support for our analysts and products

- Use cases
  - Campaign planning and analysis
  - Attribution reporting
  - Response modeling

- Many data sets for many clients from many vendors
  - Ad server and DSP event data; DMP data
  - Per day, thousands of tables and billions of rows



What we set out to deliver

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

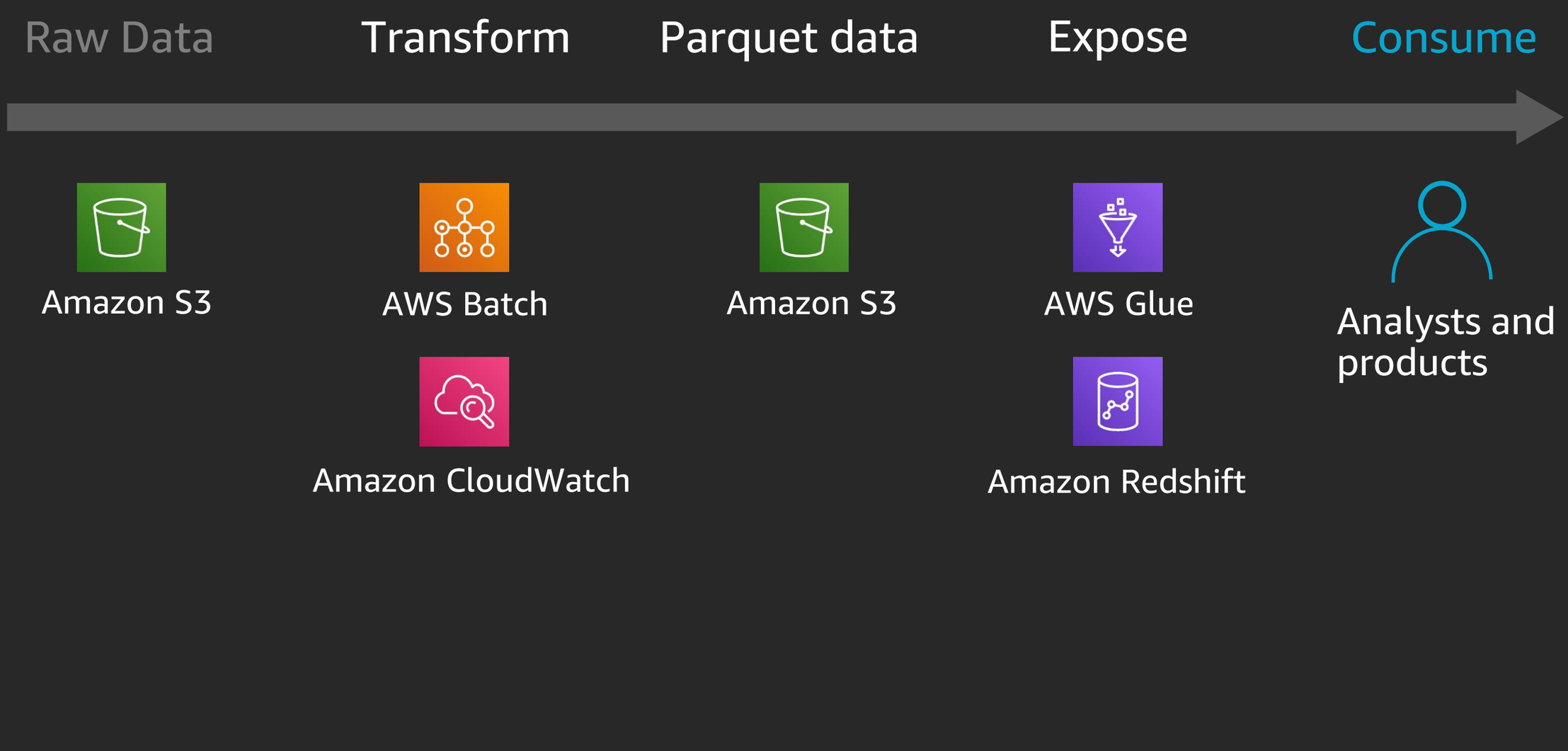
Ingredients for data-driven innovation

- Maximal coverage
  - As much data as possible
  - For current use cases and to discover and unlock new use cases
- Ease of access and rapid data availability
  - Interest wanes if you have to ask and wait
- Low costs
  - Removes barriers
  - Makes it easier to try new things

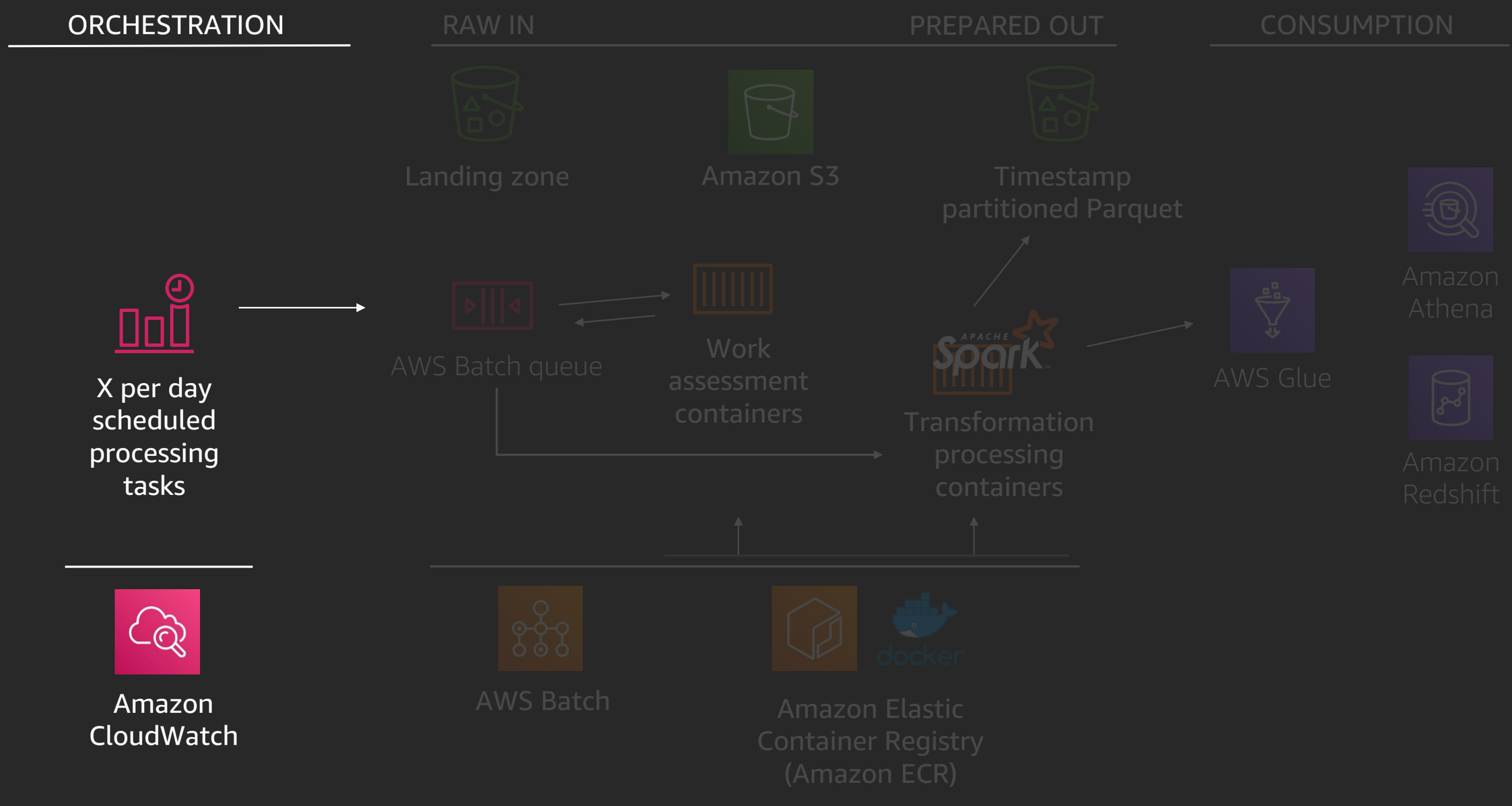
How we do it

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

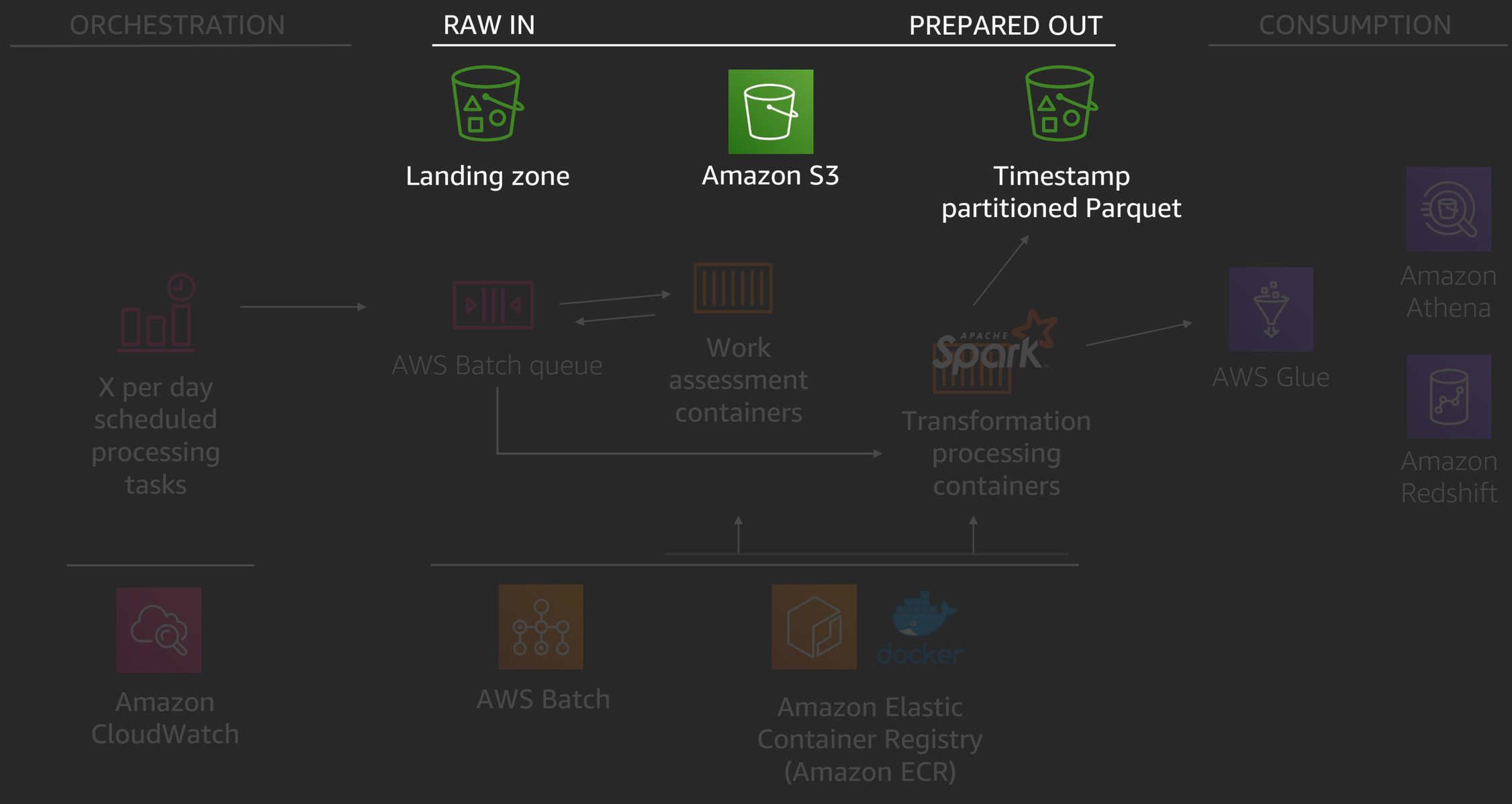
Architecture overview



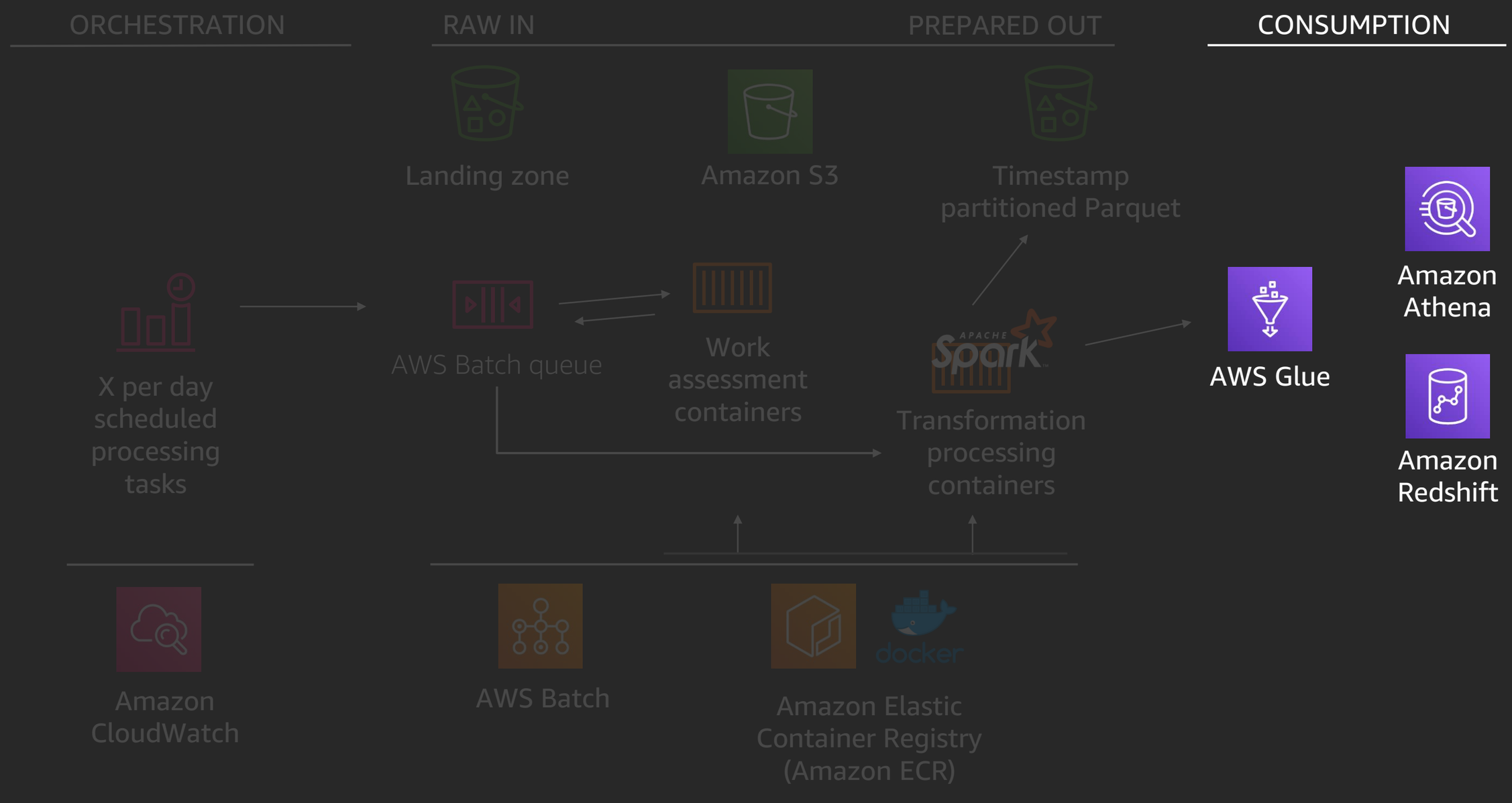
Expanded: Event file processing using AWS Batch



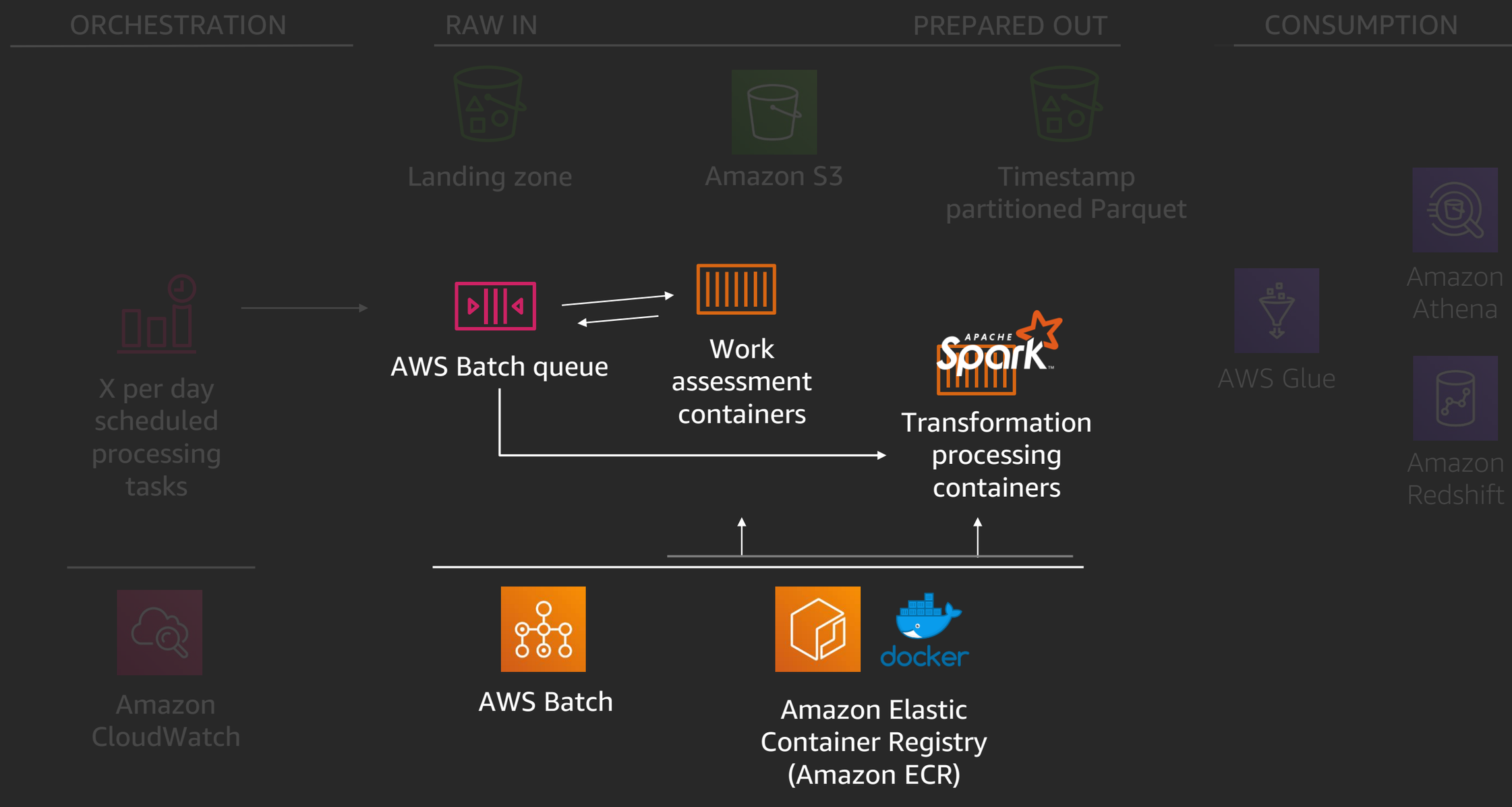
Expanded: Event file processing using AWS Batch



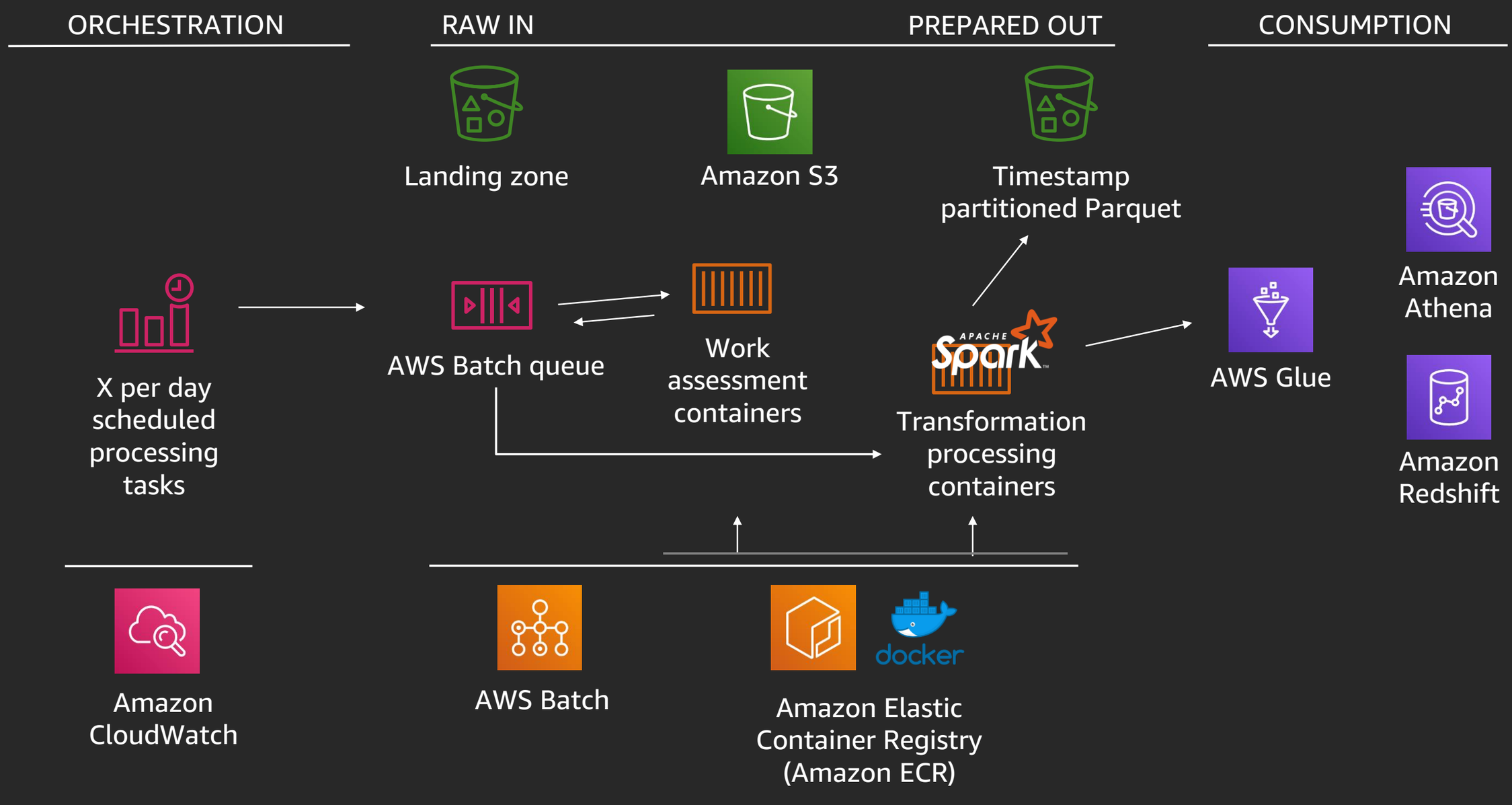
Expanded: Event file processing using AWS Batch



Expanded: Event file processing using AWS Batch



Expanded: Event file processing using AWS Batch



Tips and optimizations

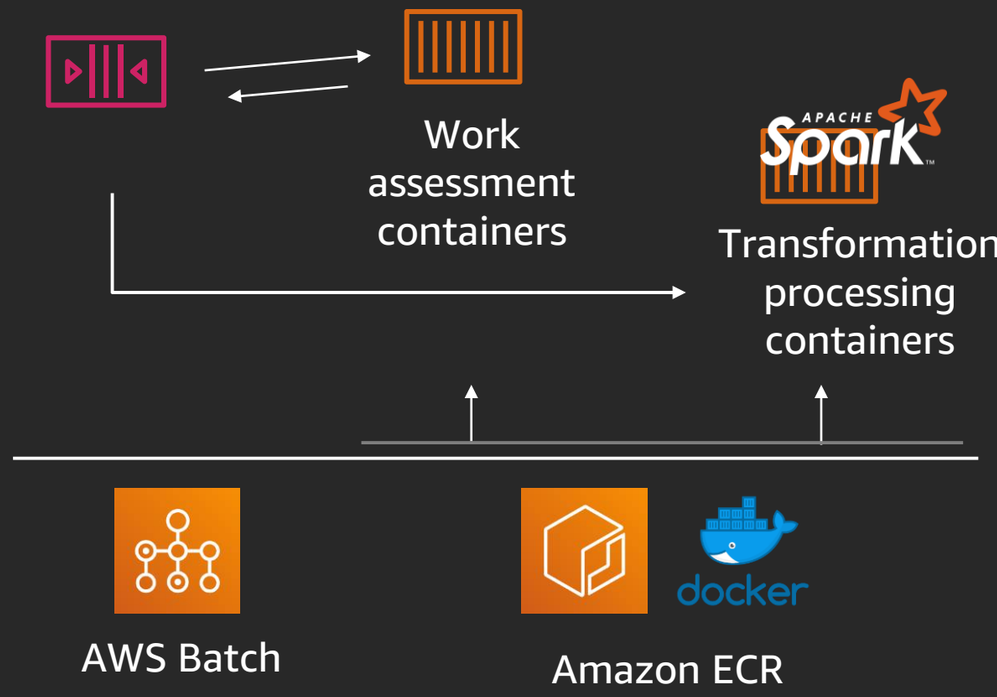
© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Leverage AWS features and fine tune

- AWS services
  - Amazon Redshift Reserved Instances
  - Amazon Redshift workload management (WLM)
  - Amazon S3 storage classes
  - Amazon Elastic Compute Cloud (Amazon EC2) Spot Instances
- Data
  - Use Parquet for most use cases
  - Partition the Amazon S3 data backing your external tables
- Architecture
  - Limiting what is surfaced
  - Self-healing data pipelines

Self-healing pipelines

- Idempotent transformation jobs
  - Avoid data duplication; reruns, no problem
- Plan for schema evolution
  - New fields, no problem
- Plan for data needing to be replaced
  - Replacement files, no problem
- Plan for job failures
  - Automatic retries, no problem

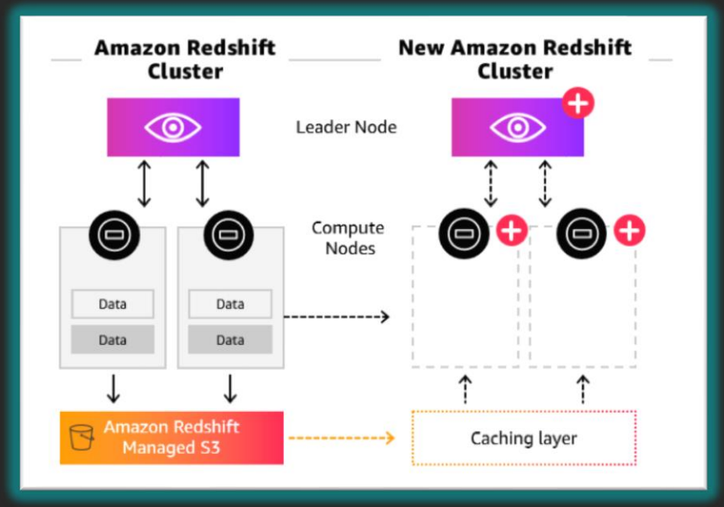


The future

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

Areas of interest

- Amazon Redshift concurrency scaling
  - One bigger cluster beats two smaller clusters in terms of query performance
  - But not in terms of local table space management and cost isolation



- Transformation engine
  - Replace bespoke AWS Batch solution?

AWS Glue

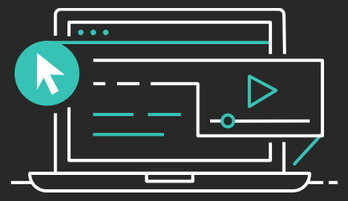
- Expanding our use of other analytic tools
  - Amazon EMR, Amazon Athena

Amazon EMR

Athena

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/

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws training and certification

Thank you!

Eric Kamm

eric.kamm@annalect.com

AWS re:Invent

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.

aws



Please complete the session survey in the mobile app.

© 2019, Amazon Web Services, Inc. or its affiliates. All rights reserved.