AWS re:Invent

ANT314-R

Extend analytics beyond your data warehouse

Martin Grund
Principal Engineer
Amazon Web Services

Kurt Larson
Technical Director
WB Games

Richard Snelling
Staff Redshift DBA
WB Games

AWS re:Invent

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

aws

What is analytics?

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



Amazon Redshift



Amazon EMR



Amazon QuickSight

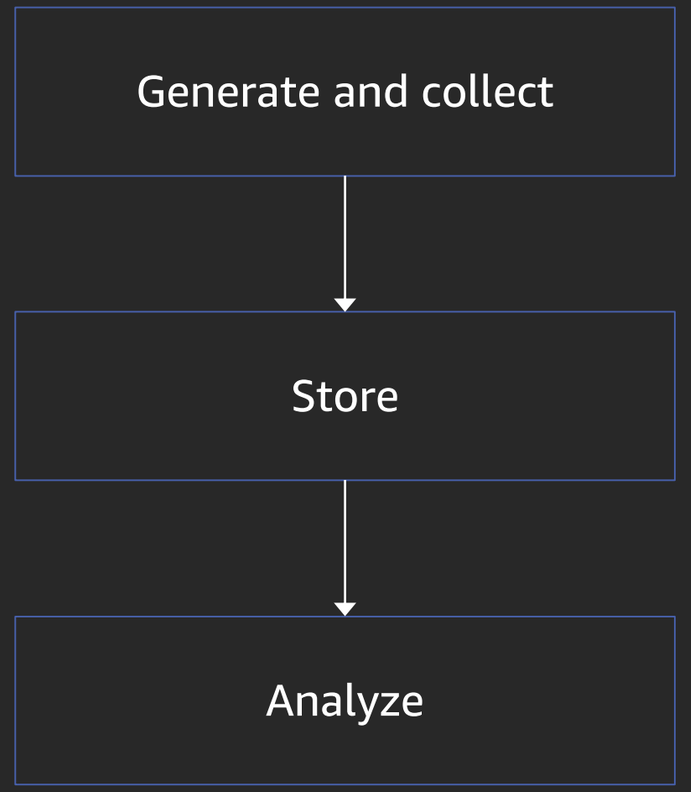


Amazon Kinesis





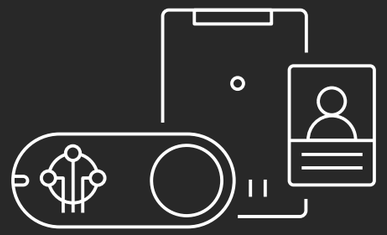


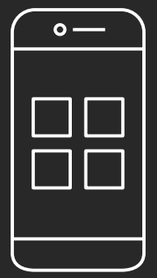


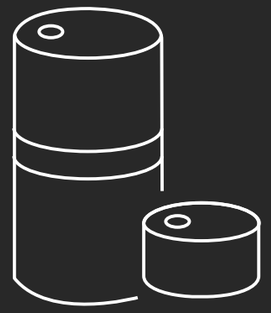
It has never been easier to collect data

New Report

First Nam	Last Nam	Action ID	Application	Click Date	Clicks	First Paramet	Second Parameter	Third Parameter
Sean	Austin	/AnswerView	End-User p	02/11/2010 12:25 PM	Gs6V	2688	Creating a report for specific user transactions	No Value
Sean	Austin	/AnswerView	End-User p	02/11/2010 12:25 PM	Gs6V	2688	Creating a report for specific user transactions	No Value
Sean	Austin	/AnswerList/Sear	End-User p	02/11/2010 12:25 PM	Gs6V	clicktrack	No Value	No Value
Sean	Austin	/AnswerList/Sear	End-User p	02/11/2010 12:25 PM	Gs6V	clicktrack	No Value	No Value
Sean	Austin	/social/ssoError	End-User p	02/11/2010 12:22 PM	Gs6V	No Value	No Value	No Value
Sean	Austin	/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/month
- S3 infrequent access: $12.8 per TB/month



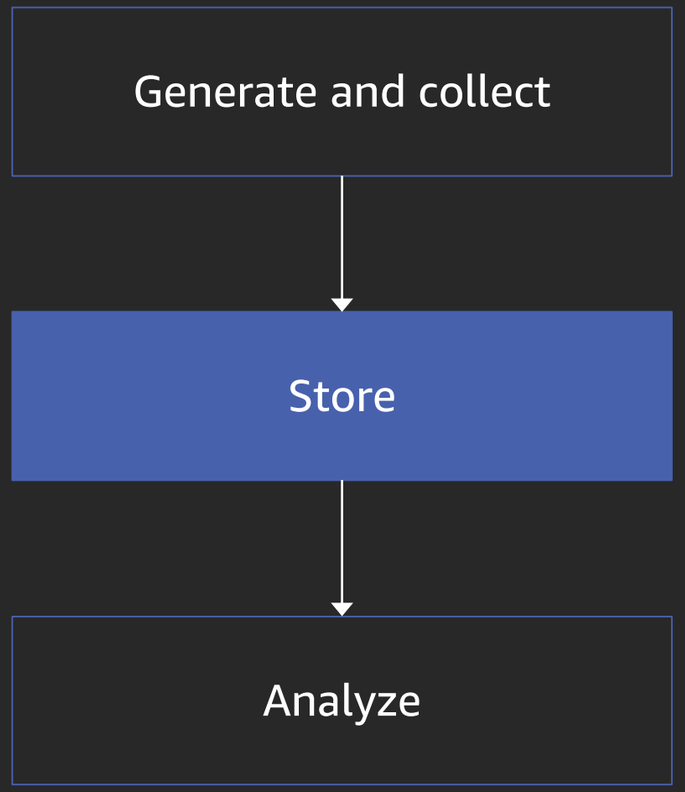
Amazon Simple Storage Service

Amazon S3 Glacier

- S3 Glacier: $4.1 per TB per month
- S3 Glacier Deep Archive: $1 per TB per month



Amazon S3 Glacier



The tyranny of “OR” in Analytics

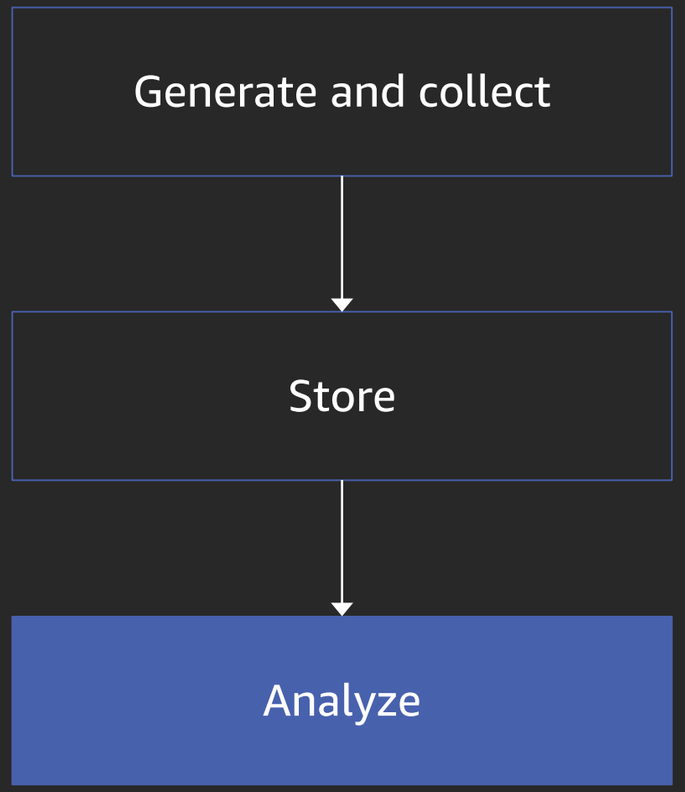
Parallel data warehouse

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

VS

Big data framework

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



Escaping the tyranny of “OR” with Amazon Redshift

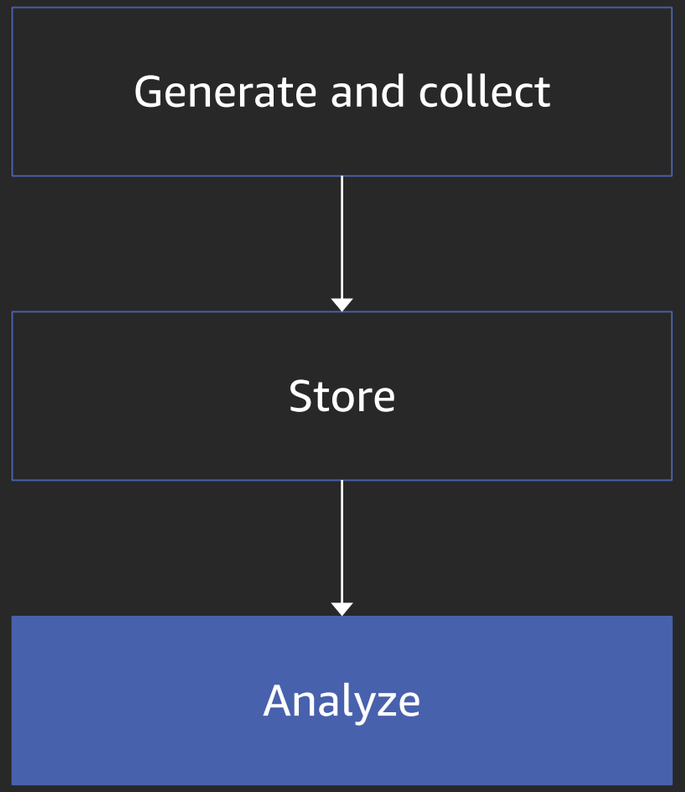
Top-of-the-line performance for any

Data scale

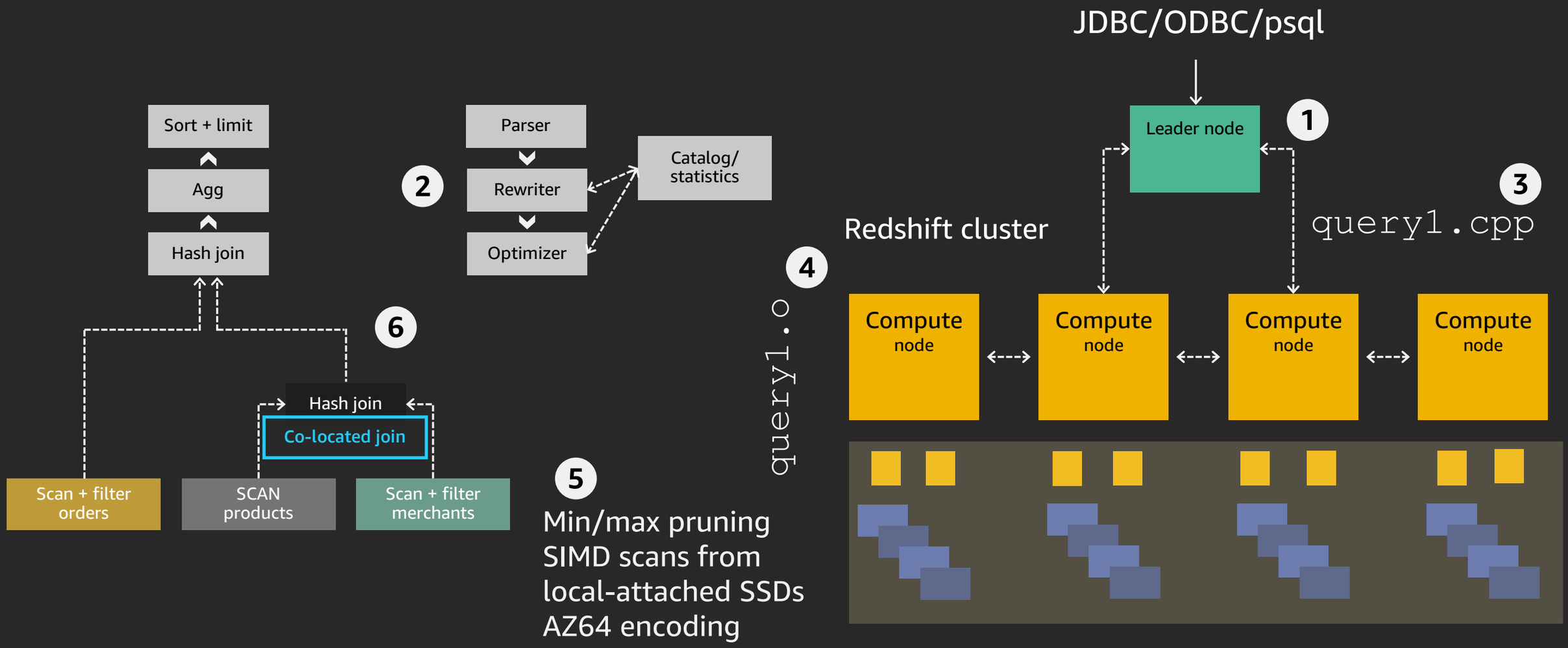
- a/ Local attached data
- b/ Very large data in Amazon S3

Concurrency

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

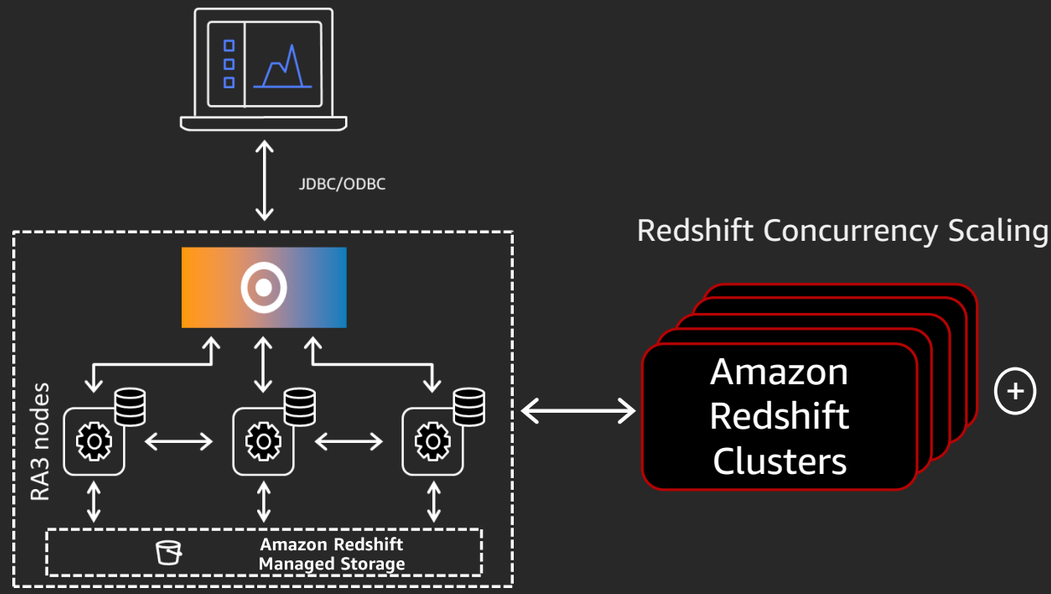


Executing this query in Amazon Redshift



Top-of-the-line data warehousing performance

Compute elasticity and scalability with concurrency scaling



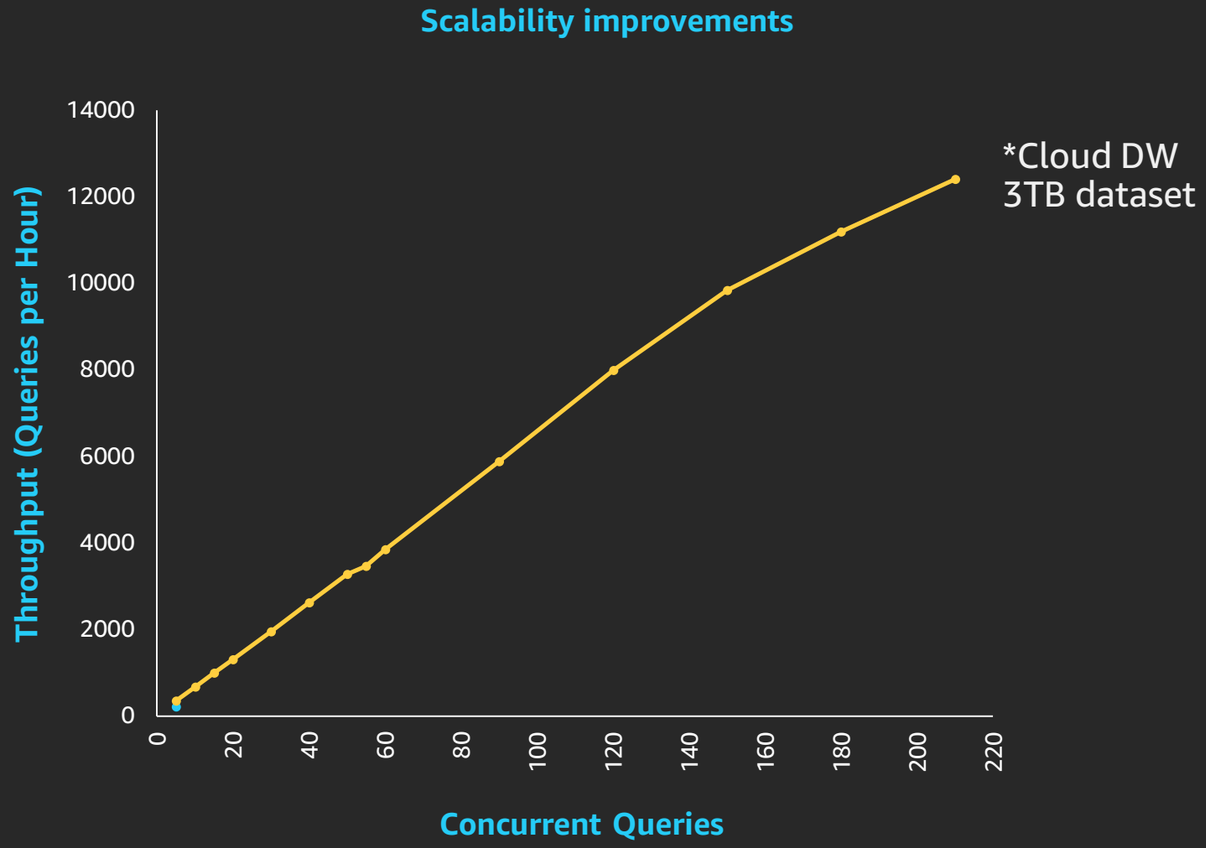
Scale-out to multiple Amazon Redshift clusters from a single endpoint in seconds

Support virtually unlimited concurrent users and queries while maintaining SLAs

Per-second billing for additional clusters used

Free 1hr usage per day (free for 97% of clusters)

>35x improvement in throughput in 2019



* Cloud DW benchmark is based on TPC-DS (v2.10) with no query modifications are done

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

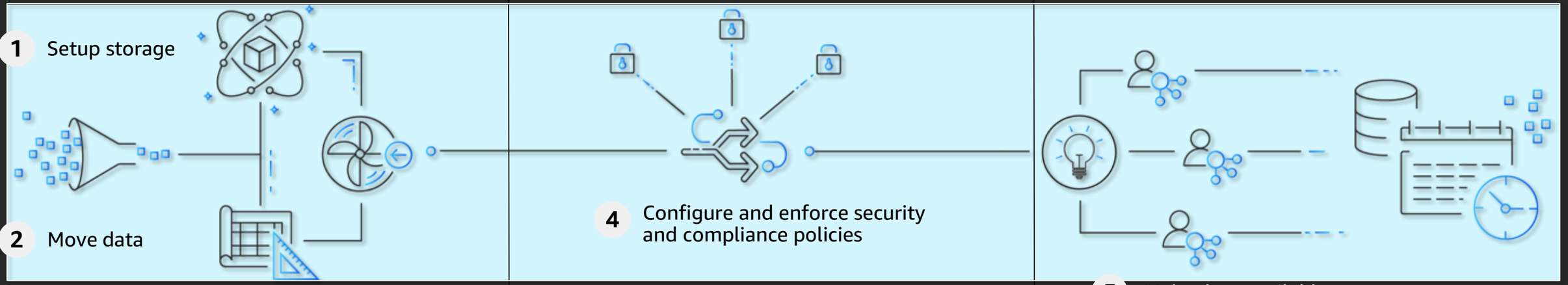


Data security officer

Analytics and ML



Data analyst



3 Cleanse, prep, and catalog data

5 Make data available for analytics

AWS Lake Formation

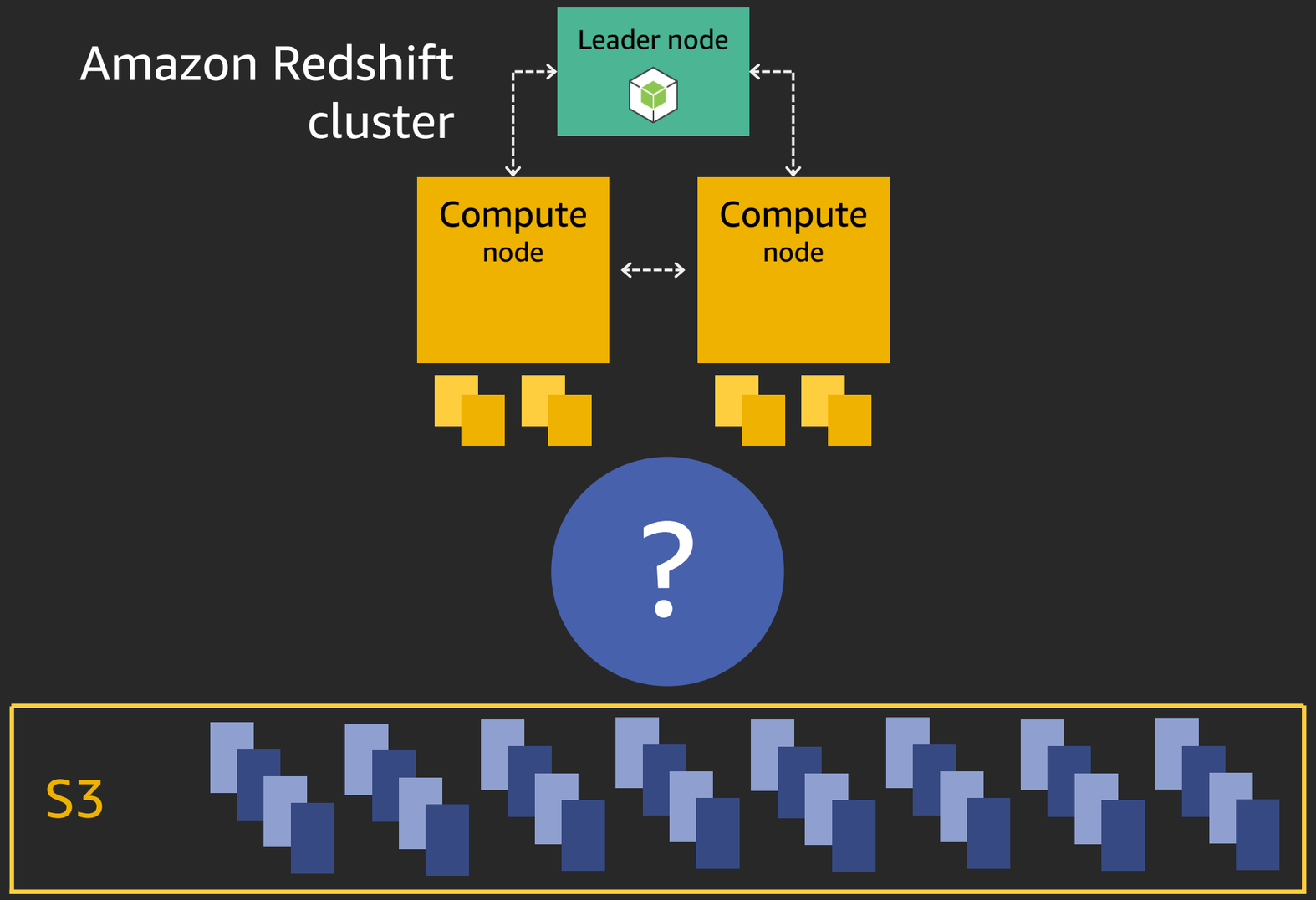
AWS Lake Formation lets you

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

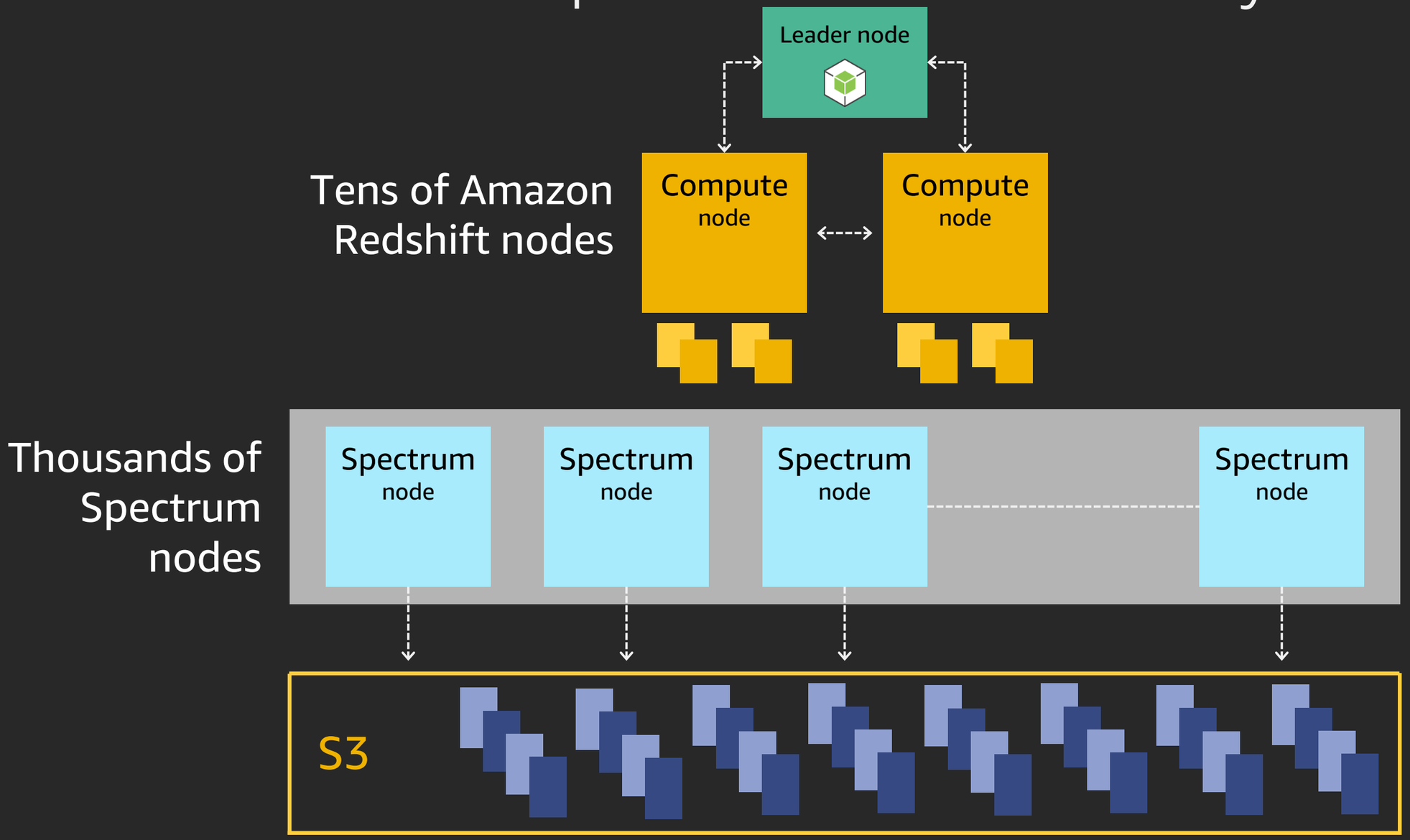
1. High bandwidth
   - Parallelism (many small straws)

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

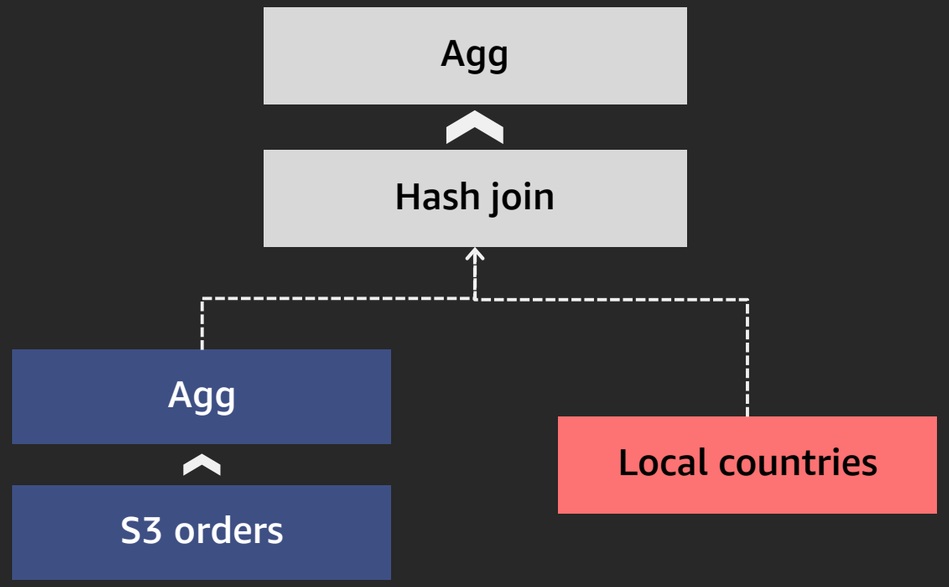
s3://Orders/ [YYYY] / [MM] / [DD] / [Country] /

20	365	195
1	1	195
1	1	12

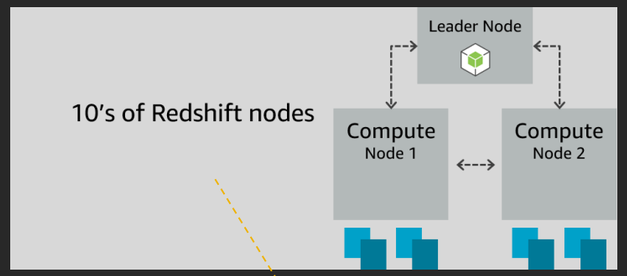
Partitions to process

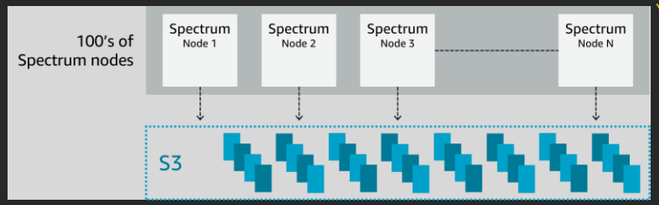
1,423,500
195
12

```
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

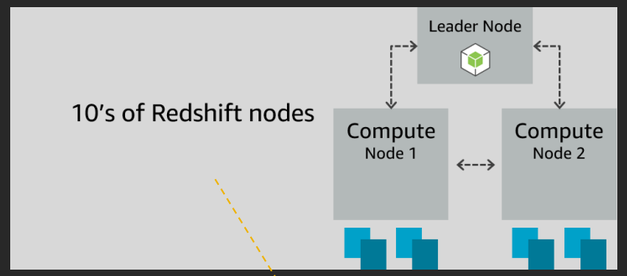
Scan products

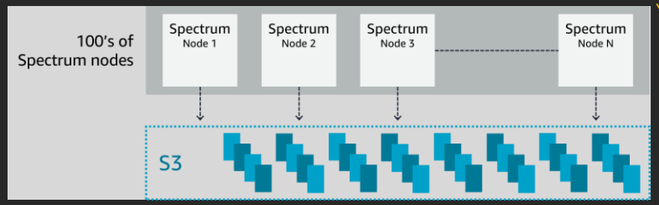
Millions

Scan + filter merchants

Hundreds of thousands

Executing this query in Amazon Redshift Spectrum





Sort + limit

Agg

Hash join

Telemetry

Up to orders of magnitude reduction

Agg

Scan orders

Tens of billions

Hash join

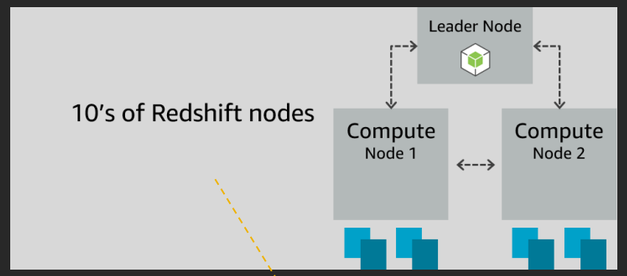
Scan products

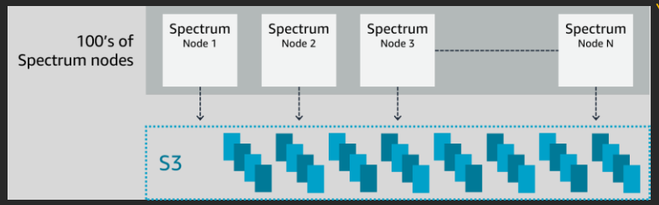
Millions

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

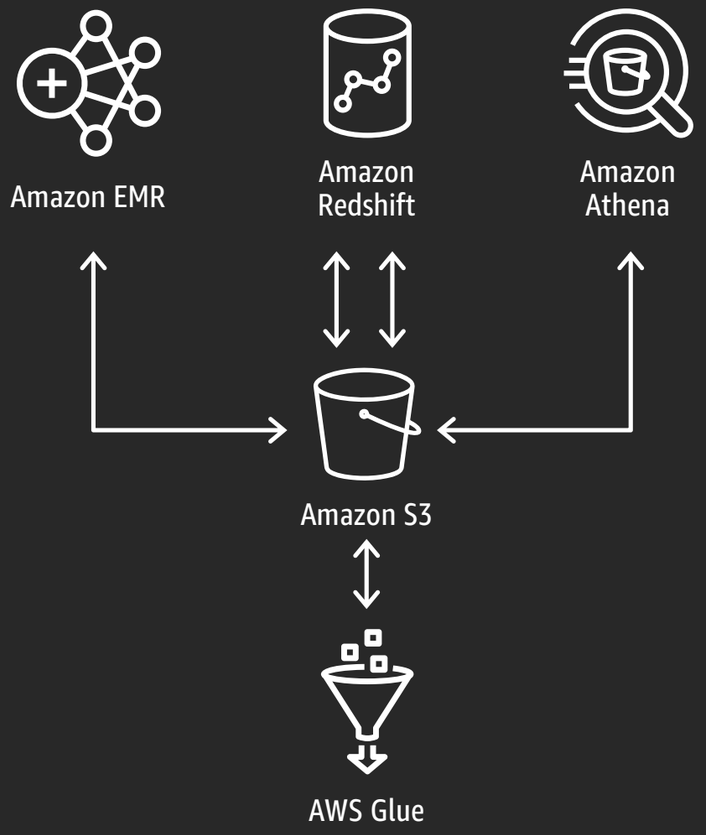
Millions

Scan + filter merchants

Hundreds of thousands

Significant acceleration of repeating queries; incremental

Data lake export: share data in Parquet format



Parquet is efficient open columnar storage format for analytics

Analyze your data with Amazon Redshift Spectrum and other AWS services such as Amazon Athena and Amazon EMR

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

Using Amazon Redshift for elastic ELT

“Smart” transformation requiring complex SQL transformations

Phase 1

- Ingest using simple file-based COPY-command, *or*
- Smart prefiltering, aggregation, and cleansing with Spectrum and Federated Query

Phase 2

- Execute of business logic and other transformations using, e.g., SQL UDFs, stored procedures,

Phase 3

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

Amazon Redshift: Key 2019 innovations

Performance & scalability

NEW! RA3 with Amazon Redshift managed storage

PREVIEW! AQUA

NEW! AZ64

PREVIEW! Materialized views

NEW! Spatial data support

NEW! Concurrency Scaling

Data lake & AWS integration

PREVIEW! Federated query across Amazon Redshift & Amazon RDS/Amazon Aurora

NEW! Data lake export in Parquet

NEW! Spectrum Request Accelerator

NEW! AWS Lake Formation integration

Fully managed

NEW! Auto-Vacuum, Auto-Analyze & Auto Table Sort

NEW! New management console

NEW! Auto WLM: Query priorities

NEW! Elastic resize scheduler

NEW! Stored procedures

NEW! Distribution and sort key advisor

NEW! Cross-instance restore

NEW! Faster cross regional copy and change

NEW! Auto Data Distribution

NEW! Deferred Maintenance

NEW! Elastic resize

Warner Bros. Games

Kurt Larson
Technical Director
WB Games

Richard Snelling
Staff Redshift DBA
WB Games

AWS re:Invent

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

aws



We believe in the power of story.

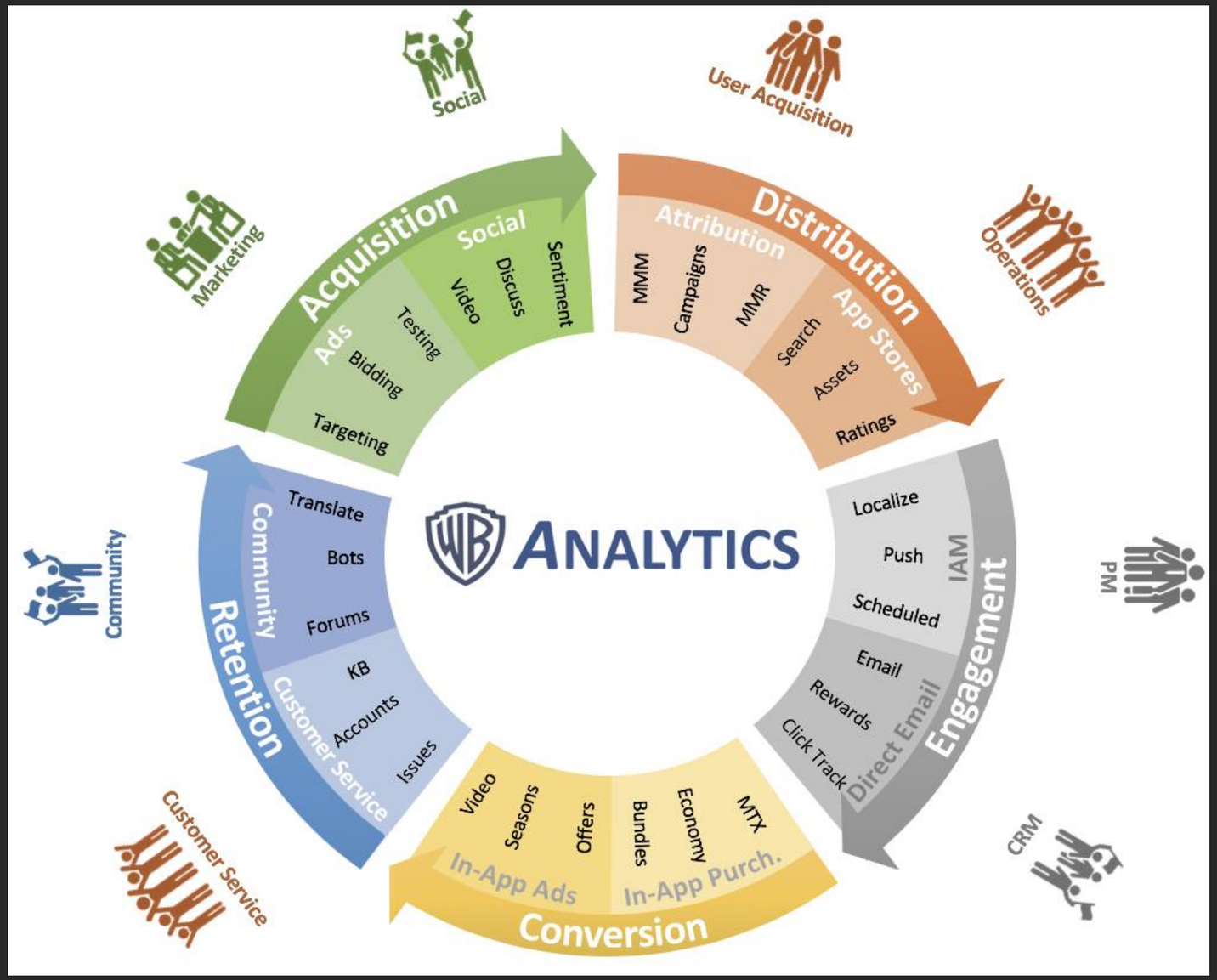

HITMAN 2
GAME OF THRONES CONQUEST
BATMAN
LEGO
DC SUPER-VILLAINS
MORTAL KOMBAT 11

What we do…

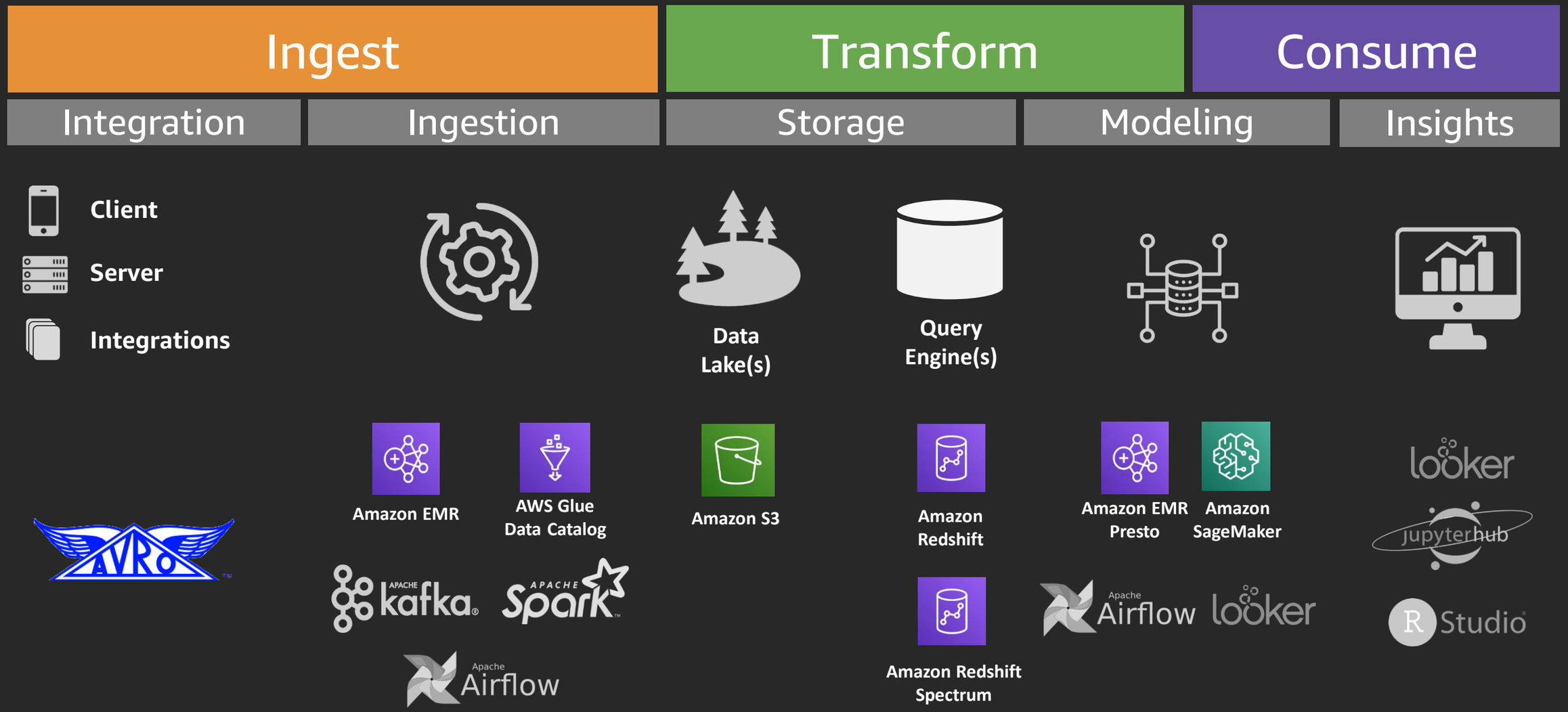
Many teams work to publish a game…

Each team brings specialized tools…

We combine these tools with client data to create a consistent, actionable view of each game.



Picking the right tools



Our Amazon Redshift Fleet

Environment

- ~20 Clusters – 6 non-production, 14 production
- 12 single tenant, 8 multi tenant
- Storage:
  - *Amazon Redshift - 200 TB*
  - *"Lake House" data lakes - 1 PB*

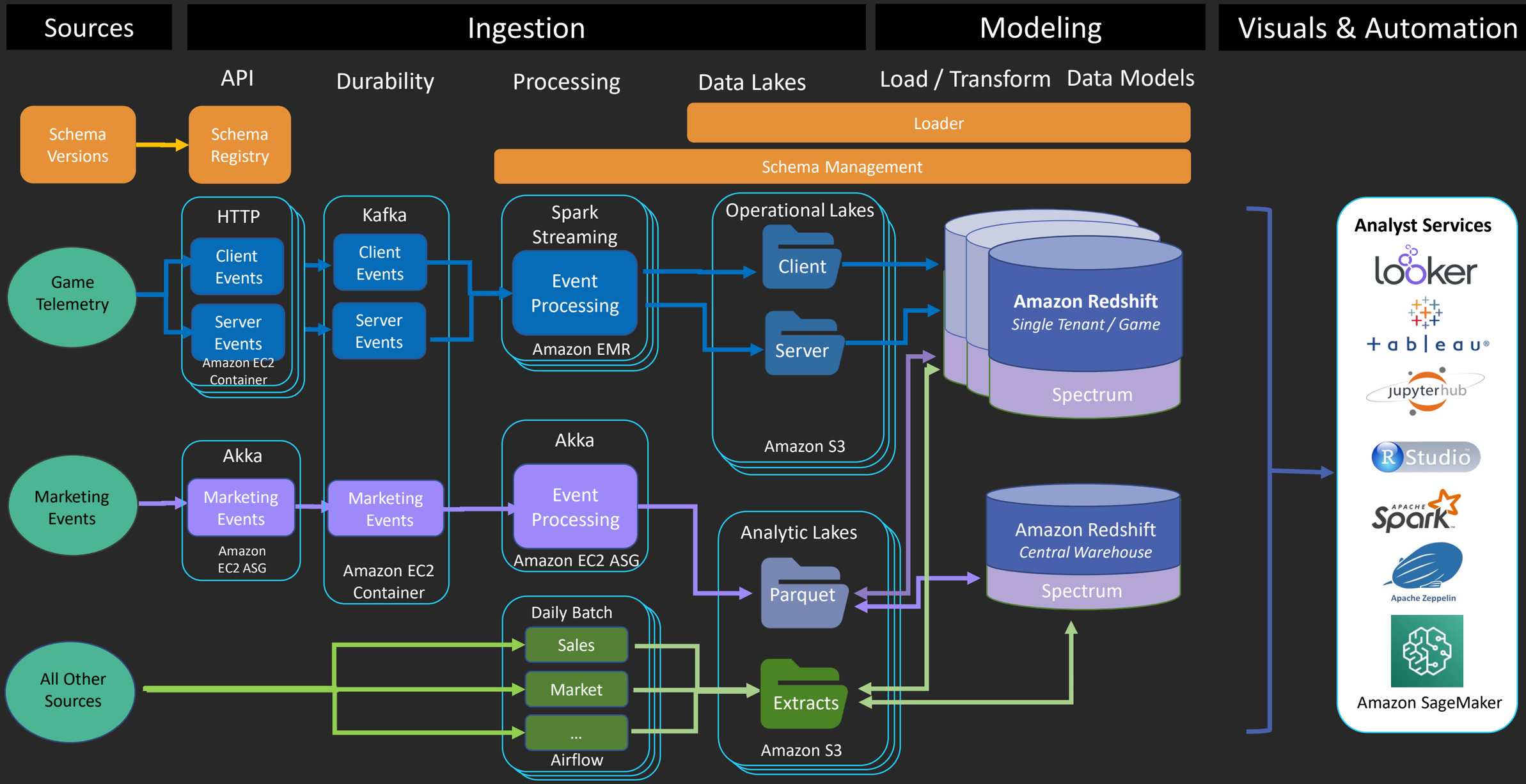
Targets

Event streams: 100k/sec peak, 20k/sec sustained

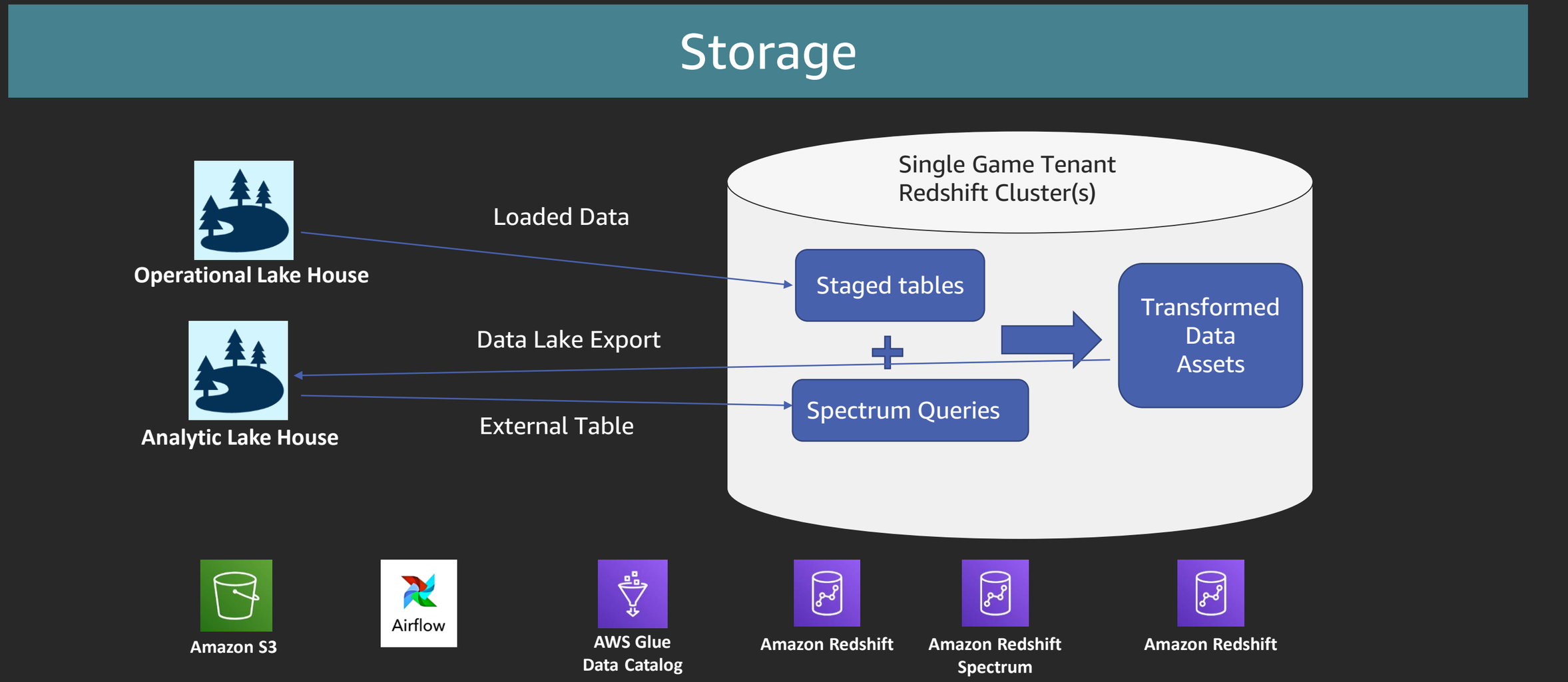
3 - 10 minute micro batches per game

- 60 - 340 tables/game
- Spectrum (scanned/mo) - ↑ 1PB

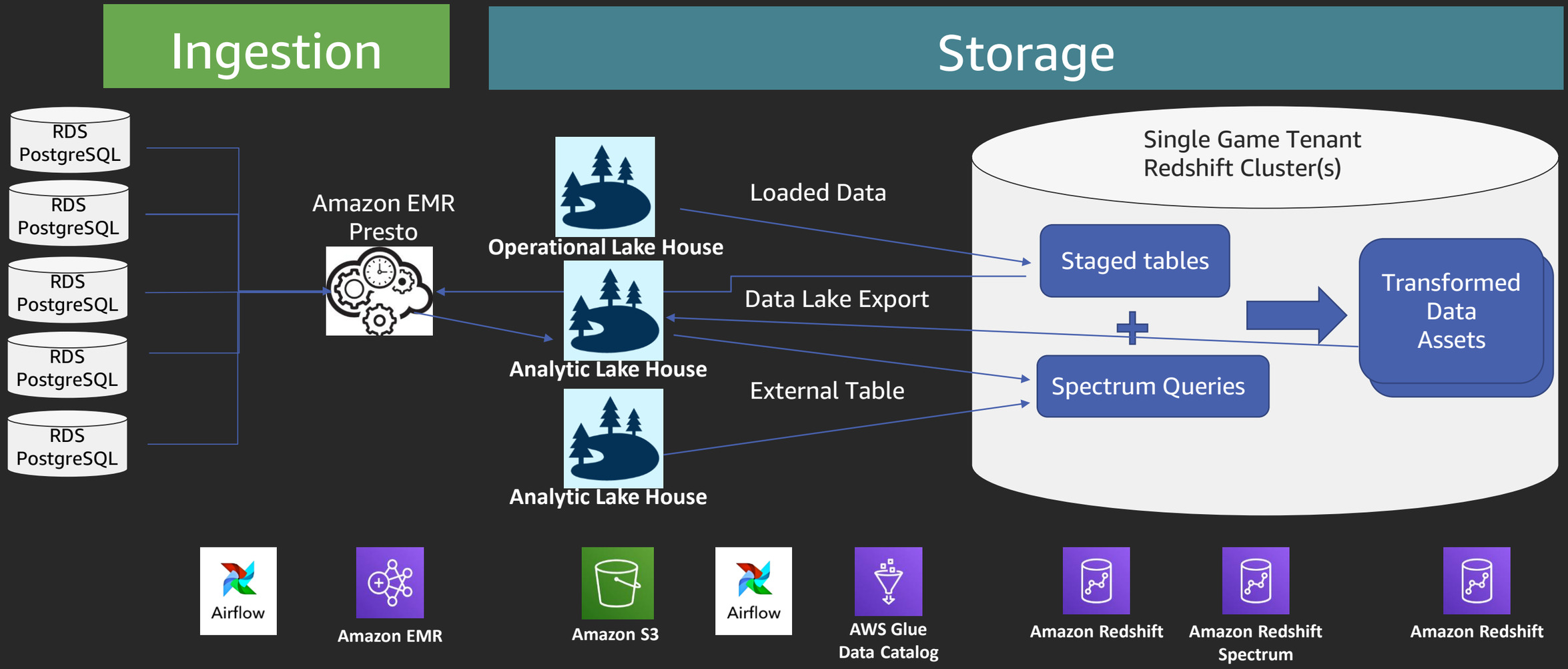
Big Data Stack



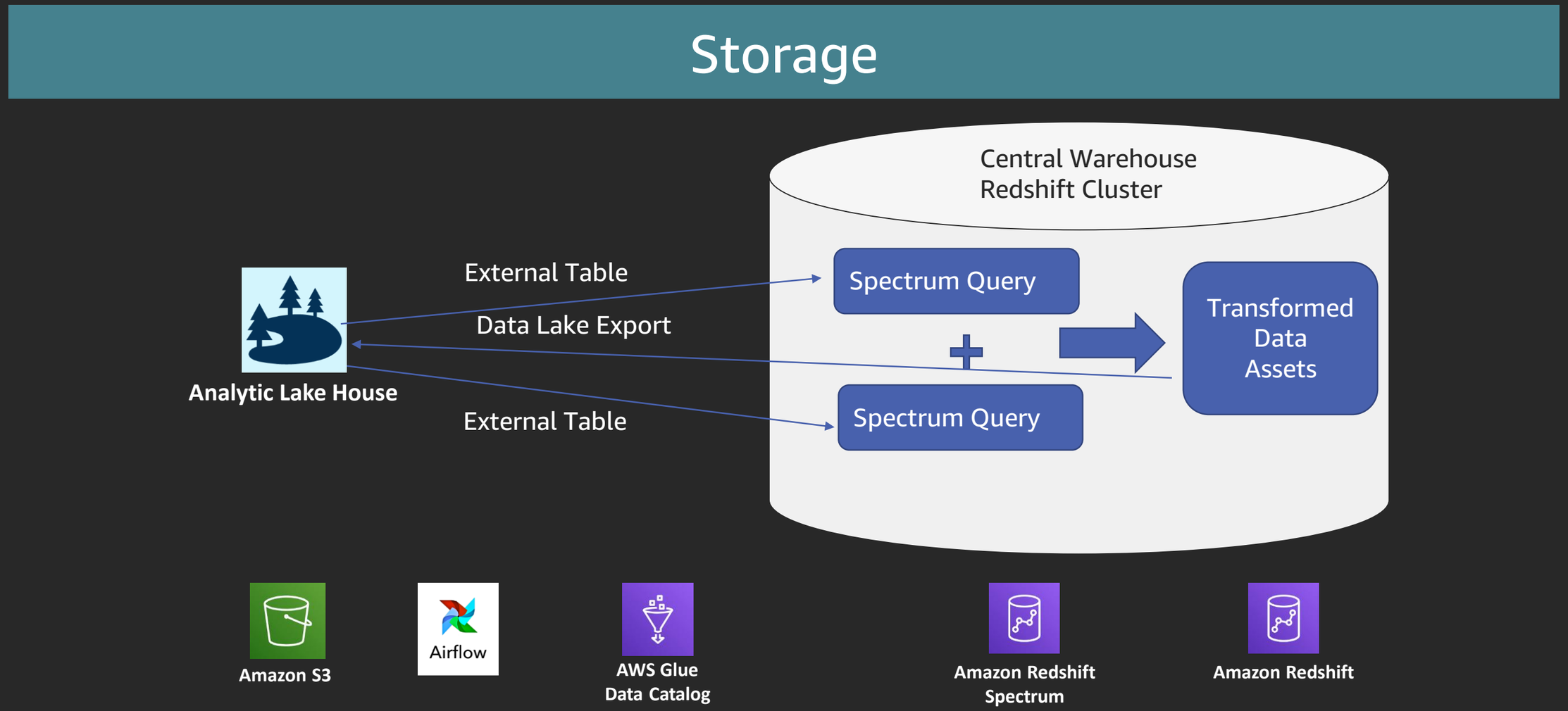
ELT in and out of a Single Tenant Game DB



ELT-ETL-ETL for data federation



ELT–ELT from Analytic Data Lake to Central Warehouse and back to Data Lake



What enables extending the data warehouse

AWS Glue Data Catalog

- Shared external catalog
- Support all AWS Big Data Stack query engines
- Managed by the service
- Hive compatible

AWS Lake Formation

- Common permissions model
- Enforced by multiple query engines
- Eases management of data lakes

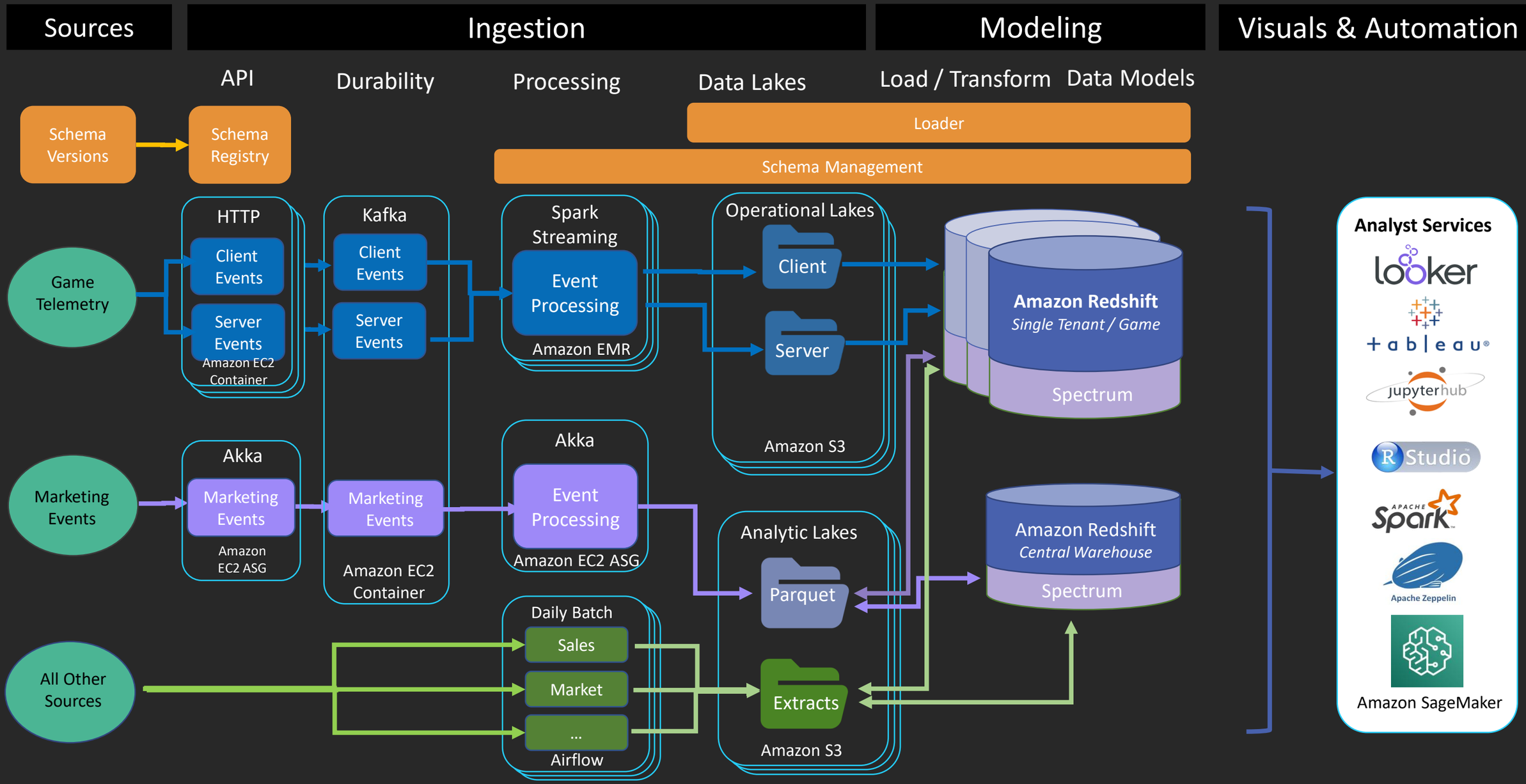
Amazon Redshift Spectrum

- Query in place
- No loads
- Share data with other tools
- Data mobility between clusters

Data Lake Export

- Size cluster for compute
- Limit cluster growth
- Data mobility between clusters
- Open source columnar format

Big Data Stack



Thank you!

Martin Grund, AWS
magrund@amazon.de

Kurt Larson, WB Games
klarson@wbgames.com

Richard Snelling, WB Games
rsnelling@wbgames.com

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



Please complete the session survey in the mobile app.

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