ADC-C3

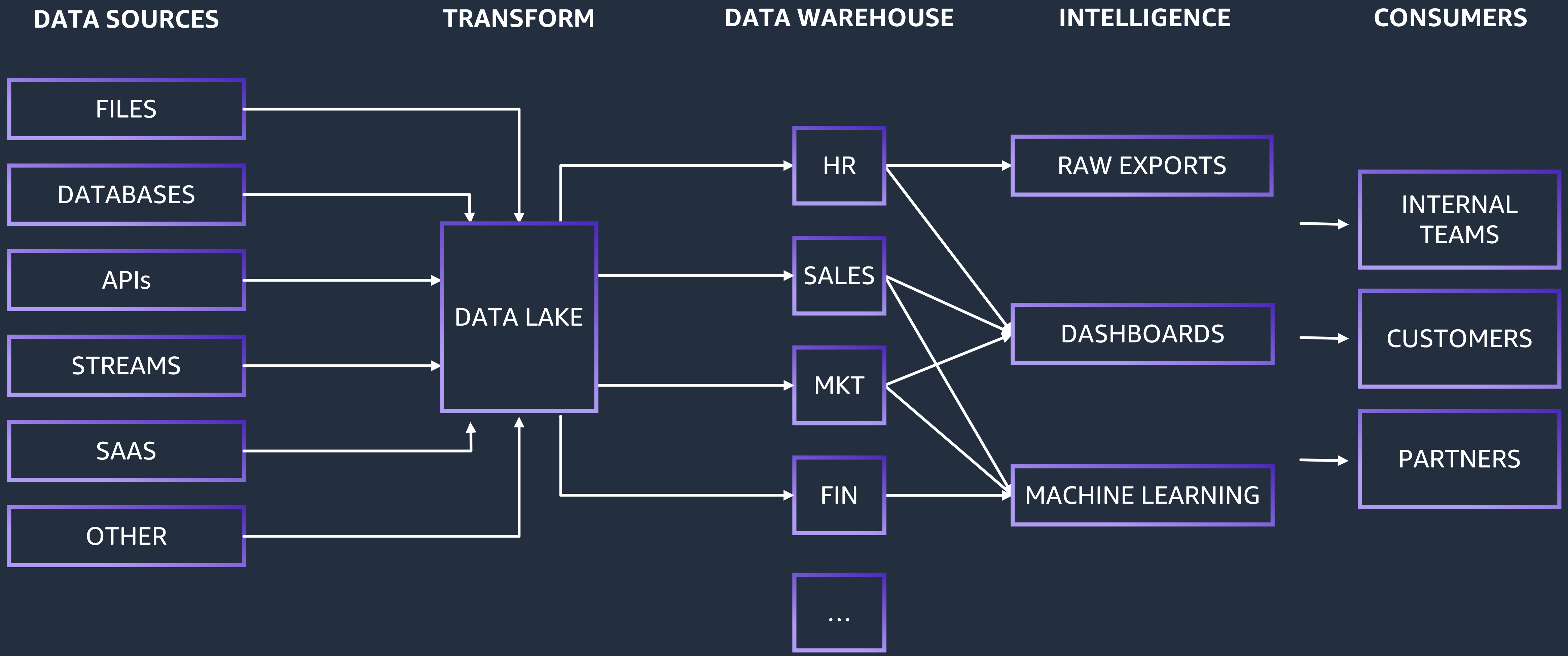
The Zero-ETL future of analytics

Solving data integration challenges

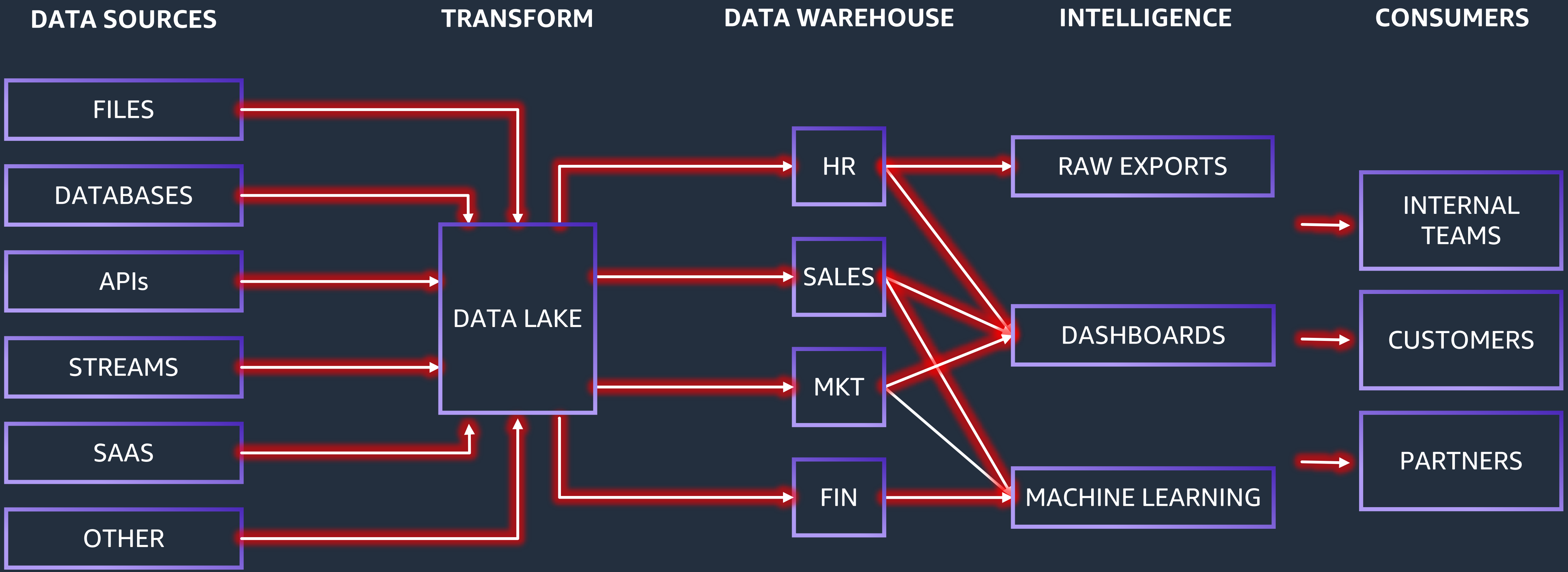
Sandipan Bhaumik (he/him)

Sr. Analytics Solutions Architect
/in/sandipanbhaumik

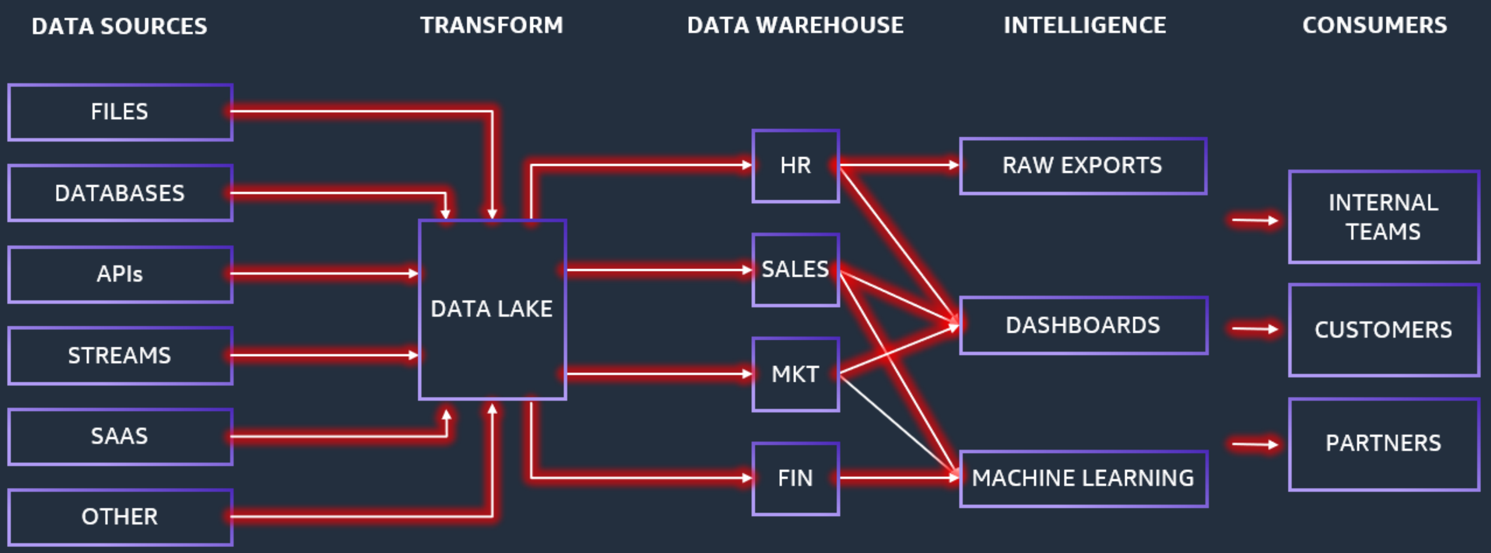
Data lifecycle in an organization



You need to build data pipelines, aka ETL



Building ETL is complex



Manage connections, secure them

Handle failure scenarios

Manage ETL infrastructure

Additional monitoring processes

Write complex code

Spend time on overheads

So, to make it easier for you we are building...

Zero ETL
future enabled
by service integrations

AWS Glue for
value-add data
transformations and more

Services for
**connecting to 100s
of data sources**

AWS Glue: Key capabilities

SERVERLESS DATA INTEGRATION SERVICE

Scalable data integration engine

Built-in data transforms

Execution engine

Monitor

Centralized and unified data governance

AWS Glue Data Catalog

AWS Glue Data Quality

AWS Glue crawlers

AWS Lake Formation

Connect and ingest data

AWS Glue connectors

AWS Glue connector marketplace

Various interfaces

User productivity and data ops

Persona specific tools

Productivity tools

Data ops tools

Services for connecting to 100s of data sources

Connect to **50+** SaaS applications with **Amazon AppFlow**

Stream data in real time from **30+** sources with **Amazon Kinesis Data Firehose**

Query **25+** data sources in place with **Amazon Athena**

Build models on **Amazon SageMaker** using data from **40+** sources

30+ sources to build interactive dashboards in **Amazon QuickSight**

Access third-party data from **300+** data providers with **AWS Data Exchange**

Connect to 100s of data sources

Adobe Analytics · ahana · AMPLITUDE · ATLASSIAN Confluence · box · Cloudant · CLOUDERA · cloudera IMPALA · Collibra · Coralogix

databricks · DATADOG · DELTA LAKE · dremio · Dropbox · dynatrace · Exasol · freshdesk · honeycomb.io · HubSpot

Informatica · LogicMonitor · MariaDB · MarkLogic · MongoDB. · new relic. · OKERA · privacera · redis · sage Intacct

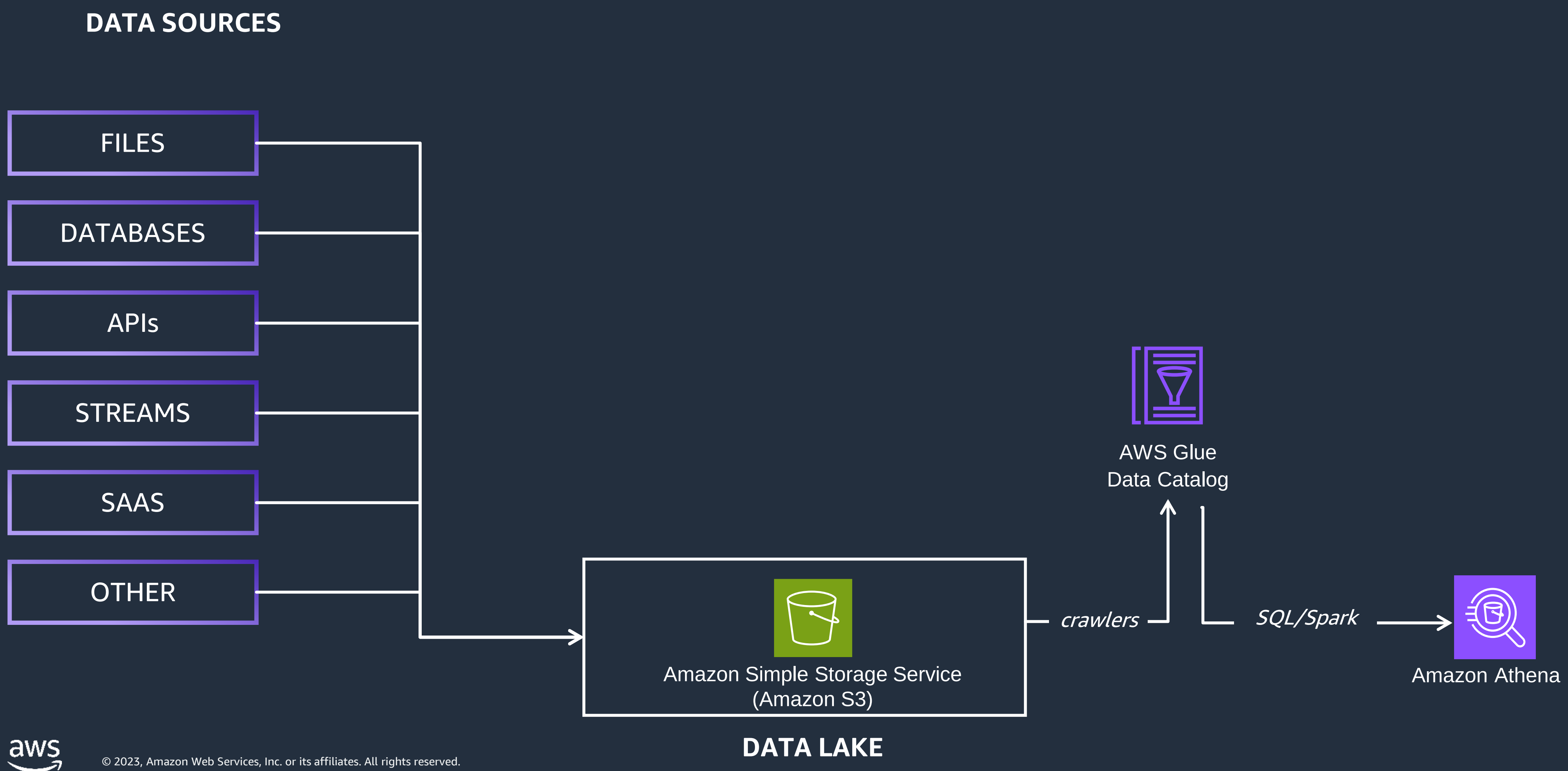
salesforce · SAP · TWILIO SendGrid · servicenow · singular · slack · snowflake · splunk> · Starburst · stripe

sumo logic · tableau · teradata. · TREND MICRO · upsolver · Veeva · VERTICA · workday. · zaloni · zendesk

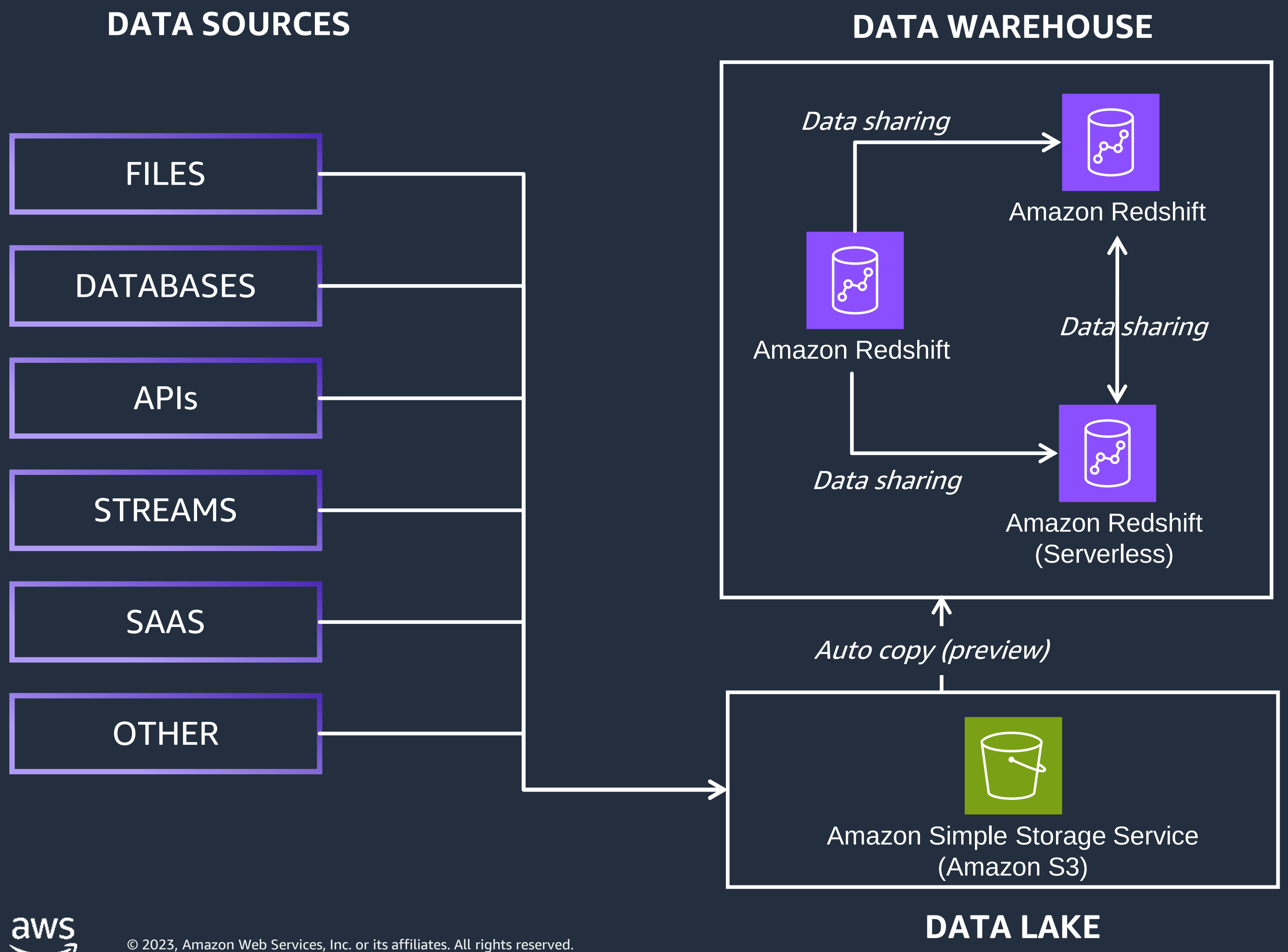
INGESTION & INTEGRATION | ANALYTICS | AI & MACHINE LEARNING | BUSINESS INTELLIGENCE | GOVERNANCE

aws

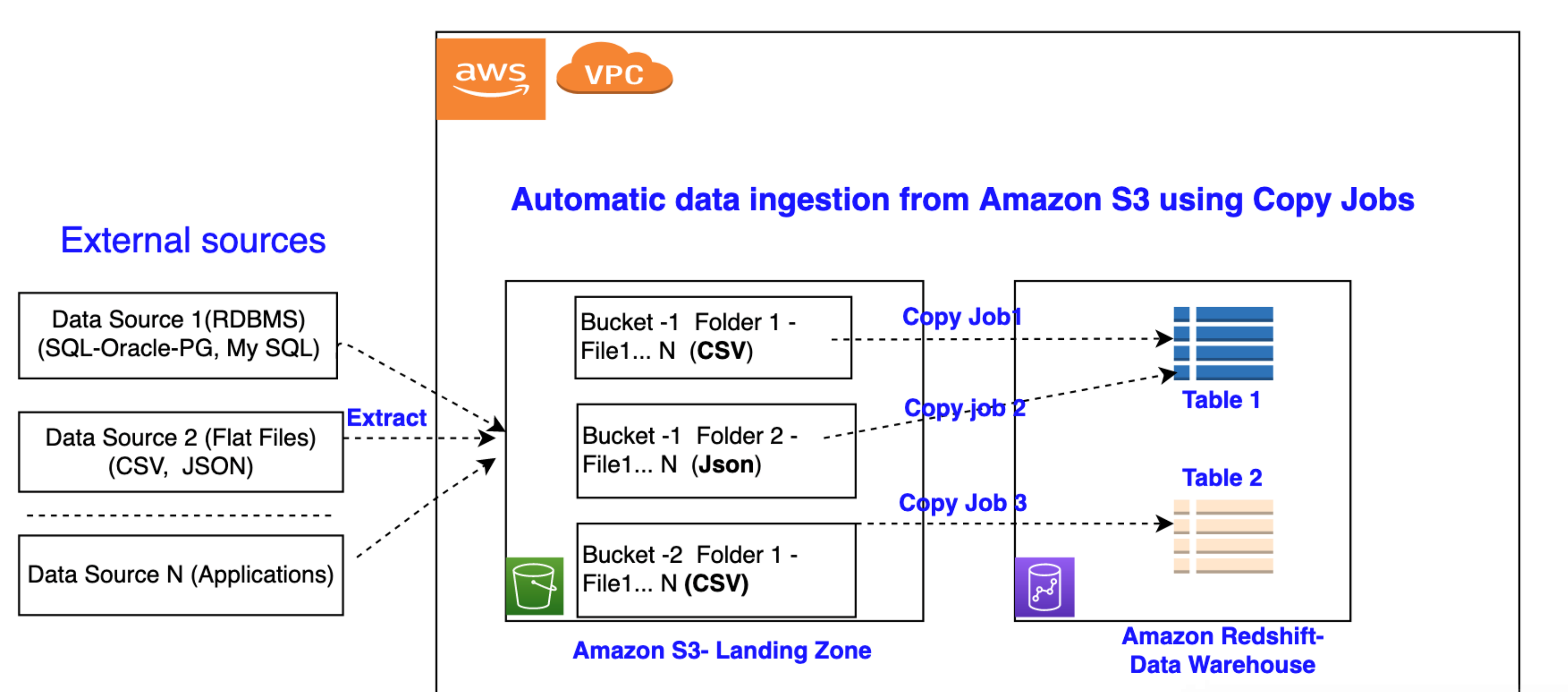
Zero ETL future enabled by service integrations



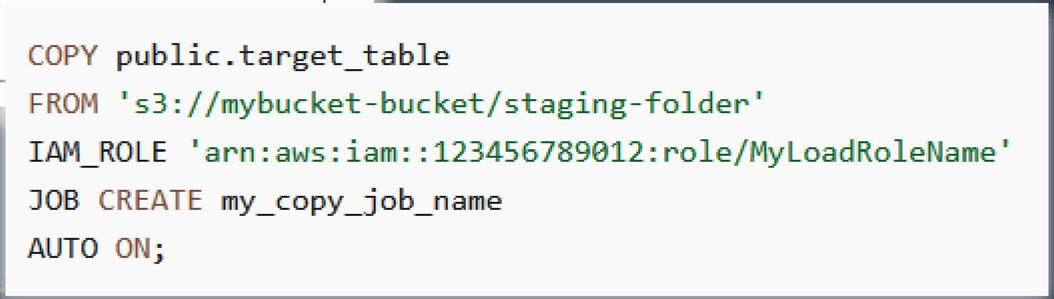
Zero ETL future enabled by service integrations



Continuous file ingestion from Amazon S3 (preview)



aws
VPC
Automatic data ingestion from Amazon S3 using Copy Jobs
External sources
Data Source 1(RDBMS) (SQL-Oracle-PG, My SQL)
Data Source 2 (Flat Files) (CSV, JSON)
Data Source N (Applications)
Extract
Bucket -1 Folder 1 - File1… N (CSV)
Bucket -1 Folder 2 - File1… N (Json)
Bucket -2 Folder 1 - File1… N (CSV)
Copy Job1
Copy-job 2
Copy Job 3
Table 1
Table 2
Amazon S3- Landing Zone
Amazon Redshift- Data Warehouse



```
COPY public.target_table
FROM 's3://mybucket-bucket/staging-folder'
IAM_ROLE 'arn:aws:iam::123456789012:role/MyLoadRoleName'
JOB CREATE my_copy_job_name
AUTO ON;
```

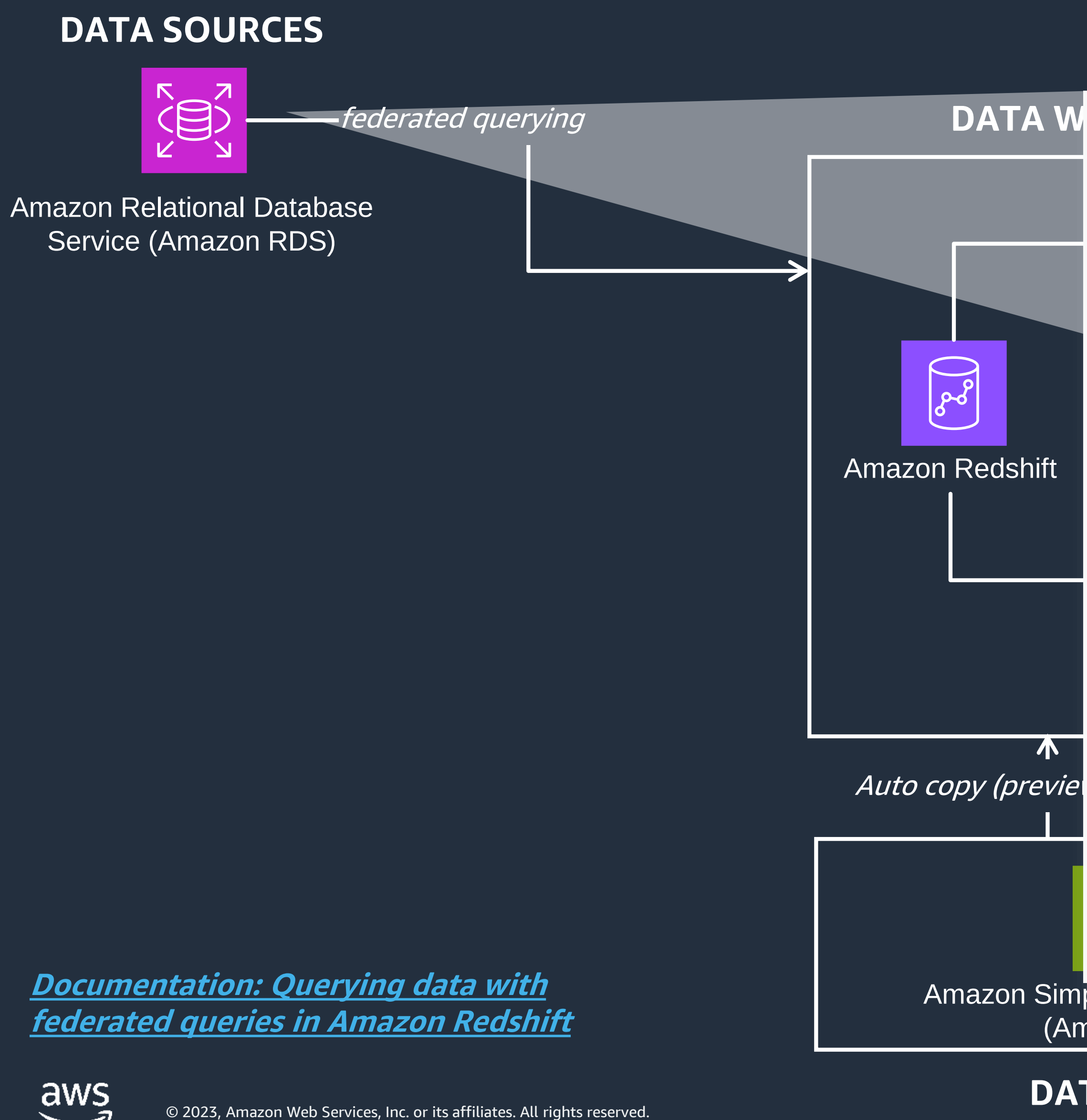
OTHER

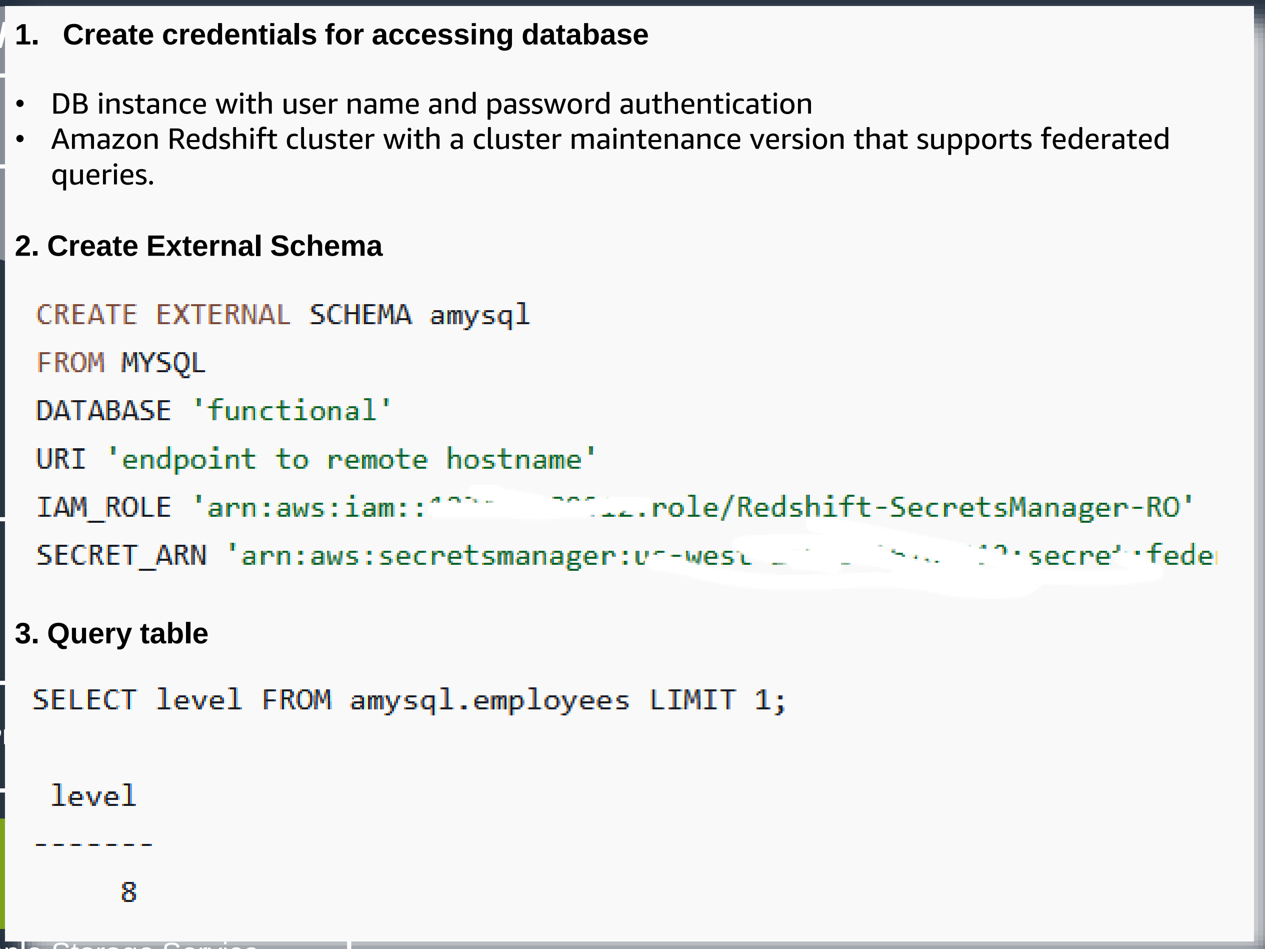
Amazon Simple Storage
(Amazon S3)

DATA LAKE

[AWS Blog: Simplify data ingestion from Amazon S3 to Amazon Redshift using auto-copy (preview)]

Zero ETL future enabled by service integrations


DATA SOURCES
federated querying
Amazon Relational Database Service (Amazon RDS)
DATA W
Amazon Redshift
Auto copy (previe
Amazon Simple Storage Service (Amazon S3)
DATA LAKE



1. **Create credentials for accessing database**

- DB instance with user name and password authentication
- Amazon Redshift cluster with a cluster maintenance version that supports federated queries.

2. Create External Schema

```
CREATE EXTERNAL SCHEMA amysql
FROM MYSQL
DATABASE 'functional'
URI 'endpoint to remote hostname'
IAM_ROLE 'arn:aws:iam::[redacted].role/Redshift-SecretsManager-RO'
SECRET_ARN 'arn:aws:secretsmanager:[redacted]:fede
```

3. Query table

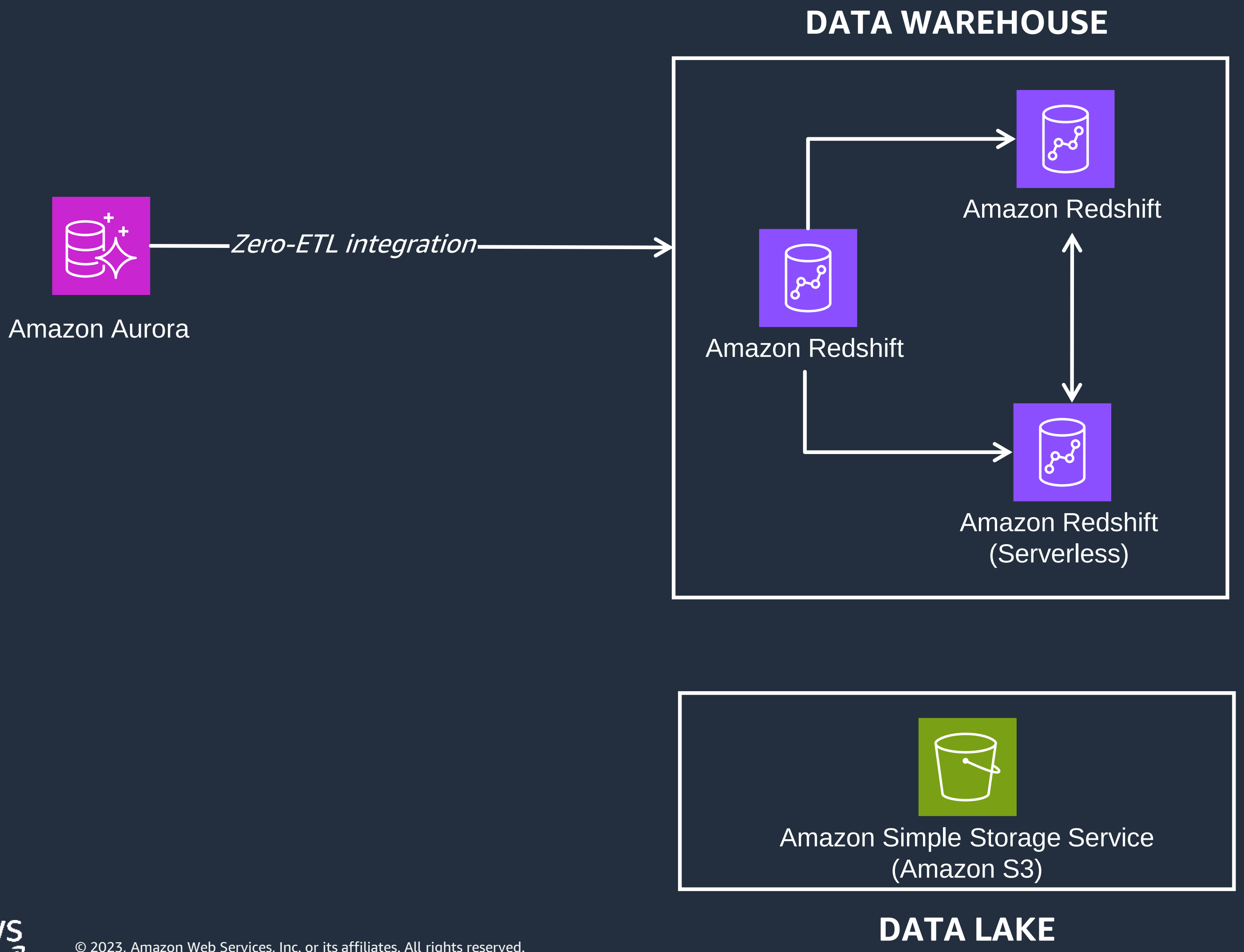
```
SELECT level FROM amysql.employees LIMIT 1;


 level
-------
     8
```

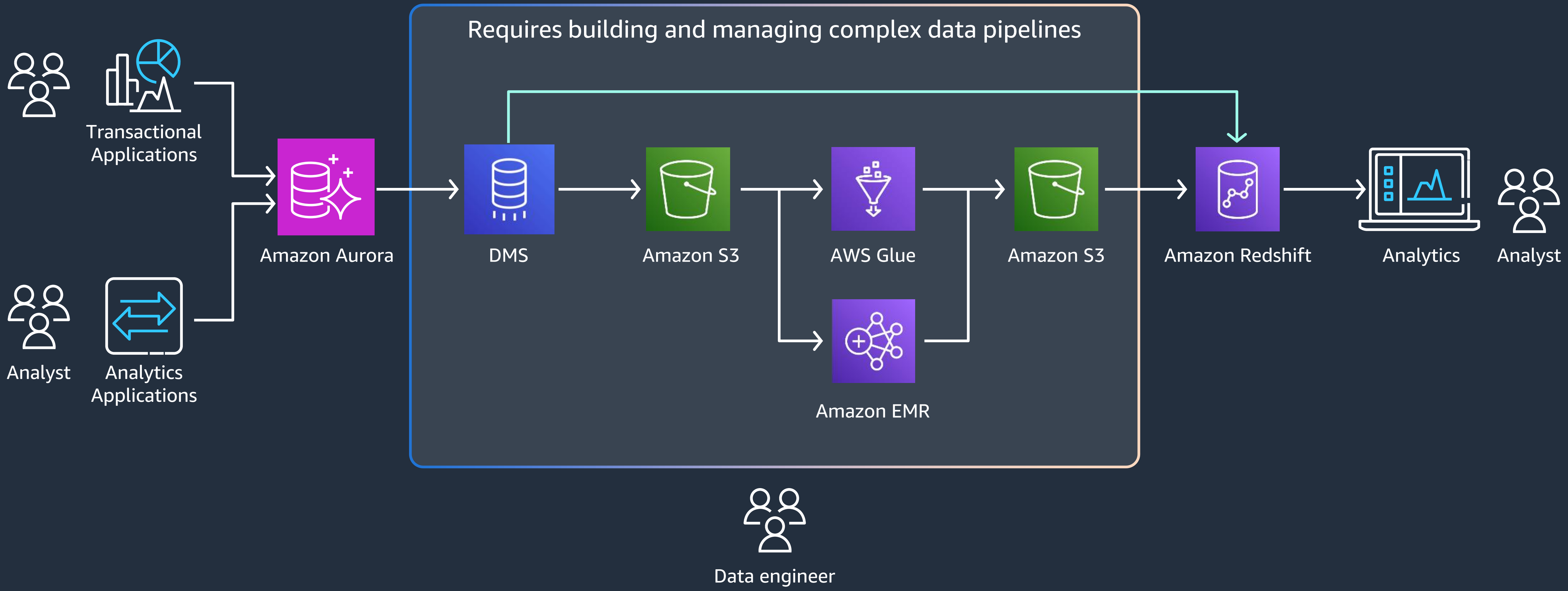
[Documentation: Querying data with federated queries in Amazon Redshift]()

Zero ETL future enabled by service integrations

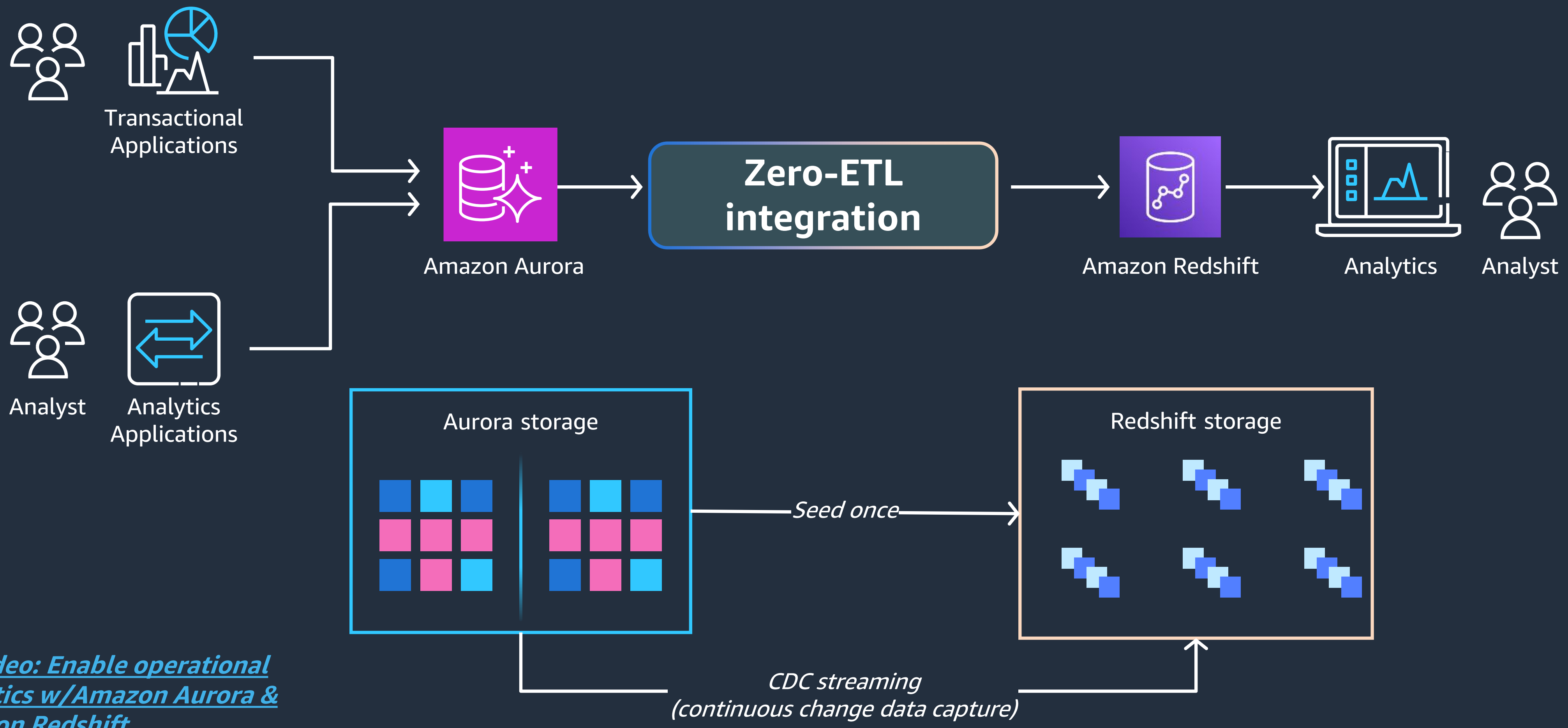
DATA SOURCES



Transaction analytics at scale

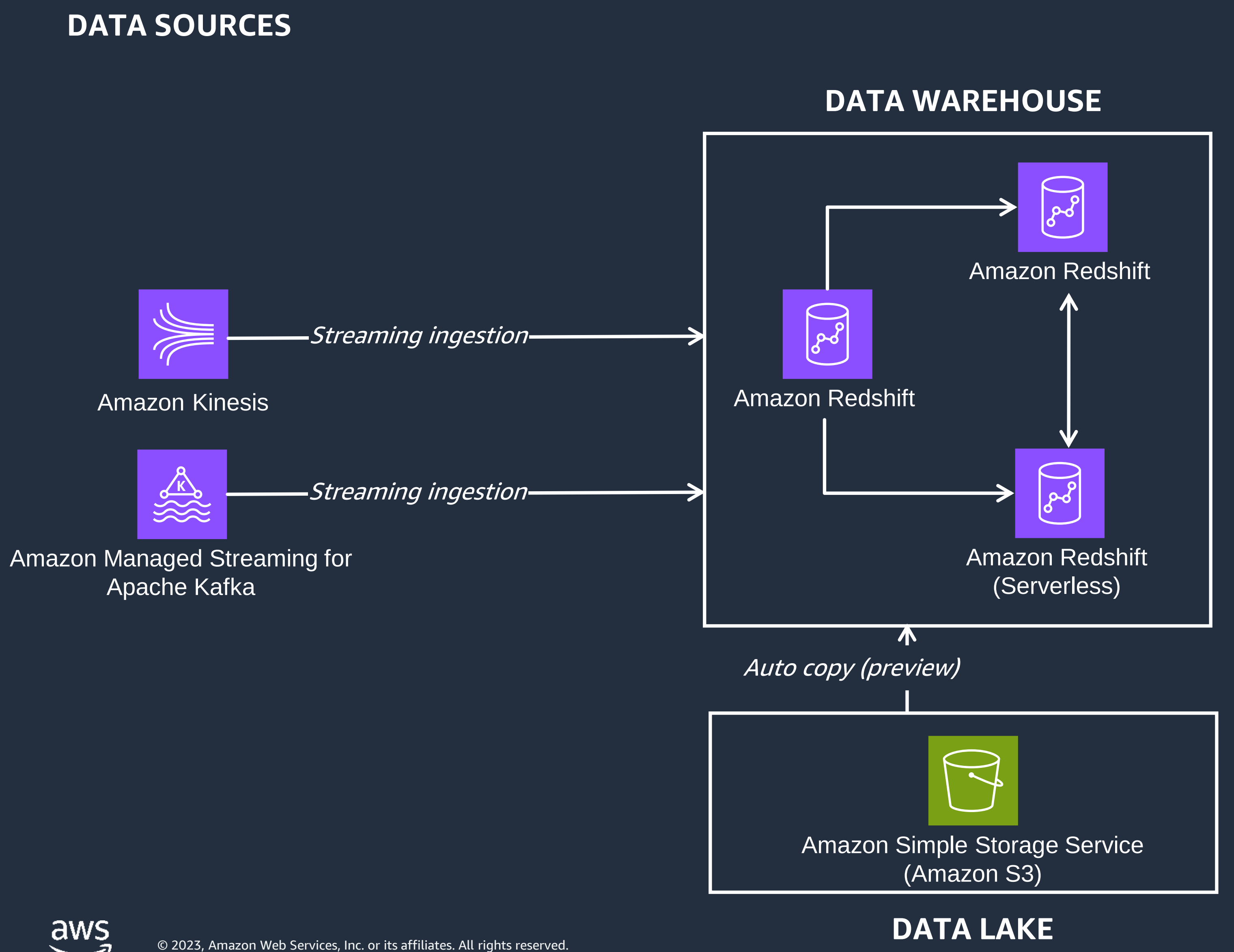


Amazon Aurora zero-ETL integration with Amazon Redshift

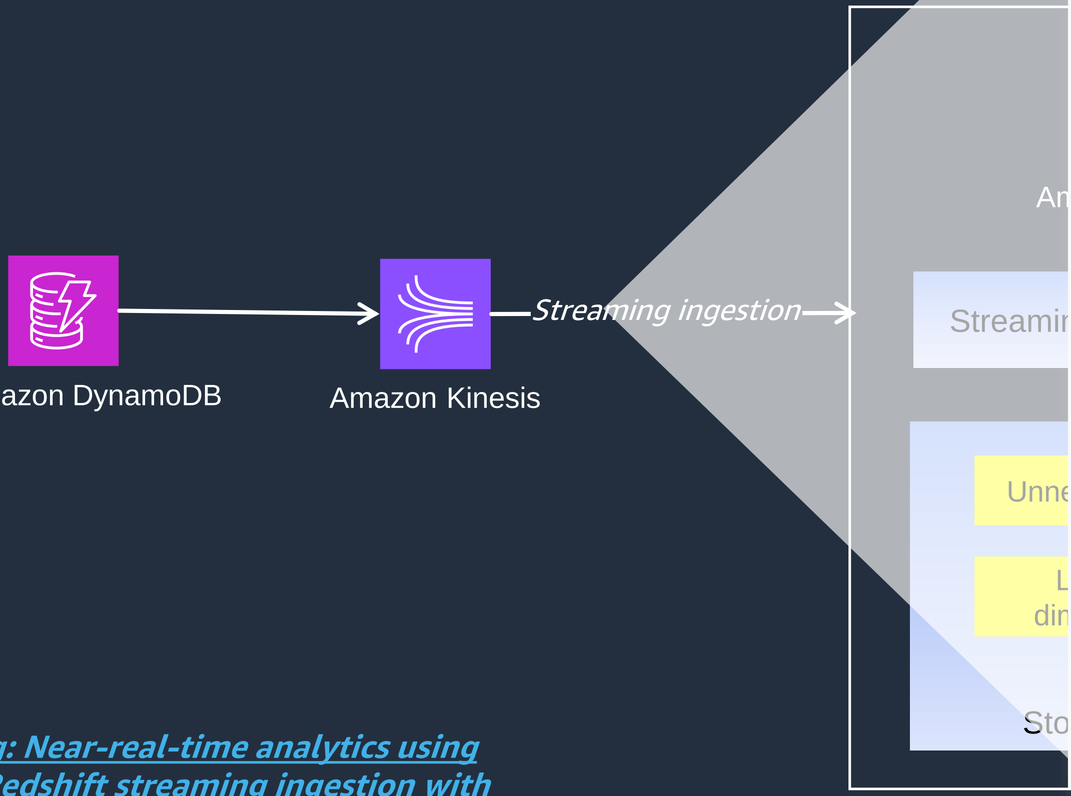


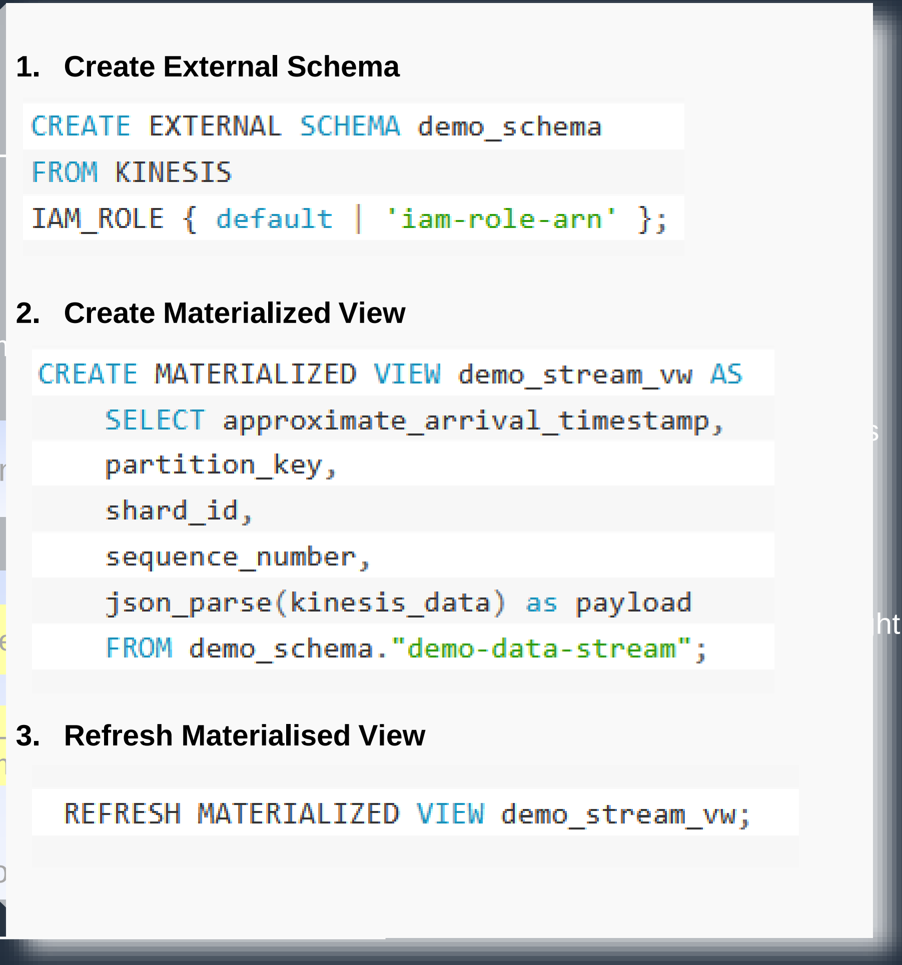
[*YT Video: Enable operational analytics w/Amazon Aurora & Amazon Redshift*]

Zero ETL future enabled by service integrations



Near-real-time analytics using Amazon Redshift streaming ingestion





1. **Create External Schema**

```
CREATE EXTERNAL SCHEMA demo_schema
FROM KINESIS
IAM_ROLE { default | 'iam-role-arn' };
```

2. **Create Materialized View**

```
CREATE MATERIALIZED VIEW demo_stream_vw AS
    SELECT approximate_arrival_timestamp,
    partition_key,
    shard_id,
    sequence_number,
    json_parse(kinesis_data) as payload
    FROM demo_schema."demo-data-stream";
```

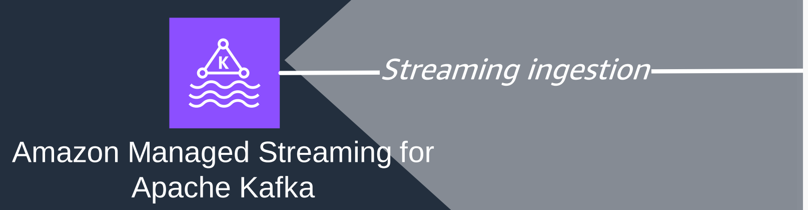
3. **Refresh Materialised View**

```
REFRESH MATERIALIZED VIEW demo_stream_vw;
```

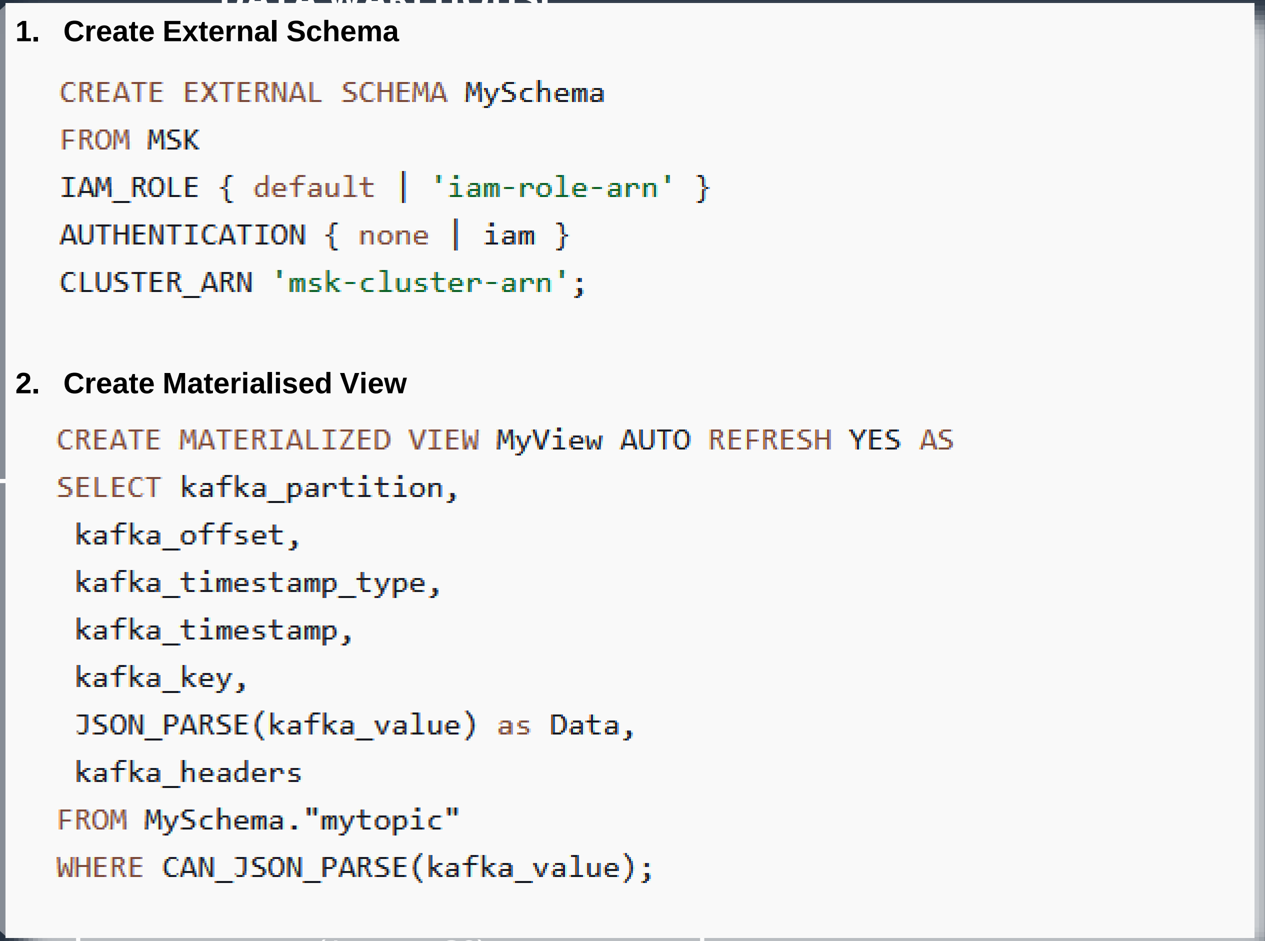
AWS Blog: Near-real-time analytics using Amazon Redshift streaming ingestion with Amazon Kinesis Data Streams and Amazon DynamoDB

Zero ETL future enabled by service integrations

DATA SOURCES



DATA WAREHOUSE



1. **Create External Schema**

```
CREATE EXTERNAL SCHEMA MySchema
FROM MSK
IAM_ROLE { default | 'iam-role-arn' }
AUTHENTICATION { none | iam }
CLUSTER_ARN 'msk-cluster-arn';
```

2. **Create Materialised View**

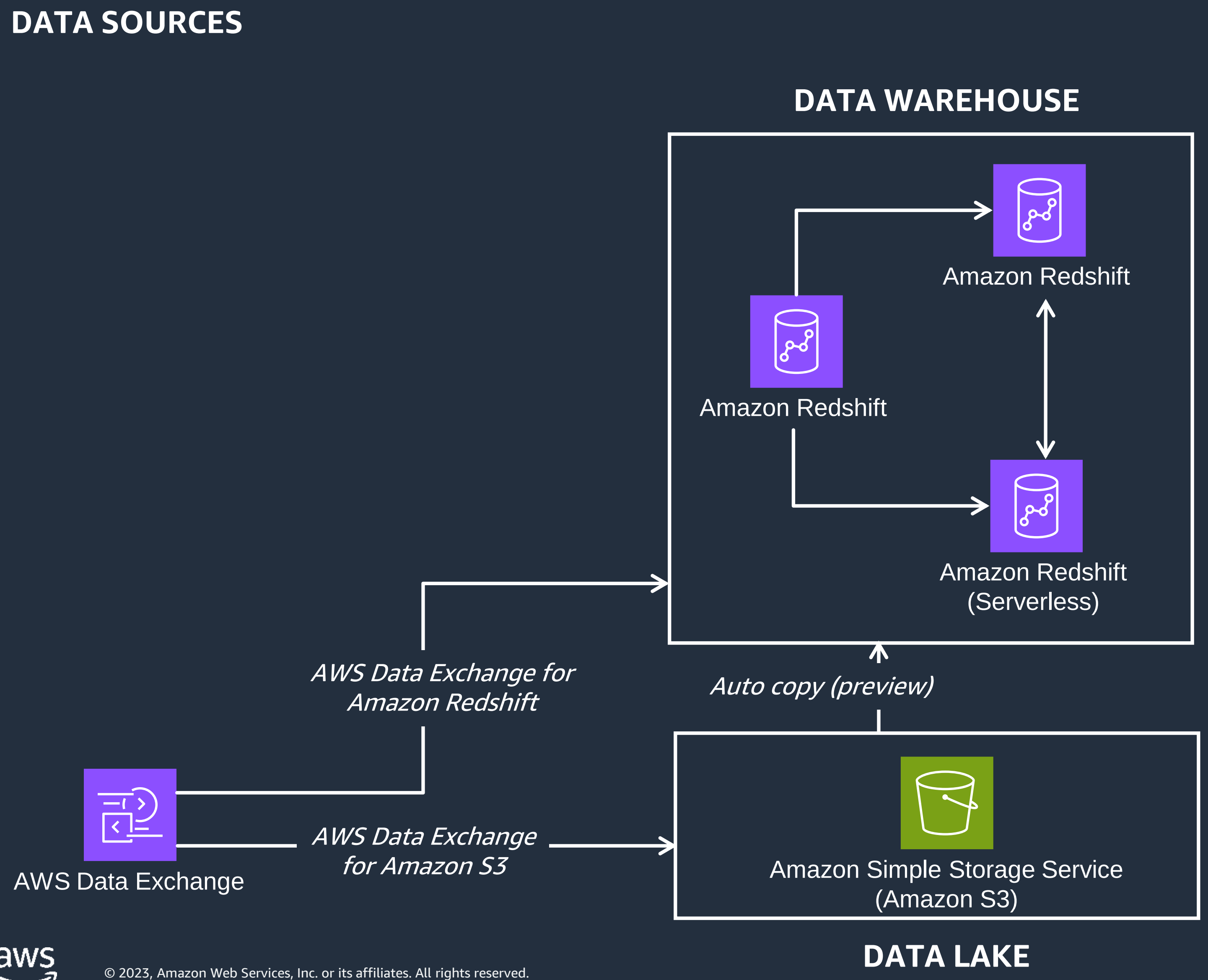
```
CREATE MATERIALIZED VIEW MyView AUTO REFRESH YES AS
SELECT kafka_partition,
 kafka_offset,
 kafka_timestamp_type,
 kafka_timestamp,
 kafka_key,
 JSON_PARSE(kafka_value) as Data,
 kafka_headers
FROM MySchema."mytopic"
WHERE CAN_JSON_PARSE(kafka_value);
```

(Amazon S3)

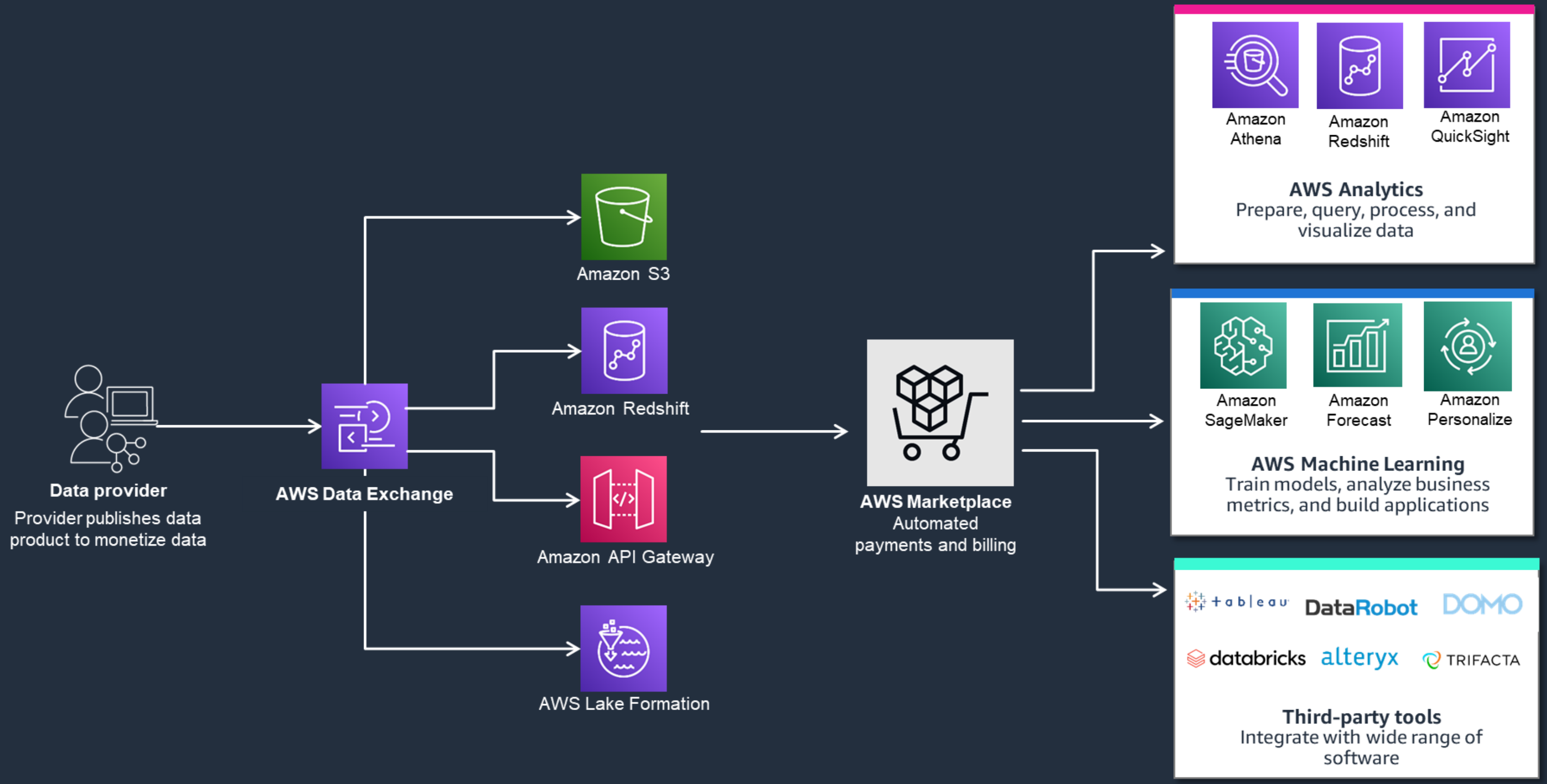
DATA LAKE

[Documentation: Getting started with streaming ingestion from Amazon Managed Streaming for Apache Kafka]

Zero ETL future enabled by service integrations



Find, subscribe to, and query third-party data without any ETL



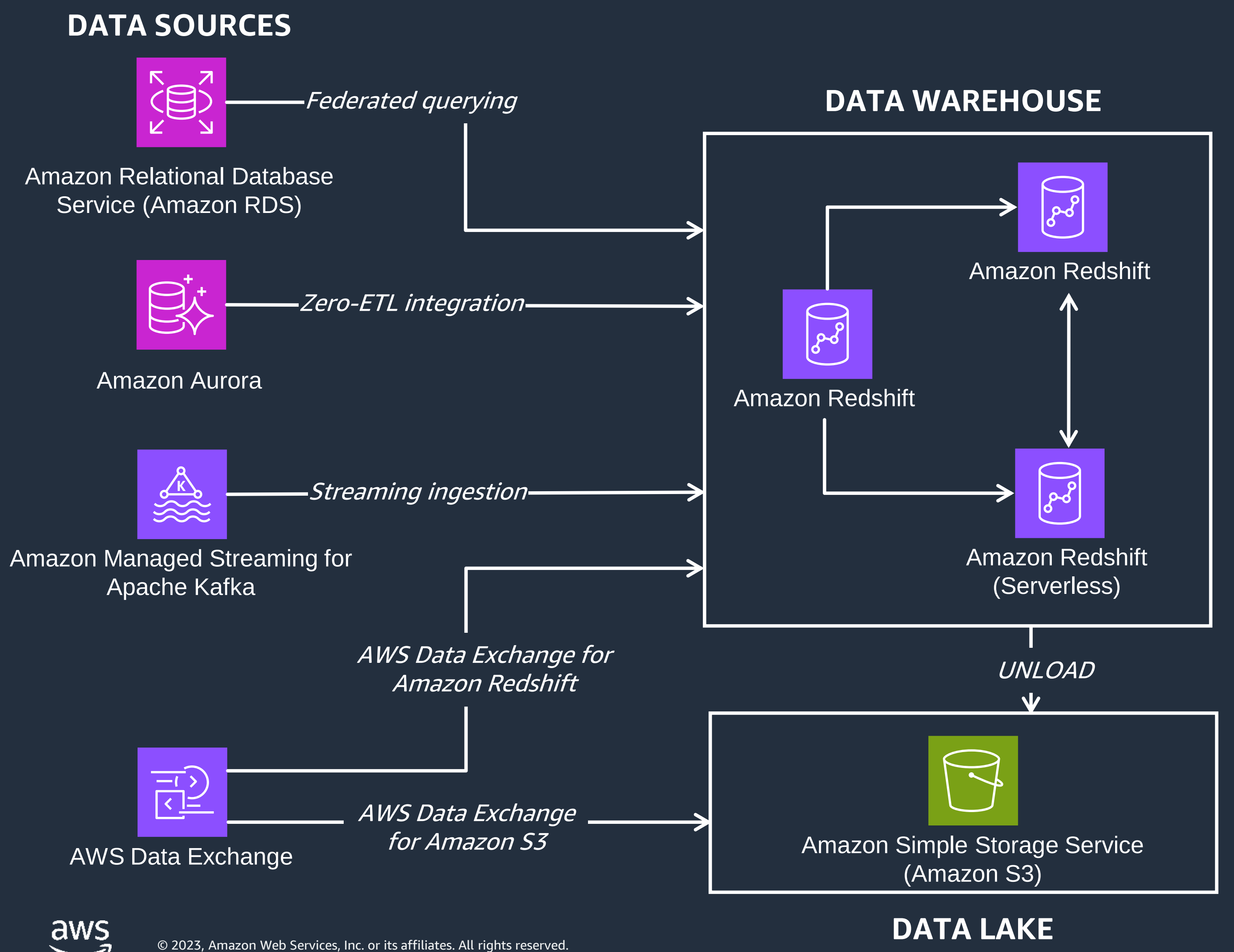
With **AWS Data Exchange for Amazon S3**, subscribers directly query provider's S3 bucket.

Data subscribers can subscribe to and directly query **Amazon Redshift data sets** without writing ETL processes.

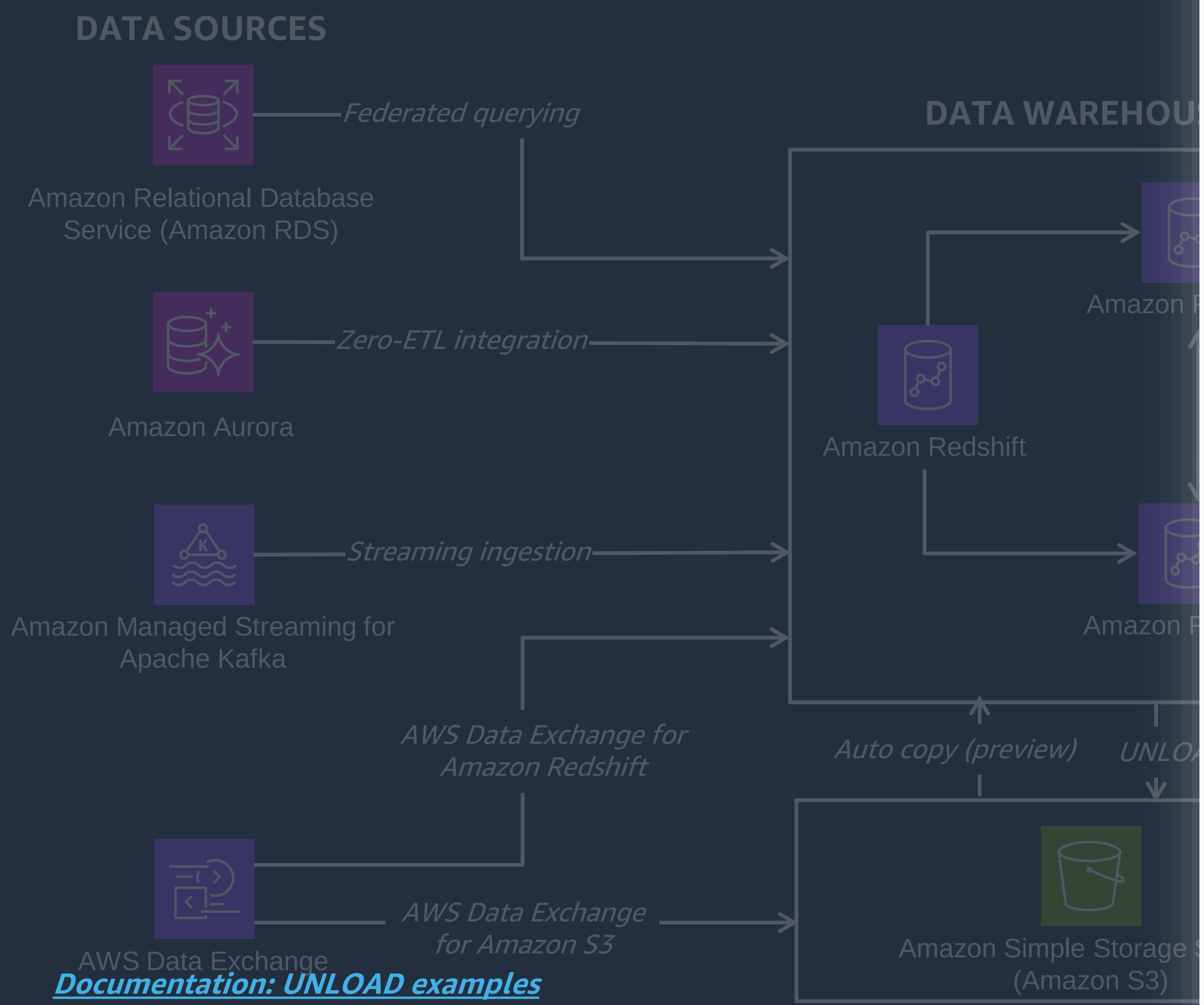
Using **AWS Data Exchange for Data APIs**, subscribers directly access data using their AWS IAM credentials and SDKs, no need for ETL pipelines.

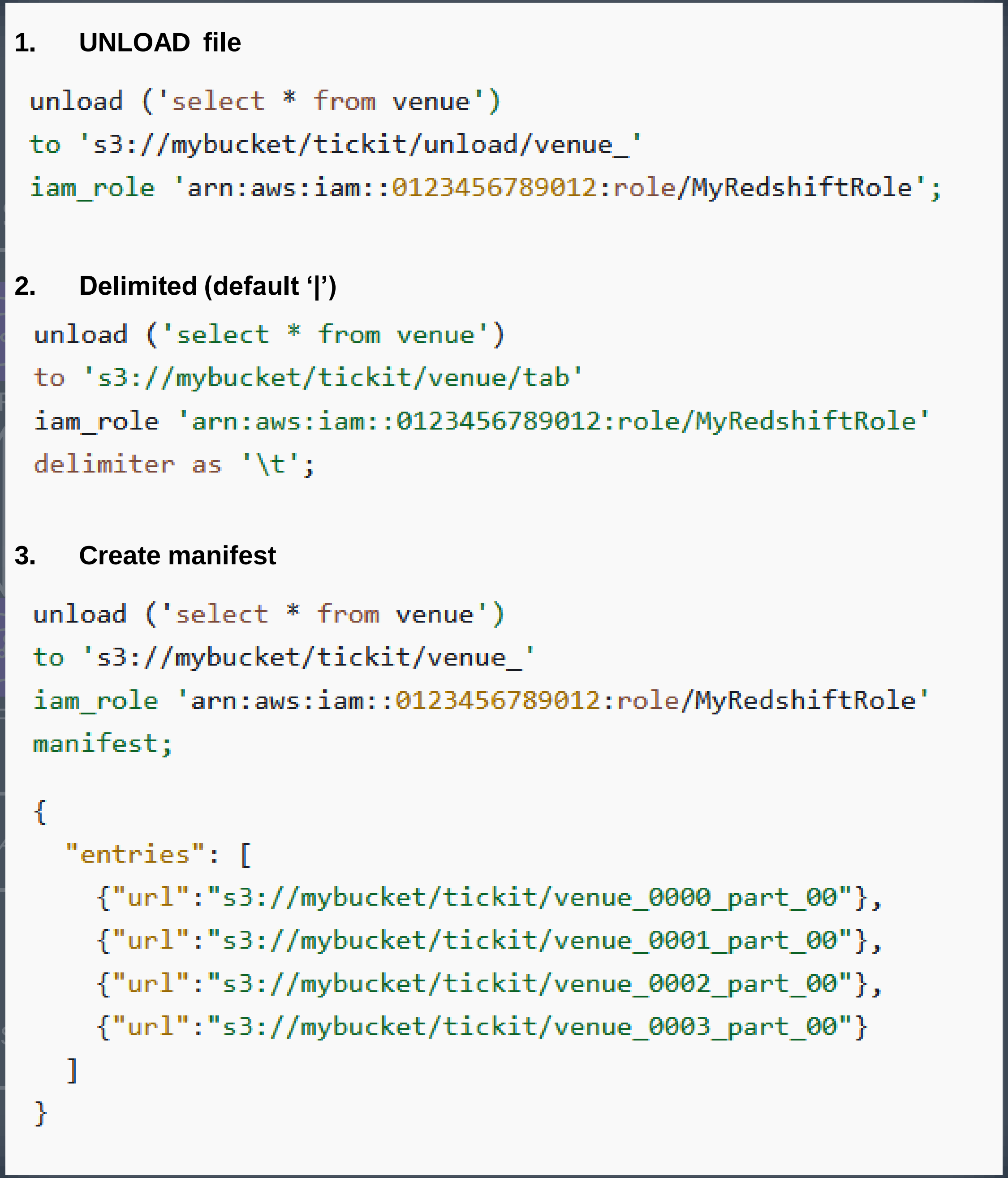
With **AWS Data Exchange for AWS Lake Formation**, subscribers can share access to the data within their AWS account or across AWS organization.

Zero ETL future enabled by service integrations


DATA SOURCES
Amazon Relational Database Service (Amazon RDS)
Federated querying
DATA WAREHOUSE
Amazon Aurora
Zero-ETL integration
Amazon Redshift
Amazon Redshift
Amazon Managed Streaming for Apache Kafka
Streaming ingestion
Amazon Redshift (Serverless)
AWS Data Exchange for Amazon Redshift
UNLOAD
AWS Data Exchange
AWS Data Exchange for Amazon S3
Amazon Simple Storage Service (Amazon S3)
DATA LAKE

Unload data to your data lake





1. **UNLOAD file**

```
unload ('select * from venue')
to 's3://mybucket/tickit/unload/venue_'
iam_role 'arn:aws:iam::0123456789012:role/MyRedshiftRole';
```

2. **Delimited (default '|')**

```
unload ('select * from venue')
to 's3://mybucket/tickit/venue/tab'
iam_role 'arn:aws:iam::0123456789012:role/MyRedshiftRole'
delimiter as '\t';
```

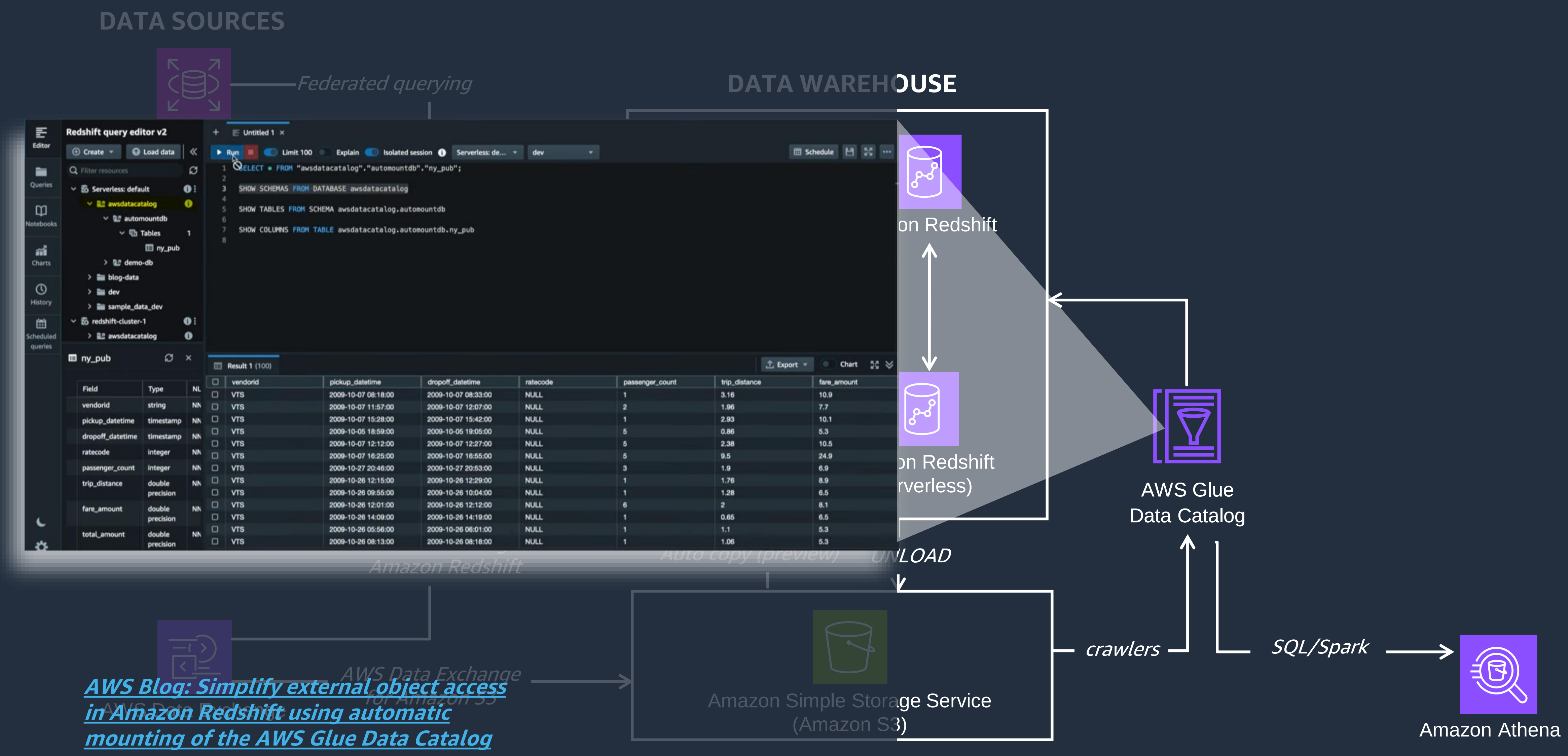
3. **Create manifest**

```
unload ('select * from venue')
to 's3://mybucket/tickit/venue_'
iam_role 'arn:aws:iam::0123456789012:role/MyRedshiftRole'
manifest;
```

```
{
  "entries": [
    {"url":"s3://mybucket/tickit/venue_0000_part_00"},
    {"url":"s3://mybucket/tickit/venue_0001_part_00"},
    {"url":"s3://mybucket/tickit/venue_0002_part_00"},
    {"url":"s3://mybucket/tickit/venue_0003_part_00"}
  ]
}
```

[Documentation: UNLOAD examples]

Zero ETL future enabled by service integrations



```
SELECT * FROM "awsdatacatalog"."automountdb"."ny_pub";

SHOW SCHEMAS FROM DATABASE awsdatacatalog

SHOW TABLES FROM SCHEMA awsdatacatalog.automountdb

SHOW COLUMNS FROM TABLE awsdatacatalog.automountdb.ny_pub

```

ny_pub

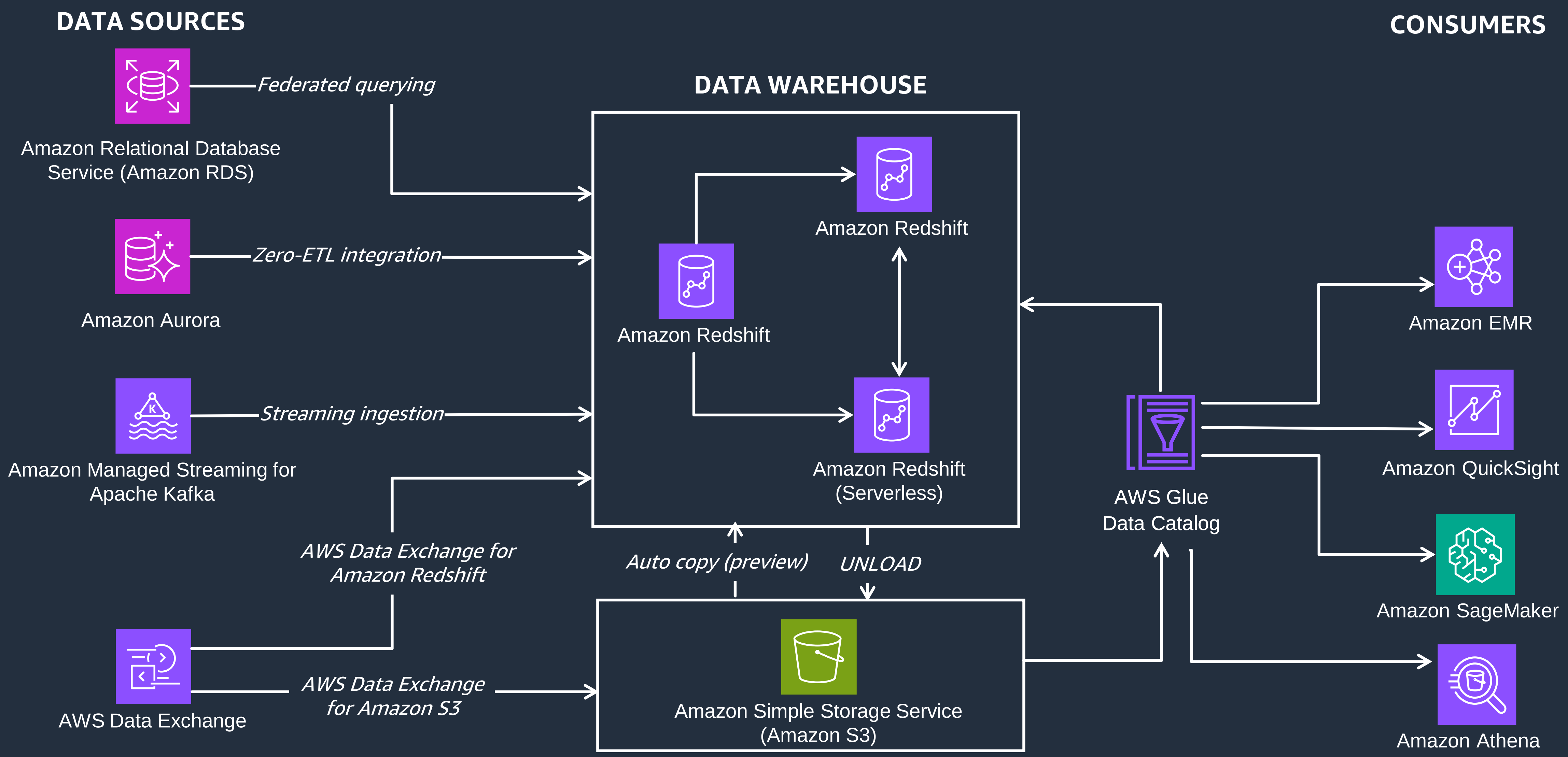
Field	Type	NL
vendorid	string	NN
pickup_datetime	timestamp	NN
dropoff_datetime	timestamp	NN
ratecode	integer	NN
passenger_count	integer	NN
trip_distance	double precision	NN
fare_amount	double precision	NN
total_amount	double precision	NN

Result 1 (100)

□	vendorid	pickup_datetime	dropoff_datetime	ratecode	passenger_count	trip_distance	fare_amount
□	VTS	2009-10-07 08:18:00	2009-10-07 08:33:00	NULL	1	3.16	10.9
□	VTS	2009-10-07 11:57:00	2009-10-07 12:07:00	NULL	2	1.96	7.7
□	VTS	2009-10-07 15:28:00	2009-10-07 15:42:00	NULL	1	2.93	10.1
□	VTS	2009-10-05 18:59:00	2009-10-05 19:05:00	NULL	5	0.86	5.3
□	VTS	2009-10-07 12:12:00	2009-10-07 12:27:00	NULL	5	2.38	10.5
□	VTS	2009-10-07 16:25:00	2009-10-07 16:55:00	NULL	5	9.5	24.9
□	VTS	2009-10-27 20:46:00	2009-10-27 20:53:00	NULL	3	1.9	6.9
□	VTS	2009-10-26 12:15:00	2009-10-26 12:29:00	NULL	1	1.76	8.9
□	VTS	2009-10-26 09:55:00	2009-10-26 10:04:00	NULL	1	1.28	6.5
□	VTS	2009-10-26 12:01:00	2009-10-26 12:12:00	NULL	6	2	8.1
□	VTS	2009-10-26 14:09:00	2009-10-26 14:19:00	NULL	1	0.65	6.5
□	VTS	2009-10-26 05:56:00	2009-10-26 06:01:00	NULL	1	1.1	5.3
□	VTS	2009-10-26 08:13:00	2009-10-26 08:18:00	NULL	1	1.06	5.3

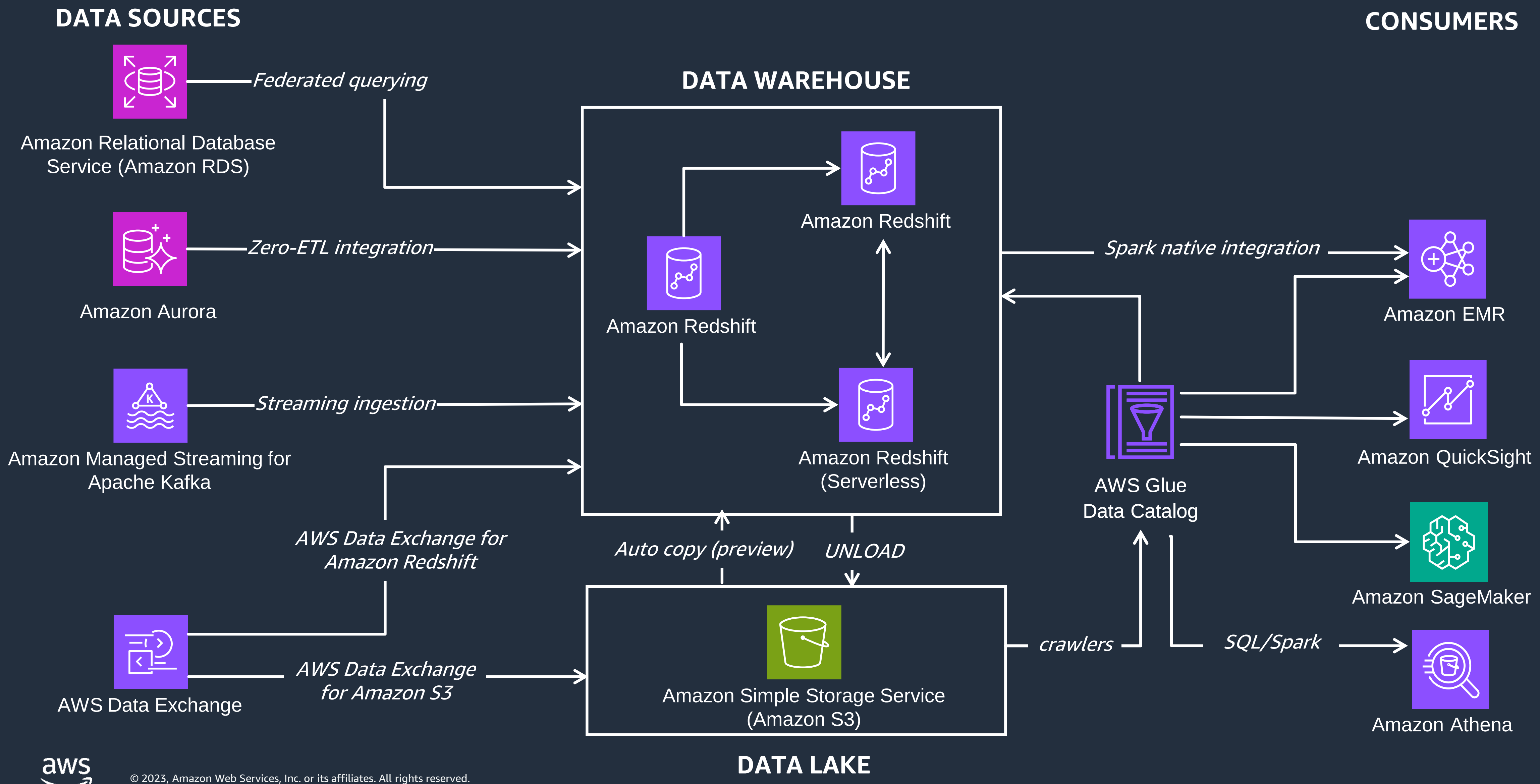
AWS Blog: Simplify external object access in Amazon Redshift using automatic mounting of the AWS Glue Data Catalog

Zero ETL future enabled by service integrations

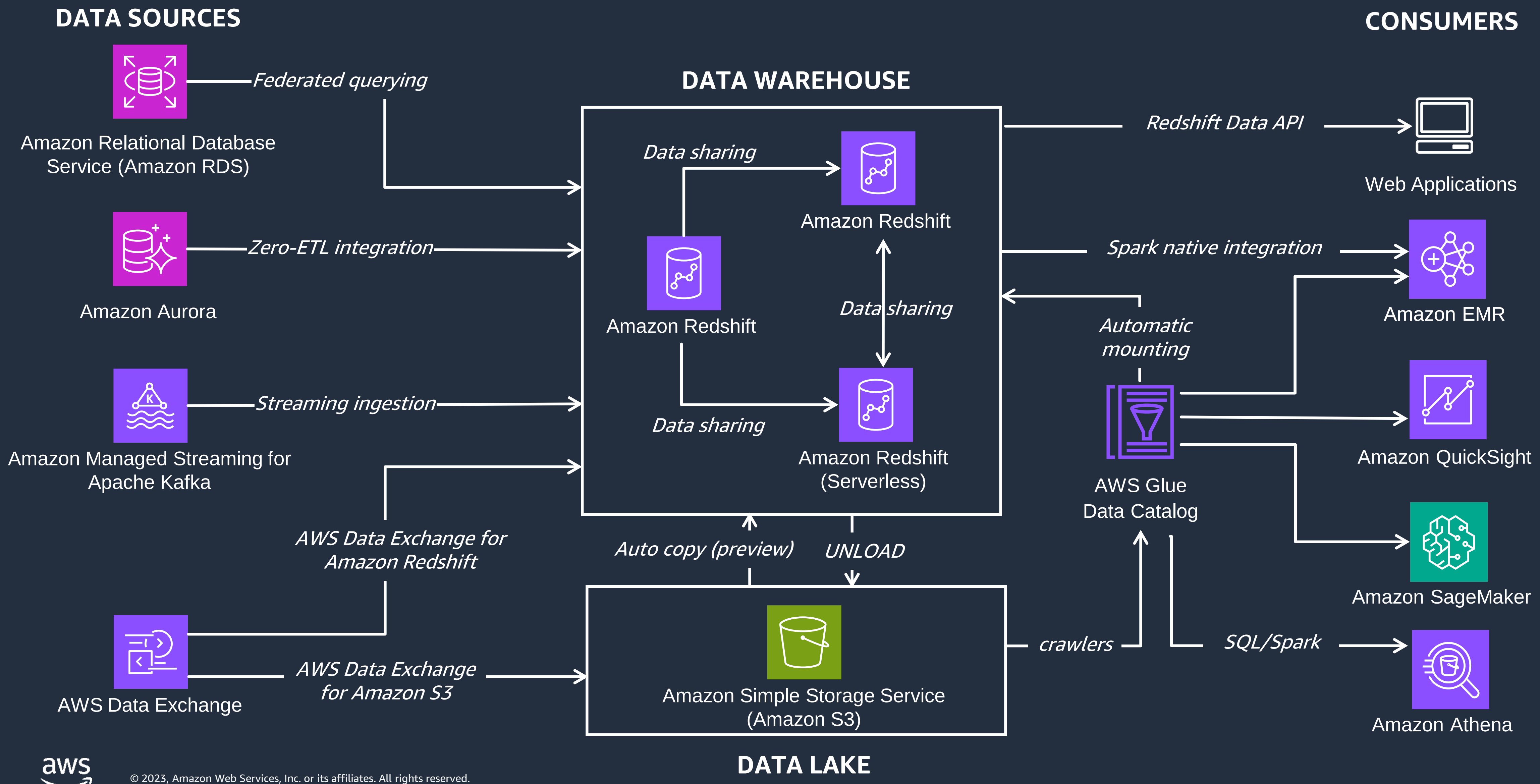


DATA LAKE

Zero ETL future enabled by service integrations



Connect all your data with native integrations



We are making data integration easier for you

Zero ETL
future enabled
by service integrations

AWS Glue for
value-add data
transformations and more

Services for
**connecting to 100s
of data sources**

We are here to help

Want to discover possibilities?

aws Well-architected Review

- ✓ Assess your workloads and learn how well your architecture aligns with cloud best practices and gain guidance for making improvements
- ✓ Use customer analytics lens to review your data workloads

Review and refine existing workloads

Want to build a data vision and strategy?

aws Data Driven Everything

- ✓ Create an organizational vision for innovation with data to drive business outcomes
- ✓ Define the first pilot, learn, and build

Jump-start the data flywheel

Want to modernize your data foundation?

aws Re:Imagine data

- ✓ Define the migrate & modernization strategy for a future data foundation
- ✓ Lower cost, increase capacity, and unlock business access

Migrate & Modernize data

Thank you!

Sandipan Bhaumik

/in/sandipanbhaumik