

# Data Warehouses in a Data Lakes Architecture

Raghu Prabhu, Global Business Development Manager for Data Lakes

# Agenda

- Introduction to Data Lake architecture on AWS
- Introduction to Redshift
- How does Redshift Spectrum work
- Example walkthrough
- Popular Redshift Spectrum use cases
- Q&A

# What has changed in the last five years?

- Cloud has changed everything
  - Limitless storage
  - Numerous compute options
  - Cost effective, no contracts
- There is a lot more data
- New breed of analysts, statisticians, and data scientists
- Applications and user experiences are guided by data



# There is **more data** than people think.

Data

grows  
**>10x**  
every 5 years

Data platforms need to

live for  
**15**  
years

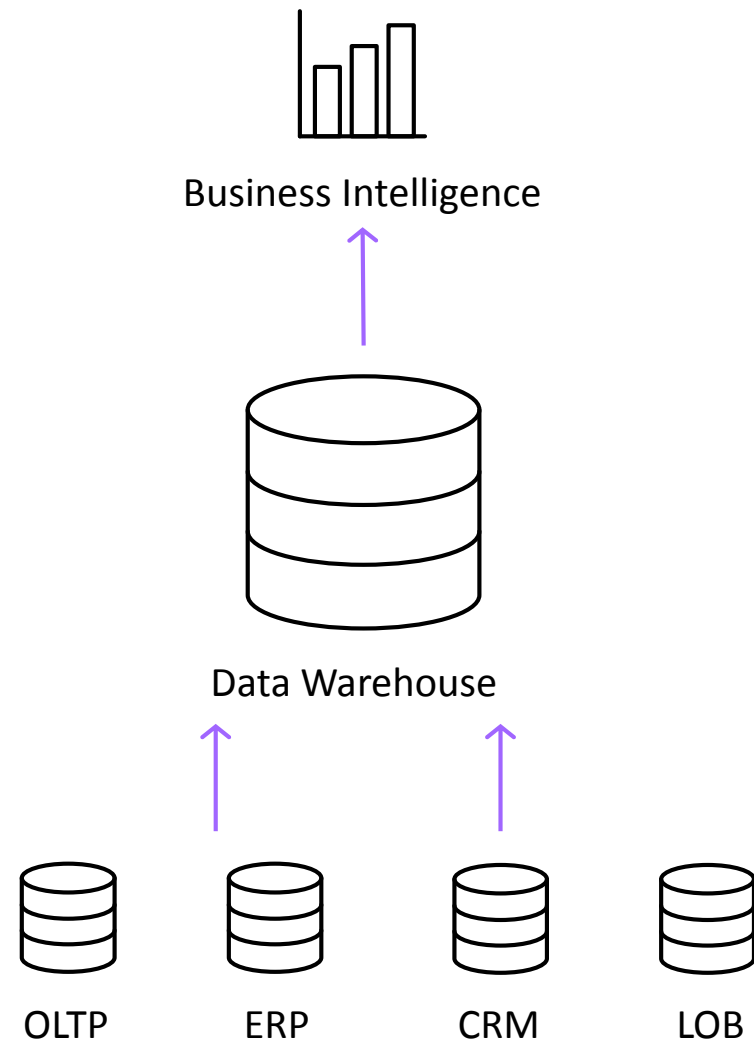
scale  
**1,000x**

\* IDC, Data Age 20215: The Evolution of Data to Life-Critical Don't Focus on Big Data, Focus on the Data That's Big, April 2017.

# What is a data lake?

A data lake is a **centralized repository** that allows you to store all your **structured and unstructured** data at any scale

# Traditionally, analytics looked like this



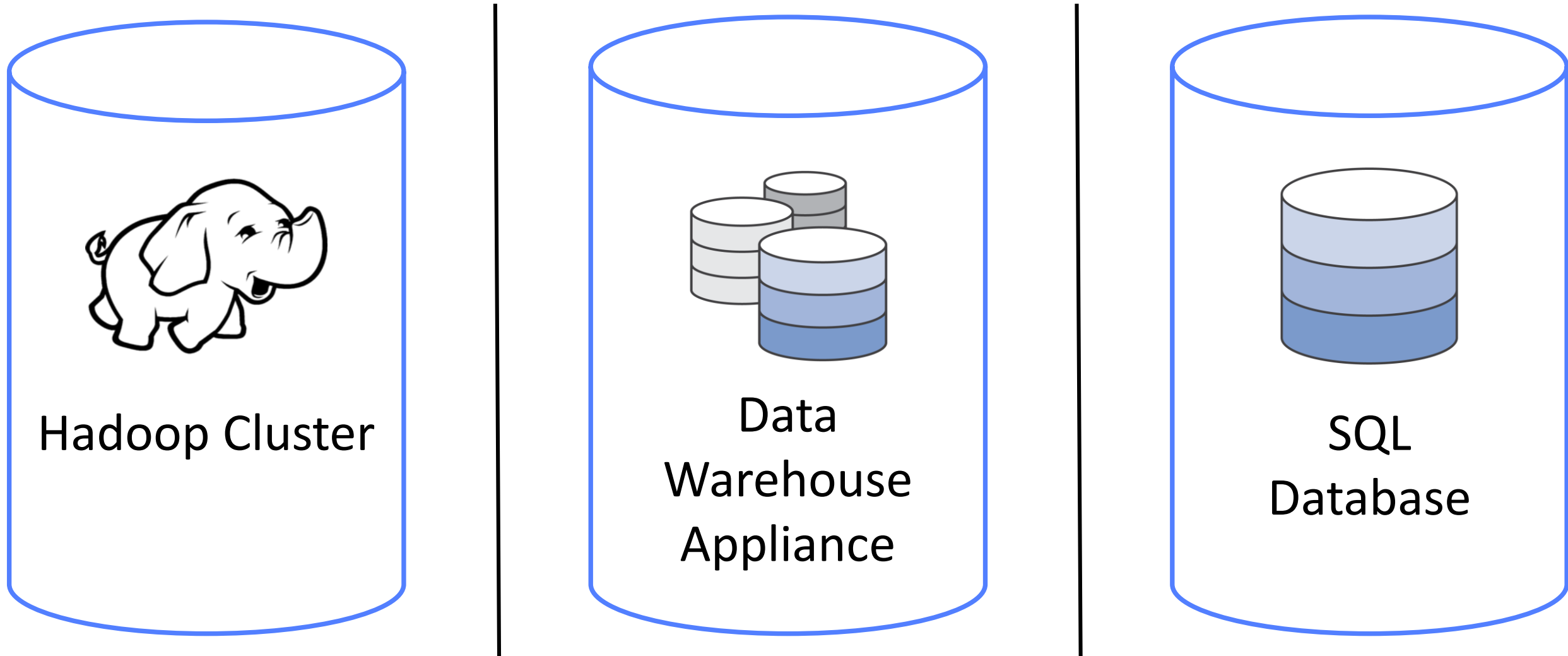
Expensive: Large initial capex + \$10k-\$50k/TB/year

GBs-TBs scale - not designed for PB/EBs

Primarily relational data

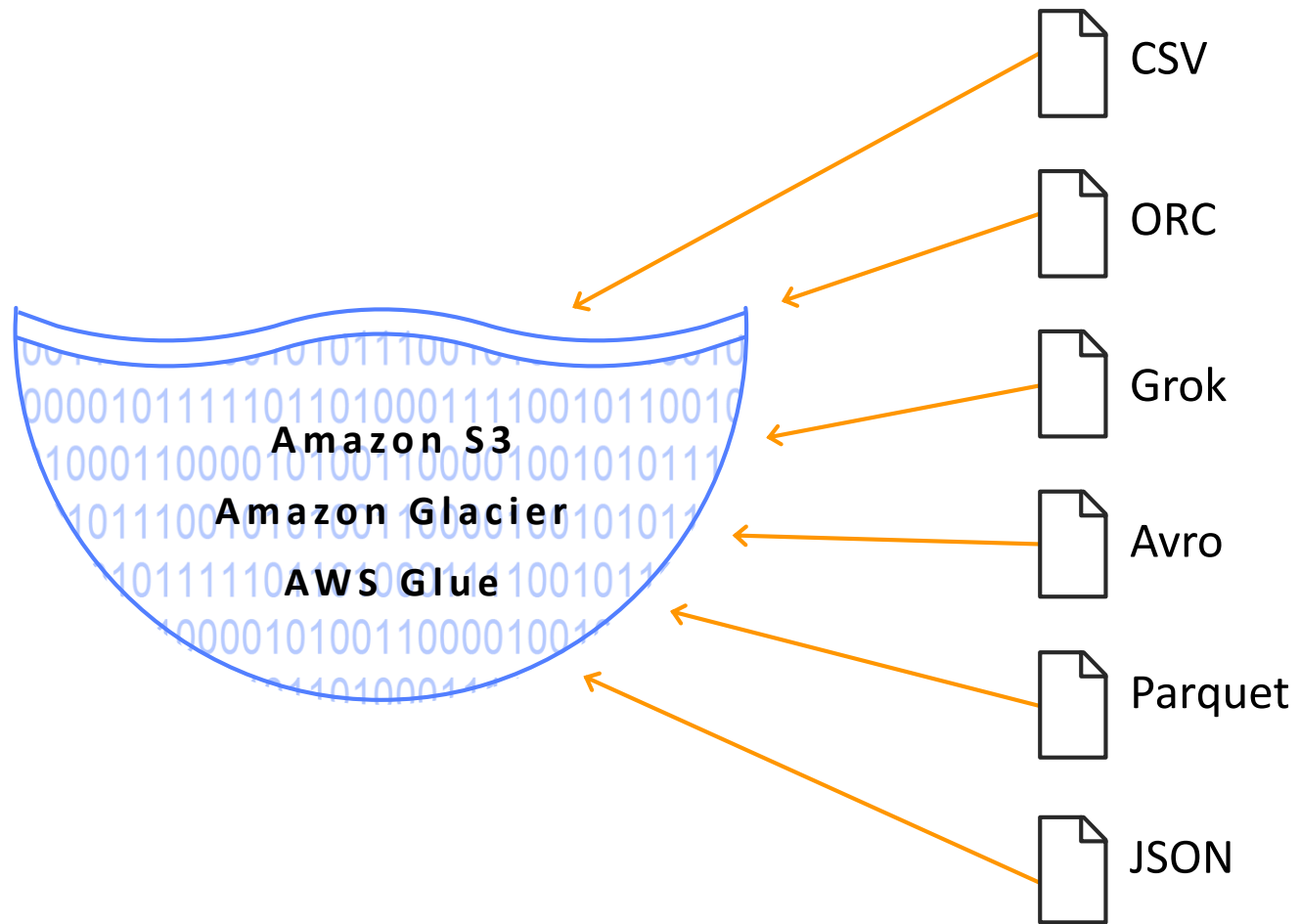
90% of data was deleted to reduce cost

# Analytics operated on isolated data silos



# Store data in the format you want

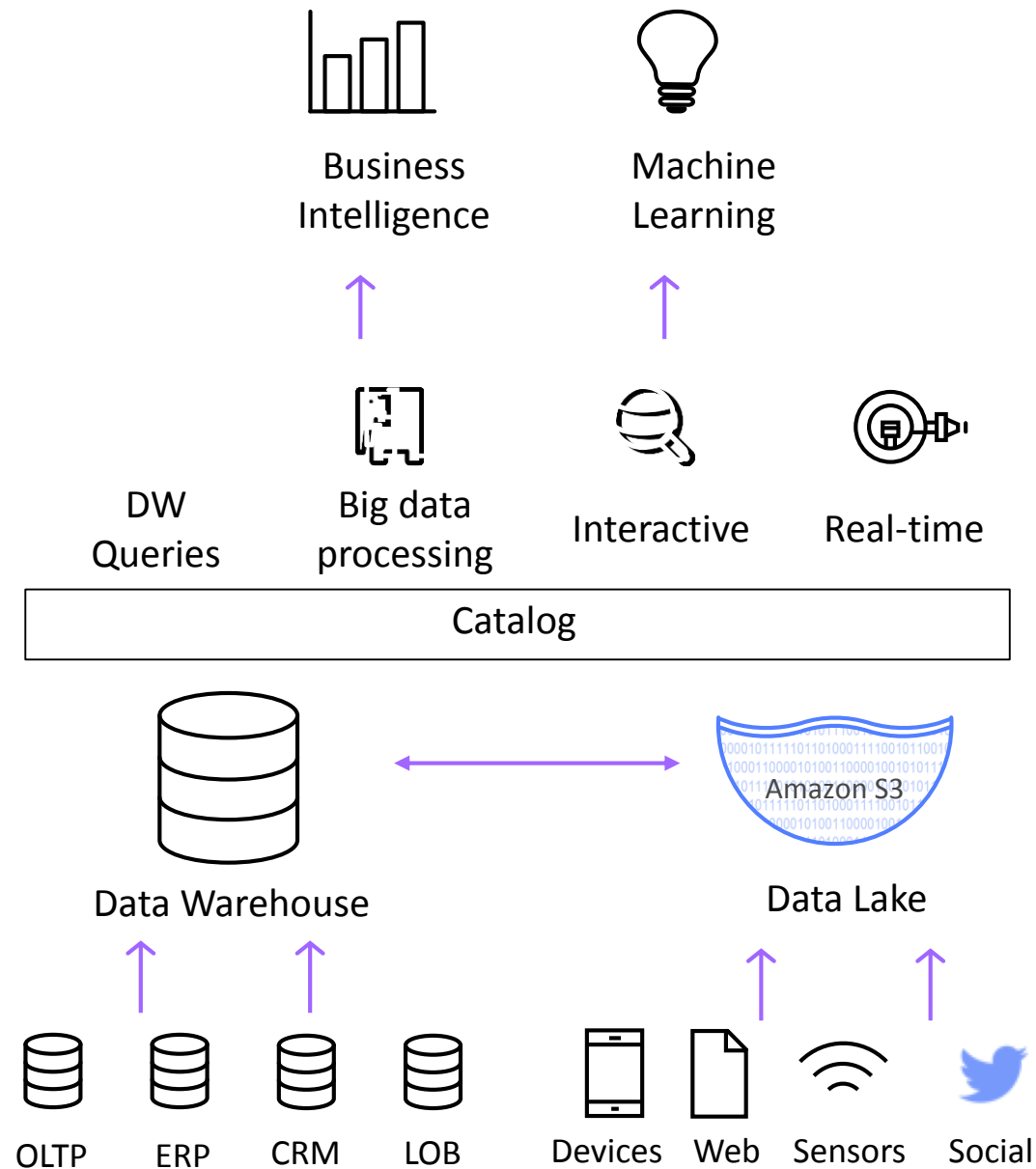
Open and comprehensive



- Store data in the format you want:
  - **Text files like CSV**
  - **Columnar like Apache Parquet, and Apache ORC**
  - **Logstash like Grok**
  - **JSON (simple, nested), AVRO**
  - **and more**



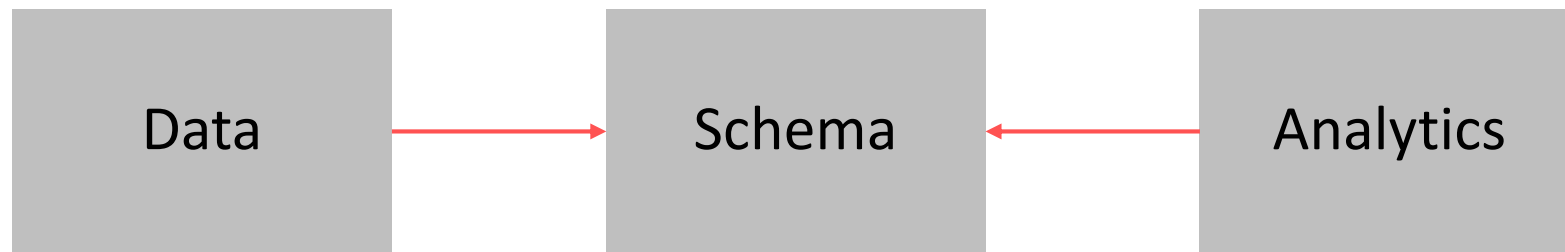
# Data lakes extend the traditional approach



$$\begin{aligned} & \checkmark \text{ The analytical power of data warehouse} \\ & \quad + \\ & \checkmark \text{ The limitless scalability of serverless compute} \\ & \quad + \\ & \checkmark \text{ The distributed processing of big data systems} \end{aligned} =$$

# What is a Data Catalog?

Data Catalog is a schema repository of all your data in one single place irrespective of the data location which may be S3, JDBC sources and NoSQL sources



- Separate your data from your schema so one can evolve independent of the other
- Separate your compute (Analytics) from your data so one can scale independent of the other

# Glue Data Catalog

AWS Glue

Data catalog

Databases

Tables

Connections

Crawlers

Classifiers

Settings

ETL

Jobs

ML

Transforms

Services ▾ Resource Groups ▾

Raghu @ raghuaws ▾ N. Virginia

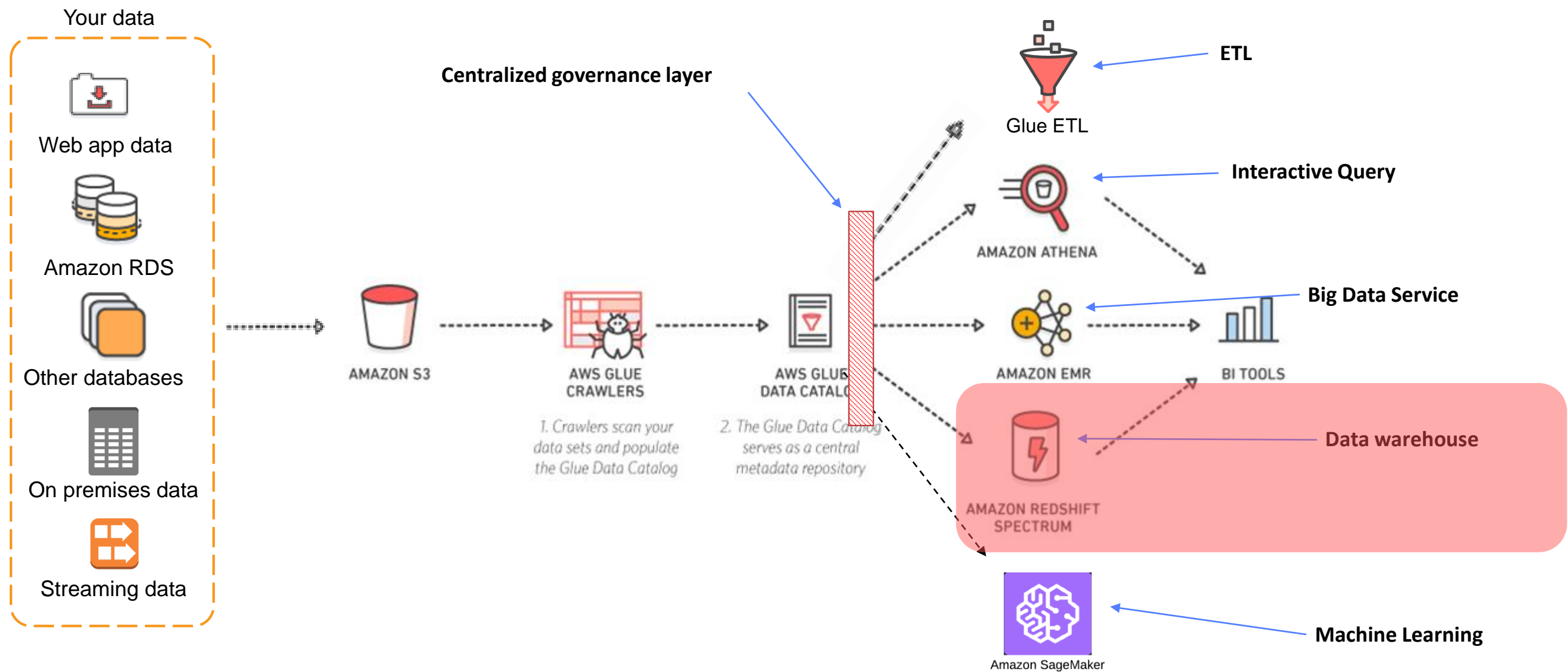
Tables A table is the metadata definition that represents your data, including its schema. A table can be used as a source or target in a job definition.

Add tables ▾ Action ▾ Filter by attributes or search by keyword Save view ▾

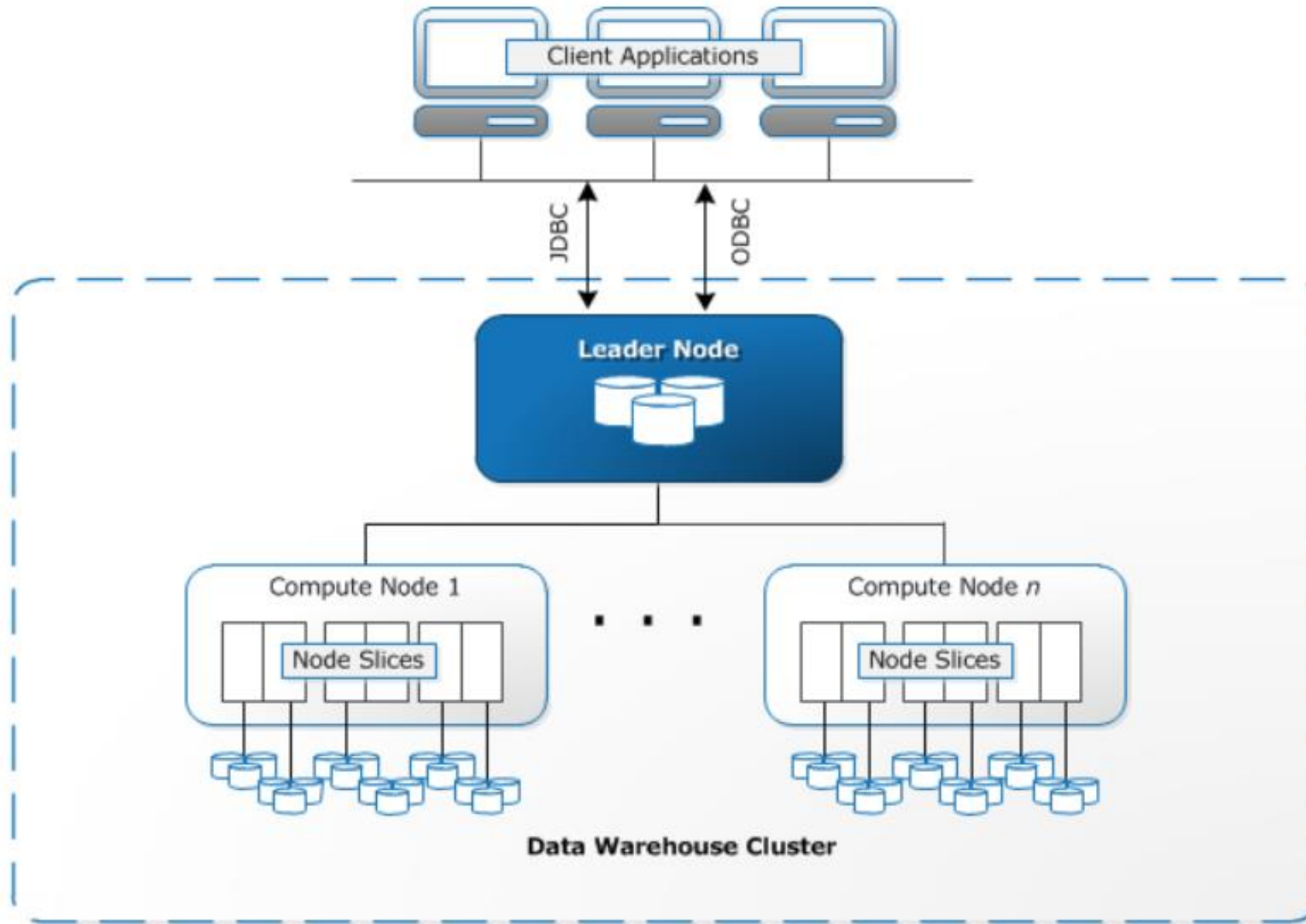
Showing: 1 - 51

Name	Database	Location	Classification	Last updated	Description
<input type="checkbox"/> example	amzn	s3://raghuoutput/example	Unknown	19 November 2018 1...	
<input type="checkbox"/> fhv	amzn	s3://nyctaxi101/fhv/	csv	30 October 2018 2:24...	
<input type="checkbox"/> green	amzn	s3://nyctaxi101/green/	csv	30 October 2018 2:24...	
<input type="checkbox"/> yellow	amzn	s3://nyctaxi101/yellow/	csv	30 October 2018 2:24...	
<input type="checkbox"/> aviation101	aviation	s3://aviation101/	csv	16 January 2018 2:49...	
<input type="checkbox"/> kevin	aviation	s3://raghuoutput/test34	Unknown	29 October 2018 9:30...	
<input type="checkbox"/> postgresdata_store_a...	aviation	data_store_api.public.aaAviation	postgresql	9 February 2018 10:4...	
<input type="checkbox"/> lakeformationdemoim...	datalake-import-staging-5znq9dkktr	wordpress.wp_commentmeta	mysql	8 January 2019 6:13 ...	

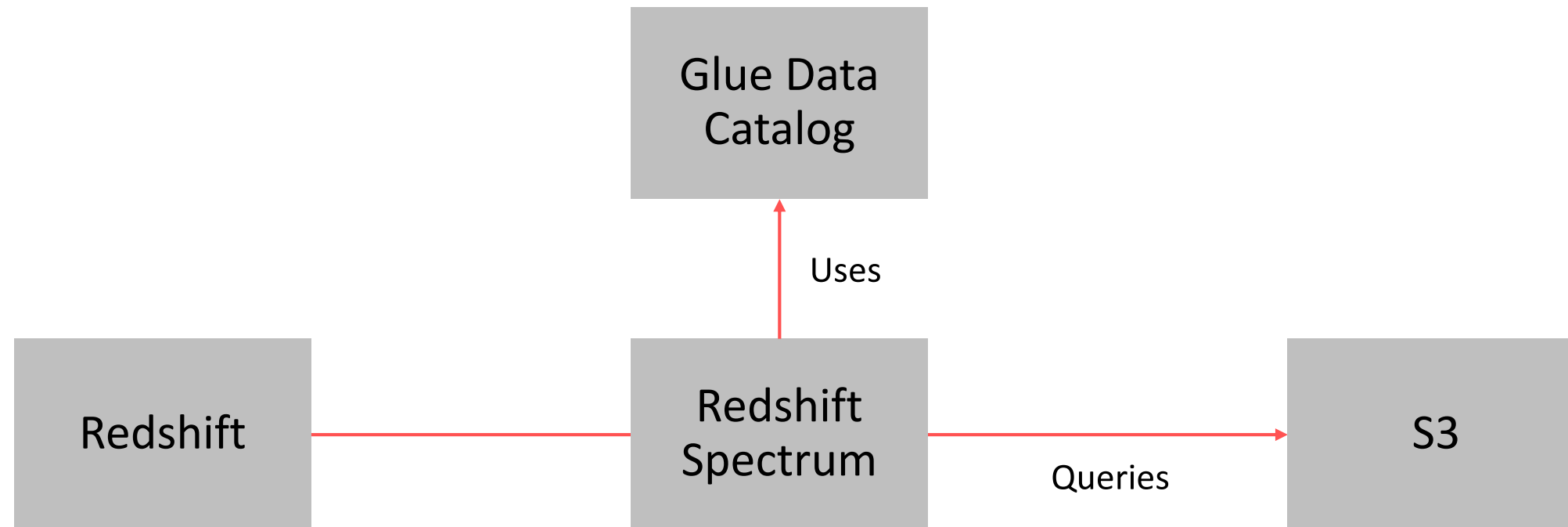
# Data Lakes on S3



# Redshift – Quick Intro



# Redshift Spectrum



# Example walkthrough

AWS Glue

Data catalog

Databases

**Tables**

Connections

Crawlers

Classifiers

Settings

ETL

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.

Add tables

Action

Database : nyctaxi

Filter or search for tables...

Save view

Showing: 1 - 4

<input type="checkbox"/>	Name	Database	Location	Classification	Last updated	Depreca
<input type="checkbox"/>	fhv	nyctaxi	s3://nyctaxi101/fhv/	csv	20 February 2018 9:06 A...	
<input type="checkbox"/>	green	nyctaxi	s3://nyctaxi101/green/	csv	20 February 2018 9:06 A...	
<input type="checkbox"/>	kevin	nyctaxi	s3://raghuoutput/test341	Unknown	29 October 2018 9:33 AM...	
<input type="checkbox"/>	yellow	nyctaxi	s3://nyctaxi101/yellow/	csv	20 February 2018 9:06 A...	



# Example walkthrough

Name	yellow
Description	
Database	nyctaxi
Classification	csv
Location	<a href="#">s3://nyctaxi101/yellow/</a>


Schema

	Column name	Data type
1	vendorid	bigint
2	tpep_pickup_datetime	string
3	tpep_dropoff_datetime	string
4	passenger_count	bigint
5	trip_distance	double
6	pickup_longitude	double
7	pickup_latitude	double
8	ratecodeid	bigint
9	store_and_fwd_flag	string
10	dropoff_longitude	double
11	dropoff_latitude	double
12	payment_type	bigint
13	fare_amount	double
14	extra	double
15	mta_tax	double



# Example walkthrough

```
create external schema ext  
from data catalog  
database 'nyctaxi'  
iam_role 'arn:aws:iam::012345678901:role/RedshiftSpectrumRole'
```



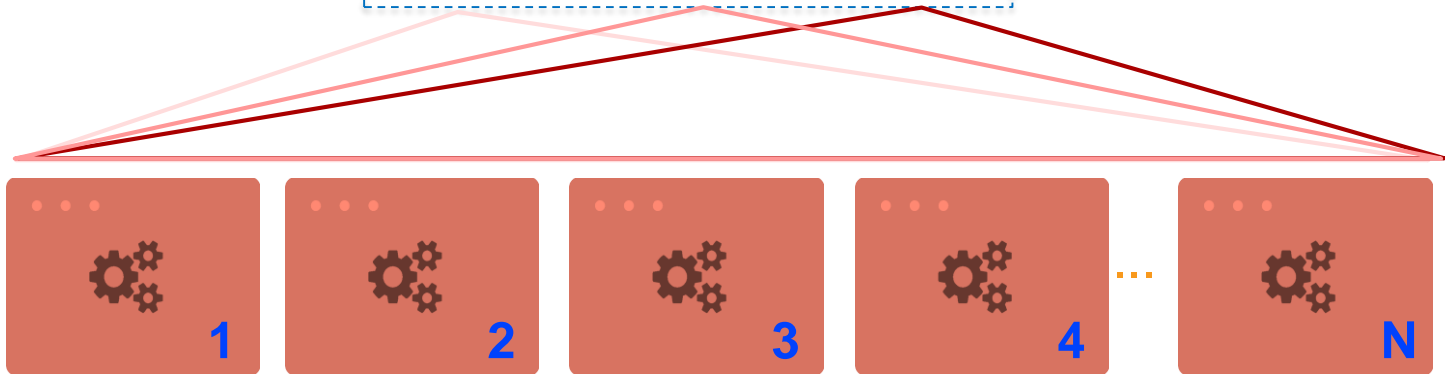
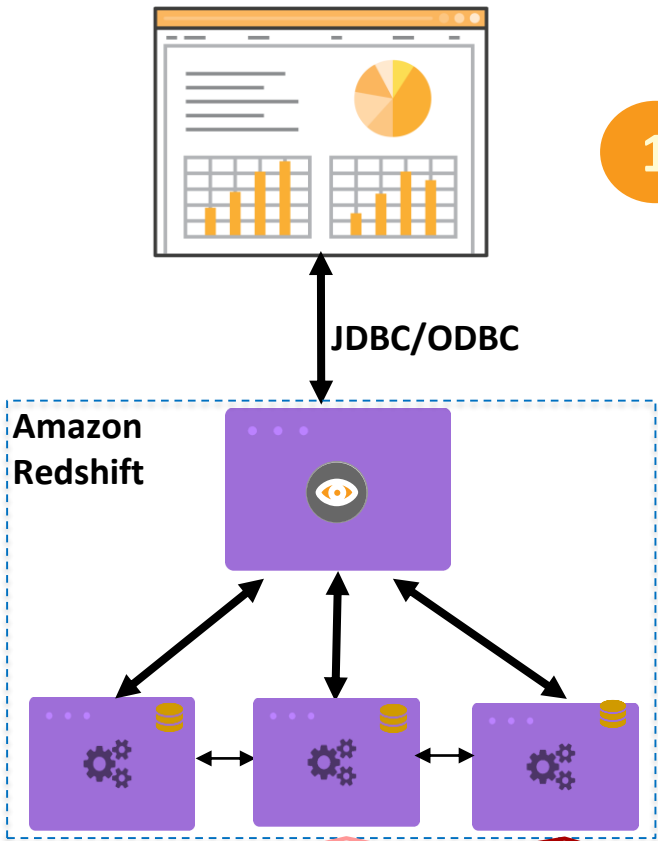
# Example walkthrough

```
Select * from ext.yellow where ...
```

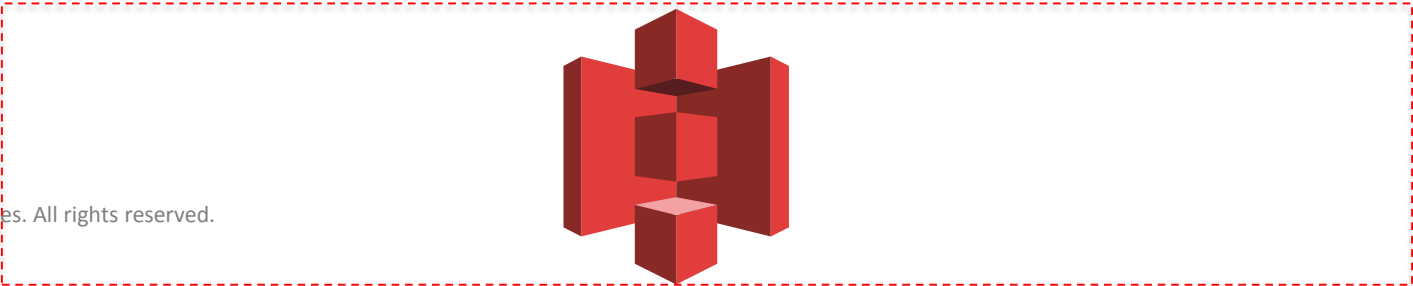
```
Select * from ext.yellow y  
Inner join redshift_native_table r on y.id = r.id  
where...
```

# Life of a query

1 Query  
select \*  
from ext.yellow  
where ...



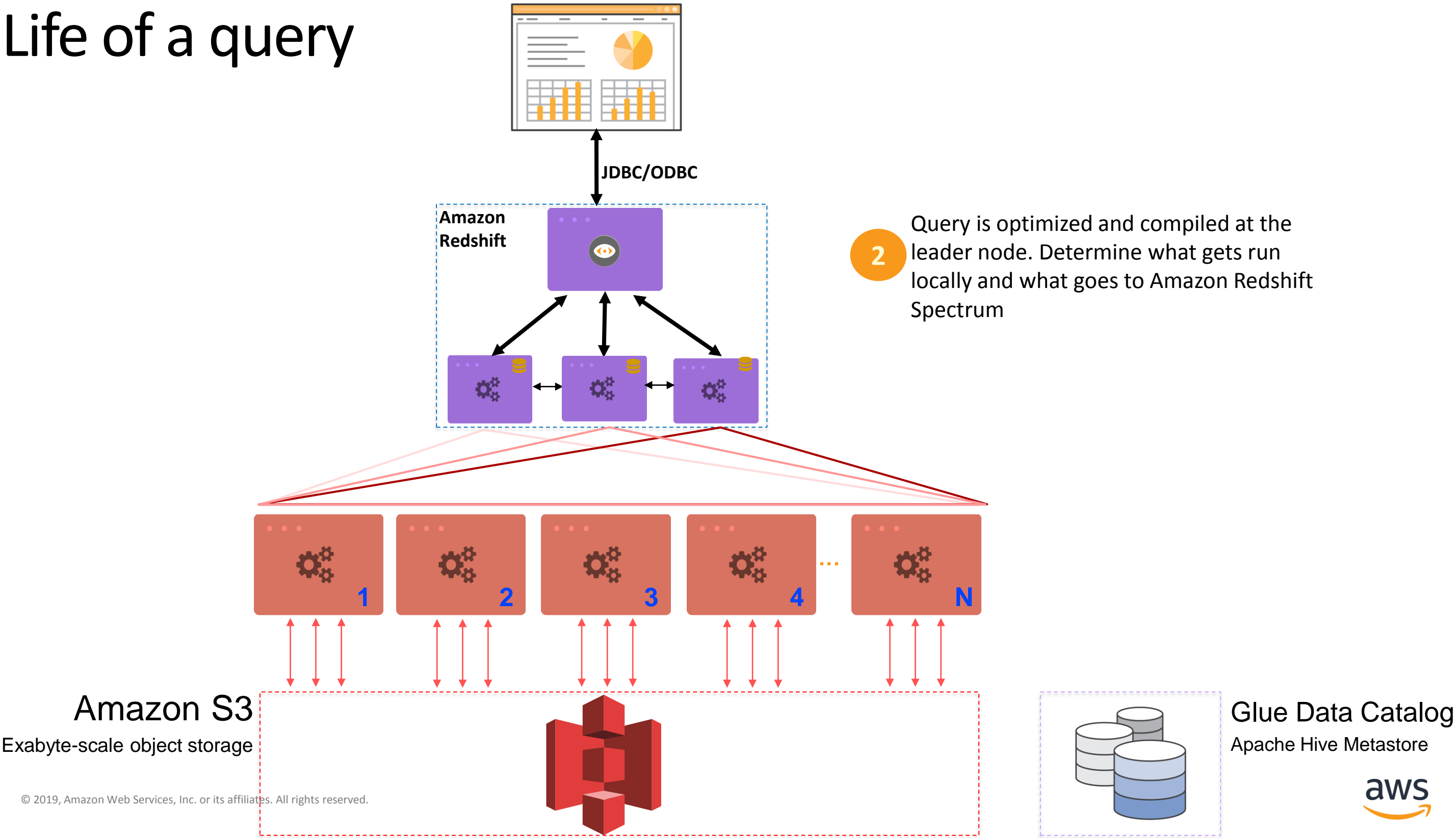
Amazon S3  
Exabyte-scale object storage



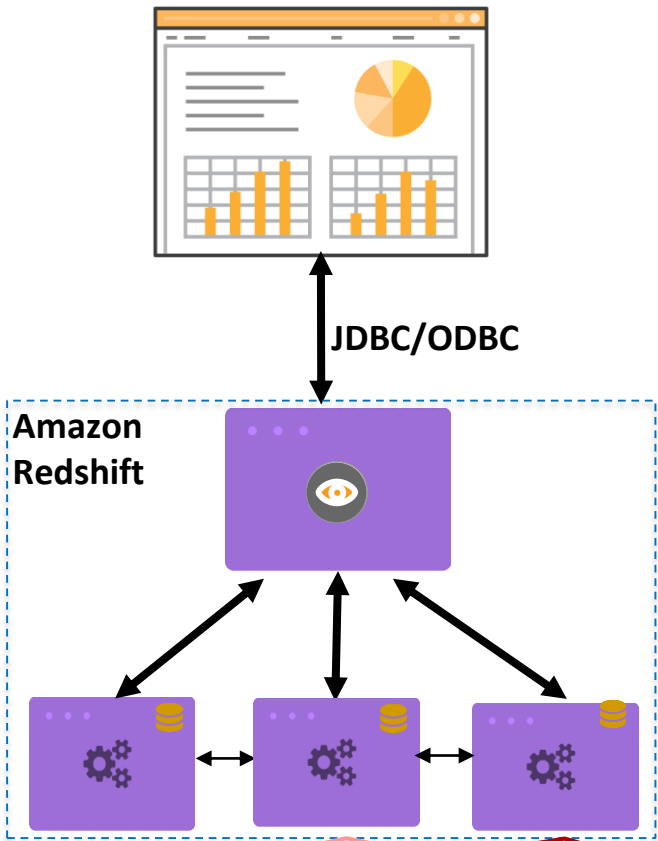
Glue Data Catalog  
Apache Hive Metastore



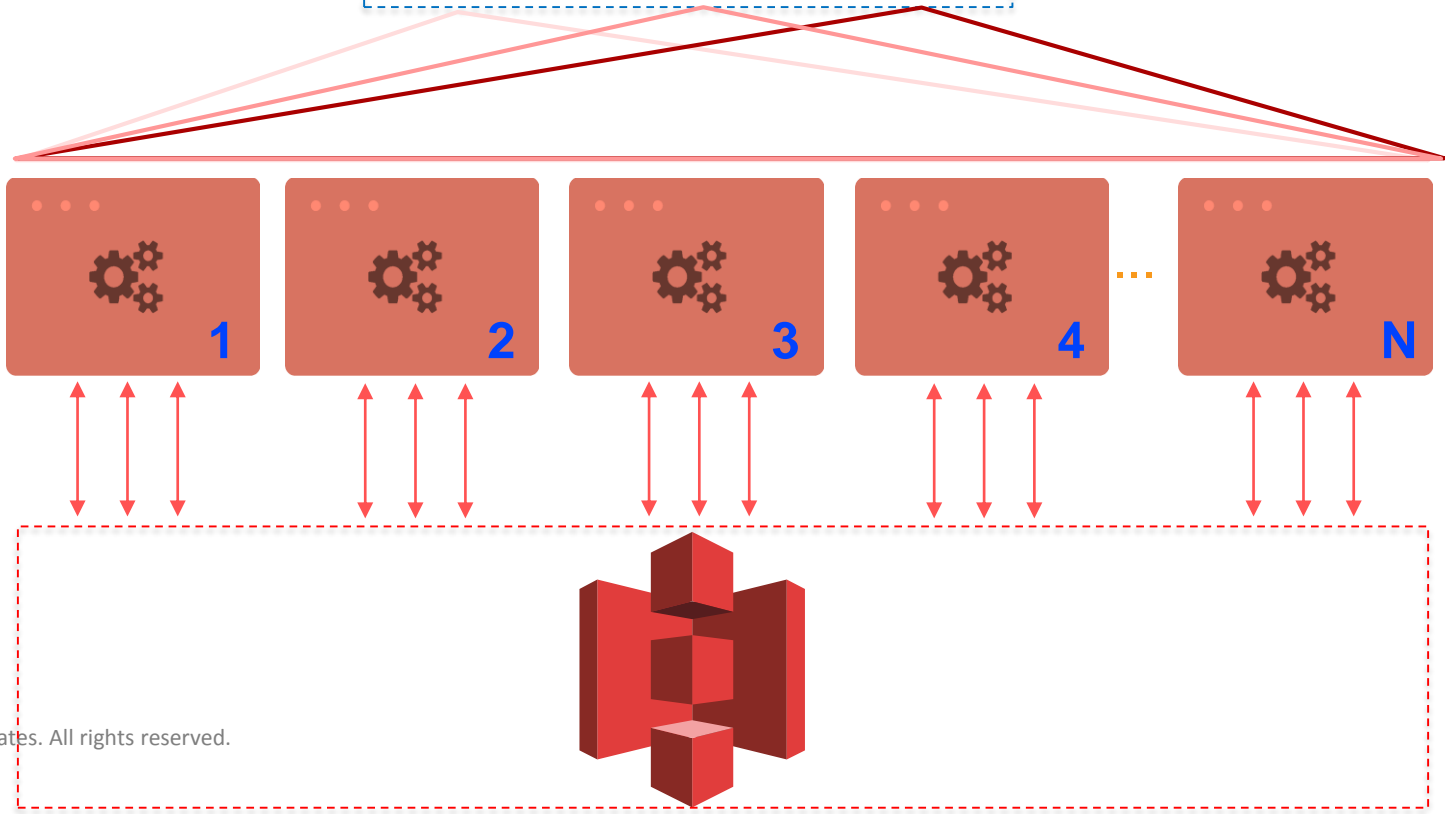
# Life of a query



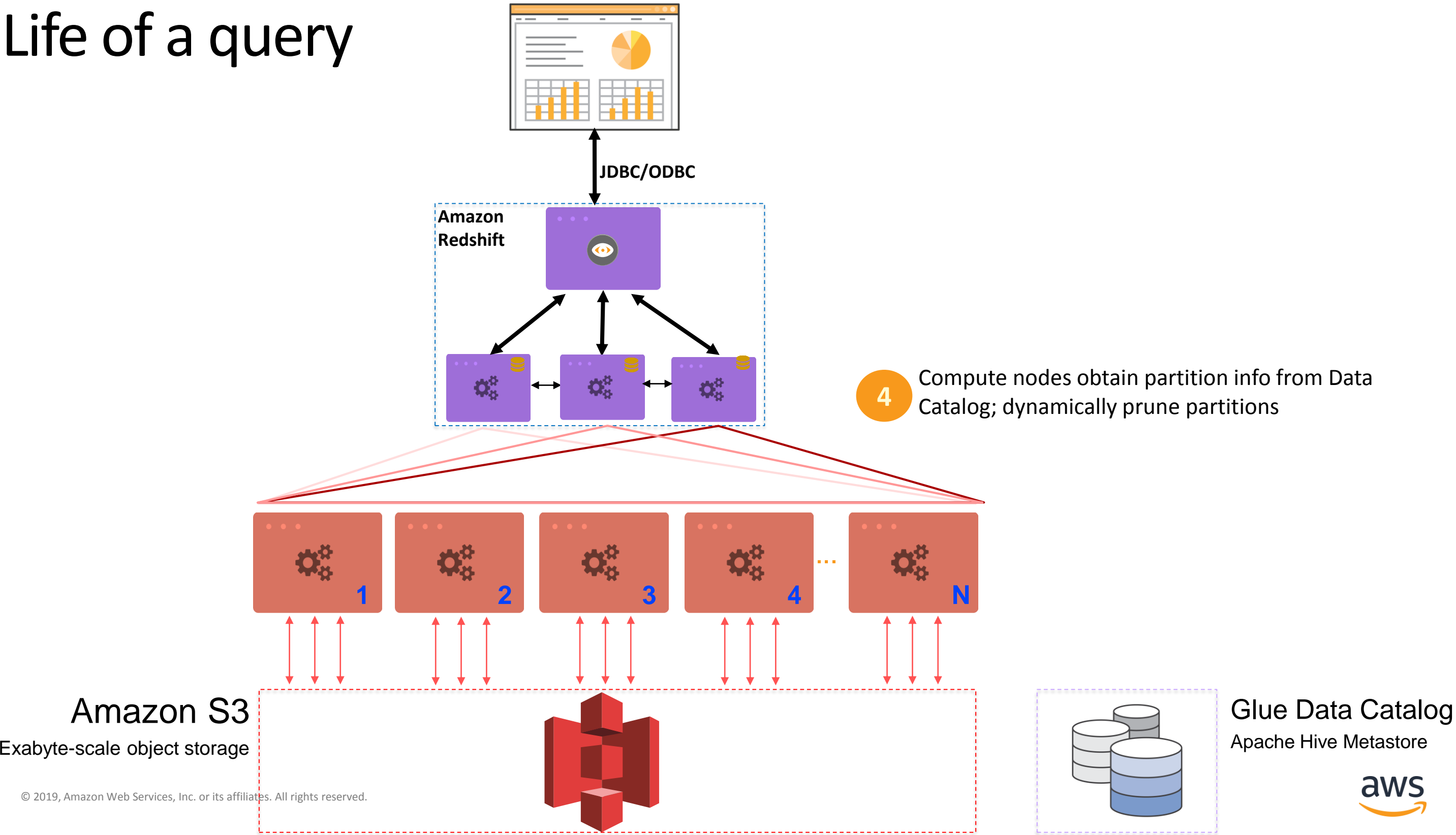
# Life of a query



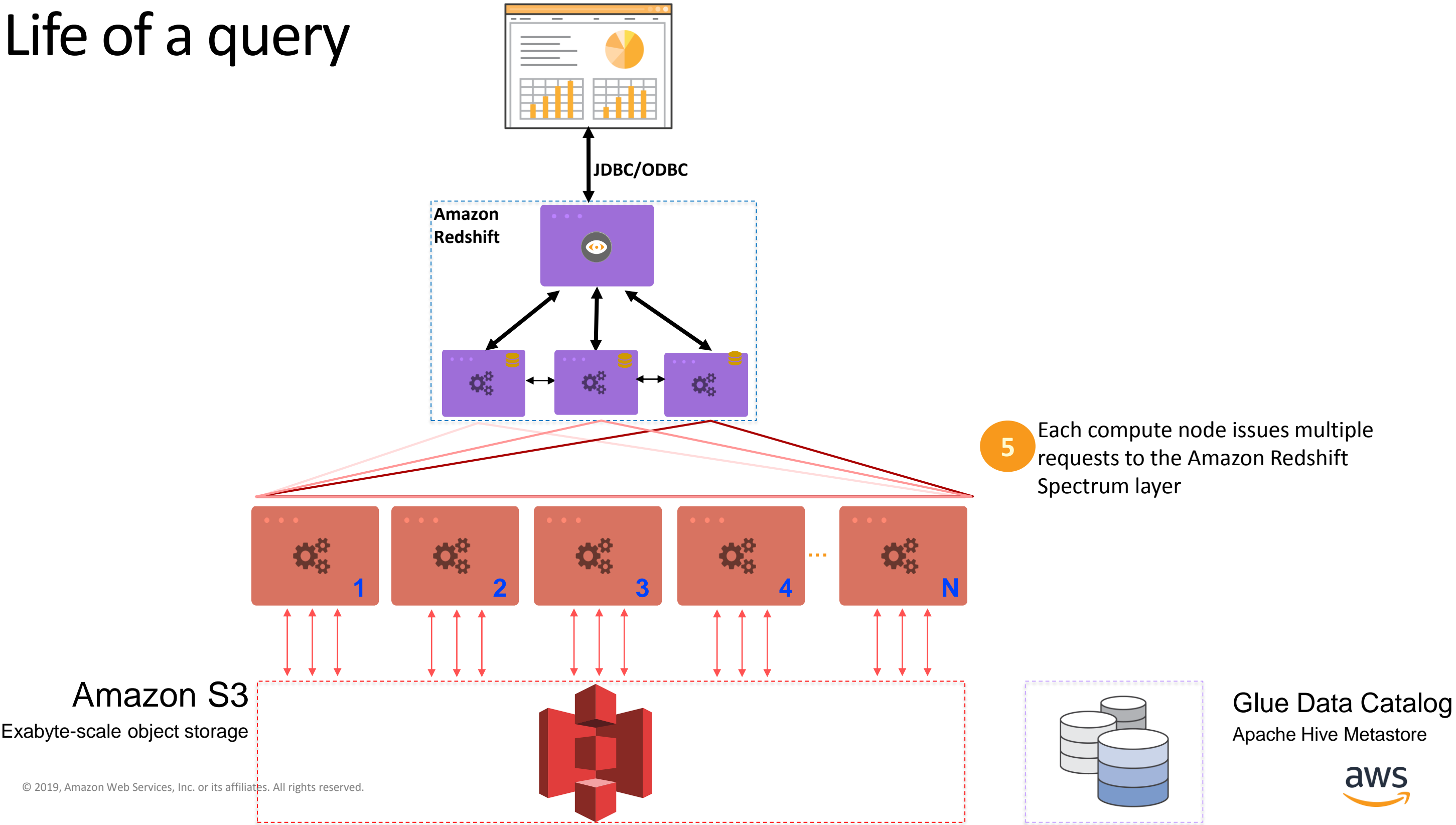
3 Query plan is sent to all compute nodes



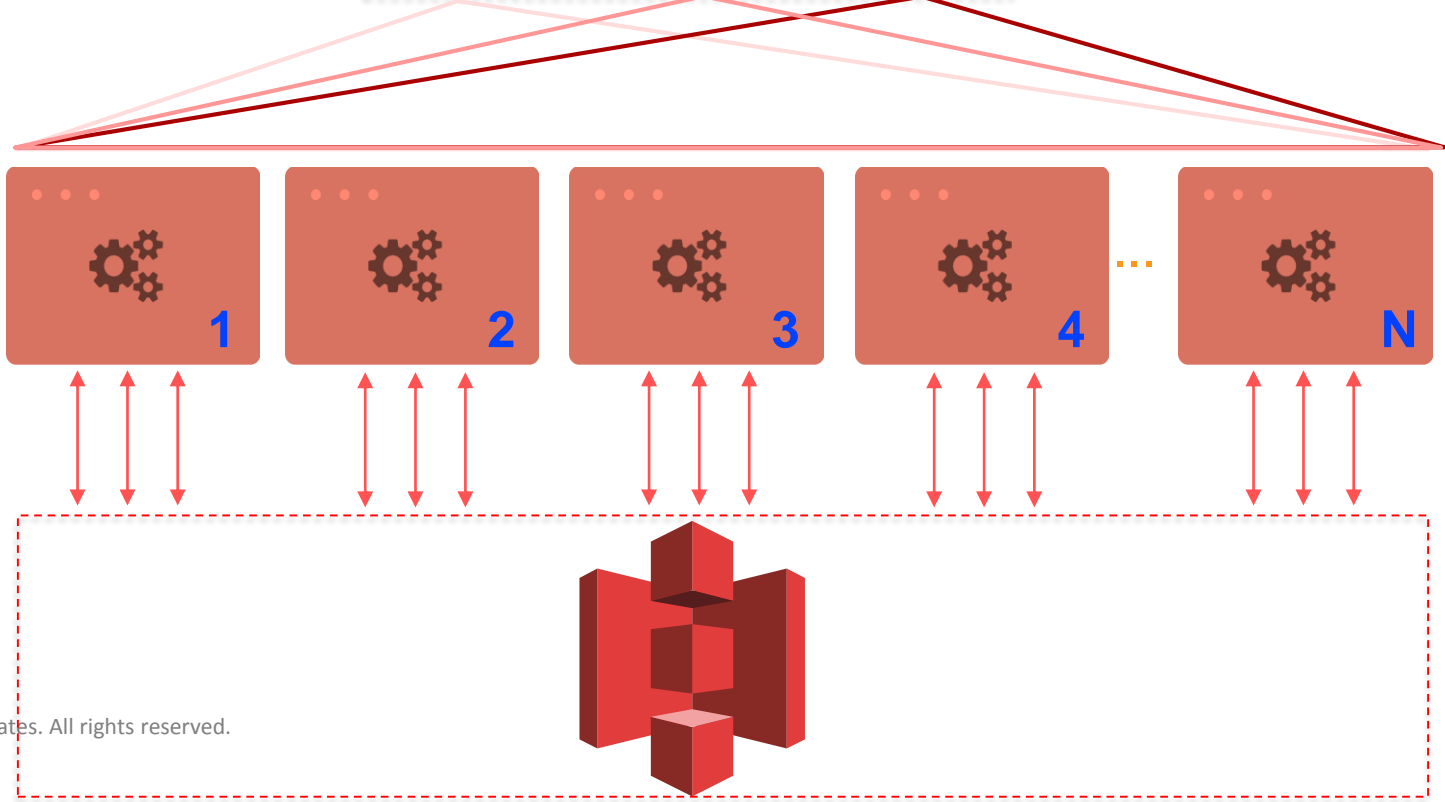
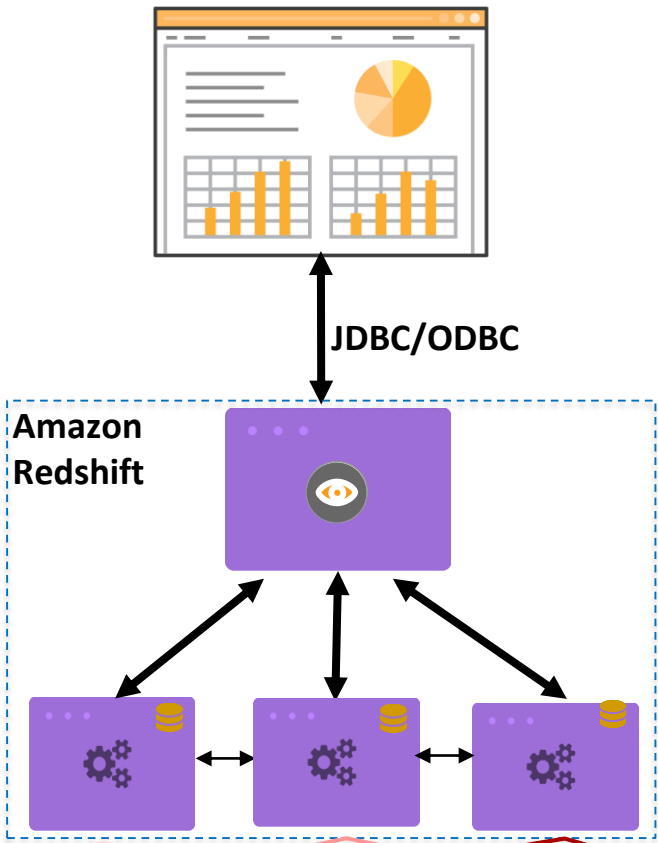
# Life of a query



# Life of a query



# Life of a query



6 Amazon Redshift Spectrum nodes scan your S3 data



Glue Data Catalog  
Apache Hive Metastore

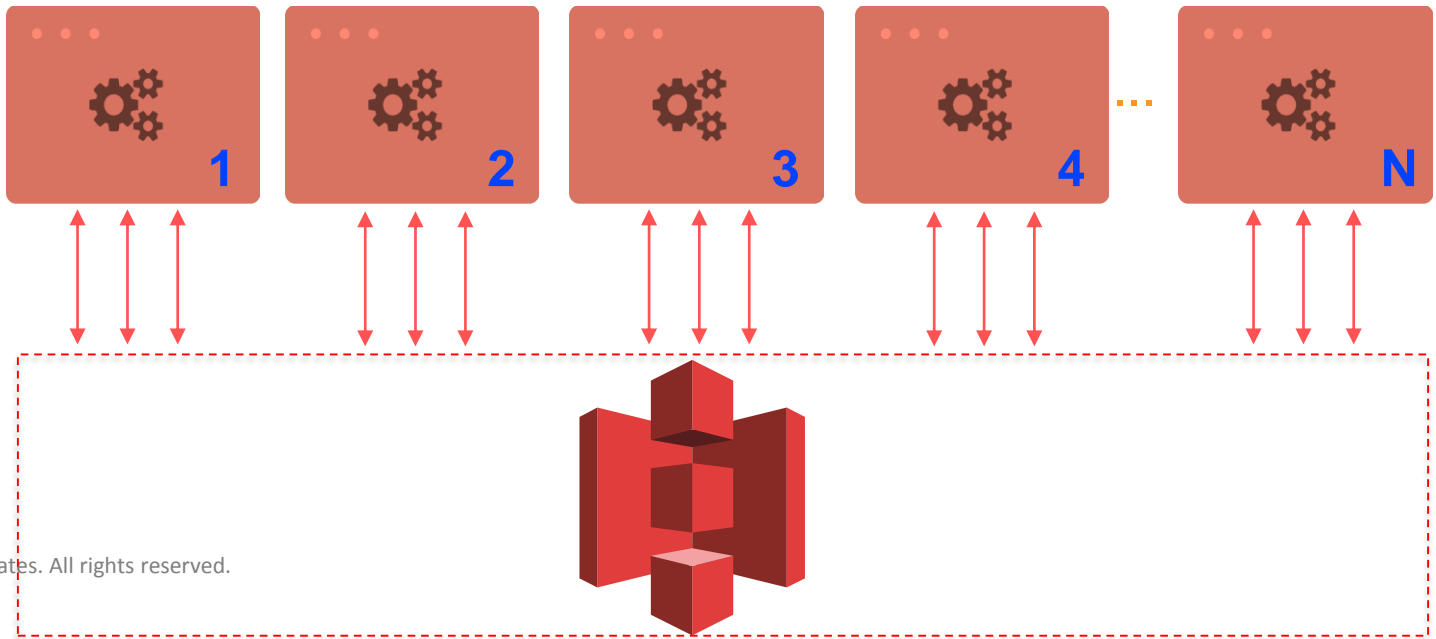
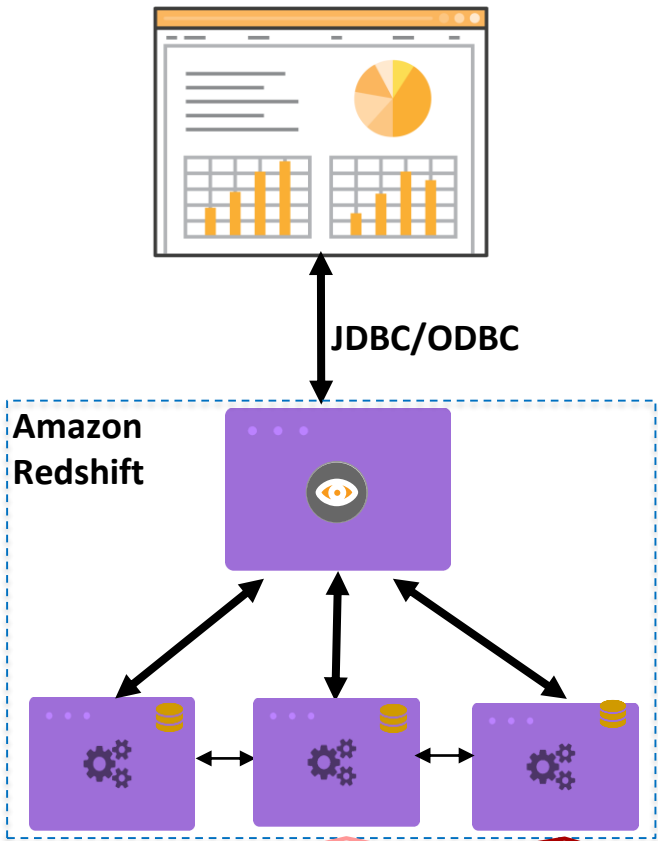


Amazon S3  
Exabyte-scale object storage



# Life of a query

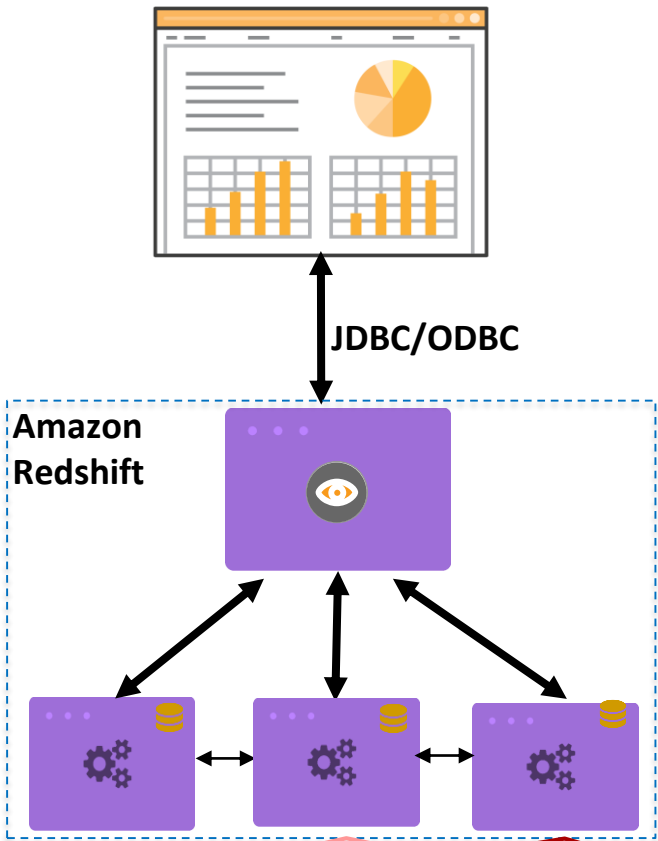
7 Amazon Redshift  
Spectrum projects, filters,  
and aggregates



Glue Data Catalog  
Apache Hive Metastore

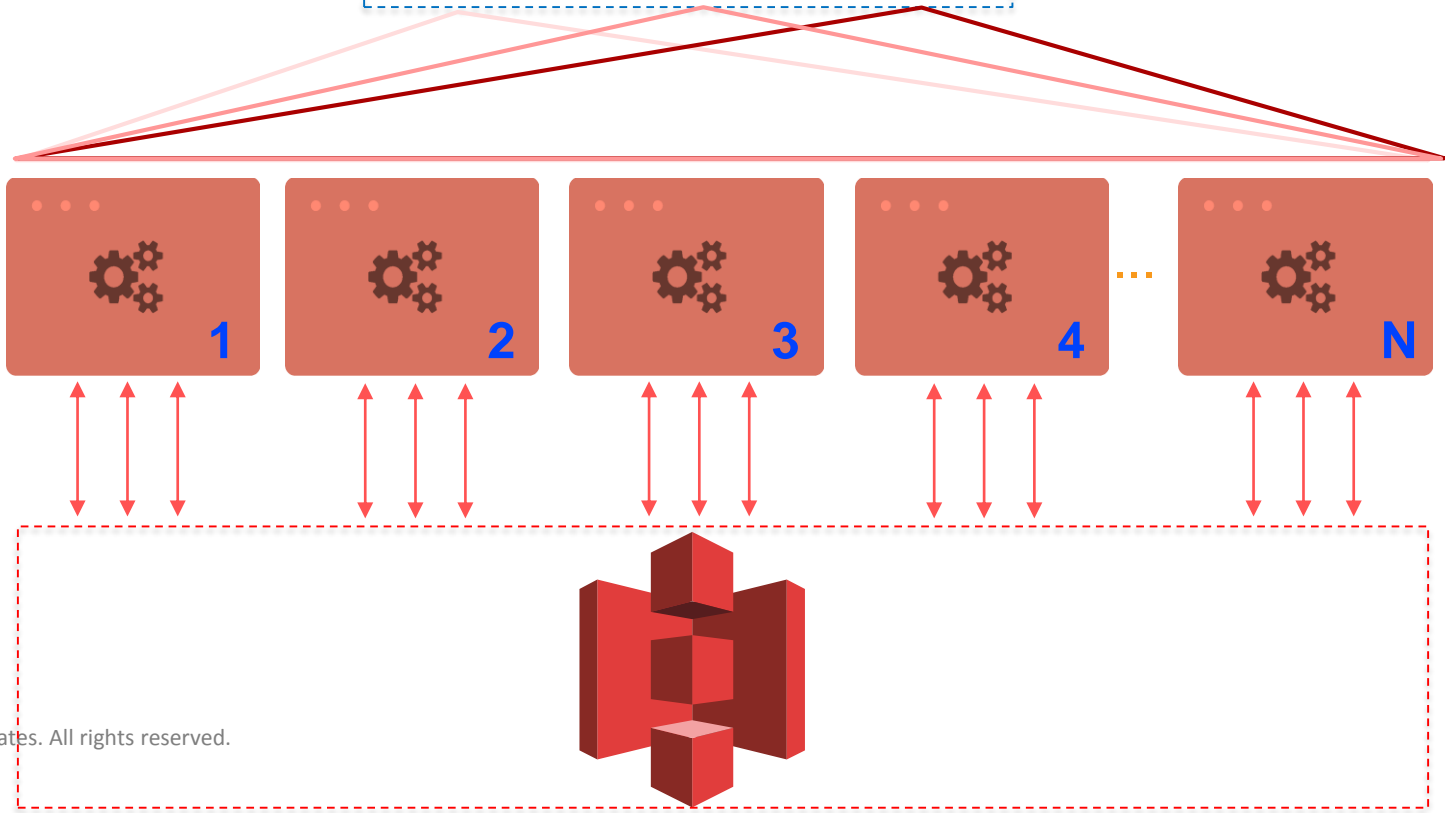


# Life of a query



8

Final aggregations and joins with local Amazon Redshift tables done in-cluster



Amazon S3  
Exabyte-scale object storage

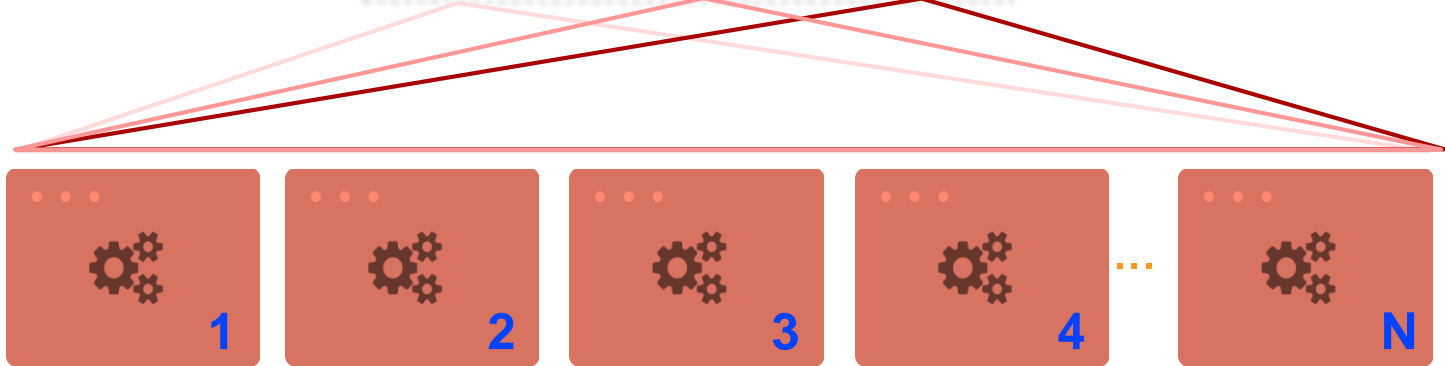
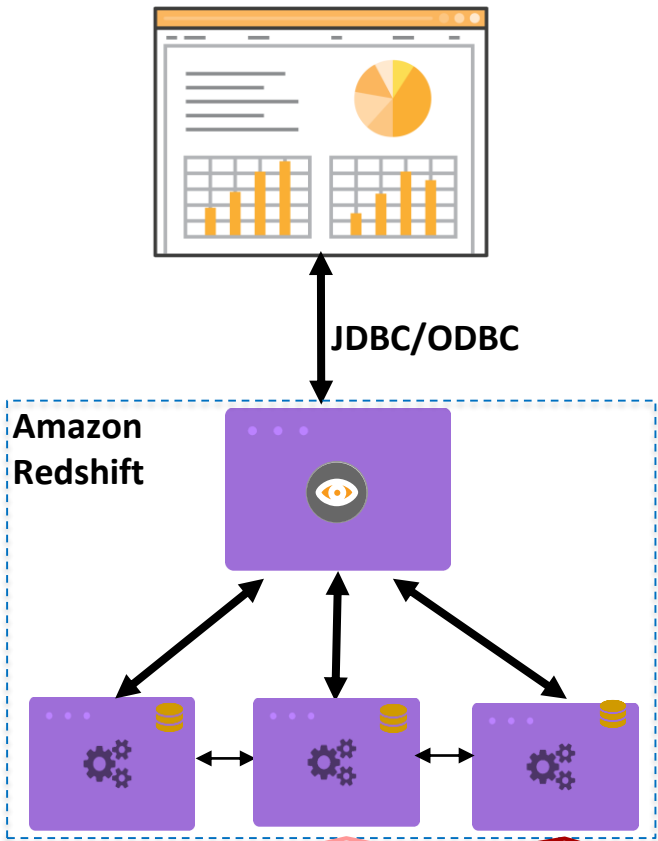


Glue Data Catalog  
Apache Hive Metastore



# Life of a query

9 Result is sent back to client



Amazon S3  
Exabyte-scale object storage



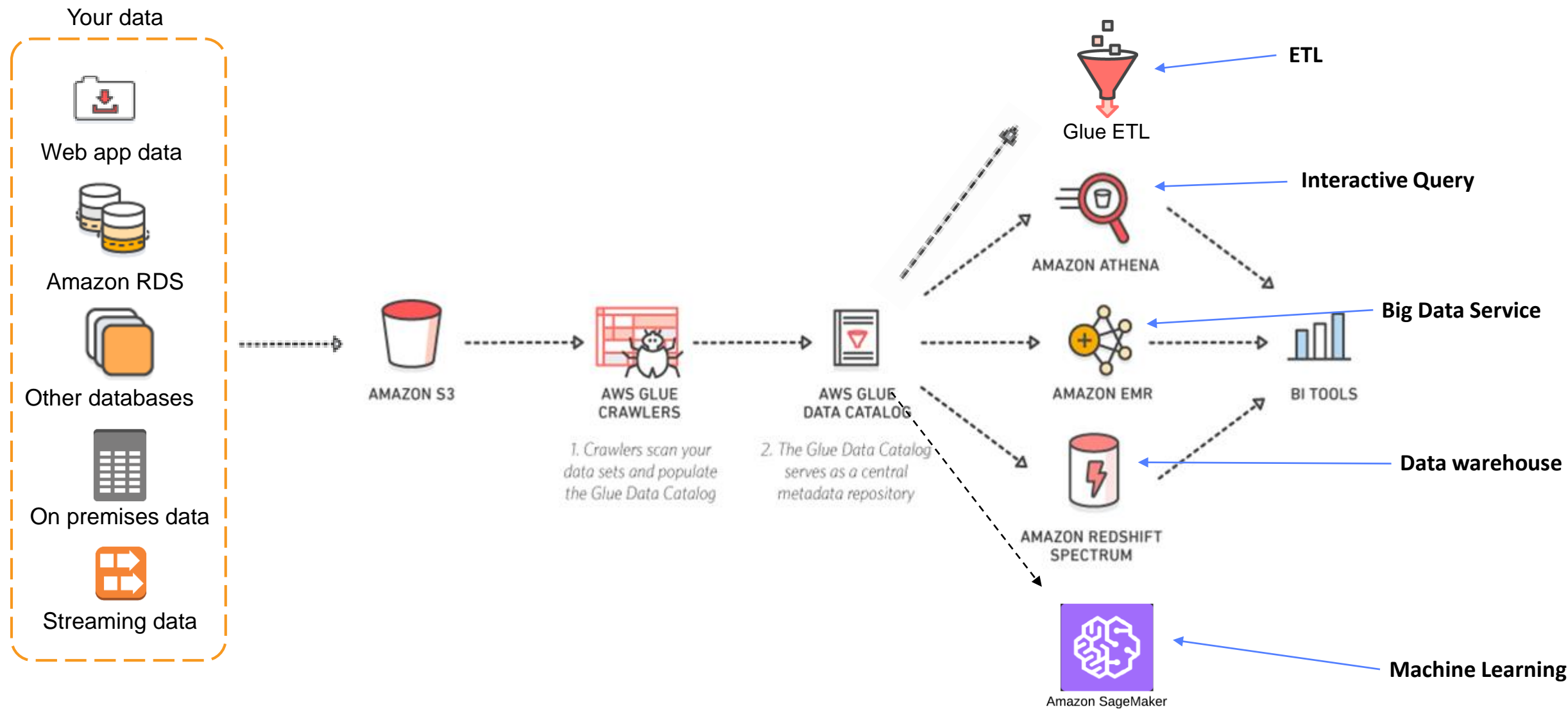
Glue Data Catalog  
Apache Hive Metastore



# Use cases for Spectrum

- Real time data analytics
  - Data is put on S3 from a streaming solution and queried through Redshift Spectrum
- Data archival solutions
  - Data is stored on S3 and moved to Redshift on a need-to-basis
  - Data is offloaded from Redshift to S3 and queried using Redshift Spectrum
- Cost control
  - Only 10% of the data in a data warehouse is actually queried. Use Redshift Spectrum to offload data to S3
- ETL
  - Use Redshift Spectrum to query and ETL data into Redshift

# Data Lakes on S3



# Thank you!