AWS Solutions Architect Associate

Session 302

Analytics & Database: Glue, Athena & Redshift Spectrum

July/2024

SERVERLESS !!!

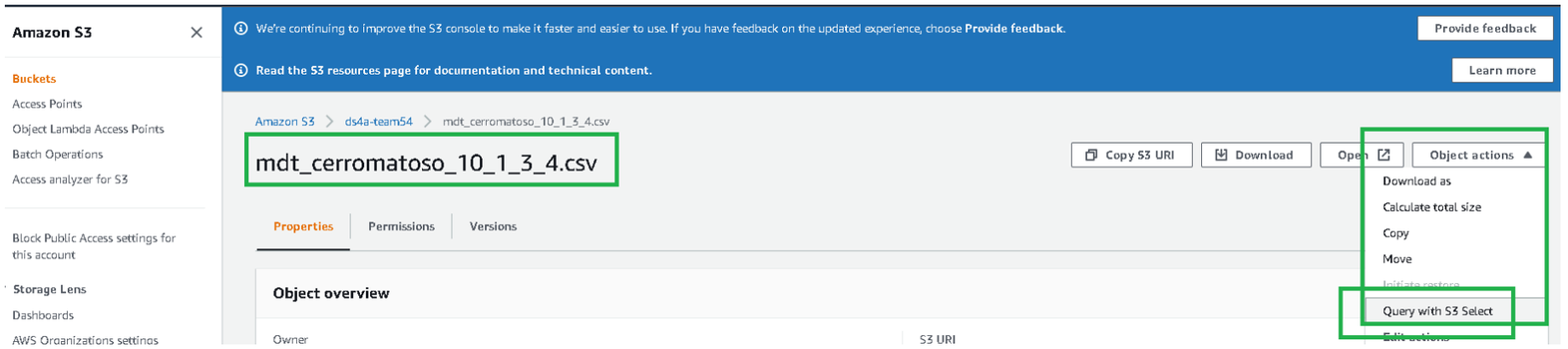
S3 Select

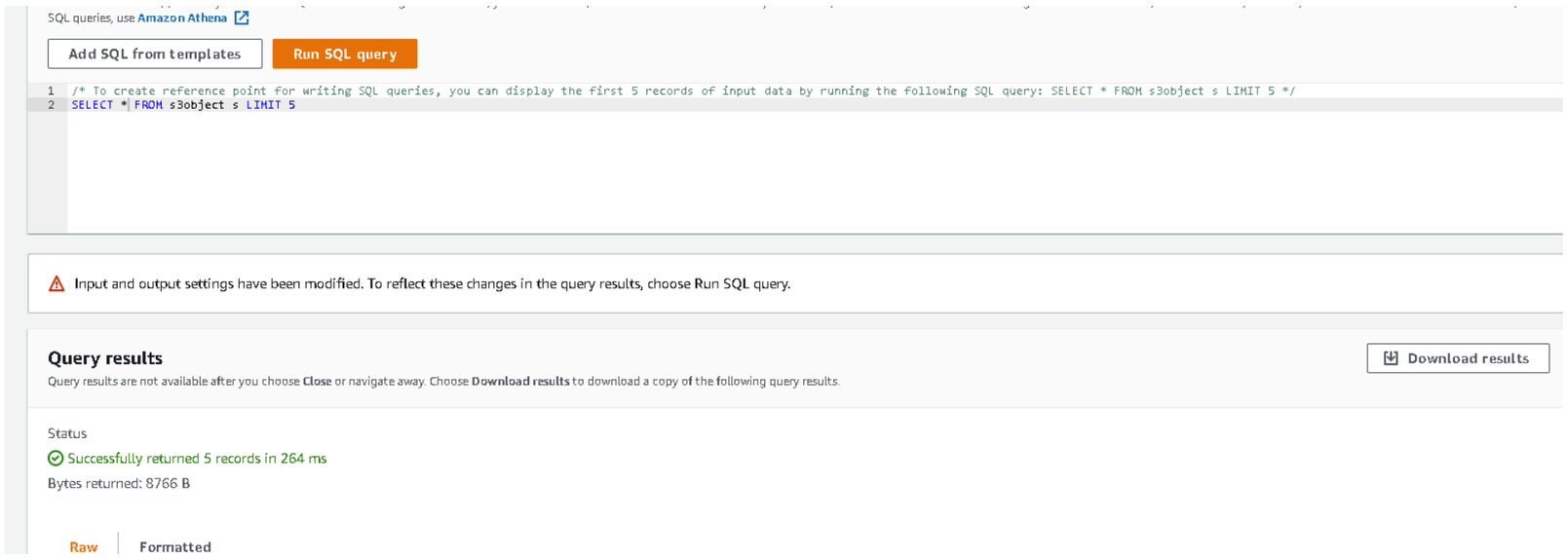
Using S3 Select Feature from S3, you can query S3 File without downloading or download file, increase efficiency.
Don't need compute service. Run on S3 on its conditions, encryption.
Very Simple.

Input: CSV, JSON or Parquet.
Input file coding: UTF-8
Input Compression: None, GZIP, BZIP2. Parquet doesn't apply.
Output: CSV, JSON

Not works for pure Big Bata because it's only work on one file.

Adv: Not parsing, apply on one file, works "on the fly", without compute unit.





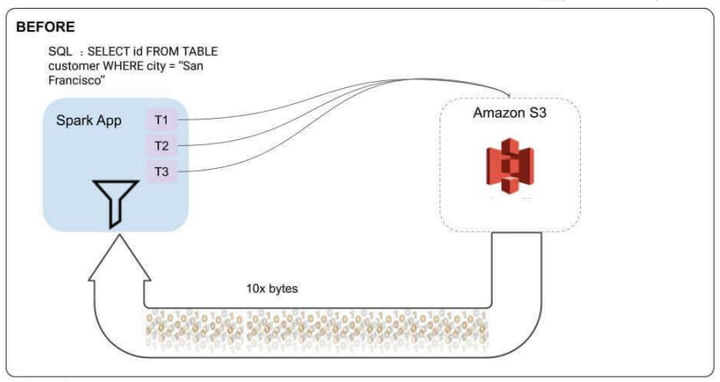


Figure 1

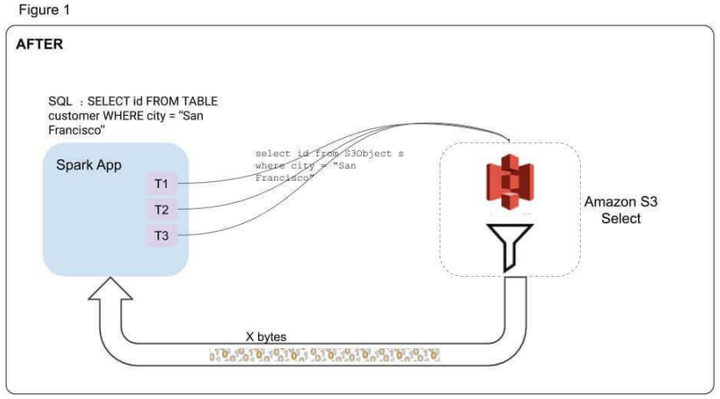


Figure 2

AWS Glue

ETL – Extract Transform and Load >>> **DATA INTEGRATION**
Data Sources: S3 (**semistructured**), DB: RDS, NoSQL: DynamoDB.
Store on target Data Stores.
Discover/Infere the data schema using classifiers.

Tool to Extract: AWS Glue Crawler.

Data Sources

AWS Glue supports the following data sources:

- Data stores
  - Amazon S3
  - Amazon Relational Database Service (Amazon RDS)
  - Third-party JDBC-accessible databases
  - Amazon DynamoDB
  - MongoDB and Amazon DocumentDB (with MongoDB compatibility)
- Data streams
  - Amazon Kinesis Data Streams
  - Apache Kafka

Data Targets

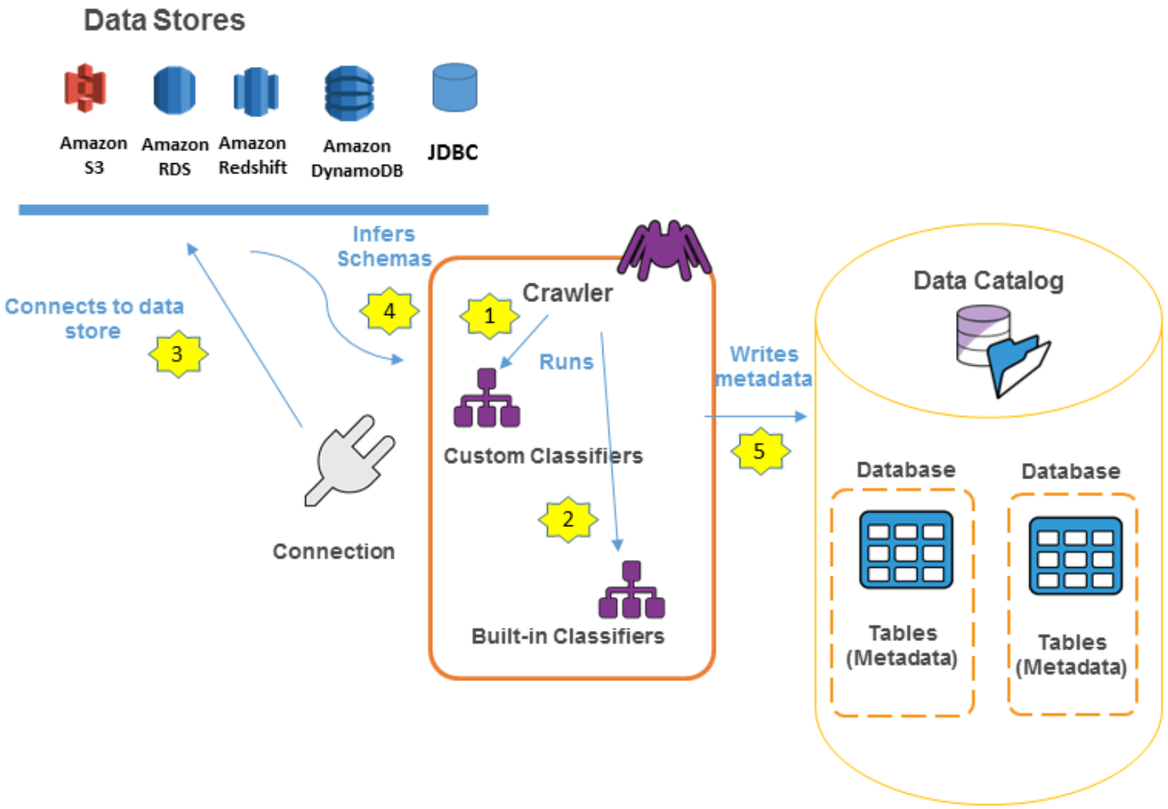
AWS Glue supports the following data targets:

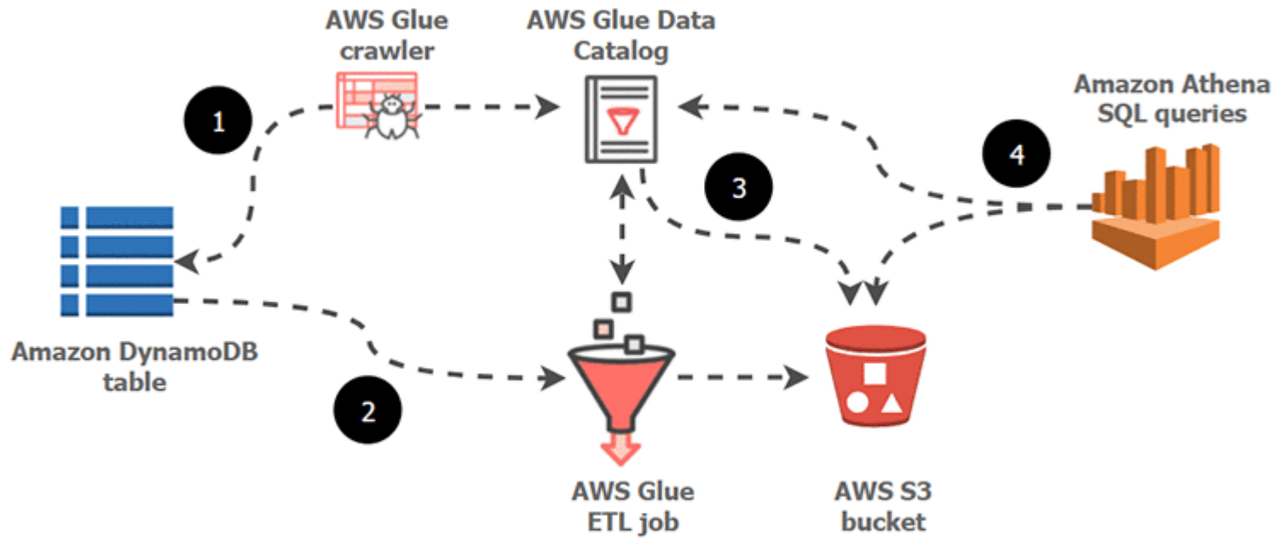
- Amazon S3
- Amazon Relational Database Service (Amazon RDS)
- Third-party JDBC-accessible databases
- MongoDB and Amazon DocumentDB (with MongoDB compatibility)

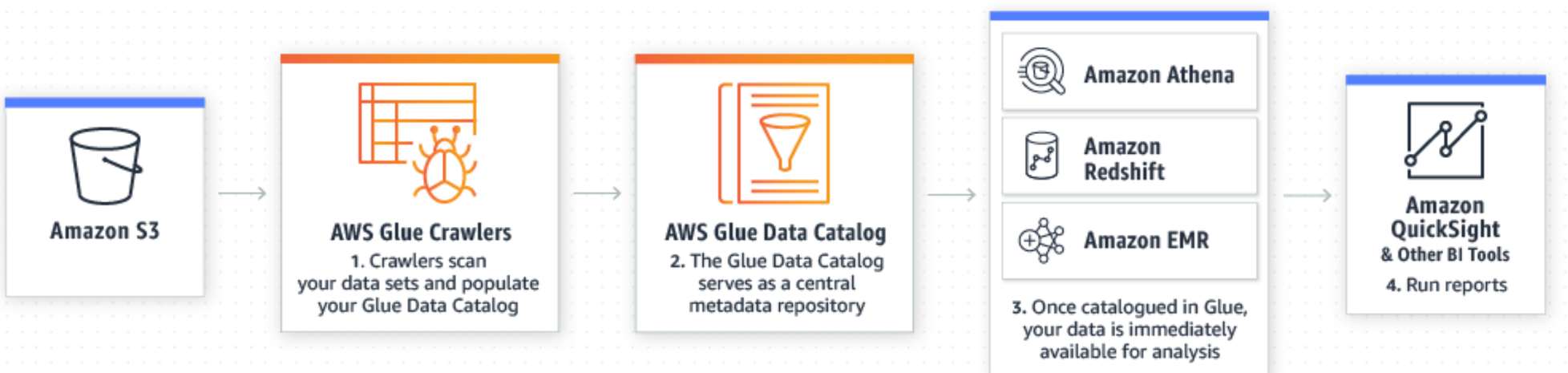
Access type that crawler uses	Data stores
Native client	• Amazon Simple Storage Service (Amazon S3) • Amazon DynamoDB
JDBC	Within Amazon Relational Database Service (Amazon RDS) or external to Amazon RDS: • Amazon Aurora • MariaDB • Microsoft SQL Server • MySQL • Oracle • PostgreSQL
MongoDB client	• MongoDB • Amazon DocumentDB (with MongoDB compatibility)

Taken from https://docs.aws.amazon.com/glue/latest/dg/how-it-works.html (12/06/2021)

Glue – Main Use Cases



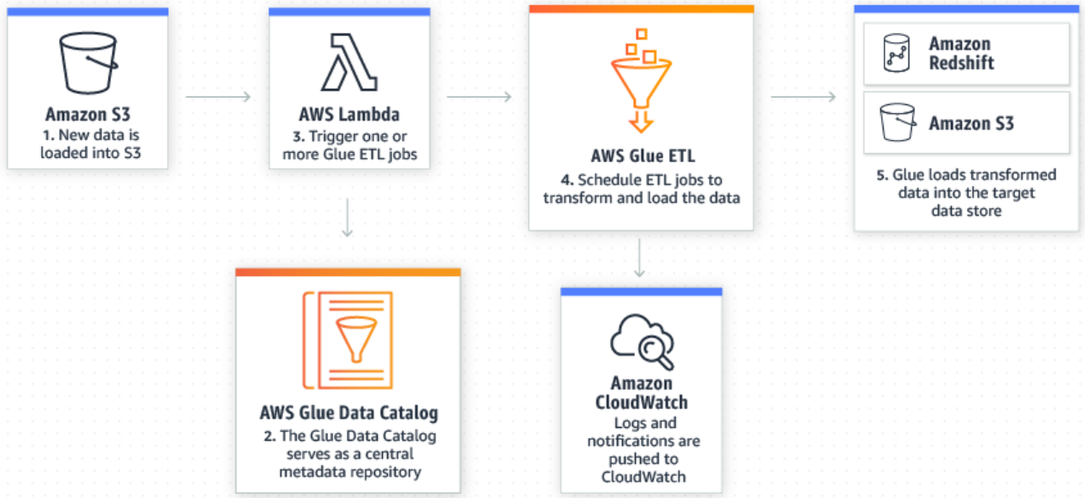




Taken from https://docs.aws.amazon.com/glue/latest/dg/populate-data-catalog.html and https://aws.amazon.com/blogs/database/simplify-amazon-dynamodb-data-extraction-and-analysis-by-using-aws-glue-and-amazon-athena/ (12/06/2021)

Glue – Use Cases

Build event-driven ETL (extract, transform, and load) pipelines



Create a unified catalog to find data across multiple data stores

Create, run, and monitor ETL jobs without coding

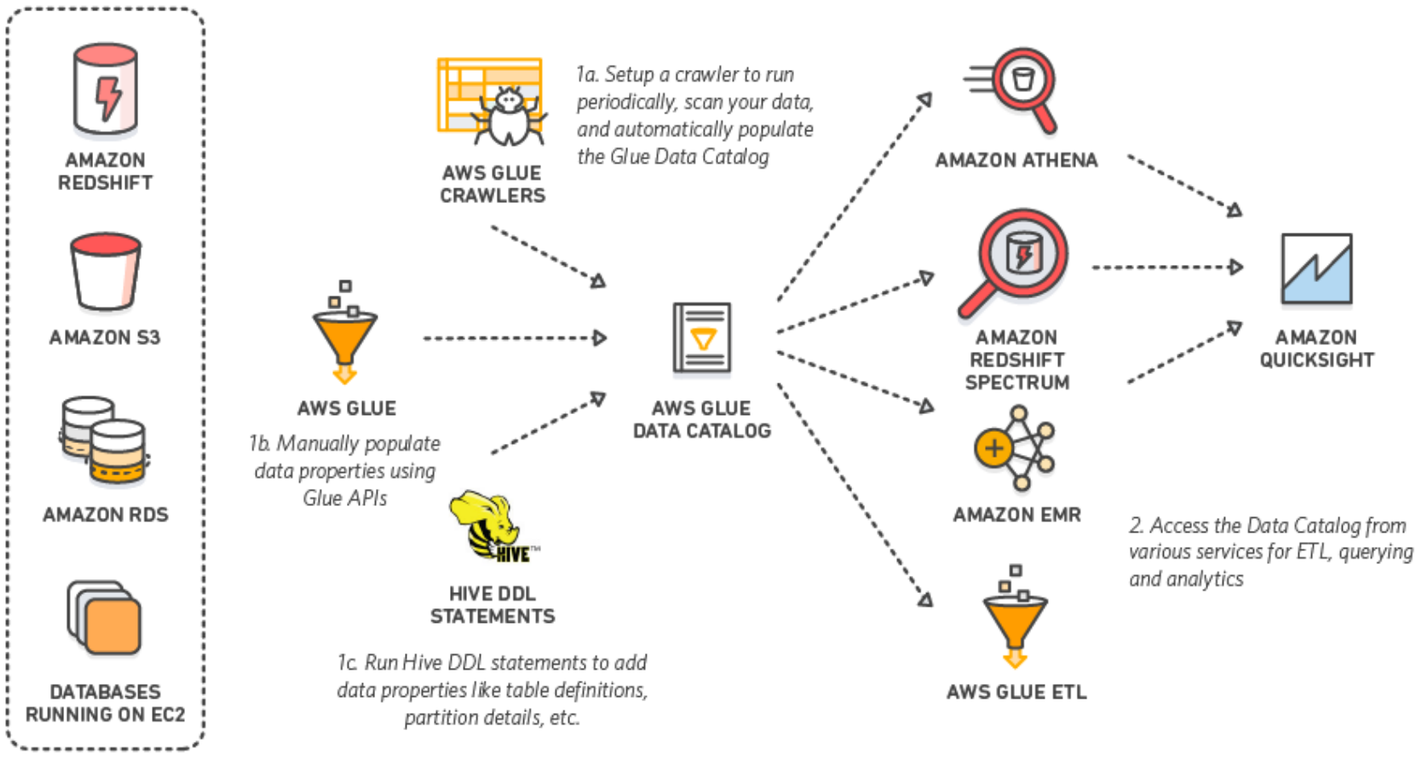
Other 2 common use cases: Explore data with self-service visual data preparation (AWS Glue DataBrew), Build materialized views to combine and replicate data (Glue Elastic Views in Beta).

Taken from https://aws.amazon.com/glue/?whats-new-cards.sort-by=item.additionalFields.postDateTime&whats-new-cards.sort-order=desc (12/06/2021)

AWS Athena

Query (ANSI SQL) service using AWS Glue Data Catalog as schema to several data sources. For native using S3 (Read and Write) and for Federated Query using Lambda Connectors: RDS, DynamoDB, Hbase, Cloudwatch Logs, etc.

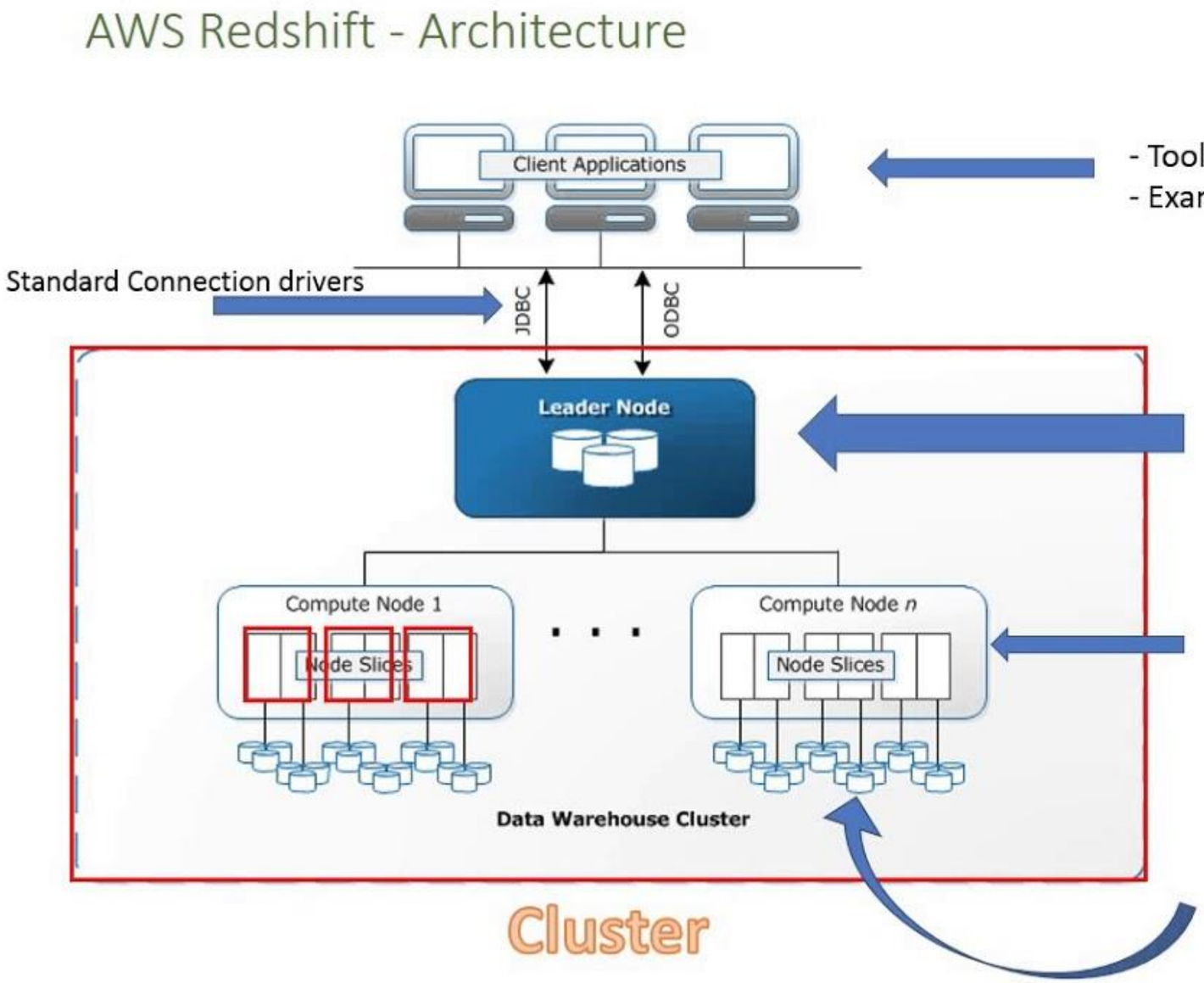
Distributed query, based on Presto.
S3 Formats: CSV, JSON, Avro, Parquet, ORC.
Serverless and Pay per Query Service (U$5 per TB Scanned). Using compressed format is more cheaper.



Taken from https://aws.amazon.com/athena/features/?nc=sn&loc=2 and https://youtu.be/Gn7lxQiSZPQ (12/06/2021)

Redshift - Overview

Redshift is Data Warehouse (DWH) preferred by AWS. Exabyte Scale for multiples data sources: DB, DataLakes, etc.



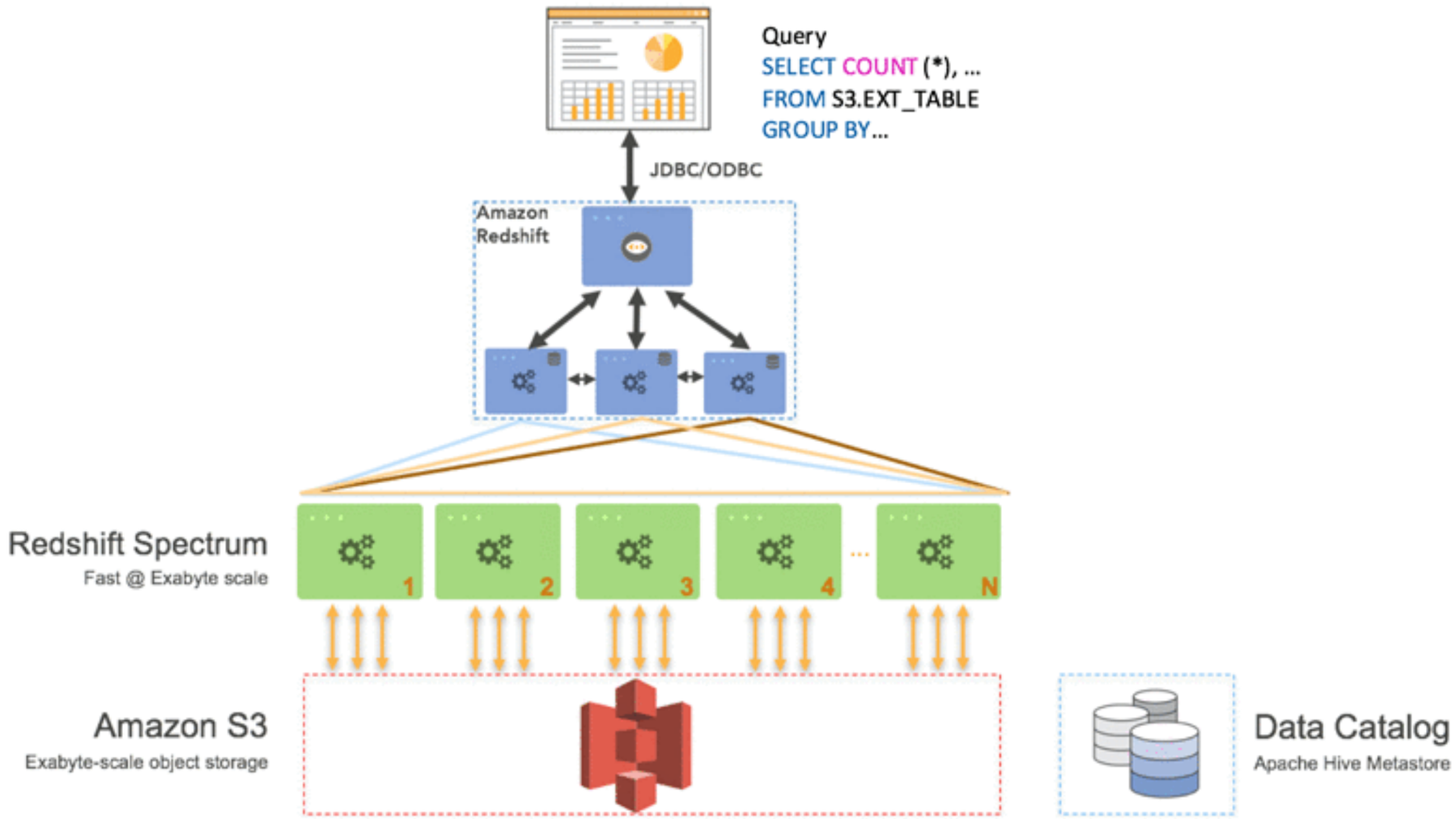
- Tools used for Interaction with Redshift
- Example: SQL Workbench, Dbeaver, ETL tool

- Communication with Client Application
- Communication with Compute Nodes
- Develops Query Execution Plans
- Distributes parts of query among compute nodes

- Executes the compiled code and send the intermediate results to Leader Node
- Node can be of 2 types:
  1) **Dense Compute (DC)** = SSD Base
  2) **Dense Storage (DS)** = HDD Based
- Processes portion of the workload assigned to the node.
- Inserted data is distributed based on the distribution key defined in table DDL

NO SERVERLESS !!!

Redshift Spectrum



Taken from https://aws.amazon.com/blogs/big-data/amazon-redshift-spectrum-extends-data-warehousing-out-to-exabytes-no-loading-required/ (12/06/2021)

Comparison Tools

The Emerging Analytics Architecture