DATA ANALYTICS ON AWS

MODULE 5 : AWS GLUE, ATHENA, QUICKSIGHT AND ELASTICSEARCH

Learning and Knowledge Management

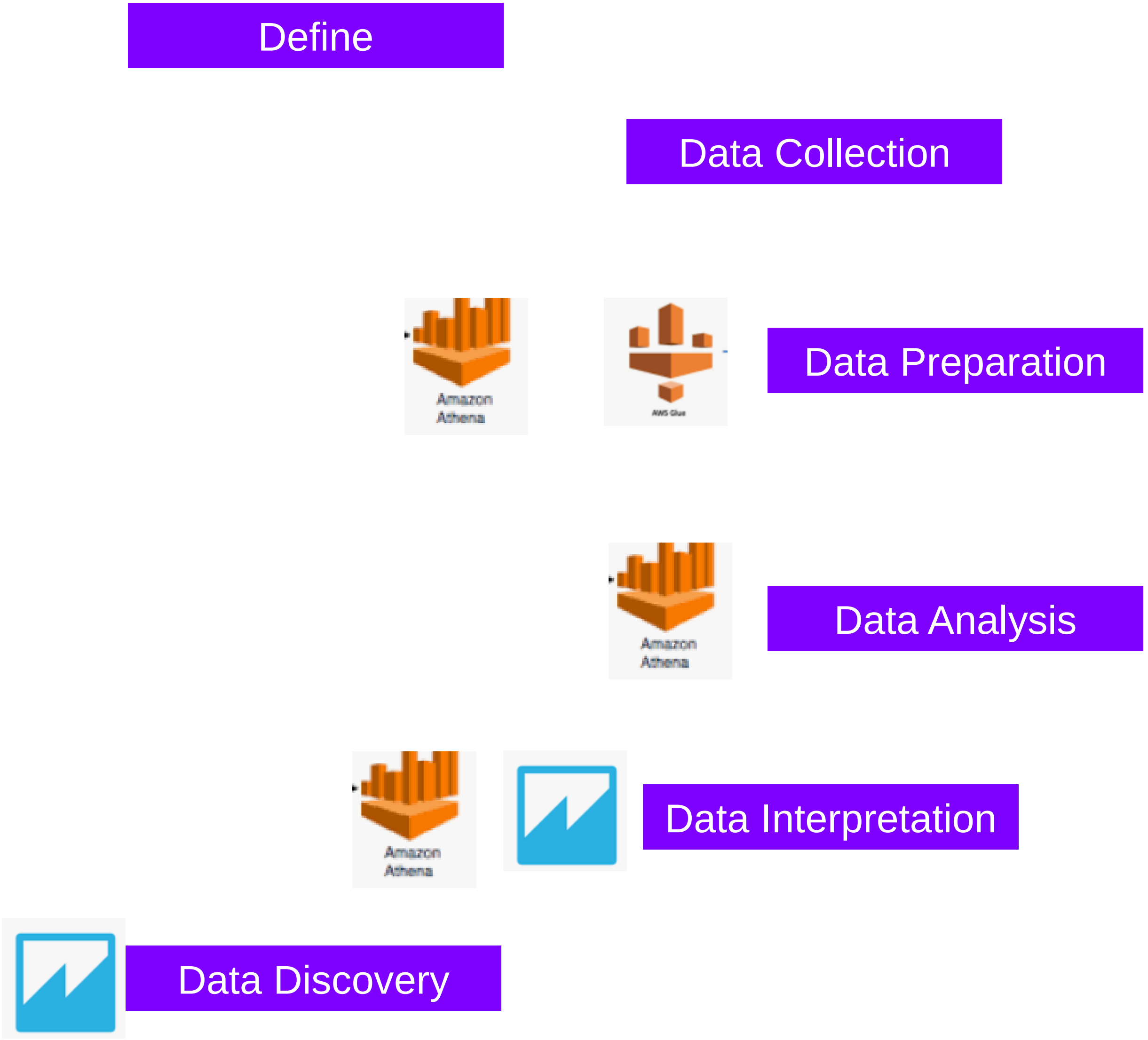
MODULE OBJECTIVES

At the end of this module, you should be able to:

- Explain Glue Data Catalog
- Explain Glue Jobs
- Perform Operations with Glue Jobs
- Explain Job Bookmarks
- Get Started with Athena
- Perform Operations with Athena
- Create Visualizations with QuickSight
- Push data to ElasticSearch and Discover data

AWS Glue Data Catalog

OVERVIEW OF AWS GLUE



Define

Data Collection

Data Preparation

Data Analysis

Data Interpretation

Data Discovery

AWS GLUE

What is AWS Glue?

.

AWS Glue is a **Serverless ETL service**

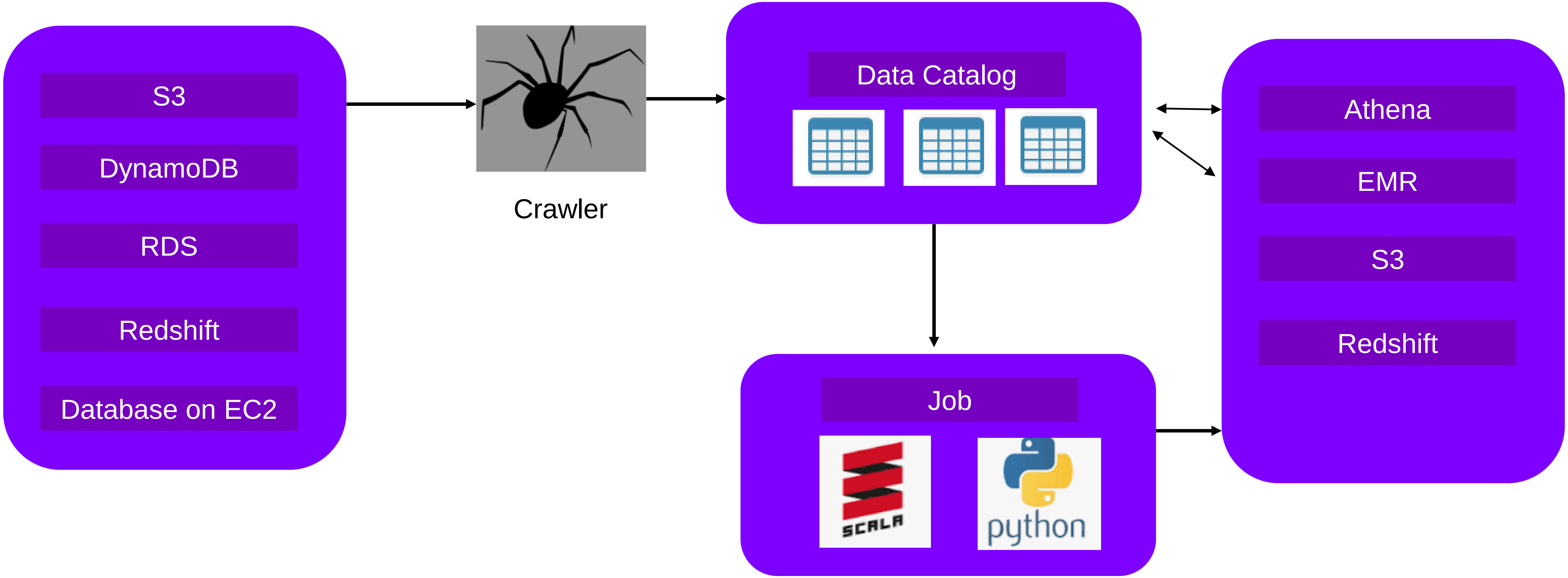
AWS GLUE

Use Cases

- Query Data in S3
- Joining data for Data Warehouse
- Creating a Centralized Data Catalog

AWS GLUE

AWS Glue Components



S3

DynamoDB

RDS

Redshift

Database on EC2

Crawler

Data Catalog

Job

Athena

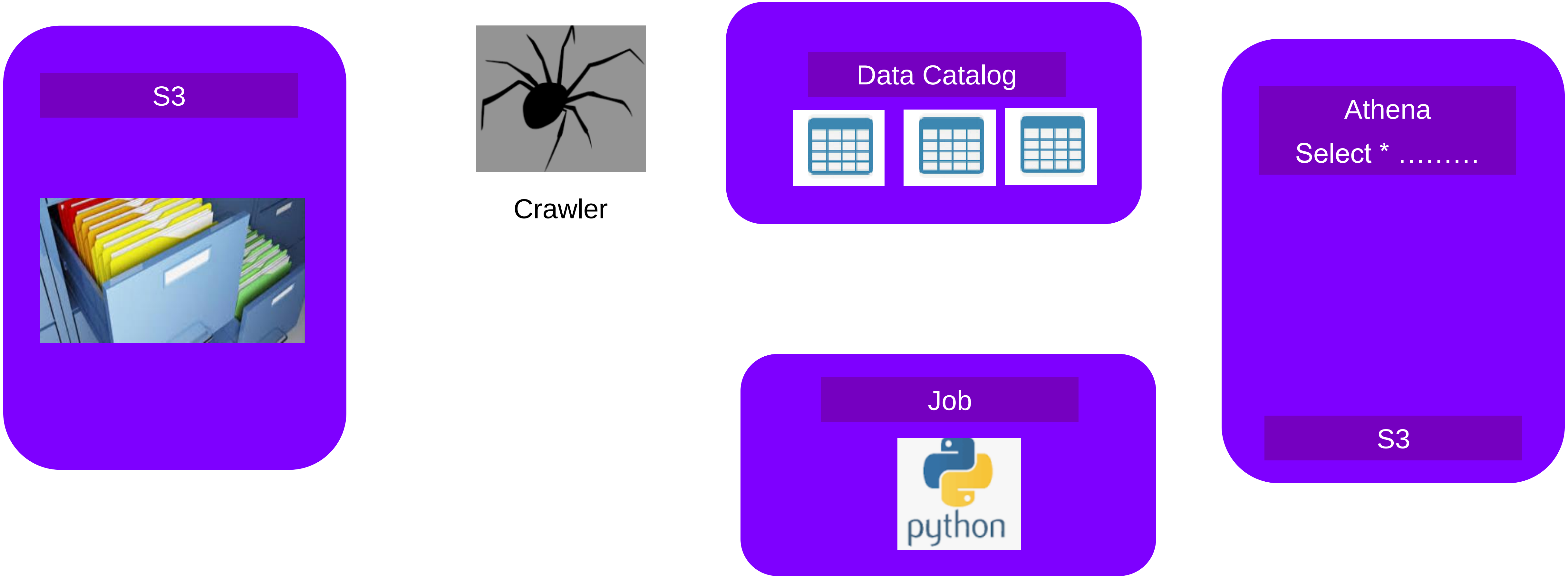
EMR

S3

Redshift

AWS GLUE

AWS Glue Components



S3

Crawler

Data Catalog

Athena

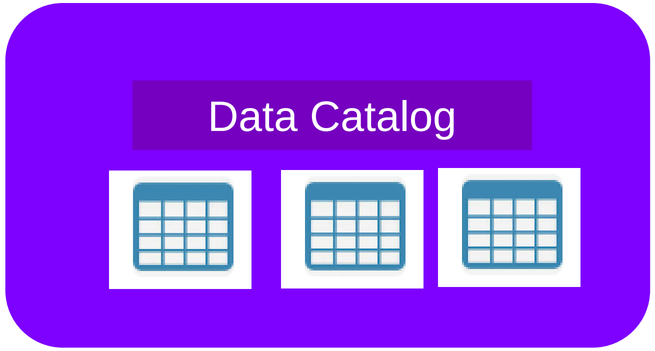
Select * ………

S3

Job

AWS GLUE

Data Catalog



Persistent Metadata Store

You can store, annotate and share metadata between AWS services (similar to Apache Hive metastore)

Centralized Repository

There is only 1 Data Catalog per AWS region, providing a uniform repository so that different systems can store and find metadata to query and transform that data.

Provides Comprehensive Audit

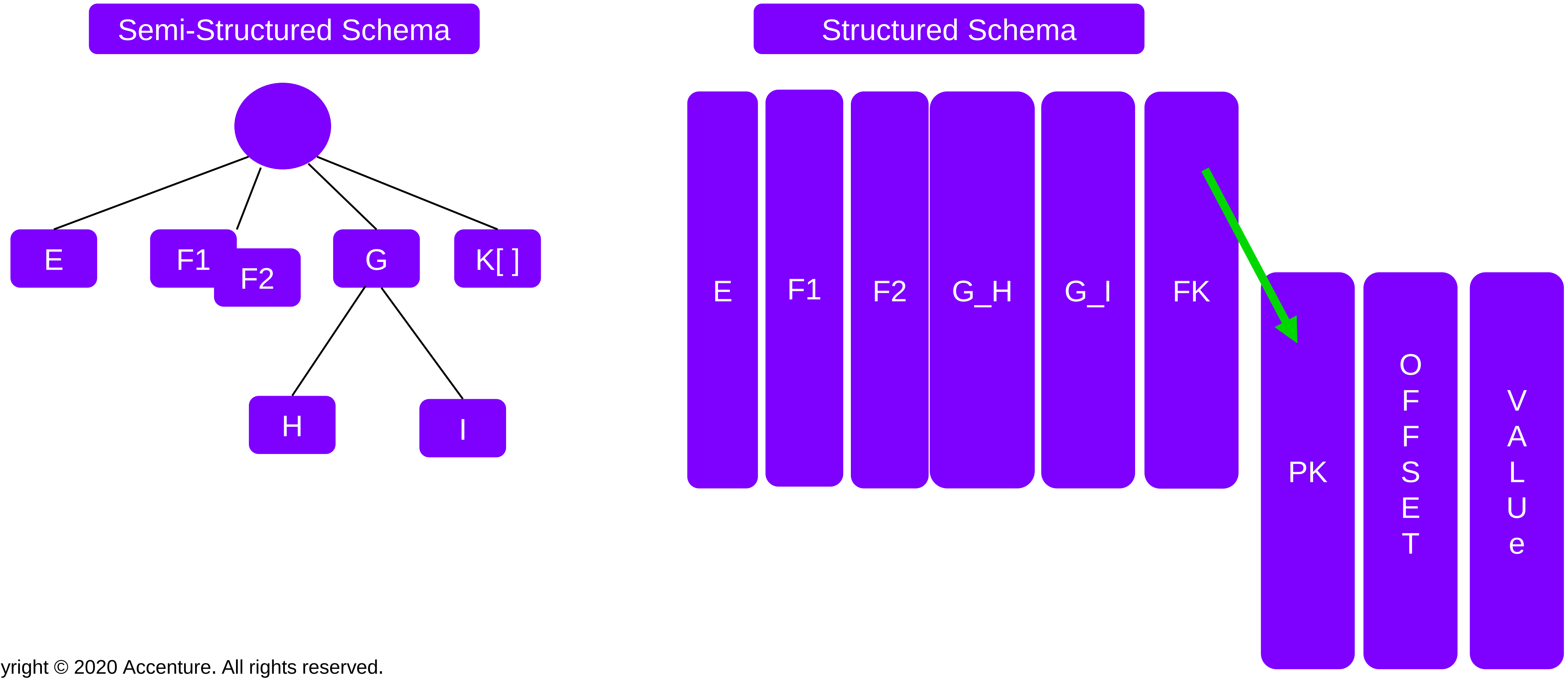
You can track schema changes and data access control. This helps ensure that its data is not inappropriately modified or shared

DEMO TO SETUP DATA CATALOG

AWS GLUE

Converting Semi-Structured Schemas to Relational Schema

- AWS Glue can do this conversion on the fly.



AWS Glue Jobs

AWS GLUE

Glue Jobs

Ingredients

Pattern, Tools, Preparing

Final Result

AWS GLUE

Glue Jobs

Ingredients

Input Data

Pattern, Tools, Preparing

Glue Job

Final Result

Output Data

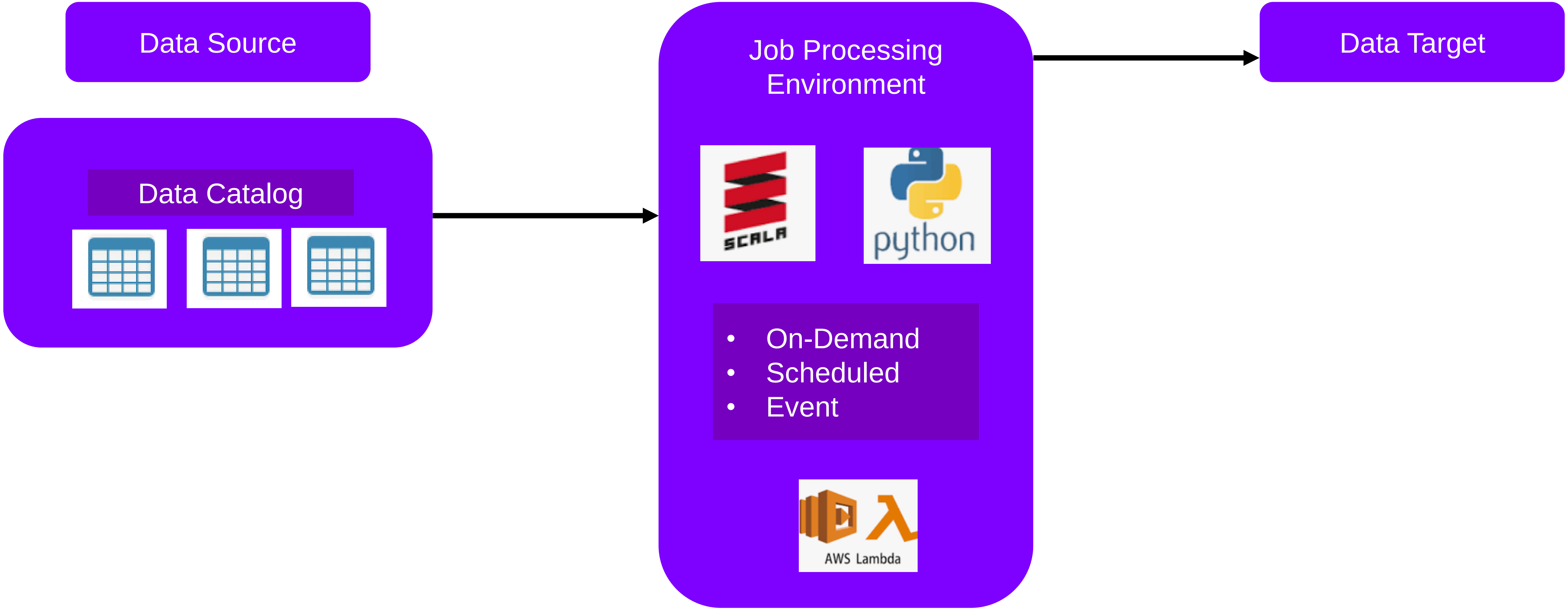
AWS GLUE

Glue Jobs

A ***job*** is the ***business logic*** that performs ***Extract, Transform and Load (ETL)*** work in AWS Glue.

AWS GLUE

Workflow Overview



Data Source

Data Catalog

Job Processing Environment

- On-Demand
- Scheduled
- Event

Data Target

DEMO TO SETUP AWS GLUE JOB

AWS GLUE

Glue Jobs : Output File Formats

- JSON *
- CSV *
- ORC
- PARQUET
- AVRO

* Optional Compression with gzip or bzip2

AWS GLUE

Glue Jobs : Data Processing Units

- Apache Spark – Min DPU – 2 | Max DPU – 100 | Default DPU - 10
- Spark Streaming - Min DPU – 2 | Max DPU – 100 | Default DPU - 5
- Python Shell - Min DPU – 0.0625 or 1 | Max DPU – 1 | Default DPU – 0.0625

.

AWS GLUE

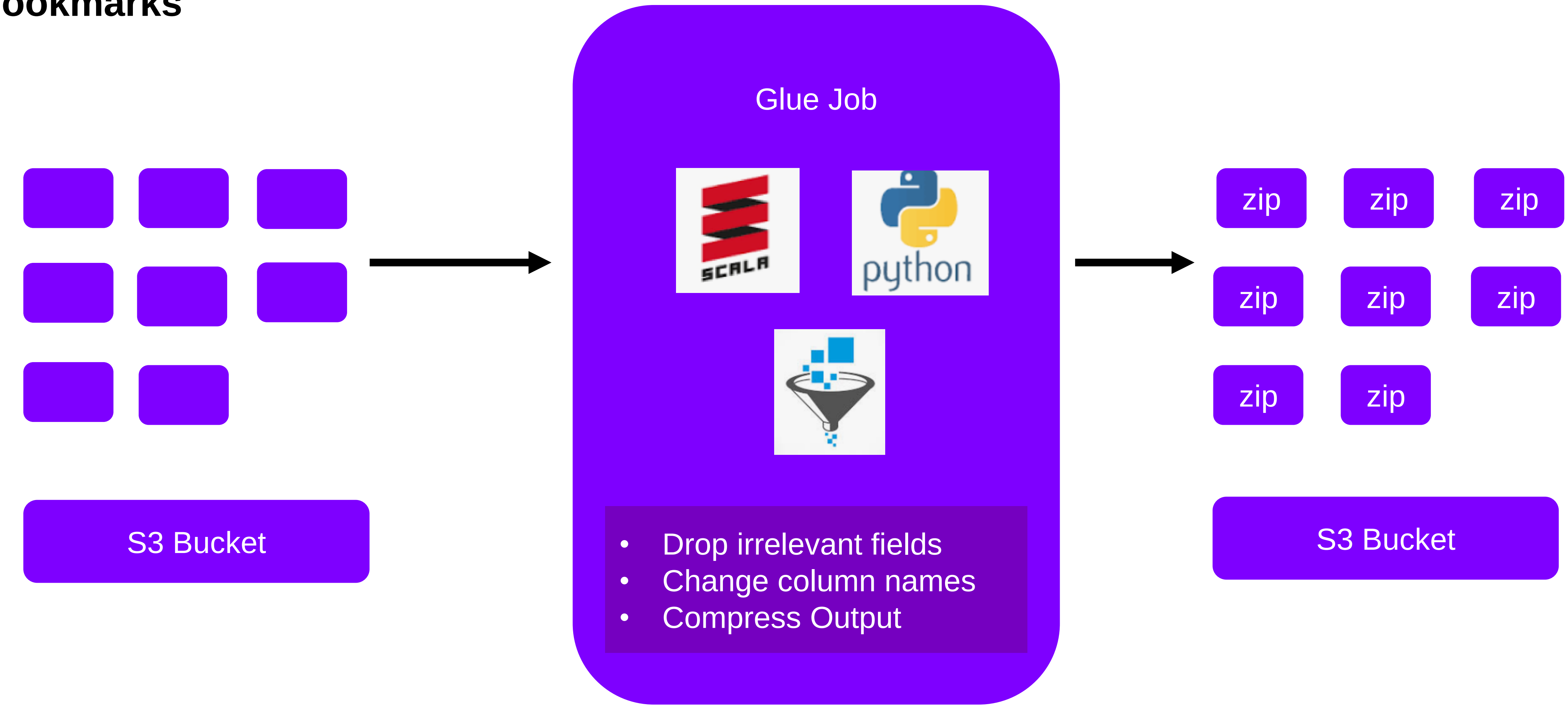
Glue Jobs : How Glue Jobs run in isolation?

- Glue Jobs run on virtual resources
- What does Glue jobs need to access the data?
- Traffic is governed by your VPC

.

AWS GLUE

Job Bookmarks



S3 Bucket

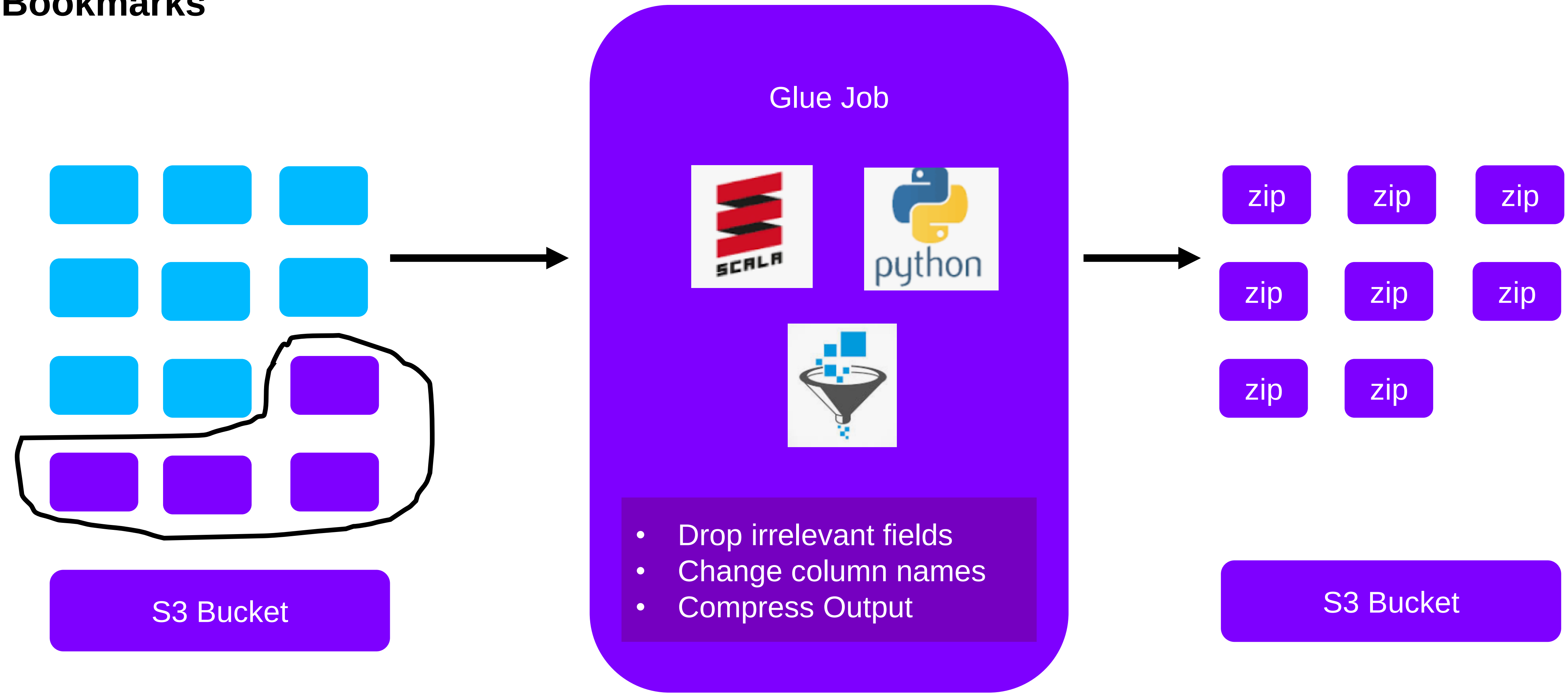
Glue Job

- Drop irrelevant fields
- Change column names
- Compress Output

S3 Bucket

AWS GLUE

Job Bookmarks



- Drop irrelevant fields
- Change column names
- Compress Output

🟦 Already processed

AWS GLUE

Job Bookmarks

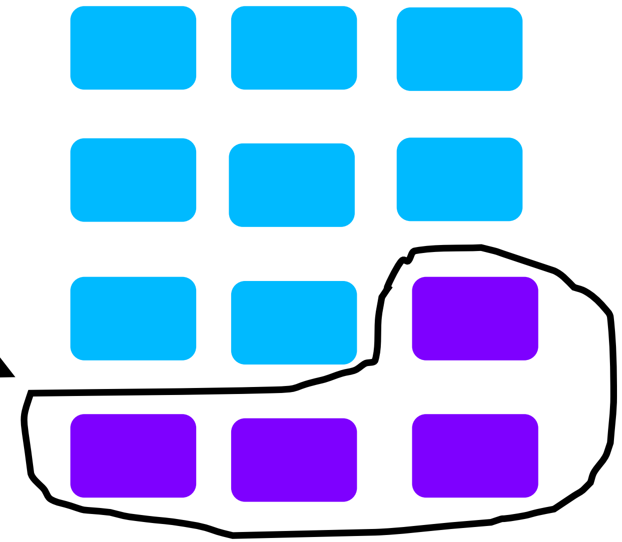
A way to process new data without reprocessing old data.

AWS GLUE

Options for Job Bookmarks

- **Enabled**
- Disabled (default)
- Pause

.



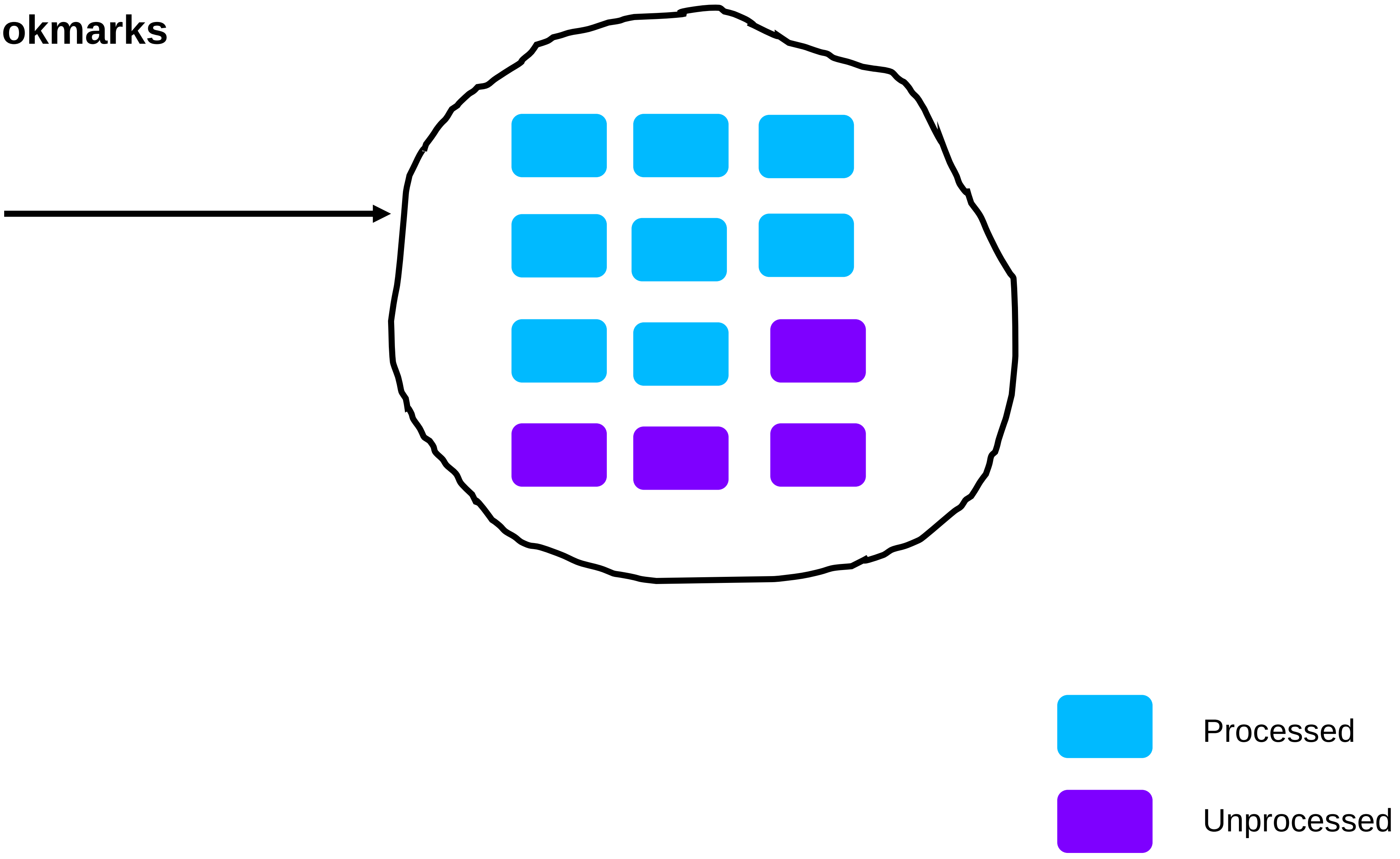
🟦 Processed

🟪 Unprocessed

AWS GLUE

Options for Job Bookmarks

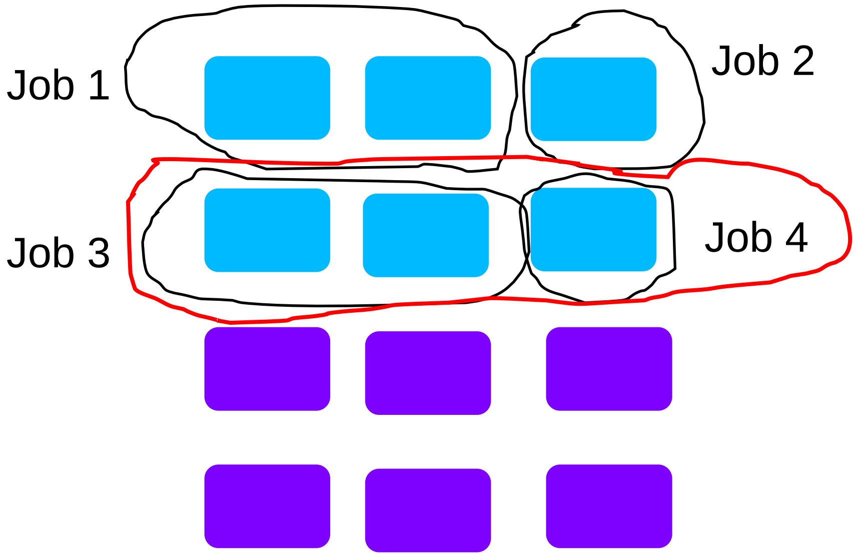
- Enabled
- **Disabled (default)**
- Pause



AWS GLUE

Options for Job Bookmarks

- Enabled
- Disabled (default)
- **Pause**



From-value : Job2
To-value : Job4

Processes everything after Job2 until Job4 (including Job4)

🟦 Processed

🟪 Unprocessed

DEMO TO SETUP AWS GLUE JOB BOOKMARKS

Athena

ATHENA

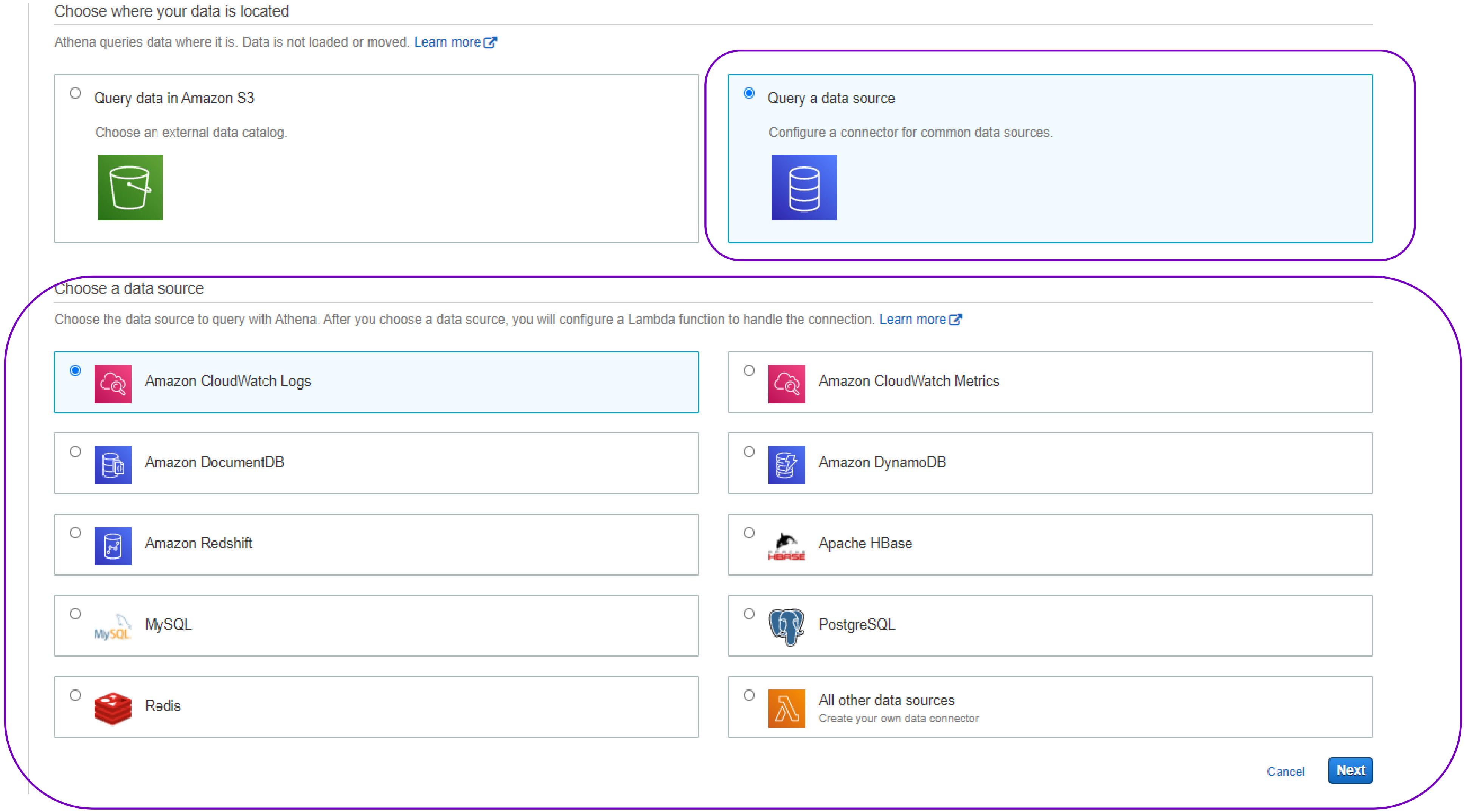
What is Athena?

- With Athena, you can easily query your data stored on S3 using SQL Queries
- Athena is serverless. You only pay for the queries that you run.
- Athena scales automatically and hence results are fast even if the dataset is large and the queries are complex.

.

ATHENA

Athena Federated queries



Choose where your data is located

Athena queries data where it is. Data is not loaded or moved. Learn more

○ Query data in Amazon S3

Choose an external data catalog.

● Query a data source

Configure a connector for common data sources.

Choose a data source

Choose the data source to query with Athena. After you choose a data source, you will configure a Lambda function to handle the connection. Learn more

● Amazon CloudWatch Logs

○ Amazon CloudWatch Metrics

○ Amazon DocumentDB

○ Amazon DynamoDB

○ Amazon Redshift

○ Apache HBase

○ MySQL

○ PostgreSQL

○ Redis

○ All other data sources
Create your own data connector

ATHENA

Athena Data Formats

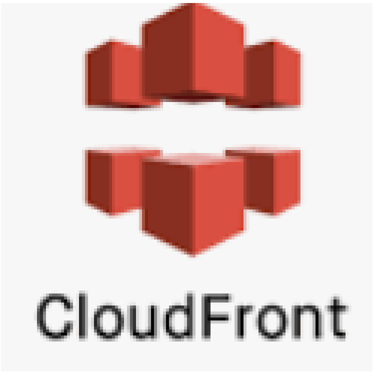
- Data Formats
  - Athena helps us to analyze structured, semi-structured and unstructured data stored in S3.
  - Example : CSV, TSV, JSON, Text files, Parquet and ORC as well as Snappy, Zlib, LZO, GZIP.
- Integrates with Quicksight
- Integrated with Glue

.

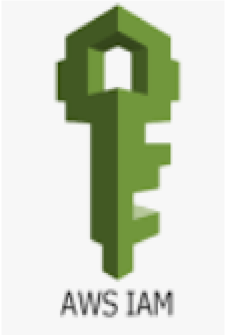
ATHENA

Integrations

Amazon CloudTrail

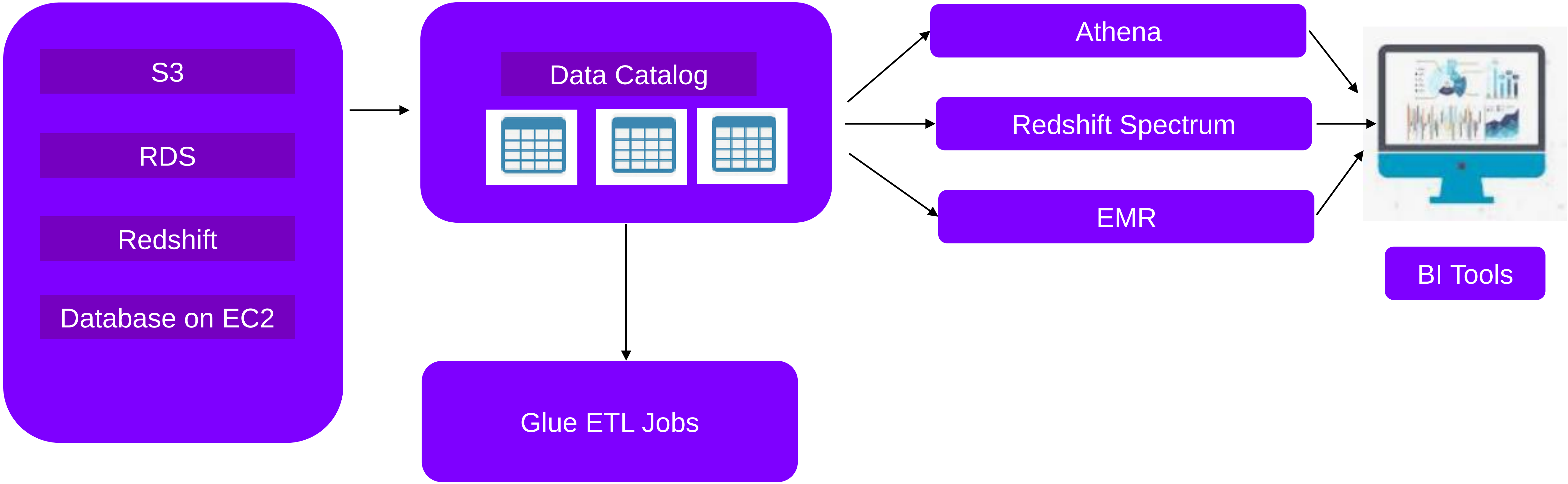


AWS CloudFormation



ATHENA

Use Cases – Ad-hoc queries and BI Tools



S3

RDS

Redshift

Database on EC2

Data Catalog

Glue ETL Jobs

Athena

Redshift Spectrum

EMR

BI Tools

DEMO TO USE ATHENA

ATHENA

Comparing Athena with other Services

Redshift
Fast Querying, Reporting, BI workloads
Very Complex SQL – multiple joins and sub queries
Brings data from various other data sources into a common format
Storing data for long period of time
Build business reports from historical data

EMR
Simple to run distributed processing frameworks like Hadoop, Spark and Presto.
Flexible to run custom applications and code
You define compute, memory and storage to optimize workloads

Athena
Easily run Ad-hoc queries for data in S3
No need to setup or manage servers
No need to format data
Storing data for long period of time
Can be used with EMR and Redshift as an integrated Data Catalog

ATHENA

Comparing Athena with other Services

S3 Select
Use SQL statements to filter the contents of S3 objects and retrieve just the subset of data that we need.
Offload filtering of data in S3 instead of our applications.
You can run Simple SQL Expressions.
Data formats are limited. CSV, JSON, Parquet. (GZIP and BZIP2 for CSV and JSON)

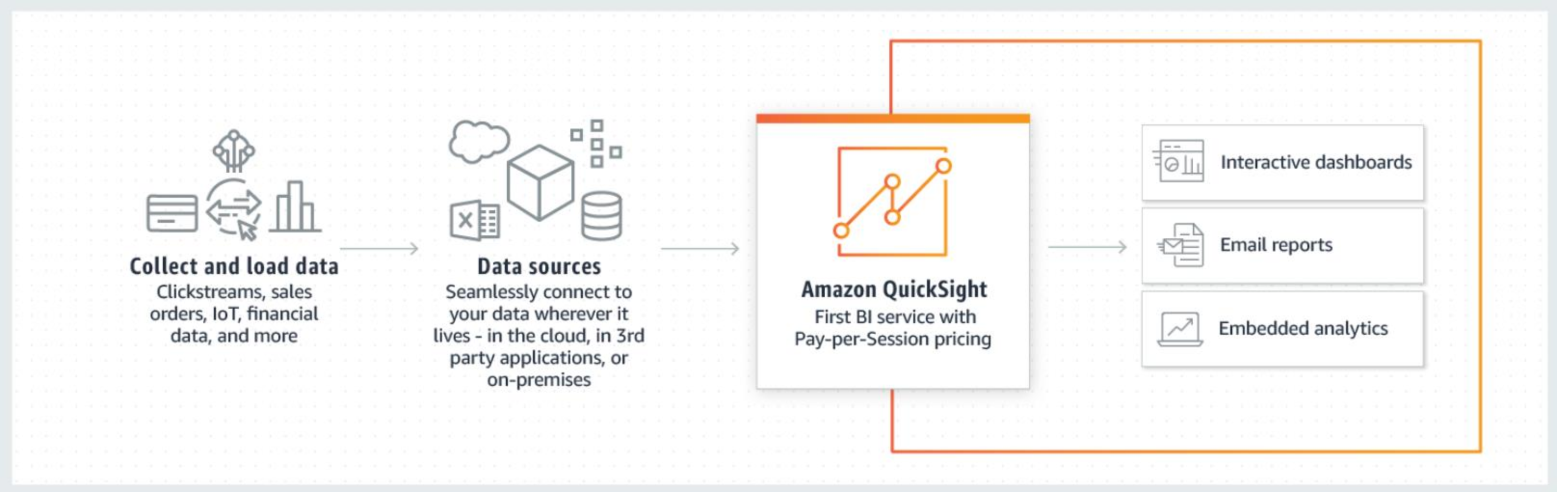
Glacier Select
Use SQL statements directly on your data in S3 Glacier without having to restore data to a more frequently accessible tier.
We can query Glacier data within minutes
We can use standard SQL statements
No need to restore data to S3

QuickSight Visualizations

QUICKSIGHT

Introduction

- Create excellent Visualizations and Dashboards with your data.
- It is a BI tool and we need to simply point QuickSight at your input data source to start creating Visualizations.
- We can create Interactive dashboards, Email reports and Embedded Analytics.



Collect and load data
Clickstreams, sales orders, IoT, financial data, and more

Data sources
Seamlessly connect to your data wherever it lives - in the cloud, in 3rd party applications, or on-premises

Amazon QuickSight
First BI service with Pay-per-Session pricing

Interactive dashboards

Email reports

Embedded analytics

.

QUICKSIGHT

How QuickSight works?

Connect to your Data

- AWS Data
- Applications
- Files

Relational Data	
• Amazon Athena	• Apache Spark
• Amazon Aurora	• MariaDB
• Amazon Redshift	• Microsoft SQL Server
• Amazon Redshift Spectrum	• MySQL
• Amazon S3	• PostgreSQL
• Amazon S3 Analytics	• Presto
• Amazon IOT Analytics	• Snowflake
	• Teradata

Importing File Data

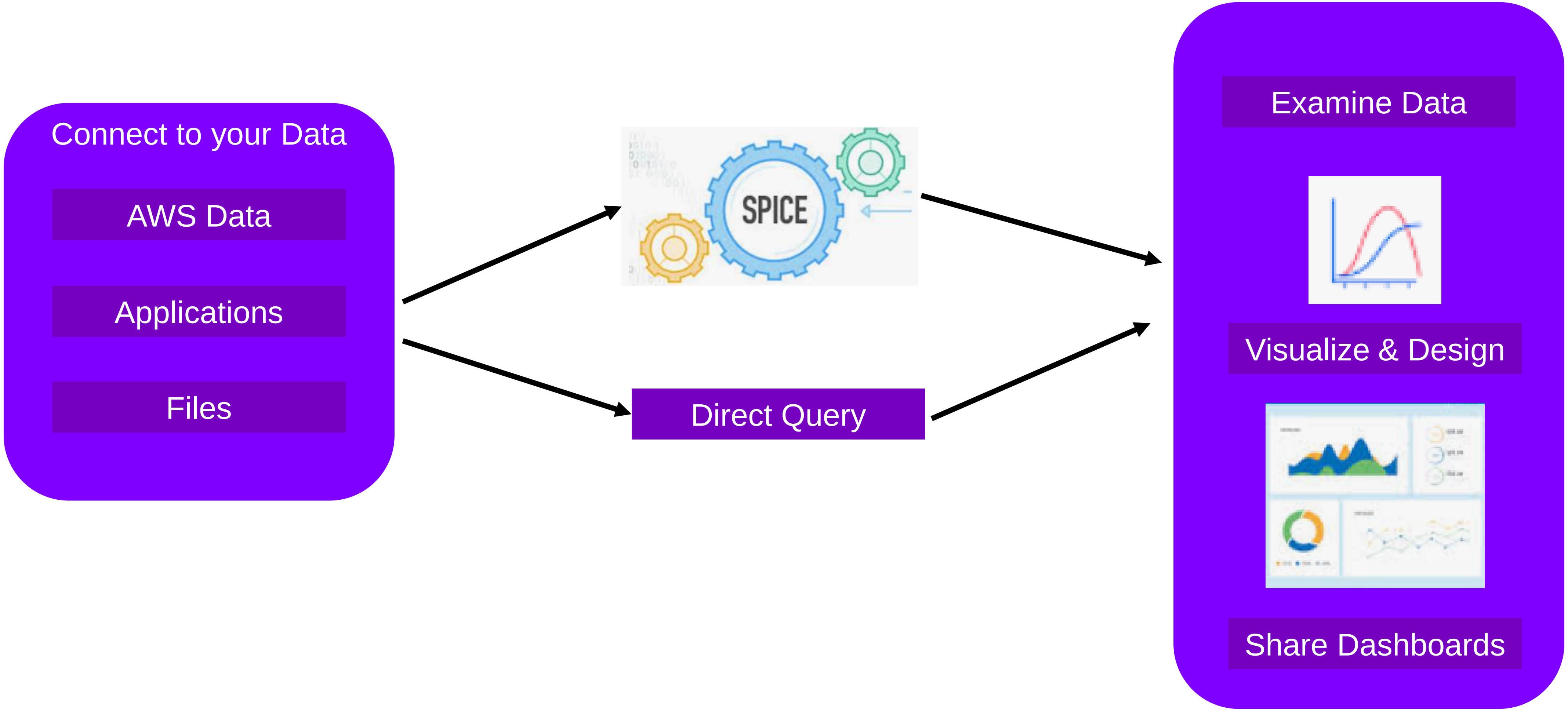
- CSV and TSV
- ELF and CLF
- JSON
- XLSX
- ZIP and GZIP (S3)

SaaS Data

- Jira
- ServiceNow
- Adobe Analytics
- GitHub
- Salesforce
- Twitter

QUICKSIGHT

How QuickSight works?



Connect to your Data

AWS Data

Applications

Files

Direct Query

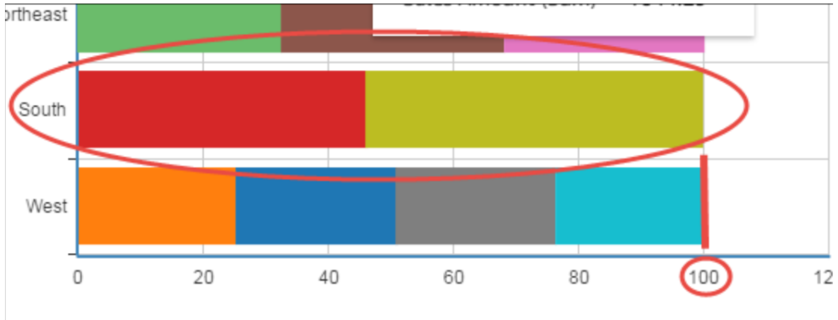
Examine Data

Visualize & Design

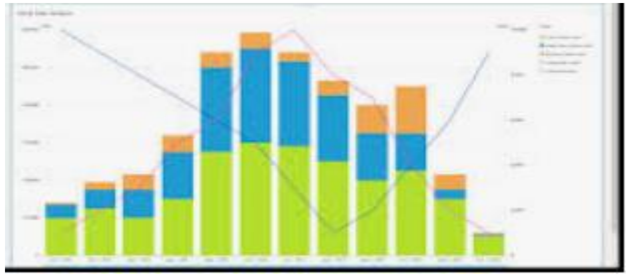
Share Dashboards

QUICKSIGHT

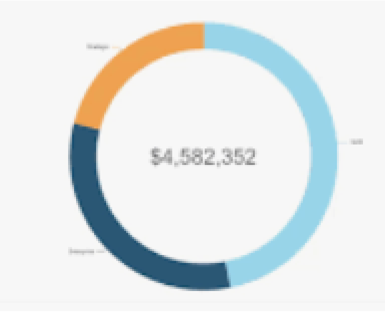
Visualization Types



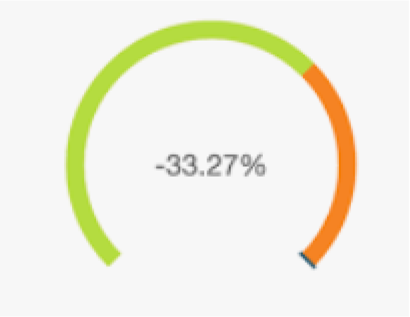
Bar Chart



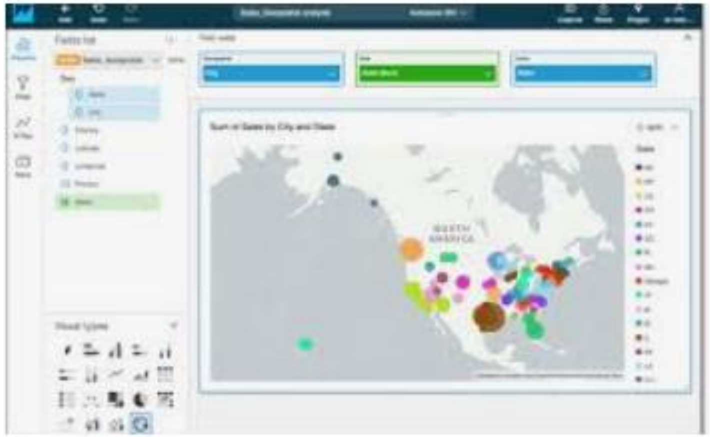
Combo Chart



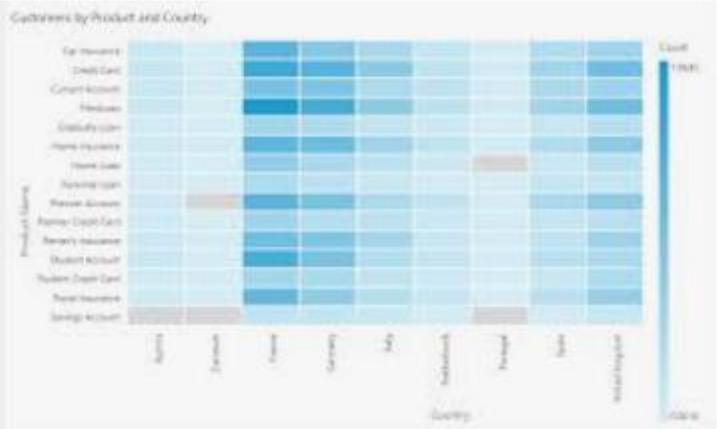
Donut Chart



Gauge Chart



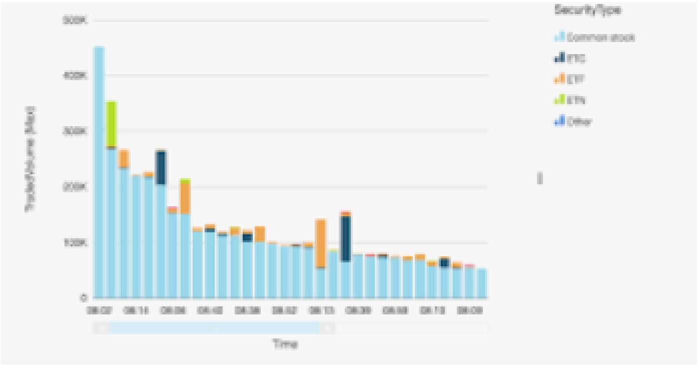
Geo-Spatial Charts (Maps)



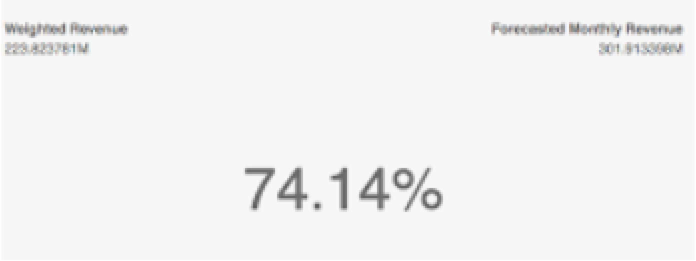
Heat Maps

QUICKSIGHT

Visualization Types



Histogram

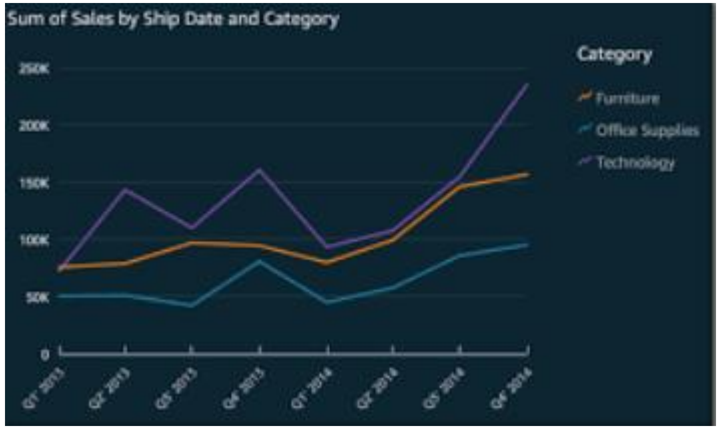


Weighted Revenue
223.823761M

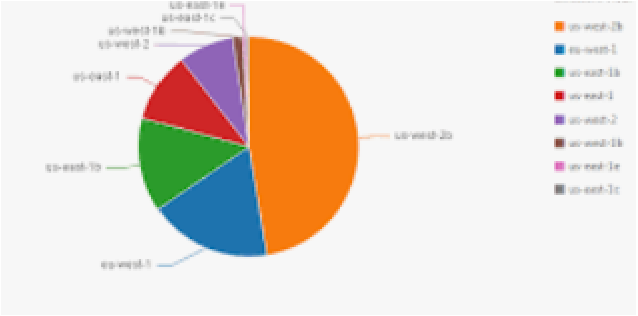
Forecasted Monthly Revenue
301.913398M

74.14%

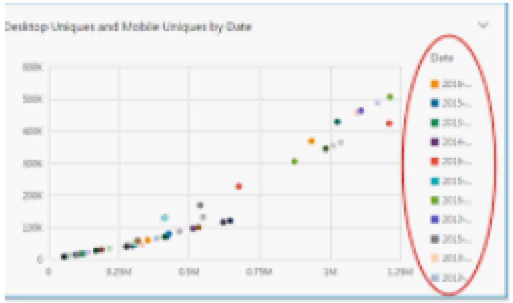
KPIs



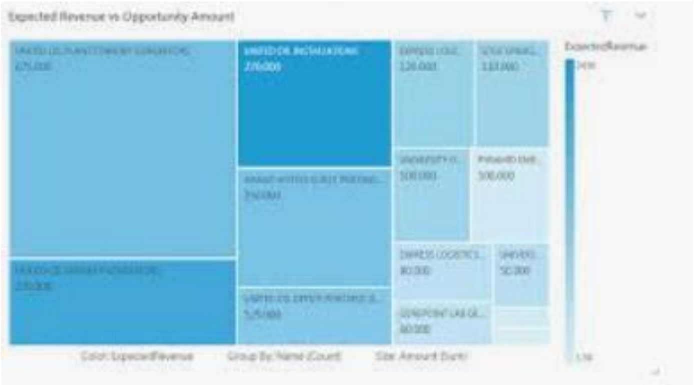
Line Chart



Pie Chart



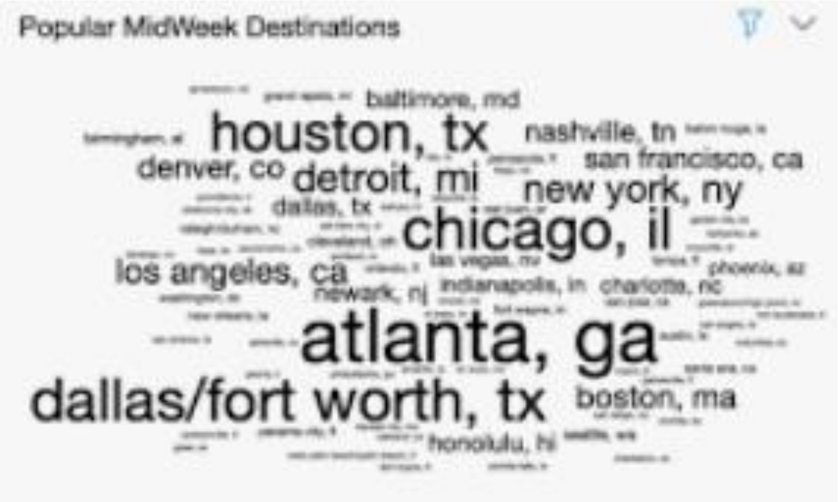
Scatter Plot



Tree Map

QUICKSIGHT

Visualization Types



Word Cloud

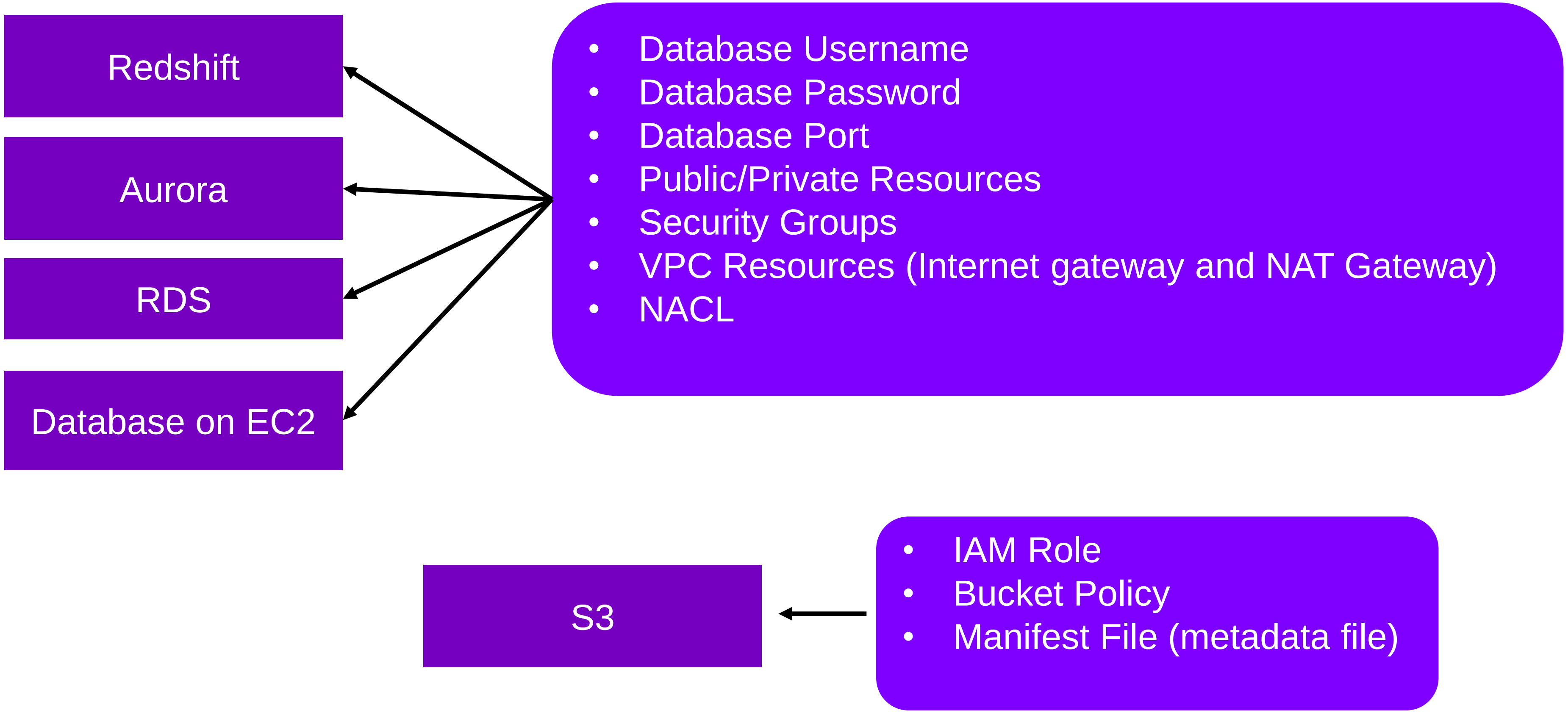
QUICKSIGHT

Security and Authentication

- Data Encryption
  - Encryption at Rest
    - Provided with Enterprise edition only
    - All metadata and data uploaded into SPICE is encrypted with AWS-managed keys
  - Encryption in Transit
    - Supported in both Standard and Enterprise edition
    - Quicksight supports encryption of data transfer using SSL.
    - This includes data to & from SPICE and from SPICE to the user interface
  - Key Management
    - AWS manages all keys associated with Quicksight
    - Database server certificates are the responsibility of the customer

QUICKSIGHT

Connecting to AWS Resources



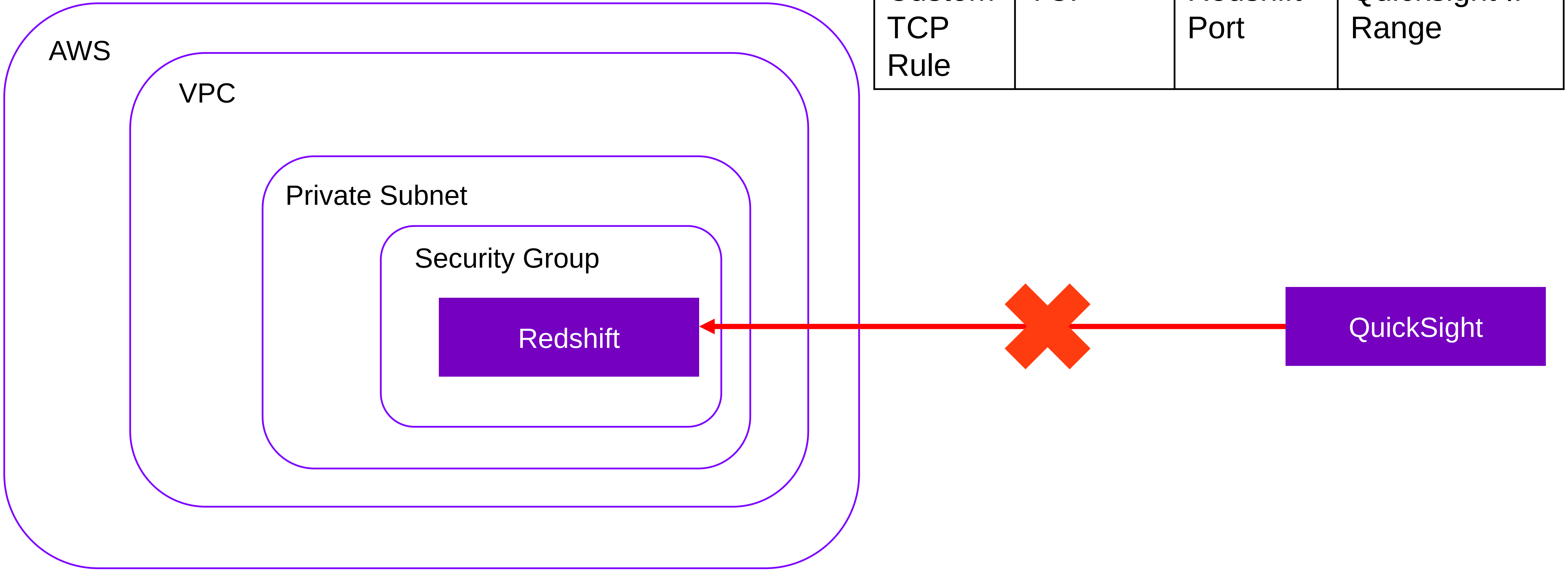
- Database Username
- Database Password
- Database Port
- Public/Private Resources
- Security Groups
- VPC Resources (Internet gateway and NAT Gateway)
- NACL

- IAM Role
- Bucket Policy
- Manifest File (metadata file)

QUICKSIGHT

Connecting to AWS Resources

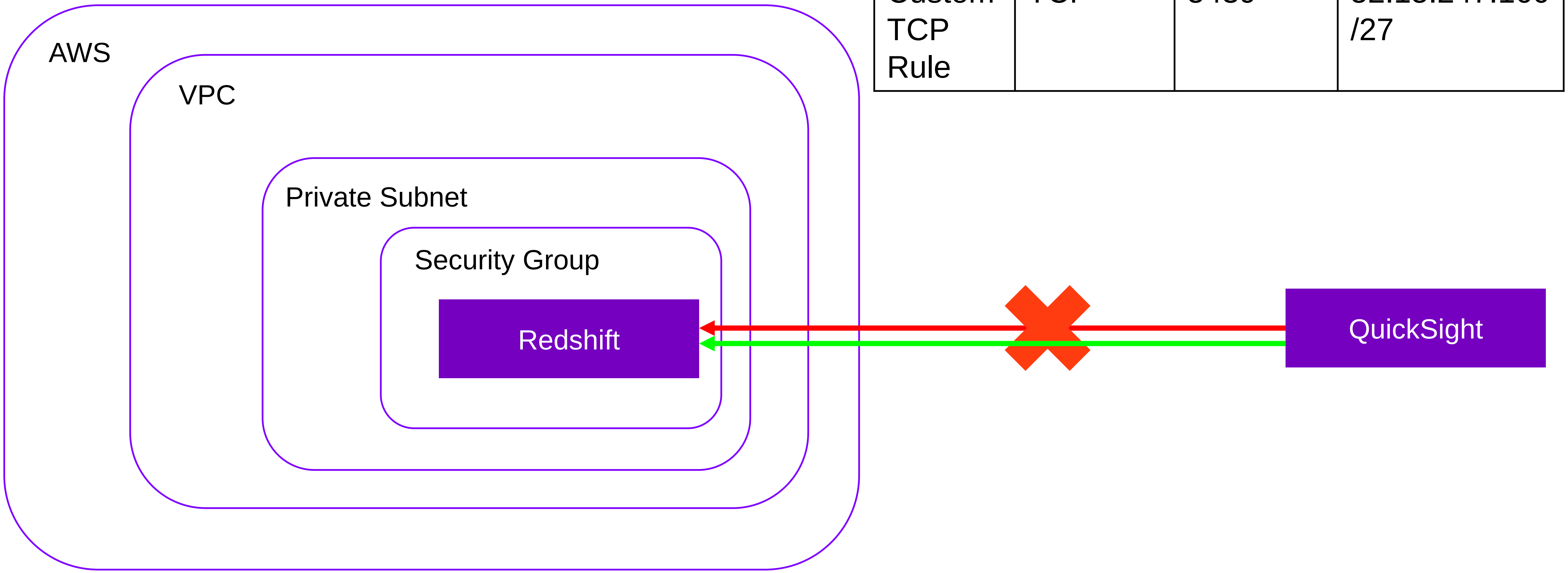
Inbound Rules			
Type	**Protocol**	**Port Range**	**Source**
Custom TCP Rule	TCP	Redshift Port	Quicksight IP Range



QUICKSIGHT

Connecting to AWS Resources

Inbound Rules			
Type	**Protocol**	**Port Range**	**Source**
Custom TCP Rule	TCP	5439	52.15.247.160 /27



QUICKSIGHT

Connecting to AWS Resources

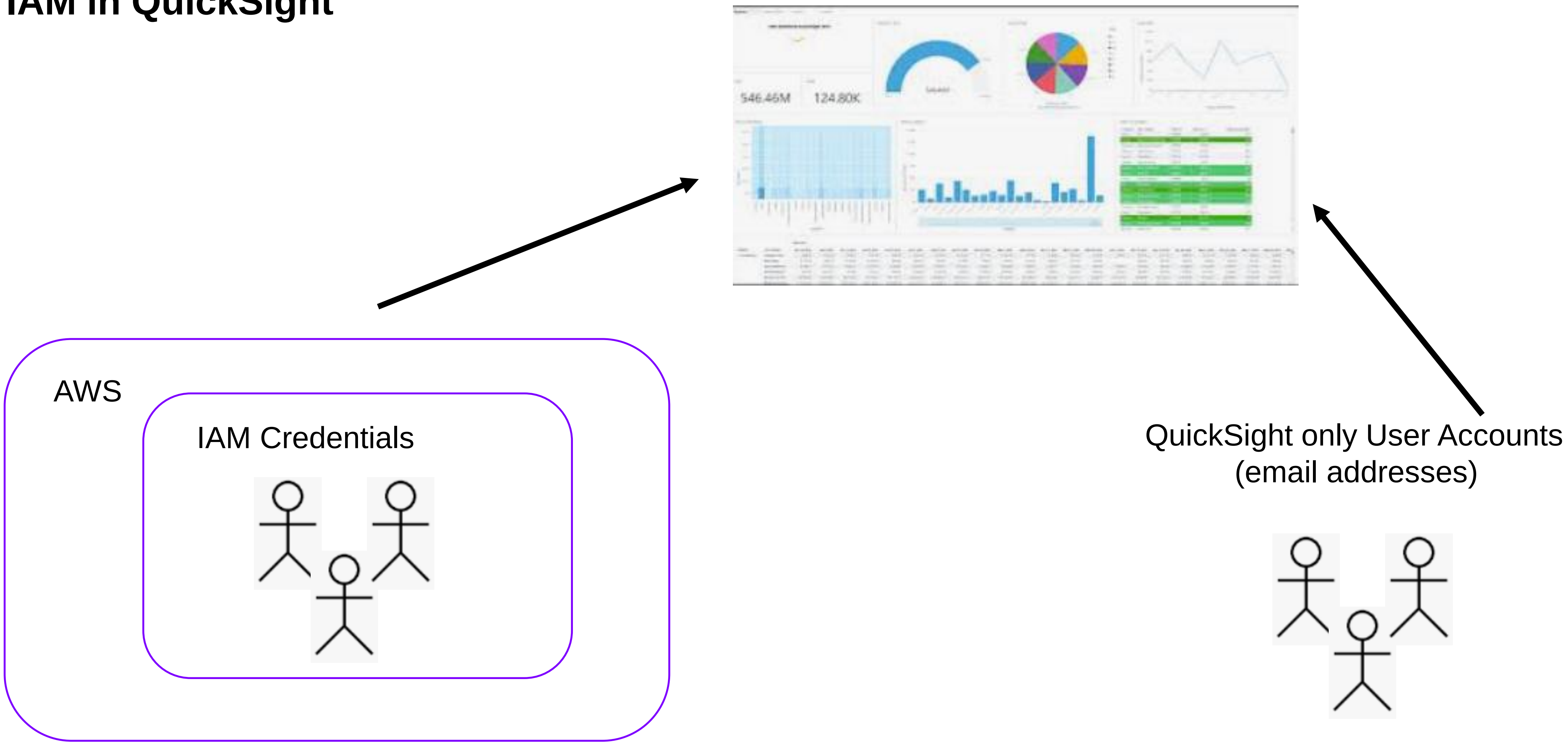
For QuickSight to connect to private Resource, the security group must contain an inbound rule authorization access from appropriate IP Address range for the QuickSight Servers in that AWS Region:

This includes Resources like:

- RDS Instance
- RedShift Clusters
- EC2 Instances

QUICKSIGHT

IAM in QuickSight



AWS

IAM Credentials

QuickSight only User Accounts
(email addresses)

QUICKSIGHT

Best Security Practices

- Firewall
- SSL
- Enhanced Security

.

DEMO TO VISUALIZE USING QUICKSIGHT

CLASS DISCUSSION

- How can I import data from existing Apache Hive metastore to the Glue Data Catalog?
- How does Glue relate to AWS Lake formation?
- How can we use custom libraries with ETL scripts in AWS Glue?
- How can I build end-to-end ETL workflow using multiple jobs in AWS Glue?
- How much do I have to pay to use Athena?

.

AWS ElasticSearch

AWS ELASTICSEARCH

What is ElasticSearch?

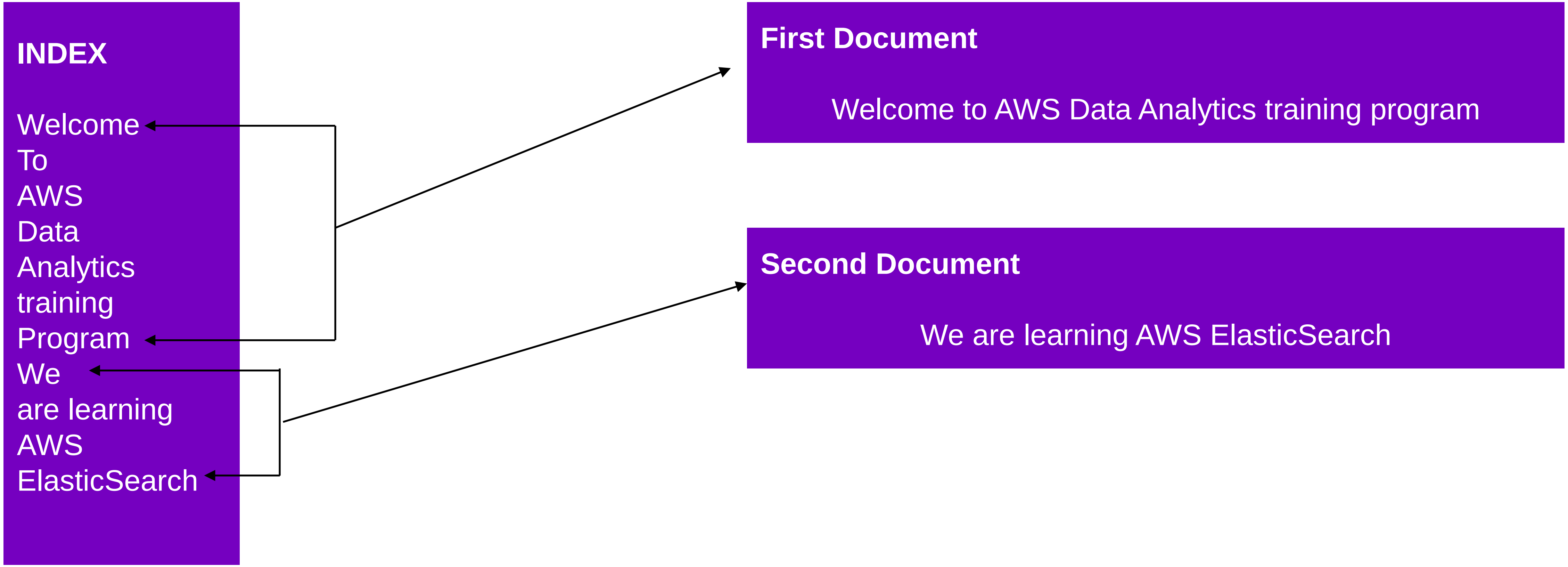
- It is a search domain which runs as a part of ELK (**E**lasticSearch **L**ogstash **K**ibana)
- A Search Engine
- An analysis tool
- Visualization tool used is Kibana

.

AWS ELASTICSEARCH

How does ElasticSearch works?

- It organizes data as indexes instead of tables and purpose of ElasticSearch is to do word searching.
- It is a search engine utility



INDEX

Welcome
To
AWS
Data
Analytics
training
Program
We
are learning
AWS
ElasticSearch

First Document

Welcome to AWS Data Analytics training program

Second Document

We are learning AWS ElasticSearch

AWS ELASTICSEARCH

Logstash and Kibana

Logstash

- Ingests, processes and stores log data.
- It is tightly integrated with ElasticSearch and can use ElasticSearch as its stores
- Powerful for managing large infrastructures that generates huge amount of log data.

Kibana

- Web utility interface for ElasticSearch and Visualization Engine.

AWS ELASTICSEARCH

Using ElasticSearch

- ElasticSearch is JSON all the way down. It organizes data in JSON documents. Everything is configured or stored in JSON documents.
- Top level organizational unit is INDEX
  - We then have Type at second level which is used to categorize data in our ES domain
    - We then have Document which has the data stored

AWS ELASTICSEARCH

Interface

- ElasticSearch uses REST API for its interface and JSON is a common format for REST APIs.
- We get a URL that we can interact with as a REST API

  https://search-mydomain-1a2a3a4a5a6a7a8a9a0a9a8a7a.us-east-1.es.amazonaws.com …
- Assuming that we have permissions, we can send a GET request to the base URL ..../my-index/type/item-id

  and that item will be returned.
- We can do a PUT to …/my-index or …/my-index/type/item-id and then it will add a document to index or item-id. If we do to an index, it will generate an id for us.
- We can do a POST to …/my-index/type (it will set the definition for that type) or …/my-index/type/item-id but we can’t resend to the index (…/my-index) because it will not generate id for us and that’s the difference between POST and PUT.
- We can send DELETE request to …/my-index/type/item-id and …/my-index to delete items or entire index. We can’t delete a type because if you do that then you will invalidate the index.

DEMO ON ELASTICSEARCH

MODULE SUMMARY

Now, you should be able to:

- Explain Glue Data Catalog
- Explain Glue Jobs
- Perform Operations with Glue Jobs
- Explain Job Bookmarks
- Get Started with Athena
- Perform Operations with Athena
- Create Visualizations with QuickSight
- Push data to ElasticSearch and Discover data

THANK YOU