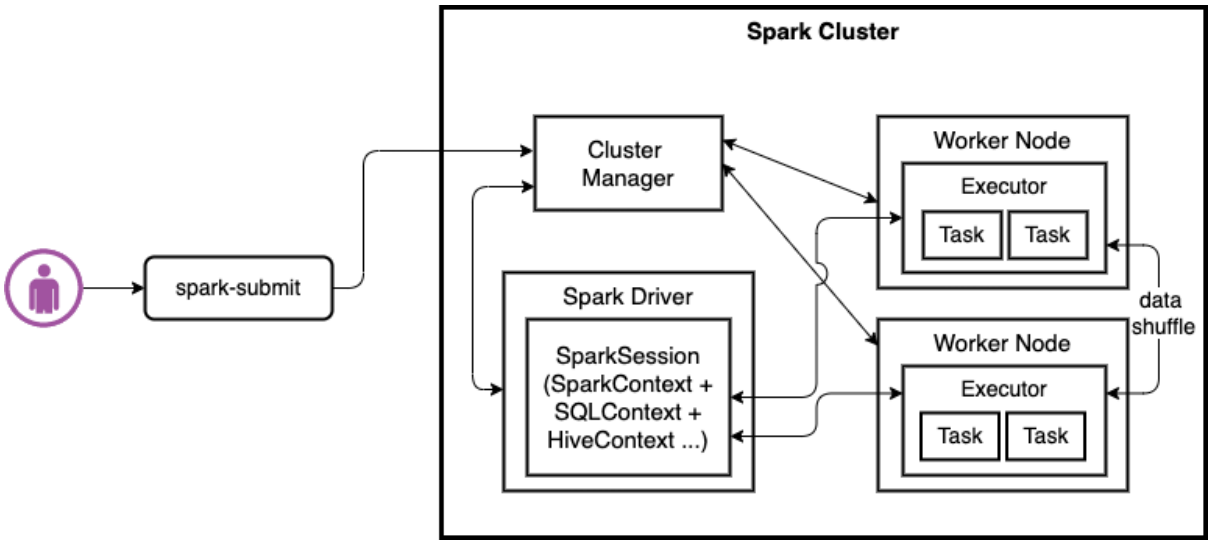
Chapter 1: Data Management – Introduction and Concepts



Chapter 2: Introduction to Important AWS Glue Features

```
# col_name              data_type               comment

id                      bigint
country                 struct<name:string,capital:string,countrycode:string,phonecode:string>

# Partition Information
# col_name              data_type               comment

currency                string
```

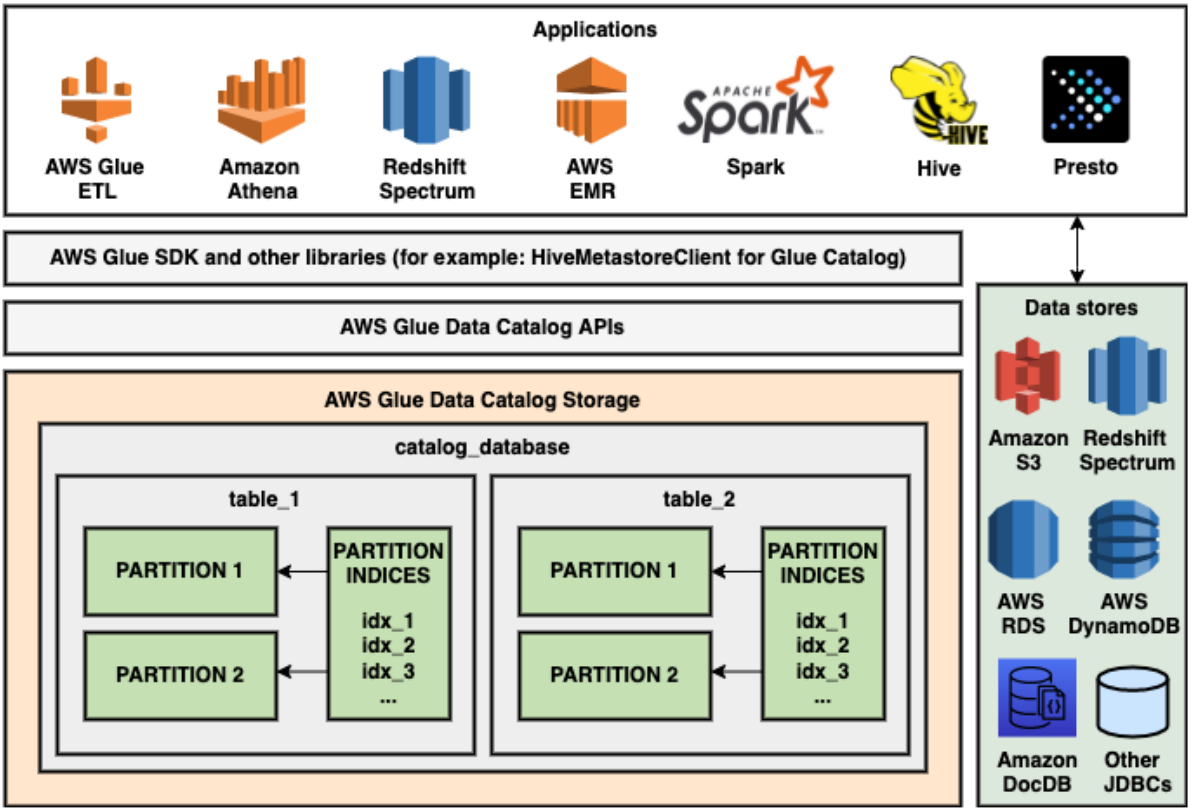
Table Schema

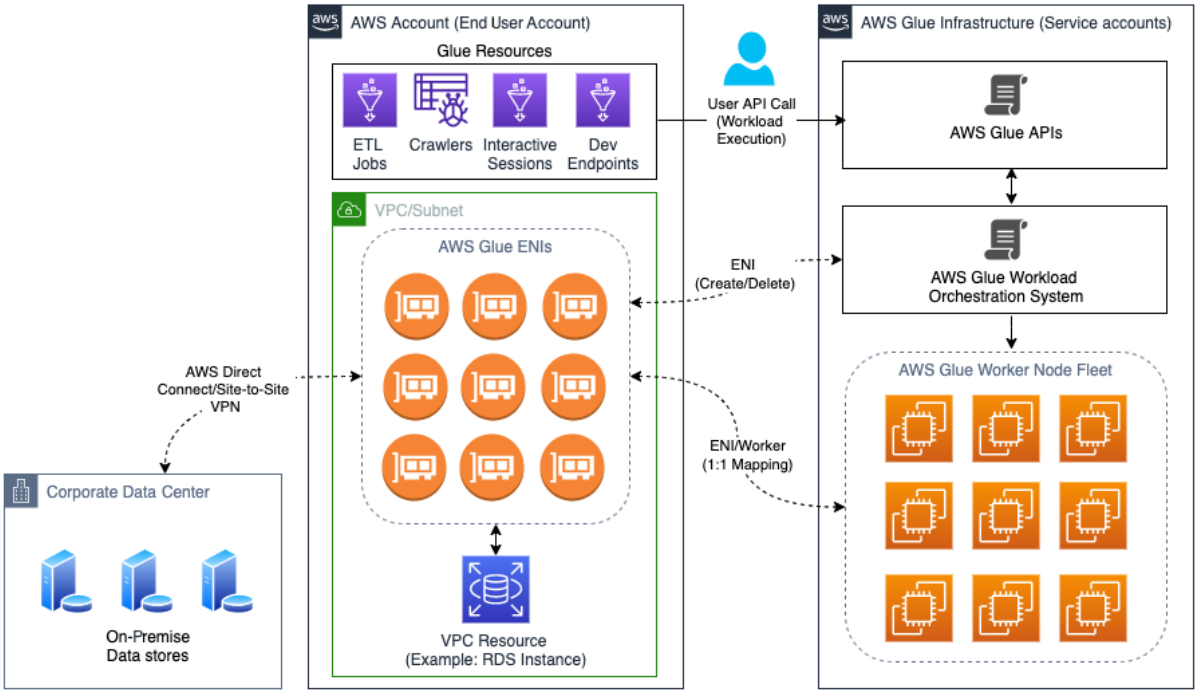
```
# Detailed Table Information
Database:               default
Owner:                  hadoop
CreateTime:             Tue Jul 12 22:01:56 UTC 2022
LastAccessTime:         UNKNOWN
Protect Mode:           None
Retention:              0
Location:               s3://BUCKET_NAME/S3_PREFIX
Table Type:             EXTERNAL_TABLE
Table Parameters:
        EXTERNAL                TRUE
        classification          parquet
        compressionType         none
        transient_lastDdlTime   1657663316
        typeOfData              file
```

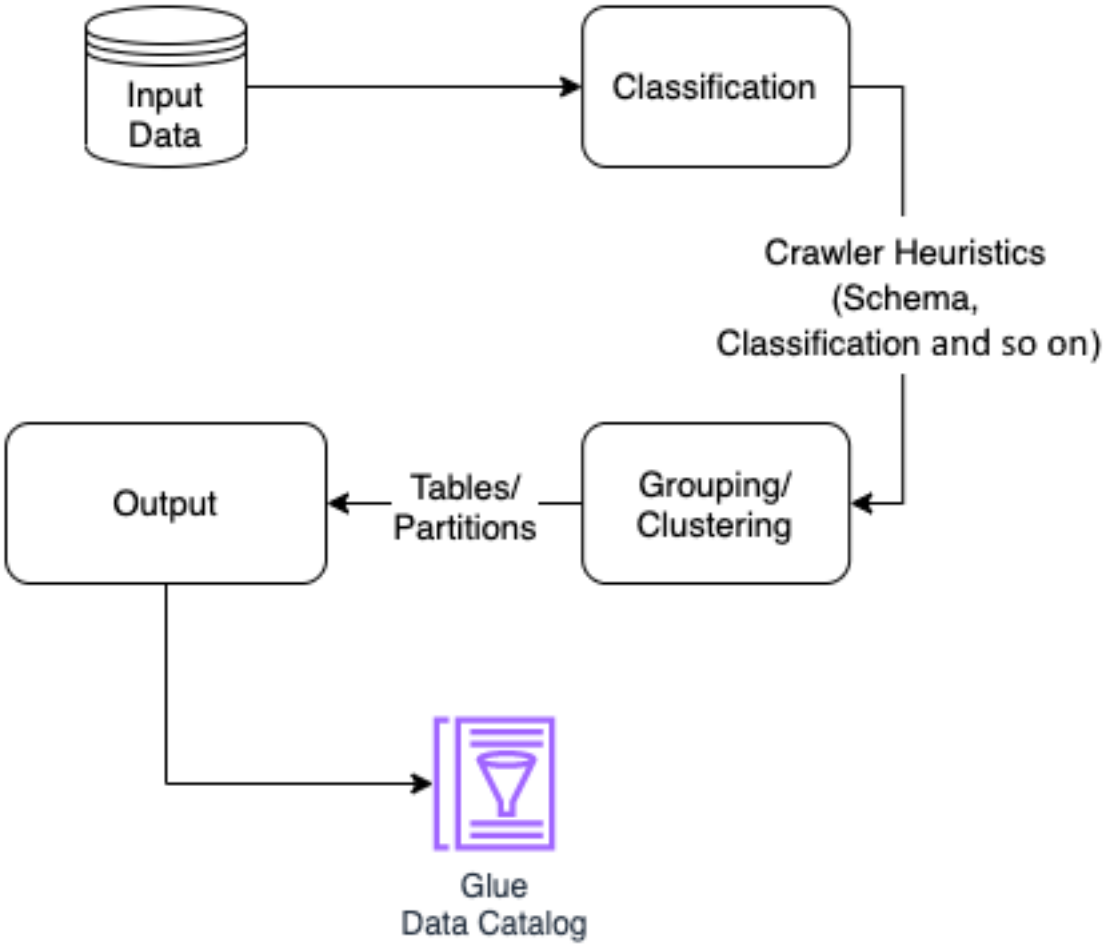
Data format, Location and other information

```
# Storage Information
SerDe Library:          org.apache.hadoop.hive.ql.io.parquet.serde.ParquetHiveSerDe
InputFormat:            org.apache.hadoop.hive.ql.io.parquet.MapredParquetInputFormat
OutputFormat:           org.apache.hadoop.hive.ql.io.parquet.MapredParquetOutputFormat
Compressed:             No
Num Buckets:            -1
Bucket Columns:         []
Sort Columns:           []
Storage Desc Params:
        serialization.format    1
```

Input/Output format, Serialization and Deserialization Library (SerDe) information



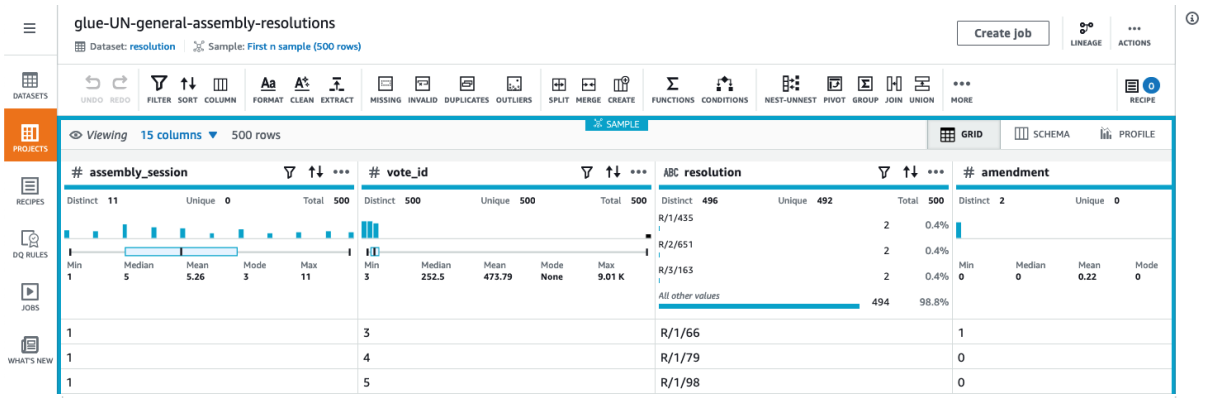

AWS Account (End User Account)
Glue Resources
ETL Jobs
Crawlers
Interactive Sessions
Dev Endpoints
VPC/Subnet
AWS Glue ENIs
VPC Resource (Example: RDS Instance)
User API Call (Workload Execution)
ENI (Create/Delete)
ENI/Worker (1:1 Mapping)
AWS Glue Infrastructure (Service accounts)
AWS Glue APIs
AWS Glue Workload Orchestration System
AWS Glue Worker Node Fleet
AWS Direct Connect/Site-to-Site VPN
Corporate Data Center
On-Premise Data stores

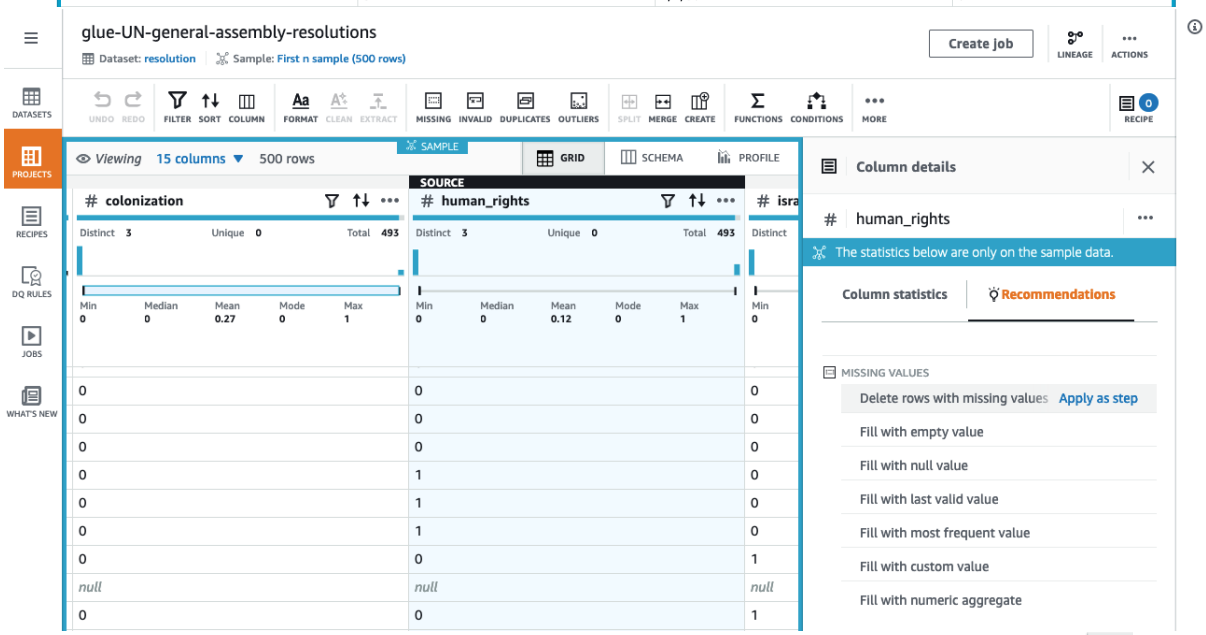

Input Data
Classification
Crawler Heuristics (Schema, Classification and so on)
Grouping/ Clustering
Tables/ Partitions
Output
Glue Data Catalog

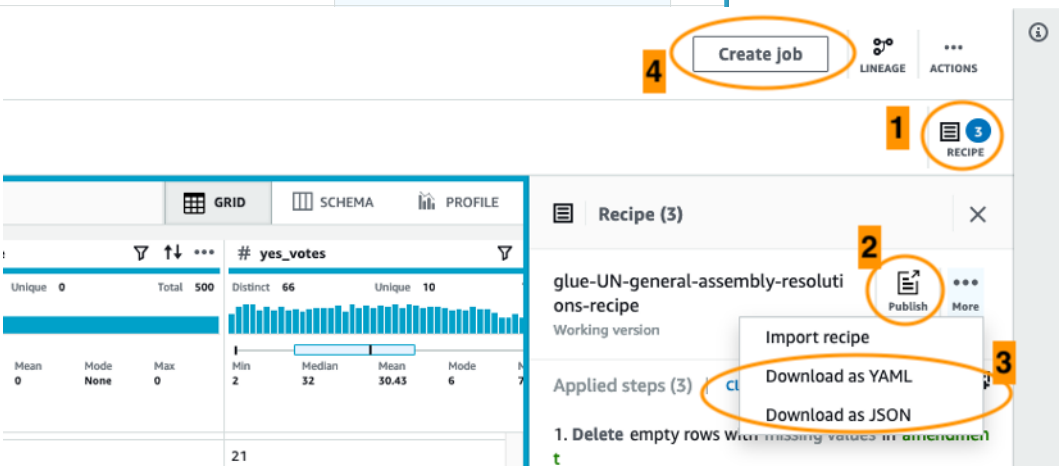
Chapter 3: Data Ingestion

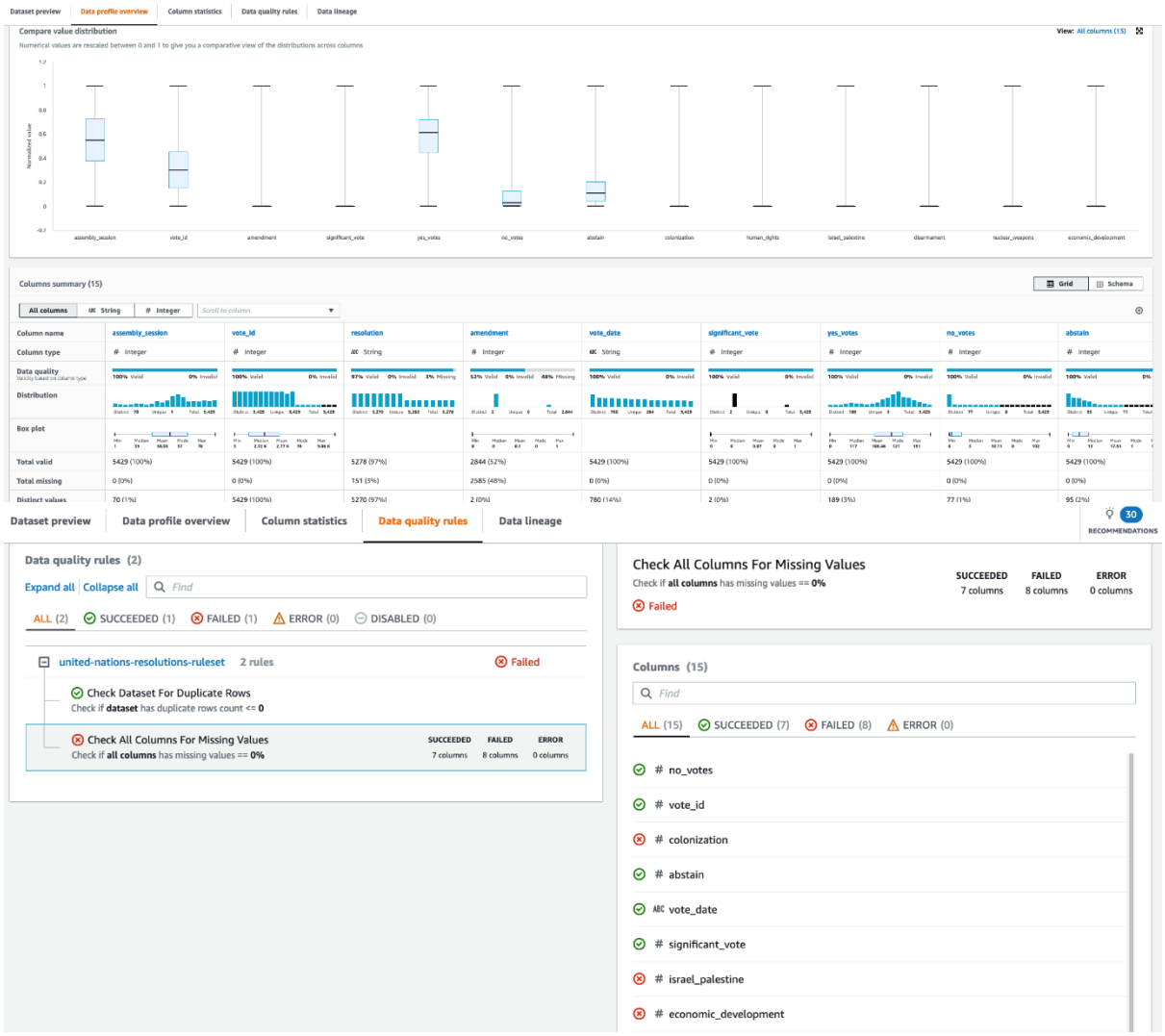
```
mysql> SELECT argument FROM mysql.general_log WHERE argument LIKE '%city%' AND
command_type = 'Prepare' ORDER BY argument;
+-----------------------------------------------------------------------+
| argument                                                              |
+-----------------------------------------------------------------------+
| SELECT * FROM (select * from city WHERE ID % 10 = 0) as city          |
| SELECT * FROM (select * from city WHERE ID % 10 = 0) as city WHERE 1=0 |
| SELECT * FROM (select * from city WHERE ID % 10 = 1) as city          |
| SELECT * FROM (select * from city WHERE ID % 10 = 1) as city WHERE 1=0 |
| SELECT * FROM (select * from city WHERE ID % 10 = 2) as city          |
| SELECT * FROM (select * from city WHERE ID % 10 = 2) as city WHERE 1=0 |
| SELECT * FROM (select * from city WHERE ID % 10 = 3) as city          |
| SELECT * FROM (select * from city WHERE ID % 10 = 3) as city WHERE 1=0 |
| SELECT * FROM (select * from city WHERE ID % 10 = 4) as city          |
| SELECT * FROM (select * from city WHERE ID % 10 = 4) as city WHERE 1=0 |
| SELECT * FROM (select * from city WHERE ID % 10 = 5) as city          |
| SELECT * FROM (select * from city WHERE ID % 10 = 5) as city WHERE 1=0 |
| SELECT * FROM (select * from city WHERE ID % 10 = 6) as city          |
| SELECT * FROM (select * from city WHERE ID % 10 = 6) as city WHERE 1=0 |
| SELECT * FROM (select * from city WHERE ID % 10 = 7) as city          |
| SELECT * FROM (select * from city WHERE ID % 10 = 7) as city WHERE 1=0 |
| SELECT * FROM (select * from city WHERE ID % 10 = 8) as city          |
| SELECT * FROM (select * from city WHERE ID % 10 = 8) as city WHERE 1=0 |
| SELECT * FROM (select * from city WHERE ID % 10 = 9) as city          |
| SELECT * FROM (select * from city WHERE ID % 10 = 9) as city WHERE 1=0 |
+-----------------------------------------------------------------------+
20 rows in set (0.20 sec)
```

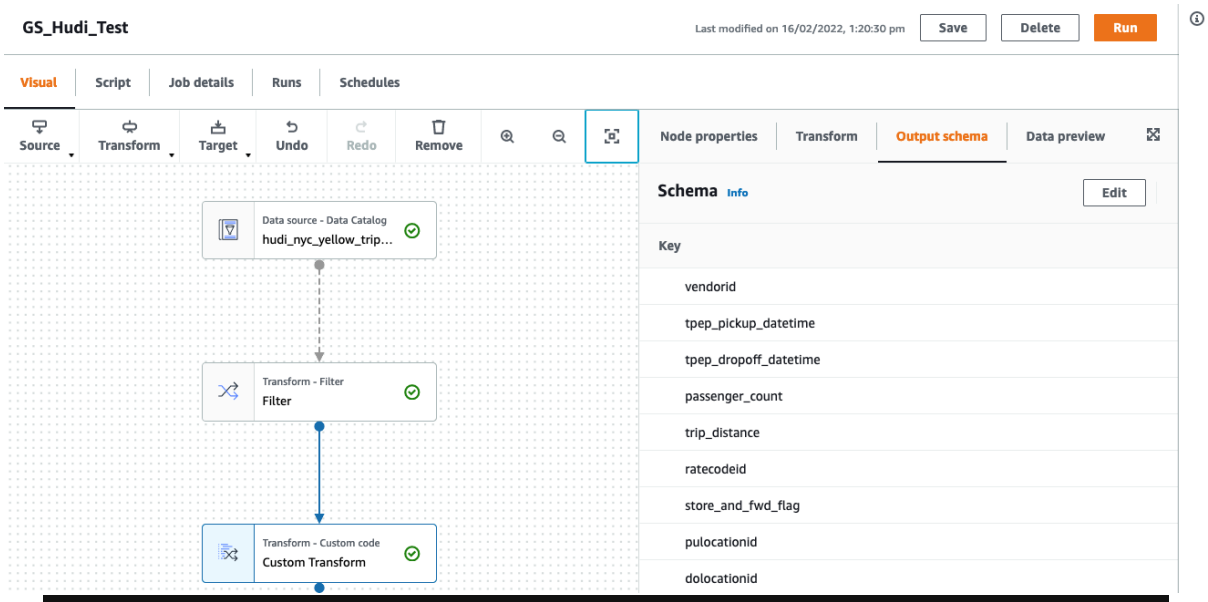
Chapter 4: Data Preparation





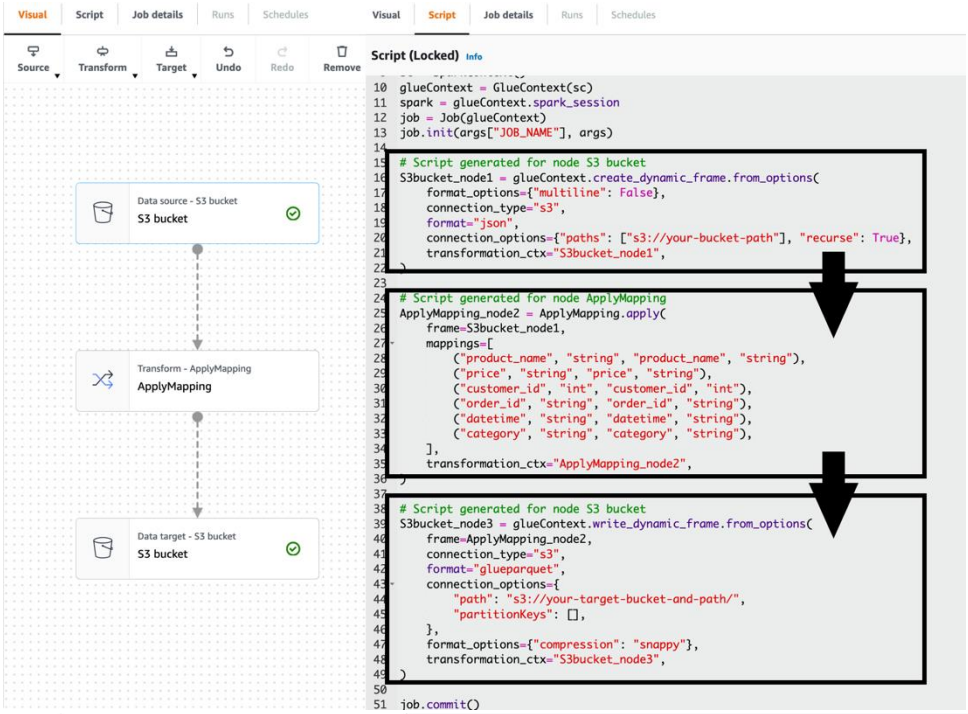


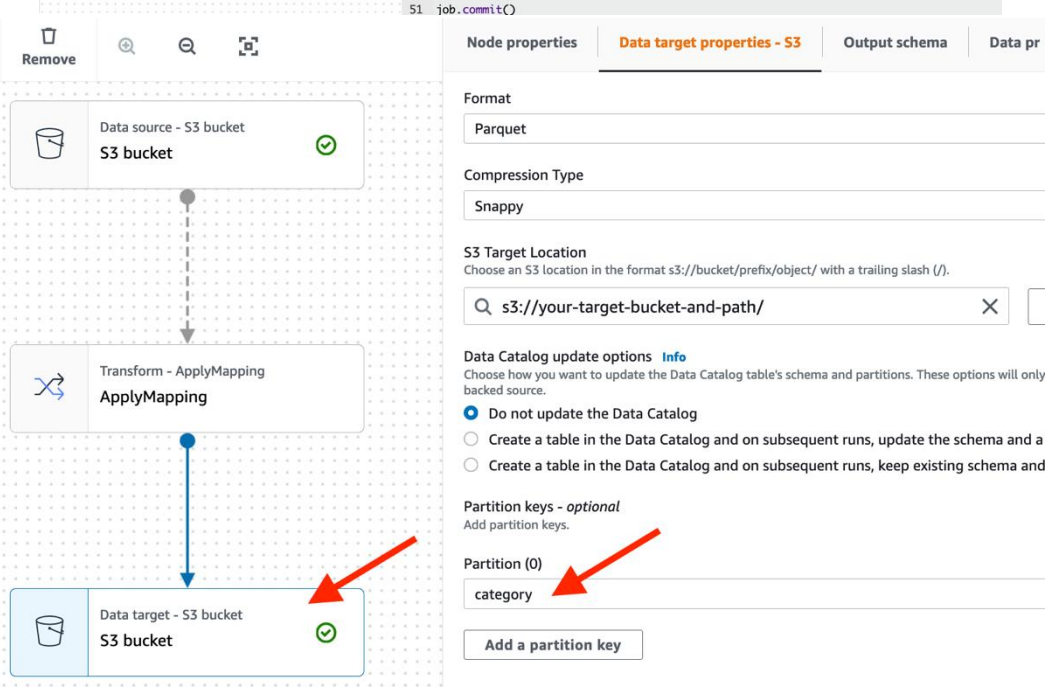

Dataset preview
Data profile overview
Column statistics
Data quality rules
Data lineage
Compare value distribution
Numerical values are rescaled between 0 and 1 to give you a comparative view of the distributions across columns
View: All columns (15)
Normalized value
assembly_session
vote_id
amendment
significant_vote
yes_votes
no_votes
abstain
colonization
human_rights
israel_palestine
disarmament
nuclear_weapons
economic_development
Columns summary (15)
Grid
Schema
All columns
String
Integer
Scroll to column
Column name
Column type
Data quality
Distribution
Box plot
Total valid
Total missing
Distinct values
assembly_session
vote_id
resolution
amendment
vote_date
significant_vote
yes_votes
no_votes
abstain
5429 (100%)
0 (0%)
70 (1%)
5429 (100%)
0 (0%)
5429 (100%)
5278 (97%)
151 (3%)
5270 (97%)
2844 (52%)
2585 (48%)
2 (0%)
5429 (100%)
0 (0%)
780 (14%)
5429 (100%)
0 (0%)
2 (0%)
5429 (100%)
0 (0%)
189 (3%)
5429 (100%)
0 (0%)
77 (1%)
5429 (100%)
0 (0%)
95 (2%)
Dataset preview
Data profile overview
Column statistics
Data quality rules
Data lineage
30
RECOMMENDATIONS
Data quality rules (2)
Expand all
Collapse all
Find
ALL (2)
SUCCEEDED (1)
FAILED (1)
ERROR (0)
DISABLED (0)
united-nations-resolutions-ruleset
2 rules
Failed
Check Dataset For Duplicate Rows
Check if dataset has duplicate rows count <= 0
Check All Columns For Missing Values
Check if all columns has missing values == 0%
SUCCEEDED
7 columns
FAILED
8 columns
ERROR
0 columns
Check All Columns For Missing Values
Check if all columns has missing values == 0%
Failed
SUCCEEDED
7 columns
FAILED
8 columns
ERROR
0 columns
Columns (15)
Find
ALL (15)
SUCCEEDED (7)
FAILED (8)
ERROR (0)
no_votes
vote_id
colonization
abstain
vote_date
significant_vote
israel_palestine
economic_development

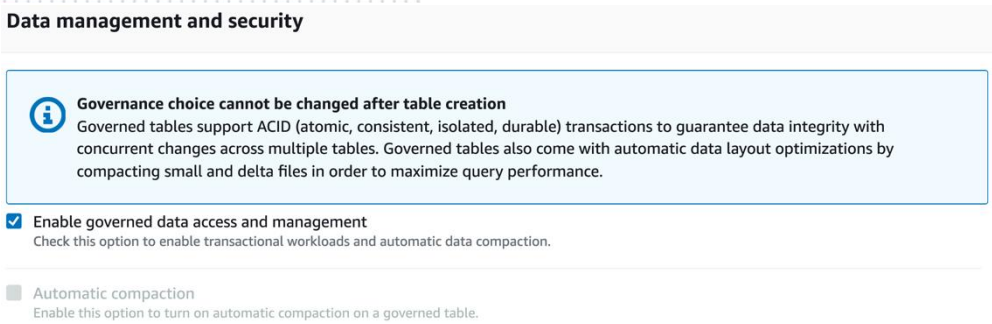



```
+-------------------+-----------+----------+-----+--------------------+---+
|employees.val.name|company    |employees|index|employees.val.email|id |
+-------------------+-----------+----------+-----+--------------------+---+
|foo2               |DummyCorp2|2         |0    |foo@company2.com    |2  |
|bar2               |DummyCorp2|2         |1    |bar@company2.com    |2  |
|foo3               |DummyCorp3|3         |0    |foo@company3.com    |3  |
|bar3               |DummyCorp3|3         |1    |bar@company3.com    |3  |
|foo1               |DummyCorp1|1         |0    |foo@company1.com    |1  |
|bar1               |DummyCorp1|1         |1    |bar@company1.com    |1  |
+-------------------+-----------+----------+-----+--------------------+---+
```

Chapter 5: Designing Data Layouts







aws Services Search for services, features, blogs, docs, and more [Option+S]

Create lifecycle rule

Lifecycle rule configuration

Lifecycle rule name

Enter rule name

Up to 255 characters

Choose a rule scope

- (●) Limit the scope of this rule using one or more filters
- ( ) Apply to all objects in the bucket

Filter type

You can filter objects by prefix, object tags, object size, or whatever combination suits your usecase.

Prefix

Add filter to limit the scope of this rule to a single prefix.

Enter prefix

Don't include the bucket name in the prefix. Using certain characters in key names can cause problems with some applications and protocols. Learn more

Object tags

You can limit the scope of this rule to the key/value pairs added below.

Add tag

Object size

You can limit the scope of this rule to apply to objects based on their size. For example, you can filter out objects that might not be cost effective to transition to Glacier Flexible Retrieval (formerly Glacier) because of per-object fees.

- [ ] Specify minimum object size
- [ ] Specify maximum object size

Lifecycle rule actions

Choose the actions you want this rule to perform. Per-request fees apply. Learn more or see Amazon S3 pricing

- [ ] Move current versions of objects between storage classes
- [ ] Move noncurrent versions of objects between storage classes
- [ ] Expire current versions of objects
- [ ] Permanently delete noncurrent versions of objects
- [ ] Delete expired object delete markers or incomplete multipart uploads

  These actions are not supported when filtering by object tags or object size.

Chapter 6: Data Management

```
root
|-- id: long
|-- index: int
|-- entries.val.id: int
|-- entries.val.values.k1: string
|-- entries.val.values.k2: string

+---+-----+--------------+---------------------+---------------------+
| id|index|entries.val.id|entries.val.values.k1|entries.val.values.k2|
+---+-----+--------------+---------------------+---------------------+
|  1|    0|             1|                  aaa|                  bbb|
|  1|    1|             2|                  ccc|                  ddd|
+---+-----+--------------+---------------------+---------------------+
```

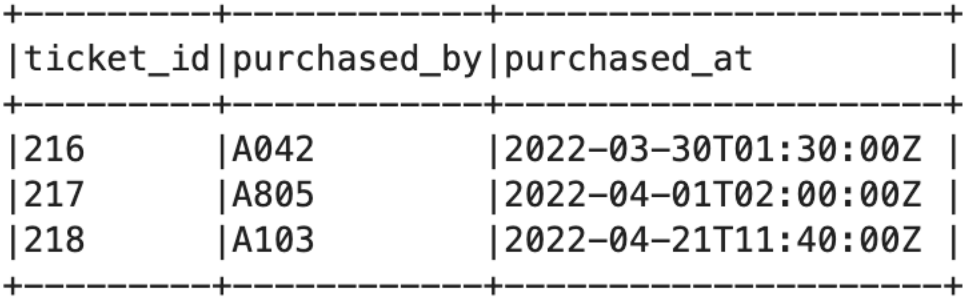
```
root
 |-- count: integer (nullable = true)
 |-- entries: long (nullable = true)
 |-- id: long (nullable = true)
 |-- index: integer (nullable = true)
 |-- entries.val.id: integer (nullable = true)
 |-- entries.val.values.k1: string (nullable = true)
 |-- entries.val.values.k2: string (nullable = true)

+-----+-------+---+-----+--------------+---------------------+---------------------+
|count|entries| id|index|entries.val.id|entries.val.values.k1|entries.val.values.k2|
+-----+-------+---+-----+--------------+---------------------+---------------------+
|    2|      1|  1|    0|             1|                  aaa|                  bbb|
|    2|      1|  1|    1|             2|                  ccc|                  ddd|
+-----+-------+---+-----+--------------+---------------------+---------------------+
```

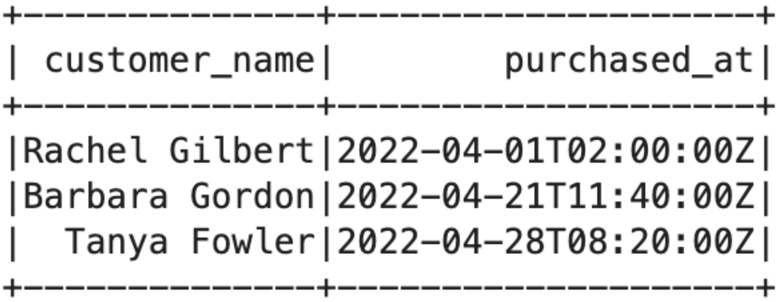
```
+----------+---------------------+------------+-----+
|product_id|        product_name|   category|price|
+----------+---------------------+------------+-----+
|        11|Introduction to C...|     Ebooks|   15|
|        12|Best practices on...|     Ebooks|   25|
|        21|          Data Quest|Video games|   30|
|        22|      Final Shooting|Video games|   20|
+----------+---------------------+------------+-----+
```

```
+----+----------------+--------------------+----------------------+
|uid |customer_name   |email               |phone                 |
+----+----------------+--------------------+----------------------+
|A103|Barbara Gordon  |gordon@example.com  |117.835.2584          |
|A042|Rebecca Thompson|thompson@example.net|001-469-964-3897x9041 |
|A805|Rachel Gilbert  |gilbert@example.com |001-510-198-4613x23986|
|A404|Tanya Fowler    |tanya@example.net   |(067)150-0263         |
+----+----------------+--------------------+----------------------+
```

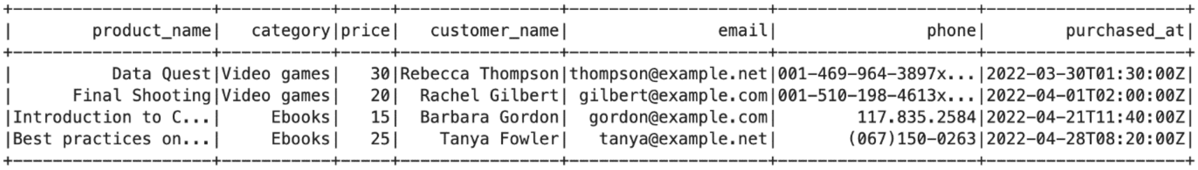
product_id	product_by	purchased_at
21	A042	2022-03-30T01:30:00Z
22	A805	2022-04-01T02:00:00Z
11	A103	2022-04-21T11:40:00Z
12	A404	2022-04-28T08:20:00Z



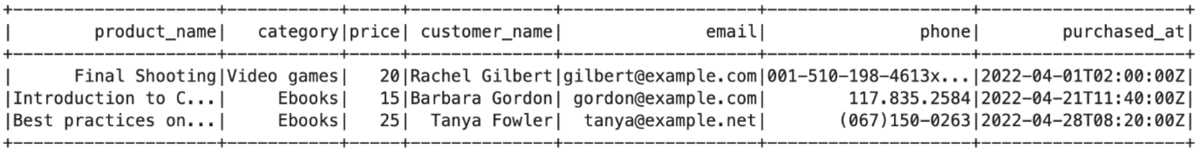
```
+---------+------------+---------------------+
|ticket_id|purchased_by|purchased_at         |
+---------+------------+---------------------+
|216      |A042        |2022-03-30T01:30:00Z |
|217      |A805        |2022-04-01T02:00:00Z |
|218      |A103        |2022-04-21T11:40:00Z |
+---------+------------+---------------------+
```



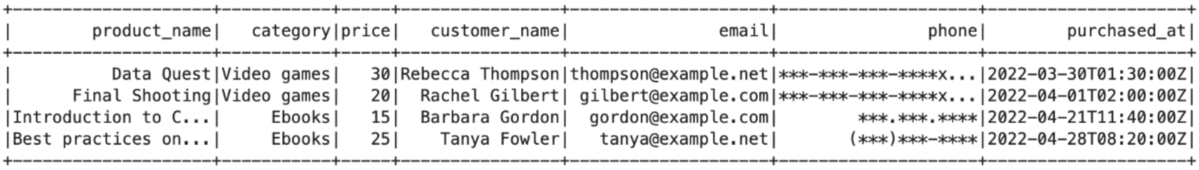
```
+--------------+--------------------+
| customer_name|        purchased_at|
+--------------+--------------------+
|Rachel Gilbert|2022-04-01T02:00:00Z|
|Barbara Gordon|2022-04-21T11:40:00Z|
|  Tanya Fowler|2022-04-28T08:20:00Z|
+--------------+--------------------+
```



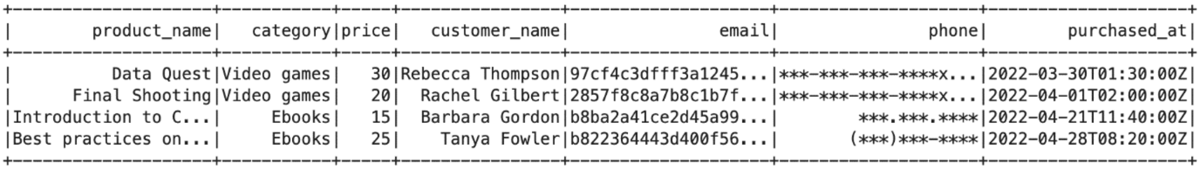
```
+--------------------+-----------+-----+----------------+--------------------+--------------------+--------------------+
|        product_name|   category|price|   customer_name|               email|               phone|        purchased_at|
+--------------------+-----------+-----+----------------+--------------------+--------------------+--------------------+
|          Data Quest|Video games|   30|Rebecca Thompson|thompson@example.net|001-469-964-3897x...|2022-03-30T01:30:00Z|
|      Final Shooting|Video games|   20|  Rachel Gilbert| gilbert@example.com|001-510-198-4613x...|2022-04-01T02:00:00Z|
|Introduction to C...|     Ebooks|   15|  Barbara Gordon|  gordon@example.com|        117.835.2584|2022-04-21T11:40:00Z|
|Best practices on...|     Ebooks|   25|    Tanya Fowler|   tanya@example.net|       (067)150-0263|2022-04-28T08:20:00Z|
+--------------------+-----------+-----+----------------+--------------------+--------------------+--------------------+
```



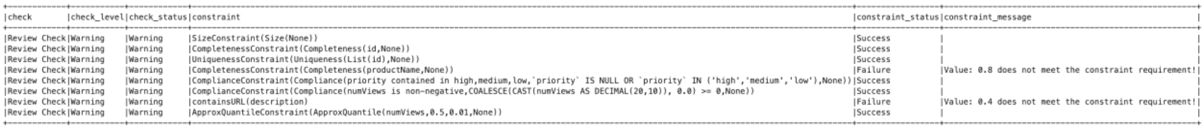
```
+--------------------+-----------+-----+--------------+-------------------+--------------------+--------------------+
|        product_name|   category|price| customer_name|              email|               phone|        purchased_at|
+--------------------+-----------+-----+--------------+-------------------+--------------------+--------------------+
|      Final Shooting|Video games|   20|Rachel Gilbert|gilbert@example.com|001-510-198-4613x...|2022-04-01T02:00:00Z|
|Introduction to C...|     Ebooks|   15|Barbara Gordon| gordon@example.com|        117.835.2584|2022-04-21T11:40:00Z|
|Best practices on...|     Ebooks|   25|  Tanya Fowler|  tanya@example.net|       (067)150-0263|2022-04-28T08:20:00Z|
+--------------------+-----------+-----+--------------+-------------------+--------------------+--------------------+
```



```
+--------------------+-----------+-----+----------------+--------------------+--------------------+--------------------+
|        product_name|   category|price|   customer_name|               email|               phone|        purchased_at|
+--------------------+-----------+-----+----------------+--------------------+--------------------+--------------------+
|          Data Quest|Video games|   30|Rebecca Thompson|thompson@example.net|***-***-***-****x...|2022-03-30T01:30:00Z|
|      Final Shooting|Video games|   20|  Rachel Gilbert| gilbert@example.com|***-***-***-****x...|2022-04-01T02:00:00Z|
|Introduction to C...|     Ebooks|   15|  Barbara Gordon|  gordon@example.com|        ***.***.****|2022-04-21T11:40:00Z|
|Best practices on...|     Ebooks|   25|    Tanya Fowler|   tanya@example.net|       (***)***-****|2022-04-28T08:20:00Z|
+--------------------+-----------+-----+----------------+--------------------+--------------------+--------------------+
```



```
+--------------------+-----------+-----+----------------+--------------------+--------------------+--------------------+
|        product_name|   category|price|   customer_name|               email|               phone|        purchased_at|
+--------------------+-----------+-----+----------------+--------------------+--------------------+--------------------+
|          Data Quest|Video games|   30|Rebecca Thompson|97cf4c3dfff3a1245...|***-***-***-****x...|2022-03-30T01:30:00Z|
|      Final Shooting|Video games|   20|  Rachel Gilbert|2857f8c8a7b8c1b7f...|***-***-***-****x...|2022-04-01T02:00:00Z|
|Introduction to C...|     Ebooks|   15|  Barbara Gordon|b8ba2a41ce2d45a99...|        ***.***.****|2022-04-21T11:40:00Z|
|Best practices on...|     Ebooks|   25|    Tanya Fowler|b822364443d400f56...|       (***)***-****|2022-04-28T08:20:00Z|
+--------------------+-----------+-----+----------------+--------------------+--------------------+--------------------+
```



```
+------------+-----------+------------+----------------------------------------------------------------------------------------------------------------------------------------+-----------------+----------------------------------------------------+
|check       |check_level|check_status|constraint                                                                                                                              |constraint_status|constraint_message                                  |
+------------+-----------+------------+----------------------------------------------------------------------------------------------------------------------------------------+-----------------+----------------------------------------------------+
|Review Check|Warning    |Warning     |SizeConstraint(Size(None))                                                                                                              |Success          |                                                    |
|Review Check|Warning    |Warning     |CompletenessConstraint(Completeness(id,None))                                                                                           |Success          |                                                    |
|Review Check|Warning    |Warning     |UniquenessConstraint(Uniqueness(List(id),None))                                                                                         |Success          |                                                    |
|Review Check|Warning    |Warning     |CompletenessConstraint(Completeness(productName,None))                                                                                  |Failure          |Value: 0.8 does not meet the constraint requirement!|
|Review Check|Warning    |Warning     |ComplianceConstraint(Compliance(priority contained in high,medium,low,`priority` IS NULL OR `priority` IN ('high','medium','low'),None))|Success          |                                                    |
|Review Check|Warning    |Warning     |ComplianceConstraint(Compliance(numViews is non-negative,COALESCE(CAST(numViews AS DECIMAL(20,10)), 0.0) >= 0,None))                    |Success          |                                                    |
|Review Check|Warning    |Warning     |containsURL(description)                                                                                                                |Failure          |Value: 0.4 does not meet the constraint requirement!|
|Review Check|Warning    |Warning     |ApproxQuantileConstraint(ApproxQuantile(numViews,0.5,0.01,None))                                                                        |Success          |                                                    |
+------------+-----------+------------+----------------------------------------------------------------------------------------------------------------------------------------+-----------------+----------------------------------------------------+
```

customer_name	email	phone
Barbara Gordon	gordon@example.com	117.835.2584
Gordon, Barbara	gordon@example.com	117-835-2584
Rebecca Thompson	thompson@example.net	001-469-964-3897x9041
Rachel Gilbert	gilbert@example.com	001-510-198-4613x23986
Gilbert, R.	gilbert@example.com	
Tanya Fowler	tanya@example.net	(067)150-0263

customer_name	email	phone	match_id
Barbara Gordon	gordon@example.com	117.835.2584	1
Gordon, Barbara	gordon@example.com	117-835-2584	1
Rebecca Thompson	thompson@example.net	001-469-964-3897x9041	2
Rachel Gilbert	gilbert@example.com	001-510-198-4613x23986	3
Gilbert, R.	gilbert@example.com		3
Tanya Fowler	tanya@example.net	(067)150-0263	4

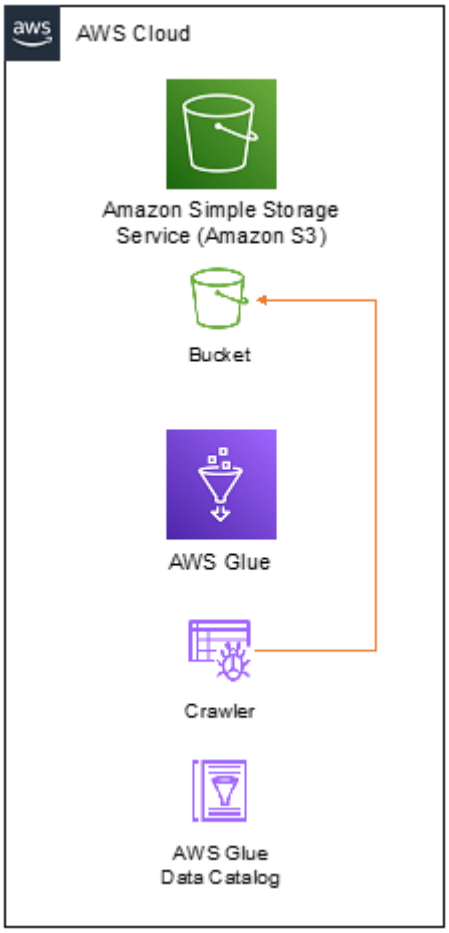
product_id	product_name	category	price
11	Introduction to Cloud	Ebooks	15
12	Best practices on data lakes	Ebooks	25
21	Data Quest	Video games	30
22	Final Shooting	Video games	20

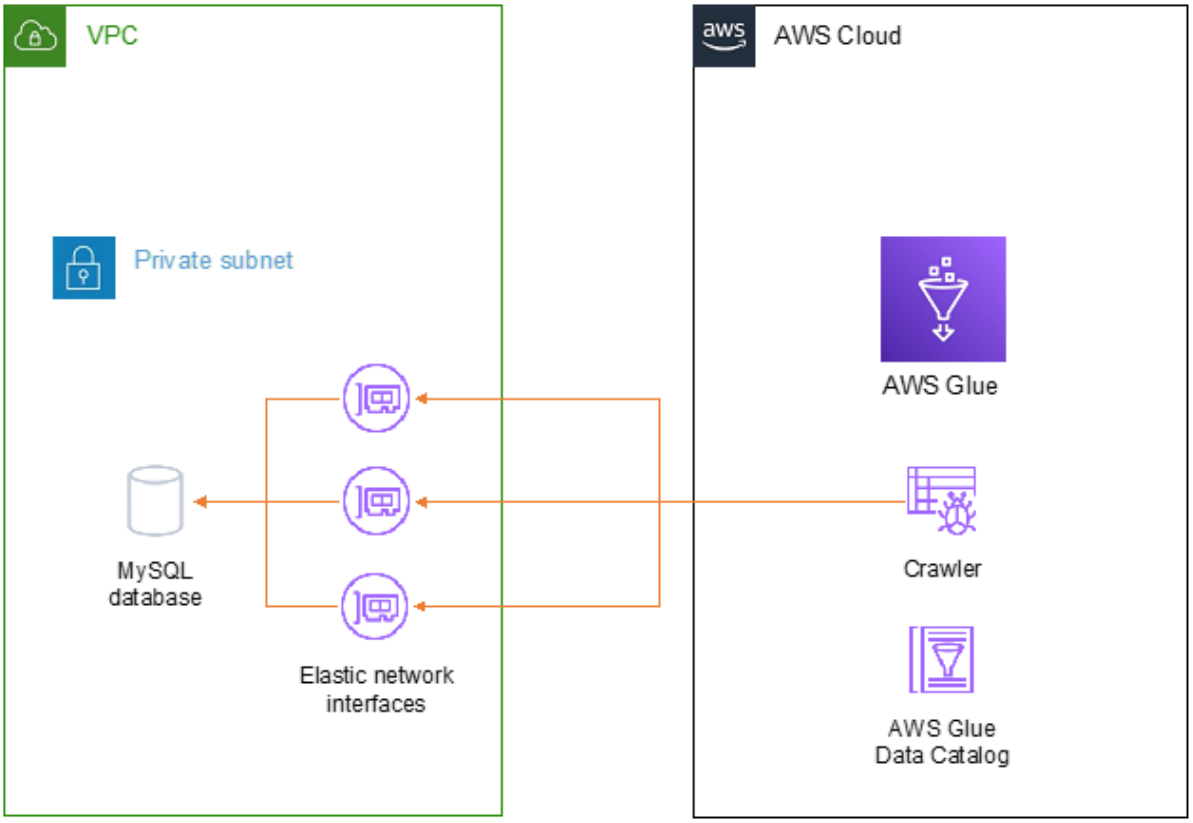
uid	customer_name	email	phone
A103	Barbara Gordon	gordon@example.com	117.835.2584
A042	Rebecca Thompson	thompson@example.net	001-469-964-3897x9041
A805	Rachel Gilbert	gilbert@example.com	001-510-198-4613x23986
A404	Tanya Fowler	tanya@example.net	(067)150-0263

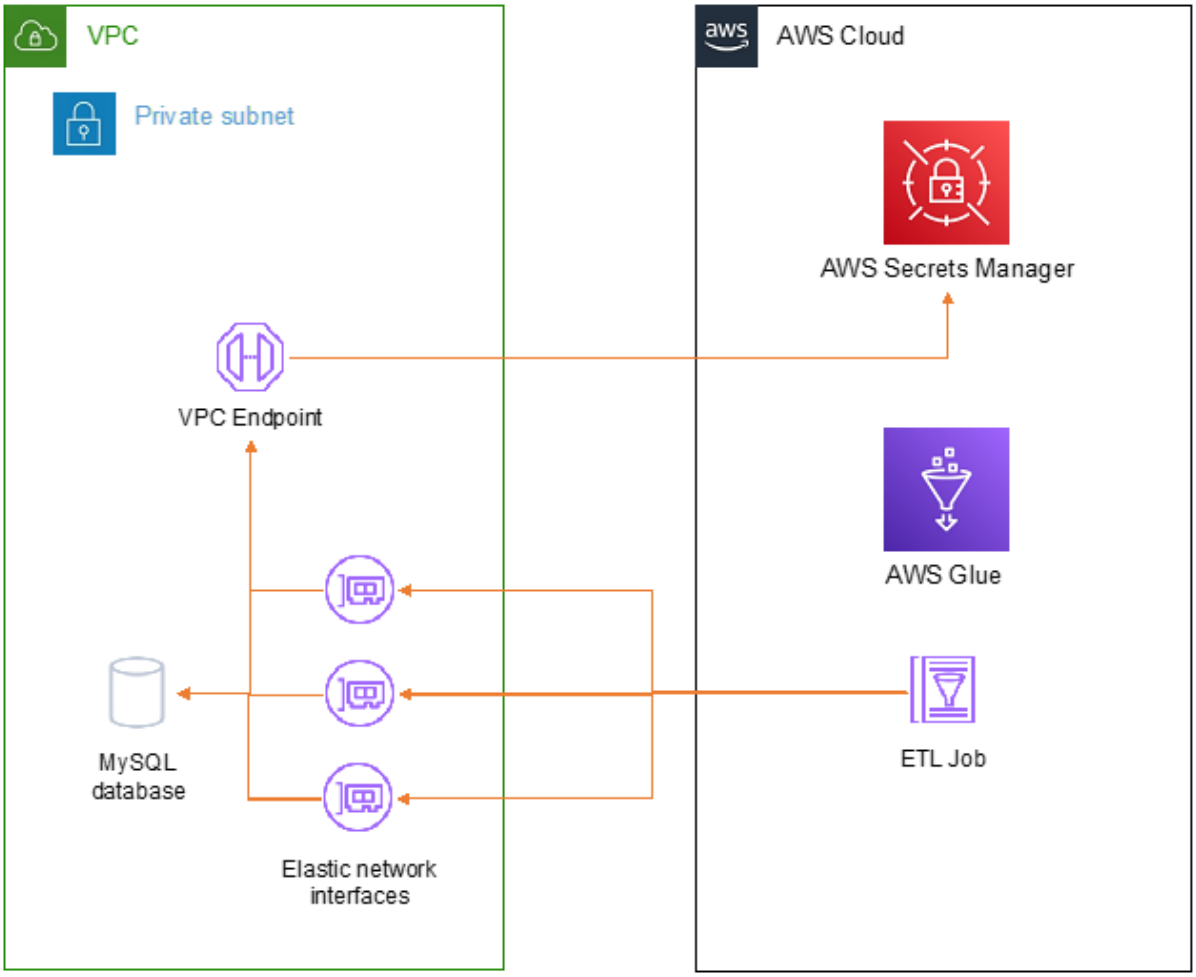
Chapter 7: Metadata Management

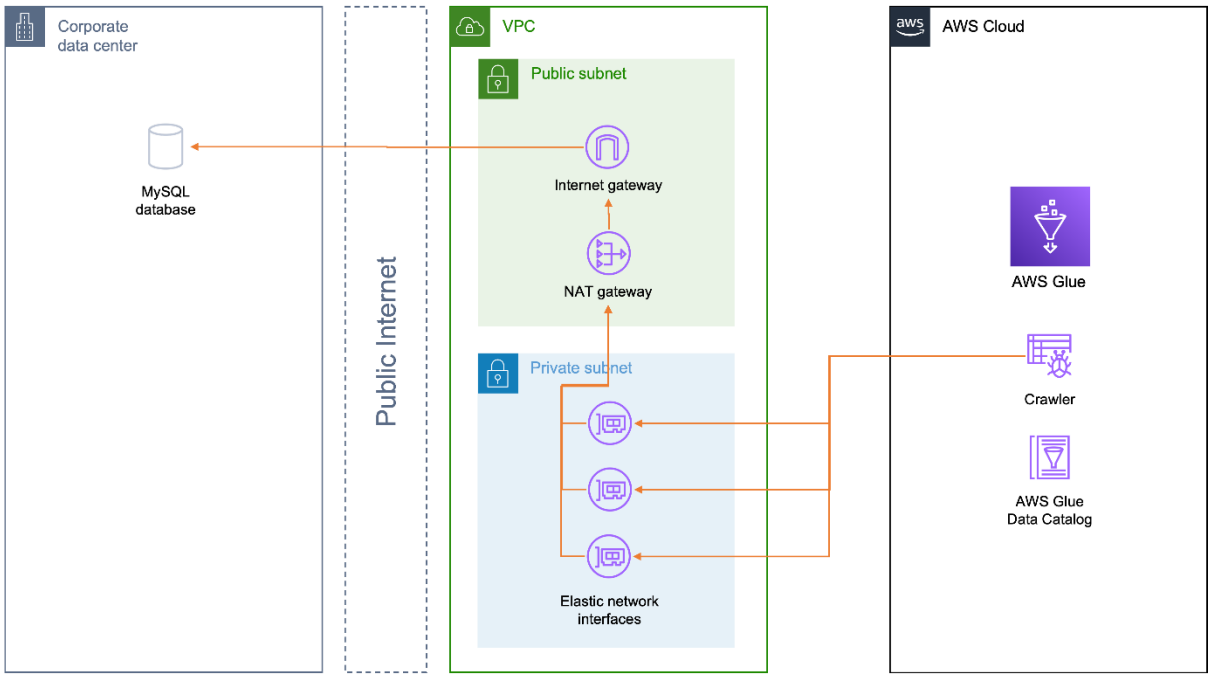
No images…

Chapter 8: Data Security

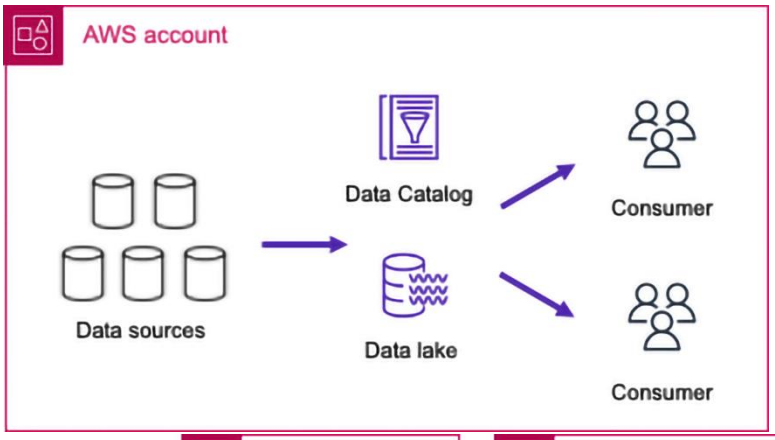

AWS Cloud
Amazon Simple Storage Service (Amazon S3)
Bucket
AWS Glue
Crawler
AWS Glue Data Catalog

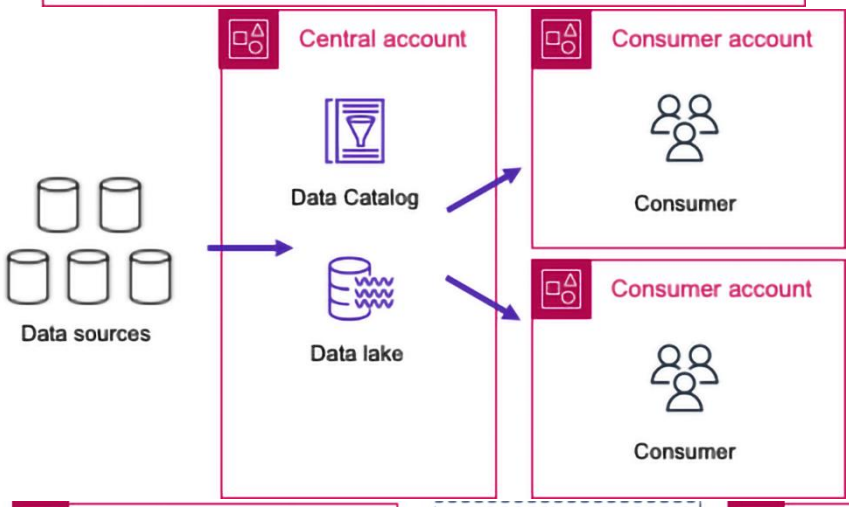

VPC
Private subnet
MySQL database
Elastic network interfaces
AWS Cloud
AWS Glue
Crawler
AWS Glue Data Catalog

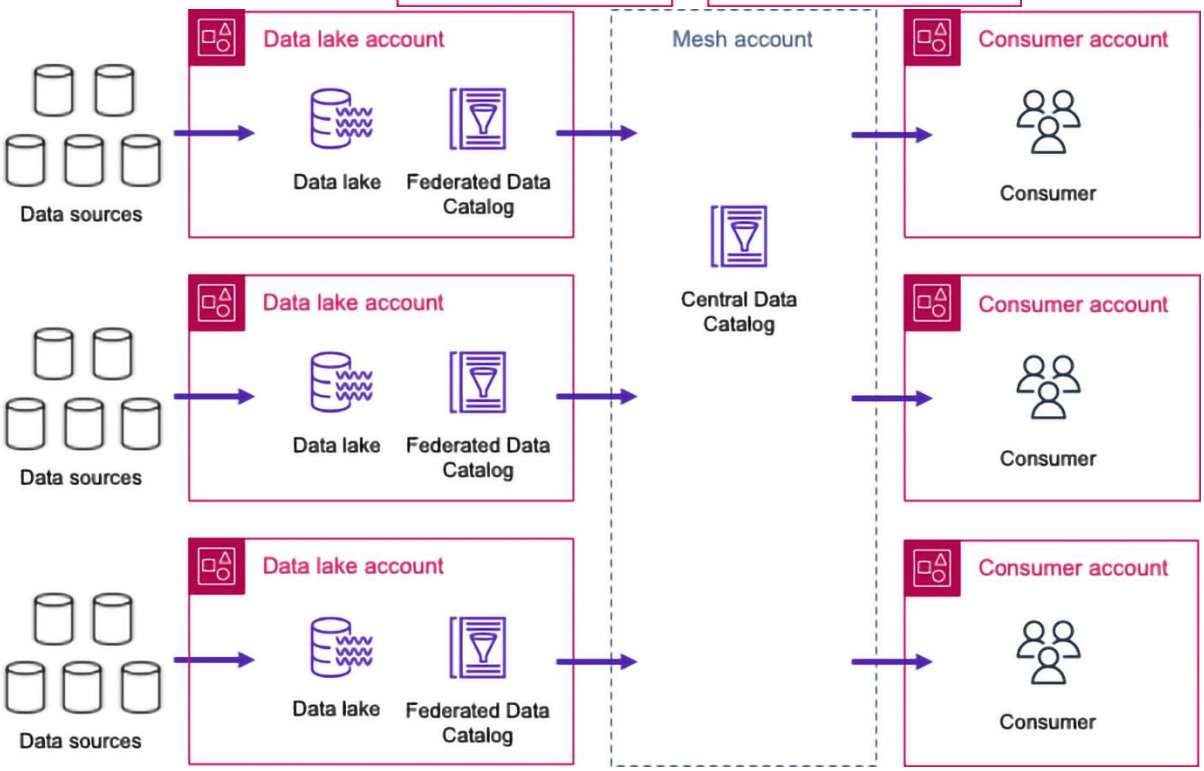

VPC
Private subnet
AWS Cloud
AWS Secrets Manager
VPC Endpoint
AWS Glue
MySQL
database
ETL Job
Elastic network
interfaces


Corporate
data center
MySQL
database
Public Internet
VPC
Public subnet
Internet gateway
NAT gateway
Private subnet
Elastic network
interfaces
AWS Cloud
AWS Glue
Crawler
AWS Glue
Data Catalog

Chapter 9: Data Sharing







Editor | Recent queries | Saved queries | Settings

Workgroup primary

Data

Data Source
AwsDataCatalog

Database
simple_datalake

Tables and views | Create

Filter tables and views

Tables (1)

pcs

Views (0)

Query 1

```
1 SELECT * FROM simple_datalake.pcs;
```

SQL Ln 1, Col 34

Run again | Cancel | Save | Clear | Create

Completed — Time in queue: 0.177 sec — Run time: 0.484 sec — Data scanned: 0.78 KB

Results (4)

Copy | Download results

Search rows

#	product_name	category	price	purchased_at	customer_name	email	phone
1	Introduction to Cloud	Ebooks	15	2022-04-21T11:40:00Z	Barbara Gordon	gordon@example.com	117.835.2584
2	Best practices on data lakes	Ebooks	25	2022-04-28T08:20:00Z	Tanya Fowler	tanya@example.net	(067)150-0263
3	Data Quest	Video games	30	2022-03-30T01:30:00Z	Rebecca Thompson	thompson@example.net	001-469-964-3897x9041
4	Final Shooting	Video games	20	2022-04-01T02:00:00Z	Rachel Gilbert	gilbert@example.com	001-510-198-4613x23986

Editor | Recent queries | Saved queries | Settings

Workgroup primary

Data

Data Source
producer_catalog

Database
simple_datalake

Tables and views | Create

Filter tables and views

Tables (1)

pcs

Views (0)

Query 1 | Query 2

```
1 SELECT * FROM simple_datalake.pcs
```

SQL Ln 1, Col 34

Run again | Cancel | Save | Clear | Create

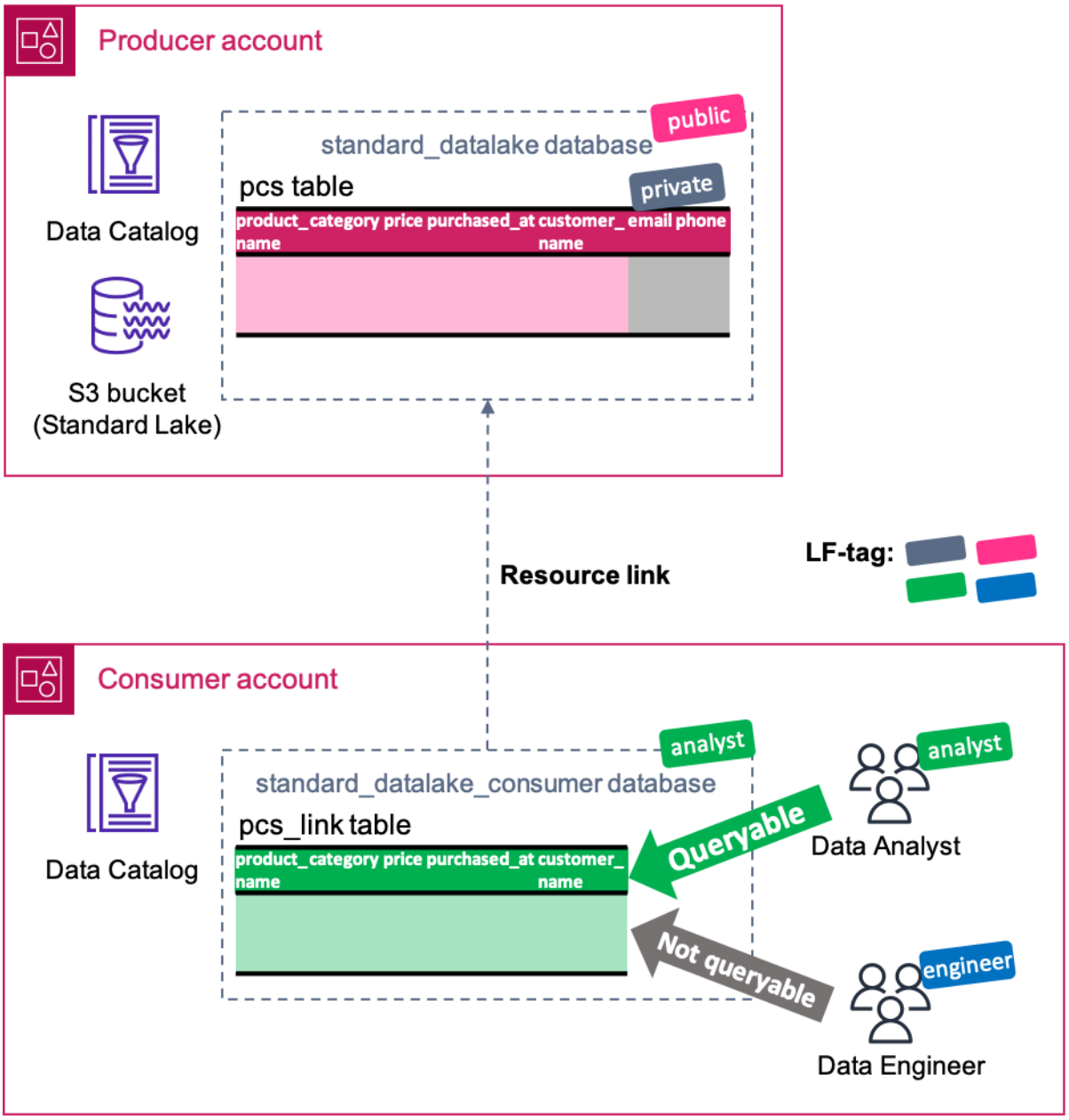
Completed — Time in queue: 0.515 sec — Run time: 0.892 sec — Data scanned: 0.78 KB

Results (4)

Copy | Download results

Search rows

#	product_name	category	price	purchased_at	customer_name	email	phone
1	Introduction to Cloud	Ebooks	15	2022-04-21T11:40:00Z	Barbara Gordon	gordon@example.com	117.835.2584
2	Best practices on data lakes	Ebooks	25	2022-04-28T08:20:00Z	Tanya Fowler	tanya@example.net	(067)150-0263
3	Data Quest	Video games	30	2022-03-30T01:30:00Z	Rebecca Thompson	thompson@example.net	001-469-964-3897x9041
4	Final Shooting	Video games	20	2022-04-01T02:00:00Z	Rachel Gilbert	gilbert@example.com	001-510-198-4613x23986



Default permissions for newly created databases and tables

These settings maintain existing AWS Glue Data Catalog behavior. You can still set individual permissions on databases and tables, which will take effect when you revoke the Super permission from IAMAllowedPrincipals. See **Changing Default Settings for Your Data Lake.**

- [ ] Use only IAM access control for new databases
- [ ] Use only IAM access control for new tables in new databases

AWS Lake Formation > Data lake locations > Register location

Register location

Amazon S3 location

Register an Amazon S3 path as the storage location for your data lake.

Amazon S3 path

Choose an Amazon S3 path for your data lake.

e.g.: s3://bucket/prefix/ Browse

Review location permissions - strongly recommended

Registering the selected location may result in your users gaining access to data already at that location. Before registering a location, we recommend that you review existing location permissions on resources in that location.

Review location permissions

IAM role

To add or update data, Lake Formation needs read/write access to the chosen Amazon S3 path. Choose a role that you know has permission to do this, or choose the **AWSServiceRoleForLakeFormationDataAccess** service-linked role. When you register the first Amazon S3 path, the service-linked role and a new inline policy are created on your behalf. Lake Formation adds the first path to the inline policy and attaches it to the service-linked role. When you register subsequent paths, Lake Formation adds the path to the existing policy.

AWSServiceRoleForLakeFormationDataAccess

⚠ Do not select the service linked role if you plan to use EMR.

Cancel Register location

Add LF-Tag Learn More

LF-Tags have a key and one or more values that can be associated with data catalog resources. Tables automatically inherit from database LF-tags, and columns inherit from table LF-tags.

Example: Key = Confidentiality | Values = private, sensitive, public

Key

Confidentiality

Key string must be less than 128 characters long, and cannot be changed once LF-tag is created.

Values

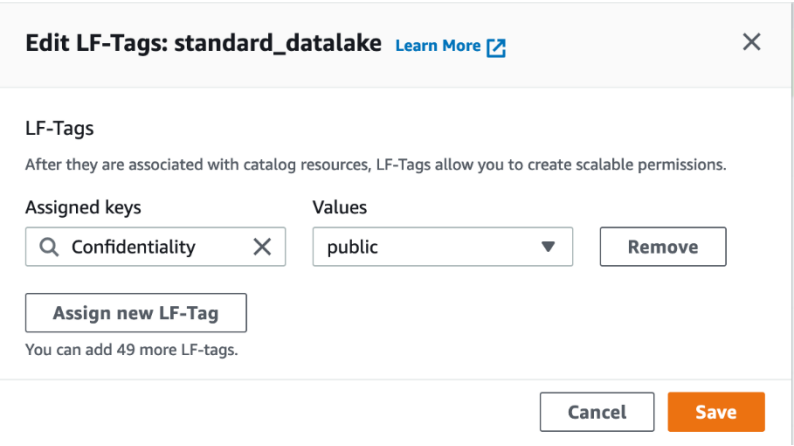
Type a single value and select [Enter] or specify multiple values separated by commas.

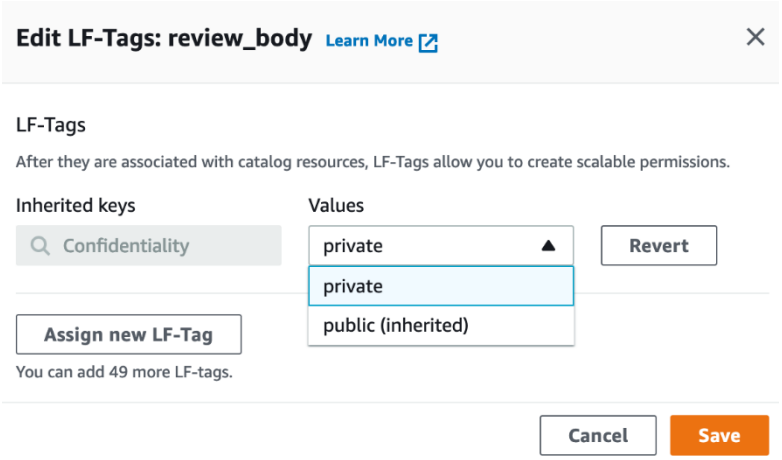
Add

private × public ×

Enter up to 15 values; each value must be less than 256 characters long.

Cancel Add LF-tag


Edit LF-Tags: standard_datalake
Learn More
LF-Tags
After they are associated with catalog resources, LF-Tags allow you to create scalable permissions.
Assigned keys
Values
Confidentiality
public
Remove
Assign new LF-Tag
You can add 49 more LF-tags.
Cancel
Save


Edit LF-Tags: review_body
Learn More
LF-Tags
After they are associated with catalog resources, LF-Tags allow you to create scalable permissions.
Inherited keys
Values
Confidentiality
private
Revert
private
public (inherited)
Assign new LF-Tag
You can add 49 more LF-tags.
Cancel
Save

AWS Lake Formation > LF-tag permissions > Grant LF-tag permissions

Grant LF-tag permissions

Select the principals to grant permissions to, the LF-Tags to grant permissions on, and the specific set of permissions.

Principals

- IAM users and roles
  Users or roles from this AWS account.
- SAML users and groups
  SAML users and group or QuickSight ARNs.
- External accounts
  AWS accounts or AWS organizations outside of this account.

AWS account or AWS organization
Enter one or more AWS account IDs or AWS organization IDs. Press Enter after each ID.

Choose AWS account ID or AWS organization ID

222222222222 ×
Account

LF-Tags

LF-tag permission scope
Choose to grant permissions on all or a subset of LF-Tags.

Key	Values	
Confidentiality ×	Choose LF-tag values ▼ public ×	Remove

Add LF-Tag

Permissions

LF-tag permissions
Select the specific access permissions to grant.

- [x] Describe
- [ ] Associate

Grantable permissions
Select the permissions that the grant recipient can grant to other principals.

- [ ] Describe
- [ ] Associate

Cancel Grant

AWS Lake Formation > Grant permissions

Grant data permissions

Principals

- ○ IAM users and roles
  Users or roles from this AWS account.
- ○ SAML users and groups
  SAML users and group or QuickSight ARNs.
- ◉ External accounts
  AWS accounts or AWS organizations outside of this account.

AWS account or AWS organization
Enter one or more AWS account IDs or AWS organization IDs. Press Enter after each ID.

Choose AWS account ID or AWS organization ID

222222222222 ×
Account

Granting data permissions to organizations is not supported when granting permissions by using LF-Tags.

LF-Tags or catalog resources

- ◉ Resources matched by LF-Tags (recommended)
  Manage permissions indirectly for resources or data matched by a specific set of LF-Tags.
- ○ Named data catalog resources
  Manager permissions for specific databases or tables, in addition to fine-grained data access.

Key	Values	
Confidentiality ×	Choose LF-tag values ▼ public ×	Remove

Add LF-Tag

Database permissions

Database permissions
Choose specific access permissions to grant.

- ☐ Create table
- ☐ Alter
- ☐ Drop
- ☑ Describe
- ☐ Super
  This permission is the union of all the individual permissions to the left, and supersedes them.

Grantable permissions
Choose the permission that may be granted to others.

- ☐ Create table
- ☐ Alter
- ☐ Drop
- ☑ Describe
- ☐ Super
  This permission allows the principal to grant any of the permissions to the left, and supersedes those grantable permissions.

AWS Lake Formation > Tables > Create resource link

Create resource link

Table resource link details

Create a table resource link in the AWS Glue Data Catalog.

Resource link name

pcs_link

Name may contain letters (A-Z), numbers (0-9), hyphens (-), or underscores (_), and must be less than 256 characters long.

Database

Resource link will be contained in this database.

standard_datalake_consumer

Shared table

Enter or choose a shared table.

pcs

Shared table's database

Enter the database containing the shared table.

standard_datalake

Shared table's owner ID

Enter the AWS account ID of the shared table owner.

244523610907

Cancel | Create

Amazon Athena > Query editor

Editor | Recent queries | Saved queries | Settings

Workgroup primary

Data

Data Source: AwsDataCatalog

Database: standard_datalake_consumer

Tables and views | Create

Filter tables and views

Tables (1)

pcs_link

- product_name string
- category string
- price int
- purchased_at string
- customer_name string

Views (0)

Query 1

```
SELECT * FROM standard_datalake_consumer.pcs_link
```

SQL Ln 1, Col 50

Run again | Cancel | Save | Clear | Create

Completed — Time in queue: 0.12 sec — Run time: 2.122 sec — Data scanned: 0.78 KB

Results (4) — Copy | Download results

Search rows

#	product_name	category	price	purchased_at	customer_name
1	Introduction to Cloud	Ebooks	15	2022-04-21T11:40:00Z	Barbara Gordon
2	Best practices on data lakes	Ebooks	25	2022-04-28T08:20:00Z	Tanya Fowler
3	Data Quest	Video games	30	2022-03-30T01:30:00Z	Rebecca Thompson
4	Final Shooting	Video games	20	2022-04-01T02:00:00Z	Rachel Gilbert

Add LF-Tag Learn More

LF-Tags have a key and one or more values that can be associated with data catalog resources. Tables automatically inherit from database LF-tags, and columns inherit from table LF-tags.
Example: Key = Confidentiality | Values = private, sensitive, public

Key

Division

Key string must be less than 128 characters long, and cannot be changed once LF-tag is created.

Values

Type a single value and select [Enter] or specify multiple values separated by commas.

Add

analyst engineer

Enter up to 15 values; each value must be less than 256 characters long.

Cancel Add LF-tag

Edit LF-Tags: standard_datalake_consumer Learn More

LF-Tags

After they are associated with catalog resources, LF-Tags allow you to create scalable permissions.

Assigned keys	Values	
Division	analyst	Remove

Assign new LF-Tag

You can add 49 more LF-tags.

Cancel Save

AWS Lake Formation > LF-tag permissions > Grant LF-tag permissions

Grant LF-tag permissions

Select the principals to grant permissions to, the LF-Tags to grant permissions on, and the specific set of permissions.

Principals

- (●) IAM users and roles — Users or roles from this AWS account.
- ( ) SAML users and groups — SAML users and group or QuickSight ARNs.
- ( ) External accounts — AWS accounts or AWS organizations outside of this account.

IAM users and roles

Add one or more IAM users or roles.

Choose IAM principals to add ▼

DataAnalyst ✕
User

LF-Tags

LF-tag permission scope

Choose to grant permissions on all or a subset of LF-Tags.

Key	Values	
Division ✕	*Choose LF-tag values* ▼ analyst ✕	Remove

Add LF-Tag

Permissions

LF-tag permissions

Select the specific access permissions to grant.

- [x] Describe
- [ ] Associate

Grantable permissions

Select the permissions that the grant recipient can grant to other principals.

- [ ] Describe
- [ ] Associate

Cancel | Grant

AWS Lake Formation > Grant permissions

Grant data permissions

Principals

- (●) **IAM users and roles** — Users or roles from this AWS account.
- ( ) **SAML users and groups** — SAML users and group or QuickSight ARNs.
- ( ) **External accounts** — AWS accounts or AWS organizations outside of this account.

IAM users and roles
Add one or more IAM users or roles.

Choose IAM principals to add

DataAnalyst ×
User

LF-Tags or catalog resources

- (●) **Resources matched by LF-Tags (recommended)** — Manage permissions indirectly for resources or data matched by a specific set of LF-Tags.
- ( ) **Named data catalog resources** — Manager permissions for specific databases or tables, in addition to fine-grained data access.

Key	Values	
Division ×	*Choose LF-tag values* ▼ analyst ×	Remove

Add LF-Tag

Database permissions

Database permissions
Choose specific access permissions to grant.

- [ ] Create table
- [ ] Alter
- [ ] Drop
- [x] Describe
- [ ] Super — This permission is the union of all the individual permissions to the left, and supersedes them.

Grantable permissions
Choose the permission that may be granted to others.

- [ ] Create table
- [ ] Alter
- [ ] Drop
- [ ] Describe
- [ ] Super — This permission allows the principal to grant any of the permissions to the left, and supersedes those grantable

Amazon Athena > Query editor

Editor | Recent queries | Saved queries | Settings

Workgroup: primary

Data

Data Source: AwsDataCatalog

Database: standard_datalake_consumer

Tables and views — Create

Filter tables and views

Tables (1)

- pcs_link
  - product_name — string
  - category — string
  - price — int
  - purchased_at — string
  - customer_name — string

Views (0)

Query 1

```
SELECT * FROM standard_datalake_consumer.pcs_link
```

SQL Ln 1, Col 50

Run again | Cancel | Save | Clear | Create

Completed — Time in queue: 0.12 sec — Run time: 2.122 sec — Data scanned: 0.78 KB

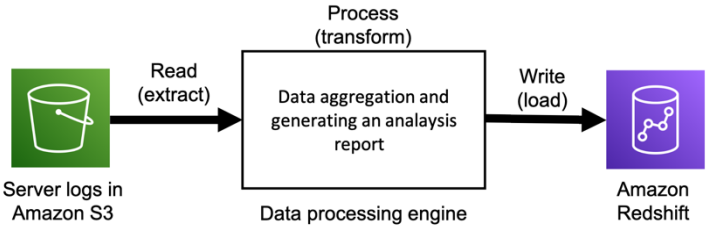
Results (4)

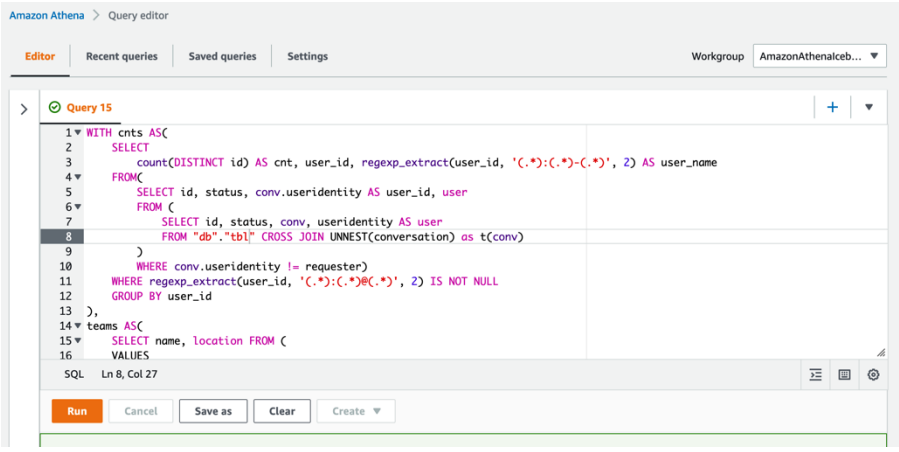
Copy | Download results

Search rows

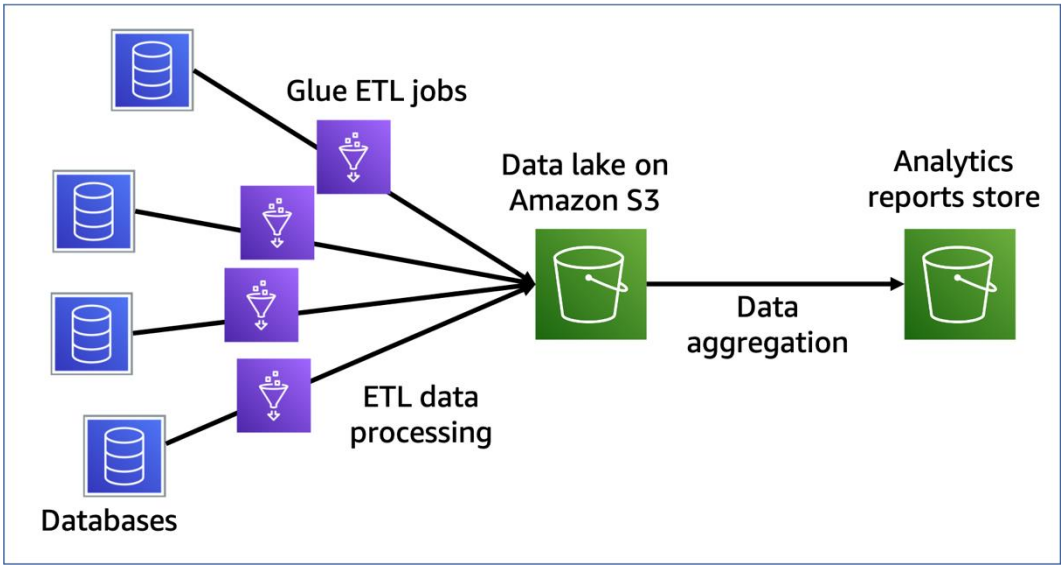
#	product_name	category	price	purchased_at	customer_name
1	Introduction to Cloud	Ebooks	15	2022-04-21T11:40:00Z	Barbara Gordon
2	Best practices on data lakes	Ebooks	25	2022-04-28T08:20:00Z	Tanya Fowler
3	Data Quest	Video games	30	2022-03-30T01:30:00Z	Rebecca Thompson
4	Final Shooting	Video games	20	2022-04-01T02:00:00Z	Rachel Gilbert

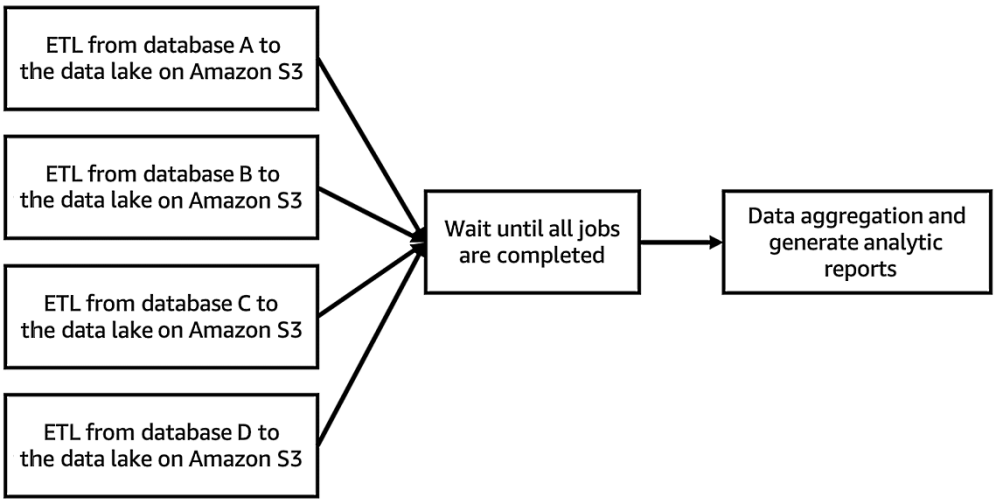
Chapter 10: Data Pipeline Management

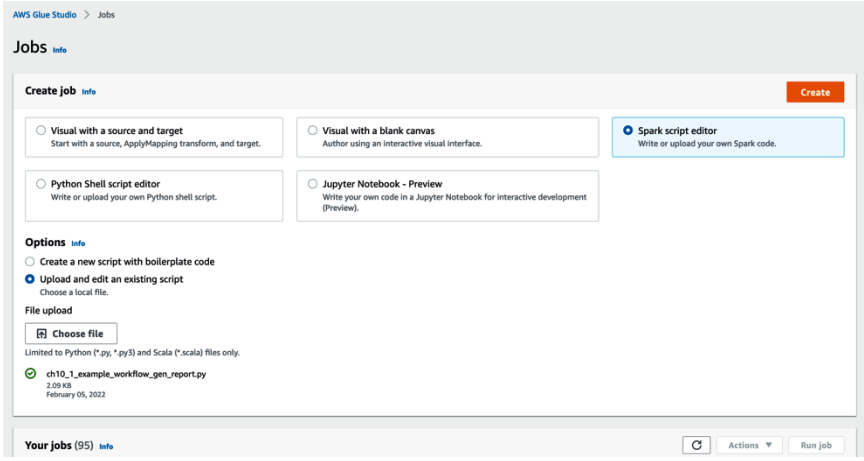


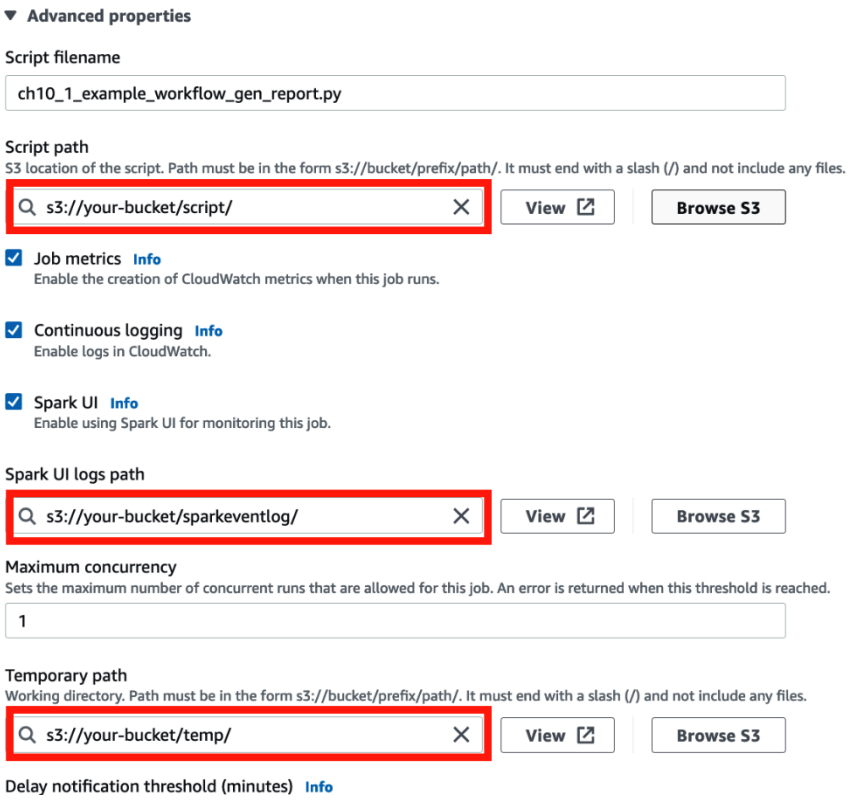


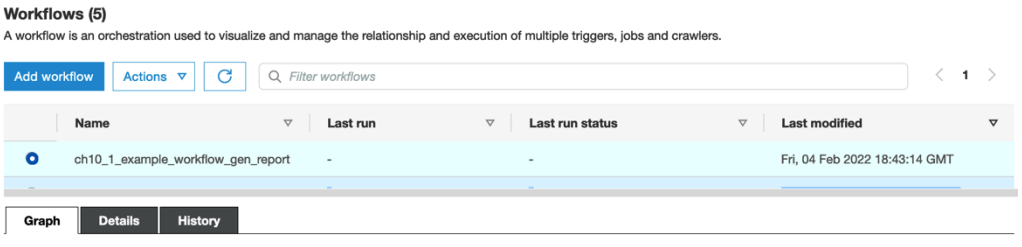
Allocated Resources per Worker (at the time of writing)	Standard	G.1X	G.2X
Memory	16 GB	16 GB	32 GB
vCPUs	4	4	8
Disk	50 GB	64 GB	128 GB

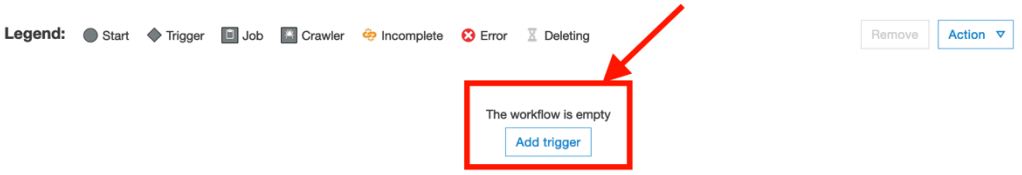


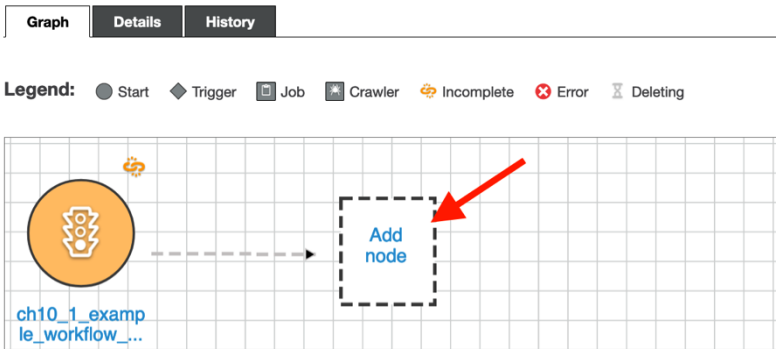

ETL from database A to the data lake on Amazon S3
ETL from database B to the data lake on Amazon S3
ETL from database C to the data lake on Amazon S3
ETL from database D to the data lake on Amazon S3
Wait until all jobs are completed
Data aggregation and generate analytic reports

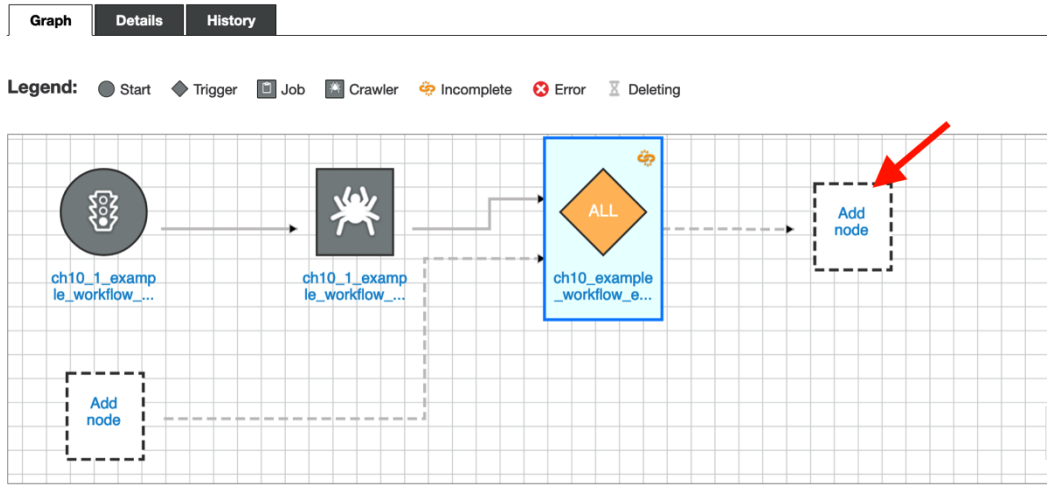

AWS Glue Studio > Jobs
Jobs Info
Create job Info
Create
Visual with a source and target
Start with a source, ApplyMapping transform, and target.
Visual with a blank canvas
Author using an interactive visual interface.
Spark script editor
Write or upload your own Spark code.
Python Shell script editor
Write or upload your own Python shell script.
Jupyter Notebook - Preview
Write your own code in a Jupyter Notebook for interactive development (Preview).
Options Info
Create a new script with boilerplate code
Upload and edit an existing script
Choose a local file.
File upload
Choose file
Limited to Python (*.py, *.py3) and Scala (*.scala) files only.
ch10_1_example_workflow_gen_report.py
2.09 KB
February 05, 2022
Your jobs (95) Info
Actions
Run job

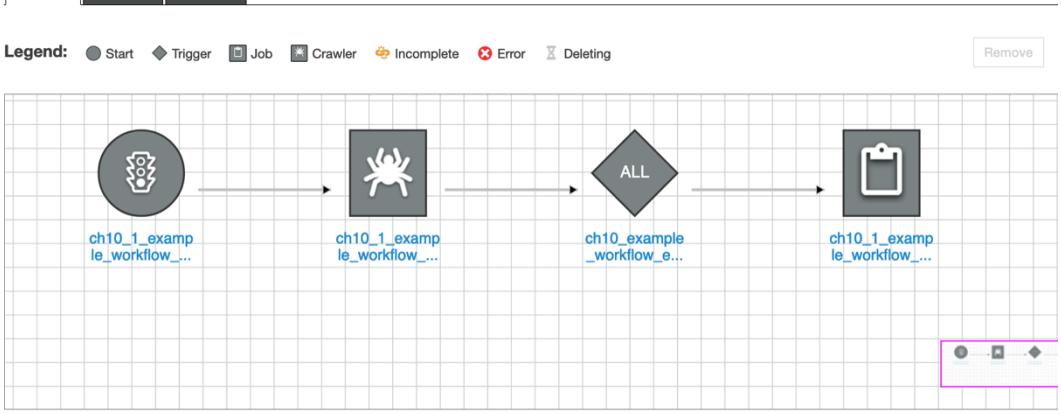

Advanced properties
Script filename
ch10_1_example_workflow_gen_report.py
Script path
S3 location of the script. Path must be in the form s3://bucket/prefix/path/. It must end with a slash (/) and not include any files.
s3://your-bucket/script/
View
Browse S3
Job metrics Info
Enable the creation of CloudWatch metrics when this job runs.
Continuous logging Info
Enable logs in CloudWatch.
Spark UI Info
Enable using Spark UI for monitoring this job.
Spark UI logs path
s3://your-bucket/sparkeventlog/
View
Browse S3
Maximum concurrency
Sets the maximum number of concurrent runs that are allowed for this job. An error is returned when this threshold is reached.
1
Temporary path
Working directory. Path must be in the form s3://bucket/prefix/path/. It must end with a slash (/) and not include any files.
s3://your-bucket/temp/
View
Browse S3
Delay notification threshold (minutes) Info

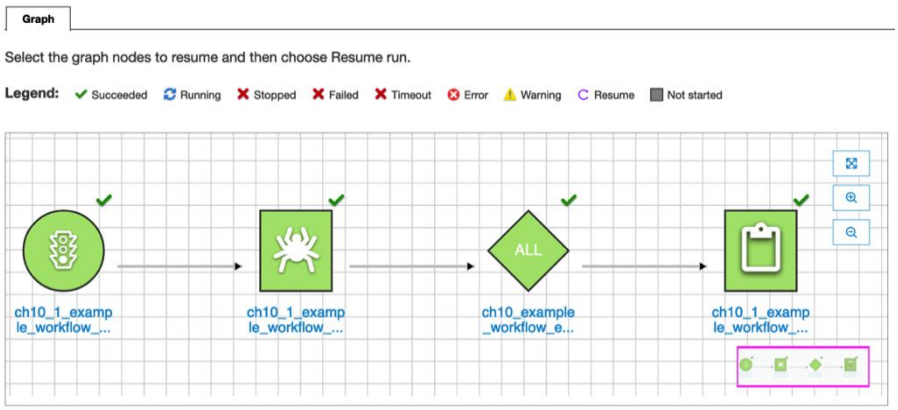

Workflows (5)
A workflow is an orchestration used to visualize and manage the relationship and execution of multiple triggers, jobs and crawlers.
Add workflow
Actions
Filter workflows
1
Name
Last run
Last run status
Last modified
ch10_1_example_workflow_gen_report
-
-
Fri, 04 Feb 2022 18:43:14 GMT
Graph
Details
History


Legend:
Start
Trigger
Job
Crawler
Incomplete
Error
Deleting
Remove
Action
The workflow is empty
Add trigger

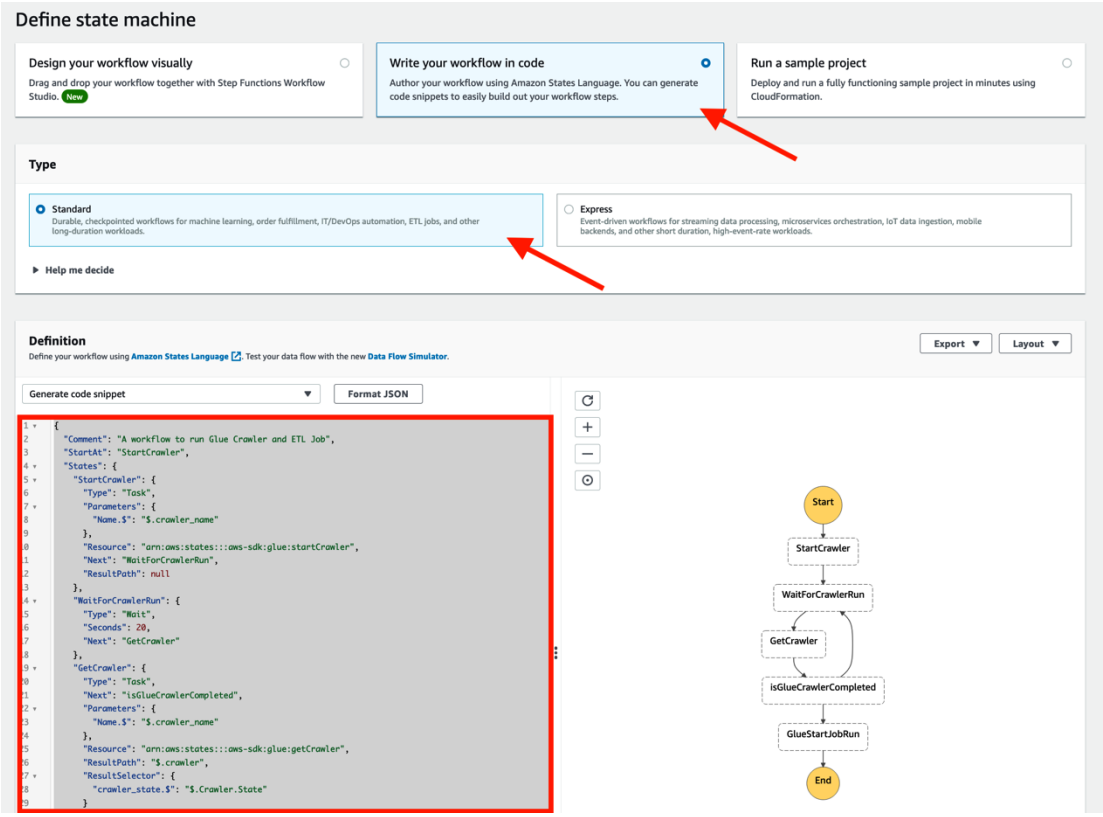

Graph
Details
History
Legend:
Start
Trigger
Job
Crawler
Incomplete
Error
Deleting
Add node
ch10_1_example_workflow_...


Graph
Details
History
Legend:
Start
Trigger
Job
Crawler
Incomplete
Error
Deleting
ch10_1_example_workflow_...
ch10_1_example_workflow_...
ALL
ch10_example_workflow_e...
Add node
Add node


Graph
Details
History
Legend:
Start
Trigger
Job
Crawler
Incomplete
Error
Deleting
Remove
ch10_1_example_workflow_...
ch10_1_example_workflow_...
ALL
ch10_example_workflow_e...
ch10_1_example_workflow_...



category	total_sales	report_year
drink	906	2021
grocery	983	2021
kitchen	4535	2021
office	1316	2021



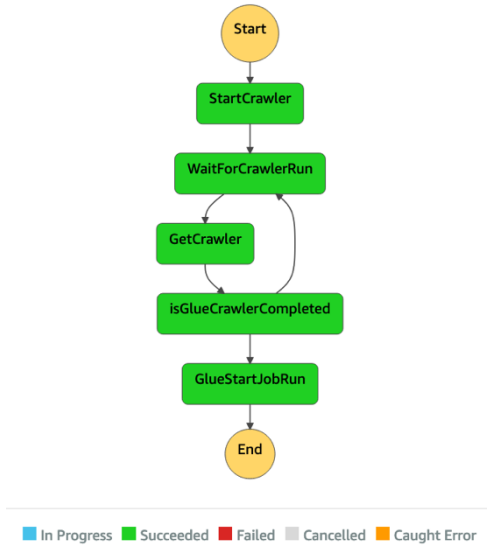
Name - *optional*

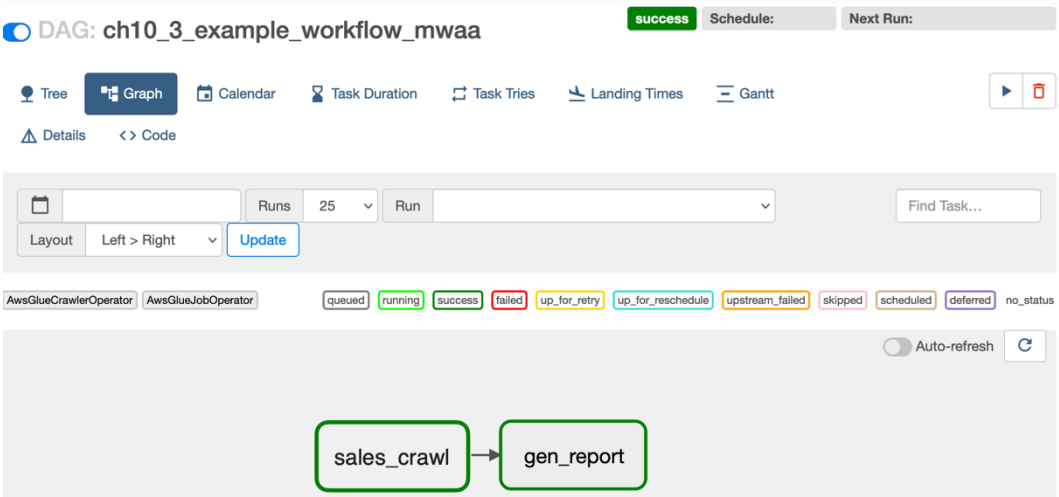
ffdb5e40-9dc9-0025-f619-63cc56083822

Input - *optional*

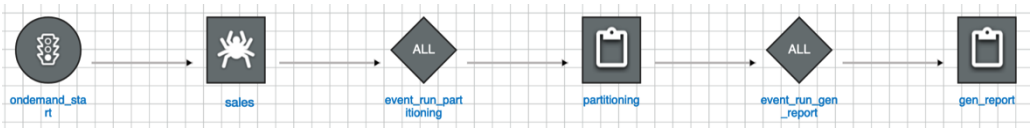
Enter input values for this execution in JSON format

```
{
    "crawler_name": "ch10_2_example_workflow",
    "etl_job_name": "ch10_2_example_workflow_gen_report",
    "etl_job_args": {
        "--datalake_location": "s3://<your-bucket-and-path>",
        "--database": "<your-database>",
        "--table": "example_workflow_sfn_sales",
        "--report_year": "2021"
    }
}
```





```yaml
Resources:
  GlueDatabase:
    Type: AWS::Glue::Database
    Properties:
      CatalogId: !Ref AWS::AccountId
      DatabaseInput:
        Name: 'glue_db'
  GlueTable:
    Type: 'AWS::Glue::Table'
    DependsOn:
      - GlueDatabase
    Properties:
      CatalogId: !Ref AWS::AccountId
      DatabaseName: 'glue_db'
      TableInput:
        Name: 'glue_table'
```



Stack name

Stack name

ch10_4_example_cf

Stack name can include letters (A-Z and a-z), numbers (0-9), and dashes (-).

Parameters

Parameters are defined in your template and allow you to input custom values when you create or update a stack.

DataLakeLocation

The combination of S3 bucket name and path that stores the analytic sales data and a sales report. This location must end with a slash (/) and not include any files.

s3://<bucket-and-path>/

DatabaseName

Database name for the table of the sales data.

GlueCrawlerRoleArn

IAM Role ARN for the Glue Crawler.

arn:aws:iam::<YOUR_ACCOUND_ID>:role/service-role/<ROLE_NAME>

GlueJobRoleArn

IAM Role ARN for the Glue ETL jobs.

arn:aws:iam::<YOUR_ACCOUND_ID>:role/service-role/<ROLE_NAME>

GlueJobScriptLocation

The combination of S3 bucket name and path that locates ch10_4_example_cf_partitioning.py and ch10_4_example_cf_gen_report.py. The combination of this location and each script name is specified in each Glue job as its script location. This location must end with a slash (/) and not include any files.

s3://<bucket-and-path-to-script>/

ReportYear

The year when you want to aggregate the dataset and generate a report.

2021

SalesDataLocation

The combination of S3 bucket name and path that stores sales-data.json that you downloaded from GitHub repository. This location must end with a slash (/) and not include any files.

s3://<bucket-and-path-to-sales-data.json>/

TableName

Table name for the table of sales data. If you use a custom table name, set the table name here.

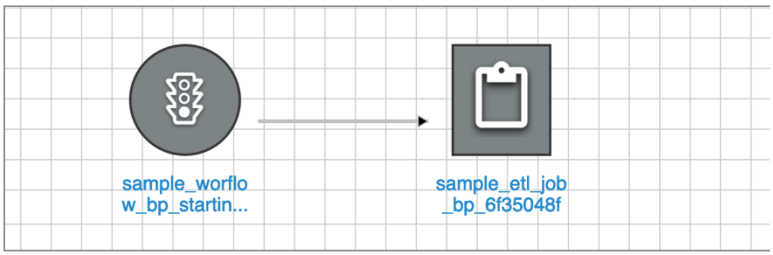
ch10_4_example_cf_sales

Cancel Previous Next

Name	Last run	Last run status
sample_worflow_bp	-	-

Graph | Details | History

Legend: Start, Trigger, Job, Crawler, Incomplete, Error, Deleting



Create a workflow from ch10_5_example_bp

AWS Glue will run the blueprint to create a workflow.

WorkflowName
Name for the workflow.
ch10_5_example_bp

ScriptLocation
Specify the S3 path to store your glue job scripts.
s3://<your-bucket-and-path> Browse

SalesDataLocation
Specify the S3 path to store the sales-data.json.
s3://<your-bucket-and-path> Browse

DataLakeLocation
Specify the S3 path to store your sales data.
s3://<your-bucket-and-path> Browse

GlueCrawlerRoleName
Choose an IAM role for Glue Crawler.

GlueJobRoleName
Choose an IAM role for Glue ETL Job.

DatabaseName
Specify a database name for the table of sales data.

ReportYear
Specify the year when you want to aggregate the dataset and generate a report.
2021

IAM role

Role assumed by AWS Glue with permission to create workflows and their AWS resources. For more information, see **Create an IAM Role for AWS Glue.**

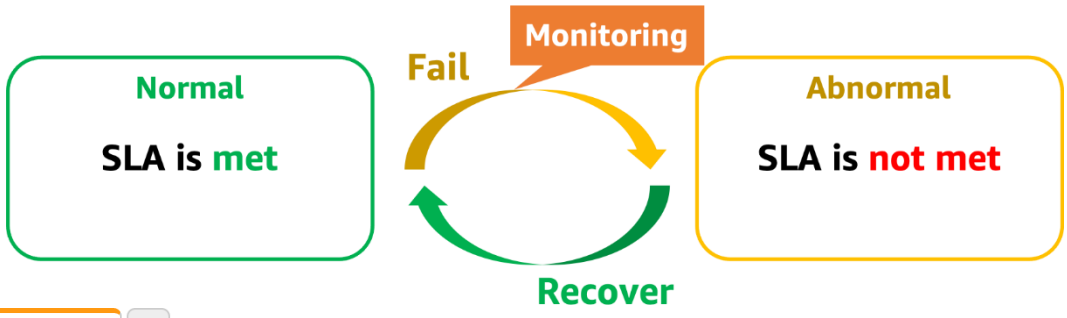
Cancel Submit

Blueprint Runs (4)

Actions | Filter blueprints

Run Id	State	Workflow name	Started on	Completed on
bpr_b26bd73939210b7a3dfd1a1044...	SUCCEEDED	ch10_5_example_bp		

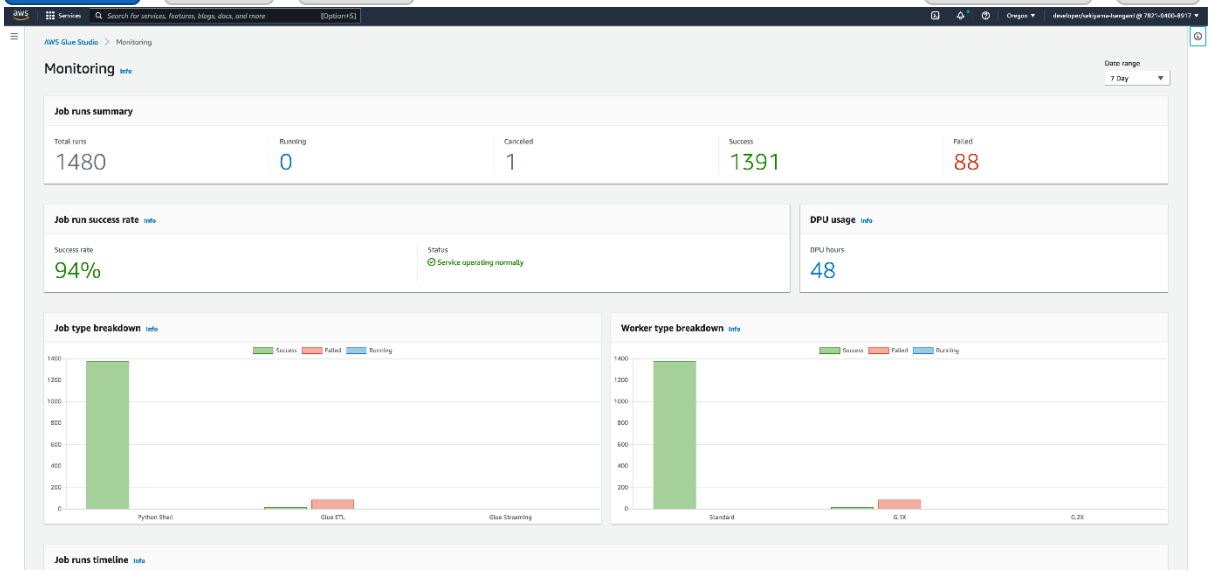
Chapter 11: Monitoring

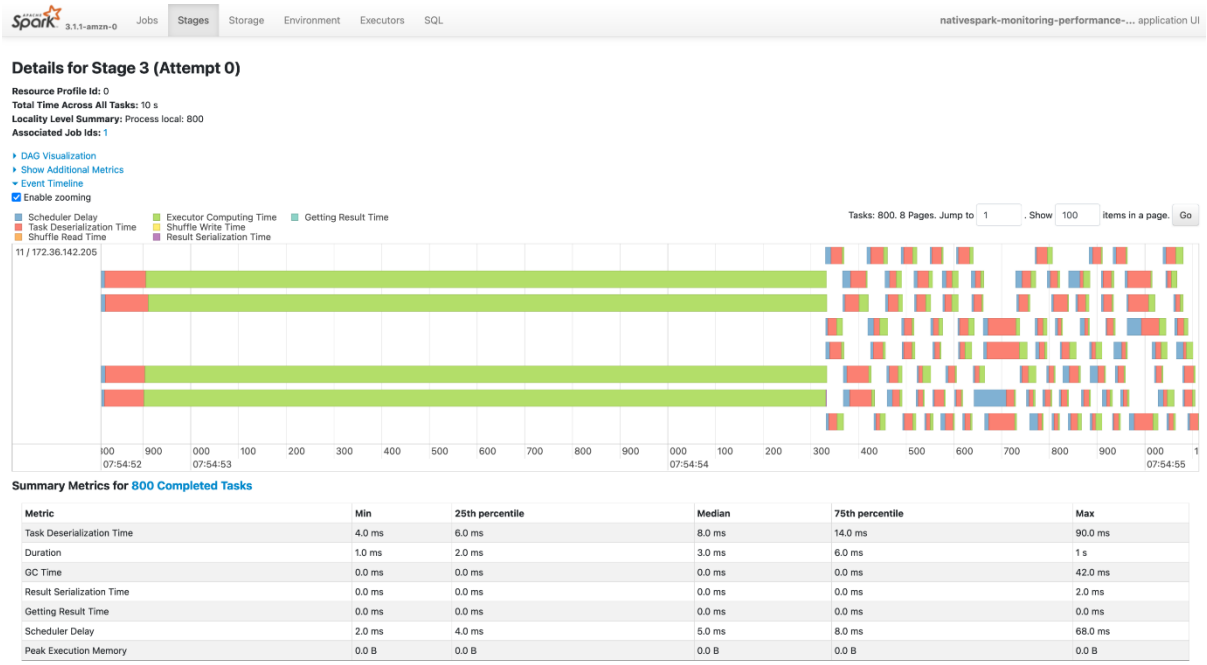


New query 1

```
SELECT s.date
FROM "covid-19"."covid_testing_us_daily" u
left outer join "covid-19"."covid_testing_states_daily" s on u.date=s.date
left outer join "covid-19"."us_state_abbreviations" abb on s.state=abb.abbreviation
GROUP BY s.date
ORDER BY s.date DESC
LIMIT 1
```

Run query | Save as | Create | (Run time: 1.86 seconds, Data scanned: 1.62 MB) | Format query | Clear



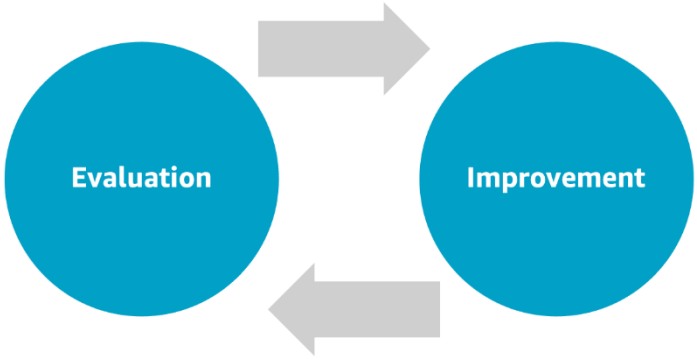


Summary Metrics for 800 Completed Tasks

Metric	Min	25th percentile	Median	75th percentile	Max
Task Deserialization Time	4.0 ms	6.0 ms	8.0 ms	14.0 ms	90.0 ms
Duration	1.0 ms	2.0 ms	3.0 ms	6.0 ms	1 s
GC Time	0.0 ms	0.0 ms	0.0 ms	0.0 ms	42.0 ms
Result Serialization Time	0.0 ms	0.0 ms	0.0 ms	0.0 ms	2.0 ms
Getting Result Time	0.0 ms	0.0 ms	0.0 ms	0.0 ms	0.0 ms
Scheduler Delay	2.0 ms	4.0 ms	5.0 ms	8.0 ms	68.0 ms
Peak Execution Memory	0.0 B	0.0 B	0.0 B	0.0 B	0.0 B

Generate job insights

AWS Glue will analyze your job runs and provide insights on how to optimize your jobs and the reasons for job failures.



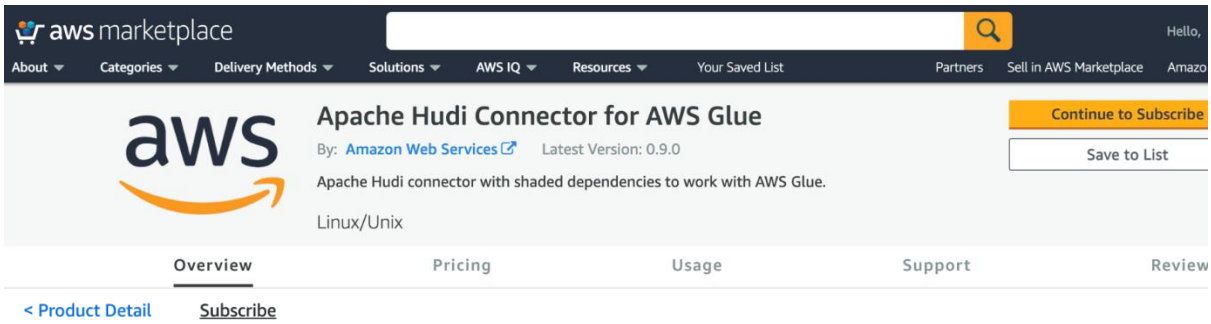
- Usage analysis
- End user feedback

- Add more datasets
- Add supported engines
- Enhance documentation
- Improve usage policy

Chapter 12: Tuning, Debugging, and Troubleshooting

No images…

Chapter 13: Data Analysis



< Product Detail Subscribe

Subscribe to this software

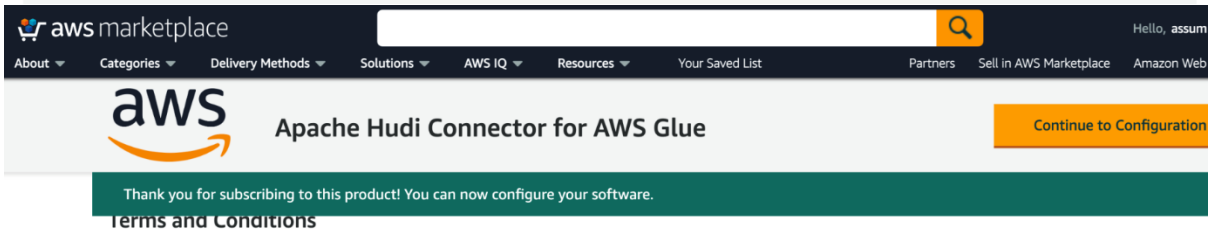
To create a subscription, review the pricing information and accept the terms for this software.

Terms and Conditions

Amazon Web Services Offer

By subscribing to this software, you agree to the pricing terms and the seller's End User License Agreement (EULA). You also agree and acknowledge that AWS may share information about this transaction (including your payment terms) with the respective seller, reseller or underlying provider, as applicable, in accordance with the AWS Privacy Notice. AWS will issue invoices and collect payments from you on behalf of the seller through your AWS account. Your use of AWS services is subject to the AWS Customer Agreement or other agreement with AWS governing your use of such services.

Accept Terms



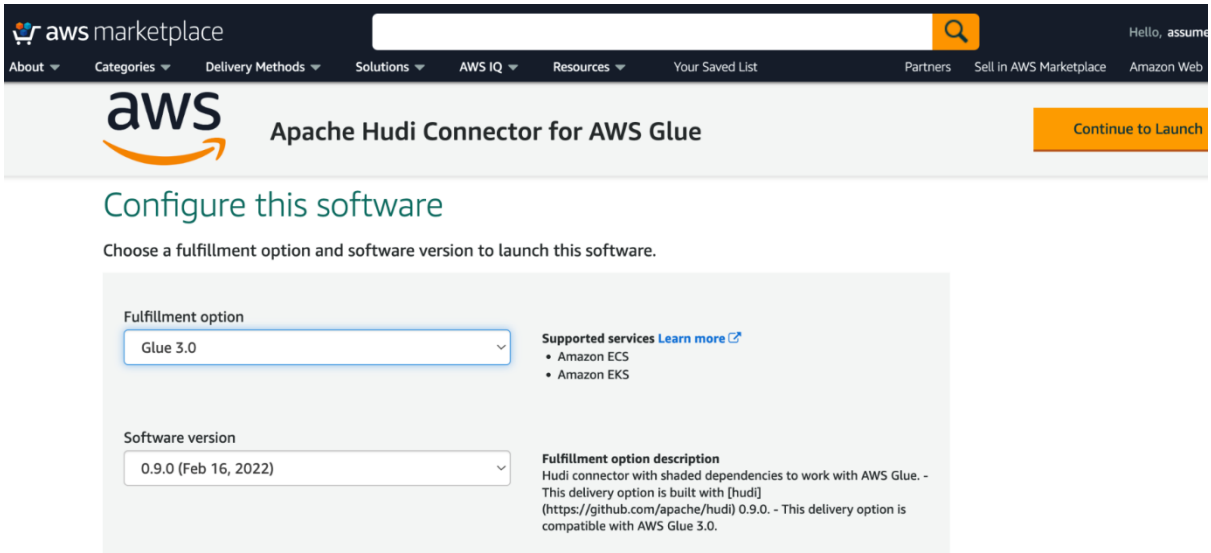
Thank you for subscribing to this product! You can now configure your software.

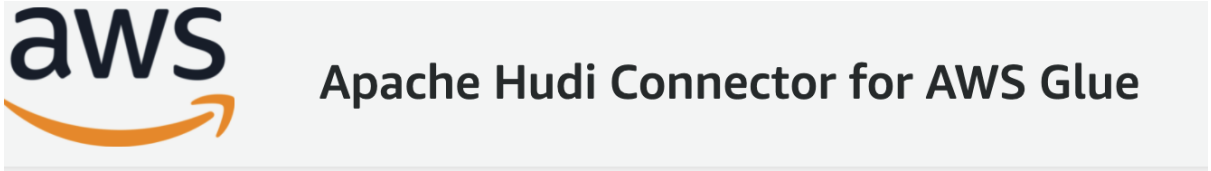
Terms and Conditions

Amazon Web Services Offer

You have subscribed to this software and agreed that your use of this software is subject to the pricing terms and the seller's End User License Agreement (EULA). You agreed that AWS may share information about this transaction (including your payment terms) with the respective seller, reseller or underlying provider, as applicable, in accordance with the AWS Privacy Notice. AWS will issue invoices and collect payments from you on behalf of the seller through your AWS account. Your use of AWS services remains subject to the AWS Customer Agreement or other agreement with AWS governing your use of such services.

Product	Effective date	Expiration date	Action
Apache Hudi Connector for AWS Glue	6/12/2022	N/A	Show Details

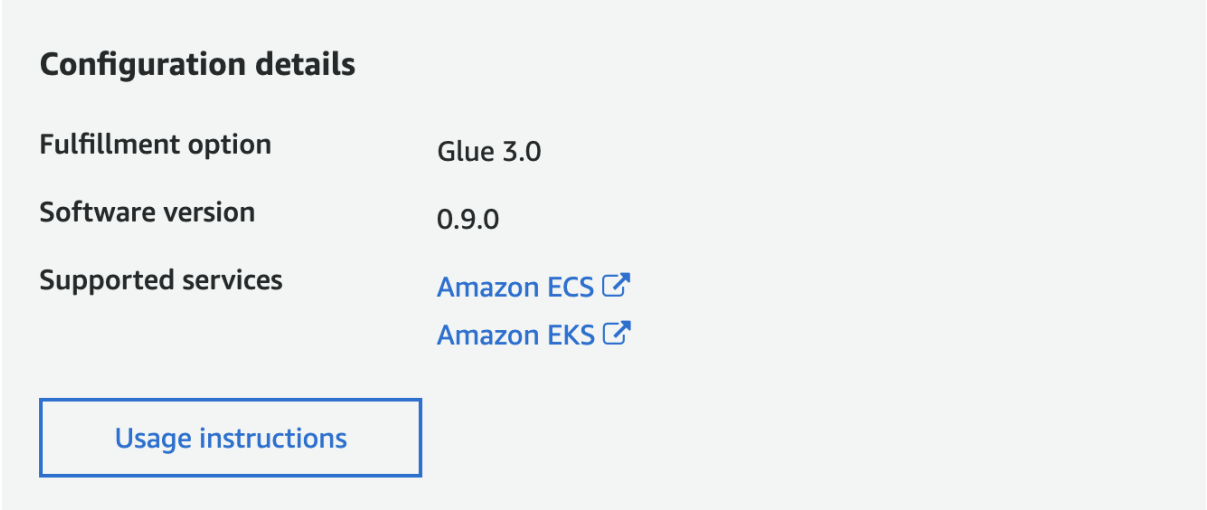




< Product Detail Subscribe Configure Launch

Launch this software

Review the launch configuration details and follow the instructions to launch this so



Usage Instructions for 0.9.0

Please subscribe to the product from AWS Marketplace and Activate the Glue connector from AWS Glue Studio.

AWS Glue Studio > Connectors > Create connection

Create connection Info

Connection properties Info

Connector

Apache Hudi Connector 0.9.0 for AWS Glue 3.0
Connect to Apache Hudi tables from AWS Glue

Name
Enter a unique name for your connection.

hudi-connection

Names can contain letters (A-Z), numbers (0-9), hyphens (-), or underscores (_), and must be less than 256 characters long.

Description - *optional*

Descriptions can be up to 2048 characters long.

Connection access Info

AWS Secret - *optional* Info
Choose a secret from **AWS Secrets Manager.** AWS secrets eliminate hardcoding sensitive information.

Choose a secret to use

Network options - *optional*

If your AWS Glue job needs to run on **Amazon Elastic Compute Cloud** (EC2) instances in a virtual private cloud (VPC) subnet, you must provide additional VPC-specific configuration information.

Cancel | Activate connector only | Create connection and activate connector

aws marketplace

Elasticsearch Connector for AWS Glue

Hello, assumed-role

About | Categories | Delivery Methods | Solutions | AWS IQ | Resources | Your Saved List | Partners | Sell in AWS Marketplace | Amazon Web Service

Refine results

Categories

Infrastructure Software (4)
Professional Services (1)

Delivery methods

Container Image (3)
Professional Services (1)

Elasticsearch Connector for AWS Glue (4 results) showing 1 - 4

1

Sort By: Relevance

Elasticsearch Connector for AWS Glue
By **Amazon Web Services** | Ver 7.13.4-2

The Elasticsearch Connector for AWS Glue helps you read from and write to ElasticSearch using Apache Spark. By using this connector, you can focus on mining meaningful business insights from your data instead of writing and maintaining the connecting logic. For more details, please refer to Glue...

AWS Glue Studio

Jobs
Monitoring
Connectors
Sensitive data detection
What's new
Glue console
Glue catalog
Crawlers
Security configurations

Python Shell script editor
Write or upload your own Python shell script.

Jupyter Notebook
Write your own code in a Jupyter Notebook for interactive development.

Source
Amazon S3
JSON, CSV, or Parquet files stored in S3.

Target
Amazon S3
S3 bucket by specifying a bucket path as the data target.

Your jobs (12) Info

Actions | Run job

Find jobs

Job name	Type	Last modified	AWS Glue version
01 - Seed data job for Data Analysis Chapter	Glue ETL	6/12/2022, 2:12:00 AM	3.0

AWS Glue Studio

Jobs
Monitoring
Connectors
Sensitive data detection
What's new
Glue console
Glue catalog
Crawlers
Security configurations
Marketplace
Documentation

Python Shell | Glue ETL | Glue Streaming

Standard | G.1X | G.2X

Job runs timeline Info

Success | Failed | Running | Canceled

2022-06-12

Job runs (1) Info

Actions | View CloudWatch logs | View run details

Filter job runs

Job name	Type	Start time	End time	Run status	Run time	Capacity	Worker type	DPU hours
01 - Seed data job for Data Analysis Chapter	Glue ETL	06/12/2022 02:29:46	-	Running	-	2	Standard	

Amazon Athena

Query editor
Workgroups
Data sources
Jobs
Workflows New
Powered by Step Functions
Enable compact mode

Data

Data source
AwsDataCatalog

Database
chapter-data-analysis-glue-database

Tables and views | Create

Filter tables and views

Tables (1)
employees
Views (0)

```
SELECT * FROM "chapter-data-analysis-glue-database"."employees" order by emp_no;
```

SQL Ln 1, Col 80

Run again | Cancel | Save | Clear | Create

Completed | Time in queue: 121 ms | Run time: 491 ms | Data scanned: 0.71 KB

Results (7)

Copy | Download results

Search rows

#	emp_no	name	department	city	salary
1	1	Adam	IT	SFO	50000
2	2	Susan	Sales	NY	60000
3	3	Jeff	Finance	Tokyo	55000

Database

Queries

Redshift query editor v2

Create | Load data

Filter resources

chapterdataanalysisredshiftcluster...

Database

dev

The database name must be 1-64 characters. Valid characters are lowercase alphanumeric characters.

User name

Authentication

- Federated user
- Temporary credentials
- Database user name and password

Password

••••••••••

Show password

Provide a Password for the database that you are connecting to. The query editor v2 creates a secret on your behalf stored in AWS Secrets Manager. This secret contains credentials to connect to your database. Learn more

Cancel | Create connection

Cluster | Serverless (admin) | Database | dev

Database

Queries

Notebooks (Preview)

Charts

Redshift query editor v2

+ Create | Load data

Filter resources

- chapterdataanalysisredshift...
  - dev
    - chapter_data_analysi...
      - Tables 1
        - employees
    - public

+ Create ▾ | Load data

Filter resources

- chapterdataanalysisredshift...
  - dev
    - chapter_data_analysi...
      - Tables 1
        - employees
    - public
  - sample_data_dev

Untitled 1 ×

Run | Limit 100 | Explain

```
1 SELECT * FROM "dev"."chapter_data_analysis_schema"."employees" order by emp_no;
```

Result 1 (7)

emp_no	name	department	city	salary
1	Adam	IT	SFO	50000
2	Susan	Sales	NY	60000
3	Jeff	Finance	Tokyo	55000
4	Bill	Manufacturing	New Delhi	70000
5	Joe	IT	Chicago	45000
6	Steve	Finance	NY	60000
7	Mike	IT	SFO	60000

Job runs (3) Info

Actions ▾ | View CloudWatch logs | View run details

Filter job runs

Job name	Type	Start time	End time	Run status	Run time	Capacity	Worker type	DPU hours
02 - Hudi Init load for Data Analysis Chapter	Glue ETL	06/12/2022 18:16:04	-	Running	-	2	Standard	
01 - Seed data job for Data Analysis Chapter	Glue ETL	06/12/2022 17:13:34	06/12/2022 17:14:40	Succeeded	1 minute	2	Standard	0.03

Query 1

```
1 SELECT emp_no, name, department, city, salary FROM "AwsDataCatalog"."chapter-data-analysis-glue-database"."employees_cow" order by emp_no;
```

SQL Ln 1, Col 97

Run again | Cancel | Save ▾ | Clear | Create ▾

Completed — Time in queue: 115 ms — Run time: 1.079 sec — Data scanned:

Results (7)

Copy | Download re

Search rows

#	emp_no	name	department	city	salary
1	1	Adam	IT	SFO	50000
2	2	Susan	Sales	NY	60000
3	3	Jeff	Finance	Tokyo	55000
4	4	Bill	Manufacturing	New Delhi	70000
5	5	Joe	IT	Chicago	45000
6	6	Steve	Finance	NY	60000
7	7	Mike	IT	SFO	60000

Job runs (4) Info

Actions ▾ | View CloudWatch logs | View run details

Filter job runs

Job name	Type	Start time	End time	Run status	Run time	Capacity	Worker type	DPU hours
03 - Hudi Incremental load for Data Analysis Chapter	Glue ETL	06/12/2022 18:25:27	-	Running	-	2	Standard	
02 - Hudi Init load for Data Analysis Chapter	Glue ETL	06/12/2022 18:16:04	06/12/2022 18:18:16	Succeeded	2 minutes	2	Standard	0.06
01 - Seed data job for Data Analysis Chapter	Glue ETL	06/12/2022 17:13:34	06/12/2022 17:14:40	Succeeded	1 minute	2	Standard	0.03

Query 1

```
1 SELECT emp_no, name, department, city, salary FROM "AwsDataCatalog"."chapter-data-analysis-glue-database"."employees_cow" order by emp_no;
```

SQL Ln 1, Col 139

Run again | Cancel | Save | Clear | Create

Completed Time in queue: 148 ms Run time: 1.061 sec Data scanned: 1.2

Results (7) Copy Download result

Search rows 1

#	emp_no	name	department	city	salary
1	1	Adam	IT	SFO	50000
2	2	Susan	Sales	NY	60000
3	3	Jeff	Finance	Cincinnati	75000
4	4	Bill	Manufacturing	New Delhi	70000
5	5	Joe	IT	Chicago	45000
6	6	Steve	Finance	NY	60000
7	7	Mike	IT	SFO	60000

Libraries Info

Python library path

/tmp/delta-core_2.12-1.0.0.jar

Dependent JARs path

Referenced files path

Job parameters Info

Key	Value - *optional*	
--DELTALAKE_CONN	deltalake-connectior	Remove
--TARGET_BUCKET	scd-targets3bucket-i	Remove

Add new parameter

You can add 48 more parameters.

Tags

Key	Value - *optional*	
Project	HandsonSeriesWithA	Remove

Query 1

```
1 SELECT * FROM "AwsDataCatalog"."chapter-data-analysis-glue-database"."employees_deltalake" order by emp_no;
```

SQL Ln 1, Col 108

Run again | Cancel | Save | Clear | Create

Completed Time in queue: 110 ms Run time: 1.265 sec Data sca

Results (7) Copy Downl

Search rows

#	emp_no	name	department	city	salary
1	1	Adam	IT	SFO	50000
2	2	Susan	Sales	NY	60000
3	3	Jeff	Finance	Tokyo	55000
4	4	Bill	Manufacturing	New Delhi	70000
5	5	Joe	IT	Chicago	45000
6	6	Steve	Finance	NY	60000
7	7	Mike	IT	SFO	60000

Query 1

```
1 SELECT * FROM "AwsDataCatalog"."chapter-data-analysis-glue-database"."employees_deltalake" order by emp_no;
```

SQL Ln 1, Col 108

Run again | Cancel | Save | Clear | Create

Completed Time in queue: 156 ms Run time: 1.531 sec Data sca

Results (7) Copy Downl

Search rows

#	emp_no	name	department	city	salary
1	1	Adam	IT	SFO	50000
2	2	Susan	Sales	NY	60000
3	3	Jeff	Finance	Cincinnati	70000
4	4	Bill	Manufacturing	New Delhi	70000

AWS Lake Formation > Data lake locations > Register location

Register location

Amazon S3 location

Register an Amazon S3 path as the storage location for your data lake.

Amazon S3 path

Choose an Amazon S3 path for your data lake.

s3://scd-targets3bucket-i7xhawru6kwi/employees_governed_table/ **Browse**

Review location permissions - strongly recommended

Registering the selected location may result in your users gaining access to data already at that location. Before registering a location, we recommend that you review existing location permissions on resources in that location.

Review location permissions

IAM role

To add or update data, Lake Formation needs read/write access to the chosen Amazon S3 path. Choose a role that you know has permission to do this, or choose the **AWSServiceRoleForLakeFormationDataAccess** service-linked role. When you register the first Amazon S3 path, the service-linked role and a new inline policy are created on your behalf. Lake Formation adds the first path to the inline policy and attaches it to the service-linked role. When you register subsequent paths, Lake Formation adds the path to the existing policy.

AWSServiceRoleForLakeFormationDataAccess

⚠ Do not select the service linked role if you plan to use EMR.

Cancel **Register location**

AWS Lake Formation > Data locations

Data locations (1)

Choose a storage location for which to review, grant or revoke user permissions.

e.g.: s3://bucket/prefix/ **Browse**

Principal	Principal type	Resource	Own
HandsonSeriesWithAWSGlueJobRole	IAM role	s3://scd-targets3bucket-i7xhawru6kwi/employees_governed_table	-

Query 1

```
1 SELECT * FROM "AwsDataCatalog"."chapter-data-analysis-glue-database"."employees_governed_table" order by emp_no;
```

SQL Ln 1, Col 95

Run again | Cancel | Save | Clear | Create

Completed Time in queue: 113 ms Run time: 5.062 sec Da

Results (7) Copy D

Search rows

#	emp_no	name	department	city	salary
1	1	Adam	IT	SFO	50000
2	2	Susan	Sales	NY	60000
3	3	Jeff	Finance	Tokyo	55000
4	4	Bill	Manufacturing	New Delhi	70000
5	5	Joe	IT	Chicago	45000
6	6	Steve	Finance	NY	60000
7	7	Mike	IT	SFO	60000

Set up your connection's properties.

For more information, see Working with Connections.

Connection name

chapter-data-analysis-msk-connection

Connection type

Kafka

- (●) Amazon managed streaming for Apache Kafka (MSK)
- ( ) Customer managed Apache Kafka

Select MSK cluster

msk-source-cluster-scd

Kafka bootstrap server URLs

b-1.msksourceclusterscd.buosan.c10.kafka.us-west-2.amazonaws.com:9094,b-2.msksourceclusters

Enter a comma-separated list of bootstrap server URLs. Include the port number. Example: b-1.vpc-test-2.o4q88o.c6.kafka.us-east-1.amazonaws.com:9094, b-2.vpc-test-2.o4q88o.c6.kafka.us-east-1.amazonaws.com:9094, b-3.vpc-test-2.o4q88o.c6.kafka.us-east-1.amazonaws.com:9094

[x] Require SSL connection

When selected, connection fails if unable to connect over SSL.

Set up access to your data store.

For more information, see Working with Connections.

VPC

Choose the VPC name that contains your data store.

vpc-00a9f717ef84e5925

Subnet

Choose the subnet within your VPC.

subnet-0361ac6939d4d2bd1

Security groups

Choose one or more security groups that allow access to the data store in your VPC. AWS Glue associates these security groups to the ENI attached to your subnet. To allow AWS Glue components to communicate and also prevent access from other networks, at least one chosen security group must specify a self-referencing inbound rule for all TCP ports.

Group ID	Group name
sg-07c9e5d071c2a40c6	chapter-data-analysis-sg

Query 1

```
SELECT emp_no,name,department,city,salary FROM "AwsDataCatalog"."chapter-data-analysis-glue-database"."employees_cow_streaming" order by emp_no;
```

SQL Ln 1, Col 1

Run again | Cancel | Save | Clear | Create

Completed — Time in queue: 141 ms — Run time: 1.12 sec — Data scanned: 0.58 KB

Results (7)

Copy | Download results

Search rows

#	emp_no	name	department	city	salary
1	1	Adam	IT	SFO	50000
2	2	Susan	Sales	NY	60000
3	3	Jeff	Finance	Tokyo	55000
4	4	Bill	Manufacturing	New Delhi	70000
5	5	Joe	IT	Chicago	45000
6	6	Steve	Finance	NY	60000
7	7	Mike	IT	SFO	60000

Query 1

```
SELECT emp_no,name,department,city,salary FROM "AwsDataCatalog"."chapter-data-analysis-glue-database"."employees_deltastreamer" order by emp_no;
```

SQL Ln 1, Col 66

Run again | Cancel | Save | Clear | Create

Completed — Time in queue: 163 ms — Run time: 926 ms — Data scanned: 0.56 KB

Results (7) — Copy | Download results

Search rows

#	emp_no	name	department	city	salary
1	1	Adam	IT	SFO	50000
2	2	Susan	Sales	NY	60000
3	3	Jeff	Finance	Tokyo	55000
4	4	Bill	Manufacturing	New Delhi	70000
5	5	Joe	IT	Chicago	45000
6	6	Steve	Finance	NY	60000
7	7	Mike	IT	SFO	60000

Query 1

```
SELECT emp_no,name,department,city,salary FROM "AwsDataCatalog"."chapter-data-analysis-glue-database"."employees_deltastreamer" order by emp_no;
```

SQL Ln 1, Col 15

Run again | Cancel | Save | Clear | Create

Completed — Time in queue: 191 ms — Run time: 1.34 sec — Data scanned: 0.56 KB

Results (7) — Copy | Download results

Search rows

#	emp_no	name	department	city	salary
1	1	Adam	IT	SFO	50000
2	2	Susan	Sales	NY	60000
3	3	Jeff	Finance	Cincinnati	70000
4	4	Bill	Manufacturing	New Delhi	70000
5	5	Joe	IT	Chicago	45000
6	6	Steve	Finance	NY	60000
7	7	Mike	IT	SFO	60000

Query 1

```
SELECT emp_no,name,department,city,salary FROM "AwsDataCatalog"."chapter-data-analysis-glue-database"."employees_cow_streaming" order by emp_no;
```

SQL Ln 1, Col 127

Run again | Cancel | Save | Clear | Create

Completed — Time in queue: 129 ms — Run time: 1.064 sec — Data scanned: 0.58 KB

Results (7) — Copy | Download results

Search rows

#	emp_no	name	department	city	salary
1	1	Adam	IT	SFO	50000
2	2	Susan	Sales	NY	60000
3	3	Jeff	Finance	Cincinnati	70000
4	4	Bill	Manufacturing	New Delhi	70000
5	5	Joe	IT	Chicago	45000
6	6	Steve	Finance	NY	60000
7	7	Mike	IT	SFO	60000

AWS Glue Studio > Connectors > opensearch-connection

opensearch-connection

Edit | Delete | Create job

Connection details Info

Connector type
MARKETPLACE

Subnet
-

Connector ECR URL
https://709825985650.dkr.ecr.us-east-1.amazonaws.com/amazon-web-services/glue/elasticsearch:7.13.4-glue3.0-2

Require SSL connection
-

Security groups
-

Description
-

Edit connection

Connection properties Info

Connector

Elasticsearch Connector 7.13.4 for AWS Glue 3.0
Connect to Elasticsearch from AWS Glue

Name (Read-only)

opensearch-connection

Description - *optional*

Descriptions can be up to 2048 characters long.

Connection access Info

AWS Secret - *optional* Info

Choose a secret from **AWS Secrets Manager.** AWS secrets eliminate hardcoding sensitive information.

ChapterDataAnalysisOSSecret

Select your tenant

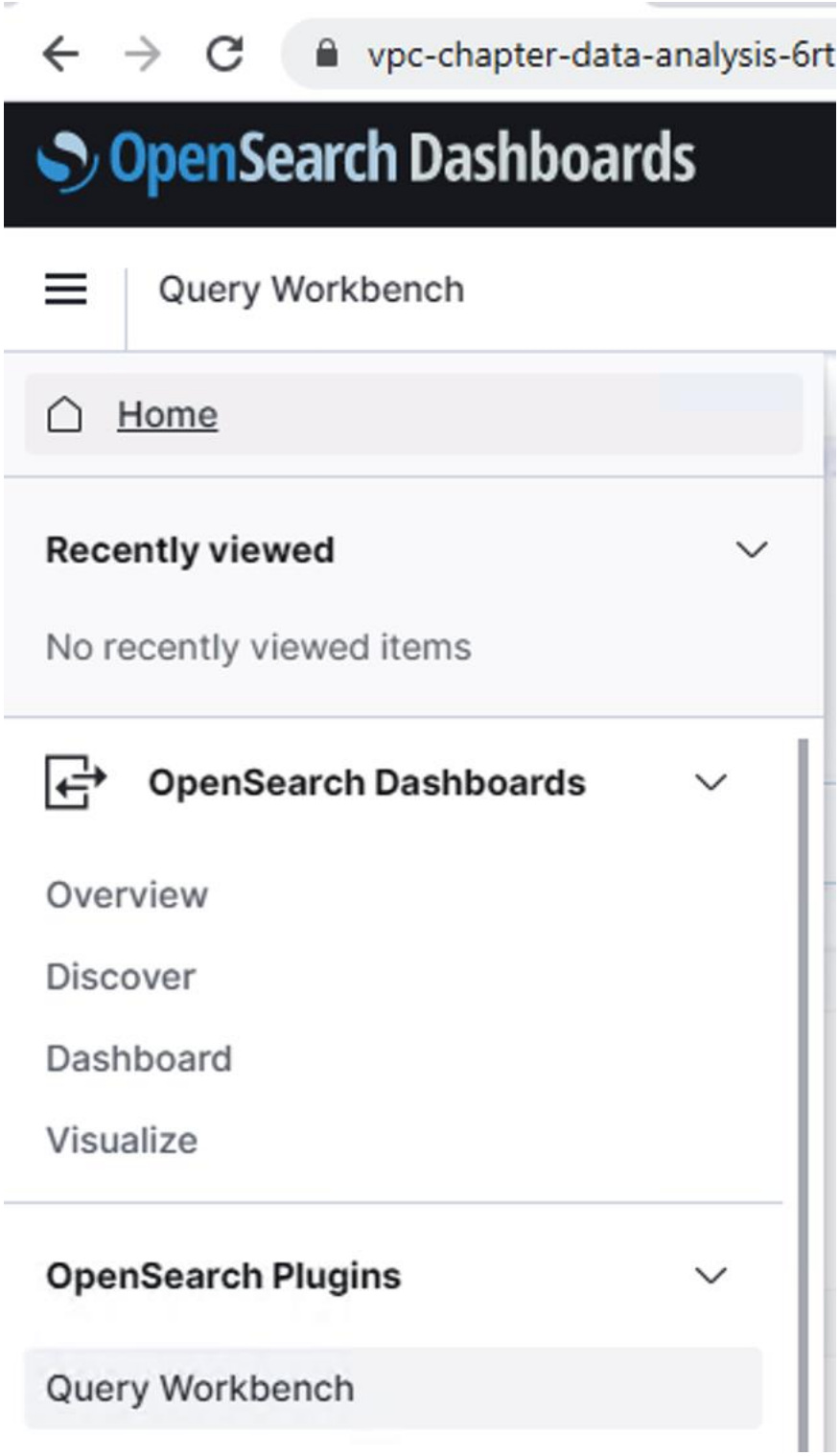
Tenants are useful for safely sharing your work with other OpenSearch Dashboards users. You can switch your tenant anytime by clicking the user avatar on top right.

- ○ Global
  The global tenant is shared between every OpenSearch Dashboards user.

- ◉ Private
  The private tenant is exclusive to each user and can't be shared. You might use the private tenant for exploratory work.

- ○ Choose from custom

Cancel Confirm


vpc-chapter-data-analysis-6rt
OpenSearch Dashboards
Query Workbench
Home
Recently viewed
No recently viewed items
OpenSearch Dashboards
Overview
Discover
Dashboard
Visualize
OpenSearch Plugins
Query Workbench

Query Workbench

Query editor

```
1  select * from employees order by emp_no;
```

Run Clear Explain

Results

Output employees

employees (7)

Search keyword

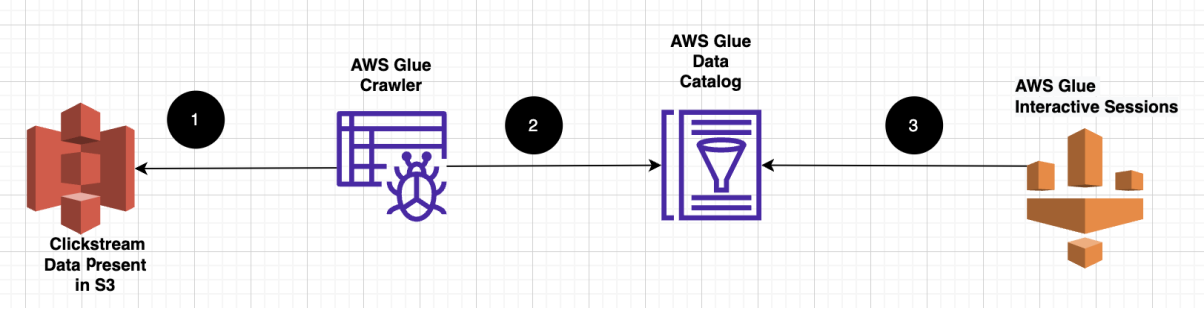
name	department	salary	city	emp_no
Adam	IT	50000	SFO	1
Susan	Sales	60000	NY	2
Jeff	Finance	55000	Tokyo	3
Bill	Manufacturing	70000	New Delhi	4

Chapter 14: Machine Learning Integration

label	book_id	book_title	authors
	12367126	1-Safe Algorithms for Symmetric Site Configurations	John Hayes, Richard B. Bauchman
	08272651	1-Safe Algorithms for Symmetric Site Configurations	John Hayes, Richard B. Bauchman
	71616223	2003 SIGMOD Innovations Award Speech	Martha Smith
	12637181	2Q: A Low Overhead High-Performance Buffer Management	Elena Garcia
	72521341	2Q: A Low Overhead High-Performance Buffer Management	Elena Garcia

label	book_id	book_title	authors
0	12367126	1-Safe Algorithms for Symmetric Site Configurations	John Hayes, Richard B. Bauchman
0	08272651	1-Safe Algorithms for Symmetric Site Configurations	John Hayes, Richard B. Bauchman
1	71616223	2003 SIGMOD Innovations Award Speech	Martha Smith
2	12637181	2Q: A Low Overhead High-Performance Buffer Management	Elena Garcia
2	72521341	2Q: A Low Overhead High-Performance Buffer Management	Elena Garcia

Chapter 15: Architecting Data Lakes for Real-World Scenarios and Edge Cases




```
%%time
%%sql
select count(*) from serverless_glue.tbl_without_index_clkstreamdata where customer = 2 and visityearmonth = 199812
```

```
CPU times: user 52.6 ms, sys: 19.1 ms, total: 71.7 ms
Wall time: 29.7 s
```

Type: Table Pie

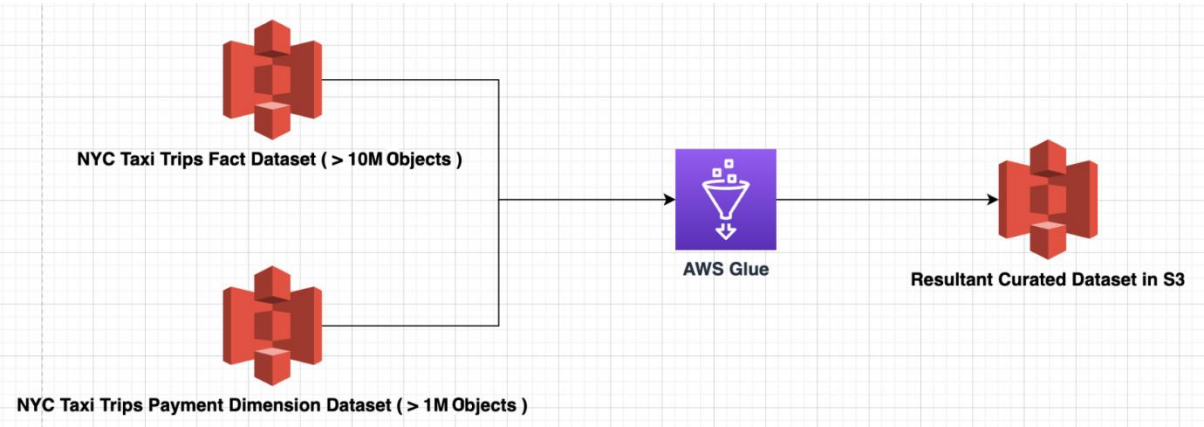
count(1)
7342094

```
%%time
%%sql
select count(*) from serverless_glue.tbl_with_index_clkstreamdata where customer = 2 and visityearmonth = 199812
```

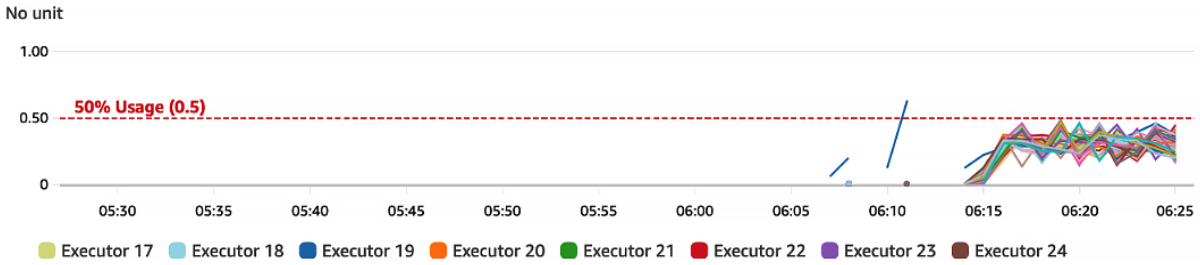
```
CPU times: user 42.2 ms, sys: 2.21 ms, total: 44.4 ms
Wall time: 9.43 s
```

Type: Table Pie

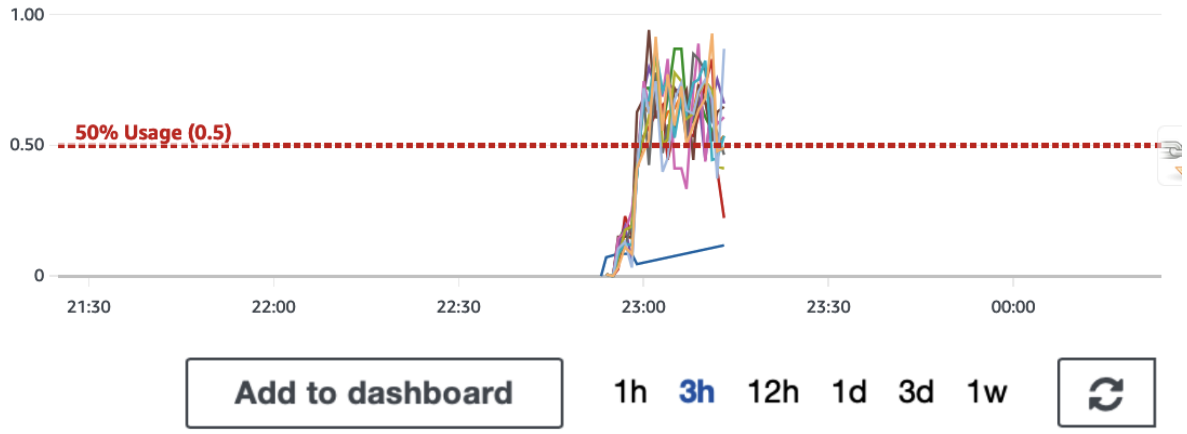
count(1)
7342094



Memory Profile: Driver and Executors



Memory Profile: Driver and Executors

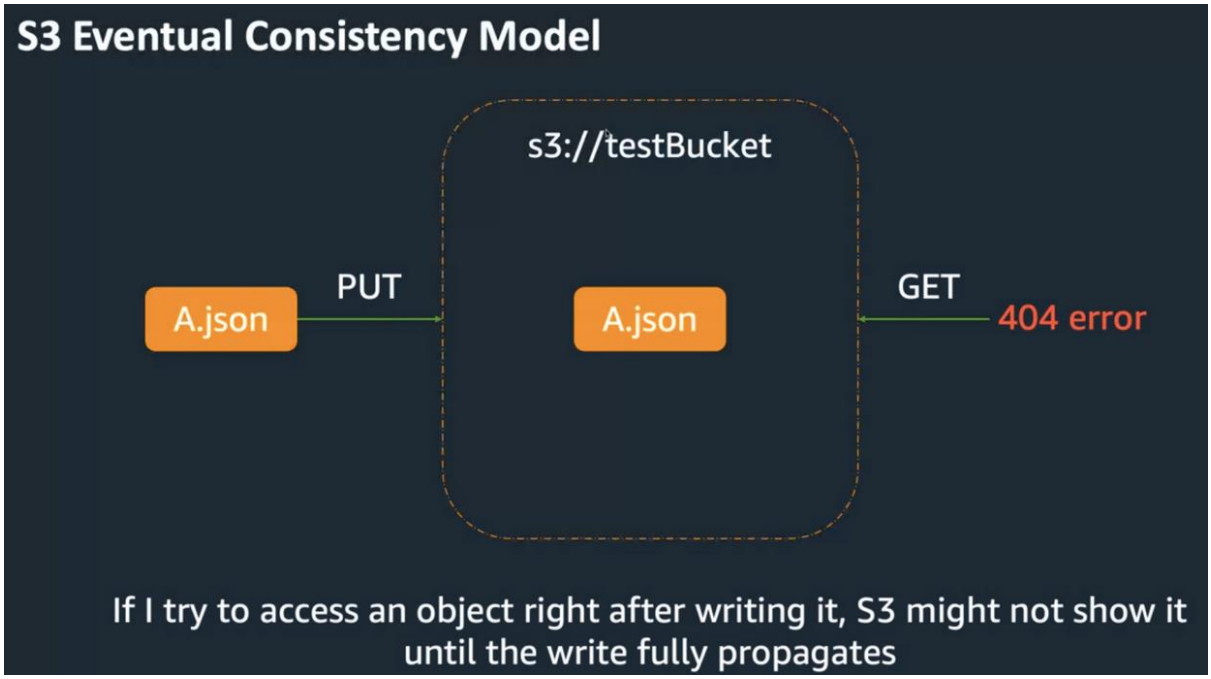


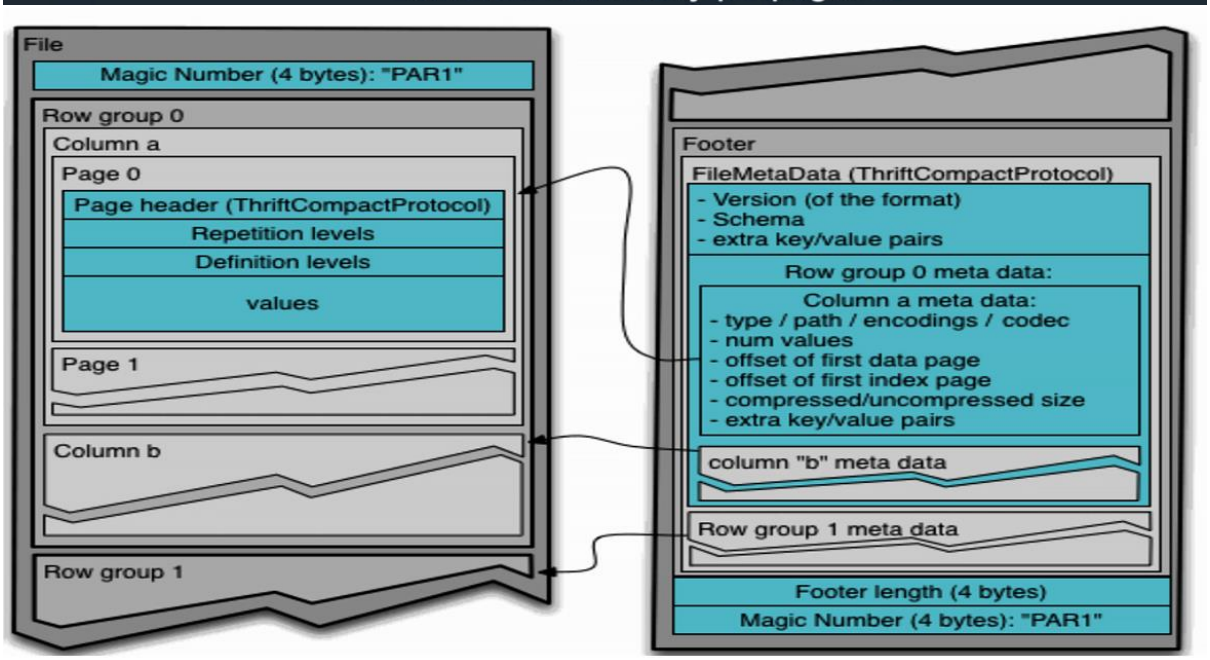
Memory Profile: Driver and Executors

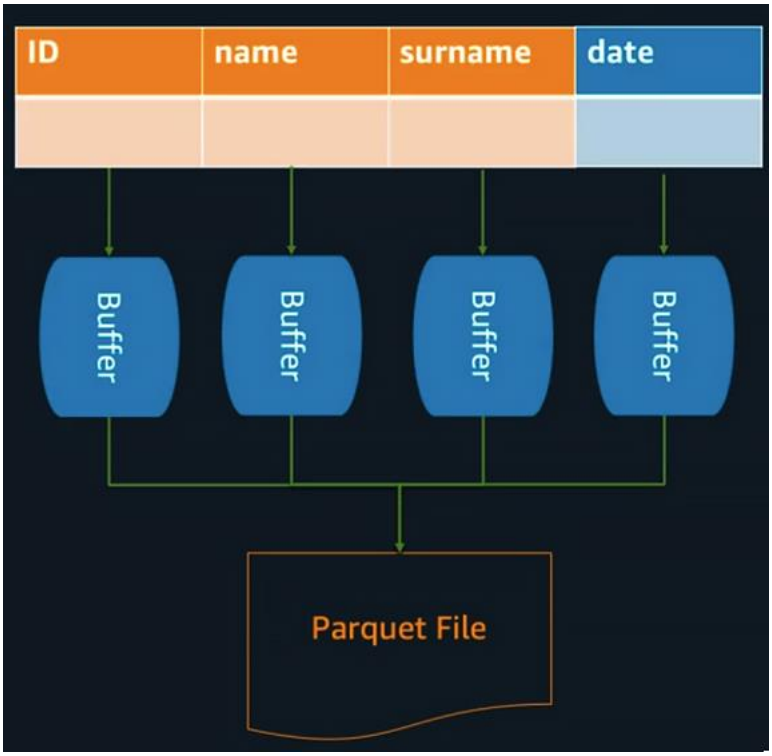
Edit table details

UPDATED_BY_CRAWLER	crawl_aib
CrawlerSchemaSerializerVer:	1.0
recordCount	20726160
averageRecordSize	841
CrawlerSchemaDeserializerV	1.0
compressionType	none
classification	parquet
typeOfData	file
groupFiles	inPartition
groupSize	1048576
recurse	true

Apply


S3 Eventual Consistency Model
s3://testBucket
A.json
PUT
A.json
GET
404 error
If I try to access an object right after writing it, S3 might not show it until the write fully propagates


File
Magic Number (4 bytes): "PAR1"
Row group 0
Column a
Page 0
Page header (ThriftCompactProtocol)
Repetition levels
Definition levels
values
Page 1
Column b
Row group 1
Footer
FileMetaData (ThriftCompactProtocol)
- Version (of the format)
- Schema
- extra key/value pairs
Row group 0 meta data:
Column a meta data:
- type / path / encodings / codec
- num values
- offset of first data page
- offset of first index page
- compressed/uncompressed size
- extra key/value pairs
column "b" meta data
Row group 1 meta data
Footer length (4 bytes)
Magic Number (4 bytes): "PAR1"


ID
name
surname
date
Buffer
Buffer
Buffer
Buffer
Parquet File