End-to-End Data Engineering Project on AWS using S3, Athena, Glue, and QuickSight

Introduction

In today's data-driven world, efficient data processing, storage, and analysis are essential for extracting insights and enabling informed decision-making. This project showcases a complete **end-to-end data engineering pipeline** using core AWS services — **Amazon S3**, **AWS Glue**, **Amazon Athena**, and **Amazon QuickSight** — to handle the full data lifecycle, from raw ingestion to rich visual analysis.

The project leverages the Data Science Job Salaries dataset from Kaggle (link), which offers real-world insights into salary trends and compensation structures within the data science profession.

The primary objective is to build a scalable, serverless data pipeline that:

- Ingests raw data into Amazon S3
- Queries the data interactively using Amazon Athena
- Transforms and processes data through AWS Glue ETL
- Visualizes the results with Amazon QuickSight dashboards

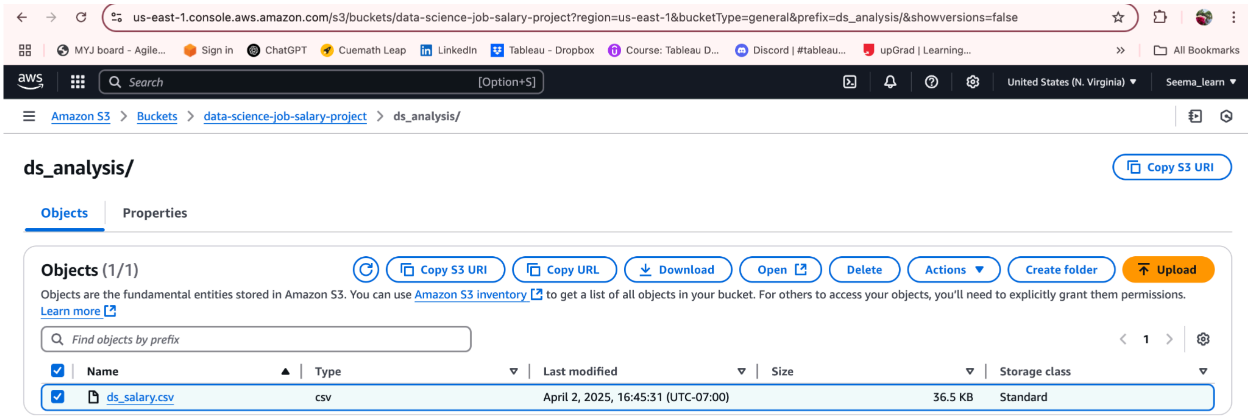
Part 1:

CSV File Uploaded to S3 Bucket:

- **Objective**: The **"ds_salary.csv"** file was uploaded to S3 for data analysis, enabling efficient storage, retrieval, and processing within AWS services.
- **Source and Destination:**
  - **Source**: The dataset was obtained from Kaggle (**Data Science Job Salaries**) and uploaded from the local system.
  - **Destination**: The file was uploaded to the S3 bucket at `s3://data-science-job-salary-project/ds_analysis/ds_salary.csv`

- **Steps Taken:**
  - The file was manually uploaded to the S3 bucket.
  - The column header was removed prior to uploading.

  S3 console screenshot:



ds_analysis/

Objects (1/1)

Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

☑	Name ▲	Type	Last modified	Size	Storage class
☑	ds_salary.csv	csv	April 2, 2025, 16:45:31 (UTC-07:00)	36.5 KB	Standard

- **S3 URI:**

  s3://data-science-job-salary-project/ds_analysis/ds_salary.csv

Part 2:

SQL Query Documentation for Athena-based Data Analysis

Introduction:

- **Purpose of the analysis**:This document outlines the SQL queries executed in AWS Athena to perform data analysis on the "ds_salary.csv" dataset, which was uploaded to an S3 bucket. The analysis aimed to derive insights on salary trends based on job roles, experience, and location.
- **Tools Used:** AWS Athena, SQL, S3, etc.

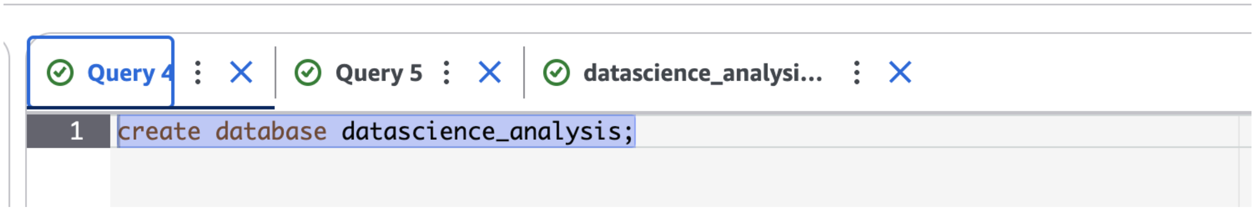
Data Source:

- **Dataset**: "ds_salary.csv" from Kaggle, uploaded to an S3 bucket.

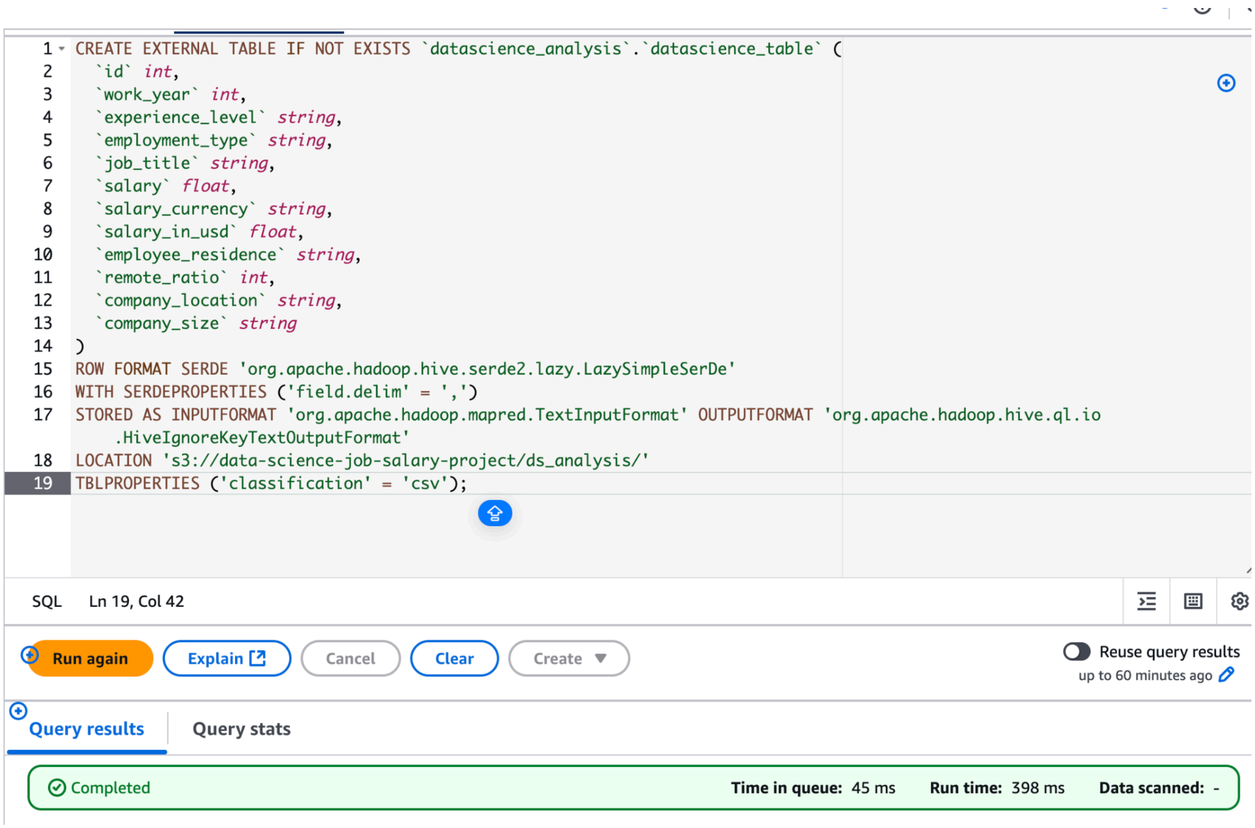
- **S3 Bucket**: `s3://data-science-job-salary-project/ds_analysis/ds_salary.csv`

SQL Query Documentation:

Database & Table Creation



```sql
create database datascience_analysis;
```

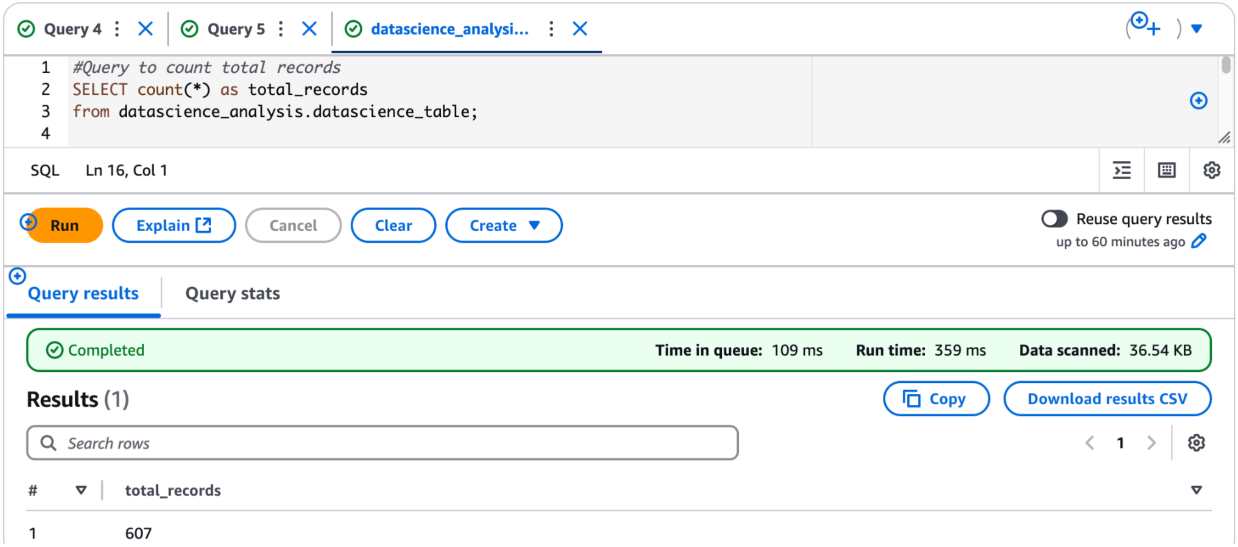


```sql
CREATE EXTERNAL TABLE IF NOT EXISTS `datascience_analysis`.`datascience_table` (
  `id` int,
  `work_year` int,
  `experience_level` string,
  `employment_type` string,
  `job_title` string,
  `salary` float,
  `salary_currency` string,
  `salary_in_usd` float,
  `employee_residence` string,
  `remote_ratio` int,
  `company_location` string,
  `company_size` string
)
ROW FORMAT SERDE 'org.apache.hadoop.hive.serde2.lazy.LazySimpleSerDe'
WITH SERDEPROPERTIES ('field.delim' = ',')
STORED AS INPUTFORMAT 'org.apache.hadoop.mapred.TextInputFormat' OUTPUTFORMAT 'org.apache.hadoop.hive.ql.io
    .HiveIgnoreKeyTextOutputFormat'
LOCATION 's3://data-science-job-salary-project/ds_analysis/'
TBLPROPERTIES ('classification' = 'csv');
```

SQL Ln 19, Col 42

Completed Time in queue: 45 ms Run time: 398 ms Data scanned: -

Query to count total records:



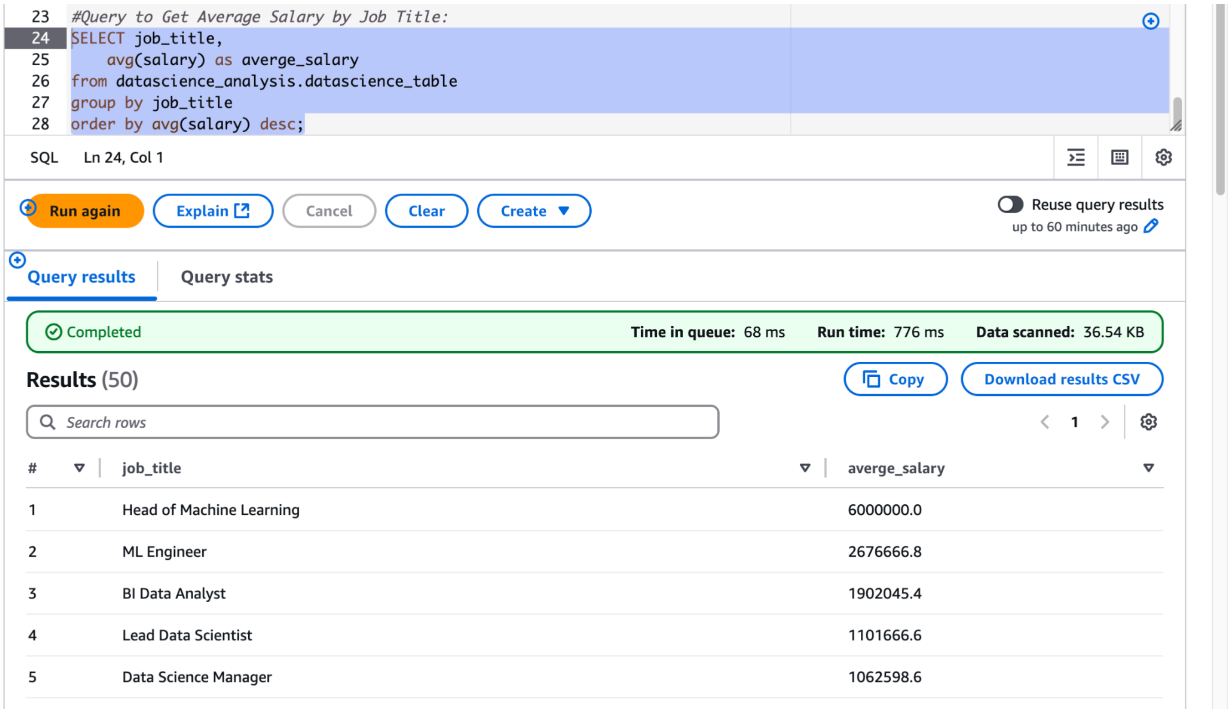
```sql
#Query to count total records
SELECT count(*) as total_records
from datascience_analysis.datascience_table;

```

Completed   **Time in queue:** 109 ms   **Run time:** 359 ms   **Data scanned:** 36.54 KB

Results (1)

#	total_records
1	607

Query to Get Average Salary by Job Title:



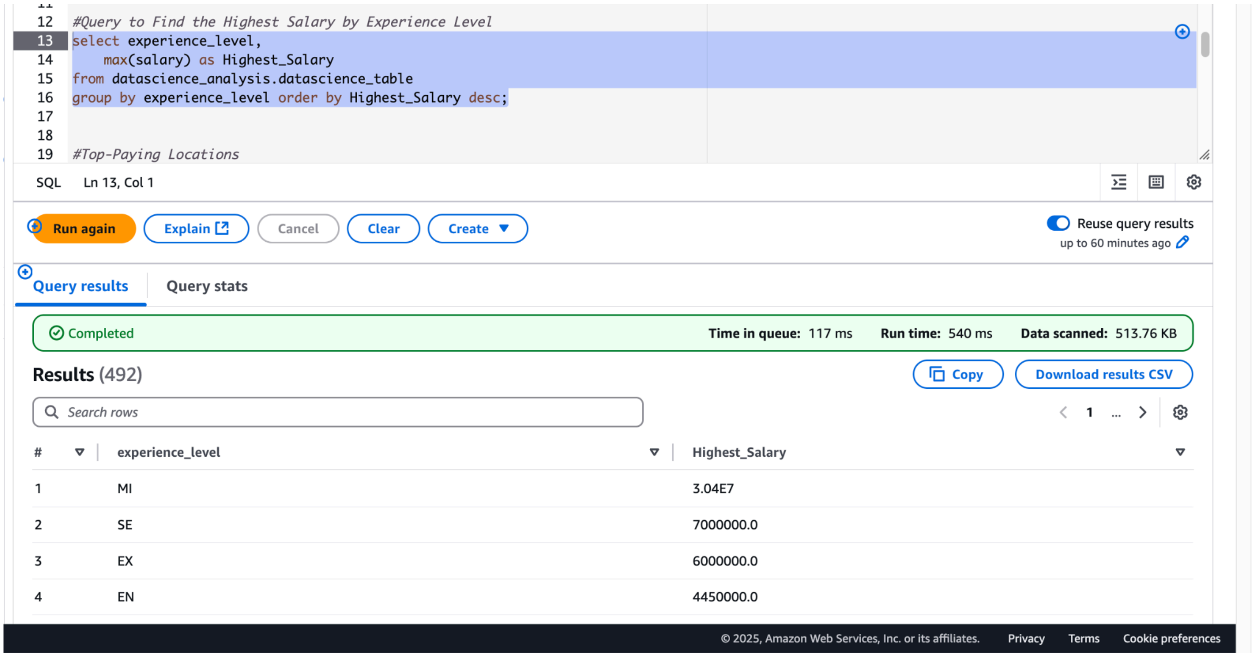
```sql
#Query to Get Average Salary by Job Title:
SELECT job_title,
    avg(salary) as averge_salary
from datascience_analysis.datascience_table
group by job_title
order by avg(salary) desc;
```

Completed   **Time in queue:** 68 ms   **Run time:** 776 ms   **Data scanned:** 36.54 KB

Results (50)

#	job_title	averge_salary
1	Head of Machine Learning	6000000.0
2	ML Engineer	2676666.8
3	BI Data Analyst	1902045.4
4	Lead Data Scientist	1101666.6
5	Data Science Manager	1062598.6

Query to Find the Highest Salary by Experience Level:



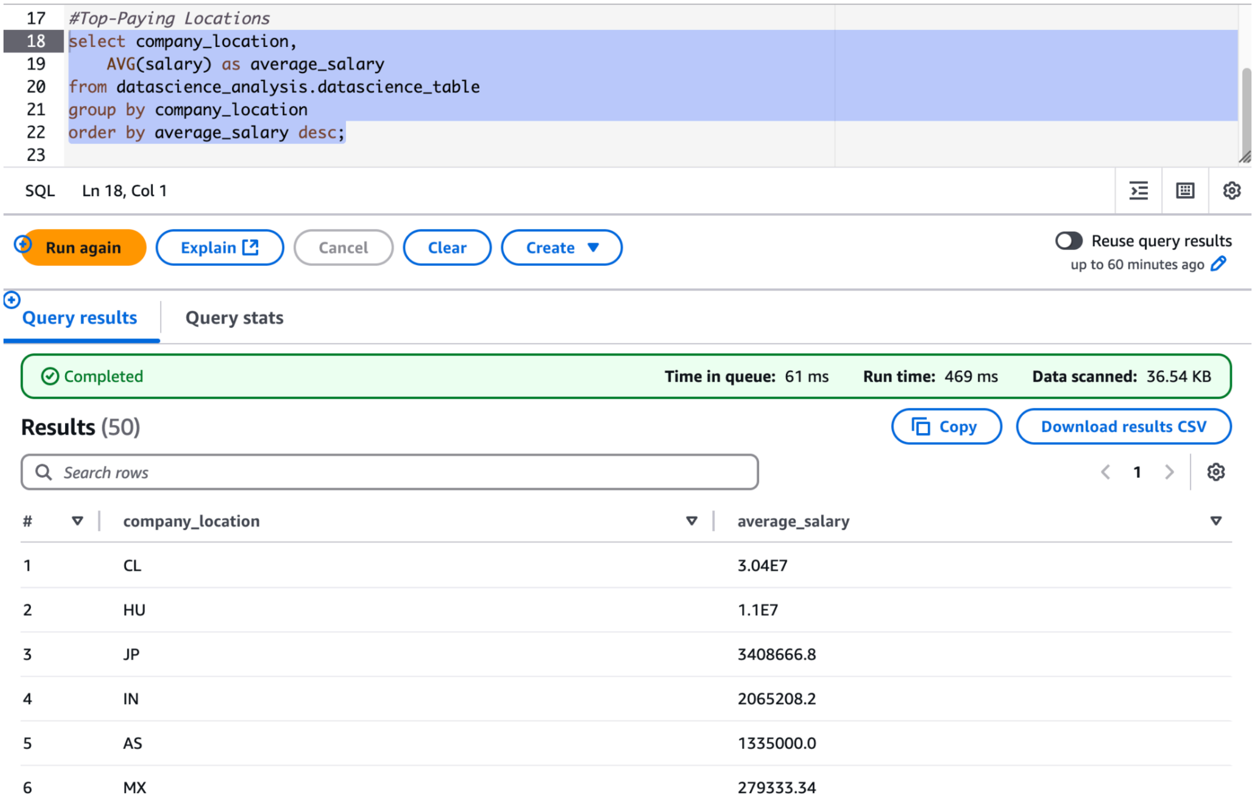
```
#Query to Find the Highest Salary by Experience Level
select experience_level,
    max(salary) as Highest_Salary
from datascience_analysis.datascience_table
group by experience_level order by Highest_Salary desc;


#Top-Paying Locations
```

Completed Time in queue: 117 ms Run time: 540 ms Data scanned: 513.76 KB

Results (492)

#	experience_level	Highest_Salary
1	MI	3.04E7
2	SE	7000000.0
3	EX	6000000.0
4	EN	4450000.0

Company Locations with the highest average salaries



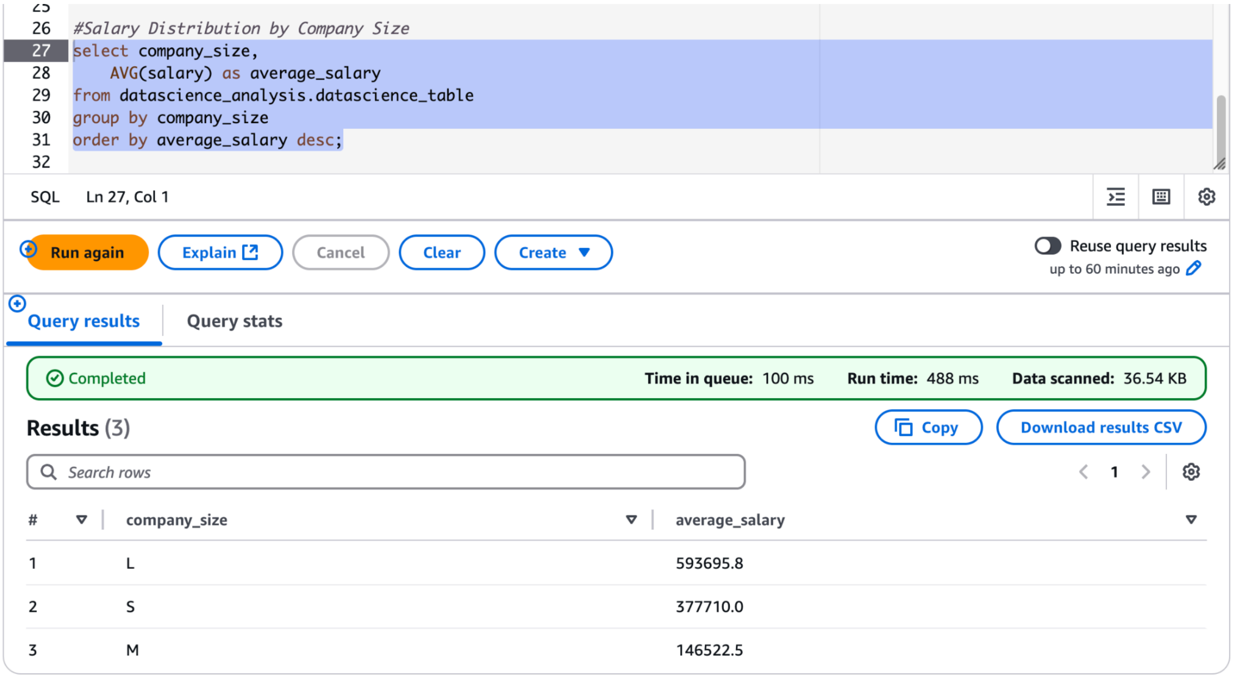
```
#Top-Paying Locations
select company_location,
    AVG(salary) as average_salary
from datascience_analysis.datascience_table
group by company_location
order by average_salary desc;

```

Completed Time in queue: 61 ms Run time: 469 ms Data scanned: 36.54 KB

Results (50)

#	company_location	average_salary
1	CL	3.04E7
2	HU	1.1E7
3	JP	3408666.8
4	IN	2065208.2
5	AS	1335000.0
6	MX	279333.34

Salary Distribution by Company Size



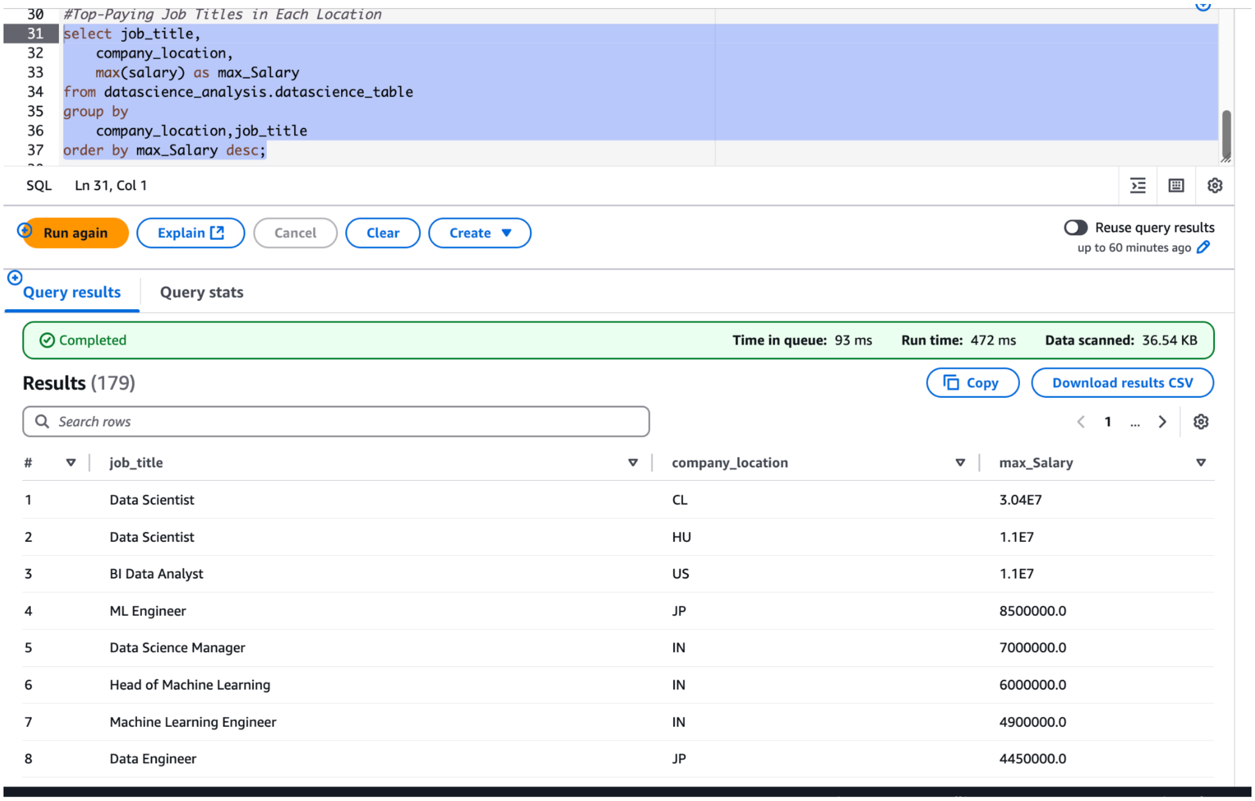
```
#Salary Distribution by Company Size
select company_size,
    AVG(salary) as average_salary
from datascience_analysis.datascience_table
group by company_size
order by average_salary desc;
```

Completed Time in queue: 100 ms Run time: 488 ms Data scanned: 36.54 KB

Results (3)

#	company_size	average_salary
1	L	593695.8
2	S	377710.0
3	M	146522.5

Top-Paying Job Titles in Each Location



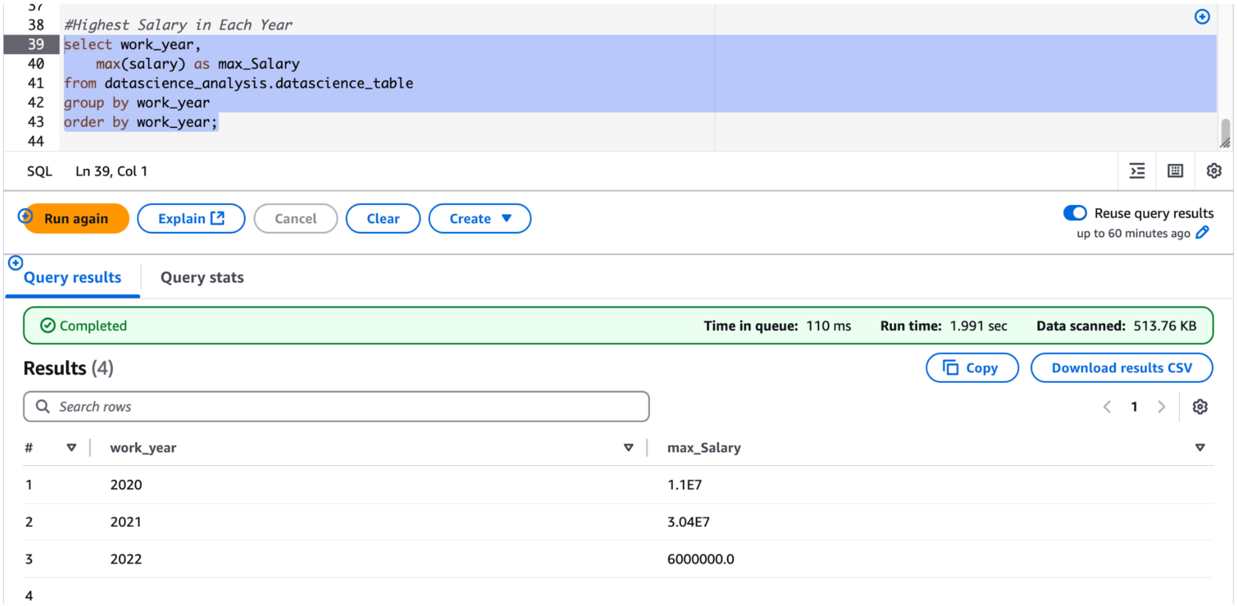
```
#Top-Paying Job Titles in Each Location
select job_title,
    company_location,
    max(salary) as max_Salary
from datascience_analysis.datascience_table
group by
    company_location,job_title
order by max_Salary desc;
```

Completed Time in queue: 93 ms Run time: 472 ms Data scanned: 36.54 KB

Results (179)

#	job_title	company_location	max_Salary
1	Data Scientist	CL	3.04E7
2	Data Scientist	HU	1.1E7
3	BI Data Analyst	US	1.1E7
4	ML Engineer	JP	8500000.0
5	Data Science Manager	IN	7000000.0
6	Head of Machine Learning	IN	6000000.0
7	Machine Learning Engineer	IN	4900000.0
8	Data Engineer	JP	4450000.0

Highest salary in each year



```
#Highest Salary in Each Year
select work_year,
    max(salary) as max_Salary
from datascience_analysis.datascience_table
group by work_year
order by work_year;
```

Completed   **Time in queue:** 110 ms   **Run time:** 1.991 sec   **Data scanned:** 513.76 KB

Results (4)

#	work_year	max_Salary
1	2020	1.1E7
2	2021	3.04E7
3	2022	6000000.0
4		

Insights and Results:

- **Head of Machine Learning** has the **highest average salary.**
- **Senior-level Software Engineer (SE)** earns the **highest salary by experience level.**
- **Chile (CL)** is the **top-paying location.**
- **Large companies (Company size L)** offer the **highest average salary.**
- **Data Scientist** is the **highest-ranking job title based on salary.**
- The **year 2020** recorded the **highest salary.**

The results are provided in the attached snapshots for each query.

Query Execution and Result Validation:

- Execution Method: Queries were executed directly in the Athena console.
- Result Location:

  `s3://data-science-job-salary-project/ds_analysis`

Part 3:

Step-by-Step ETL Implementation using AWS Glue for Data Science Job Salary Analysis

Overview

This project demonstrates an ETL pipeline built using AWS Glue to process Data Science Job Salary data. The pipeline extracts data from an S3, transforms it using AWS Glue Jobs, and loads the transformed data into S3 for reporting and analytics.

Data Flow Pipeline

> **Source: S3 (Raw data: CSV) → AWS Glue Job (Transform: Renaming columns & Normalize) → S3 (Processed Data: CSV Format) → QuickSight (for Analysis & Visualization)**

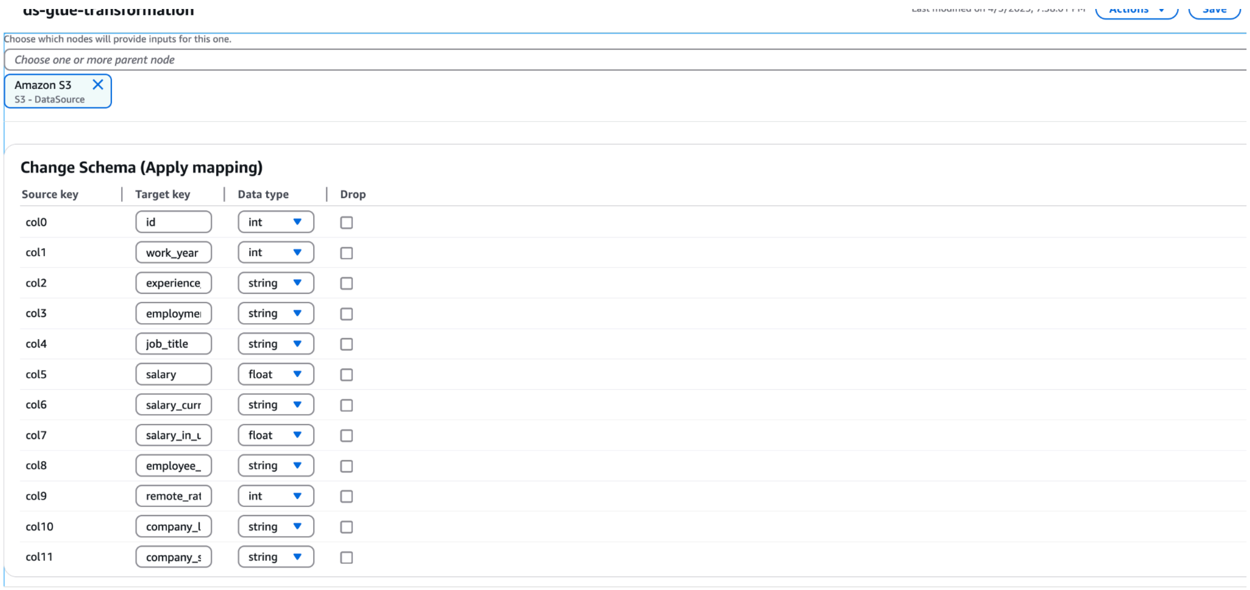
Data Extraction (Extract)

- Source: Amazon S3 (Raw Data) - The dataset includes salary details for Data Science job positions.

ETL Data Transformation (Transform) Steps in AWS Glue Studio

Steps:

1. Rename column names.



Choose which nodes will provide inputs for this one.

Choose one or more parent node

Amazon S3 ✕
S3 - DataSource

Change Schema (Apply mapping)

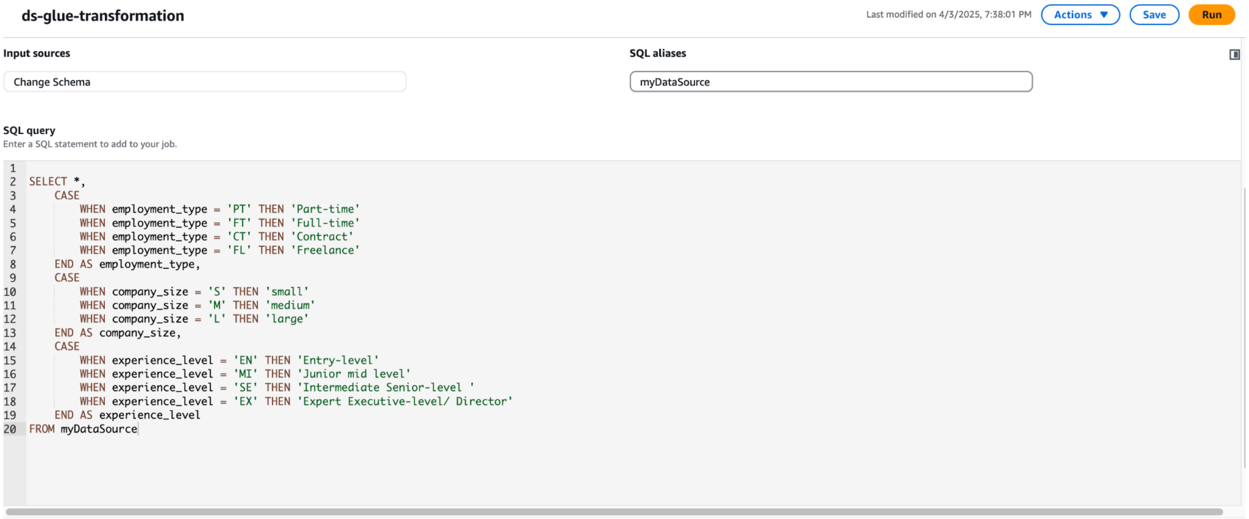
Source key	Target key	Data type	Drop
col0	id	int	□
col1	work_year	int	□
col2	experience	string	□
col3	employme	string	□
col4	job_title	string	□
col5	salary	float	□
col6	salary_curr	string	□
col7	salary_in_u	float	□
col8	employee_	string	□
col9	remote_rat	int	□
col10	company_l	string	□
col11	company_s	string	□

2. Standardizing Column Values

As part of the transformation process, the **employment_type, company_size, and experience_level** columns are standardized to improve data consistency. This includes:

- Converting **employment_type** (e.g., 'PT' → 'Part-time').
- Transforming **company_size** (e.g., 'S' → 'Small').
- Modifying **experience_level** (e.g., 'EN' → 'Entry-level').

The following SQL query is applied during the transformation to ensure uniformity:



ds-glue-transformation

Last modified on 4/3/2025, 7:38:01 PM

Input sources

Change Schema

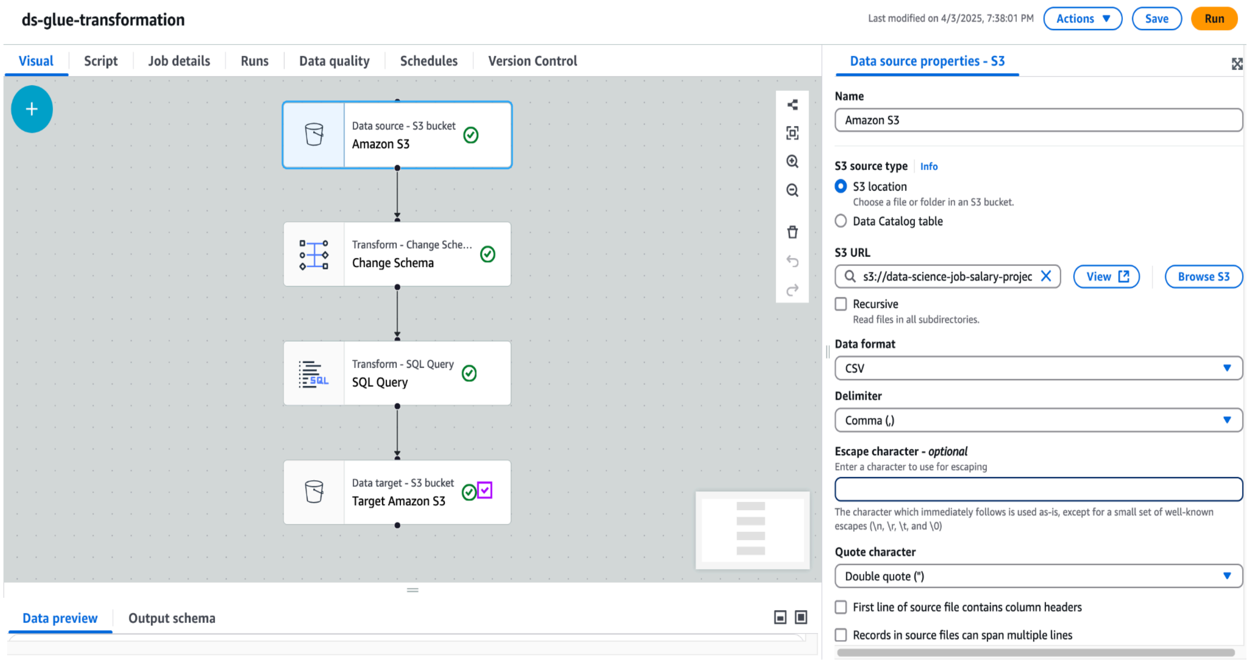
SQL aliases

myDataSource

SQL query

Enter a SQL statement to add to your job.

```sql

SELECT *,
    CASE
        WHEN employment_type = 'PT' THEN 'Part-time'
        WHEN employment_type = 'FT' THEN 'Full-time'
        WHEN employment_type = 'CT' THEN 'Contract'
        WHEN employment_type = 'FL' THEN 'Freelance'
    END AS employment_type,
    CASE
        WHEN company_size = 'S' THEN 'small'
        WHEN company_size = 'M' THEN 'medium'
        WHEN company_size = 'L' THEN 'large'
    END AS company_size,
    CASE
        WHEN experience_level = 'EN' THEN 'Entry-level'
        WHEN experience_level = 'MI' THEN 'Junior mid level'
        WHEN experience_level = 'SE' THEN 'Intermediate Senior-level '
        WHEN experience_level = 'EX' THEN 'Expert Executive-level/ Director'
    END AS experience_level
FROM myDataSource
```



Name: Amazon S3

S3 source type Info
● S3 location
Choose a file or folder in an S3 bucket.
○ Data Catalog table

S3 URL: s3://data-science-job-salary-projec
View
Browse S3

□ Recursive
Read files in all subdirectories.

Data format: CSV

Delimiter: Comma (,)

Escape character - *optional*
Enter a character to use for escaping

The character which immediately follows is used as-is, except for a small set of well-known escapes (\n, \r, \t, and \0)

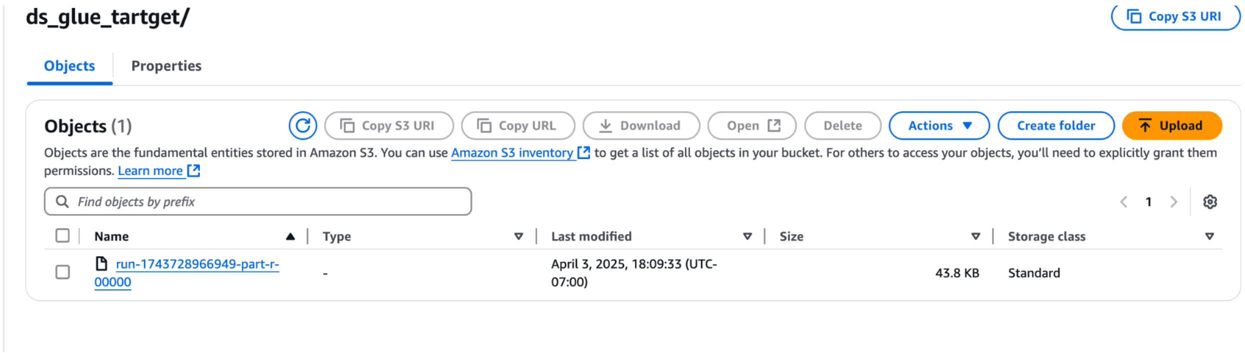
Quote character: Double quote (")

□ First line of source file contains column headers

□ Records in source files can span multiple lines

Data Loaded (Load) to S3

- **Processed Data Format:** CSV
- **Storage:** **s3://data-science-job-salary-project/ds_glue_tartget/**



Objects are the fundamental entities stored in Amazon S3. You can use Amazon S3 inventory to get a list of all objects in your bucket. For others to access your objects, you'll need to explicitly grant them permissions. Learn more

□	Name	Type	Last modified	Size	Storage class
□	run-1743728966949-part-r-00000	-	April 3, 2025, 18:09:33 (UTC-07:00)	43.8 KB	Standard

Part 4:
Visualization using QUICKSIGHT

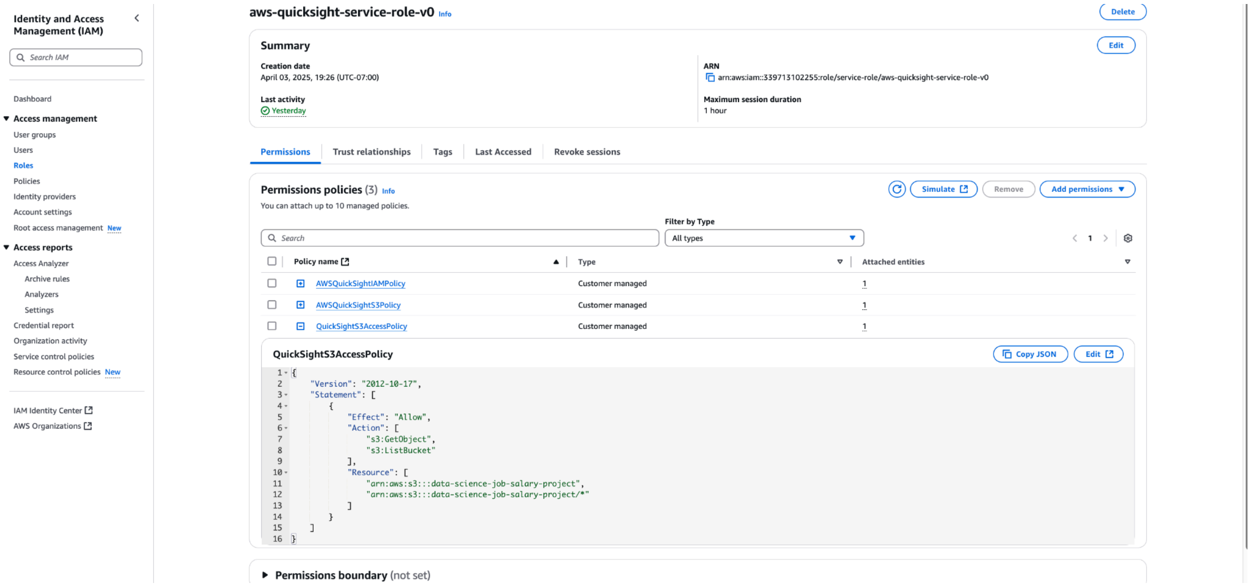
Overview

In this section, we will analyze the **Data Science Job Salaries** dataset to reveal key insights, including salary trends, job role distributions, and other significant patterns. Leveraging **Amazon QuickSight**, to build a visually engaging dashboard that delivers a clear and comprehensive overview of the data.

Configuring Permissions for QuickSight Access

To allow Amazon QuickSight to access the data stored in Amazon S3, we configured the following two settings:

1. **IAM Policy Configuration**
   Create and attach an IAM policy that grants the necessary permissions—specifically `s3:GetObject` and `s3:ListBucket`—to the IAM role (`aws-quicksight-service-role-v0`) or the user associated with QuickSight.



aws-quicksight-service-role-v0

Summary

Creation date: April 03, 2025, 19:26 (UTC-07:00)

Last activity: Yesterday

ARN: arn:aws:iam::339713102255:role/service-role/aws-quicksight-service-role-v0

Maximum session duration: 1 hour

Permissions policies (3)

You can attach up to 10 managed policies.

Filter by Type: All types

Policy name	Type	Attached entities
AWSQuickSightIAMPolicy	Customer managed	1
AWSQuickSightS3Policy	Customer managed	1
QuickSightS3AccessPolicy	Customer managed	1

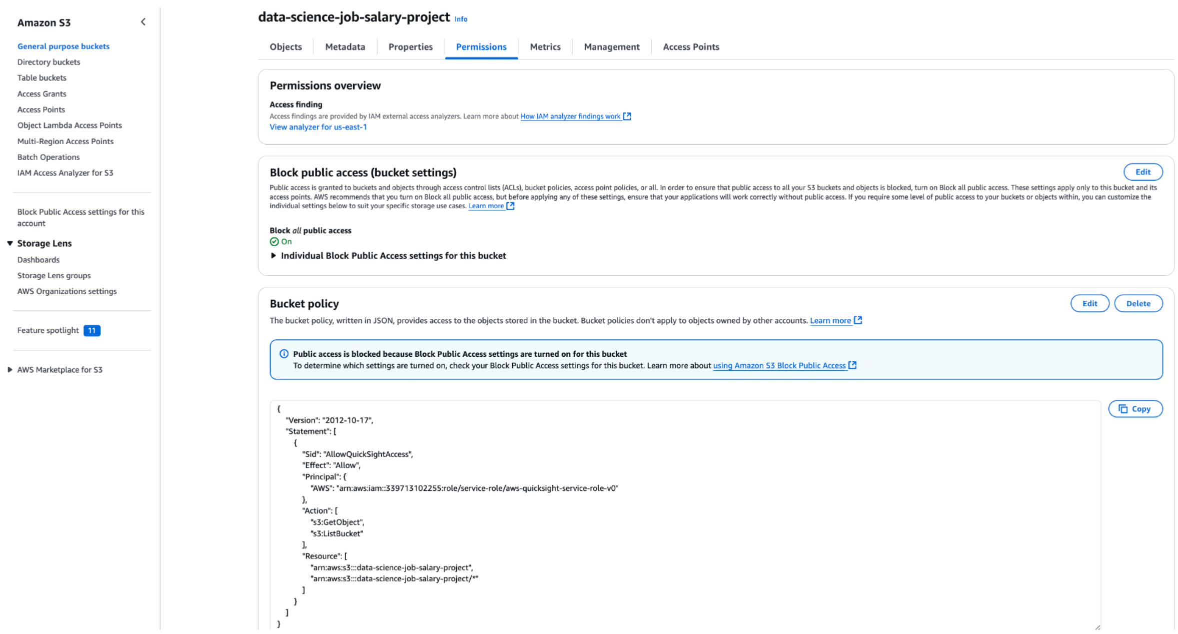
QuickSightS3AccessPolicy

```
{
    "Version": "2012-10-17",
    "Statement": [
        {
            "Effect": "Allow",
            "Action": [
                "s3:GetObject",
                "s3:ListBucket"
            ],
            "Resource": [
                "arn:aws:s3:::data-science-job-salary-project",
                "arn:aws:s3:::data-science-job-salary-project/*"
            ]
        }
    ]
}
```

Permissions boundary (not set)

2. S3 Bucket Policy Update

Modify the S3 bucket policy to explicitly grant access to QuickSight. This ensures that QuickSight can properly retrieve data from the specified S3 bucket.



data-science-job-salary-project

Permissions overview

Access finding

Access findings are provided by IAM external access analyzers. Learn more about How IAM analyzer findings work

View analyzer for us-east-1

Block public access (bucket settings)

Public access is granted to buckets and objects through access control lists (ACLs), bucket policies, access point policies, or all. In order to ensure that public access to all your S3 buckets and objects is blocked, turn on Block all public access. These settings apply only to this bucket and its access points. AWS recommends that you turn on Block all public access, but before applying any of these settings, ensure that your applications will work correctly without public access. If you require some level of public access to your buckets or objects within, you can customize the individual settings below to suit your specific storage use cases. Learn more

Block *all* public access

✓ On

▶ Individual Block Public Access settings for this bucket

Bucket policy

The bucket policy, written in JSON, provides access to the objects stored in the bucket. Bucket policies don't apply to objects owned by other accounts. Learn more

Public access is blocked because Block Public Access settings are turned on for this bucket
To determine which settings are turned on, check your Block Public Access settings for this bucket. Learn more about using Amazon S3 Block Public Access

```
{
  "Version": "2012-10-17",
  "Statement": [
    {
      "Sid": "AllowQuickSightAccess",
      "Effect": "Allow",
      "Principal": {
        "AWS": "arn:aws:iam::339713102255:role/service-role/aws-quicksight-service-role-v0"
      },
      "Action": [
        "s3:GetObject",
        "s3:ListBucket"
      ],
      "Resource": [
        "arn:aws:s3:::data-science-job-salary-project",
        "arn:aws:s3:::data-science-job-salary-project/*"
      ]
    }
  ]
}
```

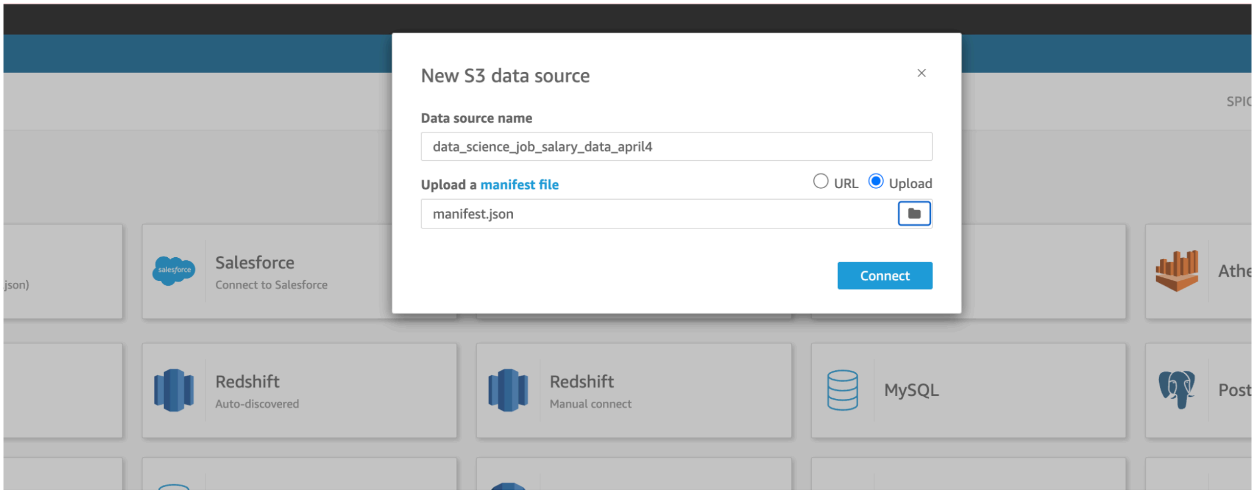
Data Source:

- Transformed data from the AWS Glue ETL job, stored in an S3 bucket.
- Data: S3 Bucket:

  s3://data-science-job-salary-project/ds_glue_tartget/run-1743783775548-part-r-00000

Connecting QuickSight to the Dataset S3:

A new dataset named data_science_job_salary_data_april4 was created in QuickSight by connecting to the S3 bucket (s3://data-science-job-salary-project/ds_glue_target/run-1743783775548-part-r-00000), where the transformed data is stored in CSV format.



New S3 data source

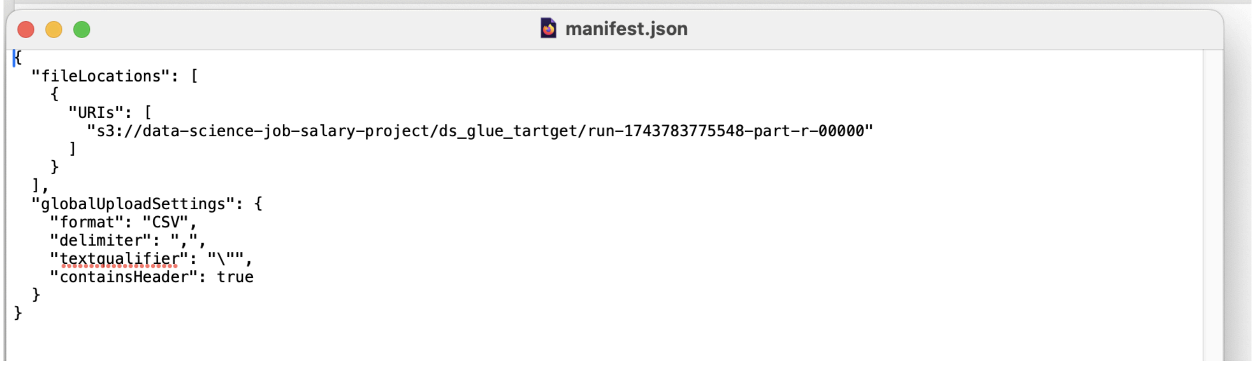
Data source name

data_science_job_salary_data_april4

Upload a manifest file ○ URL ● Upload

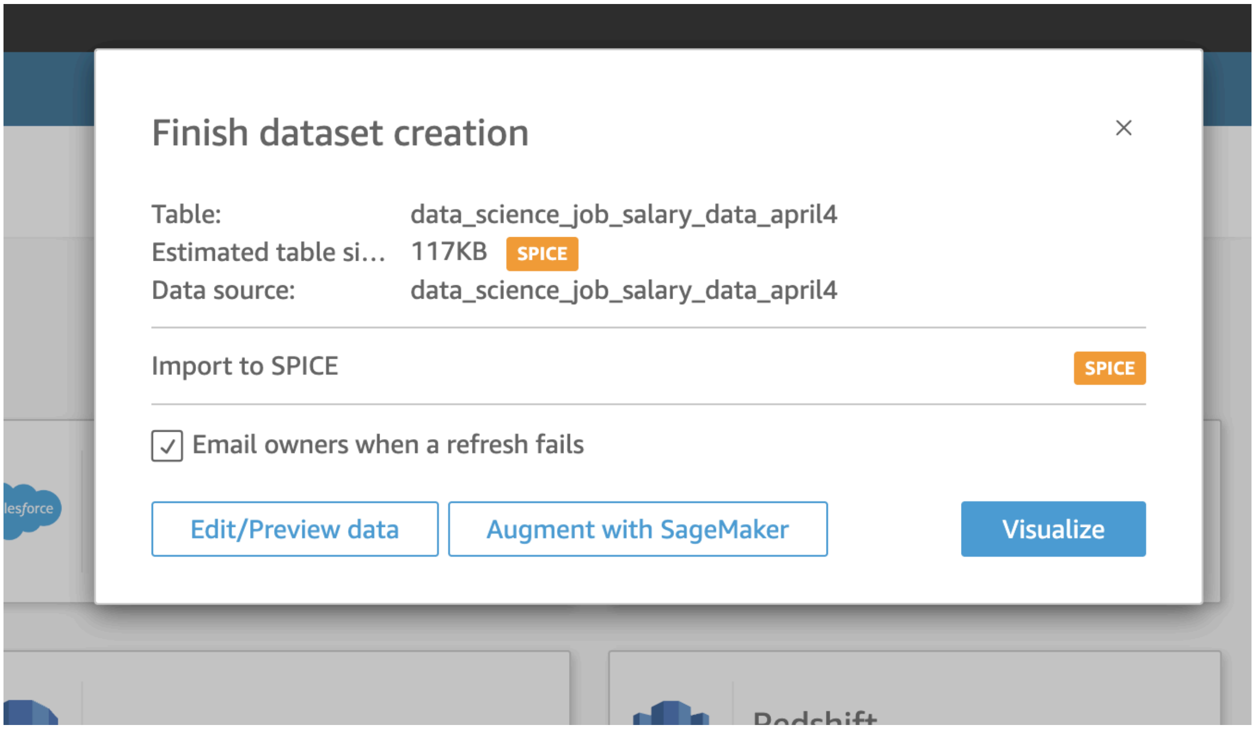
manifest.json

Manifest.json file: a manifest file is used to connect to data stored in Amazon S3. It helps QuickSight understand how to locate and interpret your files in S3.



manifest.json

```
{
  "fileLocations": [
    {
      "URIs": [
        "s3://data-science-job-salary-project/ds_glue_tartget/run-1743783775548-part-r-00000"
      ]
    }
  ],
  "globalUploadSettings": {
    "format": "CSV",
    "delimiter": ",",
    "textqualifier": "\"",
    "containsHeader": true
  }
}
```



Finish dataset creation

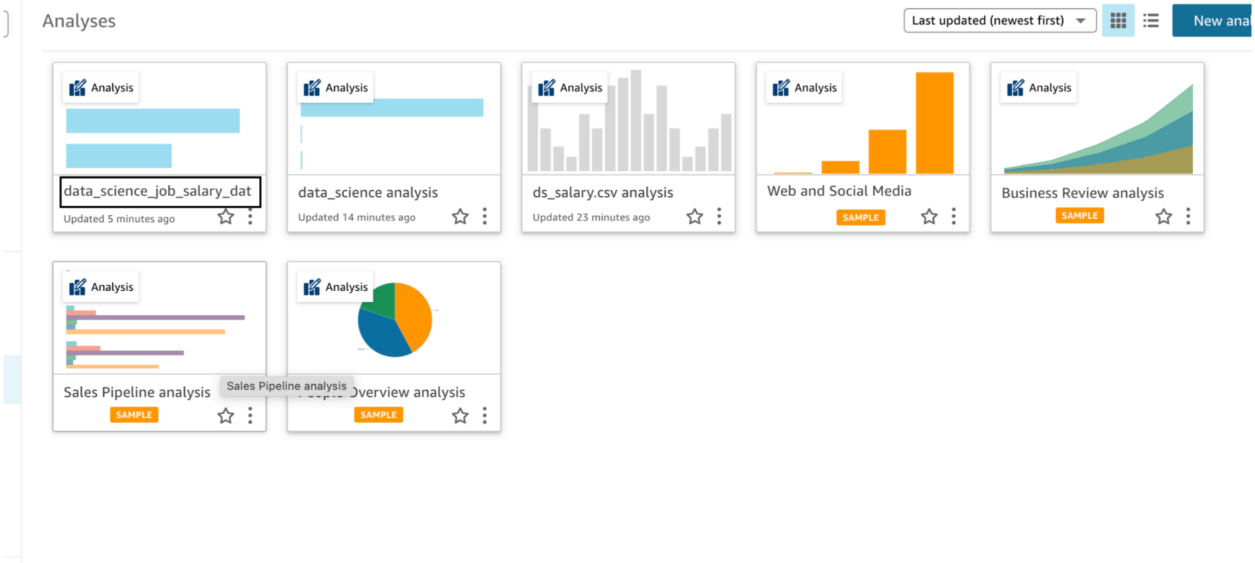
Table: data_science_job_salary_data_april4

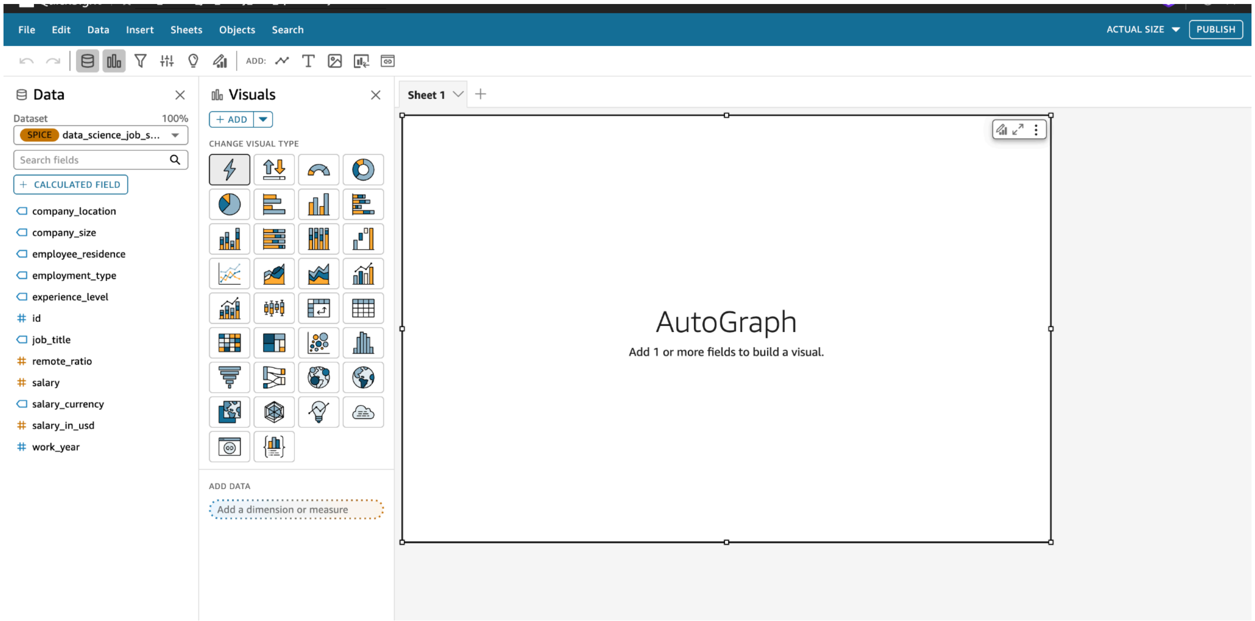
Estimated table si… 117KB SPICE

Data source: data_science_job_salary_data_april4

Import to SPICE SPICE

☑ Email owners when a refresh fails





Creating Visualizations in QuickSight:

The following visualizations were developed using Amazon QuickSight:

- Created below visualizations:
- Average Salary by Job Title
- Average Salary by Employee Country
- Average Salary by Experience Level
- Average Salary by Company Size
- Average Salary by Remote Ratio
- Average Salary Trend Over Years
- Distribution of Employees by Experience Level

- Job Title Distribution
- Funnel View of Job Titles
- Employee Distribution by Job Title and Company Size
- Experience Level Distribution by Employment Type
- Funnel View of Employees by Residence

Sharing the Dashboard:

- **Embedding:** The Amazon QuickSight dashboard embedded into a web page. Access to the embedded dashboard requires an AWS account**.**

  https://us-east-1.quicksight.aws.amazon.com/sn/accounts/339713102255/dashboards/f3b37240-37ea-4dc9-bf06-bd86298a22b1?directory_alias=seema

- **Exporting:** The dashboard exported as a PDF.

  Click here to VIEW PDF

Dashboard Insights:
Key Findings from the Data Science Job Salary Analysis Dashboard

- The majority of employees fall under the **Intermediate/Senior-level** experience category.
- **Data Scientist** is the most common job title among employees.
- **Head of Machine Learning** roles command the highest average salary among all job titles, while the J**unior/Mid-level** experience group earns the highest average salary compared to other experience levels.
- **Large companies** offer the highest average salaries compared to medium and small companies.
- Employees with a **fully remote** work setup (remote_ratio = 100%) earn the highest average salary.
- The year **2021** recorded the highest average salary across all years.
- **Full-time** employment is the most common, with the **Intermediate/Senior-level** making up the largest portion of this group.

- Among all job titles, **Data Scientist** roles are most prevalent, with the majority working in **medium-sized companies**.
- Employees residing in **CL** (Chile) have the highest average salary by country of residence.
- The **United States (US)** has the highest number of employees represented in the dataset.