**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](https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries/data)), which offers real-world insights into salary trends and compensation structures within the data science profession.

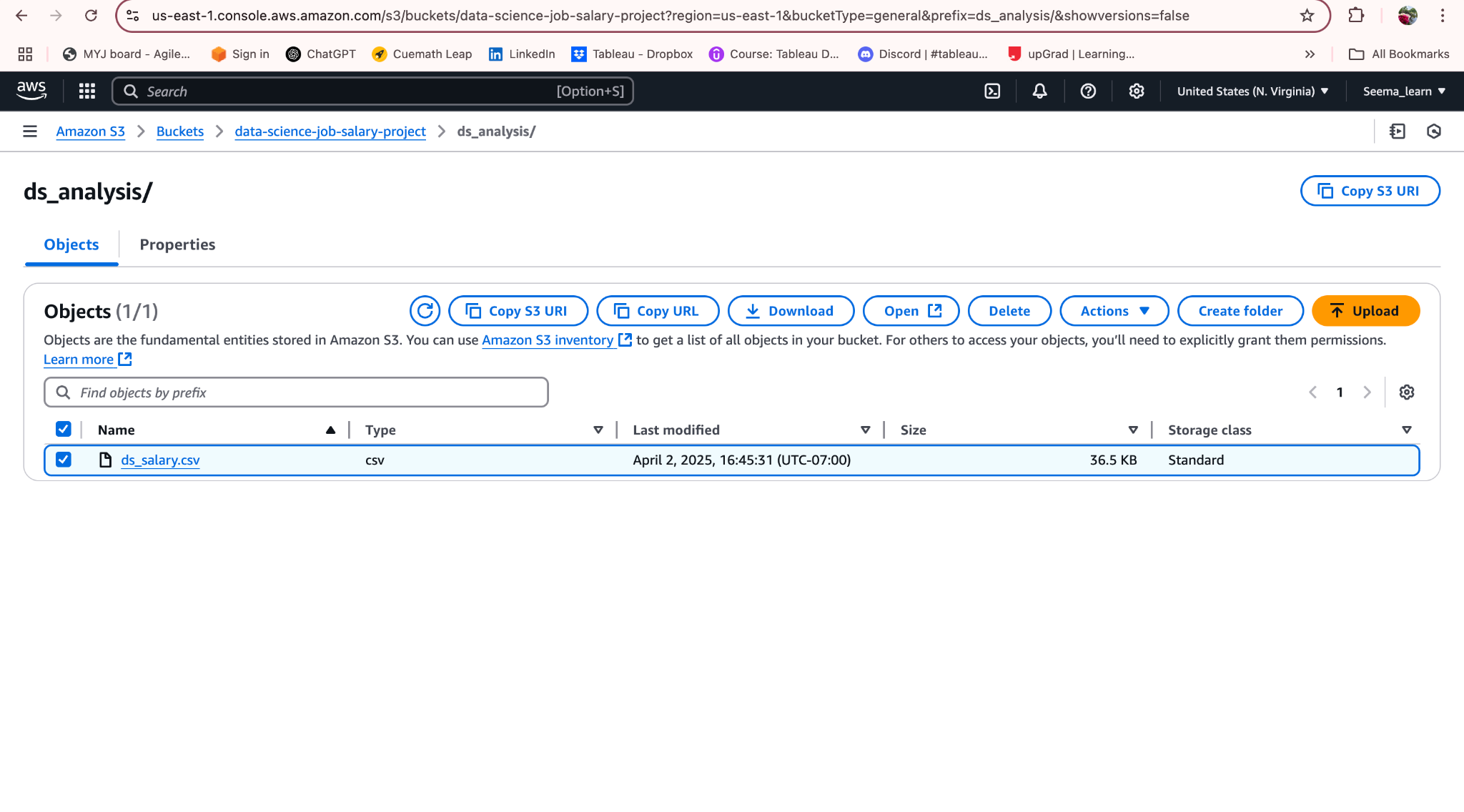
The primary objective is to build a scalable, serverless data pipeline that:

* 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**](https://www.kaggle.com/datasets/ruchi798/data-science-job-salaries/data)) 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:  


* **S3 URI:**

**s3://data-science-job-salary-project/ds\_analysis/ds\_salary.csv**

**Part 2:**

**SQL Query Documentation for Athena-based Data Analysis**

**Introduction:**

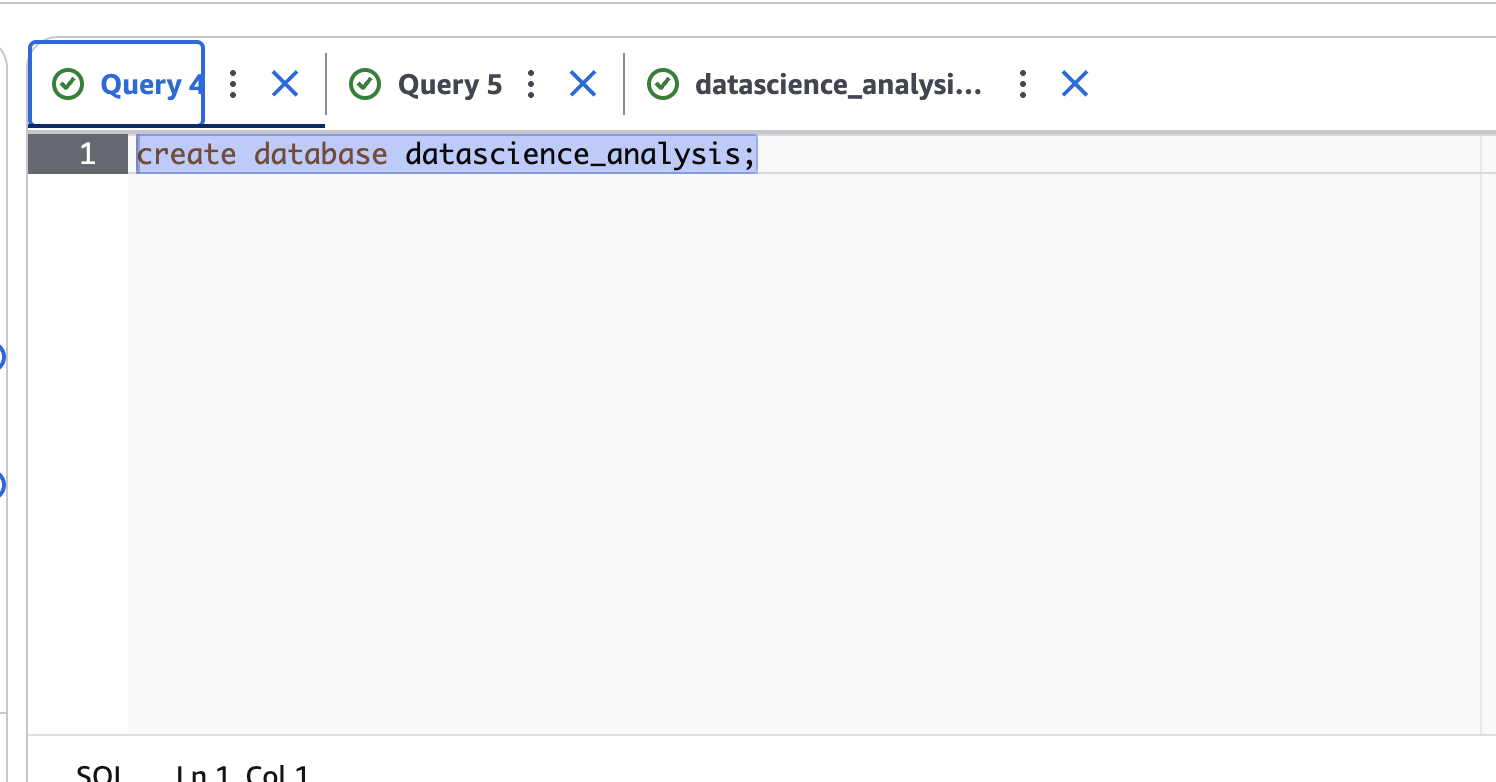
* **Purpose of the analysis**:This document outlines the SQL queries executed in AWS Athena to perform data analysis on the "ds\_salary.csv" dataset, which was uploaded to an S3 bucket. The analysis aimed to derive insights on salary trends based on job roles, experience, and location.
* **Tools Used:**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

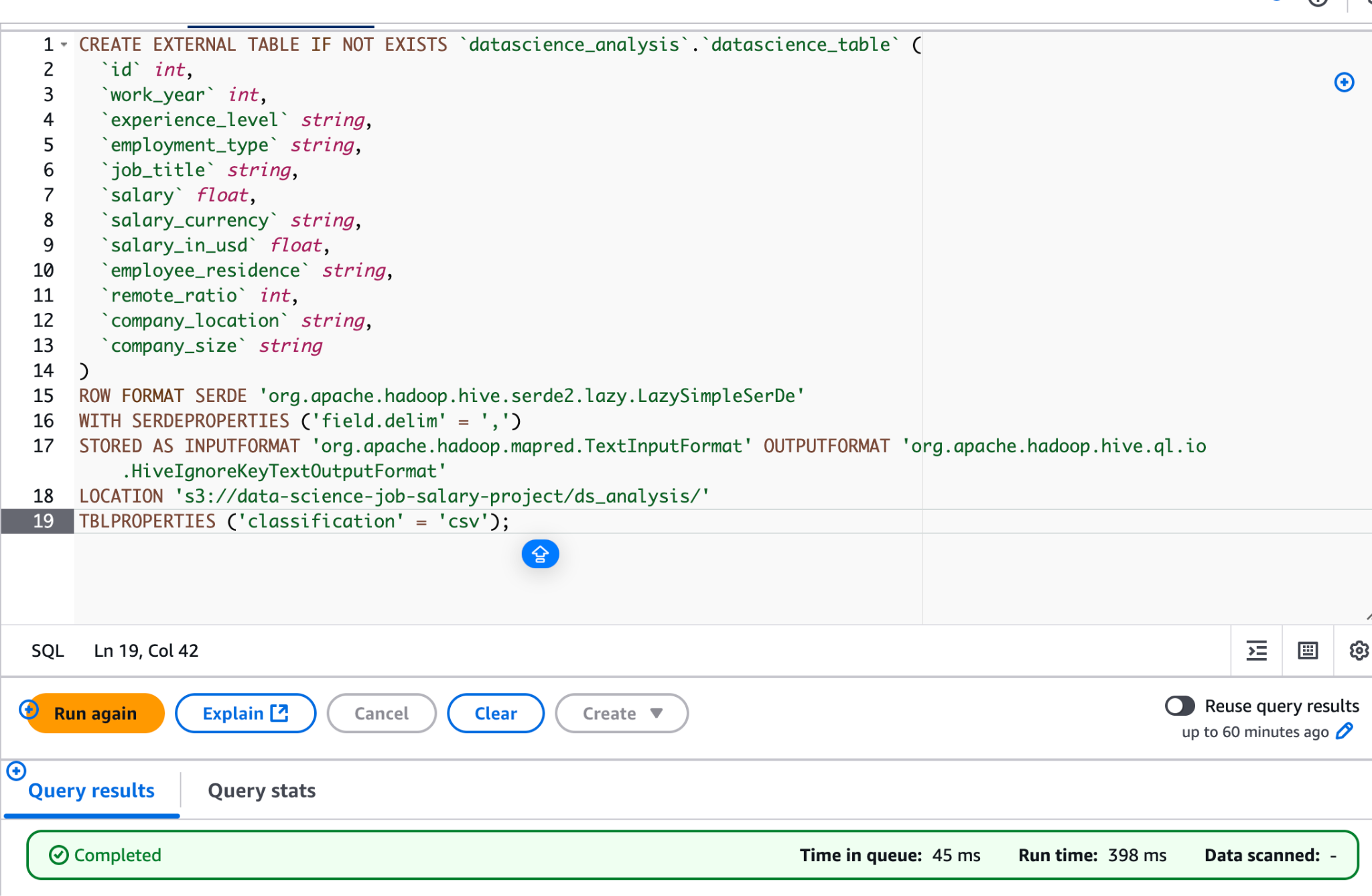
### **Data Source:**

* **Dataset**: "ds\_salary.csv" from Kaggle, uploaded to an S3 bucket.
* **S3 Bucket**: s3://data-science-job-salary-project/ds\_analysis/ds\_salary.csv

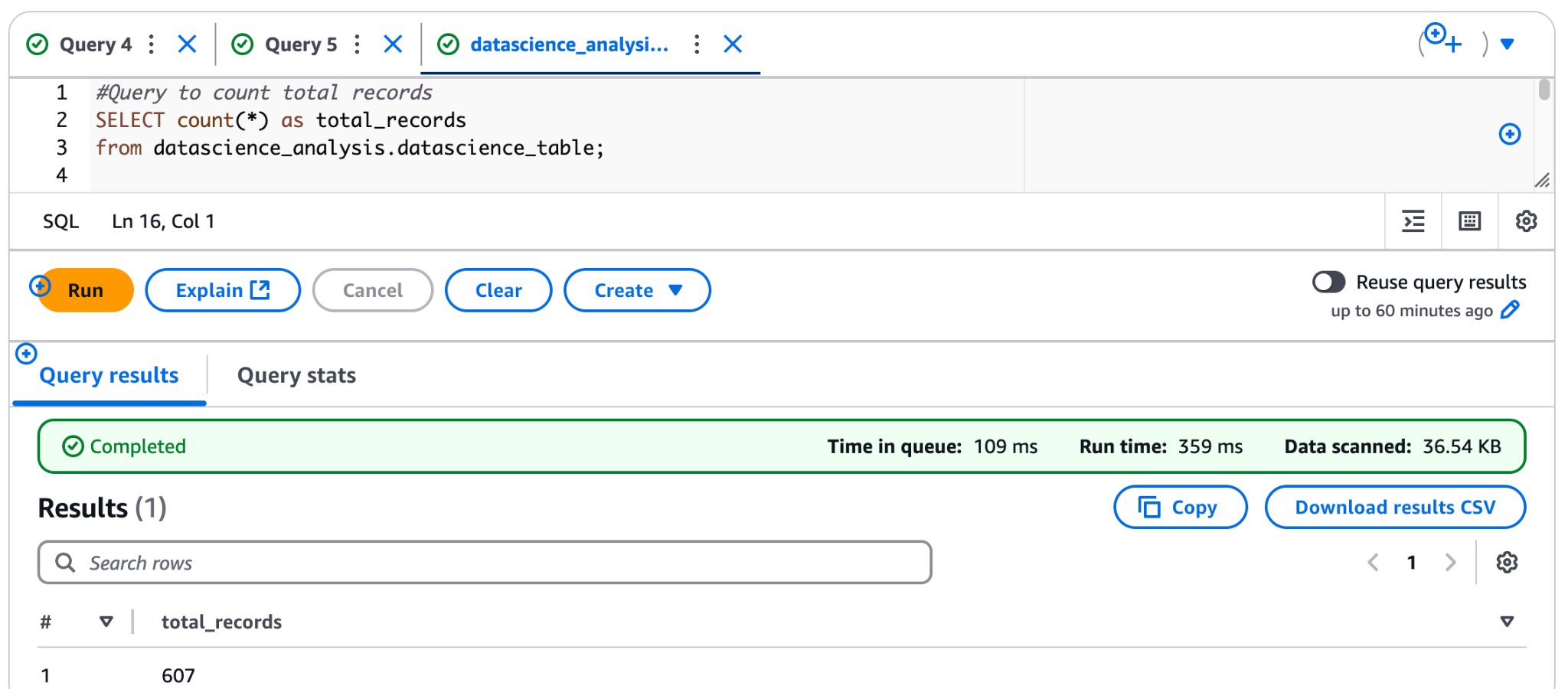
### **SQL Query Documentation:**

Database & Table Creation

****

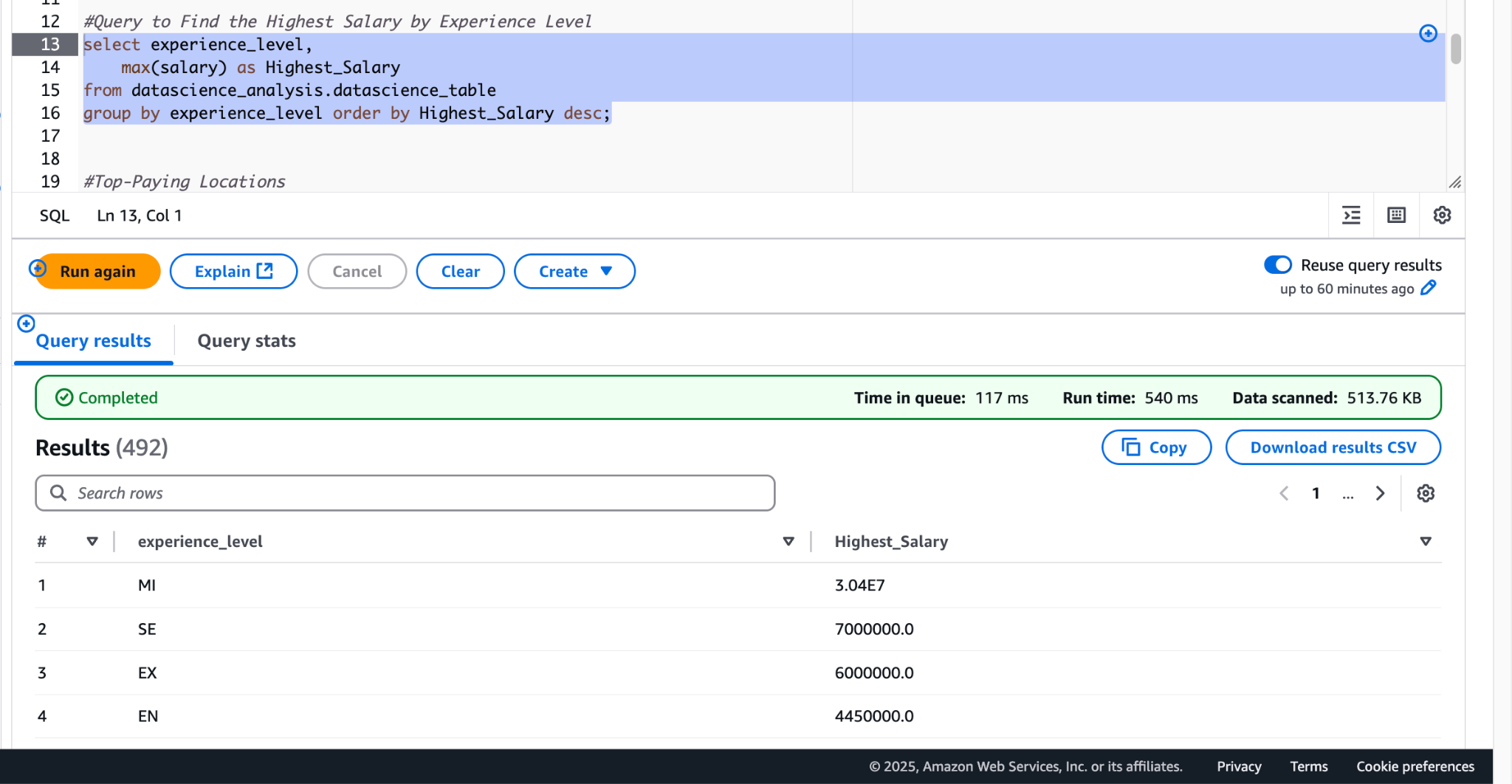
* ****

**Query to count total records:**

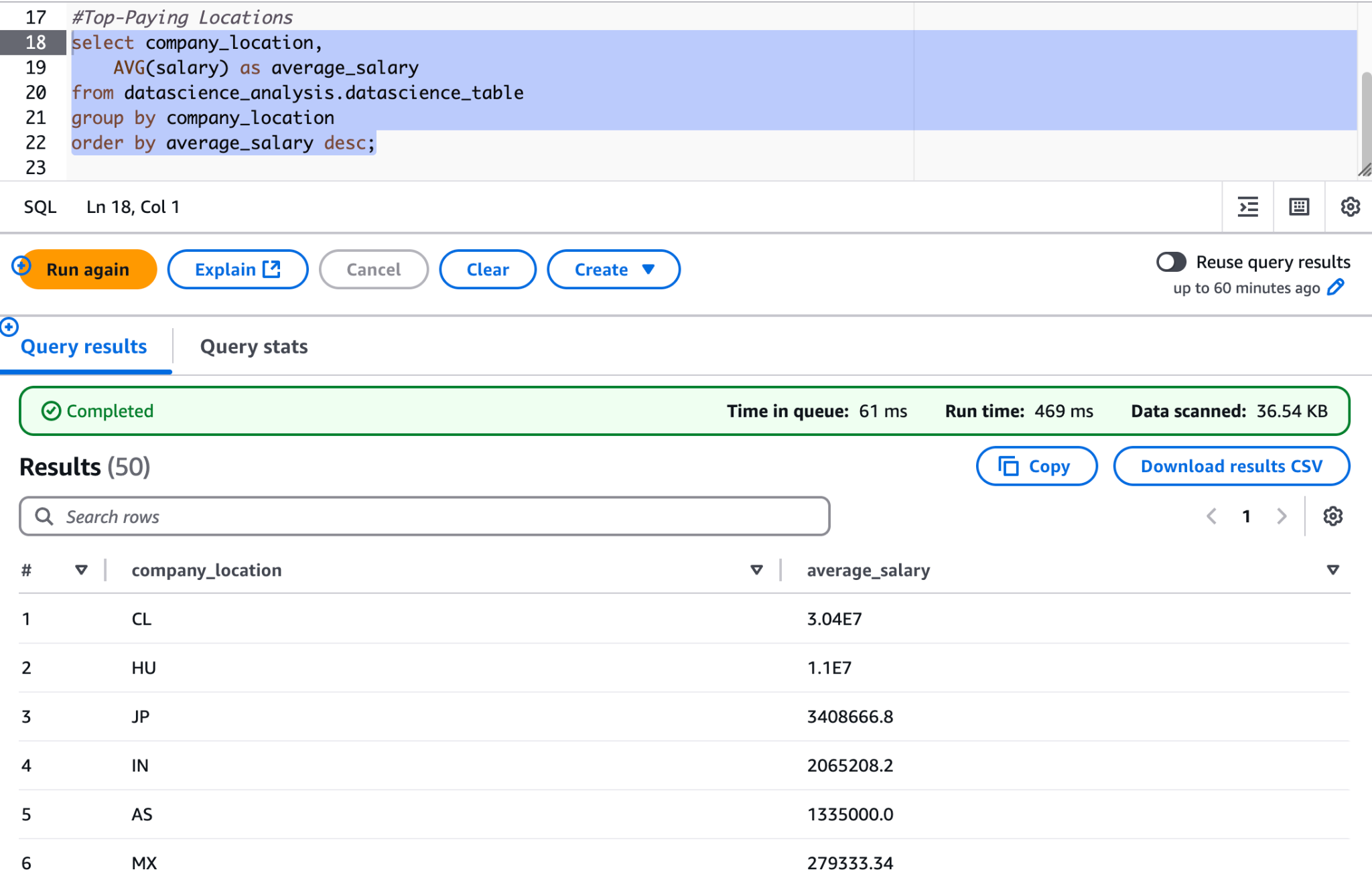
****

**Query to Get Average Salary by Job Title:**

****

**Query to Find the Highest Salary by Experience Level:  
  
**

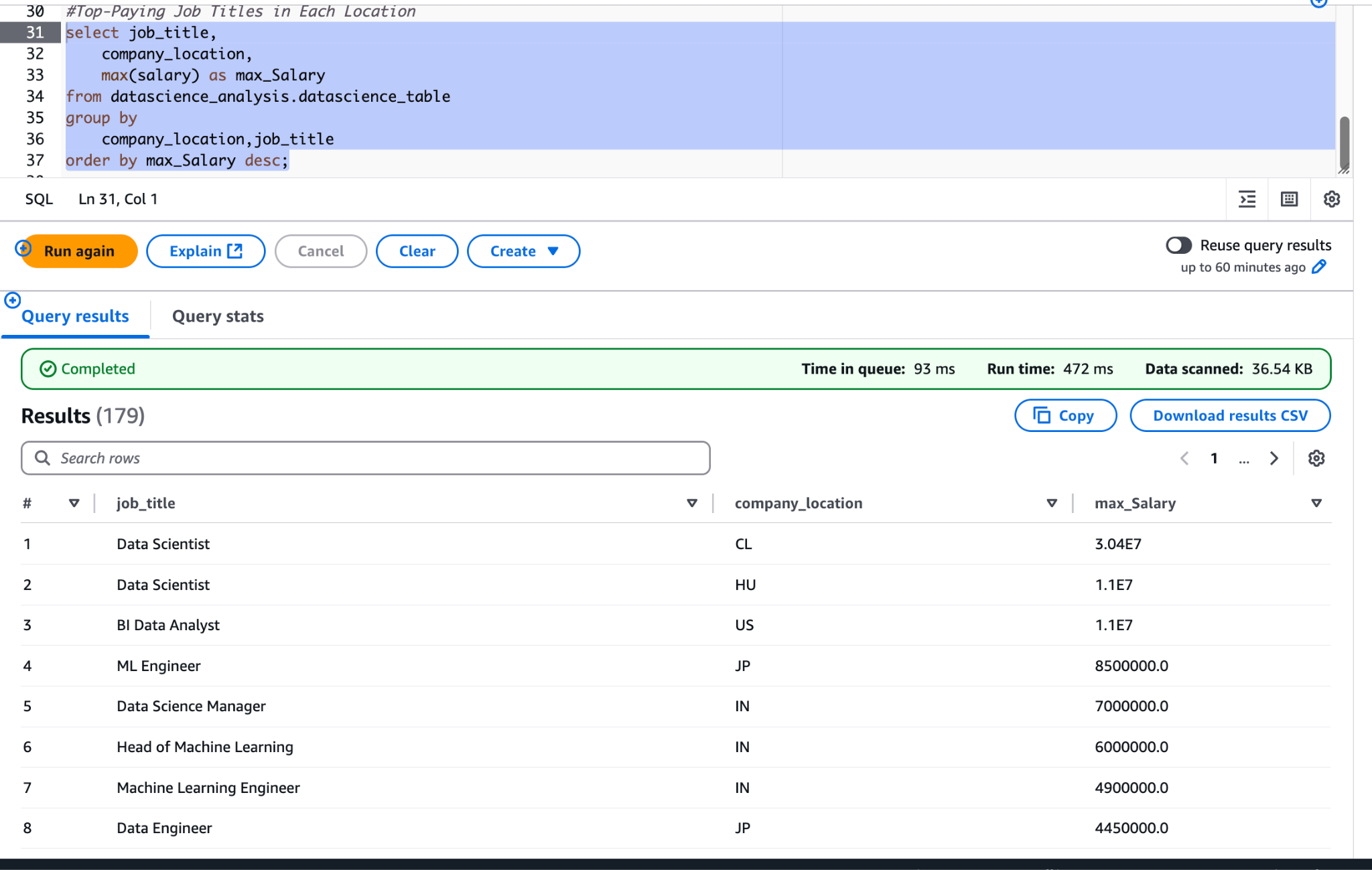
**Company Locations with the highest average salaries**

****

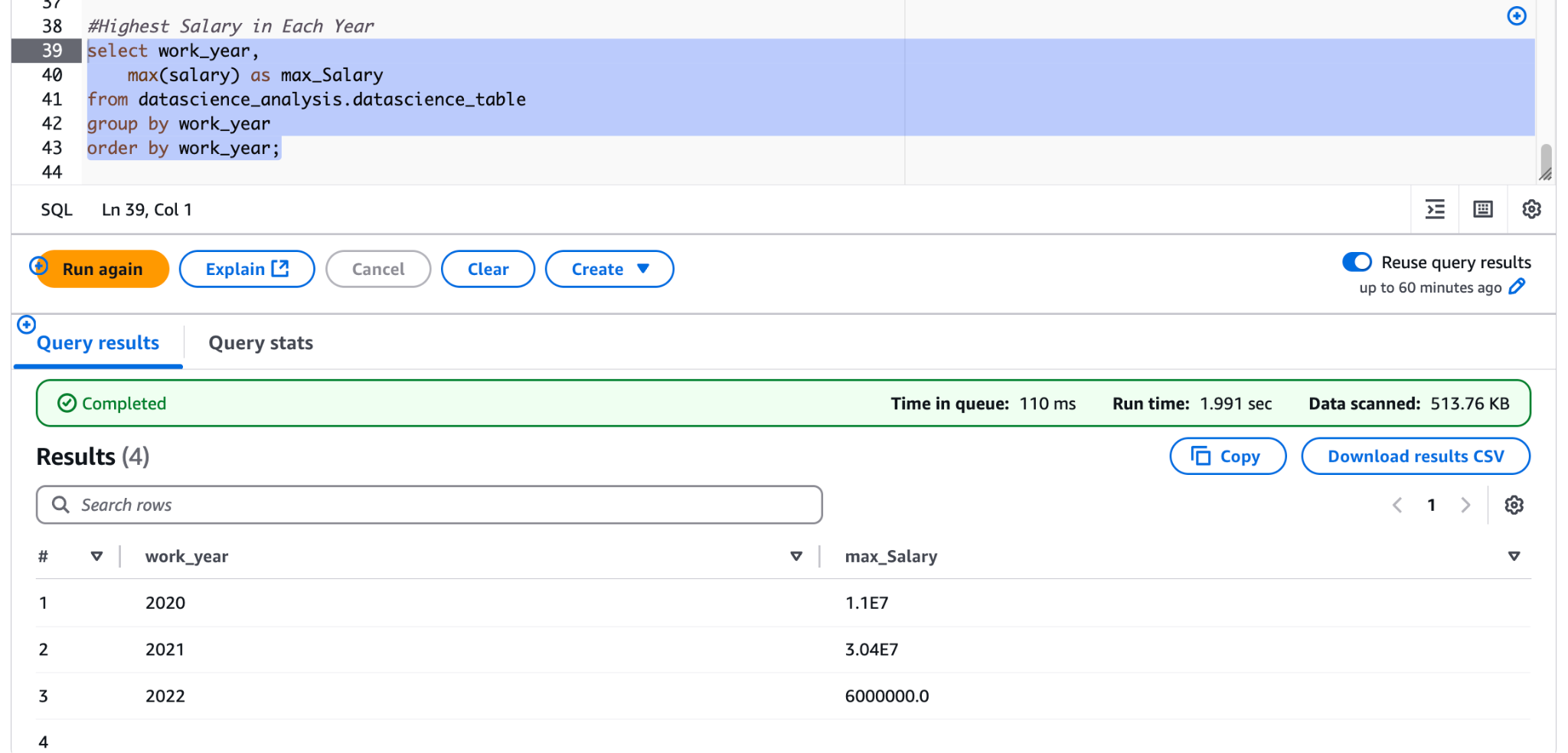
**Salary Distribution by Company Size**

****

### **Top-Paying Job Titles in Each Location**



**Highest salary in each year**



### **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** recordedthe **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.

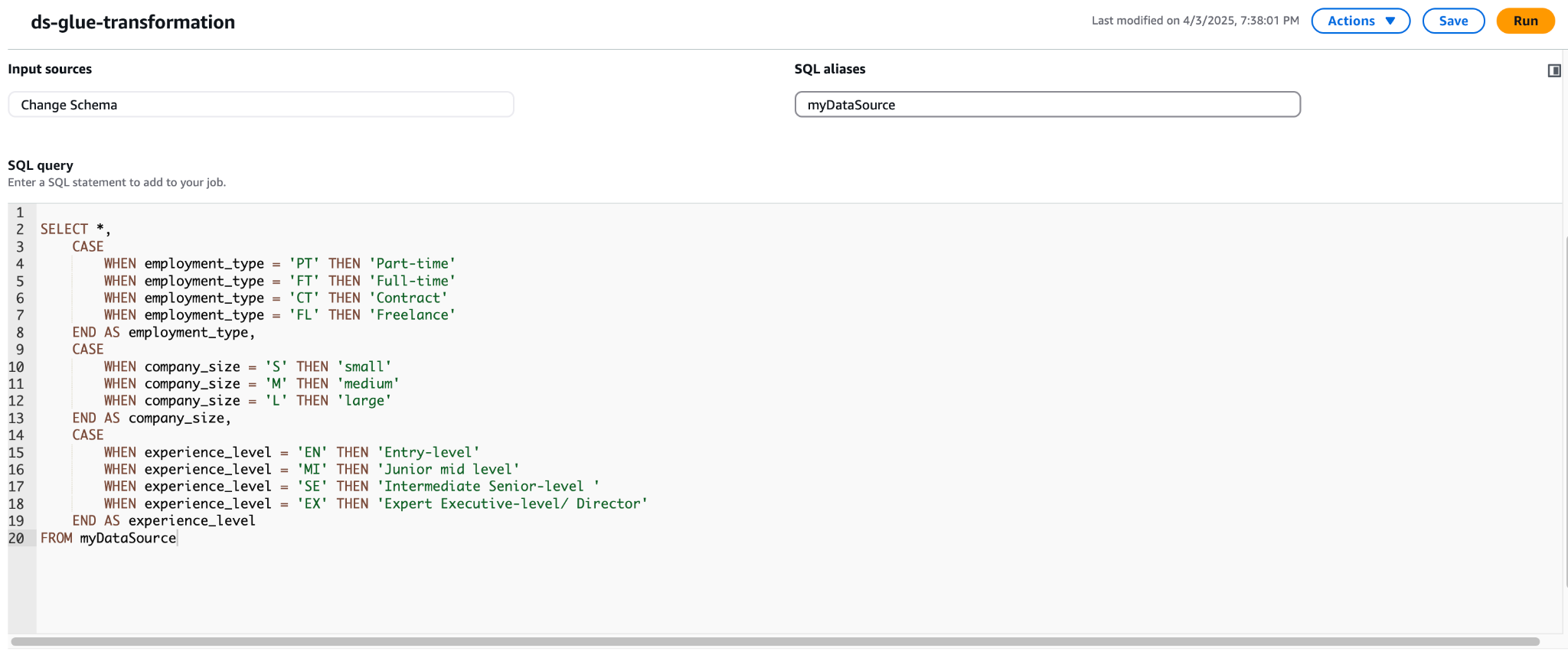


1. Standardizing Column Values

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').

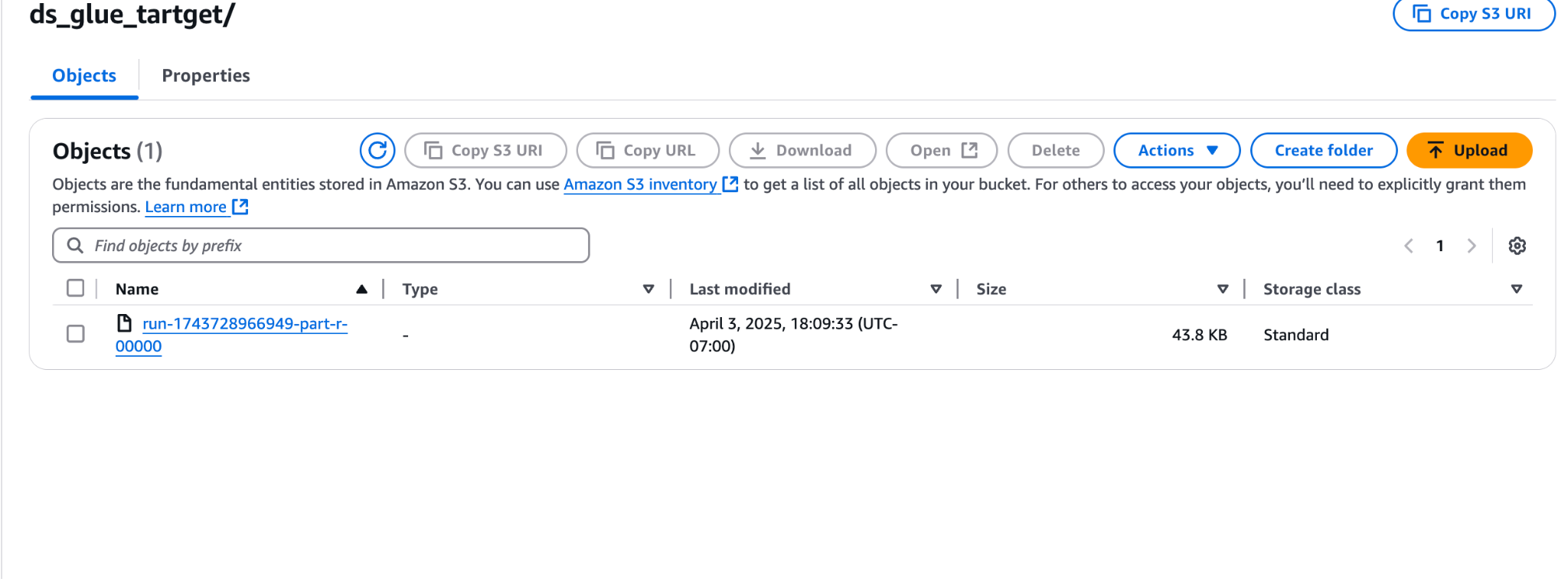
The following SQL query is applied during the transformation to ensure uniformity:



## 

## **Data Loaded (Load) to S3**

* **Processed Data Format:** CSV
* **Storage:** [**s3://data-science-job-salary-project/ds\_glue\_tartget/**](https://us-east-1.console.aws.amazon.com/s3/buckets/data-science-job-salary-project?region=us-east-1&bucketType=general&prefix=ds_glue_tartget/)

****

**Part 4:**

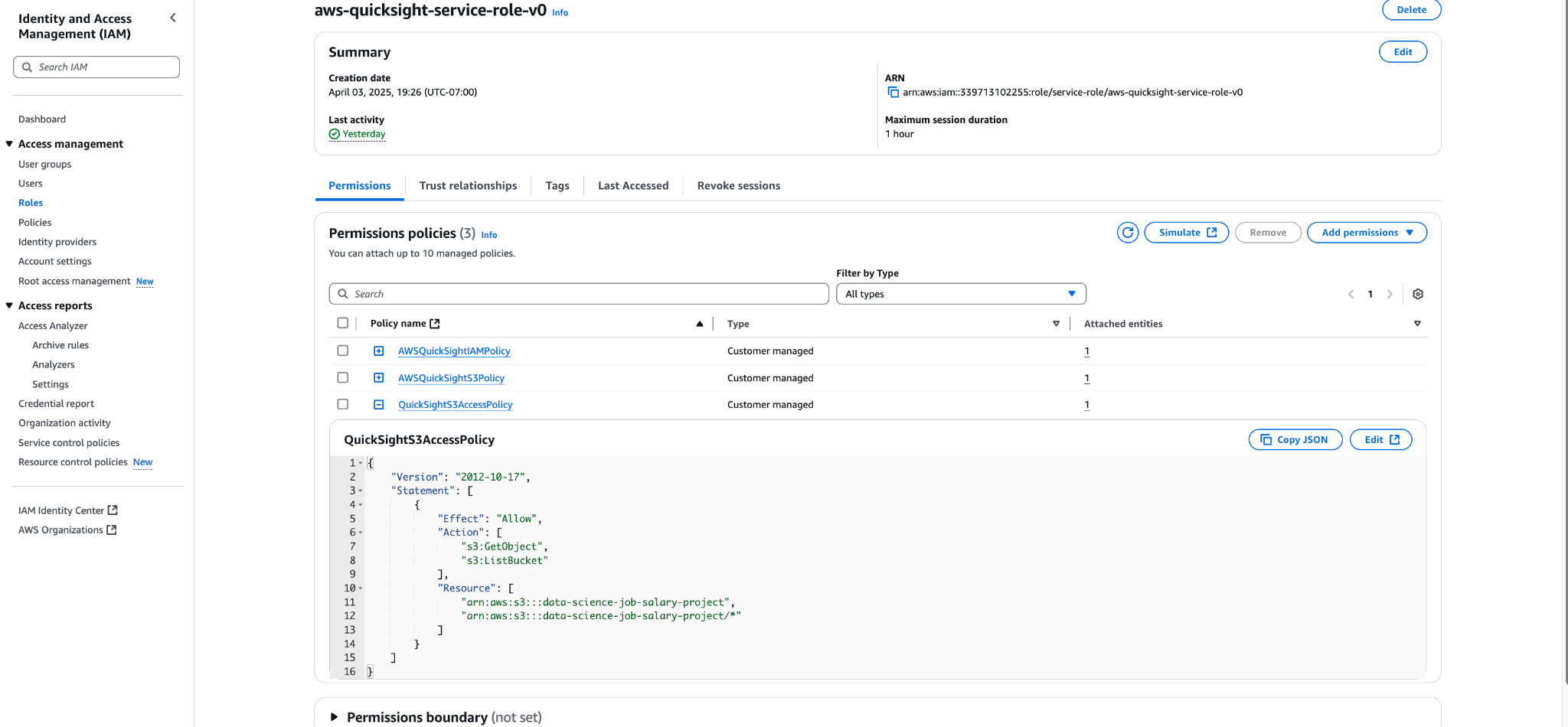
**Visualization using QUICKSIGHT**

**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.

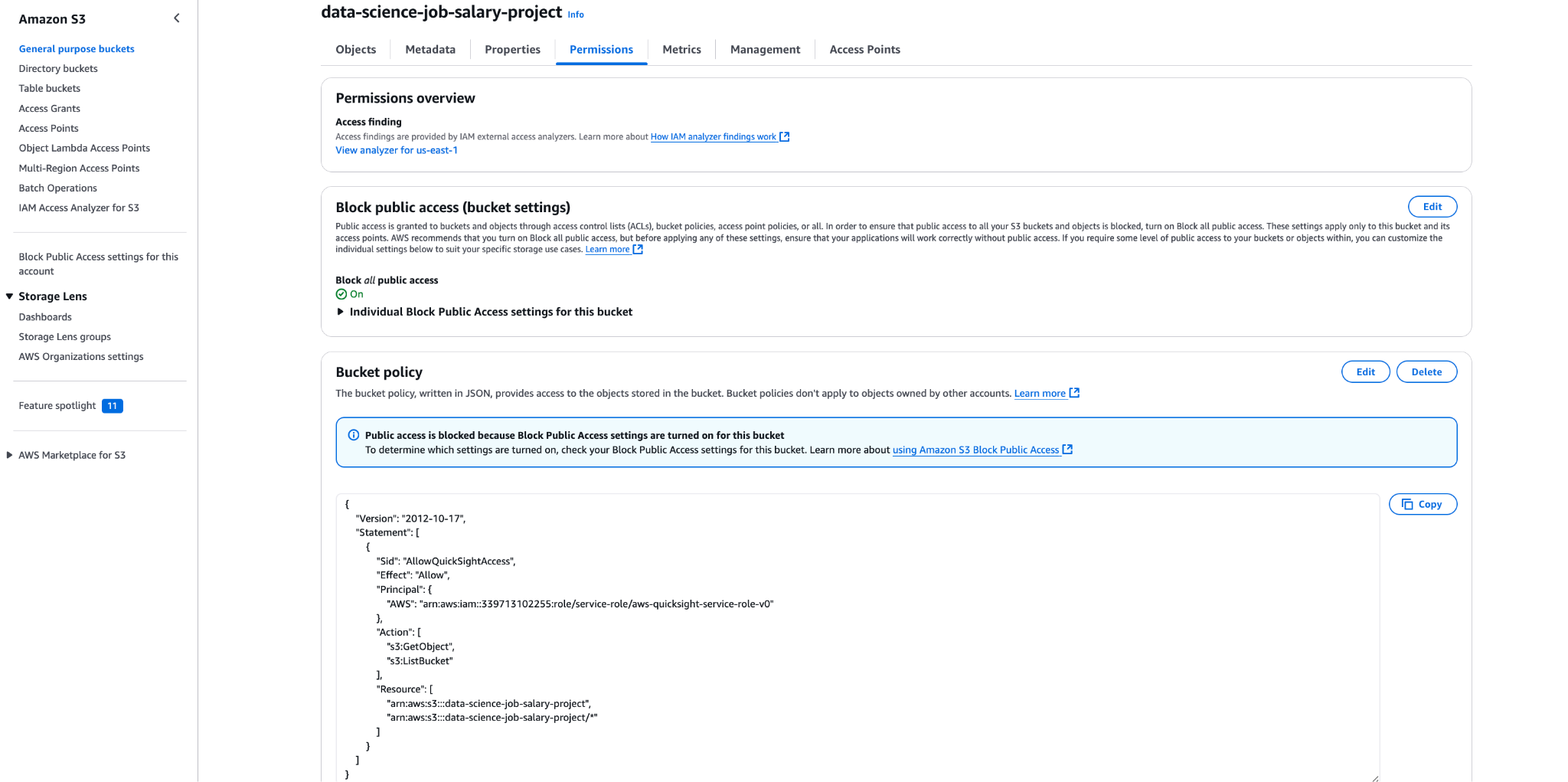
**Configuring Permissions for QuickSight Access**To allow Amazon QuickSight to access the data stored in Amazon S3, we configured the following two settings:

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.



**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.



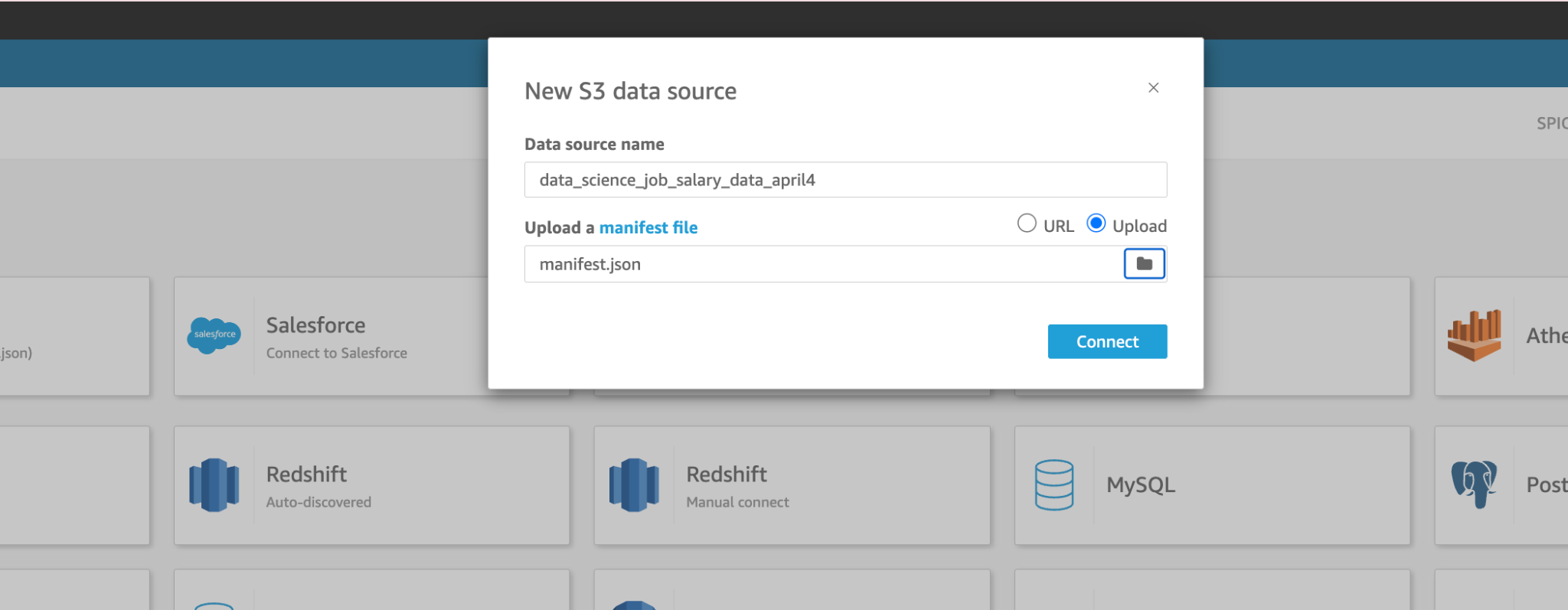
### **Data Source:**

* 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

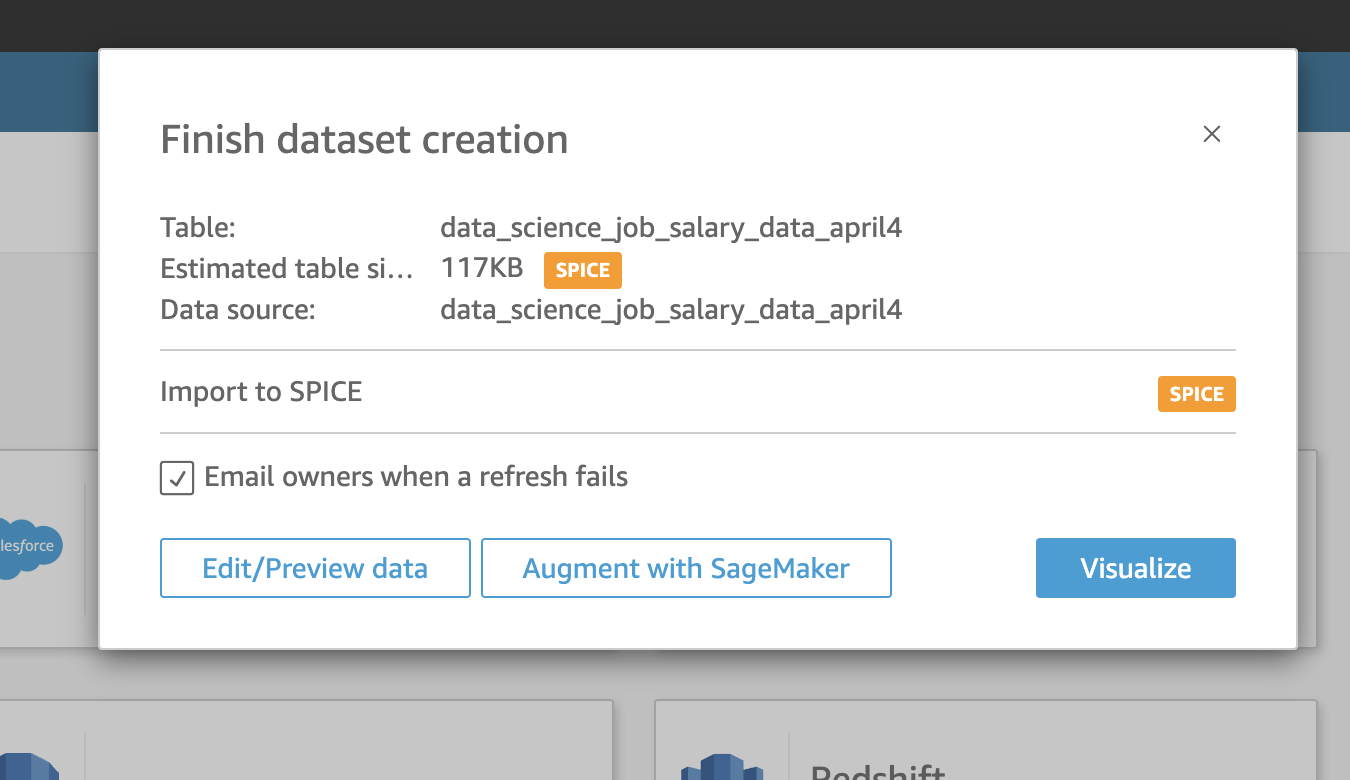
**Connecting QuickSight to the Dataset S3:**

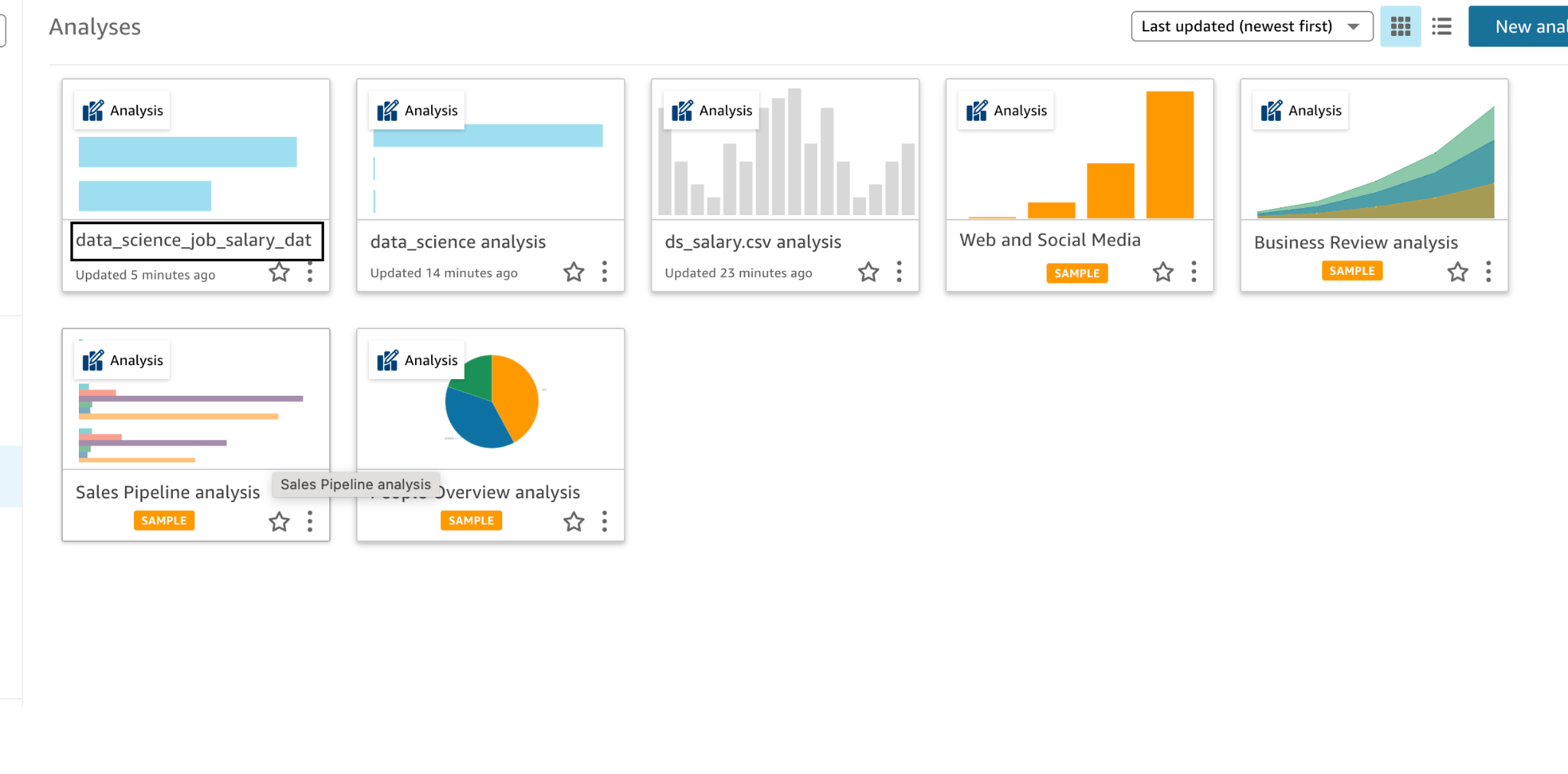
A new dataset named data\_science\_job\_salary\_data\_april4 was created in QuickSight by connecting to the S3 bucket (s3://data-science-job-salary-project/ds\_glue\_target/run-1743783775548-part-r-00000), where the transformed data is stored in CSV format.

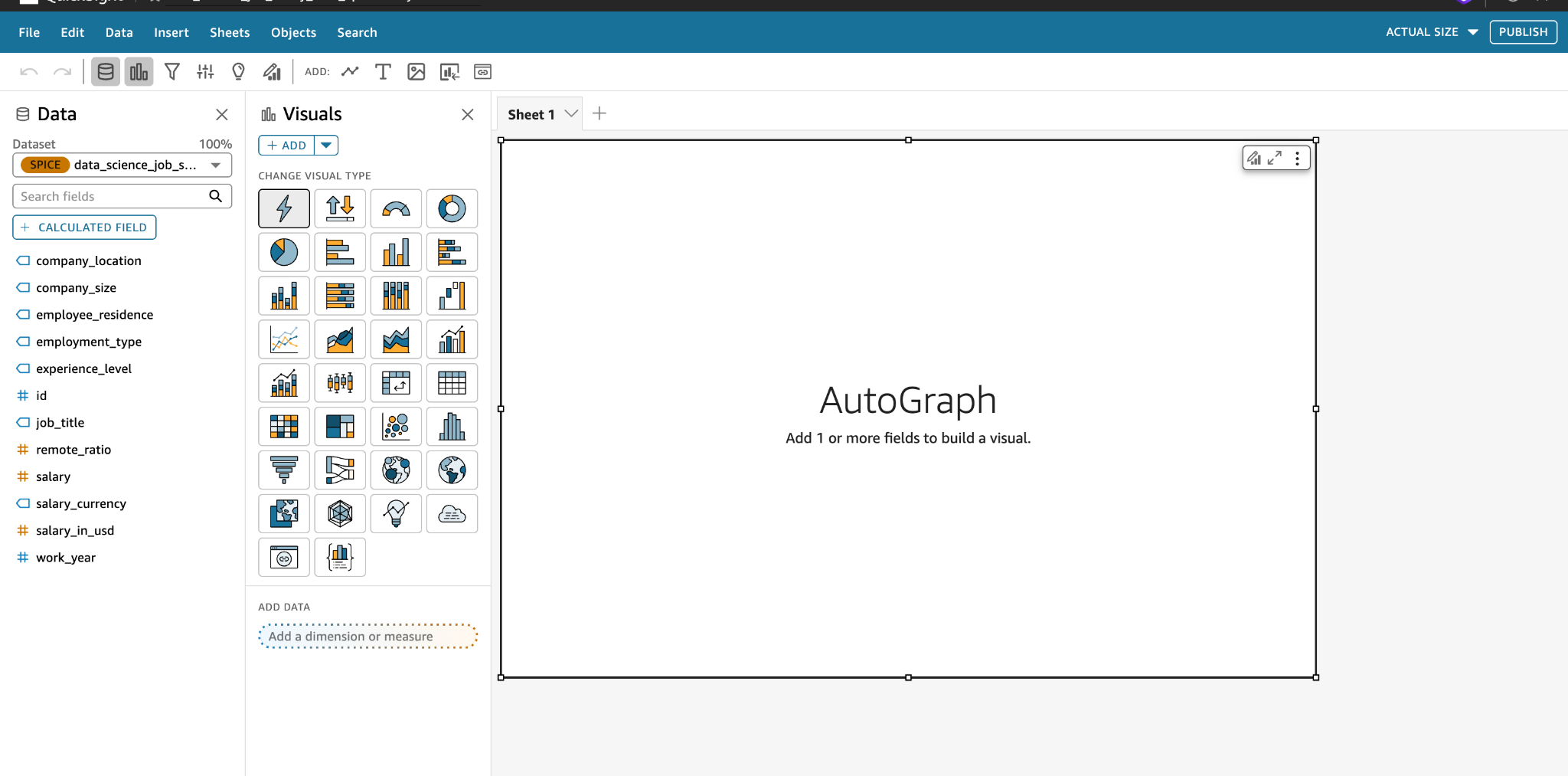


**Manifest.json file**: a manifest file is used to connect to data stored in Amazon S3. It helps QuickSight understand how to locate and interpret your files in S3.



****

****

****

**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**](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**](https://drive.google.com/file/d/1HL0ycnmlCyLB8ZV43WVqRCDPfE8jZGu-/view?usp=drive_link)

**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.