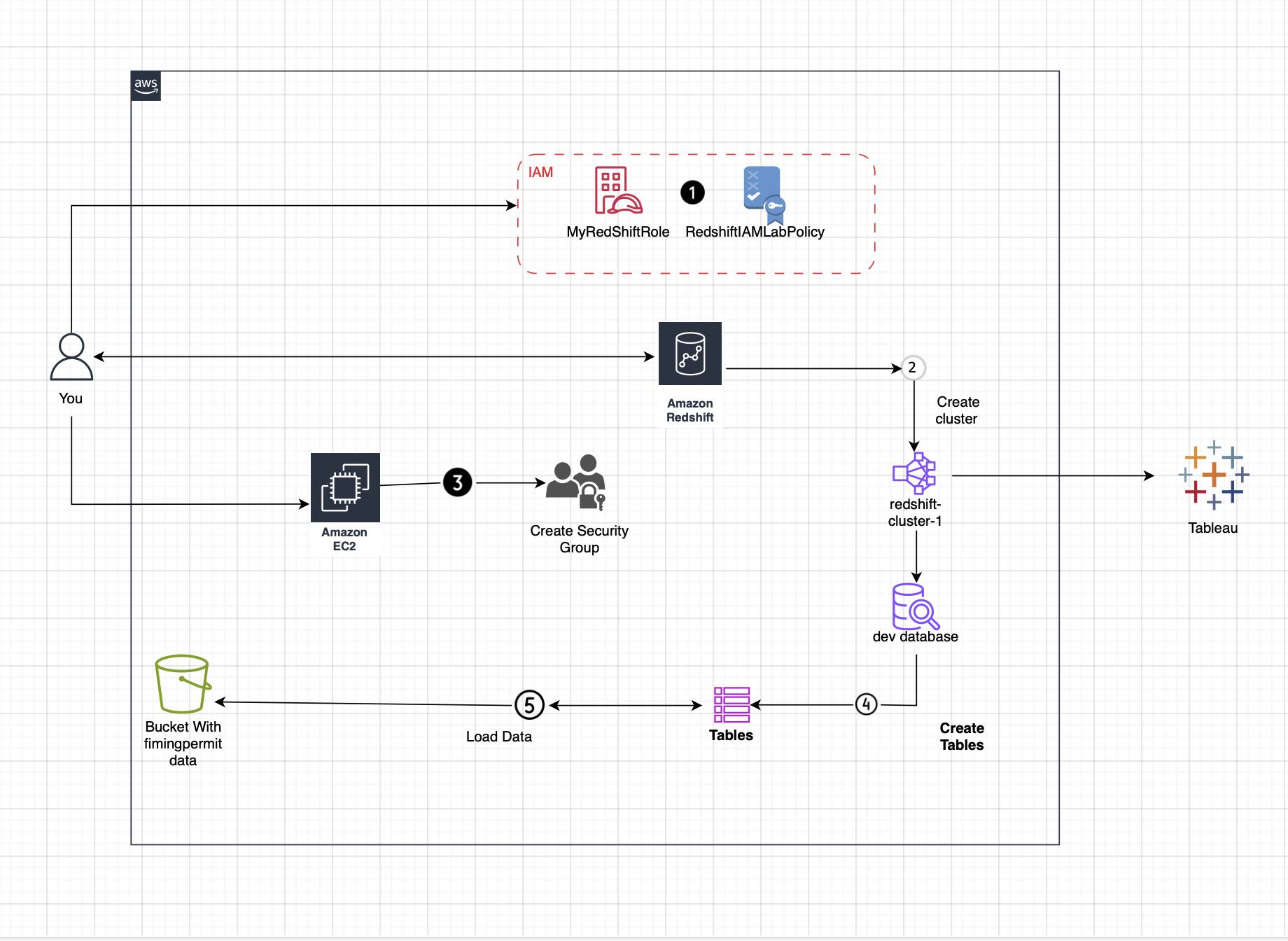
**NIT 2202 (BIG DATA)**

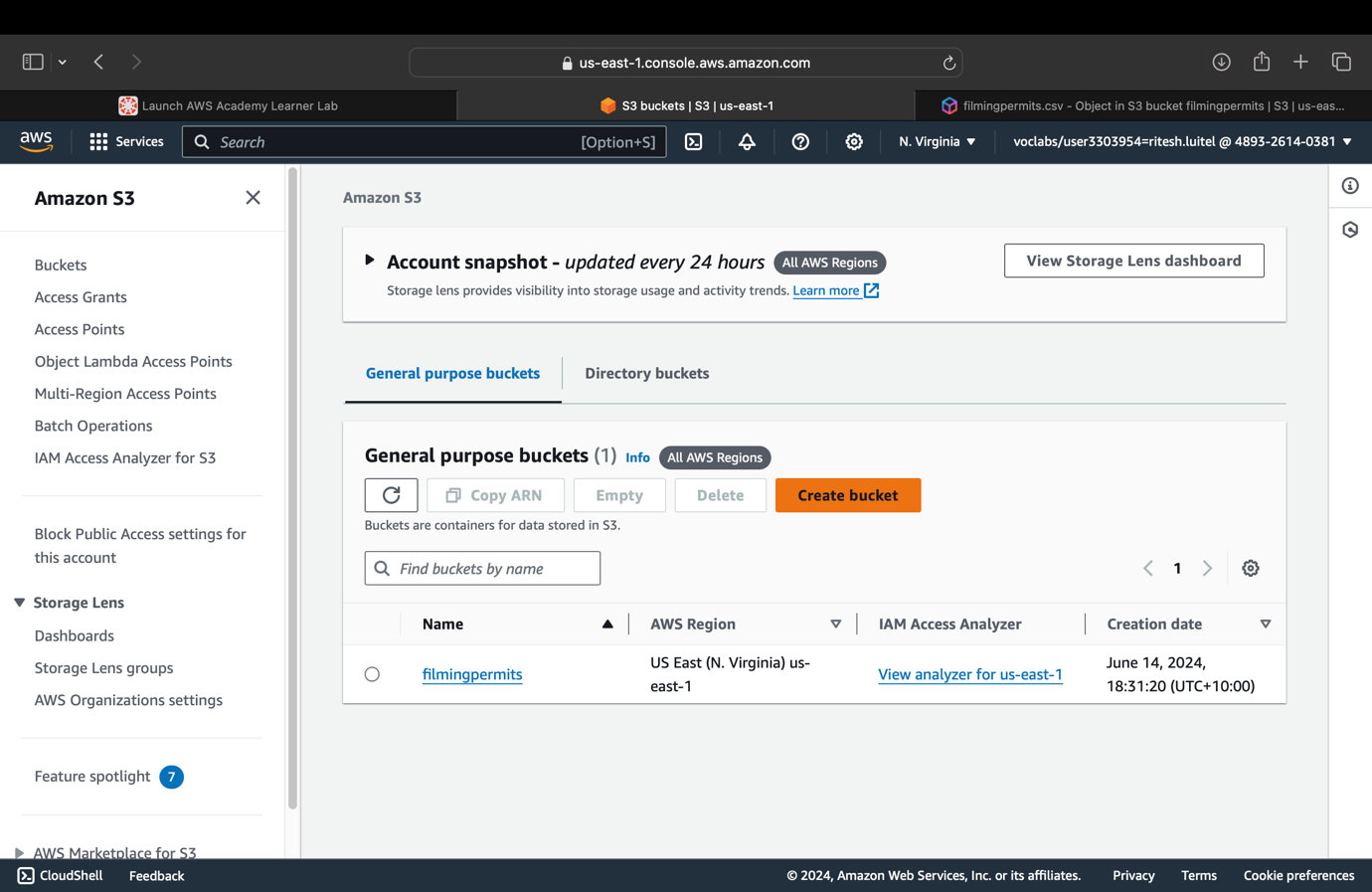
**Rites Luitel (s4682841)**

**Assessment PART I (Project 1)**

**AWS Architecture**

****

1. **Creating S3 bucket with name (filmingpermits):**

****

1. **Uploaded the project data (CSV file) in S3 bucket:**

**A screenshot of a computer

Description automatically generated**

1. **Creating Redshift cluster (redshift-cluster-1)** (with nodes = 2, node type = dc2.large, password = Passw0rd1, and “myredshiftrole” as associated IAM roles)

**A screenshot of a computer

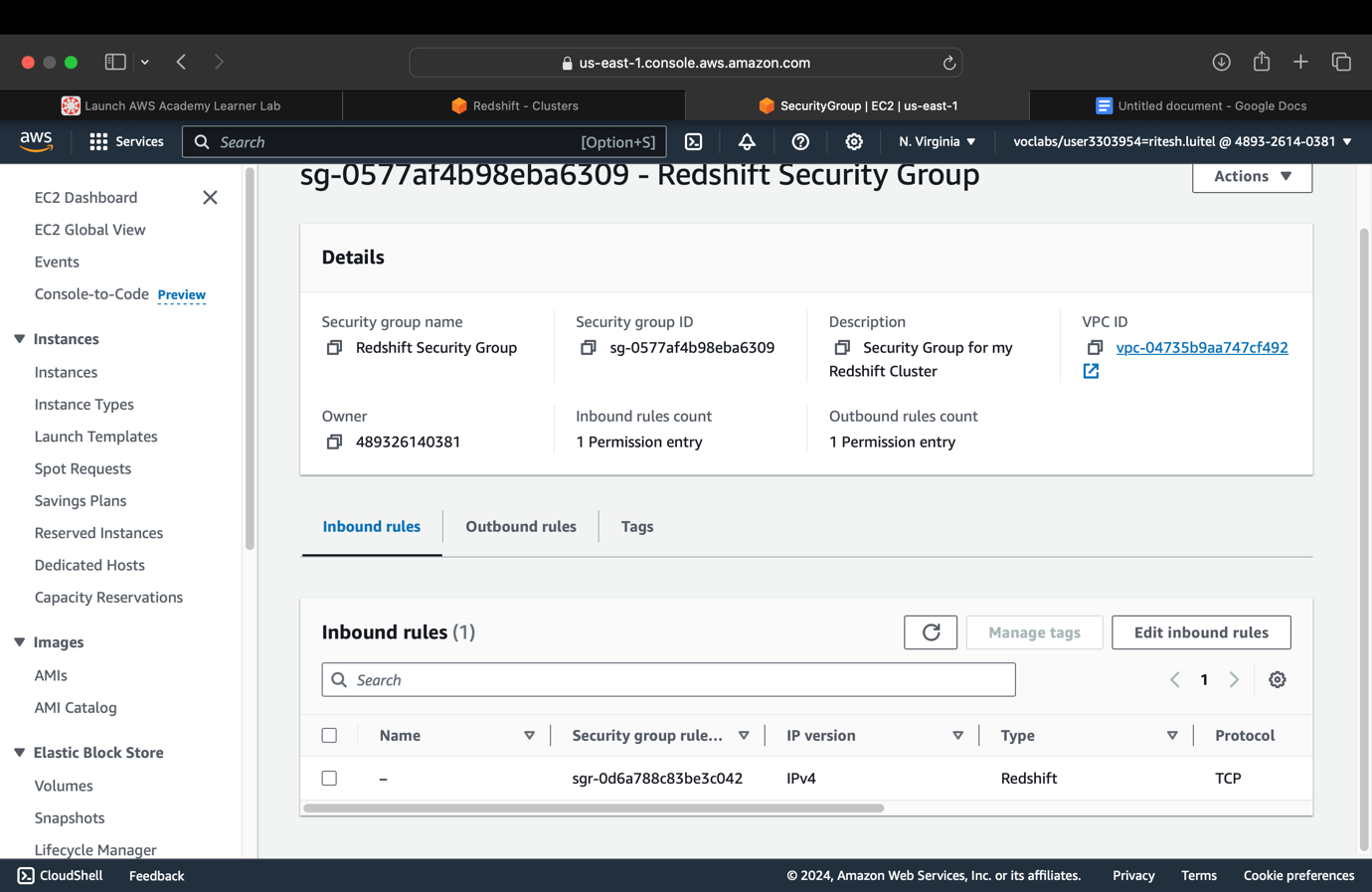
Description automatically generated**

**A screenshot of a computer

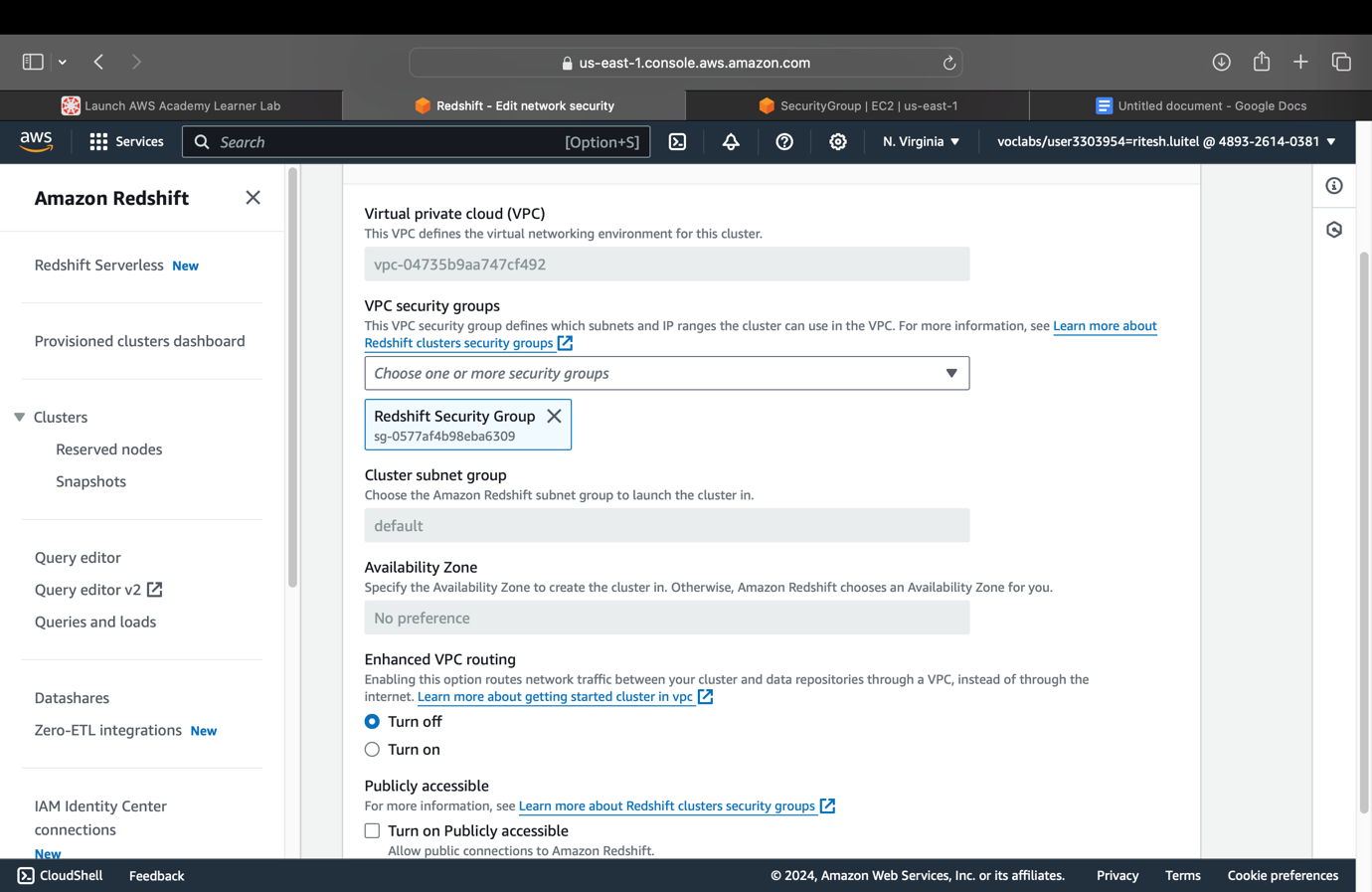
Description automatically generated**

Fig: redshift-cluster-1 created & available

1. **Creating security group with the settings** (Type - Redshift, and Source - IPv4):

****

1. **Editing security settings for redshift cluster** (Choosing Redshift Security Group and clearing the default):

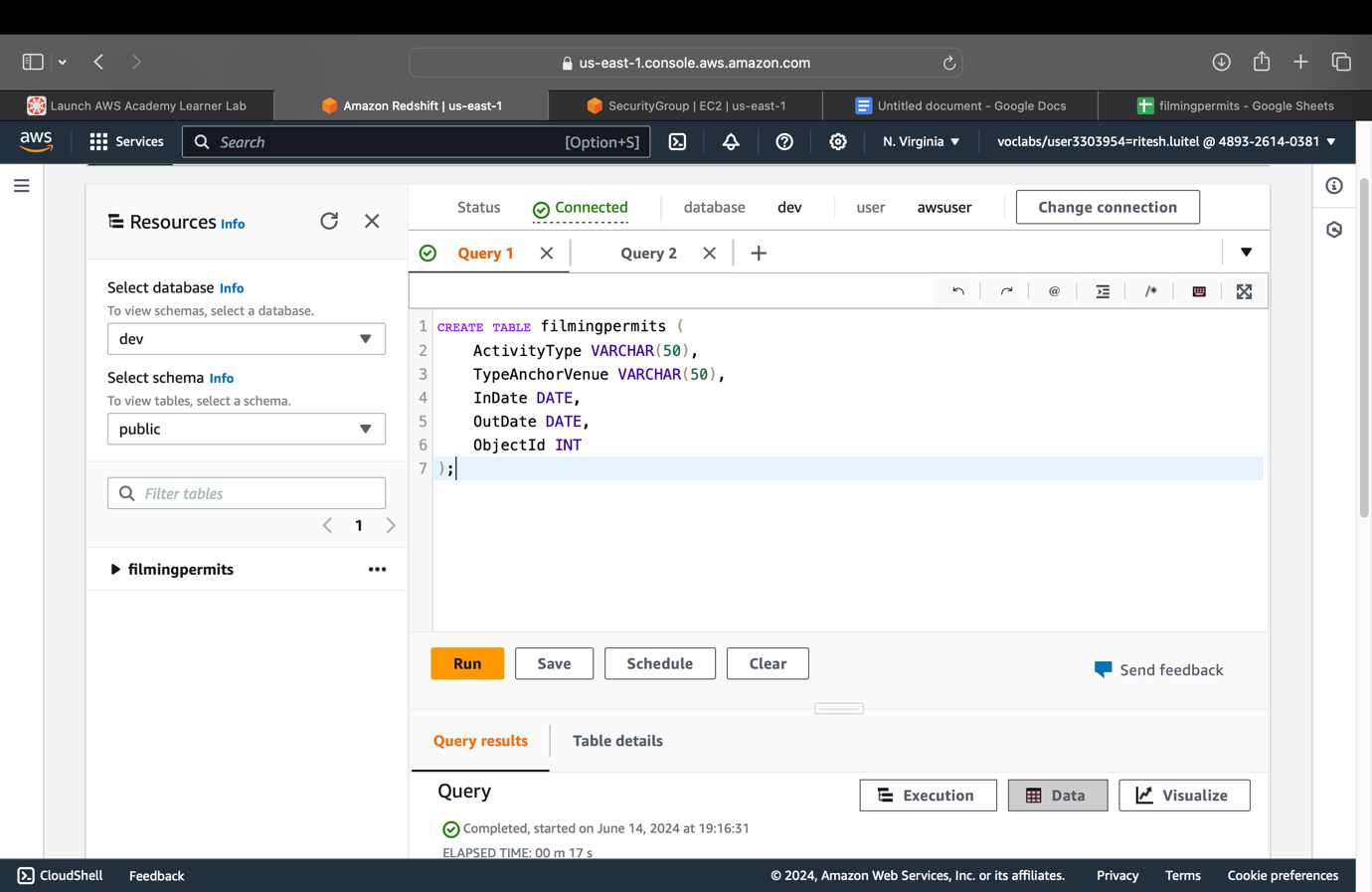
****

1. **Connecting with dev database** (Database Name - ‘dev’ & Database User - ‘awsuser’)

**A screenshot of a computer

Description automatically generated**

1. **Creating a table with SQL that matches exactly with project file in Redshift Cluster:**

****

1. **Loading the project data from S3 in redshift cluster:**

(S3 URL copied from bucket ‘filmingpermits’ & ARN copied from myRedshiftRole in EC2)

**A screenshot of a computer

Description automatically generated**

**9. Previewing the data of that table in redshift:**

**A screenshot of a computer

Description automatically generated**

Fig: Results after ‘Preview Data’

**10. Running Four queries:**

* **Two simple queries:**

Query 1: Retrieve all rows for a specific Activity Type

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

Fig: Results for Query 1

Query 2: **Retrieve rows where the event took place at a specific venue**

**A screenshot of a computer

Description automatically generated**

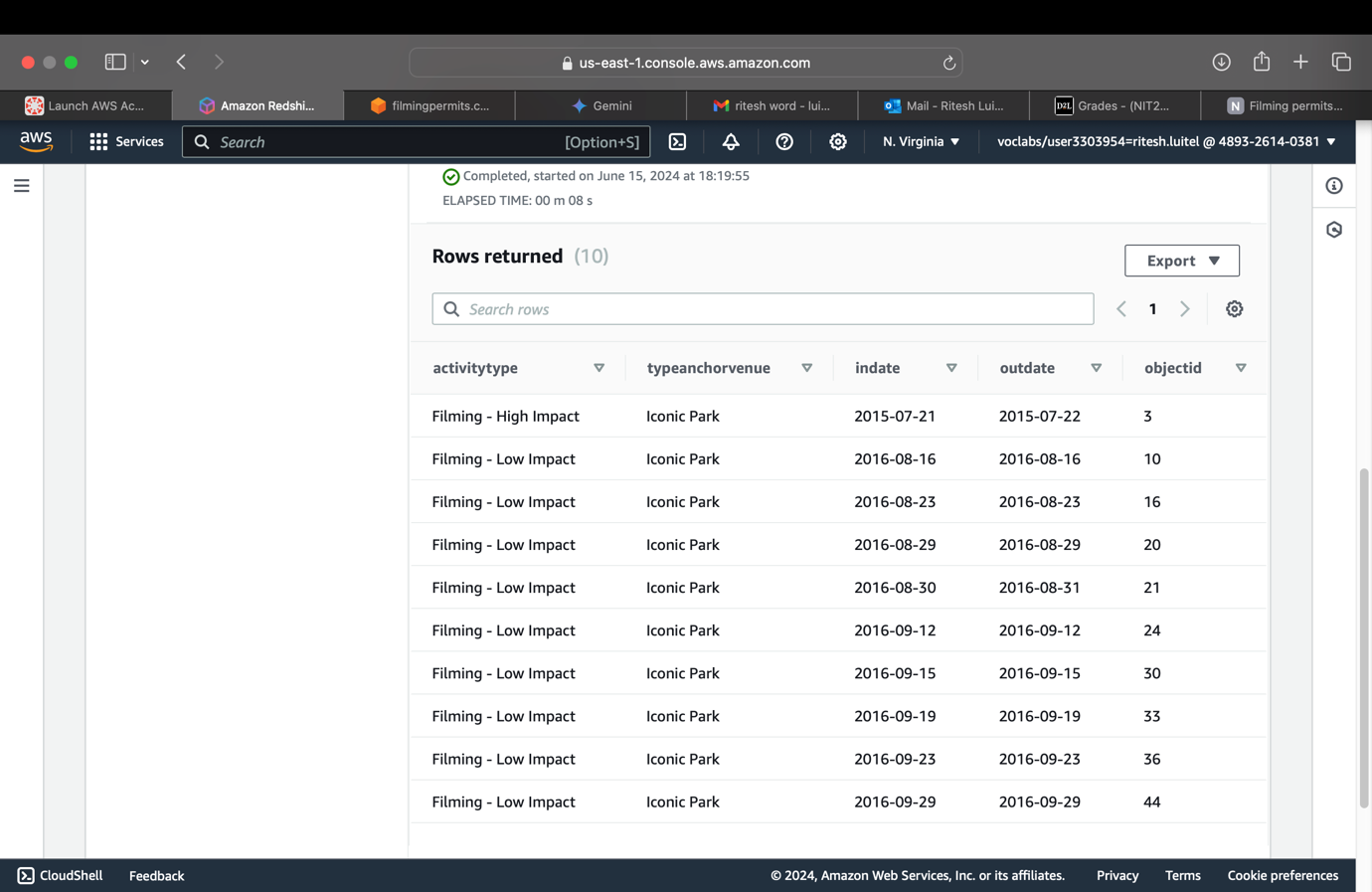
****

Fig: Results for Query 2

**Two Complex queries:**

Query 3: **Count the number of permits for each ActivityType**

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

Fig: Results for Query 3

Query 4: **Find the maximum and average duration of filming permits grouped by ActivityType**

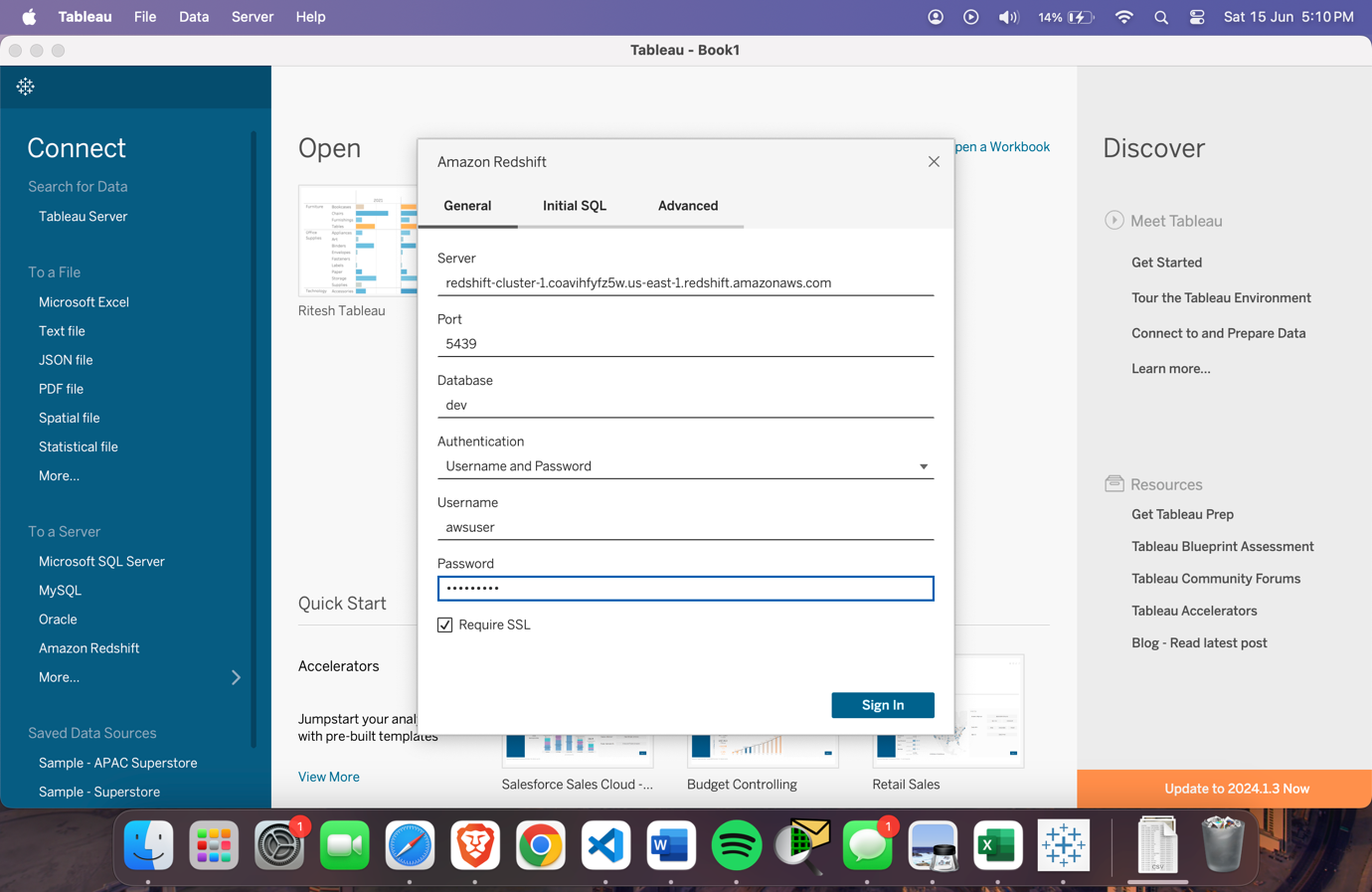
A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

**11. Connecting Tableau with Amazon Redshift:**

****

**12. Tableau – Live Data Connection and Live Data Synchronization**

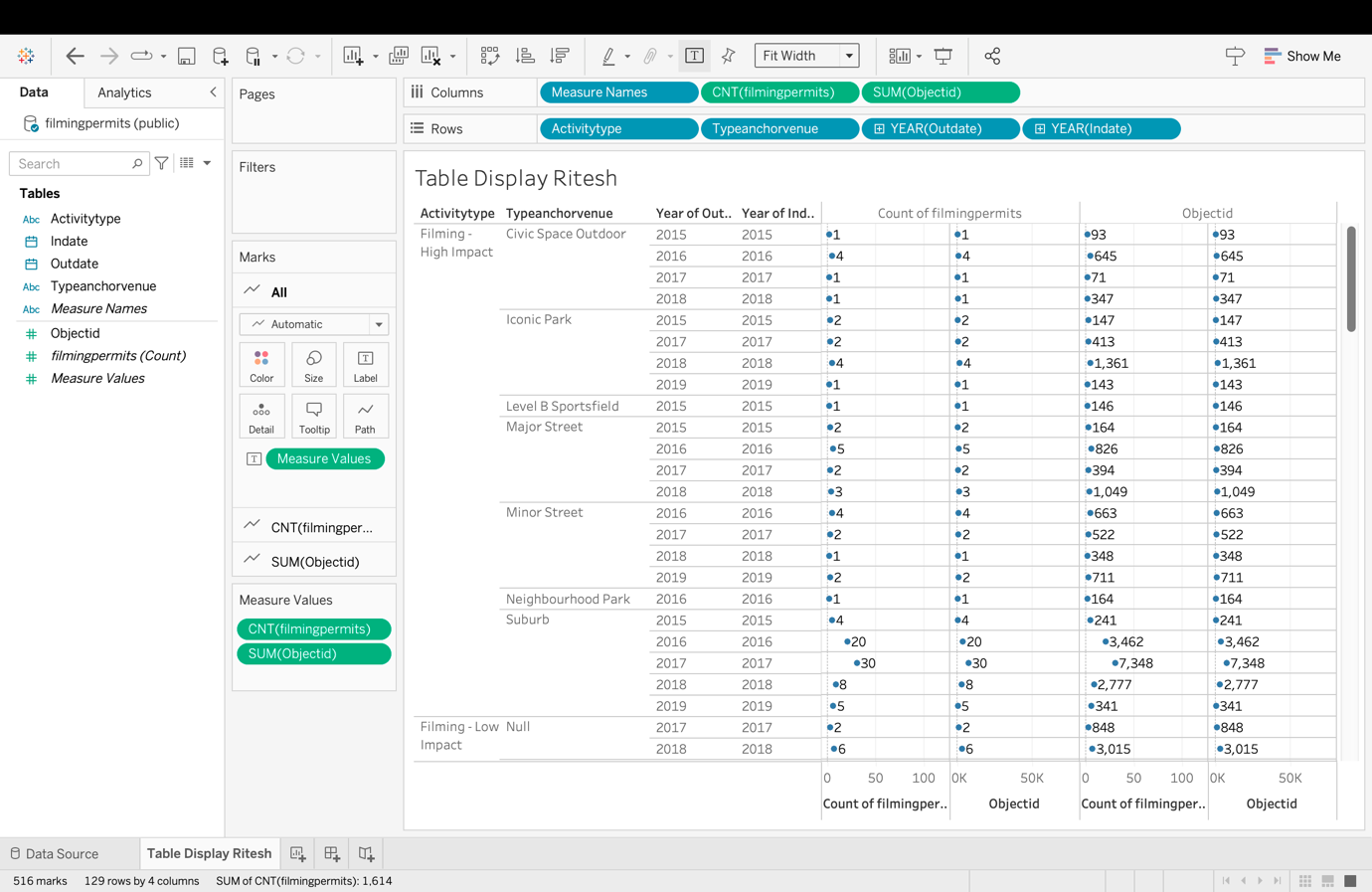
**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

**14. Tableau - Table Display**

****

**15. Creating three charts/graphs PLUS a Dashboard in Tableau.**

**Line Chart**

**A screenshot of a graph

Description automatically generated**

**Bar Chart**

**A screenshot of a computer

Description automatically generated**

**Pie Chart**

**A screenshot of a computer

Description automatically generated**

* **Data Visualization 2 – Dashboard**

**A screenshot of a computer

Description automatically generated**

**16. AWS Details: Project Hours/Cost**

**Hours Spent :6hrs**

|  |  |  |  |
| --- | --- | --- | --- |
| **Task Description** | **Date** | **Hours Spent** | **Notes** |
| Creating S3 bucket & uploading the file | 2024-06-14 | 15 minutes | |  | | --- | | Created S3 bucket and uploaded project CSV file. |  |  | | --- | |  | |
| Setting up Amazon Redshift and configuring security group | 2024-06-14 | 15  minutes | Set up Amazon Redshift and configured security group for secure access. |
| Connecting Redshift cluster to the database and creating a table using SQL | 2024-06-14 | 20 minutes | Connected Redshift cluster to database and created table using SQL. |
| Loading data into Amazon Redshift | 2024-06-14 | 40 minutes | Uploaded CSV data files to Redshift for analysis. |
| Executing SQL queries | 2024-06-15 | 0.5 hours | Ran SQL queries to validate data and do preliminary analysis |
| Connecting Tableau to Amazon Redshift and installing ODBC driver) | 2024-06-15 | 0.5 hours | Set up live data connection between Tableau and Redshift by installing ODBC driver. |
| Creating visualizations in Tableau | 2024-06-15 | 1 hour | Made charts and dashboards in Tableau to show data insights. |
| Architecting AWS infrastructure | 2024-06-16 | 1.5 hour | Planned AWS architecture for the project, detailing roles and interactions of components. |
| Reviewing project documentation | 2024-06-16 | 1 hour | Reviewed and updated project documentation for completeness and accuracy. |

**Cost incurred:**

**A screenshot of a computer

Description automatically generated**

**Queries Performed:**

**Simple and Complex Queries**

**1.Retrieve all rows for a specific ActivityType**

SELECT \*

FROM filmingpermits

WHERE ActivityType = 'Filming - Low Impact'

LIMIT 10;

**2. Retrieve rows where the event took place at a specific venue**

SELECT \*

FROM filmingpermits

WHERE TypeAnchorVenue = 'Iconic Park'

LIMIT 10;

**3. Count the number of permits for each ActivityType**

SELECT ActivityType, COUNT(\*) AS PermitCount

FROM filmingpermits

GROUP BY ActivityType

ORDER BY PermitCount DESC;

**4. Find the maximum and average duration of filming permits grouped by ActivityType**

SELECT ActivityType,

MAX(DATEDIFF(day, InDate, OutDate) + 1) AS MaxDuration,

AVG(DATEDIFF(day, InDate, OutDate) + 1) AS AvgDuration

FROM filmingpermits

GROUP BY ActivityType

ORDER BY MaxDuration DESC;