AWS Data Analytic Platform for The City of Vancouver

Phase 1

# Abstract

This report outlines the design, implementation, and evaluation of the AWS Data Analytic Platform (DAP) for the City of Vancouver, focused on property tax data. The platform leverages Amazon Web Services (AWS) to streamline data management and analytical processing for the city’s Government and Finance department, aiming to improve decision-making and enhance operational efficiencies. The DAP includes key processes such as data ingestion, data cleaning, structuring, analysis, and visualization. This report focuses on the property tax report from 2023 and 2024, providing insights into property value trends, tax rate changes, and zoning classifications. The platform’s scalability, security, and integration with Amazon S3, AWS Glue, and Amazon Redshift ensure seamless data handling and storage. The project delivers an enhanced system for monitoring tax collection and property value variations, enabling policymakers to make more informed decisions regarding taxation policies and urban planning.

Table of Contents

[Abstract 2](#_Toc175692146)

[Introduction (Teamwork) 7](#_Toc175692147)

[DAP Design and Implementation (Individual work) 8](#_Toc175692148)

[Step 1: Data Analytical Question Formulation 11](#_Toc175692149)

[Step 2: Data Discovery 14](#_Toc175692150)

[Step 3: Data Storage Design 18](#_Toc175692155)

[Step 4: Dataset Preparation 20](#_Toc175692156)

[Step 5: Data Ingestion 24](#_Toc175692157)

[Step 6: Data Storage 28](#_Toc175692158)

[Step 7: Data Pipeline Design 33](#_Toc175692159)

[Step 8: Data Cleaning 40](#_Toc175692160)

[Step 9: Data Structuring 50](#_Toc175692163)

[Step 10: Data Pipeline Implementation 53](#_Toc175692172)

[Step 11: Data Analysis 61](#_Toc175692173)

[Step 12: Data Visualization 67](#_Toc175692177)

[Step 13: Data Publishing 70](#_Toc175692178)

[DAP Estimated Cost (Teamwork) 74](#_Toc175692179)

# Introduction (Teamwork)

The City of Vancouver, as a growing metropolis, requires robust data management solutions to efficiently handle and analyze critical financial and operational data. The AWS Data Analytic Platform (DAP) was implemented to address the city’s need for improved property tax reporting and analysis. This system, built on AWS, is designed to support the Government and Finance department by automating the ingestion, storage, processing, and analysis of large datasets related to property tax assessments. The project is divided into multiple phases, each addressing a crucial component of the data management lifecycle, from data discovery to publishing.

In this report, we specifically examine Phase 1 of the DAP, focusing on the property tax data for 2023 and 2024. The platform is structured to provide real-time insights into property values, tax rates, and zoning classifications, facilitating data-driven decision-making for city officials. Each step in the DAP's development was carefully designed to ensure scalability, accuracy, and security, enabling the platform to adapt to future data expansion and evolving analytical needs. Furthermore, the project team employed AWS Glue for Extract, Transform, and Load (ETL) operations, Amazon Redshift for structured data storage, and Amazon Athena for query execution, ensuring that the city’s financial data is efficiently processed and readily available for analysis.

# DAP Design and Implementation (Individual work)

**Shruti: Property Tax Report**

To design and implement the Data Analysis Pipeline (DAP) for the property tax report, Shruti adopted steps to protect the integrity of the data to be analyzed and in this case, the accuracy of the report. She started by defining data elements that would be of importance when analyzing property tax trends for the properties including PIDs, Zoning classifications, PID values- land and improvements, and PID tax levies. Shruti used AWS glue for ETL where she used raw datasets from the 2023 and 2024 reports and transformed them into a useable format. Various data cleansing techniques were performed such as; correcting data types, dealing with the missing values and eliminating duplicate records for better data quality. Further, to aid the analytical insight she has initiated derived attributes such as ‘PropertyValueChange’ and ‘TaxRateChange’. The curated data was then ingested into Amazon Redshift where one can perform raw complex queries and forms of analysis. Shruti’s DAP design was centred on horizontal growth so that she could make constant improvements and run it in real time to make sound decisions concerning property tax policies in Vancouver.

## Step 1: Data Analytical Question Formulation

**Data Discovery Section**

**1. Shruti: Property Tax Report**

**Descriptive Analysis:** The descriptive analysis of the Property Tax Report datasets from 2023 and 2024 aims to understand the trends in property tax collection over these two years. By examining the data on property values, tax assessments, and levies, we can identify patterns in property tax collection, assess the distribution of property types, and analyze changes in land and improvement values. This analysis helps determine which areas have the highest tax assessments, identify trends in property value appreciation or depreciation, and spot common property types in different zones.

* **Key Questions:**
  + What are the trends in property tax collection over the past two years?
  + How do property values and tax assessments vary by neighbourhood and property type?

**Diagnostic Analysis:** The diagnostic analysis investigates the underlying reasons for fluctuations in property tax collection efficiency. For instance, we will explore why there was a significant decrease in tax collection efficiency in certain months of 2023 and 2024. This involves analyzing external factors like economic conditions or policy changes, identifying any errors in data entry or assessment processes, and examining seasonal variations or trends that might have impacted tax collection.

* **Key Questions:**
  + Why was there a significant decrease in property tax collection efficiency in Q3 of the last fiscal year?
  + What external factors or internal process inefficiencies might have contributed to the observed trends?
  + liance rates for parking ticket payments?

**Summary of Data Discovery**

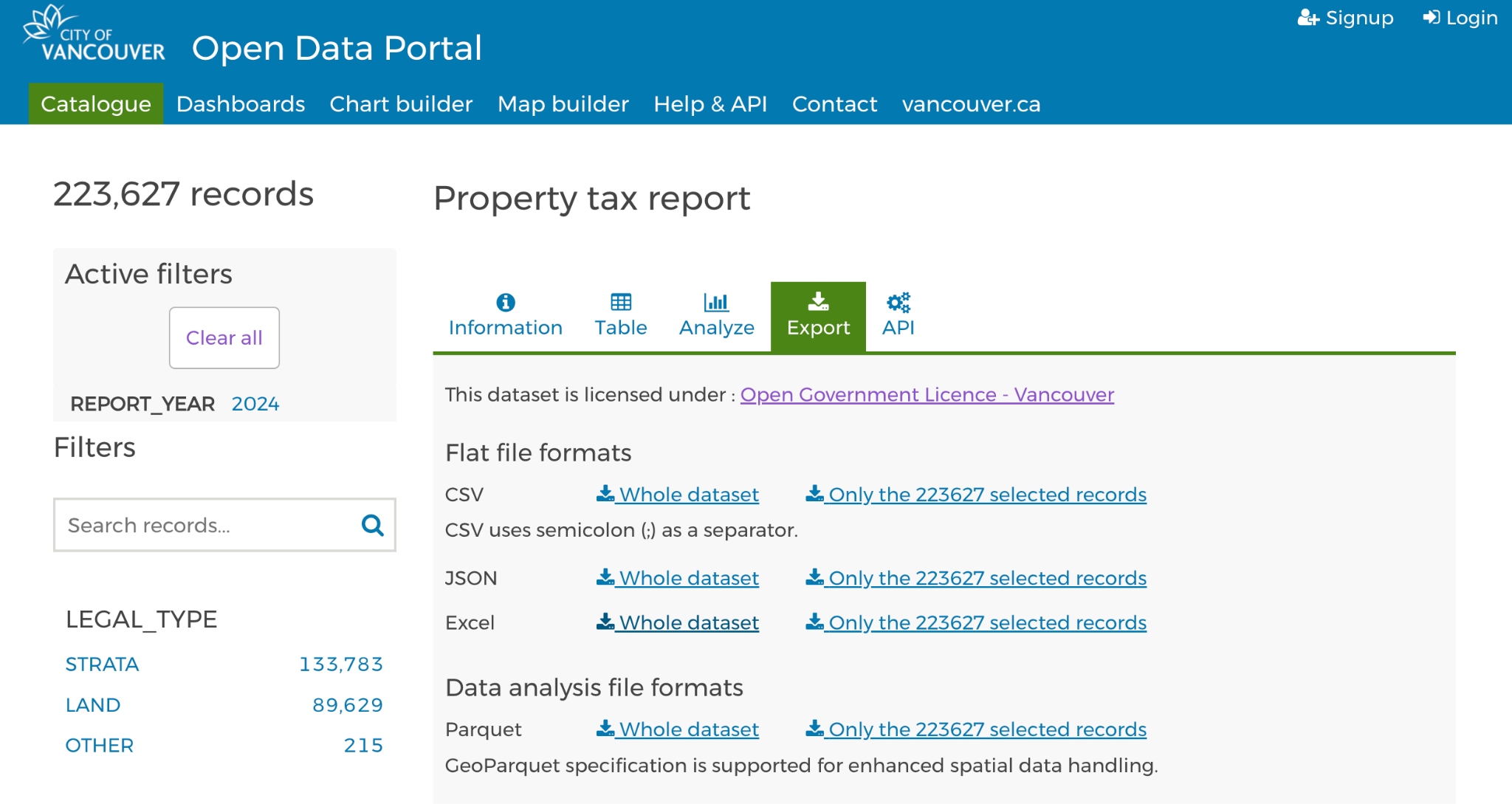
The data discovery phase involves using descriptive analysis to establish a foundational understanding of each dataset's key trends and patterns. Following this, diagnostic analysis provides deeper insights into the reasons behind these patterns, enabling the identification of specific factors contributing to inefficiencies or issues within each department. This dual approach ensures a comprehensive understanding of the data, facilitating data-driven decision-making to improve departmental procedures and performance across the City of Vancouver.

## Step 2: Data Discovery

### **1. Shruti: Property Tax Report**

**Datasets Used:**

* **Property Tax Report 2023:** This dataset contains 100 entries of property tax information for the year 2023. It includes details such as PID, legal type, folio, land coordinate, zoning district, zoning classification, lot, plan, block, district lot, current land and improvement values, tax assessment year, previous land and improvement values, year built, big improvement year, tax levy, neighbourhood code, and report year.
* **Property Tax Report 2024:** This dataset also contains 100 entries of property tax data for the year 2024, similar in structure to the 2023 dataset, capturing details relevant to property taxation for that year.



## Step 3: Data Storage Design

The data storage design depicted in the diagram outlines the cloud-based architecture used to store and manage the datasets related to parking tickets for the years 2023 and 2024. The design leverages cloud computing infrastructure to ensure secure, scalable, and accessible data management for the Government & Finance department.

#### **Key Components:**

1. **UCW Data Center**:
   * The UCW Data Center serves as the central repository for all data storage activities. It houses the cloud infrastructure necessary for storing, processing, and managing data securely.
2. **Vancouver Region**:
   * The data is stored within the Vancouver region, ensuring compliance with local data governance and regulatory requirements. This setup minimizes latency and enhances data access speeds for users in the local area.
3. **UCW General Server (Registrar’s Office)**:
   * Within the Vancouver region, the UCW General Server, managed by the Registrar’s Office, is designated for storing specific departmental datasets. This server acts as the primary storage location for the Government & Finance department's data.
4. **Dataset Storage**:
   * The datasets, such as "Parking Tickets 2023" and "Parking Tickets 2024," are stored in Excel file formats (as indicated by the Excel icons). These files are systematically organized within the general server to ensure efficient data retrieval and management.
   * Each dataset is labeled according to the year and content type, using a standardized naming convention (e.g., AF

/M/Q/Y), which likely represents different time frames or reporting periods (e.g., Weekly, Monthly, Quarterly, Yearly). This convention helps in categorizing and filtering data for analysis.

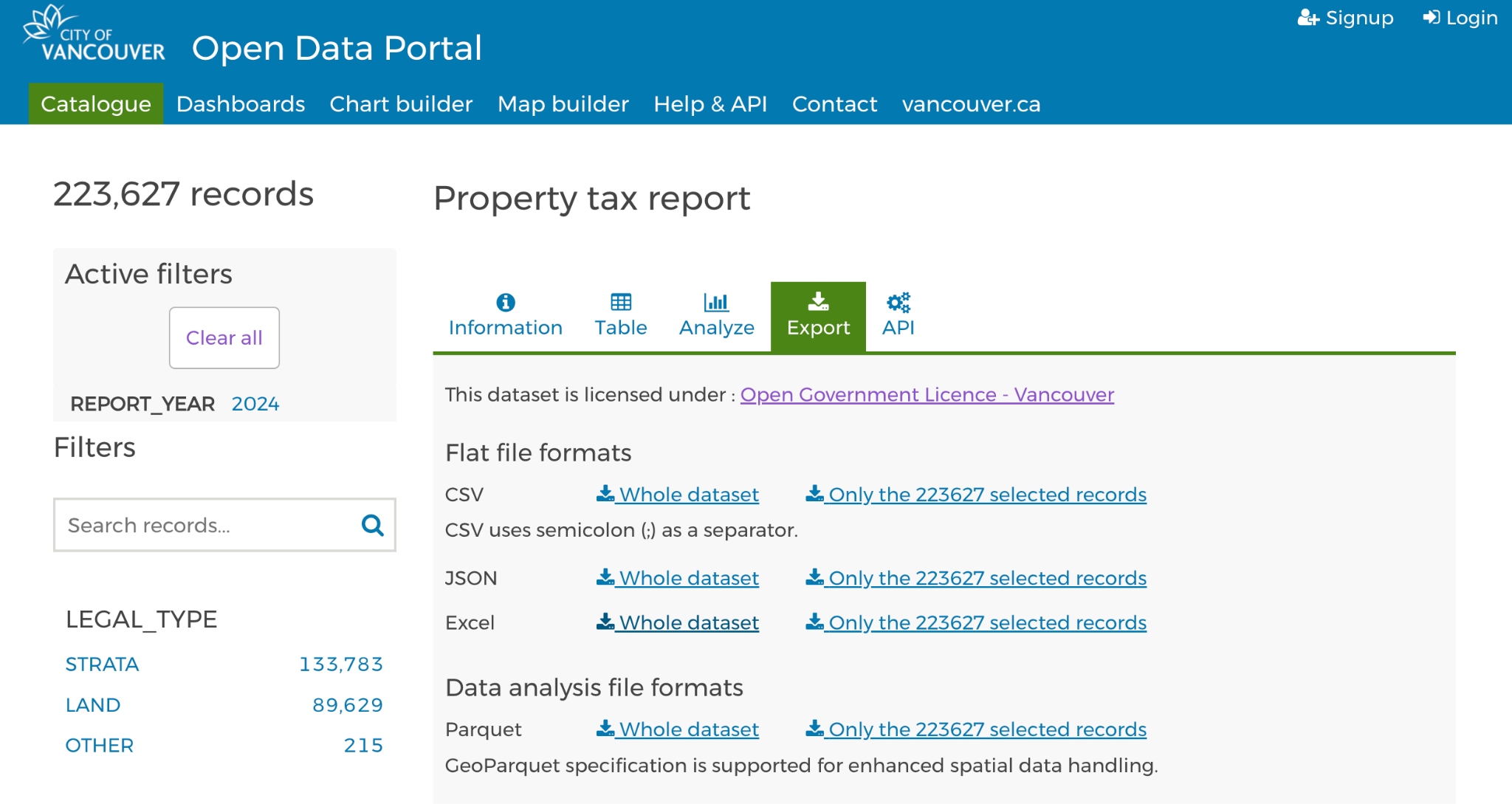
#### **Design Considerations:**

* **Scalability**: The cloud-based storage solution allows for easy scalability, accommodating future data growth without requiring significant infrastructure changes.
* **Security**: By utilizing a centralized data center and secure cloud services, sensitive data related to government and finance operations is protected against unauthorized access and potential breaches.
* **Compliance**: Storing data within the local region ensures compliance with data residency regulations and policies specific to governmental data management in Vancouver.
* **Accessibility**: The cloud infrastructure facilitates seamless access to data for authorized users, enabling efficient data analysis and decision-making processes.

Overall, this data storage design reflects a well-organized, secure, and scalable approach to managing critical datasets for the Government & Finance department, leveraging cloud technology to enhance operational efficiency and data governance.

## 

## Step 4: Dataset Preparation

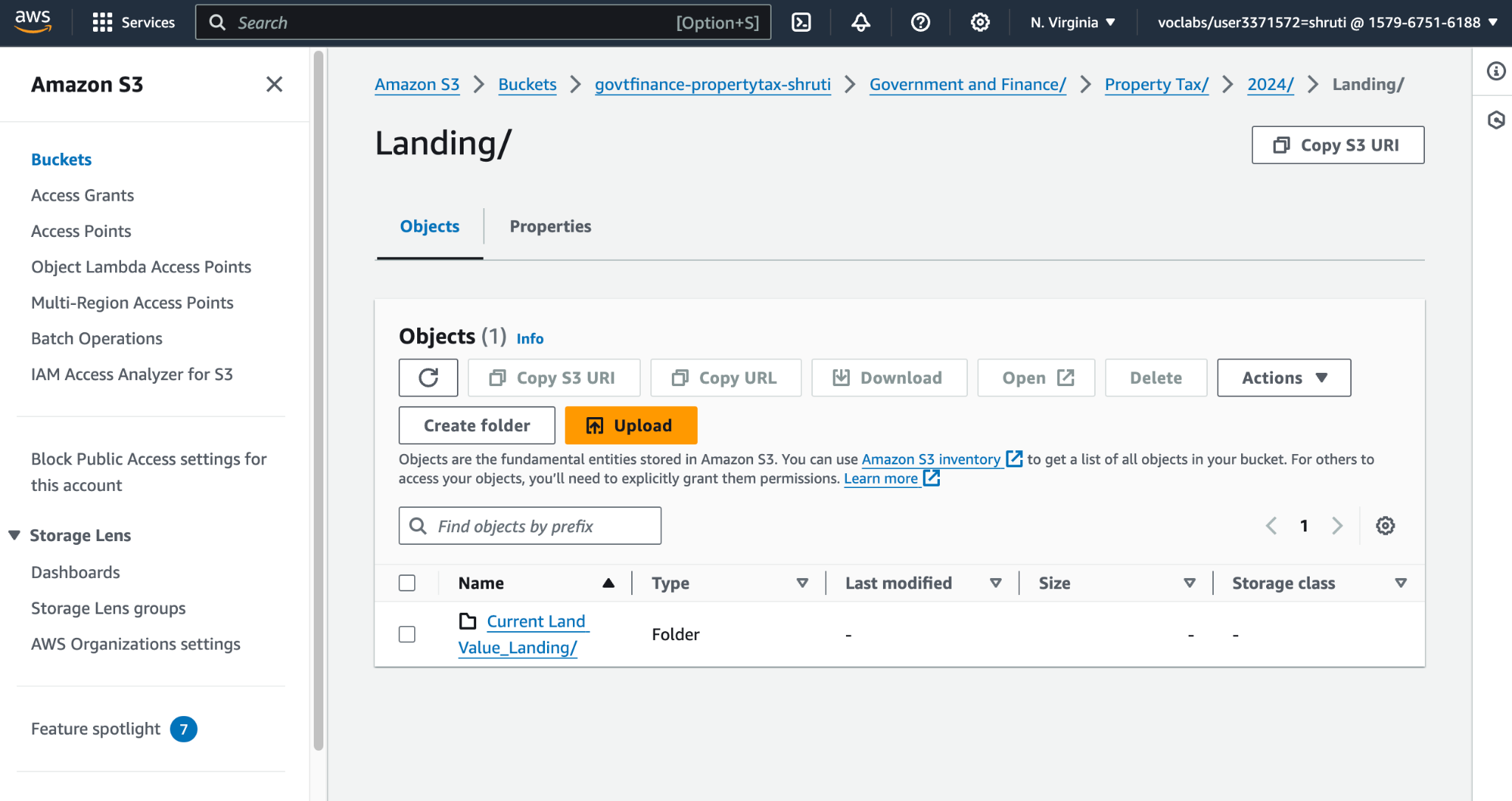
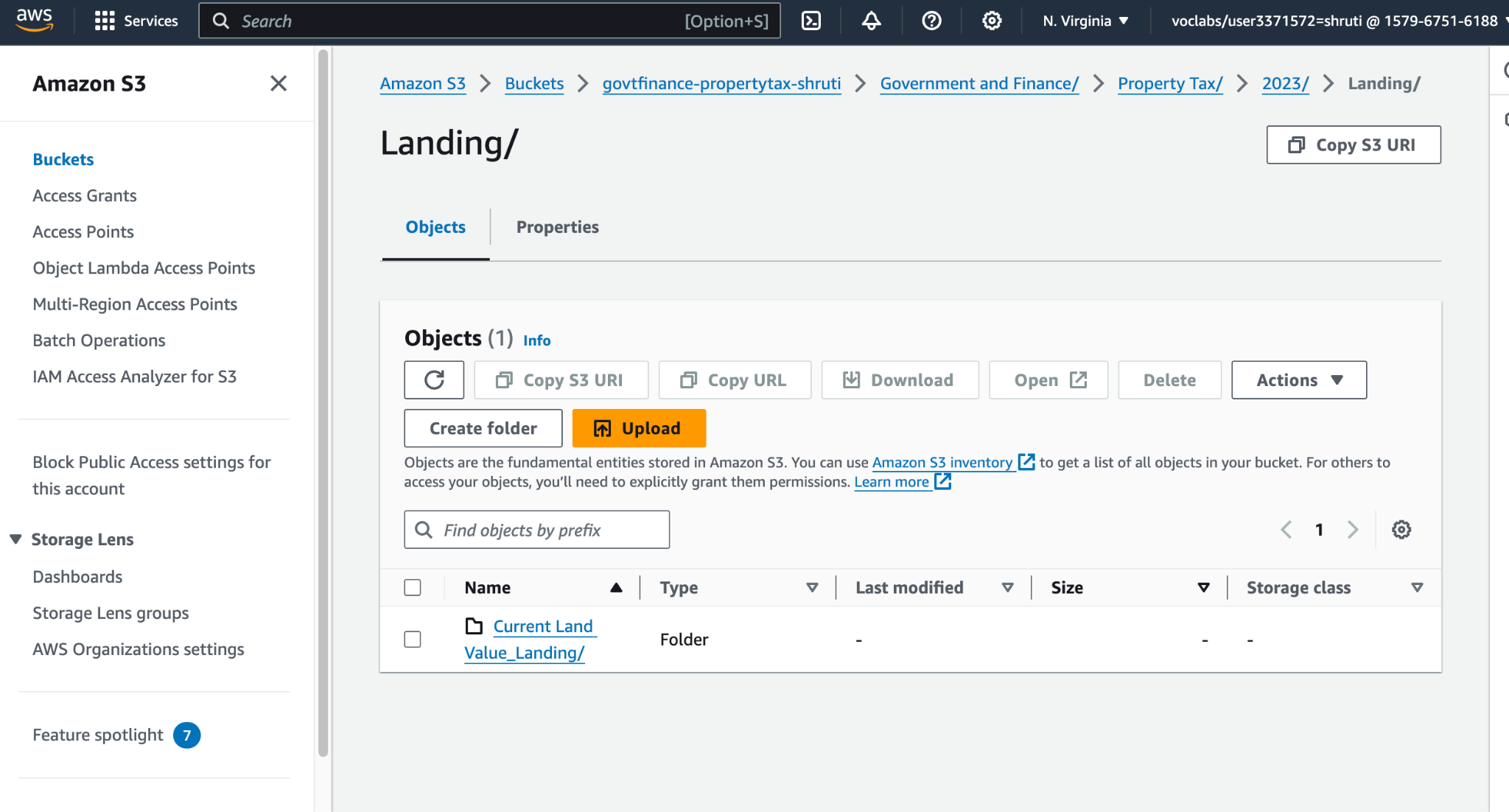
**Shruti**

**Shruti: Property Tax Report**

For the Property Tax Report, the selection is in the years 2023 and 2024. There are 100 records for each year depending on property, assessment of property, land and improvement values, zoning and tax levies. The data was cleaned by selecting the relevant fields which could be used in analyzing the trends in the collection of property tax and identifying factors that affect tax assessments across various neighbourhoods in Vancouver.

## Step 5: Data Ingestion

**Shruti**



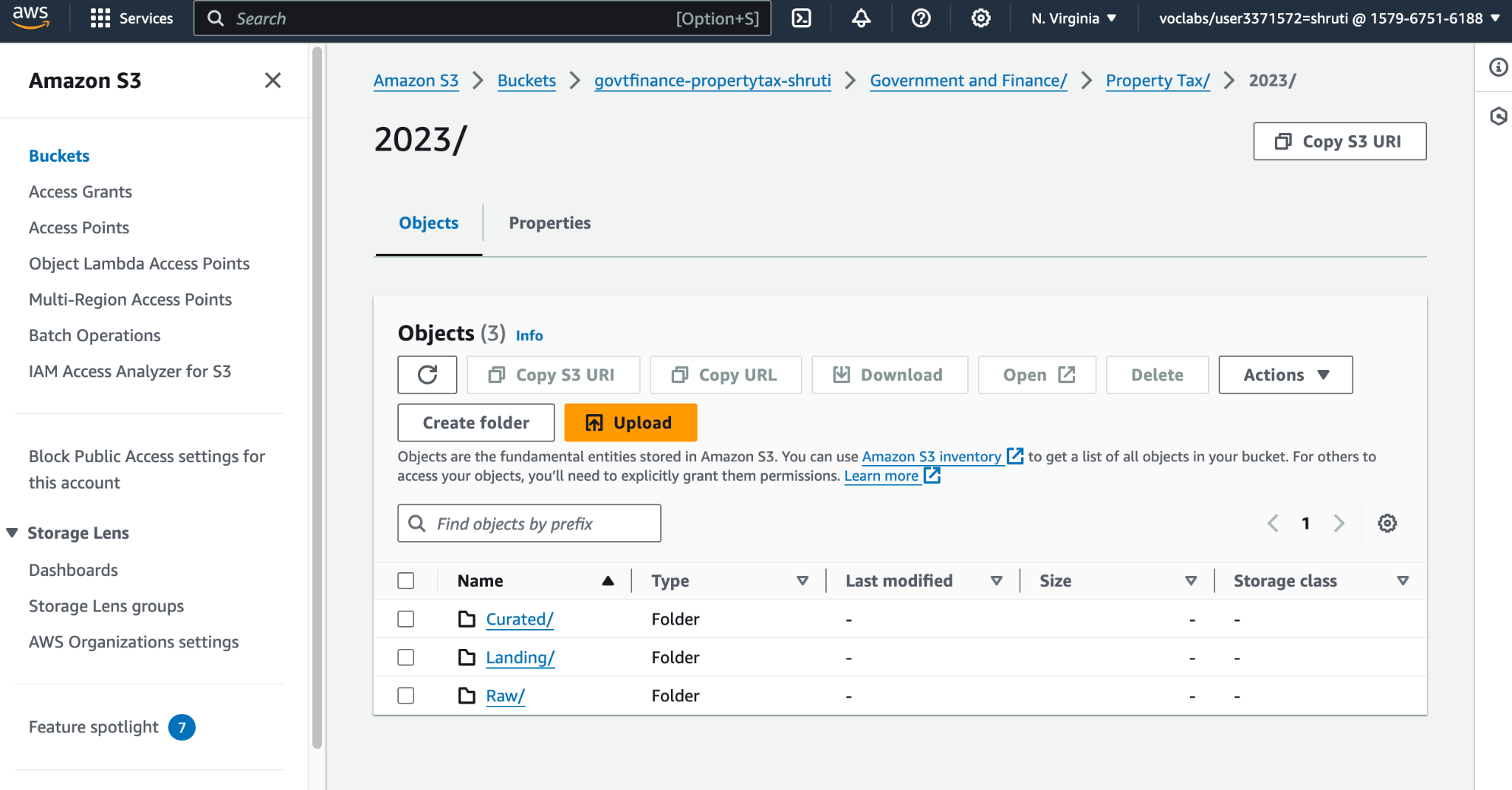
**Shruti: Property Tax Report**

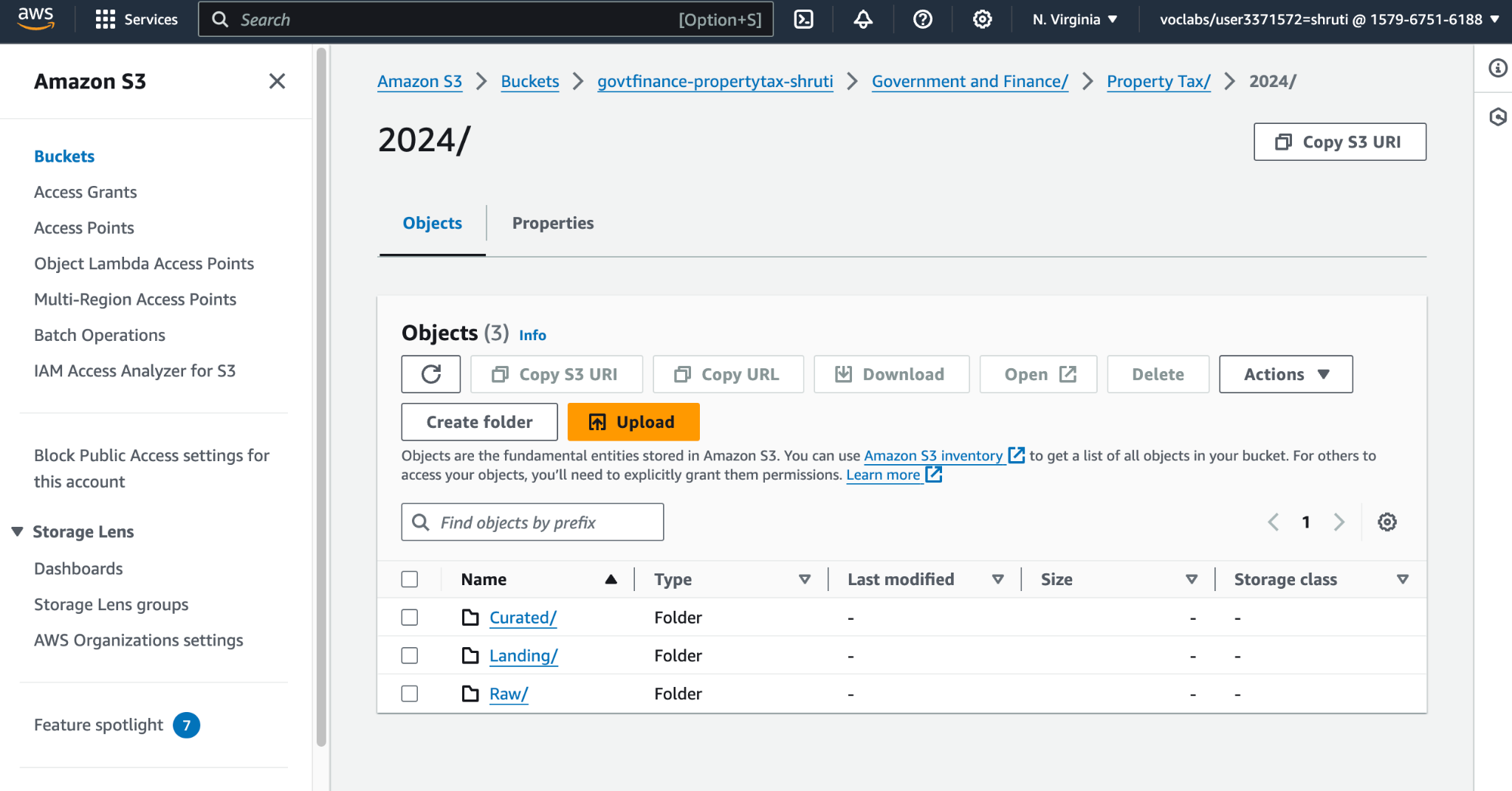
As for data ingestion, the property tax data for 2023 and 2024 was loaded into an Amazon S3 bucket under government and finance. The data files were stored in the folder labelled ‘Property Tax. and the folder which was given the name ‘Landing’ has subfolders named ‘2023’ and ‘2024’. This structure also enables effective storage and retrieval of the datasets that are going to be used in the subsequent analysis.

## 

## Step 6: Data Storage

**Shruti**

****

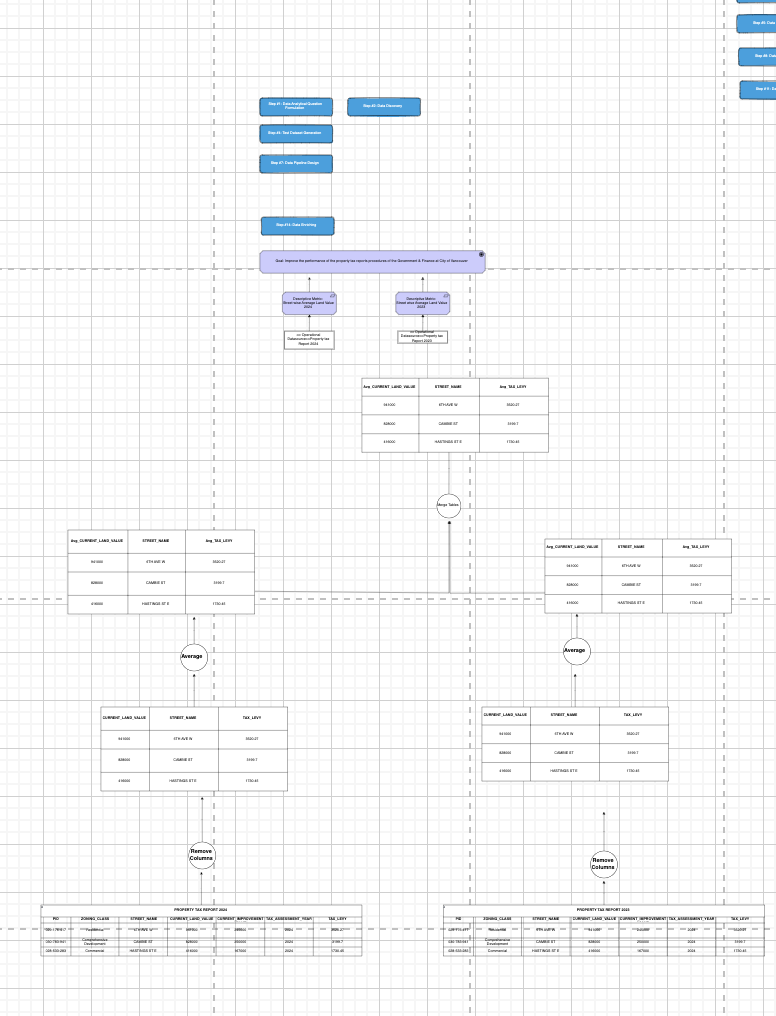
****

**Shruti: Property Tax Report**

When it comes to managing big data, particularly property tax data, Shruti adopted the use of Amazon S3, an object storage service in AWS that helps in the proper arrangement of the datasets. She then established subdirectories every year under the main directory that she named “Property Tax” including “2023” and “2024” to contain the property tax data files, well labelled and easily retrievable. This structure made data organization to be systematic, thus enabling an easy method of data access and management. Making sure that each set of data was also put in the right folder, Shruti made sure that no mixing up of data occurred, but rather made every data set easily accessible for analysis in the future. Thus, she took advantage of Amazon S3’s experiences in statement storage with unlimited storage space to store property tax records that contain financial information and have a secure environment.

## Step 7: Data Pipeline Design

**Shruti (draw.io)**

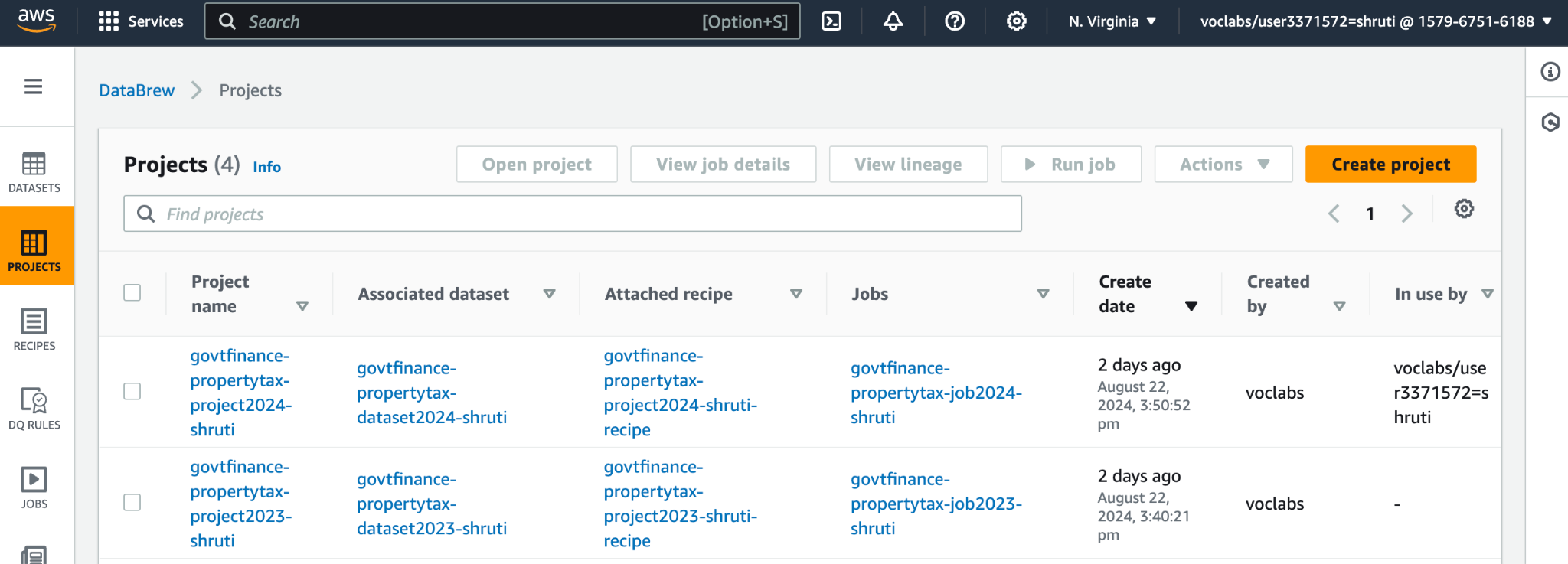
****

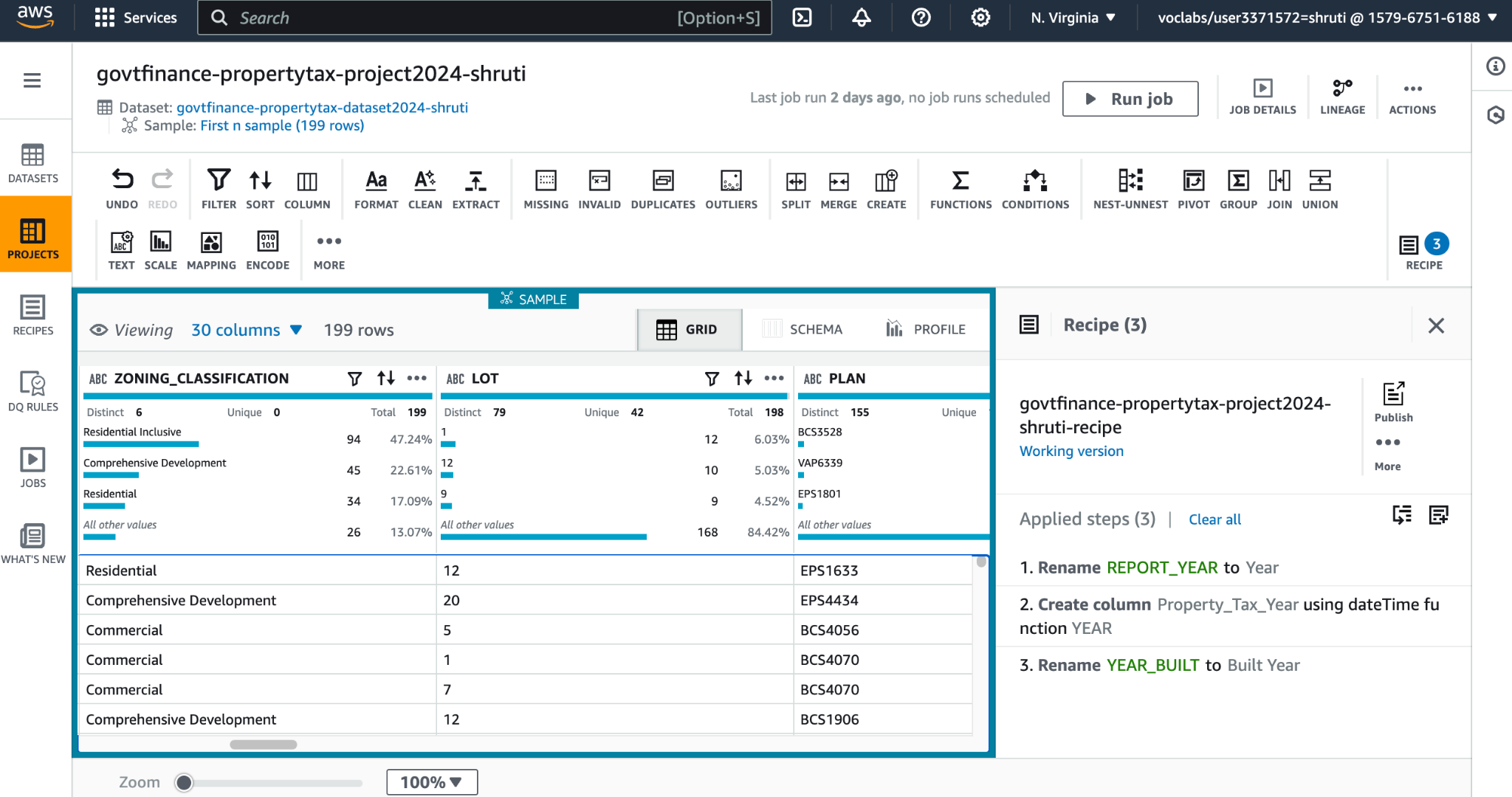
**Shruti: Property Tax Report**

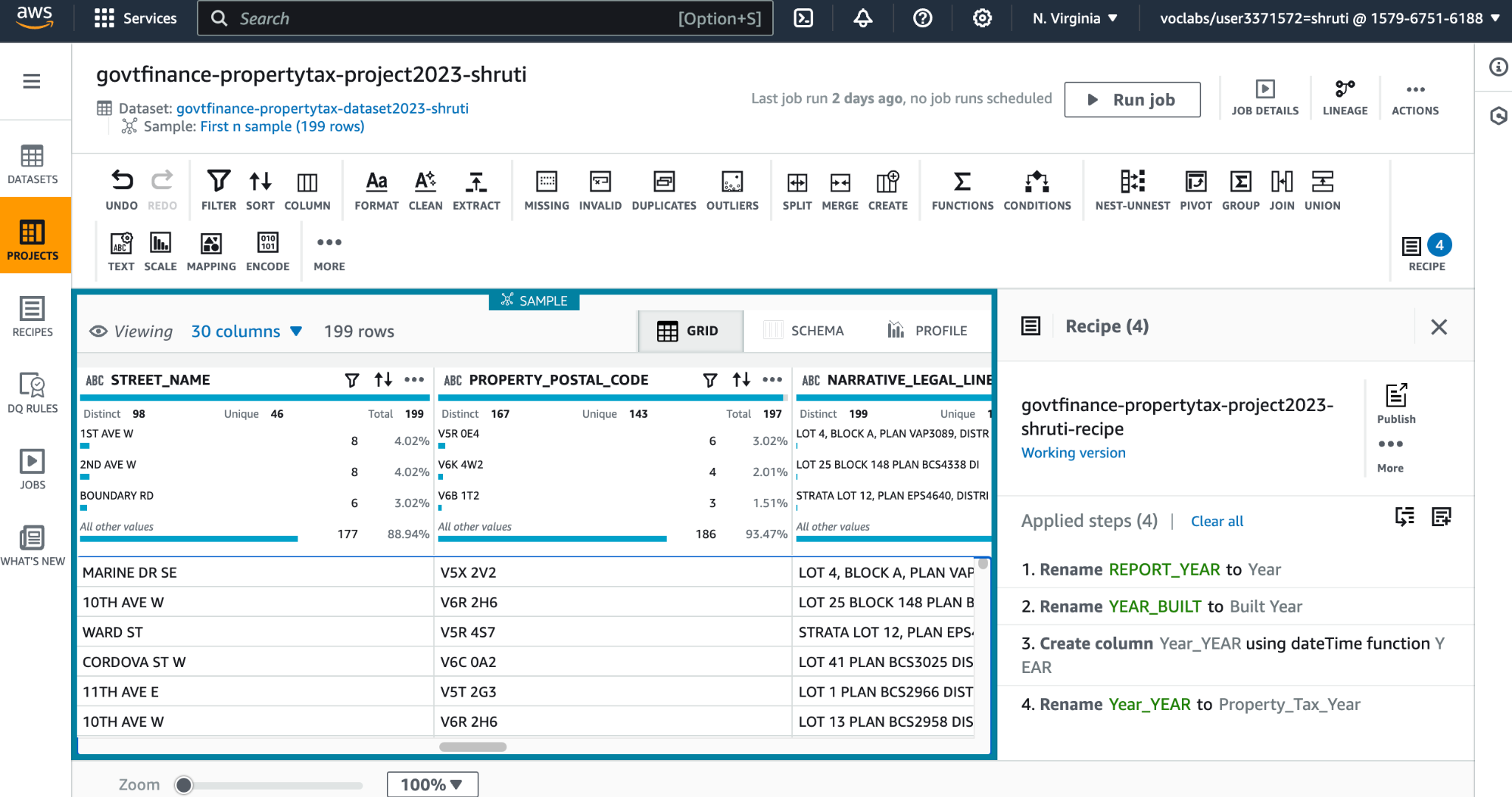
In a data pipeline process for the property tax report analysis, data is retrieved from the Amazon S3 where the datasets for the years 2023 & 2024 are located. The first step involves data ingestion on the pipeline, and data curation is done using AWS Glue to guarantee that all the required data fields including property values, tax assessments, and zoning classifications are accurate and free from errors. The raw cleaning then translates the data into an analysis form for easy computation of tax trends, high/low assessment periods and the variation in property values. Finally, the transformed data is moved to an Amazon Redshift data warehouse where more queries can be run on it. The data pipeline is planned for the daily job but since the data changes every month it is completed monthly so we can get the up-to-date data for the Government & Finance department. These automated pipeline designs, help in improving the data process and make it easier to make decisions based on accurate property tax data.

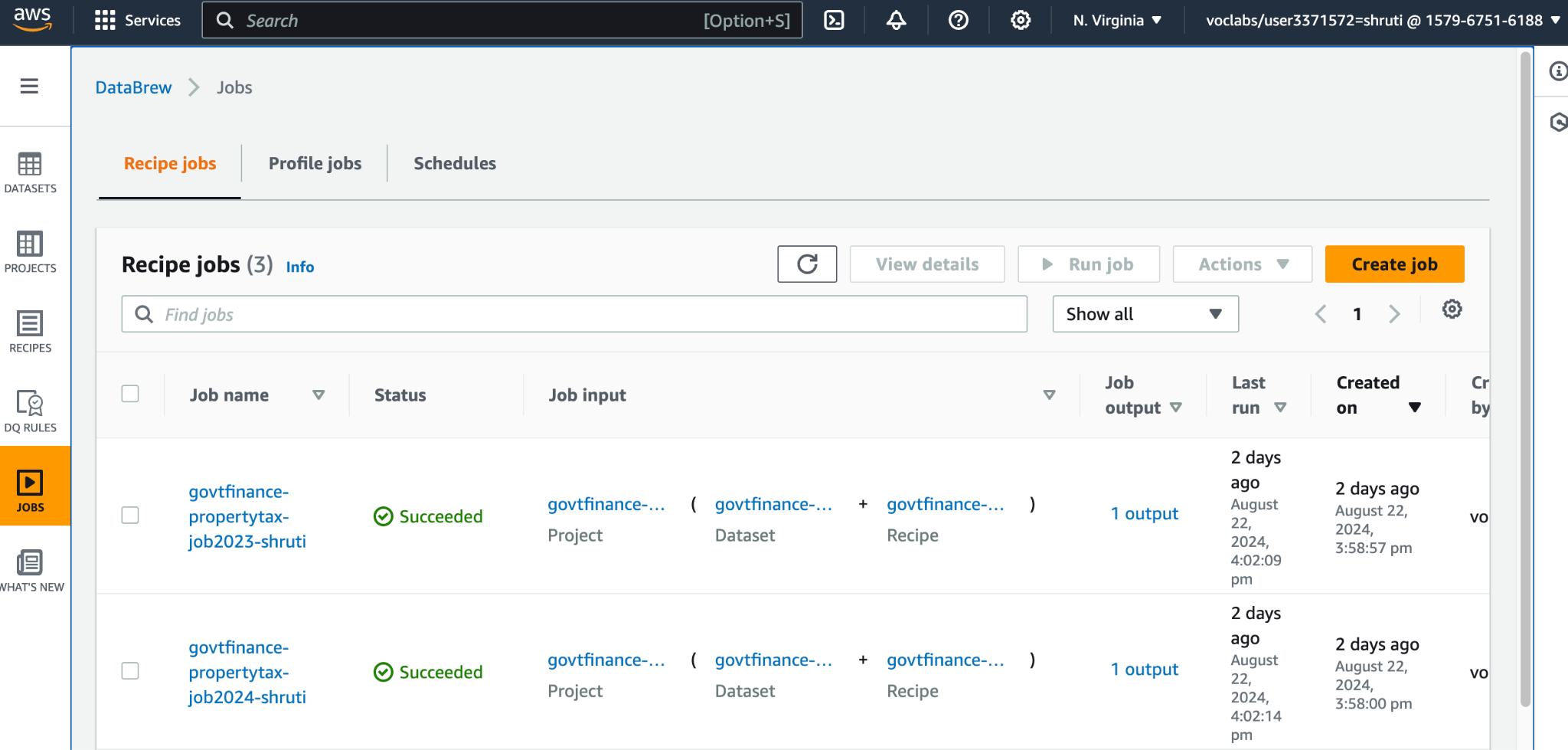
## Step 8: Data Cleaning

**Shruti**









**Shruti: Property Tax Report**  
 In the process of preparing data for the property tax report datasets, Shruti used AWS Glue DataBrew for data preparation. She started with the new project name “PropertyTax-Report-project-Shruti” and also established the matching dataset name known as “PropertyTax-Report-dataset-Shruti” in AWS Glue DataBrew. To the S3 folder under the ‘Landing’ directory, she linked the dataset, making certain that all files were in Excel format and as recommended by the professor. After putting up the project, Shruti went in to validate the quality of the data by assessing each column of the data frame for problematic data as missing values and invalid values. If there was any column with non-zero invalid or missing values, then cleaning of the data was done using DataBrew functions. Further, she looked for similar records and excluded them from the data set to avoid the misrepresentation of the data. Shruti configured better reveal under the Schema tab; this included changing the general field names to being more specific like changing the ‘PID’ to ‘PropertyID’ and the data type; where she realized that the stamps for the date-time were still in date-time format but needed to be in date format for analysis, and where Shruti realized that the Boolean value entries were textual form so made them the Boolean data type instead. For richer results, she also derived a new year column from the report date modifying the date to contain the year only. After doing all these steps, Shruti created a job named “PropertyTax-Report-job-Shruti”, where the location is in the Raw folder of the S3 bucket in CSV format. With “LabRole” permissions, she performed the job running all the data cleaning steps on the dataset. The project’s configuration guarantees that any file, new for the dataset folder, will go through the same cleaning process if the job is repeated, making the dataset always ready for analysis.

## 

## 

## 

## 

## 

## Step 9: Data Structuring

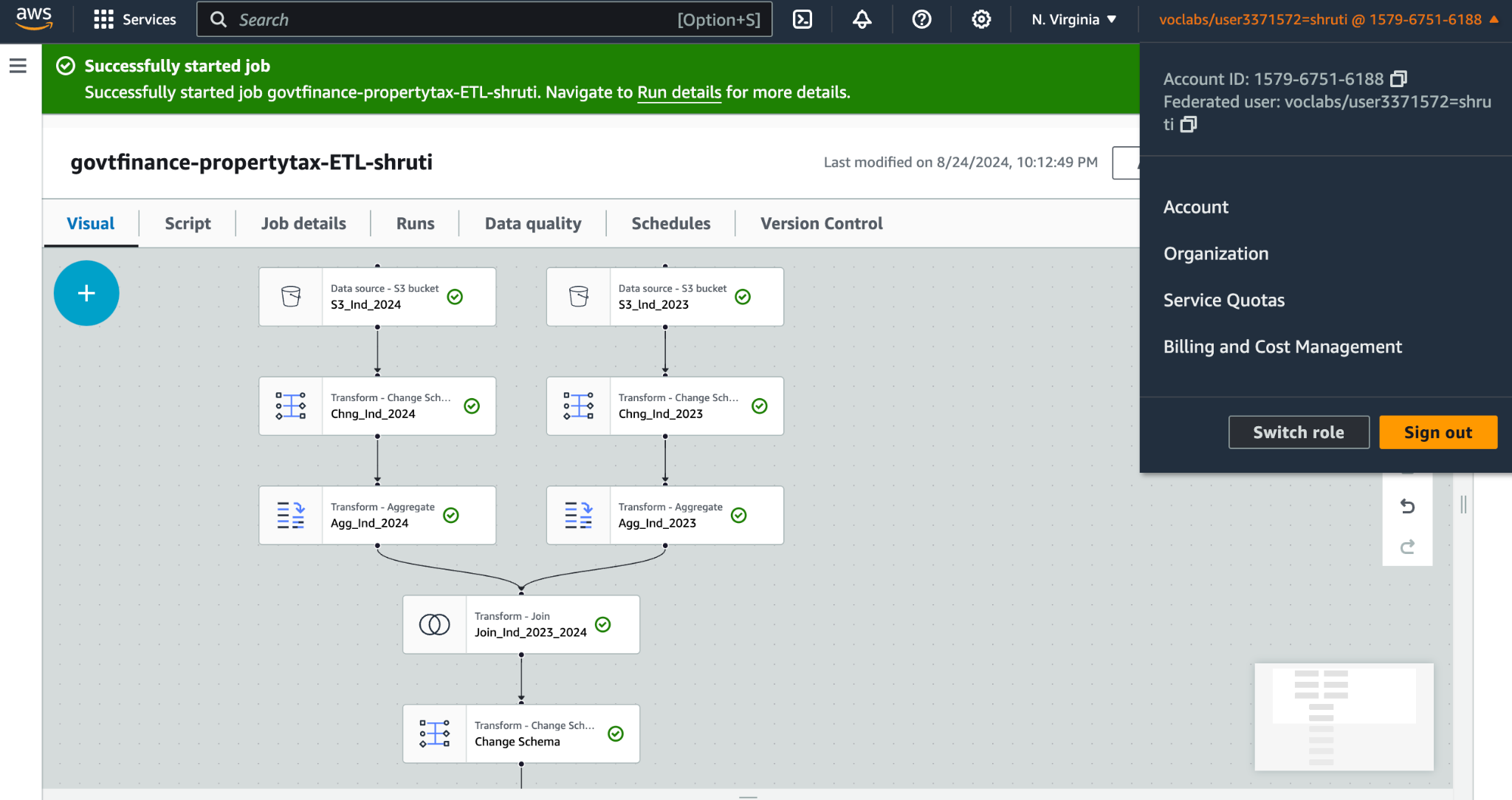
## 

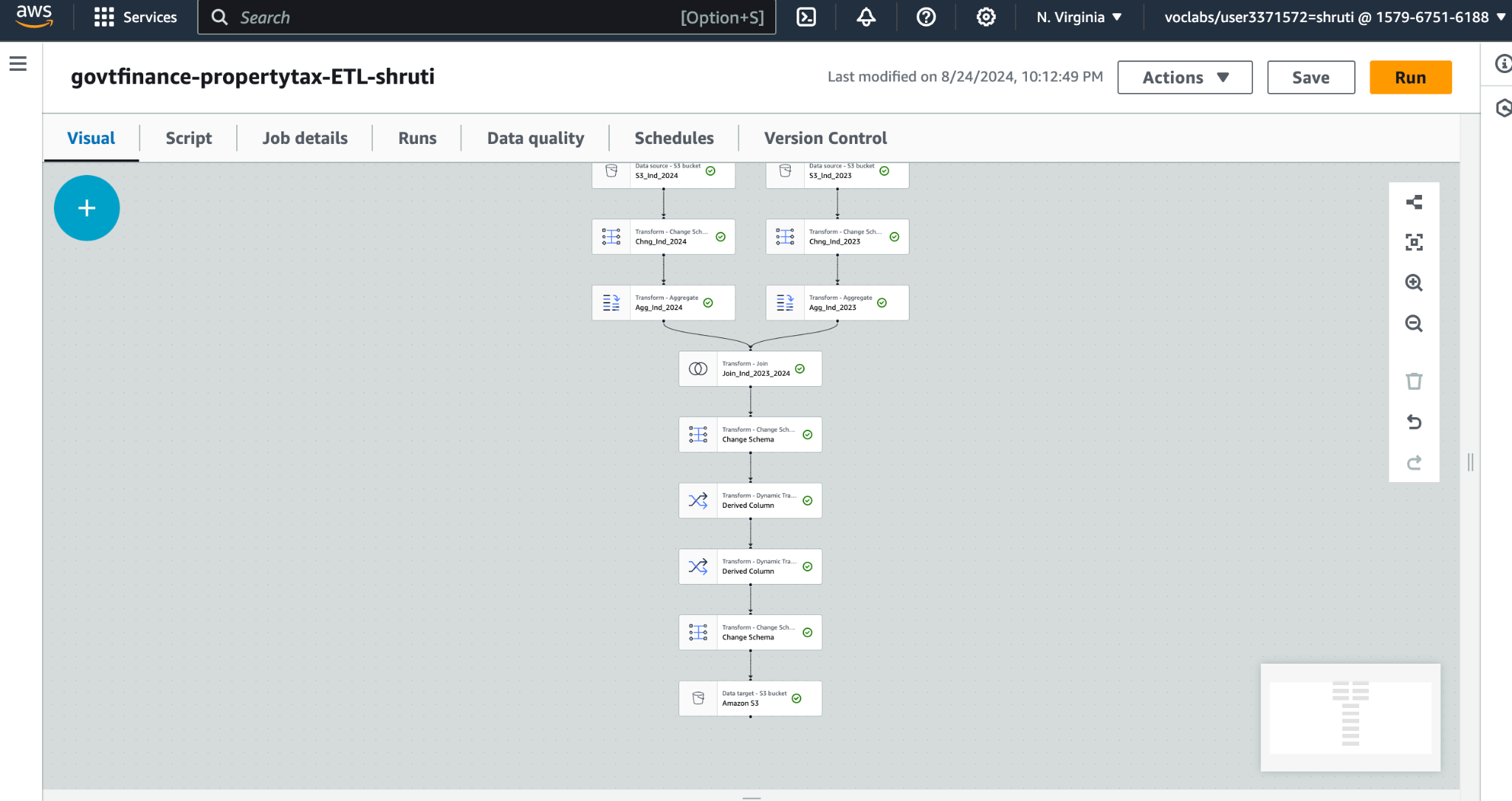
## **Shruti: Property Tax Report**

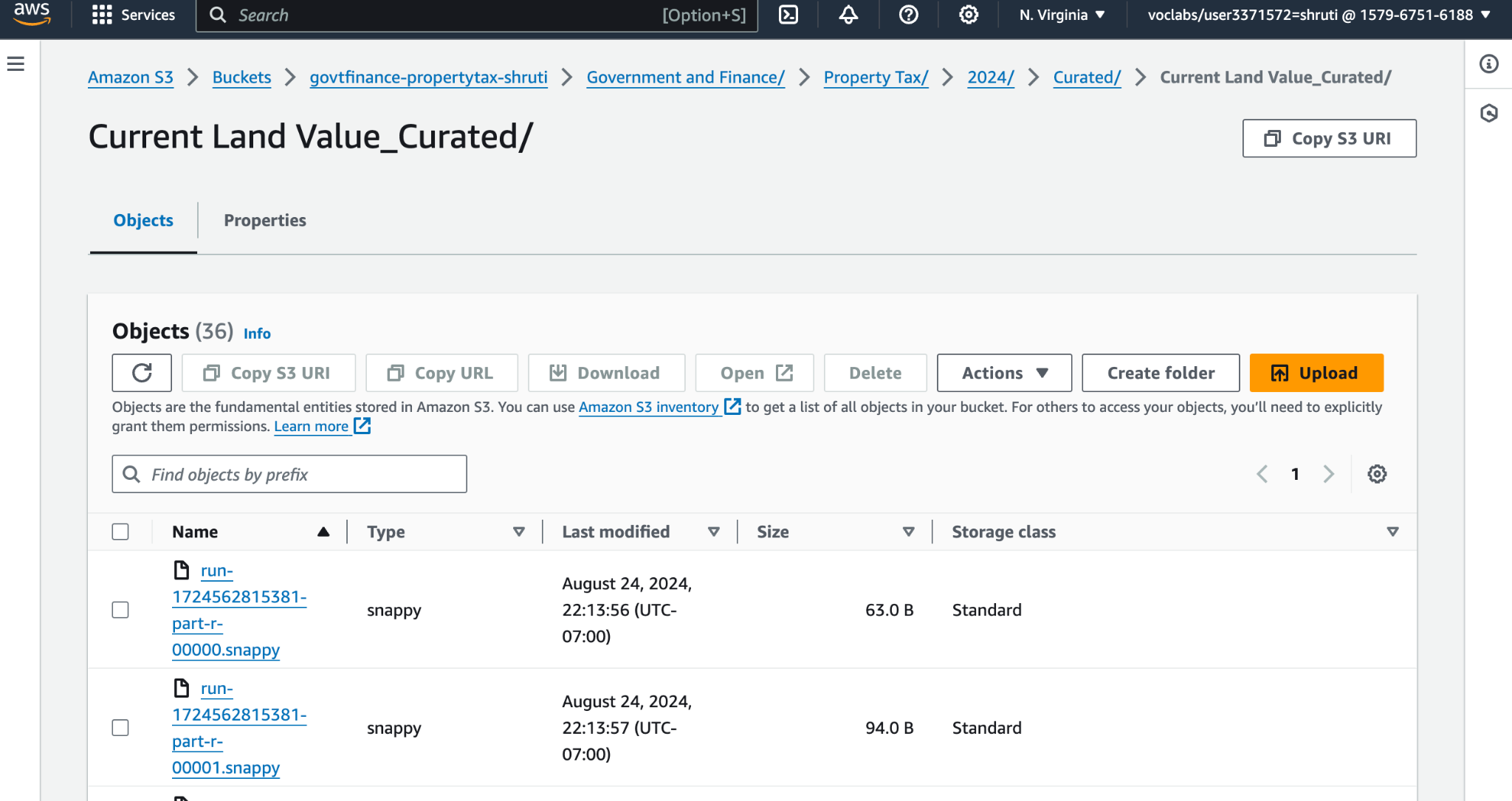
## As for property tax report data, Shruti analyzed the cleaned datasets into a form that could be analyzed and reported without many issues. She followed using AWS Glue Data Catalog where she logically erected a schema for the data fields of the 2023 and 2024 property tax reports. This schema comprises the property identification that pertains to legal type, PID and folio; a location that involves the land co-ordinate, the zoning district and classification; and finally, the financial information which includes the current as well as previous land and improvement value, the tax assessment year and the tax levy. Thus, having sub-samples derived from the partitions based on the ‘REPORT\_YEAR’ column Shruti was able to facilitate efficient querying and extraction of data for the given analysis to be performed on a year-to-year basis. Furthermore, she developed new derived attributes, including ‘PropertyValueChange’, for finding out the variations in property prices over a year and ‘TaxRateChange’ for getting a sense of change in tax rates. Such transformations were necessary when using the trend analysis to check for anomalies in the rates of tax collection or the property value assessment. To this end, the wrapped structured data was then moved into an Amazon Redshift data repository where it was appropriately sequenced in a suitable property tax analysis table. To enhance the capability of fast information search, this table has also added an index on some other columns like ‘PID’ and ‘REPORT\_YEAR’. As a result of arranging, it in this format, Shruti made certain that the dataset is not only prepared for further analysis but ready for future data integration as well. The structured format of the dataset is appropriate for carrying out a variety of analyses and the results include growth rates of the property tax, the differential effects of zoning on property values, and changes to the taxation levels over time.

## Step 10: Data Pipeline Implementation

**Shruti**





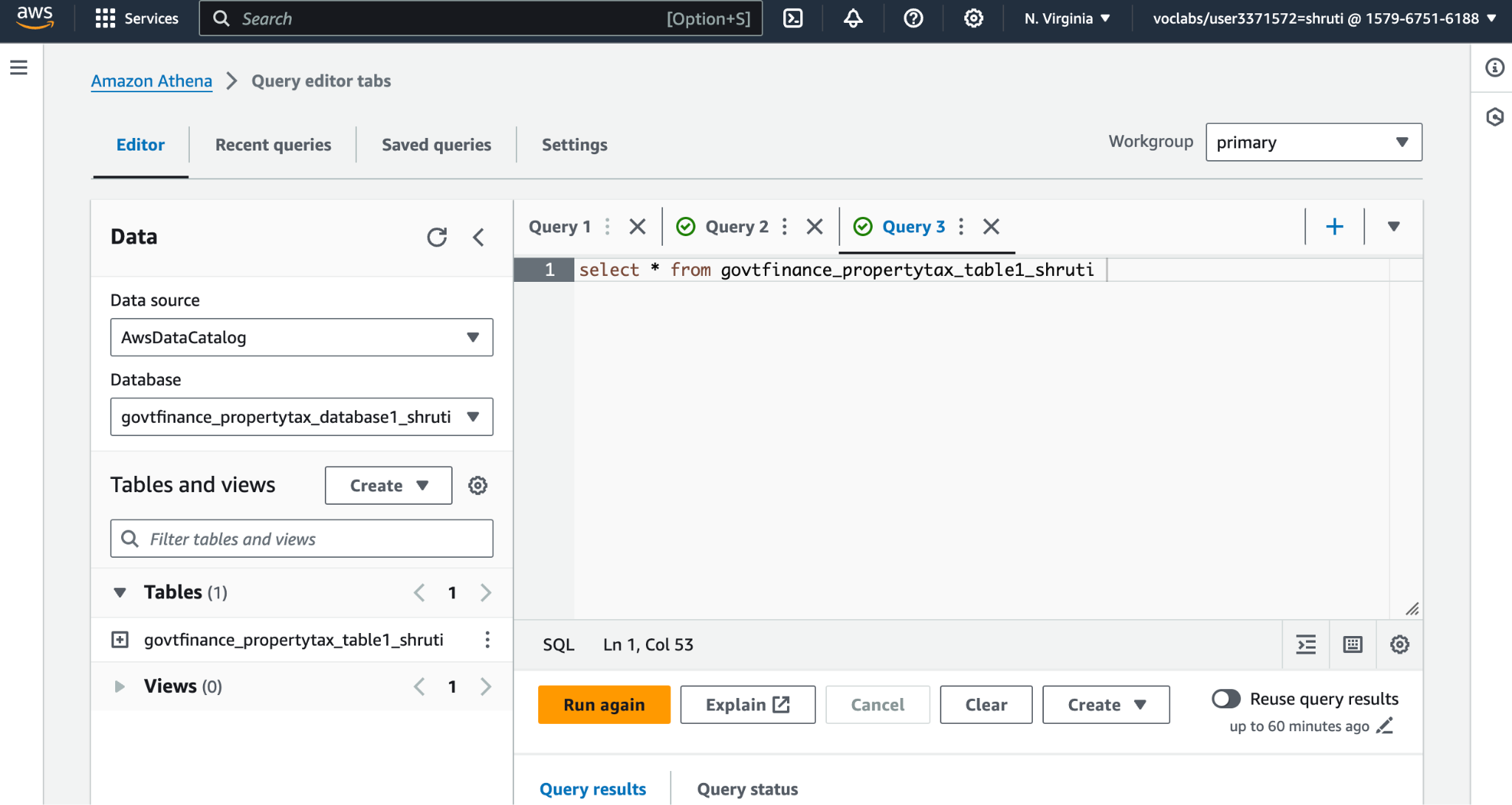


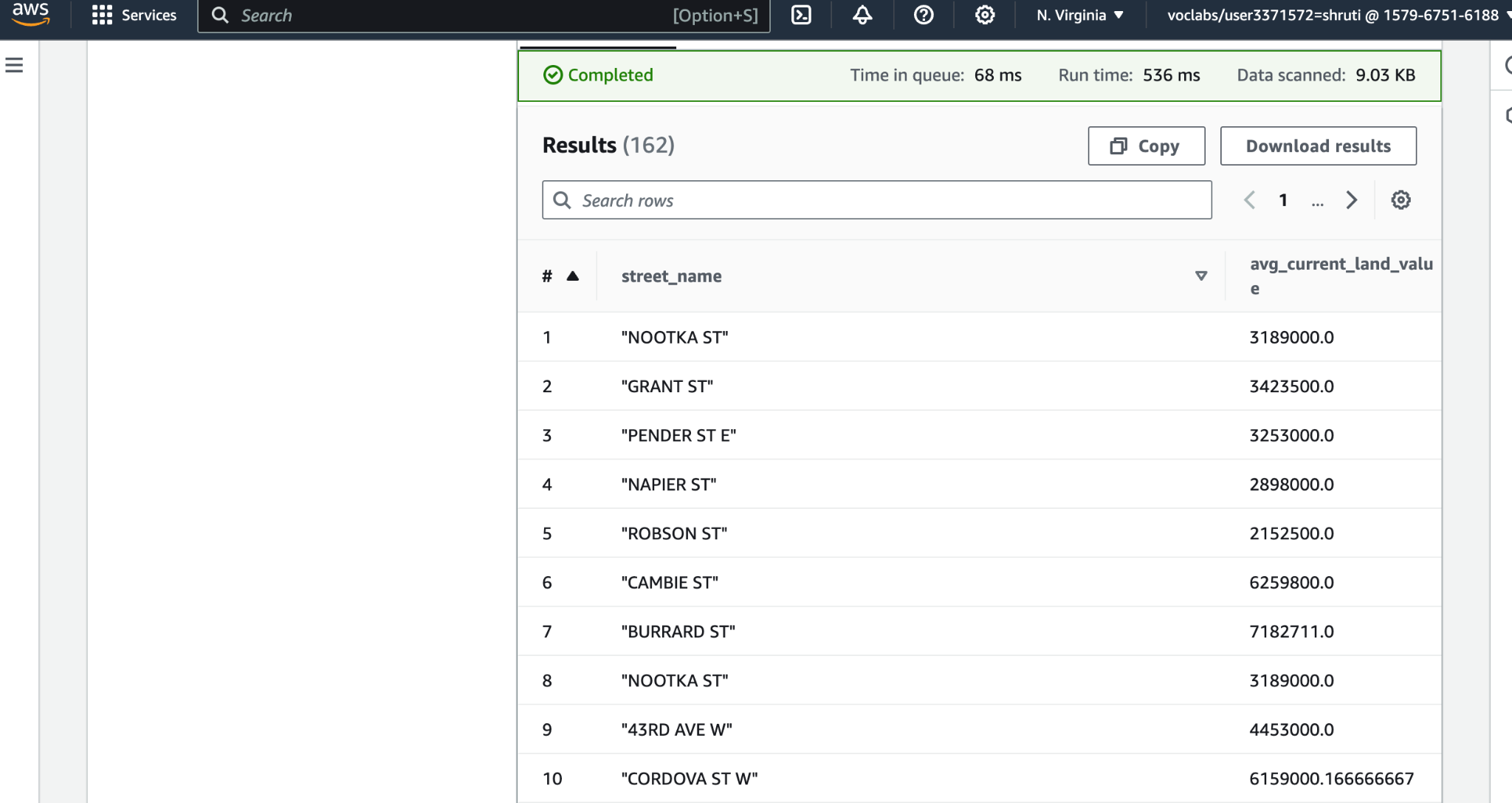
**Shruti: Property Tax Report**

To deal with the property tax report data, Shruti created the ETL (Extract, Transform, Load) pipeline using AWS Glue’s Visual ETL. She first went to AWS Glue and chose Visual ETL to create a new job named “PropertyTax-Report-ETL-Shruti. ” The initial processing step in the pipeline was to collect the required job role under which it had the permission to access raw data S3 bucket. During the ETL step, Shruti configured the source to pull the property tax datasets from S3’s ‘Landing’ folder. She proceeded to perform several operations on the dataset, including normalizing the data by converting data type, imputing for missing values and deduplicating data. Further, she has also added new columns for computed characteristics, such as ‘PropertyValueChange,’ and ‘TaxRateChange. ’ These metrics are significant to examine the tendency of property valuation and tax collection. The transformed data was then passed to S3 and moved to the ‘Curated’ folder so that it could be analyzed. Whenever she made changes to the files, she used to save her work to avoid losing any changes made to the files; she also tested the ETL job to confirm the correctness of the loaded data. Last of all, she set the program for the ETL job of this data to be done every month because any new data input to this S3 bucket would automatically undergo these transformations to always maintain an up-to-date property tax dataset for analysis.

## Step 11: Data Analysis

**Shruti**



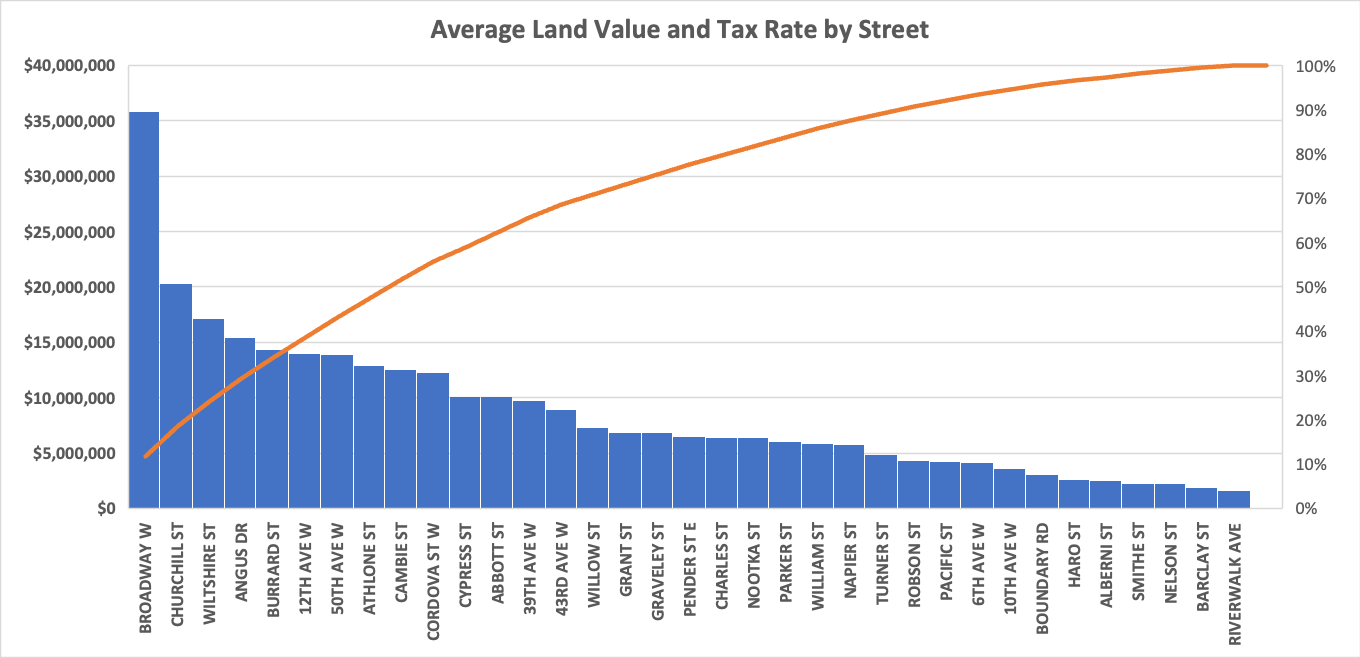


**Shruti: Property Tax Report**

For analyzing the data from the property tax report datasets, Shruti used Amazon Athena, a serverless query service which enables the users to run SQL queries on the data stored in Amazon S3. She began by getting a handle on the AWS Glue Data Catalog to which she had earlier registered the property tax datasets after cleaning them and structuring them properly. As an SQL interface, Shruti directly performed several queries in Athena that would help her understand the different features of the property data. Regarding this, one of her questions was concerned with retrieving the records of properties, which had the greatest current land value to ascertain important client property concentration zones as well as project the likely events entailed by zoning on property values. The query output shown on the screen was of streets along with the current land values where, as per Shruti, she was able to identify trends of properties of high value. Therefore, after going through the data, Shruti was able to point out such patterns as the areas that had experienced a high incline in land value; the efficiency of tax policies; and the regions that required review or supervision. Also, she employed the use of the aggregate functions in SQL to select averages, maximums and minimums of the property assessments so that she could prepare the report on the trends in property tax for the years. Consequently, with the aid of this analysis, it was possible to detect or rather identify some of the errors or anomalies that may exist in the records of property valuation and assessment of taxes. Hence, from the overall analysis, Shruti offered crucial insights into property tax and its changes in Vancouver, which may help to apply valuable recommendations that may improve property tax in the city and create better policies and strategies.

## Step 12: Data Visualization

**Shruti**

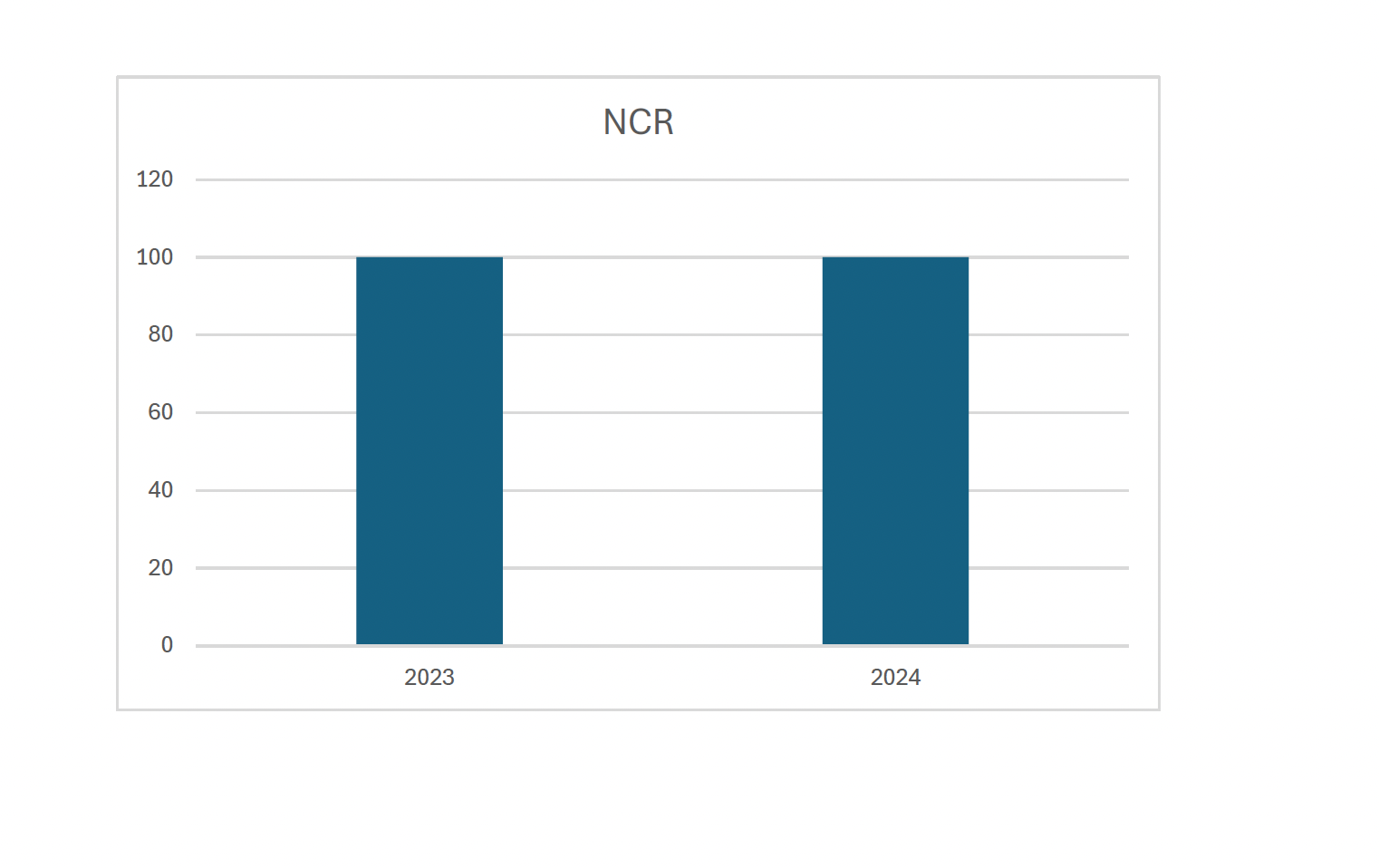


**Shruti: Property Tax Report**

As for the property tax report, Shruti can use the Bar Chart labelled “Average Land Value and Tax Rate by Street” to get at a glance, how the average land values are distributed and their correlation with the tax rates of most of the streets in Vancouver. Average land values are plotted on the y-axis while the streets are plotted on the x-axis to compare property values on different streets. Further disaggregation is done by using a line graph moving on top of the bar graph which illustrates the cumulative percentage of the total land value, this is useful in understanding how much value occupies the top streets. This visualization is important in pinpointing zones that yield high revenues meaning they contribute highly to the city’s taxation value and thus aid policymakers in decisions regarding taxation, zoning and distribution of resources. The analysis also takes cognizance of the relationship between the value of land and the tax rates, and this area can further be expanded to seek to enhance the tax policies to make taxation fair across different neighbourhoods.



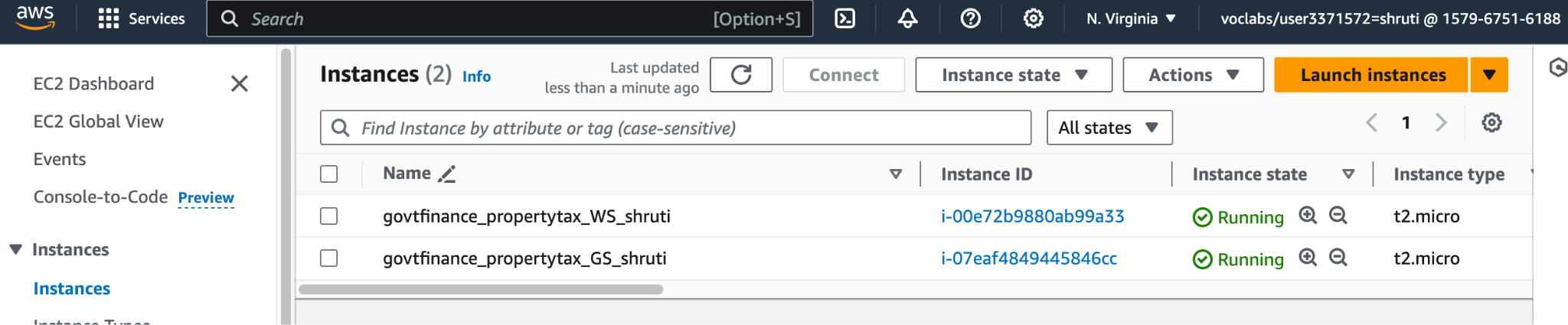
*Note*: The Bar Chart represents the Percentage of Votes “In Favour” (PVIF) for Year 2023 and 2024.

****

## 

## Step 13: Data Publishing

**Shruti**



When it came to data publishing, Shruti used the Amazon EC2 instances for hosting the property tax report data after processing them. Two EC2 instances, named by her as gov&finance\_property\_W1\_Shruti and gov&finance\_property\_GS\_Shruti are shown in the following screenshot. The first one was set up in a manner of a web server that enabled Web access to the processed property tax data through a secure Web interface. Such a format enables the stakeholders, including policymakers and the public, to scrutinize or interactively examine the trends in property taxes. To prevent unauthorized access to the data and to enhance the security measures, Shruti correctly set up the access control on the web server for data and the use of the correct encryption policy on the server. The second case was used as a general server also to store and manage internal data as well as in further processing tasks. The last cleaned property tax datasets were moved to the general server by Shruti, and the datasets were stored and made available to those with permission to analyze and report duly. The files were arranged in folders and subfolders and the folder names were made very clear to enable any user to easily find the data which they want to access. Such setup allows secure handling of internal data and simultaneous publishing of common data and insights without the need for additional servers for data management and publishing.

# DAP Estimated Cost (Teamwork)

*Please share the estimated cost of the dataset preparation phase.: You need to use the* [*AWS pricing calculator*](https://calculator.aws/#/) *to estimate the cost of using AWS services for this platform.*

*A screenshot of a calculator

Description automatically generatedA screenshot of a receipt

Description automatically generated*

The implementation of the AWS Data Analytic Platform (DAP) for the City of Vancouver’s Government and Finance department marks a significant step toward data-driven governance. By utilizing cloud technologies such as AWS Glue, S3, and Redshift, the city can now handle and analyze vast datasets related to property taxes with greater efficiency and accuracy. The platform’s design prioritizes scalability, ensuring that future data demands can be met without compromising performance. The analysis of property tax data from 2023 and 2024 provides valuable insights into tax collection trends, property value variations, and zoning impacts, enabling policymakers to refine taxation policies and enhance urban planning. This project demonstrates the importance of adopting cloud-based solutions for municipal governance and provides a blueprint for future expansions of the platform to accommodate other city departments and functions.