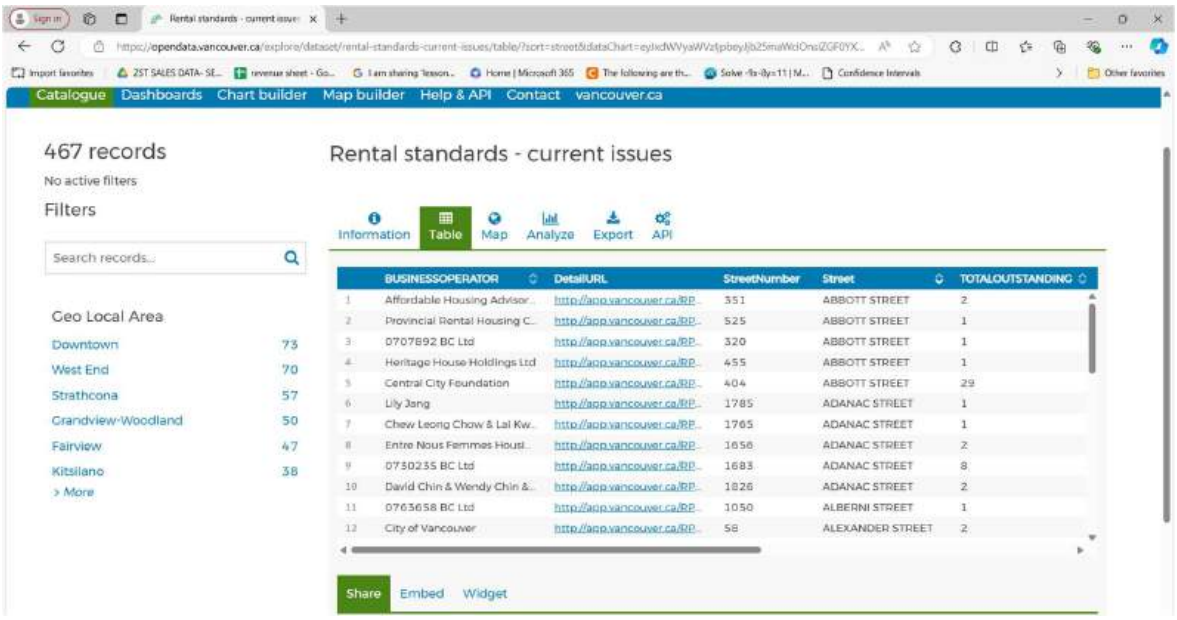
DAP Design and Implementation (Priyanka)

Step 1: Data Analytical Question Formulation

Question: What is the percentage of rental units with outstanding issues in Vancouver's Grandview-Woodland and Downtown neighborhoods?

Objective: To determine the proportion of rental units with outstanding issues in each neighborhood, allowing for a comparison of property management effectiveness and potential problem areas between Grandview-Woodland and Downtown.

Step 2: Data Discovery



Title: "Rental Standards - Current Issues"

The dataset contains information about rental properties in Vancouver that have outstanding issues related to rental standards.

Columns Displayed:

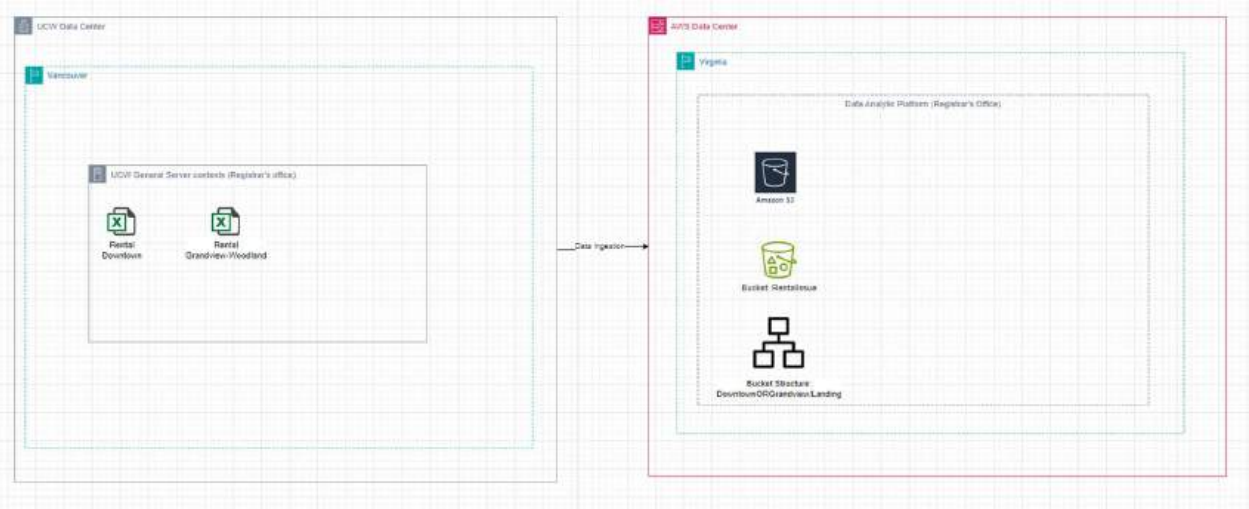
- **BUSINESSOPERATOR**: Lists the operators or companies responsible for the rental properties.
- **DetailURL**: Provides a URL link to more detailed information about each property.
- **StreetNumber**: The number assigned to the building on a specific street.

- **Street**: The name of the street where the property is located.
- **TOTALOUTSTANDING**: The number of outstanding issues reported for each property.
- **Totalunits:** The number of units reported for each property.
- **Geom**: The coordinates to the location for the property.
- **Geo Local Area:** The local area of the property.
- **geo_point_2d:** The longitude and latitude of the property

Geo Local Area Filters:

- On the left-hand side, there is a filter section for "Geo Local Area," which allows users to filter records by different neighborhoods such as Downtown, West End, Strathcona, Grandview-Woodland, etc. The numbers next to each area (e.g., Downtown - 73, Grandview-Woodland - 50) likely indicate the number of records/issues in that area.

Step 3: Data Storage Design



UCW Data Center (Vancouver):

The server in this data center now includes a "Rental Issue" designation.

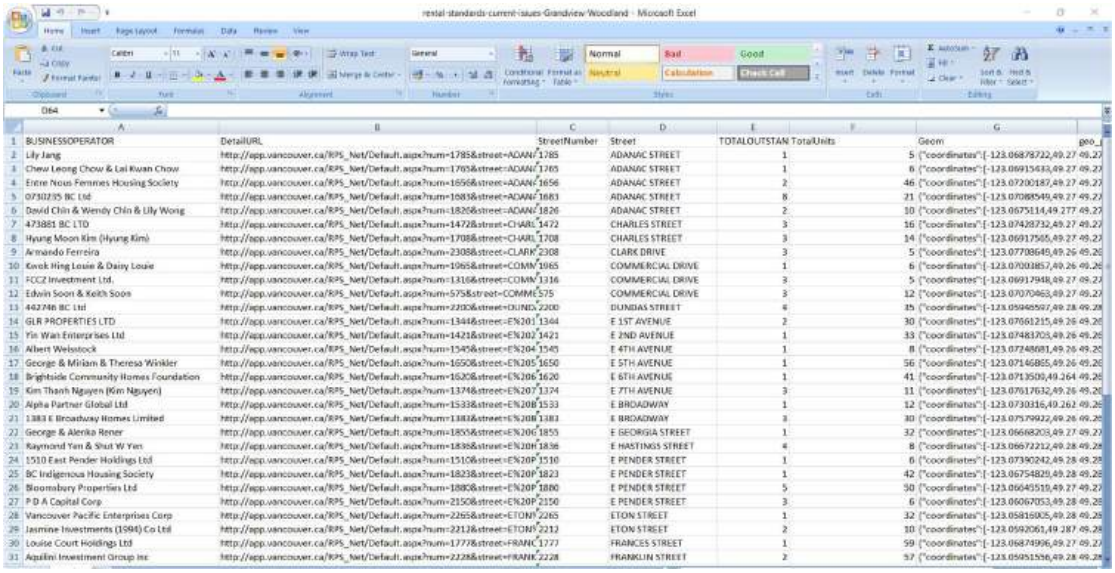
It still contains the Excel files for "Rental Downtown" and "Rental Grandview-Woodland."

AWS Data Center (Virginia):

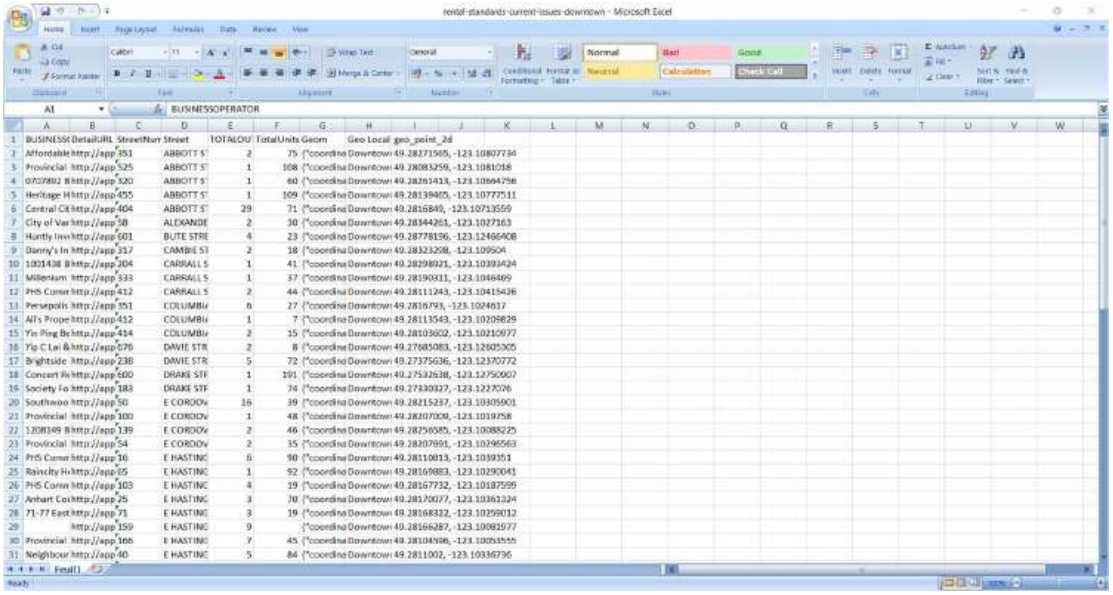
This part of the diagram remains largely the same, with the Amazon S3 bucket named RentalIssue storing the ingested data.

The bucket structure is specified to include directories (Downtown, Grandview, Landing), likely reflecting the organization of data within the S3 bucket for easier access and processing.

Step 4: Dataset Preparation



	A	B	C	D	E	F	G
1	BUSINESSOPERATOR	DetailURL	StreetNumber	Street	TOTALOUTSTAN	TotalUnits	Geom
2	Lily Jang	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1785&street=ADANA	1785	ADANAC STREET	1	5	{"coordinates":[-123.06879722,49.27
3	Chew Leong Chow & Lai Kwan Chow	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1765&street=ADANA	1765	ADANAC STREET	1	6	{"coordinates":[-123.06915433,49.27
4	Entre Nous Femmes Housing Society	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1656&street=ADANA	1656	ADANAC STREET	2	46	{"coordinates":[-123.07200187,49.27
5	0730295 BC Ltd	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1683&street=ADANA	1683	ADANAC STREET	8	21	{"coordinates":[-123.07086549,49.27
6	David Chin & Wendy Chin & Lily Wong	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1826&street=ADANA	1826	ADANAC STREET	2	10	{"coordinates":[-123.0675114,49.277
7	473881 BC LTD	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1472&street=CHARL	1472	CHARLES STREET	3	16	{"coordinates":[-123.07428732,49.27
8	Hyung Moon Kim (Hyung Kim)	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1708&street=CHARL	1708	CHARLES STREET	3	14	{"coordinates":[-123.06917565,49.27
9	Armando Ferreira	http://app.vancouver.ca/RPS_Net/Default.aspx?num=2308&street=CLARK	2308	CLARK DRIVE	3	5	{"coordinates":[-123.07708649,49.26
10	Kwok Hing Louie & Daisy Louie	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1965&street=COMM	1965	COMMERCIAL DRIVE	1	6	{"coordinates":[-123.07003857,49.26
11	FCCZ Investment Ltd.	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1316&street=COMM	1316	COMMERCIAL DRIVE	3	5	{"coordinates":[-123.06917948,49.27
12	Edwin Soon & Keith Soon	http://app.vancouver.ca/RPS_Net/Default.aspx?num=575&street=COMME	575	COMMERCIAL DRIVE	3	12	{"coordinates":[-123.07070463,49.27
13	442748 BC Ltd	http://app.vancouver.ca/RPS_Net/Default.aspx?num=2200&street=DUNDA	2200	DUNDAS STREET	4	15	{"coordinates":[-123.05946597,49.28
14	GLR PROPERTIES LTD	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1344&street=E%201	1344	E 1ST AVENUE	2	30	{"coordinates":[-123.07661215,49.26
15	Yin Wan Enterprises Ltd	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1421&street=E%202	1421	E 2ND AVENUE	1	33	{"coordinates":[-123.07481703,49.26
16	Albert Weisstock	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1545&street=E%204	1545	E 4TH AVENUE	1	8	{"coordinates":[-123.07348681,49.26
17	George & Miriam & Theresa Winkler	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1650&street=E%205	1650	E 5TH AVENUE	1	56	{"coordinates":[-123.07146865,49.26
18	Brightside Community Homes Foundation	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1620&street=E%206	1620	E 6TH AVENUE	1	41	{"coordinates":[-123.0713505,49.264
19	Kim Thanh Nguyen (Kim Nguyen)	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1374&street=E%207	1374	E 7TH AVENUE	3	11	{"coordinates":[-123.07617632,49.26
20	Alpha Partner Global Ltd	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1533&street=E%208	1533	E BROADWAY	1	12	{"coordinates":[-123.0730316,49.262
21	1383 E Broadway Homes Limited	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1383&street=E%208	1383	E BROADWAY	3	30	{"coordinates":[-123.07579922,49.26
22	George & Alenka Rener	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1855&street=E%206	1855	E GEORGIA STREET	1	32	{"coordinates":[-123.06668203,49.27
23	Raymond Yan & Shut W Yan	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1838&street=E%208	1838	E HASTINGS STREET	4	8	{"coordinates":[-123.06672212,49.28
24	1510 East Pender Holdings Ltd	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1510&street=E%20P	1510	E PENDER STREET	1	6	{"coordinates":[-123.07390242,49.28
25	BC Indigenous Housing Society	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1823&street=E%20P	1823	E PENDER STREET	1	42	{"coordinates":[-123.06754829,49.28
26	Bloomsbury Properties Ltd	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1880&street=E%20P	1880	E PENDER STREET	5	50	{"coordinates":[-123.06045519,49.27
27	P D A Capital Corp	http://app.vancouver.ca/RPS_Net/Default.aspx?num=2150&street=E%20P	2150	E PENDER STREET	3	6	{"coordinates":[-123.06067053,49.28
28	Vancouver Pacific Enterprises Corp	http://app.vancouver.ca/RPS_Net/Default.aspx?num=2265&street=ETON	2265	ETON STREET	1	32	{"coordinates":[-123.05816005,49.28
29	Jasmine Investments (1984) Co Ltd	http://app.vancouver.ca/RPS_Net/Default.aspx?num=2212&street=ETON	2212	ETON STREET	2	10	{"coordinates":[-123.0592061,49.287
30	Louise Court Holdings Ltd	http://app.vancouver.ca/RPS_Net/Default.aspx?num=1777&street=FRANC	1777	FRANCES STREET	1	59	{"coordinates":[-123.06874906,49.27
31	Aquilini Investment Group Inc	http://app.vancouver.ca/RPS_Net/Default.aspx?num=2228&street=FRANK	2228	FRANKLIN STREET	2	57	{"coordinates":[-123.05951536,49.28



	A	B	C	D	E	F	G	H	I	J	K
1	BUSINESSOPERATOR	DetailURL	StreetNum	Street	TOTALOU	TotalUnits	Geom	Geo Local	geo_point_2d		
2	Affordable	http://app	351	ABBOTT S	2	75	{"coordina	Downtown	49.28271505, -123.10807734		
3	Provincial	http://app	525	ABBOTT S	1	108	{"coordina	Downtown	49.28083259, -123.1081018		
4	0707892 B	http://app	320	ABBOTT S	1	60	{"coordina	Downtown	49.28261413, -123.10864758		
5	Heritage H	http://app	455	ABBOTT S	1	109	{"coordina	Downtown	49.28139405, -123.10777511		
6	Central Cit	http://app	404	ABBOTT S	29	71	{"coordina	Downtown	49.2816849, -123.10713559		
7	City of Var	http://app	58	ALEXANDE	2	30	{"coordina	Downtown	49.28344261, -123.1027163		
8	Huntly Inn	http://app	601	BUTE STRE	4	23	{"coordina	Downtown	49.28778196, -123.12466408		
9	Danny's In	http://app	317	CAMBIE ST	2	18	{"coordina	Downtown	49.28323298, -123.109504		
10	1001438 B	http://app	204	CARRALL S	1	41	{"coordina	Downtown	49.28298921, -123.10393424		
11	Millenium	http://app	333	CARRALL S	1	37	{"coordina	Downtown	49.28190311, -123.1046469		
12	PHS Comm	http://app	412	CARRALL S	2	44	{"coordina	Downtown	49.28111243, -123.10415426		
13	Persepolis	http://app	351	COLUMBIA	6	27	{"coordina	Downtown	49.2816793, -123.1024617		
14	Ali's Prope	http://app	412	COLUMBIA	1	7	{"coordina	Downtown	49.28113543, -123.10205829		
15	Yin Ping Bo	http://app	414	COLUMBIA	2	15	{"coordina	Downtown	49.28103602, -123.10210977		
16	Yip C Lai &	http://app	676	DAVIE STR	2	8	{"coordina	Downtown	49.27685083, -123.12605305		
17	Brightside	http://app	238	DAVIE STR	5	72	{"coordina	Downtown	49.27375636, -123.12370772		
18	Concert Pr	http://app	600	DRAKE STF	1	191	{"coordina	Downtown	49.27532638, -123.12750007		
19	Society Fo	http://app	183	DRAKE STF	1	74	{"coordina	Downtown	49.27330327, -123.1227026		
20	Southwoo	http://app	50	E CORDOV	16	39	{"coordina	Downtown	49.28215237, -123.10305901		
21	Provincial	http://app	100	E CORDOV	1	48	{"coordina	Downtown	49.28207009, -123.1019758		
22	1208149 B	http://app	139	E CORDOV	2	46	{"coordina	Downtown	49.28256585, -123.10088225		
23	Provincial	http://app	54	E CORDOV	2	35	{"coordina	Downtown	49.28207691, -123.10296563		
24	PHS Comm	http://app	16	E HASTINC	6	90	{"coordina	Downtown	49.28110813, -123.1039351		
25	Raincity H	http://app	65	E HASTINC	1	92	{"coordina	Downtown	49.28169883, -123.10290041		
26	PHS Comm	http://app	103	E HASTINC	4	19	{"coordina	Downtown	49.28167732, -123.10187599		
27	Anhart Co	http://app	25	E HASTINC	3	70	{"coordina	Downtown	49.28170077, -123.10361324		
28	71-77 East	http://app	71	E HASTINC	3	19	{"coordina	Downtown	49.28168322, -123.10259012		
29		http://app	159	E HASTINC	9		{"coordina	Downtown	49.28166287, -123.10081977		
30	Provincial	http://app	166	E HASTINC	7	45	{"coordina	Downtown	49.28104596, -123.10053555		
31	Neighbour	http://app	40	E HASTINC	5	84	{"coordina	Downtown	49.2811002, -123.10336736		

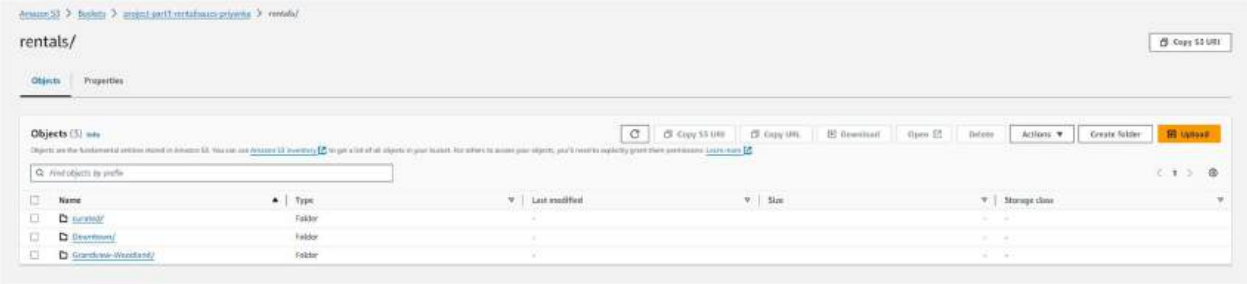
Filtered Dataset:

- The dataset is filtered by the "Geo Local Area" of **Grandview-Woodland**. This means that the data shown and ready for export is specific to rental issues in the Grandview-Woodland neighborhood, with a total of **50 records**.

Export Options:

Excel: This option allows for exporting the dataset directly into Excel format, which is particularly useful for further analysis or integration into other tools like spreadsheets.

Step 5: Data Ingestion

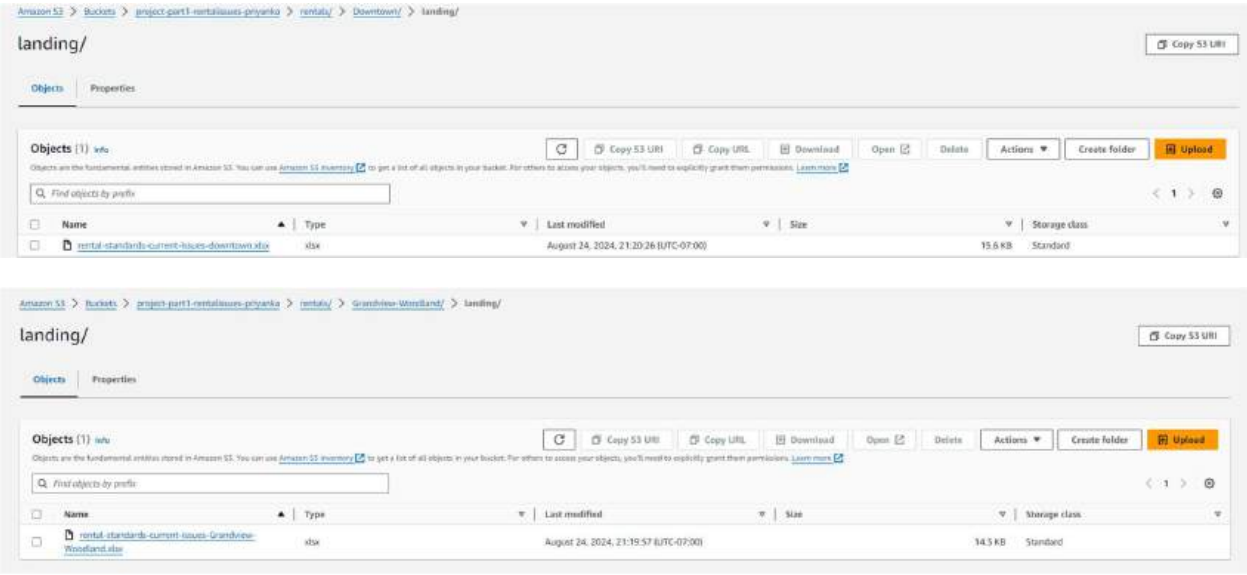


Create S3 Bucket: Create an S3 bucket (e.g., project-part1-rentalissues-priyanka) to store your data.

Create Folders: Inside the bucket, create folders for each neighborhood (Downtown and Grandview-Woodland), and within each, add a landing/ folder to hold the ingested files.

The ingested rental data files are securely stored in Amazon S3, organized by neighborhoods (Downtown and Grandview-Woodland) within specific directories. This structured storage ensures easy access and management, preparing the data for further analysis and processing.

Step 6: Data Storage

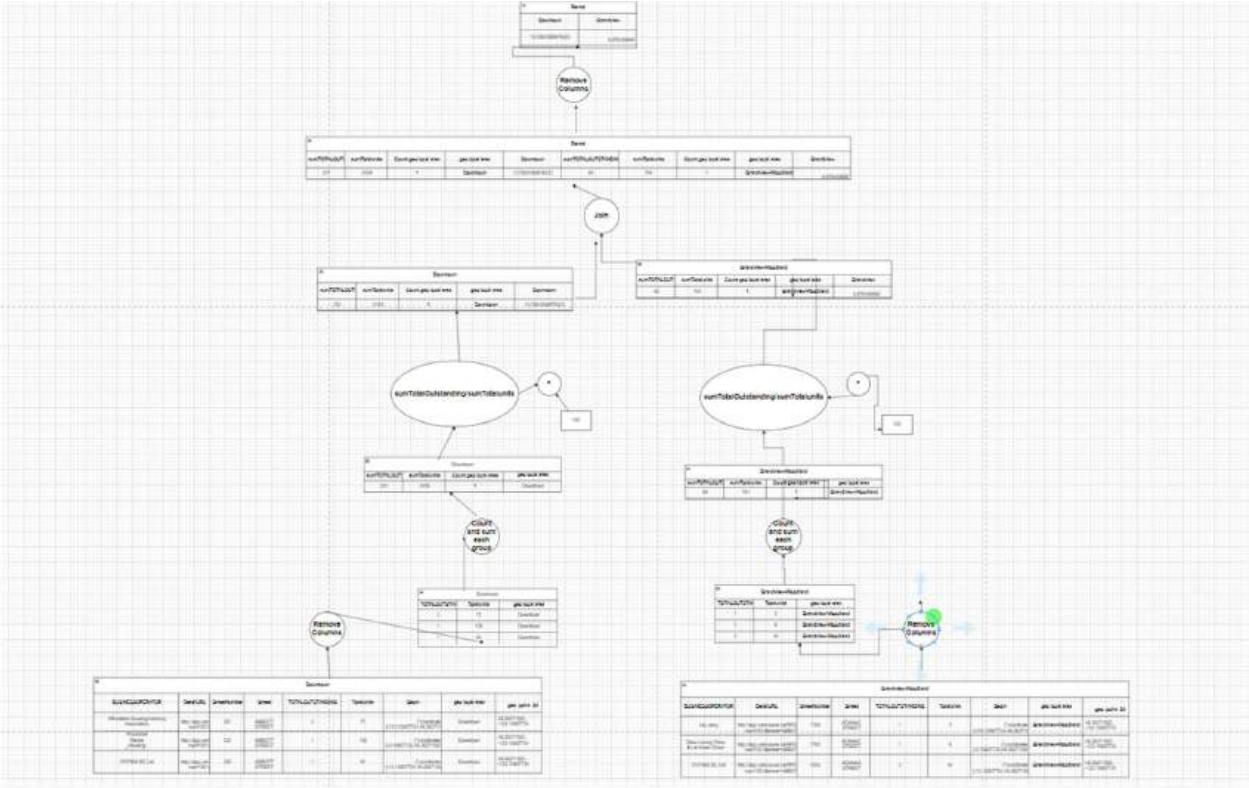


Upload Data: Upload the relevant data files to the landing/ folders, keeping everything organized and accessible for further processing.

The rental issue data files for Downtown and Grandview-Woodland have been successfully ingested into the Amazon S3 bucket within the respective directories (Downtown/landing/ and Grandview-Woodland/landing/).

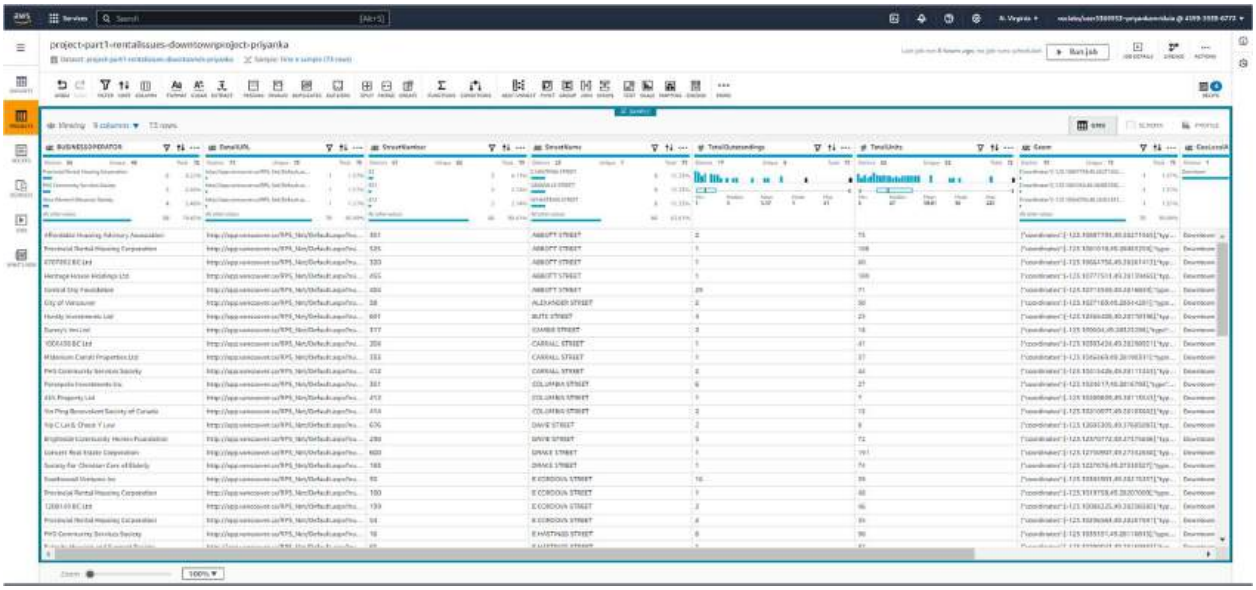
These files are now stored in a centralized, organized structure, making them ready for further processing in your Data Analytic Platform. The S3 interface also provides options for easy management, such as downloading or sharing files. This confirms that the data ingestion process is complete, and the data is accessible for analysis.

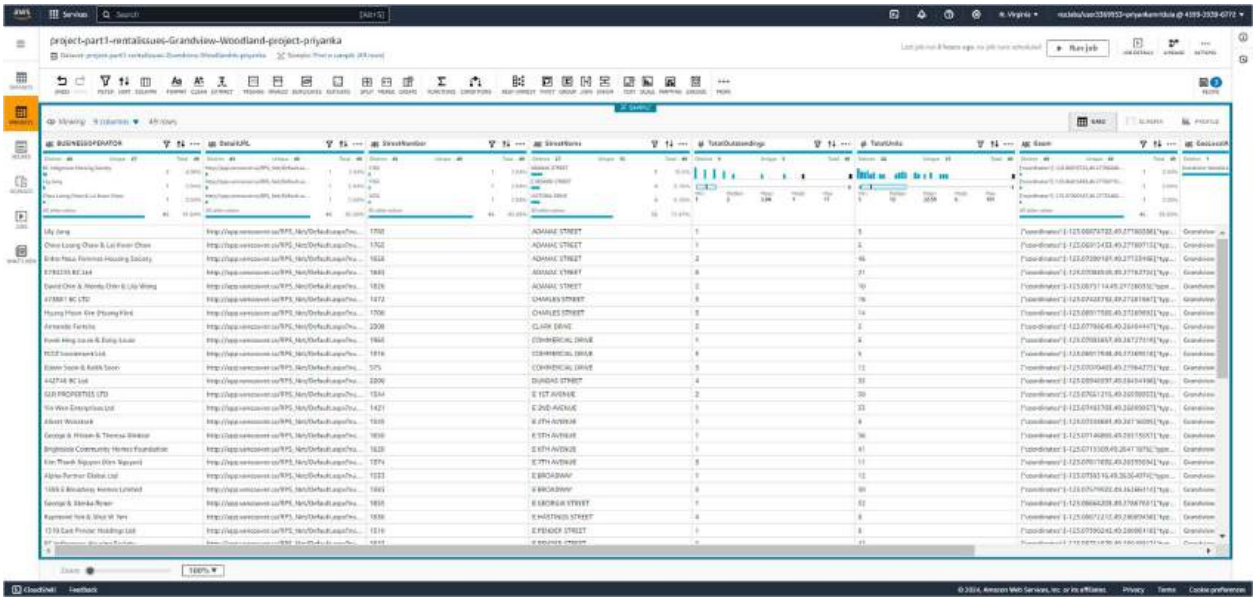
Step 7: Data Pipeline Design



- **Initial Data**: Starts with raw datasets for each neighborhood, containing columns like TotalOutstanding, TotalUnits, and geo local area.
- **Column Removal**: Unnecessary columns are removed to focus on relevant data.
- **Data Join**: The datasets from both neighborhoods are joined to allow for comparative analysis.
- **Aggregation & Calculation**: Summation of TotalOutstanding and TotalUnits is done for each neighborhood. The ratio of outstanding issues to total units is then calculated.
- **Grouping**: Data is grouped by neighborhood, and counts or sums are computed within each group.
- **Final Tables**: The final tables show the processed data, summarizing the rental issues and their ratios in each neighborhood, ready for analysis.

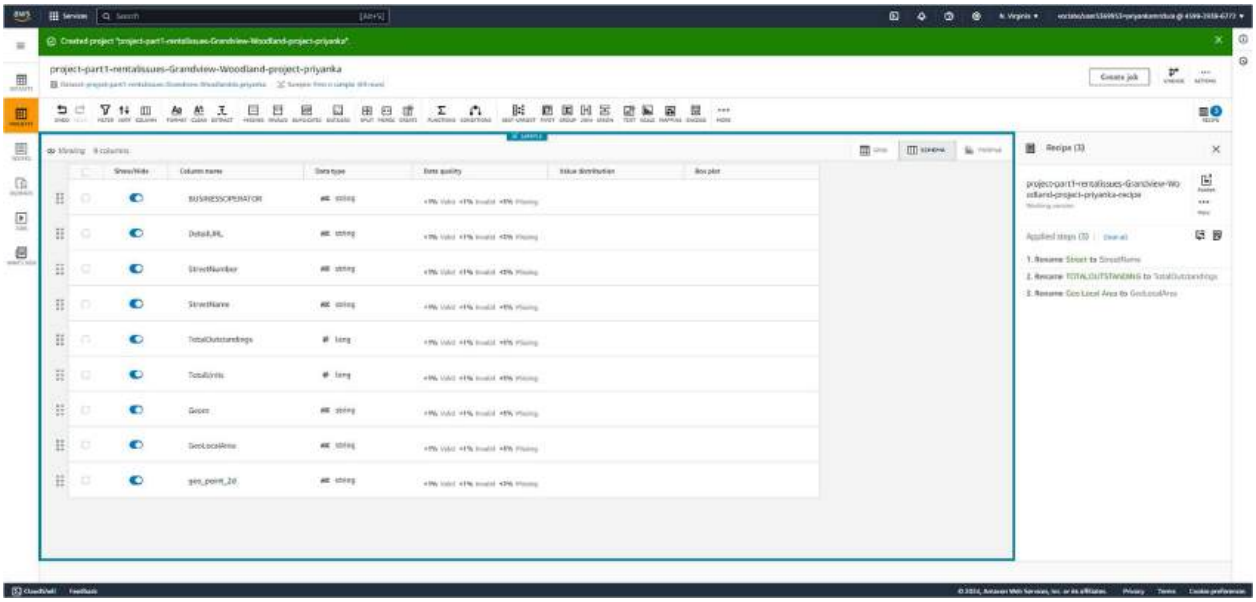
Step 8: Data Cleaning

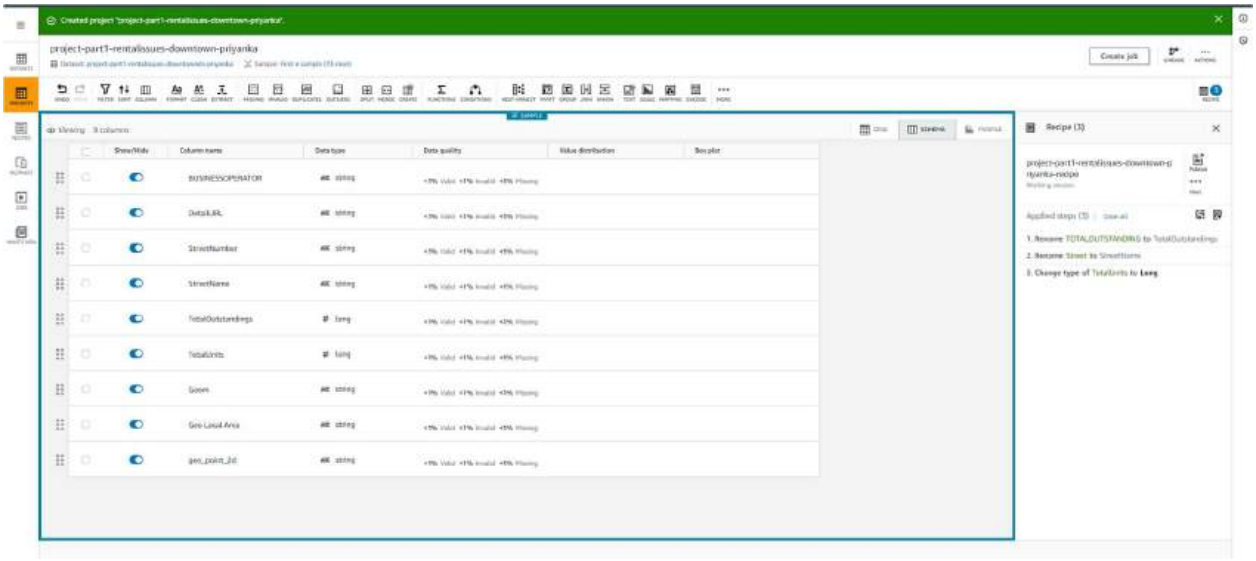




- **Downtown Dataset**: Key columns like BUSINESSOPERATOR, StreetName, TotalOutstandings, and TotalUnits are checked for data quality issues like duplicates or missing values.
- **Grandview-Woodland Dataset**: Similar cleaning is applied to ensure consistency and correctness across key columns.

Step 9: Data Structuring

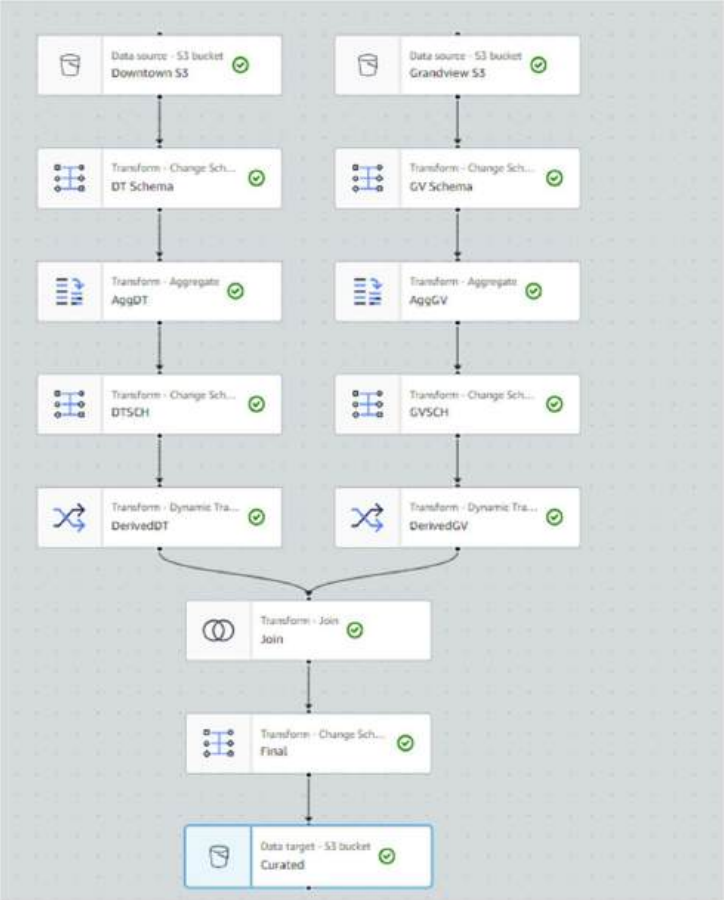




- **Renaming Columns**: Columns like Street were renamed to StreetName for clarity and consistency using AWS Glue DataBrew.
- **Data Type Adjustments**: The TotalUnits column's data type was changed to long to ensure proper numerical formatting using AWS Glue DataBrew.
- **Purpose**: These steps standardize the data, ensuring it's ready for accurate analysis and comparison across neighborhoods.

- Data is organized into a defined schema where each column has a specific name, data type, and order. This consistency ensures that datasets from different sources (e.g., Downtown and Grandview-Woodland) align perfectly when combined or compared.

Step 10: Data Pipeline Implementation

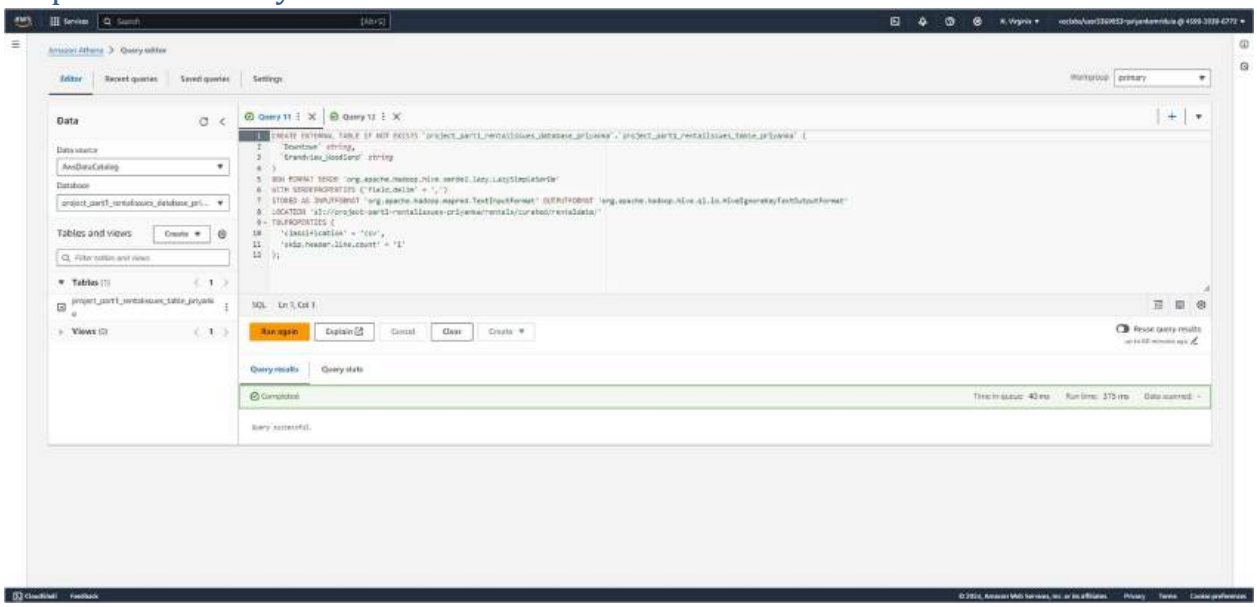


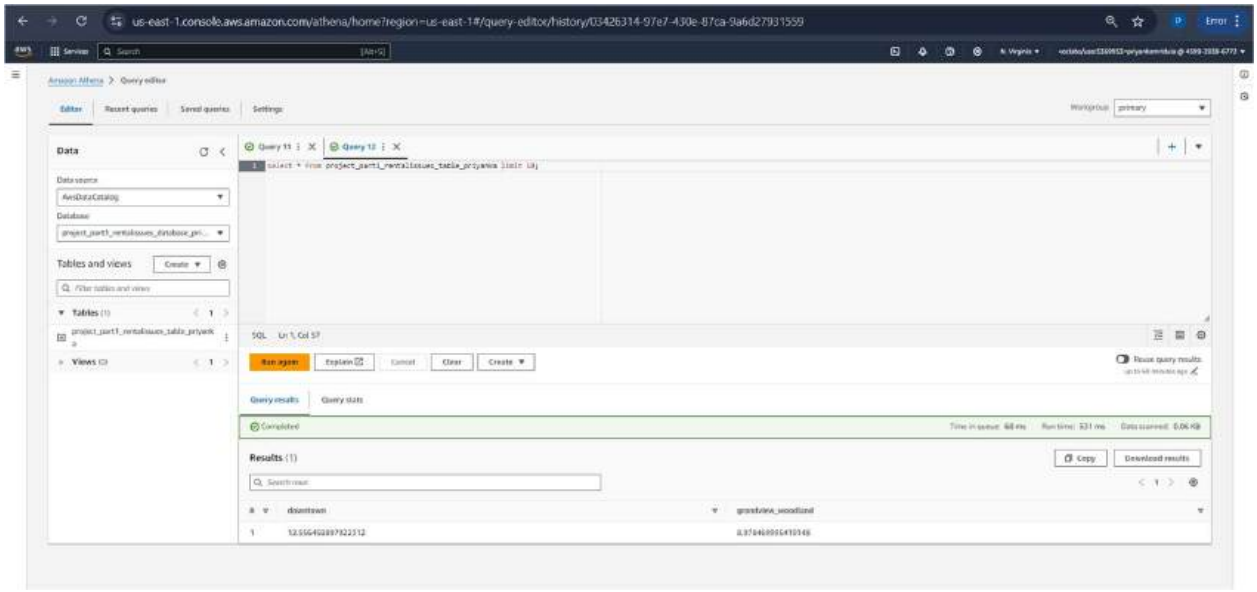
This AWS Glue ETL pipeline processes and combines data from Downtown and Grandview:

1. **Data Source**: Raw data from S3 buckets for Downtown and Grandview.
2. **Schema Transformation**: Initial schema alignment for both datasets.
3. **Aggregation**: Summarizing key metrics for each dataset.
4. **Dynamic Transformation**: Further data refinement with complex calculations.
5. **Data Join**: Merging Downtown and Grandview data into a unified dataset.
6. **Final Transformation**: Final schema adjustments.
7. **Data Target**: The processed data is stored in a curated S3 bucket.

The pipeline ensures the data is ready for accurate analysis and reporting.

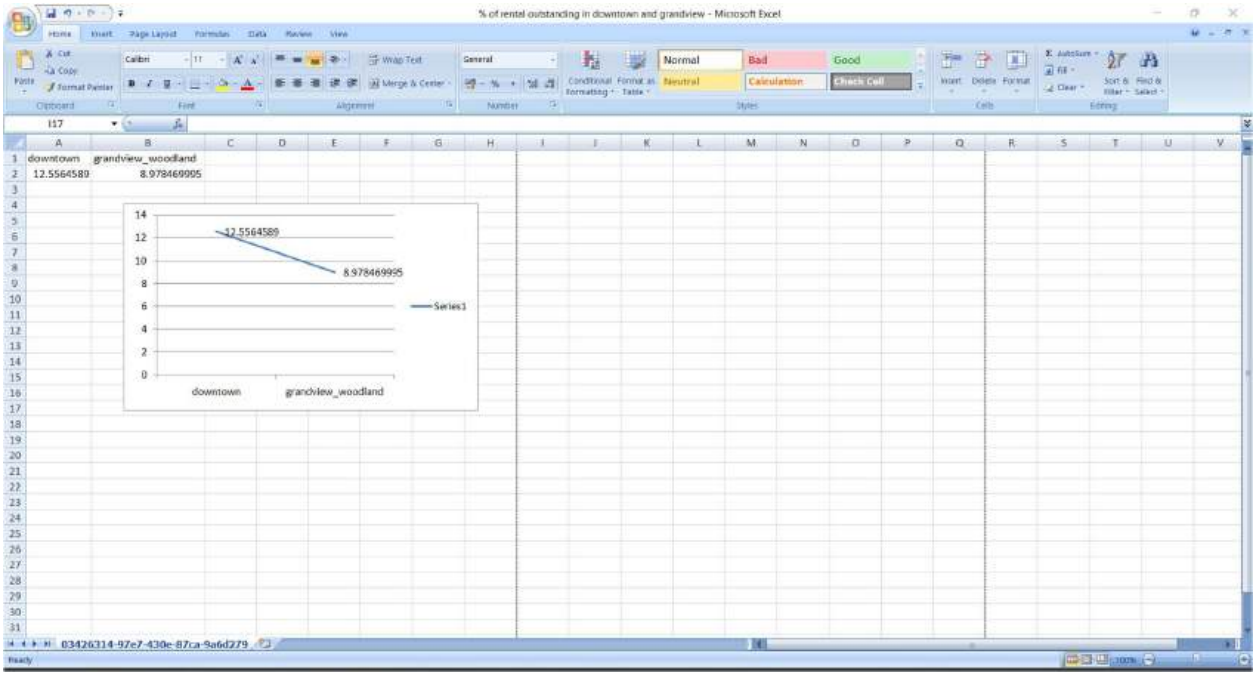
Step 11: Data Analysis

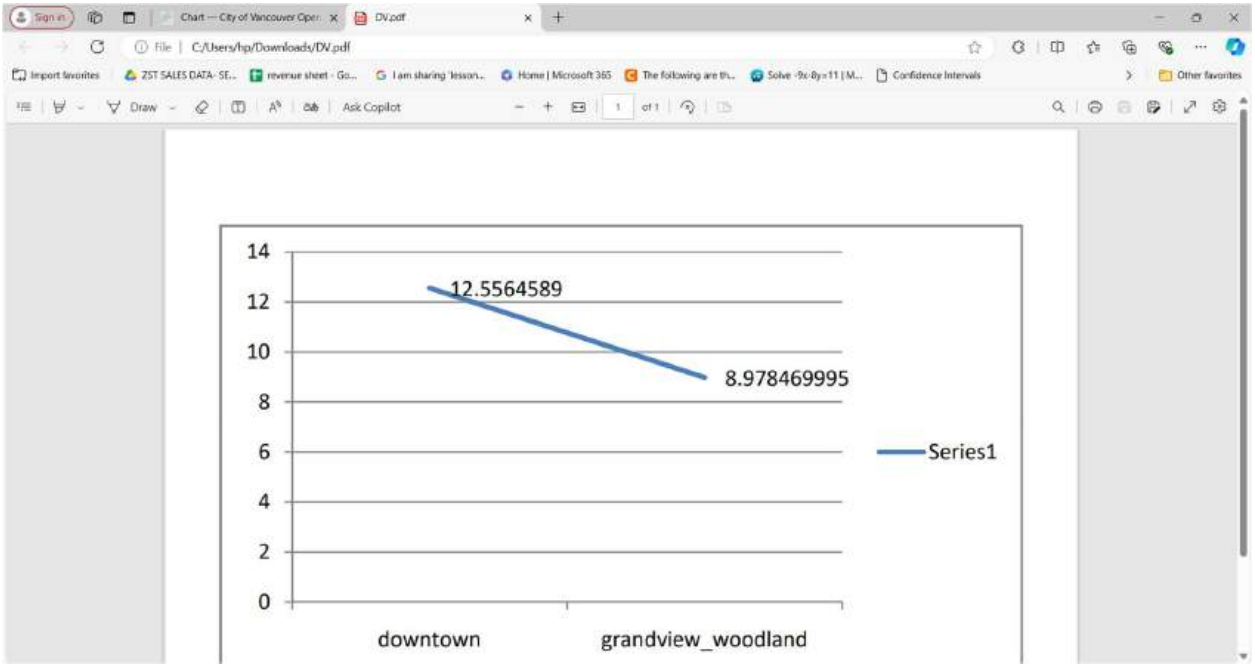




- **Create Table**: A table named project_part1_rentalissues_table_priyanka is created in Athena. It links to data stored in an S3 bucket, allowing SQL queries directly on the S3 data without moving it to a database.
- **Run Query**: A simple SQL query retrieves the first 10 rows from this table, confirming that the data is accessible, correctly linked, and properly structured.

Step 12: Data Visualization





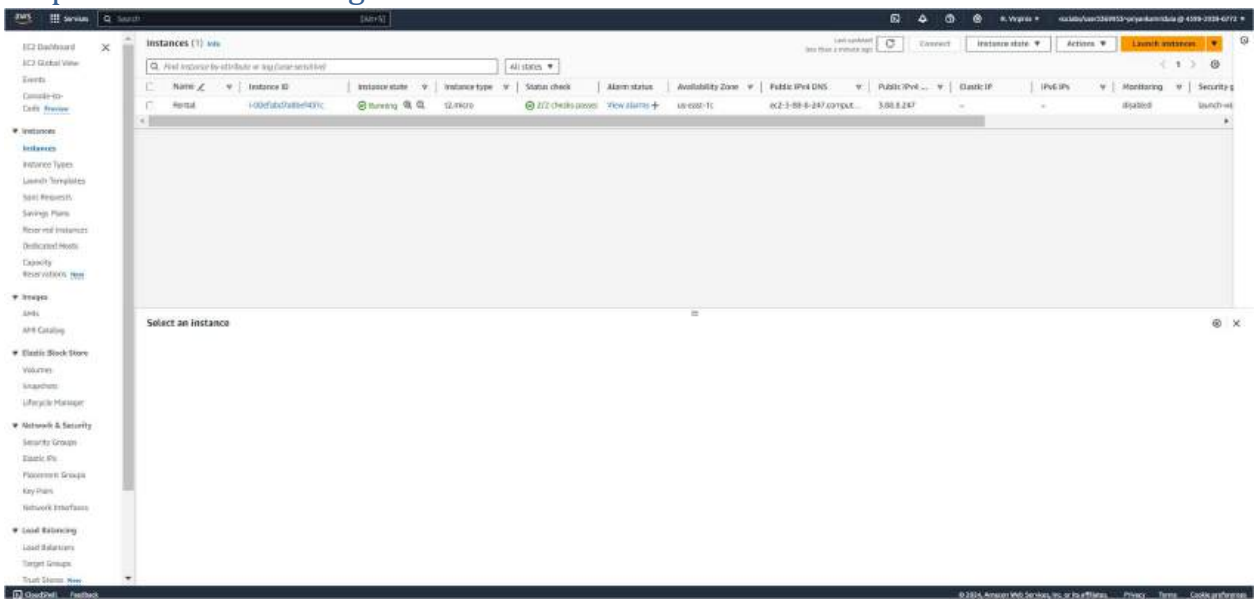
Excel Visualization:

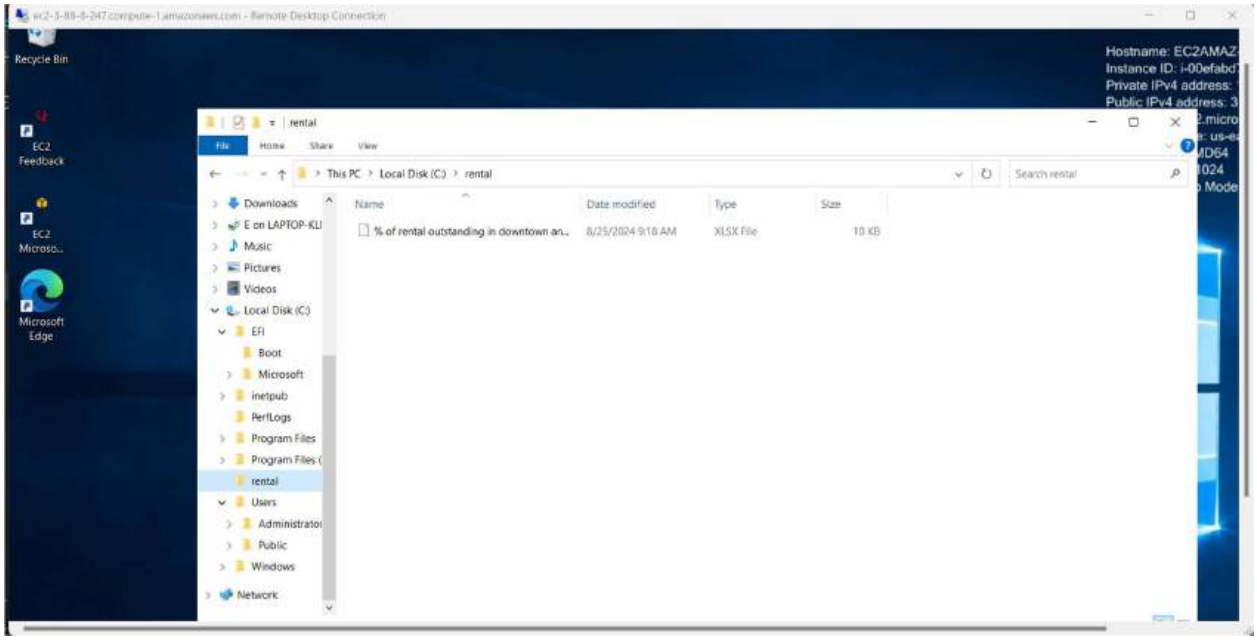
- The first screenshot shows an Excel sheet where the percentages of rental issues in Downtown and Grandview-Woodland are plotted on a simple line graph.
- The graph visually compares the percentages: Downtown has a higher percentage (12.56%) of rental issues than Grandview-Woodland (8.97%).

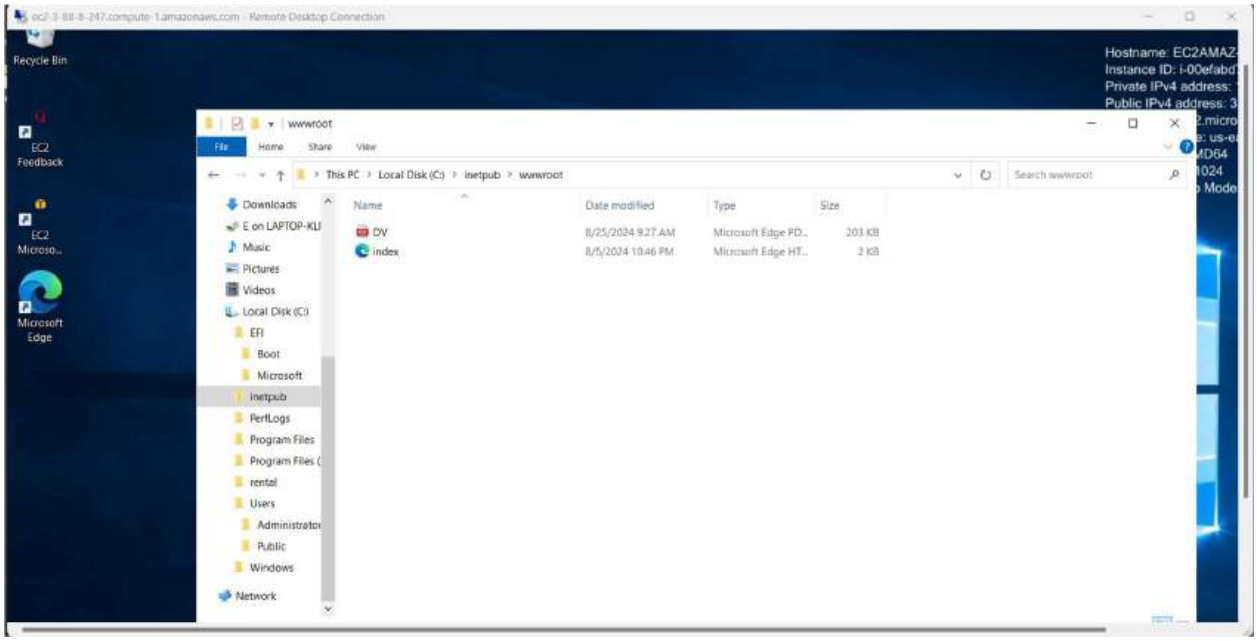
PDF Visualization:

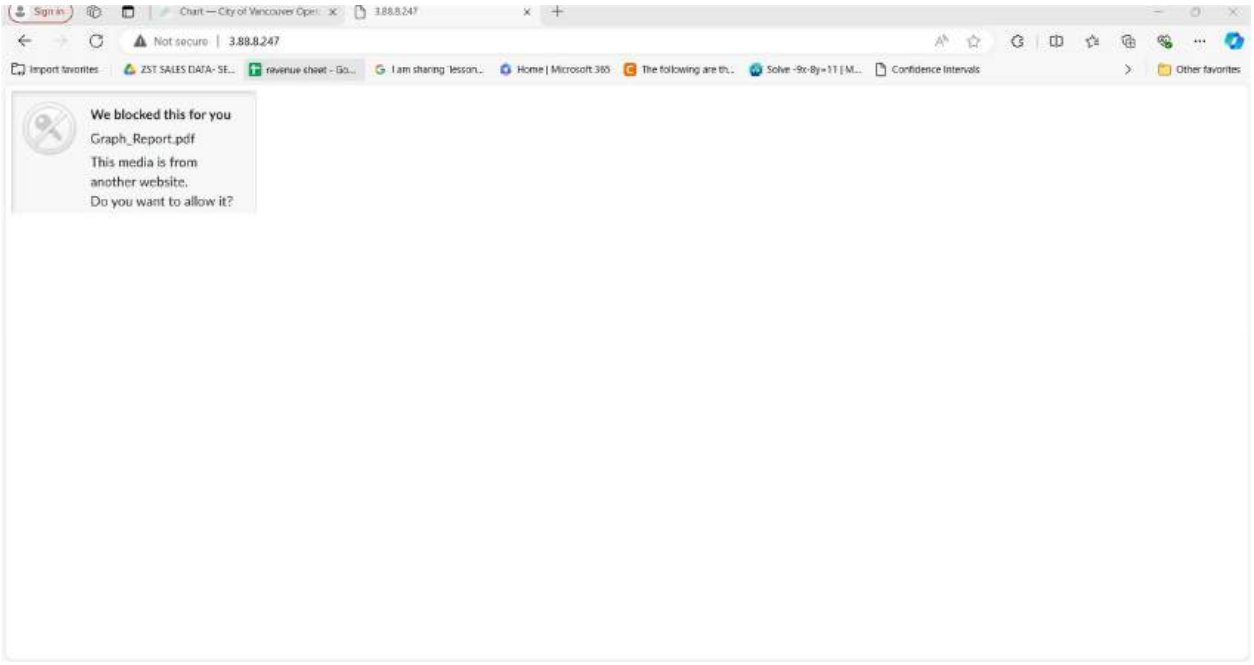
- The second screenshot shows the same graph exported to a PDF file, making the visualization shareable and easily accessible for reports or presentations.

Step 13: Data Publishing









The process of publishing data using an AWS EC2 instance.

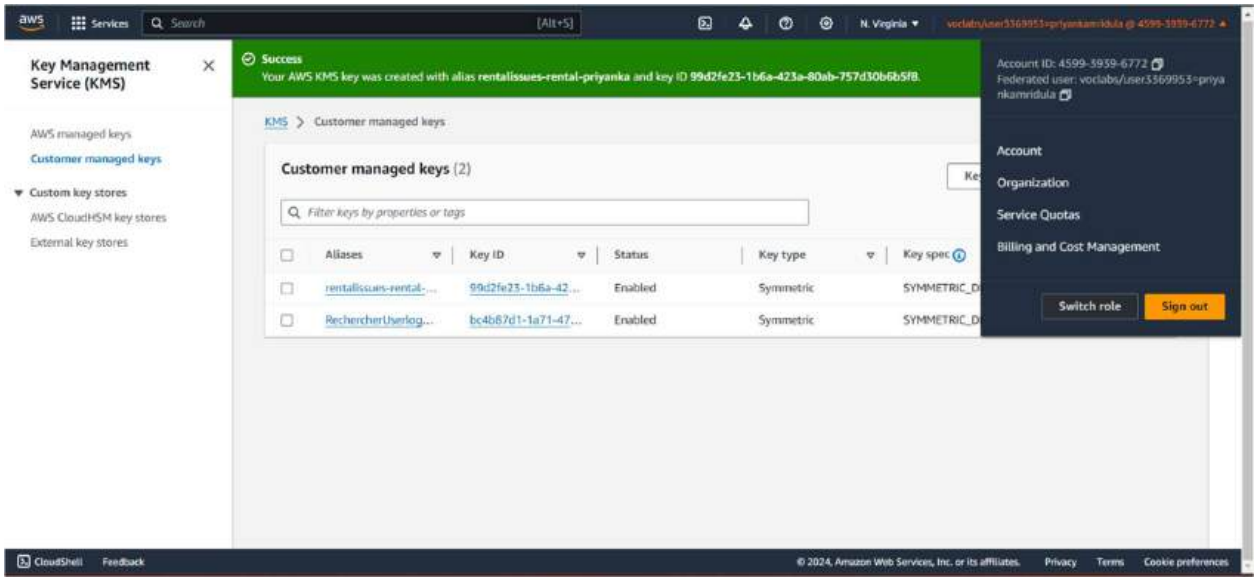
- **EC2 Setup**: A running t2.micro EC2 instance named "Rental" is used to host data files and visualizations.
- **File Storage**: Analysis results, including an Excel file and a PDF report, are stored on the EC2 instance.
- **Accessing Data**: The report is accessed via the EC2 instance's public IP, making it available for sharing through a web browser.

This setup allows easy hosting and sharing of data analysis results, making them accessible to others via the web.

Step 15: Data Protection

1. AWS Key Management Service (KMS) Configuration

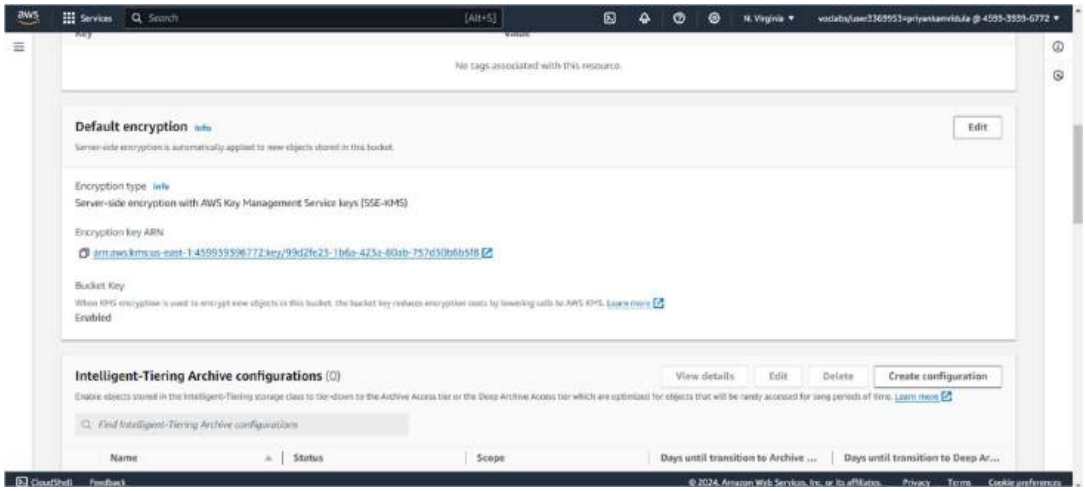
This is the successful creation of a KMS key labeled rentalissues-rental-priyanka. This key is symmetric and used for encryption and decryption of data stored in S3 buckets. By using KMS, you add an additional layer of security for sensitive data, ensuring it is encrypted at rest using a managed key that provides control over access.



Screenshot 1: AWS Key Management Service (KMS) Configuration

2. S3 Bucket Encryption with KMS

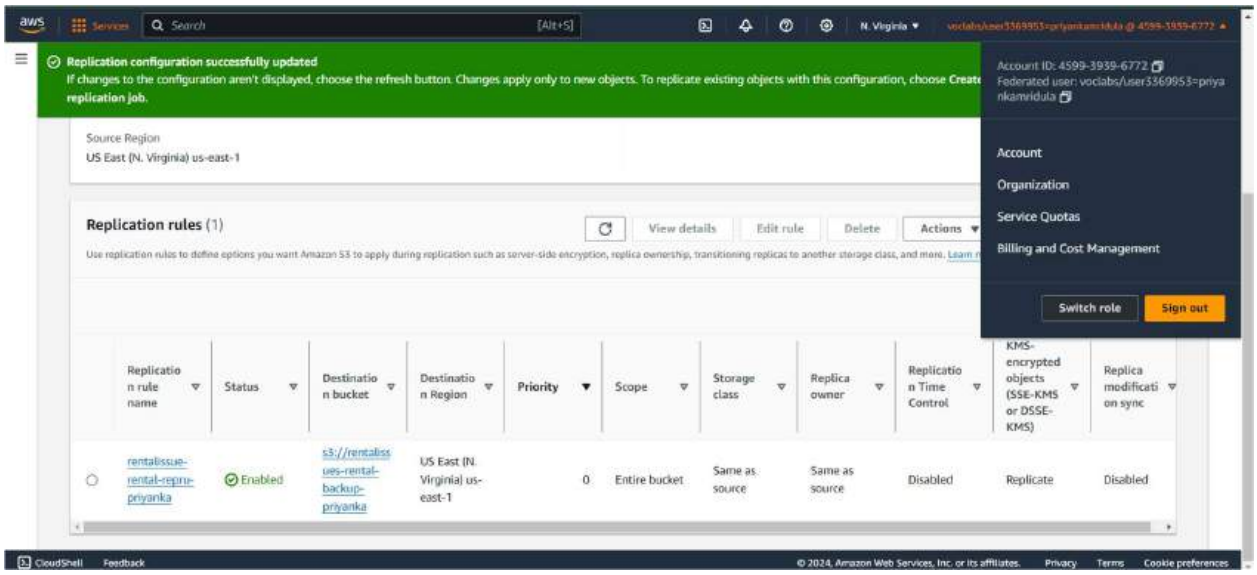
The S3 bucket in this screenshot has server-side encryption (SSE-KMS) enabled using the KMS key that was previously created. This ensures that all objects uploaded to this bucket are automatically encrypted. The bucket key option is also enabled, which reduces the cost of encryption requests by lowering the number of KMS calls.



Screenshot 2: S3 Bucket Encryption with KMS

3.Replication Rules Setup

This is for the configuration of a replication rule for the bucket. The replication rule rentalissue-rental-rep-priyanka is enabled to replicate data from the primary bucket rentalissues-rental-priyanka to the backup bucket rentalissues-rental-backup-priyanka located in the same AWS region. Replication ensures data redundancy and enhances disaster recovery capabilities by automatically copying objects from the primary to the secondary bucket.



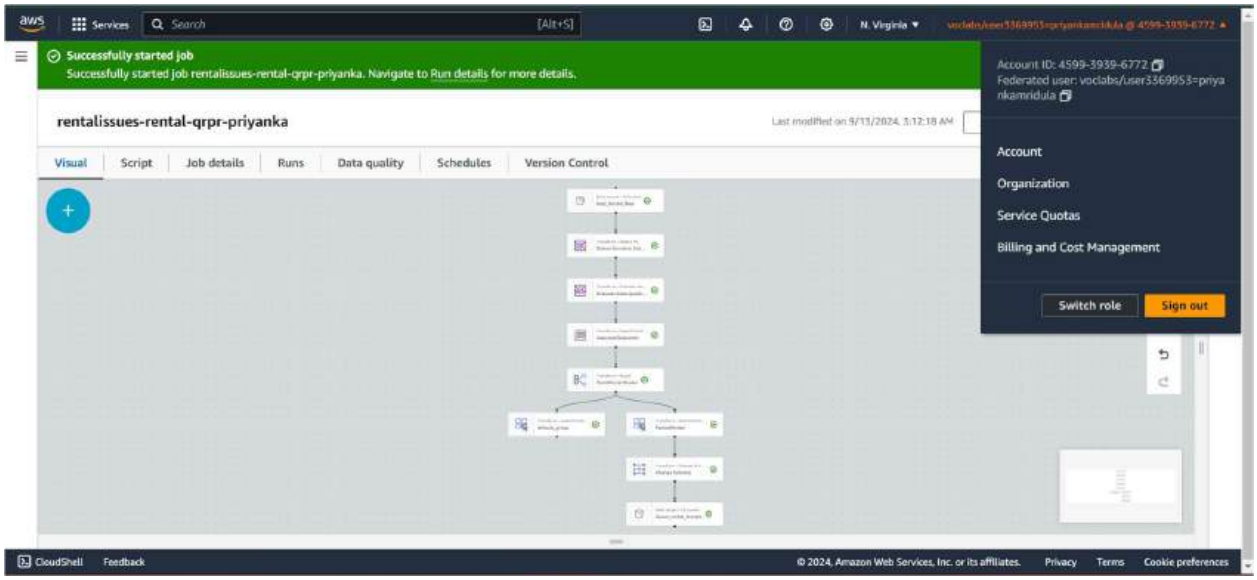
Screenshot 3: Replication Rules Setup

This setup strengthens data protection by encrypting data at rest and ensuring that it is available across different storage locations through replication. Both encryption and replication are critical for securing and managing sensitive data on AWS.

Step 16: Data Governance

1. Data Processing Job (Visual Workflow)

This displays the visual representation of a data processing job named rentalissues-rental-qprpr-priyanka in AWS Glue or another data orchestration tool. The workflow shows various nodes representing tasks like data transformation, extraction, and loading. These nodes are connected in sequence, indicating the flow of data through the pipeline.

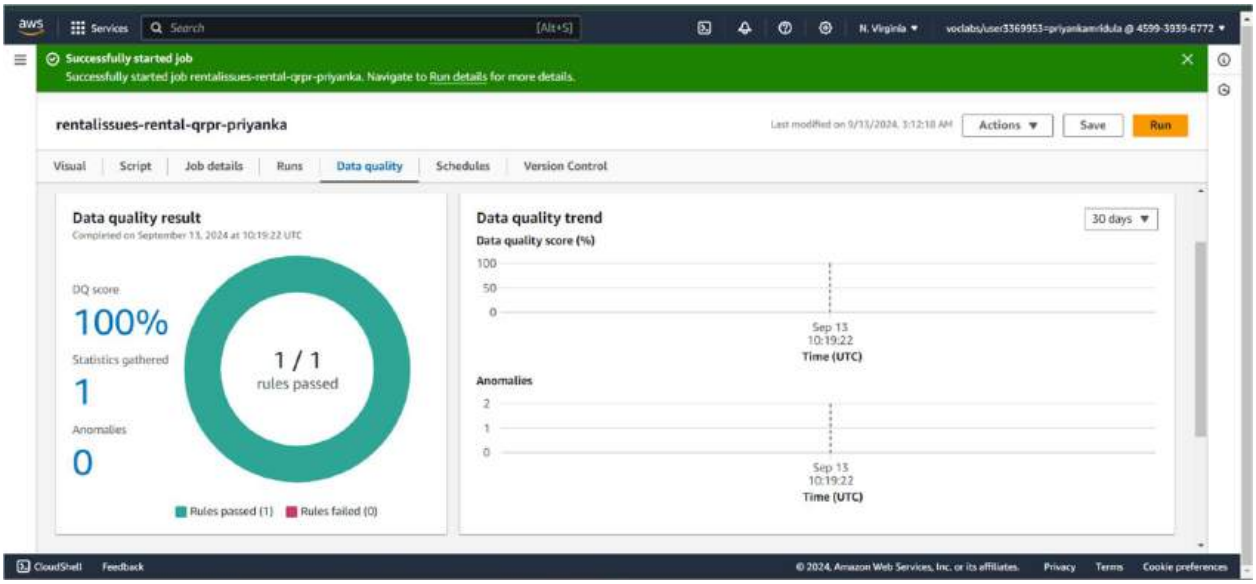


Screenshot 4: Data Processing Job Downtown(Visual Workflow)

The job successfully started, and the visual overview provides a clear understanding of each task performed, helping manage and monitor the processing of rental data for the downtown area.

2. Data Quality Check

This image shows the results of the data quality check for the job rentalissues-rental-qprpr-priyanka. The Data Quality Score is 100%, meaning that 1 out of 1 rule was passed, and no anomalies were detected. This confirms that the data processed for the downtown area meets the specified quality standards and is reliable for further use.



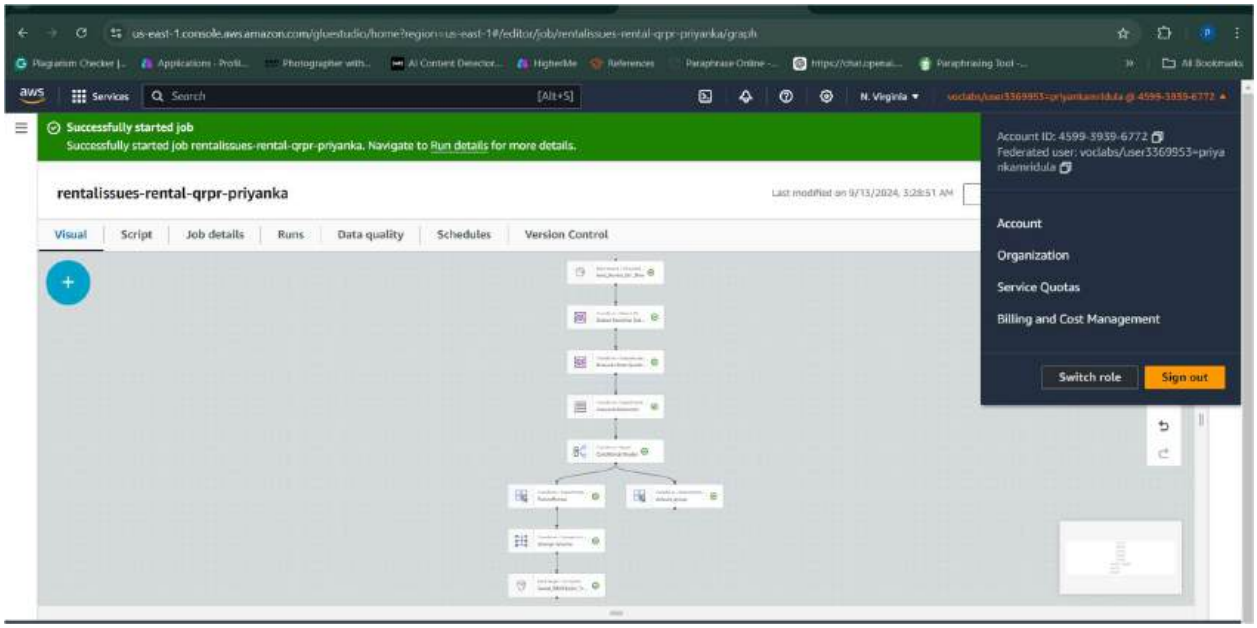
Screenshot 5: Data Quality Check downtown

The data quality trend shows a stable result for the current run, with no failed rules or detected issues, ensuring data integrity.

This demonstrates effective job execution and a thorough quality check process, ensuring the reliability of data for further analysis and decision-making in the downtown rental area context.

3. Data Processing Job (Visual Workflow)

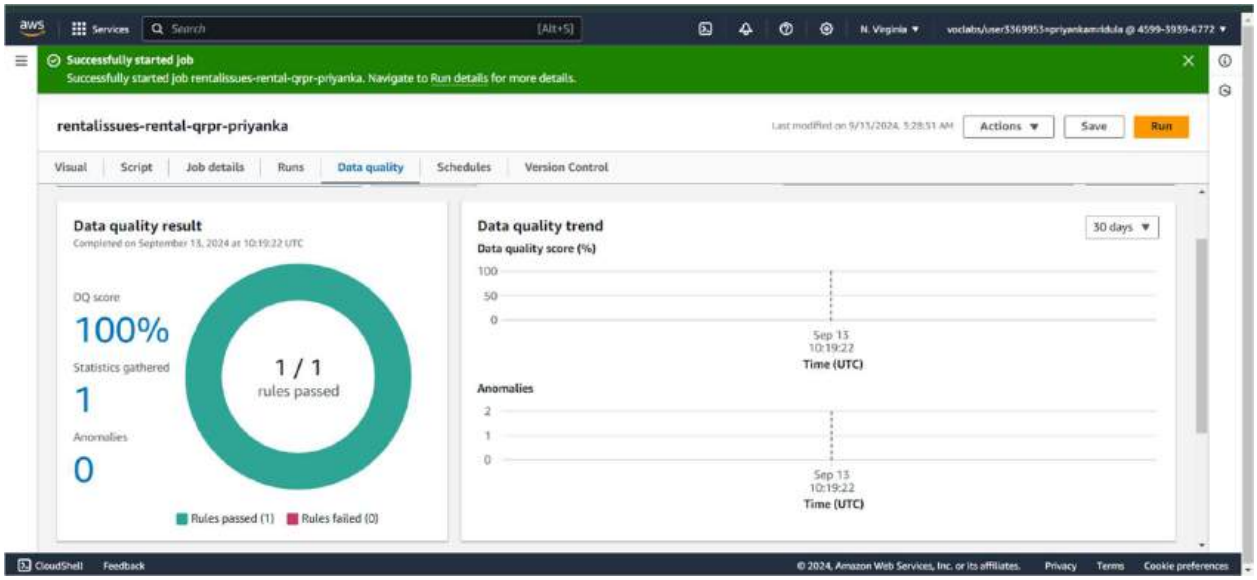
Similar to the earlier downtown setup, this image shows the visual job rentalissues-rental-qprpr-priyanka in AWS Glue or a similar service. It outlines the sequence of tasks for data extraction, transformation, and loading (ETL) related to rental issues for the Grandview-Woodland area. The visual workflow indicates that the job has started successfully.



Screenshot 6: Data Processing Job (Visual Workflow) Grandview-Woodland

4. Data Quality Check

The data quality result shows a 100% score, with 1 rule passed and no anomalies detected. This confirms that the data processed for the Grandview-Woodland area is accurate and meets the set quality standards, ensuring its integrity for further analysis and reporting.



Screenshot 7: Data Quality Check Grandview-Woodland

This setup is essential for ensuring that the rental data in the Grandview-Woodland area is processed correctly and adheres to the expected quality, much like the previous job for the downtown area.

Step 17: Data Monitoring

1. CloudWatch Dashboard for Monitoring

This is a CloudWatch dashboard named Rentalissue-rental-dashboard-priyanka. It displays:

Estimated Charges: A graph tracking the cost associated with AWS services used in the project over time. It shows a spike in charges around August, providing insight into the cost impact of data processing and storage activities.

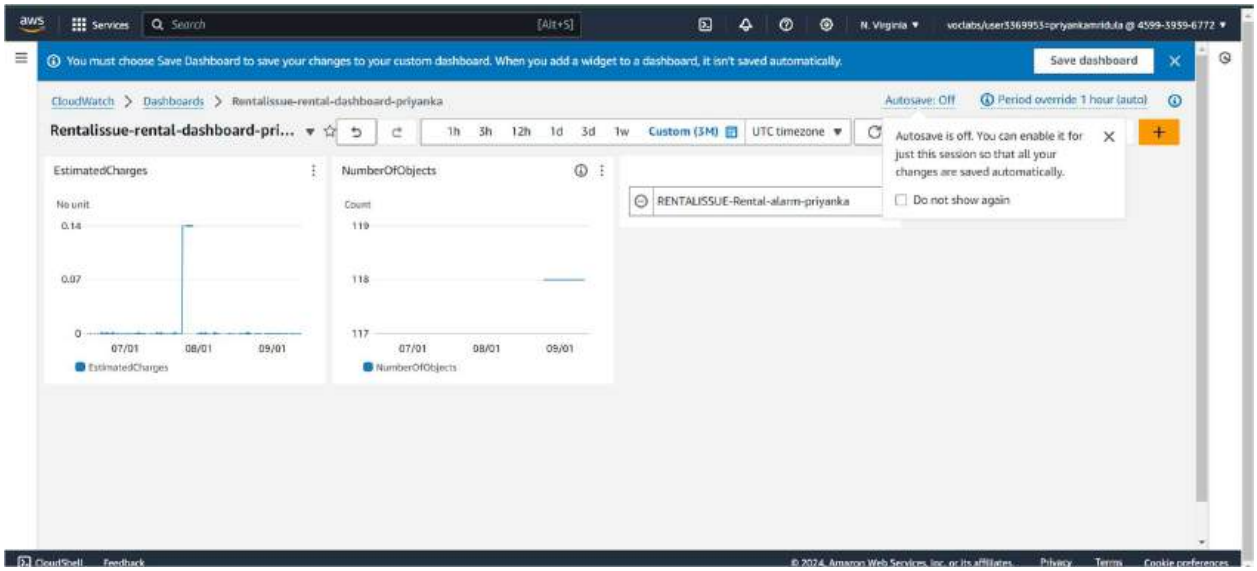
Number of Objects: This graph shows the number of objects in the monitored S3 bucket, providing a real-time count of the files stored, which is essential for tracking storage usage and identifying any unexpected changes in data volume.

2. CloudWatch Alarm Configuration:

Metric Tracked: The Number of Objects in the bucket is monitored continuously. The metric counts how many files or objects are stored in the S3 bucket at any given time.

Threshold: The alarm is configured to trigger if the Number of Objects exceeds 150. This threshold ensures that if the object count surpasses this limit, an alert is raised, possibly indicating unexpected data inflows or issues with data retention.

Action Taken: Once the threshold is crossed, the alarm automatically sends an email notification to the designated recipients. This email alert is critical for allowing the team to respond quickly to unexpected events such as unplanned uploads, data overflows, or potential security risks.



Screenshot 8: CloudWatch Dashboard for Monitoring with alarm setup

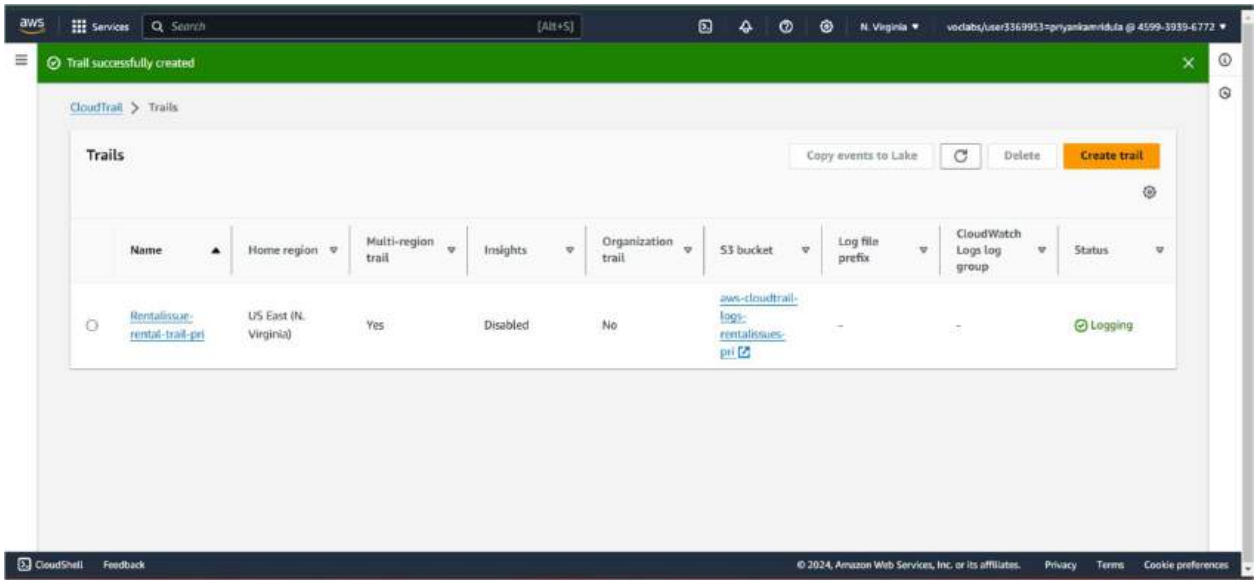
By setting this alarm, the system ensures proactive monitoring, and any deviation from the expected behavior (in terms of the number of objects) is promptly addressed through email notifications. This feature is key to maintaining the integrity of the S3 bucket and preventing potential issues like storage overload or unauthorized data uploads.

3. CloudTrail for Activity Logging

This is a CloudTrail named Rentalissue-rental-trail-pri, which has been successfully created for the project. The trail is configured to log events within the US East (N. Virginia) region, capturing all actions performed in the AWS environment. This trail is crucial for:

Tracking and Auditing: Every API call and user action in the AWS environment is logged. This helps in identifying who accessed the system, what actions they performed, and if any unexpected or unauthorized activities occurred.

Security and Compliance: The logs provide a basis for security audits and compliance with industry regulations. Logging to an S3 bucket ensures the data is retained for future analysis or audits.



Screenshot 9: CloudTrail for Activity Logging

Together, these tools ensure comprehensive monitoring of the AWS infrastructure, both in terms of system performance, cost, and security. The CloudWatch dashboard helps monitor real-time metrics, while CloudTrail provides detailed logs for auditing and security oversight.