**AWS Data Analytic Platform for The City of Vancouver – Group 5**

Aminat Eniola Bakare (2242556)

University Canada West  
BUSI 653(HBD-FALL-24-02)   
Cloud Computing TechnologiesDr. Mahmood Mortazavi Dehkordi November, 2024

# **Introduction**

Creating a Data Analytics Platform (DAP) for the City of Vancouver's **Issued Building Permits Dataset** is a transformative step toward enhancing urban administration and improving citizens' quality of life. The platform will enable the city to process large volumes of permit data in real time, delivering insights to optimize workflows and decision-making. By analyzing trends in construction, renovation, and development, it supports better urban planning, resource allocation, and sustainable development.

Key components of the platform include **data collection** from the Issued Building Permits dataset, spanning 501 records from Downtown Vancouver between March 15, 2024, and November 20, 2024. Data profiling ensures the integrity, structure, and completeness of fields such as permit types, processing times, and project values. **Data cleaning** addresses missing or inconsistent entries, improving dataset usability.

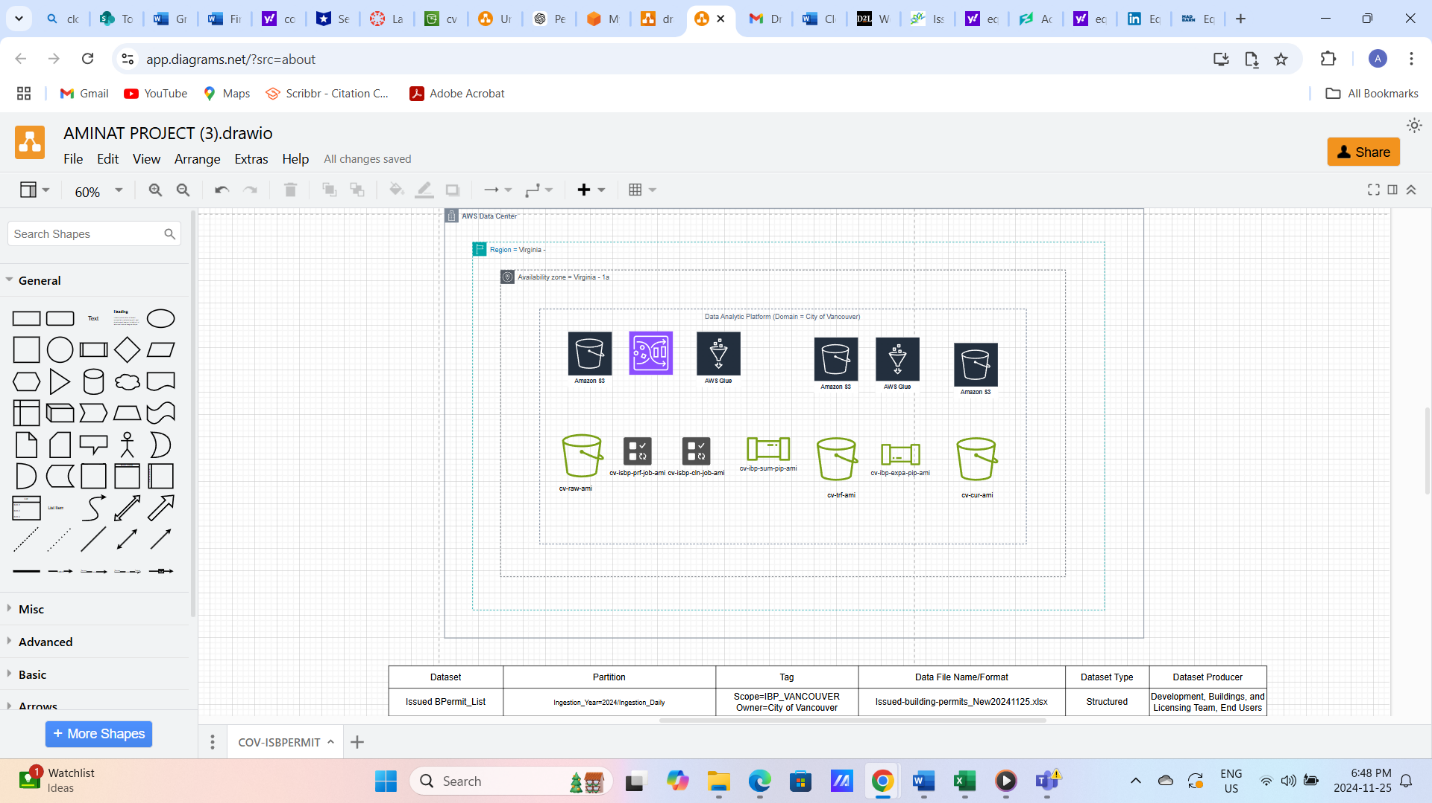
The **data pipeline design** outlines workflows for data ingestion, transformation, and aggregation. Metrics such as total permits issued, average processing times, and project value trends are calculated to deliver structured outputs for analysis. **Data enrichment** enhances analytical value by linking additional features, such as applicant interactions and project classifications. Strong **data protection policies** safeguard sensitive information, while **data governance** ensures compliance and accountability throughout the platform lifecycle.

This DAP enables Vancouver to generate actionable insights for sustainable urban governance. By leveraging the analyzed data, officials can streamline permitting processes, enhance collaboration, and make informed decisions that improve municipal services and sustainability efforts, ultimately benefiting the city's residents.

# **DAP Design and Implementation**

This transition is crucial for analyzing trends in construction, renovation, and development projects using the Issued Building Permits dataset. Spanning from 2017 onward, the dataset, enabling detailed analysis of permit issuance trends.

The **objective** is to design and implement a DAP using Amazon web services to support descriptive and exploratory analysis, enabling the City to optimize workflows and make data-driven decisions particularly for selected Dataset (Issued Building Permits). Additionally, calculate the platform's monthly AWS operational costs to ensure cost-effectiveness.

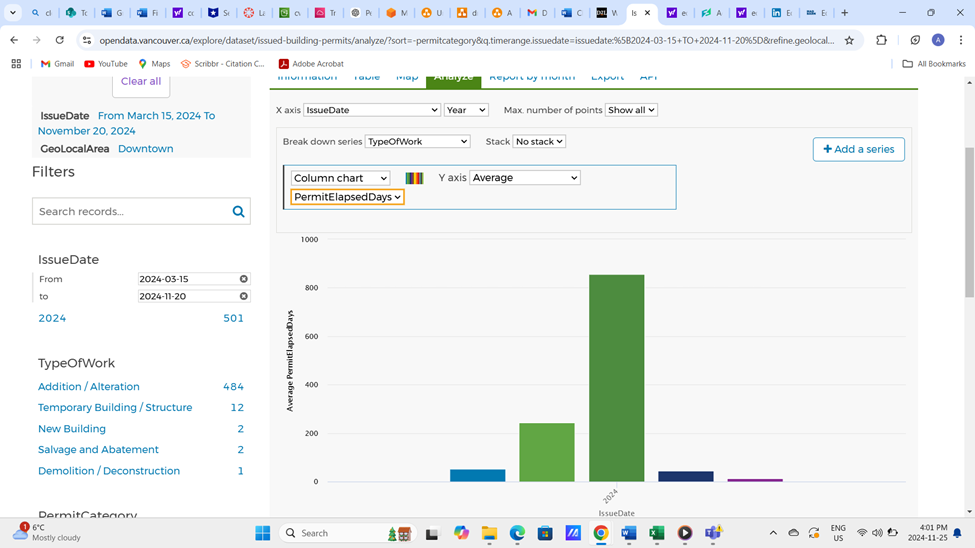


**Figure 1: Data Analytics platform for the City of Vancouver**

**Descriptive Analysis**

**Business Question -For permits issued in Downtown in 2024, what is the average processing time, for each Type Of Work?**

**Figure 2**



*Note: From City of Vancouver Data portal*

[https://opendata.vancouver.ca/explore/dataset/open-data-change-log/log/?disjunctive.datasets&amp;sort=logdate&amp;refine.datasetids=issued-building-permits](https://opendata.vancouver.ca/explore/dataset/open-data-change-log/log/?disjunctive.datasets&sort=logdate&refine.datasetids=issued-building-permits)

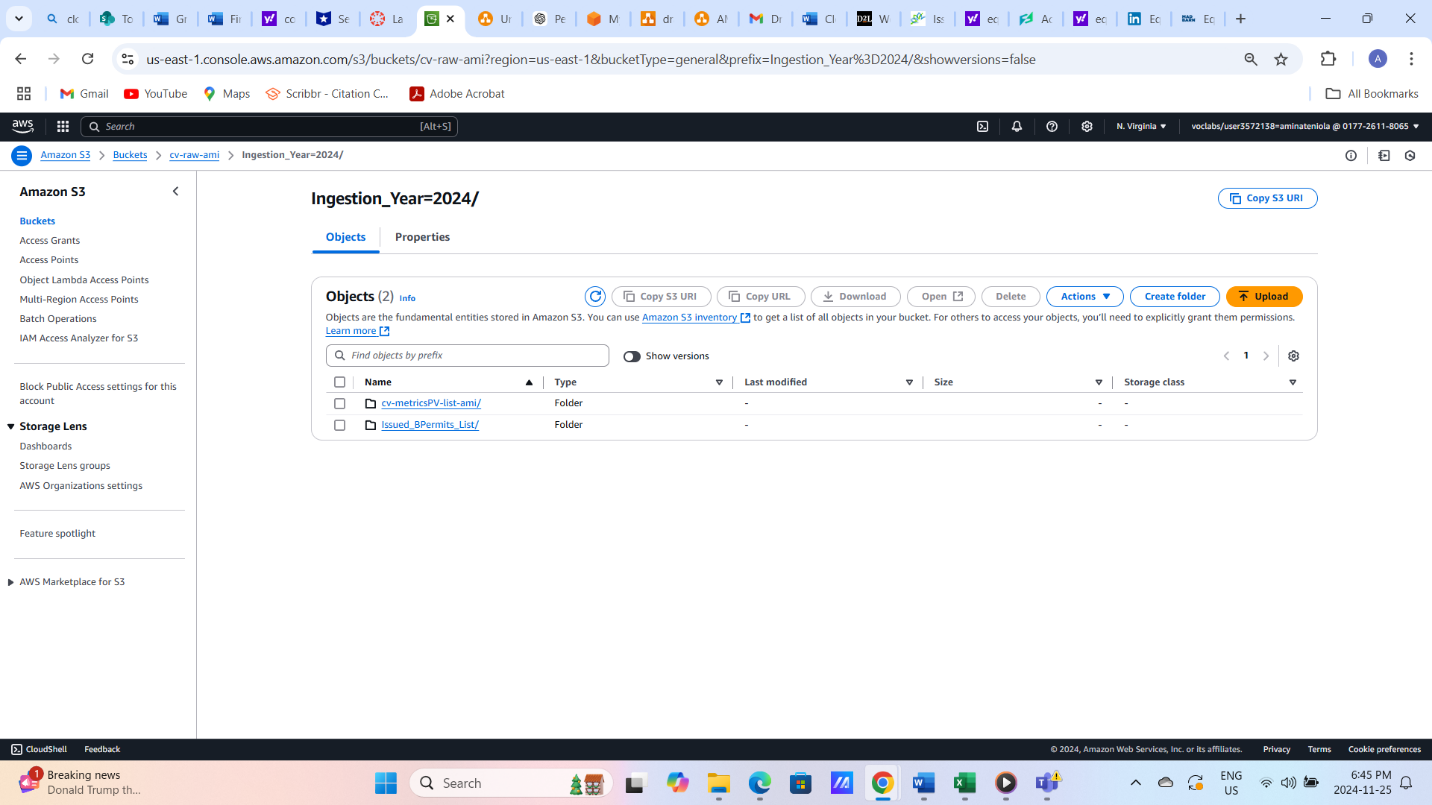
## **Step 1: Data Ingestion**

**Timeframe**: Filtered data from **March 15, 2024**, to **November 20, 2024**, from the Issued Building Permits dataset.

**Implementation**: Data was ingested in an .xlxs format to AWS S3 bucket for Storage and the storage class chosen was Standard storage because my data is ingested daily (Ingestion\_Year=2024/Issued\_BPermits\_List folder) into AWS Glue DataBrew for further processing and created CloudWatch for monitoring and logging services to track usage and job runs in the platform.

**Figure 3**

*Image of objects ingested in cv-raw-ami bucket - AWS S3*



*Note: Sreenshot from Aws S3*

## **Step 2: Data Profiling**

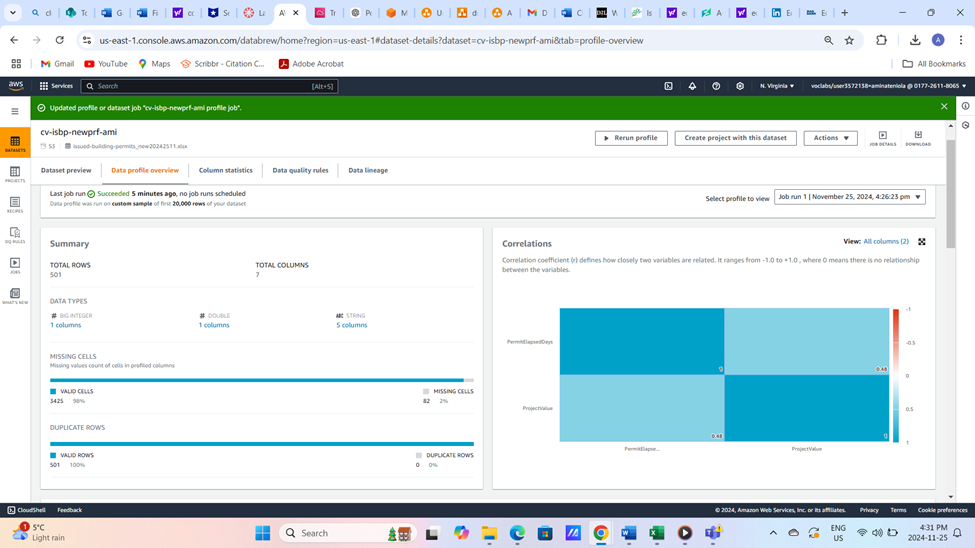
AWS Glue DataBrew profiling revealed 501 rows and 7 columns, including 1 integer, 1 double, and 5 string fields. Valid cells accounted for 98% (3425 values), while 2% were missing. No duplicate rows were found, ensuring data uniqueness. A correlation matrix highlighted key relationships, such as a moderate positive correlation (0.48) between **PermitElapsedDays** and ProjectValue, indicating that larger projects often take longer to process. Features like "Data Profile Overview" and "Column Statistics" enabled visualization of distributions and variability, aiding data cleaning and analysis. These insights provide a strong foundation for reliable descriptive and exploratory analyses.

**Implementation**: AWS Glue DataBrew was used to profile the data, ensuring accuracy and identifying issues in fields like PermitElapsedDays, Permit Number and Type Of Work.

**Reason**: Profiling ensures the dataset's quality and readiness for subsequent steps.

**Figure 4**

*Image of data profile Overview*



*Note. Screenshot from AWS Glue databrew*

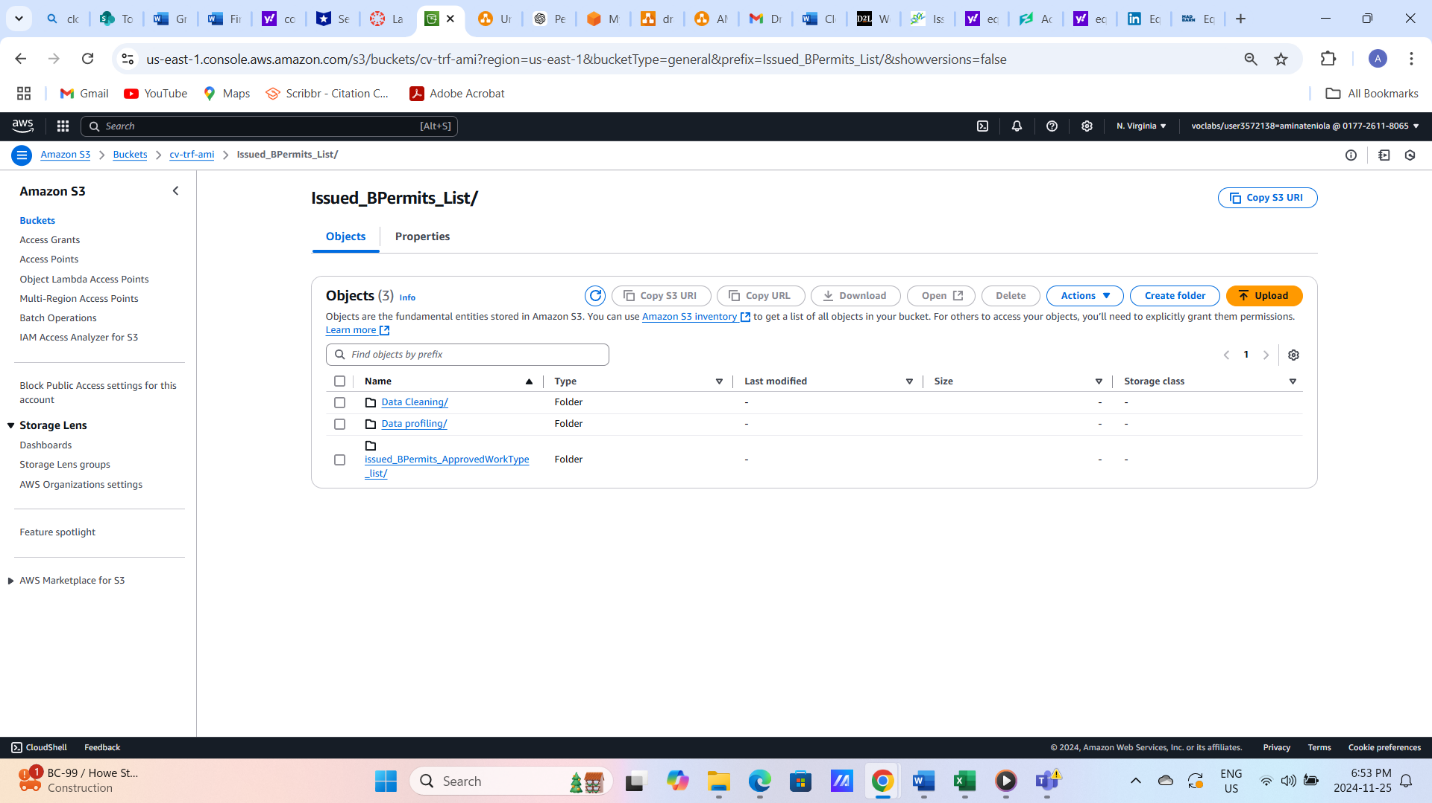
## **Step 3: Data Cleaning**

**Implementation**: Cleaned null and invalid values, ensuring PermitElapsedDays and ProjectValue have consistent, valid entries. While renaming and deletion was done, I filled the null spaces in permit category with frequently used selection Data Profiling and Cleaning are saved under the transformed bucket (cv-trf-ami)

**Reason of Action**: This ensures the analysis is based on accurate and complete data.

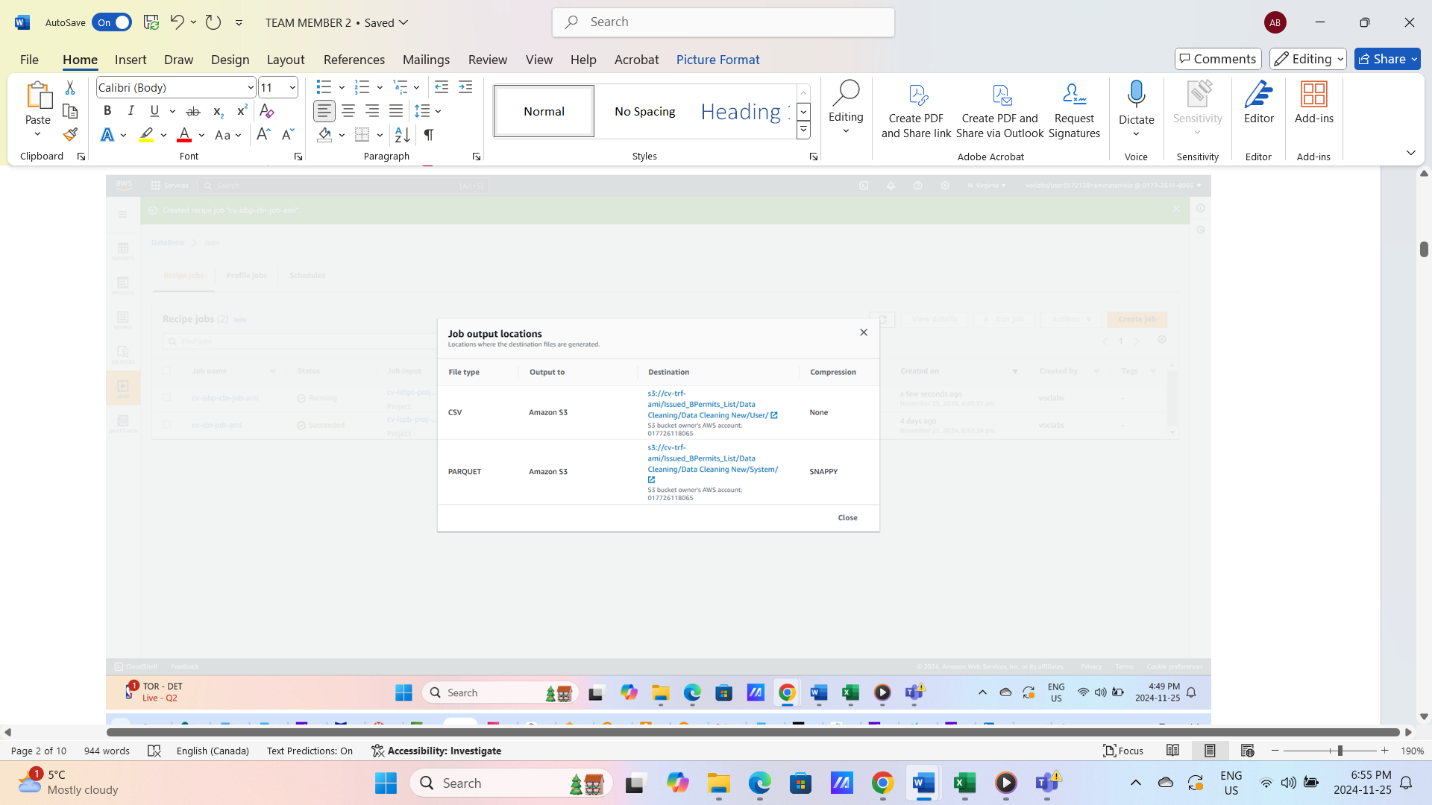
**Figure 5**

*Image of the Transformed bucket objects inside AWS S3*



**Figure 6**

*Image of data cleaning job outputs*



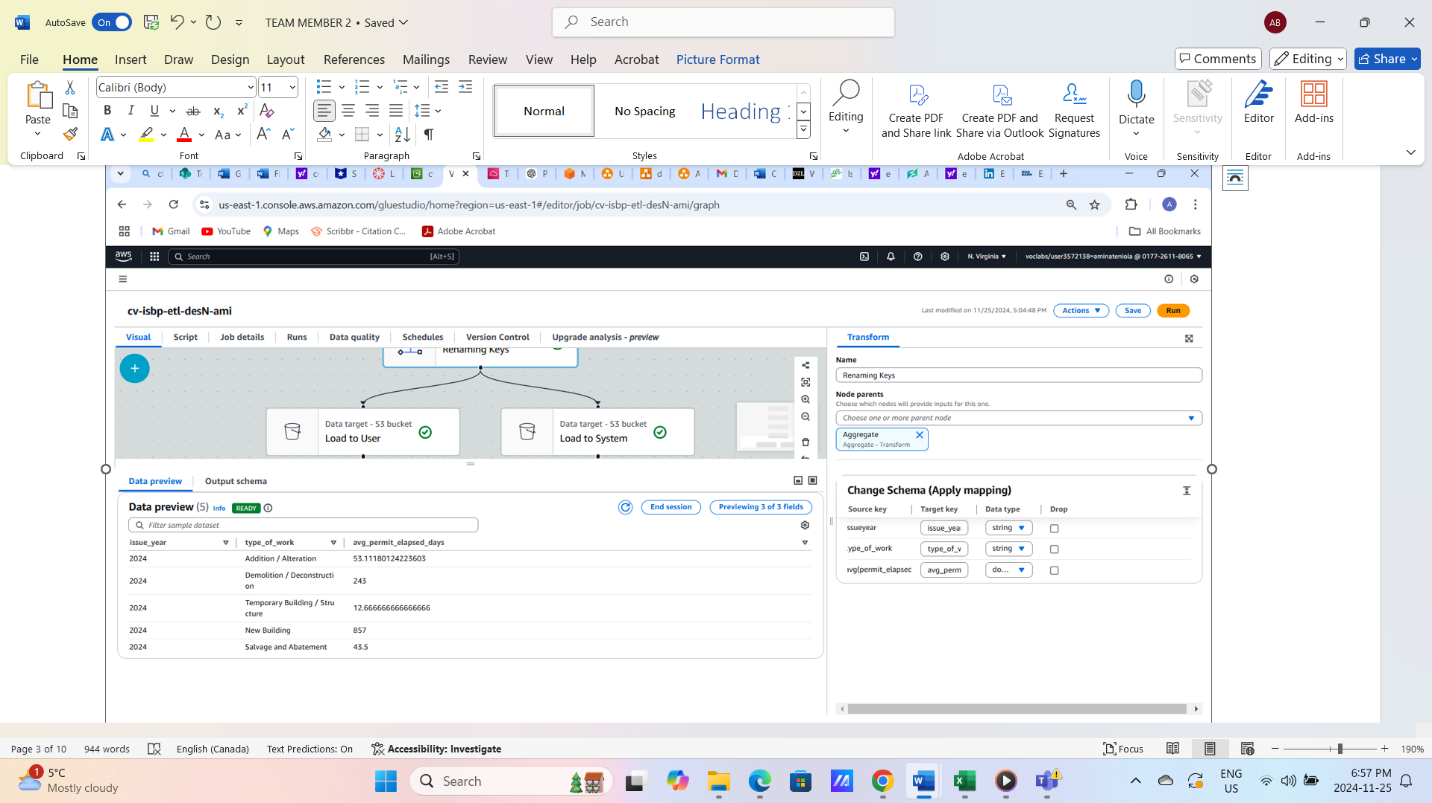
## **Step 4: Data Pipeline Design**

**Implementation**: Built a Glue ETL pipeline to automate data processing on Aws Glue

* Extracted the cleaned datasets, dropped some columns
* Then, Grouped data by Year and Type\_of\_Work
* Aggregated metrics: Average processing time (AVG(PermitElapsedDays)
* Renamed keys and loaded to System and User
* **Reason**: To generate structured outputs for descriptive analysis to answer the question.

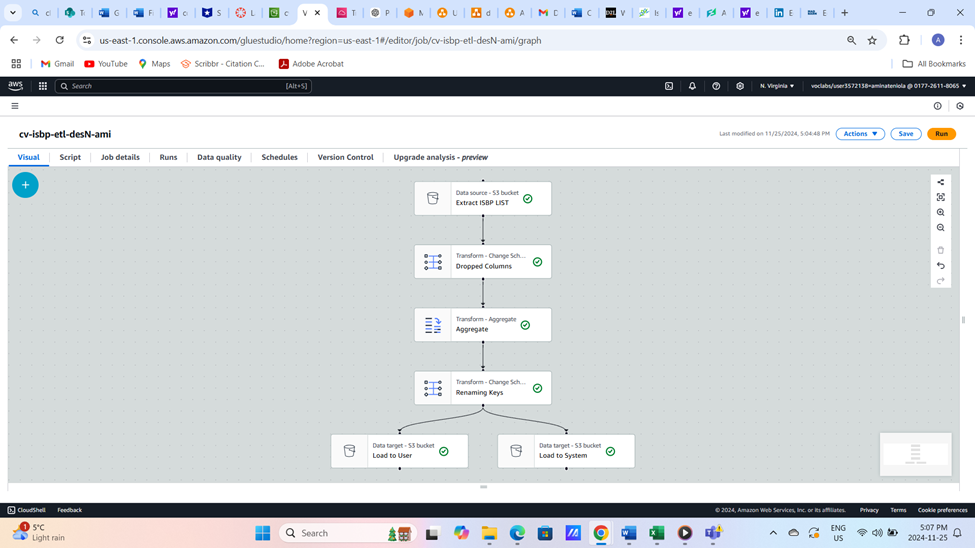
**Figure 7**

*Image of ETL pipeline for descriptive analysis output*



**Figure 8**

*Image of ETL Pipeline (Descriptive Analysis)*



*Note. Screenshot from AWS GLUE*

**Results for Descriptive Analysis**

* **Average Processing Time** - Metric: avg\_by\_permit\_elapsed\_days

Insights: Longest processing time: "New Building" (857 days), indicating significant complexity while Fastest: "Commercial - Temporary Building" (12.67 days) reflecting simpler approval processes for temporary structures.

## **Exploratory Analysis**

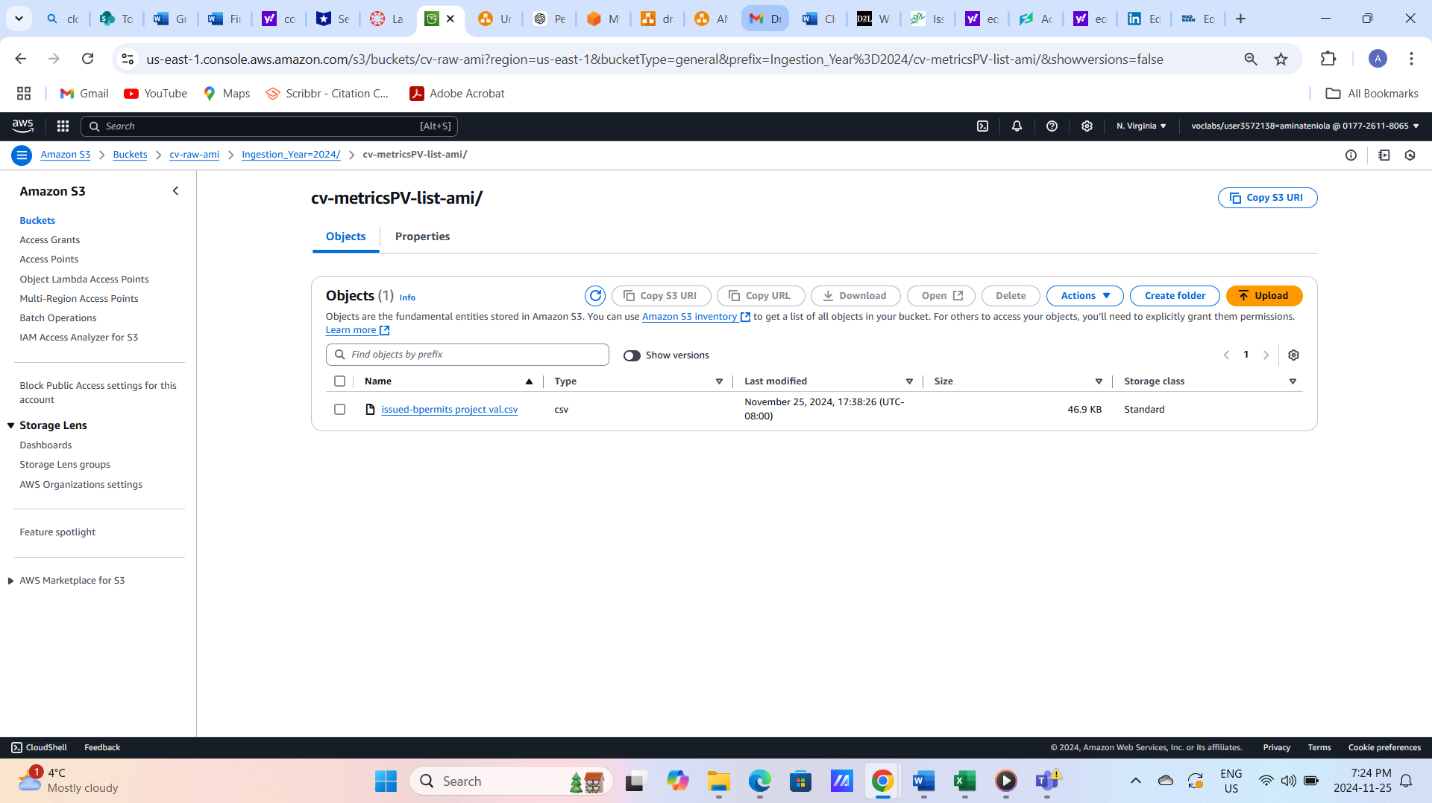
**Business Question**: " What are the average project values and processing times for different types of work issued in Downtown in 2024, and how do these correlate with the volume of permits issued?" **Exploratory analysis gives deeper insights. The new feature introduced was the Avg\_Project\_Value and Permit\_issued count to understand Applicant Behaviour.**

* **Data Ingestion**

**Implementation: Raw** [**issued-bpermits project val.csv**](https://us-east-1.console.aws.amazon.com/s3/object/cv-raw-ami?region=us-east-1&bucketType=general&prefix=Ingestion_Year%3D2024/cv-metricsPV-list-ami/issued-bpermits+project+val.csv) **Data was ingested into cv-raw-ami/Ingestion\_Year=2024/cv-metricsPV-list-ami/.**

**Reason:** To analyze permit issuance and processing time trends in conjunction with project values.

**Figure 9**



**Implementation**: For my analysis, I used the **cleaned dataset from the first analysis of** Downtown-issued permits in 2024 to focus on trends in project values, processing times, and permit volumes. A new feature, **Avg\_Project\_Value**, was introduced alongside **Permit\_Issued\_Count** to better understand permit issuance patterns and the financial dynamics of different work types. In this pipeline design process, I used AWS Glue to extract, transform, and load data efficiently for analysis. First, I extracted data from S3 buckets, pulling datasets like WorkType and Project value. These datasets were the foundation for my analysis.

**Reason**: Examining these trends helps identify whether Complex projects face processing delays or if simpler permits dominate issuance. By linking permit volumes and financial scope, we aim to optimize processes and allocate city resources effectively. **This ensures consistency and completeness for robust relationship analysis** and comparative analysis of permit issuance, project values, and processing time.

**Steps:**

**Renamed and dropped irrelevant keys for a cleaner dataset.**

**Grouped by work type and aggregated metrics:**

* Avg\_Project\_Value for financial trends.
* Avg\_Permit\_Elapsed\_Days for processing efficiency.
* SumDistinct(Permit\_Number) for permit volume trends.

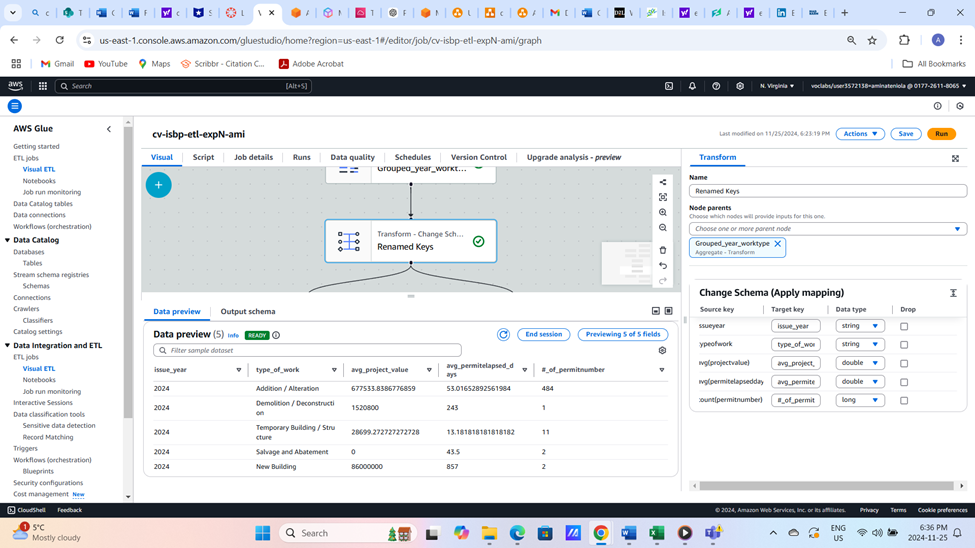
Finally, the transformed data was loaded into separate S3 buckets—one for system integration and another for user-facing analysis.

**Insights for Exploratory Analysis**

1. **Permit issued (Type of Work)**: Addition/Alteration dominates with 484 permits, moderate value ($677,533) and processing time (53 days). New Buildings, though rare, have the highest value ($86M) and longest processing time (857 days). Temporary Structures are quick (13 days) with low value ($28,699).
2. **Comparative Analysis**: High-value permits face longer processing times, while low-complexity permits like Temporary ones are faster and more frequent. Addition/Alteration balances high volume with moderate value and efficiency.

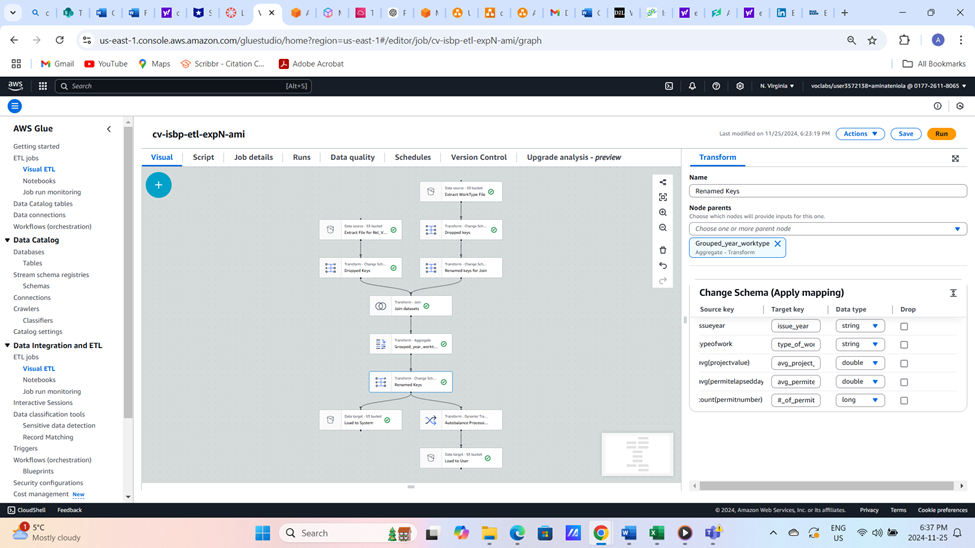
**Figure 10**

*Image of ETL pipeline output*



**Figure 11**

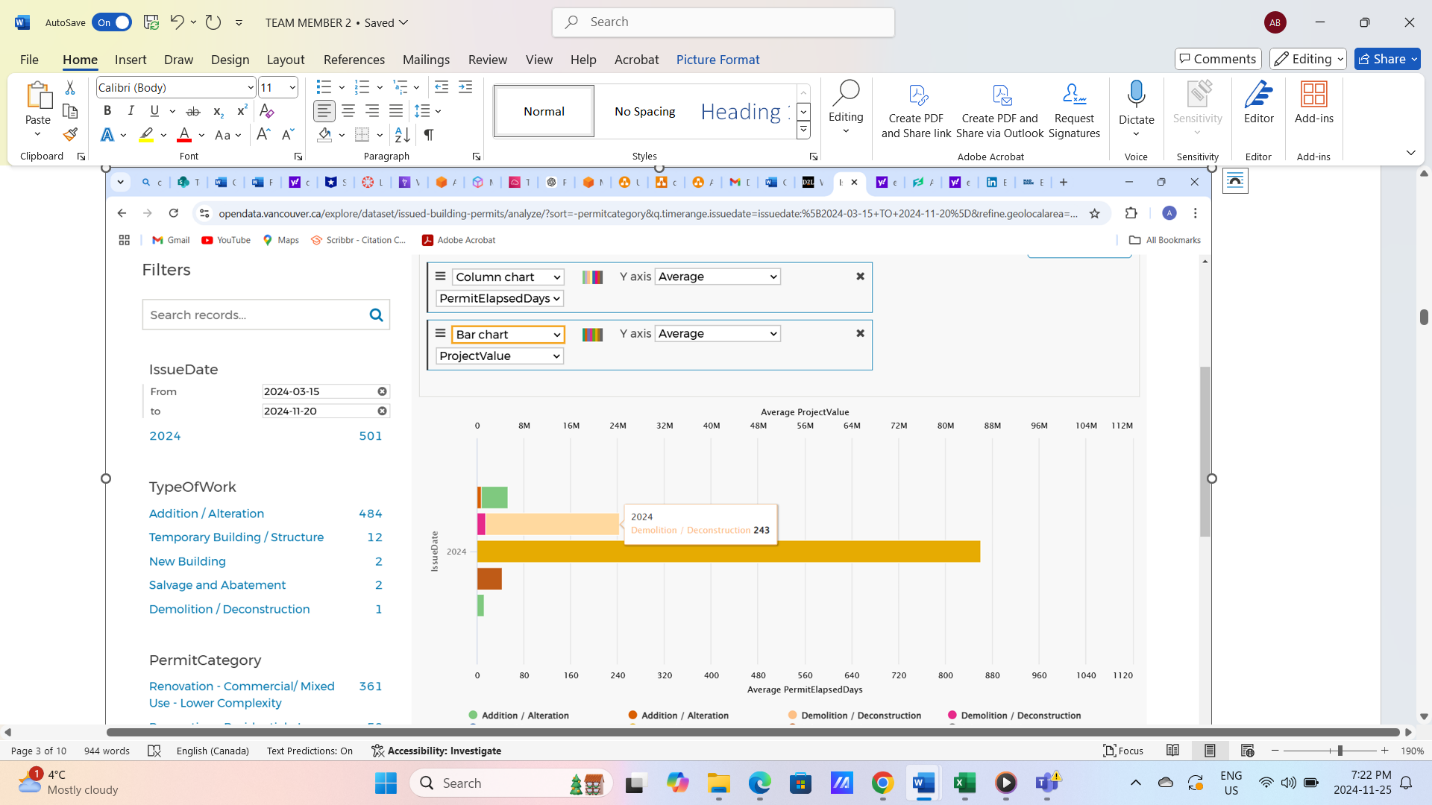
*Image of ETL pipeline for Exploratory Analysis*



*Note. Screenshots from AWS*

**Figure 12**

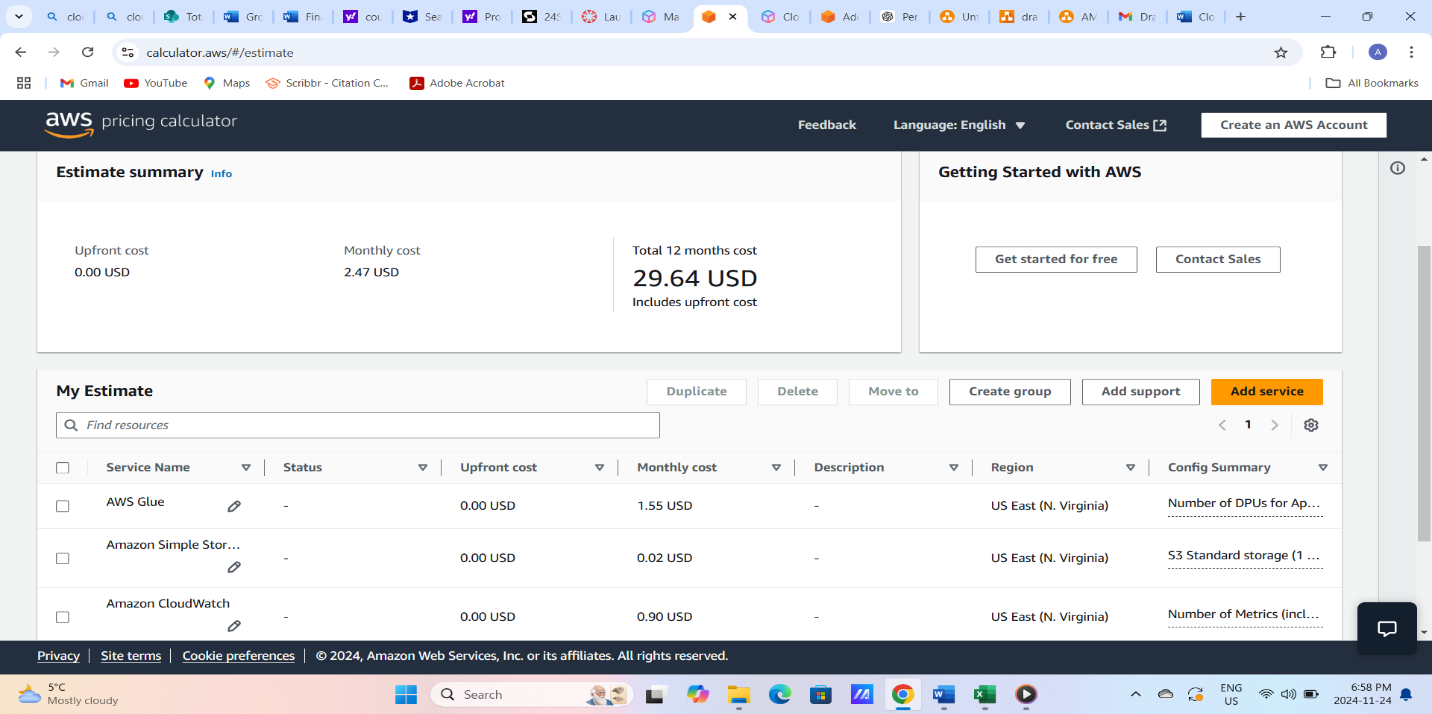
*Analysis of Permit Types by Average Project Value and Processing Time in Downtown*



*Note: From City of Vancouver Data portal*

**Figure 13**

*AWS Pricing Calculator*



AWS Pricing Calculator. (2024). *AWS Pricing Estimate for Cloud Services.* Retrieved November 23, 2024, from <https://calculator.aws/>

# **DAP Estimated Cost**

The estimated cost for this data analytics platform, utilizing **AWS Glue**, **Amazon S3**, and **Amazon CloudWatch**, totals approximately **$2.47 per month** or **$29.64 annually**. AWS Glue is the primary cost driver, accounting for **$1.55 per month** or **$18.60 annually**, due to its critical role in executing scalable ETL jobs and handling complex data transformations. Amazon S3 provides efficient storage for datasets at a minimal cost of **$0.02 per month** or **$0.24 annually**, covering PUT, COPY, POST, and GET requests. Meanwhile, Amazon CloudWatch, essential for monitoring and logging, contributes **$0.90 per month** or **$10.80 annually**. This cost structure reflects a balance between the high computational needs of AWS Glue and the budget-friendly storage capabilities of Amazon S3, ensuring an effective and cost-efficient platform for processing and analyzing building permit data.

# **Conclusion**

The Data Analytics Platform (DAP) for the City of Vancouver demonstrates the transformative potential of AWS services, particularly AWS Glue and Amazon S3, in optimizing urban governance. Focused on building permit data, the platform integrates descriptive and exploratory analysis to uncover trends in housing, urban development, and sustainability. By ensuring data quality, scalability, and cost-efficiency, it supports informed decision-making and resource optimization. The inclusion of cost analysis reflects a commitment to long-term reliability. This project highlights the power of data-driven insights to improve city planning and residents' quality of life, positioning Vancouver as a leader in smart city innovation.

# **References**

Amazon Web Services. (2024). AWS pricing calculator. AWS Pricing Calculator. <https://calculator.aws/#/estimate>