DATA ANALYTIC PLATFORM (DAP)

Exploratory Data Analysis

Project Description: This project focuses on conducting an Exploratory Data Analysis (EDA) of conflict resolution queries within the Human Resource Department. The goal is to uncover patterns, trends, and actionable insights by analyzing various attributes such as student names, gender, courses, and student IDs. The analysis aims to address interpersonal disputes, misunderstandings, and systemic issues affecting students in a university setting.

Project Title: HR department handling Conflict Resolution Queries: An Exploratory Data Analysis

Objective: To analyze and evaluate the conflict resolution process within the HR department by identifying key trends and relationships between formal and informal queries. Insights gained from the analysis will help address conflict-related challenges more effectively.

Dataset: The Conflict Resolution (CR) dataset includes formal and informal student queries related to HR conflict resolution. Key attributes in the dataset are:

- **Student ID**: Unique identifier for each student.
- **Query Type**: Specifies whether the query is formal or informal.
- **Date Query Issued**: The date the query was submitted.
- **Name**: Name of the student.
- **Resolution Status**: Indicates whether the query is resolved or unresolved.
- **Resolution Date**: The date on which the resolution was provided.

Methodology:

1- Data Collection and Preparation:
   - Inserted the csv file of conflict resolution dataset into the AWS S3 Buckets.
   - Perform initial data profiling and cleaning, which includes handling missing values, correcting data types, and renaming columns for clarity using AWS Glue databrew

2- Descriptive Statistics:
   - Generated summary statistics (e.g., means, percentages) for numerical attributes, including monthly resolution rates of formal and informal queries.

3- Data Visualization:

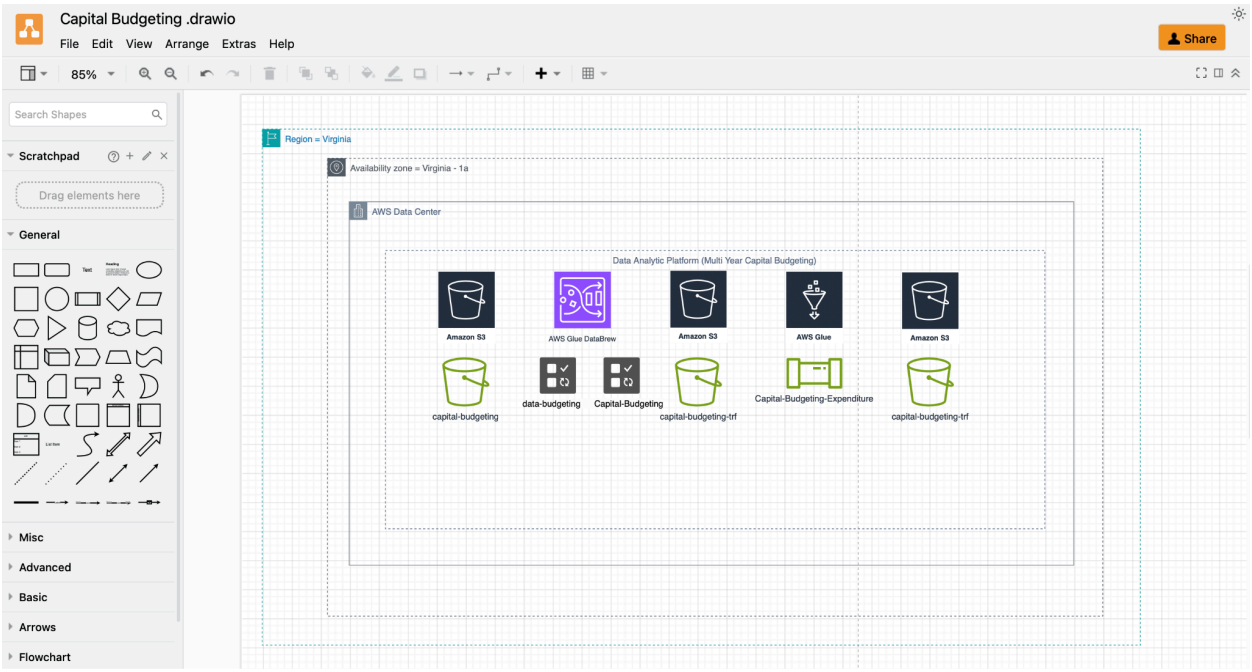
    - Create visualizations to illustrate key insights:
        - Used Draw.io for illustration of data platform.

4- Resolution Analysis:
    - Compare resolution status:
        - By Informal: How many informal queries have been resolved in a specific month and which department had the most queries.
        - By Formal: How many formal queries have been resolved in a specific month and which department had the most queries.
        - By Unresolved Status: How many formal and informal queries are still pending.
5- Insights and Findings:
    - Informal queries had a higher resolution rate than formal queries in September.
    - Specific departments received significantly higher volumes of conflict resolution queries, providing insight into potential areas for HR intervention.
6- Conclusion:
    - The analysis revealed distinct patterns and relationships between formal and informal queries and their resolution statuses. These findings can guide the HR department in optimizing conflict resolution strategies.

Tools and Technologies:

- **AWS Console**: For managing cloud resources.
- **AWS S3 Buckets**: To store and manage the dataset.
- **AWS Glue DataBrew**: For data profiling and cleaning.
- **AWS Glue**: For building and managing the ETL pipeline.
- **Draw.io**: For data platform illustrations.

This EDA project not only demonstrates your analytical and programming skills but also highlights your ability to derive meaningful insights from data, making it a valuable addition to your data analyst portfolio.



Descriptive Analysis

Project Description: The project focuses on preparing and validating datasets for the City of Vancouver's **Data Analytics Platform (DAP)**, aimed at analyzing multi-year capital budget allocations. Through **Data Wrangling** and **Data Quality Control**, the project ensures that the dataset is accurate, complete, and reliable for informed decision-making and resource prioritization.

Project Title: Descriptive Analysis of Multi-Year Capital Budget Allocations in the City of Vancouver

Background: The City of Vancouver's data analytics platform (DAP) project handles multi-year capital budget datasets. These datasets provide insights into budget allocations across various service categories, aiding in resource prioritization and financial planning. However, inconsistent formats, incomplete records, and fragmented information necessitate robust data wrangling and quality control measures to support effective governance and sustainability initiatives.

Objective: The primary goal of this project was to perform a descriptive analysis of the City of Vancouver's capital budget allocation and expenditure data. This analysis aimed to identify the service categories with the highest budget allocations and understand patterns in financial planning, aiding decision-making in urban development, sustainability, and resource management.

Dataset: The dataset consisted of the City of Vancouver's 2022 Multi-Year Capital Budget and Capital Expenditure data, which included:

- **Service Category**: Categories of services (e.g., Affordable Housing).
- **Multi-Year Budget Allocations**: Total budget planned over multiple years.
- **Annual Expenditure for 2023**: Forecasted budget for the year 2023.
- **Additional Features**: Information on previously approved budgets, current allocations, and future expenditure forecasts up to 2026.

Methodology:

1/ Data Ingestion

- **AWS S3 Buckets**: Created an S3 bucket named capital-budgeting to store the raw dataset.
- Organized data into a structured folder system with lifecycle management rules for cost-efficient storage using Glacier Interval Retrieval.

2. Data Profiling

- Conducted data profiling using AWS Glue DataBrew.
- Verified dataset quality, which required minimal corrections due to its structured and accurate format.

3. Data Cleaning

- Cleaned the dataset using AWS Glue DataBrew by:
    - Removing extra columns.
    - Renaming columns for clarity.
    - Deleting null values.
- Stored the cleaned dataset in S3, categorized into systems (Parquet format) and users (CSV format) folders.

4. Descriptive Analysis Pipeline

- Built a visual ETL pipeline in AWS Glue to identify the service category with the highest capital budget allocation. Steps included:
    - **Schema Changes**: Dropped unnecessary columns and renamed relevant fields.
    - **Filters**: Applied to rows for precision in analysis.
    - **Aggregation**: Computed maximum budget allocation by service category.
- Results were saved in S3 folders in both Parquet and CSV formats for further use.

5. Insights and Findings:

- Identified the service category with the highest multi-year capital budget allocation.
- Provided insights into budget trends and allocation priorities, helping to track financial planning and project prioritization.

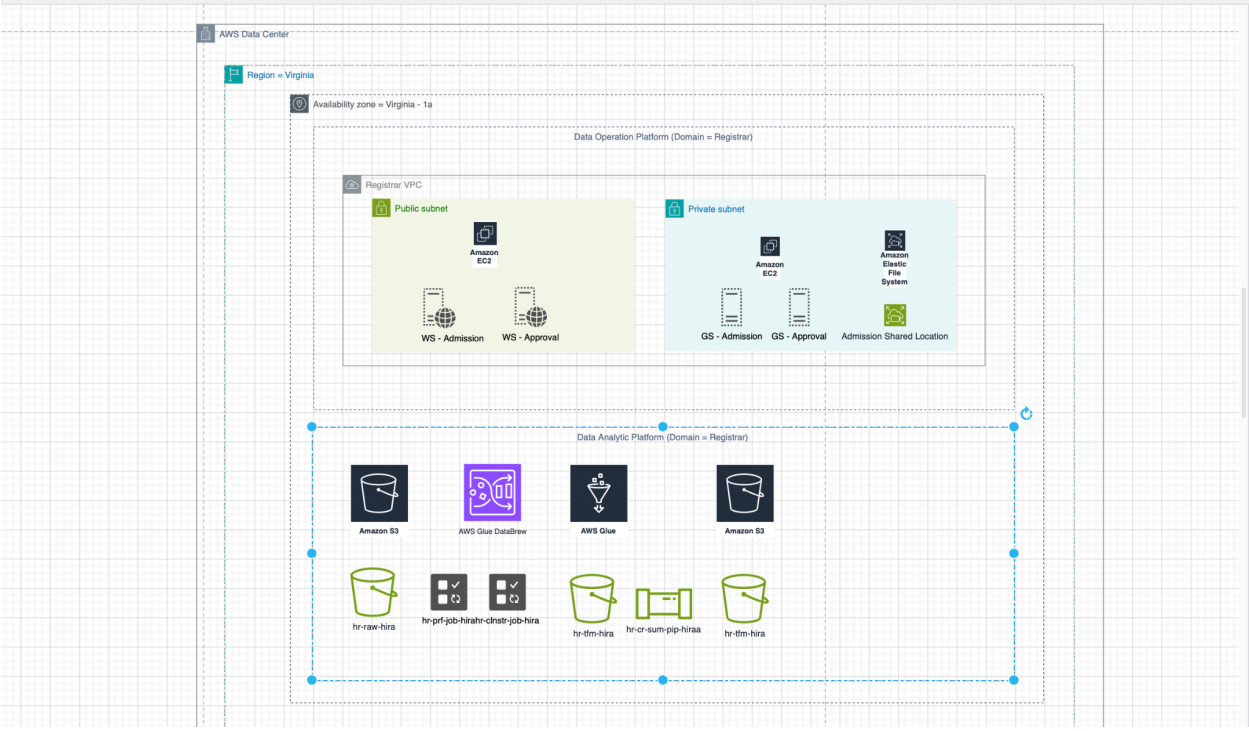
Tools and Technologies:

AWS S3: Data storage and management.

AWS Glue DataBrew: Data profiling and cleaning.

AWS Glue: Visual ETL pipeline creation.

Visualization: Draw.io



Data Wrangling

Project Title: Descriptive Analysis of Multi-Year Capital Budget Allocations in the City of Vancouver.

Objective:To clean, transform, and consolidate the City of Vancouver's multi-year capital budget dataset, ensuring its accuracy, completeness, and usability for analytical purposes.

Dataset: The dataset consisted of the City of Vancouver's 2022 Multi-Year Capital Budget and Capital Expenditure data, which included:

- **Service Categories**: Transportation, utilities, public health, and more.
- **Budget Allocations**: Multi-year and annual budget details.
- **Demographics**: Integrated datasets for enriched analysis.
- **Transaction Details**: From raw to transformed states using AWS pipelines.

Methodology:

Data Collection:

- **Tools and Technologies**:
  - **AWS S3**: Used to store raw datasets in an organized folder structure.
  - **Lifecycle Rules**: Configured for cost optimization using Glacier.
- **Steps**:
  - Ingested datasets from the City of Vancouver's portal.
  - Created capital-budgeting S3 buckets to organize raw and transformed data.

Data Profiling:

- **Tools and Technologies**:
  - **AWS Glue DataBrew**: To profile datasets and identify quality issues.
- **Steps**:
  - Analyzed data to detect missing values, duplicates, and inconsistencies.
  - Documented findings to guide cleaning processes.

Data Cleaning:

- **Tools and Technologies**:
  - **AWS Glue DataBrew**: To automate cleaning tasks.
- **Steps**:

    - Removed unnecessary columns and handled missing values.
    - Standardized formats for dates and categorical variables.
    - The cleaned dataset was stored in another S3 Bucket under the name transformation.

Data Transformation:

- **Tools and Technologies**:
    - **AWS Glue Visual ETL**: For schema adjustments, filtering, and aggregation.
- **Steps**:
    - Built ETL pipelines to process cleaned datasets.
    - Aggregated data to determine service categories with the highest budget.
    - Saved transformed data as Parquet (for internal use) and CSV (for stakeholders).

Data Integration:

1. **Tools and Technologies**:
    - **AWS Athena**: For SQL queries on integrated datasets.
2. **Steps**:
    - Merged demographic data to enrich insights.
    - Consolidated results into unified datasets.

Outcome:

- The wrangling process produced clean, structured datasets ready for exploratory analysis and reporting.
- It addressed inconsistencies and ensured data usability for decision-making.

Tools and Technologies:

- Python (using libraries like Pandas and NumPy) or R for data manipulation and cleaning.
- SQL for data extraction and initial assessment of data from relational databases.
- Jupyter Notebook or RStudio for interactive data wrangling and documentation.
- Visualization tools (like Matplotlib or Seaborn) to assist with EDA and quality checks.

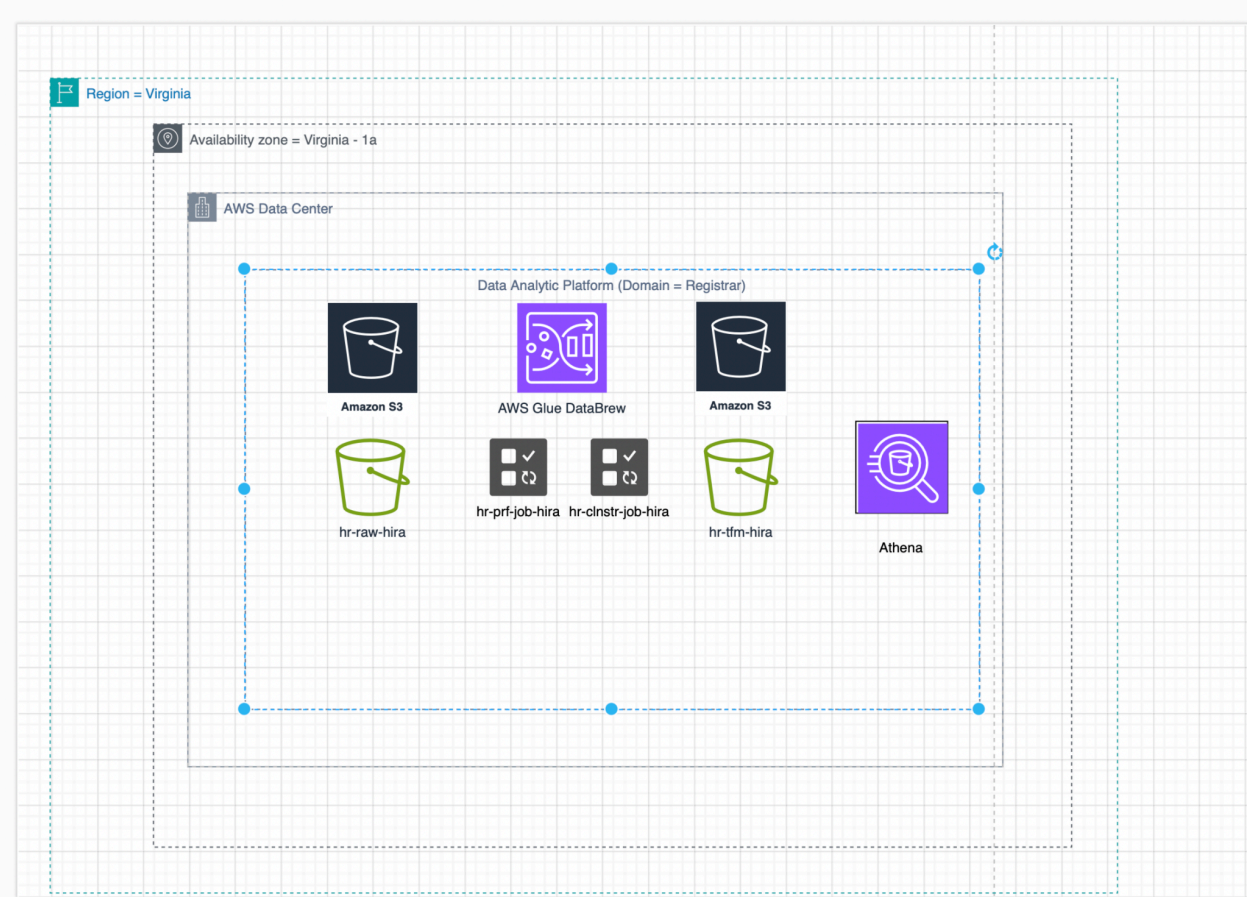
Deliverables:

1. Cleaned and transformed datasets stored in S3 (CSV and Parquet formats).
2. Documented data wrangling steps, including cleaning techniques and transformation logic.
3. Visual ETL pipeline in AWS Glue with screenshots.

Timeline:

4 weeks.

1. Week 1: Data collection and profiling.
2. Week 2: Cleaning and initial transformations.
3. Week 3: Data integration and advanced transformations.
4. Week 4: Validation and final data export.



Data Quality Control

Objective: To establish robust data quality control measures ensuring the reliability, consistency, and accuracy of the datasets, supporting reliable analytics and governance compliance.

Scope: The project will focus on the following key areas:

- Data Profiling: Analyzing existing datasets to assess quality levels.
- Data Cleansing: Developing processes to correct inaccuracies and eliminate duplicates.
- Data Validation: Implementing validation rules and checks to ensure data integrity.
- Monitoring and Reporting: Establishing ongoing monitoring processes and dashboards to track data quality metrics.
- Training and Awareness: Creating training programs for staff on data quality best practices.

Methodology:

Data Validation Rules:

- Implemented automated validation checks in AWS Glue to ensure integrity during data ingestion and transformation.
- Created **passed** and **failed** subzones in the S3 bucket to segregate valid and invalid records.

Data Cleansing Processes:

- Applied standardization techniques to correct data inconsistencies, such as naming conventions and date formats.
- Used imputation techniques to address missing values, ensuring no critical gaps in the dataset.

Monitoring and Reporting:

- Configured Amazon CloudWatch for real-time monitoring of data pipelines.
- Created dashboards to track metrics such as file size, object count, and data quality trends.
- Set up alarms for anomalies (e.g., unexpected object counts or data discrepancies).

Data Encryption and Protection:

- Encrypted data using AWS Key Management Service (KMS) for secure storage and access.
- Applied access controls to limit data exposure and ensure compliance with privacy standards.

Data Governance and Documentation:

- Defined lifecycle policies to manage data retention and transition older records to lower-cost storage (e.g., Glacier).

- Documented all data quality control processes, including validation checks and cleaning techniques, for transparency and repeatability.

Deliverables:

- Validated datasets stored in separate S3 subzones (passed and failed).
- CloudWatch dashboards and alarms for real-time monitoring.
- Encryption configuration with AWS KMS.
- Documentation of data quality metrics, validation rules, and governance policies.

Timeline:

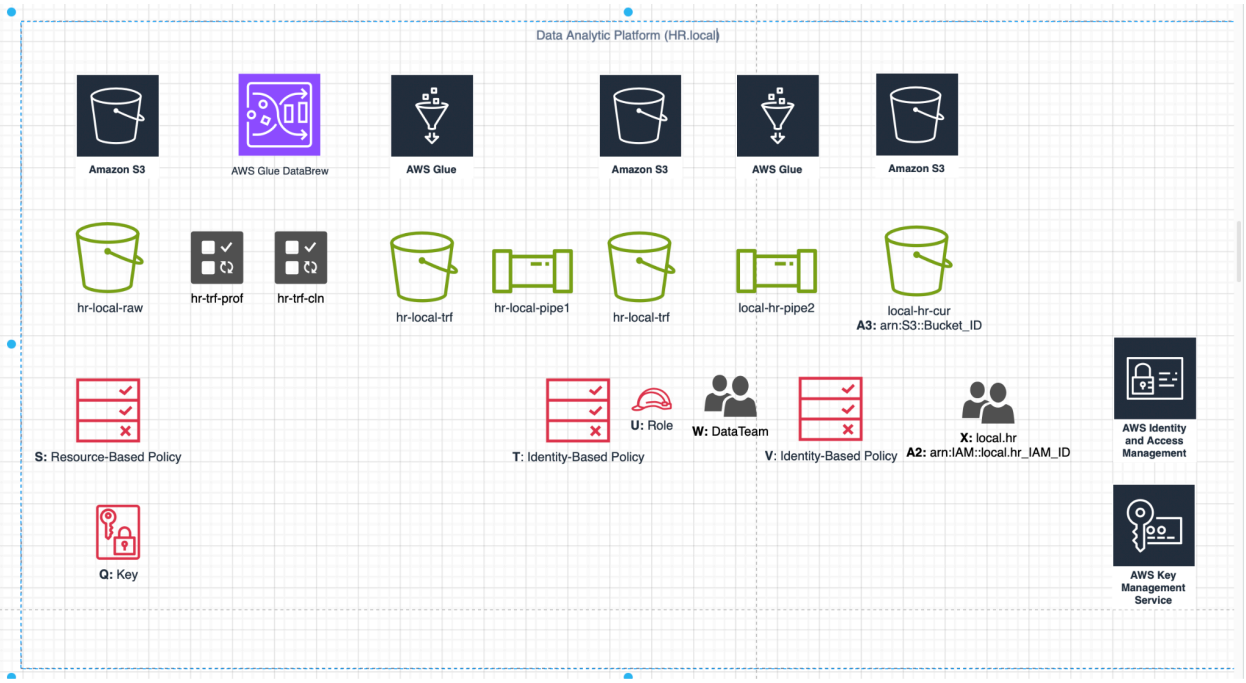
Week 1: Validation rule creation and implementation.

Week 2: Cleansing and quality monitoring setup.

Week 3: Encryption and governance configurations.

Week 4: Finalizing dashboards and documentation.

This Data Quality Control initiative aims to empower City of Vancouver Enterprise to enhance its data integrity and reliability, resulting in improved decision-making, operational efficiency, and compliance with regulatory requirements.


Data Analytic Platform (HR.local)
Amazon S3
AWS Glue DataBrew
AWS Glue
Amazon S3
AWS Glue
Amazon S3
hr-local-raw
hr-trf-prof
hr-trf-cln
hr-local-trf
hr-local-pipe1
hr-local-trf
local-hr-pipe2
local-hr-cur
A3: arn:S3::Bucket_ID
S: Resource-Based Policy
T: Identity-Based Policy
U: Role
W: DataTeam
V: Identity-Based Policy
X: local.hr
A2: arn:IAM::local.hr_IAM_ID
AWS Identity and Access Management
Q: Key
AWS Key Management Service