**Project 1: AWS-Data-Analyst-UCW Registrar Office (Academic Standing Procedure)**

* *(UCW Dataset, Class Activity, Weekly Activity)*

This document has explained the UCW Dataset which analyze the students CGPA performance metrics, identify the term risk and recommend the strategies to students for enhancing their academic performance. This Analysis has used the different AWS services like AWS S3 Bucket for data ingestion, Data Brew for cleaning and profiling, Glue fore ETL pipeline development, AWS Cloud platform tools.

This is using the dataset of UCW RewgistrarOffice-9023pAcademic Standing Procedure. It contains the students’ academic performance details. It includes information about the student’s Performance with CGPA trend and academic standings. This dataset is useful for the depth analysis to achieve the goal of student improvement outcomes. Weekly activities and class participation have been conducted on weekly basis to perform the AWS services described below.

A computer screen shot of a computer

Description automatically generated

**Descriptive Analytics**

1. **Project Description:**  
   Analyze the student academic performance w.r.t each term to identify the performance trends and propose strategies for improving students CGPA performance.
2. **Project Title:**   
   UCW Students Academic Performance Analysis and Recommendations
3. **Objective:**  
   This project goal is to identify the terms where UCW students are most at academic risks of underperformance. With the help of this UCW can highlight the CGPA trends and associate support or improvements which students require.
4. **Dataset:**  
   The dataset has following below attributes:
   * **Student ID:** A unique ID for each student.
   * **Term:** Academic Term.
   * **CGPA:** Cumulative Grade Point Average.
   * **Term GPA:** GPA achieved during the term.
   * **Academic Standing:** Classification (e.g., Good Standing, Probation, Suspension).
5. **Methodology:**
   * **Data Collection and Preparation:**
     + The raw academic datasets were ingested into AWS S3 bucket in structured format into *raw* directory ‘ro-raw-shikha’.

A screenshot of a computer

Description automatically generated

* + - Using AWS DataBrew service, we have performed the data profiling and data cleansing both to remove the NULL values, duplicate values, column renames, creating categorical column and maintain the data consistency. And then moved the cleaned data into back AWS S3 bucket under the *transform* directory ‘ro-prf-shikha’. The csv output stored underuser folder and parquet has stored under system folder.

**A screenshot of a computer

Description automatically generated**

* + - Final processed data into current directory ‘ro-cur-shikha’
  + **Data Enrichment:**
    - Additional dataset such as ‘historical academic record’ and ‘performance specific statistics’ has been used for more data enrichment for further analysis.
    - These data set has been stored under raw directory ‘ro-raw-shikha’ of AWS S3 bucket then perform the profiling and cleansing using DataBrew service and stored into the *transform* directory ‘ro-trf-shikha’ under S3 bucket.
    - Next, AWS Glue service used for making different sources schema consistent.
  + **ETL Pipeline Development:**
    - AWS Glue services used for performing the ETL (Extract, Transform, Loading) operation. It extracts the UCW dataset from transformation folder and the do the transformation to obtain the result for identify students having low CGPA as per the term registered and analyze patterns across terms and CGPA for targeted academic support.

A screenshot of a computer

Description automatically generated

* + **Data Wrangling:**
    - Extracted the schema using AWS Glue Crawlers which help the unlimited query access to the data for future analysis and load the result into ‘ro-cur-shikha’ directory under S3 bucket.

A screenshot of a computer

Description automatically generated

* + **Data Analysis:**  
    Used AWS Athena service for using the SQL queries to extract term-wise CGPA trends which help to identify at-risk terms.
  + **Insights and Findings:**
    - The Spring term displayed the lowest average CGPA which showed that during this term the students are facing higher academic challenges.
    - The significant risk of student’s performance identified in Spring enrollments.
  + **Data Encryption:**  
    Data security plays an important role for keeping data secure using AWS KMS (Key management System) service. Created the encryption key ‘reg-scdStnd-key-shikha’ which helps to protect the S3 bucket from unauthorized access. It helps to stay data securely encrypted both during transfer and at rest.

**A screenshot of a computer

Description automatically generated**

**A screenshot of a computer

Description automatically generated**

* + **Monitoring and Controlling Operations:**  
    AWS CloudWatch is used to monitor ETL jobs, S3 bucket usage, and error logs. These all metrics can be visualized in the created dashboard using AWS Cloud Watch. Cloud Watch (Monitor, Compare, Control, Metric, Logs for resource)

A screenshot of a computer

Description automatically generated

I have used the Cloud Trail service. This service helps me to monitor the activities by integrating with the CloudWatch logs. Created the cloud trail with the name ‘reg-scdmStnd-dap-users-shikha’. A screenshot of a computer

Description automatically generated

1. **Tools and Technology: In this project only AWS cloud services are used. Services names are as below.** 
   * **S3:** For data storage and management.
   * **DataBrew:** For cleaning and profiling the dataset.
   * **Glue:** For building and managing the ETL pipeline.
   * **Athena:** For querying and analyzing the data.
   * **KMS:** For encrypting and securing the data.
   * **CloudWatch:** For monitoring and ensuring smooth operations.

**Project 2: AWS-Data-Analyst-Vancouver City Greenest City Projects**

* *(City of Vancouver- Open Data Portal- Greenest City Project)*

This document has analyzed the Vancouver’s greenest City Projects dataset to evaluate the sustainability initiatives with their distributed goals across different categories and sub-categories. It also helps us to understand the percentage of goal achievement u der the Greenest City Action Plan and its improvement outcomes. This project uses the different AWS services for efficient data storage, management, processing and analysis which are described below.

A diagram of a data flow

Description automatically generated

**Descriptive Analytics**

1. **Project Description:**  
   The project involves the analyzing of Vancouver’s Greenest City Projects dataset. This help to evaluate the distribution and impact of sustainability goals aligned with the Greenest City Action Plan. It helps to identify the trends, gaps and opportunity for improvement in achieving sustainability goals such as Clean Air, Local Foor, Green Economy etc. The AWS services are used here to leverage the actionable insights for resource allocation.
2. **Project Title:**   
   Analysis for Sustainability Initiatives by Vancouver’s Greenest City projects.
3. **Objective:**  
   To analyze the Vancouver’s Greenest City project dataset to evaluate the progress, underperforming category like clean air, local food and underperforming sectors like public or city of sustainability initiatives and provide the recommendations for target achievement.
4. **Dataset:**  
   The dataset has following below attributes:
   * **MAPID:** A unique ID for each project.
   * **Name:** Project name.
   * **Category1:** Project Category (City project, City studio project, Greenest city fund projects, Private projects).
   * **Category2:** Project Goal (‘Lighter-Footprint’, ‘Green Transportation, ‘Local-Food’, ‘Green-Economy’, ‘Access-to-Nature’,’ Clean-Air’, ‘Climate-Leadership’, ‘Green-Building’, ‘Green-Building’).
   * **Address:** Project physical location
   * **Sort Description:** Project description
   * **URL, URL2, URL3-** Links for more information
   * **GEOM-** Geographical coordinates. (JSON format)
   * **Geo Local Area-** Project Local Area
5. **Methodology:**
   * **Data Collection and Preparation:**
     + The raw greencity datasets were ingested into AWS S3 bucket in structured format into *raw* directory ‘shikha-greencity-raw’.

A screenshot of a computer

Description automatically generated

* + - Using AWS DataBrew service, we have performed the data profiling and data cleansing both to remove the NULL values, duplicate values, column renames, creating categorical column and maintain the data consistency. And then moved the cleaned data into back AWS S3 bucket under the *transform* directory ‘shikha-greencity-trf’. The csv output stored under user folder and parquet has stored under system folder.

A screenshot of a computer

Description automatically generated

* + - Final processed data into current directory ‘ro-cur-shikha’
  + **Data Enrichment:**
    - Additional dataset such as ‘Greenest City Action Plan (GCAP) Metrics’ has been used for more data enrichment for further analysis. This metric provides me the more depth information related the Vancouver’s environmental and sustainability goal progress w.r.t sustainability goals (climate, green buildings, transportation, waste, and water). It helps to highlight the baseline level, target achievement percentage, challenges and future action plan w.r.t each sustainability goal.
    - These data set has been stored under raw directory ‘shikha-greencity-raw’ of AWS S3 bucket then perform the profiling and cleansing using DataBrew service and stored into the *transform* directory ‘shikha-greencity-trf’ under S3 bucket.
    - Next, AWS Glue service used for making different sources schema consistent.
  + **ETL Pipeline Development:** AWS Glue services used for performing the ETL (Extract, Transform, Loading) operation. It extracts the Greencity dataset from transformation folder and do the transformation to obtain the result for identify the current distribution of projects across different categories, and how does each category contribute to the total project count. 

A graph with different colored bars

Description automatically generated

* + **Data Wrangling:** Extracted the schema using AWS Glue Crawlers which help the unlimited query access to the data for future analysis and load the result into ‘ro-cur-shikha’ directory under S3 bucket.

A screenshot of a computer

Description automatically generated

* + **Data Analysis:**  
    Used AWS Athena service for using the SQL queries to to calculate how many goals are below the baselines. With the query result, I got to know that the ‘Clean-Air’ goal is the only goal which is not achievable out of all goals.
  + **Insights and Findings:**
    - The Clean-Air goal is the goal which is not achievable out of all goas in Sustainability Action Plan Goal 2020. This can be due to increase of vehicle and related pollution.
    - The goal is transferred to next year goal with ore initiatives to achieve.
  + **Data Protection – Encryption Key:**  
    Data security plays an important role for keeping data secure using AWS KMS (Key management System) service. Created the encryption key ‘Protection\_key\_shikha’ which helps to protect the S3 bucket from unauthorized access. It helps to stay data securely encrypted both during transfer and at rest.

A screenshot of a computer

Description automatically generatedA screenshot of a computer

Description automatically generated

Next, created the replication rule with the name ‘greencity-rule-shikha’ under ‘Management’ tab on the original bucket “ro-raw-shikha”.A screenshot of a computer

Description automatically generated

* + **Data Governance:** In data governance, I ensured the data follows all the compliance like quality, completeness, uniqueness, privacy, protection and control. Next, I have used the AWS Glue service for creating ETL. This governance pipeline ensures the privacy, uniqueness and completeness of the dataset.

A screenshot of a computer

Description automatically generated

This ETL stored the Passed and Failed result on the defined folder path” s3://greencity-shikha-rawtrfcurr/greencity-trf-shikha/data-quality-shikha

A screenshot of a computer

Description automatically generated

* + **Monitoring and Controlling Operations:**  
    AWS CloudWatch is used to monitor ETL jobs, S3 bucket usage, and error logs. These all metrics can be visualized in the created dashboard using AWS Cloud Watch. Created service name “CloudWatch service” to Monitor, Compare, Control, Metric, Logs for resources.A screenshot of a computer

    Description automatically generated

I have used the Cloud Trail service. This service helps me to monitor the activities by integrating with the CloudWatch logs. Created the cloud trail with the name ‘greencity-UserTrail-shikha’. A screenshot of a computer

Description automatically generated

Lastly, I have used the event history of the user under Event History of Cloud Trail to monitor the events executed by the users.A screenshot of a computer

Description automatically generated

1. **Tools and Technology: In this project only AWS cloud services are used. Services names are as below.** 
   * **S3:** For data storage and management.
   * **DataBrew:** For cleaning and profiling the dataset.
   * **Glue:** For building and managing the ETL pipeline.
   * **Athena:** For querying and analyzing the data.
   * **KMS:** For encrypting and securing the data.
   * **CloudWatch:** For monitoring and ensuring smooth operations.