****

**NORTHEASTERN UNIVERSITY  
ITC 6460 Fall 2023**

**Section 01**

**Cloud Analytics Project: NK Inc. Carbon Emission Reporting**

**By Husky Group 5:**

**Nsikak-Abasi Una**

**Keevin Peirson**

**INTRODUCTION**

In recent years, the global community has witnessed a growing awareness of environmental issues, particularly the impact of carbon emissions on climate change. As organizations strive to adopt sustainable practices, the need for accurate and transparent carbon emissions reporting has become imperative. This paper explores how Amazon Web Services (AWS) Modern Data Architecture can be leveraged to streamline and enhance carbon emissions reporting for organizations, contributing to a more sustainable and accountable future.

**IMPORTANCE OF CARBON EMISSIONS REPORTING**

Carbon emissions reporting plays a crucial role in measuring and managing an organization's environmental impact. Governments, investors, and consumers are increasingly demanding transparency and accountability in terms of carbon footprints. Accurate reporting not only helps organizations comply with regulatory requirements but also fosters a positive corporate image, attracts socially responsible investors, and meets the expectations of environmentally conscious consumers.

**AWS MODERN DATA ARCHITECTURE OVERVIEW**

The monthly fleet usage and energy consumption data of NK Inc are uploaded into an Amazon S3 container named 'Raw Bucket.' AWS Glue, a serverless data integration service, is then employed to transform this raw data and store it in a different S3 bucket known as 'Transformed Bucket.' AWS Glue DataBrew, a visual data preparation tool, comes into play to cleanse and standardize the data, preparing it for carbon emissions calculations. To extract energy usage details from the monthly bill PDF files provided by the utility provider, a Machine Learning service named Amazon Textract is utilized. The carbon emissions factor data from GHG Protocol finds its home in an Amazon DynamoDB table, a flexible NoSQL database service designed for lookup calculations. The construction of the carbon emissions calculator is carried out using the serverless computing service Amazon Lambda, which effectively computes the equivalent carbon emissions. The calculated emissions data is subsequently stored in a cloud data warehouse, Amazon Redshift, for downstream analytics. Lastly, Amazon QuickSight is harnessed to craft interactive dashboards showcasing carbon emissions data in a user-friendly manner.

**A diagram of data processing

Description automatically generated**

Amazon Web Services (AWS) offers comprehensive cloud computing services that empower organizations to build scalable, secure, and flexible solutions. AWS Modern Data Architecture encompasses a suite of services designed to handle modern data processing, storage, and analytics complexities. Organizations can efficiently manage and analyze large volumes of data by leveraging this architecture, making it an ideal platform for carbon emissions reporting.

Critical Components of AWS Modern Data Architecture for Carbon Emissions Reporting

1. **Data Collection and Ingestion**:
   * **Amazon S3:** object storage solution known for its exceptional scalability, data reliability, top-notch security measures, and impressive performance capabilities. It empowers users to securely store and safeguard vast volumes of data, catering to a diverse array of applications, including data lakes, websites, cloud-native apps, backup systems, archives, machine learning, and analytics.
2. **Data Processing and Transformation**:
   * **AWS Glue**: is a serverless data integration service that simplifies the process of locating, readying, transferring, and amalgamating data from various origins to facilitate analytics, machine learning (ML), and application development.
   * **AWS Glue Data Brew**: is a visual data preparation tool designed to streamline the task of cleaning and standardizing data, ultimately simplifying the process for data analysts and data scientists who are gearing up to use this data for analytics and machine learning (ML) applications.
   * **AWS Lambda**: is an event-driven, serverless compute service that empowers you to execute code for a wide range of applications and backend services, all without the need to provision or oversee servers. It seamlessly responds to code execution requests across a broad spectrum of scales, handling anything from a mere dozen events per day to several hundred thousand events per second.
3. **Data Storage**:
   * **Amazon Redshift**: employs SQL for the analysis of structured and semi-structured data across data warehouses, operational databases, and data lakes. It utilizes AWS-designed hardware and harnesses the capabilities of machine learning to consistently provide optimal price-to-performance ratios, irrespective of the scale of the operation.
   * **Amazon DynamoDB**: Stores metadata and supports fast and scalable queries.
4. **Data Analysis and Reporting**:
   * **Amazon Athena**: is an interactive analytics service that operates in a serverless manner, and it is constructed upon open-source frameworks while accommodating a variety of open-table and file formats. Athena offers a streamlined and adaptable approach for the analysis of vast data volumes, enabling you to examine data directly in its original location.
   * **Amazon QuickSight**: empowers organizations driven by data, delivering a comprehensive business intelligence (BI) solution at an extensive scale. QuickSight enables users to address diverse analytical requirements using a single, authoritative source, achieved through contemporary interactive dashboards, structured reports, integrated analytics, and natural language inquiries.
5. **AWS Identity and Access Management (IAM):** allows you to define which individuals or entities have access to AWS services and resources, offering centralized administration of precise permissions and the ability to evaluate access for the purpose of enhancing and fine-tuning permissions across the AWS environment.

**IMPLEMENTATION STEPS**

1. **Data Collection**: the raw fleet usage (csv) & utility bill (pdf) data was uploaded into an s3 bucket
2. **Data Processing**:
   1. Data Transformation of the fleet usage data: we selected s3 bucket as the source csv file, transformed it using mapping, and then returned it to the target destination s3 bucket as a JSON file.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* 1. Data Cleansing of the fleet usage data: The carbon emissions calculator module mandates a particular data model for input data. In this segment, we will employ AWS Glue DataBrew to preprocess the data for subsequent emissions computations. AWS Glue DataBrew, a visual data preparation tool, simplifies the process of data cleaning and standardization for data analysts and data scientists, making it suitable for analytical purposes. As part of this data preparation, we will eliminate the "miles\_driven" column, which does not belong to the emissions calculator data model. We have successfully cleaned the transformed JSON files related to fleet usage.
* Created a dataset

A screenshot of a computer

Description automatically generated

* Created a project and then deleted the miles-driven column

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* 1. Data Quality of the fleet usage data: The absence of dependable emissions data presents a significant obstacle on the path to achieving a net-zero and low-carbon economy. For organizations committed to sustainability, data quality holds paramount significance. In this segment, we will explore the utilization of AWS Glue DataBrew to consistently assess data quality, thereby ensuring the preservation of data integrity throughout the emissions data calculation process.
* To check for quality, use the dq rules option in databrew to create a ruleset. rule 1: that data isn’t missing from the raw\_data column rule 1

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* The rule was created and ran.

A screenshot of a computer

Description automatically generated

* 1. Data Quality of the fleet usage data: Monitoring emissions data changes over time, identifying their sources, and understanding the underlying reasons is vital for ensuring data traceability and fulfilling audit prerequisites. Establishing a data governance strategy is imperative to comprehensively trace emissions datasets' lineage across the platform. Within this section, we will delve into the process of examining the data lineage of emissions data within the platform.

A screenshot of a computer

Description automatically generated

A screenshot of a computer

Description automatically generated

* 1. Utility bill processing: we have developed an AWS Lambda function responsible for employing the Amazon Textract Intelligent Document Processing API to extract unstructured data from these PDFs. This extracted data is then stored in the Transformed S3 bucket, where it will be utilized for emissions calculations. Please proceed to upload, deploy, and evaluate the Lambda code. Notably, all the essential attributes from the PDF bills, essential for emissions calculations, have been effectively extracted using Amazon Textract.

A screenshot of a computer

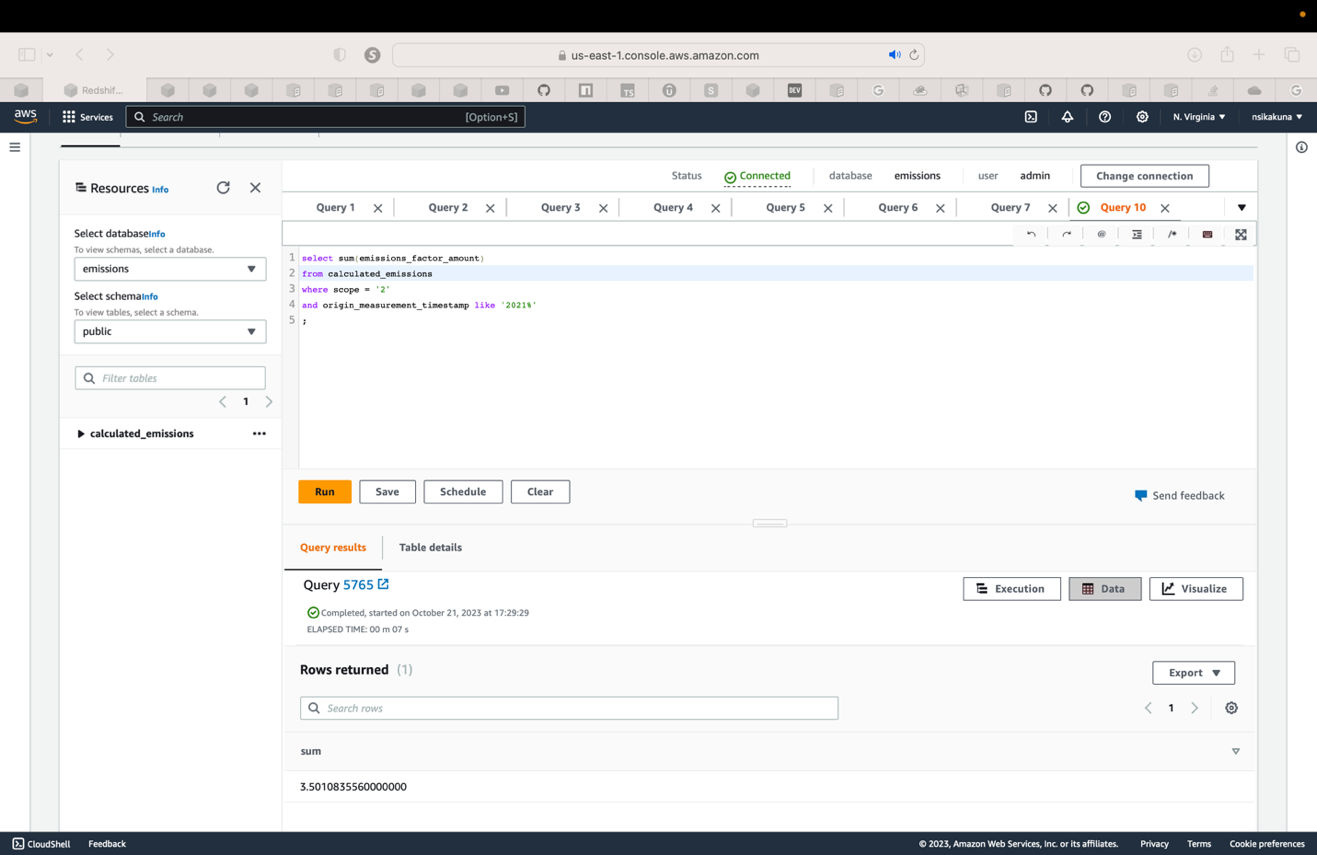
Description automatically generated

1. **Carbon Emissions Calculator**: We've developed a Lambda function specifically designed to make estimations of Greenhouse Gas emissions, utilizing the GHG Emissions Calculation Tool. This function is designed to read the previously generated transformed fleet usage and utility data, perform emissions calculations, and subsequently record the outcomes in both Amazon DynamoDB and our Amazon Redshift data warehouse, where it can be readily accessed for querying purposes. The code was successfully deployed and tested.

A screenshot of a computer

Description automatically generated

1. **Emissions Reporting**: we examined the computed emissions through Amazon Redshift and employed Amazon QuickSight to construct business intelligence dashboards. These dashboards serve the purpose of visualizing the emissions that have been calculated and projecting future emission trends.
   1. Analyze the calculated emissions: we used Amazon Redshift to analyze the calculated fleet usage and utility bill carbon emissions.
      * Query 1: The overall utility bill emissions from the electricity consumption for the year 2021



* + - Query 2: The carbon emissions of (*Motor Gasoline - Gasoline Heavy-duty Vehicles*) for 2020 and 2021

A screenshot of a computer

Description automatically generated

* 1. Build an interactive dashboard using Amazon Quicksight:
     1. visualize the fleet usage emissions for the last two years (2020-2012): The dashboard below describes the fuel usage of different types of vehicles on each month and corresponding overall equivalent carbon emissions. Using a stacked bar combo.

A screenshot of a computer

Description automatically generated

* + 1. Forecast the carbon emissions for the next 12months. This line chart describes the forecasted CO2 emissions for the next 12 months.

A screenshot of a computer

Description automatically generated

**Benefits of AWS Modern Data Architecture for Carbon Emissions Reporting**

1. **Scalability**:
   * Effortless Scalability: AWS enables easy scalability, allowing organizations to adapt their infrastructure based on the volume of emissions data. This ensures consistent performance and reliability even as data volumes grow.
   * Handling Large Data Volumes: The platform's scalable infrastructure capabilities make it well-suited for managing the substantial data generated through carbon emissions monitoring.
2. **Cost-effectiveness**:
   * Pay-as-You-Go Model: AWS employs a cost-effective pay-as-you-go model, which means organizations only incur charges for the specific resources and services they utilize.
   * Serverless Computing: Serverless computing options, like AWS Lambda, enhance cost efficiency by eliminating the need for provisioning and managing servers.
3. **Flexibility**:
   * Tailored Architecture: The flexibility of AWS enables the customization of architecture to meet the unique needs and requirements of different organizations, ensuring optimal performance.
4. **Security and Compliance**:
   * Robust Security: AWS provides robust security measures, which, coupled with various compliance certifications, guarantees the integrity and protection of data, essential for handling sensitive emissions information.
5. **Real-time Insights:**
   * Timely Decision-Making: AWS facilitates real-time monitoring and analysis capabilities, ensuring organizations have access to up-to-the-minute insights, allowing for timely decision-making and interventions in response to emissions data.

**CONCLUSION**

AWS Modern Data Architecture stands as a robust and adaptable answer for organizations striving to elevate their carbon emissions reporting capacities. Through the utilization of AWS services spanning data collection, processing, storage, and analysis, NK Inc not only met regulatory standards but has also unearthed crucial insights into its environmental footprint. In an era where sustainability takes center stage, embracing such contemporary data architectures emerges as imperative for organizations dedicated to transparency, accountability, and environmental stewardship.

**REFERENCE**

* *Amazon S3*. (n.d.). Amazon Web Services, Inc. Retrieved October 24, 2023, from <https://aws.amazon.com/pm/serv-s3/?trk=20e04791-939c-4db9-8964-ee54c41bc6ad&sc_channel=ps&ef_id=CjwKCAjw1t2pBhAFEiwA_-A-NDDs7CxMxVISS1NB_hHdOx9vUV9dsgq5E-vyDJ-58P9jgVjIzj0raBoCSK4QAvD_BwE:G:s&s_kwcid=AL>
* *Serverless Data Integration – AWS Glue – Amazon Web Services*. (n.d.). Amazon Web Services, Inc. Retrieved October 24, 2023, from <https://aws.amazon.com/glue/?trk=30a5c4e0-0d3e-437e-b82b-d5463ae61868&sc_channel=ps&ef_id=CjwKCAjw1t2pBhAFEiwA_-A-NO8VVGJOrg0xHIn7JvGbjYPo-tAw4c919frh767ai_SRq3LOWGJwmhoCZ44QAvD_BwE:G:s&s_kwcid=AL>
* *AWS Glue DataBrew - visual data preparation - Amazon Web Services*. (n.d.). Amazon Web Services, Inc. <https://aws.amazon.com/glue/features/databrew/>
* AWS. (2019a). *AWS Lambda – Serverless Compute - Amazon Web Services*. Amazon Web Services, Inc. <https://aws.amazon.com/lambda/>
* *Greenhouse Gas Protocol |*. (n.d.). Ghgprotocol.org. <https://ghgprotocol.org>
* AWS. (2019a). *Amazon Redshift - Amazon Web Services*. Amazon Web Services, Inc. <https://aws.amazon.com/redshift/>
* *Amazon DynamoDB*. (n.d.). Amazon Web Services, Inc. Retrieved October 24, 2023, from <https://aws.amazon.com/pm/dynamodb/?trk=390f2f77-1064-4521-bd83-27d9213b65c9&sc_channel=ps&ef_id=CjwKCAjw1t2pBhAFEiwA_-A-NJIBt0RcE6CuGsMvK_5dPPz-rVLUGZ8HnN3WF4KqBwTWEYAMeHdgIBoCmfYQAvD_BwE:G:s&s_kwcid=AL>
* *Amazon Athena — Serverless Interactive Query Service - AWS. (2019). Amazon Web Services, Inc.* [*https://aws.amazon.com/athena/*](https://aws.amazon.com/athena/)
* *Amazon. (n.d.). Amazon QuickSight. Amazon Web Services, Inc.* [*https://aws.amazon.com/quicksight/*](https://aws.amazon.com/quicksight/)
* *AWS. (2017). AWS Identity & Access Management. Amazon Web Services, Inc.* [*https://aws.amazon.com/iam/*](https://aws.amazon.com/iam/)