ANALYSIS OF ENERGY CONSUMPTION PATTERNS IN THE UNITED STATES

A Project Report

Presented to

DATA 228-11

Spring, 2023

By

Manidedeepya Chennapragada (016045276)

Kalyan Vikkurthi (016663257)

Dharnidhar Reddy Banala (016671226)

Nikhil Mylarusetty (016656393)

Copyright © 2021

Manidedeepya Chennapragada,Kalyan Vikkurthi,Dharnidhar Reddy Banala,

Nikhil Mylarusetty

ALL RIGHTS RESERVED

ABSTRACT

Analysis Of Energy Consumption Patterns In The United States

By Manidedeepya Chennapragada,Kalyan Vikkurthi,Dharnidhar Reddy Banala, Nikhil Mylarusetty

This project involves the implementation of a comprehensive data pipeline for real-time data collection, processing, and analysis of electricity consumption across different sectors with the objective of identifying trends, patterns to help related personnel make decisions based on the data. To achieve this goal, we utilized various technologies and tools such as AWS CLI, Kafka, AWS S3, AWS Glue, AWS Crawler, and AWS Athena to manage the sizable dataset, clean and transform the data, and prepare it for further analysis. The solution involved establishing a Kafka producer to stream the collected data to the Kafka messaging system, which was then consumed by the AWS Glue and processed in a defined schema. The resulting data was stored in an S3 bucket and crawled by the AWS Crawler for metadata extraction and cataloging. Finally, the implementation of AWS Athena allowed for efficient querying and analysis of the stored data without the need for complex infrastructure management. By integrating Tableau with Athena, authorized users gained access to a personalized dashboard that facilitated effective data visualization and exploration. The results of the project include the successful implement the data pipeline, enabling real-time data collection and processing capabilities. The project successfully addressed the problem statement by developing an efficient and scalable system for real-time data analysis, and the resulting solution can be used in various industries and applications to facilitate data-driven decision-making based on analyzed and processed real-time data.

Acknowledgements

For the successful development of our project, we would like to thank Prof Andrew H. Bond for his continued guidance and assistance throughout the course work.

Table of Contents

Appendix

Table of Figures

Table of Tables

Chapter 1. Introduction

1.1 Project Goals and Objectives

The project proposes using big data analytics to examine energy consumption patterns in the United States. The primary objective is to provide valuable insights into electricity generation, consumption, and pricing trends by leveraging data from the US Energy Information Administration (EIA). The project aims to gather data from multiple sources such as real-time data feeds, APIs, and web scraping techniques are included to create a comprehensive dataset that accurately captures energy consumption metrics at both national and state levels. Additionally, the project seeks to employ big data technologies such as Apache Kafka and Amazon S3 for efficient data streaming, storage, and querying using Amazon Athena.

1.2 Problem and Motivation

The project is motivated by the critical issues associated with energy consumption, which have significant impacts on both the environment and the economy. By analyzing energy consumption patterns, the project aims to address these challenges and provide insights that can inform decision-making processes. The project recognizes the importance of understanding electricity generation, consumption, and pricing trends to promote sustainable energy management practices. By leveraging big data analytics and integrating data from various sources, the project seeks to overcome the limitations of traditional approaches and contribute to a more comprehensive and accurate understanding of energy consumption dynamics in the United States.

1.3 Project Application and Impact

The application of big data analytics in this project has significant potential for generating valuable outcomes. By creating a comprehensive dataset through the integration of various data sources, the project enables the examination of energy consumption patterns at

both national and state levels. The utilization of big data technologies like Apache Kafka and Amazon S3 allows for efficient data processing, storage, and querying, facilitating the generation of meaningful insights. The project aims to provide interactive visualizations and reports that highlight energy consumption trends, enabling stakeholders to make informed decisions regarding energy management and conservation. Ultimately, the project seeks to have a positive impact on promoting sustainable energy practices and contributing to the understanding of energy consumption dynamics in the United States.

1.4 Data Description

The EIA possesses a vast collection of datasets covering various aspects of energy production, consumption, and distribution in the United States. These datasets help policymakers, researchers, businesses, and the general public acquire a deeper understanding of the energy industry and make more informed decisions. Annual Energy Outlook, Monthly Energy Review, State Energy Data System, and International Energy Statistics are frequently consulted EIA datasets. The amount of data collected from this website was around 2-3 GB.

1.5 Project Deliverables

Table 1

Project Deliverables

Phases	**Deliverables**	**Scheduled Date**
Planning & designing	Topic	02-03-2023
	Finding Resources for Dataset	02-08-2023
System Implementation	AWS Sign Up	02-18-2023
	Boto3	02-18-2023
	Kafka	03-03-2023
	S3 bucket Creation	03-18-2023
	Sending API data to an AWS Kafka server using a producer	03-23-2023
	Loading data into an S3 bucket using a Kafka consumer	03-29-2023
	Extracting information from an S3 bucket using Crawler Jobs	04-04-2023
	AWS Glue ETL Data Catalog for pre-processing	04-10-2023

	Querying Using Amazon Athena	04-12-2023
	Visualization Using Tableau	04-20-2023
	Hosting Static Website	05-01-2023
Documentation	Project Report	05-07-2023
	Project Power Point Presentation	05-08-2023

Chapter 2. Project Background and Related Work

2.1 Background and Used Technologies

The US Energy Information Administration (EIA) is a government agency which is collecting, analyzing, and disseminating energy-related data and information. This website provides information about energy which is independent and impartial policymaking, efficient market, and public understanding of energy resources and their interaction with the economy as well as the environment. To collect and analyze energy-related data, the EIA utilizes various technologies and tools. These include data collection and processing software, statistical analysis software, and database management systems. Additionally, the EIA makes use of various energy data sources, such as surveys, public data sources, and data provided by energy companies.

2.2 State-of-the-Art

Several products are currently available in the market that analyzes energy consumption patterns and provides valuable insights into electricity generation, consumption, and pricing trends. Some of the popular products in the market include:

- EnergyCAP - EnergyCAP is an energy management software that provides an automated, streamlined approach to energy tracking, utility bill processing, and reporting. This software is designed to help organizations reduce energy consumption, lower costs, and meet sustainability goals.
- Schneider Electric EcoStruxure Power Monitoring Expert - EcoStruxure Power Monitoring Expert is an energy management system that provides real-time

monitoring and analysis of energy usage data. The system is designed to help organizations improve energy efficiency, reduce energy consumption, and lower costs.

- Siemens EnergyIP Analytics - Siemens EnergyIP Analytics is a of software applications that help utilities and energy companies optimize energy management and grid operations. The software provides real-time analysis of energy consumption data, enabling users to make informed decisions about energy generation, distribution, and pricing.
- Opower - Opower is a cloud-based energy management platform that provides utilities and energy companies with real-time energy usage data. The platform is designed to help organizations reduce energy consumption, lower costs, and improve customer engagement.

While these products offer valuable insights into energy consumption patterns, our project aims to overcome the limitations of traditional approaches by leveraging big data analytics and integrating data from various sources to provide a comprehensive and accurate understanding of energy consumption dynamics in the United States.

2.3 Literature Survey

The paper "Understanding Household Energy Consumption Behavior: The Contribution of Energy Big Data Analytics" discusses the importance of energy big data analytics in comprehending and analyzing the behavior of households in terms of energy consumption. The authors highlight the increasing accessibility of energy data derived from smart meters and sensors, which facilitates the utilization of big data analytics methodologies to obtain discernment into energy consumption patterns and conduct within households. The authors engage in a discourse regarding the difficulties that arise in comprehending energy consumption behavior, including but not limited to the complexity of data, the magnitude of data, and the necessity for data preprocessing methodologies. The article emphasizes the

significance of big data analytics in revealing consumption patterns, detecting opportunities for energy conservation, and forecasting energy demand. Furthermore, the authors investigate diverse data mining and machine learning algorithms utilized in scrutinizing energy data, including clustering, classification, and regression. The authors provide instances of research works that have employed energy big data analytics to assess consumer conduct, enhance energy efficiency, and devise focused interventions. In summary ,this study makes a noteworthy contribution to the understanding of the role of big data analytics in gaining insights into the behavior of household energy consumption.

The paper titled "KAFKA: The modern platform for data management and analysis in the big data domain" addresses the utilization of Apache Kafka as a contemporary platform for managing and analyzing data in the field of big data. The authors analyze the salient characteristics and functionalities of Kafka, including its decentralized messaging infrastructure, resilience to errors, capacity for expansion, and ability to handle data in real-time. The authors emphasize Kafka's capability to effectively and dependably manage substantial quantities of data streams, rendering it appropriate for a range of applications such as real-time analytics, event-driven architectures, and data integration. The article additionally discusses the structure and constituents of Kafka, including producers, consumers, topics, partitions, and brokers, giving an in-depth understanding of the system. Additionally, the authors provide multiple instances and illustrations of Kafka's implementation across diverse sectors, showcasing its capacity to tackle significant data-related obstacles.

The paper "Kafka: a Distributed Messaging System for Log Processing" focuses on Apache Kafka's function as a distributed messaging system tailored to log processing. The importance of an effective and scalable messaging system is highlighted, as is the difficulty of dealing with enormous volumes of log data created by distributed systems. They

emphasize the architecture and distinguishing characteristics of Kafka, including its fault tolerance, high throughput, and real-time data processing capabilities, among others. Kafka's scalability, capacity to store and replicate log data across distant nodes, and fault tolerance are highlighted in this study. Furthermore, the authors explore various use cases of Kafka in log processing, including real-time monitoring, data integration, and data analysis. They discuss the benefits of Kafka's publish-subscribe model and partitioning mechanism for parallel processing and scalability. The paper also provides insights into the performance benchmarks and case studies that demonstrate the effectiveness of Kafka in log processing scenarios. Overall, this paper serves as a valuable resource for understanding the significance of Kafka as a distributed messaging system has been developed to enhance the efficiency of log processing.

The paper "Big data analytics in power distribution systems" focuses on the application of big data analytics in the field of power distribution systems. The authors analyze the difficulties presented by the growing quantity, diversity, and speed of data produced within power distribution networks. The utilization of big data analytics techniques, has been identified as a promising approach to tackle the mentioned challenges and enhance the effectiveness, dependability, and safety of power distribution systems. This paper investigates diverse applications of big data analytics in power distribution, encompassing load prediction, fault identification, outage administration, and demand reaction. The authors provide a comprehensive overview of the data sources, data preprocessing methods, and analytics algorithms that are frequently employed in this field.

Chapter 3 System Requirements and Analysis

3.1 Domain and Business Requirements

This chapter focuses on identifying the domain and business requirements for our project, which aims to analyze energy consumption patterns in the United States using big

data analytics. In the domain, we need to collect data from various sources like real-time data feeds, APIs, and web scraping techniques were included to create a comprehensive dataset that accurately captures energy consumption metrics at the national and state levels. From a business perspective, our goal is to provide valuable insights into electricity generation, consumption, and pricing trends by leveraging data from the US Energy Information Administration (EIA). To achieve this, we will utilize big data technologies like Apache Kafka and Amazon S3 for efficient data streaming, storage, and querying using Amazon Athena.

3.2 Customer-Oriented Requirements

Our project is driven by the desire to address critical issues related to energy consumption and provide insights that inform decision-making processes. By analyzing energy consumption patterns, we aim to contribute to sustainable energy management practices. Understanding electricity generation, consumption, and pricing trends is crucial for achieving this goal. Therefore, our project aims to leverage big data analytics and integrate data from various sources to overcome the limitations of traditional approaches and provide a comprehensive and accurate understanding of energy consumption dynamics in the United States.

3.3 System Function Requirements

To meet the system function requirements, we will utilize the AWS CLI for API data upload and perform Exploratory Data Analysis (EDA) and Extract Load Transform (ELT) processes. These processes involve cleaning and transforming the dataset by removing duplicates, handling missing values, and ensuring data type consistency and accuracy. Additionally, we will implement Kafka producers to stream data from Jupyter Notebook to the Kafka cluster, and use Boto3 and AWS S3 to securely store the collected data.

3.4 System Behavior Requirements

Our system needs to efficiently scan and organize the data stored in the S3 bucket using AWS Crawler, extracting metadata to facilitate easy discovery and cataloging. We will also utilize AWS Glue to generate a schema that defines the structure and organization of the data, enabling efficient data processing, transformation, and analysis. For querying the data, AWS Athena will be used, providing users with the ability to execute SQL queries on the data stored in S3 for fast and interactive analysis.

3.5 System Performance and Non-Functional Requirements

To ensure system performance, we need to establish smooth data transfer, reliable transmission, and continuous data flow between Jupyter Notebook and the Kafka messaging system. Our system should also provide durability, scalability, and secure storage for the collected data in the AWS S3 bucket. Furthermore, efficient data ingestion, exploration, and extraction should be facilitated through AWS Crawler and AWS Glue. AWS Athena will offer a serverless query service with fast and interactive analysis capabilities.

3.6 System Context and Interface Requirements

To enable seamless integration and data visualization, we will establish a connection between Amazon Athena and Tableau. This integration will empower users to create dynamic visualizations, interactive dashboards, and insightful reports based on the queried data.

3.7 Technology and Resource Requirements

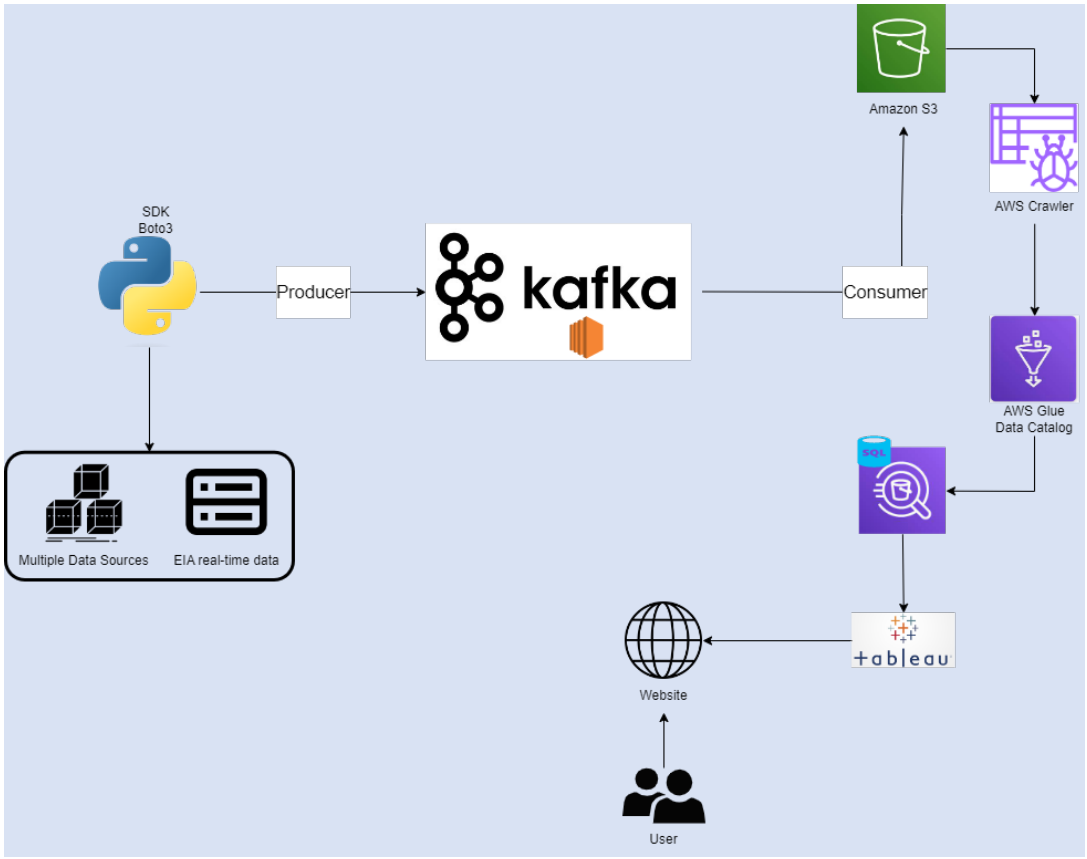
Our project relies on several key technologies, including Apache Kafka, Amazon S3, AWS Crawler, AWS Glue, AWS Athena, Jupyter Notebook, and Tableau. These technologies, along with the necessary infrastructure and resources, form the foundation of our project.

Chapter 4 System Design

4.1 System Architecture Design

Figure 1

ETL Architecture



Web scraping using RESTful APIs was employed to extract data from the Energy Information Administration (EIA). The process involved configuring the API endpoints and implementing logic to retrieve the desired data efficiently. By leveraging RESTful APIs, specific data points were accessed from the EIA's vast repository of energy-related information. This approach ensured that the extracted data was up-to-date and reliable, as it was obtained directly from the source.

To streamline the data flow and enable seamless data storage, a Kafka setup was established. The producer side was equipped with the necessary logic to efficiently input the extracted data into Amazon Web Services (AWS). Kafka acted as a message broker, ensuring

reliable communication between the producer and consumer. This setup facilitated the transfer of data from the scraping process to AWS, allowing for further processing and analysis.

Upon saving the data in the S3 container, AWS crawler tasks were executed to perform analytics using AWS Athena. The saved data was retrieved from the S3 container, and AWS crawler tasks were initiated to extract meaningful insights. AWS Athena, a serverless query service, was utilized to efficiently query and analyze the data stored in S3. By leveraging the power of AWS services, valuable insights were derived, contributing to the organization's decision-making processes and overall understanding of energy-related trends and patterns.

4.2 System Interface And Connectivity Design

The client-side application was developed to seamlessly save the extracted data to an S3 container within AWS. This storage solution provided a scalable and reliable infrastructure for data persistence. By leveraging AWS services, such as S3, the extracted data could be securely stored and easily accessible. The client-side application ensured that the data was properly formatted and saved in the designated S3 container, enabling future retrieval and analysis.

The results obtained from AWS Athena were transformed into visually appealing charts, graphs, and dashboards in Tableau. These visualizations were hosted on a static webpage, accessible to authorized users, creating a user-friendly environment for data exploration and analysis.

4.3 Design Problems, Solutions, And Patterns

The following table presents the design challenges we faced during the system implementation and the corresponding solutions and design patterns we applied to overcome them.

Table 2

Design Problem, Solutions and Patterns

Design Problems	Solutions	Patterns
Extracting Data from EIA	Configuring API Endpoints and Implementing Efficient Retrieval Logic	RESTful API
Ensuring Reliable Data Transfer and Storage	Establishing a Kafka Setup as a Message Broker	Message Broker Pattern
Enabling Efficient Querying and Analysis of Data Stored in S3	Using AWS Athena, a Serverless Query Service	Serverless Pattern
Saving Extracted Data to S3 Container in AWS	Developing a Client-Side Application with Proper Formatting and Storage Logic	AWS S3 Pattern
Transforming Query Results into Visualizations	Using Tableau to Create Charts, Graphs, and Dashboards	Visualization Pattern

Chapter 5 System Implementation

5.1 System Implementation Summary

The team utilized the CLI of AWS for API data upload. Initially, the AWS CLI was set up in our local environment using the terminal and was able to create the user successfully in our AWS account with the help of SSH key provided by AWS. Credentials for the user were generated, including the AWS access key and access secret Key, to save the files in standard delimiter format also known as CSV. These credentials were then imported into the AWS CLI. The configuration file was verified to ensure that all the settings were correctly applied. This setup allowed us to effectively manage the upload of the sizable dataset using the AWS CLI.

During the Exploratory Data Analysis (EDA) and Extract Load Transform (ELT) process, we conducted various cleaning and transformation steps on the dataset. This included removing duplicate entries, handling missing values, and correcting data types.

In particular, we focused on modifying the columns with appropriate data types to ensure consistency and accuracy. One key adjustment involved the column "period," which initially had a string data type and included the "UTC" identifier along with the timestamp. To improve the data quality, we removed the "UTC" identifier and converted the "period" column to the timestamp data type, enabling easier analysis and manipulation of time-related data.

These cleaning and transformation steps were crucial in preparing the dataset for further analysis and ensuring the data was accurate, consistent, and suitable for subsequent processing and insights extraction.

Real-Time Data Sample

The Datasets we have gathered are from EIA website using API and have real-time data. The below figure is the sample data of EIA for electricity data for all sectors.

Figure 2

Dataset Sample From EIA

eia + Sources & Uses + Topics + Geography Search eia.gov

chart	period	location	stateDescription	sectorid	sectorDescription	fueltypeid	fuelTypeDescription	ash-content	ash-content-units	consumption-for-eg	consumption-for-eg-units	consumption-for-eg-btu	consumption-for-eg-btu-units	consumption-uto	consumption-uto-units	consumption-uto-btu
	2022-12	OH	Ohio	99	All Sectors	OOG	other gases	0	percent	933.186	thousand Mcf	0.33958	million MMBtu	3855.073	thousand Mcf	1.20195
	2022-12	IL	Illinois	99	All Sectors	DPV	estimated small scale solar photovoltaic		percent		thousand physical units		million MMBtu		thousand physical units	
	2022-12	IL	Illinois	99	All Sectors	DFO	distillate fuel oil	0	percent		thousand short tons		million MMBtu		thousand short tons	
	2022-12	IL	Illinois	99	All Sectors	COW	all coal products	9.89	percent	2155.959	thousand short tons	37.73309	million MMBtu	121.666	thousand short tons	2.51565
	2022-12	IL	Illinois	99	All Sectors	COL	coal, excluding waste coal	9.89	percent	2155.959	thousand short tons	37.73309	million MMBtu	121.666	thousand short tons	2.51565
	2022-12	IL	Illinois	99	All Sectors	BIT	bituminous coal	19.14	percent	726.47	thousand short tons	12.99405	million MMBtu	109.49	thousand short tons	2.305
	2022-12	IL	Illinois	99	All Sectors	BIS	bituminous coal and synthetic coal	19.14	percent	726.47	thousand short tons	12.99405	million MMBtu	109.49	thousand short tons	2.305

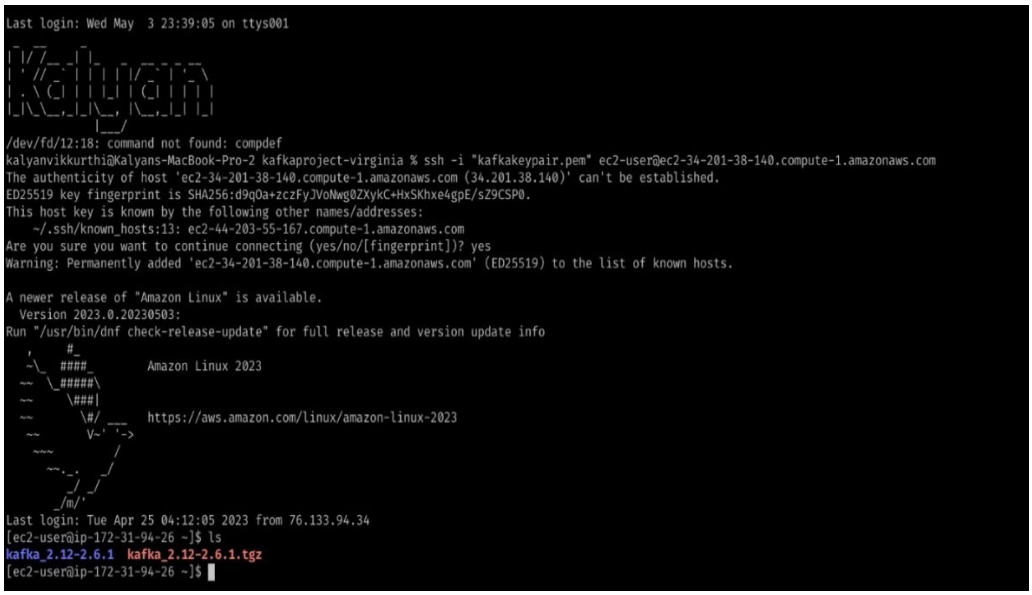
Kafka Installation on AWS EC2 Instance

The initial step involved importing the data obtained from the Energy Information Administration (EIA) into the Jupyter Notebook. The Jupyter Notebook provided an interactive environment that facilitated efficient data transformation, manipulation, and

exploratory analysis.To ensure smooth data transfer, a Kafka producer was implemented. Acting as a bridge between the Jupyter Notebook and the Kafka messaging system, the producer component effectively streamed the collected data from the notebook to the Kafka cluster. This seamless communication ensured reliable transmission and continuous data flow.

Figure 3

Kafka Connection

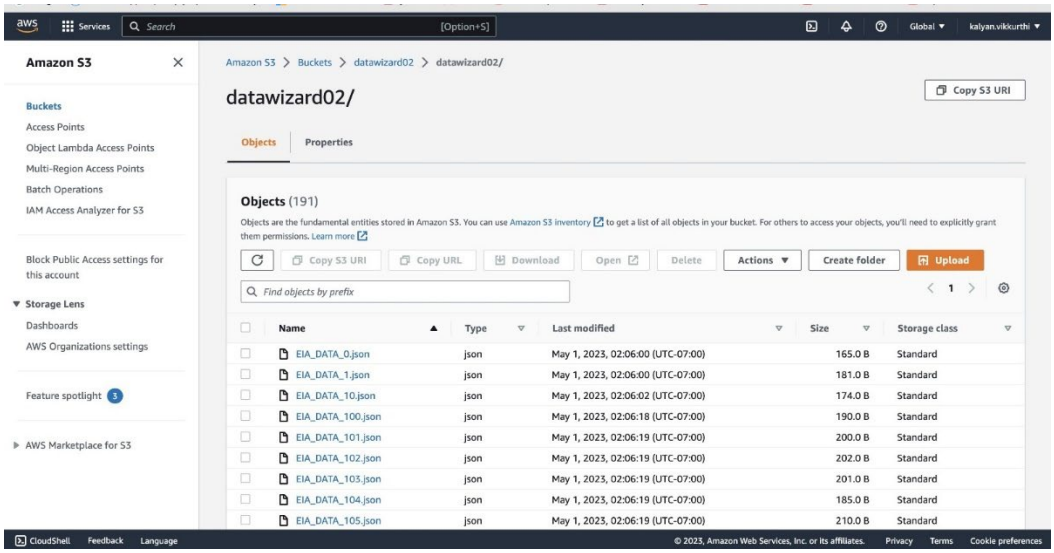


AWS S3 Bucket to store data

Utilizing Boto3, the consumer component efficiently processed the incoming data received from Kafka and performed the necessary operations to store it in the designated S3 bucket. This process involved establishing a connection with the S3 service, ensuring proper authentication and authorization to access the bucket. Once the connection was established, the consumer component used Boto3 to write the data received from Kafka into the S3 bucket. This storage mechanism provided durability, scalability, and secure storage for the collected data.

Figure 4

AWS S3 Bucket

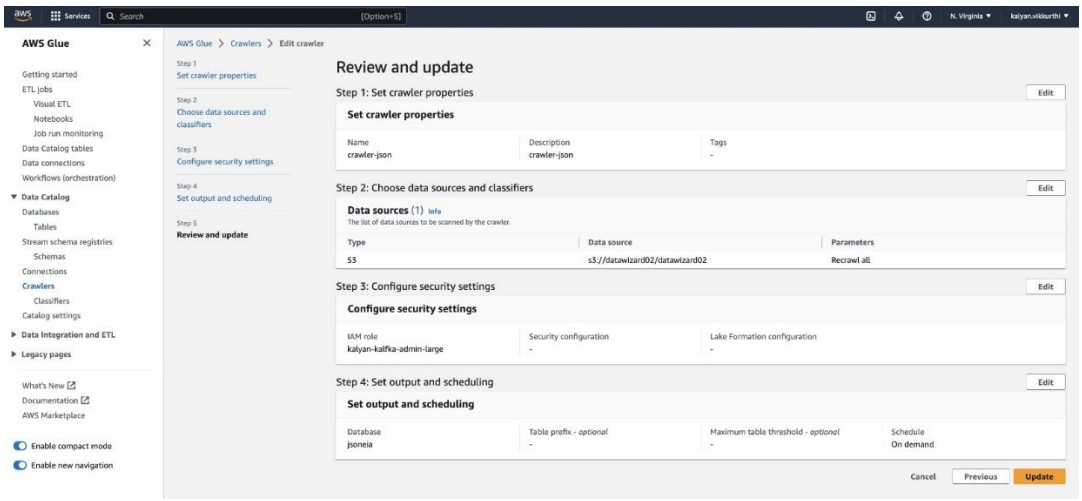


AWS Crawler

The data stored in the S3 bucket is then crawled into AWS Crawler for further processing and analysis. AWS Crawler scans the data in the S3 bucket, extracting metadata and organizing the dataset to facilitate easy discovery and cataloging. This automated process streamlines data ingestion and prepares the dataset for subsequent analytics tasks, enabling efficient data exploration and extraction of valuable insights.

Figure 5

AWS Crawler

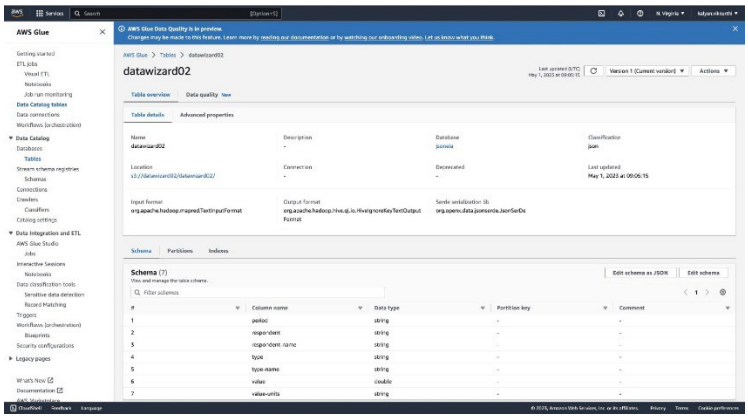


AWS Glue for Catalogs

In AWS Glue, a schema is generated to define the structure and organization of the data. This schema creation process in AWS Glue allows for a standardized representation of the data, enabling efficient data processing, transformation, and analysis tasks.

Figure 6

AWS Glue for Catalogs



AWS Athena for Querying

AWS Athena is utilized for querying the data. With AWS Athena, users can execute SQL queries on the stored data in S3, enabling fast and interactive analysis without the need for infrastructure management. This serverless query service simplifies the data exploration process, allowing users to derive valuable insights and perform data-driven decision-making.

Figure 7

Data Querying in Athena

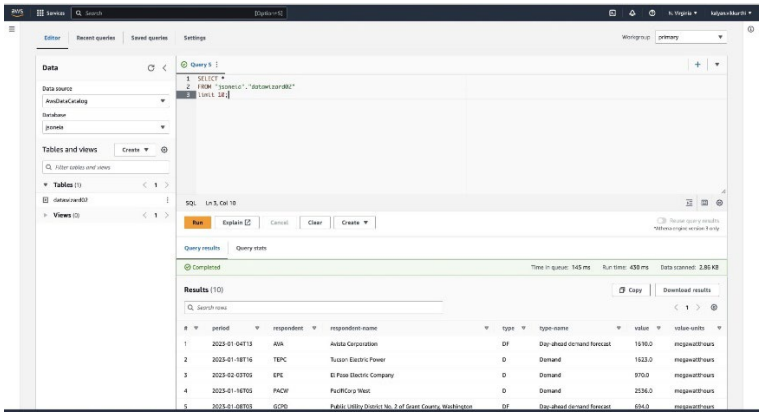


Tableau Connection from Amazon Athena

The connection between Amazon Athena and Tableau is established, enabling seamless integration for data visualization and analysis. By connecting Tableau to Amazon Athena, users gain the ability to create dynamic visualizations, interactive dashboards, and insightful reports based on the queried data. This integration streamlines the process of visualizing and exploring data, empowering users to uncover meaningful patterns, trends, and insights for effective decision-making.

Figure 8

Connecting Tableau to Athena

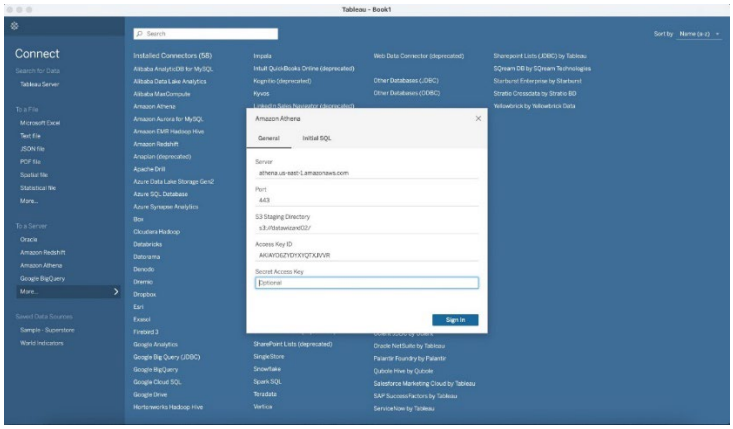
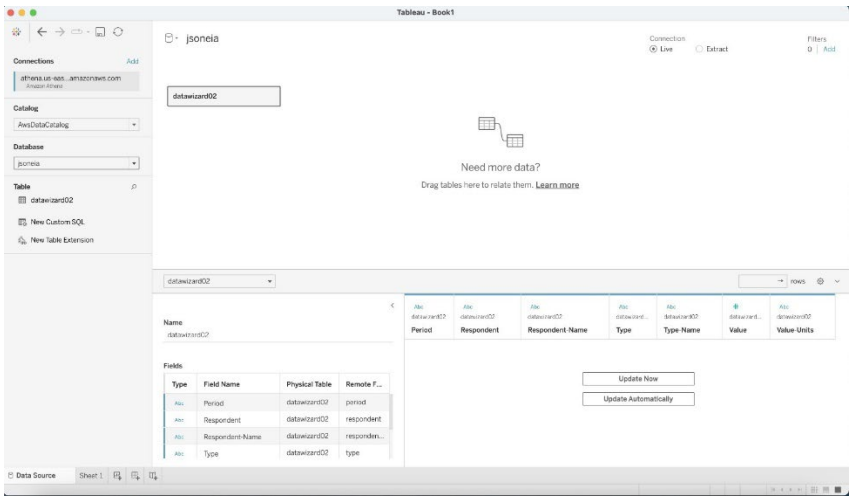


Figure 9

Data Loaded in Tableau Desktop

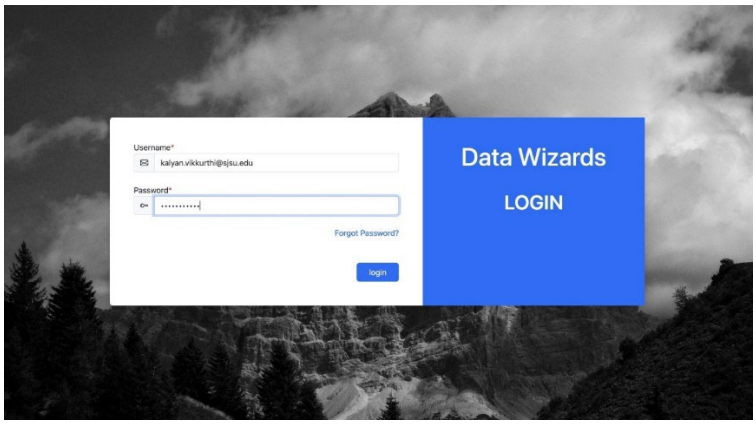


Website Login Page

An authentication-enabled login page is developed for authorized users to access the website. This login page ensures that only authorized individuals can gain entry to the website, providing a secure and controlled environment for accessing the hosted content.

Figure 10

User Login Page

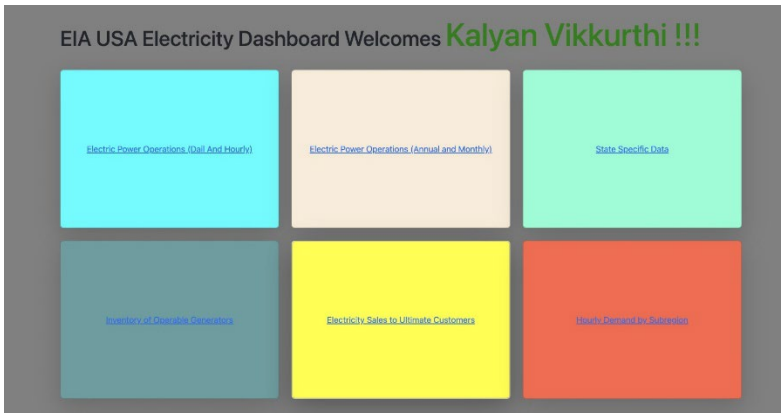


Electricity Analysis User Dashboard

An electricity analysis user dashboard is designed, providing a dedicated interface for users to access and explore electricity-related data. This dashboard presents visualizations, analytics, and insights, empowering users to gain a comprehensive understanding of electricity trends, patterns, and metrics.

Figure 11

User Dashboard

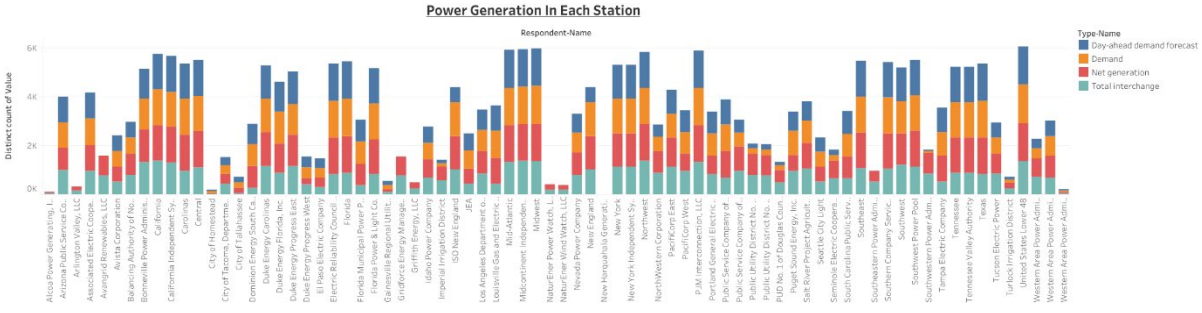


Data Visualization and analysis

The stacked bar chart in this scenario provides a visual representation of different count values on the Y-axis, while the X-axis displays various power generating plants in the US. Each category of data is assigned a distinct color for easy differentiation. The blue bars represent the "Day-Ahead Demand Forecast," which indicates the predicted power demand calculated in advance. The orange bars represent the "Demand," representing the actual power demand on a given day. The red bars represent "No Generation," indicating instances where power generation is absent at specific power plants. Finally, the aqua green bars represent the "Total Interchange," representing the total power exchange or interchange between different power grids or plants. By analyzing the stacked bar chart, you can observe and compare the distribution and frequency of these categories across different power generating plants in the US.

Figure 12

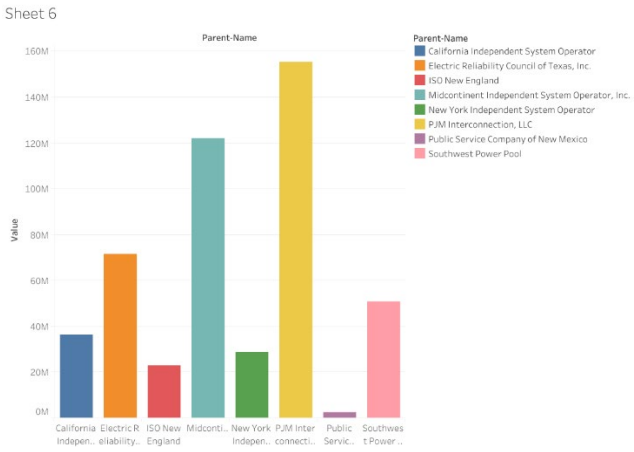
Visualization of Power Generation in Different Power Plants



The bar graph in this scenario displays megawatts values on the Y-axis and represents different parent organizations involved in electricity generation on the X-axis. Each organization is distinguished by a unique color. The Y-axis values represent various metrics related to electricity generation, specifically the megawatts of power generated. By examining the graph, patterns, differences, and anomalies can be identified among the organizations, providing insights into their performance in electricity generation. The visual nature of the graph enhances the clarity and attractiveness of the data presentation.

Figure 13

Visualization of Power Generation in Parent Power Plants



In the below bar graph the megawatts is taken on the Y-axis and the X- axis represents the information of different power generation types namely coal, wind, solar, petroleum, nuclear, natural gas, hydro, and other resources. This allows us for easy comparison and analysis of power generation levels across different sources. By examining the graph, insights can be gained regarding the relative contributions of each power generation type, enabling the identification of the most and least amount of power generated by various sources. Overall, a clear understanding of the energy mix and diversity of the power generation landscape is provided by this visualization.

Figure 14

Visualization of Hourly Power Generation Based on Type

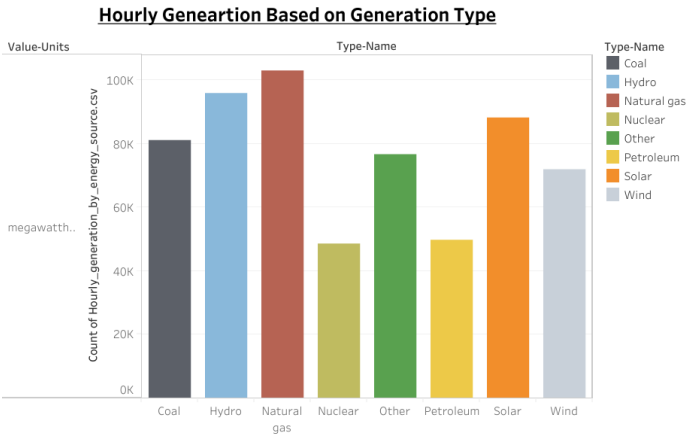
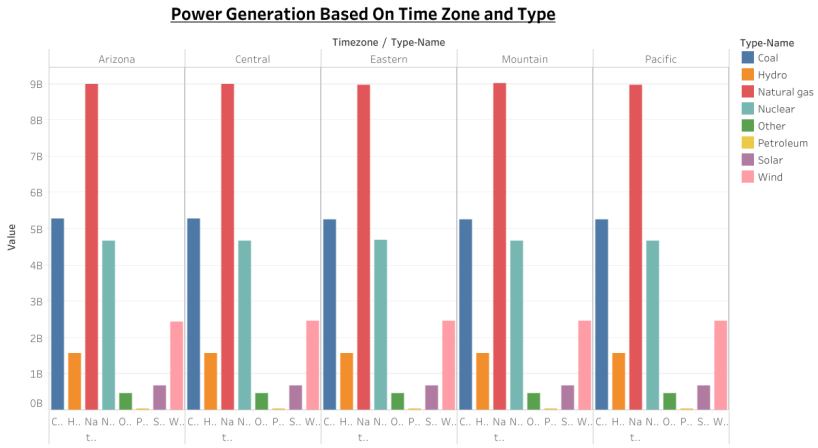


Figure 15

Visualization of Power Generation Based On Time-Zone



5.2 System Implementation Issues And Resolutions

Below is a table outlining the challenges we encountered during the system implementation process and the corresponding solutions we applied to address them.

Table 3

System Implementation Issues

System Component	Title	Issues/Problems	Resolutions/Solutions
RESTful APIs	Authentication and Authorization	Security vulnerabilities due to lack of authentication and authorization mechanisms	Implementation of authentication and authorization protocols such as OAuth or JSON Web Tokens (JWTs) to secure API endpoints
Kafka	Scalability and Performance	Performance issues caused by a high volume of messages being produced and consumed	Scaling of Kafka cluster horizontally by adding more brokers to distribute message load and optimization of configuration settings such as the number of partitions

Amazon S3	Data Management and Access Control	Inefficient data retrieval and management due to improper data organization and bucket design	Proper organization of data into logical partitions and implementation of bucket policies and lifecycle rules to automate data management tasks
AWS Crawler	Data Integration and Quality	Inaccurate data extraction and parsing due to the complexity of the source data	Fine-tuning of crawler settings to customize the extraction process and implementation of data transformation scripts to parse and transform extracted data
AWS Athena	Query Performance and Cost	Slow query performance due to large datasets and inefficient query design	Optimization of query design by using partitioning, indexing, and other techniques, and implementation of data compression and format conversion to reduce query time
Tableau	Data Visualization and Reporting	Performance issues and data visualization problems due to complex and poorly designed dashboards	Simplification of dashboard design, reduction of data granularity, and implementation of caching and data extracts to improve dashboard performance and responsiveness.

5.3 Used Technologies and Tools

The process involved using RESTful APIs to extract data from the EIA website, with Kafka facilitating data transfer. Amazon S3 served as a scalable storage container. AWS Athena enabled querying and analysis of data stored in S3, while Tableau created interactive visualizations. Automated AWS Crawler tasks were employed for insight extraction. These tools formed an efficient pipeline for data extraction, storage, analysis, and visualization.

RESTful APIs and Kafka facilitated data retrieval and transfer. Amazon S3 ensured reliable storage. AWS Athena enabled powerful analysis, while Tableau created interactive visualizations. Automated AWS Crawler tasks enhanced insight extraction. Together, these tools supported data-driven decision-making and enhanced understanding of energy-related trends.

Table 4

Tools and Technologies Used

Tool / Technology	Use/Description
RESTful APIs	Used to extract data from the Energy Information Administration (EIA) website.
Kafka	The system is set up as a message broker with the purpose of optimizing the transmission of data between the producer and consumer.
Amazon S3	Storage container within AWS to save the extracted data.
AWS Crawler	Automated tasks to extract valuable insights from the data stored in the S3 container.
AWS Athena	Serverless query service employed for querying and analyzing the data stored in S3.
Tableau	Data visualization tool integrated with AWS Athena to create interactive and insightful visualizations.

Chapter 6 System Testing and Experiment

6.1. Testing report

Table 5

Test Case 1

Test No	1
Test Title	Pulling data from EIA website using RESTful APIs
Test Purpose	To extract data from the EIA website using RESTful APIs
Test Setup	Jupyter Notebook
Prerequisites	Internet connection, EIA API credentials
Procedure	Utilize appropriate RESTful APIs to extract data from the EIA website

Checks	Verify the retrieved data in the Jupyter Notebook for completeness and accuracy
Expected Results	Able to load the complete data into the Jupyter notebook without any data leakage as per the provided parameters
Result	Able to load the complete data into the Jupyter notebook without any data leakage as per the provided parameters
Reason for Failure	-
Remarks	Successfully completed

Table 6

Test Case 2

Test No	2
Test Title	Sending data to Kafka cluster with the help of producer
Test Purpose	To test the ability to send obtained data to the Kafka cluster using a producer
Test Setup	Jupyter Notebook
Prerequisites	Kafka cluster, producer component setup
Procedure	Data is sent to Kafka Cluster with the help of producer component
Checks	Verify the successful transfer of complete data from the producer to the consumer component
Expected Results	To receive complete data sent by the producer to the consumer component without any data loss
Result	Able to receive complete data sent by the producer to the consumer component without any data loss
Reason for Failure	-
Remarks	Successfully completed

Table 7

Test Case 3

Test No	3
Test Title	Storing results into S3 bucket

Test Purpose	To test the ability to store the obtained results into an S3 bucket
Test Setup	Jupyter Notebook
Prerequisites	AWS S3 setup, consumer component setup
Procedure	Utilize the consumer component to write the obtained data into the designated S3 bucket
Checks	Verify the successful storage of the data by checking if it is correctly written into the S3 bucket
Expected Results	Consumer component should successfully write the data into the S3 bucket
Result	Consumer component is able successfully write the data into the S3 bucket
Reason for Failure	-
Remarks	Successfully completed

Table 8

Test Case 4

Test No	4
Test Title	Querying results in AWS Athena
Test Purpose	To test the querying capabilities in AWS Athena
Test Setup	AWS Athena, S3
Prerequisites	AWS S3 bucket, crawler job, schema setup
Procedure	Create a crawler job in AWS to read the data from the S3 bucket and generate a schema in AWS Data Catalog
Checks	Check if the crawler job successfully reads the data and creates a schema in AWS Data Catalog
Expected Results	Crawler job should successfully read the data from the S3 bucket and create a successful schema in AWS Data Catalog
Result	-
Reason for Failure	-
Remarks	Successfully completed

Table 9

Test Case 5

Test No	5
Test Title	Viewing data statistics in deployed website
Test Purpose	To test the ability to view data statistics in the deployed website
Test Setup	Deployed website
Prerequisites	Authorized user login credentials
Procedure	Login to the deployed website as each team member and verify if they can see their user dashboard
Checks	Ensure that each team member can successfully log in and view their own user dashboard
Expected Results	Each team member should successfully log in and can see their own user dashboard
Result	-
Reason for Failure	-
Remarks	Successfully completed

Chapter 7 Conclusion and Future Work

7.1 Conclusion

In summary, the experiment shows the effectiveness of big data analytics in comprehending and examining patterns of energy consumption. The project successfully extracted, converted, and analyzed data from the EIA website by employing cutting-edge technologies including Apache Kafka, Amazon S3, AWS Athena, and Tableau. The created user dashboard offers a simple interface for perusing and gleaning useful information from data pertaining to electricity. These discoveries could lead to significant adjustments in consumer behavior, policy-making procedures, and methods for managing energy. The

project contributes significantly to improving our knowledge of the dynamics of energy usage and laying the foundation for a more sustainable and effective energy future by utilizing the capabilities of big data analytics.

7.2 Future Scope

The future scope of this project involves expanding the data sources to include renewable energy data, weather patterns, economic indicators, and consumer behavior. This expansion will provide a more comprehensive understanding of energy consumption trends and enable accurate forecasting. Integration of machine learning algorithms will enhance the project's predictive capabilities, allowing for advanced energy forecasting models and optimization of energy distribution. Real-time monitoring systems and alerts can be implemented to capture and analyze real-time energy usage data, enabling proactive energy management and waste reduction. Additionally, fostering collaborative energy management among stakeholders can lead to the implementation of energy-saving initiatives and policy interventions for a more sustainable energy ecosystem. Overall, continuous improvement, advanced technologies, and collaboration are key aspects of the future scope of this project.

Git

The Project report, Project Plan, Presentation, source code, website embed code, Tableau Workbook will be uploaded on GitHub repository.

GitHub link: https://github.com/KalyanVikkurthi002/Datawizards-kafka

References

Zhou, K., & Yang, S. (2016). Understanding household energy consumption behavior: The contribution of energy big data analytics. *Renewable & Sustainable Energy Reviews*, *56*, 810–819. https://doi.org/10.1016/j.rser.2015.12.001

Yu, N., Shah, S., Johnson, R. L., Sherick, R., Hong, M., & Loparo, K. A. (2015). *Big data analytics in power distribution systems*. https://doi.org/10.1109/isgt.2015.7131868

Shree, R., Choudhury, T., Gupta, S., & Kumar, P. (2017). KAFKA: The modern platform for data management and analysis in big data domain. In *2017 2nd International Conference on Telecommunication and Networks (TEL-NET)*. https://doi.org/10.1109/tel-net.2017.8343593

Kreps, J., Narkhede, N., & Rao, J. (2011). Kafka: a Distributed Messaging System for Log Processing. Proceedings of the NetDB, 11(10), 1-7. https://pages.cs.wisc.edu/~akella/CS744/F17/838-CloudPapers/Kafka.pdf

https://www.eia.gov/opendata/browser/electricity/electric-power-operational-data?frequency=monthly&data=ash-content;consumption-for-eg;consumption-for-eg-btu;consumption-uto;consumption-uto-btu;cost;cost-per-btu;generation;heat-content;receipts;receipts-btu;stocks;sulfur-content;total-consumption;total-consumption-btu;&start=2020-01&end=2022-12&sortColumn=period;&sortDirection=desc;