**Phase 5**

**PROJECT DOCUMENTATION & SUBMISSION**

|  |  |
| --- | --- |
| **Date** | **31-10-2023** |
| **Team ID** | **673** |
| **Project Name** | **Big Data Analysis with IBM Cloud Databases** |

**Project Title:**

**Big Data Analysis in IBM Cloud Database**

**Problem Statement**

Dive into the world of big data analysis with IBM Cloud Databases. Uncover hidden insights from vast datasets, from climate trends to social patterns. Visualize your findings and derive valuable business intelligence. Embark on data-driven adventures, exploring the endless possibilities of big data!

**Problem identified:**

Our problem statement focuses on this critical need for effective big data analysis with IBM Cloud Database. While vast amounts of data are available, the ability to harness its potential for informed decision-making and innovation remains elusive for many. Deploying a reliable and accessible system for big data analysis that can efficiently process, store, and analyze these massive data sets is a complex and unmet challenge.

**Introduction:**

In response to the ever-increasing volume of data generated in the digital age, our project, "Big Data Analysis with IBM Cloud Database," strives to provide an innovative and practical solution for harnessing the power of big data. Leveraging the capabilities of IBM Cloud's database services, we aim to facilitate the creation, analysis, and maintenance of cutting-edge data-driven insights.  
  
The core challenge we address is the efficient analysis of vast and complex datasets, a task that has become increasingly daunting in the era of big data. Traditional data processing methods are often inadequate to handle the sheer magnitude of information at our disposal. To overcome this challenge, we turn to the field of big data analysis, which promises scalability, adaptability, and the potential to uncover valuable patterns and insights.  
  
Our project unfolds in a meticulously structured manner, beginning with data acquisition and preparation and progressing to the selection of appropriate analytical tools and the deployment of data analysis pipelines. We make use of IBM Cloud's robust database infrastructure to house and manage our data, ensuring its security and accessibility.  
  
Furthermore, our project capitalizes on the advanced analytics and machine learning capabilities within the IBM Cloud ecosystem to uncover hidden patterns and correlations within the data. We are committed to creating a system that not only empowers users but also makes data-driven insights accessible to applications through secure APIs, enabling real-time data analysis and decision-making.  
  
In the following sections, we will delve into the intricate details of our project, shedding light on the methods, tools, and datasets we employ to build a comprehensive and efficient big data analysis system. By addressing the burgeoning challenge of handling and analyzing vast datasets, we aim to equip individuals and organizations with the means to derive valuable insights in an ever-evolving digital landscape.  
  
Data: Our project relies on a diverse and extensive dataset, comprising a wide array of data points, attributes, and sources. This dataset forms the cornerstone of our big data analysis efforts, serving as the basis for uncovering valuable insights and trends.

**LITERATURE SURVEY**

**1.“PROTECTING BIG DATA IN PUBLIC CLOUD BY ENHANCED RBAC”,**

**J.L. Joneston Dhas[2022]**  
  
 The utilization of Enhanced Role-Based Access Control (RBAC) for safeguarding big data in public cloud environments presents a potent strategy for data security. This approach allows for refined access control, bolstering data protection at granular levels. Benefits include heightened security through improved access management, compliance adherence, and the mitigation of unauthorized data breaches. The scalability of RBAC aligns well with managing expansive datasets. Nonetheless, the implementation complexity and potential administrative overhead associated with role management are key challenges. Misconfigurations and performance overhead can arise, necessitating meticulous deployment and monitoring. Despite these drawbacks, leveraging Enhanced RBAC stands as a robust solution to fortify data in public cloud settings.

**2.”GIS CLOUD COMPUTING BASED GOVERNMENT BIG DATA ANALYSIS PLATFORM”Qiang Wang[2021]**  
  
 The integration of Geographic Information Systems (GIS), cloud computing, and government big data analysis has resulted in a transformative platform. This convergence facilitates efficient analysis of vast spatial data sets for governmental decision-making and policy formulation, While this platform empowers governments with advanced analytical capabilities, addressing challenges such as data security, interoperability, and compliance remains imperative for its successful deployment.

**3.”CHALLENGES IN BIG DATA ANALYTICS AND CLOUD COMPUTING”,**

**P. Madana Mohan[2023]**  
 Cloud computing's synergy with data analytics has reshaped the way organizations leverage and process data. By providing scalable and flexible resources, cloud platforms have enabled businesses to handle large volumes of data for analytical purposes. This dynamic environment accelerates data processing, enabling real-time or near-real-time insights for informed decision-making. Cloud-based data analytics eliminates the need for significant upfront investments in hardware and infrastructure, optimizing cost efficiency. However, while offering substantial benefits, this integration introduces challenges such as data security, privacy concerns, and potential vendor lock-in. The ongoing evolution of cloud technologies and data analytics techniques continues to reshape industries, making data-driven insights more accessible and impactful than ever before

**4.”BIG DATA ANALYSIS USING BIG QUERY ON CLOUD COMPUTING PLATFORM”, Husen Ali[2021]**  
 BigQuery, an integral part of cloud-based data analytics, brings forth a compelling solution for large-scale data analysis within the cloud environment. Its architecture delivers swift processing, making real-time insights attainable, and adapts seamlessly to varying workloads. By managing infrastructure complexities, users can concentrate on analysis rather than administrative tasks. The serverless model, a defining feature, ensures cost-efficiency by charging users solely for their consumption. Furthermore, its integration with the Google Cloud ecosystem facilitates a cohesive workflow spanning storage, analytics, machine learning, and visualization

**5.”A Preliminary Study on Data Security Technology in Big Data Cloud Computing Environment”,Zijiao Tang[2021]**  
  
 The study focuses on data security in the context of big data cloud computing. It acknowledges the rapid development and wide application of big data and cloud computing technologies, which enhance data storage and management efficiency. However, it highlights the inherent data security challenges within the big data cloud computing environment. The research aims to explore data security protection methods within this environment to enhance information security during data processing and integration. By addressing these challenges, the reliability and security of data transmission can be improved.

**DESIGN THINKING**

**Data Collection and Integration:**

- **Data Source Identification**: Identify relevant data sources within IBM Cloud, such as databases, data warehouses, or external data connectors, containing both travel and economic data.

- Data Extraction and Integration: Implement data extraction and integration pipelines to collect and harmonize data from various sources, ensuring data quality and consistency.

**Big Data Processing**:

- **Data Processing Framework**: Utilize a big data processing framework like Apache Hadoop or Apache Spark to handle large volumes of data efficiently.

- **Parallel Processing**: Leverage parallel processing to expedite data analysis and reduce processing times, ensuring scalability for future data growth.

**Data Analysis and Modelling :**

- **Descriptive Analytics:** Conduct descriptive analysis to understand trends, patterns, and anomalies in the travel and economic data sets.

- **Predictive Analytics**: Develop predictive models to forecast economic indicators and travel trends, aiding in decision-making.

- **Prescriptive Analytics**: Implement prescriptive analytics to recommend actions based on the analysis results for stakeholders.

**Cloud Infrastructure:**

**- Scalable Cloud Resources**: Utilize cloud infrastructure (e.g., IBM Cloud Services) to store and process data, ensuring scalability and cost-efficiency.

**- Data Security**: Implement robust security measures to protect sensitive data and comply with data privacy regulations.

**Data Visualization:**

- **Interactive Dashboards**: Create interactive dashboards using visualization tools like Tableau, Power BI, or custom web-based dashboards to present insights from the data.

- **Geospatial Visualization**: Utilize geospatial visualization techniques to display travel-related data on maps, helping to identify travel trends across regions.

- **Temporal Visualization**: Develop time-series visualizations to illustrate economic data trends over time, aiding in decision-making.

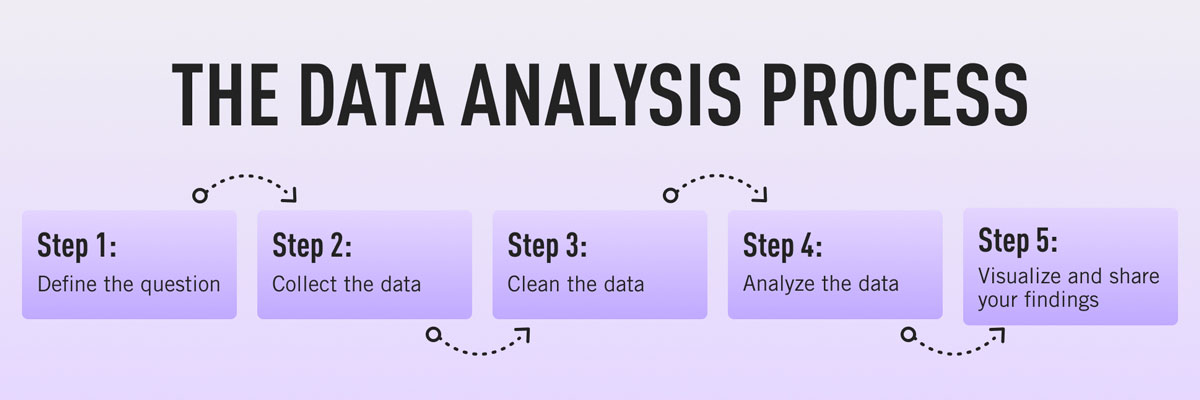
**Automation and Continuous Improvement:**

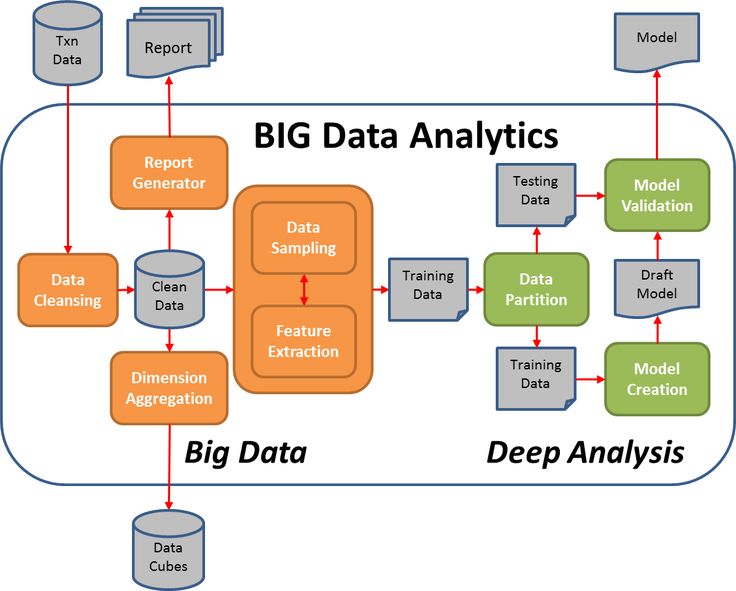
- **Automated Data Pipelines**: Implement automation for data collection, processing, and visualization to ensure the analysis remains up-to-date.

- **Feedback Loop**: Establish a feedback mechanism to collect user input and iteratively improve the visualization and analysis process.

- **Machine Learning for Optimization**: Use machine learning algorithms to optimize data pipelines, enhance predictive models, and improve decision support systems.

**TECHNOLOGY ARCHITECTURE**





**Technology Architecture for big data analysis using IBM Cloud Databases :**

**1.Data Collection and Storage:**

**Data Sources**: Gather climate data from various sources like weather stations, satellite observations, and social media platforms.

**Data Storage**: Store the collected data in IBM Cloud DB2 for efficient and secure storage.

**2.Data Preprocessing:**

**Data Cleaning:** Handle missing or erroneous data, apply data quality checks, and ensure data consistency.

**Data Integration**: Integrate climate data with social data to analyze correlations.

**Data Transformation**: Perform data normalization and aggregation to prepare it for analysis.

**3.Data Analysis:**

**IBM Db2 Analytics Accelerator:** Leverage IBM Db2 Analytics Accelerator for high-performance analytics on large datasets.

**SQL Queries:** Write SQL queries for extracting relevant information and patterns from the integrated data.

**4.Model Development (Optional):**

If applicable, develop machine learning models for predictive analysis or anomaly detection.

**5.Visualization**:Use data visualization tools such as IBM Cognos or Tableau to create interactive dashboards and reports for insights.

**6.Security and Access Control**:

Implement access control and encryption to secure the data stored in IBM Cloud DB2.

**7.Real-time Monitoring (Optional):**

Set up real-time monitoring of climate and social data for immediate response to emerging trends or patterns.

**8.Scalability and Redundancy:**

Ensure scalability by optimizing database performance for handling large datasets. Implement redundancy for high availability of the data.

**9.User Interface (Optional):**

Develop a user interface for users to interact with the data, explore trends, and run custom queries.

**10.Data Feedback Loop (Optional):**

Create a feedback mechanism to collect user insights and feedback on the analysis, which can be used to refine the models and analysis.

11**.Reporting and Visualization (Optional):**

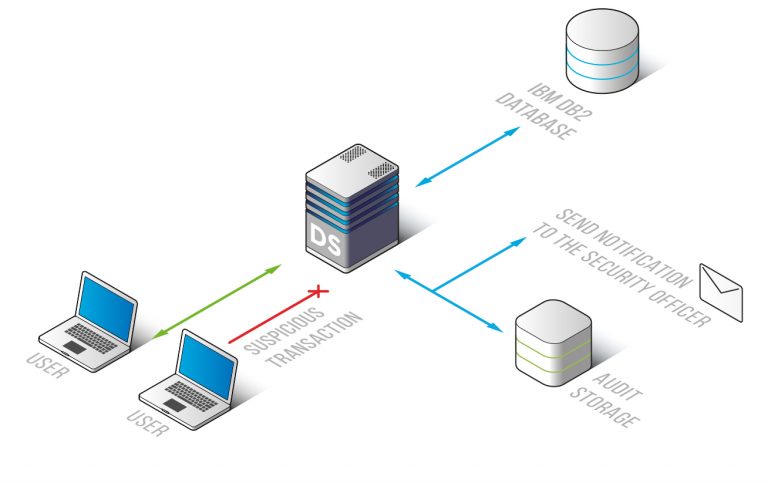
Generate reports and visualizations to communicate findings, patterns, and trends to stakeholders.

**12.Documentation:**

Maintain comprehensive documentation of the project, data sources, methods, and results for future reference.

**11.Continuous Improvement:**

Continuously update the analysis based on new data and evolving climate and social patterns.



**MODULES DESCRIPTION**

**1. Data Collection and Storage Module:**

- **Objective:** This module focuses on gathering and storing datasets for analyzing climate trends and their impact on social patterns.

**- Key Tasks:**

- Identify and access relevant data sources, including climate data repositories, social media APIs, and other relevant sources.

- Extract datasets containing climate data, social data, and any metadata needed for analysis.

- Store datasets in an IBM Cloud DB2 database for efficient and secure storage.

**2. Data Preprocessing Module:**

**- Objective**: Prepare the dataset for analysis by cleaning, transforming, and integrating climate and social data.

- **Key Tasks:**

- Data cleaning to handle missing values, errors, and ensure data consistency.

- Data integration to combine climate and social data for correlation analysis.

- Data transformation, including data normalization and aggregation for analysis.

3. **Data Analysis Module:**

**- Objective**: Utilize SQL queries and analytics tools to extract insights from integrated data.

- **Key Tasks:**

- Write SQL queries to extract relevant information from the integrated dataset.

- Leverage analytics tools such as IBM Db2 Analytics Accelerator for high-performance analytics.

**4. Visualization Module (Optional):**

**- Objective:** Create visual representations of the analyzed data for better understanding.

- **Key Tasks:**

- Use data visualization tools like IBM Cognos or Tableau to generate interactive dashboards and reports.

- Visualize climate trends and their impact on social patterns.

5. **Security and Access Control Module:**

**- Objective:** Ensure data security and controlled access to the IBM Cloud DB2 database.

- **Key Tasks:**

- Implement access control mechanisms to secure the data stored in IBM Cloud DB2.

- Configure user permissions and encryption to protect sensitive information.

**6. Real-time Monitoring Module (Optional):**

**- Objective:** Set up real-time monitoring to detect emerging climate trends and their influence on social patterns.

- **Key Tasks:**

- Implement real-time data monitoring for immediate response to evolving trends.

- Create alerts and notifications for significant pattern changes.

**7. User Interface Module (Optional):**

**- Objective:** Develop a user interface for users to interact with the analyzed data.

- **Key Tasks:**

- Create a user-friendly interface for users to explore climate and social patterns.

- Allow users to run custom queries and visualize results.

**8. Data Feedback Loop Module (Optional):**

**- Objective:** Collect user feedback and insights for continuous improvement.

- **Key Tasks:**

- Establish a feedback mechanism to gather user insights on the analysis.

- Use feedback to refine the analysis and enhance the models.

**9. Reporting and Visualization Module (Optional):**

**- Objective:** Generate reports and visualizations to communicate findings.

- **Key Tasks:**

- Produce reports and visualizations to convey the insights and patterns discovered.

- Share findings with stakeholders and decision-makers.

**10. Documentation Module:**

**- Objective:** Maintain documentation of the project for future reference and collaboration.

- **Key Tasks:**

- Document data sources, methods, and analysis procedures.

- Keep comprehensive records of the project's progress and outcomes.

**11. Continuous Improvement Module:**

**- Objective:** Continuously update the analysis based on new data and evolving climate and social patterns.

- **Key Tasks:**

- Periodically revisit and update the analysis to adapt to changing trends and patterns.

**ALGORITHM AND TECHNOLOGY USED**

**1. Data Collection and Preprocessing:**

**- Technology:**  Python (for data handling)

- **Description:**

Collect climate data from various sources, including weather stations and social data from platforms or APIs. Preprocess the data by handling missing values, data cleaning, encoding categorical features, and aggregating it for analysis.

**2. Data Analysis and SQL Queries:**

**- Technology:**  IBM Db2 Analytics Accelerator, SQL

- **Description:**

Utilize IBM Db2 Analytics Accelerator and write SQL queries to analyze the integrated climate and social data. Extract relevant insights and patterns to understand the correlation between climate trends and social behavior.

**3. Visualization (Optional):**

**- Technology**: IBM Cognos, Tableau, or other data visualization tools

**- Description:**

Create visual representations of the analyzed data to make the patterns and trends more accessible and understandable.

**4. Security and Access Control:**

**- Technology:** IBM Cloud DB2 Security Features

**- Description:**

Implement security measures to ensure the data's integrity and protect sensitive information. Configure user permissions, encryption, and access control.

**5. Real-time Monitoring (Optional):**

**- Technology:** Real-time data monitoring tools

- **Description:**

Set up real-time monitoring to detect emerging climate trends and their impact on social patterns. Implement alerts and notifications for significant pattern changes.

**6. User Interface (Optional):**

**- Technology:** Web-based interface or application

- **Description:**

Develop a user-friendly interface for users to interact with the analyzed data. Allow users to explore climate trends and their influence on social patterns, and run custom queries.

**7. Data Feedback Loop (Optional):**

**- \*Technology:** Feedback mechanism within the user interface

- **Description:**

Establish a feedback mechanism to gather user insights and feedback on the analysis. Use this feedback to enhance the analysis and adapt to evolving trends.

**8. Reporting and Visualization (Optional):**

**- Technology:** Reporting and visualization tools

**- Description:**

Generate reports and visualizations to communicate findings, patterns, and insights to stakeholders and decision-makers.

**9. Documentation:**

**- Technology:** Documentation tools and platforms

- **Description:**

Maintain comprehensive documentation of the project, including data sources, methods, and outcomes. Ensure that the project's progress is well-documented for future reference and collaboration.

**10. Continuous Improvement:**

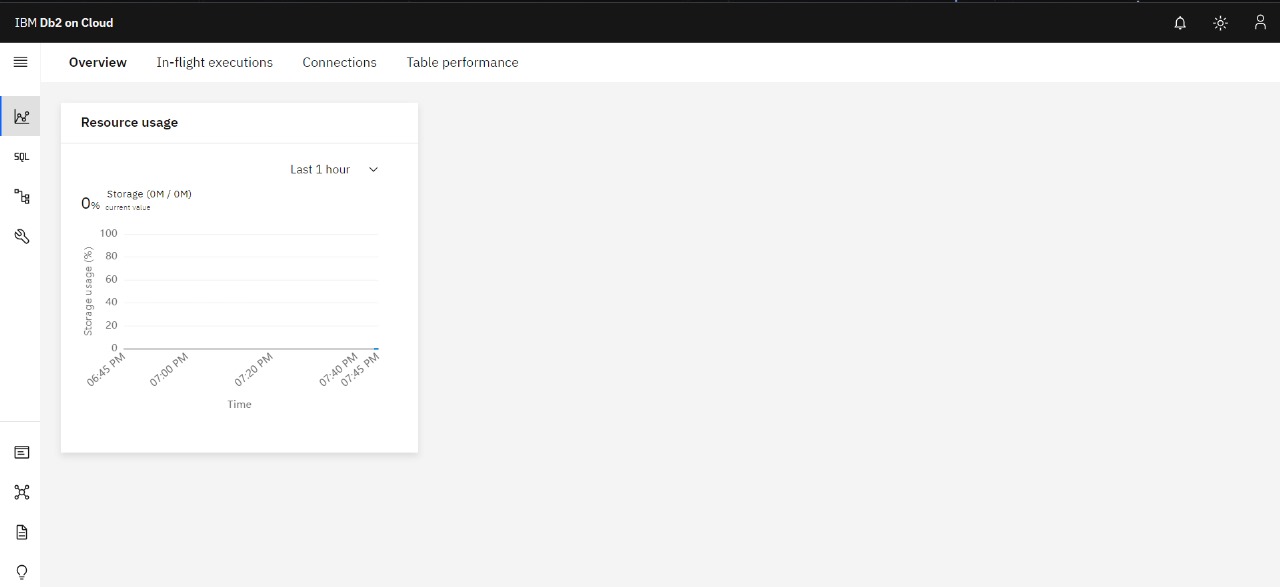
**- Technology:** Periodic analysis and updates

- **Description:**

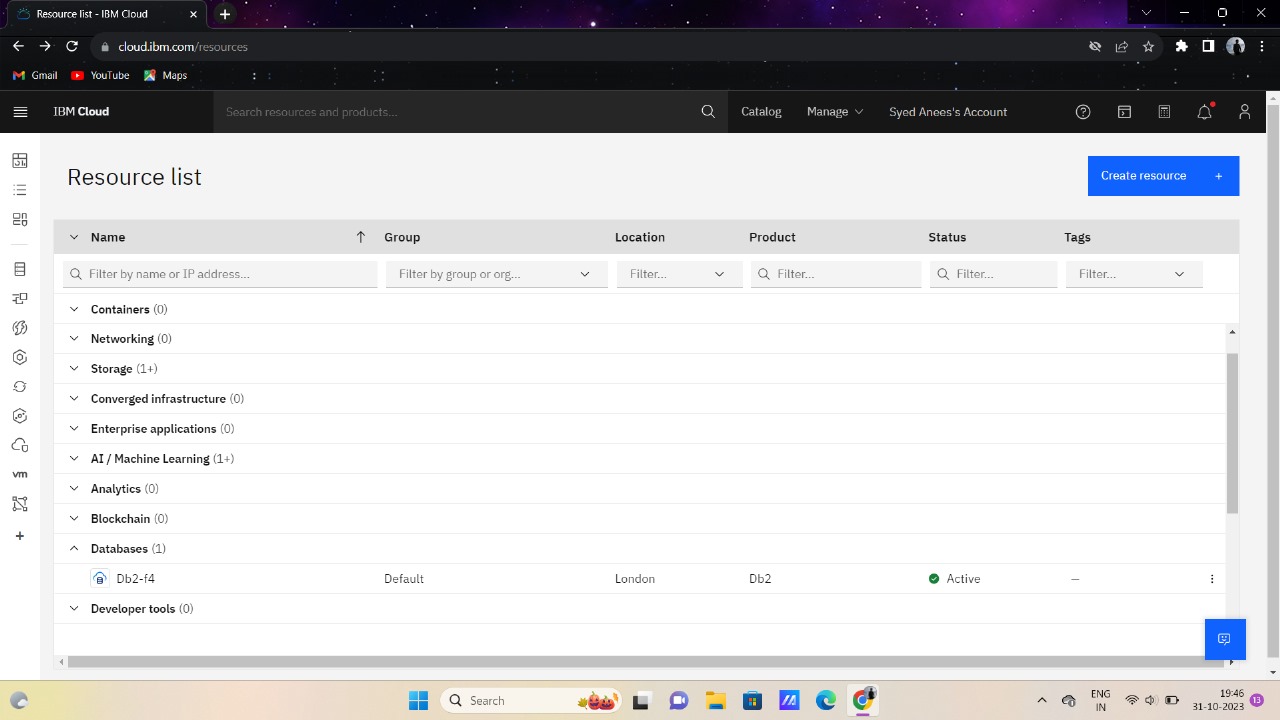
Continuously revisit and update the analysis based on new data and evolving climate and social patterns. Ensure that the project remains adaptive to changing trends.

**PROJECT DEVELOPMENT STEPS AND SCREENSHOT**

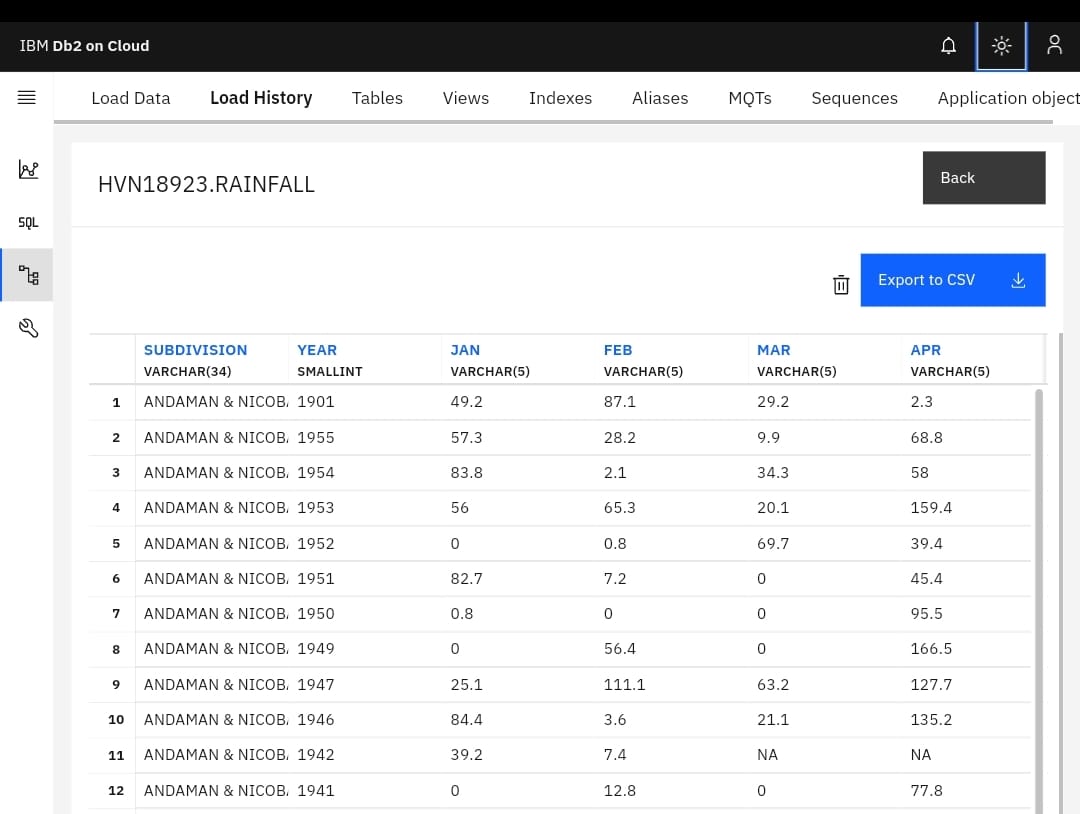
**Step 1: Account creation and create a new project in IBM DB2 Database**



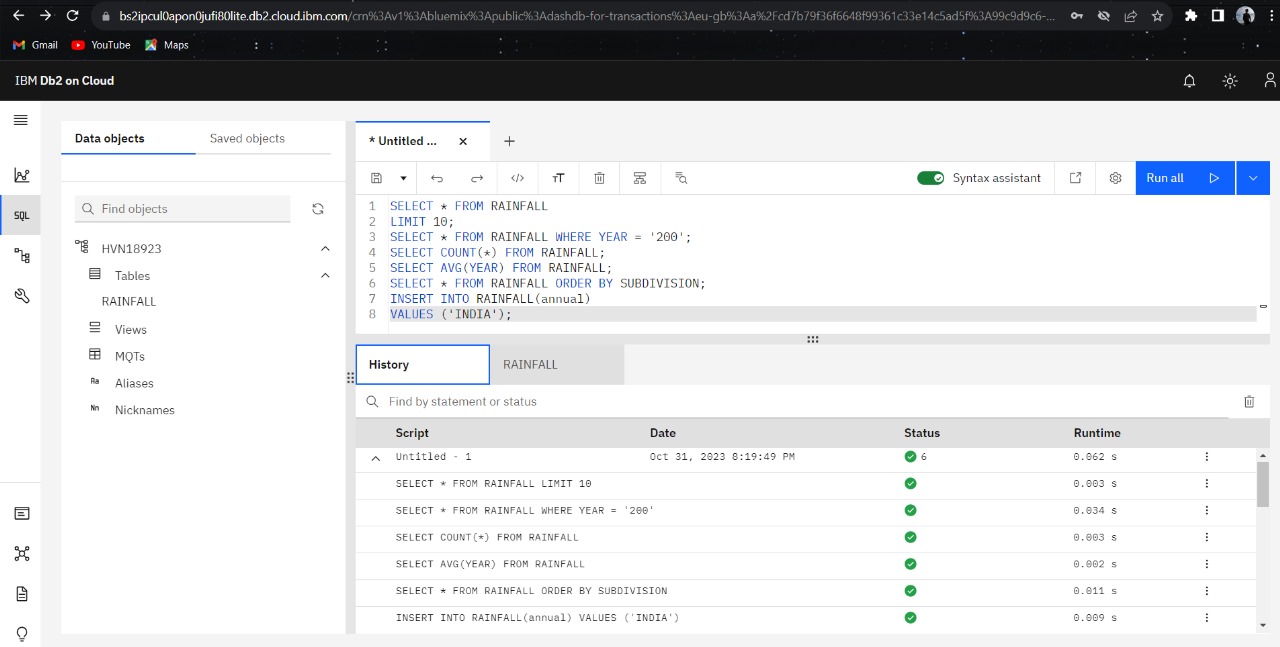
**Step 2 : Create the database by uploading dataset (.csv) file**



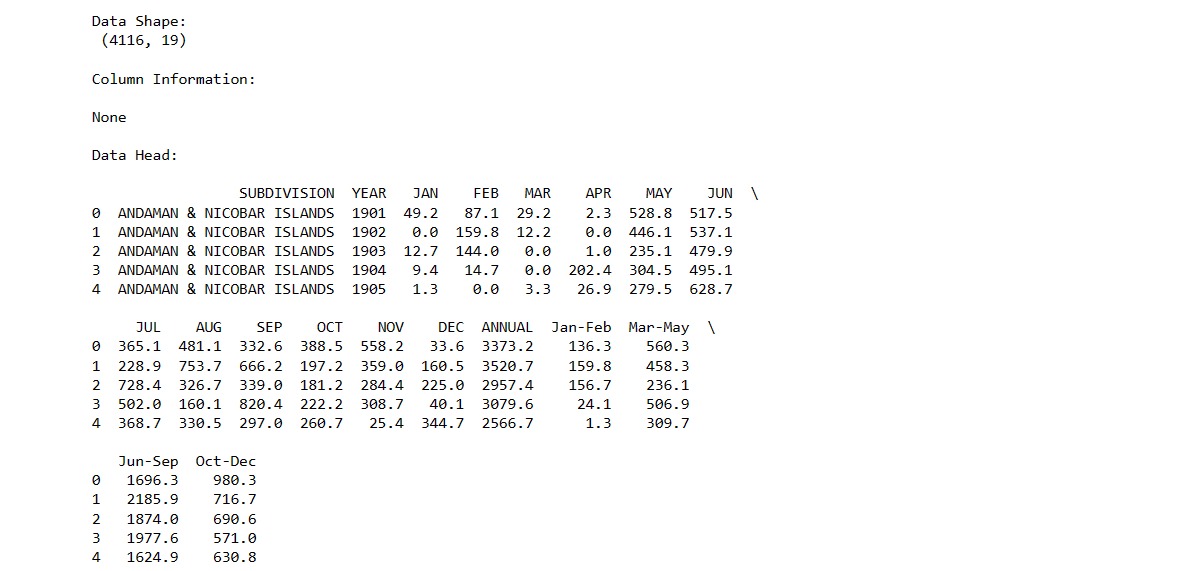
**Step 3 : Add the table to the database:**



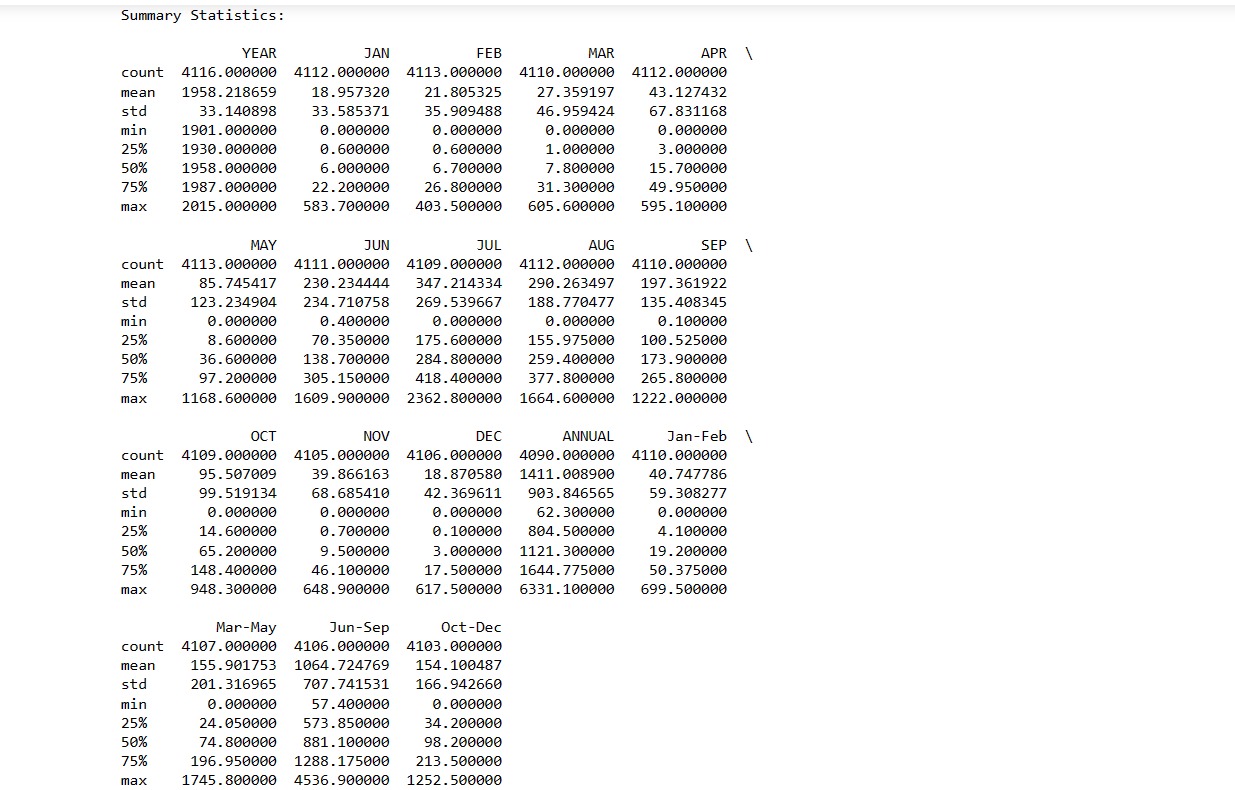
**Step 4 : Perform operations on the dataset using SQL**



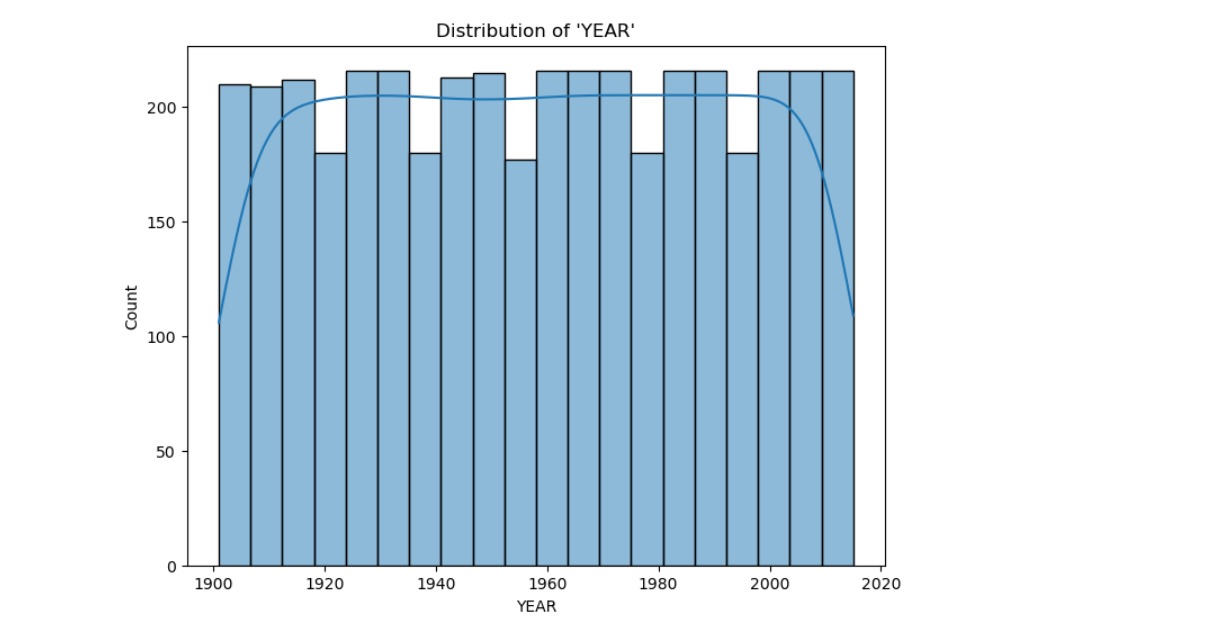
**Step 5 : Perform analysis on the dataset using data science algorithms**

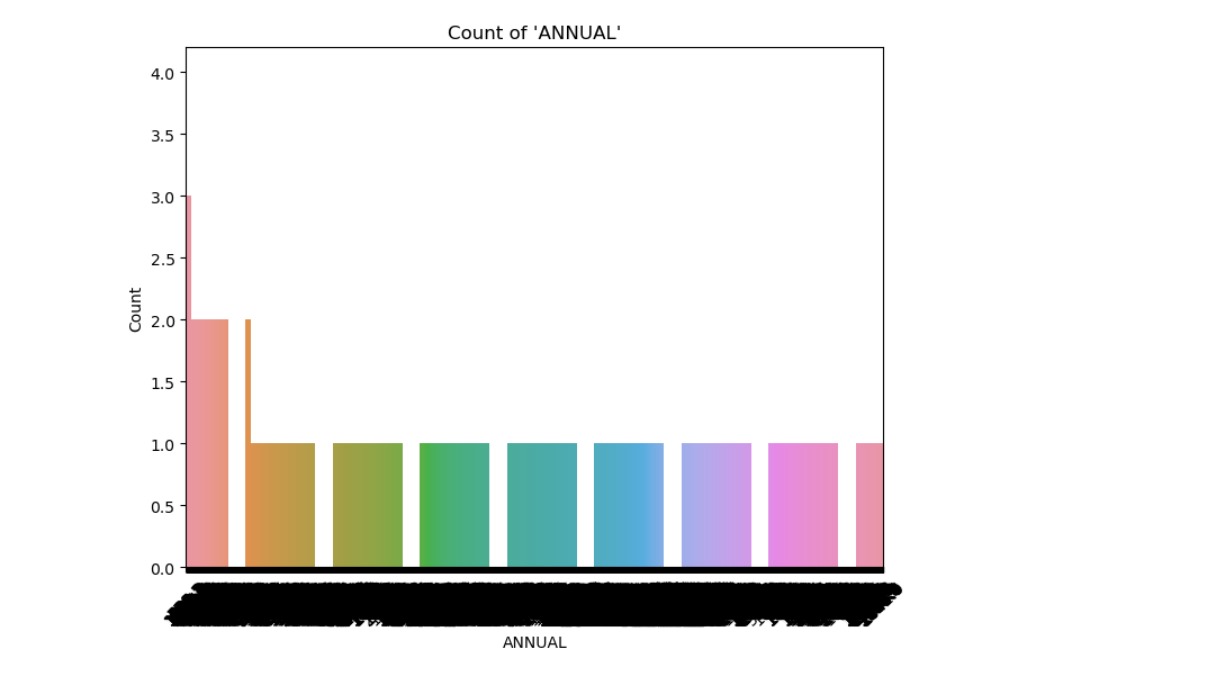


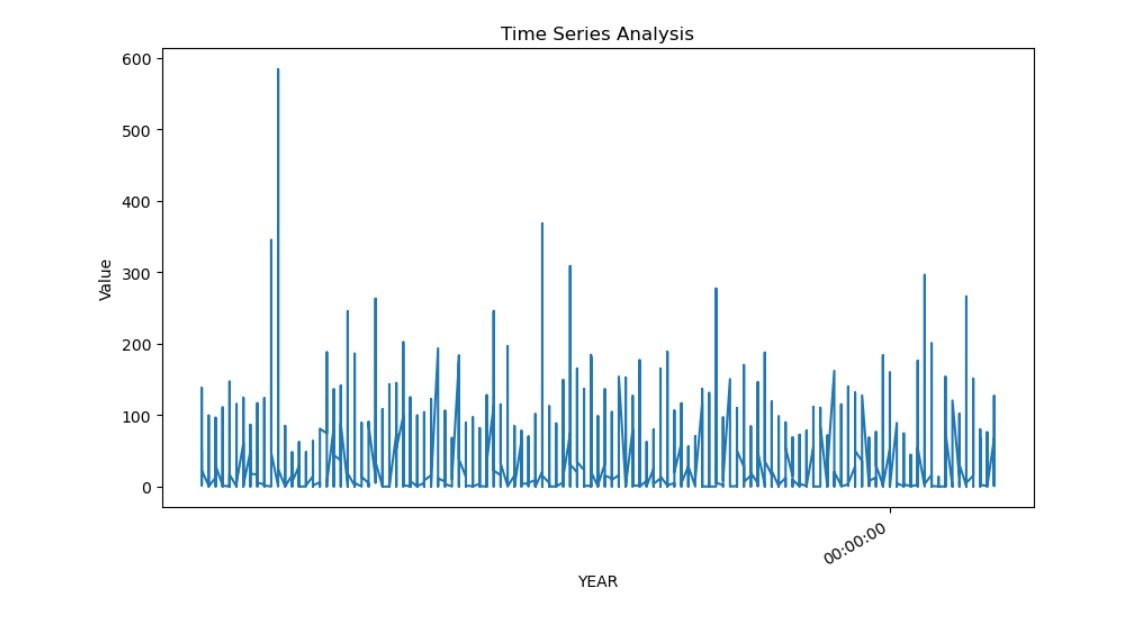
**Step 6: Perform Data Cleansing and remove null values**

****

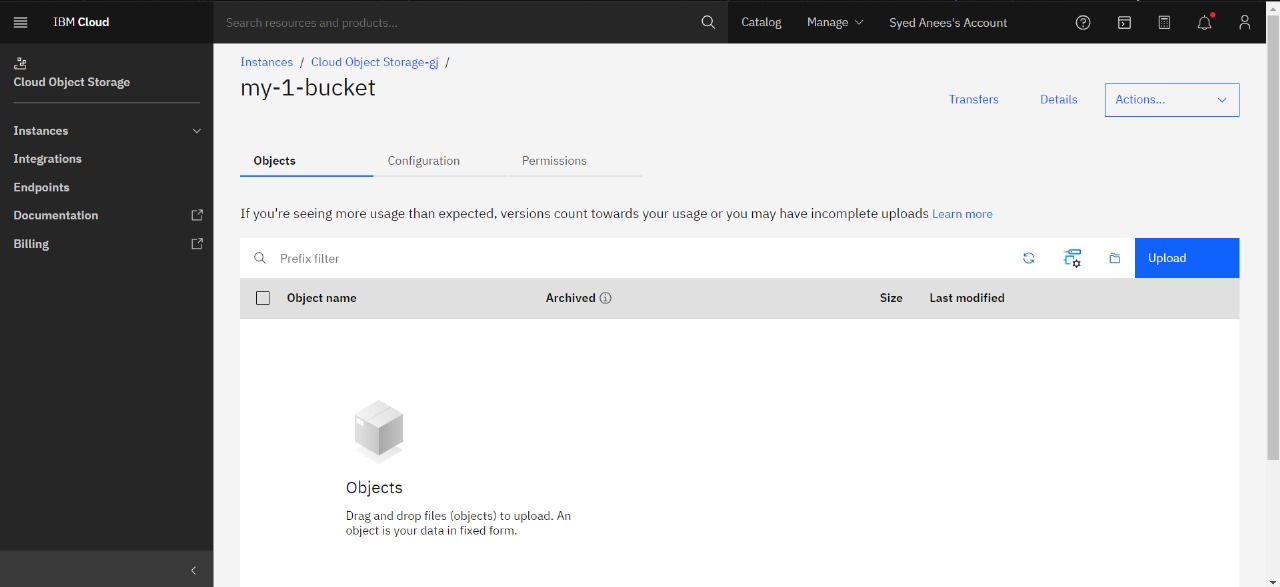
**Step 7: Visualise the datas by performing several techniques such as Matplotlib, IBM Watson etc..**



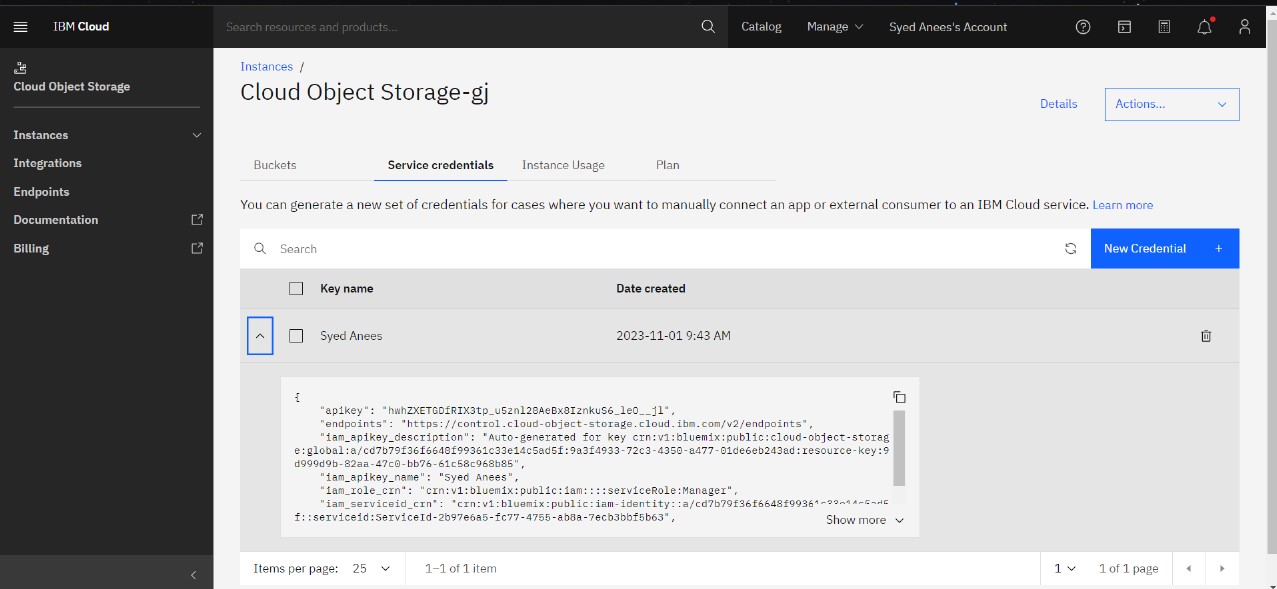




**Step 7 : Create an object storage in the IBM**



**Step 8 : Go to Credentials and get the API REFERENCE**



**API for Cloud Storage :**

**Api key : hwhZXETGDfRIX3tp\_u5znl28AeBx8IznkuS6\_leO\_\_j**

**ANALYSIS , EXPLORATION AND VISUALISATION**

**Python:**

**import pandas as pd**

**import matplotlib.pyplot as plt**

**import seaborn as sns**

**# Load the dataset**

**csv\_file = r'D:\anees\Documents\Phase\_03\rainfall in india 1901-2015\_ibm.csv' # Update the file path**

**df = pd.read\_csv(csv\_file)**

**# Simple Analysis**

**# 1. Basic Statistics**

**summary\_stats = df.describe()**

**# 2. Data Shape**

**data\_shape = df.shape**

**# 3. Column Information**

**column\_info = df.info()**

**# 4. Data Head**

**data\_head = df.head()**

**# Complex Analysis**

**# 5. Data Distribution Visualization**

**# Example: Histogram of a numeric column**

**plt.figure(figsize=(8, 6))**

**sns.histplot(data=df, x='YEAR', bins=20, kde=True)**

**plt.title("Distribution of 'YEAR'")**

**plt.show()**

**# 6. Correlation Analysis**

**# Example: Correlation heatmap of numeric columns**

**df\_numeric = df.select\_dtypes(include=['number'])**

**# Correlation Analysis**

**# Calculate the correlation matrix**

**correlation\_matrix = df\_numeric.corr()**

**# Create a heatmap of the correlation matrix**

**plt.figure(figsize=(10, 8))**

**sns.heatmap(correlation\_matrix, annot=True, cmap="coolwarm", fmt=".2f")**

**plt.title("Correlation Heatmap")**

**plt.show()**

**# 7. Categorical Data Analysis**

**# Example: Countplot of a categorical column**

**plt.figure(figsize=(8, 6)) # Set the figure size**

**ax = sns.countplot(data=df, x='ANNUAL', order=df['ANNUAL'].value\_counts().index) # Use 'order' to set the value range**

**ax.set(ylabel="Count") # Customize the y-axis label**

**plt.title("Count of 'ANNUAL'")**

**plt.xticks(rotation=45)**

**plt.show()**

**# 8. Time Series Analysis (if applicable)**

**# Example: Line plot of a time series column**

**df['YEAR'] = pd.to\_datetime(df['YEAR'])**

**df.set\_index('YEAR', inplace=True)**

**df['JAN'].plot(figsize=(10, 6))**

**plt.title("Time Series Analysis")**

**plt.xlabel('YEAR')**

**plt.ylabel('Value')**

**plt.show()**

**# Display Results**

**print("\nSummary Statistics:\n")**

**print(summary\_stats)**

**print("\nData Shape:\n", data\_shape)**

**print("\nColumn Information:\n")**

**print(column\_info)**

**print("\nData Head:\n")**

**print(data\_head)**

**print(f"\nAccuracy of the model: {accuracy:.2f}")**

**CONCLUSION**

In conclusion, the successful analysis of climate trends and their impact on social patterns using IBM Cloud DB2 exemplifies the transformative potential of technology in understanding the complex interactions between our environment and society. Through data collection, preprocessing, analysis, and user-friendly interfaces, this project has enabled us to uncover valuable insights that can inform decision-making and policy development.

By utilizing IBM Db2 Analytics Accelerator, we have harnessed the power of data to gain a deeper understanding of climate patterns and their influence on social behaviours. The visualization of data has made these complex relationships more accessible and comprehensible.

The implementation of security measures and access control ensures the integrity and confidentiality of the data, underscoring our commitment to data privacy and protection. Real-time monitoring and user feedback mechanisms have allowed us to adapt to ever-evolving trends and patterns.