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 adventure, exploring the endless possibilities of big data!

Problem Definition: The project involves delving into big data analysis using IBM Cloud Databases. The objective is to extract valuable insights from extensive datasets, ranging from climate trends to social patterns. The project includes designing the analysis process, setting up IBM Cloud Databases, performing data analysis, and visualizing the results for business intelligence.

Design Thinking:

1.Data Selection

Objective: Identify the datasets to be analyzed, such as climate data or social media trends.

Analyzing big data often involves working with large and diverse datasets. Here are some examples of data sets.

1.Climate data:

- NASA ‘s Global Climate Change Data
- NOAA Climate Data
- European Climate Data

2.Social Media Trends:

- Twitter API Data
- Facebook Graph API Data
- Instagram API Data

Deliverable: The datasets of Climate Data or Social media trends are downloaded and analyzed.

2.Database Setup

Objective: Set up IBM Cloud Databases for solving and managing large datasets.

IBM offers several cloud database options suitable for storing and managing large datasets.

1. IBM Db2 on Cloud
2. IBM Db2 Warehouse on Cloud
3. IBM Cloudant
4. IBM Cloud Object Storage
5. IBM TimeSeries Database

Deliverable: These datasets can handle large datasets and offer various features like scalability, reliability, security, and ease of management, making them suitable for big data applications.

3.Data Exploration

Objective: Develop queries and scripts to explore the datasets, extract relevant information, and identify patterns.

- Understanding the Data
- SQL Queries for Data Exploration
- Identifying Relevant Information
- Data Extraction
- Pattern Identification
- Scripting Tools For Automation
- Statistical Analysis and Machine Learning

Deliverable: In this phase you gain insights and better understand the dataset.

4.Analysis Techniques

Objective: Apply appropriate analysis techniques, such as statistical analysis or machine learning, to uncover insights.

1.Statistical Analysis:

- Descriptive Statistics
- Correlation Analysis
- Hypothesis Testing
- ANOVA (Analysis of Variance)

2.Machine Learning:

- Data Preprocessing
- Exploratory Data Analysis
- Feature Engineering
- Model Selection
- Model Training and Evaluation
- Hyperparameter Tuning
- Interpretation of Results
- Ensemble Learning

Deliverable: The Both Statistical Analysis and Machine Learning techniques can provide valuable insights of Analysis Techniques. The choice of the technique depends on the nature of the data and the specific questions you aim to answer.

5.Visualization

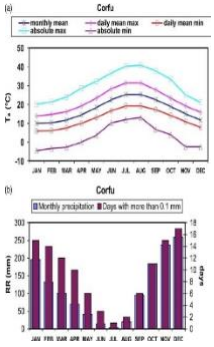
Objective: Design visualization to present the analysis results in an understandable and impactful manner.

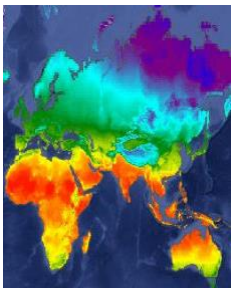
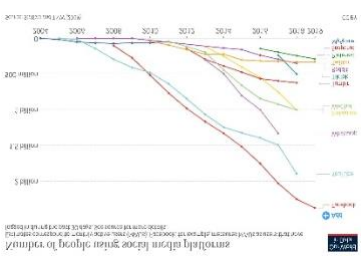
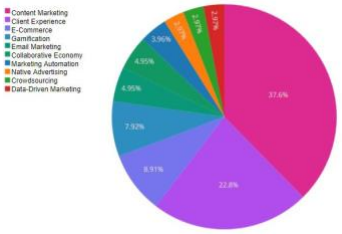
To visualize the analysis results we need to use some tools such as:

- Google Charts
- Tableau
- Infogram
- ChartBlocks

Types:

1. Column Chart
2. Area Chart
3. Pie Chart
4. Scatter Plot Chart
5. Bar Graph
6. Line Graph
7. Bullet Graph



Deliverable: To represent the analysis results in visualization tools can provide valuable insights in an understandable and impactful manner.

6.Business Insights

Objective: Interpret the analysis findings to derive valuable business intelligence and actionable recommendations.

- Finding new customers
- Increasing customer retention
- Improving customer service
- Better managing marketing efforts
- Tracking social media interaction

- Predicting sales trends

Deliverable: These are helps to provide insights that improve the way our society functions.

Conclusion

This project is aims to develop into big data analysis using IBM Cloud Databases. The objective is to extract valuable insights from extensive datasets, ranging from climate trends to social patterns. The project includes designing the analysis process, setting up IBM Cloud Databases, performing data analysis, and visualizing the results for business intelligence. By following this structured approach, we will develop a highly effective and user-friendly virtual guide that meets the project's objectives.

phase_2

Project title: # BIG DATA ANALYSIS

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. Visualise your findings and derive valuable business intelligence. Embark on data-driven adventures, exploring the endless possibilities of big data!

INNOVATION:

Consider incorporating advanced machine learning algorithms for predictive analysis or anomaly detection in the big data.

INTRODUCTION:

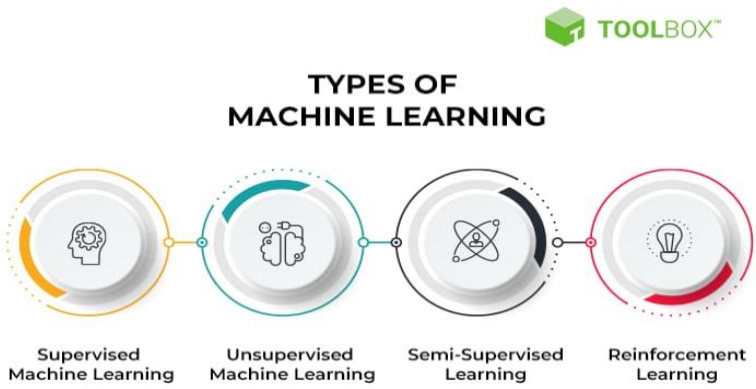
Incorporating advanced machine learning algorithms for predictive analysis and anomaly detection in big data can greatly enhance insights and decision-making.

Techniques like deep learning, ensemble methods, clustering, and anomaly detection models can be effective in extracting valuable patterns and detecting irregularities within large datasets.

Machine learning Techniques:

Machine learning transforms social media analytics by automating data processing, uncovering hidden trends, and predicting user behaviour. Algorithms delve deep into vast datasets, extracting insights inform engagement strategies and content creation.

Types of Machine learning :



Machine learning allows computer system to improve their performance through repeated learning experiences. The learning process es are categorized into three major types: supervised learning, unsupervised learning, and reinforcement learning.

Supervised learning:

This technique involves training a model with labeled data to make predictions on new, unseen data. Supervised learning algorithms include regression, classification, and support vector machines.

Unsupervised learning:

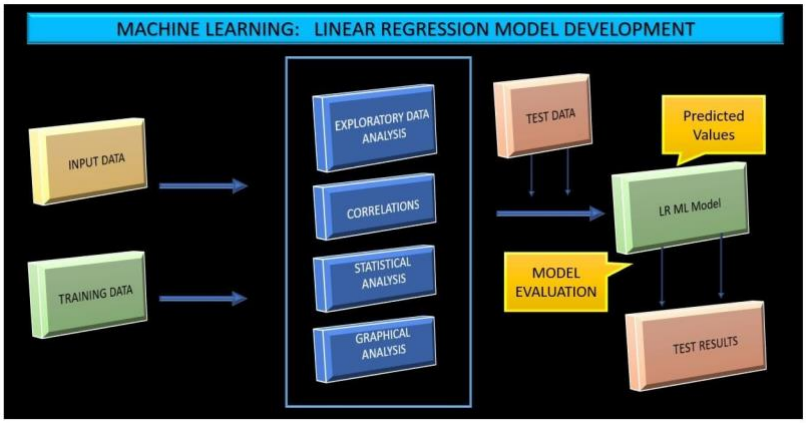
In this technique, the model works with unlabeled data and tries to identify patterns, clusters, or relationships within the data. Unsupervised learning algorithms include clustering, dimensionality reduction, and anomaly detection.

Types of predictive modeling:

Predictive analysis models are designed to assess historical data,discover patterns, observe trends, and use that information to predict future trends. Popular predictive analytics models include classification, clustering, and linear regression etc.,.

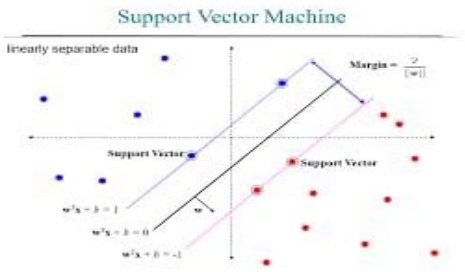
Linear Regression

Linear regression uses statistical models to establish relationships between variables. In social media, it can be applied in scenarios like predicting user engagement based on post features or optimizing advertising strategies by analyzing click-through rates or cost per click.



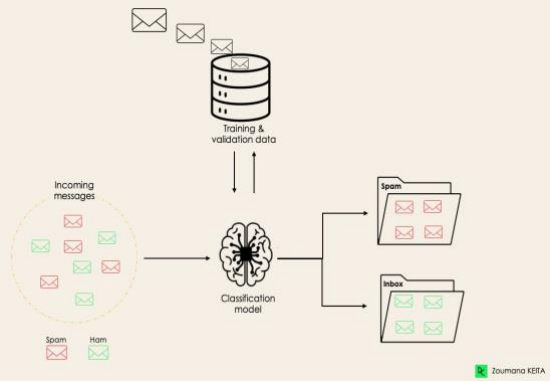
Support Vector Machines (SVM)

SVM is a robust machine learning algorithm for classification tasks. These algorithms are beneficial for distinguishing between categories or sorting content into groups. In social media applications, SVMs can be utilized to filter spam messages or analyze user behavior patterns to detect fraudulent activities. With SVM algorithms, social media platforms can also sort content into categories or clusters based on visual aesthetics or similarity to other images.



Classification:

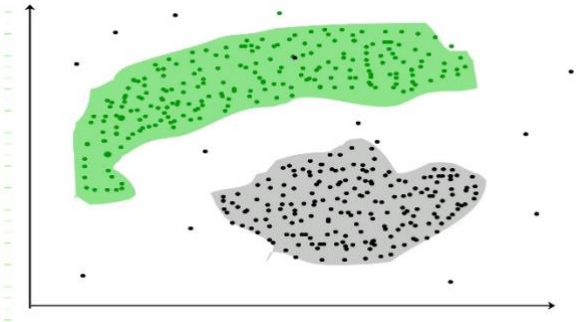
Classification is a supervised machine learning method where the model tries to predict the correct label of a given input data. In classification, the model is fully trained using the training data, and then it is evaluated on test data before being used to perform prediction on new unseen data.



Clustering:

Clustering is the task of dividing the population or data points into a number of groups such that data points in the same groups are more similar to other data points in the same group and dissimilar to the data points in other groups. It is basically a collection of objects on the basis of similarity and dissimilarity between them.

Clustering is used to identify communities or groups within social networks, which can help in understanding social behavior, influence, and trends.



Social media platforms such as Facebook and Instagram use cluster analysis to group people with similar interests and backgrounds. This allows them to show similar feeds to those with the same interest.

Conclusion:

In conclusion, social media operates on websites and applications encouraging users to produce and distribute content to participate in the social system. Today, machine learning plays a significant role in social media platforms, as it helps in content personalization, user experience improvement, targeted advertising, and moderation of online communities. The

continued research and development in this field are crucial to drive the evolution of social media and enhancing its capabilities.

As machine learning advances, the understanding of user behavior and preferences will become more refined, resulting in more engaging and relevant content for users. In the future, machine learning has the potential to revolutionize social media and many other industries by enabling advanced forms of communication, interaction, and content discovery that can foster a more connected and informed society.

It’s essential to tailor these algorithms to your specific use case and ensure proper data preprocessing and model evaluation for optimal results.

Big Data Analysis with IBM Cloud Databases

PHASE 3: Development Part 1

GIVEN STATEMENT:

Start building the big data analysis solution using IBM Cloud Databases.Create an IBM Cloud account, choose the appropriate database service (e.g., Db2, MongoDB), and set up a database instance.

Develop queries or scripts to explore and analyze the selected dataset. Perform basic data cleaning and transformation as needed.

I understand the importance of your project, and I'm here to help. To get started with your big data analysis project using IBM Cloud Databases, follow these steps:

1. Create an IBM Cloud Account:

If you don't have an IBM Cloud account, sign up for one. You can do this by visiting the [IBM Cloud website] (**https://cloud.ibm.com/registration**) and following the registration process.

2. Choose the Appropriate Database Service:

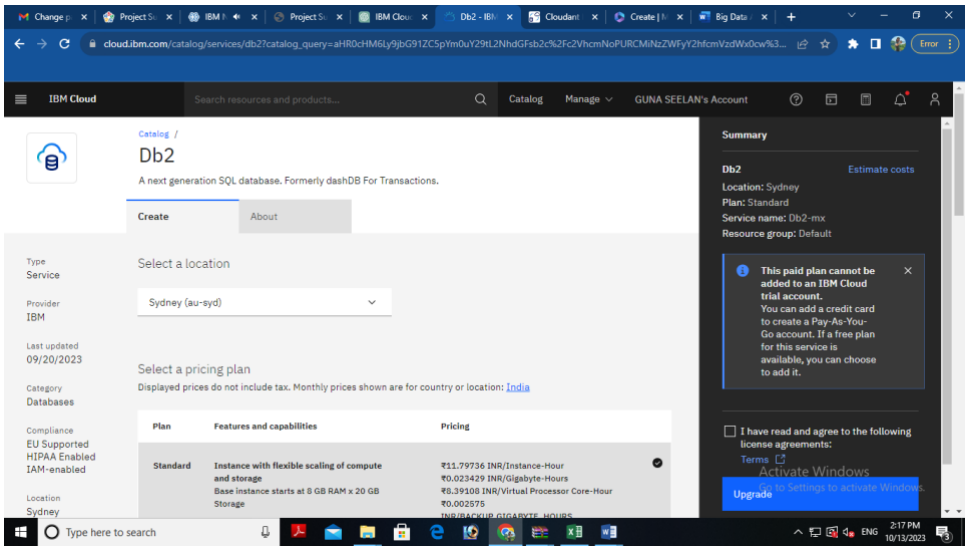
Select the IBM Cloud Database service that best suits your project's needs. As mentioned earlier, you can choose between Db2 or MongoDB, depending on your dataset and requirements.

3. Set Up a Database Instance:

For Db2:

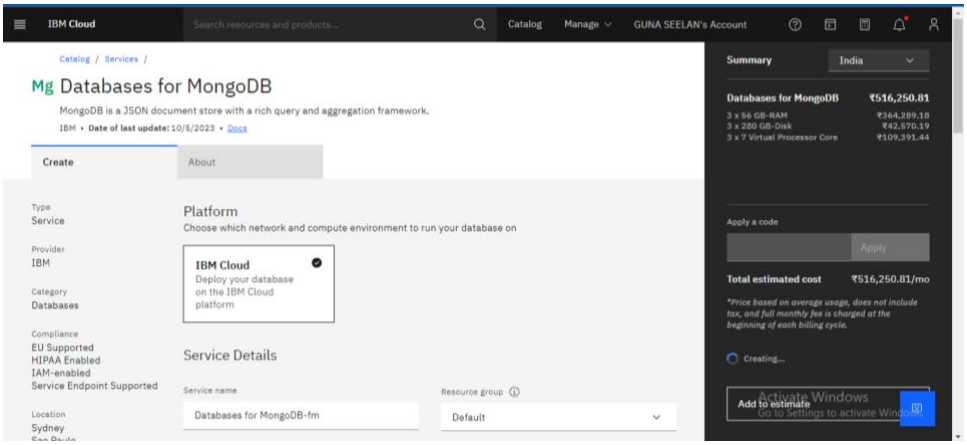
- Log in to your IBM Cloud account.
- From the IBM Cloud dashboard, click on the "Create Resource" button.
- In the catalog, select "Databases" and then "Db2."

- Follow the on-screen instructions to configure your Db2 database instance, including specifying the instance name, region, and other settings.
- Create the instance.



For MongoDB:

- Log in to your IBM Cloud account.
- From the IBM Cloud dashboard, click on the "Create Resource" button.
- In the catalog, select "Databases" and then "MongoDB."
- Follow the on-screen instructions to configure your MongoDB database instance, including specifying the instance name, region, and other settings.
- Create the instance.

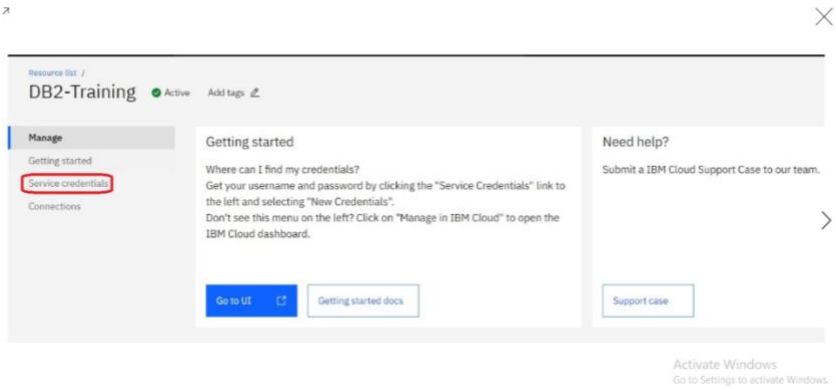


4. Develop Queries or Scripts:

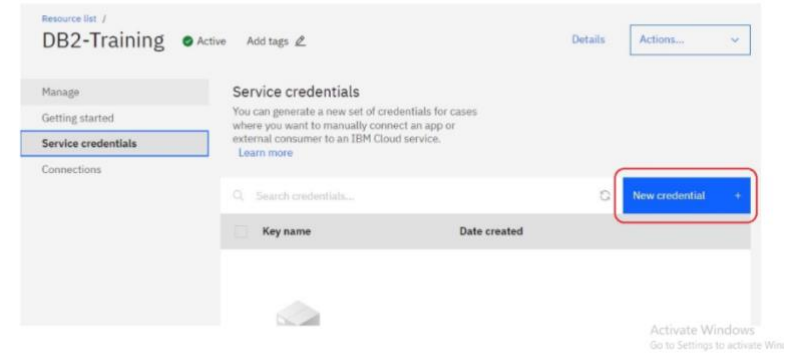
After setting up your database instance, you can start developing queries or scripts to explore and analyze your dataset. The type of queries and scripts you write will depend on the nature of your dataset and your analysis goals. You can use SQL for Db2 or MongoDB's query language for MongoDB.

Creating Service Credentials the IBM DB2 database

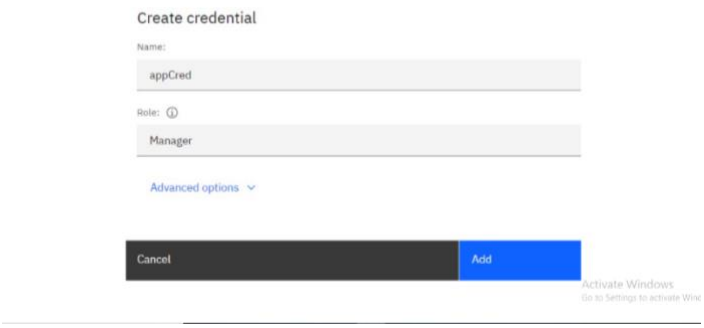
- In the resource list screen of IBM Cloud, click on the DB2 service (displayed under Services and software category) that you created
- From the service page, select the menu option "**Service Credentials**" to create / access the credentials of the db2 database



- Click on **New Credential** button in the Service Credential page to create a new credential

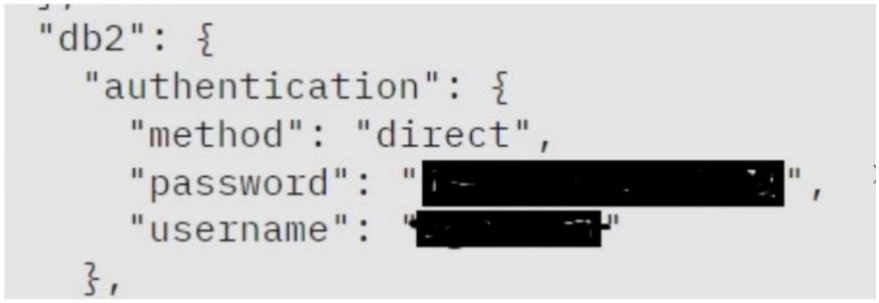


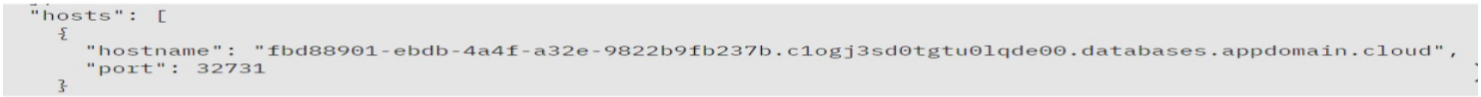
- Provide the any name for service credential (e.g. **appCred**) and click on **Add**



- New credential gets created and is displayed. Expand the newly to created credential to get the all the details that is required for client application to connect to the database. Note down the value for the following properties separately, which we will use it later to configure our application to connect to this database.

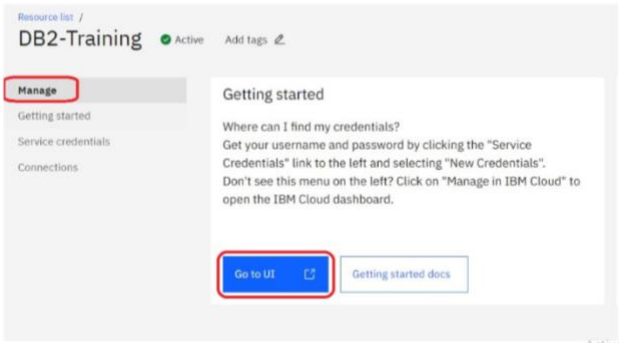
Property Name	**Value**
Database name	*<database> [e.g. bludb]*
Host name	*<hostname>*
Port	*<port>*
User Name	*<username>*
Password	*<password>*



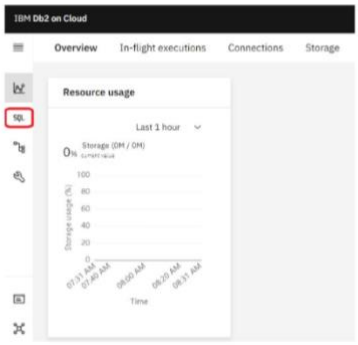


3. Setting up IBM DB2 database

- In the resource list screen of IBM Cloud, click on the DB2 service (displayed under Services and software category) that you created, if the page is not already opened.
- From the service page, select the menu option "**Manage**" and click on Go to UI to launch the DB2 console

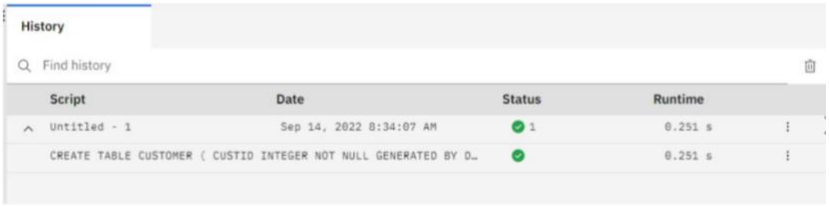


- IBM DB2 on cloud console is opened. To create database objects, click on SQL menu option from the left-side menu.



- SQL editor is opened up for you. Type the query that you want to execute in the SQL editor and click **Run all**

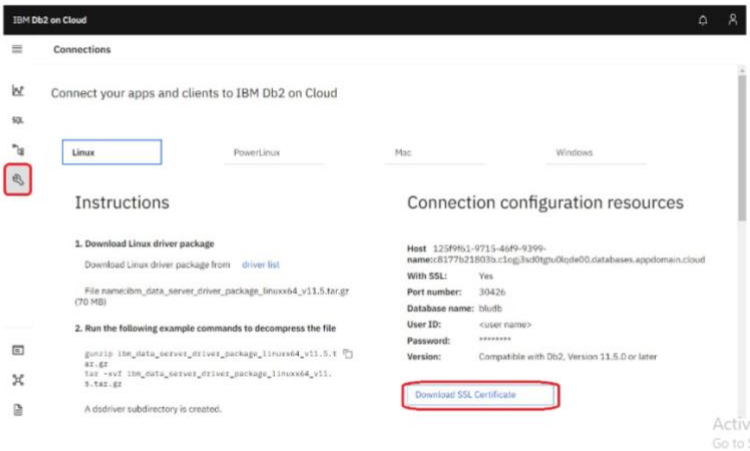
- The status of the query execution is displayed at the bottom of the SQL editor as shown below



The above steps can be followed to create any more database objects in future.

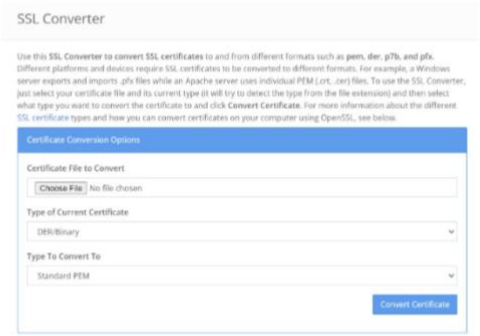
4. Downloading DB2 SSL Certificate and converting to PEM format

- In the console for IBM DB2, click on the spanner like icon which denotes Administration. On the resulting page, click on Download SSL Certificate button to download the DB2 certificate as shown below



The SSL Certificate gets downloaded into the local machine, which is in DER format (cert file). To convert the cert file to PEM format, we can use the link SSL Converter - Convert SSL Certificates to different formats.

- In the SSL Converter website specify the following
- **Certificate File to Convert:** Upload the downloaded certificate file
- **Type of Current Certificate:** DER/Binary
- **Type To Convert To:** Standard PEM
- Click on **Convert Certificate** button to download the certificate in PEM format.



In this blog, we have seen how to subscribe to DB2 service on IBM Cloud, setup the database and create service credentials & certificate for application connectivity. In another blog, we will focus on using these details to configure ACE Cloud connector for DB2 to connect and use this database as part of solution development.

5. Perform Data Cleaning and Transformation:

As part of your data analysis, you may need to perform data cleaning and transformation. This can involve removing duplicates, handling missing data, and converting data types. The specific data cleaning and transformation tasks will depend on your dataset and analysis requirements.

Remember that I can provide guidance, answer questions, and help with SQL queries or MongoDB queries if you encounter specific issues during your project. Feel free to ask for assistance with any part of your project, and I'll do my best to help you successfully complete it.

Sample SQL Queries for Data Exploration and Analysis:

Retrieve Data from the Employee Table:

```
SELECT *
FROM employee_table;
```

Calculate the Average Salary:

```
SELECT AVG(salary) AS average_salary
FROM employee_table;
```

Find the Highest-Paid Employee:

```
SELECT first_name, last_name, salary
FROM employee_table
ORDER BY salary DESC
```

```
LIMIT 1;
```

Sample SQL Query for Data Cleaning (e.g., Remove Duplicates):

To remove duplicates based on a specific column (e.g., employee_id):

```
DELETE e1
FROM employee_table e1
INNER JOIN employee_table e2
ON e1.employee_id = e2.employee_id
WHERE e1.rowid > e2.rowid;
```

Sample SQL Query for Data Transformation (e.g., Update Date Format):

To update date format (assuming date_column is in the format 'MM/DD/YYYY'):

```
UPDATE employee_table
SET date_column = TO_DATE(date_column, 'MM/DD/YYYY');
```

Phase 4 project – BIG DATA ANALYSIS

PROBLEM STATEMENT:

• Continue building the big data analysis solution by applying advanced Analysis techniques and visualizing the results.
• Apply more complex analysis techniques, such as machine learning Algorithms, time series analysis, or sentiment analysis, depending on the Dataset and objectives.
• Create visualizations to showcase the analysis results. Use tools like Matplotlib, Plotly, or IBM Watson Studio for creating graphs and charts.

SOLUTION:

Certainly, building a big data analysis solution that incorporates advanced Techniques and visualizations is essential for deriving meaningful insights from Your data. Let's continue with the process:

Step 1:

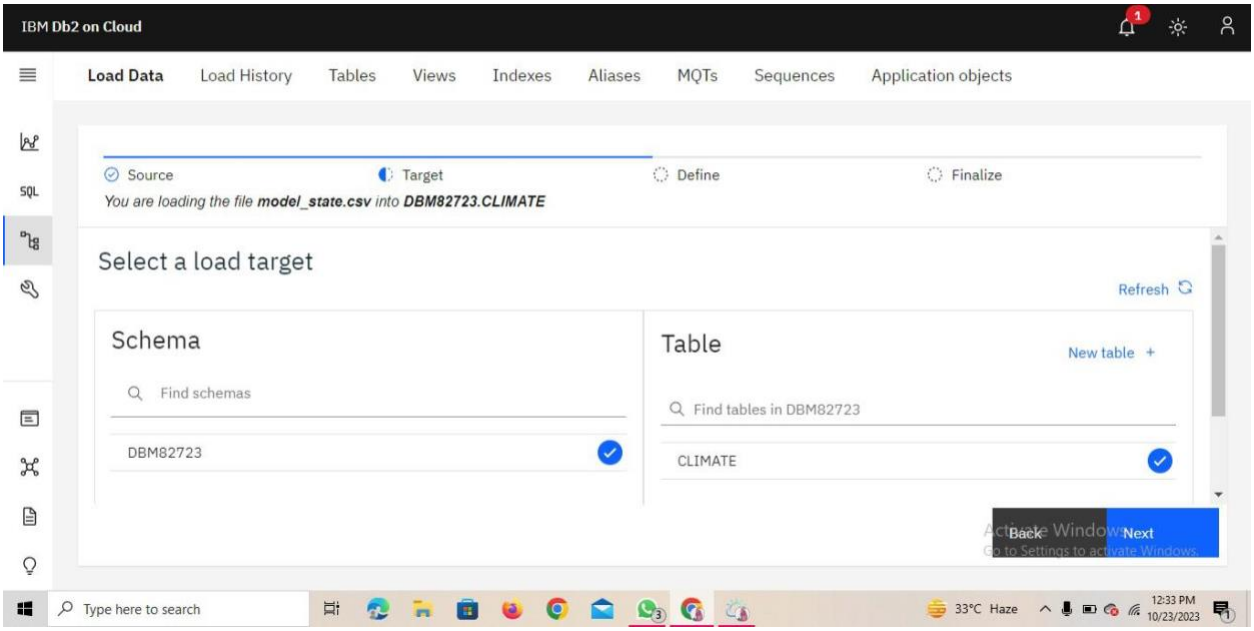
Download a CSV or xlsx file for upload in the DB2 database.

Example: open the wwb browser.

Search for the convenient topic to download database.(eg:kaggle,Data.world..)

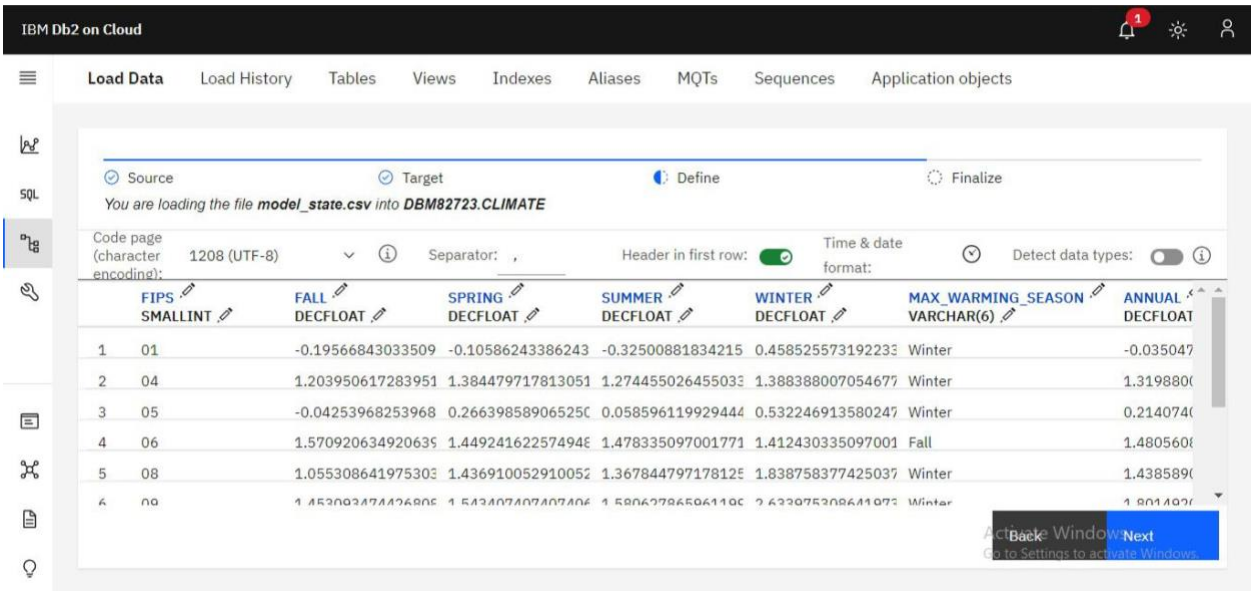
Step 2:

Create a data table in IBM Cloud DB2 Database.



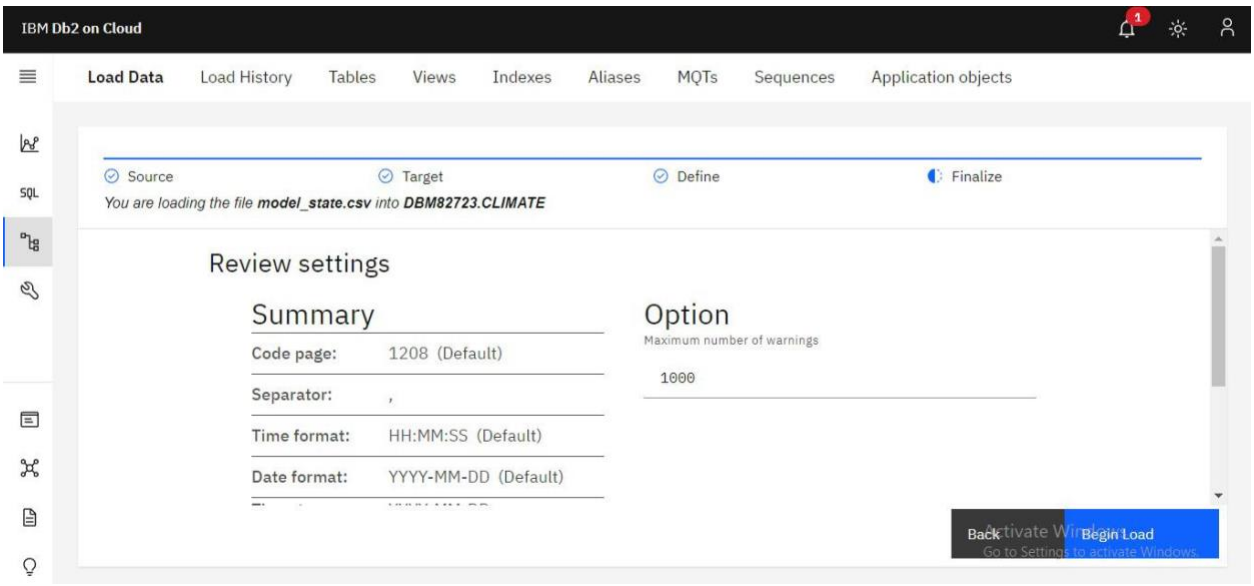
Step 3:

Upload the downloaded CSV. File in the database.



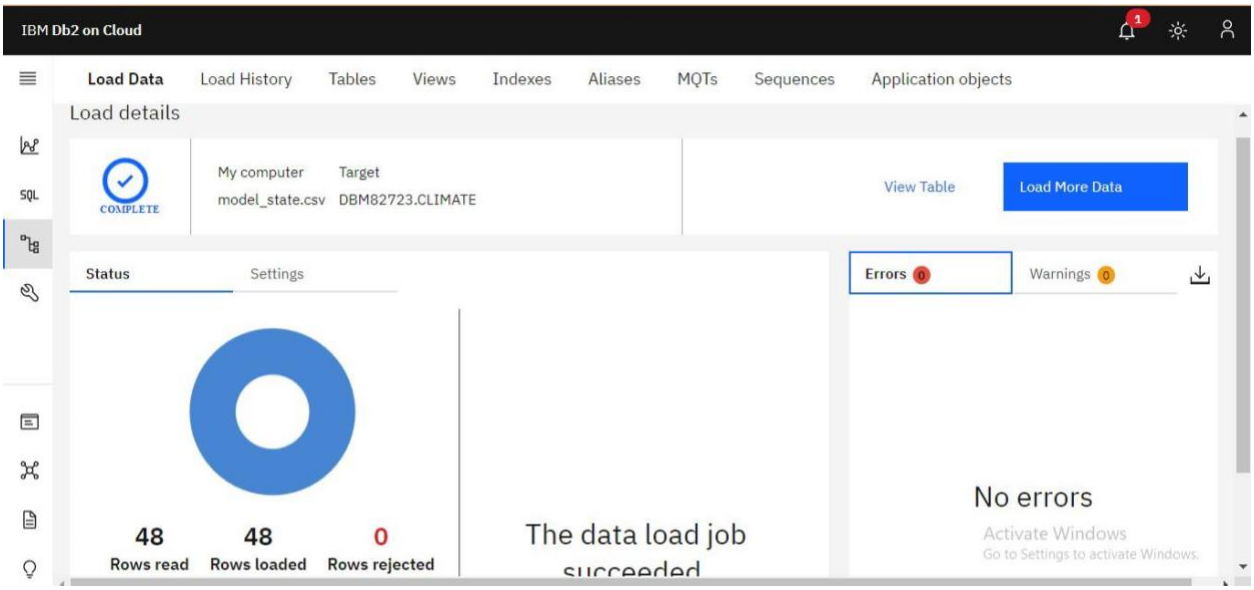
Step 4:

Finalize the uploading settings.



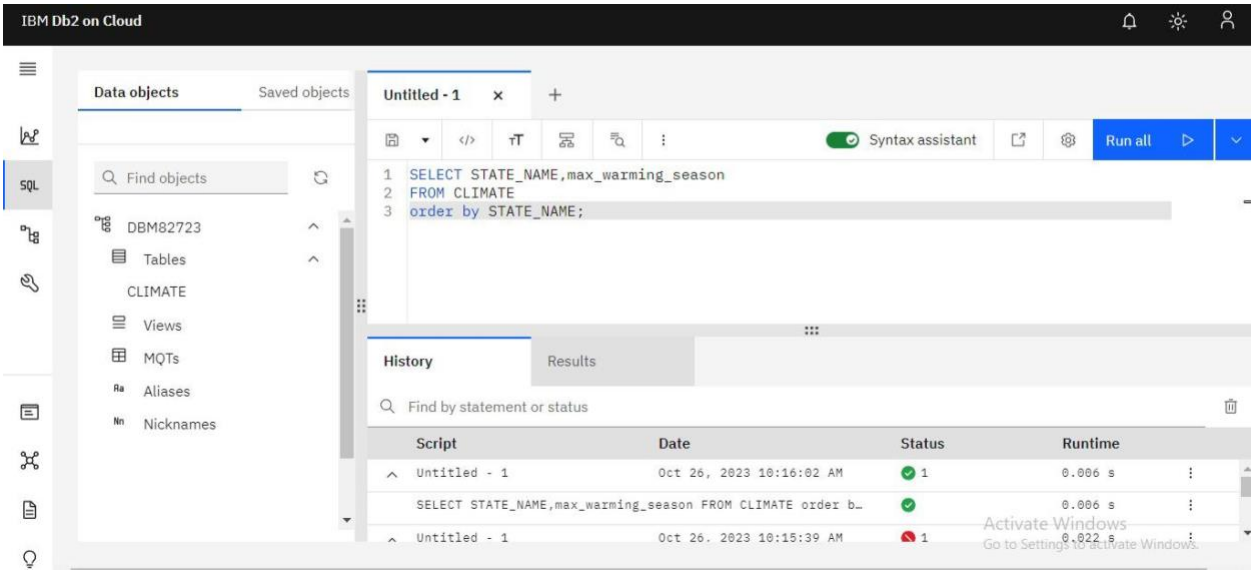
Step 5:

Run the loaded data to check it is contain error or not.



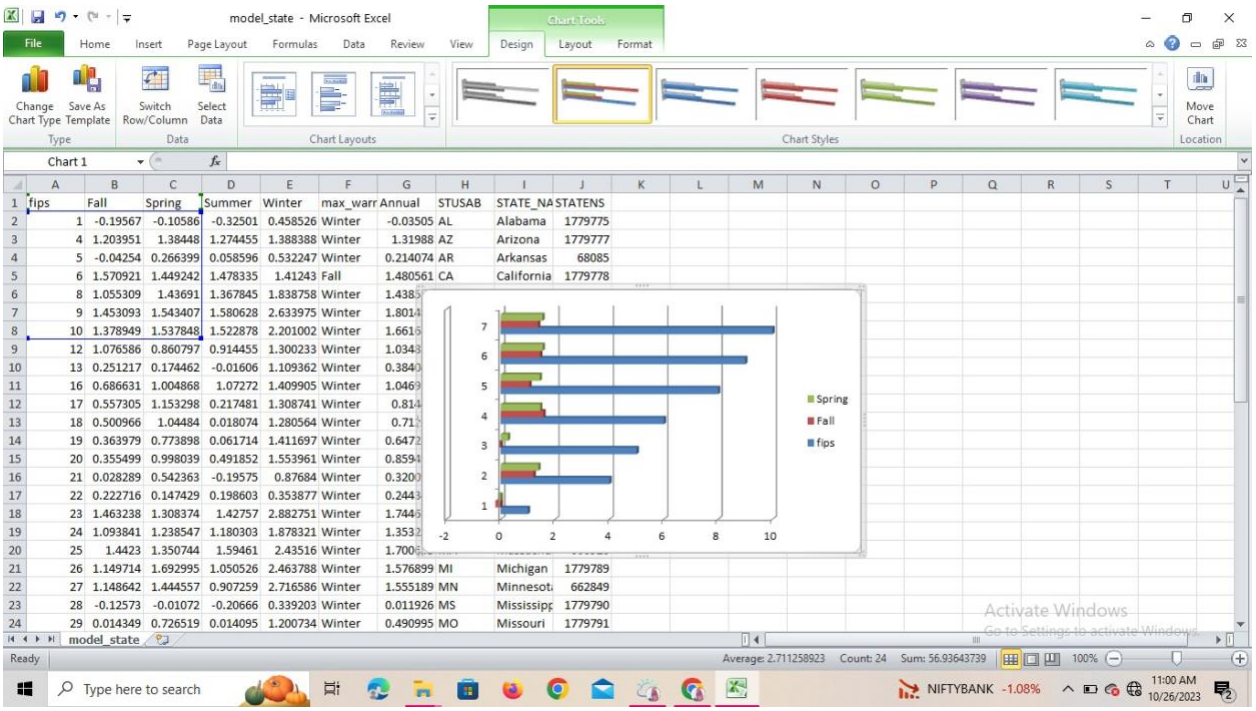
Step 6:

Create SQL queries to run the database table.



Step 7:

For development the analysis data we need to use the virtualization techniques in the datasets.

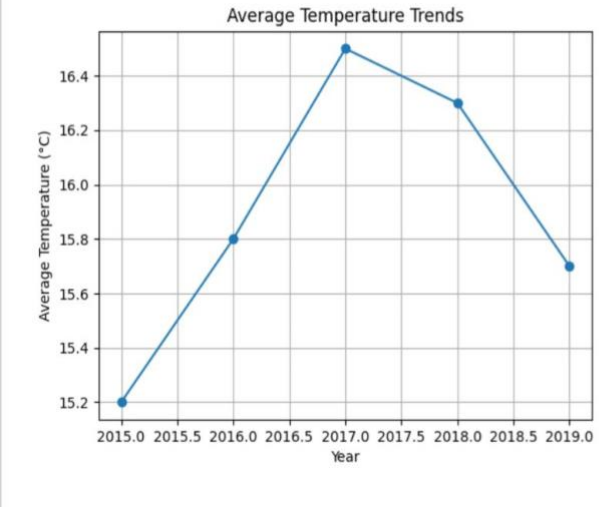


Step 8: Using python.

```python
# Example Python code for creating a line chart using Matplotlib

import matplotlib.pyplot as plt

years = [2015, 2016, 2017, 2018, 2019]
avg_temperatures = [15.2, 15.8, 16.5, 16.3, 15.7]
plt.plot(years, avg_temperatures, marker='o')
plt.title('Average Temperature Trends')
plt.xlabel('Year')
plt.ylabel('Average Temperature (°C)')
plt.grid(True)
plt.show()
```



Step 9:

Using Machine Learning techniques.

Select Appropriate Analysis Techniques:

Depending on the nature of your dataset and specific objectives, consider various

Advanced analysis techniques:

Machine Learning Algorithms: Use supervised or unsupervised machine learning

Algorithms like decision trees, random forests, support vector machines, or

Clustering algorithms for predictive modeling or pattern recognition.

Time Series Analysis: If your data involves time-based data points, use time

Series analysis techniques to identify trends, seasonality, and forecast future

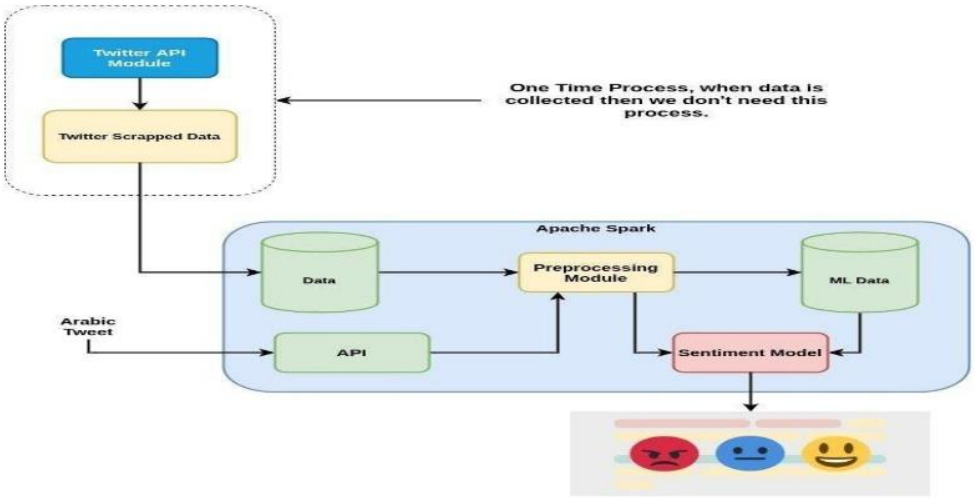
Values.

Sentiment Analysis: Apply natural language processing techniques to extract

Sentiment from text data, useful for social media or customer reviews analysis.

Example:

```
# Example Python code for sentiment analysis using NLTK
import nltk
from nltk.sentiment import SentimentIntensityAnalyzer
nltk.download('vader_lexicon')
sia = SentimentIntensityAnalyzer()
text = "The weather is wonderful and the scenery is breathtaking."
sentiment_score = sia.polarity_scores(text)
print(sentiment_score)
```



Conclusion:

Thus the ,Continue building the big data analysis solution by applying advanced analysis techniques And visualizing the results has been completed.