

CLOUD APPLICATION DEVELOPMENT:

PROJECT NAME: Big Data Analysis with IBM Cloud Databases

NAME: R.Pojan Kumar

REGISTER NO: 912421104027

PHASE 5: PROJECT DOCUMENTATION AND SUBMISSION



Introduction to Big Data and IBM Cloud

Big data refers to the large and complex sets of data that are difficult to process and analyze using traditional data processing methods. With the advent of new technologies like IBM Cloud, big data analysis has become more accessible and efficient. IBM Cloud provides a scalable and secure platform for storing, processing, and analyzing big data. With IBM Cloud, organizations can leverage advanced machine learning algorithms for predictive analysis or anomaly detection in big data.

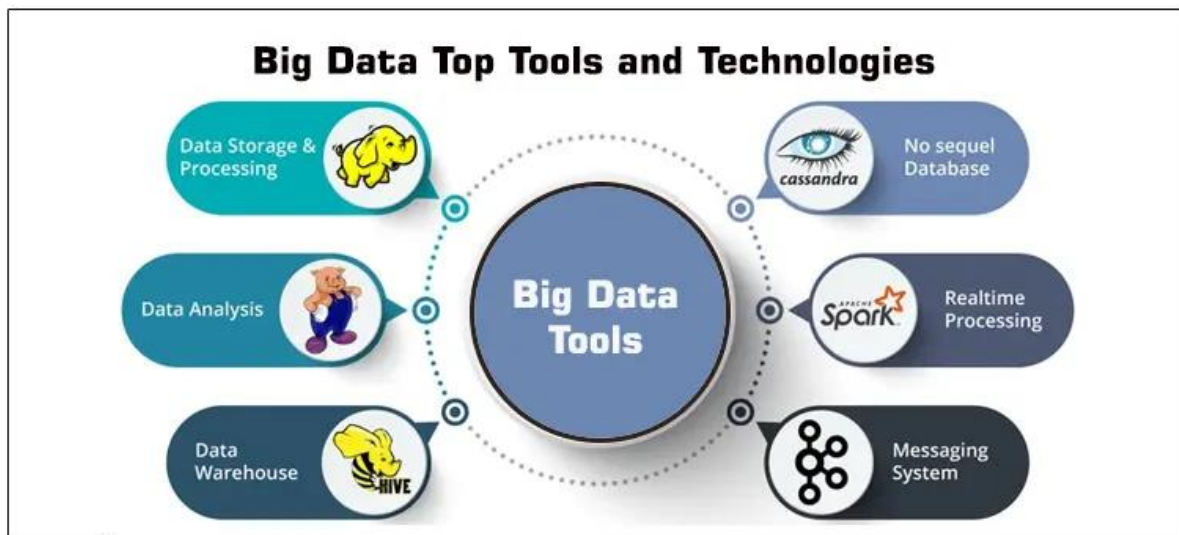
Project Objective:

The primary objective of this big data analysis project is to gain actionable insights from a large dataset to support informed decision-making and add value to the business. The project is aimed at identifying patterns, trends, and correlations within the data that can be used to improve business operations, optimize processes, or develop new strategies.



Design thinking:

- Understand the business context and stakeholders: Identify the key stakeholders involved, their roles, and their expectations from the analysis.
- Conduct interviews and workshops: Engage with stakeholders to understand their challenges, goals, and the questions they seek to answer using data.
- Define user personas: Create profiles representing different user types within the organization to understand their unique needs and perspectives.



DEVELOPMENT PHASES:



Development Part 1

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.



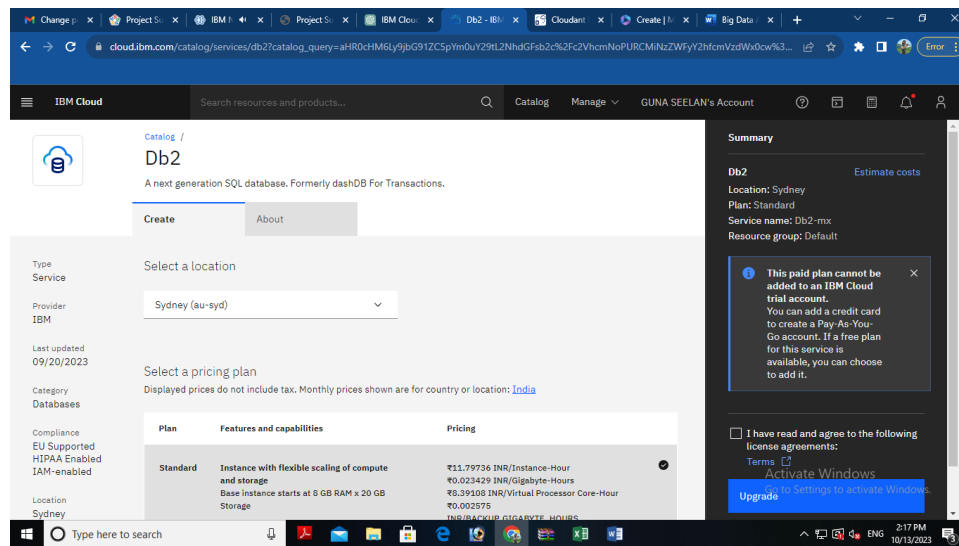
2. Choose the Appropriate Database Service:

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:

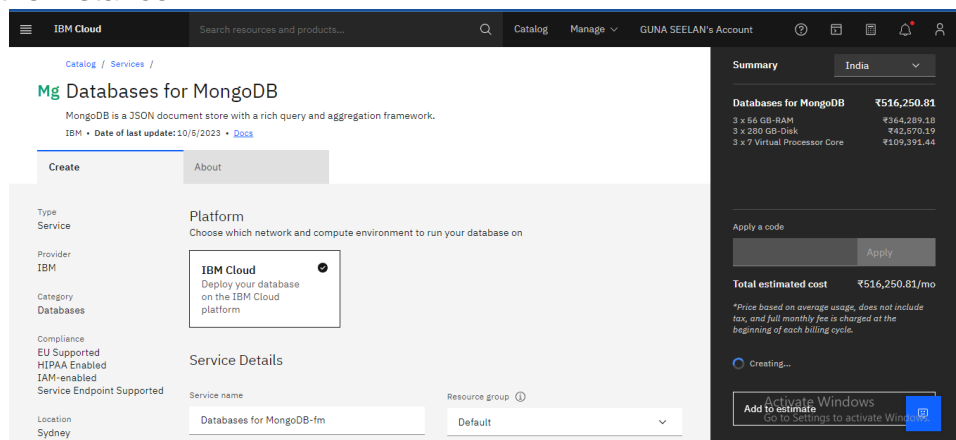
For Db2:

- ◆ 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.



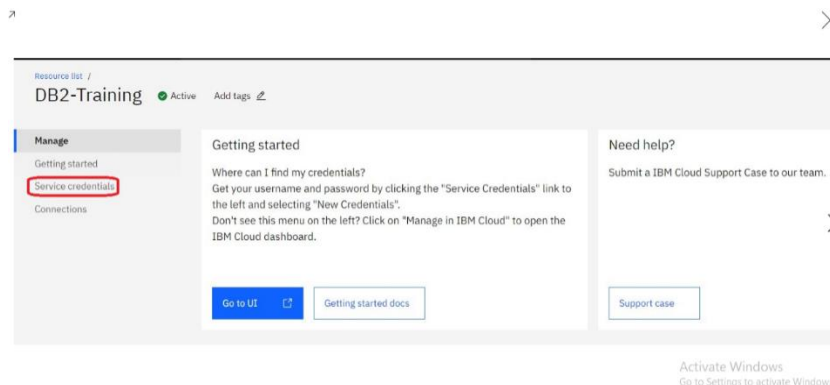
4. Develop Queries or Scripts:

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.

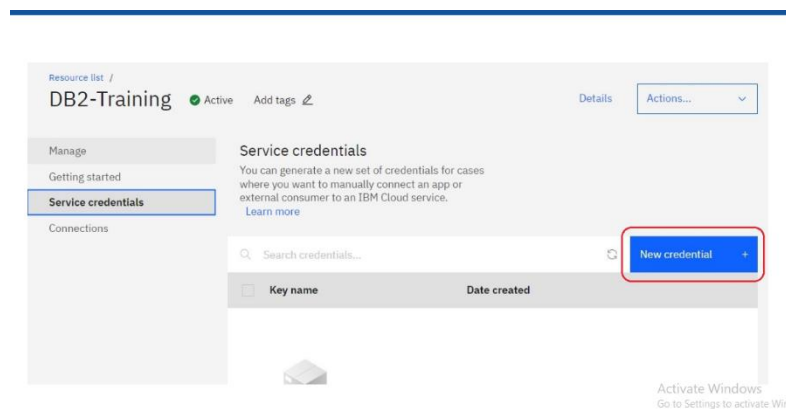
Creating Service Credentials the IBM DB2 database

- ◆ 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**

A screenshot of the 'Create credential' form in the IBM Cloud console. The form has two main input fields: 'Name' and 'Role'. The 'Name' field contains the text 'appCred'. The 'Role' field is set to 'Manager'. Below these fields is an 'Advanced options' section with a dropdown arrow. At the bottom of the form, there are two buttons: 'Cancel' and 'Add'. The 'Add' button is highlighted in blue.

- ◆ 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>

```

"db2": {
  "authentication": {
    "method": "direct",
    "password": "XXXXXXXXXX",
    "username": "XXXXXXXXXX"
  },

```

Activate Windows

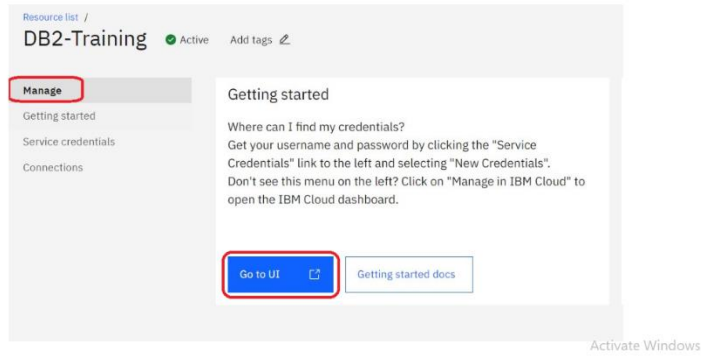
```

"hosts": [
  {
    "hostname": "fbd88901-ebdb-4a4f-a32e-9822b9fb237b.clogj3sd0tgtu0lqde00.databases.appdomain.cloud",
    "port": 32731
  }
]

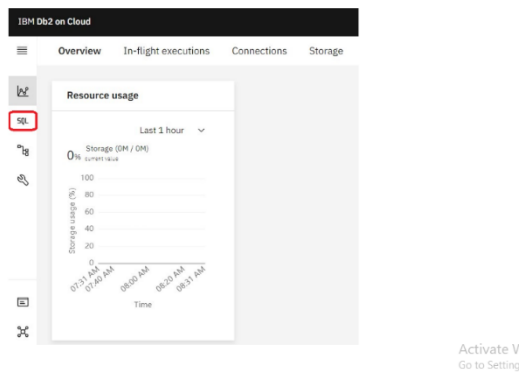
```

3. Setting up IBM DB2 database

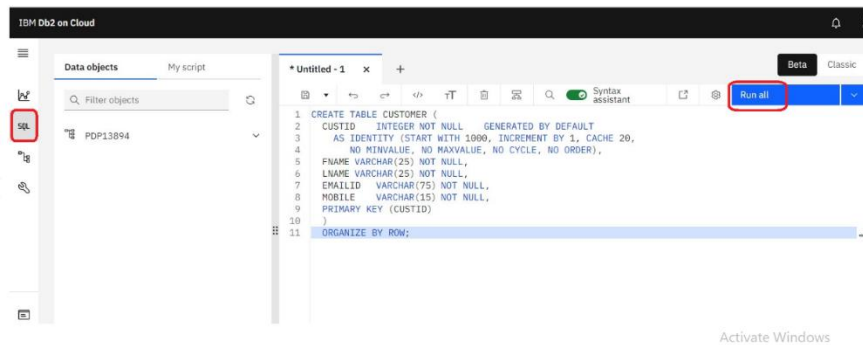
- ◆ In the resource list screen of IBM Cloud, click on the DB2 service (displayed under Services and software category) that you created, if the page is not already opened.
- ◆ From the service page, select the menu option "**Manage**" and click on Go to UI to launch the DB2 console



- ◆ 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

History			
Find history			
Script	Date	Status	Runtime
Untitled - 1	Sep 14, 2022 8:34:07 AM	✓ 1	0.251 s
CREATE TABLE CUSTOMER (CUSTID INTEGER NOT NULL GENERATED BY D...		✓	0.251 s

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 screenshot shows the IBM Db2 on Cloud console. The 'Connections' page is active, displaying instructions for connecting to the database. The 'Linux' tab is selected. A red box highlights the spanner icon in the left sidebar, which is used to access the Administration page. Another red box highlights the 'Download SSL Certificate' button in the 'Connection configuration resources' section, which is used to download the DB2 certificate.

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.

SSL Converter

Use this SSL Converter to convert SSL certificates to and from different formats such as pem, der, p7b, and pfx. Different platforms and devices require SSL certificates to be converted to different formats. For example, a Windows server exports and imports .pfx files while an Apache server uses individual PEM (.cer, .cert) files. To use the SSL Converter, just select your certificate file and its current type (it will try to detect the type from the file extension) and then select what type you want to convert the certificate to and click **Convert Certificate**. For more information about the different SSL certificate types and how you can convert certificates on your computer using OpenSSL, see below.

Certificate Conversion Options

Certificate File to Convert

No file chosen

Type of Current Certificate

DER/Binary

Type To Convert To

Standard PEM

Act
Go to

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:

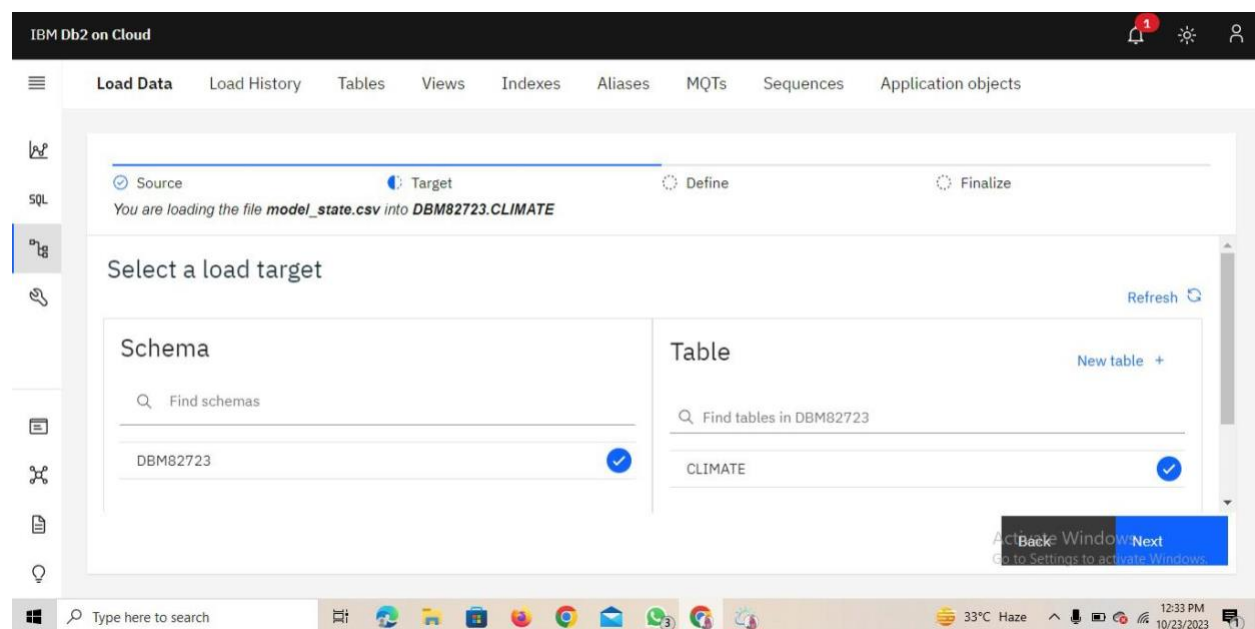
Download a CSV or xlsx file for upload in the DB2 database.

Example: open the wwv 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.

IBM Db2 on Cloud

Load Data | Load History | Tables | Views | Indexes | Aliases | MQTs | Sequences | Application objects

Source | Target | Define | Finalize

You are loading the file **model_state.csv** into **DBM82723.CLIMATE**

Code page (character encoding): 1208 (UTF-8) | Separator: , | Header in first row: ☒ | Time & date format: | Detect data types: ☐

	FIPS SMALLINT	FALL DECFLOAT	SPRING DECFLOAT	SUMMER DECFLOAT	WINTER DECFLOAT	MAX_WARMING_SEASON VARCHAR(6)	ANNUAL DECFLOAT
1	01	-0.19566843033509	-0.10586243386243	-0.32500881834215	0.458525573192233	Winter	-0.035047
2	04	1.203950617283951	1.384479717813051	1.274455026455033	1.388388007054677	Winter	1.319880
3	05	-0.04253968253968	0.266398589065250	0.058596119929444	0.532246913580247	Winter	0.214074
4	06	1.570920634920635	1.44924162257494E	1.478335097001771	1.412430335097001	Fall	1.480560
5	08	1.055308641975303	1.436910052910052	1.36784479717812E	1.838758377425037	Winter	1.438589
6	09	1.452003777777778	1.543707777777778	1.59067786506110E	2.622075208641073	Winter	1.901407

Back Window Next
Go to Settings to activate Windows

Step 4:

Finalize the uploading settings.

IBM Db2 on Cloud

Load Data | Load History | Tables | Views | Indexes | Aliases | MQTs | Sequences | Application objects

Source | Target | Define | Finalize

You are loading the file **model_state.csv** into **DBM82723.CLIMATE**

Review settings

Summary	
Code page:	1208 (Default)
Separator:	,
Time format:	HH:MM:SS (Default)
Date format:	YYYY-MM-DD (Default)

Option

Maximum number of warnings: 1000

Back Activate Windows Begin Load
Go to Settings to activate Windows

Step 5:

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

The screenshot shows the 'Load Data' interface in IBM Db2 on Cloud. The 'Load details' section indicates the job is 'COMPLETE' with a status of 'My computer' and 'Target' 'model_state.csv' and 'DBM82723.CLIMATE'. A large blue donut chart shows the progress: 48 Rows read, 48 Rows loaded, and 0 Rows rejected. The text 'The data load job succeeded' is displayed. On the right, there are buttons for 'View Table' and 'Load More Data'. Below the chart, there are tabs for 'Status' and 'Settings'. To the right of the chart, there are tabs for 'Errors' (0) and 'Warnings' (0). A message 'No errors' is displayed with a note to 'Activate Windows'.

Step 6:

Create SQL queries to run the database table.

The screenshot shows the 'SQL' interface in IBM Db2 on Cloud. The 'Data objects' panel on the left shows a tree view with 'DBM82723' expanded, showing 'Tables', 'Views', 'MQTs', 'Aliases', and 'Nicknames'. The 'CLIMATE' table is selected. The 'Untitled - 1' editor shows the following SQL query:

```
1 SELECT STATE_NAME,max_warming_season
2 FROM CLIMATE
3 order by STATE_NAME;
```

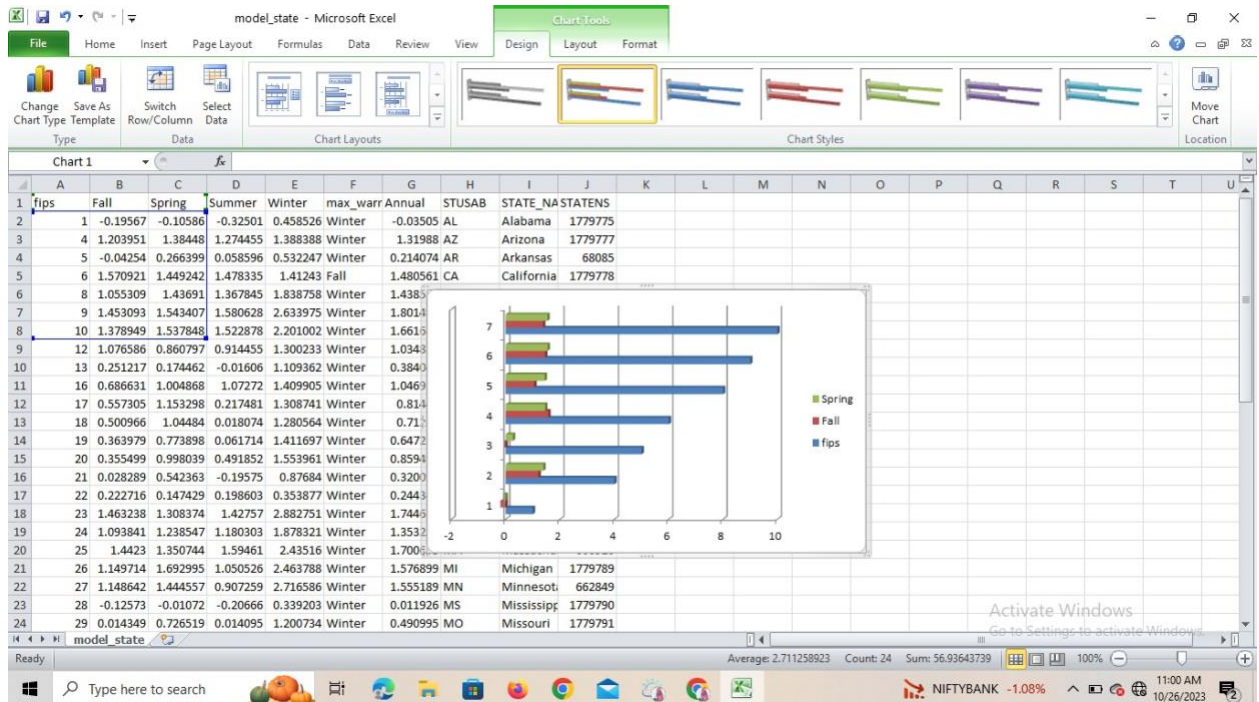
The 'Run all' button is visible. Below the editor, the 'History' tab shows a table of executed queries:

Script	Date	Status	Runtime
Untitled - 1	Oct 26, 2023 10:16:02 AM	✓ 1	0.006 s
SELECT STATE_NAME,max_warming_season FROM CLIMATE order b...		✓	0.006 s
Untitled - 1	Oct 26, 2023 10:15:39 AM	✗ 1	0.022 s

An 'Activate Windows' watermark is visible in the bottom right corner.

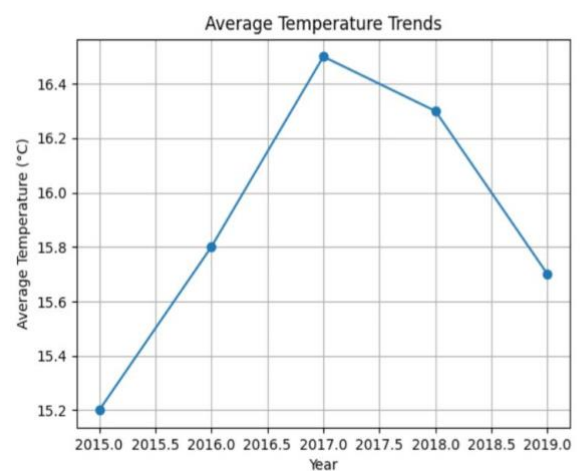
Step 7:

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



Step 8: Using python.

```
1 # Example Python code for creating a
  line chart using Matplotlib
2
3 import matplotlib.pyplot as plt
4
5 years = [2015, 2016, 2017, 2018, 2019]
6 avg_temperatures = [15.2, 15.8, 16.5,
  16.3, 15.7]
7 plt.plot(years, avg_temperatures,
  marker='o')
8 plt.title('Average Temperature Trends')
9 plt.xlabel('Year')
10 plt.ylabel('Average Temperature (°C)')
11 plt.grid(True)
12 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.

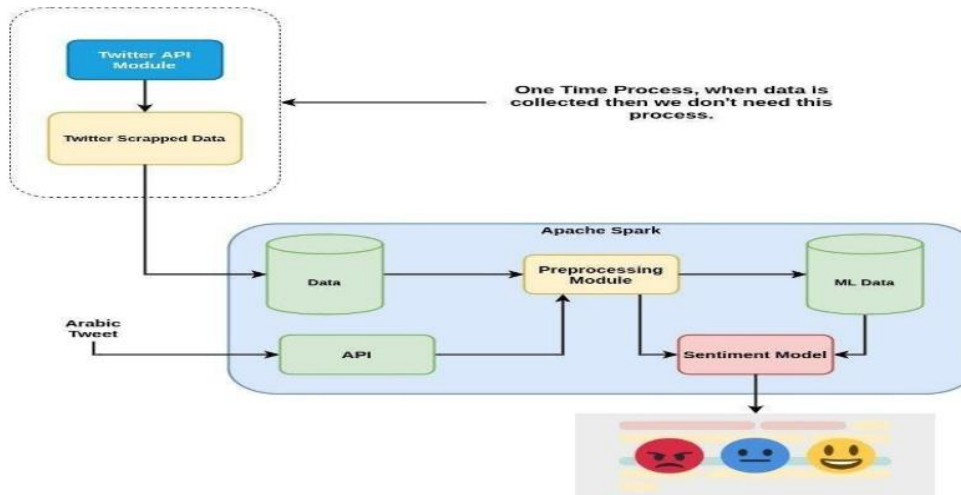
Sentiment Analysis: Apply natural language processing techniques to extract Sentiment from text data, useful for social media or customer reviews analysis.

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.

ADVANTAGES:

1. **Scalability:** IBM Cloud Databases are designed to handle the scalability needs of big data workloads. They allow you to easily scale your database resources up or down based on demand, ensuring that your data infrastructure can handle the increasing volume of data as your business grows.
2. **Managed Service:** IBM provides fully managed database services, which means they handle tasks such as provisioning, patching, backups, and high availability. This allows your data engineering and analysis teams to focus on the actual analysis work rather than database management.
3. **Data Security:** IBM Cloud Databases come with robust security features. They offer data encryption both in transit and at rest, access control, and authentication mechanisms. This is crucial for safeguarding sensitive and valuable data used in big data analysis.
4. **High Availability:** IBM Cloud Databases are built with high availability in mind. They typically offer failover and data redundancy capabilities to minimize downtime, ensuring that your analysis workloads are not disrupted.
5. **Multi-Cloud and Hybrid Deployment:** IBM Cloud provides the flexibility to deploy databases across multiple cloud providers, on-premises, or in a hybrid cloud environment. This is advantageous for organizations with complex data architectures or multi-cloud strategies.
6. **Support for Various Database Engines:** IBM Cloud Databases support various database engines, including relational databases (e.g., Db2), NoSQL databases (e.g., CouchDB), and time-series databases (e.g., InfluxDB). This enables you to choose the database engine that best suits your specific analysis needs.
7. **Integration with IBM Watson:** If you're interested in leveraging AI and machine learning for big data analysis, IBM Cloud Databases can easily integrate with IBM Watson services. This allows you to build AI-powered applications and derive deeper insights from your data.
8. **Developer-Friendly:** IBM Cloud Databases come with developer-friendly features, including RESTful APIs, SDKs, and connectors. This simplifies the development and integration of applications and services with the databases, facilitating data analysis.
9. **Cost Management:** IBM Cloud Databases offer various pricing options, including pay-as-you-go and reserved instances. This flexibility allows you to manage costs effectively, especially in big data scenarios where resource requirements may fluctuate.

10. **Global Data Centers:** IBM's global network of data centers allows you to deploy databases in regions that are geographically close to your users or data sources, reducing latency and improving data access performance.
11. **Data Analytics Services:** IBM Cloud provides integrated data analytics and AI services, such as IBM Watson Studio and IBM Cognos Analytics, which can be used in conjunction with IBM Cloud Databases for advanced big data analysis.


DISADVANTAGES:


1. **Cost:** The use of cloud services, including IBM Cloud Databases, can result in ongoing operational costs that may be significant, especially for large-scale big data analysis. Organizations need to carefully manage their cloud expenses to avoid unexpected budget overruns.
2. **Data Transfer Costs:** Moving large volumes of data into and out of the cloud can incur data transfer costs, particularly when dealing with big data. These costs can add up, so organizations should be mindful of data egress fees, especially for data-intensive analysis workloads.
3. **Latency:** While IBM Cloud offers global data centers, data access latency may still be a concern, especially for real-time or low-latency applications. For some big data analysis scenarios, local, on-premises solutions may offer lower latency.
4. **Data Sovereignty and Compliance:** Depending on your industry and the regions where you operate, you may face regulatory and compliance challenges. Data sovereignty, data residency, and compliance with local data protection laws can be complex issues when using a cloud provider's services.
5. **Vendor Lock-In:** Using IBM Cloud Databases may potentially lock your organization into their ecosystem, making it more challenging to migrate to another cloud provider or bring your database in-house. Migrating data between cloud providers can be complex and costly.
6. **Limited Customization:** While managed services are convenient, they may come with limitations in terms of database customization. If your big data analysis requires highly specialized configurations or database tuning, you might find that managed services do not provide the level of control you need.
7. **Security Concerns:** While cloud providers like IBM offer robust security features, data breaches and security incidents are always a concern. It's essential to take appropriate measures to secure your data and applications, including proper access controls and encryption.
8. **Resource Scaling Challenges:** Although cloud databases are designed for scalability, there can be challenges when scaling resources up or down to

meet specific analysis demands. Poorly managed scaling can lead to unexpected performance issues or costs.

9. **Data Ownership and Control:** Organizations may have concerns about relinquishing control of their data to a third-party cloud provider. Understanding the terms and conditions of data ownership is crucial, as well as having data backup and recovery plans in place.
10. **Downtime and Outages:** While cloud providers strive to offer high availability, there is always a risk of downtime or outages. Organizations should consider strategies for business continuity in the event of service disruptions.

CONCLUSION:

 In conclusion, leveraging IBM Cloud databases for big data analysis equips businesses with the tools and infrastructure necessary to extract actionable insights from vast datasets, enabling data-driven decision-making and fostering innovation and growth. The flexibility, scalability, security, and integration capabilities of IBM Cloud databases make them a compelling choice for organizations seeking to harness the power of big data.

 The overall conclusion of big data analysis using IBM Cloud databases is typically positive, given the robust tools and services IBM provides.