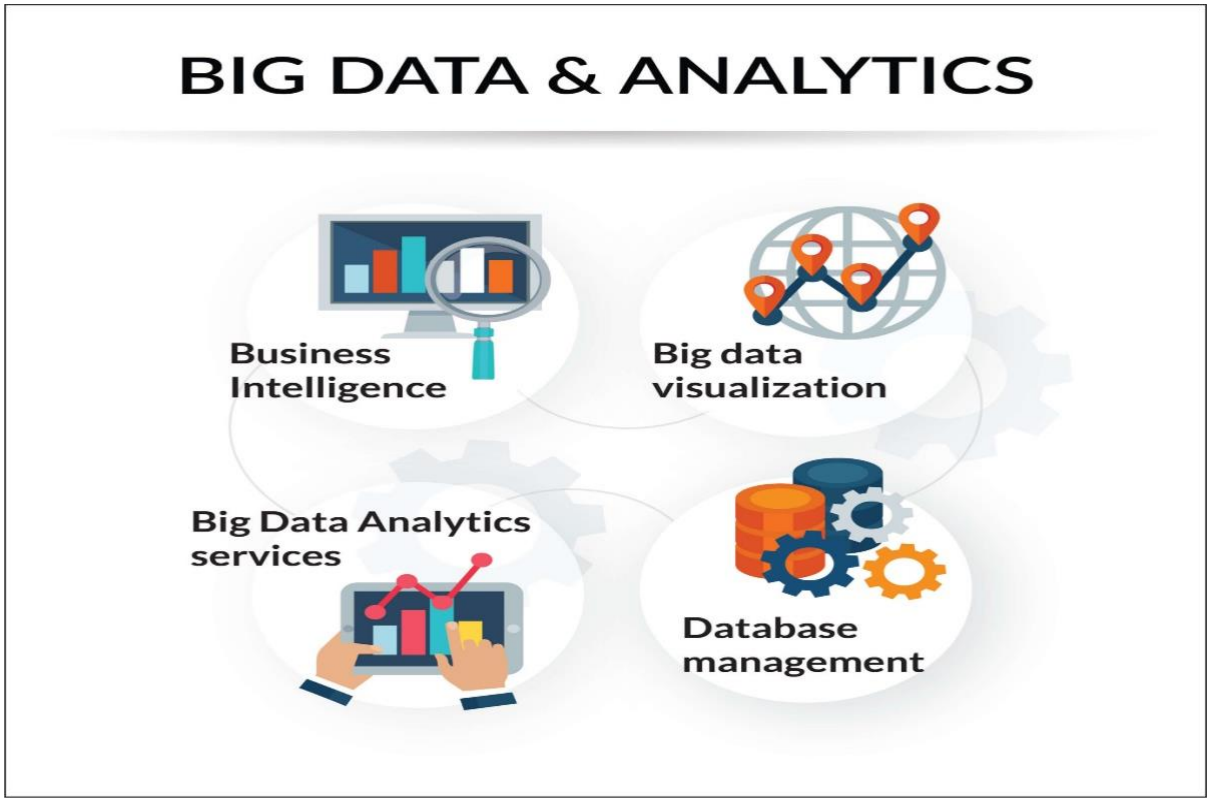
Big Data Analysis with IBM Cloud Databases

TEAM MEMBER

111421104091 : Praveen Kumar A

Phase-1 Document Submission

Project 5 : Big Data Analysis



Problem Definition:

The challenge we face today revolves around the effective handling and utilization of vast datasets, spanning fields like climate patterns and social trends. This challenge encompasses several key aspects: accessing the data itself, discovering valuable insights hidden within it, presenting complex data in a clear manner, extracting meaningful business insights, and fostering a structured approach for data exploration. In essence, the problem lies in the complexity of managing, analyzing, and deriving practical benefits from extensive and diverse datasets, which hinders our ability to tap into the vast potential of big data for informed decision-making and innovation.

Objective:

The objective for the problem statement is to leverage IBM Cloud Databases to perform comprehensive big data analysis, enabling the discovery of concealed insights within extensive datasets, such as climate trends and social patterns. This analysis should culminate in the visualization of these findings and the extraction of valuable business intelligence. The ultimate goal is to promote data-driven decision-making and exploration of the limitless potential of big data in various domains.

Design Thinking:

Abstract:

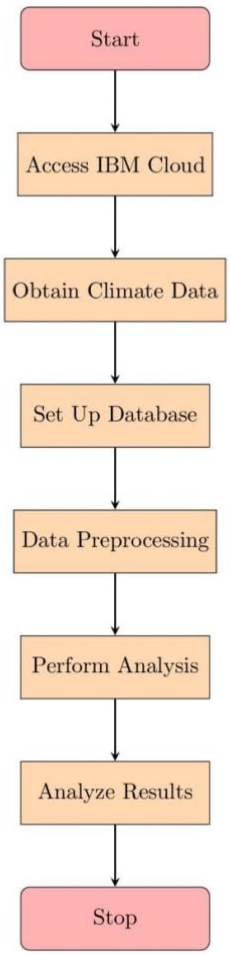
Big data analysis is the process of examining and analyzing large and complex datasets to uncover hidden patterns and insights. This can be a challenging task, as big data sets are often unstructured, semi-structured, and distributed across multiple sources.

IBM Cloud Databases offers a wide range of services that can be used for big data analysis. These services include:

- IBM Cloud Data Engine: A fully managed data lake service that provides a scalable and secure platform for storing, processing, and analyzing big data.
- IBM Cloud Object Storage: A highly scalable and durable object storage service that can be used to store any type of data, including big data.
- IBM Cloudant: A fully managed NoSQL database that is optimized for high performance and scalability.
- IBM Watson Studio: A cloud-based integrated development environment (IDE) for data scientists and machine learning engineers.

These services can be used together to build and deploy big data analytics solutions. For example, you could use IBM Cloud Data Engine to store and process your big data, IBM Cloud Object Storage to store your raw data, IBM Cloudant to store your processed data, and IBM Watson Studio to develop and deploy your analytics models.

Flowchart:



Modules:

Module 1: Data selection

Dataset : Climate dataset

When selecting data for big data analysis with IBM Cloud Database using a climate dataset, it is important to consider the following factors:

- Relevance: The data should be relevant to the specific business questions or problems that you are trying to answer.
- Completeness: The data should be complete and accurate.
- Variety: The data should be diverse, including both structured and unstructured data.
- Volume: The data should be large enough to provide meaningful insights.

Some examples of data that could be selected for a climate dataset include:

- Temperature data
- Precipitation data
- Sea level data
- Greenhouse gas emissions data
- Land use data

Link : [https://www.kaggle.com/datasets/nadanassershalaby/climate-dataset/download?datasetVersionNumber=1]

Module 2: Database setup

Once you have selected the data for your climate dataset, you need to set up a database to store the data. IBM Cloud Database offers a variety of database services, including Cloud SQL, Cloud Databases for PostgreSQL, and Cloud Databases for MongoDB.

The type of database that you choose will depend on the specific needs of your analysis. For example, if you are doing a lot of ad hoc queries, you may want to choose a database that is optimized for performance. If you are storing a large amount of unstructured data, you may want to choose a NoSQL database.

Module 3: Data Exploration

Once you have set up your database, you need to explore the data to get a better understanding of it. This can be done using a variety of tools, such as SQL queries, data visualization tools, and machine learning algorithms.

Some of the things that you should look for when exploring the data include:

- Outliers: Are there any data points that are significantly different from the rest of the data?
- Missing values: Are there any data points that are missing?
- Correlations: Are there any relationships between different variables?
- Trends: Are there any trends in the data over time?

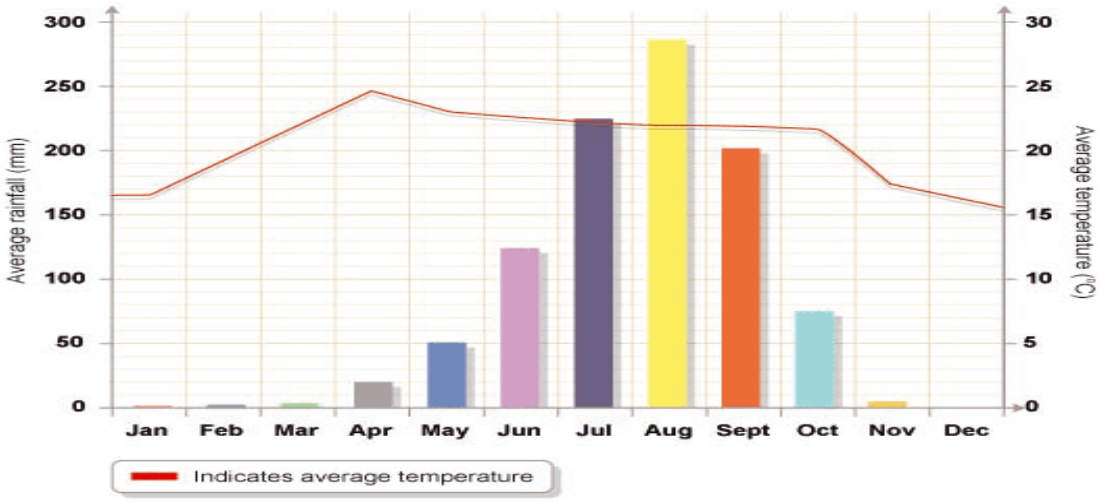
Module 4: Analysis Techniques

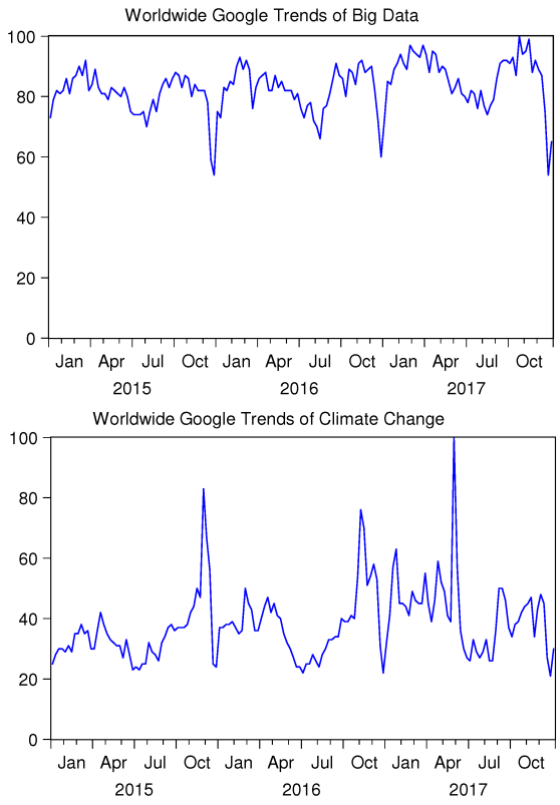
Once you have explored the data, you can start to apply various analysis techniques to answer your business questions. Some common analysis techniques include:

- Statistical analysis: Statistical analysis can be used to identify trends, correlations, and patterns in the data.
- Machine learning: Machine learning algorithms can be used to build predictive models and identify anomalies in the data.
- Data mining: Data mining techniques can be used to extract hidden insights from the data.

Module 5: Visualization

Data visualization is a powerful way to communicate insights from big data analysis to a wide audience. There are a variety of data visualization tools available, such as Tableau, QlikView, and Power BI.





When choosing a data visualization tool, it is important to consider the following factors:

- Audience: Who is the target audience for the visualization?
- Purpose: What is the purpose of the visualization?
- Data type: What type of data is being visualized?

Module 6: Business Insights

Big data analysis can be used to generate a variety of business insights, such as:

- Identifying trends: Big data analysis can be used to identify trends in customer behavior, market conditions, and environmental factors.
- Predicting outcomes: Big data analysis can be used to build predictive models that can be used to forecast future events.
- Improving efficiency: Big data analysis can be used to identify areas where processes can be improved and costs can be reduced.
- Developing new products and services: Big data analysis can be used to identify new customer needs and develop new products and services to meet those needs.

Conclusion :

Big data analysis with IBM Cloud Database can be used to generate valuable insights from climate datasets. By following the steps outlined above, you can select the right data, set up the right database, explore the data effectively, apply the right analysis techniques, visualize the results in a meaningful way, and generate business insights that can drive your business forward.