CLOUD APPLICATION DEVELOPMENT

GROUP 3

PROJECT 1: Big Data Analysis with IBM Cloud Databases

PROBLEM DEFINITION:

The project aims to use the extensive potential of data analysis and IBM cloud databases to unlock hidden patterns and trends within extensive datasets, spanning diverse fields such as climate trends and social patterns to increase the profits of businesses. We also aim to extract business insights that would be profitable in the near future for the company.

DATA WAREHOUSING:

CHOOSING THE RIGHT DATABASE: IBM Cloud offers various database options, including Db2, Db2 on Cloud, and Cloudant. Among these DB2, db2 on cloud and DB2 warehouse are used more significantly for structured data. As we are doing big data analytics, we choose to use IBM Cloudant and IBM databases for MongoDB, they are NoSQL database service that can handle semi-structured and unstructured data in large volumes and are ideal for storing huge quantity of JSON data and it is more flexible to use.

IMPORTING THE DATA: As the data would be of large amount of structured, semi-structured and unstructured data, importing those data is not a big task as we are going to use IBM Cloudant and MongoDB which are flexible for large volumes of data without any issues.

DATA PRE-PROCESSING:

We can use visualization techniques or statistical methods to detect irregular data points in order to explore the data and to identify the anomalies or outliers once the outliers in our data are identified we can use machine learning algorithms or outliers' detection techniques such as Z-score and IQR techniques. After identifying the outliers, we need to modify them with a reasonable value. It can be achieved by removing the outlier data or replacing them with a reasonable data.

PERFORMANCE AND SECURITY:

SCALABILITY AND PERFORMANCE: IBM cloud databases are known for their remarkable scalability and performance. Vertical scaling allows you to effortlessly boost resources, such as CPU and memory, as your data and workloads expand, resulting in improved system performance. Additionally, horizontal scaling ensures seamless scalability

by distributing data across multiple nodes, making it easier to handle increased workloads. Automated resource management further enhances performance as it dynamically scales and allocates resources, all while requiring minimal manual intervention. High availability is a key feature, with built-in redundancy and failover mechanisms that guarantee system availability, reducing interruptions to performance. Moreover, advanced caching techniques are employed to optimize query response times, ultimately bolstering the overall system performance and user experience.

SECURITY AND MAINTANENCE:

Regularly applying security patches and updates to database systems, which can help automate this process, but it's crucial to stay on top of updates to address known vulnerabilities and ensure the latest security features are in place. Implementing strict access controls and authorization mechanisms to restrict who can access and modify the database. Utilizing role-based access control (RBAC) to grant appropriate permissions to users and ensure that only authorized personnel can interact with the data. Configuring the default security policies offered by the IBM Cloud Databases and setting up monitoring and auditing tools to continuously track database activity and identify any suspicious or unauthorized access.

DATA ANALYSIS:

VISUALIZATION: Enhancing the communication of the data insights by leveraging visualization tools such as IBM Cognos or Tableau by showcasing the optimal insights in the form of a graphical representation.

AUTOMATION: Utilize IBM Cloud's automation features to automate routine tasks and effectively scale resources as needed at any point of time to reduce the complexity and time involved to every work from the initial stage.

EXPLORATORY DATA ANALYSIS (EDA): EDA is a crucial step in data analysis as it helps in understanding the structure, patterns, and relationships within the dataset. By visualizing and summarizing the data, EDA can reveal any outliers, missing values, or potential problems with the data quality. It also helps in identifying the appropriate statistical techniques and models for further analysis.