**NANMUDHALVAN**

**-** **CAD101 Cloud Application Development - Group 1**

**COLLEGE NAME: JEPPIAAR INSTITUTE OF TECHNOLOGY**

**TITLE: 2106-Data Warehousing with IBM Cloud Db2 Warehouse**

**TEAM NAME: Proj\_228503\_Team\_1**

**TEAM MEMBERS:**

1. **Jayasutharasan S**
2. **Manoj M**
3. **Prakash raj M**
4. **Santhosh V**
5. **Smaronshiju**
6. **Justin anto F**

**PHASE2: Innovation**

In this Phase we must put our design into innovation to solve the problem.

So, we must Consider incorporating advanced analytics tools or machine learning models for predictive analysis within the data warehouse.

**Objectives:**

The goal of this innovative design is to enhance the existing data warehousing project by incorporating Flask, IBM Db2, and predictive analytics. The objective is to provide data architects with powerful tools to not only analyse historical data but also predict future trends, thereby empowering informed decision-making.

**1. Choose Analytics Tools or Models:**

In this phase, we decided to enhance our data warehousing project by introducing advanced analytics. To do this, we chose the Scikit-learn library in Python. Scikit-learn is a versatile and widely used library that provides a range of machine learning algorithms and tools. We selected it as our analytics tool to implement a machine learning model for predictive analysis within our data warehouse.

**2. Data Preparation:**

Before we could apply advanced analytics, it was crucial to ensure that our data was well-prepared and suitable for analysis. Data preparation involves tasks such as cleaning, transformation, and handling missing values. One critical step was to remove duplicate records from our dataset to maintain data integrity and prevent skewed results in our analysis.

**3. Feature Engineering:**

Feature engineering is a critical step in building machine learning models. While we kept this example simple, feature engineering often involves creating new features or modifying existing ones to enhance the predictive power of the model. In real-world scenarios, feature engineering might include tasks like scaling, one-hot encoding, or creating complex derived features based on domain knowledge.

**4. Model Training:**

With our data prepared and features engineered, we proceeded to train a machine learning model. In this case, we used a Random Forest Classifier. Model training involves feeding the algorithm with historical data, allowing it to learn patterns and relationships within the data. We divided our data into a training set (used for model training) and a testing set (used for model evaluation).

**5. Model Evaluation:**

To assess the performance of our model, we evaluated it using accuracy. Accuracy measures the proportion of correctly predicted outcomes. However, in more complex scenarios, you may need to employ additional metrics such as precision, recall, F1-score, or area under the ROC curve (AUC) for a more comprehensive assessment of model performance.

**6. Model Deployment:**

Model deployment refers to the process of making your trained machine learning model accessible for use in a production environment. The specifics of deployment can vary depending on your data warehouse technology. For instance, you might export the trained model as a serialized object and then load it within the data warehouse when needed. The goal is to seamlessly integrate the model into your data processing pipeline.

**7. Integration with Data Warehouse:**

To derive value from our predictive model, we made necessary modifications to our data warehouse queries and data processing. This allowed us to incorporate the results of predictive analysis into our data warehouse. For example, we might have added columns with predictions from the model or used the model's output to make data-driven decisions within the data warehouse.

**Conclusion:**

In summary, Phase 2 involved enhancing our data warehousing project by incorporating advanced analytics using a machine learning model. We selected Scikit-learn, prepared our data, engineered features, trained the model, evaluated its performance, and integrated it into our data warehouse. This addition of advanced analytics empowers data-driven decision-making and unlocks valuable insights within our data warehousing project.