## Phase 2: Innovation

In this phase , we Consider incorporating advanced machine learning algorithms for predictive analysis or anomaly detection in the big data.

### 1. Choose Analytics Tools or Models:

In this phase, we decided to enhance our Big data analytics project by introducing advanced Predictive 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 .

### 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 check for anomaly in our dataset, Anomaly detection in big data is a crucial task that involves identifying patterns or data points that deviate significantly from the norm or expected behavior.

### 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 Big data technology. The model is deployed in IBM cloud database. The goal is to seamlessly integrate the model into your data processing pipeline.

In summary, Phase 2 involved enhancing our Big data analysis 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 . This addition of advanced analytics empowers data-driven decision-making and unlocks valuable insights within our Big data analysis project. Feel free to use this detailed explanation to communicate your work effectively to your staff.