**Description:** As a consultant specializing in risk management for financial institutions, you've been assigned to develop a loan default prediction system using Apache Spark. The lending institution wants to assess the creditworthiness of loan applicants and identify potential defaulters to minimize loan losses and optimize risk management practices. Your task is to leverage Apache Spark to analyze historical loan data posted [here](https://www.kaggle.com/datasets/nikhil1e9/loan-default), build predictive models, and deploy a real-time prediction pipeline for loan default detection.

**Tasks:**

1. **Data Acquisition and Preprocessing:**
   * Gather historical loan data from the lending institution's databases or data providers.
   * Preprocess the data by handling missing values, encoding categorical variables, and normalizing numerical features.
2. **Feature Engineering:**
   * Extract relevant features from the loan application data that may influence the likelihood of default.
   * Engineer additional features such as debt-to-income ratio, credit utilization, payment history, and borrower demographics.
3. **Model Development:**
   * Split the historical loan data into training and testing sets, ensuring a balanced distribution of defaulted and non-defaulted loans.
   * Build loan default prediction models using machine learning algorithms supported by Apache Spark, such as logistic regression, decision trees, random forests, or gradient-boosted trees.
   * Experiment with ensemble methods and hybrid models to improve prediction accuracy.
4. **Model Evaluation:**
   * Evaluate the performance of each prediction model using appropriate metrics (e.g., accuracy, precision, recall, F1-score, ROC AUC) on the test dataset.
   * Conduct cross-validation to assess the robustness of the models and identify potential overfitting.
5. **Real-Time Prediction Pipeline:**
   * Develop a Spark streaming application to ingest loan application data in real-time and generate predictions on the fly.
   * Integrate the trained prediction models into the streaming pipeline to classify loan applications as high or low risk.
   * Implement alerting mechanisms to notify loan officers or risk managers when high-risk applications are detected.
6. **Performance Optimization:**
   * Optimize Spark job configurations and parallelism settings to enhance the scalability and efficiency of the real-time prediction pipeline.
   * Explore techniques for distributed model training and inference to handle large volumes of data and improve processing speed.
7. **Documentation and Deployment:**
   * Document the entire loan default prediction system, including data preprocessing steps, feature engineering techniques, model selection criteria, and deployment architecture.
   * Provide guidelines for maintaining and updating the prediction pipeline over time, including retraining models and adapting to changing lending practices.

**Deliverables:**

1. Jupyter notebook or Python script containing the Spark code for data preprocessing, modeling, and real-time prediction.
2. Report documenting the loan default prediction process, including data exploration, feature engineering techniques, model evaluation results, and performance optimizations.
3. Visualization of model performance metrics, real-time monitoring dashboards, and loan default alerts.
4. Deployment package for the prediction system, including configuration files and setup instructions.

**Additional Notes:**

* Understand the factors influencing loan defaults and refine the prediction models accordingly.
* Emphasize the importance of continuous monitoring and model recalibration to adapt to changing economic conditions and regulatory requirements.
* Consider the ethical implications of loan default prediction and ensure fairness and transparency in the model development process.
* Provide training and support document to the lending institution's technical team for maintaining and operating the loan default prediction system.