**Description:** As a cybersecurity consultant, you've been tasked with developing a web page phishing detection system using Apache Spark. Phishing attacks pose a significant threat to internet users by tricking them into disclosing sensitive information or installing malware. Your objective is to leverage Apache Spark to analyze web page features and build a classification model to differentiate between legitimate and phishing web pages. The data is posted [here](https://www.kaggle.com/datasets/danielfernandon/web-page-phishing-dataset)

**Tasks:**

1. **Data Investigation and Preprocessing:**
   * Explore and EDA for the phishing data.
   * Preprocess the data by extracting relevant features from HTML content, such as URL length, domain age, presence of HTTPS, and HTML form submission.
2. **Feature Engineering:**
   * Extract additional features from the web page data that may indicate phishing behavior, such as presence of suspicious keywords, redirects,…
   * Normalize numerical features and encode categorical variables for model compatibility.
3. **Model Development:**
   * Split the dataset into training and testing sets, ensuring a balanced distribution of legitimate and phishing web pages.
   * Build a phishing classification model using machine learning algorithms supported by Apache Spark, such as logistic regression, decision trees, random forests, or gradient-boosted trees.
   * Experiment with different feature combinations and model architectures to improve classification accuracy.
4. **Model Evaluation:**
   * Evaluate the performance of the classification 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 model and identify potential overfitting.
5. **Real-Time Detection Pipeline:**
   * Develop a Spark streaming application to ingest web page data in real-time and classify pages as legitimate or phishing.
   * Integrate the trained classification model into the streaming pipeline to make predictions on incoming web pages.
   * Implement alerting mechanisms to notify users or administrators when phishing activity is detected.
6. **Performance Optimization:**
   * Optimize Spark job configurations and parallelism settings to enhance the scalability and efficiency of the real-time detection pipeline.
   * Explore techniques for distributed feature extraction and model inference to handle large volumes of web page data and improve processing speed.
7. **Documentation and Deployment:**
   * Document the entire web page phishing classification system, including data preprocessing steps, feature engineering techniques, model selection criteria, and deployment architecture.
   * Provide guidelines for maintaining and updating the detection pipeline over time, including retraining models and adapting to emerging phishing tactics.

**Deliverables:**

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

**Additional Notes:**

* Understand the characteristics of phishing web pages and refine the classification model accordingly.
* Emphasize the importance of continuous monitoring and model updates to adapt to evolving phishing techniques and patterns.
* Consider the ethical implications of web page classification and ensure fairness and transparency in the model development process.
* Provide training and support document to users or administrators for interpreting detection results and responding to phishing threats effectively.