**Problem Statement:** Smart News Summarizer for UPSC Aspirants

**Objective:** UPSC preparation requires staying updated on current affairs, but the flood of daily news can be overwhelming. Our mobile app simplifies this by summarizing key headlines into a single page and allowing users to filter news by categories like sports, politics, and economy. You can also view monthly news summaries to stay on top of important developments. Leveraging advanced Machine Learning (ML), Natural Language Processing (NLP), Big Data technologies, and Blockchain, the app ensures accurate summaries, organized content, and reliable data.

**Requirements:**

1. **Data Collection:**

* Newspaper PDFs uploaded by users.
* Metadata such as upload time, user preferences, and selected categories.

1. **Infrastructure:**

* High-performance computing resources for training ML models.
* Secure and compliant data storage solutions (e.g., MongoDB, Amazon S3).
* Blockchain platform for secure transaction records (e.g., Hyperledger, Ethereum).

1. **Software and Tools:**

* Big Data processing frameworks (e.g., Apache Hadoop, Apache Spark).
* ML libraries (e.g., TensorFlow, PyTorch, Scikit-Learn).
* NLP libraries (e.g., NLTK, SpaCy, BERT).
* Blockchain development tools (e.g., Solidity, Hyperledger Composer).
* Data processing tools (e.g., Pandas, NumPy).
* Real-time data processing platforms (e.g., Apache Kafka, Spark Streaming).

**Processing Steps:**

1. **Data Ingestion and Preprocessing:**

* Collect and ingest newspaper PDFs uploaded by users.
* Extract text from PDFs using libraries like PyMuPDF or pdfminer.
* Preprocess text data (handling missing values, normalization, tokenization).

1. **Blockchain Integration:**

* Develop smart contracts to record and verify summaries and user interactions on the blockchain.
* Implement consensus mechanisms to ensure data integrity and security.
* Store hashes of summarized content on the blockchain for immutability.

1. **Feature Engineering:**

* Extract key phrases and entities from text using NLP techniques.
* Create composite features combining text data and contextual information.

1. **Model Development:**

* **Text Summarization Model:**
  + Train a model for summarization.
* **Text Classification Model:**
* Train ML model to categorize news into sports, politics, economy, etc.

1. **Big Data Processing:**

* Use Apache Spark for large-scale data processing and analysis.
* Implement ETL (Extract, Transform, Load) processes for efficient data handling.

1. **System Integration:**

* Integrate ML models and blockchain components into the backend system.
* Develop a dashboard for real-time monitoring of summaries and user interactions.
* Implement an alert system to notify users of new summaries and category updates.

1. **Testing and Validation:**

* Conduct extensive testing using historical data and simulated scenarios.
* Validate models’ performance using metrics such as accuracy, precision, recall, and F1 score.
* Perform scalability and stress testing to ensure the system can handle large volumes of data and user interactions.

**Expected Outcomes:**

1. **Enhanced User Experience:**

* Easy-to-use mobile application for summarizing and categorizing news.
* Real-time updates and notifications based on user preferences.

1. **Improved Summarization and Categorization:**

* Accurate and concise summaries of newspaper articles.
* Effective categorization of news into relevant sections.

1. **Data Security and Integrity:**

* Immutable records of summaries ensuring data integrity.
* Secure and transparent user interactions using blockchain.

**Deliverables:**

1. **Mobile Application:**

* Fully functional mobile app with PDF upload, summarization, and category filtering features.
* User-friendly interface and seamless user experience.

1. **Technical Documentation:**

* Detailed documentation of data ingestion, preprocessing, feature engineering, ML models, and blockchain integration.
* API documentation for system integration.

1. **Performance Report:**

* Comprehensive report on model performance metrics and validation results.
* Insights from scalability and stress testing.

1. **Deployment Plan:**

* Step-by-step guide for deploying the system in the production environment.
* Maintenance and update schedules for continuous improvement.

1. **User Training:**

* Training materials and sessions for end-users.
* FAQs and troubleshooting guide for users.