**A structured overview of the problem-solving approach**:

### **1.** Understanding the Problem Statement

* Analyze news articles related to 10 major textile dye suppliers.
* Identify risks impacting the company’s supply chain.
* Classify risks into five categories:
  1. Geopolitical and Regulatory Risks
  2. Agricultural and Environmental Risks
  3. Financial and Operational Risks
  4. Supply Chain and Logistics Risks
  5. Market and Competitive Risks
* Provide actionable insights for the supply chain management team.

### 2. Data Extraction & Processing

#### **Step 1: Loading the JSON File**

* Read the JSON file containing news article links.
* Extract relevant metadata (Filtered data based on Publication year of 2023 and 2024)

#### **Step 2: Web Scraping**

* Did web scraping for missing articles.
* However, did not get articles for all missing articles due to inaccessibility of websites.
* Removed those records which does not have full articles.
* Created a new column for suppliers and filter out the records where these supplier names present

#### **Step 3: Data Cleaning & Preprocessing**

* Remove HTML tags, special characters, and unnecessary spaces.
* Convert text to lowercase for consistency.
* Tokenization: Split text into words/sentences. Tokenize the text and perform lemmatization/stemming.
* Stopword removal to focus on meaningful words.
* Lemmatization for reducing words to their root form.
* Named Entity Recognition (NER) to extract supplier names, locations, and risk-related terms.
* Create features such as word frequency, sentiment scores, and topic modeling outputs.

### 3. Risk Classification Using NLP & Machine Learning

**Define Risk Categories**

Create a labeled dataset by manually tagging a subset of 30 articles into the five risk categories:

* 1. Geopolitical and Regulatory Risks
  2. Agricultural and Environmental Risks
  3. Financial and Operational Risks
  4. Supply Chain and Logistics Risks
  5. Market and Competitive Risks

Include sentiment analysis to identify the direction of risk (positive or negative).

### Optionally, use zero-shot classification models (e.g., Hugging Face Transformers) to pre-label articles for semi-supervised learning.

#### Step 4: Feature Engineering

* TF-IDF (Term Frequency-Inverse Document Frequency) for keyword extraction.
* Word embeddings (Word2Vec, BERT) for better text representation.
* Sentiment Analysis (VADER, TextBlob) to determine risk direction (+ve or -ve).

#### Step 5: Model Selection & Training

* **Supervised Learning Approach:**
  + Use machine learning models (Logistic Regression, Random Forest, SVM) to classify risks.
  + Fine-tune hyperparameters and evaluate performance.
  + Applied SMOTE

#### Step 6: Risk Categorization & Visualization

* Assign risk labels to articles based on model predictions.