**🧭 Step 1: Understanding the Problem**

**Q1:** What is the business objective of this project? What kind of fraud are we trying to detect?

The goal is to build a machine learning model that can identify potentially fraudulent e-commerce transactions in real-time with high precision and low false negatives. This helps the business reduce financial losses due to fraud while minimizing disruption for legitimate users.

If a fraudster uses a stolen card and transacts via a known fraudulent device or an unusual email domain from an IP far from the cardholder’s usual location, the model should flag this transaction in real time before approval.

**🧭 Step 2: Understanding the Data**

**Q2:** What files do you have in the dataset? How do you plan to merge or use them?

The dataset is split into training and testing sets for both transactions and identity information. The key files are:

* train\_transaction.csv & test\_transaction.csv — contain transaction-level data including the target variable isFraud.
* train\_identity.csv & test\_identity.csv — contain identity and device metadata for a subset of transactions.

Transaction and identity datasets can be merged using the common key TransactionID. This will create a unified view for both training and testing data, where each row represents a transaction enriched with identity-based features such as device type, browser, IP address, etc.

**Q3:** Which columns will help you identify customer behavior vs. device behavior?

**Q4:** What is the proportion of fraud vs. non-fraud in the dataset? Why does this matter?

**🧭 Step 3: Exploratory Data Analysis (EDA)**

**Q5:** What types of missing data issues do you expect? How will you deal with them?

**Q6:** Which variables could be good fraud indicators? Think: amount, time, location, email, device, etc.

**Q7:** How would you check for data leakage or suspiciously powerful variables?

**🧭 Step 4: Feature Engineering**

**Q8:** How would you engineer time-based features from TransactionDT?

**Q9:** Can you group users by cards or addresses to build aggregate behavior features (like average amount, frequency, etc.)?

**Q10:** How might you infer whether a transaction is Card Present (CP) vs. Card Not Present (CNP)?

**🧭 Step 5: Modeling Strategy**

**Q11:** What models would you use for this task? Why?

**Q12:** How would you deal with class imbalance during model training?

**Q13:** What metrics would you choose for model evaluation — and why are Accuracy and AUC not always sufficient?

**🧭 Step 6: Post-Model Analysis**

**Q14:** How will you explain your model’s predictions to a non-technical stakeholder?

**Q15:** What would you do if your model starts degrading in real-time (i.e., concept drift)?

**🧭 Step 7: Deployment & Real-Time Use Case**

**Q16:** How would you simulate real-time fraud detection using this batch dataset?

**Q17:** How would you make your model fast and efficient enough for real-time predictions?