PLANNING LOGIC :

DATA COLLECTION :

Data Collection is the process of gathering required data from reliable sources to train and test a machine learning model. In this project, data is collected in the form of online payment transaction records, which include details like transaction type, amount, sender and receiver balances, and fraud labels.

This collected dataset is used for preprocessing, analysis, model training, and fraud prediction.

```
df = pd.read_csv('/content/final data.csv')

df.head()
```

	step	type	amount	nameOrig	oldbalanceOrg	newbalanceOrig	nameDest	oldbalanceDest	newbalanceDest	isFraud	isFlaggedFrau
0	1	PAYMENT	9839.64	C1231006815	170136.0	160296.36	M1979787155	0.0	0.0	0.0	0
1	1	PAYMENT	1864.28	C1666544295	21249.0	19384.72	M2044282225	0.0	0.0	0.0	0
2	1	TRANSFER	181.00	C1305486145	181.0	0.00	C553264065	0.0	0.0	1.0	0
3	1	CASH_OUT	181.00	C840083671	181.0	0.00	C38997010	21182.0	0.0	1.0	0
4	1	PAYMENT	11668.14	C2048537720	41554.0	29885.86	M1230701703	0.0	0.0	0.0	0

DATA PREPROCESSING :

Data preprocessing is an important step in machine learning where raw transaction data is cleaned and transformed into a suitable format for model training. Since real-world online payment data may contain missing values, categorical values, noise, or imbalance, preprocessing ensures the dataset becomes accurate and consistent.

In this fraud detection project, preprocessing includes:

- **Steps Involved**

1. **Handling Missing Values**
    - **Checking for null/NaN values and filling or removing them.**
2. **Removing Duplicate Records**
    - **Duplicate transactions are removed to avoid bias.**
3. **Encoding Categorical Data**
    - **Converting non-numeric columns like transaction type into numerical form using Label Encoding or One-Hot Encoding.**
4. **Feature Selection**
    - **Removing unnecessary columns and keeping only useful features for fraud prediction.**

5. **Scaling / Normalization (Optional)**
    - **Scaling features like transaction amount to improve model performance.**
6. **Handling Imbalanced Data**
    - **Fraud cases are usually very less compared to normal transactions, so techniques like SMOTE or class balancing may be used.**
7. **Splitting Dataset**
    - **Dividing data into training set and testing set for model evaluation.**