**Capstone Project Proposal: Fraud Detection System for JPMorgan Chase**

**1. Project Title:**  
Fraud Detection System for JPMorgan Chase

**2. Problem Statement:**  
JPMorgan Chase, a global financial leader, processes millions of transactions each day. With the rise of sophisticated cyber threats, traditional fraud detection methods struggle to identify new and evolving fraud patterns in real-time. This project will develop a real-time fraud detection system leveraging machine learning models to enhance security and reduce financial losses.

**3. Objectives:**

* Ingest and preprocess transaction data from varied sources.
* Develop and evaluate supervised and unsupervised fraud detection models.
* Create a real-time anomaly detection and alerting pipeline.
* Build interactive dashboards for fraud monitoring.
* Optimize models to maximize recall while minimizing false positives.

**4. Scope:**

* Data ingestion from CSV, SQL, API, and NoSQL.
* ML techniques: Logistic Regression, Random Forest, Isolation Forest, Autoencoders.
* Real-time scoring simulation with streaming tools.
* Dashboard creation using Power BI or Tableau.
* Deployment on local or cloud-based infrastructure.

**5. Methodology:**

1. **Data Ingestion & Cleaning** – Handle missing data, normalize formats, and remove inconsistencies.
2. **Exploratory Data Analysis** – Identify fraud-prone transaction types, locations, devices.
3. **Feature Engineering** – Transaction velocity, IP-location mismatches, device change flags.
4. **Model Development** – Train supervised models on labeled fraud data; apply unsupervised methods for anomaly detection.
5. **Real-Time Pipeline** – Process incoming transactions, compute fraud risk scores, trigger alerts.
6. **Visualization** – Develop dashboards for transaction trends, fraud alerts, and KPIs.
7. **Evaluation & Optimization** – Assess using precision, recall, F1-score, ROC-AUC; fine-tune hyperparameters.

**6. Deliverables:**

* Processed dataset ready for analysis.
* EDA report with key insights.
* Fraud detection models with documented metrics.
* Real-time monitoring pipeline.
* Interactive dashboard.
* Final project report and presentation.

**7. Expected Outcomes:**

* Robust real-time fraud detection with high detection accuracy.
* Actionable intelligence for security teams.
* Scalable and adaptable architecture for evolving fraud tactics.

**8. Timeline:**

* Weeks 1–2: Data ingestion and preprocessing.
* Weeks 3–4: EDA and feature engineering.
* Weeks 5–6: Model building and evaluation.
* Weeks 7–8: Real-time pipeline and dashboard.
* Week 9: Model optimization.
* Week 10: Report writing and presentation.

####

If you’ve been given just one dataset with the project files, then yes — all your ingestion, preprocessing, EDA, feature engineering, and model building will be based on that dataset.

To align with the proposal’s “multiple sources” requirement:

1. **Treat the given dataset as your master source**
   * This is your primary historical transactions dataset.
   * It likely contains all the columns from the project’s data dictionary.
2. **Simulate multiple sources for completeness**
   * **CSV source** → the provided dataset file.
   * **SQL source** → Load the same data into a local SQLite/MySQL DB and fetch from there.
   * **API source** → Serve a subset of the dataset via a Flask/FastAPI endpoint, then consume it.
   * **NoSQL source** → Store a portion in MongoDB and query it back.
   * This is for demonstration purposes, showing you can handle different pipeline types.
3. **Handle real-time simulation**
   * Stream rows from the dataset in chunks or at time intervals to simulate live transactions.
   * No new data is needed; just replay existing records.
4. **Benefits of sticking to one dataset**
   * Consistent preprocessing and model building.
   * Reduced scope creep.
   * More focus on feature quality, model performance, and dashboard design.

Next Step: Inspect the dataset to confirm its structure, column names, and mapping to the project’s data dictionary so we can create a cleaning and modeling plan.

The dataset has **6,362,620 rows** and **11 columns**:

**Columns:**

1. step – Time step (likely in hours or days from a start point)
2. type – Transaction type (PAYMENT, TRANSFER, CASH\_OUT, etc.)
3. amount – Transaction amount
4. nameOrig – Originating customer ID
5. oldbalanceOrg – Sender’s balance before transaction
6. newbalanceOrig – Sender’s balance after transaction
7. nameDest – Destination customer ID
8. oldbalanceDest – Recipient’s balance before transaction
9. newbalanceDest – Recipient’s balance after transaction
10. isFraud – Fraud indicator (1 = Fraud, 0 = Legit)
11. isFlaggedFraud – Flag indicator for unusual large transfers

**Sample:**

step type amount nameOrig oldbalanceOrg newbalanceOrig nameDest oldbalanceDest newbalanceDest isFraud isFlaggedFraud

0 1 PAYMENT 9839.64 C1231006815 170136.0 160296.36 M1979787155 0.0 0.0 0 0

1 1 PAYMENT 1864.28 C1666544295 21249.0 19384.72 M2044282225 0.0 0.0 0 0

...

This matches a known **synthetic fraud detection dataset** and will be your sole working dataset.  
You can now:

* Map these columns to the **data dictionary** in your proposal.
* Use it for **EDA**, **feature engineering**, **model training**, and **real-time simulation**.

If you want, I can **map these exact columns to your original project data dictionary** so your methodology stays consistent. This will also help when writing your report. Would you like me to do that?