**Proposal for Dataset Design & Generation Strategy**

**Generate a** multi-modal telecom metadata **which can be reproduced and used for AI based validation strategies adhering to following instructions:**

1. 03 different multi-modal telecom datasets, one each encompassing CDR, IPDR, and EDR metadata streams.
2. The Modalities of the datasets are:
   1. CDR (Call Detail Records) – voice/SMS metadata
   2. IPDR (Internet Protocol Detail Records) – app/web activity
   3. EDR (Event Detail Records) – device, mobility, network context
3. **The multi-modal telecom dataset** haa realistic diversity and embedded fraud typologies for advanced modeling, including:
4. The dataset is **Syntheticaly Generated and each dataset should have atleast 5000 records.**
5. The dataset simulates environments that model diverse user behaviors (benign and fraudulent) and network conditions, incorporating realistic noise levels and data distributions.
6. Ateast 10% dataset should include Fraud scenarios such as: including Vishing coordination, SIM swap sequences, DDoS traffic patterns, and covert communication indicative of espionage based on established typologies and emerging threat intelligence.
7. Parameters controlling fraud intensity, coordination complexity, and temporal dynamics be varied to assess model sensitivity.
8. Dataset has:
   1. A baseline of **5,000 records per stream** with ~10% embedded fraud
   2. Control over key parameters (fraud intensity, coordination complexity, temporal density)

**Schema Design for Each Modal Stream is given below**

**A. CDR Fields**

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| caller\_id | string | Unique MSISDN |
| callee\_id | string | Destination MSISDN |
| call\_start | timestamp | Start of call |
| call\_duration | int (sec) | Duration in seconds |
| call\_type | enum | [voice\_in, voice\_out, sms\_in, sms\_out] |
| cell\_id | string | Source tower |
| location | string | Latitude, longitude |
| imei | string | Device identifier |
| is\_fraud | bool | Label fraud/non-fraud |

**B. IPDR Fields**

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| user\_id | string | MSISDN/IP Map |
| timestamp | datetime | Session time |
| domain | string | Accessed website/app |
| ip\_dst | string | Destination IP |
| port | int | TCP/UDP port |
| protocol | enum | TCP/UDP |
| duration | int | Seconds |
| bytes\_sent | int |  |
| bytes\_received | int |  |
| vpn\_usage | bool | VPN or not |
| is\_fraud | bool | Fraud label |

**C. EDR Fields**

|  |  |  |
| --- | --- | --- |
| **Field** | **Type** | **Description** |
| event\_id | string | Unique |
| user\_id | string | MSISDN |
| device\_model | string | Device type |
| os\_type | string | OS info |
| roaming\_status | bool | In roaming |
| network\_type | string | LTE, 5G, WiFi |
| event\_time | datetime | Event time |
| location | string | Lat/Lon |
| event\_type | string | reboot, SIM switch, etc. |
| is\_fraud | bool | Label |

|  |  |
| --- | --- |
| **Component** | **Description** |
| **3 Modal Datasets** | CDR, IPDR, and EDR datasets, each with ≥ 5,000 records |
| **10% Fraud Injection** | Vishing, SIM Swap, DDoS, Covert Communications |
| **Control Parameters** | fraud\_intensity, coordination\_complexity, temporal\_dynamics |
| **Preprocessing Modules** | Cleaning, Normalization, Feature Engineering, Graph Mapping |
| **Documentation** | Codebase, Schema, Fraud Generation Rules, Instructions |
| **Repository** | GitHub-compatible folder with Jupyter notebooks & CSV output |
| **Model-ready Output** | Ready for graph-based ML, anomaly detection, and time-series modeling |

**Fraud Injection Scenarios are stated below:**

**Vishing Coordination**

* Burst of outbound calls from single device to multiple recipients
* Short-duration, repeat calls
* Associated with unusual geolocation changes

**SIM Swap Patterns**

* Sudden IMEI change
* Drop in usual activity followed by high-value SMS/OTP activity
* New device with different OS

**DDoS-like Traffic**

* From IoT device IPs
* Short sessions, high connection counts
* Repeated destination port usage

**Covert Communication**

* Encrypted DNS/Tor addresses
* Short TCP sessions with non-standard ports
* Unusual time-of-day activity patterns

**Parameter Controls**

Include sliders or categorical bins for:

* fraud\_intensity: low, medium, high (affects % of fraud + pattern visibility)
* coordination\_complexity: independent, semi-coordinated, botnet-like
* temporal\_dynamics: single event, recurring weekly, time-of-day aligned

### **Dataset Modal Breakdown**

|  |  |  |  |
| --- | --- | --- | --- |
| **Modal** | **Records** | **Fraud Types** | **Control Parameters** |
| **CDR** | 5,000+ | Vishing, SIM Swap | Fraud Intensity, Call Time Window |
| **IPDR** | 5,000+ | DDoS, Covert Comms | Protocol, Ports, Burst Frequency |
| **EDR** | 5,000+ | SIM Swap, Device Tampering | OS Change, Reboot Events, Roaming |

### **Fraud Controls (Parameter-Sensitive)**

|  |  |
| --- | --- |
| **Parameter** | **Description** |
| fraud\_intensity | % of records and detectability (Low: 5%, High: 25%) |
| coordination\_complexity | Independent → Coordinated → Botnet-style |
| temporal\_dynamics | One-time → Time-of-day → Weekly patterns |

**Tools for further Dataset Generation**

* Python + Faker for metadata realism (names, locations)
* NumPy / SciPy for statistical distributions (e.g., Pareto for call durations)
* Pandas for schema joining
* scikit-learn or custom logic for labeling fraud based on rule patterns
* Optionally, SDV (Synthetic Data Vault) for probabilistic generative modeling