# Technical Documentation: Feature Engineering Pipeline (Notebook 2)

Project: Aadhaar Societal Trends Analysis

Notebook: Feature_Engineering.ipynb

Objective: To generate high-value strategic metrics and Key Performance Indicators (KPIs) from the merged dataset, enabling predictive modeling and deep diagnostic analysis.

## 1. Data Loading & Validation

The cleaned master dataset created in the previous step was ingested and validated before processing.

- **Input File:** Aadhar_Data.csv (Output from Notebook 1).
- **Initial Shape:** 2,234,840 rows × 17 columns.
- **Integrity Check:**
  - Validated Missing Values = 0.
  - Validated Duplicates = 0.
  - Confirmed date column conversion to datetime objects for time-series compatibility.

## 2. Advanced Feature Engineering (The "Gold Standard" Metrics)

This phase involved creating calculated fields to measure system performance, saturation, and risk. These metrics are critical for the "Impact" score in the hackathon.

### A. Workload Aggregation

We created base metrics to quantify total operational volume.

- **total_updates**: Sum of biometric and demographic updates.
  - *Formula:* total_biometric + total_demography
- **total_workload**: The grand total of all Aadhaar transactions in a region.
  - *Formula:* total_enrollment + total_updates

### B. Strategic KPIs (Key Performance Indicators)

These are the custom metrics designed to answer specific problem statements:

1. **Maturity Score (maturity_pct)**
   - *Definition:* Measures if a region is in "Growth Phase" (high enrollment) or "Maintenance Phase" (high updates).
   - *Formula:* $(\frac{\text{Total Updates}}{\text{Total Enrollment} + 1}) \times 100$
   - *Insight:* High scores ($>100\%$) indicate a saturated market where the primary workload is maintenance, not new IDs.
2. **Migration Index (migration_pct)**

    - *Definition:* Uses demographic updates (address/phone changes) as a proxy for tracking population movement.
    - *Formula:* $(\frac{\text{Adult Demographic Updates}}{\text{Total Enrollment} + 1}) \times 100$
    - *Insight:* High scores signal "Migrant Hotspots" where people are moving in/out frequently.

3. **Biometric Backlog Risk (biometric_backlog_vol)**
    - *Definition:* Identifies the gap between infant enrollment and mandatory biometric updates for school-age children.
    - *Formula:* $\text{Infant Enrollment (0-5)} - \text{Child Biometric Updates (5-17)}$
    - *Insight:* A positive number indicates a **Backlog**—children who have IDs but haven't updated their biometrics, risking service denial.
4. **Maintenance Intensity (maintenance_intensity_pct)**
    - *Definition:* The percentage of total workload dedicated to fixing/updating existing IDs rather than creating new ones.
    - *Formula:* $(\frac{\text{Total Updates}}{\text{Total Workload}}) \times 100$
5. **National Workload Share (workload_share_pct)**
    - *Definition:* Measures how much pressure a specific district puts on the national infrastructure.
    - *Formula:* $(\frac{\text{District Workload}}{\text{Total National Workload}}) \times 100$

**3. Data Cleaning (Post-Engineering)**

After calculating ratios, some edge cases (like division by zero) were handled.

- **Handling Nulls/Infinity:** The code uses + 1 in denominators (Smoothing) to prevent DivisionByZero errors.
- **Output Verification:** The final dataset was checked again for nulls created during calculation.
    - *Result:* 0 Missing values across all new 25 columns.

**4. Final Schema & Output**

The dataset was expanded from 17 raw columns to **25 analytical features**.

- **New Schema:**
    - maturity_pct (Float)
    - migration_pct (Float)
    - adult_update_pct (Float)
    - maintenance_intensity_pct (Float)

  - workload_share_pct (Float)
  - biometric_backlog_vol (Float)
- **Output File:** Aadhar_Updated_Data.csv
- **Description:** A feature-rich dataset ready for Machine Learning (Clustering) and Statistical Anomaly Detection.