SafeData 2.0 - Implementation & Tech Blueprint

Implementation Steps (High-level)

1. Data Inventory & Sensitivity Mapping: catalog datasets, detect quasi-identifiers and rare slices.
2. Prototype SDG: train tabular synthesizer (CTGAN/TVAE) on a pilot dataset; validate utility.
3. Integrate DP: implement DP-SGD on the synthetic trainer; select initial epsilon (e.g., 1-5).
4. Targeted SDC: implement detection for rare QI combos and apply minimal perturbation/generalization.
5. Risk Simulation Module: build linkage and inference attack simulators to compute re-identification risk.
6. Optimization Loop: implement Bayesian optimizer to balance privacy parameters and utility objectives.
7. API & Orchestration: wrap pipeline as microservices; schedule with Airflow/Prefect.
8. Audit & Reporting: automated Privacy-Utility Report generation and logging.
9. Pilot & Evaluate: run multiple rounds with stakeholder tasks; tune parameters.
10. Productionize & Deploy: containerize, secure, and deploy (on-prem or hybrid cloud).

Ideal Tech Stack

- Programming: Python
- Data: Pandas, Dask, PySpark
- Synthetic Data: SDV (CTGAN, TVAE), CTABGAN, Copulas
- Differential Privacy: PyDP, TensorFlow Privacy, Opacus (PyTorch)
- Orchestration: Apache Airflow or Prefect
- Serving/API: FastAPI + NGINX
- Containers: Docker + Kubernetes
- Storage: PostgreSQL / MinIO / S3 (encrypted)
- Monitoring: Prometheus + Grafana
- CI/CD: GitHub Actions or GitLab CI
- Security: Vault (secrets), RBAC, TLS, Disk encryption (AES-256)

Prototype Implementation Table

Component	Tool/Library	Purpose
Data Ingest	Python + Pandas	Read & validate uploaded datasets
SDG Trainer	SDV / CTGAN	Generate synthetic datasets
DP Mechanism	TensorFlow Privacy / Opacus	Add DP to training
SDC Toolkit	Custom/Pandas	Detect & perturb rare combos
Attack Simulator	Custom scripts	Re-identification tests
Optimizer	Scikit-Optimize / Ax	Tune parameters
Orchestration	Airflow/Prefect	Batch & workflow scheduling
API Service	FastAPI	Expose anonymization endpoints

All Tools & Techniques (detailed)

- Quasi-Identifier Detection: regex rules + frequency analysis
- SDG Methods: CTGAN, TVAE, Copula-based, Bayesian networks
- DP Techniques: DP-SGD, output perturbation for queries
- SDC Methods: generalization, suppression, microaggregation, data swapping

- Attack Types Simulated: linkage, membership inference, attribute inference
- Utility Metrics: AUC, MAE, KS-test, JS divergence, histogram overlap
- Logging & Audit: immutable logs, versioned configs, privacy ledger

Where to take datasets (MoSPI portals)

- Official MoSPI portal: https://mospi.gov.in/ (Download Tables & Data)
- e-Sankhyiki: https://esankhyiki.mospi.gov.in/ (catalogue & microdata access; may require registration)
- Data Innovation Lab: https://datainnovation.mospi.gov.in/ (data experiments & challenge datasets)

Note: Microdata sometimes requires registration and acceptance of terms; follow the 'Guide to download microdata' on MoSPI.

Deployment Checklist (Quick)

- Confirm dataset access permissions & approvals
- Provision secure staging environment (on-prem or private cloud)
- Containerize components (Docker), prepare k8s manifests
- Implement monitoring & logs
- Run privacy & utility evaluation suite on pilot datasets
- Document configurations and generate Privacy-Utility Report templates