GENERATING REALISTIC SYNTHETIC DATA FOR ML

CHALLENCE

Data Scarcity
Limited availability of real-world data makes it challenging to train machine learning models effectively.

Privacy Concerns
Using real data risks exposing sensitive information, especially in regulated industries like healthcare and finance.

Statistical Accuracy
Generating synthetic data that maintains the complex relationships and statistical properties of real data is difficult.

APPROACH

Data Scarcity
Generate synthetic data using TVAEs and SAGANs to fill gaps in real data.

Privacy Concerns
Use differential privacy to protect sensitive information in the generated data.

Statistical Accuracy
Validate synthetic data with real-time checks to ensure it matches real data patterns.

VALUE PROPOSITION

PROS

- Solves data scarcity by generating realistic synthetic data.
- Protects privacy with differential privacy techniques.
- Ensures statistical accuracy for reliable ML model training.

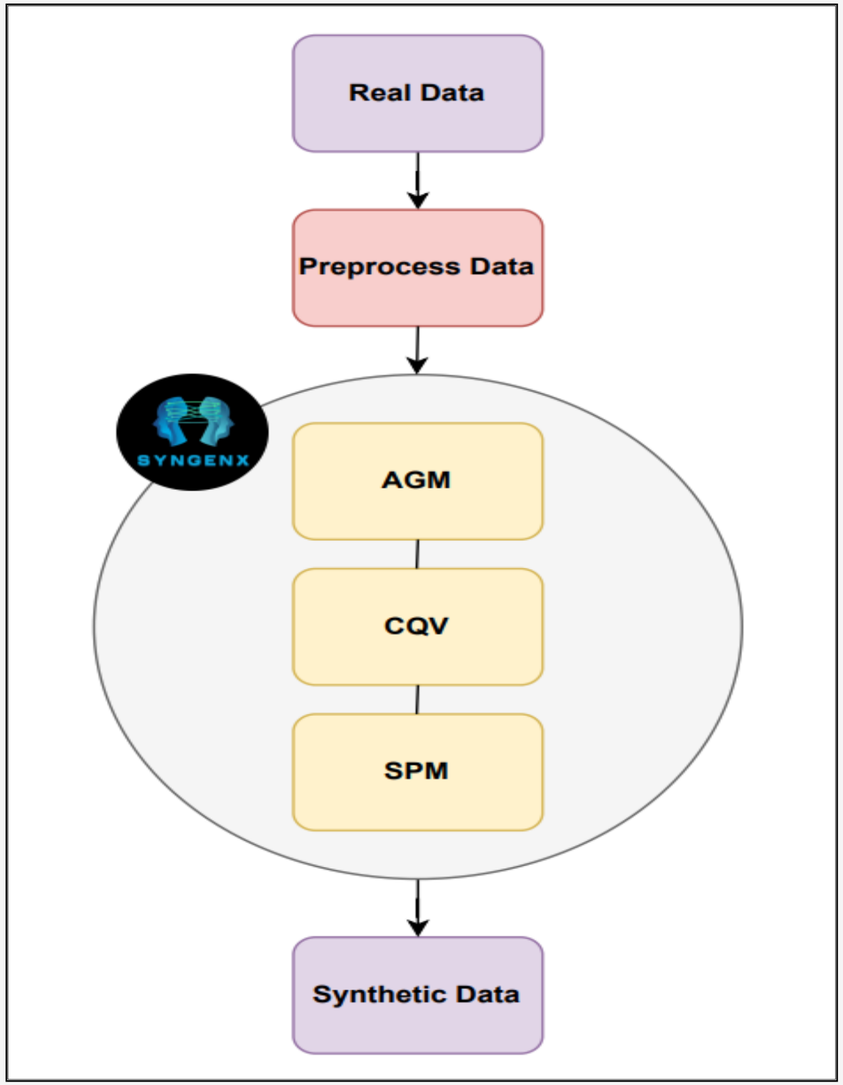
CONS

- Requires advanced AI expertise for implementation.
- High computational cost for training models.
- Quality depends on the real data used for training.

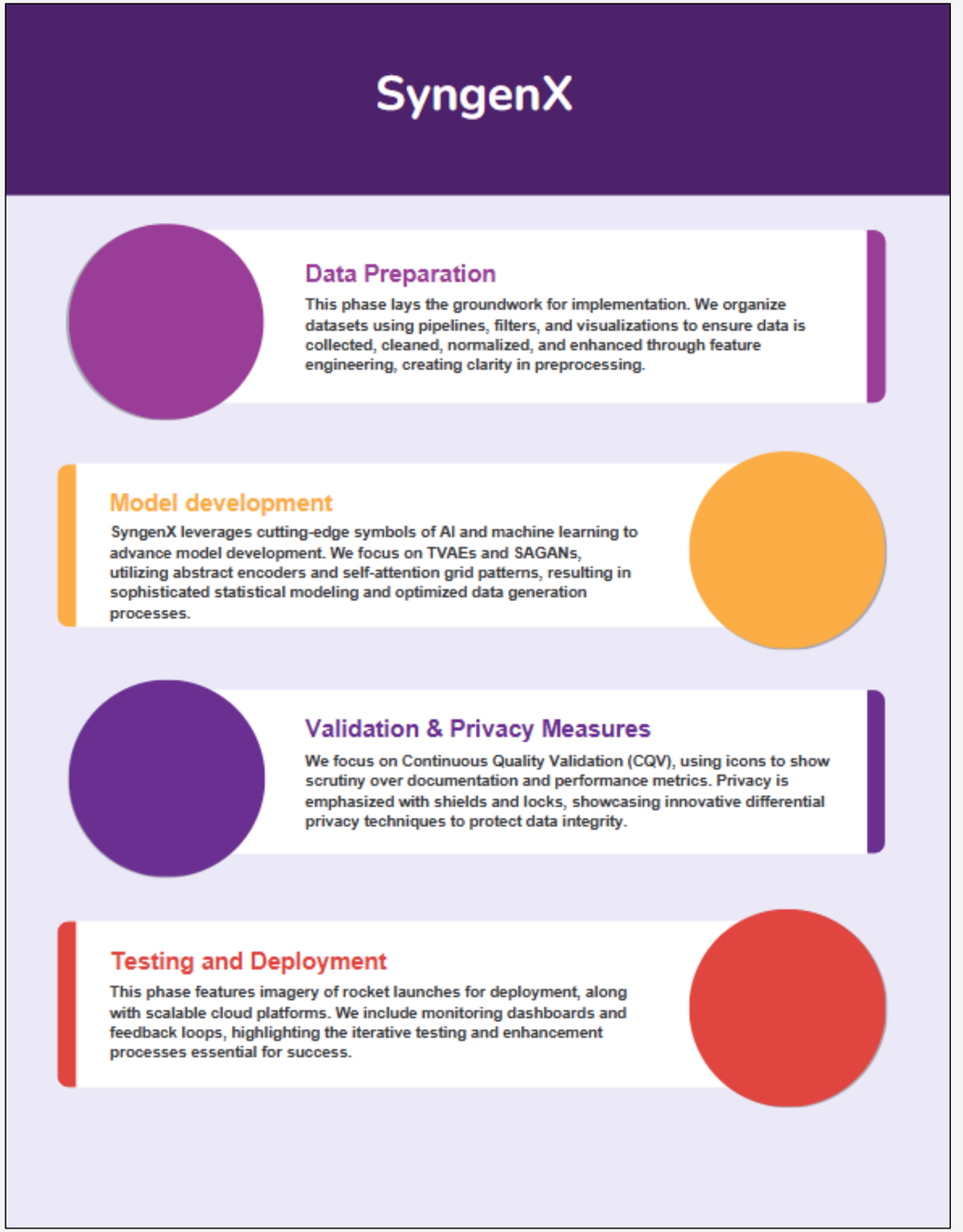
TECHNICAL STACK

Technical Stack	
Programming	Python
Frameworks	TensorFlow, PyTorch
Synthetic Data Tools	SDV, scikit-learn
Privacy Tools	TensorFlow Privacy, PySyft
Infrastructure	AWS/Google Cloud, Docker, Kubernetes
Visualization	Matplotlib, Seaborn
Validation	SciPy, MLFlow

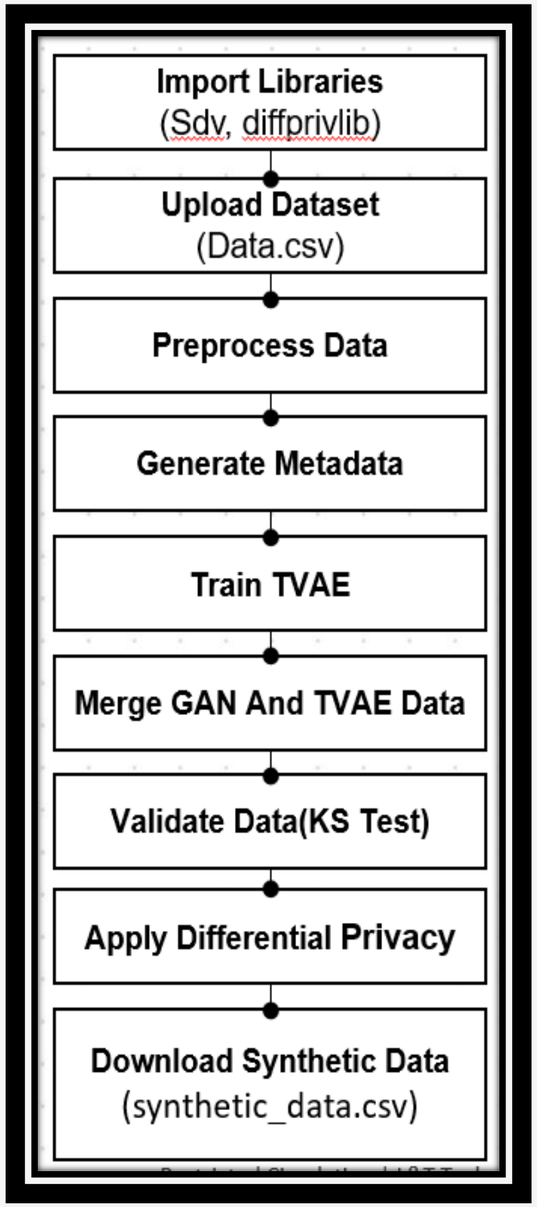
WORKFLOW



IMPLEMENTATION PLAN



EXECUTION FLOW



Execution Drive Link

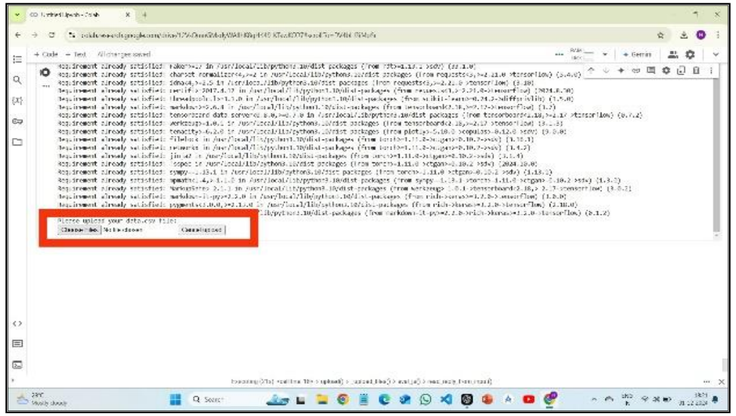
https://drive.google.com/file/d/1v3GuY_DR-HywatYUxn0sRx0_qSLCHy1b/view?usp=sharing

CSV Drive Link

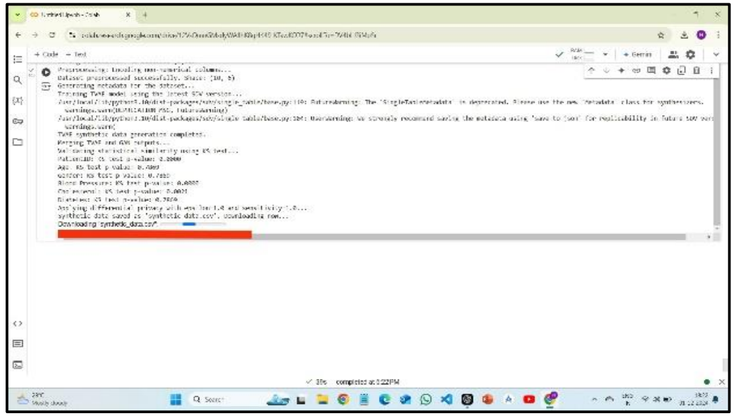
https://drive.google.com/file/d/1dakfoFQmEFj6ULkRWVOakt_XpCF5CFWf/view?usp=drivesdk

VALIDATION

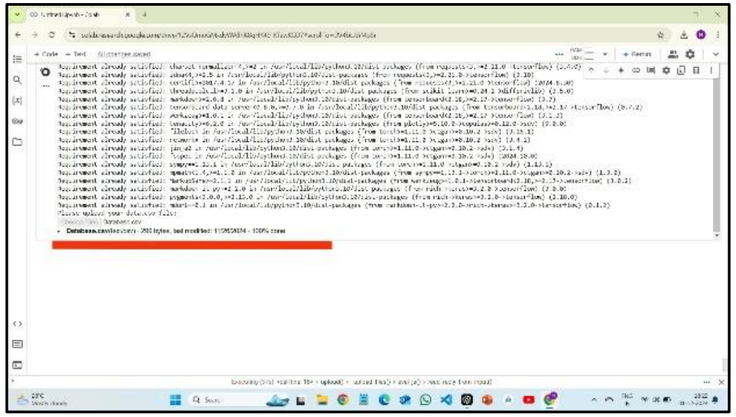
(1) Data Upload Process



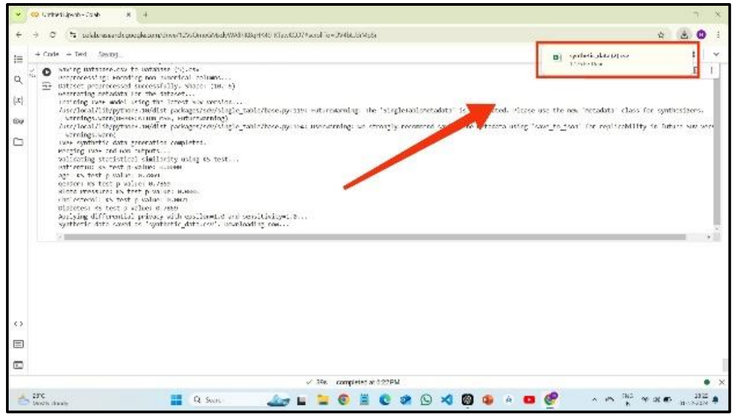
(3) Synthetic Data Generation Process



(2) Real Data Uploaded



(4) Synthetic Data - Download Confirmation



COST ESTIMATE

ESTIMATION	
Development Costs	
Data Scientists & Engineers	₹4,150–₹12,450/hour
Team of 4 (3–6 months)	₹83,00,000–₹1,66,00,000
Infrastructure Costs	
Cloud Computing (GPU)	₹4,15,000–₹12,45,000/month
6 months	₹25,00,000–₹75,00,000
Data Management Systems	₹4,15,000–₹8,30,000
Licensing and Software	
Proprietary Tools	₹8,30,000–₹16,60,000
Validation and Testing	
Quality Assurance	₹16,60,000–₹33,20,000
Operational Costs	
Maintenance and Scaling	₹8,30,000–₹16,60,000 annually
Miscellaneous Costs	
Legal Compliance	₹8,30,000–₹12,45,000
Training	₹8,30,000–₹12,45,000
Initial Development Phase	**₹1,50,00,000–₹3,12,00,000.**
Annual Ongoing Costs	**₹16,60,000–₹41,50,000**

ASSUMPTIONS

Availability of Real Data
Sufficient real-world data is available for training the models.
Computational Resources: Adequate hardware or cloud infrastructure (e.g., GPUs) is available for model training.

Privacy Compliance
Differential privacy techniques will effectively safeguard sensitive data and comply with regulations like GDPR and HIPAA.

User Expertise
The team implementing and managing the system has the required expertise in AI, machine learning, and privacy techniques.

CONCLUSION

SynGenX provides a robust solution for generating high-quality, privacy-compliant synthetic data. It addresses data scarcity, ensures statistical accuracy, and protects privacy using advanced AI models like TVAEs and SAGANs. Despite the need for expertise and resources, it offers significant value for scalable and reliable AI-driven innovation.

REFERENCES

Kingma & Welling (2014) on Variational Autoencoders, Zhang et al. (2019) on Self-Attention GANs, and Patki et al. (2016) on Synthetic Data Vault.Abadi et al. (2016) on Deep Learning with Differential Privacy, TensorFlow Privacy GitHub.

THANK YOU!