**Project Proposal**

**Nền tảng Tạo Dữ liệu Tổng hợp với Bảo toàn Quyền riêng tư (Synthetic Data Generation Platform with Privacy Preservation)**

Người Thực Hiện: Nguyễn Thị Minh Thư

Trường Đại học Công Nghệ TP.HCM

MSSV: 2180603043

Email: nguyenthiminhthu92003@gmail.com

GVHD: Trương Thị Minh Châu

**1. Executive Summary**

Trong môi trường doanh nghiệp hiện nay, đặc biệt là các lĩnh vực như ngân hàng và tài chính, vấn đề bảo mật dữ liệu và tuân thủ quyền riêng tư cá nhân (PII) đang trở thành rào cản lớn đối với việc ứng dụng các mô hình AI/ML. Từ trải nghiệm thực tế khi thực tập tại một ngân hàng thương mại, em nhận ra rằng việc không thể sử dụng dữ liệu thật khiến nhiều dự án học máy phải dừng bước từ rất sớm.

Để giải quyết vấn đề này, em đề xuất xây dựng một nền tảng tạo dữ liệu tổng hợp (synthetic data platform) trên AWS, giúp tạo ra các tập dữ liệu mô phỏng lại phân phối thống kê của dữ liệu thật, không chứa thông tin định danh và có thể chia sẻ một cách an toàn cho các bên liên quan (Data Scientist, Developer, Researcher).

Key Features:

* Tự động hóa pipeline xử lý và kiểm duyệt dữ liệu gốc.
* Sinh dữ liệu bằng mô hình deep generative model (TVAE/CTGAN).
* Đánh giá thống kê và kiểm thử dữ liệu đầu ra.
* Tối ưu chi phí sử dụng AWS Free Tier.

Business Benefits & ROI:

* Giảm rủi ro pháp lý khi chia sẻ dữ liệu nội bộ hoặc đào tạo AI.
* Tiết kiệm thời gian xử lý dữ liệu (không cần làm thủ công hoặc xin phép truy cập).
* Mở rộng khả năng hợp tác nghiên cứu và thực hành AI cho cộng đồng.
* Tăng năng suất R&D trong các nhóm kỹ thuật, đảm bảo dữ liệu sạch và hợp lệ.

Investment & Timeline:

* Tổng chi phí triển khai: ~$0.45
* Thời gian thực hiện: 8 tuần, theo từng sprint (1 tuần / block)

Success Metrics:

* Dữ liệu sinh ra đạt ≥ 90% độ tương đồng về phân phối.
* 100% không có thông tin định danh theo regex và kiểm tra mẫu thủ công.
* Có thể tái sử dụng pipeline cho ≥ 2 dataset khác nhau mà không cần viết lại toàn bộ mã nguồn.

**2. Problem Statement**

Phân tích hiện trạng:

Trong quá trình thực tập tại ngân hàng, em nhận thấy rõ nhu cầu rất lớn về việc áp dụng học máy để phân tích và ra quyết định. Tuy nhiên, dữ liệu thực thường chứa thông tin nhạy cảm, không thể chia sẻ cho team kỹ thuật hoặc cộng tác bên ngoài. Việc ẩn danh thủ công tốn thời gian và dễ sai sót.

Các vấn đề cụ thể:

* Không thể truy cập dữ liệu thật để huấn luyện mô hình.
* Không có công cụ xác minh dữ liệu đã thật sự "sạch" hay chưa.
* Thiếu workflow đơn giản, dễ tái sử dụng cho người học và nhà phát triển.
* Stakeholders bị ảnh hưởng:
* Data Scientist: Không có dữ liệu để phát triển mô hình.
* CTO / DPO: Quan ngại về rủi ro pháp lý khi lộ dữ liệu cá nhân.
* Mentor / Người học: Khó triển khai bài tập thực hành AI sát thực tế.

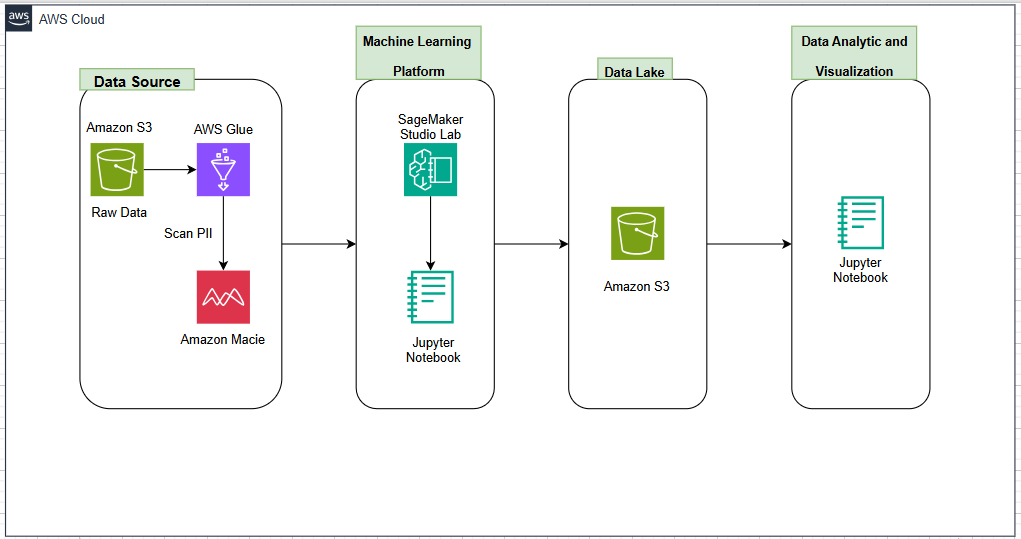
**Nếu không giải quyết?**

* Việc huấn luyện AI sẽ tiếp tục bị bóp nghẹt vì thiếu dữ liệu.
* Các nhóm nghiên cứu không thể chia sẻ dữ liệu cho nhau.
* Rủi ro rò rỉ dữ liệu thật tăng cao nếu ẩn danh sai.

**3. Solution Architecture**

Mô tả tổng thể:

Dự án đề xuất một hệ thống sinh dữ liệu tổng hợp chạy hoàn toàn trên AWS với luồng xử lý như sau:

Thành phần chính và chức năng:

| **Dịch vụ** | **Vai trò** |
| --- | --- |
| Amazon S3 | Lưu trữ dữ liệu gốc & dữ liệu sinh |
| AWS Glue | Làm sạch dữ liệu, chuẩn hóa schema |
| Amazon Macie | Tự động quét dữ liệu chứa thông tin định danh |
| SageMaker Studio Lab | Huấn luyện mô hình sinh dữ liệu (TVAE) |
| Jupyter Notebook | Kiểm thử, phân tích và trực quan hóa |

Bảo mật và compliance:

* Bật mã hóa toàn bộ bucket S3.
* Phân quyền IAM chỉ cho phép đúng dịch vụ truy cập.
* Không chia sẻ dữ liệu thật ra khỏi môi trường an toàn.
* Không sử dụng credential trực tiếp trong Studio Lab.
* Khả năng mở rộng:
* Có thể thêm pipeline kiểm bias bằng AWS Clarify.

**4. Technical Implementation**

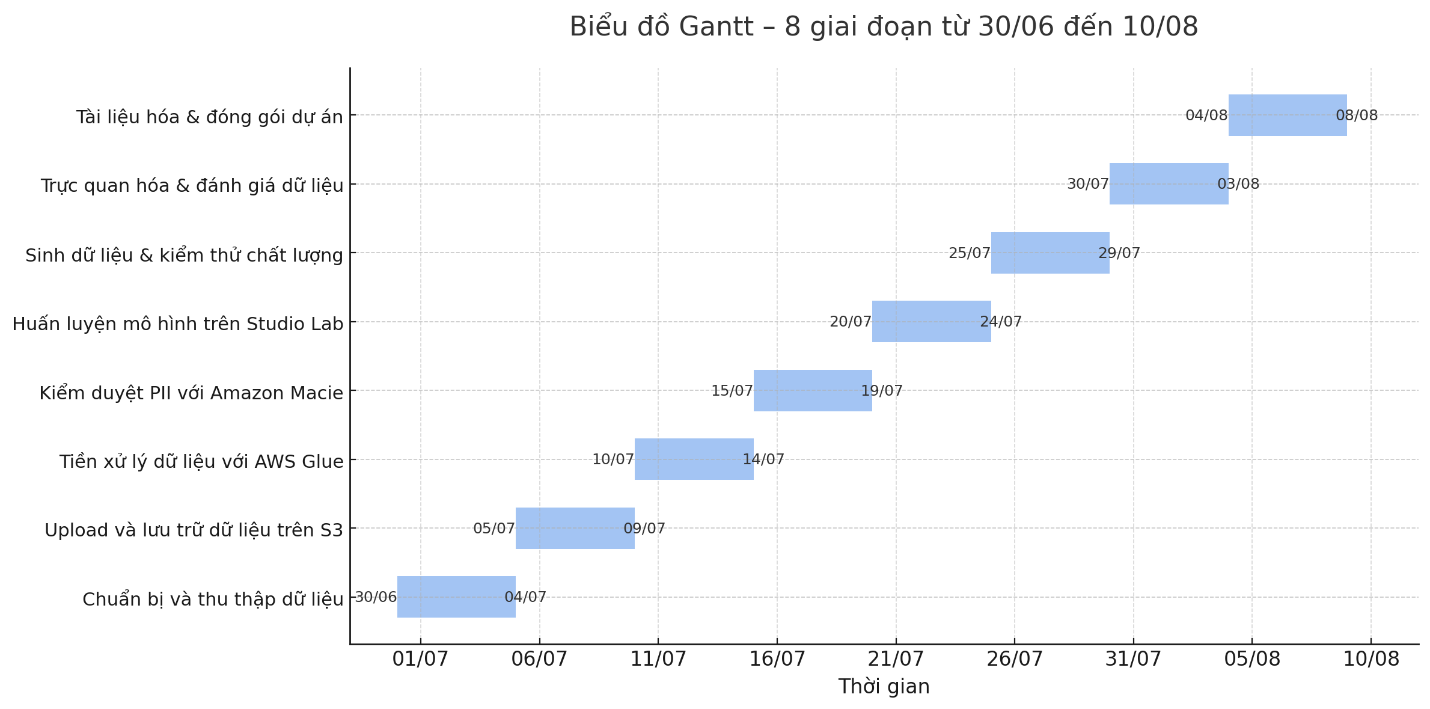
**Giai đoạn triển khai:**

1. Chuẩn bị dữ liệu
2. Thu thập dữ liệu tabular mẫu mô phỏng môi trường tài chính.
3. Upload lên S3 bằng giao diện web.
4. Tiền xử lý với AWS Glue
5. Tạo Glue Job (PySpark): chuẩn hóa format ngày, xử lý missing value, rename column.
6. Output dạng Parquet hoặc CSV.
7. Kiểm duyệt thông tin nhạy cảm với Macie
8. Cấu hình Macie quét mẫu (email, số tài khoản, địa chỉ...)
9. Kiểm tra báo cáo và export log đánh giá.
10. Huấn luyện mô hình trong Studio Lab
11. Tải dữ liệu về local và upload lên Studio Lab.
12. Cài SDV, chọn mô hình TVAE, thiết lập epochs, batch\_size phù hợp.
13. Theo dõi loss và trực quan hóa tiến trình học bằng matplotlib.
14. Sinh dữ liệu và kiểm thử
15. Tạo bảng dữ liệu mới và so sánh với dữ liệu gốc về histogram, correlation.
16. Viết script tự động kiểm tra sự hiện diện của PII bằng regex.
17. Trực quan hóa phân phối, kiểm tra outlier.
18. Tài liệu hóa và đóng gói
19. Viết Notebook tổng hợp luồng làm việc.
20. Viết Workshop hướng dẫn cài đặt và chạy lại từ đầu.

**5. Timeline & Milestones**

|  |  |  |
| --- | --- | --- |
| **Tuần** | **Mục tiêu** | **Kết quả mong đợi** |
| 1 | Phân tích yêu cầu, đọc tài liệu SDV,  AWS | Ghi chú & bản phác thảo kiến trúc |
| 2 | Thiết kế pipeline, chọn dữ liệu mẫu | Upload dữ liệu lên S3 |
| 3 | Xử lý dữ liệu bằng Glue | Có output sạch, schema chuẩn |
| 4 | Quét PII với Macie | Báo cáo kiểm tra & log |
| 5 | Huấn luyện mô hình đầu tiên | Sinh dữ liệu thử nghiệm |
| 6 | Đánh giá, trực quan hóa | So sánh phân phối, correlation |
| 7 | Tài liệu hóa lại pipeline | Notebook + hướng dẫn chạy lại |
| 8 | Chuẩn bị báo cáo tổng kết | Gửi demo, PDF, sơ đồ |

Biểu đồ Gantt:



**6. Budget Estimation**

Ước tính chi phí:

|  |  |
| --- | --- |
| Thành phần | Chi phí ước tính |
| S3 | $0.10 |
| Glue | $0.15 |
| Macie | $0.20 |
| Studio Lab | Miễn phí |
| **Tổng cộng** | ~$0.45 |

ROI & Lợi ích kinh doanh:

* Cắt giảm 100% chi phí mua dữ liệu thử nghiệm.
* Tiết kiệm thời gian xử lý: từ 1 tuần xuống còn 1 ngày.
* Có thể mở rộng thành công cụ chia sẻ nội bộ hoặc teaching tool.
* Chiến lược tối ưu hóa:
* Dùng AWS Free Tier đúng hạn mức.
* Xoá dữ liệu sau khi xử lý để tránh phát sinh thêm.
* Không dùng dashboard đắt tiền (chỉ dùng Notebook).

**7. Risk Assessment**

|  |  |  |  |
| --- | --- | --- | --- |
| **Rủi ro** | **Mức độ ảnh hưởng** | **Khả năng xảy ra** | **Hành động giảm thiểu** |
| Thiếu tài nguyên Studio Lab | Trung bình | Trung bình | Giảm batch size, epochs |
| Credential bị lộ nếu dùng boto3 | Cao | Thấp | Không hardcode key, dùng local upload |
| Bias trong dữ liệu sinh | Trung bình | Trung bình | So sánh thống kê, thêm Clarify nếu cần |

**8. Expected Outcomes**

Chỉ số thành công:

* Dữ liệu sinh ra có histogram gần giống dữ liệu thật (≥ 90%).
* Không có thông tin định danh trong 100% bản ghi.
* Có thể chia sẻ pipeline lại cho người khác làm theo.

Lợi ích ngắn và trung hạn:

* Giúp học viên AI thực hành thực tế mà không vi phạm bảo mật.
* Giúp tổ chức kiểm thử pipeline huấn luyện không cần dữ liệu thật.

Giá trị dài hạn:

* Là nền tảng mở cho workshop về AI, data privacy.
* Làm nền tảng cho việc tích hợp CI/CD pipeline tạo synthetic data an toàn.

Phụ lục

Cấu hình TVAE:

Epochs: 300

Batch size: 500

Optimizer: Adam

Sampling: Gaussian noise injection

Công cụ:

AWS: S3, Glue, Macie

Python: SDV, matplotlib, seaborn, pandas

Viết tài liệu: Jupyter, Markdown, VSCode

Tham khảo:

[https://sdv.dev](https://sdv.dev/)

AWS Free Tier Guide

AWS Security Best Practices