

AnonimaData


Scalable Service for Privacy-Preserving Dataset Anonymization



Scalable and Reliable Systems

DANILA MELELEO
VALENTINA DE RESPINIS

Agenda

- 
- 01. Problema e soluzione
 - 02. Obiettivo del progetto
 - 03. Panoramica dell'architettura
 - 04. Flusso operativo utente
 - 05. Algoritmi di anonimizzazione
 - 06. Scelte tecniche ed implementative
 - 07. Deployment e scalabilità su GCP
 - 08. Conclusioni e prossimi passi

Problema e soluzione



Il problema

- Condividere i dati è essenziale ma **rischioso**
- Le persone possono essere **ri-identificate** da attributi semplici
- I dati sensibili richiedono **protezioni migliori**



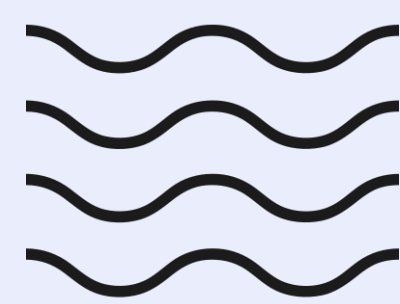
Perchè anonimizzare?

- Rimuovere i nomi **non è sufficiente**
- **87%** dei cittadini USA è identificabile in questo modo
- Le tecniche tradizionali, come la pseudonimizzazione, sono **deboli**



La soluzione

- **AnonimaData**: una piattaforma web **scalabile**
- Algoritmi di anonimizzazione **configurabili**
- Cloud-based, user-friendly



Obiettivo del Progetto



Obiettivo

Sviluppare una piattaforma web **scalabile** e **affidabile** per l'anonimizzazione di dataset



Funzionalità Principali

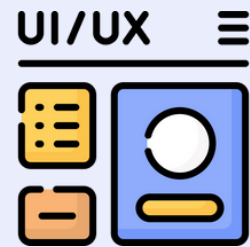
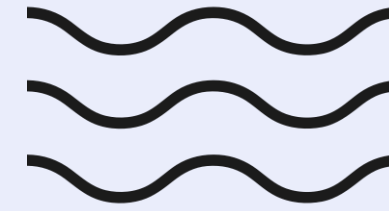
- Consentire agli utenti di **caricare dataset** in modo sicuro (CSV o JSON)
- Offrire una selezione di **algoritmi di anonimizzazione** configurabili
- Generare versioni anonimizzate dei dati da scaricare e **salvare** su database e storage cloud
- Fornire un'**interfaccia semplice** e accessibile per gestire il processo



Caratteristiche Tecniche

- Applicazione completamente **deployata su Google Cloud Platform**
- Infrastruttura gestita tramite **Terraform** per garantire scalabilità e riproducibilità
- Supporto per **dataset** tabellari di qualsiasi schema o tipo di dato

Panoramica dell'architettura



Frontend (React + Vite)



Backend (FastAPI – Python)



Cloud Infrastructure (GCP + Terraform)

- Sviluppata in **React**, bundleizzata con **Vite**
- Build statica servita da **Nginx** in container **Docker**
- Containerizzata con Docker, deployata su **Cloud Run**
- Comunicazione diretta con **Firebase Authentication** lato client

- API REST sviluppata con **FastAPI**
- Autenticazione tramite **OAuth 2.0** (Google) con verifica token Firebase
- Logging centralizzato con **Cloud Logging**
- Salvataggio metadati e gestione dati persistenti nel DB









- Deploy su **Google Cloud Platform (Cloud Run)**
- Gestione infrastruttura cloud e risorse tramite **Terraform**

Flusso operativo utente

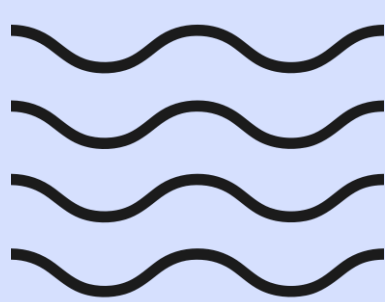
01	Login	L'utente si autentica mediante Google Account
02	Upload Dataset	L'utente carica un file CSV o JSON tramite l'interfaccia web
03	Selezione Algoritmo	L'utente sceglie uno dei quattro algoritmi disponibili (k-anonymity, l-diversity, t-closeness, differential privacy) e imposta il parametro
04	Invio al Backend	Il file e il parametro vengono inviati tramite una chiamata POST al server
05	Elaborazione e Anonimizzazione	Il backend classifica automaticamente le colonne, applica l'algoritmo selezionato e salva il risultato in GCS e i metadati in Firestore
06	Visualizzazione Anteprima	Il backend restituisce un'anteprima delle prime righe del dataset anonimizzato
07	Download Dataset	L'utente può scaricare il file anonimizzato direttamente dalla piattaforma

Algoritmi di anonimizzazione



Algoritmo	Descrizione	Pro e contro	
● k-anonymity	● Ogni individuo è indistinguibile da almeno k-1 persone	 Semplice, scalabile  Può esporre dati sensibili	
● l-diversity	● Garantisce la diversità degli attributi sensibili in ogni gruppo	 Protezione più forte rispetto a k-anonymity  Non protegge sempre la distribuzione degli attributi	
● t-closeness	● Limita la differenza tra la distribuzione degli attributi sensibili nei gruppi e quella dell'intero dataset	 Mantiene la distribuzione dei dati  Più complesso da calcolare	
● differential privacy	● Aggiunge rumore statistico per prevenire l'identificazione, anche con dati esterni	 Garantisce privacy matematica  Può ridurre la precisione dei dati	

Scelte tecniche ed implementative



Algoritmo	Approccio adottato	Caratteristiche distintive
● k-anonymity	<ul style="list-style-type: none">• Generalizzazione progressiva basata su gerarchie modulari per i quasi-identificatori	<ul style="list-style-type: none">• Adattabile a diversi dataset• Classificazione automatica delle colonne
● l-diversity	<ul style="list-style-type: none">• Estensione diretta di k-anonymity con controllo della varietà dei valori sensibili nei gruppi	<ul style="list-style-type: none">• Elevata riusabilità del codice• Uniformità nell'applicazione delle generalizzazioni
● t-closeness	<ul style="list-style-type: none">• Calcolo della distanza tra distribuzioni per garantire la coerenza statistica tra gruppi e dataset globale	<ul style="list-style-type: none">• Preserva la distribuzione dei dati• Ideale per dati sensibili categoriali
● differential privacy	<ul style="list-style-type: none">• Aggiunta di rumore calibrato a livello di dataset (input perturbation) in base al tipo di colonna (numerico/categoriale)	<ul style="list-style-type: none">• Output realistico e coerente• Automazione completa• Adattivo e type-aware per ogni colonna

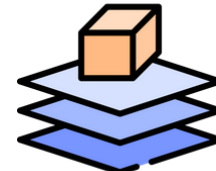


Deployment e scalabilità su GCP



Deploy su Google Cloud Platform

- **Frontend** e **Backend** pubblicamente accessibili e protetti tramite **Firebase Authentication**
- Backend integrato con **Cloud Logging** per visibilità su errori, performance e tracing delle richieste
- Servizi **stateless** con **autoscaling** e **load balancing** automatico



Infrastructure as Code

- Infrastruttura gestita con **Terraform**, per versionamento, riproducibilità e provisioning automatizzato
- Risorse principali:
 - Cloud Run**: esecuzione container scalabili
 - Cloud Storage**: archiviazione file
 - Firestore**: salvataggio metadati
 - Firebase Auth**: gestione autenticazione
 - Cloud Logging**: visibilità su errori e performance



Servizi GCP utilizzati



- **Firebase Auth**: autenticazione sicura via Google
- **Cloud Storage (GCS)**: archiviazione dei file CSV caricati e anonimizzati
- **Firestore**: salvataggio dei metadati dei dataset anonimizzati
- **Cloud Logging**: logging centralizzato e monitoraggio delle performance

Conclusioni e prossimi passi



Conclusioni

- Soluzione **semplice**, **veloce** e **scalabile** per anonimizzare dataset
- Supporto **multi-algoritmo** configurabile
- Deploy completo su **Google Cloud Platform**, con infrastruttura cloud-native e auto-scalabile



Sviluppi futuri

- Possibilità di **migliorare gli algoritmi** implementati o aggiungerne di nuovi
- Elaborazione asincrona via **Pub/Sub** per gestire job pesanti
- Espansione della classificazione automatica, con **supporto multilingua**
- Valutazione della **qualità** post-anonimizzazione



Grazie per l'attenzione!



Scalable and Reliable Systems

DANILA MELELEO
VALENTINA DE RESPINIS