



# AnonimaData

# Agenda

- Requisiti
- Tecnologie usate
- Modelli di anonimizzazione
- Struttura distribuita
- Test e risultati

# Requisiti di AnonimaData

## **Formato dei dataset**

---

Dataset in formato  
JSON e CSV

## **Varietà di algoritmi di anonimizzazione**

---

K-anonymity  
L-anonymity  
Differential Privacy

## **Personalizzazione degli algoritmi**

---

Ogni algoritmo deve  
essere  
personalizzabile  
dall'utente in base  
alle proprie esigenze

# Requisiti di AnonimaData

## **Anteprima dell'anonimizzazione**

---

L'utente deve poter analizzare l'esito dell'anonimizzazione direttamente dall'applicazione

## **Archiviazione dei dataset**

---

Salvataggio persistente dei dataset anonimizzati per poter essere recuperati anche in un momento successivo dell'anonimizzazione

## **Interfaccia web**

---

Utilizzo dell'applicativo tramite una piattaforma web

# Requisiti di AnonimaData

## **Semplicità d'uso**

---

L'utente inesperto deve riuscire ad usare l'applicazione

## **Autenticazione utente**

---

Ogni utente può accedere solamente ai propri dataset

## **Generalizzabilità dei dataset**

---

Devono poter essere presentati in input vari tipi di tabelle e il sistema deve riuscire a gestirle

# Tecnologie utilizzate

## **Google Cloud Platform**

---

Utilizzo di Cloud Run e Cloud Storage, comunicazione tra i servizi via Pub/Sub

## **Terraform**

---

IaC che descrive l'infrastruttura realizzata

## **Docker**

---

Ogni servizio è stato containerizzato e pubblicato su GCP

# Tecnologie utilizzate

## **Python**

---

Utilizzato con Flask per realizzare i servizi di backend

## **React**

---

Framework utilizzato per realizzare la Single Page Application con cui l'utente si interfaccia

## **Firebase**

---

Servizio di Google per gestire esternamente l'autenticazione utente

# Modelli di anonimizzazione

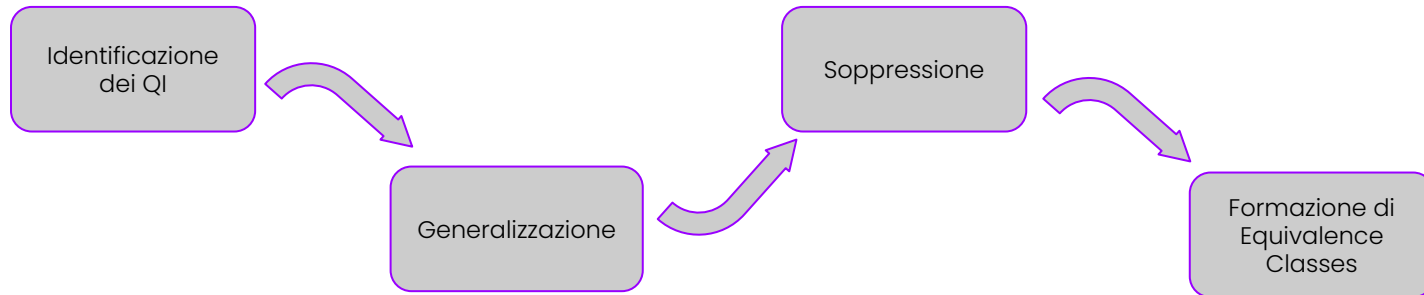
- K-anonymity
- L-diversity
- Differential Privacy



# K-anonymity

---

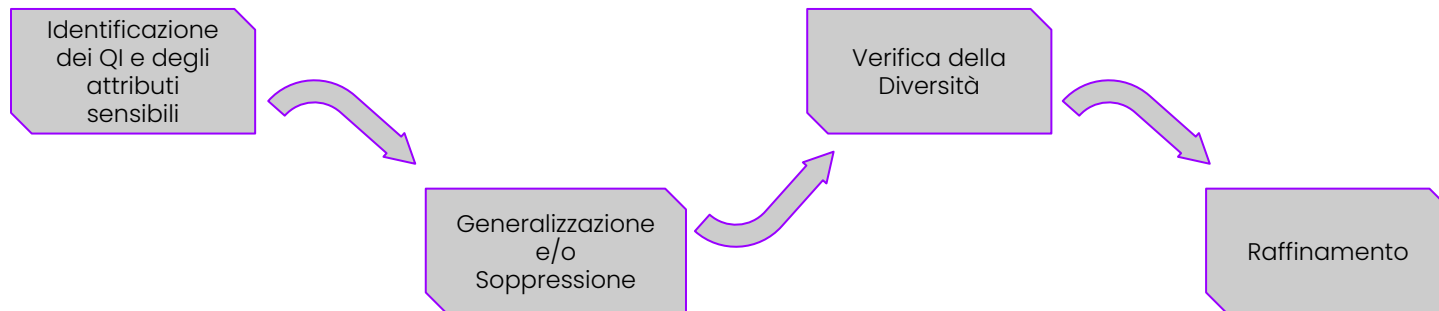
Garantisce che ogni record nel dataset anonimizzato sia **indistinguibile da almeno altri  $(k-1)$  record** rispetto a un insieme di attributi "quasi-identificatori"



# L-diversity

---

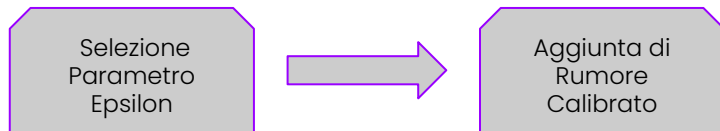
Garantisce che, all'interno di ogni gruppo di  $k$  record indistinguibili (classe di equivalenza), ci siano **almeno  $L$  valori distinti** per ogni attributo sensibile



# Differential Privacy

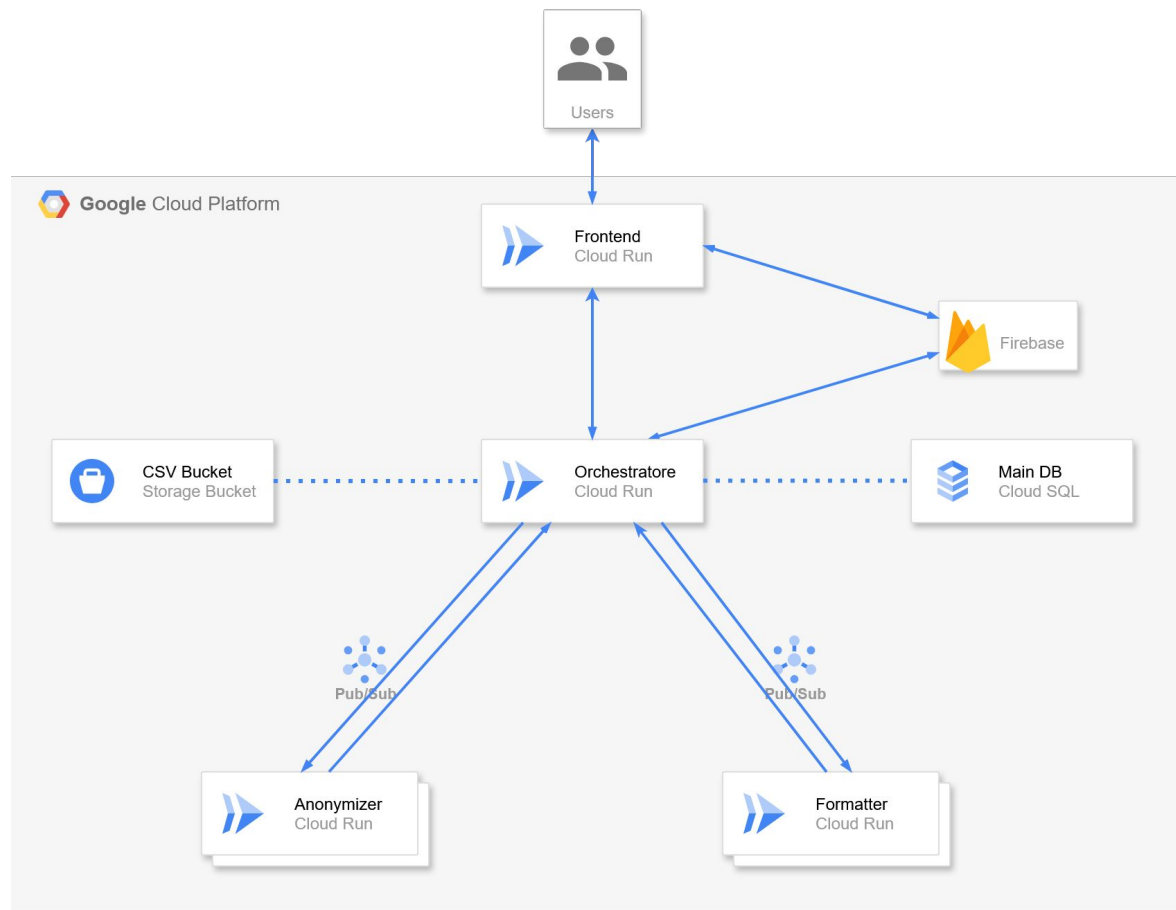
---

Garantisce che la presenza o assenza di un singolo individuo nel dataset **non influenzi significativamente l'output** di un'analisi o di una query



# Architettura del sistema

- Frontend
- Orchestratore
- Formatter
- Anonymizer



---

Lo schema architetturale completo su GCP

# Orchestratore

---

L'Orchestratore è il cuore del backend di AnonimaData e si pone come punto d'ingresso principale per le interazioni degli utenti e la gestione del flusso di lavoro complessivo

# Orchestratore: funzionalità

## **Interfaccia con il frontend**

---

Espone una serie di endpoint API REST utili al frontend per interagire con il sistema

## **Comunicazione interservizio**

---

Funge da hub per la comunicazione di tipo pub sub

## **Verifica dei permessi**

---

Ogni richiesta è autenticata via Firebase per garantire accesso solo ai dati dell'utente.

# Orchestratore: funzionalità

## **Gestione errori**

---

Riceve le notifiche di errori dal pub/sub, aggiornando in caso lo stato del job

## **Persistenza**

---

responsabile della persistenza gestita tramite DB relazionale e Storage ad oggetti (Google Cloud Storage)

## **Gestione stato dei job**

---

Si occupa di salvare uno stato dettagliato completo di informazioni per ogni job presente



# API REST

- */upload\_and\_analyze*
- */get\_status/<job\_id>*
- */request\_anonymization*
- */get\_files*
- */delete/<job\_id>*
- */download/<job\_id>*

# Formatter

---

Primo servizio backend che elabora i dataset caricati dagli utenti. Il suo scopo è quello di preparare i dati grezzi per le successive fasi di anonimizzazione.

# Formatter: funzionalità

## **Standardizzazione del Formato**

---

Riceve i dataset (CSV, JSON) e li converte in un formato standard facilmente manipolabile (DataFrame Pandas)

## **Analisi Approfondita delle Colonne**

---

Scansiona ogni colonna del dataset, rilevando automaticamente i diversi tipi di dato presenti in ciascuna di esse


## **Generazione di Metadati Estesi**

---

Includono informazioni critiche che guidano il processo di anonimizzazione

# Formatter: Metadati

---

- *column\_name*
  - *data\_type*
  - *is\_quasi\_identifier*
  - *should\_anonymize*
- 
- user input**

# Anonymizer

---

Componente che si occupa di trasformare l'output del Formatter, ossia dati potenzialmente sensibili, in una versione protetta e anonimizzata

# Anonymizer: funzionalità

## **Ricezione Richiesta di Anonimizzazione**

---

Dal topic Pub/Sub riceve una richiesta contenente jobId, output Formatter e metadati user-dependant

## **Esecuzione degli Algoritmi di Privacy**

---

Applica tecniche di generalizzazione, soppressione e/o aggiunta di rumore casuale ai dati in chiaro

## **Generazione del Dataset Anonimizzato**

---

Produce un nuovo dataset che rispetta le proprietà di privacy definite dall'algoritmo e dai suoi parametri

# Una prima versione non distribuita

## **Divisione in moduli**

---

Divisione logica sotto forma di moduli.

Frontend visualizzato direttamente dal backend

## **Rilascio come container singolo**

---

Immagine docker unica per tutti i servizi

Scopo principale di testare i moduli di anonimizzazione e il workflow

## **Deploy su server di test**

---

Deploy su macchina con [accesso pubblico](#)

## Dataset Anonymization Tool

### 1. Upload Dataset

Select Dataset (CSV, Excel, JSON, TXT):

Sfogliare... production\_dataset.csv

Upload & Analyze

File analyzed successfully

### 2. Configure Columns

Review detected column types and select Quasi-Identifiers (QI) and columns to anonymize.

Column Name (Type)	Is QI?	Anonymize?
ID (numeric)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
NAME (text)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
BIRTH (date)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
CELLPHONE (phone_number)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
EMAIL (email)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
CODE (alphanumeric)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
ALIAS (text)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
PASSWORD (alphanumeric)	<input type="checkbox"/>	<input checked="" type="checkbox"/>
AGE (numeric)	<input type="checkbox"/>	<input checked="" type="checkbox"/>

### 3. Select Anonymization Method

Choose Method:

k-Anonymity

k-Value (e.g., 3):

3

Start Anonymization

Anonymization completed

Download your anonymized data:

[Full Anonymized Data](#)

[Anonymized Data Sample \(First 10 rows\)](#)

---

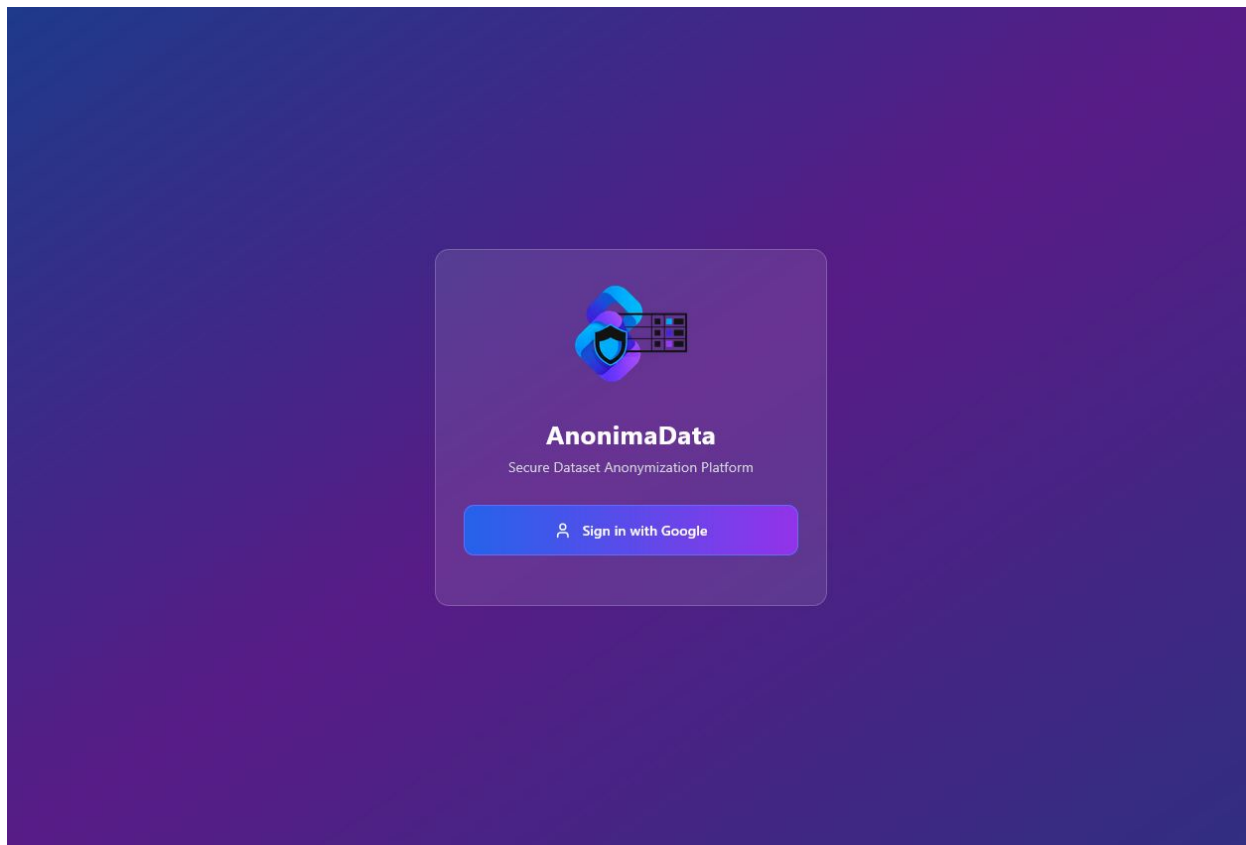
Versione di test di AnonimaData (soluzione AIO)



# Frontend

---

Applicazione web basata su React che permette agli utenti di interagire con il servizio di anonimizzazione. La sua funzione principale è quella di tradurre le operazioni complesse del backend in un'esperienza utente semplice e intuitiva



---

Schermata di login dell'applicazione

## Dashboard

[Upload New Dataset](#)

Total Datasets

**2**


Total Protected Rows

**10741**


### Your Anonymized Datasets



DATASET	ALGORITHM	ROWS	COMPLETED	STATUS	ACTIONS
production_dataset.csv	differential-privacy	4557	12/07/2025, 19:31:24	Anonymized	
company_salary_data_2025.csv	k-anonymity	6184	30/06/2024, 20:51:59	Anonymized	

Dashboard dell'applicazione



## Configure Anonymization

### Column Configuration

Select the type for each column in your dataset


Column Name	Quasi-Identifier	Sensitive
ID	<input type="checkbox"/>	<input type="checkbox"/>
NAME	<input type="checkbox"/>	<input type="checkbox"/>
BIRTH	<input type="checkbox"/>	<input type="checkbox"/>
CELLPHONE	<input type="checkbox"/>	<input type="checkbox"/>
EMAIL	<input type="checkbox"/>	<input type="checkbox"/>
CODE	<input type="checkbox"/>	<input type="checkbox"/>
ALIAS	<input type="checkbox"/>	<input type="checkbox"/>
PASSWORD	<input type="checkbox"/>	<input type="checkbox"/>
AGE	<input type="checkbox"/>	<input type="checkbox"/>

☒ **Quasi-Identifier:** Columns that could be used to identify individuals (e.g., age, zip code) ☐ **Sensitive:** Columns containing sensitive information (e.g., medical data, salary)

---

Selezione dei parametri di anonimizzazione

## Dashboard

 Upload New Dataset

Total Datasets

3










Total Protected Rows

10741



## Your Anonymized Datasets



DATASET	ALGORITHM	ROWS	COMPLETED	STATUS	ACTIONS
users_data.csv				⌚ Waiting for input	 
production_dataset.csv	differential-privacy	4557	12/07/2025, 19:31:24	✅ Anonymized	  
company_salary_data_2025.csv	k-anonymity	6184	30/06/2024, 20:51:59	✅ Anonymized	  

## STATUS

## ACTIONS

⌚ Waiting for input



✅ Anonymized



Dashboard dell'applicazione con un job in sospeso



## Anonymization Results



### Anonymization Completed

Your dataset has been successfully anonymized using k-anonymity.

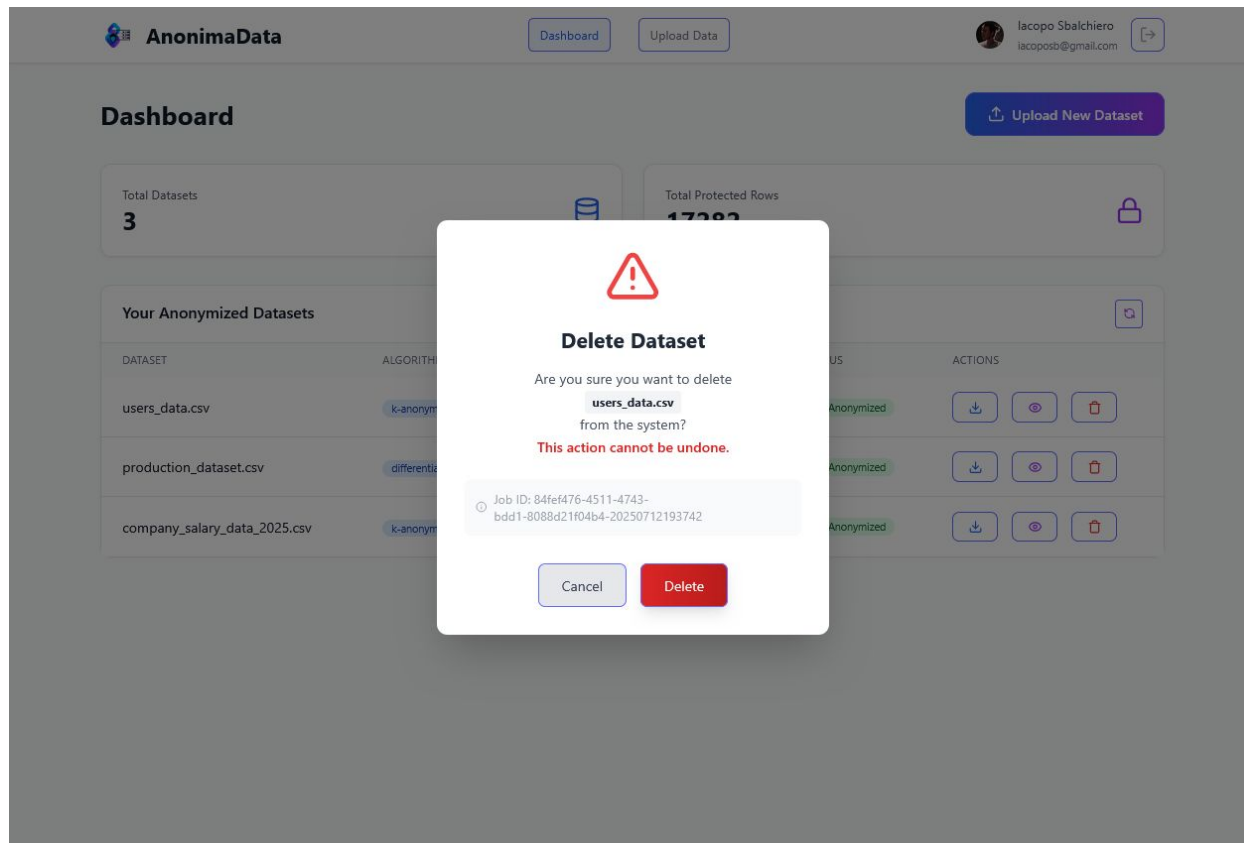
#### Anonymized Data Preview

ID	NAME	BIRTH	CELLPHONE	EMAIL	CODE	ALIAS	PASSWORD	AGE
1	Grace Jones	1970-03-05	[390000226582.00-399994607154.00]	***SUPPRESSED***	0CGGZZ7W	cjpGjX	***SUPPRESSED***	51.00-57.00
2	Alice Miller	2000-06-04	[390000226582.00-399994607154.00]	***SUPPRESSED***	L370HIE1	z5mbHl	***SUPPRESSED***	20.00-26.00
3	Diana Jones	1962-06-03	[390000226582.00-399994607154.00]	***SUPPRESSED***	14NTFC2U	TwjZKz	***SUPPRESSED***	63.00-70.00
4	Eve Johnson	1978-01-22	[390000226582.00-399994607154.00]	***SUPPRESSED***	NPTD4NCW	qYicAa	***SUPPRESSED***	44.00-51.00
5	Alice Davis	2003-01-13	[390000226582.00-399994607154.00]	***SUPPRESSED***	XC3L6OH5	uAzjPI	***SUPPRESSED***	20.00-26.00
6	Frank Williams	1978-04-24	[390000226582.00-399994607154.00]	***SUPPRESSED***	IFIBCj0S	JwgpQc	***SUPPRESSED***	44.00-51.00
7	Ivy Johnson	1956-06-05	[390000226582.00-399994607154.00]	***SUPPRESSED***	X1ZCSYK7	nYrFry	***SUPPRESSED***	63.00-70.00
8	Grace Brown	1976-12-11	[390000226582.00-399994607154.00]	***SUPPRESSED***	4A6DXOK7	kDhJxq	***SUPPRESSED***	44.00-51.00
9	Jack Williams	1953-06-01	[390000226582.00-399994607154.00]	***SUPPRESSED***	7QPU800S	jgiHeD	***SUPPRESSED***	70.00-75.00

[Download Anonymized Data](#)[Close](#)

Job ID: 84fef476-4511-4743-bdd1-8088d21f04b4-20250712193742

Anteprima dei dati anonimizzati



---

Rimozione di un dataset

## Processing Dataset



### An error occurred during processing

Invalid anonymization parameters: Parameter k must be at least 2

 Job ID: ba991b22-f4ce-419e-94db-c40552455368-20250712193938

Return to Dashboard

---

# Gestione degli errori



# Integrazione con Firebase

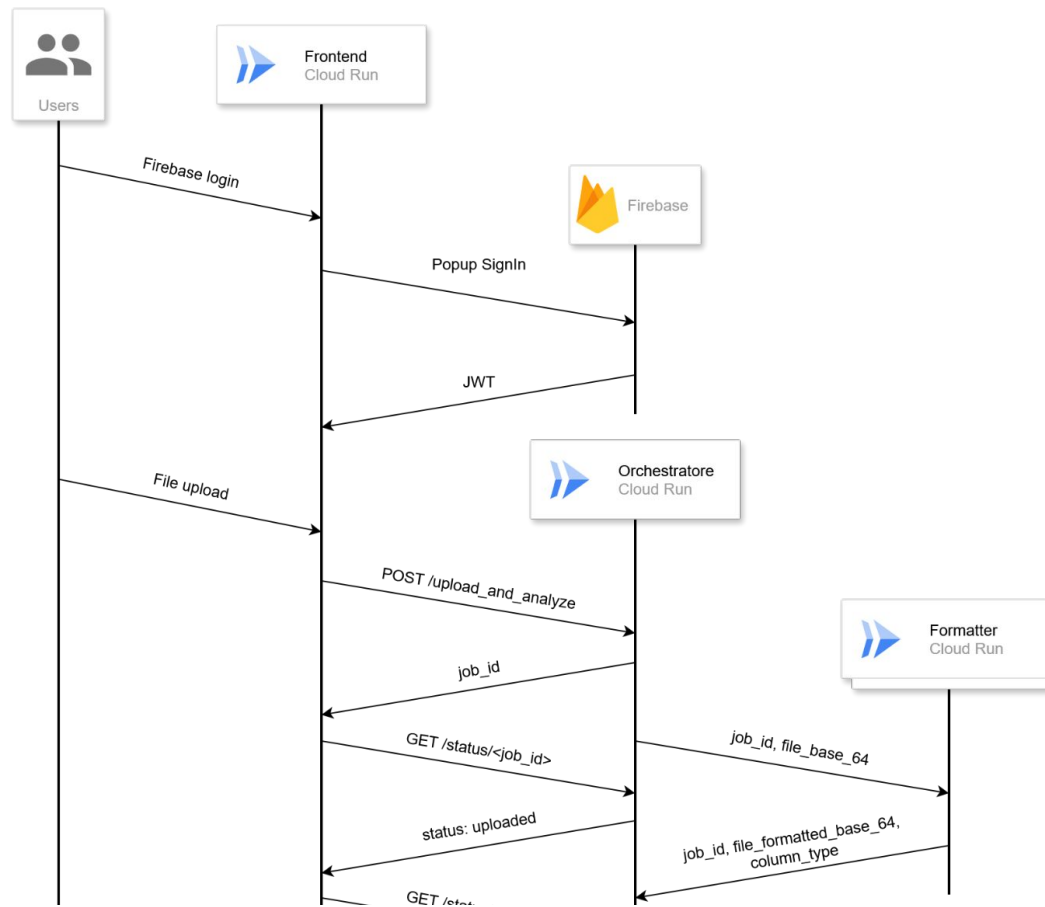
---

- Firebase utilizzato per l'accesso alla piattaforma
- Tutte le richieste devono essere autenticate
- Possibilità di accedere solamente ai propri job
- Frontend gestisce il login e l'autorizzazione
- Orchestratore riceve JWT e restituisce solo i dati collegati all'utente

# Persistenza dei dati

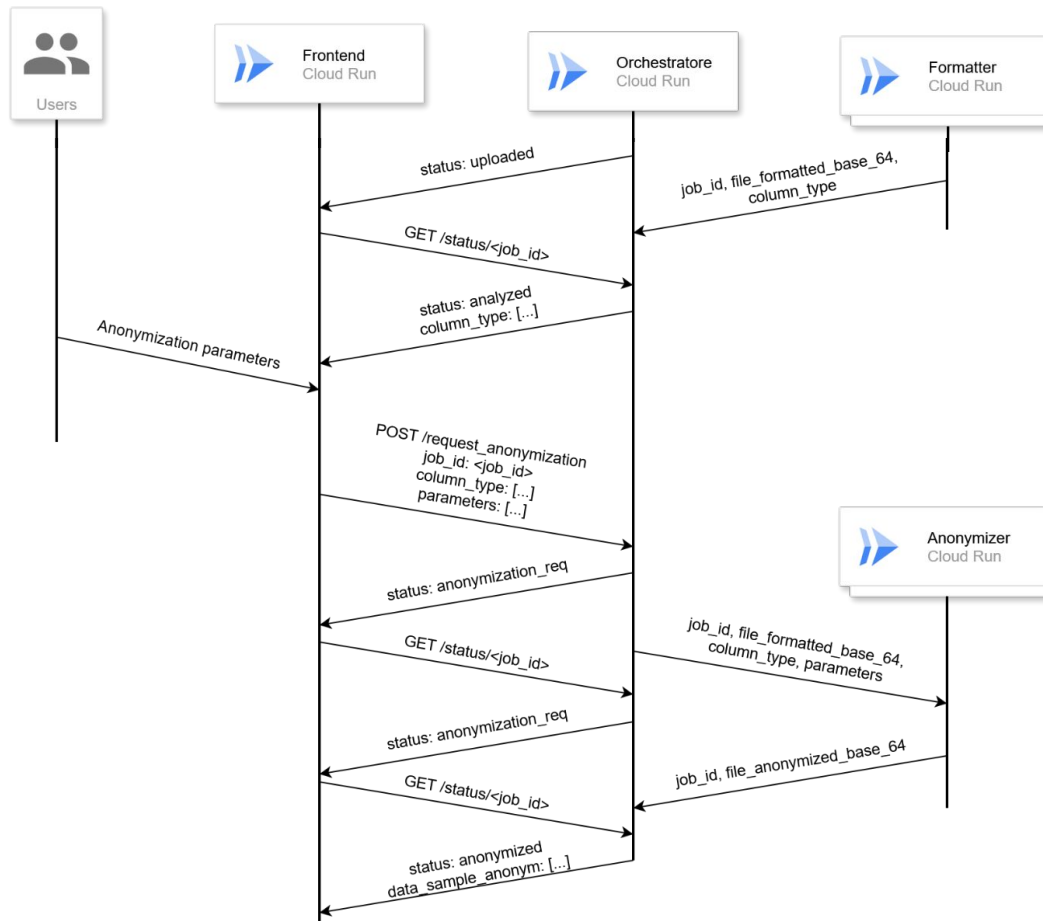
---

- Una entry nel DB per ogni job
- Contiene lo stato, i metadati, i parametri di anonimizzazione, anteprema formato JSON dei dati anonimizzati
- Mantiene il percorso dei file presenti nel bucket
- Salvati in memoria solo i dati anonimizzati

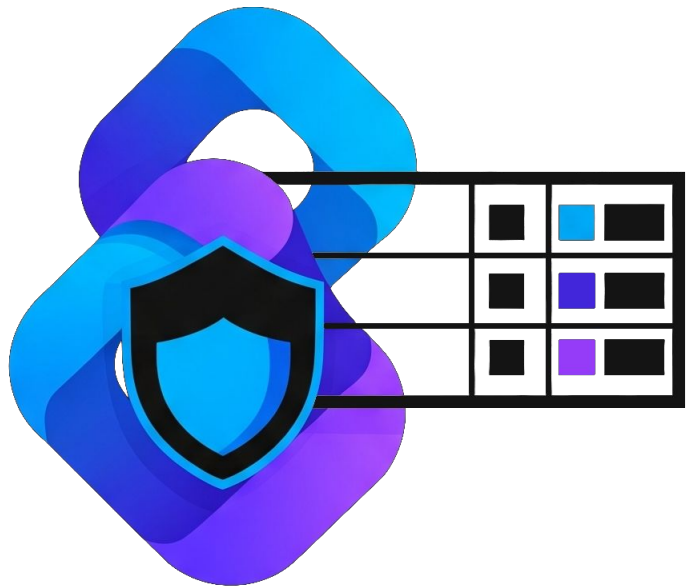


---

Anonimizzazione di un set di dati (1/2)



Anonimizzazione di un set di dati (2/2)



Demo

# Test

- Soak test
- Stress test
- Spike test

# Definizione del test

---

Ogni utente emulato:

- carica un dataset da 1000 righe sulla piattaforma
- sceglie un algoritmo di anonimizzazione
- scarica il file completato

# Difficoltà riscontrate e bottleneck da superare

---

- Problemi di accesso concorrente al database centrale
- Consumo elevato di risorse a causa della ricerca degli stati



# Stress test

## Definizione del test

---

- 50 utenti per 5 min.
- 100 utenti per 5 min.
- 150 utenti per 5 min.
- 200 utenti per 5 min.

## Esito dei test (6282 iter.)

---

Upload del file	98,64%
Job id creato	100,00%
File analizzato	100,00%
Parametri caricati	99,50%
File anonimizzato	100,00%
Download del file	99,92%

# Stress test

## Esito richieste http

---

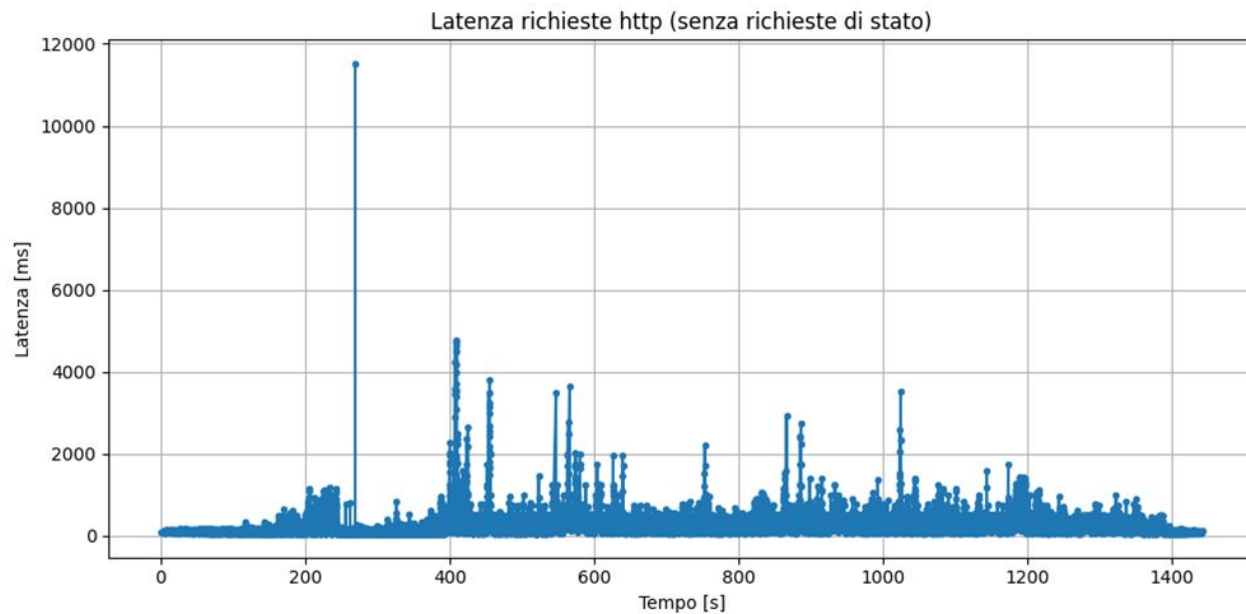
Medio	326.14ms
P(95)	8,48s
Massimo	11,52s

## Durata delle iterazioni

---

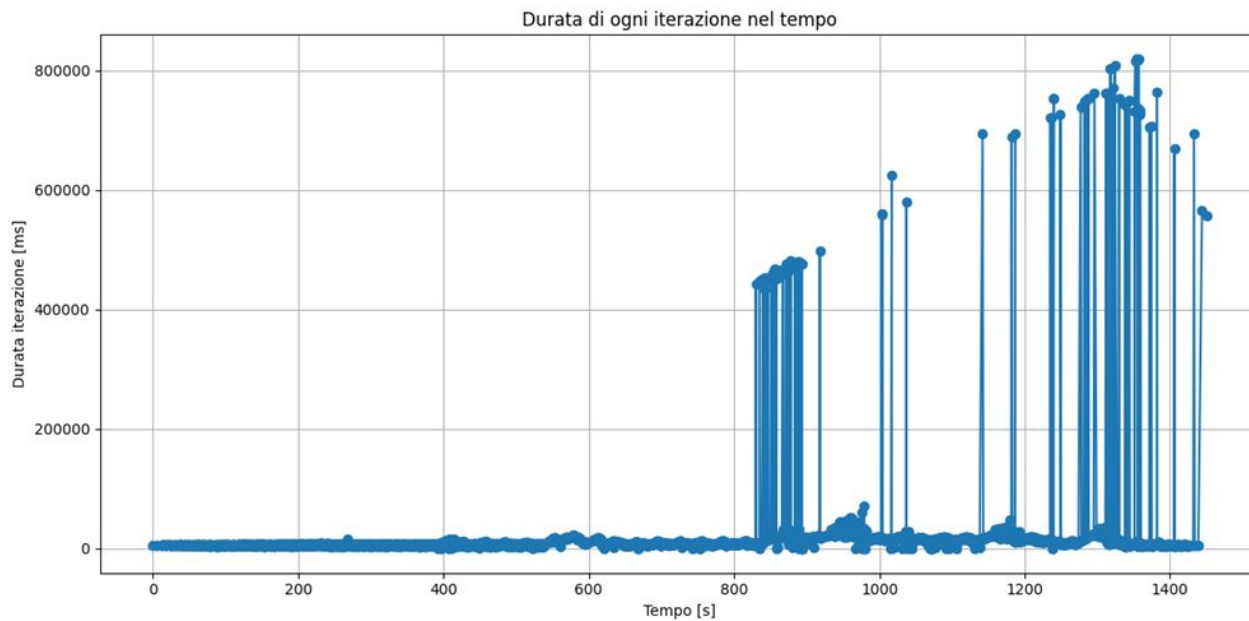
Medio	7s
P(95)	20,75s
Massimo	13m39s





---

Stress test di AnonimaData



Stress test di AnonimaData

# Spike test

## Definizione del test

---

400 utenti in 1 minuto

## Esito dei test (231 iter.)

---

Upload del file	99,80%
Job id creato	100,00%
File analizzato	100,00%
Parametri caricati	99,11%
File anonimizzato	100,00%
Download del file	91,06%



# Spike test

## Esito richieste http

---

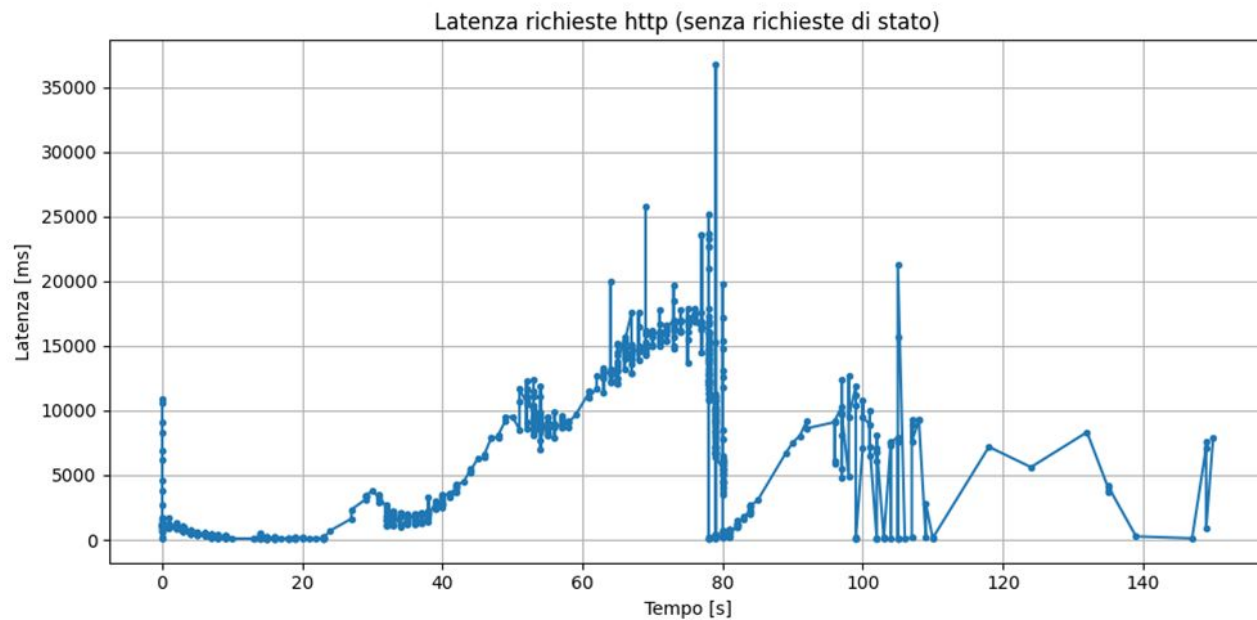
Medio	3,73s
P(95)	18,49s
Massimo	38,97s

## Durata delle iterazioni

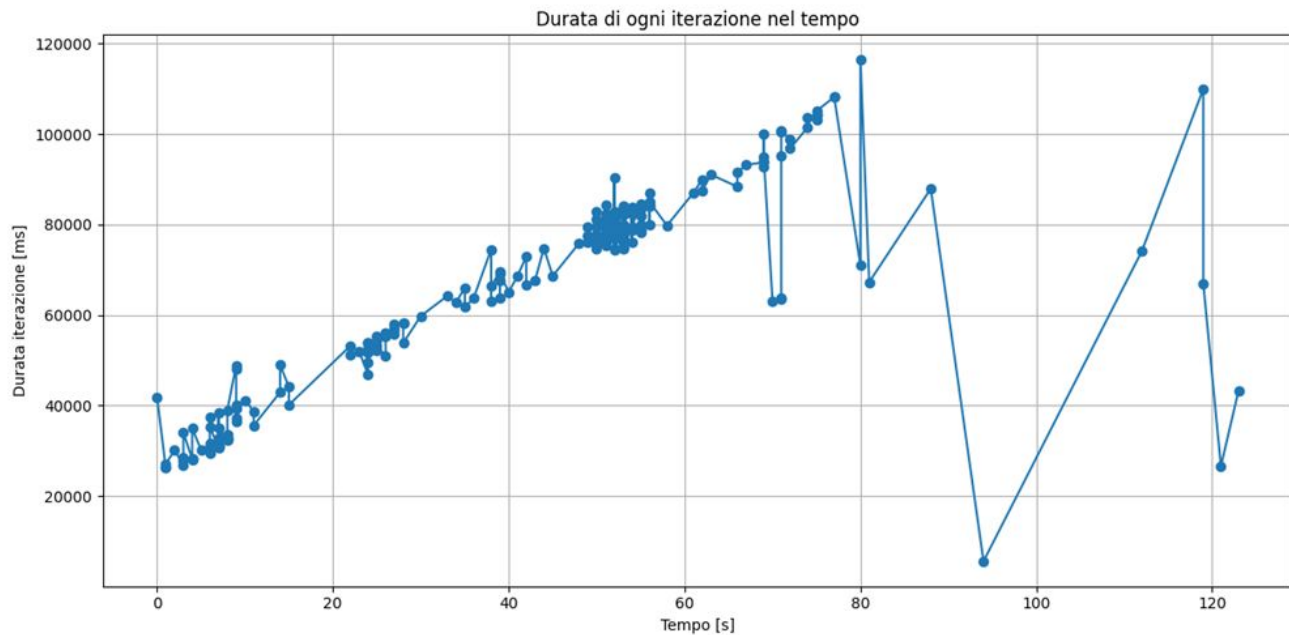
---

Medio	1m16s
P(95)	1m40s
Massimo	1m56s





Spike test di AnonimaData



Spike test di AnonimaData



# Soak test

## Definizione del test

---

50 utenti per 60 minuti

## Esito dei test (16079 iter.)

---

Upload del file	99,32%
Job id creato	100,00%
File analizzato	100,00%
Parametri caricati	99,50%
File anonimizzato	100,00%
Download del file	99,83%



# Soak test

## Esito richieste http

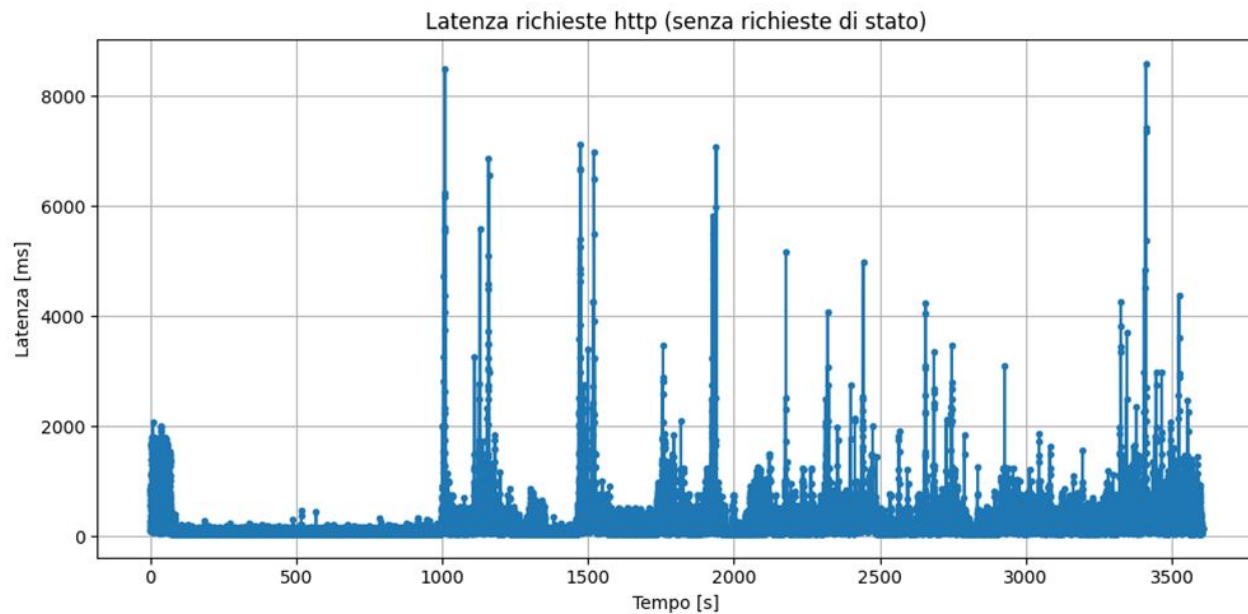
---

Medio	93,55ms
P(95)	2,26s
Massimo	10,89s

## Durata delle iterazioni

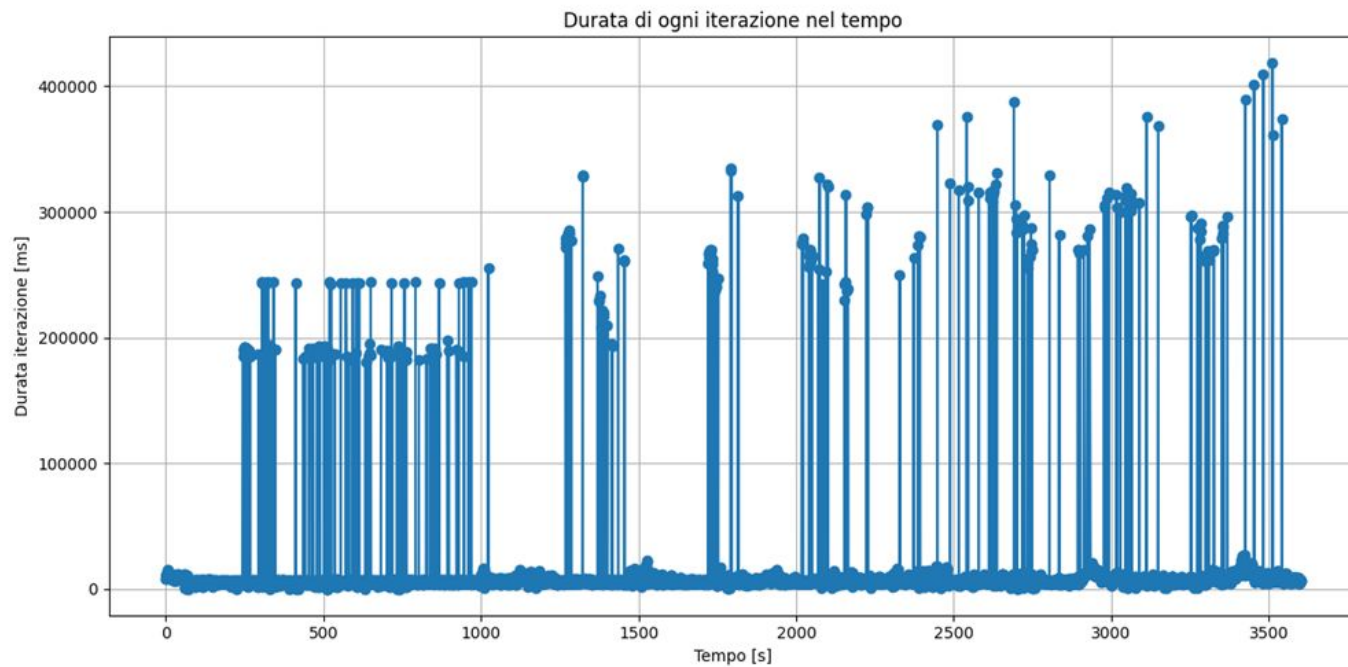
---

Medio	6,06s
P(95)	12,21s
Massimo	6m58s

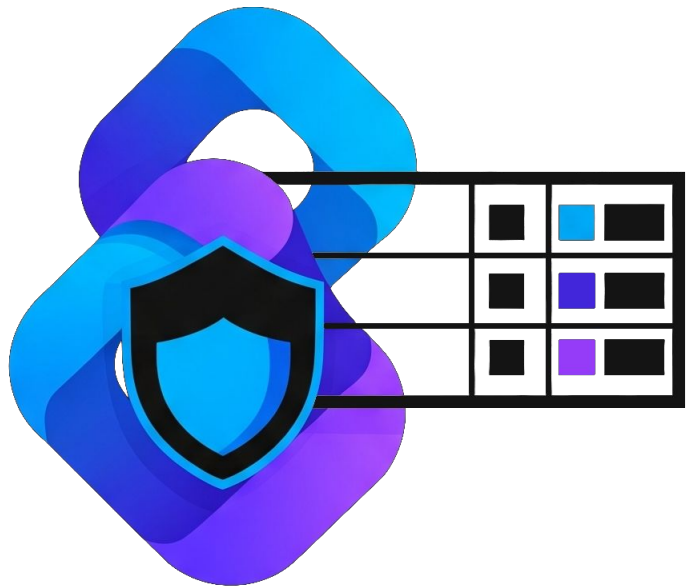


---

Soak test di AnonimaData



Soak test di AnonimaData



# AnonimaData

Fine