Modelli di Processo e Stratificazione a Larga Scala per l'Analisi Comportamentale della Collaborazione su GitHub

Relazione per la prova finale

Candidato: **Federico Mondelli**
Tutore interno: Prof.ssa Stefania Montani

Università del Piemonte Orientale
Dipartimento di Scienze e Innovazione Tecnologica
Corso di Laurea in Informatica

Anno Accademico 2024/2025

Scopo, Strumenti e Specifiche

Scopo e Pregresso

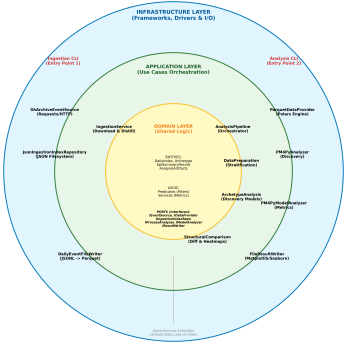
- Obiettivo: Pipeline analitica scalabile per trasformare stream grezzi in modelli di processo.
- Contesto: Analisi Big Data Open Source (GitHub) su **singola workstation**.
- Necessità: Superare il divario tra dato grezzo e modello interpretabile.

Strumenti e Specifiche

- **Sorgente:** GHArchive (JSON massivi).
- **Engine:** Python + PM4Py.
- **Vincolo Metodologico:** Prospettiva *Actor-Centric* (Traccia: Utente + Repository).

Architettura: Disaccoppiamento e Doppia Clean Architecture

Il sistema separa l'Acquisizione (I/O Bound) dall'Analisi (CPU Bound) tramite Storage Locale.



Schema logico del sistema: Doppia Clean Architecture e collocazione dei componenti.

1. Ingestor (Resilienza)

- Streaming riga per riga ($O(1)$ RAM).
- Idempotenza e gestione errori 404.
- Consolidamento automatico JSONL → **Parquet**.

2. Analyzer (Metodologia)

- Elaborazione **Out-of-Core** (Polars Lazy).
- Pipeline di Stratificazione (Quantili).
- Process Mining (Aggregazione + Discovery).
- Confronto Strutturale (Metriche e Heatmap).

Sottosistema di Analisi: Pipeline di Stratificazione

Per confrontare progetti eterogenei, il sistema riduce l'entropia creando cluster omogenei.

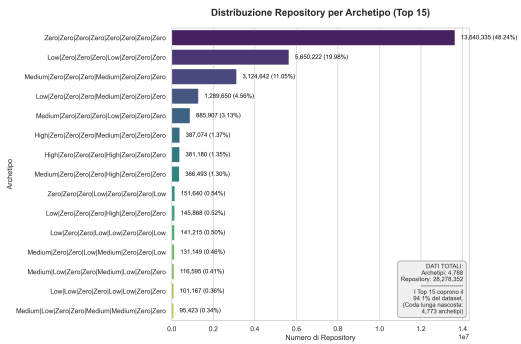
A. Metriche Normalizzate

$$Valore_{Norm} = \frac{Metrica_{Cumulativa}}{Età\ Repository\ (gg)}$$

- **Workload** (Attività tecnica)
- **Collaboration** (Attori unici)
- **Engagement** (Interazioni sociali)
- **Popularity** (Interesse esterno)

B. Quantili Dinamici (Power Law)

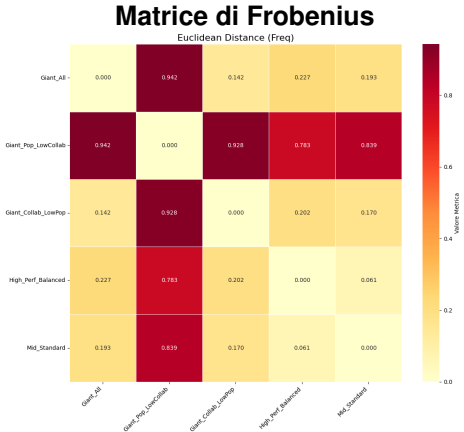
- **Q50:** Attività base.
- **Q90:** Attività strutturata.
- **Q99:** Outlier ("Giant").



Distribuzione Repository per Archetipo

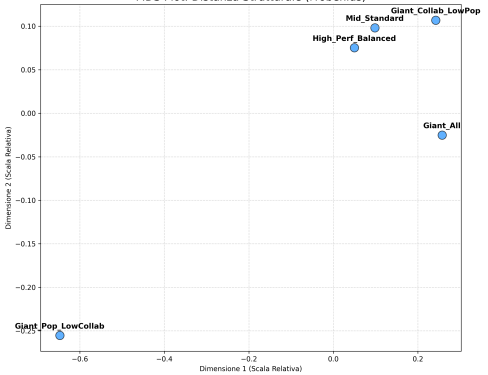
Risultati: Analisi Comparativa Globale

Matrice di Frobenius



Quantificazione esatta delle distanze comportamentali tra tutti i cluster.

MDS Plot

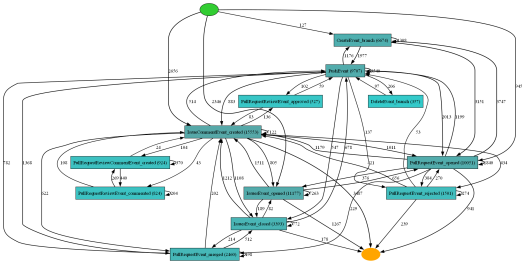


Rappresentazione spaziale dei cluster (Distanza relativa).

Risultati: Process Mining Differenziale (Actor-Centric)

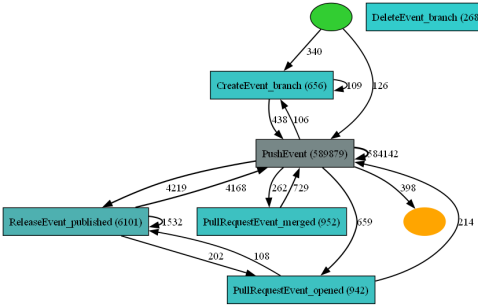
Metodologia: Traccia = *Attore + Repository*. Aggregazione per Archetipo.

Archetipo: ogni metrica “Giant’



Flusso complesso, alta interazione sociale.

Archetipo: metriche sociali “Zero” o “Low”

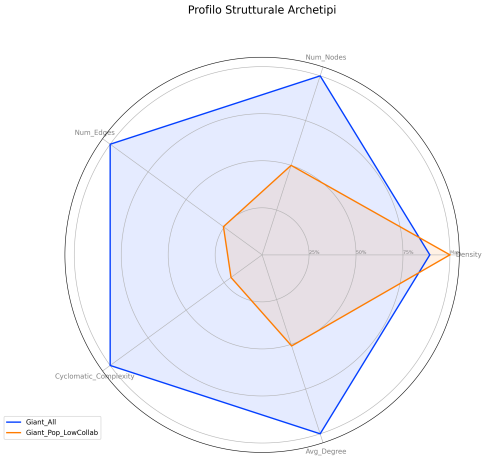


Flusso lineare focalizzato sul codice (Push), scarsa interazione.

L’analisi visiva evidenzia come la complessità diminuisce su Archetipi non collaborativi.

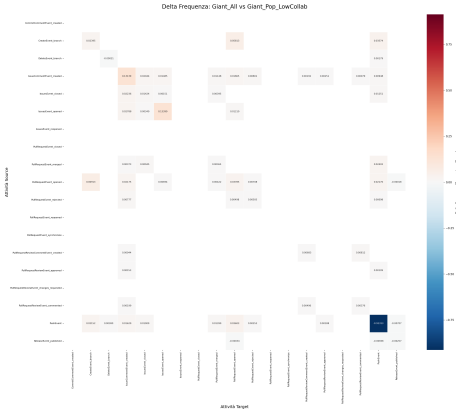
Risultati: Profilazione e Differenze

Profilo Morfologico (Radar)



Confronto di Densità e Complessità Ciclomatica.

Heatmap Differenziale



Analisi puntuale delle divergenze operative.

Conclusioni e Sviluppi Futuri

Il lavoro ha validato l'efficacia di un approccio scalabile per l'analisi di processi su larga scala.

Risultati Conseguiti

- **Architettura:** Resilienza del Data Lake e gestione Out-of-Core.
- **Metodologia:** La stratificazione ha ridotto l'entropia, isolando cluster omogenei.
- **Analisi:** Ricostruzione fedele dei flussi trasversali degli sviluppatori (Actor-Centric).

Sviluppi Futuri

- Migrazione verso framework distribuiti (Apache Spark).
- Conformance Checking Comparativo: Determinare a quale Archetipo corrisponde meglio il comportamento del singolo sviluppatore.
- Integrazione di altre fonti (GitLab) per generalizzare i pattern.

Grazie per l’attenzione