Data Science e Tecnologie per le Basi di Dati
Homework 1

La fase di sviluppo di software di medie o grandi dimensioni è tipicamente svolta mediante l'utilizzo di servizi di versionamento del codice (ad esempio SVN, Git, ecc.). Questi servizi permettono di tenere traccia delle modifiche apportate al codice durante il suo sviluppo. Si supponga di voler utilizzare i dati di un sito di versionamento del codice per realizzare un data wharehouse ed ottenere statistiche sull'utilizzo da parte delle aziende e dei loro collaboratori.
Il codice generato dagli utenti del sito viene suddiviso in repository. Ogni repository rappresenta un progetto in sviluppo ed è associata ad una specifica azienda software. Una repository è inoltre caratterizzata dalla visibilità nel sito web (pubblica o privata), da una categoria di software (ad esempio: "app", "videogame", "driver", ecc...) e da uno o più linguaggi di programmazione. I linguaggi di programmazione riconosciuti dal sito web sono 15 ('Scala', 'Python', 'Java', ecc...).
Ogni repository è composta da uno o più branch paralleli. Un branch è definibile come un flusso di lavoro in cui uno o più collaboratori eseguono gli update del codice. Ogni set atomico di modifiche del codice salvate sul sito detto commit. Per ogni branch sono note la data di creazione e la repository a cui appartiene (una sola). Branch e repository sono caratterizzati da un nome univoco.
Il sistema tiene traccia di alcune informazioni sui collaboratori che apportano modifiche al codice. In particolare ne memorizza il ruolo (ad esempio: 'test', 'sviluppo', 'design', ...) il gruppo di lavoro al quale appartengono (ad esempio: 'sviluppo backend', 'sviluppo UI', ...). Un collaboratore appartiene ad un solo gruppo di lavoro. Un collaboratore lavora generalmente su diversi branch e diverse repository, quindi potenzialmente anche per diverse aziende.
Le modifiche del codice vengono tracciate nel tempo per ogni ora del giorno e data.
Si vogliono realizzare delle statistiche in base al numero di commit apportati nel tempo dai collaboratori, il numero di inserimenti e di rimozioni di linee di codice.
L'analisi deve essere condotta in base a:

- branch, nome del branch, data di creazione, anno di creazione
- nome della repository, azienda, visibilità, categoria, linguaggi
- collaboratore, ruolo collaboratore, gruppo di lavoro
- ora, data, mese, bimestre, quadrimestre, trimestre, anno, mese dell'anno, quadrimestre dell'anno

Homework Tasks

1. Progettare il data warehouse in modo da soddisfare le richieste descritte nelle specifiche del problema. Disegnare lo schema concettuale del datawarehouse e definire lo schema logico (tabelle dei fatti e delle dimensioni).
2. Esprimere le due interrogazioni seguenti utilizzando il linguaggio SQL esteso.
   (a) Considerare le repository private. Separatamente per mese e nome repository, calcolare:
      i. il numero di commit eseguiti in media al giorno
      ii. il numero di commit eseguiti in media in un branch
      iii. il numero mensile cumulativo di commit dall'inizio dell'anno
   (b) Considerare i dati relativi a repository che includono il linguaggio 'Scala'. Separatamente per branch e gruppo di lavoro, eseguire le seguenti analisi:
      i. rapporto tra il numero di inserimenti ed il numero di rimozioni
      ii. percentuale di inserimenti rispetto al totale della repository a cui appartiene il branch
      iii. assegnare un rank ai gruppi di lavoro in base al rapporto tra il numero di inserimenti ed il numero di rimozioni, separatamente per branch