TOP-TED

COMPITO 2

BARCELLA GABRIELE (1058426)

BORDOGNA FILIPPO (1058427)

CASSINA ANDREA (1057831)

MORÉ GABRIELE (1058401)

JOB

- Abbiamo realizzato un job pyspark che permette di aggregare i dati contenuti in 3 file csv distinti per poi caricarli su un database che ci permetta tramite query di:
  - Ricavare i video correlati a quello appena visto
  - Mostrare i video di tendenza
  - Filtrare sui vari campi

DATI

- Siamo partiti dal dataset che ci è stato fornito contenente:
  - ID
  - Nome dello speaker
  - Titolo del talk
  - Dettagli del talk
  - Data di caricamento (mese e anno)
  - Url

DATI(2)

- Siccome il campo visualizzazione non conteneva dati significativi (0 o null) lo abbiamo popolato
- Inoltre abbiamo convertito il formato della data da (mmm-aaaa) a (aaaa-mm) in modo da facilitare l'ordinamento
- Abbiamo inoltre aggiunto la durata dei talk
- Abbiamo creato un altro dataset contenente le trascrizioni presenti per ogni talk
- Viste le difficoltà nel reperire dati direttamente da TEDx abbiamo deciso in questa prima fase di utilizzare dati verosimili ma non reali
- Laddove necessario (numero di visualizzazioni e durata) abbiamo usato un generatore di numeri casuali

SCHEMA FINALE

- n_views: double (nullable = true)
- durate_sec: double (nullable = true)
- _id: string (nullable = true)
- main_speaker: string (nullable = true)
- title: string (nullable = true)
- details: string (nullable = true)
- posted: string (nullable = true)
- url: string (nullable = true)
- tags: array (nullable = true)
  - element: string (containsNull = true)
- watch_next: array (nullable = true)
  - element: struct (containsNull = true)
    - url: string (nullable = true)
    - watch_next_idx: string (nullable = true)
    - main_speaker: string (nullable = true)
    - title: string (nullable = true)
    - details: string (nullable = true)
    - posted: string (nullable = true)
    - n_views: double (nullable = true)
    - durate_sec: double (nullable = true)
    - tags: array (nullable = true)
      - element: string (containsNull = true)
- transcriptions: array (nullable = true)
  - element: struct (containsNull = true)
    - language: string (nullable = true)
    - abbreviation: string (nullable = true)
    - phrase: string (nullable = true)

CRITICITÀ E POSSIBILI EVOLUZIONI

- Difficoltà nel reperire le informazioni da TEDx
- Manipolazione di file CSV

- Possibilità di interazione con le API di TEDx
- Aggiunta di nuovi campi per funzionalità ancora non implementate
- Trasformare i dataset da CSV a JSON (in modo da mantenere le strutture gerarchiche)

LINK UTILI

- Trello
- GitHub