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

tedx_data	
PK	_id String NOT NULL
	n_views Integer NOT NULL
	durate_sec Integer NOT NULL
	main_speaker String NOT NULL
	title String NOT NULL
	details String NOT NULL
	posted String NOT NULL
	url String NOT NULL
	tags Object Array
	watch_next Object Array
	trascriptions Object Array


```
_id: "0aa852c7bf5073f58864742f15c086f8"
n_views: 72239748
durate_sec: 2045
main_speaker: "Wobbly World"
title: "Diversity in harmony"
details: "As 12 musicians from disparate cultures harmonize their eastern and we..."
posted: "2018-02"
url: "https://www.ted.com/talks/wobbly_world_diversity_in_harmony"
tags: Array
  0: "TED"
  1: "talks"
  2: "music"
  3: "culture"
  4: "arts"
watch_next: Array
  0: Object
    url: "https://www.ted.com/talks/abigail_washburn_building_us_china_relations..."
    watch_next_idx: "827e60e4f85d9c815c5278bc24a4ca5a"
    main_speaker: "Abigail Washburn"
    title: "Building US-China relations ... by banjo"
    details: "Abigail Washburn wanted to be a lawyer improving US-China relations --..."
    posted: "2012-04"
    n_views: 658660
    durate_sec: 3289
  > tags: Array
  > 1: Object
  > 2: Object
transcriptions: Array
  0: Object
    language: "Italian"
    abbreviation: "IT"
    phrase: "Questa è solo una frase di prova"
  > 1: Object
  > 2: Object
```

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