TEDcoffee

HOMEWORK 2

AWS Glue

Daniel Hernan Altamirano
Davide Cesani
Federico Nespoli

TEDcoffee – Dati necessari

- Per realizzare il servizio progettato, sono necessari i dati contenuti nei dataset:
  - *Tedx_dataset*: informazioni generali sui talks
  - *Tags_dataset*: tags relativi ai talks
  - *Watch_next_dataset*: riferimenti al talk da visualizzare successivamente al corrente
  - *duration_dataset*: informazioni relative alla durata dei talks

In questa sede, non si considera l'utilizzo del servizio AWS cognito per l'autenticazione.

Daniel Hernan Altamirano
Davide Cesani
Federico Nespoli

TEDcoffee – Utilizzo dati watch_next_dataset

- I dati presenti nel dataset watch_next_dataset vengono utilizzati, congiuntamente a quelli di duration_dataset, per permettere all'utente di fruire di un ulteriore talk inerente a quello corrente.
  - In particolare, vengono mostrati i talk *più brevi* tra quelli inerenti.

Daniel Hernan Altamirano
Davide Cesani
Federico Nespoli

TEDcoffee – Job PySpark – Modifiche - Roadmap

Watch_next_dataset

- Leggere file presente in AWS S3
- Processare il file (.csv)
- Pulizia dataframe (l'oggetto creato dopo il processo del file .csv):
  - Sono presenti delle righe duplicate e degli URL che non sono relativi a dei talk
- Join con il dataframe contenente le informazioni sui video e i tags.

Duration_dataset

- Leggere file presente in AWS S3
- Processare il file (.csv)
- Join con il dataframe contenente le informazioni sui video, i tags e le informazioni di watch_next aggiunte precedentemente.

Daniel Hernan Altamirano
Davide Cesani
Federico Nespoli

TEDcoffee – Job PySpark – modifiche - codice

```
# LETTURA FILE watch_next_dataset
next_dataset_path = "s3://ted-coffee-data/watch_next_dataset.csv"
next_dataset = spark.read.option("header","true").csv(next_dataset_path)

# CREATE THE AGGREGATE MODEL
# selezione delle sole tuple con URL di un talk.
# collect_set per evitare duplicati.
next_dataset = next_dataset.select("*").where(col("url").like("%com/talks/%"))
next_dataset_agg = next_dataset.groupBy(col("idx").alias("idx_ref")).agg(collect_set("watch_next_idx").alias("next_idx"), collect_set("url").alias("next_url"))
next_dataset_agg.printSchema()

# ADD watch_next info TO TEDX_DATASET
# (left) join con il dataset che verrà importato in MongoDB
tedx_dataset_agg = tedx_dataset_agg.join(next_dataset_agg, tedx_dataset_agg.idx == next_dataset_agg.idx_ref, "left") \
    .drop("idx_ref") \
    .select(col("*"))
tedx_dataset_agg.printSchema()

# LETTURA FILE duration_dataset
duration_dataset_path = "s3://ted-coffee-data/duration_dataset.csv"
duration_dataset = spark.read.option("header","true").csv(duration_dataset_path)

# ADD duration info TO TEDX_DATASET
# (left) join con il dataset che verrà importato in MongoDB
tedx_dataset_agg = tedx_dataset_agg.join(duration_dataset, tedx_dataset_agg.idx == duration_dataset.idx_ref, "left") \
    .drop("idx_ref") \
    .select(col("idx").alias("_id"), col("*")) \
    .drop("idx")
tedx_dataset_agg.printSchema()
```

- I commenti del codice descrivono lo scopo delle istruzioni che li seguono.
- *Il restante codice, non visibile in figura, è rimasto invariato rispetto a quanto scritto durante la lezione.*

Daniel Hernan Altamirano
Davide Cesani
Federico Nespoli

TEDcoffee – Schema dati


```
_id: "8d2005ec35280deb6a438dc87b225f89"
main_speaker: "Alexandra Auer"
title: "The intangible effects of walls"
details: "More barriers exist now than at the end of World War II, says designer..."
posted: "Posted Apr 2020"
url: "https://www.ted.com/talks/alexandra_auer_the_intangible_effects_of_wal..."
tags: Array
    0: "TED"
    1: "talks"
    2: "design"
    3: "society"
    4: "identity"
    5: "social change"
    6: "community"
    7: "humanity"
    8: "TEDx"
next_idx: Array
    0: "5bd34fcc55d9e1267f605fa0c060d54e"
    1: "8576654442b6633b1dc0eb48a989172a"
    2: "078766d6cc461cf71d45dc268b66db95"
    3: "d9896b41b372ec60cdd3c662e57caad3"
    4: "fe35edd737282ab3a325f2387cf1b50b"
    5: "5134ae81a27c94354173f38e84289ad5"
next_url: Array
    0: "https://www.ted.com/talks/ronald_rael_an_architect_s_subversive_reimag..."
    1: "https://www.ted.com/talks/megan_campisi_and_pen_pen_chen_what_makes_th..."
    2: "https://www.ted.com/talks/anna_heringer_the_warmth_and_wisdom_of_mud_b..."
    3: "https://www.ted.com/talks/will_hurd_a_wall_won_t_solve_america_s_borde..."
    4: "https://www.ted.com/talks/alex_honnold_how_i_climbed_a_3_000_foot_vert..."
    5: "https://www.ted.com/talks/julia_dhar_how_to_disagree_productively_and_..."
duration: "11:40"
```


- In figura si nota il risultato dell'esecuzione del Job.
  - In particolare, si nota l'aggregazione dei *next_idx* e dei *next_url*.
  - È presente inoltre, il campo duration che indica la durata del talk.

Daniel Hernan Altamirano
Davide Cesani
Federico Nespoli

TEDcoffee – Criticità ed Evoluzioni

Criticità

- Tempistiche di esecuzione del job su AWS.
- Individuazione errore nel log non immediata.
- Necessità di pulizia del dataset watch_next_dataset da tuple inutili.
- Creazione manuale del dataset duration_dataset causa la mancanza dello script di scraping dedicato.

Evoluzioni

- Scraping scripts per l'aggiornamento automatico dei dati.
- Integrazione dei dati utente raccolti con AWS Cognito:
  - Conteggio delle views degli utenti di TEDcoffee
    - *Idea*: creazione di un file .csv denominato views_dataset, aggiornato in tempo reale dalle informazioni raccolte da AWS Cognito. Il salvataggio di tale file in S3 produce l'esecuzione della lambda function di trigger che avvia il job per l'integrazione dei dati. Ciò comporterà l'aggiunta di un campo «views» nello schema dati.

Daniel Hernan Altamirano
Davide Cesani
Federico Nespoli