GEOX

UNA STRADA VERSO IL FUTURO

AGGIUNTA DATASET

WATCH_NEXT

- Si è aggiunto il dataset watch_next_tedx nel bucket S3.
- Si è modificato lo script del processo contenuto in AWS Glue.
  - Si è aggiunto il dataset, indicandone il path. (I duplicati vengono rimossi)
  - Raggruppa i talk per id, creando successivamente il modello di aggregazione.
  - Infine, abbiamo collegato il dataset principale con il nuovo set di dati, associandone l'ID.

```
next_dataset_agg2 = next_dataset
.groupBy(col("idx").alias("idx_ref_n"))
.agg(collect_list("watch_next_idx")
                  .alias("next_idx"))

tedx_dataset_agg2 = tedx_dataset_agg
.join(next_dataset_agg2, tedx_dataset_agg._id ==
next_dataset_agg2.idx_ref_n, "left")
.drop("idx_ref_n")
```

AGGIUNTA DATASET #2

GEO_TALK

- Si è aggiunto il dataset geo_tedx_dataset nel bucket S3.
- Si è modificato lo script del processo contenuto in AWS Glue.
  - Si è aggiunto il dataset, indicandone il path.
  - Infine, si è collegato il dataset principale con il nuovo set di dati, associandone l'ID.
  - In aggiunta, abbiamo generato una struttura dati per contenere le informazioni relative all'area geografica.

```
tedx_dataset_agg3 = tedx_dataset_agg2
.join(geo_dataset_agg3, tedx_dataset_agg2._id ==
geo_dataset_agg3.idx, "left")
.select(col("*"), struct(col("continent"),
col("nation"),col("city")).alias("geo_area"))
.drop("idx","continent","nation","city")
```

AGGIUNTA DATASET #3

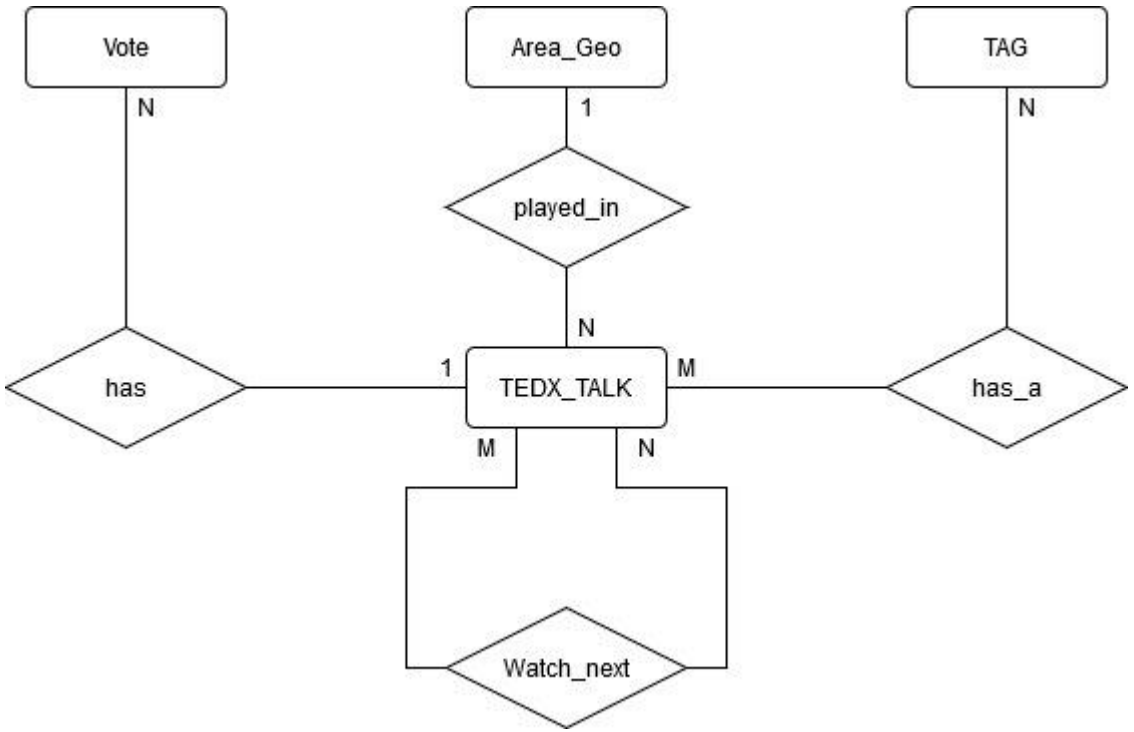
VOTE_USER

- Si è aggiunto il dataset vote_user_dataset nel bucket S3.
- Si è modificato lo script del processo contenuto in AWS Glue.
    - Si è aggiunto il dataset, indicandone il path.
    - Abbiamo raggruppato i voti per ID del Talk.
    - Inoltre abbiamo creato una struttura per la visualizzazione dei voti.
    - Abbiamo effettuato una left join con il dataset principale.

```
vote_dataset_agg4 =
vote_dataset.groupBy(col("idx_tedx")).agg(collect
_list(struct(col("date"),col("time"),col("mail_us
er"),col("vote"))).alias("vote_user"))

tedx_dataset_agg4 =
tedx_dataset_agg3.join(vote_dataset_agg4,
tedx_dataset_agg3._id ==
vote_dataset_agg4.idx_tedx, "left") \
    .drop("idx_tedx")
```

ER SCHEMA



MongoDB Example


```
_id: "4adc9fee977fa04c357ed4c9b52aa3cc"
main_speaker: "Butterscotch"
title: ""Accept Who I Am""
details: "Firing off her formidable beatboxing skills, musician Butterscotch ser..."
posted: "Posted Apr 2020"
url: "https://www.ted.com/talks/butterscotch_accept_who_i_am"
num_views: "0"
tags: Array
  0: "TED"
  1: "talks"
  2: "live music"
  3: "music"
  4: "performance"
next_idx: Array
  0: "9f7b1654e792011b7e1c6f4288520226"
  1: "edb909effab1896976984a06df06f94e"
  2: "8e6129177f808f12381d5db92813d878"
  3: "090a8f3b93c36209b3b3a6a19bfeede5"
geo_area: Object
  continent: "Europe"
  nation: "Italy"
  city: "Benevento"
vote_user: Array
  0: Object
    date: "2019-08-03"
    time: "2:30:18"
    mail_user: "wsautter0@nifty.com"
    vote: "3"
  1: Object
    date: "2019-09-12"
    time: "4:09:36"
    mail_user: "kgaydenk@photobucket.com"
    vote: "2"
  2: Object
  3: Object
  4: Object
  5: Object
  6: Object
```

USO DEI DATI

?

WATCH_NEXT	Guardare i video correlati a quello che si sta guardando.
GEO_TALK	Fornisce informazioni aggiuntive ai video. Permette di espandere le possibilità di ricerca. Mostra contenuti relativi all'area geografica.
VOTE_USER	Fornisce una valutazione complessiva della conferenza. Permette di attribuire un giudizio personale. È essenziale per la creazione di una classifica basata sul parere degli utenti.

CRITICITÀ TECNICHE

- L’aggiornamento dei dati relativi alla correlazione dei talk è svolto in maniera manuale.
- Un secondo aggiornamento da tenere in considerazione è quello relativo all’area geografica, siccome sono dati aggiuntivi.
  Anche in questo caso, l’aggiornamento risulta statico.
- Correlazione dei video effettuata staticamente.
  Non avviene una correlazione basata su altri parametri e/o algoritmi.
- Ogni minimo update implica un aggiornamento di tutto il database.
- Possibilità di incongruenze nel dataset.
- Alto tempo di importazione dei nuovi dataset.

POSSIBILI EVOLUZIONI

- Script per aggiornamento automatico del dataset dei Talk, basato sui dati ufficiali del sito TED.
- Implementazione algoritmo per correlazione video basata su parametri.
- Script per aggiornamento automatico del dataset dei Talk, relativo all'area geografica e ai voti.
- Gestione migliorata del dataset in termini di nuovi elementi e di update di poco impatto.
- Implementazione di voto con commento, per specificare meglio il proprio parere.
- Analisi legate alle valutazioni e alle zone geografiche.