TEDays

2ª consegna

A cura di: (1085186) Davide Mai, Haaim Syed (1086229)

CONTENUTI

Gestione video consigliati

Abbiamo modificato il job PySpark iniziale per aggiungere alcuni campi al dataset. In particolare, abbiamo aggiunto i campi next_id, watch_next_id e watch_next_title. Abbiamo sfruttato i dati presenti nel dataset related_videos.

```
# MANAGE THE RELATED VIDEOS DATASET

related_dataset_path = "s3://tedx-2025-data-dado/related_videos.csv"
related_dataset = spark.read.option("header","true").csv(related_dataset_path)

related_dataset_agg = related_dataset.groupBy(col("id").alias("next_id")).agg(collect_list("related_id").alias("watch_next_id"),
                                                                              collect_list("title").alias("watch_next_title"))

tedx_dataset_agg = tedx_dataset_agg.join(related_dataset_agg, tedx_dataset_agg._id == related_dataset_agg.next_id, "left")
```

Gestione video consigliati

Su MongoDB otteniamo quindi i due array watch_next_id e watch_next_title, che rappresentano gli id e i titoli dei video consigliati.


```
_id: "567505"
slug : "ben_proudfoot_the_true_story_of_the_iconic_tagline_because_i_m_worth_i…"
speakers : "Ben Proudfoot"
title : "The true story of the iconic tagline "Because I'm worth it." | The Fin…"
url : "https://www.ted.com/talks/ben_proudfoot_the_true_story_of_the_iconic_t…"
description : "From two-time Oscar winner Ben Proudfoot comes THE FINAL COPY OF ILON …"
duration : "1059"
publishedAt : "2025-03-07T13:49:56Z"
▸ tags : Array (8)
next_id : "567505"
▾ watch_next_id : Array (3)
    0: "121643"
    1: "87043"
    2: "88795"
▾ watch_next_title : Array (3)
    0: "Why are women still taken less seriously than men?"
    1: "What if women built the world they want to see?"
    2: """"A seat at the table"" isn't the solution for gender equity""
```

I nostri dati

Cercando in Internet abbiamo trovato un dataset pubblico su kaggle (https://www.kaggle.com/datasets/dhavalrupapara/world-countries-holidays-dataset-2023) rappresentante una moltitudine di festività nazionali in ogni Paese del mondo.

Esempio di dataset per un Paese

```
Date,Name,Type,Country Name,Country Code
2023-02-15,Liberation Day,['National holiday'],Afghanistan,AF
2023-03-21T01:54:20+04:30,March Equinox,['Season'],Afghanistan,AF
2023-03-21,Nauruz,['Observance'],Afghanistan,AF
2023-03-23,First Day of Ramadan,['Observance'],Afghanistan,AF
2023-04-22,Eid al-Fitr,['National holiday'],Afghanistan,AF
2023-04-23,Eid al-Fitr Holiday,['National holiday'],Afghanistan,AF
2023-04-24,Eid al-Fitr Holiday,['National holiday'],Afghanistan,AF
2023-04-28,Afghan Victory Day,['National holiday'],Afghanistan,AF
2023-05-01,Labor Day,['National holiday'],Afghanistan,AF
2023-06-21T19:27:49+04:30,June Solstice,['Season'],Afghanistan,AF
2023-06-27,Day of Arafat,['National holiday'],Afghanistan,AF
2023-06-28,Eid al-Qurban,['National holiday'],Afghanistan,AF
```

I nostri dati

La nostra applicazione permetterà all'utente di scegliere di vedere festività native al proprio paese oppure le giornate internazionali stilate dall'ONU, o entrambe le cose!

Abbiamo per questo motivo usato la lista di queste giornate (https://www.un.org/en/observances/list-days-weeks) per creare un dataset completo di tutto unendolo insieme a tutti i Paesi.

dataset delle festività internazionali, dopo essere stato formattato come quello pubblico

```
Date,Event,Type
2025-01-04,World Braille Day,['Observance']
2025-01-24,International Day of Education,['Observance']
2025-01-26,International Day of Clean Energy,['Observance']
2025-01-27,International Day of Commemoration in Memory of the Victims of the Holocaust,['Observance']
2025-01-28,International Day of Living Together in Peace,['Observance']
2025-02-01,World Interfaith Harmony Week,['Observance']
2025-02-02,World Wetlands Day,['Observance']
2025-02-04,International Day of Human Fraternity,['Observance']
2025-02-06,International Day of Zero Tolerance for Female Genital Mutilation,['Observance']
2025-02-10,World Pulses Day,['Observance']
2025-02-10,International Day of the Arabian Leopard,['Observance']
2025-02-11,International Day of Women and Girls in Science,['Observance']
2025-02-12,International Day for the Prevention of Violent Extremism as and when Conducive to Terrorism,['Observance']
2025-02-13,World Radio Day,['Observance']
2025-02-17,World Tourism Resilience Day,['Observance']
2025-02-20,World Day of Social Justice,['Observance']
2025-02-21,International Mother Language Day,['Observance']
2025-03-01,Zero Discrimination Day,['Observance']
```

Il Job Pyspark

Un estratto del job

```python
holidays_dataset_path = "s3://tedx-holidays/ONU_modified.csv"

args = getResolvedOptions(sys.argv, ['JOB_NAME'])

sc = SparkContext()

glueContext = GlueContext(sc)
spark = glueContext.spark_session

job = Job(glueContext)
job.init(args['JOB_NAME'], args)

#### READ INPUT FILES TO CREATE AN INPUT DATASET
holidays = spark.read \
    .option("header","true") \
    .option("quote", "\"") \
    .option("escape", "\"") \
    .csv(holidays_dataset_path)

holidays.printSchema()

write_mongo_options = {
    "connectionName": "TEDX1",
    "database": "unibg_tedx_2025",
    "collection": "holidays",
    "ssl": "true",
    "ssl.domain_match": "false"}
from awsglue.dynamicframe import DynamicFrame
tedx_dataset_dynamic_frame = DynamicFrame.fromDF(holidays, glueContext, "nested")

glueContext.write_dynamic_frame.from_options(tedx_dataset_dynamic_frame, connection_type="mongodb", connection_options=write_mongo_options)
```

La collezione su MongoDB

Il risultato su MongoDB è una collezione di 7499 festività, strutturate come nell'immagine.

```
_id: ObjectId('680255b2996fa90e5afaf575')
Date : "2023-12-31"
Name : "New Year's Eve"
Type : "['Observance']"
Country Name : "Austria"
Country Code : "AT"
```

Criticità

Abbiamo riscontrato alcuni problemi:

- I dati relativi alle festività provengono da fonti diverse, alcune festività sono ripetute in diversi Paesi, e abbiamo dovuto trovare un formato adeguato da utilizzare per tutti i dati.
- Il consiglio dei video in base alla data attuale e alle festività presenti in quella data sarà poco preciso, in quanto alcune feste riportano la data del 2023 e altre la data del 2025. Questo problema sarà presente soprattutto in occasione di feste la cui data varia di anno in anno, come per esempio Pasqua.

Grazie!