ConnecTEDx

Inspirating ideas, connecting people

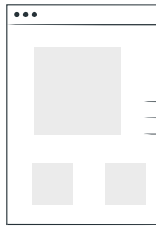
Daniele Gotti - 1079011
Filippo Bolis - 1079493

Job Pyspark – *Load-Data*

```
# JOB ADD WATCH NEXT:
## LETTURA DEI DATI DALLA TABELLA "related_videos.csv" + JOIN AL DATASET MAIN DEI WATCH-NEXT

#   1. Collego l'URI S3 dei dati che si trovano nel bucket
watch_next_dataset_path = "s3://prog-connectedx-data/related_videos.csv"

#   2. Leggo i dati dal csv usando option per gestire la lettura su diverse righe
watch_next_dataset = spark.read \
    .option("header","true") \
    .option("quote", "\"") \
    .option("escape", "\"") \
    .csv(watch_next_dataset_path)

# 3. Seleziona le colonne utili all'app dalla tabella dei video collegati

watch_next_dataset = watch_next_dataset.groupBy(col("id").alias("id_ref")).agg(
    collect_list(struct(
        col("related_id").alias("id"),
        col("title").alias("title"),
        col("presenterDisplayName").alias("presenter")
    )).alias("related_videos")
)

# 4. Svolge il join tra il dataset principale e il nuovo watch_next_dataset

tedx_dataset_main = tedx_dataset_main.join(watch_next_dataset, tedx_dataset_main.id == watch_
    .drop("id_ref")

tedx_dataset_main.printSchema()
```

I dataset, contenenti informazioni sui video TEDx, vengono uniti, arricchiti con dettagli, immagini e video correlati e infine scritti in un database MongoDB per un facile accesso e analisi


```
_id: "528495"
slug : "ryan_panchadsaram_anjali_grover_and_david_biello_an_updated_action_pla…"
speakers : "Ryan Panchadsaram, Anjali Grover and David Biello"
title : "An updated action plan for solving the climate crisis — and a look at …"
url : "https://www.ted.com/talks/ryan_panchadsaram_anjali_grover_and_david_bi…"
description : "When it comes to climate, what are we doing right and where should we …"
duration : "722"
publishedAt : "2024-04-29T13:28:48Z"
url_image : "https://talkstar-photos.s3.amazonaws.com/uploads/585acd28-d676-449a-a0…"
▾ related_videos : Array (6)
  ▾ 0: Object
      id : "83767"
      title : "An action plan for solving the climate crisis"
      presenter : "John Doerr and Ryan Panchadsaram"
  ▸ 1: Object
  ▸ 2: Object
  ▸ 3: Object
  ▸ 4: Object
  ▸ 5: Object
▾ tags : Array (8)
    0: "climate change"
    1: "environment"
```

Codice completo *Click qui*

Job Pyspark – *User-Video*

Il codice utilizza AWS Glue con PySpark per caricare dati da un nuovo file CSV su S3 relativo ai video visti dagli utenti. I dati vengono aggregati per l'id dell'utente, creando una lista di video visti per ciascuno

	id_user,id_video
2	0,52688
3	0,5285
4	0,52778
5	1,52688

```
_id: "2"
video_visti : Array (5)
  0: "52688"
  1: "528495"
  2: "527782"
  3: "526878"
  4: "522927"
```

```python
### AGGREGO I DATI PER id_user
user_video_agg = user_video_dataset.groupBy(col("id_user")).agg(
    collect_list(col("id_video")).alias("video_visti")
)

### SELEZIONO SOLO LE COLONNE id_user E video_visti
user_video_agg = user_video_agg.select(col("id_user").alias("_id"), "video_visti")

### STAMPA LO SCHEMA DEL DATASET AGGREGATO
user_video_agg.printSchema()

### CONVERTE IN DYNAMICFRAME E SCRIVI IN MONGODB
write_mongo_options = {
    "connectionName": "CONNECTEDX_Connection",
    "database": "connectedx_database",
    "collection": "user_video",
    "ssl": "true",
    "ssl.domain_match": "false"
}
```

Codice completo *Click qui*

Job Pyspark – *User-Data*

Codice completo su *Click qui*

```
### PROCESSO I DATI
user_dataset = user_dataset.withColumn("_id", col("id")).drop("id")

### AGGREGO I DATI VIDEO
user_video_agg = user_video_dataset.groupBy(col("id_user")).agg(
    collect_list(col("id_video")).alias("video_visti")
)

user_video_agg = user_video_agg.select(col("id_user").alias("_id"), "video_visti")

### "ESPLODO" VIDEO_VISTI PER FARE IL JOIN CON I TAG
user_video_exploded = user_video_agg.withColumn("id_video", explode(col("video_visti")))

### JOIN TRA TAGS DATASET CON VIDEO DATA "ESPLOSI"
tags_with_user = tags_dataset.join(user_video_exploded, tags_dataset.id == user_video_exploded.id_video, "inner")

### FACCIO LA CONTA DEI TAGS PER USER
tags_count_per_user = tags_with_user.groupBy(user_video_exploded["_id"], tags_dataset["tag"]).agg(
    count("*").alias("tag_count")
)

### RI-AGGREGO IL TUTTO PER PRENDERE I TAGS E LA LORO CONTA
tags_aggregated = tags_count_per_user.groupBy(col("_id")).agg(
    collect_list(struct(col("tag"), col("tag_count").alias("tag_count"))).alias("tags_viewed")
)

### JOIN USER DATA CON VIDEO DATA
user_data_full = user_dataset.join(user_video_agg, "_id", "left")

### JOIN CON AGGREGATED TAGS DATASET
user_data_full = user_data_full.join(tags_aggregated, "_id", "left")
```

Aggrega i video visti dagli utenti, **calcola la frequenza di visualizzazione dei tag** e **unisce queste informazioni** con il dataset degli utenti


```
_id: "2"
name : "Gabriele"
surname : "Mazzoleni"
position : "Mariano (BG)"
coordinateX : "45.640513"
coordinateY : "9.580923"
email : "mazzoleni@gmail.com"
password : "9101"
url_user_img : "https://static-00.ic
▾ video_visti : Array (5)
    0: "52688"
    1: "528495"
    2: "527782"
    3: "526878"
    4: "522927"
▾ tags_viewed : Array (27)
  ▾ 0: Object
      tag : "math"
      tag_count : 1
```

API Eventbrite

Le API di Eventbite che volevamo utilizzare per ottenere i dati degli eventi sono state **disabilitate** nel 2019

Eventbrite API v3

Disabled in December 12, 2019

Abbiamo quindi optato per un operazione di **Scraping del sito**

Per effettuare lo scraping abbiamo:

01

Aperto il sito di Eventbrite

02

Premuto F12 per ispezionare la pagina

03

Scaricato i dati Json inviati degli eventi

04

Dal sito *Stevesie* rimappato e aggregato i dati

05

Copiato i dati in un file Excel eliminando le colonne a noi inutili

06

Convertito il file in csv tramite una macro e caricato su S3

Abbiamo ottenuto 248 dati:

	A	B	C	D	E	F	G	H	I	J	K	L	M
1	name,language,event_url,summary,img_url,start_date,ticket_url,price,end_date,city,longitude,latitude,address												
2	COLTIVARE L'UMANO SEMINARIO RESIDENZIALE,en-us,https://www.eventbrite.it/e/coltivare-lumano-seminario-residenziale-tickets-923899887												
3	Bergamo Outdoor Escape Game: Atalanta the Hero,en-us,https://www.eventbrite.com/e/bergamo-outdoor-escape-game-atalanta-the-hero-ticke												
4	Crema - Workshop Fotografia Ritratto,it-it,https://www.eventbrite.it/e/biglietti-crema-workshop-fotografia-ritratto-778306805607,workshop in for												
5	SUMMER PARTNER,en-us,https://www.eventbrite.com/e/summer-partner-tickets-927770335047,IL PRIMO PARTY DELLA CONSULENZA ITALIANA												
6	Luminous Gardens - Alessandro Martire,en-us,https://www.eventbrite.it/e/luminous-gardens-alessandro-martire-tickets-898651429617,Il pianis												
7	INNERHERO® - Il Viaggio dell'Eroe per il tuo successo nella vita,it-it,https://www.eventbrite.it/e/registrazione-innerhero-il-viaggio-delleroe-per-il-tu												
8	Fusion 360 | Corso Completo,it-it,https://www.eventbrite.it/e/biglietti-fusion-360-corso-completo-938189218207,Il corso affronta le tecniche di r												
9	Super sunday,en-us,https://www.eventbrite.it/e/super-sunday-tickets-878423216487,Evento formativo di 5 ore sul network,,2024-09-15,https://w												
10	Ramada Summer Festival 2024,it-it,https://www.eventbrite.it/e/biglietti-ramada-summer-festival-2024-913846157457,8 serate sotto le stelle 8 ba												
11	Coffee Talk: A colazione con Gea Smith,it-it,https://www.eventbrite.it/e/biglietti-coffee-talk-a-colazione-con-gea-smith-932353082167,Coffee Tal												
12	Shakuntala - Il dramma indiano a Milano,it-it,https://www.eventbrite.it/e/biglietti-shakuntala-il-dramma-indiano-a-milano-95673629401,Shakunta												
13	Il Codice delle Emozioni & del Corpo con il dottor Bradley Nelson,en-us,https://www.eventbrite.it/e/il-codice-delle-emozioni-del-corpo-con-il-dott												
14	DRINK & FLOWERS,en-us,https://www.eventbrite.it/e/drink-flowers-tickets-932701494277,Floral Sips Blooming Moments,https://img.evbuc.com												
15	Welcome Summer! Aperitivo di networking...,it-it,https://www.eventbrite.it/e/biglietti-welcome-summer-aperitivo-di-networking-932402048627,A												
16	Performance Parco Lambro,en-us,https://www.eventbrite.com/e/performance-parco-lambro-tickets-205824876747,performance acrobatica p												
17	CANDLE PAINTING,en-us,https://www.eventbrite.it/e/candle-painting-tickets-932671344097,Color Your Candle Light Your World,https://img.evbu												
18	POT PAINTING,en-us,https://www.eventbrite.it/e/pot-painting-tickets-932665536727,Pot Painting: Where Creativity Blooms,https://img.evbuc.co												
19	Evento Cashflow Club Milano 21 Settembre 2024,it-it,https://www.eventbrite.it/e/biglietti-evento-cashflow-club-milano-21-settembre-2024-92688												
20	TWO DAYS,en-us,https://www.eventbrite.com/e/two-days-tickets-920129701707,Due giorni di formazione per la tua crescita professionale,https:												
21	Techno Beats 001,en-us,https://www.eventbrite.com/e/techno-beats-001-tickets-912410383017,Techno Beats is the first EDM event in Bolzano,h												
22	WORSHOP: Gestione contabile del condominio senza stress,en-us,https://www.eventbrite.it/e/worshop-gestione-contabile-del-condominio-senz												
23	Milano: La Farfalla un aperitivo Appennello,it-it,https://www.eventbrite.it/e/biglietti-milano-la-farfalla-un-aperitivo-appennello-906621528397,Dip												
24	DRINK & BEADING,en-us,https://www.eventbrite.it/e/drink-beading-tickets-932673851597,Where Beads Become Masterpieces,https://img.evbu												

Job Pyspark – *Event-Data*

```python
spark = glueContext.spark_session
job = Job(glueContext)
job.init(args['JOB_NAME'])

### LEGGO IL FILE CSV PER CREARE IL DATASET DI INPUT
events_dataset = spark.read \
    .option("header","true") \
    .option("quote", "\"") \
    .option("escape", "\"") \
    .csv(events_dataset_path)

### STAMPO LO SCHEMA DEL DATASET
events_dataset.printSchema()

### CONVERTE IN DYNAMICFRAME E SCRIVI IN MONGODB
write_mongo_options = {
    "connectionName": "CONNECTEDX_Connection",
    "database": "connectedx_database",
    "collection": "event_data",
    "ssl": "true",
    "ssl.domain_match": "false"
}

### CONVERTE IN DYNAMICFRAME
events_dynamic_frame = DynamicFrame.fromDF(events_dataset, glueContext, "nested")

### SCRIVO IN MONGODB UTILIZZANDO LE OPZIONI SPECIFICATE
glueContext.write_dynamic_frame.from_options(events_dynamic_frame, connection_type="mongodb
```

Dopo aver letto e verificato lo schema del dataset, lo converte in un DynamicFrame e lo scrive in un database MongoDB

```
_id: ObjectId('668d56f5bc57800dff417938')
name : "COLTIVARE L'UMANO SEMINARIO RESIDENZIALE"
language : "en-us"
event_url : "https://www.eventbrite.it/e/coltivare-
summary : "Percorsi di autodeterminazione per un ap
img_url : "https://img.evbuc.com/https%3A%2F%2Fcdn.
start_date : "2024-08-17"
ticket_url : "https://www.eventbrite.com/checkout-e
price : "139.93"
end_date : "2024-08-20"
city : "Barzio"
longitude : "9.453116099999999"
latitude : "45.9424051"
address : "Valsassina 23816 Barzio"
```

Codice completo *Click qui*

Criticità

Dati statici degli eventi	In mancanza dell'API di Eventbrite siamo stati costretti a svolgere uno scraping, tuttavia i dati ottenuti sono utili solo nel breve periodo
Password esposte	Le password degli utenti sono pubbliche nel database MongoDB – S3
Doppio DB	In caso di aggiornamento dei dati su S3, è necessario riavviare ogni volta il job di AWS Glue, il che può causare ritardi.

For more info:
GitHub | Trello

Possibili evoluzioni

Dati statici degli eventi	Se in futuro Eventbrite offrirà API per l'accesso ai dati degli eventi, sarà possibile implementarle. In alternativa, si potrebbe considerare il cambio del provider dei dati.
Password esposte	Si potrebbe rendere l'accesso più sicuro, ad esempio con Cognito

For more info:
GitHub | Trello