TEDx IN PILLOLE

LORENZO MAGNI
1073257

MARTINA RASMO
1072480

MARIANNA ROMELLI
1072382

Job pyspark base

```
## READ WATCH NEXT DATASET
watch_next_dataset_path = "s3://unibg-2023-data-tedx-lm/watch_next_dataset.csv"
watch_next_dataset = spark.read.option("header", "true").csv(watch_next_dataset_path)

watch_next_dataset = watch_next_dataset.drop("url")
watch_next_dataset = watch_next_dataset.groupBy(col("idx").alias("idx_ref")) \
.agg(array_distinct(collect_list("watch_next_idx")).alias("watch_next"))
```

Lettura del dataset dal file CSV e aggregazione dei WatchNext in base agli ID

```
## READ WATCH NEXT DATASET
tedx_dataset_agg_2 = tedx_dataset_agg.join(watch_next_dataset, tedx_dataset_agg._id == watch_next_dataset.idx_ref, "left") \
    .drop("idx_ref") \
    .select(col("_id"), col("*")) \
```

Unione dei due dataset sulla base dell'ID

MONGO DB

Il Job produce un documento contenente i dati del video, una lista di tag e una lista di ID di video consigliati da guardare successivamente


```
{
    "_id": "4adc9fee977fa04c357ed4c9b52aa3cc",
    "main_speaker": "Butterscotch",
    "title": "\"Accept Who I Am\"",
    "details": "Firing off her formidable beatboxing skills, musician Butterscotch serenades...",
    "posted": "Posted Apr 2020",
    "url": "https://www.ted.com/talks/butterscotch_accept_who_i_am",
    "num_views": "0",
    "tags": [
        "TED",
        "talks",
        "live music",
        "music",
        "performance"
    ],
    "watch_next": [
        "edb909effab1896976984a06df06f94e",
        "9f7b1654e792011b7e1c6f4288520226",
        "090a8f3b93c36209b3b3a6a19bfeede5",
        "8e6129177f808f12381d5db92813d878"
    ]
}
```

Criticità

Testing del codice

Non è stato trovato uno strumento semplice per debaggare il codice da errori

Duplicazione dei dati

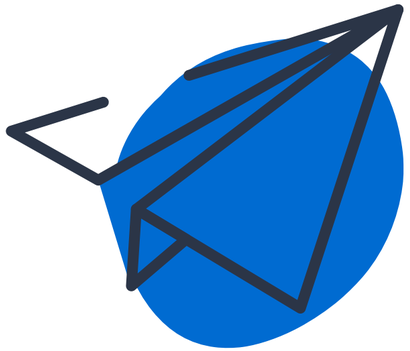
I dati presentavano diversi duplicati dunque abbiamo utilizzato la funzione array_disctinct

Tempo di computazione

Non è stato possibile testare il codice in locale, quindi abbiamo dovuto utilizzare AWS con tempi di computazione lunghi e consumo di crediti

POSSIBILI EVOLUZIONI

Il nostro obbiettivo è sfruttare i dati presenti nel dataset per creare eyelights personalizzati




Link
GitHub
Trello
TED