



# TEDxComments



DWH - AWS Glue



Next



# Table of contents



01 Job PySpark

02 Data & Schema

03 Potential Flaws

04 Future Development



Back



Next

# Job PySpark



```
## READ WATCH NEXT DATASET
wn_dataset_path = "s3://fc-tedx-2024-data/csv/related_videos.csv"
wn_dataset = spark.read.option("header", "true").csv(wn_dataset_path)

wn_real_id = wn_dataset.select(col("internalId").alias("internal"), col("id").alias("real_id")).distinct()

wn_dataset = wn_dataset.join(wn_real_id, wn_dataset.related_id==wn_real_id.internal, "left")

# CREATE THE AGGREGATE MODEL, ADD WATCH NEXT TO TEDX_DATASET
wn_dataset_agg = \
    wn_dataset.groupBy(col("id").alias("id_ref")).agg(collect_list("real_id").alias("related_videos"))

tedx_dataset_agg_img_wn = tedx_dataset_agg_img.join(wn_dataset_agg, tedx_dataset_agg_img._id == \
    wn_dataset_agg.id_ref, "left").drop("id_ref")
```

Sono stati “tradotti” i related\_id in *real\_id* realizzando una tabella intermedia di lookup (*wn\_real\_id*), successivamente sono stati aggregati in una lista ed inseriti nel singolo video

Back



Next

# Job PySpark



*#VIEWS DATASET*

```
wn_dataset_views = wn_dataset.select(col("real_id"), col("viewedCount").alias("views")).distinct()
```

*# CREATE THE AGGREGATE MODEL, ADD VIEWS TO TEDX\_DATASET*

```
tedx_dataset_agg_img_wn_views = tedx_dataset_agg_img_wn.join(wn_dataset_views, \  
    tedx_dataset_agg_img_wn._id == wn_dataset_views.real_id, "left").drop("real_id")
```

Dal dataset creato in precedenza, sono state estratte le visualizzazioni e sono state aggiunte ai documenti dei video tramite i nuovi ID aggiornati.

Back



Next

# Schema

```
_id: "102008"  
slug: "chris_kluwe_how_augmented_reality_will_change_sports_and_build_empathy"  
speakers: "Chris Kluwe"  
title: "How augmented reality will change sports ... and build empathy"  
url: "https://www.ted.com/talks/chris_kluwe_how_augmented_reality_will_chang..."  
description: "Chris Kluwe wants to look into the future of sports and think about ho..."  
duration: "537"  
publishedAt: "2014-05-22T15:00:59Z"  
▼ tags: Array (3)  
  0: "technology"  
  1: "sports"  
  2: "virtual reality"  
img_url: "https://pe.tedcdn.com/images/ted/9b43acda94d7e3743aa877ca5b492862b860f..."  
▼ related_videos: Array (6)  
  0: "33592"  
  1: "60796"  
  2: "77920"  
  3: "16891"  
  4: "79546"  
  5: "117240"  
views: "1361406"
```



Back



Next

# Comments Support

```
_id: "10005"
slug: "isabel_allende_tales_of_passion"
speakers: "Isabel Allende"
title: "Tales of passion"
url: "https://www.ted.com/talks/isabel_allende_tales_of_passion"
description: "Author and activist Isabel Allende discusses women, creativity, the de_"
duration: "1062"
publishedAt: "2008-01-03T06:00:00Z"
tags: Array
img_url: "https://talkstar-photos.s3.amazonaws.com/uploads/b1850a2f-47f3-47ec-bb"
related_videos: Array
views: "5366013"
comments: Object
  info: Array
  disc: Array
  extra: Array
```

Aggiunto supporto per i commenti che avranno la seguente struttura:

```
{
  comment_id: uuid,
  user_id: string,
  timestamp: int,
  title: string,
  body: string,
  upvote: int
}
```

Back



Next

# Materialized View

Forma normale vs Ridondanza? Soluzione: **Materialized View**

```
[
  {
    $lookup: {
      from: "tedx_data",
      localField: "related_videos",
      foreignField: "_id",
      as: "related_videos"
    },
    {
      $project: {
        ...
      }
    }
  }
]
```



Una materialized view in MongoDB è un risultato pre-calcolato di una pipeline di aggregazione che viene memorizzato su disco e può essere letto direttamente. A differenza delle viste standard, le materialized views offrono prestazioni di lettura migliori poiché sono già calcolate e pronte all'uso.

Back



Next

# MongoDB View Creation

Stage 1: \$lookup

```
1 {
2   from: "tedx_data",
3   localField: "related_videos",
4   foreignField: "_id",
5   as: "related_videos"
6 }
```

Output after \$lookup stage (Sample of 10 documents)

```
{
  "related_videos": [
    {
      "_id": "1096241",
      "comments": [
        {
          "text": "that they",
          "into": "..."
        }
      ]
    }
  ]
}
```

Stage 2: \$project

```
1 {
2   "related_videos.slug": false,
3   "related_videos.url": false,
4   "related_videos.comments": false,
5   "related_videos.related_videos": false,
6   "related_videos.tags": false,
7   "related_videos.publishedAt": false,
8   "related_videos.description": false
9 }
```

Output after \$project stage (Sample of 10 documents)

```
{
  "_id": "108710",
  "slug": "rosie_king_how...",
  "speakers": "Rosie King",
  "title": "How autism free...",
  "url": "https://www.ted.c...",
  "description": "People...",
  "duration": "3:59"
}
```


Back





Next






# MongoDB Materialized View Result



▼  tedx\_comments

▼  tedx\_data

▼  **tedx\_watch\_next** ...



```
_id: "138559"
slug: "peter_weinstock_lifelike_simulations_that_make_real_life_surgery_safer"
speakers: "Peter Weinstock"
title: "Lifelike simulations that make real-life surgery safer"
url: "https://www.ted.com/talks/peter_weinstock_lifelike_simulations_that_ma..."
description: "Critical care doctor Peter Weinstock shows how surgical teams are usin..."
duration: "1009"
publishedAt: "2017-03-20T14:57:11Z"
tags: Array (11)
img_url: "https://talkstar-photos.s3.amazonaws.com/uploads/837e0028-e5bb-480e-bf..."
related_videos: Array (6)
  0: Object
    _id: "124490"
    speakers: "Jocelyne Bloch"
    title: "The brain may be able to repair itself -- with help"
    duration: "685"
    img_url: "https://talkstar-photos.s3.amazonaws.com/uploads/4575bffc-5f15-4a82-bc..."
    views: "3785132"
  1: Object
    _id: "150820"
    speakers: "Nadine Hachach-Haram"
    title: "How augmented reality could change the future of surgery"
    duration: "654"
    img_url: "https://talkstar-photos.s3.amazonaws.com/uploads/6e46ae8a-2bf7-49fa-81..."
    views: "1334397"
  2: Object
  3: Object
```



Back



Next

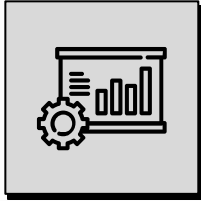


# Potential Flaws



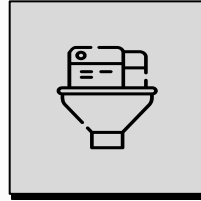
01

02



## Aggiornamento dei Dati

I dati delle visualizzazioni devono essere costantemente aggiornati per essere up-to-date.



## Aggregazione dei Dati

Per evitare ridondanza nei dati è stato necessario aggiungere una chiave referenziale che potrebbe rallentare le query, problema parzialmente risolto dalla view

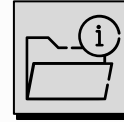
Back



Next



# Future development



Inserimento di commenti	Valutazione dei commenti	Tagging in base al topic
Le persone potranno inserire commenti ad un video	I commenti verranno valutati dagli utenti così da far emergere i più inerenti	I commenti dovranno essere categorizzati in base alla loro tipologia ed argomento

Back



Next



# Thanks!

Colpani Filippo - 1078874  
Foglieni Luca - 1081399



TEDxComments

**Credits:** This presentation template was created by **Slidesgo**, and includes icons by **Flaticon**, and infographics & images by **Freepik**

Back

