Homework 2 - Tecnologie Cloud & Mobile 2020

1040886 - Marco Adobati
1040394 - Valerio Sabato

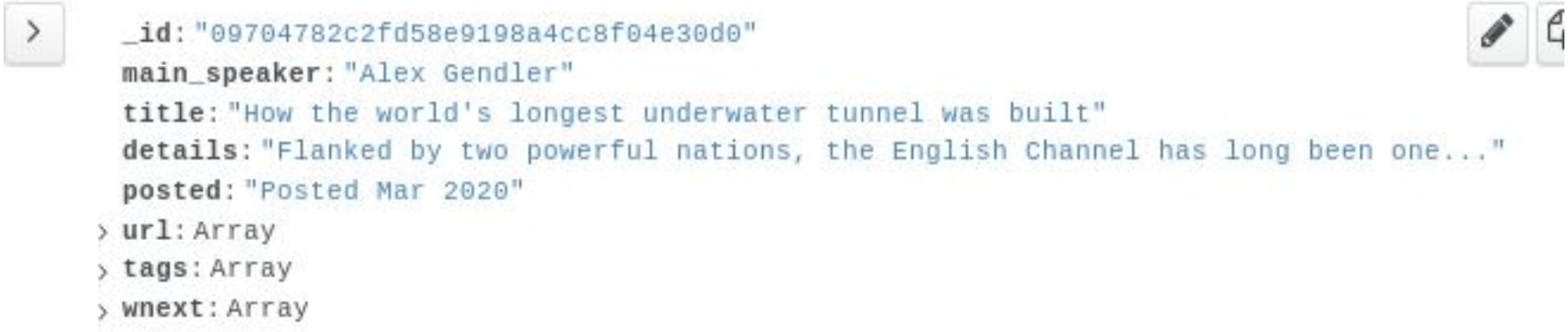
Job TEDX Watch next

Abbiamo creato il job *TEDX_watch_next* in modo da aggiungere allo schema sia gli url dei video successivi oltre ai tag.

1) Per prima cosa abbiamo aggiunto sul bucket dell'S3 il csv contenente idx, *watch_next_idx*, ed *url*
2) Abbiamo raggruppato per *idx* i dati nel *watch_next_dataset*, e abbiamo aggregato i *watch_next_idx* ed i relativi *url*
3) Successivamente, nel job, abbiamo fatto un *LEFTJOIN* tra *tedX_Dataset* e il *Watch_next*, utilizzando come chiave l'index, si è usato il leftjoin in modo che se un video non avesse dei *watch_next* esso avrebbe avuto dei *null*

Jobs TEDX Watch next

Il risultato ottenuto è il seguente:



possiamo notare che ogni video ha una categoria url dove sono contenuti gli url dei video successivi e anche una categoria(*wnext*) che contiene i gli id sempre dei video successivi. Con ciò fornito dal file *watch_next.csv* possiamo fornire all'utente la possibilità di vedere il prossimo talk

Jobs TEDX Watch next

Ad esempio in un talk abbiamo i seguenti *url* e *wnext*:

per URL


```
url: Array
  0: "https://www.ted.com/talks/alex_gendler_epic_engineering_building_the_b..."
  1: "https://www.ted.com/talks/alex_gendler_epic_engineering_building_the_b..."
  2: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
  3: "https://www.ted.com/talks/congrui_jin_what_if_cracks_in_concrete_could..."
  4: "https://www.ted.com/talks/congrui_jin_what_if_cracks_in_concrete_could..."
  5: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
  6: "https://www.ted.com/talks/benedetta_berti_and_evelien_borgman_what_doe..."
  7: "https://www.ted.com/talks/benedetta_berti_and_evelien_borgman_what_doe..."
  8: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
  9: "https://www.ted.com/talks/elon_musk_the_future_we_re_building_and_bori..."
 10: "https://www.ted.com/talks/elon_musk_the_future_we_re_building_and_bori..."
 11: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
 12: "https://www.ted.com/talks/alan_eustace_i_leapt_from_the_stratosphere_h..."
 13: "https://www.ted.com/talks/alan_eustace_i_leapt_from_the_stratosphere_h..."
 14: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
 15: "https://www.ted.com/talks/mick_mountz_what_happens_inside_those_massiv..."
 16: "https://www.ted.com/talks/mick_mountz_what_happens_inside_those_massiv..."
 17: "https://www.ted.com/session/new?context=ted.www%2Fwatch-later"
```


per Tag


```
wnext: Array
  0: "322fec9c9085abe109b66c421847b5b0"
  1: "322fec9c9085abe109b66c421847b5b0"
  2: "9f7b1654e792011b7e1c6f4288520226"
  3: "4cb3c8986dbf3eef7a997fc89660a487"
  4: "4cb3c8986dbf3eef7a997fc89660a487"
  5: "9f7b1654e792011b7e1c6f4288520226"
  6: "e8274bf7ed4f4e9bf1c024d18445822a"
  7: "e8274bf7ed4f4e9bf1c024d18445822a"
  8: "9f7b1654e792011b7e1c6f4288520226"
  9: "c74260d99f67a5a2660221976aedd8ef"
 10: "c74260d99f67a5a2660221976aedd8ef"
 11: "9f7b1654e792011b7e1c6f4288520226"
 12: "70654f763a4f087262c97304209434f0"
 13: "70654f763a4f087262c97304209434f0"
 14: "9f7b1654e792011b7e1c6f4288520226"
 15: "9f7bf917901864c7a143877c862b7cfd"
 16: "9f7bf917901864c7a143877c862b7cfd"
 17: "9f7b1654e792011b7e1c6f4288520226"
```

Modifica job per cambio formato data

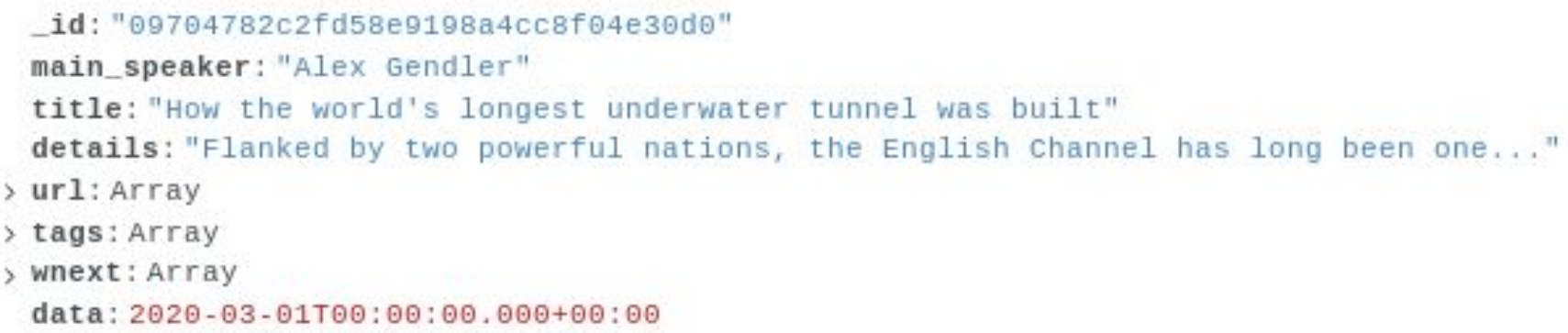
Modificando il job precedente abbiamo convertito l'attributo *posted* in formato *unix_timestamp* in quanto le nostre ***API*** devono mostrare i video tra due date, in questo modo abbiamo ottenuto un attributo che rappresenta la data ordinabile

Job change data format

1) Creando delle colonne di supporto nel dataframe abbiamo inserito il campo *data* di tipo *unix_timestamp*
2) Per fare questa operazione abbiamo utilizzato alcune funzioni di pyspark tra cui: *substring*, *col*, *lit*, e *unix_timestamp*

Job change data format

Infine abbiamo ottenuto il seguente schema all'interno della nostra collezione:



In modo da dare poi la possibilità ai nostri utenti di vedere i talk tra due particolari date.

Criticità tecniche

1. Alcuni talk sono formattati male all’interno del dataset, per cui alcuni campi tra non sono correttamente visualizzati

Possibili evoluzioni

Una naturale evoluzione di ciò sarebbe lo sviluppo delle **API** decise nell'homework precedente sfruttando quanto è stato fornito dal job qua costruito, ovvero gli *id* ed *url* per il watch next talk, e la data in *unix_timestamp*