



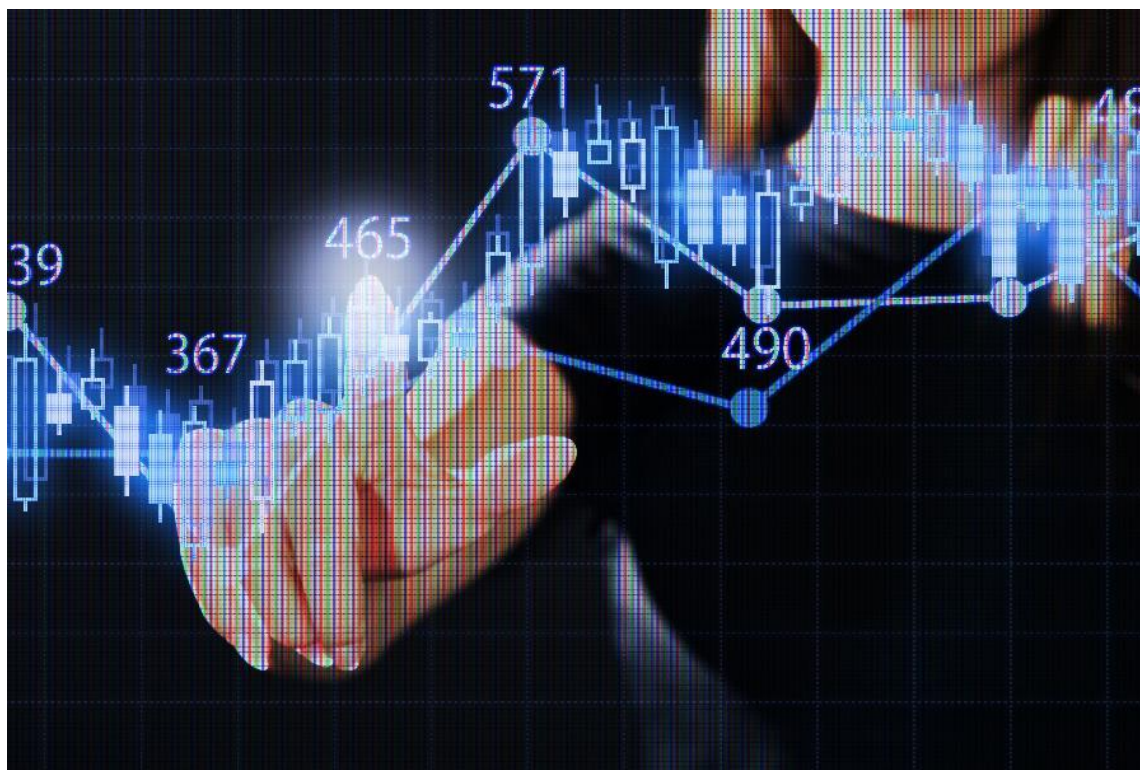
# TEDTOK

PARTE 2 – DATABASE WATCH-NEXT SU MONGODB

# WATCH\_NEXT DATASET

- `'watch_next_dataset.csv'` contiene i dati dei video consigliati per la visione dopo il video corrente
  1. **COSA FARNE** : come possiamo sfruttare queste informazioni?
  2. **COME FARLO**: come aggiungiamo queste informazioni al nostro database in modo intuitivo e leggibile?


## ‘COSA FARNE’ : SOLUZIONE



Usare queste informazioni come ‘**prima scrematura**’ per la ricerca di video

- L'algoritmo di ricerca, o l'IA, per trovare il prossimo video da proporre all'utente potrà ottenere una **prima pool di candidati** direttamente dal database, per poi scegliere il video più adatto a seconda delle preferenze dell'utente

## ‘COME FARLO’ : SOLUZIONE

An abstract graphic on the left side of the slide. It features a dark blue background with a glowing, interconnected network of points and lines, resembling a data visualization or a molecular structure. The lines and points are a lighter shade of blue, creating a sense of depth and complexity.

**Nuovo job** su AWS Glue in PySpark che, in prima battuta, ricalca quello fatto a lezione, costruendo il dataset congiunto dei video e dei loro tag

A questo, però, aggiungiamo anche l'import di *watch\_next\_dataset.csv*, e **puliamo i dati** (il csv contiene tuple duplicate e numerosi rimandi al link *watch-later*)

Infine, con un'operazione di join creiamo **un unico dataset** in cui abbiamo **tutti i video e relativi URL catalogati per id, ed ad ognuno di essi associati i propri tag e video consigliati**





# CODICE SU:

GITHUB\_REPOSITORY