# **Big Data Challenge Questions**

## **YouTube Challenge**

1. Display the first 10 records in the US dataset?
2. How many records are in CA dataset?
3. How many records are in US dataset?
4. How many records in CA dataset do not have any tags assigned?
5. Between CA and US datasets which one has the most releases in 2018?
6. Number of total records in both US and CA datasets that have the title:
   * A Very Cool Christmas - Movie Review - brutalmoose
7. How many records in US dataset have both Comment and Rating disabled?
8. How many records in US dataset have more Dislikes than Likes?
9. In CA dataset are there any records that have Comments\_Disabled set to true, but have Comments?
10. In US dataset are there any records that the total of Dislikes and Likes are equal to Views count?
11. In CA dataset, how many records have minimum number of Likes?
    * Note: we are not interested in equal ones.
12. In CA dataset, Publish\_Time includes date and time (2017-11-13T17:13:01.000Z). We want to separate date and time for better readability and future processing, and we want to convert the date format to mm/dd/yyyy. Create a new dataset that includes 4 fields: Video\_ID, Title, PublishedDate and PublishedTime.

* ECL standard library has a ConvertDateFormat() function that can be used to reformat date: STRING ConvertDateFormat(STRING date\_text, VARSTRING from\_format='%m/%d/%Y', VARSTRING to\_format='%Y%m%d')
* <https://github.com/hpcc-systems/HPCC-Platform/blob/master/ecllibrary/std/Date.ecl>
* Name your new dataset NewDS and output the first 25 records.
* The result dataset should look like following:

|  |  |  |  |
| --- | --- | --- | --- |
| Video\_ID | Title | PublishedDate | PublishedTime |
| CpU72eM8vCo | Operation: Dry Tank | 11/13/2017 | 07:13:54 |

1. How many videos have the most tags in US?
   * In ECL CountWords function is used to count the number of words using a separator

* STD.Str.CountWords(source, separator)
* <https://hpccsystems.com/training/documentation/standard-libraryreference/html/CountWords.html>
* Note: Keep in mind, that a video can be trended in more than one day, which will result in duplicated Video\_ID and Title.

1. In CA dataset how many records have the same dates (day, month, year) for Publish\_Time and Trending\_Date?

* Note: Keep in mind that Publish\_Time and Trending\_Date do not have the same format. For comparison you need the exact same format on both fields.

1. Now we want to know "total releases per year" in US. Using the NewDS dataset you just created, or finding another solution, show number of releases per year. Looking at the result, how many years is included in this dataset?
2. From "total releases per year" for US, what is the minimum release total and what year it belongs to?
3. From "total releases per year" for US, what is the maximum release total and what year it belongs to?
4. Find the number of videos that went viral on the same Trending\_Date in both US and CA.
5. Which Channel\_Title released the most videos in year 2017 in CA?
6. How many videos have different Channel\_Title between US and CA.

* Note: Since we are looking at videos, remember to eliminate duplicates.