Big Data
Small Project

Overview

In this project I'm answering the question of " The percentage of likes for each speaker in Ted " I used a **dataset** from **kaggle** which has nearly 8000 lines of data. With this data I calculated the sum of views and likes for each speaker, or author as it's named in the code, and calculated the percentage of likes of all the views he got.

The dataset

The dataset is about various ted speeches from all over the world composed from

Title > Auther / Speaker > Date > Number of views > Number of likes > Link of the speech

Technologies

I used hadoop's mapreduce model to calculate the sum and the percentage.

The code

Mapper :

```
import sys

for line in sys.stdin:
    data = line.strip().split(",")
    if len(data) == 6:
        id_x, auth, day, views, likes,something = data
        print ("{0},{1},{2}".format(auth, views,likes))
```

Reducer

```python
#!/usr/bin/python
import sys

views = 0
likes = 0
percentage = 0
oldId = None


for line in sys.stdin:
    data_mapped = line.strip().split(",")
    if len(data_mapped) != 3:
        continue

    thisId, view, like = data_mapped

    if oldId and oldId != thisId:
        print (oldId, ",", views,",",likes,",",percentage)
        oldId = thisId
        views = 0
        likes = 0
        percentage = 0


    oldId = thisId
    views += float(view)
    likes+= float(like)
    percentage = (likes*100)/views

if oldId != None:
    print (oldId, ",", views,",",likes,",",percentage)
```

Conclusion

As we see in the results file, most of the videos have an average of 3% likes of the total views.