NTNU

TDT4225 Store, distribuerte datamengder

Assignment #4

Øving 4

Medvirkende:
Brede Yabo Kristensen

March 12, 2019

1 RDD API Tasks

1.1 Task 1

The files we are reading are rows with columns seperated by , .

I first split each column using , as a deliminator. This is done by using the map() function as it applies a function to all of the rows. The result is yet another RDD. When splitting the rows we get an array for each row, we then want the fourth column as it is the genre column. We could of used flatmap(), but we want to an rdd that has the same number of rows.

We then use the function distinct() to remove all duplicate from the rdd. Finally we count the number of rows using the count() function, which returns number.

1.2 Task 2

We start the same way as last time, only now we get the fourth column from the artist.csv file (year). We also make sure that the rows are converted to integers as we will compare them later. We then reduce the rdd using the reduce() function. We want to return the oldest artist and do this by comparing all the rows.

Finally we print the oldest artist birth date.

1.3 Task 3