Trending YouTube Video Analysis

Sai Krishna Veeravelly(S544724)
Sai Krishna Vuppala(S545040)
Gayatri Devabhaktuni(S545030)
Laxman kumar Alugubelli (S545015)
Pavan Kumar Koppuravuri(S545220)

Northwest Missouri State University, Maryville MO 64468, USA

1 Introduction

This dataset includes several months of data on daily basis of the uploaded YouTube videos. This is about the data of different countries all over the world with up to 200 listed trending videos per day. The data file includes the video title, channel title,publish time, tags, views, likes and dislikes, description, and comment count. The data also includes a coulumn of category id to differentiate the videos. In this Project We use variety of tools, to analyze users interactions, to select the year's top-trending videos based on number of views, shares, comments and likes from different categories like Music videos, celebrity and/or reality TV performances, and the viral videos which are among the top performers on the YouTube trending list.we can also analyze channel's performance analytics.

2 5V's respect to project

2.1 Variety:

The data used is of four different countries. The data gives the complete youtube videos of different channels from all the four countries.

2.2 Velocity:

There are about 200 records of data which is recorded everyday over the period of several months. This vast data that streams shows the power of velocity in data processing technique.

2.3 Volume:

The total volume of the data contains around 140000 records from four different countries. Huge volume of data is streamed everyday from different channels.

2.4 Value:

We can generate a statistical form of data from the big dataset and also we can bring desired output from the dataset which we have taken. The output generated gives clear analysis of the youtube videos with respect to the goal in accordings to the user.

2.5 Veracity:

The dataset is extracted from a reputed organization,it is trustworthy and we can say that this data is accurate. The accuracy in the data gives the desired output, hence veracity plays a vital role in utilizing the data.

3 Goals

We have set following goals for the dataset which we choose

1. The goal is to know the average likes a channel got in all 4 countries.(Sai krishna Vuppala)
2. The goal is to know the average number of views a channel got in all 4 countries.(Sai krishna Vuppala)
3. The goal is to determine the total number of comments received by a channel for all videos in each of the four countries.(Gayatri Devabhaktuni)
4. The goal is to know the total number of likes received by all videos in each channel for the four countries.(Gayatri Devabhaktuni)
5. The goal is to know the total dislikes received by all videos in each channel for the four countries.(Sai Krishna Veeravelly)
6. The goal is to know the total number of views for each channel got in all 4 countries.(Sai Krishna Veeravelly)
7. The goal is to know the number of videos uploaded by each channel.(Pavan Kumar Koppuravuri)
8. The goal is to know the number of likes a channel got on a particular day.(Pavan Kumar Koppuravuri)
9. The goal is to know the total views on YouTube on a particular day.(Laxman kumar Alugubelli)
10. The goal is to know the total views a channel got on a particular day.(Laxman kumar Alugubelli)

4 Tools and Technologies

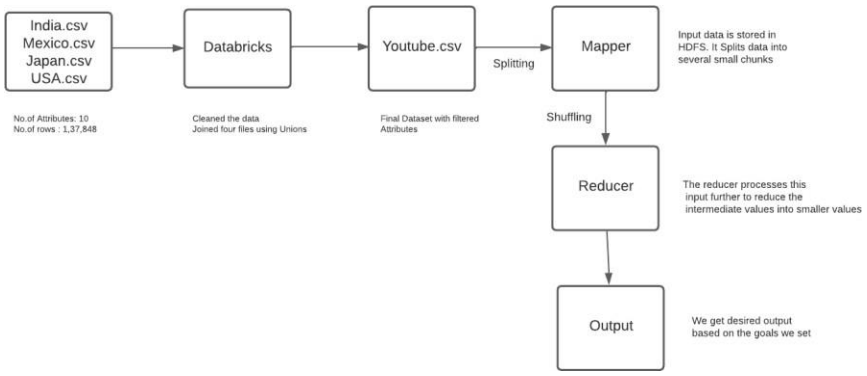
4.1 Data bricks

Data bricks is an unified and open source platform which helps to run interactive and scheduled data analysis. It is also extremely flexible and easy to get started on, making distributed analytics much easier to use. The initial dataset we took has huge amount of data from multiple countries. The data is segregated with required fields of four countries using groups/joins in the data bricks. The filtered data is used as a final dataset for processing the project.

4.2 Hadoop

In Hadoop, we used Map Reduce technique, which is a Hadoop framework for writing applications that can process vast amounts of data on large clusters. There are two primary tasks in Map Reduce: Map, reduce. In the map job, we split the input dataset into chunks. Map task processes these chunks in parallel. We used map-reduce techniques to analyze the YouTube dataset of four countries to get the different outputs based on the goals we set.

5 Block Diagram



6 Implementation

6.1 Map reduce

Steps for MapReduce Program Execution

1. Download the MapReduceDemo.zip file. To download, click on code button and select 'Download ZIP'(https://github.com/bandiajay/MapReduce—zip file can be downloaded from here)
2. Extract the MapReduceDemo.zip file and copy the MapReduceDemo folder to the Eclipse-workspace (You can find Eclipse-workspace in C:¡Your SID¿)(if your system doesn't have Eclipse IDE follow the steps for Eclipse Installation given below)
3. In the Eclipse, File-Open Projects from File System Then, click on "Directory" in "Import Projects from File System or Archive" window, then browse the project folder from Eclipse workspace.
4. Change the program arguments by following below steps

(a) Right click on the project select Properties
(b) Select Run/Debug settings in the ”Properties for MapReduceDemo” window, and click ”New” button. Then, select Java Application, click OK. Give name as ”ViewCount”. Under the ’Main’ tab, and in the ”Main class:” search for ”ViewCount-mapreducedemo”, select it and click OK.
(c) Select the ViewCount and click on ’Edit’ button
(d) Select the ”Arguments” tab, and in the Program Arguments give path for input file and output file as
  – text C:-Users-YourS-ID-eclipse-workspace-MapReduceDemo-data-input-rawViews.txt C:-Users-YourS-ID-eclipse-workspace-MapReduceDemo-data-output and Click Ok.,

5. Delete the output Folder from the MapReduceDemo folder (Location: C:/Users - Your SID/eclipse-workspace/MapReduceDemo/data)
6. To check for output go to data folder inside project an output folder will be generated after successfully executing program
7. Open part-r-00000 file inside output folder in your project where the output of the program is stored. Run the ViewCount.java file to get desired output

6.2 Steps for Eclipse Installation

1. Eclipse: https://www.eclipse.org/
2. Download and install the latest version of Eclipse installer.
3. Get the Installer for your platform (student/faculty laptops are windows-x64).
4. Install Eclipse IDE for Enterprise Java and Web Developers from the installer menu.

6.3 Steps to add winutils.exe

1. Download winutils.exe and Hadoop.dll from https://github.com/steveloughran/winutils/tree /master /hadoop-2.7.1/bin
2. Create a folder hadoop in windows C in that create bin folder and place winutils.exe file and Hadoop.dll in bin folder
3. Create a new system variable give name as Hadoop-Home and give variable value as C:/hadoop and click ok
4. Click on path in system variables and click on edit option the click on new and give C:/hadoop/bin the click ok
5. Now close Exlipse IDE once and open and run ViewCount.java

6.4 DataBricks

1. Have an account in Data bricks community edition, after logging in you need to go to workspace and create a new one.
2. To run the workspace successfully we need to have cluster attached and running.
3. After attaching the cluster we can run the code in the workspace and get the desired output

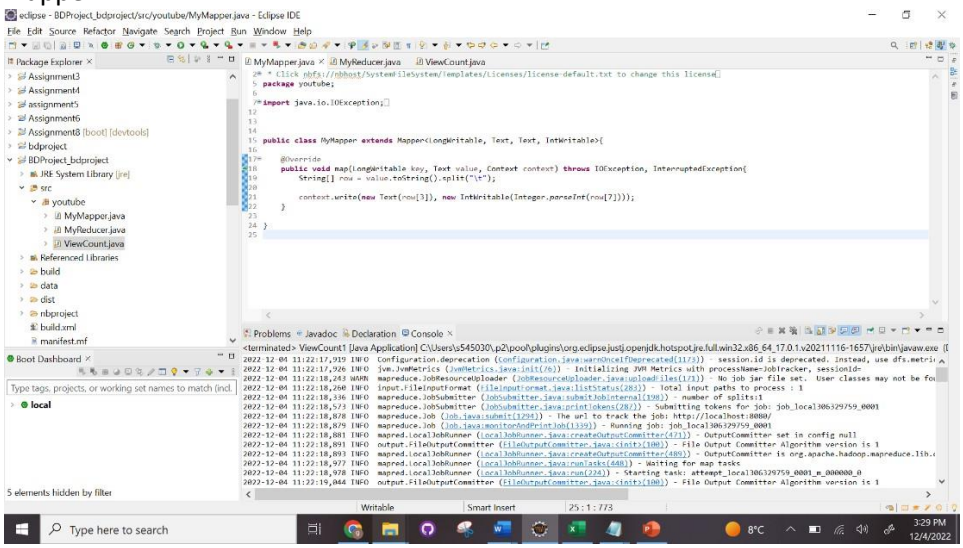
7 Results

7.1 Goal 1

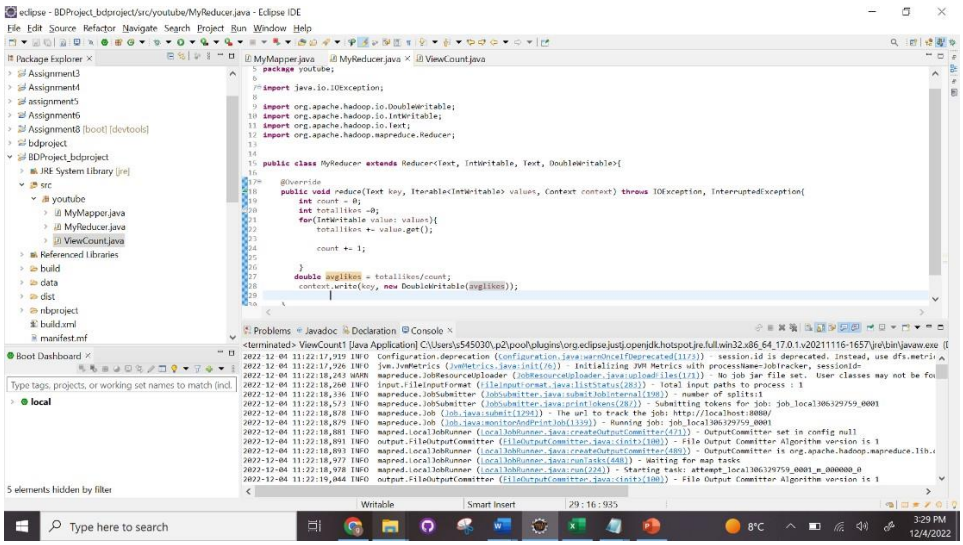
The goal is to know the average likes a channel got in all 4 countries.

Here, we are able to get the average number of likes a channel got for the 4 countries we considered. The time taken to achieve this goal is 30 minutes. To attain this goal we combined all the data from 4 countries and cleaned it using Databricks. After cleaning we downloaded the data in .csv format and changed it to a .txt file, as it is needed for the map-reduce function in Hadoop.

Mapper:

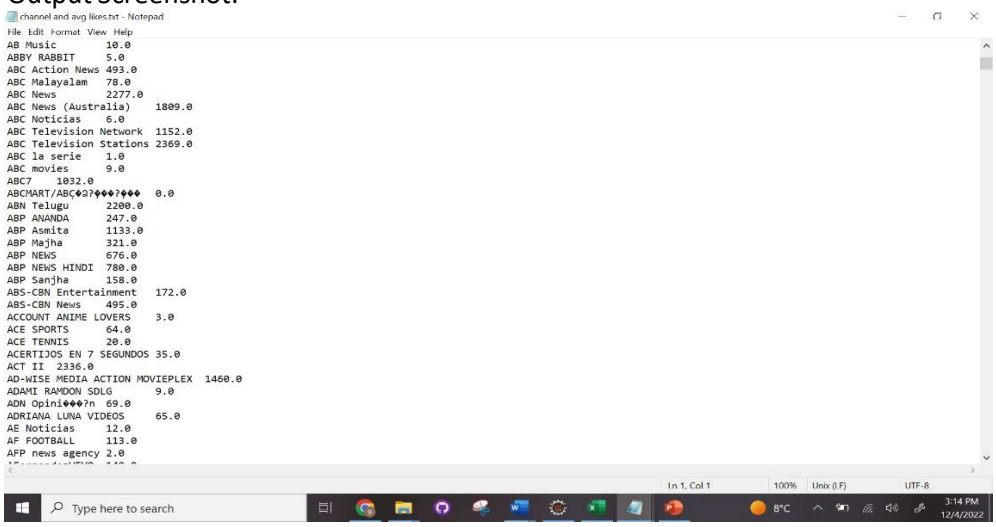


Reducer:



6

Output Screenshot:

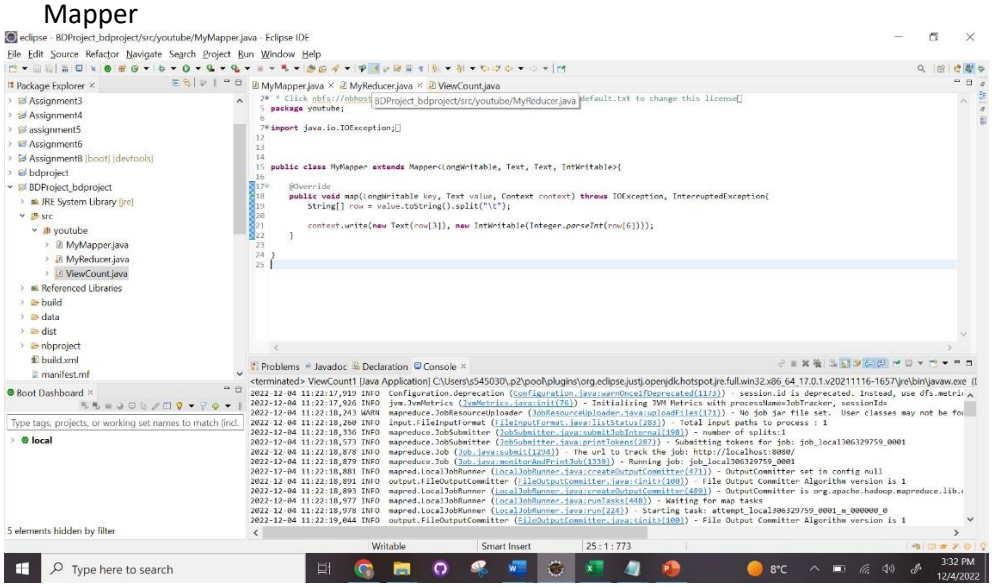


7.2 Goal 2

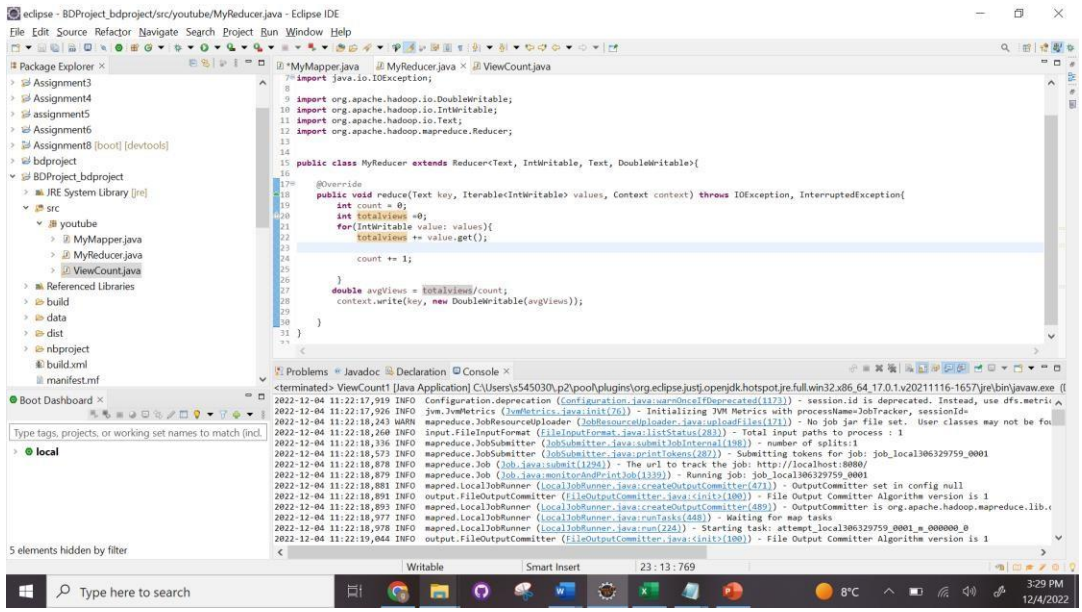
The goal is to know the average number of views a channel got in all 4 countries.

We got the desired output for this goal by applying map reduce techniques on channel row and views row, we have written mapper and reducer class in a way that it should display the average views for each channel in all 4 countries. The time taken to achieve this goal is 15 minutes.

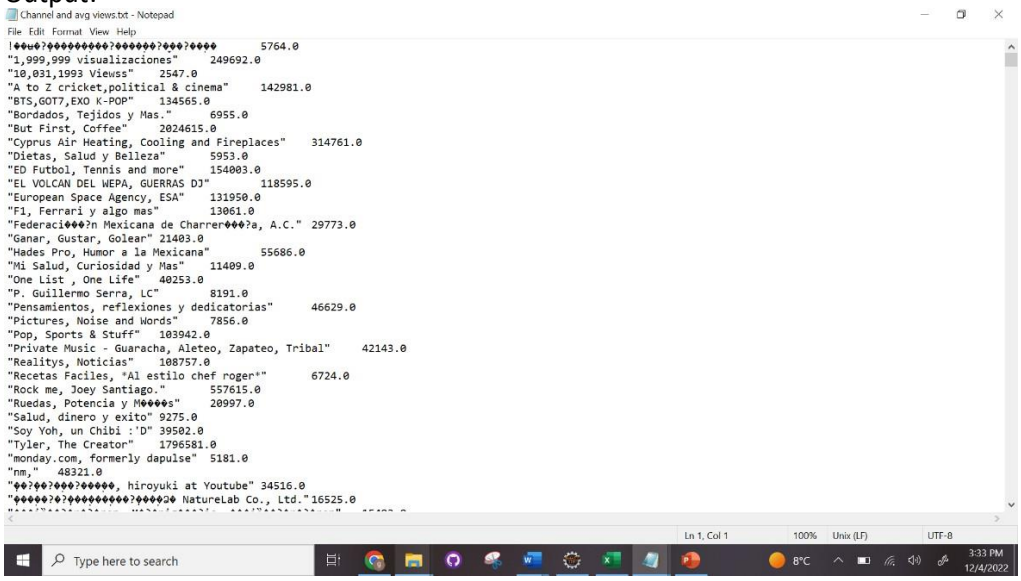
Mapper



Reducer



Output:



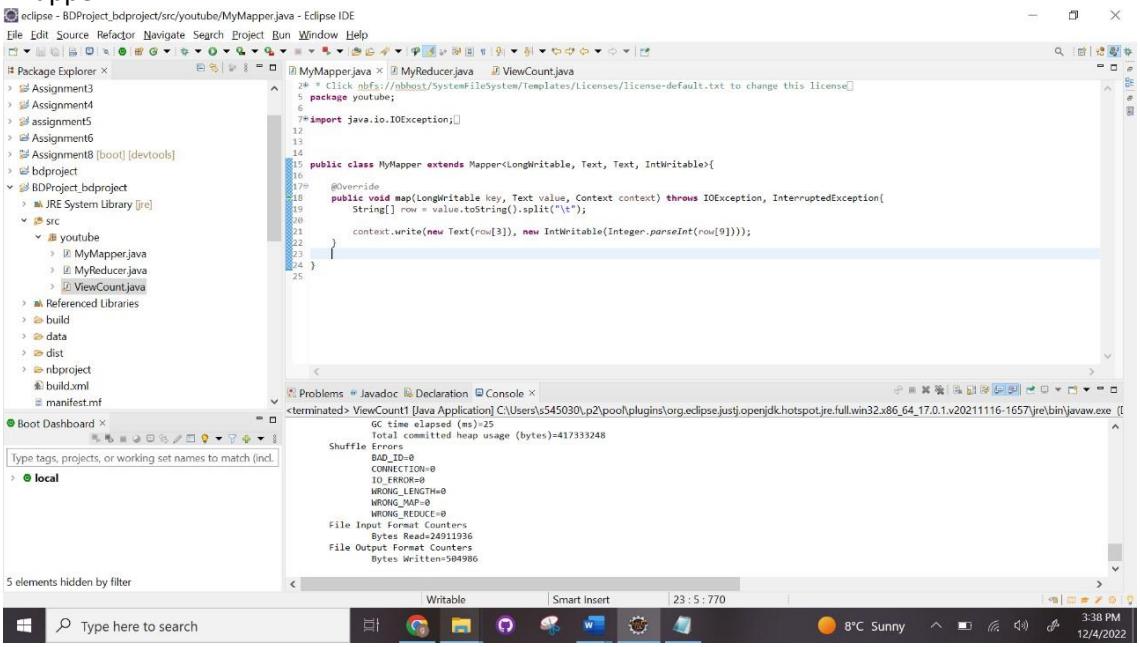
7.3 Goal 3

The goal is to determine the total number of comments received by a channel for all videos in each of the four countries.

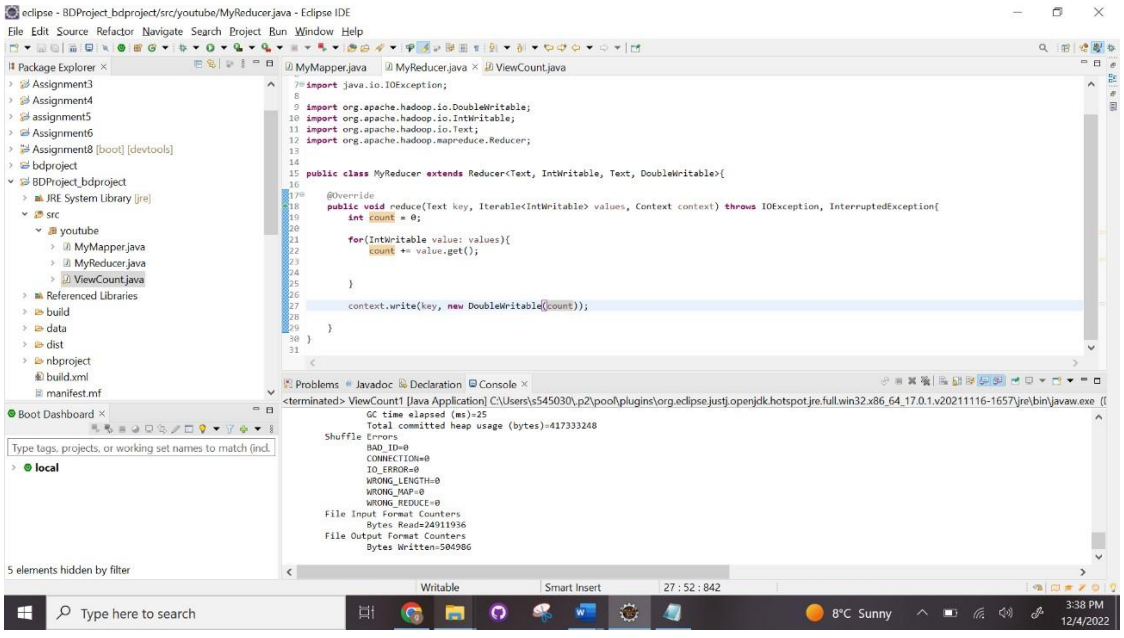
We got the desired output for this goal by applying map reduce techniques on channel row and comments row, we have written mapper and reducer classes in such a way that it should display the number of comments received by a channel for all videos they uploaded. The time taken to achieve this goal is 15 minutes.

8

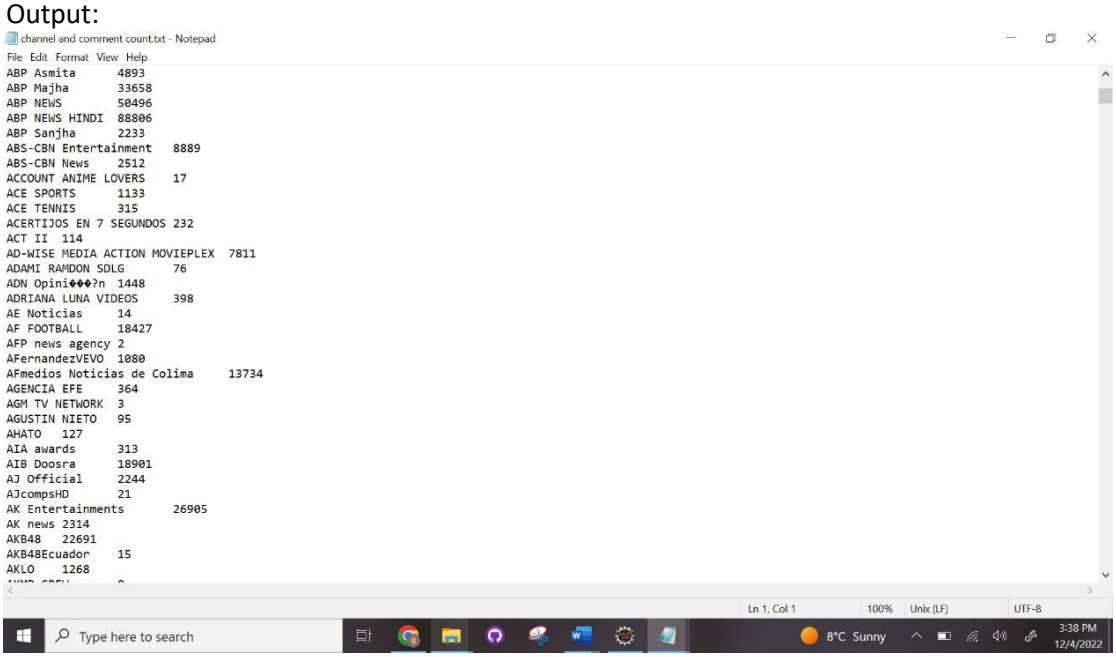
Mapper



Reducer



Output:

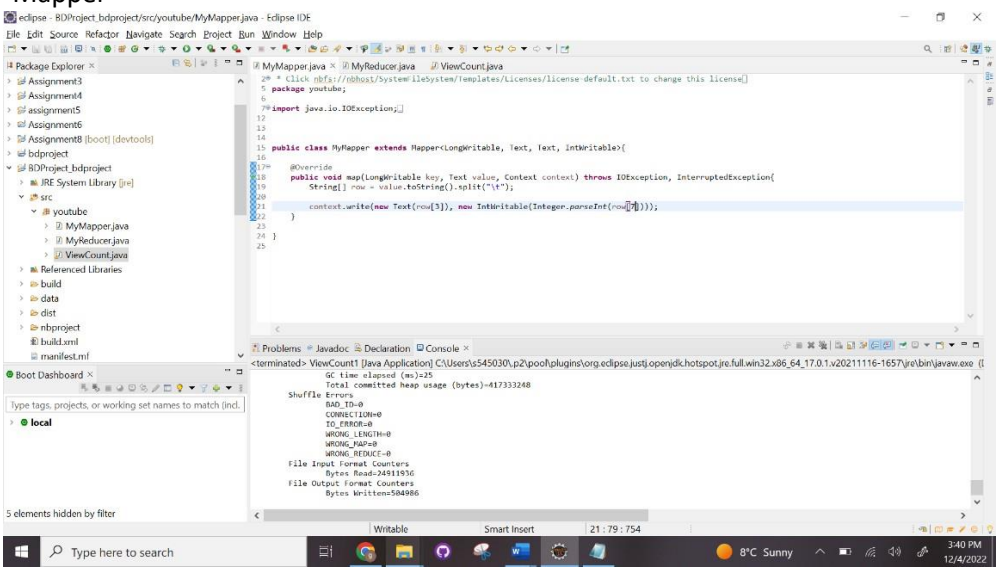


7.4 Goal 4

The goal is to know the total number of likes received by all videos in each channel for the four countries.

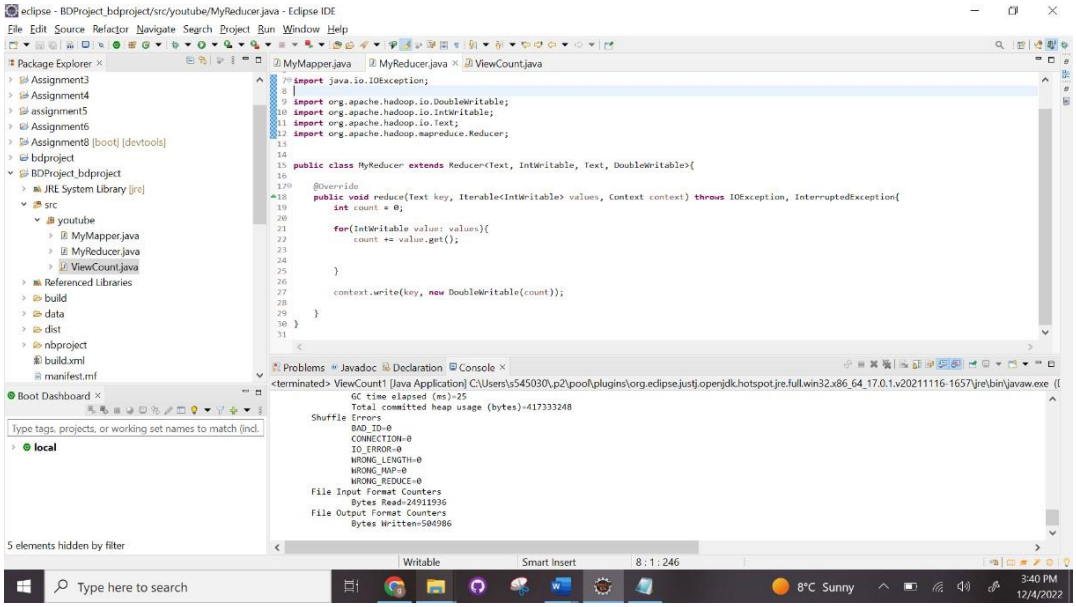
We got the desired output for this goal by applying map reduce techniques on channel row and likes row, we have written mapper and reducer classes in such a way that it should display the total number of likes received by a channel for all videos they uploaded.The time taken to achieve this goal is 15 minutes.

Mapper

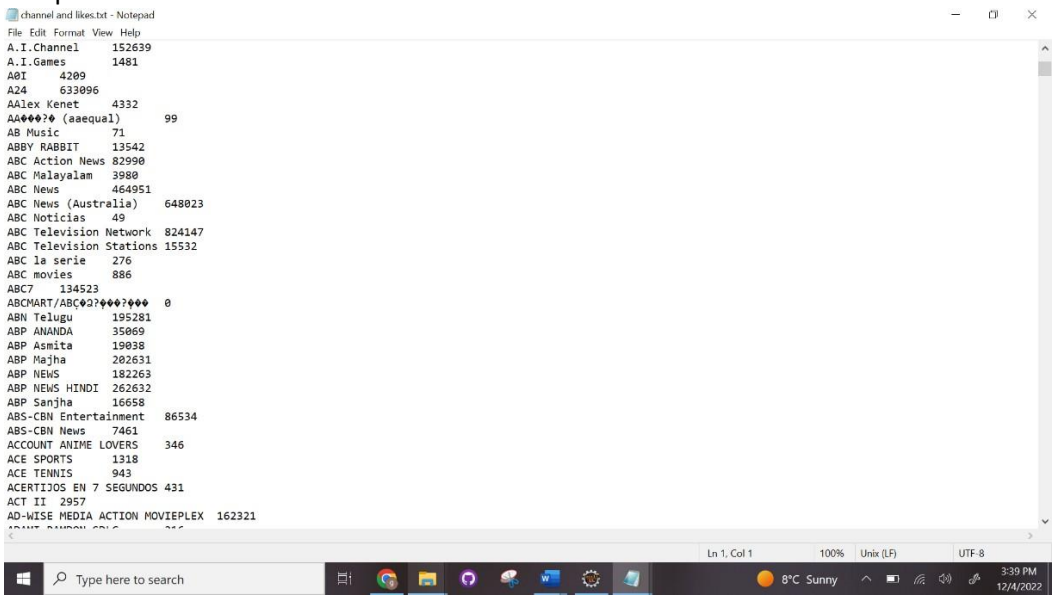


1

Reducer



Output:

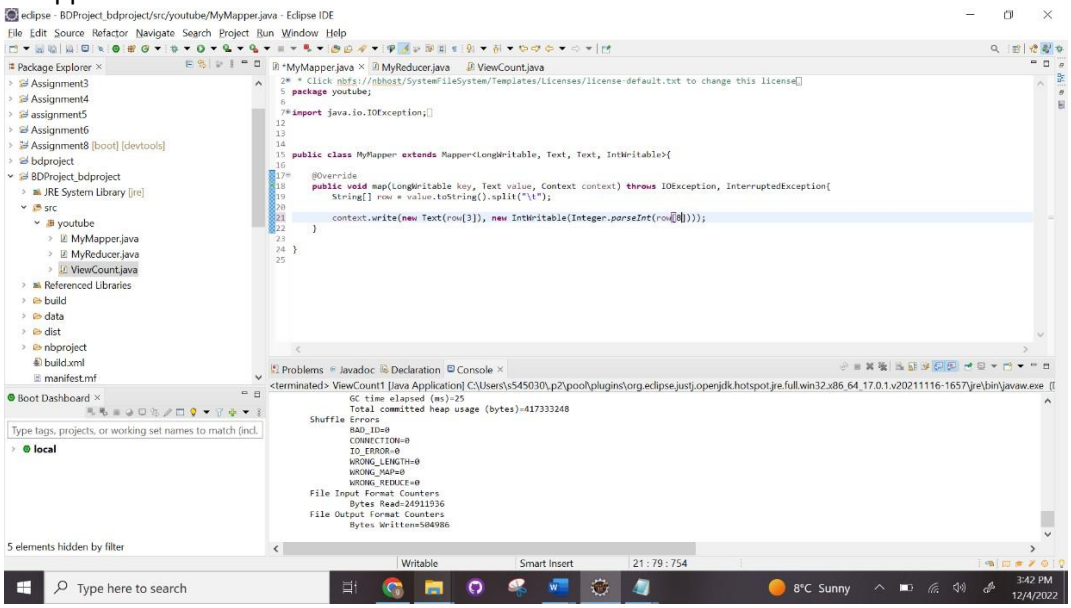


7.5 Goal 5

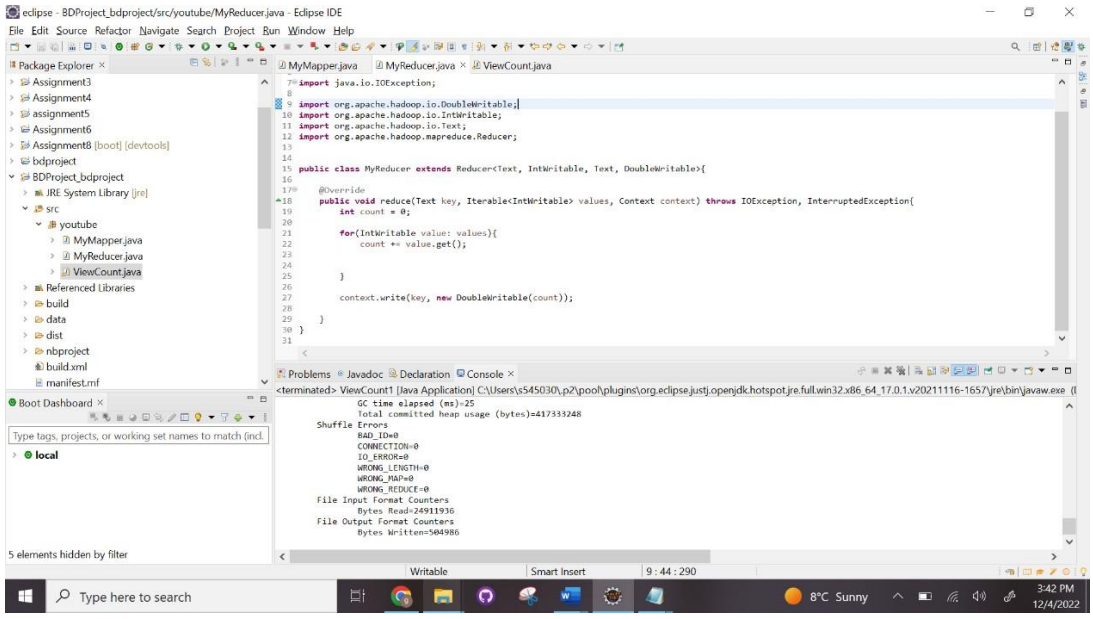
The goal is to know the total dislikes received by all videos in each channel for the four countries.

We got the desired output for this goal by applying map reduce techniques on channel row and dislikes row, we have written mapper and reducer classes in such a way that it should display the total number of dislikes received by a channel for all videos they uploaded.The time taken to achieve this goal is 15 minutes.

Mapper:

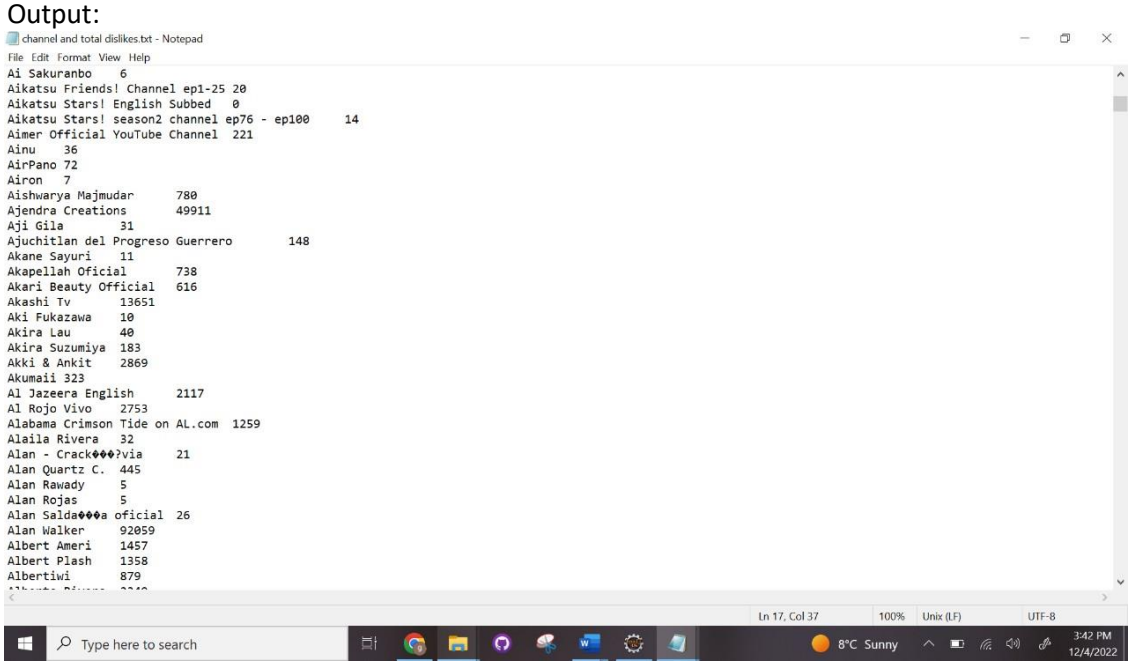


Reducer:



1

Output:

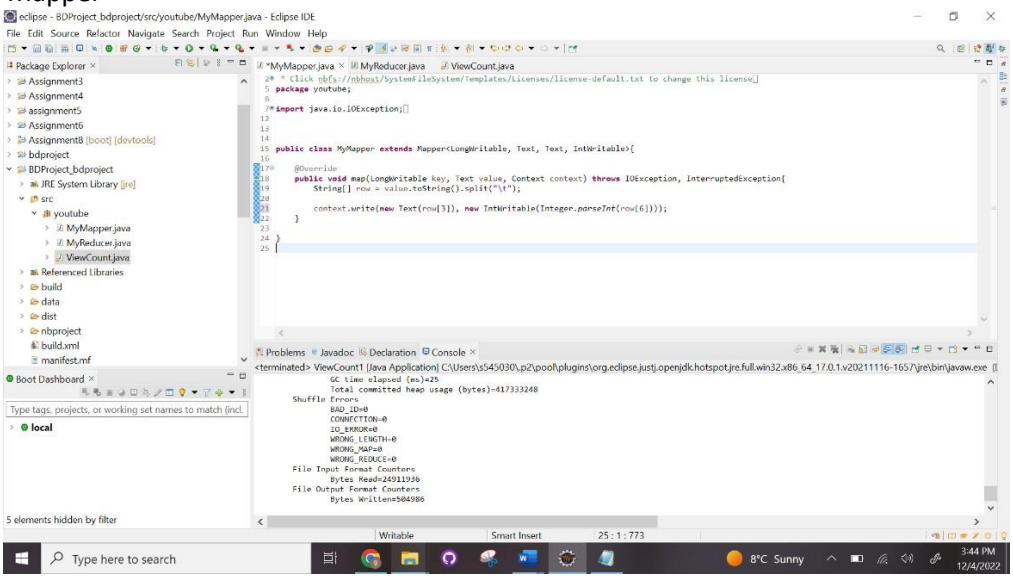


7.6 Goal 6

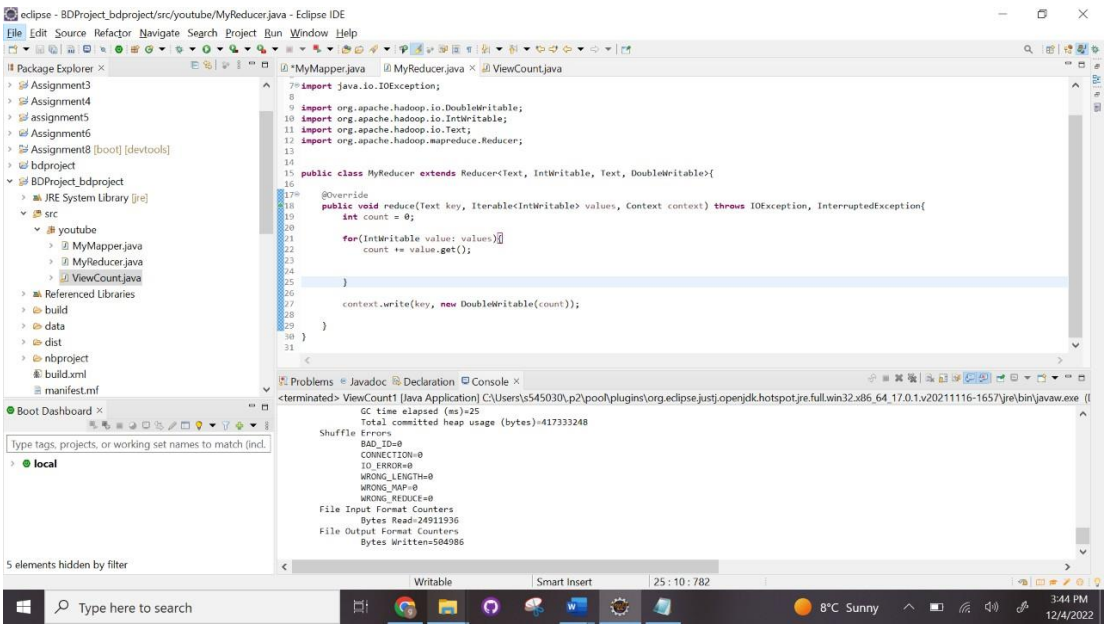
The goal is to know the total number of views for each channel got in all 4 countries.

We got the desired output for this goal by applying map reduce techniques on channel row and views row, we have written mapper and reducer classes in such a way that it should display the total number of views received by a channel for all videos they uploaded.The time taken to achieve this goal is 15 minutes.

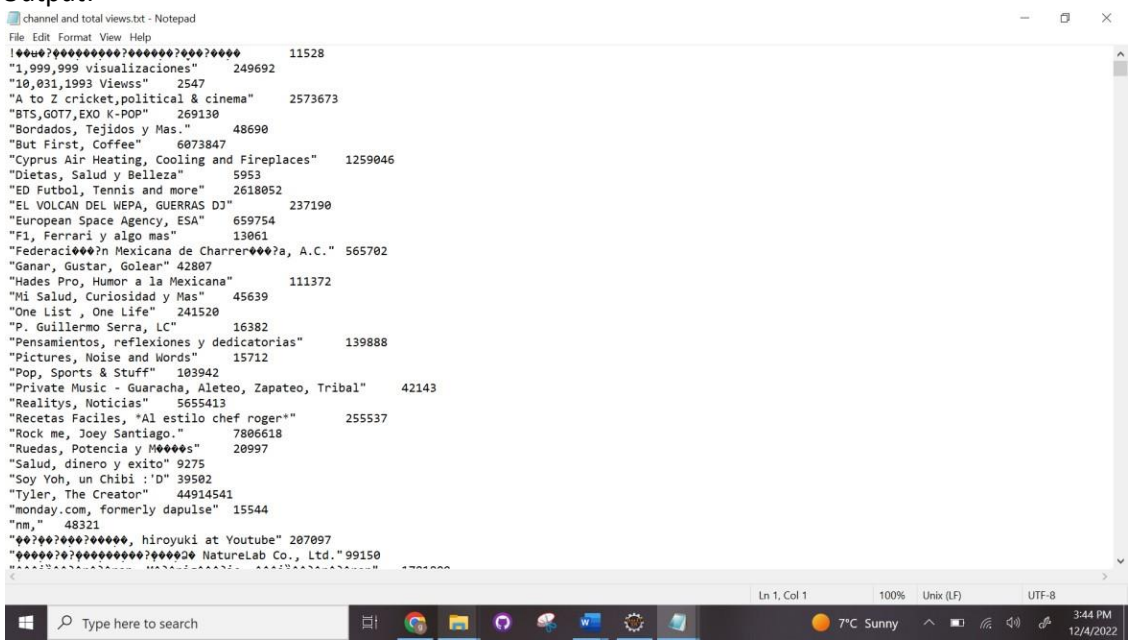
Mapper



Reducer:



Output:



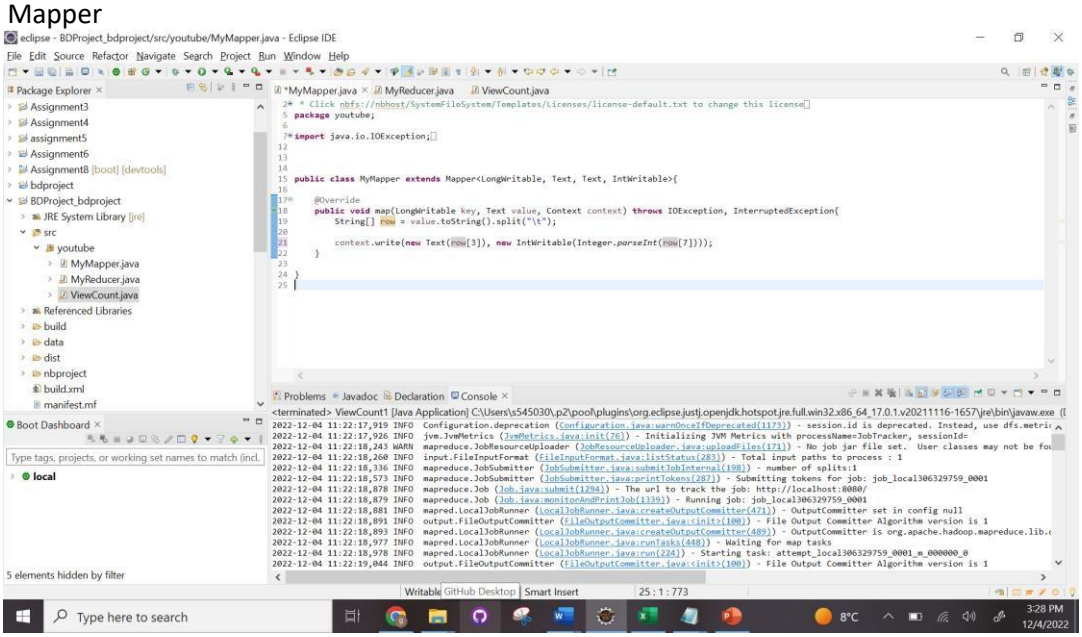
7.7 Goal 7

The goal is to know the number of videos uploaded by each channel.

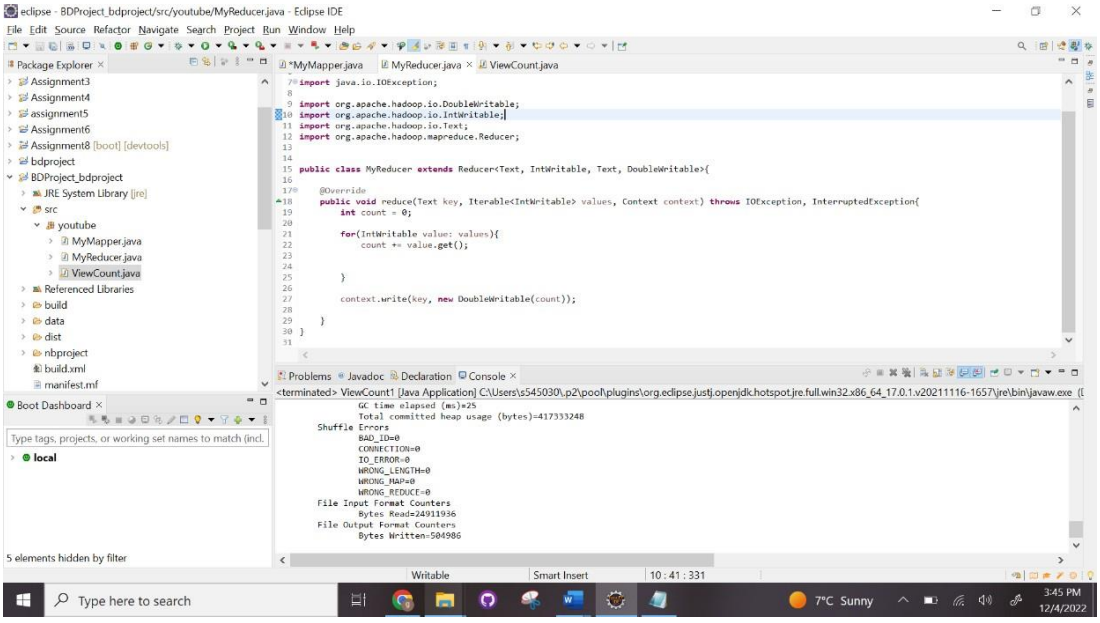
We got the desired output for this goal by applying map reduce techniques on channel row, we have written mapper and reducer classes in such a way that it should display the total number of videos uploaded by a channel of all 4 countries.The time taken to achieve this goal is 15 minutes.

1

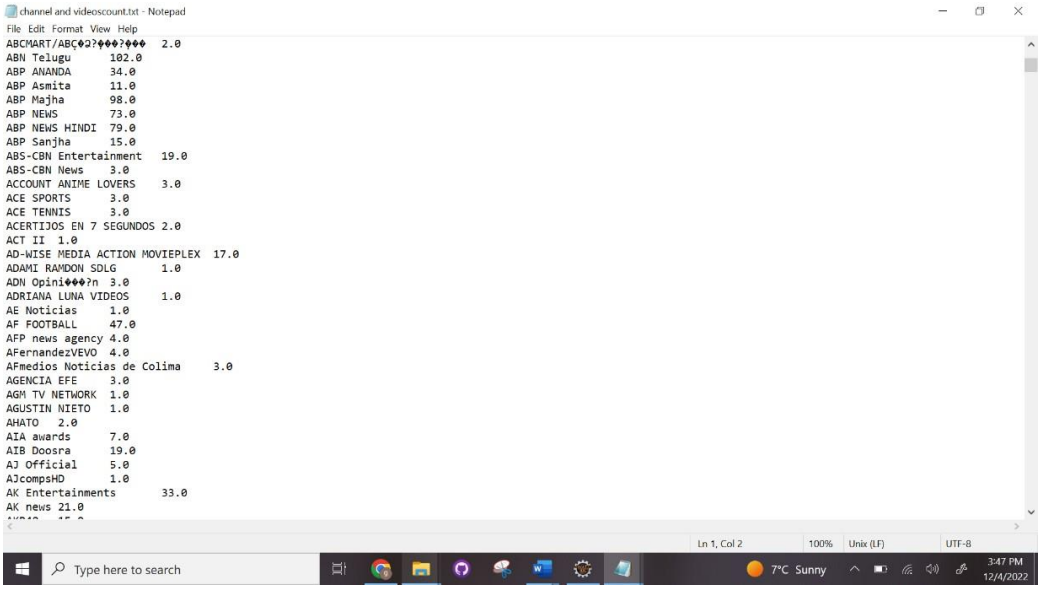
Mapper



Reducer:



Output:

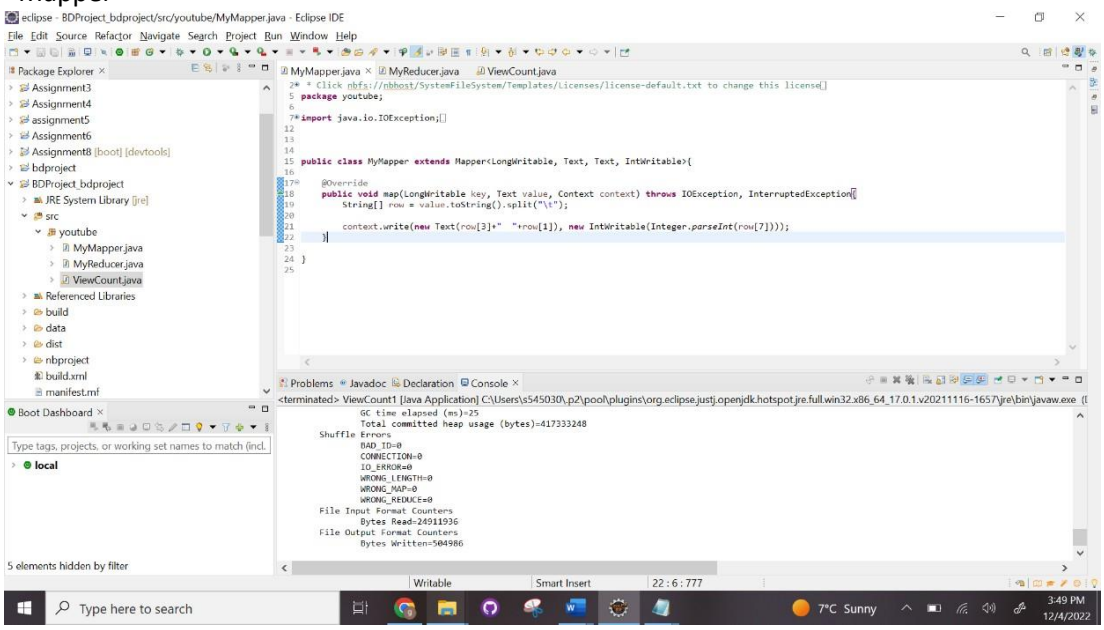


7.8 Goal 8

The goal is to know the number of likes a channel got on a particular day.

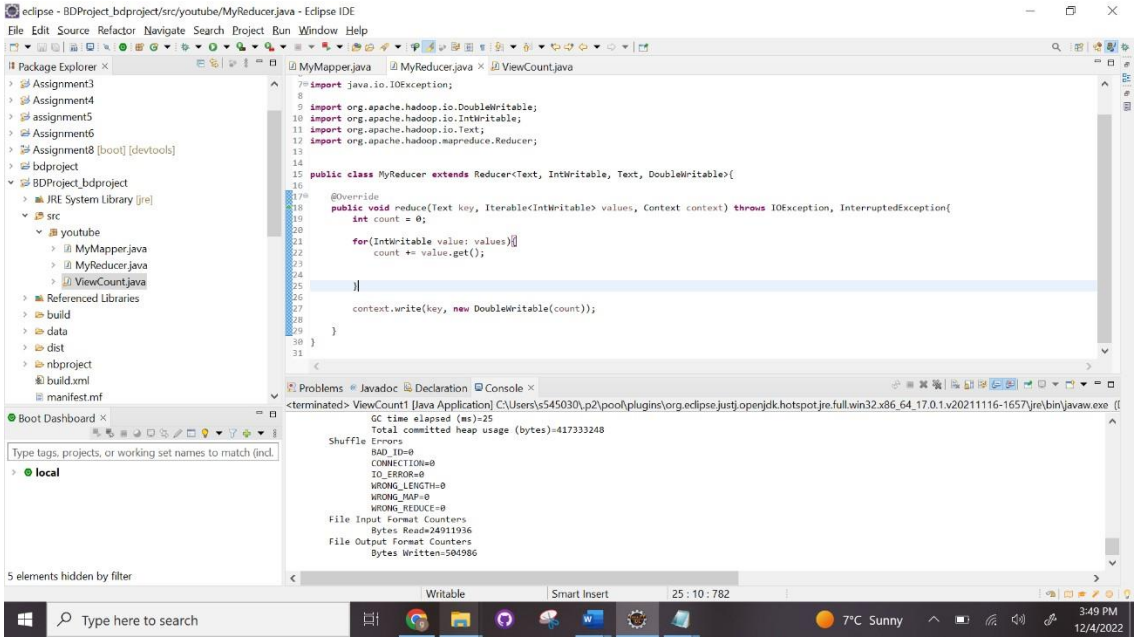
We got the desired output for this goal by applying map reduce techniques on likes row, channel row and dates row, we have written mapper and reducer classes in such a way that it should display the total number of likes received by a channel on a particular day for all videos they uploaded.The time taken to achieve this goal is 15 minutes.

Mapper

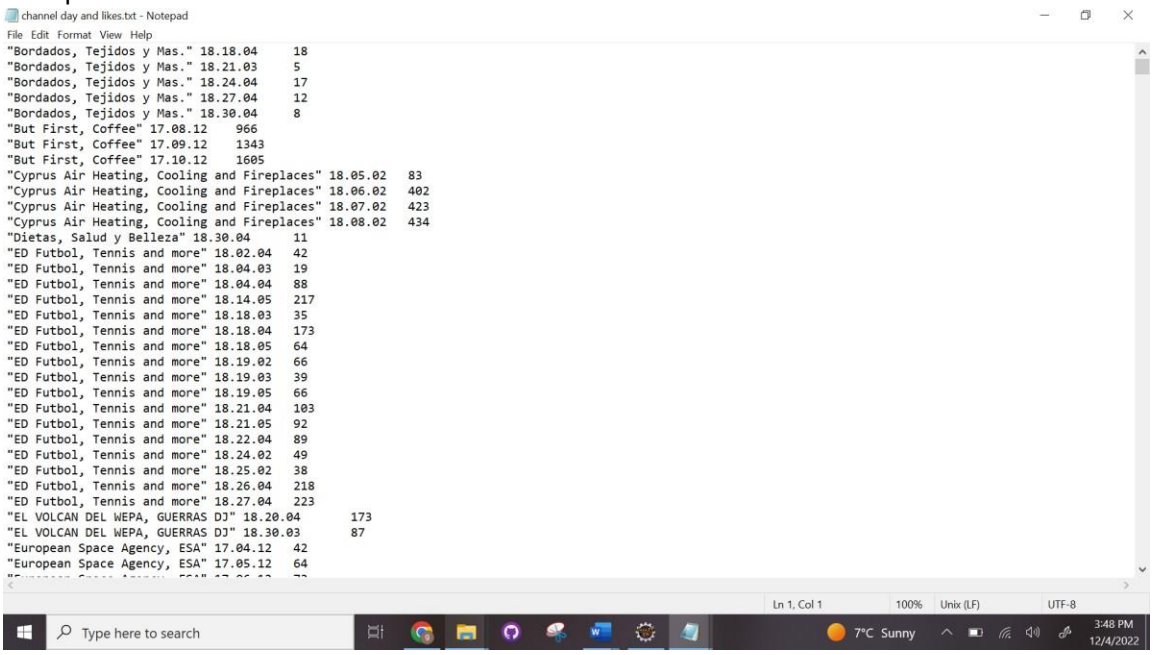


1

Reducer:



Output:

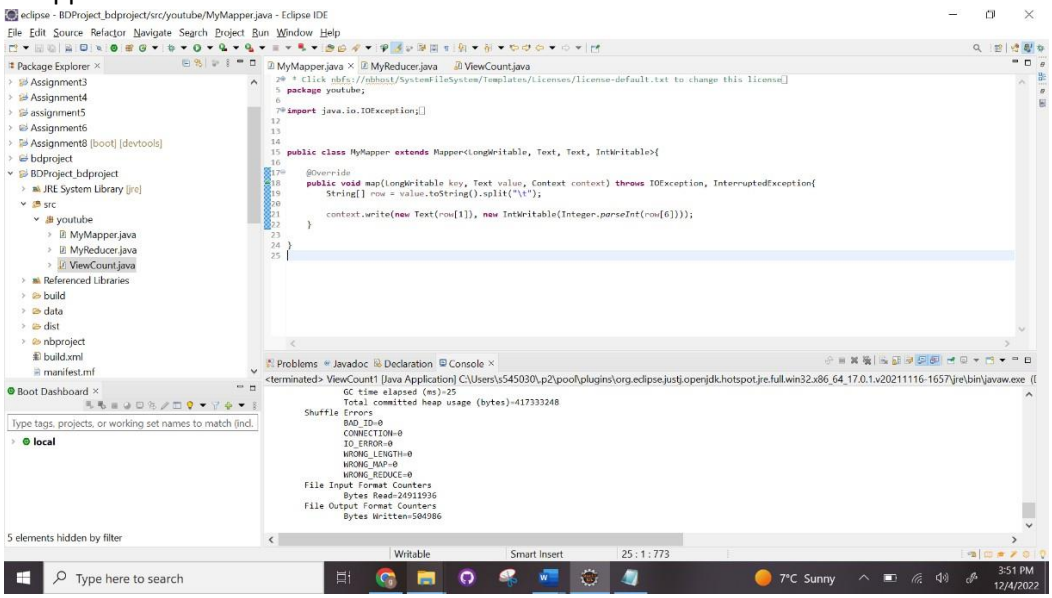


7.9 Goal 9

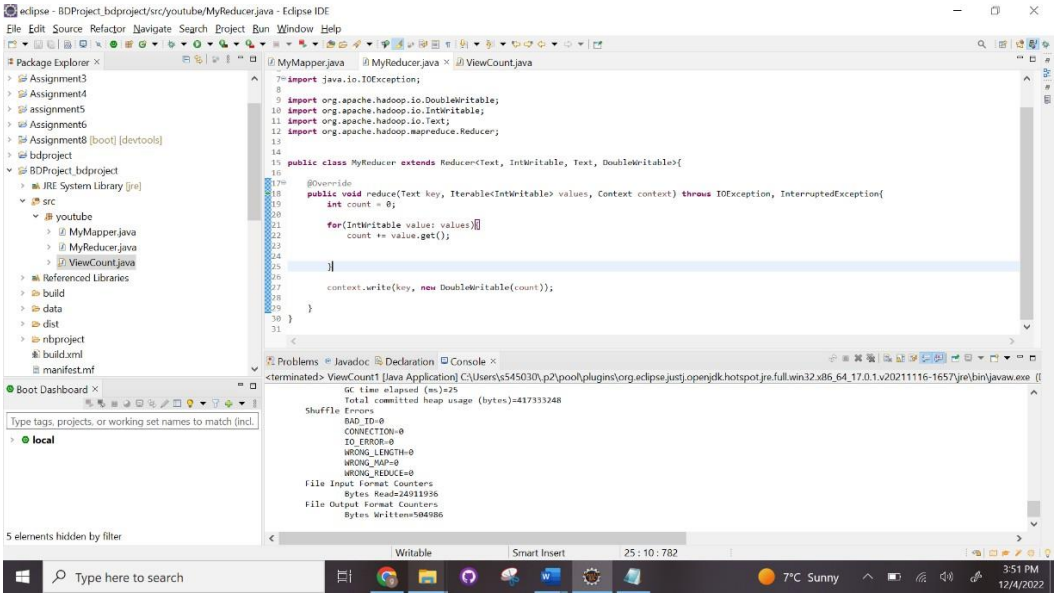
The goal is to know the total views on YouTube on a particular day.

We got the desired output for this goal by applying map reduce techniques on date row and views row, we have written mapper and reducer classes in such a way that it should display the total number of views received by all videos in the youtube on a particular day.The time taken to achieve this goal is 15 minutes.

Mapper

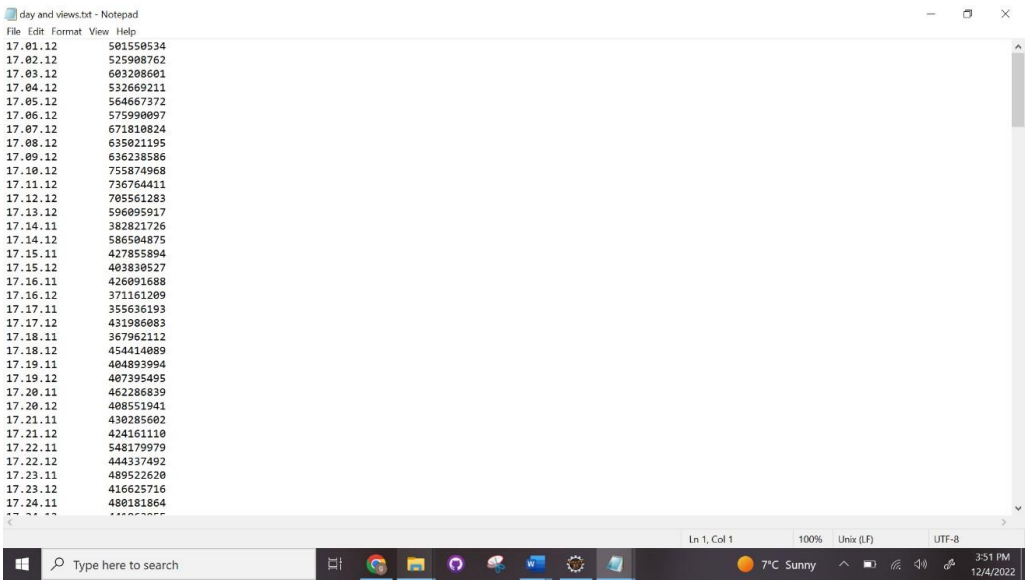


Reducer



1

Output:

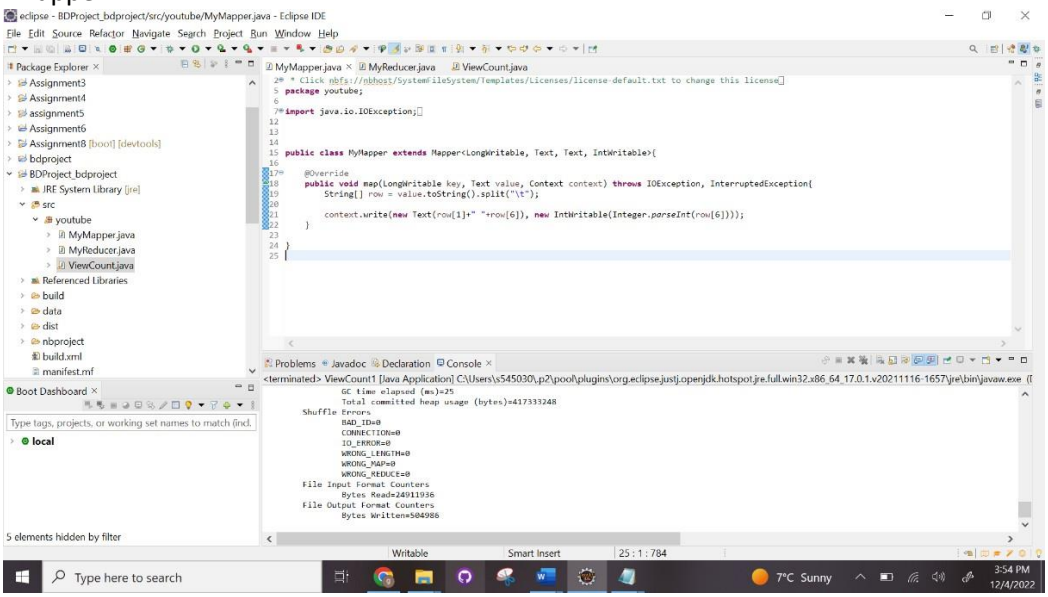


7.10 Goal 10

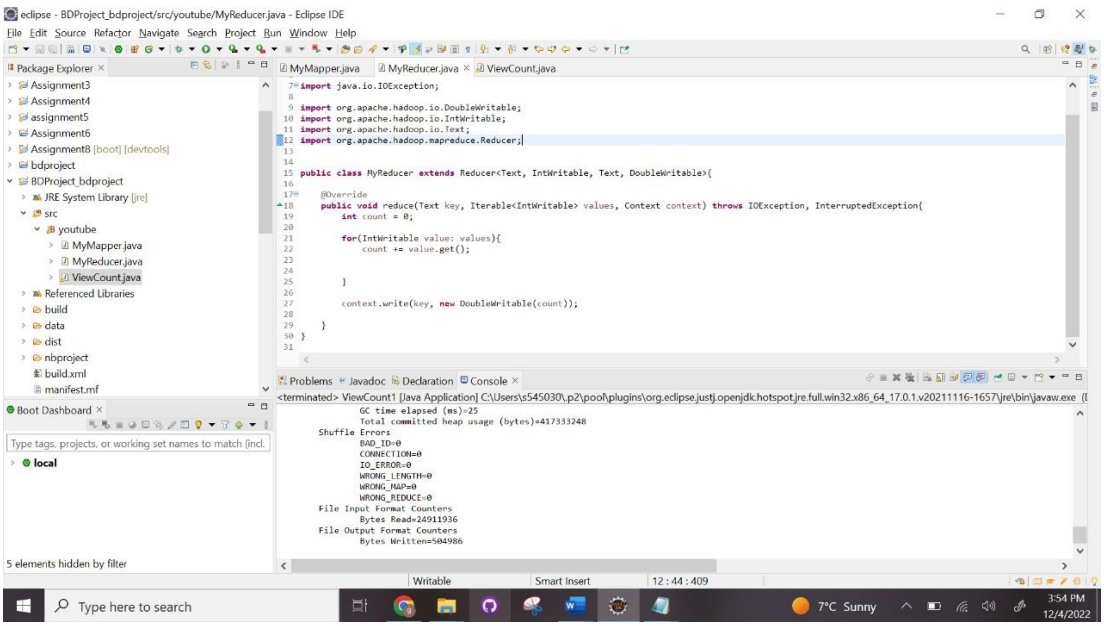
The goal is to know the total views a channel got on a particular day.

We got the desired output for this goal by applying map reduce techniques on channel row,date row and views row, we have written mapper and reducer classes in such a way that it should display the total number of views received by all videos in the youtube for each channel.The time taken to achieve this goal is 15 minutes.

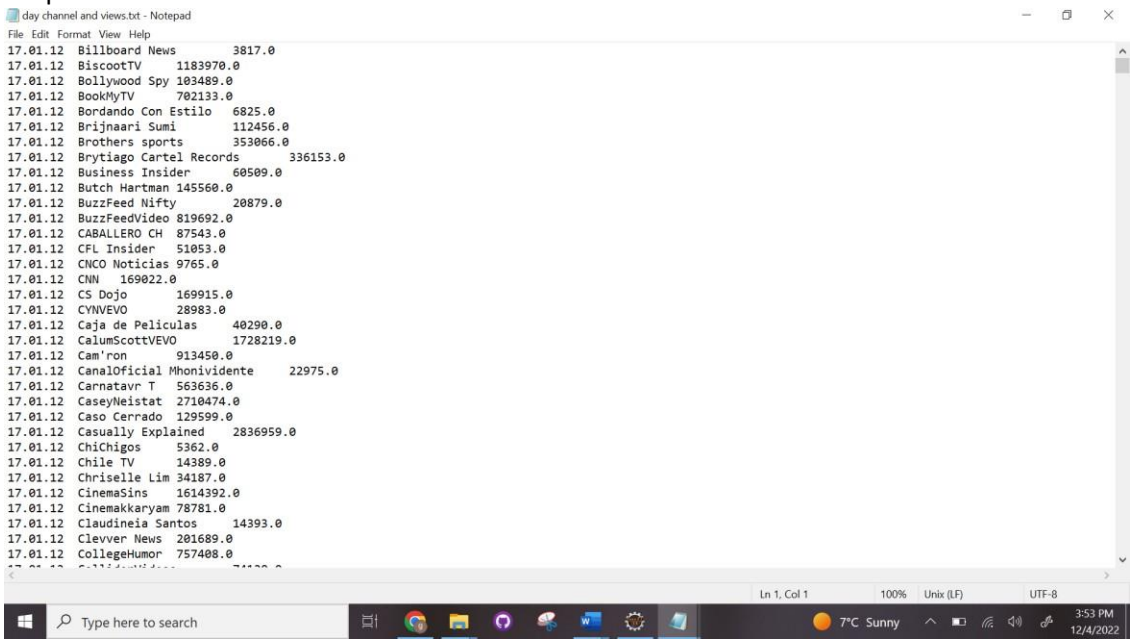
Mapper



Reducer:



Output:



8 Conclusion

This project teaches you how to leverage the power of YouTube Data and draw crucial insights into the performance of YouTube channels of four countries(India, Japan, Mexico, USA). We have achieved the goals that we set for the project using the mentioned technologies.Carrying out an in-depth analysis of the channels requires you to integrate data from a diverse set of data sources. Integrating such diverse data can be challenging and we have used map reduce techniques and data bricks to overcome this challenge and we have got the desired output from the goals we set as expected in a stipulated time

9 Github Repository

https://github.com/Saiv0711/Bigdata-Project