**INDUSTRIAL TRAINING REPORT**

BIG DATA PROJECT REPORT

Submitted in partial fulfilment of the

Requirements for the award of

**Degree of Bachelor of Technology**

**in**

**Computer Science & Engineering**

****

Submitted By :

Name: Pranjal Jain

University Registration No. 10316210174

**SUBMITTED TO:**

**Department of Computer Science & Engineering**

SRM UNIVERSITY DELHI-NCR, SONIPAT

HARYANA-131029

­­­­

**TABLE OF CONTENTS**

Declaration.......................................................................................................................I

Acknowledgement......................................................................................................... II

Company’s Profile......................................................................................................... III

About the platform.......................................................................................................... IV

Data Analysis Using Big Data Hadoop....................................................................................V

* Tools and Technologies deployed
* Snapshots

Conclusion and Scope**.**

**DECLARATION**

I hereby declare that the Industrial Training Report entitled ("Youtube Data Analysis") is an authentic record of my own work as requirements of Industrial Training during the period from 10th June 2018 to 20th July 2018 for the award of degree of B.Tech. (Computer Science & Engineering), SRM University, Delhi-NCR, Haryana, under the guidance of Aptech Positive Solutions partners.

**(Signature of student)**

**Pranjal Jain**

**10316210174**

**Date: 14/11/2018**

**ACKNOWLEDGEMENT**

The internship opportunity I had with APTECH COMPUTER EDUCATION was a great chance for learning and professional development. Therefore, I consider myself as a very lucky individual as I was provided with an opportunity to be a part of it. I am also grateful for having a chance to meet so many wonderful people and professionals who led me though this internship period.

Bearing in mind previous I am using this opportunity to express my deepest gratitude and special thanks to the CTO of APTECH COMPUTER EDUCATION, Mr. Himanshu Sharma who in spite of being extraordinarily busy with his duties, took time out to hear, guide and keep me on the correct path and allowing me to carry out my project at their esteemed organization and extending during the training.

I express my deepest thanks to Mr. Nikhil, Intern Manager for taking part in useful decision & giving necessary advices and guidance and arranged all facilities to make life easier. I choose this moment to acknowledge her contribution gratefully.

I perceive as this opportunity as a big milestone in my career development. I will strive to use gained skills and knowledge in the best possible way, and I will continue to work on their improvement, in order to attain desired career objectives. Hope to continue cooperation with all of you in the future.

Sincerely,

Pranjal Jain

Delhi

Date:14/11/2018

Company Profile

**Overview**

Aptech is a premier education institute. Aptech is a pioneer in providing career-orientated courses to aspiring professionals since 1986.

The institute has trained 68 lakh (6.8 million) students through a wide network of education centres located in over 40 countries.

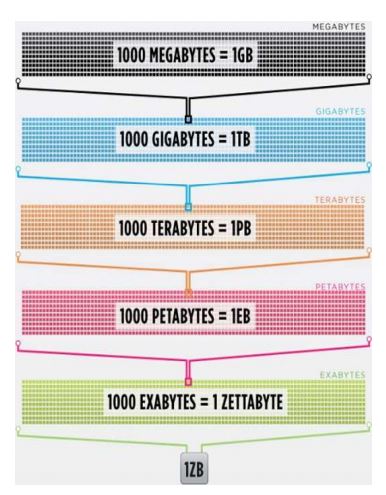
**Quality** **Education**

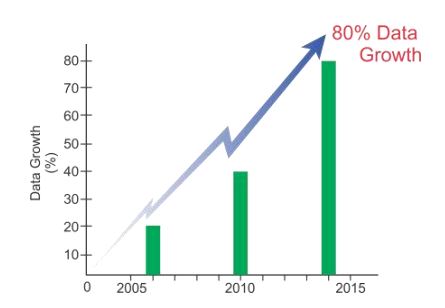
Aptech gives great importance to providing high-quality education to students. By joining a career course with Aptech, you get:

* A wide variety of career, professional, short-term & certification courses designed for the learning & career needs of students, working professionals & others.
* Workshops, Events & other activities to encourage student-industry interaction, prepare them for their job interviews & make them industry-ready.
* Job fairs, campus placement drives, etc. conducted regularly for students to get job placements on successful course completion.

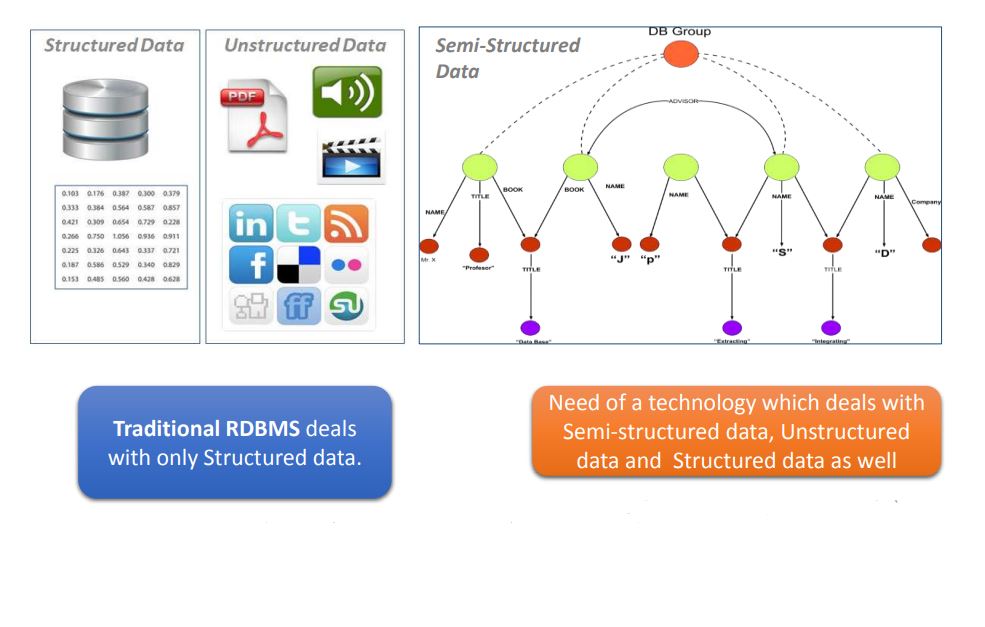
**About the platform**

**What is Big Data?**

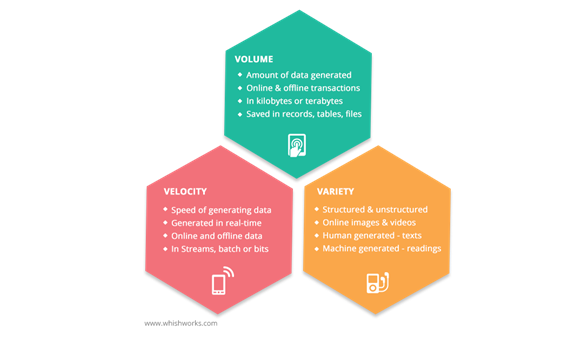
Big data is the term for a collection of data sets so large and complex that it becomes difficult to process using on-hand database management tools or traditional data processing applications.



**Types of Big Data**



**3V’s of Big Data**

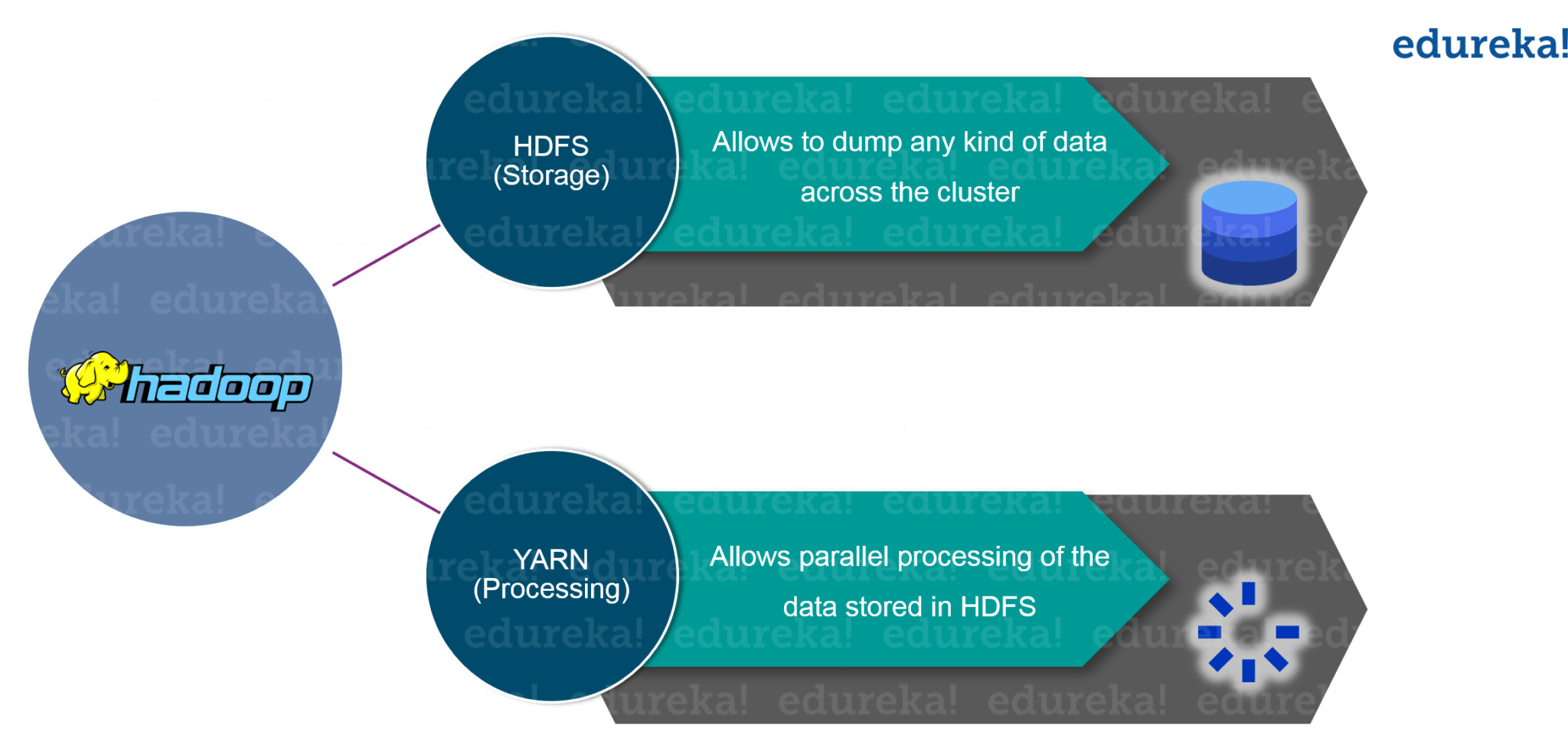


**Sources of Data**



**What is Hadoop?**

Hadoop is a framework that allows you to first store Big Data in a distributed environment, so that, you can process it parallely. There are basically two components in Hadoop:



The first one is ***HDFS*** for storage (Hadoop distributed File System), that allows you to store data of various formats across a cluster. The second one is ***YARN***, for resource management in Hadoop. It allows parallel processing over the data, i.e. stored across HDFS.

**Difference between Hadoop & RDBMS**

1. Hadoop is not a database.  Hbase or Impala may be considered databases but Hadoop is just a file system (hdfs) with built in redundancy, parallelism.
2. Hadoop offers massive scale in processing power and storage at a very low comparable cost to an RDBMS.
3. Hadoop offers tremendous parallel processing capabilities.  You can run jobs in parallel to crunch large volumes of data.
4. Typically RDBMS will manage a large chunk of the data in its cache for faster processing while at the same time maintaining read consistency across sessions.  I would argue Hadoop does a better job at using the memory cache to process the data without offering any other items like read consistency.
5. Hadoop is very good for parallel processing problems - like finding a set of keywords in a large set of documents (this operation can be parallelized).  However typically RDBMS implementations will be faster for comparable data sets.
6. **Hardware failure:** as soon as you start using many pieces of hardware, the chance that one will fail is very high. A common way of avoiding data loss is through replication. Redundant copies of the data are kept by the system so that in the event of failure, there is another copy available. Node failure and disk failure efficient handle in Hadoop framework.

|  |  |  |
| --- | --- | --- |
|  | **Traditional RDBMS** | **Hadoop** |
| Data Size | Gigabytes | Petabytes |
| Access | Interactive and Batch | Batch |
| Updates | Read & Write many times | Write once, read many times |
| Structure | Static Schema | Dynamic Schema |
| Integrity | High | Low |
| Scaling | Non Linear | Linear |

Other software components that can run on top of or alongside Hadoop and have achieved top-level Apache project status include:

|  |  |
| --- | --- |
| **Framework** | **Description** |
| Flume | Software that collects, aggregates and moves large amounts of streaming data into HDFS. |
| Hive | A data warehousing and SQL like query language that presents data in the form of tables. Hive programming is similar to database programming. |
| Oozie | A Hadoop job scheduler. |
| Pig | A platform for manipulating data stored in HDFS that includes a compiler for MapReduce programs and a high-level language called PigLatin. It provides away to perform data extractions, transformations and loading, and basic analysis without having to write MapReduce programs. |
| Spark | An open-source cluster computing framework within-memory analytics. |
| Sqoop | A connection and transfer mechanism that moves data between Hadoop and relational databases. |

**Data Analysis Using Big Data Hadoop**

**Project Objective**

To analyse the YouTube dataset, and find out top 5 most likely and profitable categories(industries).

**Software and Framework Requirement**

**Softwares Used:**

1. Virtual Box -  VirtualBox allows you to run more than one operating system at a time. This way, you can run software **written** for one operating system on another (for example, Linux on Windows or a Mac) without having to **reboot** to use it.
2. Winscp - WinSCP is a free and open-source SFTP, FTP, WebDAV, Amazon S3 and SCP client for Microsoft Windows. Its main function is secure file transfer between a local and a remote computer. Beyond this, WinSCP offers basic file manager and file synchronization functionality.
3. Putty - PuTTY is a versatile terminal program for Windows. It is the world's most popular free SSH client. It supports [SSH](https://www.ssh.com/ssh/protocol/), [telnet](https://www.ssh.com/ssh/telnet), and raw socket connections with good terminal emulation.

**Framework Used**

Apache Pig

**Apache Pig Framework**

Pig is a high level programming language useful for analyzing large data sets. Pig was a result of development effort at Yahoo!

Pig enables people to focus more on analyzing bulk data sets and to spend less time in writing Map-Reduce programs.

**Pig Components**

Pig consists of two components:

* **Pig Latin,** which is a language
* **Runtime environment,** for running PigLatin programs.

A Pig Latin program consist of a series of operations or transformations which are applied to the input data to produce output. These operations describe a data flow which is translated into an executable representation, by Pig execution environment. Underneath, results of these transformations are series of MapReduce jobs which a programmer is unaware of. So, in a way, Pig allows programmer to focus on data rather than the nature of execution.

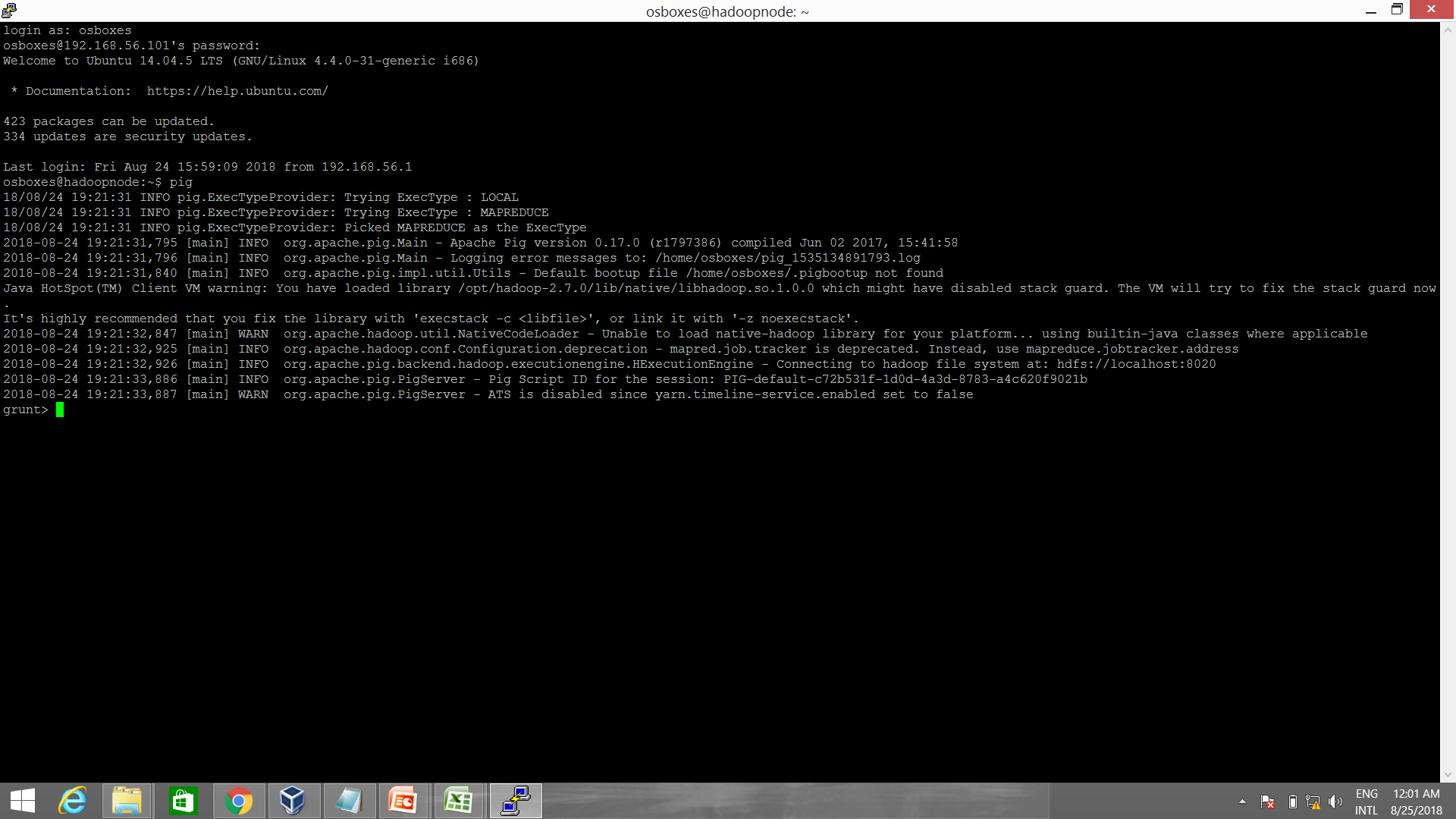
PigLatin is a relatively stiffened language which uses familiar keywords from data processing e.g., Join, Group and Filter.

**YouTube Data-Set Schema**

**Data-Set Description**

Column 1: Video id of 11 characters.  
Column 2: uploader of the video  
Column 3: Interval between the day of establishment of Youtube and the date of uploading of the video.  
Column 4: Category of the video.  
Column 5: Length of the video.  
Column 6: Number of views for the video.  
Column 7: Rating on the video.  
Column 8: Number of ratings given for the video  
Column 9: Number of comments done on the videos.

**Running Pig Framework**

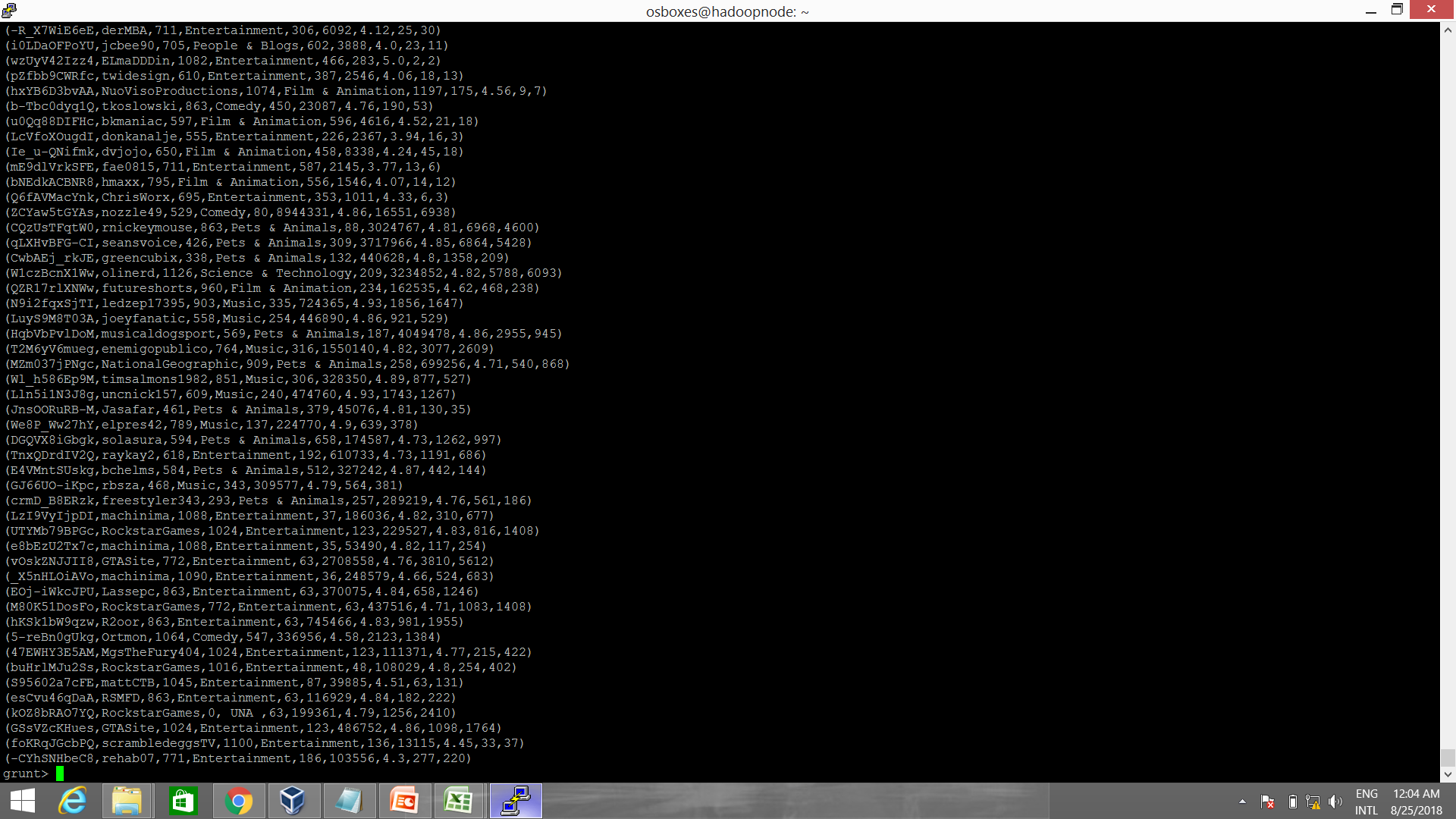
****

**Command to load data in Pig**

grunt> youtube = load ‘/pig/youtubedata.txt’ using PigStorage(‘\t’) AS (video\_id:chararray, channel:chararray, days:int, category:chararra y, lenght:chararray, views:int, rating:float, no\_rate:int, no\_cmmnt:int);

grunt> dump youtube;

**Data Loaded in Pig**

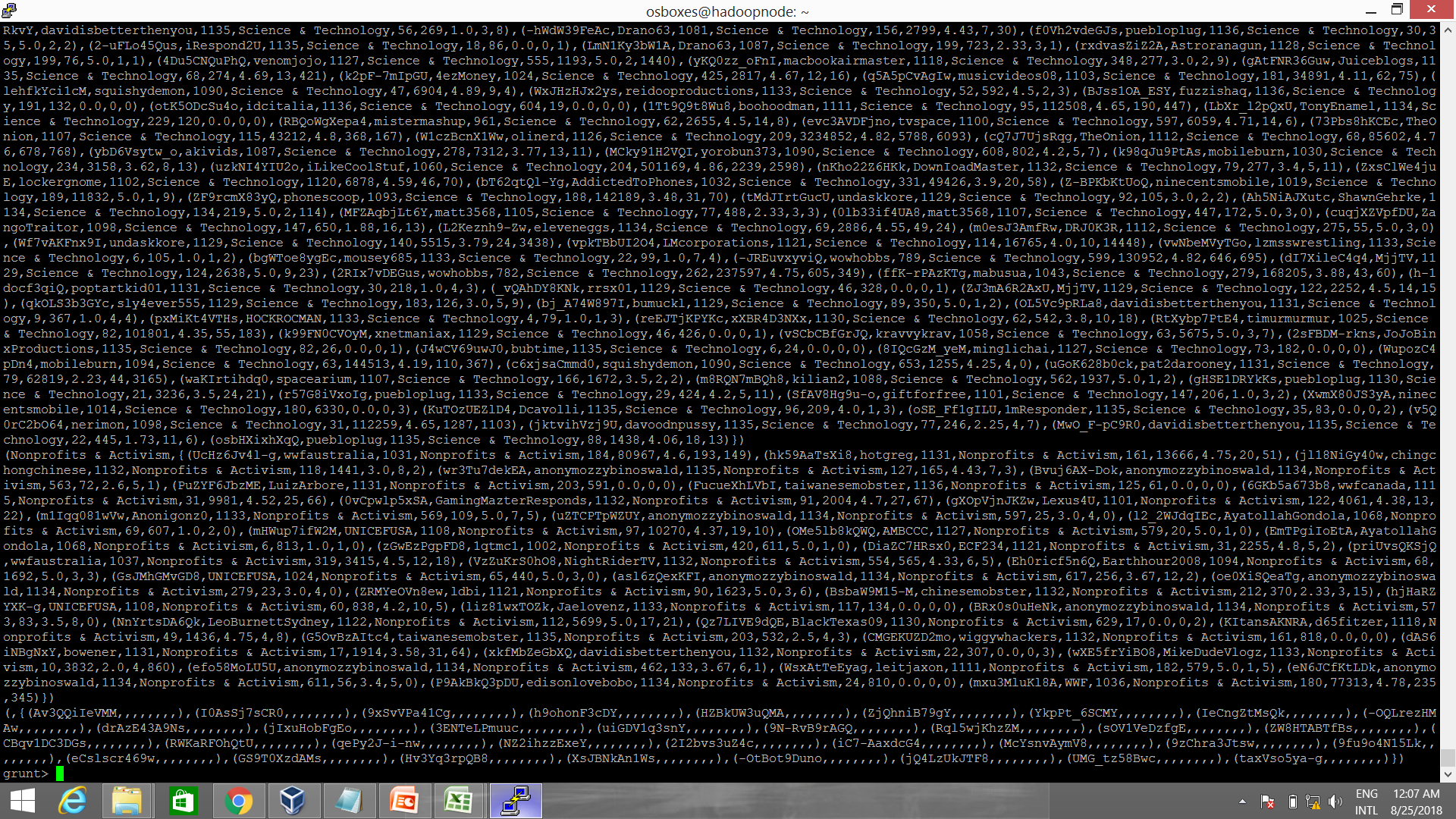
****

**Group Command**

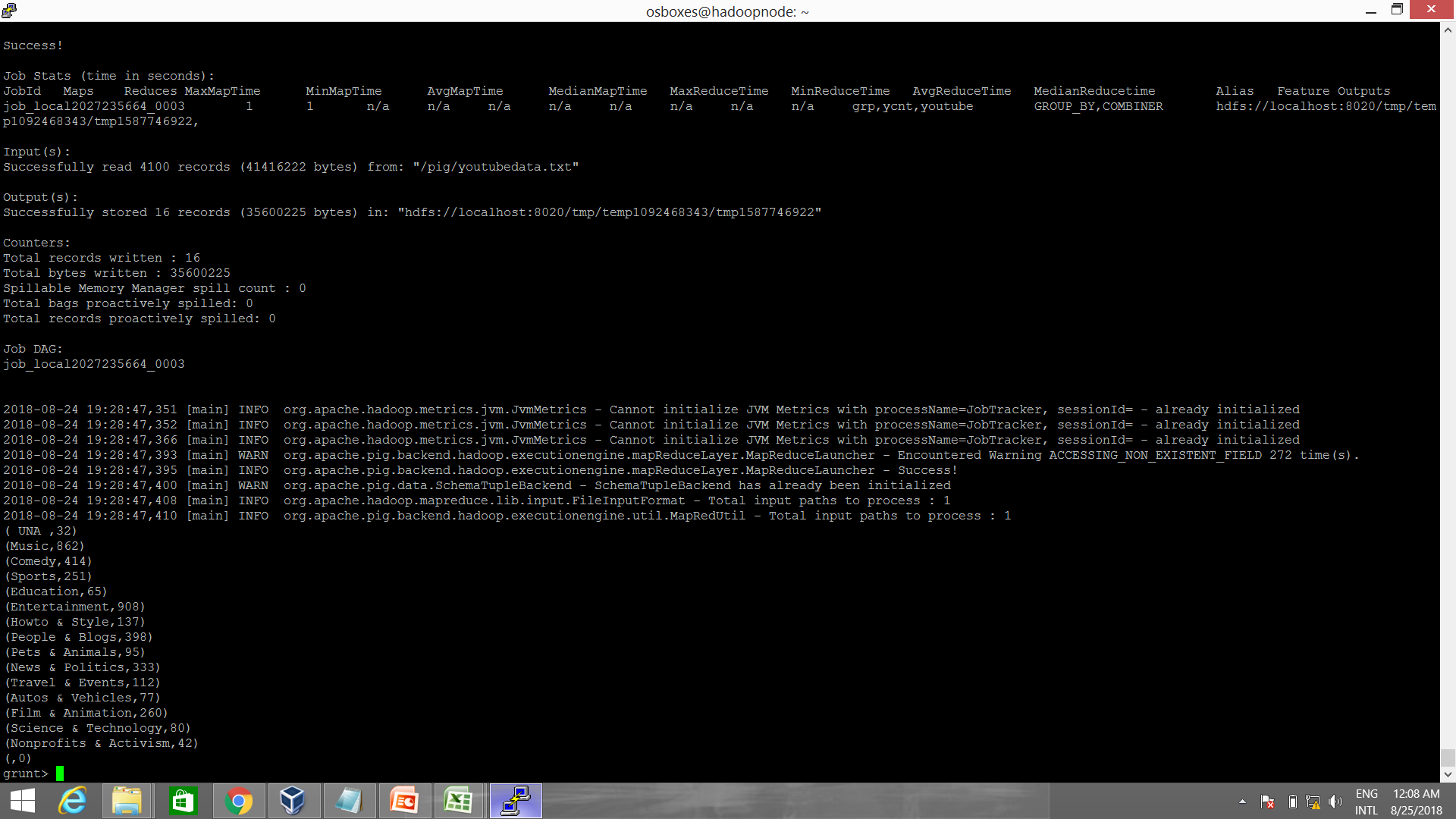
grunt> grp = GROUP youtube BY category;

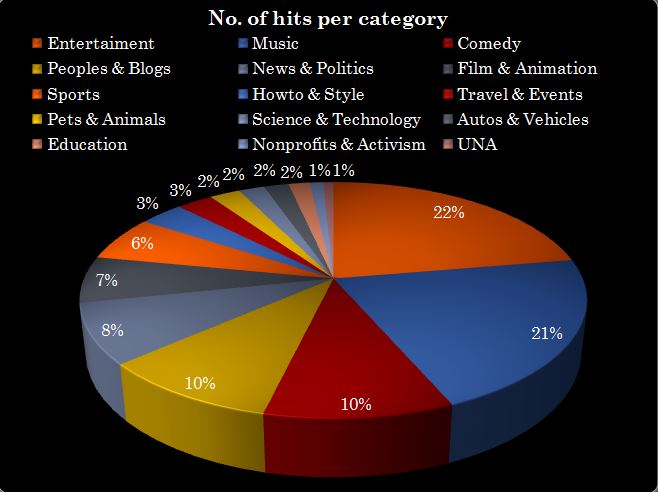
grunt> dump grp;

**Showing Grouped Data**

****

**Showing different categories using count command**

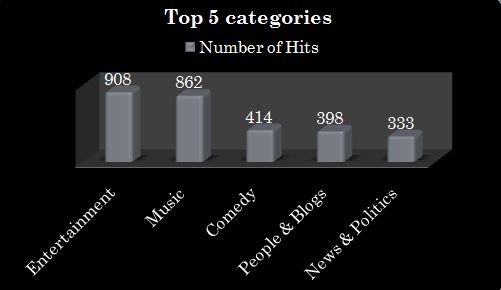
****

****

**Analysing the pie chart**

Here we can observe 15 categories with their percentage. Now we will select only top 5 categories holding maximum percentage. They are:-

* Entertainment
* Music
* Comedy
* People & Blogs
* News & Politics

****

**Comparison Conditions**

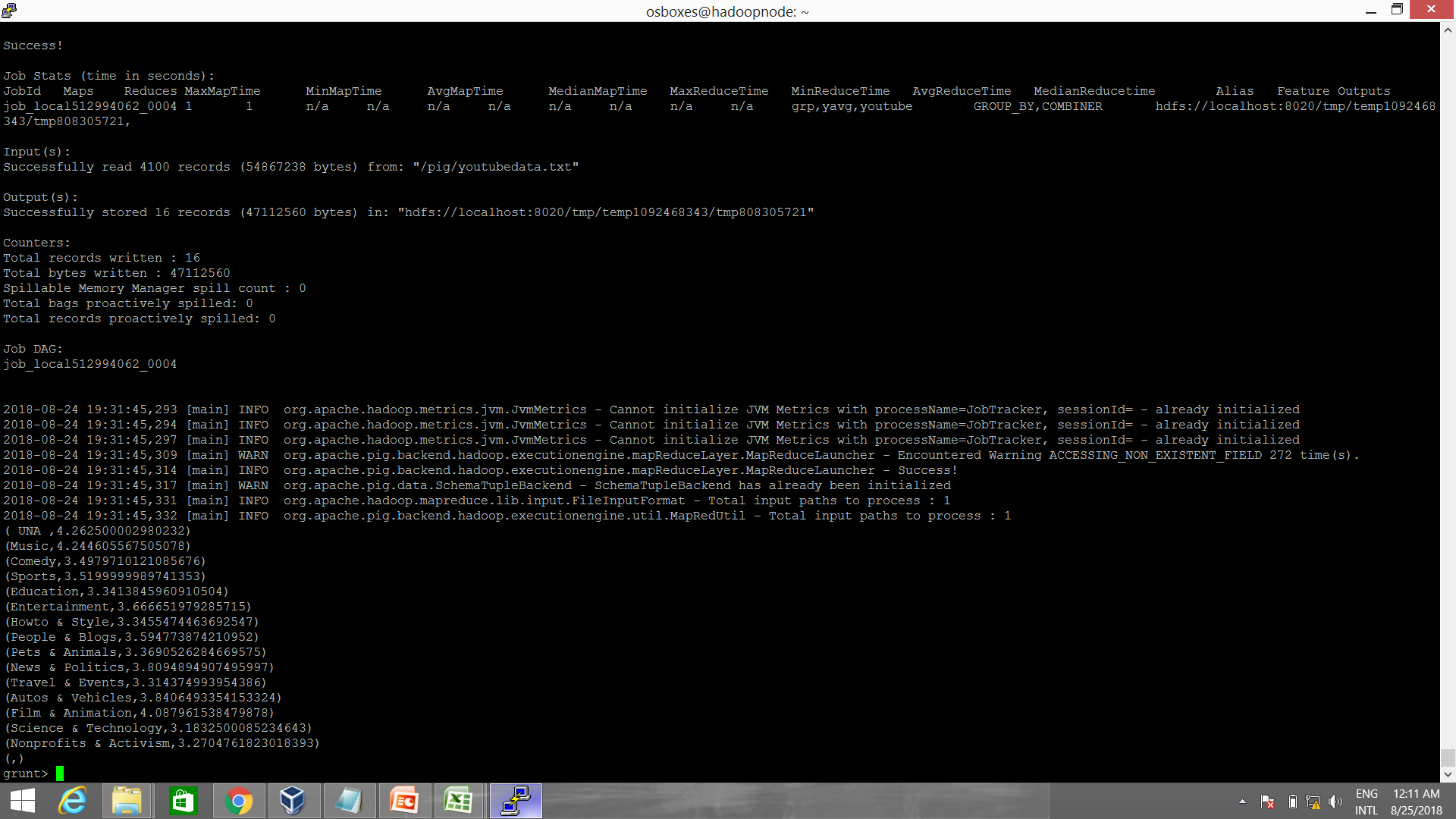
Here we compare these categories on the basis of their:-

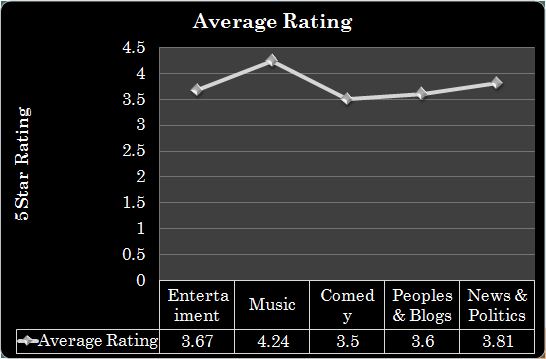
* Count of a category.
* Average user rating(5 star).
* Total no. of views.
* Maximum views on an video.
* No. of rating given to video per 1000.

**Average User Rating**

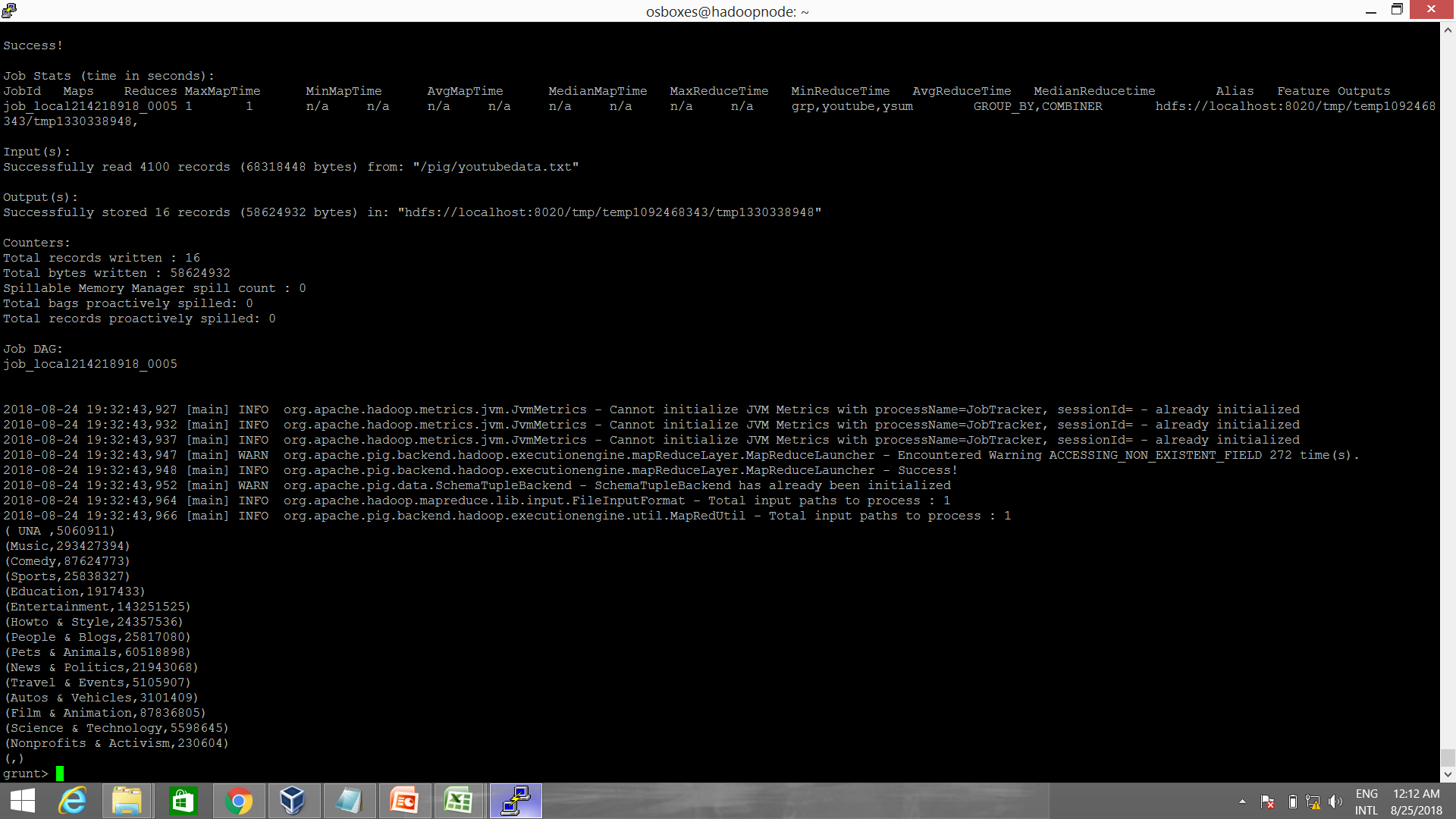
grunt> yavg = foreach grp generate group as category,AVG(youtube.rating) as avg\_rate;

grunt> dump yavg;

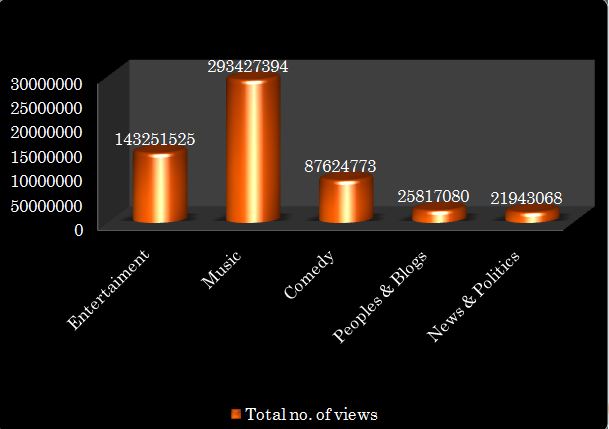
****

****

**Total number of views**

grunt> ysum = foreach grp generate group as category,SUM(youtube.views) as views;

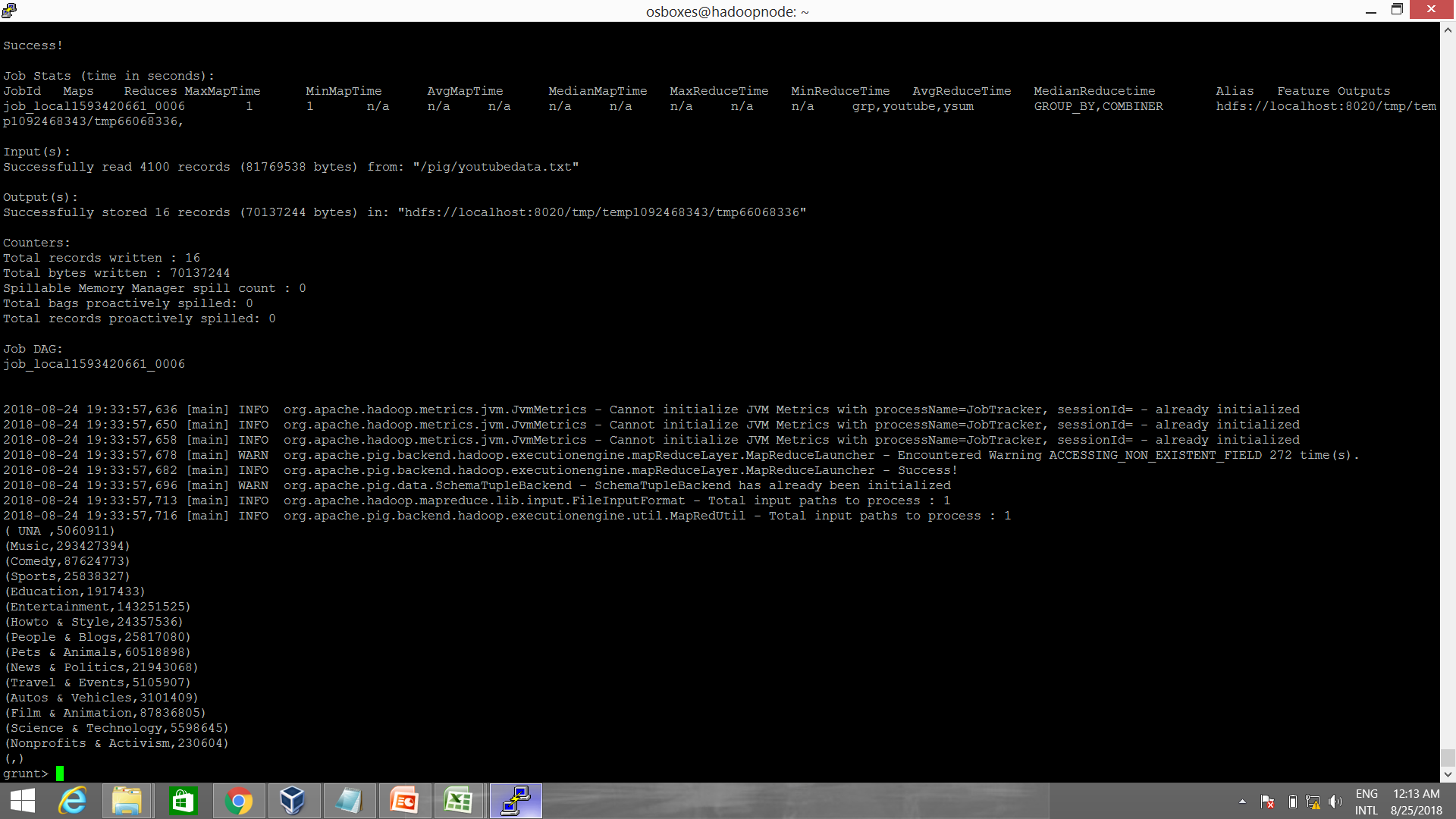
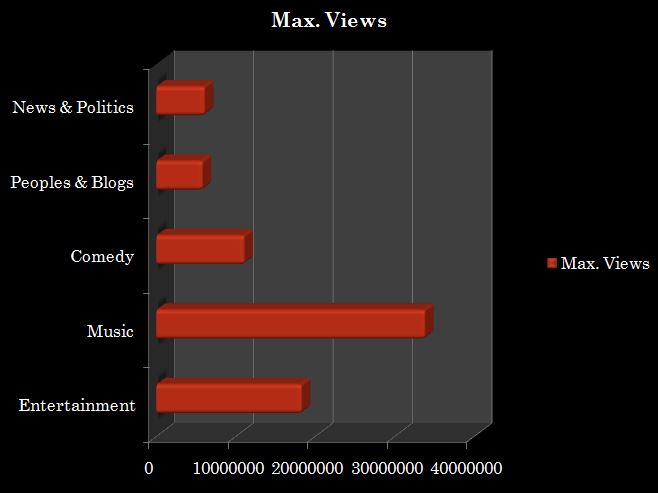
grunt> dump ysum;



**Maximum views on an video**

grunt> ymax = foreach grp generate group as category,MAX(youtube.views) as max\_views;

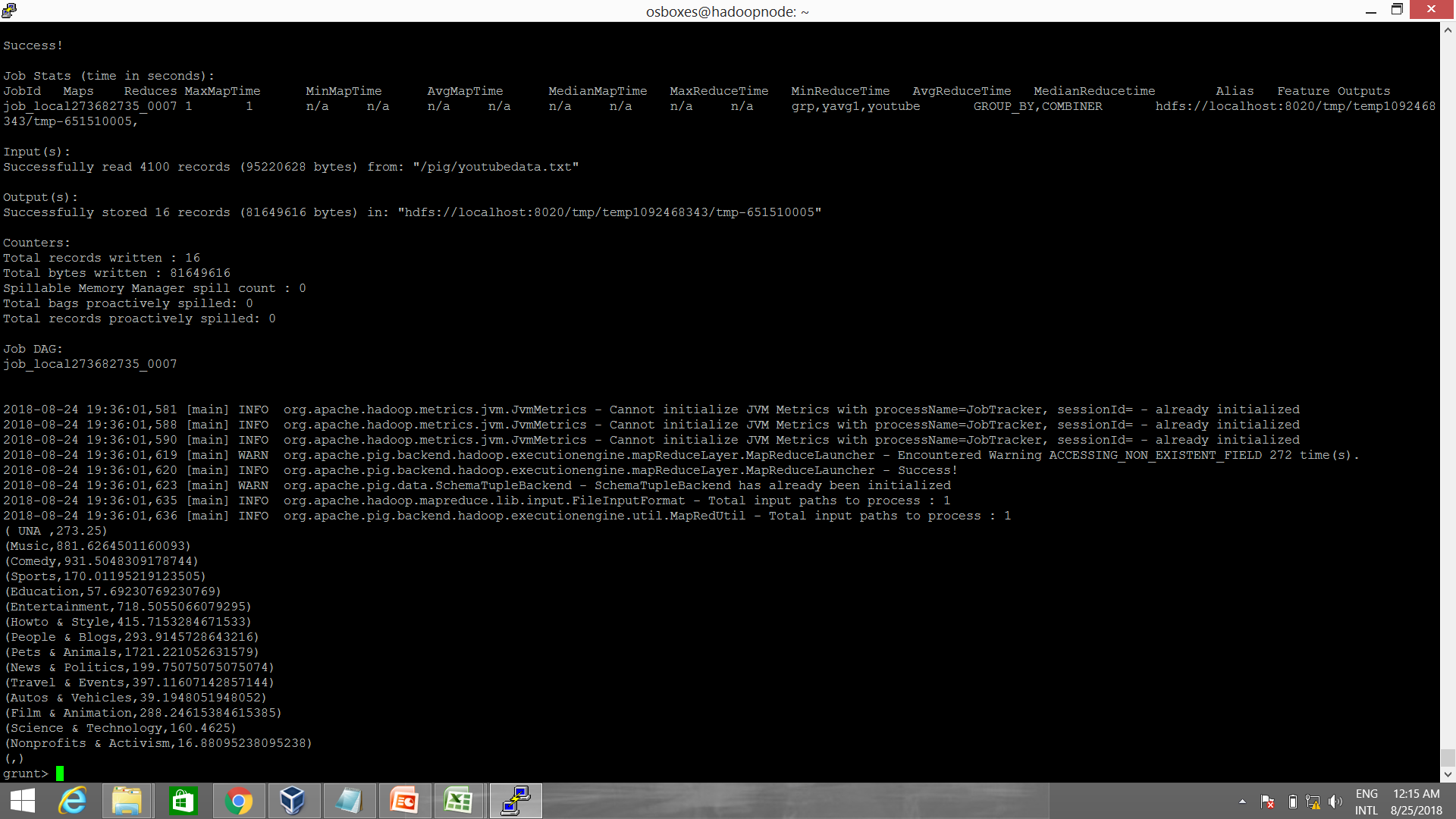
grunt> dump ymax;

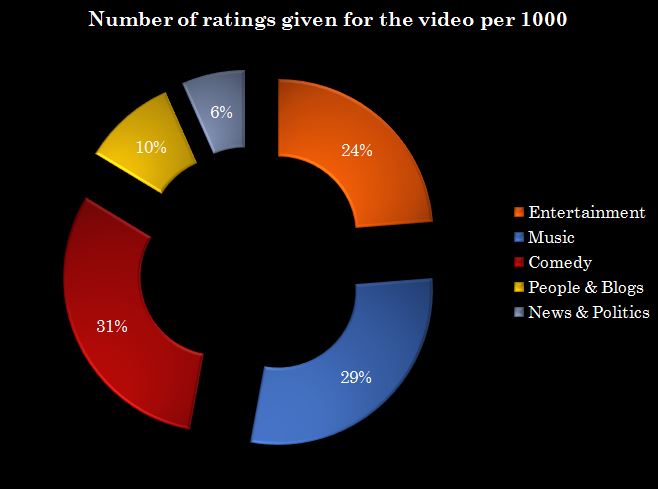
****

**No. of ratings given for the video per 1000 comments**

grunt> yavg1 = foreach grp generate group as category,AVG(youtube.no\_rate) as avgrate;

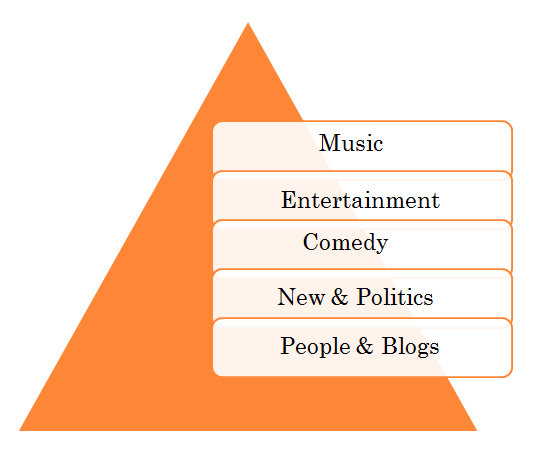
grunt> dump yavg1;





On the basis of above conditions, Youtube dataset analysis is over, thus I got the result on the basis of applied conditions.

Now I’ll proceed towards the conclusion of the project.

**Conclusion -** On the basis of above comparison, Thus I can conclude that the most likely and profitable category(industry):

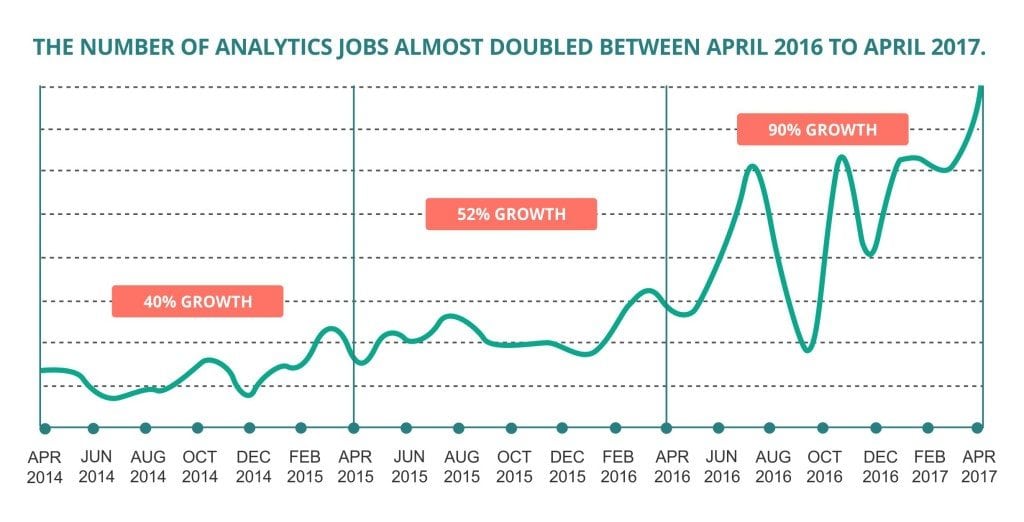
**Scope of this Technology**

**Hadoop and Big Data**

Hadoop is the supermodel of Big Data. To be skilled in Hadoop is a deciding factor in getting a springboard to your career or getting left behind. If you are a fresher there is a huge scope if you are skilled in Hadoop. Amongst the open source framework, there is almost no other alternative which can deal with petabytes of data as Hadoop can. In 2015 was it was predicted that Indian [Big Data Hadoop](https://imarticus.org/future-of-big-data-hadoop-developer-in-india/) industry will grow five folds in the analytics centre.

**Job Market in Analytics on the Rise in India**

Research suggests that by the end of 2018 India alone will face a shortage of about two lac data scientist. The probable growth of Big Data in India is because of the awareness of the benefits that insights from unstructured data can impact businesses and increase its ROI. Another fact is that India is considered a hub for outsourcing such operational efficiencies at low cost. One can see Bangalore emerging as a hub for such outsourcing capabilities.



**Salary Structure on Big Data and Hadoop professionals in India**

The salary structure for a trained professional in Big Data Hadoop is quite lucrative with an average start-up at 6 – 9 lac and a manager with 7-10 years getting anywhere close to 15-20 lac and in some cases above 15 years of experience drawing almost or more than a 1 crore.