

***A Mini-Project Report On***

**“Comparison between Apache Spark and Apache Hadoop using YouTube Data”**

***Submitted By***

**Your Name**

**Roll No: 056**

**F.Y. M.Sc. (Data Science and Big Data Analytics)**

**School of Computer Science**

**Faculty of Science**

**MIT – World Peace University**

**Pune - 411038**

**Academic Year 2019-2020**

**APRIL – 2020**

**MIT WORLD PEACE UNIVERSITY, PUNE**

**SCHOOL OF COMPUTER SCIENCE**

***Certificate***

This is to certify that

**Your Name, Roll No. 056,**

Of ***M.Sc. (Data Science and Big Data Analytics)*** successfully completed his/her Mini-Project in Hadoop

**“Comparison between Spark and Hadoop using YouTube Data”**

to our satisfaction and submitted the same during the academic year 2019- 2020 towards the partial fulfillment of degree of **Master of Science in Data Science and Big Data Analytics** of MIT World Peace University under the School of Computer Science, MIT WPU, Pune.

|  |  |  |
| --- | --- | --- |
| **Dr. Shubhalaxmi Joshi** | **Prof. Sachin Bhoite** | **Prof. Prajakta Soman** |
| **Associate Dean** | **Program Head** | **Assistant Professor** |
| **Faculty of Science** | **School of Computer Science** | **School of Computer Science** |
| **MITWPU** | **MIT WPU** | **MIT WPU** |

**ABSTRACT**

In this Big Data mini project, our main aim was to do comparison of Apache Hadoop tools and Apache Spark tools. So, we have performed data analysis on YouTube Data using two Hadoop tools i.e Pig and Hive and PySpark, this short mini project will help us compare between technologies.

**ACKNOWLEDGEMENT**

It gives us a great pleasure to present our project on Big Data. This is our

milestone in M.Sc Data Science and Big Data Analytics.

We would like to express our sincere thanks to all the Mentors who helped us throughout the project. I would like to acknowledge the help and guidance provided by our Project Guide Mrs Prajakta Soman in all places during the

presentation of this project.

We are also thankful to MIT-WPU management for providing infrastructure and environment. Onwards our project works, we are also thankful to the staff member of the Faculty of Science for their moral supports towards the project.

Your Name

Roll No

Introduction

**Domain Name**

The Domain we are working on is Social Media. We analysed the data from YouTube and found out some key insight through it.

**Motivation**

The reason we choose this topic is because, there’s a lot of debate over which technology is better in the Big Data World, Apache Spark or Hadoop. So to finally come to a conclusion on this subject, we ourself thought to find it out by using both these technologies and comparing them.

**Solution to the problem**

We compared both the technologies, Apache Spark and Apache Hadoop using the same dataset. But as these technologies have various components, we used them to compare which one has an edge over another. Also we analysed the data and found some insights from it.

**Formal Problem Statement**

The formal problem statement is to compare and see which technology outperforms the other in different aspects.

We have also created Problem statements of our own to be performed on both these technologies:

1. Find out the top 10 categories in which the most number of videos are uploaded
2. Find top 10 rated videos in each category
3. Find top 10 commented videos in each category
4. Find top 10 most viewed videos in each category
5. Find top 10 lengthy videos in each category.
6. Find sum of video ratings in each category.
7. Find sum of video comments in each category.
8. Find sum of video views in each category.
9. Find sum of video rating, video comments and video views based on category.

**Dataset Details**

The attributes of the dataset along with their description :-

|  |  |
| --- | --- |
| **Attributes** | **Description** |
| video ID | an 11-digit string, which is unique |
| Uploader | a string of the video uploader's username |
| Age | an integer number of days between the date when the video was uploaded and Feb.15, 2007 (YouTube's establishment) |
| category | a string of the video category chosen by the uploader |
| length | an integer number of the video length |
| views | an integer number of the views |
| Rate | a float number of the video rate |
| ratings | an integer number of the ratings |
| comments | an integer number of the comments |
| related IDs | up to 20 strings of the related video IDs |

The size of the YouTube dataset is 203 MB. The dataset has 7,49,316 rows and 11 columns. But in this project we excluded related IDs column as it is not important for the analysis. The data set was collected from the website <http://ybhavesh.blogspot.com/2015/08/proof-of-concept-or-poc-on-youtube-data.html> and the link to directly download the dataset - <http://netsg.cs.sfu.ca/youtubedata/0222.zip>.

Before this, we also explored two datasets which were from WittyFeed and another one from Amazon.

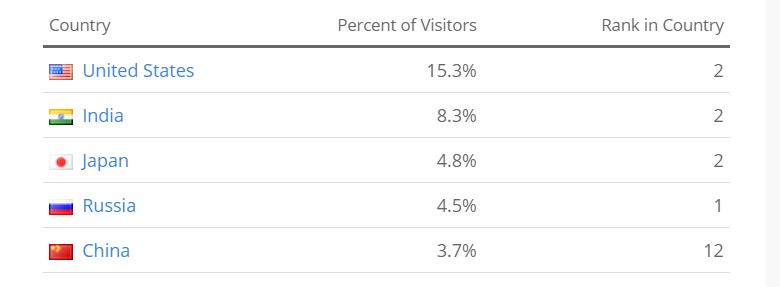
The WittyFeed dataset had data about the likes, comments, shares and subscribe from the pages of various Social Media website like Facebook, Instagram, Twitter etc. The reason we rejected that dataset because it was too small, 1 lakh rows approx.

We also explored Amazon Dataset and even downloaded it. It had 35 million rows approx. The reason we didn’t use it for project is because we were denied permission from Amazon for using the dataset.

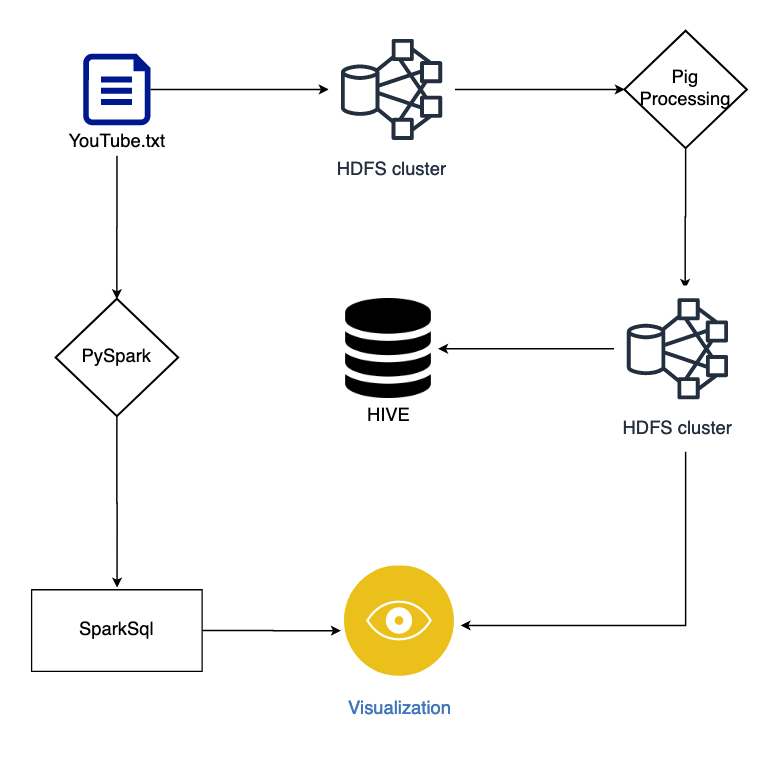
**Related Theory and Facts**

YouTube is a video sharing service where users can watch, like, share, comment and upload their own videos. The video service can be accessed on PCs, laptops, tablets and via mobile phones. It is one of the most popular video hosting platforms out there with almost [5 billion videos watched on the platform per day and 1,300,000,000 users](https://merchdope.com/youtube-stats/) as of March 2020. The type of videos/content available on the platform are:- Product reviews video, Vlogs, Gaming videos, Comedy/skit videos, Haul videos, Memes/tags, Educational videos, Unboxing videos, Q&A videos, Prank videos.

YouTube is localised in 91 countries, and can be accessed in 80 different languages according to its official press assets. This reportedly covers 95% of the internet population.

*Top countries for YouTube usage*

*Source:* [Alexa](https://www.alexa.com/siteinfo/youtube.com?ver=classic)

**Architecture and Design of the system**

**Hadoop Implementation**

1. YouTube dataset is loaded in HDFS.
2. The dataset is then transferred to PIG Processing, where 9 different tables are created.
3. The 9 tables created in HIVE are send to HDFS. We created parquet file for every table in Hadoop, as parquet is more efficient in terms of storage and performance.
4. Then the tables are send to HIVE where different queries would be performed and the data would be analysed.
5. The data is visualised using Power BI.

**PySpark Implementation**

1. YouTube dataset is loaded in PySpark.
2. 5 different dataframes are created in PySpark.
3. The dataframes are then manipulated using SparkSql queries and the data is analysed.
4. Also created parquet files to store the metadata.
5. The data is then visualised using python library called Plotly.

**Technology Stack**

Python programming language is used in the project. Various python libraries are also used like Pandas for creating and manipulating dataframes and Plotly for interactive data visualisation. Other than that, scripting languages like SQL and PIG Latin are also used for quering on databases.

[Jupyter Notebook](http://jupyter.org/) which is a [development environment](https://www.fullstackpython.com/development-environments.html) for writing and executing Python code is also used in the project. All the PySpark and SparkSql commands are run and executed on this IDE.

The Frameworks used in this project are Hadoop and Apache Spark. We used various components from Hadoop ecosystem like HDFS, HIVE and PIG. Also components like PySpark and SparkSql were used from Apache Spark framework.

**Software and Hardware Requirements**

**Software Requirement -**

Operating System - Windows 10, MacOs 14.2 and later (64 bit)

IDE - Jupyter Notebook (versions - 6.0.3, 6.0.2)

Framework - Apache Spark version - 2.4.5

Cloudera version - CDH 5.13.x

**Hardware Requirement -**

Processor - Intel core i5, i7, i9 (2.3 Ghz, 2.6 Ghz, 2.9 Ghz)

RAM - 8GB or more

**Implementation Details**

**Pre-processing:**

The YouTube data which was downloaded, was in the format of txt files. There were in total 5 txt files in the zip folder. We then merged all the files and created a single txt file.

**Final Output and Findings:**

According to our analysis we found that:-

1. The video category Music, Entertainment and Comedy videos has the most number of videos on YouTube.
2. The video category Music and Comedy videos has the highest number of ratings 120k and 87.52k.
3. The video category Music and Comedy has the highest number of comments which is 38.39k and 35.40k.
4. Comedy videos has the highest number of views which is 42 million approx.

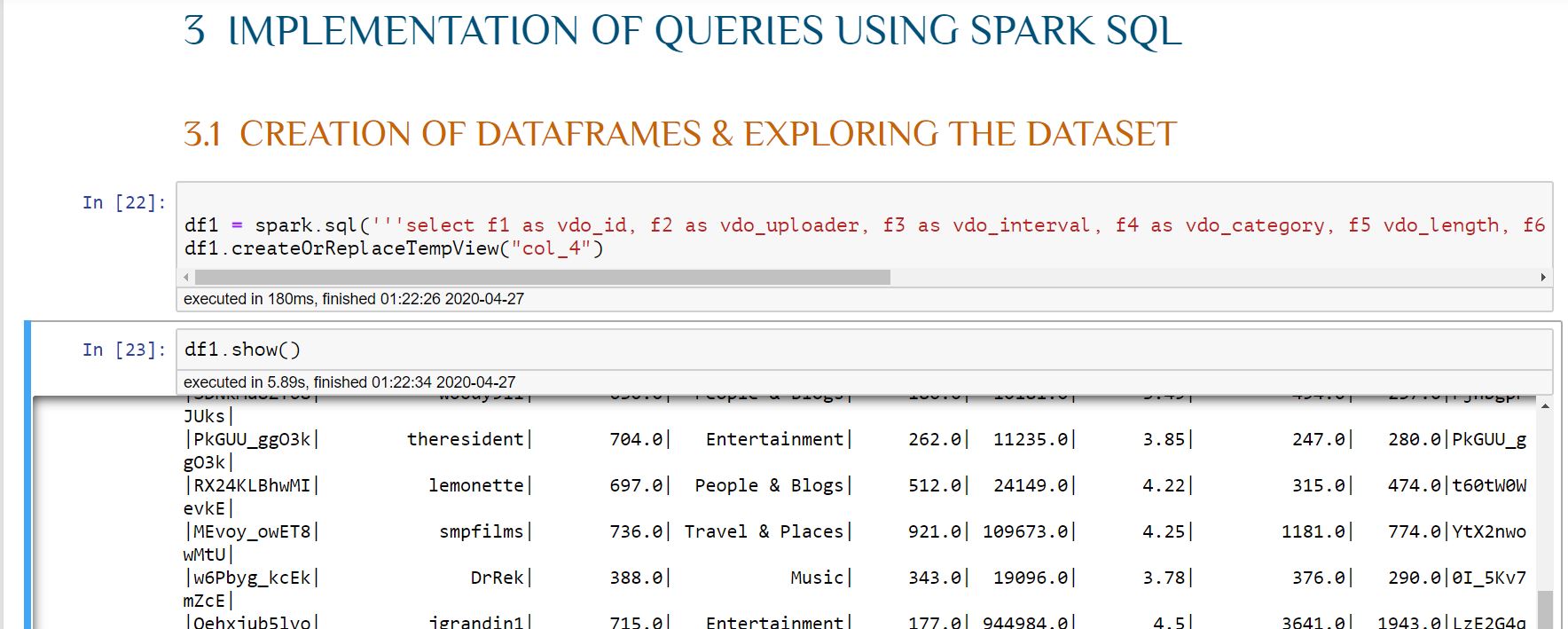
**Other Details**

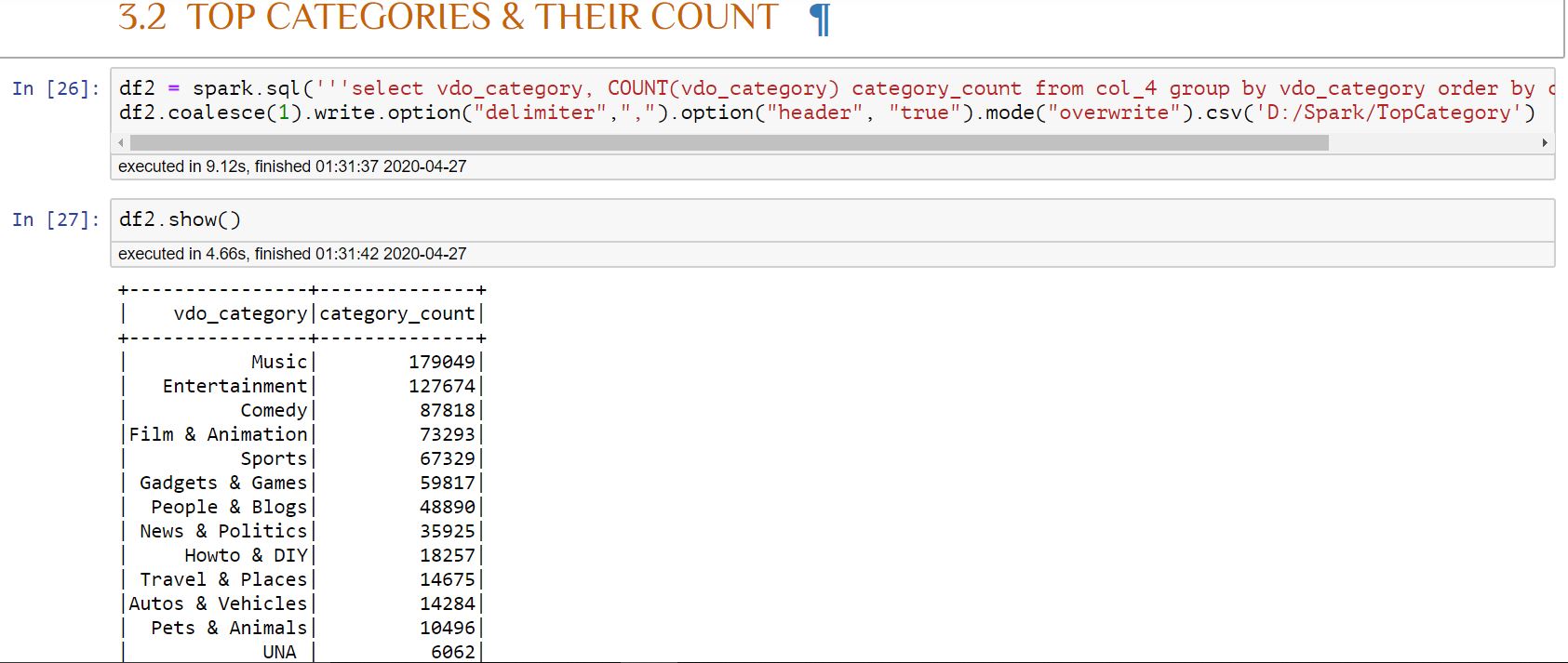
It took 62 days to complete the project, starting from 26th February 2020 and ended on 26th April 2020.

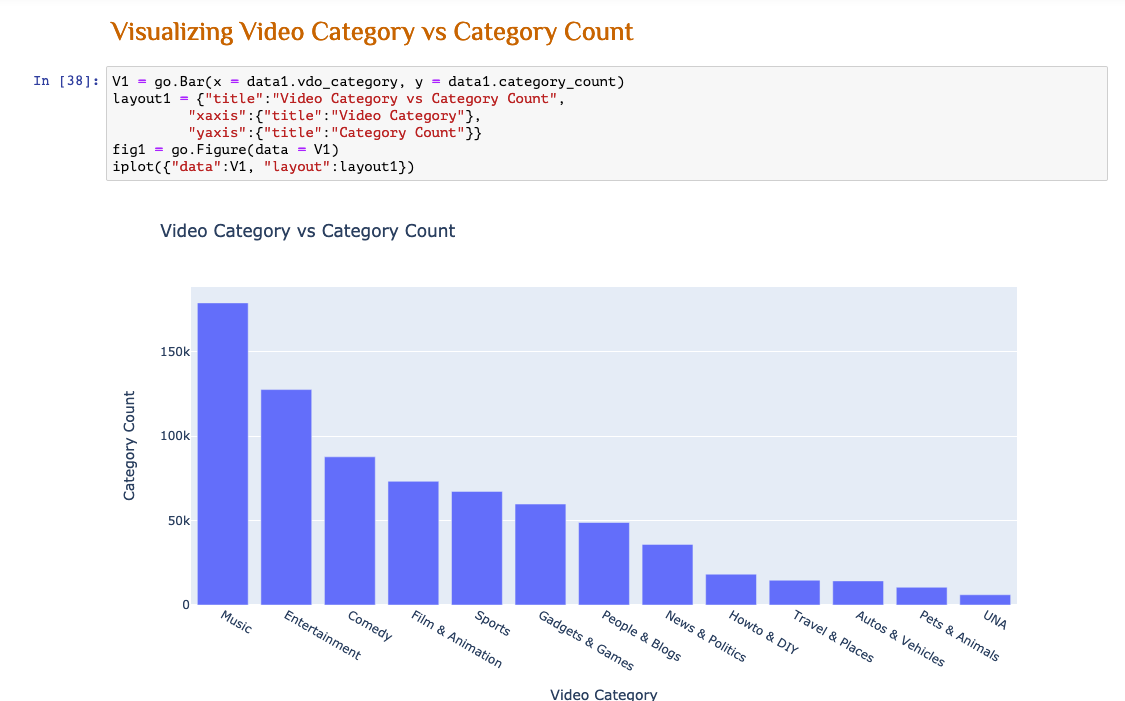
173 lines of code is written in Apache Spark and 121 lines is written in Hadoop.

There was no Monetory Cost involved in the project as open source tools were in the project.

**Apache Spark and Plotly Visualization Screenshots:**

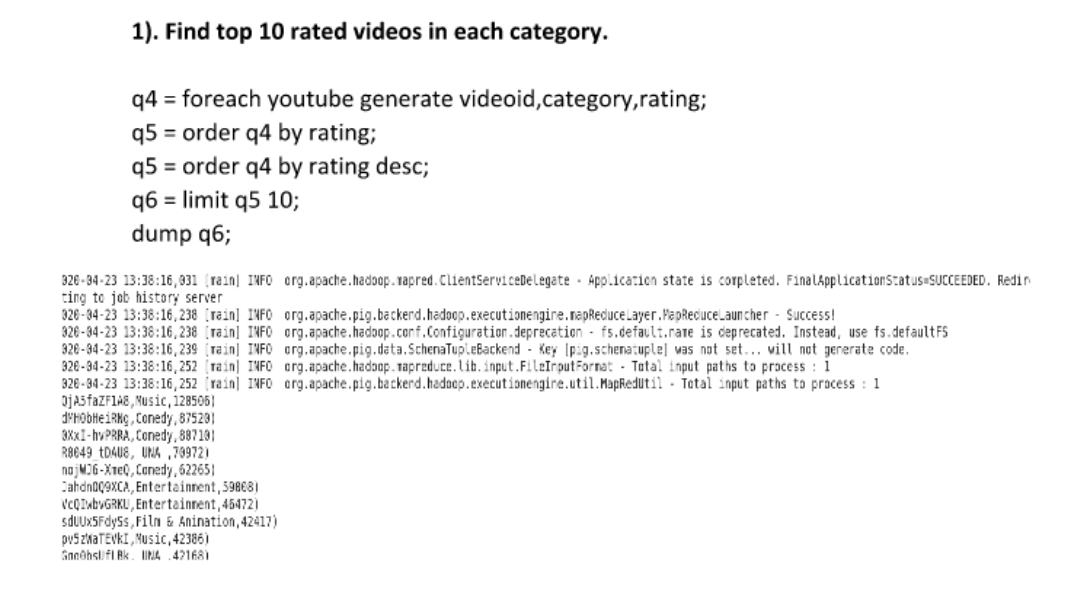


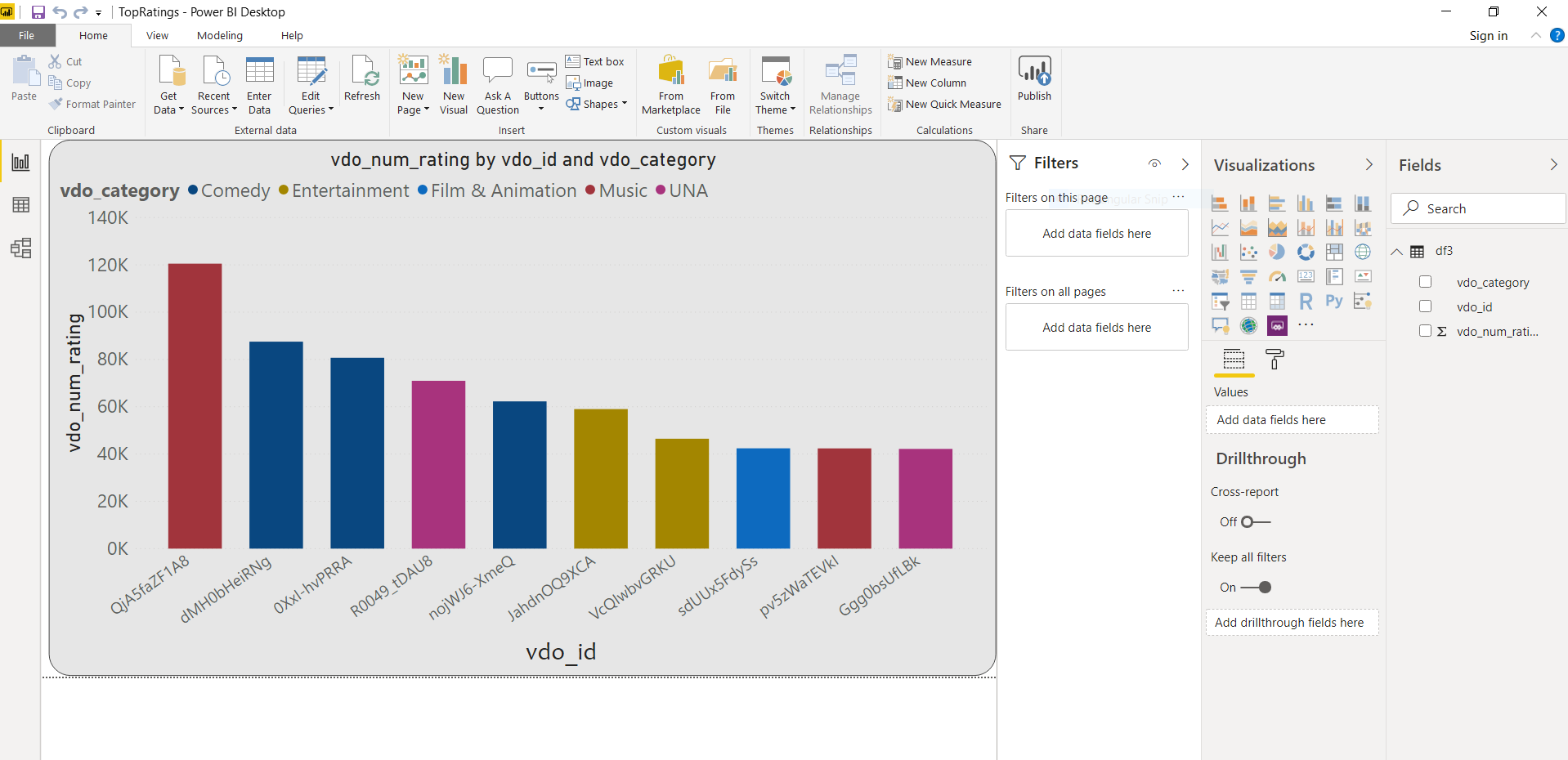


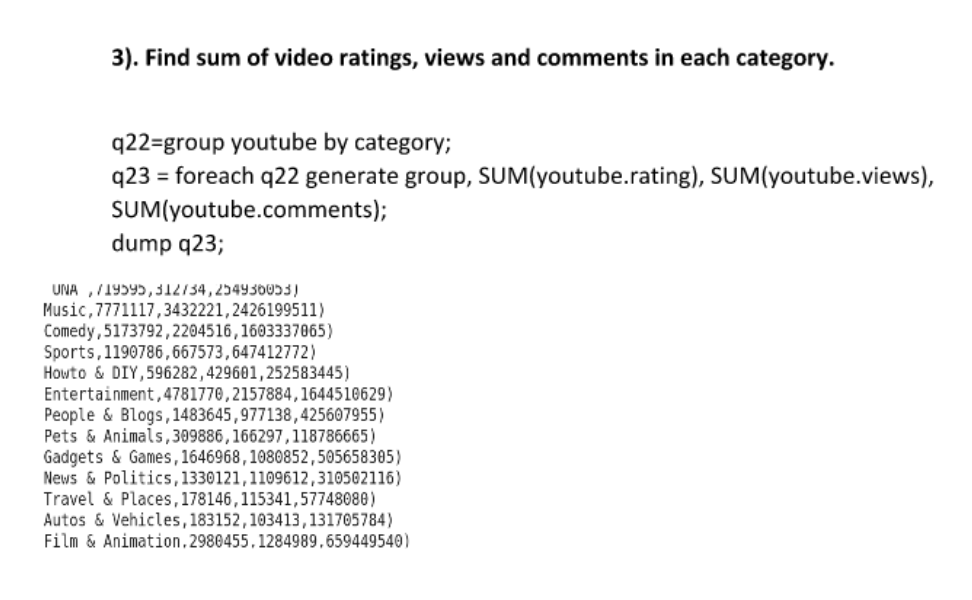
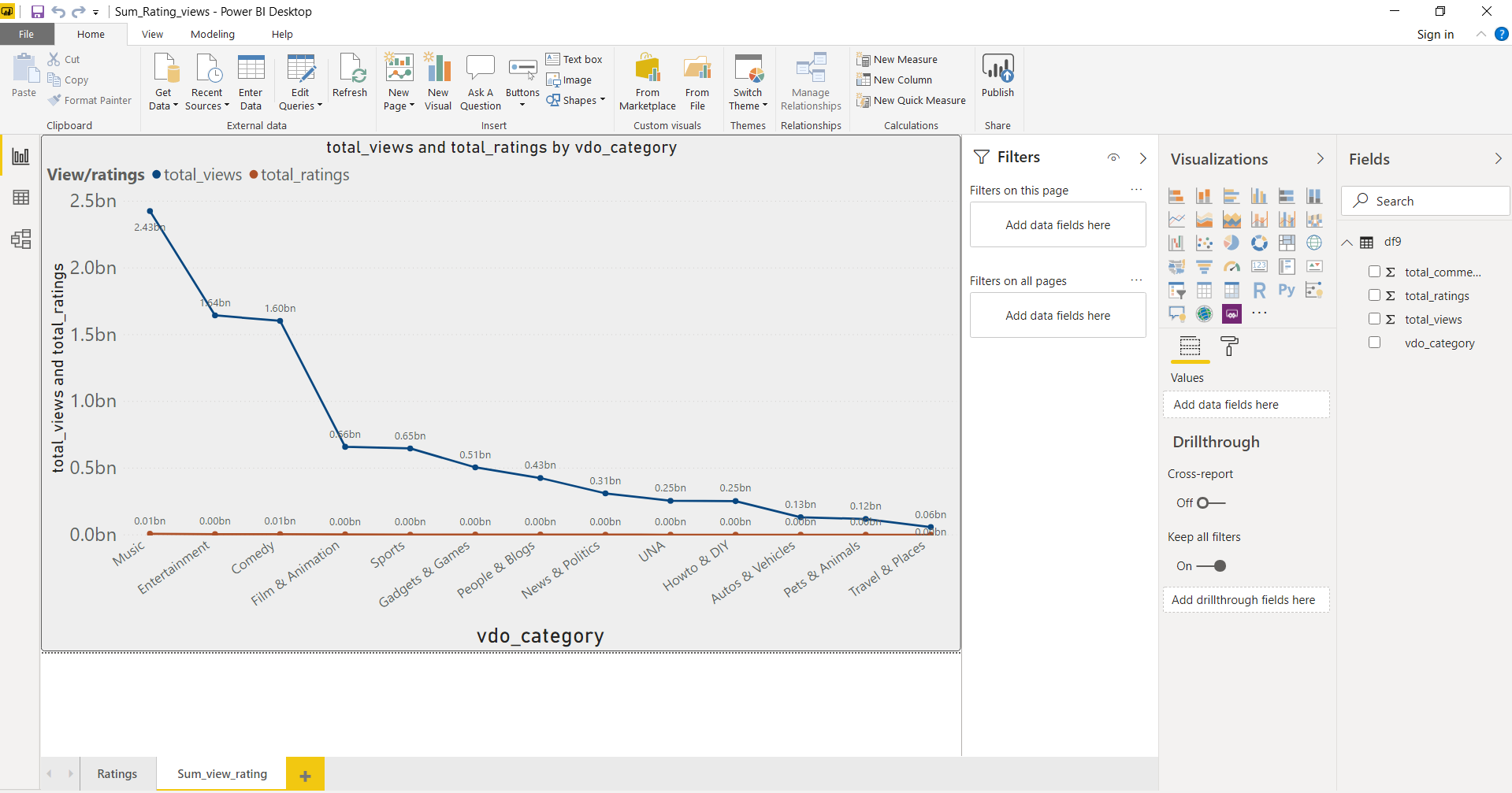


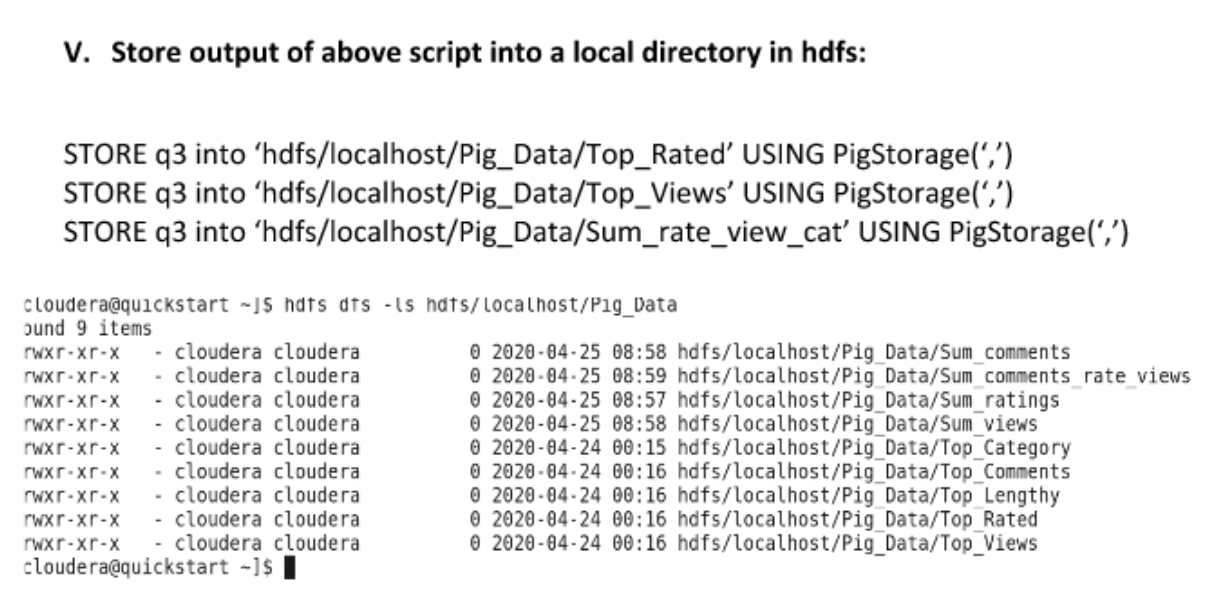
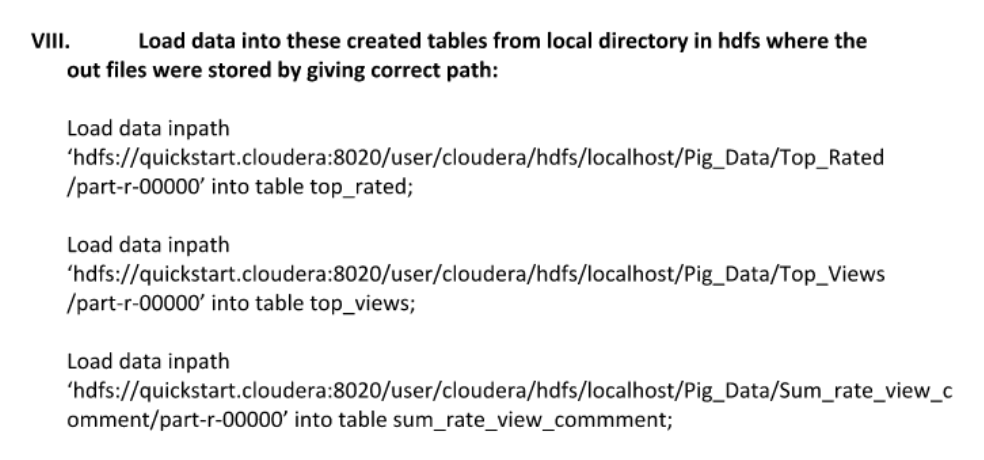
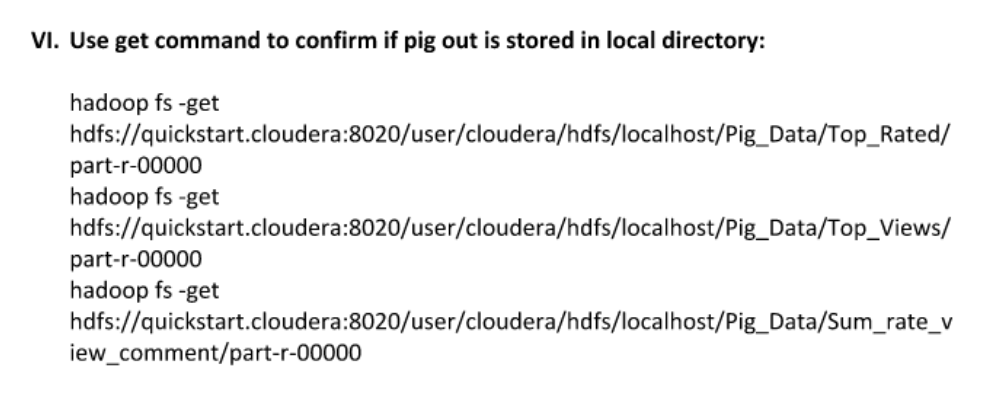
****

**Hadoop and PowerBI Screenshots:**







**Below are the lists of points, describe the comparisons Between Pig vs Spark:**

|  |  |  |
| --- | --- | --- |
| **BASIS FOR**  **COMPARISON** | **PIG** | **SPARK** |
| Availability | Open Source Framework by Apache Open Source Projects | Open source clustering framework provided by Apache Open Source projects |
| Implementation | Provided by Hortonworks and Cloudera providers etc., | A framework used for a distributed environment. |
| Performance | Provides good performance for distributed pipelines | Spark is preferred over Pig for great performance. |
| Scalability | Limitations in scalability | Faster runtimes are expected for Spark framework. |
| Pricing | Open Source and depends on the scripts efficiency | Open Source and depends on the efficiency of algorithms implemented. |
| Speed | Faster but slower compared to Spark but productive for smaller scripts | Many times Faster than Pig and provides greater runtime capacity. |
| Query Speed | Multi Query execution capacity. | Spark SQL query performance is very high with SQL Tuning. |
| Data Integration | Fast and Flexible with different tools. | Can load data and manipulate from different external applications. |
| Data Format | All data formats are supported for data operations. | Supports complex data formats such as JSON, NoSql, parquets etc. |
| Ease of Use | Easier to frame pig scripts like SQL queries. | Handles complex operations using frameworks in-built features. |

**Conclusion**

The final statement to conclude the comparison between Pig and Spark is that Spark wins in terms of ease of operations, maintenance and productivity whereas Pig lacks in terms of performance scalability and the features,

integration with third-party tools and products in the case of a large volume of data sets. As both Pig and Spark projects belong to Apache Software Foundation, both Pig and Spark are open source and can be used and integrated with Hadoop environment and can be deployed for data applications based on the amount and volumes of data to be operated upon.

In most of the cases, Spark has been the best choice to consider for the large-scale business requirements by most of the clients or customers in order to handle the large-scale and sensitive data of any financial institutions or public information with more data integrity and security.

Apart from the existing benefits Spark has its own advantages being open source project and has been evolving recently more sophistically with great clustering operational features that replace existing systems to reduce cost incurring processes and reduces the complexities and run time.

**Bibliography**

<https://developer.mozilla.org/en-US/docs/Web/CSS/color_value>

<http://ybhavesh.blogspot.com/2015/08/proof-of-concept-or-poc-on-youtube-data.html> - Dataset Link

<https://docs.microsoft.com/en-us/azure/hdinsight/hadoop/apache-hadoop-connect-hive-power-bi>

<https://community.cloudera.com/t5/Community-Articles/How-to-Connect-To-Hive-Using-ODBC-in-Tableau/ta-p/246730>

<https://community.cloudera.com/>

<https://www.tutorialspoint.com/apache_pig/index.htm>

<https://intellipaat.com/blog/tutorial/hadoop-tutorial/apache-pig/>

<https://www.tutorialspoint.com/hive/index.htm>

<https://spark.apache.org/docs/latest/api/python/pyspark.html>

**Appendix**

Challenges faced:

Due to Covid - 19 lockdown we faced many challenges like.

1. Internet issues because there was no maintenance in the company. We overcame that issue by raising a ticket online for the company and the issue was solved within 2 days.
2. Our laptops used to overheat because of heavy software used for the project and working for almost 10 hours everyday.
3. Once our Hadoop system completely crashed because of overload. So we had to again write all the codes in different system from start.
4. We faced a lot of challenges while connecting HIVE with Power BI, but we got help from Microsoft support and the problem was solved.
5. We tried connecting Tableau with HIVE but Tableau had many restrictions related to Database access privilege.

**Publications(Blog):**

Connecting Apache Hive To Microsoft Power BI

<https://medium.com/@rahulpathak1996/connecting-apache-hive-to-microsoft-power-bi-d460e2278720?sk=5eb21cdfc9d53096a26e9367667ab464>

A POC on YouTube Data Analysis using Pig & Hive

<https://medium.com/@rahulpathak1996/a-poc-for-youtube-data-analysis-using-pig-hive-e8dcb8c3cda>

**Certificates related to the project:**

IBM Data Science Professional Certificate - Python for Data Science

<https://www.coursera.org/account/accomplishments/certificate/LKCGY3EAMGBD>

<https://www.coursera.org/account/accomplishments/certificate/Q2WJ2R6DM6DV>

<https://coursera.org/share/dfd92357feba1946f2535f0644d03c49>

IBM Data Science Professional Certificate - Databases and SQL for Data Science

<https://www.coursera.org/account/accomplishments/certificate/GAVTLNFPXTY3>

<https://www.coursera.org/account/accomplishments/certificate/GV46NW6PLX3N>

<https://coursera.org/share/7e6b160bccd2ba0981f2b5a0e38a0a17>

IBM Data Science Professional Certificate - Data Visualisation with Python

<https://www.coursera.org/account/accomplishments/certificate/EHKX53VSR49G>

<https://www.coursera.org/account/accomplishments/certificate/FDBMVXS42WHF>

<https://coursera.org/share/4203909459e08973414df0848c5ef8ca>

The Ultimate Hands-On Hadoop - Tame your Big Data by Udemy

<https://www.udemy.com/certificate/UC-W9P2L15L/>

<https://www.udemy.com/certificate/UC-f0a6aba9-f7d8-4717-8234-663e6cc89bec/>