

**BIG DATA ON -**

**Web-Based Management of Apache** *Hive*

**HADOOP LAB (CBD 3113)**

**SUBMITTED BY :**

**SUBMITTED TO:**

**Mr.Mehtab Mehdi**

**ANALYSIS OF**

**YOUTUBE**

**USING APACHE HIVE**

**What is Big Data?**

Wikipedia defines Big Data as "a collection of data sets so large and complex that it becomes difficult to process using the available database management tools. The challenges include how to capture, curate, store, search, share, analyze and visualize Big Data”. In today's environment, we have access to more types of data. These data sources include online transactions, social networking activities, mobile device services, internet gaming etc.

Big Data is a collection of data sets that are large and complex in nature. They constitute both structured and unstructured data that grow large so fast that they are not manageable by traditional relational database systems or conventional statistical tools.

Big Data is defined as any kind of data source that has at least three shared characteristics:

· Extremely large Volumes of data

· Extremely high Velocity of data

· Extremely wide Variety of data

According to Big Data: Concepts, Methodologies, Tools, and Applications, Volume I by Information Resources Management Association (IRMA), "organizations today are at the tipping point in terms of managing data. Data sources are ever expanding.

Data from Facebook, Twitter, YouTube, Google etc., are to grow 50X in the next 10 years. Over 2.5 exabytes of data is generated every day. Some of the sources of huge volume of data are:

1. A typical large stock exchange captures more than 1 TB of data every day.

2. There are over 5 billion mobile phones in the world which are producing enormous amount of data on daily basis.

3. YouTube users upload more than 48 hours of video every minute.

4. Large social networks such as Twitter and Facebook capture more than 10 TB of

data daily.

5. There are more than 30 million networked sensors in the world which further produces TBs of data every day.

Structured and semi-structured formats have some limitations with respect to handling large quantities of data. Hence, in order to manage the data in the Big Data world, new emerging approaches are required, including document, graph, columnar, and geospatial database architectures. Collectively, these are referred to as NoSQL, or not only SQL, databases. In essence the data architectures need to be

mapped to the types oftransactions. Doing so will help to ensure the right data is available when you need it.

**What is Hadoop?**

As organizations are getting flooded with massive amount of raw data, the challenge here is that traditional tools are poorly equipped to deal with the scale and complexity of such kind of data. That's where Hadoop comes in. Hadoop is well suited to meet many Big Data challenges, especially with high volumes of data and data with a variety of structures.

At its core, Hadoop is a framework for storing data on large clusters of commodity hardware — everyday computer hardware that is affordable and easily available — and running applications against that data. A cluster is a group of interconnected computers (known as nodes) that can work together on the same problem. Using networks of affordable compute resources to acquire business insight is the key value proposition of Hadoop.

Hadoop consists of two main components:-

1. A distributed processing framework named MapReduce (which is now supported by a component called YARN(Yet Another Resource Negotiator) and

2. A distributed file system known as the Hadoop Distributed File System, or HDFS. In Hadoop you can do any kind any kind of aggregation of data whether it is one month old data or one-year-old data. Hadoop provides a mechanism called MapReduce model to do distributed processing of large data which internally takes care of data even if one machine goes down.

**Hadoop Ecosystem**

Hadoop is a shared nothing system where each node acts independently throughout the system. A framework where a piece of work is divided among several parallel MapReduce task. Each task operated independently on cheap commodity servers. This enables businesses to generate values from data that was previously considered too expensive to be stored and processed in a traditional data warehouse or OLTP (Online Transaction Processing) environment. In the old paradigm, companies would use a traditional enterprise data warehouse system and would buy the biggest data warehouse they could afford and store the data on a single machine. However, with the increasing amount of data, this approach is no longer affordable nor practical. Some of the components of Hadoop ecosystem are HDFS (Hadoop Distributed File System), MapReduce, Yarn, Hive and Hbase. Hadoop has two core components. ‘Storage’ part to store the data and ‘Processing’ part to process the data. The storage part is called ‘HDFS’ and the processing part is called as ‘YARN’.

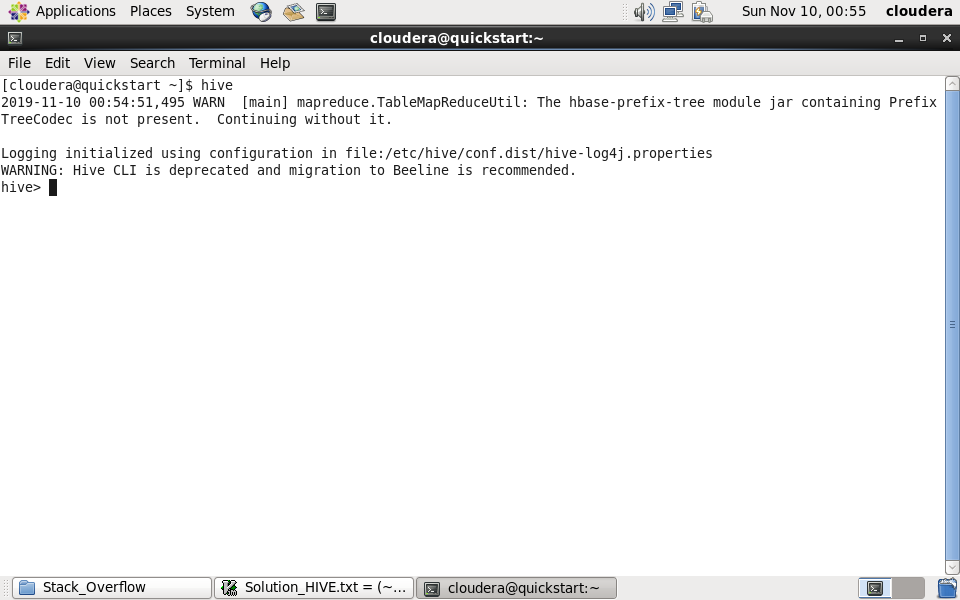
**TECHNOLOGIES USED IN PROJECT**

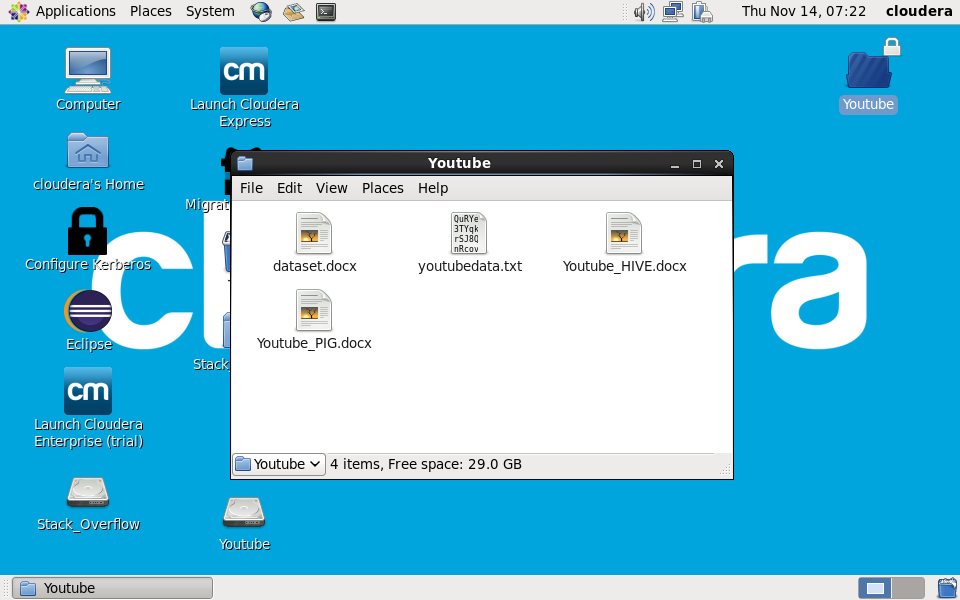
**APACHE HIVE:- Apache Hive** is a [data warehouse](https://en.wikipedia.org/wiki/Data_warehouse) software project built on top of [Apache Hadoop](https://en.wikipedia.org/wiki/Apache_Hadoop) for providing data summarization, query and analysis.[[2]](https://en.wikipedia.org/wiki/Apache_Hive#cite_note-2) Hive gives an [SQL](https://en.wikipedia.org/wiki/SQL)-like interface to query data stored in various databases and file systems that integrate with Hadoop. Traditional SQL queries must be implemented in the [MapReduce](https://en.wikipedia.org/wiki/MapReduce) Java API to execute SQL applications and queries over distributed data. Hive provides the necessary SQL abstraction to integrate SQL-like queries ([HiveQL](https://en.wikipedia.org/wiki/Apache_Hive#HiveQL)) into the underlying Java without the need to implement queries in the low-level Java API. Since most data warehousing applications work with SQL-based querying languages, Hive aids portability of SQL-based applications to Hadoop.While initially developed by [Facebook](https://en.wikipedia.org/wiki/Facebook), Apache Hive is used and developed by other companies such as [Netflix](https://en.wikipedia.org/wiki/Netflix) and the [Financial Industry Regulatory Authority](https://en.wikipedia.org/wiki/Financial_Industry_Regulatory_Authority) (FINRA).Amazon maintains a software fork of Apache Hive included in [Amazon Elastic MapReduce](https://en.wikipedia.org/wiki/Apache_Hadoop#Amazon_Elastic_MapReduce) on [Amazon Web Services](https://en.wikipedia.org/wiki/Amazon_Web_Services).

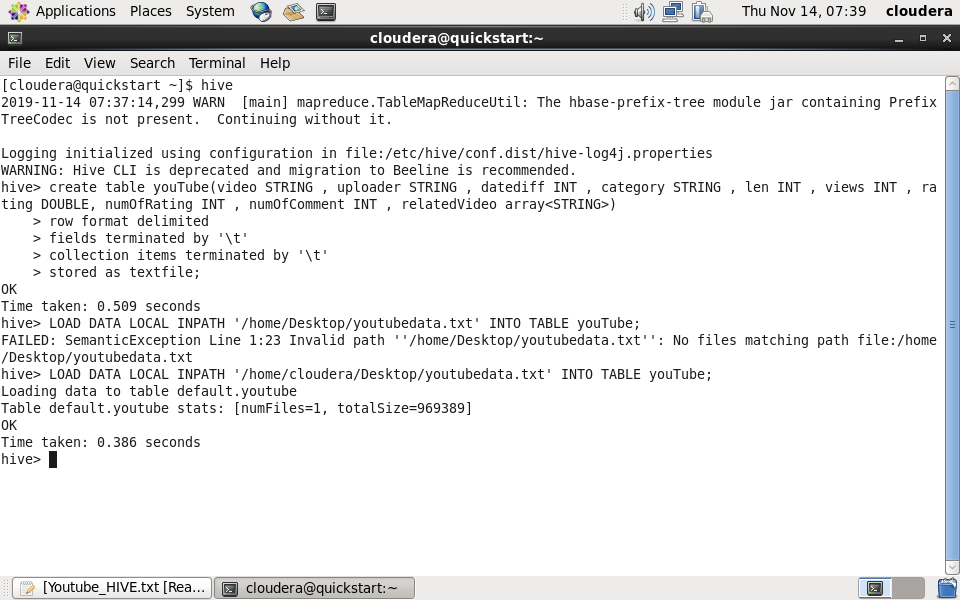


**First Step (Historical Load)**

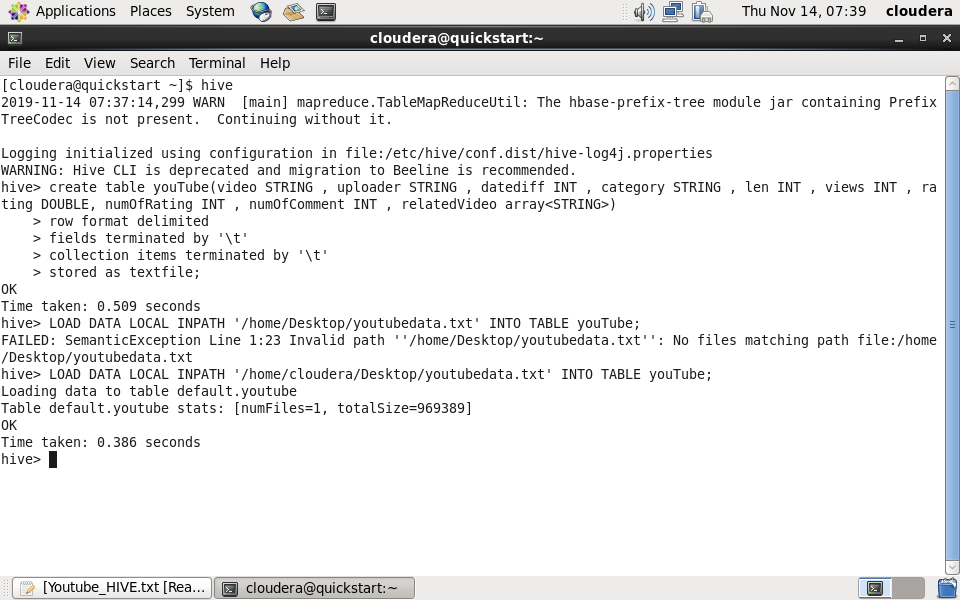
1. Connect to HIVE from Cloudera.



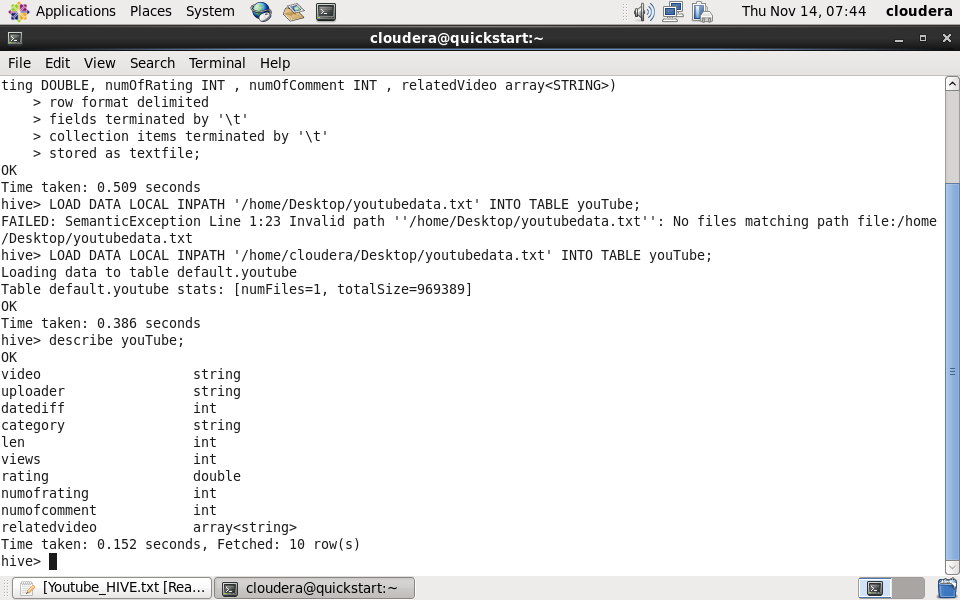
2) Create Table for the datasets given for the analysis of Stack Overflow



3) Loading the dataset present on Desktop of Cloudera into the HIVE TABLES created .



(You can also describe the structure of tables created in HIVE)

****

**Problem Statement**

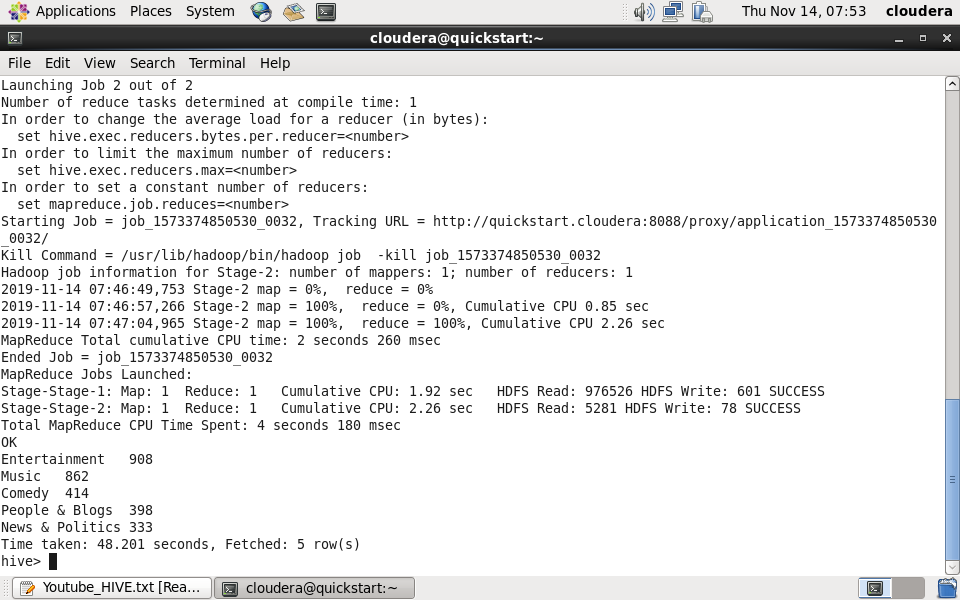
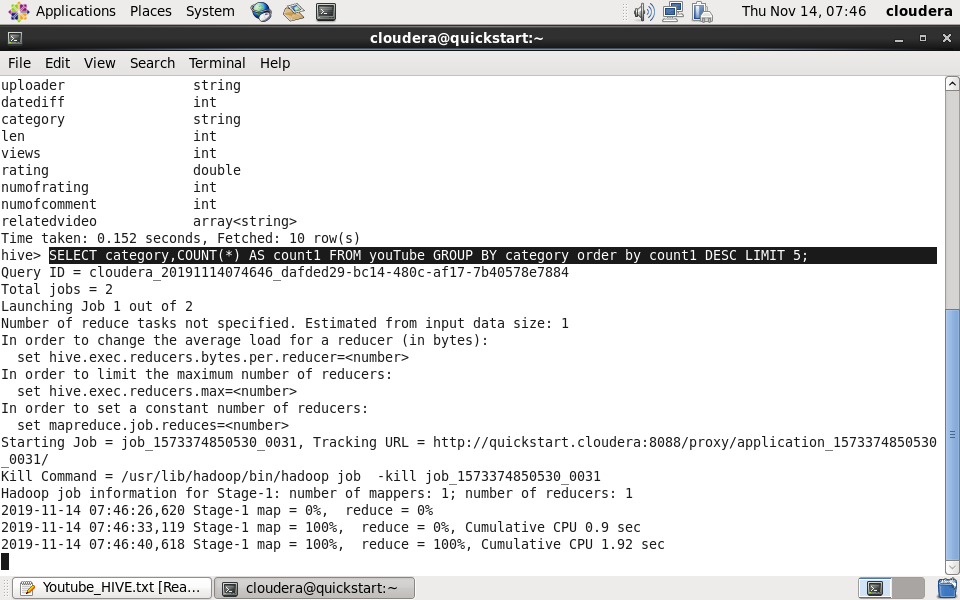
1. Find out the top 5 categories with maximum number of videos uploaded.
2. Find out the top 10 rated videos.
3. Find out the most viewed videos.

1. **Find out the top 5 categories with maximum number of videos uploaded.**

SELECT category,COUNT(\*) AS count1 FROM youTube GROUP BY category order by count1 DESC LIMIT 5;

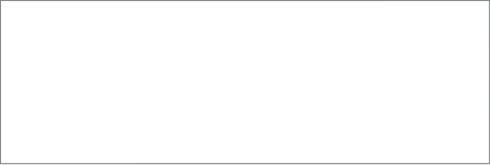
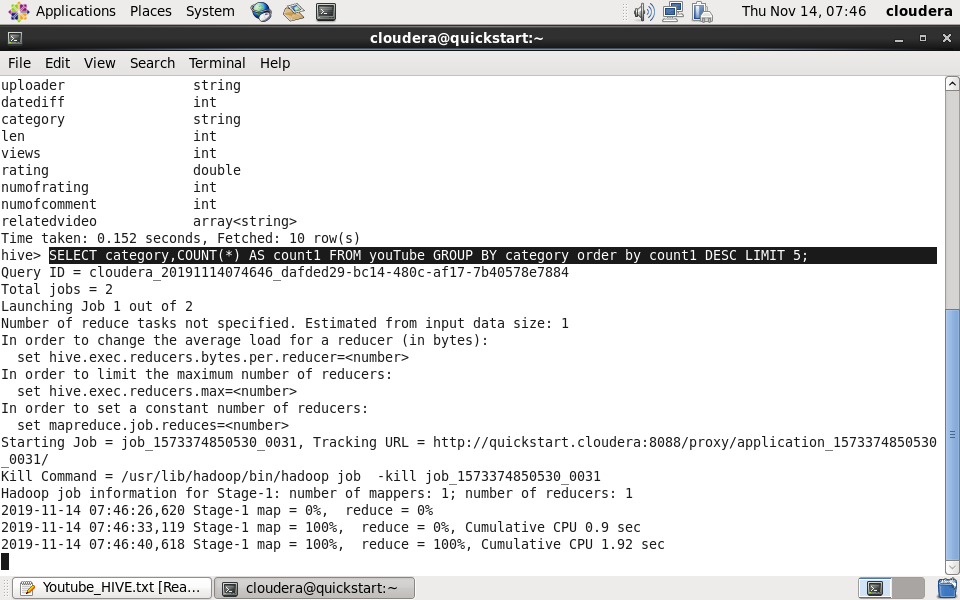
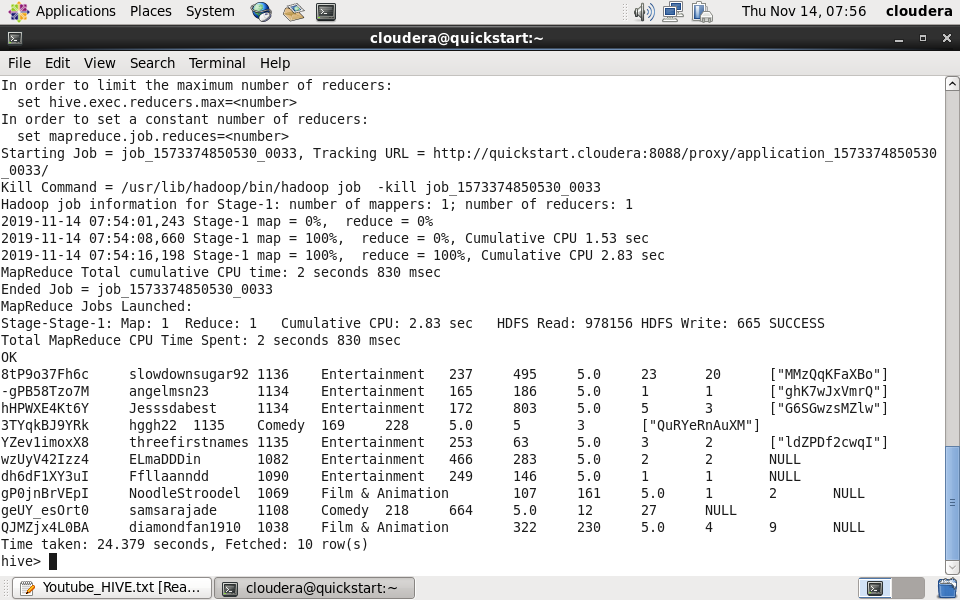
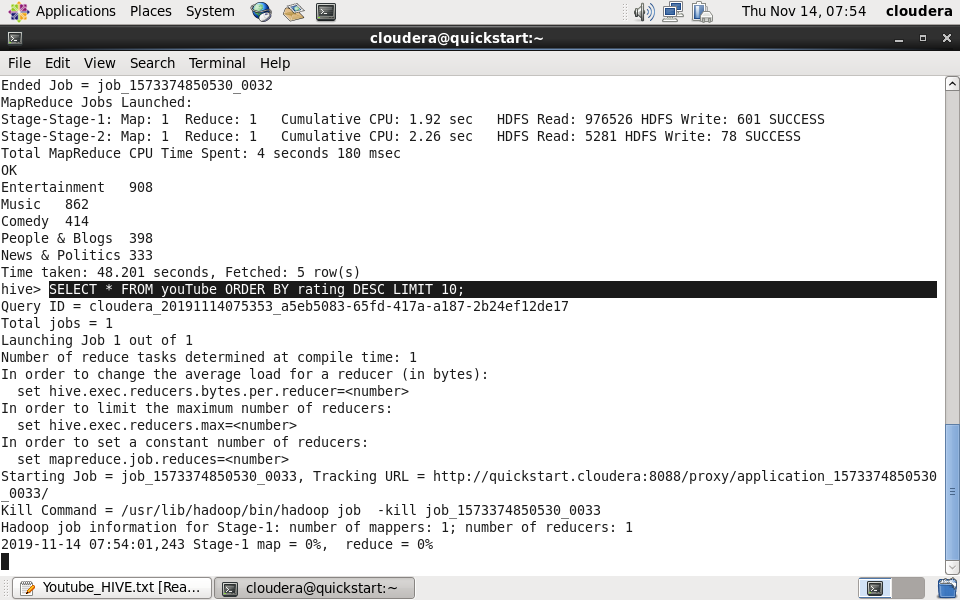
Find out the most viewed videos.

3) SELECT \* FROM youTube ORDER BY views DESC LIMIT 10;



**2. Find out the top 10 rated videos.**

SELECT \* FROM youTube ORDER BY rating DESC LIMIT 10;



**3. Find out the most viewed videos.**

SELECT \* FROM youTube ORDER BY views DESC LIMIT 10;

