ANALYSIS OF MOVIES DATABASE

A report submitted in partial fulfilment of the requirements for

The award of the degree of

B.Tech in

Computer Science and Engineering

Submitted by:

Shivam Agrawal

Roll No. : 1302081048

Batch : 2013-17

COMPUTER SCIENCE AND ENGINEERING

DEPARTMENT

DIT UNIVERSITY

Annexure 2

CANDIDATE/S DECLARATION

I hereby certify that the work, which is being presented in the industrial training report, entitled **Analysis Of Movies Database**, in partial fulfilment of the requirement for the award of the Degree of **Bachelor of Technology** and submitted to the institution is an authentic record of my/our own work carried out during the period *June-2016* to *July-2016* under the supervision of supervisor(s) name. I/we also cited the reference about the text(s)/figure(s)/table(s) from where they have been taken.

Date: Signature of the Candidate

This is to certify that the above statement made by the candidate is correct to the best of my /our knowledge.

Date: Signature(s) of the Supervisor (s)

All India Council for Robotics and Automation

(In association with Times Globacom (P) Ltd. and IIT Kharagpur**)**

All India Council for Robotics & Automation (AICRA) is a leading, global and non-profit organization that is setting the standard for robotics and automation by helping over 35,000 worldwide members and other professionals to solve difficult technical problems, while enhancing their leadership and personal career capabilities. It aims to promote robotics, automations and other new technologies skills development by catalyzing creation of large, quality, for profit vocational institutions. AICRA provides technical support systems to institutions such as quality assurance, Information systems and train the trainer (TTT) academies either directly or through partnerships. To strengthen supplementary skill development, AICRA focuses on fostering private sector led efforts that include both non-profit and for-profit initiatives with the goal of building models that are scalable.

The executive board of AICRA is the senior governing body that's composed of the President, President – elect Secretary, Vice President, Treasurer, 8 members with geographic, technical, and operational experience, the Parliamentarian (non-voting), and up to 5 at-large, competency-based members.

The Executive Board works closely with the AICRA staff, made up of dozens of seasoned experts in areas like non-profit management, finance, event planning, marketing, training, publishing and more.

Vision

AICRA was set up to promote, setting up standards for advance robotics & automation industry as well as developing skill set of latest technologies as part of skill development mission to fulfill the growing need in India for skilled manpower across sectors and narrow the existing gap between the demand and supply of technical skills. The Union Finance Minister announced in his Budget Speech (2008-09): "...There is a compelling need to launch a world-class skill development programme in a mission mode that will address the challenge of imparting the skills required by a growing economy. Both the structure and the leadership of the mission must be such that the programme can be scaled up quickly to cover the whole country."

Mission

Upgrade Robotics & Automation skills to international standards through significant industry involvement and develop necessary frameworks for standards, curriculum and quality assurance

Enhance, support and coordinate private sector initiatives for skill development through appropriate engagement models; strive for significant operational and financial involvement from the private sector

Play the role of a "market-maker" by bringing financing, particularly in sectors where market mechanisms are ineffective or missing. Prioritize initiatives that can have a multiplier or catalytic effect as opposed to one-off impact.

Abstract

The project is based on the analysis of Movies database that consists of details of all the movies released till date including both Hollywood and Bollywood. During the project an analysis was performed to segregate the movies in to different categories depending on there income, year of release, count of the number of screens on which the movies launched etc. using some of the very famous technologies of the Big Data technology such as Hadoop, Pig, Scoop, Hive and HBase.

Hadoop and Hive enabled the processing of the data. Hbase was used to store data to the database or to retrieve a subset of the data on to the local machine. Pig enabled the querying of the database to retrieve the data.

After the analysis of the data, a descriptive chart was prepared which depicted the criteria and the variation that took place on that criteria according to all the analysis of the database.

Acknowledgement

I have taken efforts in this project. However, it would not have been possible without the kind support and help of many individuals and organizations. I would like to extend my sincere thanks to all of them.

I am highly indebted to Mr. Dharmendra for his guidance and constant supervision as well as for providing necessary information regarding the project & also for their support in completing the project.

I would like to express my gratitude towards my parents & member of **All India Council for Robotics and Automation** for their kind co-operation and encouragement which help me in completion of this project.

I would like to express my special gratitude and thanks to industry persons for giving me such attention and time.

My thanks and appreciations also go to my colleague in developing the project and people who have willingly helped me out with their abilities.

Table of Contents

List of Figures

Chapter 1

Introduction

1.1. Purpose: Movies are an integral part of the entertainment market. Even though there are many other sources to get entertained such as Games, Songs, and Amusement parks etc. but, movies are preferred by most of the people to loosen their minds from the day to day stress.

Not only that movies are of great importance to the general people for entertainment but they even are of great importance to the government as movies share a great part of the profit that government make through the entertainment stuffs.

With the growing demand for watching movies by the public the importance to have a control on the sector has also increased.

The project majorly does that part of analysis of a database which contains a record of movies that are released till date in various countries. The analysis is done on various parameters in which the rating is the prime.

1.2. Objective: To analyze the ups and downs in the movies, released in various countries based on various factors such as year of release, rating, cost and in the end to prepare a brief graph depicting the scenario on the movies in the market with prime analysis of Bollywood and Hollywood database.

1.3. Motivation: With the growing demand for movies the number of movies produced is also increasing day by day. And hence the database containing the data about these movies is also growing to an alarming rate. The analysis of this data is a challenging task as the amount of data cannot be analyzed by traditional databases.

The demand to mine important facts and trends in the movies and there demand were the prime reasons that motivated towards this project.

1.4. Definition and Overview: Today, we're surrounded by data. People upload videos, take pictures on their cell phones, text friends, update their Facebook status, leave comments around the web, click on ads, and so forth. Machines, too, are generating and keeping more and more data. The exponential growth of data first presented challenges to cutting-edge businesses such as Google, Yahoo, Amazon, and Microsoft. They needed to go through terabytes and petabytes of data to figure out which websites were popular, what books were

in demand, and what kinds of ads appealed to people. Existing tools were becoming inadequate to process such large data sets. Google was the first to publicize MapReduce—a system they had used to scale their data processing needs. This system aroused a lot of interest because many other businesses were facing similar scaling challenges, and it wasn't feasible for everyone to reinvent their own proprietary tool. Doug Cutting saw an opportunity and led the charge to develop an open source version of this MapReduce system called Hadoop . Soon after, Yahoo and others rallied around to support this effort. Today, Hadoop is a core part of the computing infrastructure for many web companies, such as Yahoo , Facebook , LinkedIn , and Twitter. Many more traditional businesses, such as media and telecom, are beginning to adopt this system too. Hadoop is an open source framework for writing and running distributed applications that process large amounts of data. Distributed computing is a wide and varied field, but the key distinctions of Hadoop are that it is

1.4.1. **Accessible:** Hadoop runs on large clusters of commodity machines or on cloud computing services such as Amazon's Elastic Compute Cloud (EC2).

1.4.2. **Robust:** Because it is intended to run on commodity hardware, Hadoop is architected with the assumption of frequent hardware malfunctions. It can gracefully handle most such failures.

1.4.3. **Scalable:** Hadoop scales linearly to handle larger data by adding more nodes to the cluster.

1.4.4. **Simple:** Hadoop allows users to quickly write efficient parallel code. Hadoop's accessibility and simplicity give it an edge over writing and running large distributed programs. Even college students can quickly and cheaply create their own Hadoop cluster. On the other hand, its robustness and scalability make it suitable for even the most demanding jobs at Yahoo and Facebook. These features make Hadoop popular in both academia and industry.

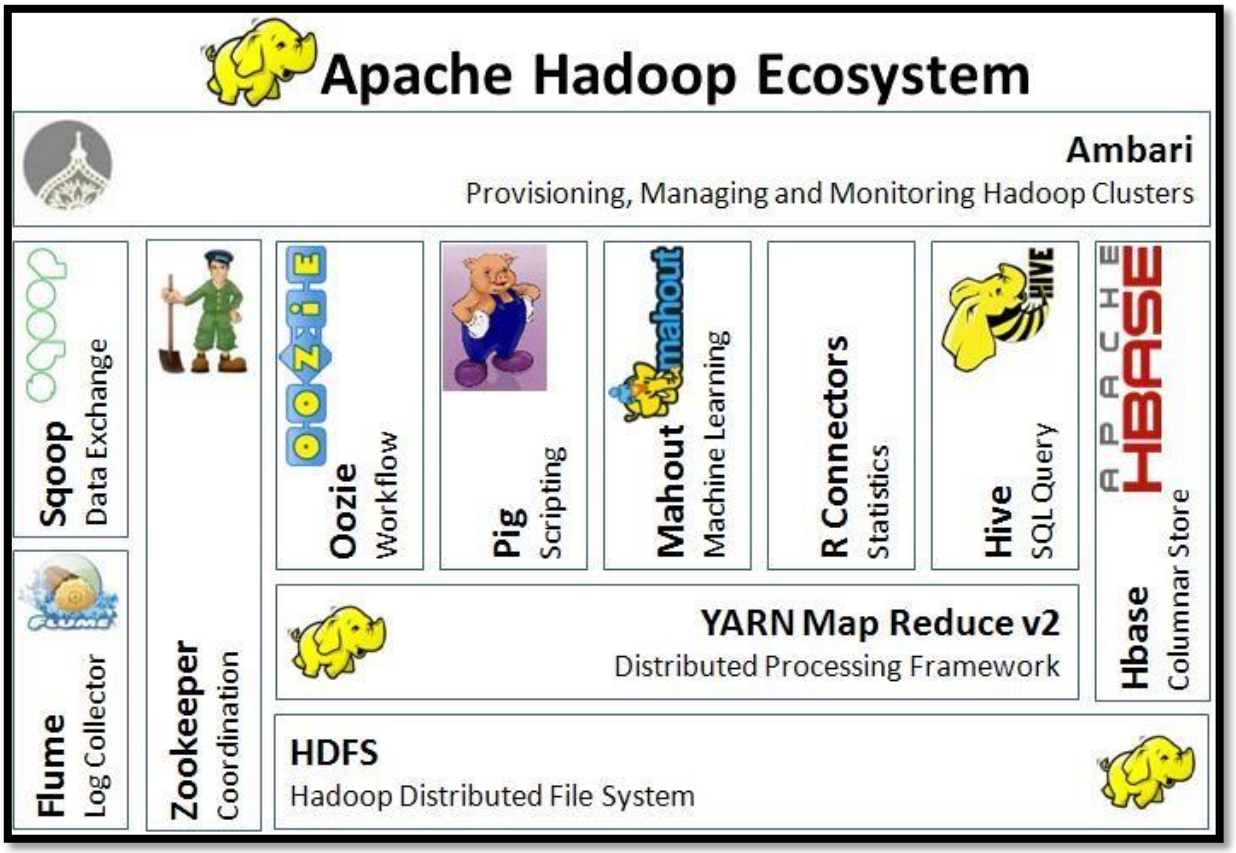


Figure 1.1. - Apache Hadoop Ecosystem

Chapter 2

System Requirements

2.1. External Interface Requirement:

2.1.1. Hardware Interface:

Intel Core i3

40GB Hard disc.

4 GB RAM

System with all standard accessories like monitor, keyboard, mouse, etc.

2.1.2. Software Interface:

Linux Operating System.

Apache Hadoop.

Mozilla Firefox: (or any browser).

Microsoft Excel or Open office.

MySQL

2.2. Functional Requirements:

2.2.1. Technical Feasibility:

Evaluating the technical feasibility is the trickiest part of a feasibility study. This is because, at this point in time, not too many detailed design of the system, making it difficult to access issues like performance, costs on (on account of the kind of technology to be deployed) etc.

A number of issues have to be considered while doing a technical analysis. Understand the different technologies involved in the proposed system.

Before commencing the project, we have to be very clear about what are the technologies that are to be required for the development of the new system.

Find out whether the organization currently possesses the required technologies. Is the required technology available with the organization?

If so is the capacity sufficient?

For instance –"Will the current printer be able to handle the new reports and forms required for the new system?"

2.2.2. Operational Feasibility

Proposed projects are beneficial only if they can be turned into information systems that will meet the organizations operating requirements. Simply stated, this test of feasibility asks if the system will work when it is developed and installed. Are there major barriers to Implementation? Here are questions that will help test the operational feasibility of a project.

2.2.2.1. Is there sufficient support for the project from management from users? If the current system is well liked and used to the extent that persons will not be able to see reasons for change, there may be resistance.

2.2.2.2. Are the current business methods acceptable to the user? If they are not, Users may welcome a change that will bring about a more operational and useful systems.

2.2.2.3. Have the user been involved in the planning and development of the project? Early involvement reduces the chances of resistance to the system and in General and increases the likelihood of successful project.

Since the proposed system was to help reduce the hardships encountered in the existing manual system, the new system was considered to be operational feasible.

Chapter 3

Big Data

Big data is a popular term used to describe the exponential growth and availability of data, both structured, unstructured and Semi Structured. And big data may be as important to business – and society – as the Internet has become.

- Lots of Data (Terabytes or Petabytes)
- 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. The challenges include capture, storage, search, sharing, transfer, analysis, and visualization.
- Systems / Enterprises, Internet users, generate huge amount of data from Terabytes to and even Petabytes of information.

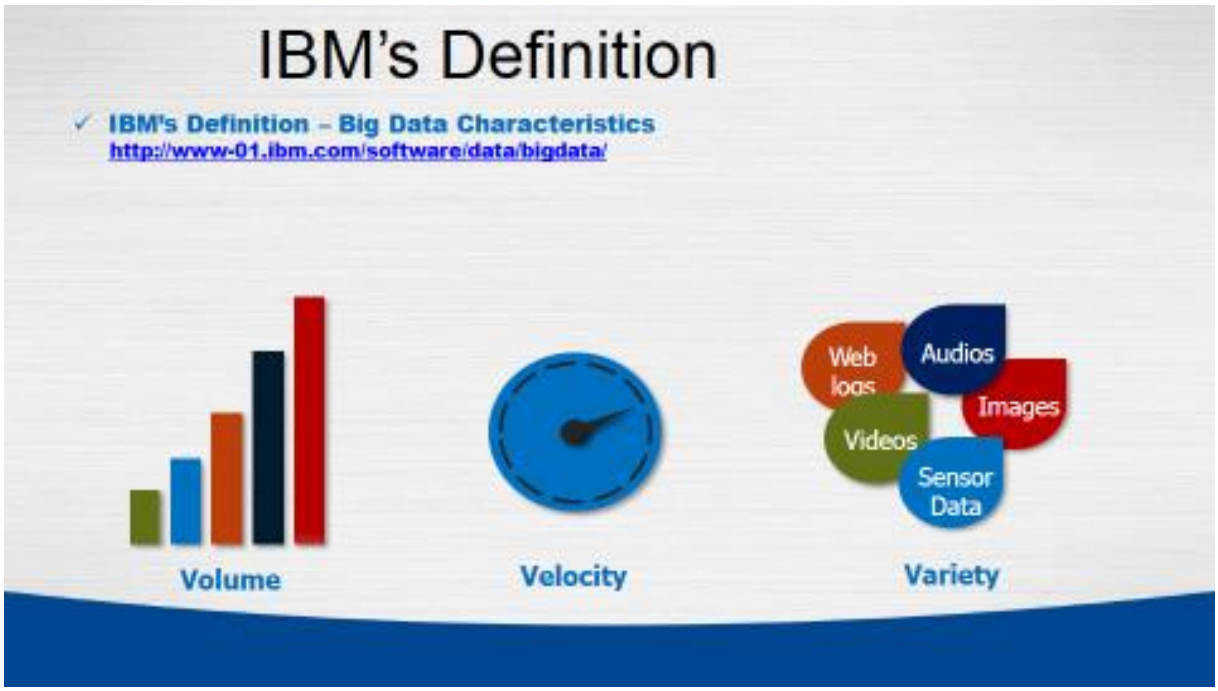


Figure 3.1 IBM Definition

- **Volume:** Many factors contribute to the increase in data volume. Transaction-based data stored through the years. Unstructured data streaming in from social media. Increasing amounts of sensor and machine-to-machine data being collected. In the past, excessive data volume was a

storage issue. But with decreasing storage costs, other issues emerge, including how to determine relevance within large data volumes and how to use analytics to create value from relevant data.

- **Velocity.** Data is streaming in at unprecedented speed and must be dealt with in a timely manner. RFID tags, sensors and smart metering are driving the need to deal with torrents of data in near-real time. Reacting quickly enough to deal with data velocity is a challenge for most organizations.
- **Variety.** Data today comes in all types of formats. Structured, numeric data in traditional databases. Information created from line-of-business applications. Unstructured text documents, email, video, audio, stock ticker data and financial transactions. Managing, merging and governing different varieties of data is something many organizations still grapple with.

Figure 3.2 Big Data Customers

3.1. Big Data challenges:

3.1.1. Need for speed

Now This Time hypercompetitive business environment, companies not only have to find and analyze the relevant data they need, they must find it quickly. Visual-ization helps organizations perform analyses and make decisions much more rapidly, but the challenge is going through the sheer volumes of data and accessing the level of detail needed, all at a high speed.

3.1.2. Data quality

Analyze data quickly and put it in the proper context for the audience that will be consuming the information, the value of data for decision-making purposes if the data is not accurate or timely. This is a challenge with any data analysis, but when considering the volumes of information involved in big data projects, it becomes even more pronounced.

3.1.3. Understanding the data

It takes a lot of understanding to get data in the right shape so that you can use visualization as part of data analysis For example, if the data comes from social media content, you need to know who the user is in a general sense – such as a customer using a particular set of products – and understand what it is you're trying to visualize out of the data. Without some sort of context, visualization tools are likely to be of less value to the user.

3.2. Types of Data

There are three types of Data

3.2.1. Unstructured data

Unstructured data files often include text and multimedia content. Examples include e-mail messages, word processing documents, videos, photos, audio files, presentations, webpages and many other kinds of business documents. Note that while these sorts of files may have an internal structure, they are still considered "unstructured" because the data they contain doesn't fit neatly in a database. Experts estimate that 80 to 90 percent of the data in any organization is unstructured. And the amount of unstructured data in enterprises is growing significantly often many times faster than structured databases are growing.

Unstructured data is all those things that can't be so readily classified and fit into a neat box: photos and graphic images, videos, streaming instrument data, webpages, PDF files, PowerPoint presentations, emails, blog entries, wikis and word processing documents.

3.2.2. Structured data

Data that resides in a fixed field within a record or file is called structured data. This includes data contained in relational databases and spreadsheets.

3.2.3. Semi-Structured Data

Semi-structured data is a cross between the two. It is a type of structured data, but lacks the strict data model structure. XML is example.

3.3. Characteristics of Structured Data

Structured data first depends on creating a data model – a model of the types of business data that will be recorded and how they will be stored, processed and accessed. This includes defining what fields of data will be stored and how that data will be stored: data type (numeric, currency, alphabetic, name, date, address).

Chapter 4

Big data tools

Software like Hadoop can process stores of both unstructured and structured data that are extremely large, very complex and changing rapidly.

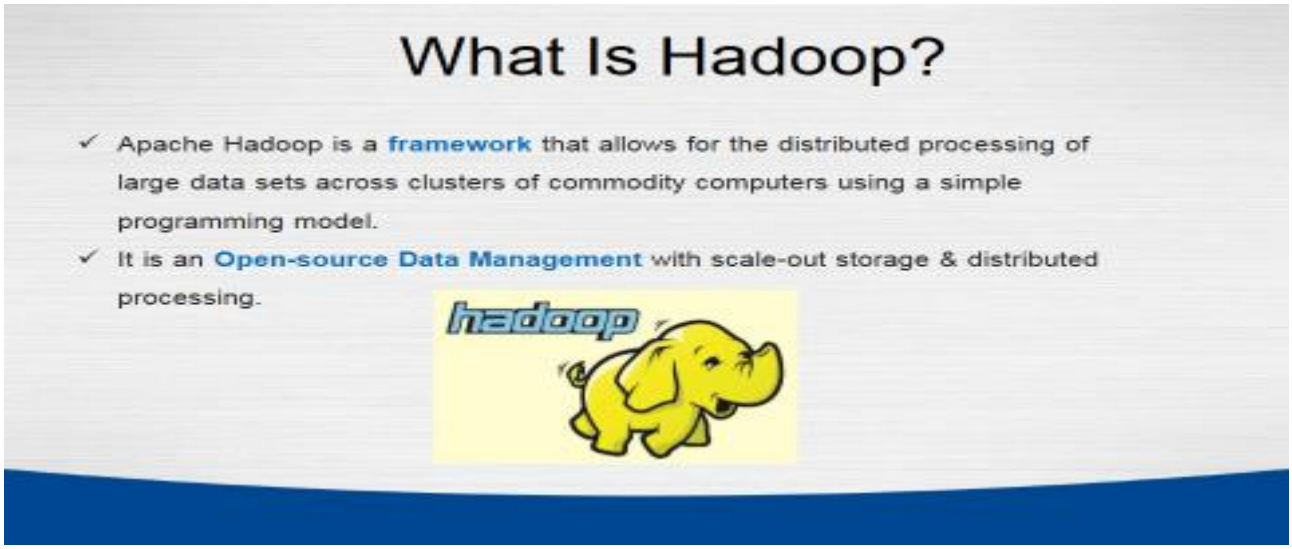


Figure 4.1 What is Hadoop

4.1. What is Hadoop?

Apache Hadoop is an open-source software framework written in Java for distributed storage and distributed processing of very large data sets on computer clusters built from commodity hardware. All the modules in Hadoop are designed with a fundamental assumption that hardware failures (of individual machines, or racks of machines) are commonplace and thus should be automatically handled in software by the framework.

4.1.1. History Of Hadoop

Google invented the basic frameworks that constitute what is today popularly called as Hadoop. They faced the future first with the problem of handling billions of searches and indexing millions of web pages. When they could not find any large scale, distributed, scalable computing platforms for their needs, they just went ahead and created their own.

Doug Cutting was inspired by Google's white papers and decided to create an open source project called "Hadoop".

Doug Cutting, Cloudera's Chief Architect, helped create Apache Hadoop out of necessity as data from the web exploded, and grew far beyond the ability of traditional systems to handle it.

Yahoo further contributed to this project and played a key role in developing Hadoop for enterprise applications. Since then many companies such as Facebook, Linkedin, ebay, Hortonworks, Cloudera etc have contributed to the Hadoop project.

- Hadoop provides a reliable shared storage (HDFS) and analysis system (MapReduce).
- Hadoop is highly scalable and unlike the relational databases, Hadoop scales linearly. Due to linear scale, a Hadoop Cluster can contain tens, hundreds, or even thousands of servers.
- Hadoop is very cost effective as it can work with commodity hardware and does not require expensive high-end hardware.

4.2.When To Use Hadoop

4.2.1. Data Size and Data Diversity

When you are dealing with huge volumes of data coming from various sources and in a variety of formats then you can say that you are dealing with Big Data. In this case, Hadoop is the right technology for you.

4.2.2. Future Planning

It is all about getting ready for challenges you may face in future. If you anticipate Hadoop as a future need then you should plan accordingly. To implement Hadoop on you data you should first understand the level of complexity of data and the rate with which it is going to grow. So, you need a cluster planning. It may begin with building a small or medium cluster in your industry as per data (in GBs or few TBs) available at present and scale up your cluster in future depending on the growth of your data.

4.2.3. Multiple Frameworks for Big Data

There are various tools for various purposes. Hadoop can be integrated with multiple analytic tools to get the best out of it, like Mahout for Machine-Learning,

R and Python for Analytics and visualization, Python, Spark for real time processing, MongoDB and Hbase for Nosql database, Pentaho for BI etc.

4.2.4. Lifetime Data Availability

When you want your data to be live and running forever, it can be achieved using Hadoop's salability. There is no limit to the size of cluster that you can have. You can increase the size anytime as per your need by adding datanodes to it with minimal cost.

4.2.5. Some Other Use as follows

- Analytics
- Search
- Data Retention
- Log file processing
- Analysis of Text, Image, Audio, & Video content
- Recommendation systems like in E-Commerce Websites

4.3. When Not To Use Hadoop

4.3.1. Real Time Analytics

If you want to do some Real Time Analytics, where you are expecting result quickly, Hadoop should not be used directly. It is because Hadoop works on batch processing, hence response time is high.

4.3.2. Real Time Analytics – Industry Accepted Way

Since Hadoop cannot be used for real time analytics, people explored and developed a new way in which they can use the strength of Hadoop (HDFS) and make the processing real time. So, the industry accepted way is to store the Big Data in HDFS and mount Spark over it. By using spark the processing can be done in real time and in a flash (real quick).

4.3.3. Industry accepted way:

All the historical big data can be stored in Hadoop HDFS and it can be processed and transformed into a structured manageable data. After processing the data in Hadoop you need to send the output to relational database technologies for BI, decision support, reporting etc.

4.3.4. Multiple Smaller Datasets

Hadoop framework is not recommended for small-structured datasets as you have other tools available in market which can do this work quite easily and at a fast pace than Hadoop like MS Excel, RDBMS etc. For a small data analytics, Hadoop can be costlier than other tools.

4.4. Architecture of Hadoop

Below is a high-level architecture of multi-node Hadoop Cluster.

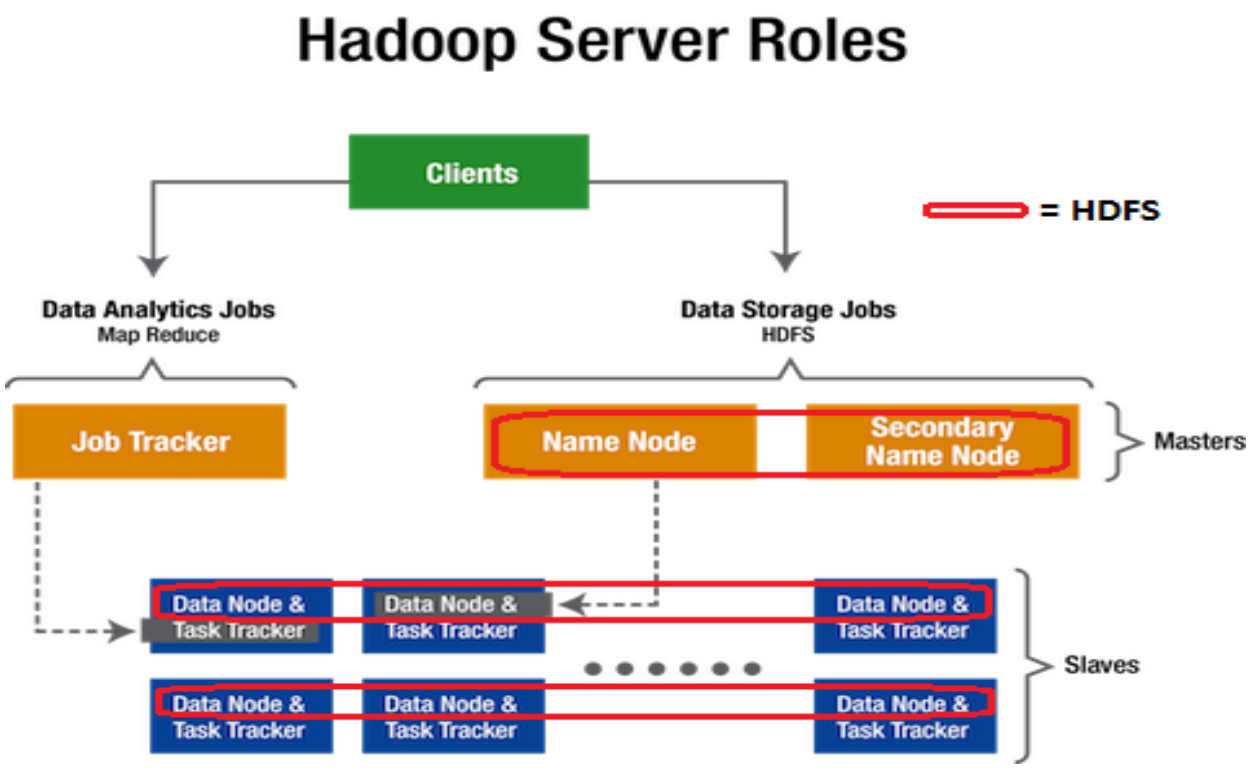


Figure 4.2 Hadoop Server Roles

4.4.1. Hadoop works in a master-worker / master-slave fashion.

4.4.2. Hadoop has two core components: HDFS and MapReduce.

4.4.3. **HDFS (Hadoop Distributed File System)** offers a highly reliable and distributed storage, and ensures reliability, even on a commodity hardware, by replicating the data across multiple nodes. Unlike a regular file system, when data is pushed to HDFS, it will automatically split into multiple blocks (configurable parameter) and stores/replicates the data across various datanodes. This ensures high availability and fault tolerance.

4.4.4. **MapReduce** offers an analysis system which can perform complex computations on large datasets. This component is responsible for performing all the computations and works by breaking down a large complex computation into multiple tasks and assigns those to individual worker/slave nodes and takes care of coordination and consolidation of results.

4.4.5. The master contains the Namenode and Job Tracker components.

4.4.5.1. **Namenode** holds the information about all the other nodes in the Hadoop Cluster, files present in the cluster, constituent blocks of files and their locations in the cluster, and other information useful for the operation of the Hadoop Cluster.

4.4.5.2. **Job Tracker** keeps track of the individual tasks/jobs assigned to each of the nodes and coordinates the exchange of information and results.

4.4.6. Each Worker / Slave contains the Task Tracker and a Datanode components.

4.4.6.1. **Task Tracker** is responsible for running the task / computation assigned to it.

4.4.6.2. **Datanode** is responsible for holding the data.

4.4.7. The computers present in the cluster can be present in any location and there is no dependency on the location of the physical server.

4.5. Hadoop Ecosystem

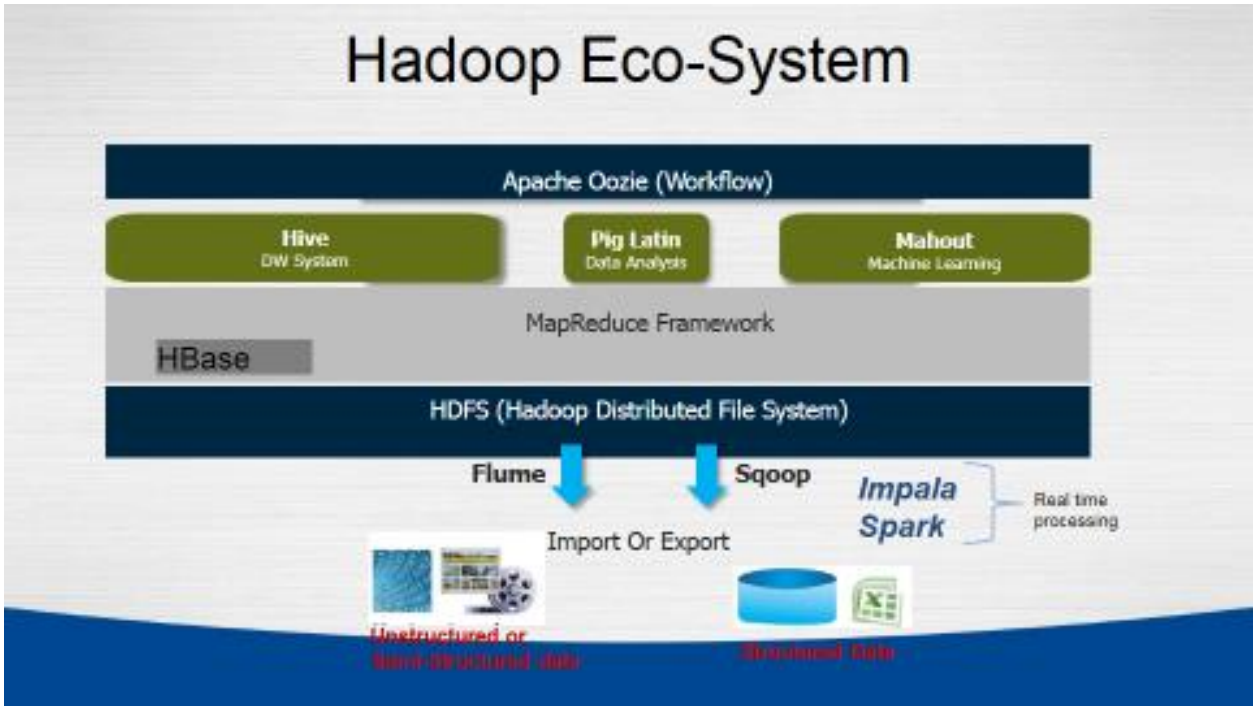


Figure 4.3 Hadoop Eco- System

4.6.Hive

Hive provides a warehouse structure and SQL-like access for data in HDFS and other Hadoop input sources (e.g. Amazon S3). Hive's query language, HiveQL, compiles to MapReduce. It also allows user-defined functions (UDFs). Hive is widely used, and has itself become a "sub-platform" in the Hadoop ecosystem.

4.7.PIG

Pig is a framework consisting of a high-level scripting language (Pig Latin) and a run-time environment that allows users to execute MapReduce on a Hadoop cluster. Like HiveQL in Hive, Pig Latin is a higher-level language that compiles to MapReduce.

4.8. Mahout

Mahout is a scalable machine-learning and data mining library. There are currently four main groups of algorithms in Mahout:

4.8.1. Recommendations, a.k.a. collective filtering

4.8.2. classification, a.k.a categorization

4.8.3. clustering

4.8.4. frequent itemset mining, a.k.a parallel frequent pattern mining

4.8.5. MapReduce

4.8.5.1. The MapReduce paradigm for parallel processing comprises two sequential steps: map and reduce.

4.8.5.2. In the map phase, the input is a set of key-value pairs and the desired function is executed over each key/value pair in order to generate a set of intermediate key/value pairs.

4.9.HBase

Based on Google's Bigtable, HBase "is an open-source, distributed, versioned, column-oriented store" that sits on top of HDFS. HBase is column-based rather than row-based, which enables high-speed execution of operations performed over similar values across massive data sets, e.g. read/write operations that involve all rows but only a small subset of all columns. HBase does not provide its own query or scripting language, but is accessible through Java, Thrift, and REST APIs.

4.9.1. In the reduce phase, the intermediate key/value pairs are grouped by key and the values are combined together according to the reduce code provided by the user; for example, summing. It is also possible that no reduce phase is required, given the type of operation coded by the user.

4.10. Sqoop:

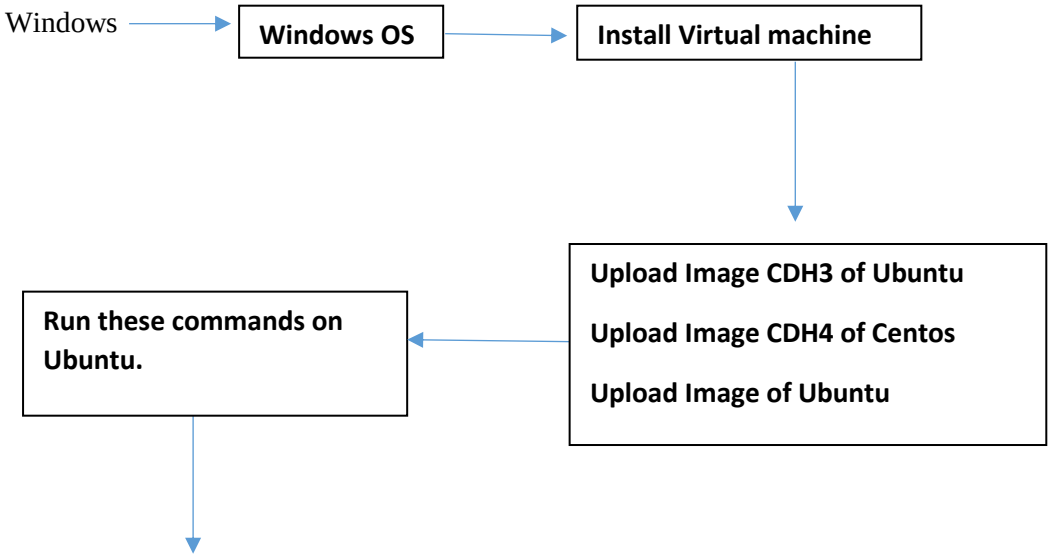
Sqoop ("SQL-to-Hadoop") is a tool which transfers data in both directions between relational systems and HDFS or other Hadoop data stores, e.g. Hive or HBase.

According to the Sqoop blog, "You can use Sqoop to import data from external structured datastores into Hadoop Distributed File System or related systems like Hive and HBase. Conversely, Sqoop can be used to extract data from Hadoop and export it to external structured datastores such as relational databases and enterprise data warehouses.

Chapter 5

Installing Hadoop

5.1. Steps To Installation Hadoop



5.2.Follow These Steps To install Hadoop On Ubuntu

5.2.1. Update the repository:

*#Command: sudo apt-get **update***

5.2.2. Once the Update is complete :

#Command: sudo apt-get install openjdk-6-jdk

After Java has been Installed, to check whether Java is installed on your system or not give the below command:

5.2.3.

#Command:java -version

5.2.4. Install openssh-server:

#Command: sudo apt-get install openssh-server

5.2.5. Download and extract Hadoop:

#Command: wget http://archive.apache.org/dist/hadoop/core/hadoop-1.2.0/hadoop-1.2.0.tar.gz

tar -xvf hadoop-1.2.0.tar.gz

5.2.6. Edit core-site.xml:

The **core-site.xml** file contains information such as the port number used for Hadoop instance, memory allocated for the file system, memory limit for storing the data, and size of Read/Write buffers.

Open the core-site.xml and add the following properties in between <configuration>, </configuration> tags.

#Command: sudo gedit hadoop-1.2.0/conf/core-site.xml

<property>

<name>fs.default.name</name>

<value>hdfs://localhost:8020</value>

</property>

5.2.7. Edit hdfs-site.xml:

#Command: sudo gedit hadoop-1.2.0/conf/hdfs-site.xml

The **hdfs-site.xml** file contains information such as the value of replication data, namenode path, and datanode paths of your local file systems. It means the place where you want to store the Hadoop infrastructure.

Open this file and add the following properties in between the <configuration> </configuration> tags in this file. In the above file, all the property values are user-defined and you can make changes according to your Hadoop infrastructure.

<property>

<name>dfs.replication</name>

<value>1</value>

</property>

<property>

<name>dfs.permissions</name>

<value>false</value>

</property>

5.2.8. Edit mapred-site.xml:

#Command: sudo gedit hadoop-1.2.0/conf/mapred -site.xml

This file is used to specify which MapReduce framework we are using. By default, Hadoop contains a template of yarn-site.xml. First of all, it is required to copy the file from **mapred-site,xml.template** to **mapred-site.xml** file using the following command.

Open mapred-site.xml file and add the following properties in between the <configuration>, </configuration>tags in this file.

<property>

<name>mapred.job.tracker</name>

<value>localhost:8021</value>

</property>

Get your ip address:

5.2.9. Check IP Address

#Command: ifconfig

5.2.10.

#Command: sudo gedit /etc/hosts

Create a ssh key:

5.2.11.

#Command: ssh-keygen -t rsa –P ""

Chapter 6

Sqoop

Using Sqoop and Mysql and HDFS we can handle the data and Transfer the data MYsql to HDFS and Annalise the Data

Sqoop is a tool designed to transfer data between Hadoop and relational database servers. It is used to import data from relational databases such as MySQL, Oracle to Hadoop HDFS, and export from Hadoop file system to relational databases. It is provided by the Apache Software Foundation. SQL to Hadoop and Hadoop to SQL. Through Sqoop we can Import full table, part of table and selected value into Hadoop.

Figure 6.1 Sqoop as Interface

Hadoop for analytics requires loading data into Hadoop clusters and processing it in conjunction with data that resides on enterprise application servers and databases. Loading GBs and TBs & Petabyte of data into HDFS from production databases or accessing it from map reduce applications is a challenging task. While doing so, we have to consider things like data consistency, overhead of running these jobs on production systems and at the end if this process would be efficient or not. Using batch scripts to load data is an inefficient way to go with.

6.1. How Sqoop Works?

The following image describes the workflow of Sqoop.

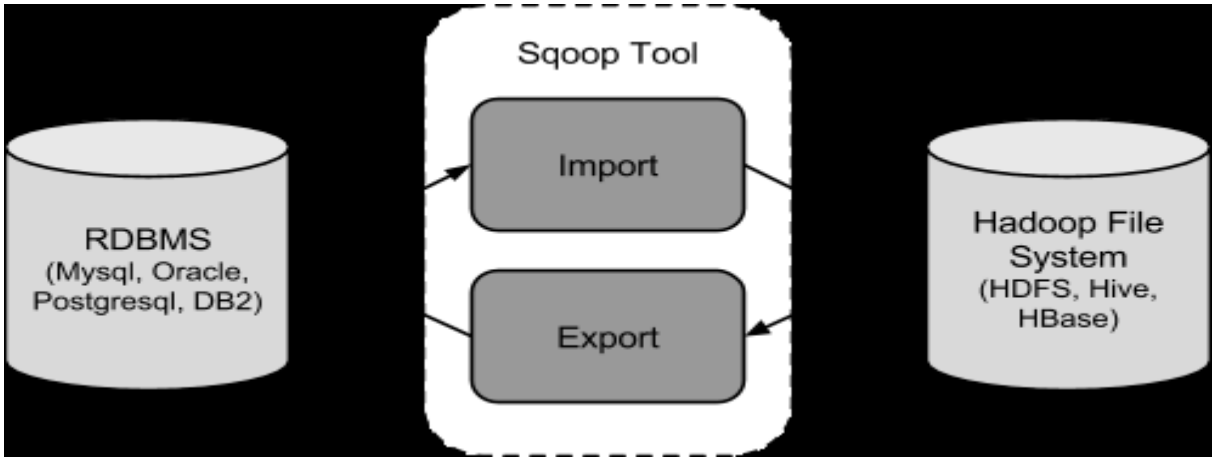


Figure 6.2 Interfacing Hadoop and Sql

6.2. Sqoop Import

The import tool imports individual tables from RDBMS to HDFS. Each row in a table is treated as a record in HDFS. All records are stored as text data in text files or as binary data in Avro and Sequence files.

6.3. Sqoop Export

The export tool exports a set of files from HDFS back to an RDBMS. The files given as input to Sqoop contain records, which are called as rows in table. Those are read and parsed into a set of records and delimited with user-specified delimiter

6.4. Step To Install and Create Table on Operating system.

Example

Let us take an example of 1tables named as **H_info** which are in a database called **HadoopGyan** in a MySQL database server.

The 1 tables and their data are as follows.

Do Following Steps:-

Step 1 Create Database.

Step 2 Create Table schema into MYSQL.

Step 3 Insert the values into table one by one or Dump full data into MYSQL Table.

Step 4 Now your full data into your table.

H_info

ID	Name	Country
101	Map reduce	USA
102	Pig	UK
103	Hive	India
104	HBase	Africa

6.5. Importing full Table To HDFS

This string will connect to a MySQL database named Hadoopgyan The connect string you supply will be used on TaskTracker nodes throughout your MapReduce cluster; if you specify the literal name localhost, each node will connect to a different database (or more likely, no database at all). Instead, you should use the full hostname or IP address of the database host that can be seen by all your remote nodes.

You might need to authenticate against the database before you can access it. You can use the --username and --password or -P parameters to supply a username and a password to the database.

Sqoop automatically supports several databases, including MySQL. Connect strings beginning with jdbc:mysql:// are handled automatically in Sqoop.

6.6. Write this way your command to import data FROM MYSQL to HDFS this is called IMPORT in SQOOP

$ Sqoop Import --connect jdbc:mysql://local host/Database name -- username -- table name --target-dir /HDFS -M 1

$ Sqoop Import –-connect jdbc:mysql://local host/ Hadoopgyan –username root –table H_info

-–Target-dir /user/cloudera/HadoopGyan –- m 1

Through above command our table H_info import to the hadoop (HDFS)

Type of file to be stored in HDFS

1) Text Files ::----- -- as- Textfile
2) Sequence File ::----- --as-sequencefile
3) Binary or Avro file::----- --as-avrofile

NOTE in Sqoop we can define their own number of mapper to increase speed and better file performance and understanding of output such as partitioning in Map reduce. We can also import the Table structure as well as their limited data and part of data to HDFS. We can also import data into HIVE and HBASE. For more knowledge please visit Hadoopgyan Institute.

6.7. Sqoop Export

This string will connect to a MySQL database named Hadoopgyan The connect string you supply will be used on TaskTracker nodes throughout your MapReduce cluster; if you specify the literal name localhost, each node will connect to a different database (or more likely, no database at all). Instead, you should use the full hostname or IP address of the database host that can be seen by all your remote nodes.

You might need to authenticate against the database before you can access it. You can use the --username and --password or -P parameters to supply a username and a password to the database.

Sqoop automatically supports several databases, including MySQL. Connect strings beginning with jdbc:mysql:// are handled automatically in Sqoop.

Let us take an example of 1tables named as **H_info** which are in a database called **HadoopGyan** in a MySQL database server.
The 1 tables and their data are as follows.

Do Following Steps:-

Step 1 Create Database.

Step 2 Create Table schema into MYSQL.

Step 3 Insert the values into table one by one or Dump full data into MYSQL Table.

Step 4 Now your full data into your table.

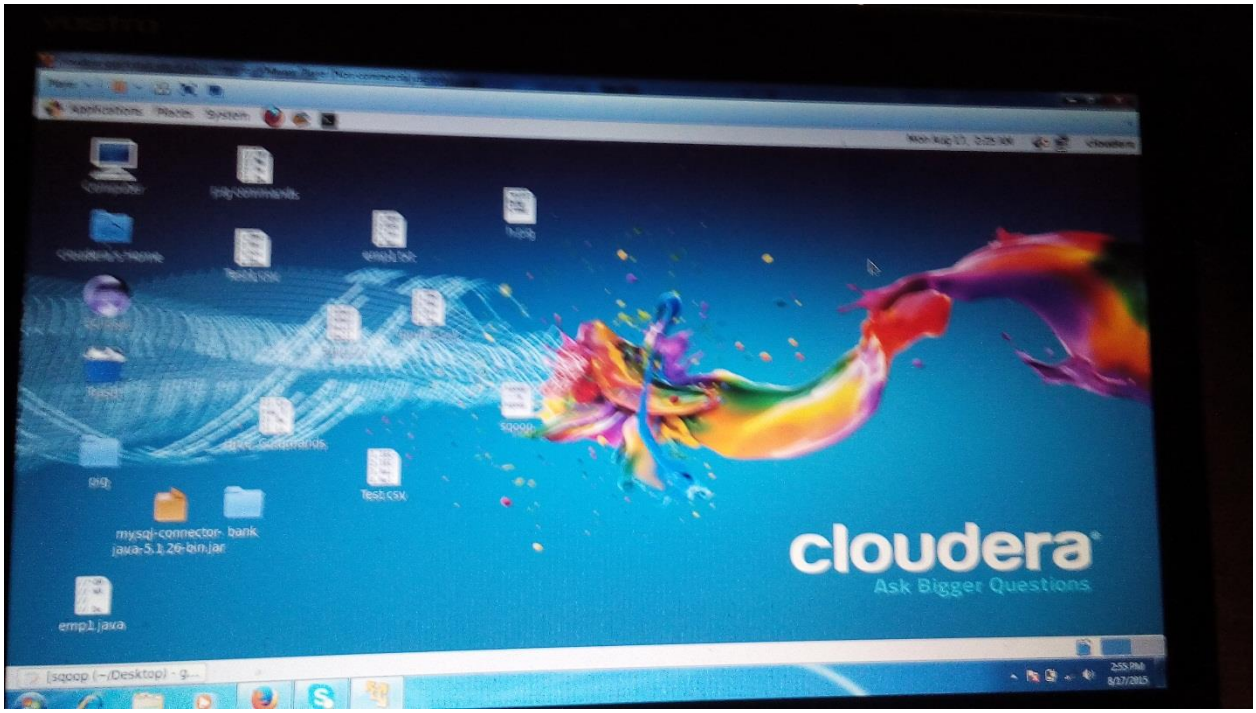
6.8. Write this way your command to Export data Into MYSQL From HDFS this is called Export in SQOOP

$ Sqoop export –connect jdbc:mysql://localhost:/portnumber/databasename
--username root -P –table name
--export-dir "/path of table "

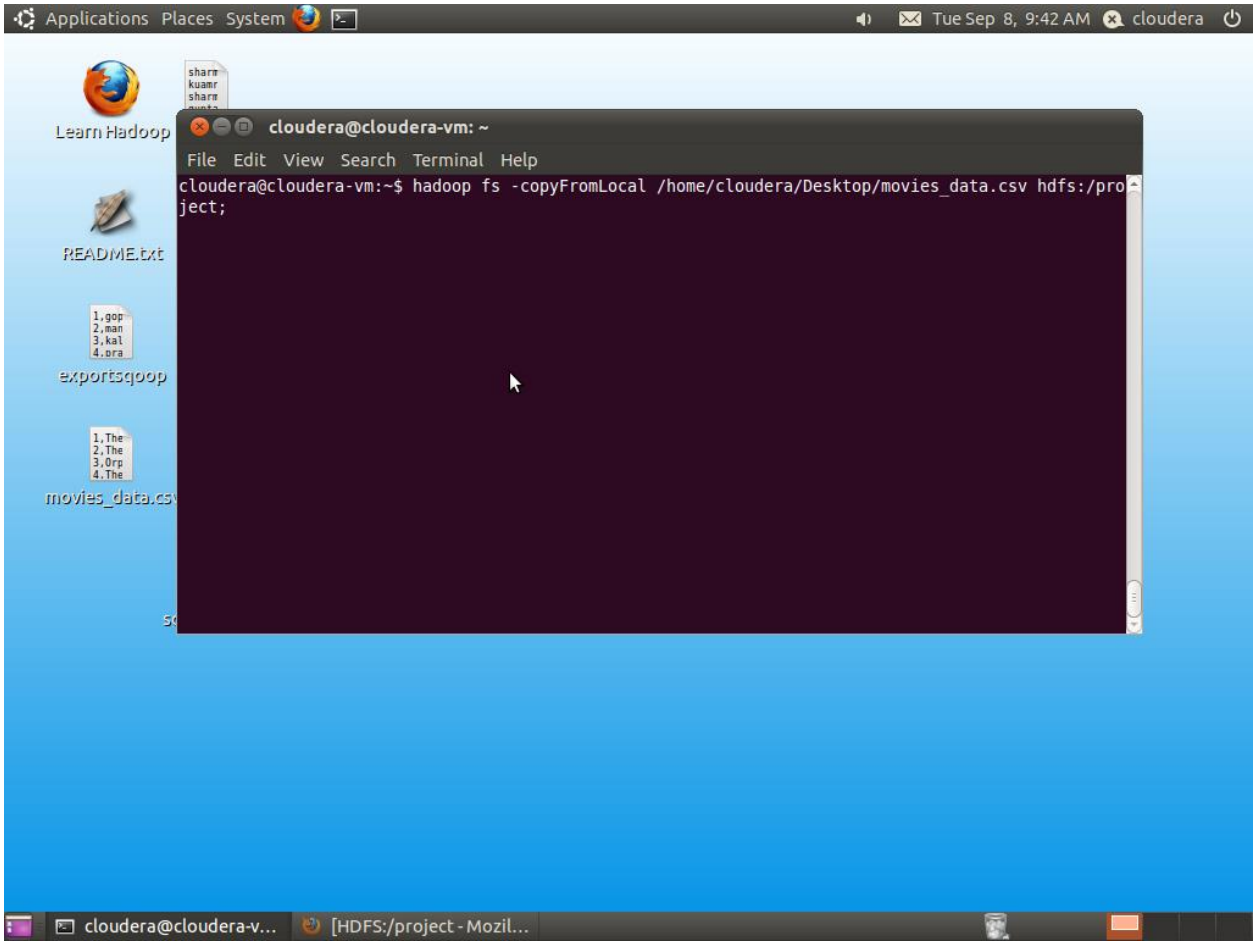
OR

$ sqoop export –-connect jdbc:mysq://localhost/Portnumber/databasename
--username root -Password Password --table name
--export-dir "Path of table"

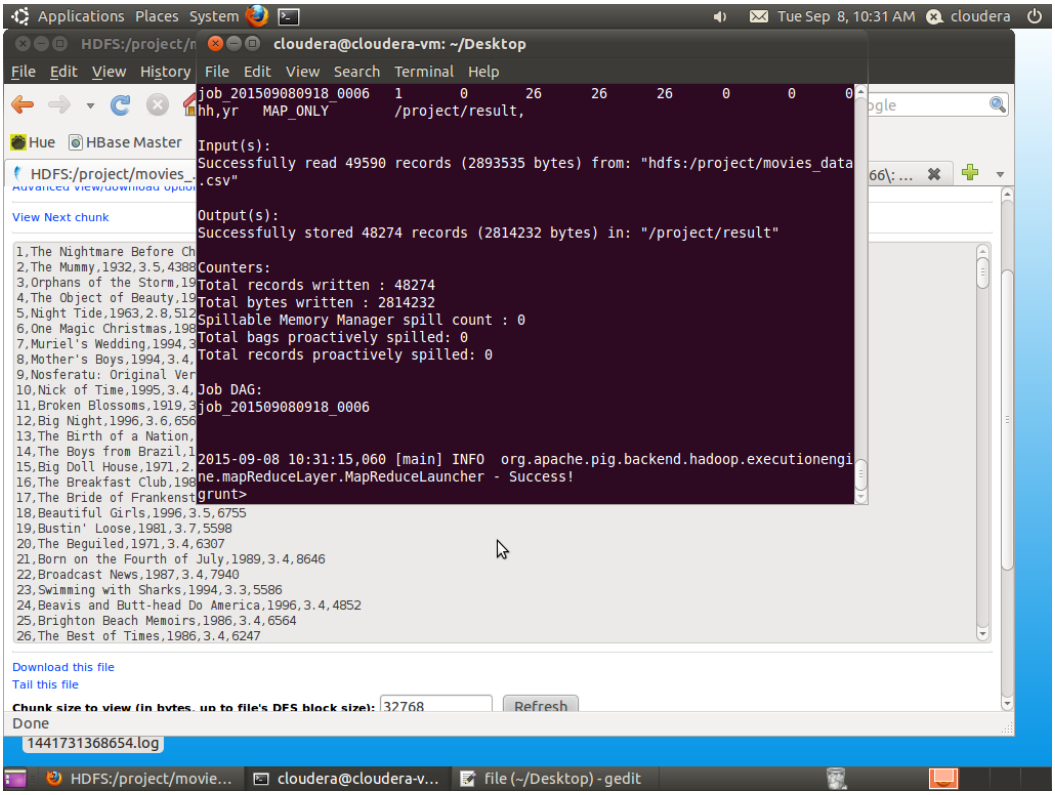
6.8.1. Open CDH4 Centos for Analyzing Big Data



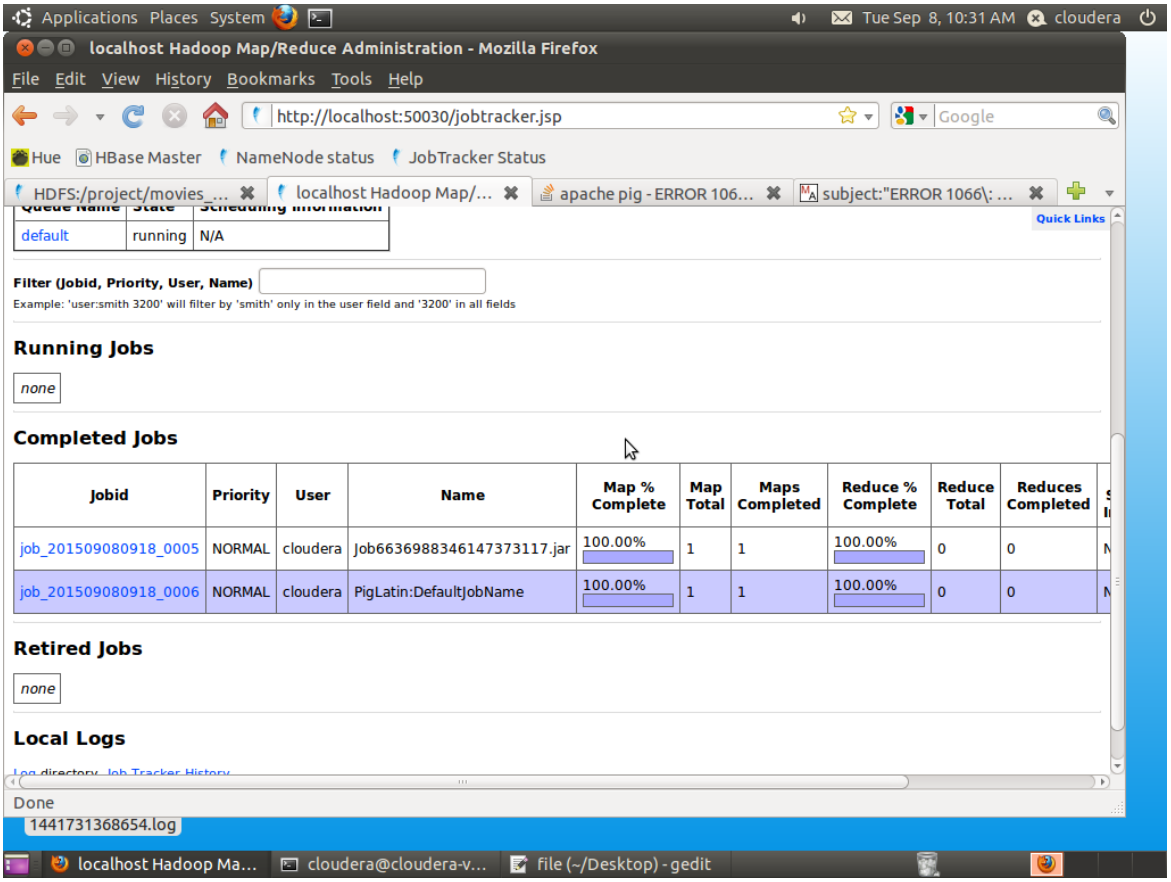
4.8.1. Copy the Input file into HDFS



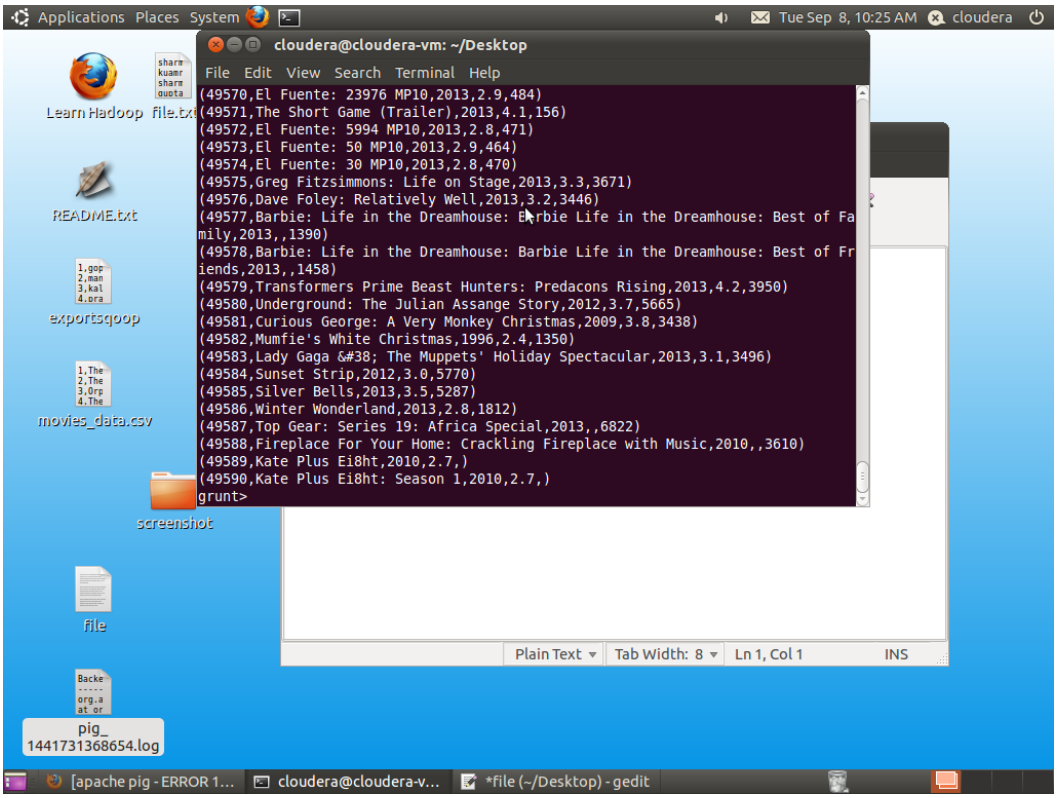
6.8.2. Copying Data into HDFS .



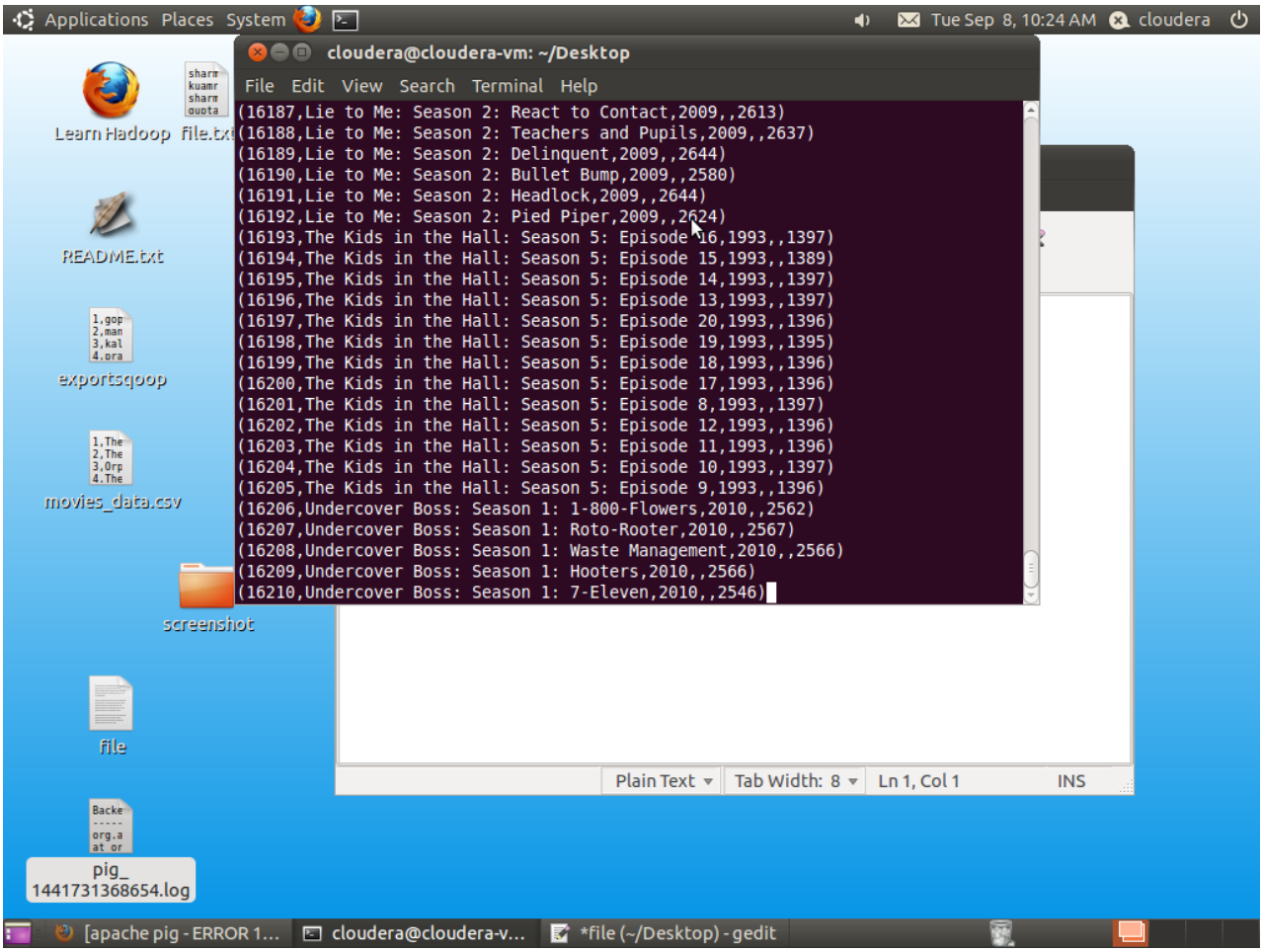
6.8.3. Resulted Data into The HDFS

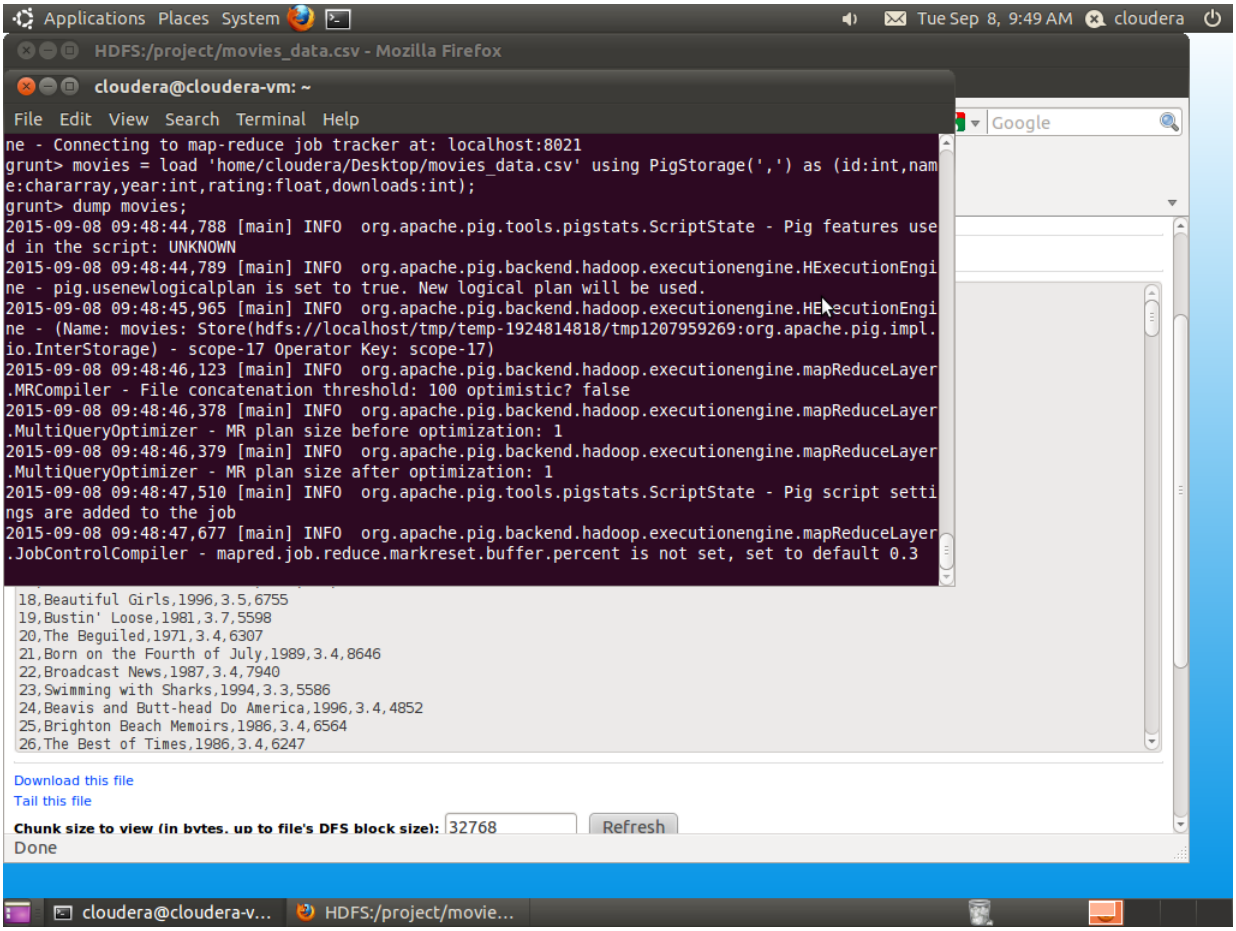


6.9. Create Schema In the Pig Latin and Filter through year

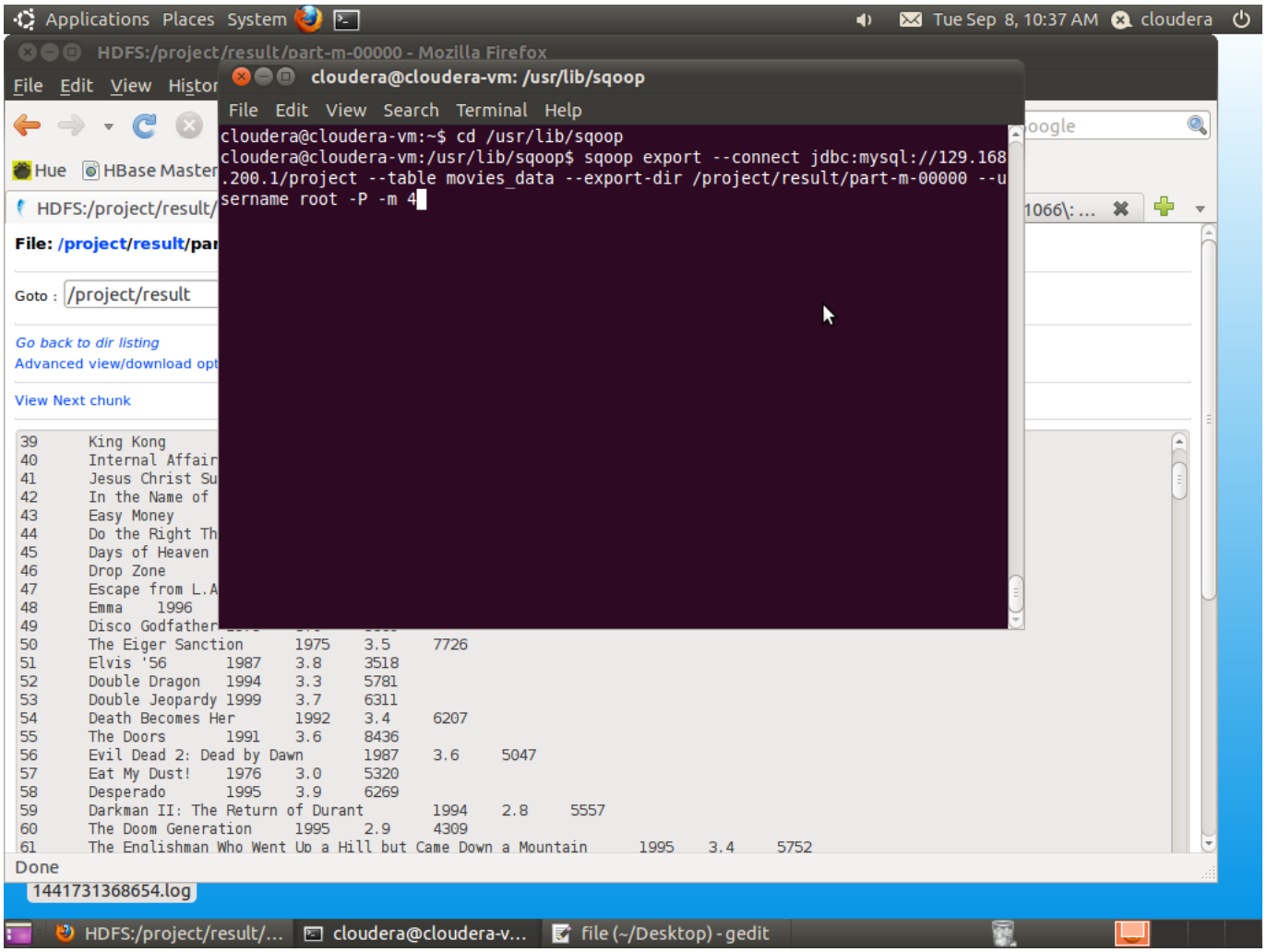


6.10. Show the Total Record of HDFS into Pig Schema

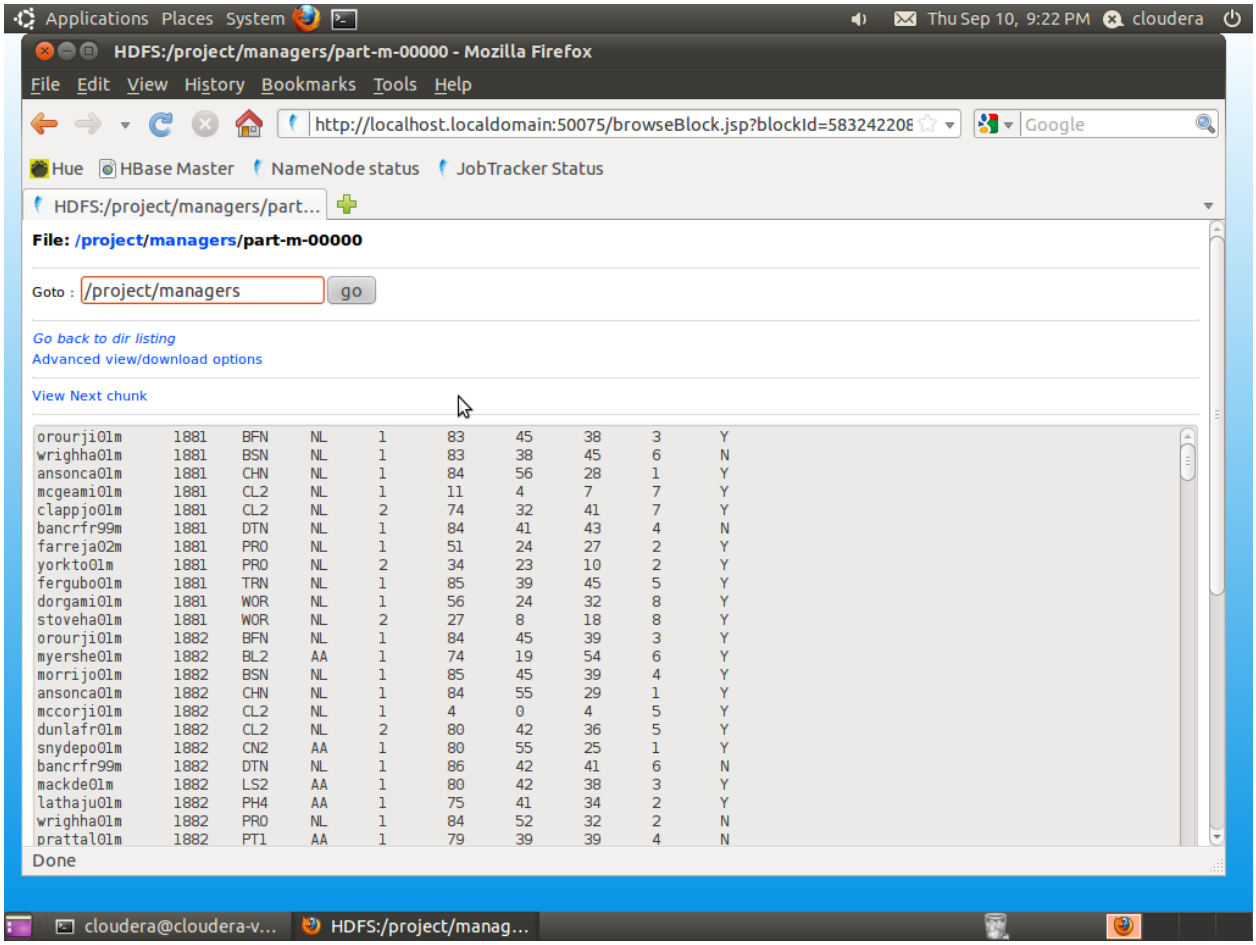



Applications Places System
Tue Sep 8, 9:49 AM cloudera
HDFS:/project/movies_data.csv - Mozilla Firefox
cloudera@cloudera-vm: ~
File Edit View Search Terminal Help
ne - Connecting to map-reduce job tracker at: localhost:8021
grunt> movies = load 'home/cloudera/Desktop/movies_data.csv' using PigStorage(',') as (id:int,nam
e:chararray,year:int,rating:float,downloads:int);
grunt> dump movies;
2015-09-08 09:48:44,788 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features use
d in the script: UNKNOWN
2015-09-08 09:48:44,789 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngi
ne - pig.usenewlogicalplan is set to true. New logical plan will be used.
2015-09-08 09:48:45,965 [main] INFO org.apache.pig.backend.hadoop.executionengine.HExecutionEngi
ne - (Name: movies: Store(hdfs://localhost/tmp/temp-1924814818/tmp1207959269:org.apache.pig.impl.
io.InterStorage) - scope-17 Operator Key: scope-17)
2015-09-08 09:48:46,123 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.MRCompiler - File concatenation threshold: 100 optimistic? false
2015-09-08 09:48:46,378 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.MultiQueryOptimizer - MR plan size before optimization: 1
2015-09-08 09:48:46,379 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.MultiQueryOptimizer - MR plan size after optimization: 1
2015-09-08 09:48:47,510 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script setti
ngs are added to the job
2015-09-08 09:48:47,677 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer
.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
Google
18,Beautiful Girls,1996,3.5,6755
19,Bustin' Loose,1981,3.7,5598
20,The Beguiled,1971,3.4,6307
21,Born on the Fourth of July,1989,3.4,8646
22,Broadcast News,1987,3.4,7940
23,Swimming with Sharks,1994,3.3,5586
24,Beavis and Butt-head Do America,1996,3.4,4852
25,Brighton Beach Memoirs,1986,3.4,6564
26,The Best of Times,1986,3.4,6247
Download this file
Tail this file
Chunk size to view (in bytes, up to file's DFS block size): 32768
Refresh
Done
cloudera@cloudera-v...
HDFS:/project/movie...

6.11. Filter the Big data of Movies Data



6.12. Output of Pig Stored into the HDFS Using Store Keyword



6.13. Install MySQL on Windows

Figure 6.13.1

MySQL to HDFS – Using Sqoop

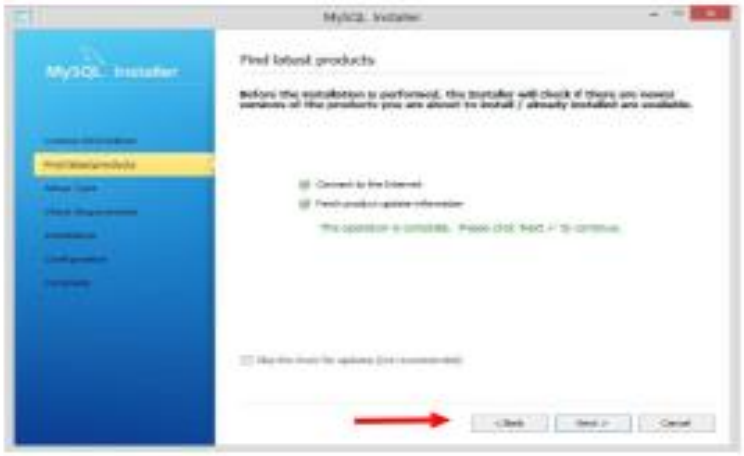
3 - Click **Next:**

4 - Click **Execute:**

Figure 6.13.2

MySQL to HDFS – UsinSqoop

5 - Click **Next:**



6 - Click
Next:

Figure 6.13.3

MySQL to HDFS – Using Sqoop

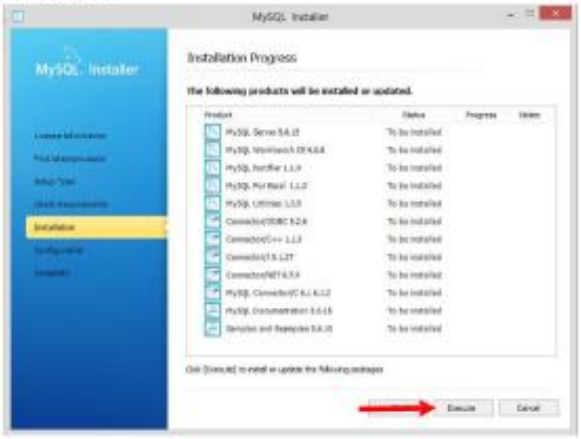
7 - Click **Execute:**

8 - Click **Next:**

Figure 6.13.4

MySQL to HDFS – Using Sqoop

9 - Click **Execute:**

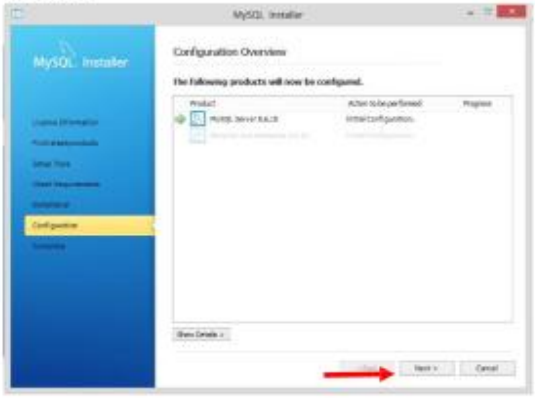


10 - Click **Next:**

Figure 6.13.5

MySQL to HDFS – Using Sqoop

11 - Click **Next:**



12 - Click **Next:**

Figure 6.13.6

MySQL to HDFS – Using Sqoop

13 - Enter in **MySQL Root Password** –> root
Enter in **Repeat Password** –> root
Click **Next:**

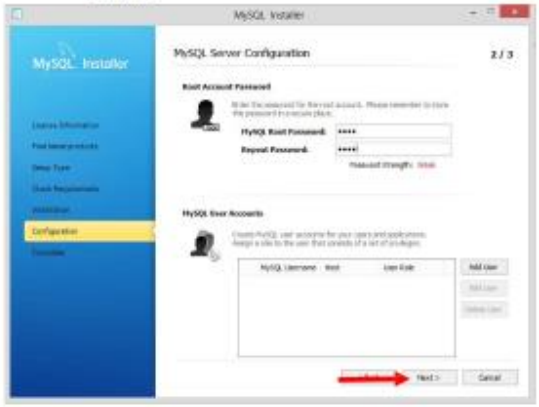


Figure 6.13.7

MySQL to HDFS – Using Sqoop

14 - Click **Next:**

15 - Click **Next:**

MySQL to HDFS – Using Sqoop

16 - Click **Next:**

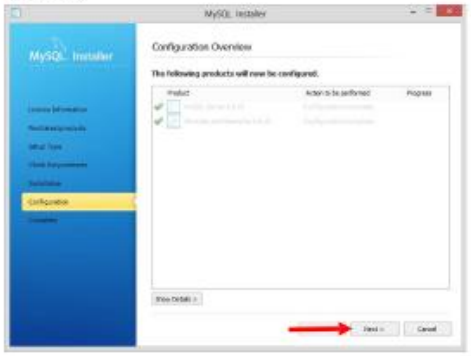


Figure 6.13.8

MySQL to HDFS – Using Sqoop

17 - Uncheck the check-box (Start MySQL Workbench after Setup) and Click **Finish:**

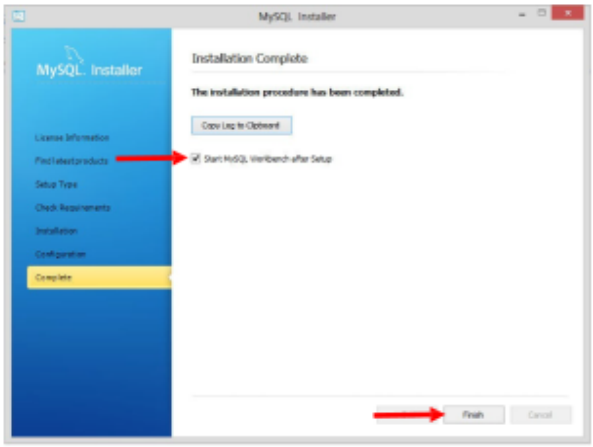


Figure 6.13.9

MySQL to HDFS – Using Sqoop

24 - Download MySQL connector using the below link:

https://drive.google.com/file/d/0B2-rICGKD40N5VJ3QlpWSk95OFE/edit?usp=sharing

25 - Open Cloudera cdh3 and **move MySQL connector to Cloudera cdh3 (To Desktop) using FileZilla.**

Use the below link to understand how to move a file from Windows to cloudera cdh3 vm.

http://www.edureka.in/blog/transfer-files-windows-cloudera-demo-vm/

26 - Once the MySQL connector is present on Cloudera Cdh3 Desktop, move it to the lib folder of sqoop by executing the below command:

27 **Command:** sudo cp /home/cloudera/Desktop/mysql-connector-java-5.1.26-bin.jar /usr/lib/sqoop/lib

27 - **Change the directory to Sqoop** by executing the below command:

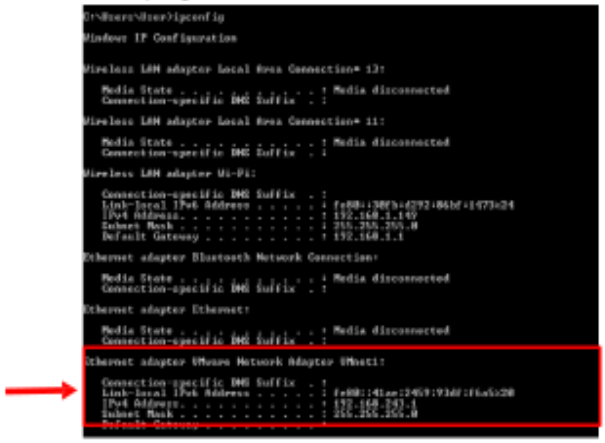
Command: cd /usr/lib/sqoop

Figure 6.13.10

MySQL to HDFS – Using Sqoop

28 - Open Command Prompt (CMD) on Windows and check the IPv4 Address by executing the below command:

Command: ipconfig



29 - Grant all privileges to root@your_ipv4_address by executing the below command

(In MySQL 5.6 Command Line Client):

Required items for the command:

Ip – Find out the IPv4 address of your system using the above step. In my case it is

192.168.243.1

Command:

grant all privileges on *.* to root@192.168.243.1 IDENTIFIED BY root WITH GRANT OPTION;

Chapter 7

Project Screenshots

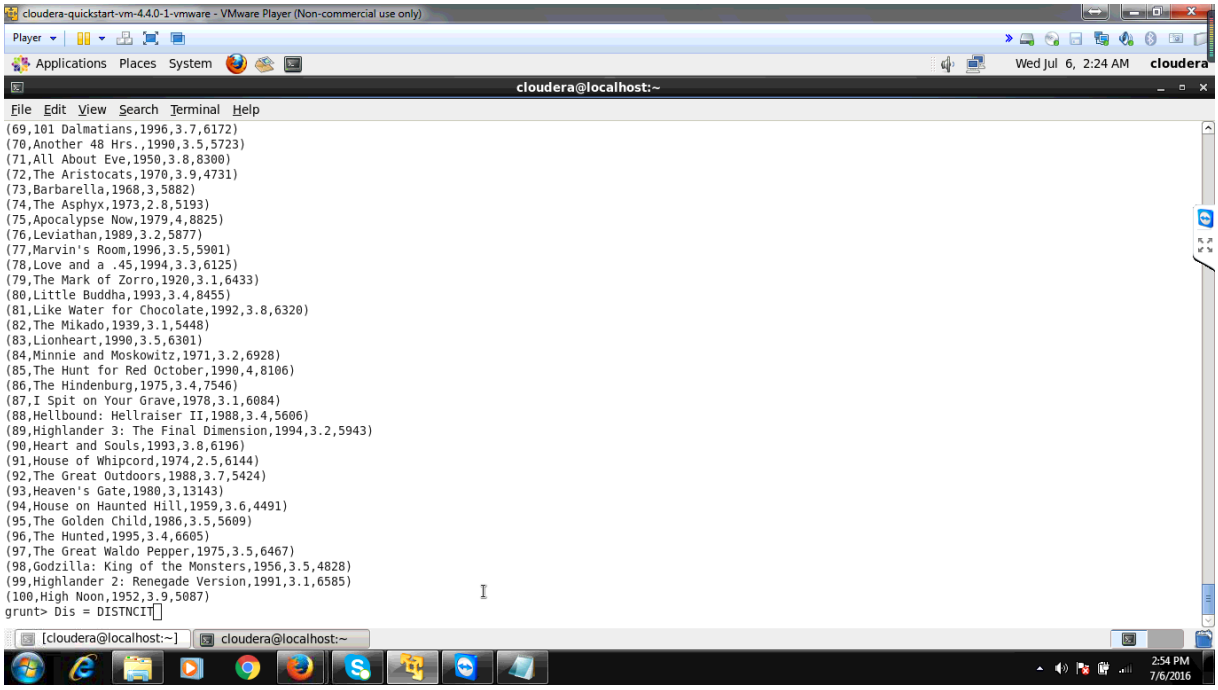


Figure-7.1

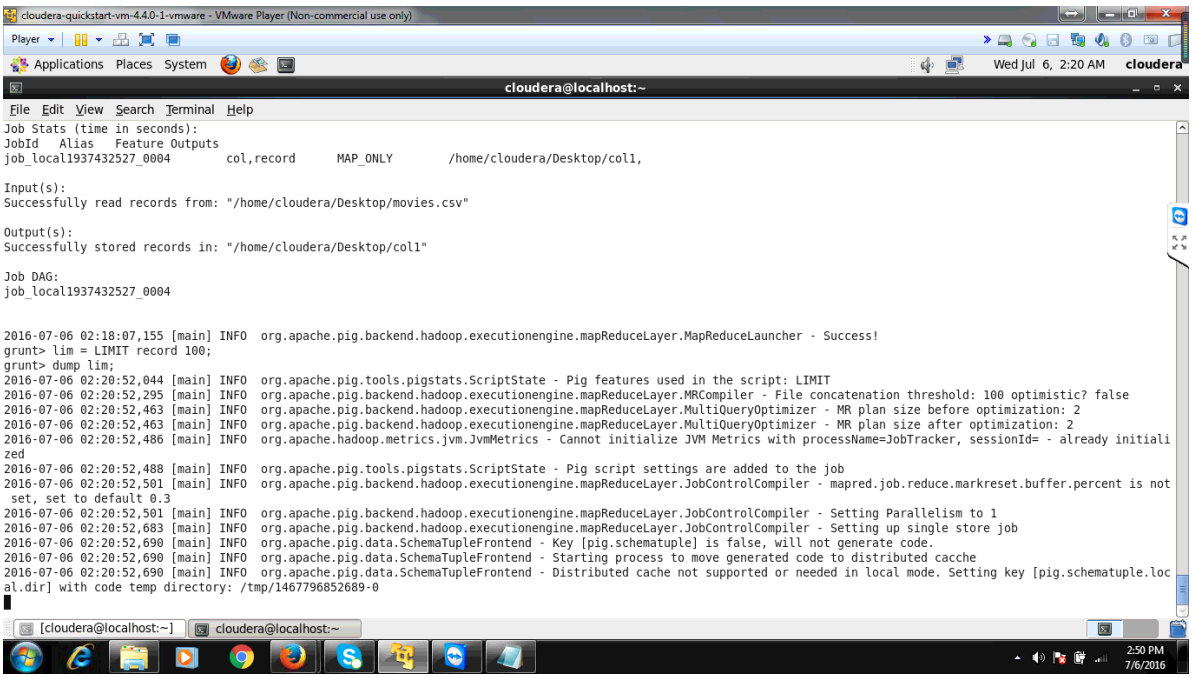

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 2:20 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local1937432527_0004 col,record MAP_ONLY /home/cloudera/Desktop/col1,
Input(s):
Successfully read records from: "/home/cloudera/Desktop/movies.csv"
Output(s):
Successfully stored records in: "/home/cloudera/Desktop/col1"
Job DAG:
job_local1937432527_0004
2016-07-06 02:18:07,155 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> lim = LIMIT record 100;
grunt> dump lim;
2016-07-06 02:20:52,044 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: LIMIT
2016-07-06 02:20:52,295 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MRCompiler - File concatenation threshold: 100 optimistic? false
2016-07-06 02:20:52,463 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size before optimization: 2
2016-07-06 02:20:52,463 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MultiQueryOptimizer - MR plan size after optimization: 2
2016-07-06 02:20:52,486 [main] INFO org.apache.hadoop.metrics.jvm.JvmMetrics - Cannot initialize JVM Metrics with processName=JobTracker, sessionId= - already initialized
2016-07-06 02:20:52,488 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig script settings are added to the job
2016-07-06 02:20:52,501 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - mapred.job.reduce.markreset.buffer.percent is not set, set to default 0.3
2016-07-06 02:20:52,501 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting Parallelism to 1
2016-07-06 02:20:52,683 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.JobControlCompiler - Setting up single store job
2016-07-06 02:20:52,690 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Key [pig.schematuple] is false, will not generate code.
2016-07-06 02:20:52,690 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Starting process to move generated code to distributed cacche
2016-07-06 02:20:52,690 [main] INFO org.apache.pig.data.SchemaTupleFrontend - Distributed cache not supported or needed in local mode. Setting key [pig.schematuple.local.dir] with code temp directory: /tmp/1467796852689-0
[cloudera@localhost:~] cloudera@localhost:~
2:50 PM
7/6/2016

Figure-7.2

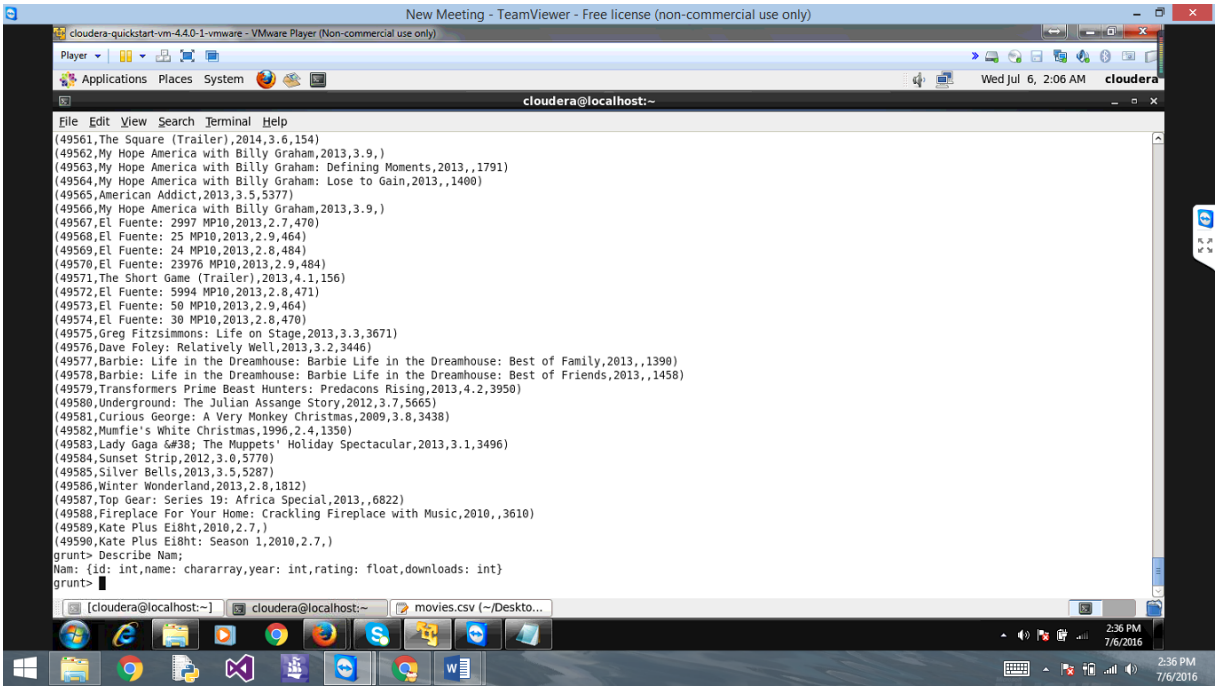

New Meeting - TeamViewer - Free license (non-commercial use only)
cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 2:06 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
(49561,The Square (Trailer),2014,3.6,154)
(49562,My Hope America with Billy Graham,2013,3.9,)
(49563,My Hope America with Billy Graham: Defining Moments,2013,,1791)
(49564,My Hope America with Billy Graham: Lose to Gain,2013,,1400)
(49565,American Addict,2013,3.5,5377)
(49566,My Hope America with Billy Graham,2013,3.9,)
(49567,El Fuente: 2997 MP10,2013,2.7,470)
(49568,El Fuente: 25 MP10,2013,2.9,464)
(49569,El Fuente: 24 MP10,2013,2.8,484)
(49570,El Fuente: 23976 MP10,2013,2.9,484)
(49571,The Short Game (Trailer),2013,4.1,156)
(49572,El Fuente: 5994 MP10,2013,2.8,471)
(49573,El Fuente: 50 MP10,2013,2.9,464)
(49574,El Fuente: 30 MP10,2013,2.8,470)
(49575,Greg Fitzsimmons: Life on Stage,2013,3.3,3671)
(49576,Dave Foley: Relatively Well,2013,3.2,3446)
(49577,Barbie: Life in the Dreamhouse: Barbie Life in the Dreamhouse: Best of Family,2013,,1390)
(49578,Barbie: Life in the Dreamhouse: Barbie Life in the Dreamhouse: Best of Friends,2013,,1458)
(49579,Transformers Prime Beast Hunters: Predacons Rising,2013,4.2,3950)
(49580,Underground: The Julian Assange Story,2012,3.7,5665)
(49581,Curious George: A Very Monkey Christmas,2009,3.8,3438)
(49582,Mumfie's White Christmas,1996,2.4,1350)
(49583,Lady Gaga & The Muppets' Holiday Spectacular,2013,3.1,3496)
(49584,Sunset Strip,2012,3.0,5770)
(49585,Silver Bells,2013,3.5,5287)
(49586,Winter Wonderland,2013,2.8,1812)
(49587,Top Gear: Series 19: Africa Special,2013,,6822)
(49588,Fireplace For Your Home: Crackling Fireplace with Music,2010,,3610)
(49589,Kate Plus Ei8ht,2010,2.7,)
(49590,Kate Plus Ei8ht: Season 1,2010,2.7,)
grunt> Describe Nam;
Nam: {id: int,name: chararray,year: int,rating: float,downloads: int}
grunt>
[cloudera@localhost:~]
cloudera@localhost:~
movies.csv (~/Deskto...
2:36 PM
7/6/2016
2:36 PM
7/6/2016

Figure-7.3.1

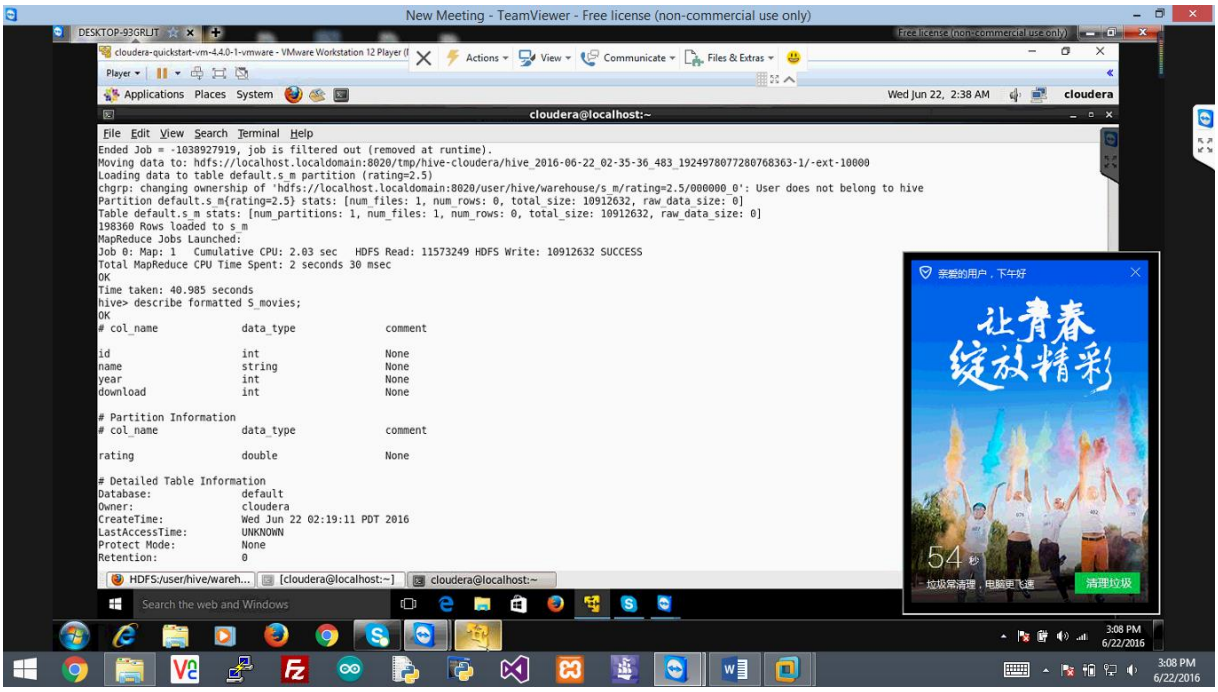

New Meeting - TeamViewer - Free license (non-commercial use only)
cloudera@localhost:~
File Edit View Search Terminal Help
Ended Job = -1038927919, job is filtered out (removed at runtime).
Moving data to: hdfs://localhost.localdomain:8020/tmp/hive-cloudera/hive_2016-06-22_02-35-36_483_1924978077280768363-1/-ext-10000
Loading data to table default.s_m partition (rating=2.5)
chgrp: changing ownership of 'hdfs://localhost.localdomain:8020/user/hive/warehouse/s_m/rating=2.5/000000_0': User does not belong to hive
Partition default.s_m{rating=2.5} stats: [num_files: 1, num_rows: 0, total_size: 10912632, raw_data_size: 0]
Table default.s_m stats: [num_partitions: 1, num_files: 1, num_rows: 0, total_size: 10912632, raw_data_size: 0]
198360 Rows loaded to s_m
MapReduce Jobs Launched:
Job 0: Map: 1 Cumulative CPU: 2.03 sec HDFS Read: 11573249 HDFS Write: 10912632 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 30 msec
OK
Time taken: 40.985 seconds
hive> describe formatted S_movies;
OK
col_name data_type comment
id int None
name string None
year int None
download int None
Partition Information
col_name data_type comment
rating double None
Detailed Table Information
Database: default
Owner: cloudera
CreateTime: Wed Jun 22 02:19:11 PDT 2016
LastAccessTime: UNKNOWN
Protect Mode: None
Retention: 0
让青春绽放精彩
清理垃圾

Figure 7.3.2

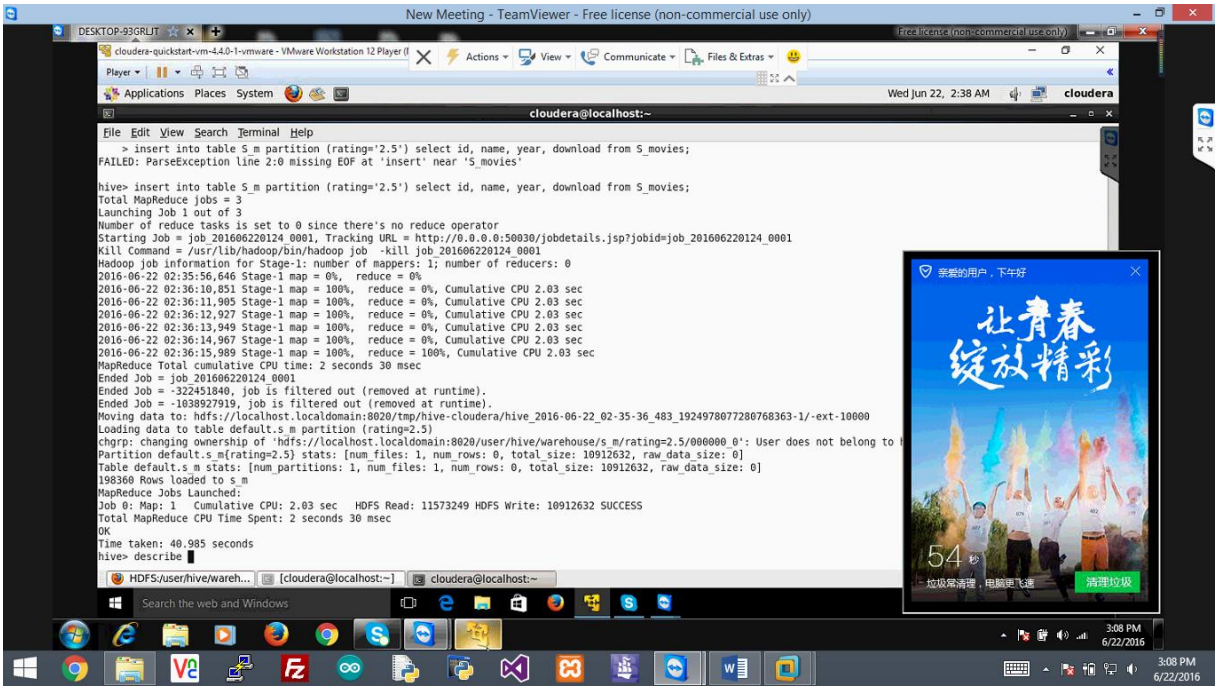

New Meeting - TeamViewer - Free license (non-commercial use only)
cloudera@localhost:~
File Edit View Search Terminal Help
> insert into table S_m partition (rating='2.5') select id, name, year, download from S_movies;
FAILED: ParseException line 2:0 missing EOF at 'insert' near 'S_movies'
hive> insert into table S_m partition (rating='2.5') select id, name, year, download from S_movies;
Total MapReduce jobs = 3
Launching Job 1 out of 3
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_201606220124_0001, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_201606220124_0001
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_201606220124_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2016-06-22 02:35:56,646 Stage-1 map = 0%, reduce = 0%
2016-06-22 02:36:10,851 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.03 sec
2016-06-22 02:36:11,905 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.03 sec
2016-06-22 02:36:12,927 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.03 sec
2016-06-22 02:36:13,949 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.03 sec
2016-06-22 02:36:14,967 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 2.03 sec
2016-06-22 02:36:15,989 Stage-1 map = 100%, reduce = 100%, Cumulative CPU 2.03 sec
MapReduce Total cumulative CPU time: 2 seconds 30 msec
Ended Job = job_201606220124_0001
Ended Job = -322451840, job is filtered out (removed at runtime).
Ended Job = -1038927919, job is filtered out (removed at runtime).
Moving data to: hdfs://localhost.localdomain:8020/tmp/hive-cloudera/hive_2016-06-22_02-35-36_483_1924978077280768363-1/-ext-10000
Loading data to table default.s_m partition (rating=2.5)
Partition default.s_m{rating=2.5} stats: [num_files: 1, num_rows: 0, total_size: 10912632, raw_data_size: 0]
Table default.s_m stats: [num_partitions: 1, num_files: 1, num_rows: 0, total_size: 10912632, raw_data_size: 0]
198360 Rows loaded to s_m
MapReduce Jobs Launched:
Job 0: Map: 1 Cumulative CPU: 2.03 sec HDFS Read: 11573249 HDFS Write: 10912632 SUCCESS
Total MapReduce CPU Time Spent: 2 seconds 30 msec
OK
Time taken: 40.985 seconds
hive> describe
让青春绽放精彩
清理垃圾

Figure-7.4

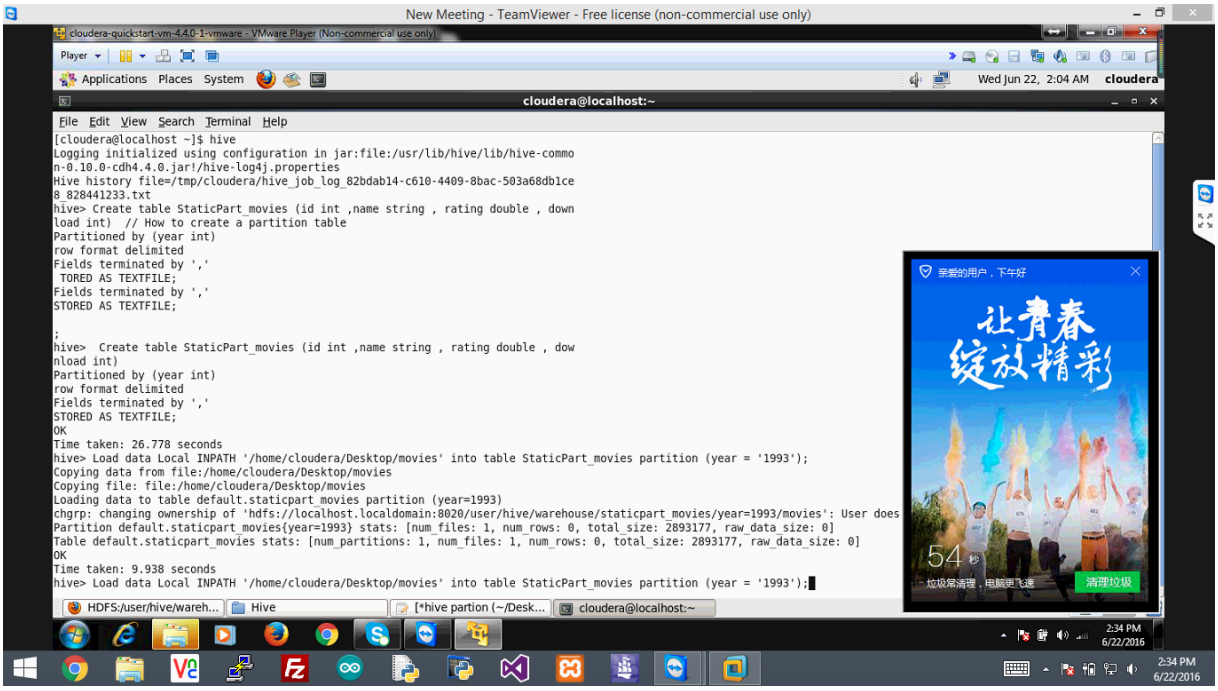

New Meeting - TeamViewer - Free license (non-commercial use only)
cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jun 22, 2:04 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
[cloudera@localhost ~]$ hive
Logging initialized using configuration in jar:file:/usr/lib/hive/lib/hive-commo
n-0.10.0-cdh4.4.0.jar!/hive-log4j.properties
Hive history file=/tmp/cloudera/hive_job_log_82bdab14-c610-4409-8bac-503a68db1ce
8_828441233.txt
hive> Create table StaticPart_movies (id int ,name string , rating double , down
load int) // How to create a partition table
Partitioned by (year int)
row format delimited
Fields terminated by ','
TORED AS TEXTFILE;
Fields terminated by ','
STORED AS TEXTFILE;
;
hive> Create table StaticPart_movies (id int ,name string , rating double , dow
nload int)
Partitioned by (year int)
row format delimited
Fields terminated by ','
STORED AS TEXTFILE;
OK
Time taken: 26.778 seconds
hive> Load data Local INPATH '/home/cloudera/Desktop/movies' into table StaticPart_movies partition (year = '1993');
Copying data from file:/home/cloudera/Desktop/movies
Copying file: file:/home/cloudera/Desktop/movies
Loading data to table default.staticpart_movies partition (year=1993)
chgrp: changing ownership of 'hdfs://localhost.localdomain:8020/user/hive/warehouse/staticpart_movies/year=1993/movies': User does
Partition default.staticpart_movies{year=1993} stats: [num_files: 1, num_rows: 0, total_size: 2893177, raw_data_size: 0]
Table default.staticpart_movies stats: [num_partitions: 1, num_files: 1, num_rows: 0, total_size: 2893177, raw_data_size: 0]
OK
Time taken: 9.938 seconds
hive> Load data Local INPATH '/home/cloudera/Desktop/movies' into table StaticPart_movies partition (year = '1993');
亲爱的用户，下午好
让青春绽放精彩
54 秒
垃圾常清理，电脑更飞速
清理垃圾
HDFS://user/hive/wareh...
Hive
[*hive partion (~/Desk...
cloudera@localhost:~
2:34 PM
6/22/2016

Figure-7.5

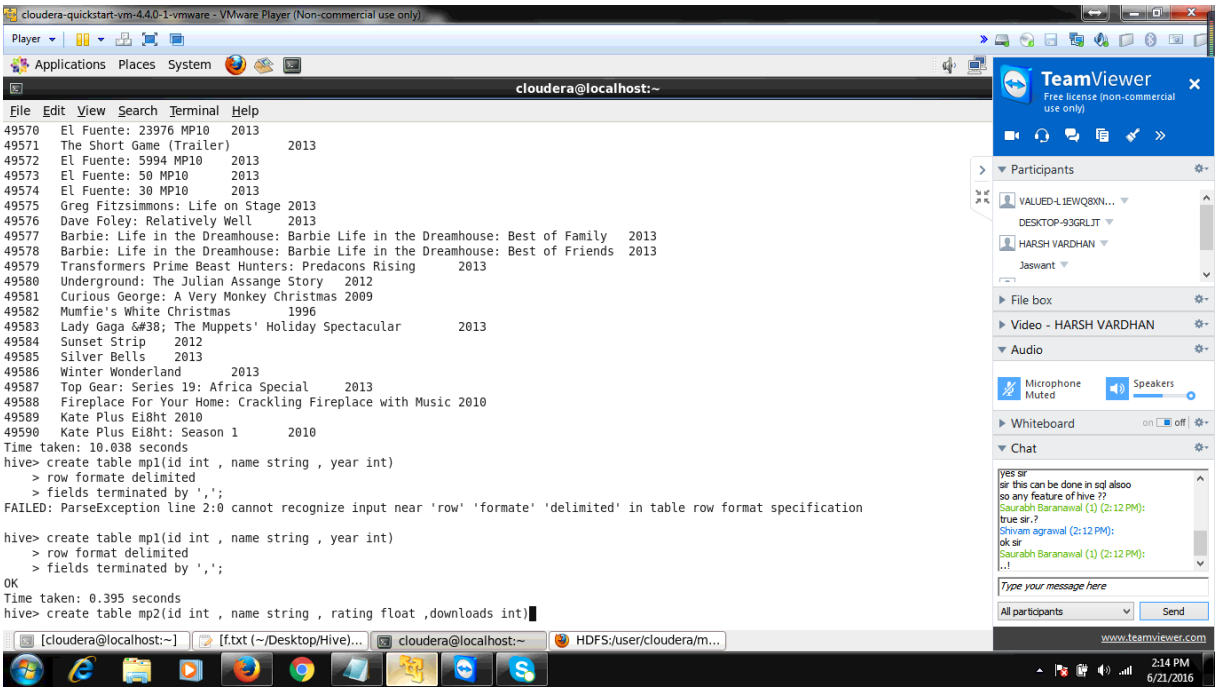

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
cloudera@localhost:~
File Edit View Search Terminal Help
49570 El Fuente: 23976 MP10 2013
49571 The Short Game (Trailer) 2013
49572 El Fuente: 5994 MP10 2013
49573 El Fuente: 50 MP10 2013
49574 El Fuente: 30 MP10 2013
49575 Greg Fitzsimmons: Life on Stage 2013
49576 Dave Foley: Relatively Well 2013
49577 Barbie: Life in the Dreamhouse: Barbie Life in the Dreamhouse: Best of Family 2013
49578 Barbie: Life in the Dreamhouse: Barbie Life in the Dreamhouse: Best of Friends 2013
49579 Transformers Prime Beast Hunters: Predacons Rising 2013
49580 Underground: The Julian Assange Story 2012
49581 Curious George: A Very Monkey Christmas 2009
49582 Mumfie's White Christmas 1996
49583 Lady Gaga & The Muppets' Holiday Spectacular 2013
49584 Sunset Strip 2012
49585 Silver Bells 2013
49586 Winter Wonderland 2013
49587 Top Gear: Series 19: Africa Special 2013
49588 Fireplace For Your Home: Crackling Fireplace with Music 2010
49589 Kate Plus Ei8ht 2010
49590 Kate Plus Ei8ht: Season 1 2010
Time taken: 10.038 seconds
hive> create table mp1(id int , name string , year int)
> row formate delimited
> fields terminated by ',';
FAILED: ParseException line 2:0 cannot recognize input near 'row' 'formate' 'delimited' in table row format specification
hive> create table mp1(id int , name string , year int)
> row format delimited
> fields terminated by ',';
OK
Time taken: 0.395 seconds
hive> create table mp2(id int , name string , rating float ,downloads int)
TeamViewer
Free license (non-commercial use only)
Participants
VALUED-L1EWQ8XN...
DESKTOP-93GRLJT
HARSH VARDHAN
Jaswant
File box
Video - HARSH VARDHAN
Audio
Microphone Muted
Speakers
Whiteboard
Chat
yes sir
sir this can be done in sql alsoo
so any feature of hive ??
Saurabh Baranawal (1) (2:12 PM):
true sir.?
Shivam agrawal (2:12 PM):
ok sir
Saurabh Baranawal (1) (2:12 PM):
..!
Type your message here
All participants
Send
www.teamviewer.com
[cloudera@localhost:~]
[f.txt (~/Desktop/Hive)...
cloudera@localhost:~
HDFS:/user/cloudera/m...
2:14 PM
6/21/2016

Figure -7.6

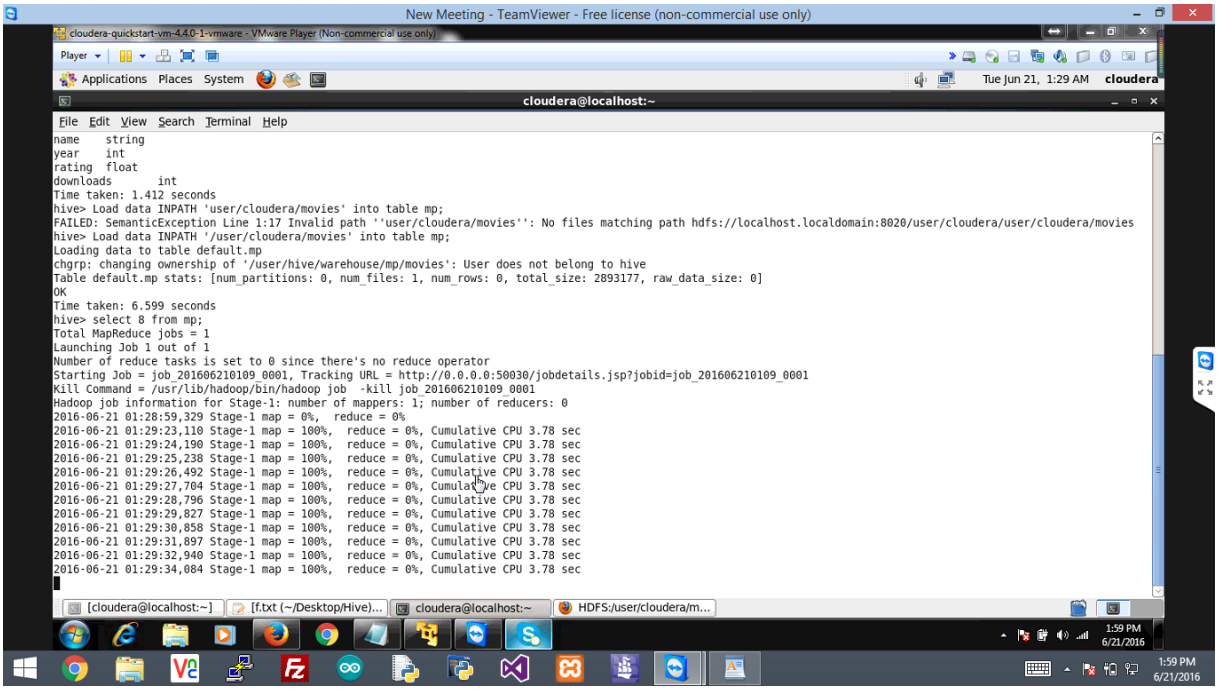

New Meeting - TeamViewer - Free license (non-commercial use only)
cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Tue Jun 21, 1:29 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
name string
year int
rating float
downloads int
Time taken: 1.412 seconds
hive> Load data INPATH 'user/cloudera/movies' into table mp;
FAILED: SemanticException Line 1:17 Invalid path ''user/cloudera/movies'': No files matching path hdfs://localhost.localdomain:8020/user/cloudera/user/cloudera/movies
hive> Load data INPATH '/user/cloudera/movies' into table mp;
Loading data to table default.mp
chgrp: changing ownership of '/user/hive/warehouse/mp/movies': User does not belong to hive
Table default.mp stats: [num_partitions: 0, num_files: 1, num_rows: 0, total_size: 2893177, raw_data_size: 0]
OK
Time taken: 6.599 seconds
hive> select 8 from mp;
Total MapReduce jobs = 1
Launching Job 1 out of 1
Number of reduce tasks is set to 0 since there's no reduce operator
Starting Job = job_201606210109_0001, Tracking URL = http://0.0.0.0:50030/jobdetails.jsp?jobid=job_201606210109_0001
Kill Command = /usr/lib/hadoop/bin/hadoop job -kill job_201606210109_0001
Hadoop job information for Stage-1: number of mappers: 1; number of reducers: 0
2016-06-21 01:28:59,329 Stage-1 map = 0%, reduce = 0%
2016-06-21 01:29:23,110 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.78 sec
2016-06-21 01:29:24,190 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.78 sec
2016-06-21 01:29:25,238 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.78 sec
2016-06-21 01:29:26,492 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.78 sec
2016-06-21 01:29:27,704 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.78 sec
2016-06-21 01:29:28,796 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.78 sec
2016-06-21 01:29:29,827 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.78 sec
2016-06-21 01:29:30,858 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.78 sec
2016-06-21 01:29:31,897 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.78 sec
2016-06-21 01:29:32,940 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.78 sec
2016-06-21 01:29:34,084 Stage-1 map = 100%, reduce = 0%, Cumulative CPU 3.78 sec
[cloudera@localhost:~]
[f.txt (~/Desktop/Hive)...
cloudera@localhost:~
HDFS:/user/cloudera/m...
1:59 PM
6/21/2016

Figure-7.7

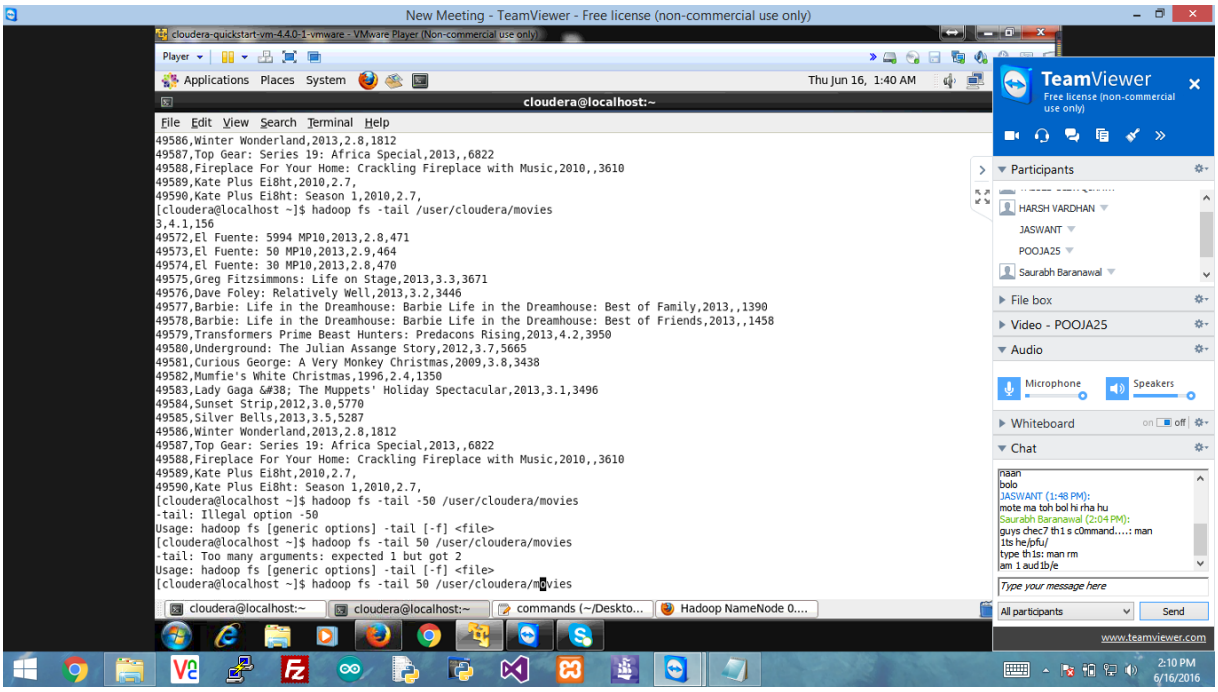

New Meeting - TeamViewer - Free license (non-commercial use only)
cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Applications Places System
Thu Jun 16, 1:40 AM
cloudera@localhost:~
File Edit View Search Terminal Help
49586,Winter Wonderland,2013,2.8,1812
49587,Top Gear: Series 19: Africa Special,2013,,6822
49588,Fireplace For Your Home: Crackling Fireplace with Music,2010,,3610
49589,Kate Plus Ei8ht,2010,2.7,
49590,Kate Plus Ei8ht: Season 1,2010,2.7,
[cloudera@localhost ~]$ hadoop fs -tail /user/cloudera/movies
3,4.1,156
49572,El Fuente: 5994 MP10,2013,2.8,471
49573,El Fuente: 50 MP10,2013,2.9,464
49574,El Fuente: 30 MP10,2013,2.8,470
49575,Greg Fitzsimmons: Life on Stage,2013,3.3,3671
49576,Dave Foley: Relatively Well,2013,3.2,3446
49577,Barbie: Life in the Dreamhouse: Barbie Life in the Dreamhouse: Best of Family,2013,,1390
49578,Barbie: Life in the Dreamhouse: Barbie Life in the Dreamhouse: Best of Friends,2013,,1458
49579,Transformers Prime Beast Hunters: Predacons Rising,2013,4.2,3950
49580,Underground: The Julian Assange Story,2012,3.7,5665
49581,Curious George: A Very Monkey Christmas,2009,3.8,3438
49582,Mumfie's White Christmas,1996,2.4,1350
49583,Lady Gaga & The Muppets' Holiday Spectacular,2013,3.1,3496
49584,Sunset Strip,2012,3.0,5770
49585,Silver Bells,2013,3.5,5287
49586,Winter Wonderland,2013,2.8,1812
49587,Top Gear: Series 19: Africa Special,2013,,6822
49588,Fireplace For Your Home: Crackling Fireplace with Music,2010,,3610
49589,Kate Plus Ei8ht,2010,2.7,
49590,Kate Plus Ei8ht: Season 1,2010,2.7,
[cloudera@localhost ~]$ hadoop fs -tail -50 /user/cloudera/movies
-tail: Illegal option -50
Usage: hadoop fs [generic options] -tail [-f] <file>
[cloudera@localhost ~]$ hadoop fs -tail 50 /user/cloudera/movies
-tail: Too many arguments: expected 1 but got 2
Usage: hadoop fs [generic options] -tail [-f] <file>
[cloudera@localhost ~]$ hadoop fs -tail 50 /user/cloudera/movies
cloudera@localhost:~
cloudera@localhost:~
commands (~/Deskto...
Hadoop NameNode 0....
TeamViewer
Free license (non-commercial use only)
Participants
HARSH VARDHAN
JASWANT
POOJA25
Saurabh Baranawal
File box
Video - POOJA25
Audio
Microphone
Speakers
Whiteboard
Chat
naan
bolo
JASWANT (1:48 PM):
mote ma toh bol hi rha hu
Saurabh Baranawal (2:04 PM):
guys chec7 th1 s c0mmand.....: man
1ts he/pfu/
type th1s: man rm
am 1 aud1b/e
Type your message here
All participants
Send
www.teamviewer.com
2:10 PM
6/16/2016

Figure-7.8

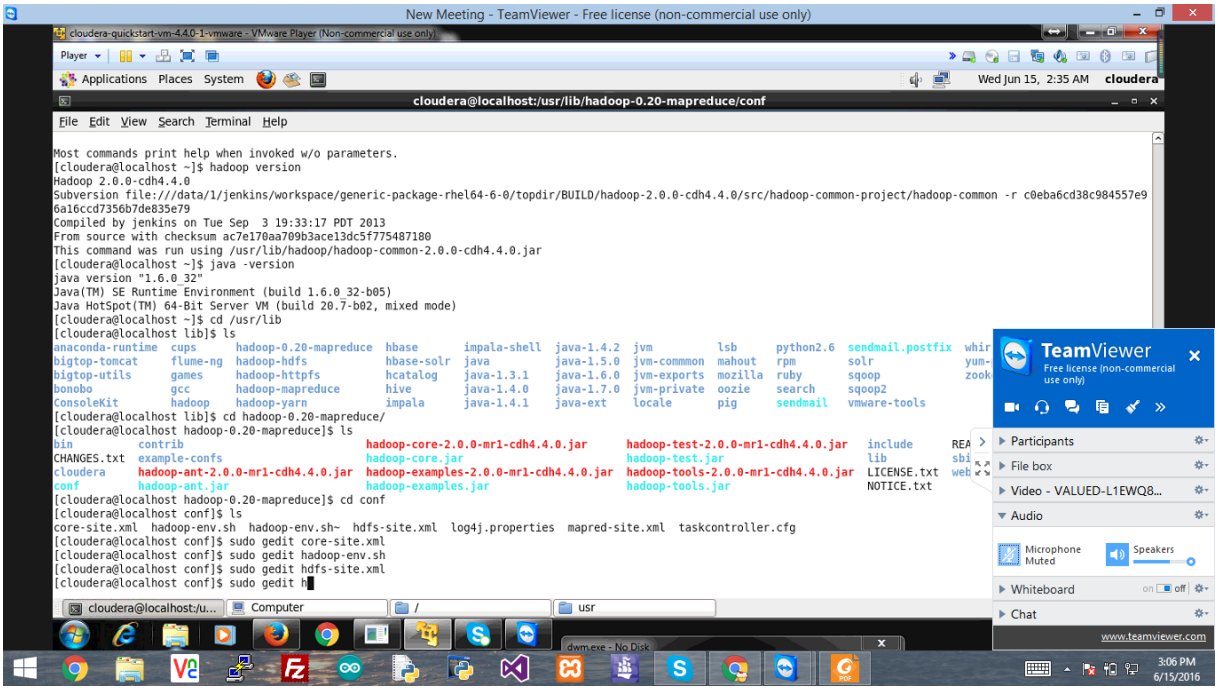

New Meeting - TeamViewer - Free license (non-commercial use only)
cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Applications Places System
Wed Jun 15, 2:35 AM cloudera
cloudera@localhost:/usr/lib/hadoop-0.20-mapreduce/conf
File Edit View Search Terminal Help
Most commands print help when invoked w/o parameters.
[cloudera@localhost ~]$ hadoop version
Hadoop 2.0.0-cdh4.4.0
Subversion file:///data/1/jenkins/workspace/generic-package-rhel64-6-0/topdir/BUILD/hadoop-2.0.0-cdh4.4.0/src/hadoop-common-project/hadoop-common -r c0eba6cd38c984557e96a16ccd7356b7de835e79
Compiled by jenkins on Tue Sep 3 19:33:17 PDT 2013
From source with checksum ac7e170aa709b3ace13dc5f775487180
This command was run using /usr/lib/hadoop/hadoop-common-2.0.0-cdh4.4.0.jar
[cloudera@localhost ~]$ java -version
java version "1.6.0_32"
Java(TM) SE Runtime Environment (build 1.6.0_32-b05)
Java HotSpot(TM) 64-Bit Server VM (build 20.7-b02, mixed mode)
[cloudera@localhost ~]$ cd /usr/lib
[cloudera@localhost lib]$ ls
anaconda-runtime cups hadoop-0.20-mapreduce hbase impala-shell java-1.4.2 jvm lsb python2.6 sendmail.postfix
bigtop-tomcat flume-ng hadoop-hdfs hbase-solr java java-1.5.0 jvm-commmon mahout rpm solr
bigtop-utils games hadoop-httpfs hcatalog java-1.3.1 java-1.6.0 jvm-exports mozilla ruby sqoop
bonobo gcc hadoop-mapreduce hive java-1.4.0 java-1.7.0 jvm-private oozie search sqoop2
ConsoleKit hadoop hadoop-yarn impala java-1.4.1 java-ext locale pig sendmail vmware-tools
[cloudera@localhost lib]$ cd hadoop-0.20-mapreduce/
[cloudera@localhost hadoop-0.20-mapreduce]$ ls
bin contrib hadoop-core-2.0.0-mr1-cdh4.4.0.jar hadoop-test-2.0.0-mr1-cdh4.4.0.jar include
CHANGES.txt example-confs hadoop-core.jar hadoop-test.jar lib
cloudera hadoop-ant-2.0.0-mr1-cdh4.4.0.jar hadoop-examples-2.0.0-mr1-cdh4.4.0.jar hadoop-tools-2.0.0-mr1-cdh4.4.0.jar LICENSE.txt
conf hadoop-ant.jar hadoop-examples.jar hadoop-tools.jar NOTICE.txt
[cloudera@localhost hadoop-0.20-mapreduce]$ cd conf
[cloudera@localhost conf]$ ls
core-site.xml hadoop-env.sh hadoop-env.sh~ hdfs-site.xml log4j.properties mapred-site.xml taskcontroller.cfg
[cloudera@localhost conf]$ sudo gedit core-site.xml
[cloudera@localhost conf]$ sudo gedit hadoop-env.sh
[cloudera@localhost conf]$ sudo gedit hdfs-site.xml
[cloudera@localhost conf]$ sudo gedit h
cloudera@localhost:/u...
Computer
/
usr
dwm.exe - No Disk
TeamViewer
Free license (non-commercial use only)
Participants
File box
Video - VALUED-L1EWQ8...
Audio
Microphone Muted
Speakers
Whiteboard
Chat
www.teamviewer.com
3:06 PM
6/15/2016

Figure-7.9

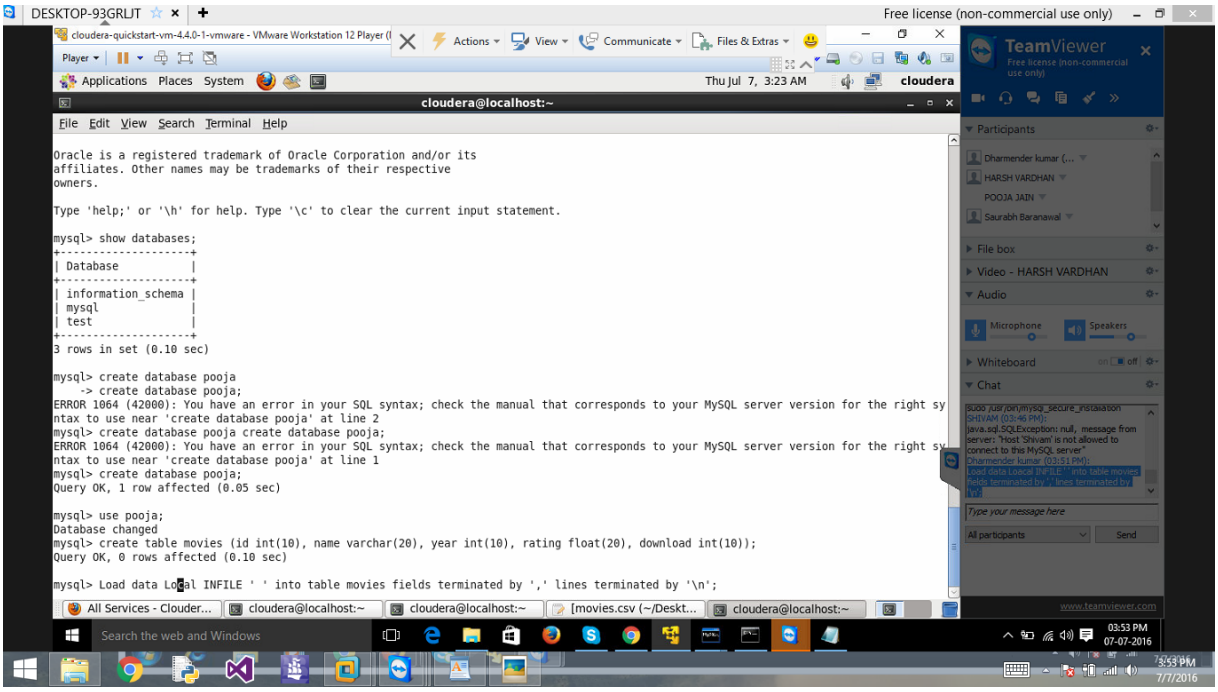

cloudera@localhost:~
File Edit View Search Terminal Help
Oracle is a registered trademark of Oracle Corporation and/or its
affiliates. Other names may be trademarks of their respective
owners.
Type 'help;' or '\h' for help. Type '\c' to clear the current input statement.
mysql> show databases;
| Database |
| information_schema |
| mysql |
| test |
3 rows in set (0.10 sec)
mysql> create database pooja
-> create database pooja;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right sy
ntax to use near 'create database pooja' at line 2
mysql> create database pooja create database pooja;
ERROR 1064 (42000): You have an error in your SQL syntax; check the manual that corresponds to your MySQL server version for the right sy
ntax to use near 'create database pooja' at line 1
mysql> create database pooja;
Query OK, 1 row affected (0.05 sec)
mysql> use pooja;
Database changed
mysql> create table movies (id int(10), name varchar(20), year int(10), rating float(20), download int(10));
Query OK, 0 rows affected (0.10 sec)
mysql> Load data Local INFILE ' ' into table movies fields terminated by ',' lines terminated by '\n';
TeamViewer
Participants
File box
Video - HARSH VARDHAN
Audio
Whiteboard
Chat

Figure-7.11

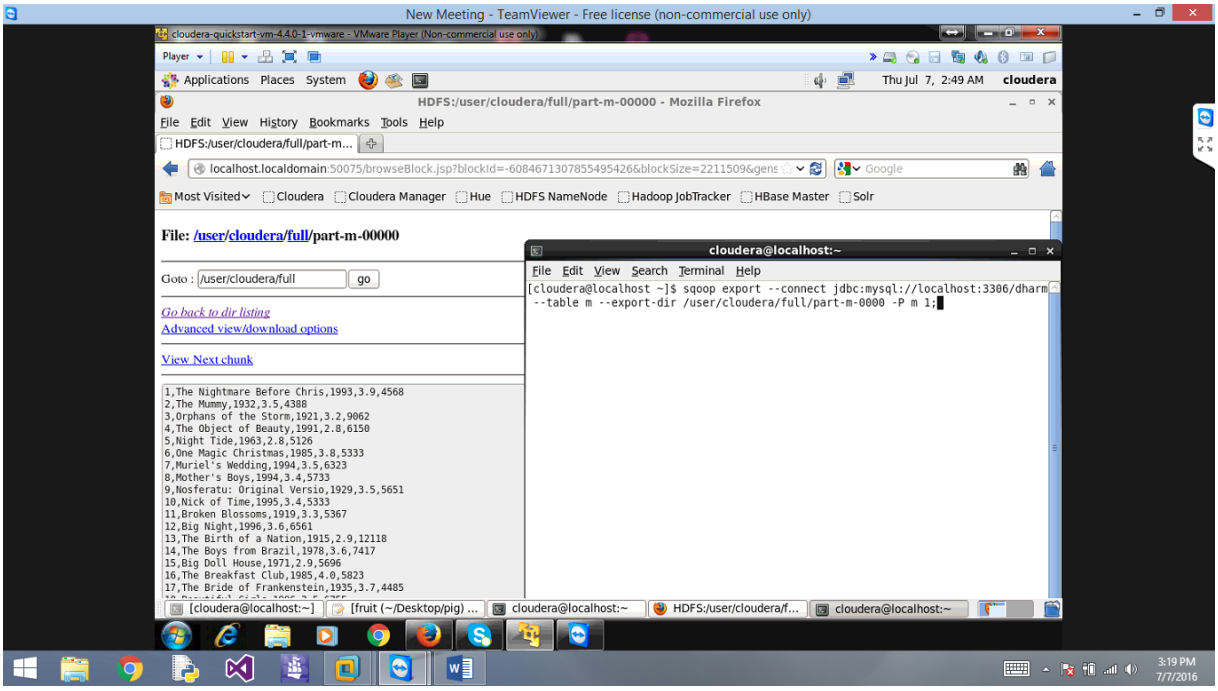

New Meeting - TeamViewer - Free license (non-commercial use only)
cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Applications Places System
Thu Jul 7, 2:49 AM cloudera
HDFS:/user/cloudera/full/part-m-00000 - Mozilla Firefox
File Edit View History Bookmarks Tools Help
HDFS:/user/cloudera/full/part-m...
localhost.localdomain:50075/browseBlock.jsp?blockId=-6084671307855495426&blockSize=2211509&gens
Most Visited Cloudera Cloudera Manager Hue HDFS NameNode Hadoop JobTracker HBase Master Solr
File: /user/cloudera/full/part-m-00000
Goto : /user/cloudera/full go
Go back to dir listing
Advanced view/download options
View Next chunk
1,The Nightmare Before Chris,1993,3.9,4568
2,The Mummy,1932,3.5,4388
3,Orphans of the Storm,1921,3.2,9062
4,The Object of Beauty,1991,2.8,6150
5,Night Tide,1963,2.8,5126
6,One Magic Christmas,1985,3.8,5333
7,Muriel's Wedding,1994,3.5,6323
8,Mother's Boys,1994,3.4,5733
9,Nosferatu: Original Versio,1929,3.5,5651
10,Nick of Time,1995,3.4,5333
11,Broken Blossoms,1919,3.3,5367
12,Big Night,1996,3.6,6561
13,The Birth of a Nation,1915,2.9,12118
14,The Boys from Brazil,1978,3.6,7417
15,Big Doll House,1971,2.9,5696
16,The Breakfast Club,1985,4.0,5823
17,The Bride of Frankenstein,1935,3.7,4485
cloudera@localhost:~
File Edit View Search Terminal Help
[cloudera@localhost ~]$ sqoop export --connect jdbc:mysql://localhost:3306/dharm
--table m --export-dir /user/cloudera/full/part-m-0000 -P m 1;
[cloudera@localhost:~] [fruit (~/Desktop/pig) ... cloudera@localhost:~ HDFS:/user/cloudera/f... cloudera@localhost:~
3:19 PM
7/7/2016

Figure-7.12

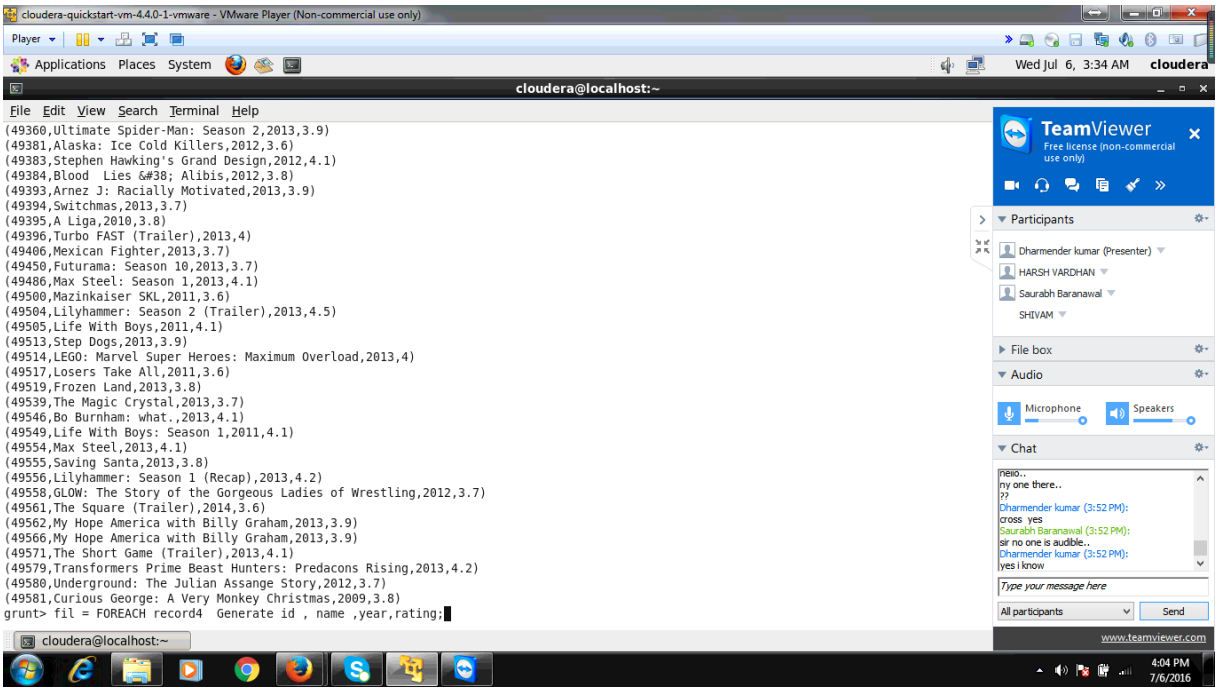

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 3:34 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
(49360,Ultimate Spider-Man: Season 2,2013,3.9)
(49381,Alaska: Ice Cold Killers,2012,3.6)
(49383,Stephen Hawking's Grand Design,2012,4.1)
(49384,Blood Lies & Alibis,2012,3.8)
(49393,Arnez J: Racially Motivated,2013,3.9)
(49394,Switchmas,2013,3.7)
(49395,A Liga,2010,3.8)
(49396,Turbo FAST (Trailer),2013,4)
(49406,Mexican Fighter,2013,3.7)
(49450,Futurama: Season 10,2013,3.7)
(49486,Max Steel: Season 1,2013,4.1)
(49500,Mazinkaiser SKL,2011,3.6)
(49504,Lilyhammer: Season 2 (Trailer),2013,4.5)
(49505,Life With Boys,2011,4.1)
(49513,Step Dogs,2013,3.9)
(49514,LEGO: Marvel Super Heroes: Maximum Overload,2013,4)
(49517,Losers Take All,2011,3.6)
(49519,Frozen Land,2013,3.8)
(49539,The Magic Crystal,2013,3.7)
(49546,Bo Burnham: what.,2013,4.1)
(49549,Life With Boys: Season 1,2011,4.1)
(49554,Max Steel,2013,4.1)
(49555,Saving Santa,2013,3.8)
(49556,Lilyhammer: Season 1 (Recap),2013,4.2)
(49558,GLOW: The Story of the Gorgeous Ladies of Wrestling,2012,3.7)
(49561,The Square (Trailer),2014,3.6)
(49562,My Hope America with Billy Graham,2013,3.9)
(49566,My Hope America with Billy Graham,2013,3.9)
(49571,The Short Game (Trailer),2013,4.1)
(49579,Transformers Prime Beast Hunters: Predacons Rising,2013,4.2)
(49580,Underground: The Julian Assange Story,2012,3.7)
(49581,Curious George: A Very Monkey Christmas,2009,3.8)
grunt> fil = FOREACH record4 Generate id , name ,year,rating;
TeamViewer
Free license (non-commercial use only)
Participants
Dharmender kumar (Presenter)
HARSH VARDHAN
Saurabh Baranawal
SHIVAM
File box
Audio
Microphone
Speakers
Chat
Type your message here
All participants
Send
www.teamviewer.com
cloudera@localhost:~
4:04 PM
7/6/2016

Figure-7.13

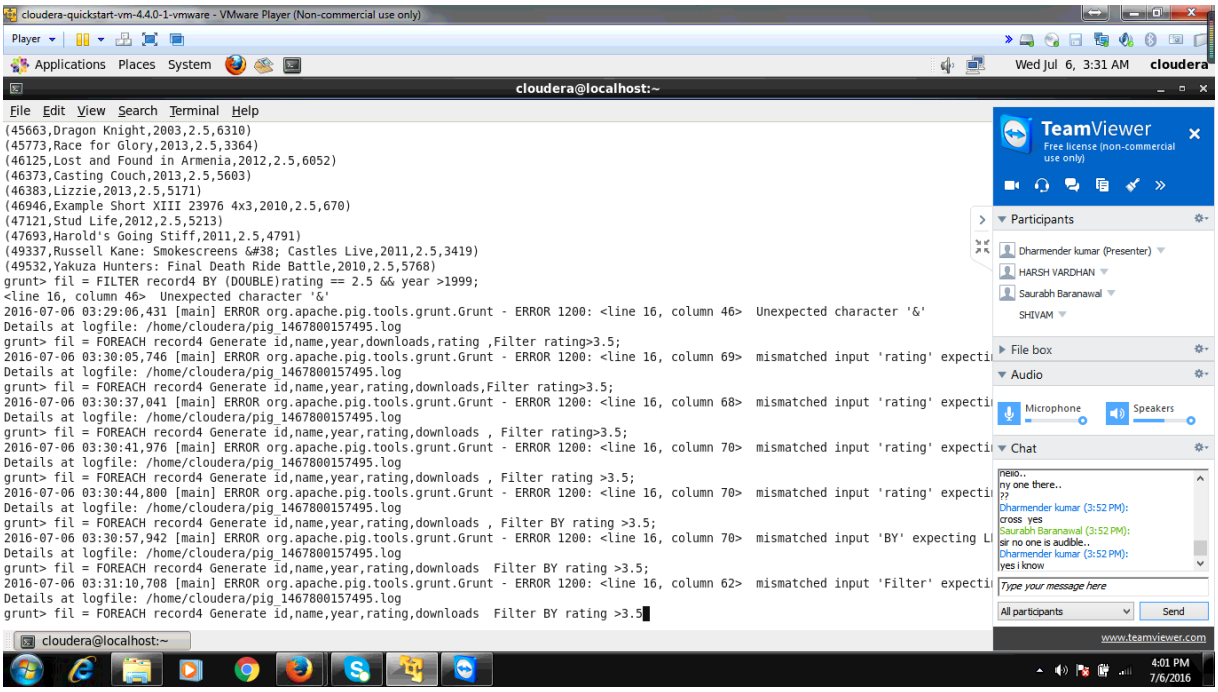

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 3:31 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
(45663,Dragon Knight,2003,2.5,6310)
(45773,Race for Glory,2013,2.5,3364)
(46125,Lost and Found in Armenia,2012,2.5,6052)
(46373,Casting Couch,2013,2.5,5603)
(46383,Lizzie,2013,2.5,5171)
(46946,Example Short XIII 23976 4x3,2010,2.5,670)
(47121,Stud Life,2012,2.5,5213)
(47693,Harold's Going Stiff,2011,2.5,4791)
(49337,Russell Kane: Smokescreens & Castles Live,2011,2.5,3419)
(49532,Yakuza Hunters: Final Death Ride Battle,2010,2.5,5768)
grunt> fil = FILTER record4 BY (DOUBLE)rating == 2.5 && year >1999;
<line 16, column 46> Unexpected character '&'
2016-07-06 03:29:06,431 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 16, column 46> Unexpected character '&'
Details at logfile: /home/cloudera/pig_1467800157495.log
grunt> fil = FOREACH record4 Generate id,name,year,downloads,rating ,Filter rating>3.5;
2016-07-06 03:30:05,746 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 16, column 69> mismatched input 'rating' expecti
Details at logfile: /home/cloudera/pig_1467800157495.log
grunt> fil = FOREACH record4 Generate id,name,year,rating,downloads,Filter rating>3.5;
2016-07-06 03:30:37,041 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 16, column 68> mismatched input 'rating' expecti
Details at logfile: /home/cloudera/pig_1467800157495.log
grunt> fil = FOREACH record4 Generate id,name,year,rating,downloads , Filter rating>3.5;
2016-07-06 03:30:41,976 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 16, column 70> mismatched input 'rating' expecti
Details at logfile: /home/cloudera/pig_1467800157495.log
grunt> fil = FOREACH record4 Generate id,name,year,rating,downloads , Filter rating >3.5;
2016-07-06 03:30:44,800 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 16, column 70> mismatched input 'rating' expecti
Details at logfile: /home/cloudera/pig_1467800157495.log
grunt> fil = FOREACH record4 Generate id,name,year,rating,downloads , Filter BY rating >3.5;
2016-07-06 03:30:57,942 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 16, column 70> mismatched input 'BY' expecting LI
Details at logfile: /home/cloudera/pig_1467800157495.log
grunt> fil = FOREACH record4 Generate id,name,year,rating,downloads Filter BY rating >3.5;
2016-07-06 03:31:10,708 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 16, column 62> mismatched input 'Filter' expecti
Details at logfile: /home/cloudera/pig_1467800157495.log
grunt> fil = FOREACH record4 Generate id,name,year,rating,downloads Filter BY rating >3.5
TeamViewer
Free license (non-commercial use only)
Participants
Dharmender kumar (Presenter)
HARSH VARDHAN
Saurabh Baranawal
SHIVAM
File box
Audio
Microphone
Speakers
Chat
nello..
ny one there..
??
Dharmender kumar (3:52 PM):
cross yes
Saurabh Baranawal (3:52 PM):
sir no one is audible..
Dharmender kumar (3:52 PM):
yes I know
Type your message here
All participants
Send
www.teamviewer.com
cloudera@localhost:~
4:01 PM
7/6/2016

Figure-7.14

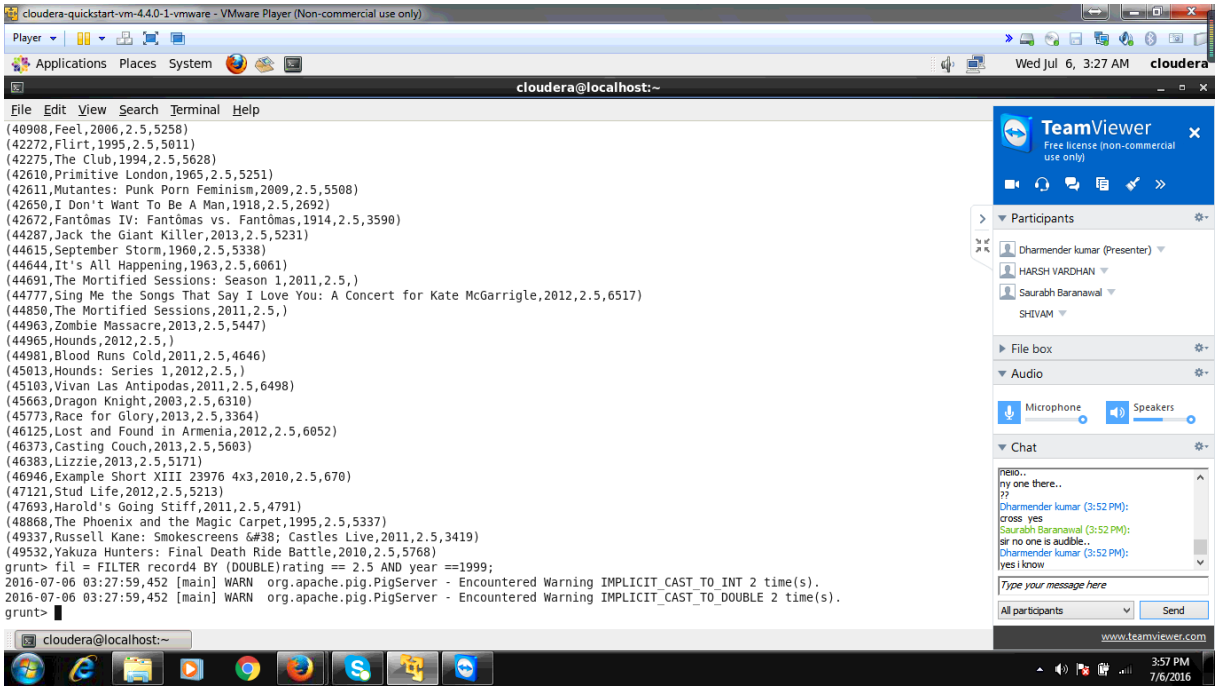

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 3:27 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
(40908,Feel,2006,2.5,5258)
(42272,Flirt,1995,2.5,5011)
(42275,The Club,1994,2.5,5628)
(42610,Primitive London,1965,2.5,5251)
(42611,Mutantes: Punk Porn Feminism,2009,2.5,5508)
(42650,I Don't Want To Be A Man,1918,2.5,2692)
(42672,Fantômas IV: Fantômas vs. Fantômas,1914,2.5,3590)
(44287,Jack the Giant Killer,2013,2.5,5231)
(44615,September Storm,1960,2.5,5338)
(44644,It's All Happening,1963,2.5,6061)
(44691,The Mortified Sessions: Season 1,2011,2.5,)
(44777,Sing Me the Songs That Say I Love You: A Concert for Kate McGarrigle,2012,2.5,6517)
(44850,The Mortified Sessions,2011,2.5,)
(44963,Zombie Massacre,2013,2.5,5447)
(44965,Hounds,2012,2.5,)
(44981,Blood Runs Cold,2011,2.5,4646)
(45013,Hounds: Series 1,2012,2.5,)
(45103,Vivan Las Antipodas,2011,2.5,6498)
(45663,Dragon Knight,2003,2.5,6310)
(45773,Race for Glory,2013,2.5,3364)
(46125,Lost and Found in Armenia,2012,2.5,6052)
(46373,Casting Couch,2013,2.5,5603)
(46383,Lizzie,2013,2.5,5171)
(46946,Example Short XIII 23976 4x3,2010,2.5,670)
(47121,Stud Life,2012,2.5,5213)
(47693,Harold's Going Stiff,2011,2.5,4791)
(48868,The Phoenix and the Magic Carpet,1995,2.5,5337)
(49337,Russell Kane: Smokescreens & Castles Live,2011,2.5,3419)
(49532,Yakuza Hunters: Final Death Ride Battle,2010,2.5,5768)
grunt> fil = FILTER record4 BY (DOUBLE)rating == 2.5 AND year ==1999;
2016-07-06 03:27:59,452 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_INT 2 time(s).
2016-07-06 03:27:59,452 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 2 time(s).
grunt>
TeamViewer
Free license (non-commercial use only)
Participants
Dharmender kumar (Presenter)
HARSH VARDHAN
Saurabh Baranawal
SHIVAM
File box
Audio
Microphone
Speakers
Chat
Dharmender kumar (3:52 PM):
cross yes
Saurabh Baranawal (3:52 PM):
sir no one is audible..
Dharmender kumar (3:52 PM):
yes i know
Type your message here
All participants
Send
www.teamviewer.com
cloudera@localhost:~
3:57 PM
7/6/2016

Figure-7.15

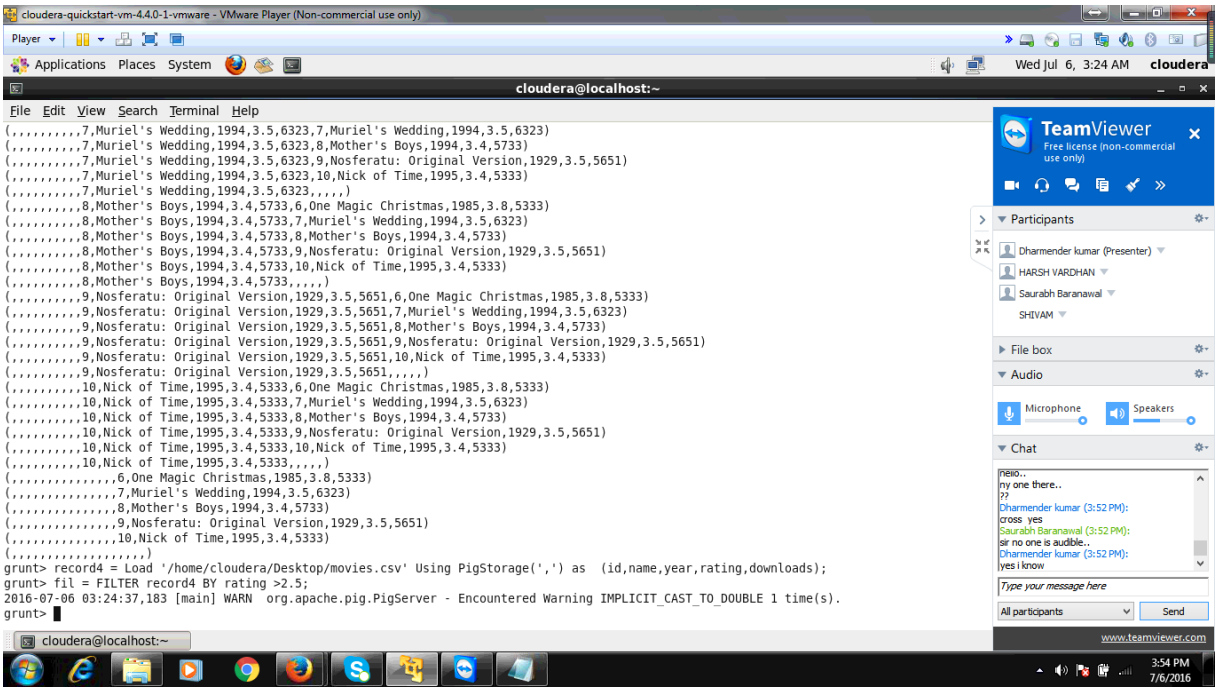

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 3:24 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
(,,,,,,,,,,7,Muriel's Wedding,1994,3.5,6323,7,Muriel's Wedding,1994,3.5,6323)
(,,,,,,,,,,7,Muriel's Wedding,1994,3.5,6323,8,Mother's Boys,1994,3.4,5733)
(,,,,,,,,,,7,Muriel's Wedding,1994,3.5,6323,9,Nosferatu: Original Version,1929,3.5,5651)
(,,,,,,,,,,7,Muriel's Wedding,1994,3.5,6323,10,Nick of Time,1995,3.4,5333)
(,,,,,,,,,,7,Muriel's Wedding,1994,3.5,6323,,,,,)
(,,,,,,,,,,8,Mother's Boys,1994,3.4,5733,6,One Magic Christmas,1985,3.8,5333)
(,,,,,,,,,,8,Mother's Boys,1994,3.4,5733,7,Muriel's Wedding,1994,3.5,6323)
(,,,,,,,,,,8,Mother's Boys,1994,3.4,5733,8,Mother's Boys,1994,3.4,5733)
(,,,,,,,,,,8,Mother's Boys,1994,3.4,5733,9,Nosferatu: Original Version,1929,3.5,5651)
(,,,,,,,,,,8,Mother's Boys,1994,3.4,5733,10,Nick of Time,1995,3.4,5333)
(,,,,,,,,,,8,Mother's Boys,1994,3.4,5733,,,,,)
(,,,,,,,,,,9,Nosferatu: Original Version,1929,3.5,5651,6,One Magic Christmas,1985,3.8,5333)
(,,,,,,,,,,9,Nosferatu: Original Version,1929,3.5,5651,7,Muriel's Wedding,1994,3.5,6323)
(,,,,,,,,,,9,Nosferatu: Original Version,1929,3.5,5651,8,Mother's Boys,1994,3.4,5733)
(,,,,,,,,,,9,Nosferatu: Original Version,1929,3.5,5651,9,Nosferatu: Original Version,1929,3.5,5651)
(,,,,,,,,,,9,Nosferatu: Original Version,1929,3.5,5651,10,Nick of Time,1995,3.4,5333)
(,,,,,,,,,,9,Nosferatu: Original Version,1929,3.5,5651,,,,,)
(,,,,,,,,,,10,Nick of Time,1995,3.4,5333,6,One Magic Christmas,1985,3.8,5333)
(,,,,,,,,,,10,Nick of Time,1995,3.4,5333,7,Muriel's Wedding,1994,3.5,6323)
(,,,,,,,,,,10,Nick of Time,1995,3.4,5333,8,Mother's Boys,1994,3.4,5733)
(,,,,,,,,,,10,Nick of Time,1995,3.4,5333,9,Nosferatu: Original Version,1929,3.5,5651)
(,,,,,,,,,,10,Nick of Time,1995,3.4,5333,10,Nick of Time,1995,3.4,5333)
(,,,,,,,,,,10,Nick of Time,1995,3.4,5333,,,,,)
(,,,,,,,,,,,,,,,6,One Magic Christmas,1985,3.8,5333)
(,,,,,,,,,,,,,,,7,Muriel's Wedding,1994,3.5,6323)
(,,,,,,,,,,,,,,,8,Mother's Boys,1994,3.4,5733)
(,,,,,,,,,,,,,,,9,Nosferatu: Original Version,1929,3.5,5651)
(,,,,,,,,,,,,,,,10,Nick of Time,1995,3.4,5333)
(,,,,,,,,,,,,,,,,,,)
grunt> record4 = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') as (id,name,year,rating,downloads);
grunt> fil = FILTER record4 BY rating >2.5;
2016-07-06 03:24:37,183 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_DOUBLE 1 time(s).
grunt>
cloudera@localhost:~
TeamViewer
Free license (non-commercial use only)
Participants
Dharmender kumar (Presenter)
HARSH VARDHAN
Saurabh Baranawal
SHIVAM
File box
Audio
Microphone
Speakers
Chat
hello..
ny one there..
??
Dharmender kumar (3:52 PM):
cross yes
Saurabh Baranawal (3:52 PM):
sir no one is audible..
Dharmender kumar (3:52 PM):
yes I know
Type your message here
All participants
Send
www.teamviewer.com
3:54 PM
7/6/2016

Figure-7.16

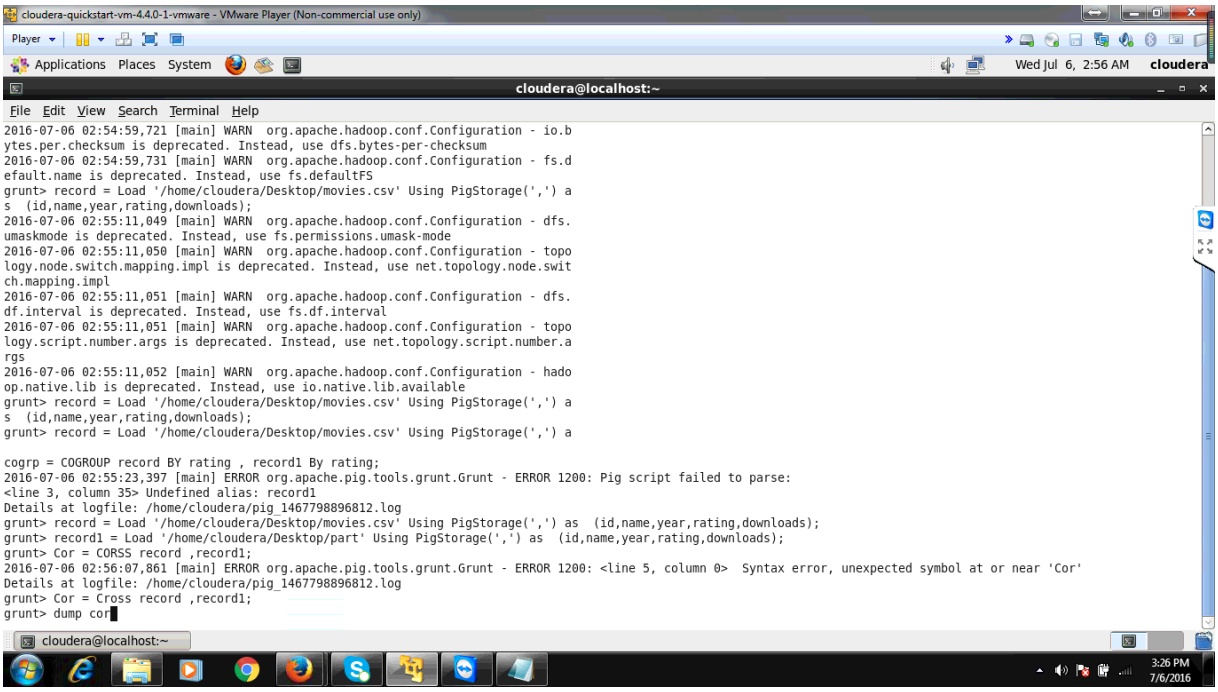

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 2:56 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
2016-07-06 02:54:59,721 [main] WARN org.apache.hadoop.conf.Configuration - io.b
ytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2016-07-06 02:54:59,731 [main] WARN org.apache.hadoop.conf.Configuration - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') a
s (id,name,year,rating,downloads);
2016-07-06 02:55:11,049 [main] WARN org.apache.hadoop.conf.Configuration - dfs.
umaskmode is deprecated. Instead, use fs.permissions.umask-mode
2016-07-06 02:55:11,050 [main] WARN org.apache.hadoop.conf.Configuration - topo
logy.node.switch.mapping.impl is deprecated. Instead, use net.topology.node.swit
ch.mapping.impl
2016-07-06 02:55:11,051 [main] WARN org.apache.hadoop.conf.Configuration - dfs.
df.interval is deprecated. Instead, use fs.df.interval
2016-07-06 02:55:11,051 [main] WARN org.apache.hadoop.conf.Configuration - topo
logy.script.number.args is deprecated. Instead, use net.topology.script.number.a
rgs
2016-07-06 02:55:11,052 [main] WARN org.apache.hadoop.conf.Configuration - hado
op.native.lib is deprecated. Instead, use io.native.lib.available
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') a
s (id,name,year,rating,downloads);
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') a
cogrp = COGROUP record BY rating , record1 By rating;
2016-07-06 02:55:23,397 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: Pig script failed to parse:
<line 3, column 35> Undefined alias: record1
Details at logfile: /home/cloudera/pig_1467798896812.log
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') as (id,name,year,rating,downloads);
grunt> record1 = Load '/home/cloudera/Desktop/part' Using PigStorage(',') as (id,name,year,rating,downloads);
grunt> Cor = CORSS record ,record1;
2016-07-06 02:56:07,861 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 5, column 0> Syntax error, unexpected symbol at or near 'Cor'
Details at logfile: /home/cloudera/pig_1467798896812.log
grunt> Cor = Cross record ,record1;
grunt> dump cor
cloudera@localhost:~
3:26 PM
7/6/2016

Figure-7.17

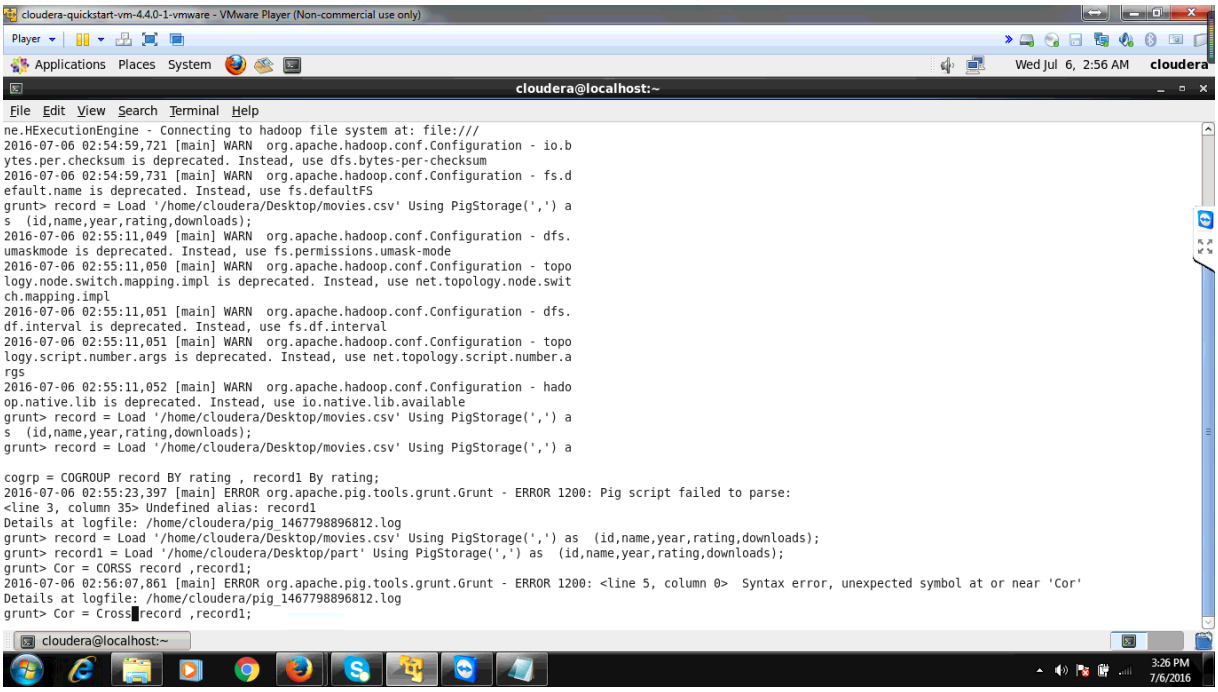

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 2:56 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
ne.HExecutionEngine - Connecting to hadoop file system at: file:///
2016-07-06 02:54:59,721 [main] WARN org.apache.hadoop.conf.Configuration - io.b
ytes.per.checksum is deprecated. Instead, use dfs.bytes-per-checksum
2016-07-06 02:54:59,731 [main] WARN org.apache.hadoop.conf.Configuration - fs.d
efault.name is deprecated. Instead, use fs.defaultFS
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') a
s (id,name,year,rating,downloads);
2016-07-06 02:55:11,049 [main] WARN org.apache.hadoop.conf.Configuration - dfs.
umaskmode is deprecated. Instead, use fs.permissions.umask-mode
2016-07-06 02:55:11,050 [main] WARN org.apache.hadoop.conf.Configuration - topo
logy.node.switch.mapping.impl is deprecated. Instead, use net.topology.node.swit
ch.mapping.impl
2016-07-06 02:55:11,051 [main] WARN org.apache.hadoop.conf.Configuration - dfs.
df.interval is deprecated. Instead, use fs.df.interval
2016-07-06 02:55:11,051 [main] WARN org.apache.hadoop.conf.Configuration - topo
logy.script.number.args is deprecated. Instead, use net.topology.script.number.a
rgs
2016-07-06 02:55:11,052 [main] WARN org.apache.hadoop.conf.Configuration - hado
op.native.lib is deprecated. Instead, use io.native.lib.available
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') a
s (id,name,year,rating,downloads);
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') a
cogrp = COGROUP record BY rating , record1 By rating;
2016-07-06 02:55:23,397 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: Pig script failed to parse:
<line 3, column 35> Undefined alias: record1
Details at logfile: /home/cloudera/pig_1467798896812.log
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') as (id,name,year,rating,downloads);
grunt> record1 = Load '/home/cloudera/Desktop/part' Using PigStorage(',') as (id,name,year,rating,downloads);
grunt> Cor = CORSS record ,record1;
2016-07-06 02:56:07,861 [main] ERROR org.apache.pig.tools.grunt.Grunt - ERROR 1200: <line 5, column 0> Syntax error, unexpected symbol at or near 'Cor'
Details at logfile: /home/cloudera/pig_1467798896812.log
grunt> Cor = Cross record ,record1;
cloudera@localhost:~
3:26 PM
7/6/2016

Figure-7.18

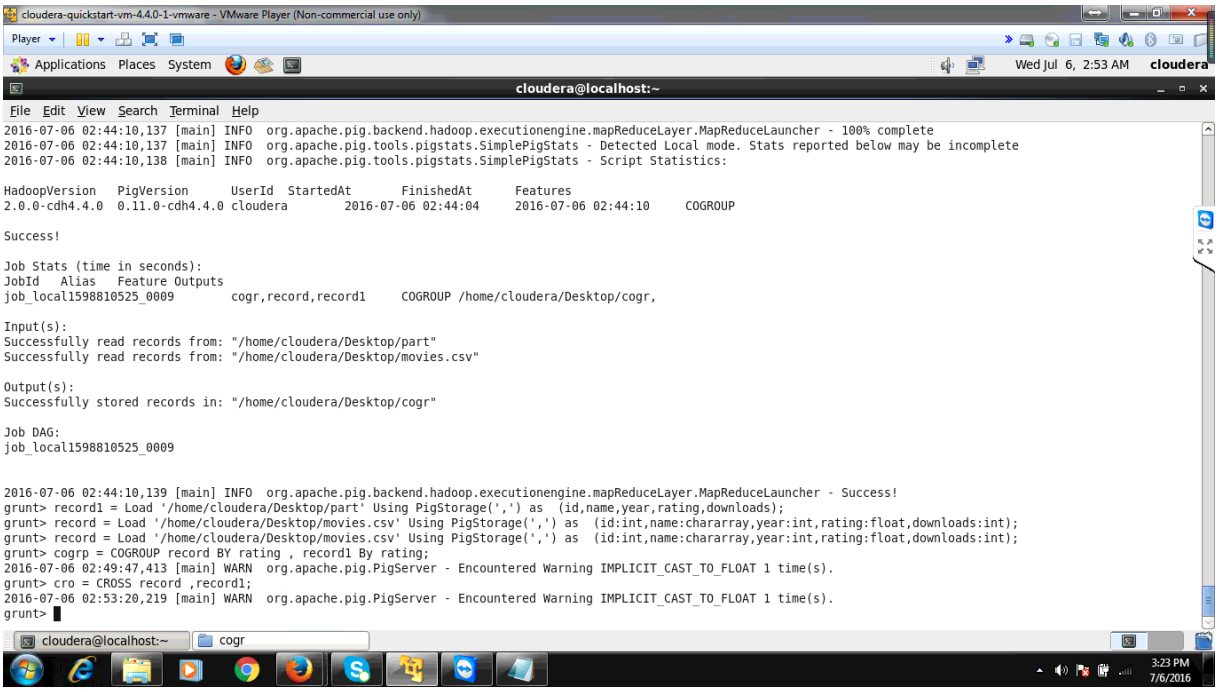

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 2:53 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
2016-07-06 02:44:10,137 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-07-06 02:44:10,137 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Detected Local mode. Stats reported below may be incomplete
2016-07-06 02:44:10,138 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.0.0-cdh4.4.0 0.11.0-cdh4.4.0 cloudera 2016-07-06 02:44:04 2016-07-06 02:44:10 COGROUP
Success!
Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local1598810525_0009 cogr,record,record1 COGROUP /home/cloudera/Desktop/cogr,
Input(s):
Successfully read records from: "/home/cloudera/Desktop/part"
Successfully read records from: "/home/cloudera/Desktop/movies.csv"
Output(s):
Successfully stored records in: "/home/cloudera/Desktop/cogr"
Job DAG:
job_local1598810525_0009
2016-07-06 02:44:10,139 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> record1 = Load '/home/cloudera/Desktop/part' Using PigStorage(',') as (id,name,year,rating,downloads);
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') as (id:int,name:chararray,year:int,rating:float,downloads:int);
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') as (id:int,name:chararray,year:int,rating:float,downloads:int);
grunt> cogrp = COGROUP record BY rating , record1 By rating;
2016-07-06 02:49:47,413 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_FLOAT 1 time(s).
grunt> cro = CROSS record ,record1;
2016-07-06 02:53:20,219 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_FLOAT 1 time(s).
grunt>
cloudera@localhost:~
cogr
3:23 PM
7/6/2016

Figure-7.19

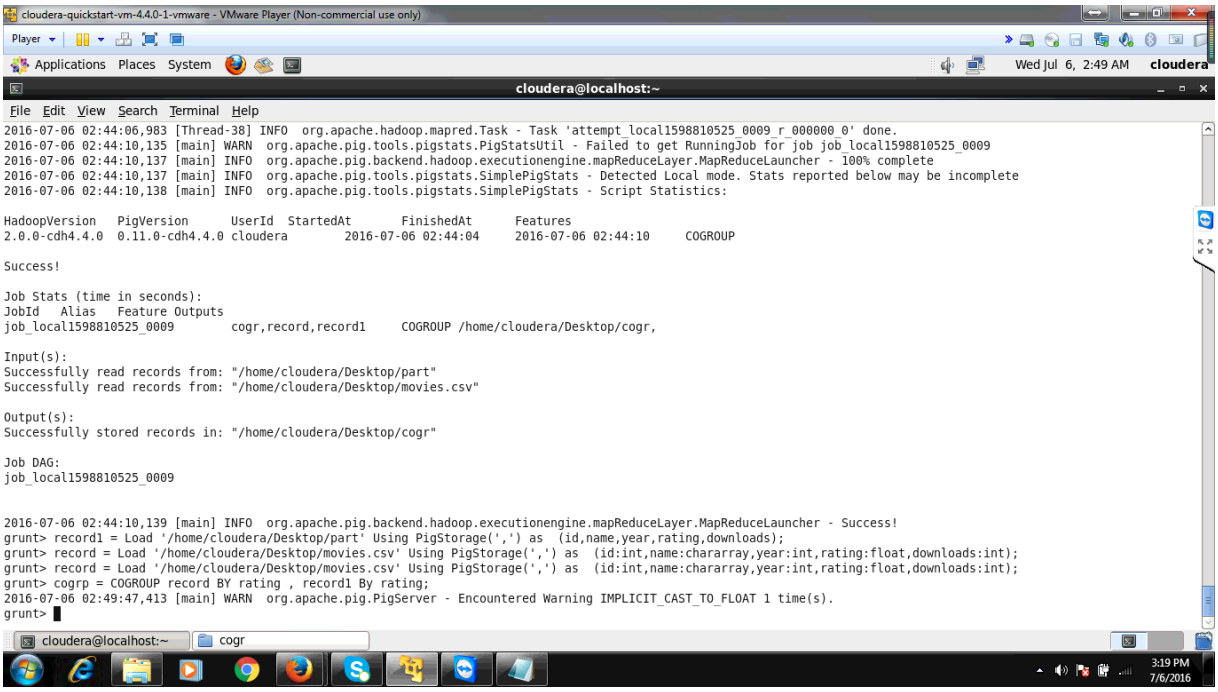

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 2:49 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
2016-07-06 02:44:06,983 [Thread-38] INFO org.apache.hadoop.mapred.Task - Task 'attempt_local1598810525_0009_r_000000_0' done.
2016-07-06 02:44:10,135 [main] WARN org.apache.pig.tools.pigstats.PigStatsUtil - Failed to get RunningJob for job job_local1598810525_0009
2016-07-06 02:44:10,137 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-07-06 02:44:10,137 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Detected Local mode. Stats reported below may be incomplete
2016-07-06 02:44:10,138 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.0.0-cdh4.4.0 0.11.0-cdh4.4.0 cloudera 2016-07-06 02:44:04 2016-07-06 02:44:10 COGROUP
Success!
Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local1598810525_0009 cogr,record,record1 COGROUP /home/cloudera/Desktop/cogr,
Input(s):
Successfully read records from: "/home/cloudera/Desktop/part"
Successfully read records from: "/home/cloudera/Desktop/movies.csv"
Output(s):
Successfully stored records in: "/home/cloudera/Desktop/cogr"
Job DAG:
job_local1598810525_0009
2016-07-06 02:44:10,139 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> record1 = Load '/home/cloudera/Desktop/part' Using PigStorage(',') as (id,name,year,rating,downloads);
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') as (id:int,name:chararray,year:int,rating:float,downloads:int);
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') as (id:int,name:chararray,year:int,rating:float,downloads:int);
grunt> cogrp = COGROUP record BY rating , record1 By rating;
2016-07-06 02:49:47,413 [main] WARN org.apache.pig.PigServer - Encountered Warning IMPLICIT_CAST_TO_FLOAT 1 time(s).
grunt>
cloudera@localhost:~
cogr
3:19 PM
7/6/2016

Figure-7.20

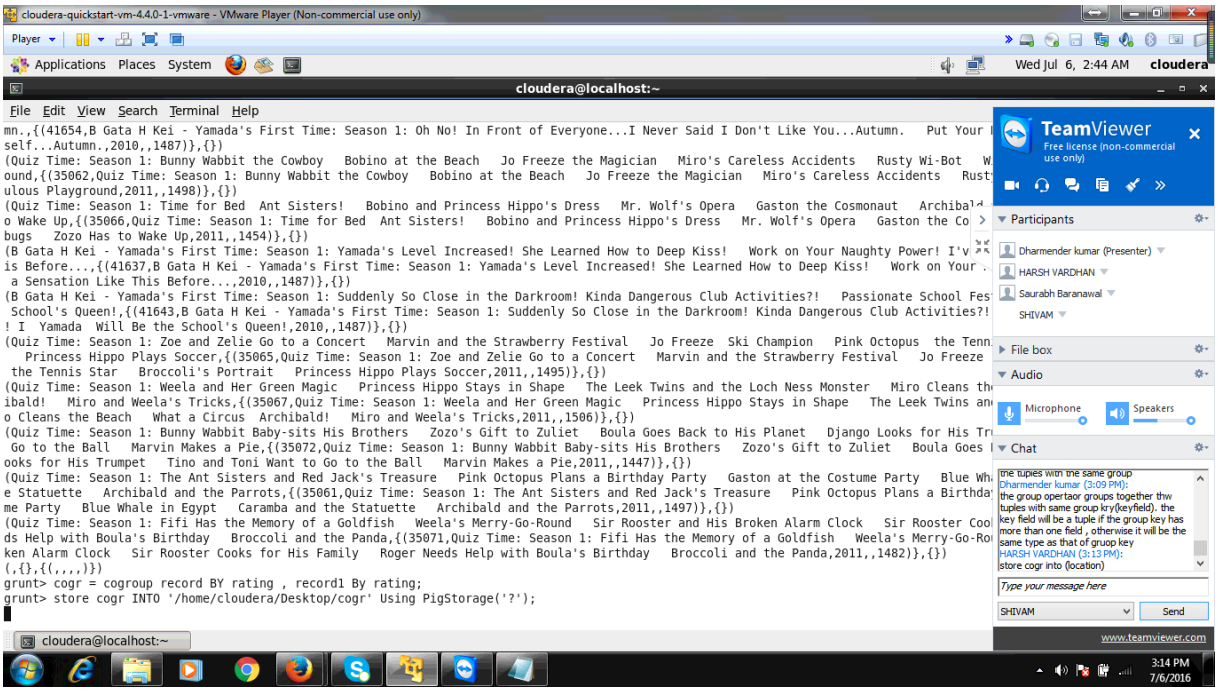

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 2:44 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
mn.,{(41654,B Gata H Kei - Yamada's First Time: Season 1: Oh No! In Front of Everyone...I Never Said I Don't Like You...Autumn. Put Your
self...Autumn.,2010,,1487)},{})
(Quiz Time: Season 1: Bunny Wabbit the Cowboy Bobino at the Beach Jo Freeze the Magician Miro's Careless Accidents Rusty Wi-Bot
ound,{(35062,Quiz Time: Season 1: Bunny Wabbit the Cowboy Bobino at the Beach Jo Freeze the Magician Miro's Careless Accidents
ulous Playground,2011,,1498)},{})
(Quiz Time: Season 1: Time for Bed Ant Sisters! Bobino and Princess Hippo's Dress Mr. Wolf's Opera Gaston the Cosmonaut
o Wake Up,{(35066,Quiz Time: Season 1: Time for Bed Ant Sisters! Bobino and Princess Hippo's Dress Mr. Wolf's Opera Gaston the Co
bugs Zozo Has to Wake Up,2011,,1454)},{})
(B Gata H Kei - Yamada's First Time: Season 1: Yamada's Level Increased! She Learned How to Deep Kiss! Work on Your Naughty Power! I'v
is Before...,{(41637,B Gata H Kei - Yamada's First Time: Season 1: Yamada's Level Increased! She Learned How to Deep Kiss! Work on Your
a Sensation Like This Before...,2010,,1487)},{})
(B Gata H Kei - Yamada's First Time: Season 1: Suddenly So Close in the Darkroom! Kinda Dangerous Club Activities?! Passionate School Fes
School's Queen!,{(41643,B Gata H Kei - Yamada's First Time: Season 1: Suddenly So Close in the Darkroom! Kinda Dangerous Club Activities?!
! I Yamada Will Be the School's Queen!,2010,,1487)},{})
(Quiz Time: Season 1: Zoe and Zelie Go to a Concert Marvin and the Strawberry Festival Jo Freeze Ski Champion Pink Octopus the Tenn
Princess Hippo Plays Soccer,{(35065,Quiz Time: Season 1: Zoe and Zelie Go to a Concert Marvin and the Strawberry Festival Jo Freeze
the Tennis Star Broccoli's Portrait Princess Hippo Plays Soccer,2011,,1495)},{})
(Quiz Time: Season 1: Weela and Her Green Magic Princess Hippo Stays in Shape The Leek Twins and the Loch Ness Monster Miro Cleans th
ibald! Miro and Weela's Tricks,{(35067,Quiz Time: Season 1: Weela and Her Green Magic Princess Hippo Stays in Shape The Leek Twins an
o Cleans the Beach What a Circus Archibald! Miro and Weela's Tricks,2011,,1506)},{})
(Quiz Time: Season 1: Bunny Wabbit Baby-sits His Brothers Zozo's Gift to Zuliet Boula Goes Back to His Planet Django Looks for His Tr
Go to the Ball Marvin Makes a Pie,{(35072,Quiz Time: Season 1: Bunny Wabbit Baby-sits His Brothers Zozo's Gift to Zuliet Boula Goes
ooks for His Trumpet Tino and Toni Want to Go to the Ball Marvin Makes a Pie,2011,,1447)},{})
(Quiz Time: Season 1: The Ant Sisters and Red Jack's Treasure Pink Octopus Plans a Birthday Party Gaston at the Costume Party Blue Wh
e Statuette Archibald and the Parrots,{(35061,Quiz Time: Season 1: The Ant Sisters and Red Jack's Treasure Pink Octopus Plans a Birthda
me Party Blue Whale in Egypt Caramba and the Statuette Archibald and the Parrots,2011,,1497)},{})
(Quiz Time: Season 1: Fifi Has the Memory of a Goldfish Weela's Merry-Go-Round Sir Rooster and His Broken Alarm Clock Sir Rooster Coo
ds Help with Boula's Birthday Broccoli and the Panda,{(35071,Quiz Time: Season 1: Fifi Has the Memory of a Goldfish Weela's Merry-Go-Ro
ken Alarm Clock Sir Rooster Cooks for His Family Roger Needs Help with Boula's Birthday Broccoli and the Panda,2011,,1482)},{})
(,{},{(,,,,)})
grunt> cogr = cogroup record BY rating , record1 By rating;
grunt> store cogr INTO '/home/cloudera/Desktop/cogr' Using PigStorage('?');
cloudera@localhost:~
TeamViewer
Free license (non-commercial use only)
Participants
Dharmender kumar (Presenter)
HARSH VARDHAN
Saurabh Baranawal
SHIVAM
File box
Audio
Microphone
Speakers
Chat
the tuples with the same group
Dharmender kumar (3:09 PM):
the group opertaor groups together thw tuples with same group kry(keyfield). the key field will be a tuple if the group key has more than one field , otherwise it will be the same type as that of gruop key
HARSH VARDHAN (3:13 PM):
store cogr into (location)
Type your message here
SHIVAM
Send
www.teamviewer.com
3:14 PM
7/6/2016

Figure-7.21

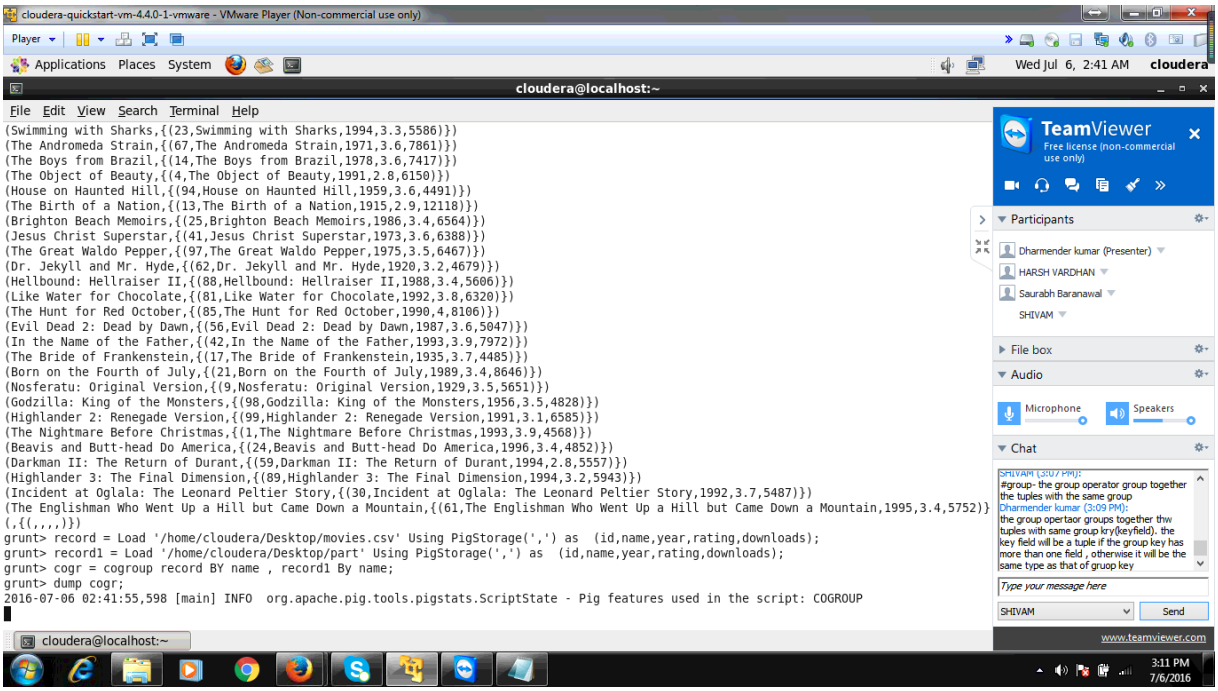

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 2:41 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
(Swimming with Sharks,{(23,Swimming with Sharks,1994,3.3,5586)})
(The Andromeda Strain,{(67,The Andromeda Strain,1971,3.6,7861)})
(The Boys from Brazil,{(14,The Boys from Brazil,1978,3.6,7417)})
(The Object of Beauty,{(4,The Object of Beauty,1991,2.8,6150)})
(House on Haunted Hill,{(94,House on Haunted Hill,1959,3.6,4491)})
(The Birth of a Nation,{(13,The Birth of a Nation,1915,2.9,12118)})
(Brighton Beach Memoirs,{(25,Brighton Beach Memoirs,1986,3.4,6564)})
(Jesus Christ Superstar,{(41,Jesus Christ Superstar,1973,3.6,6388)})
(The Great Waldo Pepper,{(97,The Great Waldo Pepper,1975,3.5,6467)})
(Dr. Jekyll and Mr. Hyde,{(62,Dr. Jekyll and Mr. Hyde,1920,3.2,4679)})
(Hellbound: Hellraiser II,{(88,Hellbound: Hellraiser II,1988,3.4,5606)})
(Like Water for Chocolate,{(81,Like Water for Chocolate,1992,3.8,6320)})
(The Hunt for Red October,{(85,The Hunt for Red October,1990,4,8106)})
(Evil Dead 2: Dead by Dawn,{(56,Evil Dead 2: Dead by Dawn,1987,3.6,5047)})
(In the Name of the Father,{(42,In the Name of the Father,1993,3.9,7972)})
(The Bride of Frankenstein,{(17,The Bride of Frankenstein,1935,3.7,4485)})
(Born on the Fourth of July,{(21,Born on the Fourth of July,1989,3.4,8646)})
(Nosferatu: Original Version,{(9,Nosferatu: Original Version,1929,3.5,5651)})
(Godzilla: King of the Monsters,{(98,Godzilla: King of the Monsters,1956,3.5,4828)})
(Highlander 2: Renegade Version,{(99,Highlander 2: Renegade Version,1991,3.1,6585)})
(The Nightmare Before Christmas,{(1,The Nightmare Before Christmas,1993,3.9,4568)})
(Beavis and Butt-head Do America,{(24,Beavis and Butt-head Do America,1996,3.4,4852)})
(Darkman II: The Return of Durant,{(59,Darkman II: The Return of Durant,1994,2.8,5557)})
(Highlander 3: The Final Dimension,{(89,Highlander 3: The Final Dimension,1994,3.2,5943)})
(Incident at Oglala: The Leonard Peltier Story,{(30,Incident at Oglala: The Leonard Peltier Story,1992,3.7,5487)})
(The Englishman Who Went Up a Hill but Came Down a Mountain,{(61,The Englishman Who Went Up a Hill but Came Down a Mountain,1995,3.4,5752)})
(,{(,,,,)})
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') as (id,name,year,rating,downloads);
grunt> record1 = Load '/home/cloudera/Desktop/part' Using PigStorage(',') as (id,name,year,rating,downloads);
grunt> cogr = cogroup record BY name , record1 By name;
grunt> dump cogr;
2016-07-06 02:41:55,598 [main] INFO org.apache.pig.tools.pigstats.ScriptState - Pig features used in the script: COGROUP
TeamViewer
Free license (non-commercial use only)
Participants
Dharmender kumar (Presenter)
HARSH VARDHAN
Saurabh Baranawal
SHIVAM
File box
Audio
Microphone
Speakers
Chat
#group- the group operator group together the tuples with the same group
Dharmender kumar (3:09 PM):
the group opertaor groups together thw tuples with same group kry(keyfield). the key field will be a tuple if the group key has more than one field , otherwise it will be the same type as that of gruop key
Type your message here
SHIVAM
Send
www.teamviewer.com
cloudera@localhost:~
3:11 PM
7/6/2016

Figure-7.22

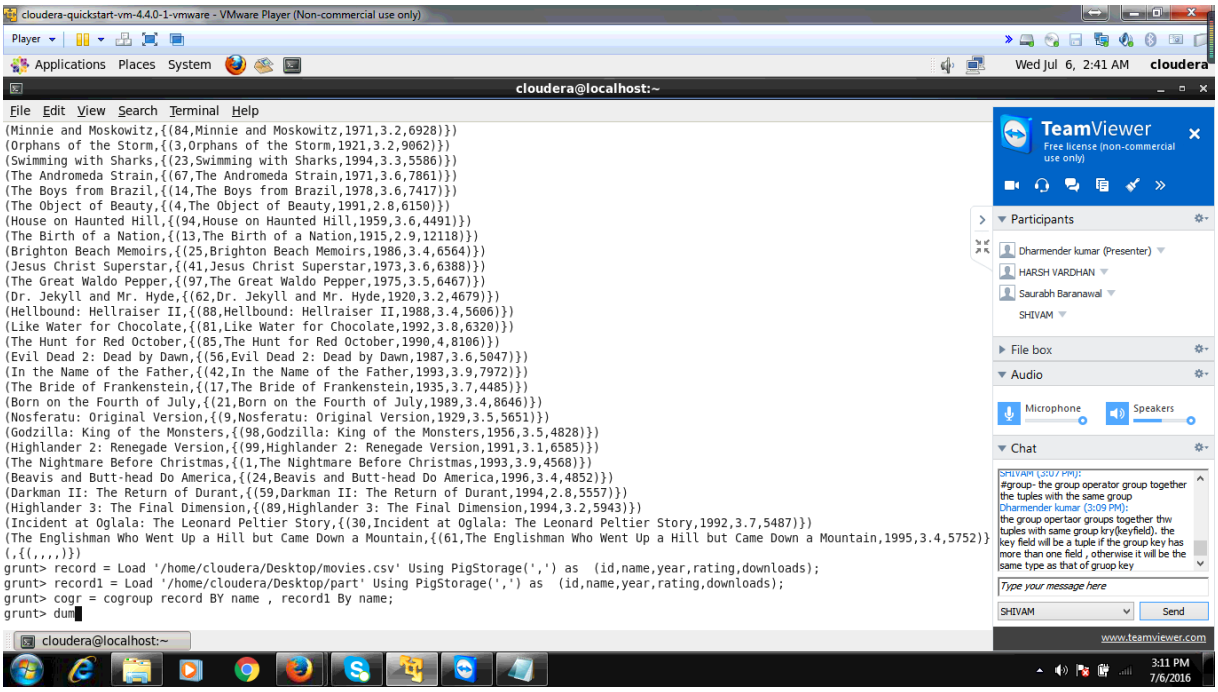

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 2:41 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
(Minnie and Moskowitz,{(84,Minnie and Moskowitz,1971,3.2,6928)})
(Orphans of the Storm,{(3,Orphans of the Storm,1921,3.2,9062)})
(Swimming with Sharks,{(23,Swimming with Sharks,1994,3.3,5586)})
(The Andromeda Strain,{(67,The Andromeda Strain,1971,3.6,7861)})
(The Boys from Brazil,{(14,The Boys from Brazil,1978,3.6,7417)})
(The Object of Beauty,{(4,The Object of Beauty,1991,2.8,6150)})
(House on Haunted Hill,{(94,House on Haunted Hill,1959,3.6,4491)})
(The Birth of a Nation,{(13,The Birth of a Nation,1915,2.9,12118)})
(Brighton Beach Memoirs,{(25,Brighton Beach Memoirs,1986,3.4,6564)})
(Jesus Christ Superstar,{(41,Jesus Christ Superstar,1973,3.6,6388)})
(The Great Waldo Pepper,{(97,The Great Waldo Pepper,1975,3.5,6467)})
(Dr. Jekyll and Mr. Hyde,{(62,Dr. Jekyll and Mr. Hyde,1920,3.2,4679)})
(Hellbound: Hellraiser II,{(88,Hellbound: Hellraiser II,1988,3.4,5606)})
(Like Water for Chocolate,{(81,Like Water for Chocolate,1992,3.8,6320)})
(The Hunt for Red October,{(85,The Hunt for Red October,1990,4,8106)})
(Evil Dead 2: Dead by Dawn,{(56,Evil Dead 2: Dead by Dawn,1987,3.6,5047)})
(In the Name of the Father,{(42,In the Name of the Father,1993,3.9,7972)})
(The Bride of Frankenstein,{(17,The Bride of Frankenstein,1935,3.7,4485)})
(Born on the Fourth of July,{(21,Born on the Fourth of July,1989,3.4,8646)})
(Nosferatu: Original Version,{(9,Nosferatu: Original Version,1929,3.5,5651)})
(Godzilla: King of the Monsters,{(98,Godzilla: King of the Monsters,1956,3.5,4828)})
(Highlander 2: Renegade Version,{(99,Highlander 2: Renegade Version,1991,3.1,6585)})
(The Nightmare Before Christmas,{(1,The Nightmare Before Christmas,1993,3.9,4568)})
(Beavis and Butt-head Do America,{(24,Beavis and Butt-head Do America,1996,3.4,4852)})
(Darkman II: The Return of Durant,{(59,Darkman II: The Return of Durant,1994,2.8,5557)})
(Highlander 3: The Final Dimension,{(89,Highlander 3: The Final Dimension,1994,3.2,5943)})
(Incident at Oglala: The Leonard Peltier Story,{(30,Incident at Oglala: The Leonard Peltier Story,1992,3.7,5487)})
(The Englishman Who Went Up a Hill but Came Down a Mountain,{(61,The Englishman Who Went Up a Hill but Came Down a Mountain,1995,3.4,5752)}
(,{(,,,,)})
grunt> record = Load '/home/cloudera/Desktop/movies.csv' Using PigStorage(',') as (id,name,year,rating,downloads);
grunt> record1 = Load '/home/cloudera/Desktop/part' Using PigStorage(',') as (id,name,year,rating,downloads);
grunt> cogr = cogroup record BY name , record1 By name;
grunt> dum
cloudera@localhost:~
TeamViewer
Free license (non-commercial use only)
Participants
Dharmender kumar (Presenter)
HARSH VARDHAN
Saurabh Baranawal
SHIVAM
File box
Audio
Microphone
Speakers
Chat
#group- the group operator group together the tuples with the same group
Dharmender kumar (3:09 PM):
the group opertaor groups together thw tuples with same group kry(keyfield). the key field will be a tuple if the group key has more than one field , otherwise it will be the same type as that of gruop key
Type your message here
SHIVAM
Send
www.teamviewer.com
3:11 PM
7/6/2016

Figure-7.23

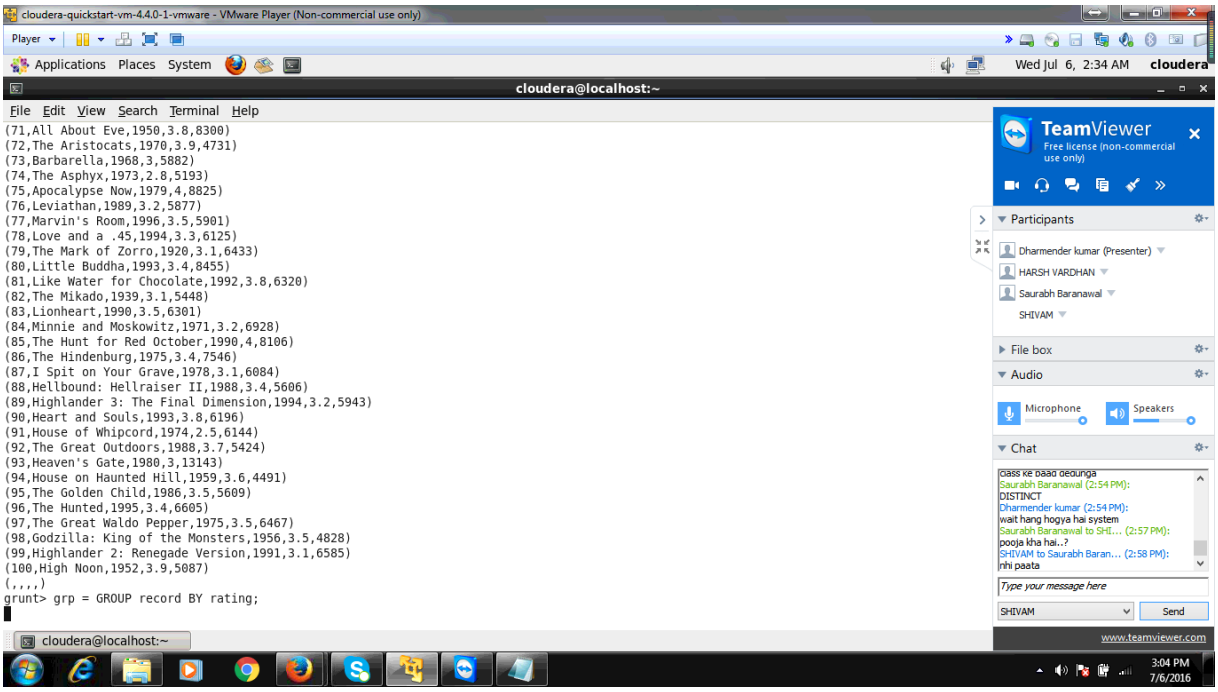

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 2:34 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
(71,All About Eve,1950,3.8,8300)
(72,The Aristocats,1970,3.9,4731)
(73,Barbarella,1968,3,5882)
(74,The Asphyx,1973,2.8,5193)
(75,Apocalypse Now,1979,4,8825)
(76,Leviathan,1989,3.2,5877)
(77,Marvin's Room,1996,3.5,5901)
(78,Love and a .45,1994,3.3,6125)
(79,The Mark of Zorro,1920,3.1,6433)
(80,Little Buddha,1993,3.4,8455)
(81,Like Water for Chocolate,1992,3.8,6320)
(82,The Mikado,1939,3.1,5448)
(83,Lionheart,1990,3.5,6301)
(84,Minnie and Moskowitz,1971,3.2,6928)
(85,The Hunt for Red October,1990,4,8106)
(86,The Hindenburg,1975,3.4,7546)
(87,I Spit on Your Grave,1978,3.1,6084)
(88,Hellbound: Hellraiser II,1988,3.4,5606)
(89,Highlander 3: The Final Dimension,1994,3.2,5943)
(90,Heart and Souls,1993,3.8,6196)
(91,House of Whipcord,1974,2.5,6144)
(92,The Great Outdoors,1988,3.7,5424)
(93,Heaven's Gate,1980,3,13143)
(94,House on Haunted Hill,1959,3.6,4491)
(95,The Golden Child,1986,3.5,5609)
(96,The Hunted,1995,3.4,6605)
(97,The Great Waldo Pepper,1975,3.5,6467)
(98,Godzilla: King of the Monsters,1956,3.5,4828)
(99,Highlander 2: Renegade Version,1991,3.1,6585)
(100,High Noon,1952,3.9,5087)
(,,,,)
grunt> grp = GROUP record BY rating;
TeamViewer
Free license (non-commercial use only)
Participants
Dharmender kumar (Presenter)
HARSH VARDHAN
Saurabh Baranawal
SHIVAM
File box
Audio
Microphone
Speakers
Chat
Saurabh Baranawal (2:54 PM):
DISTINCT
Dharmender kumar (2:54 PM):
wait hang hogya hai system
Saurabh Baranawal to SHI... (2:57 PM):
pooja kha hai..?
SHIVAM to Saurabh Baran... (2:58 PM):
nhi paata
Type your message here
SHIVAM
Send
www.teamviewer.com
cloudera@localhost:~
3:04 PM
7/6/2016

Figure-7.24

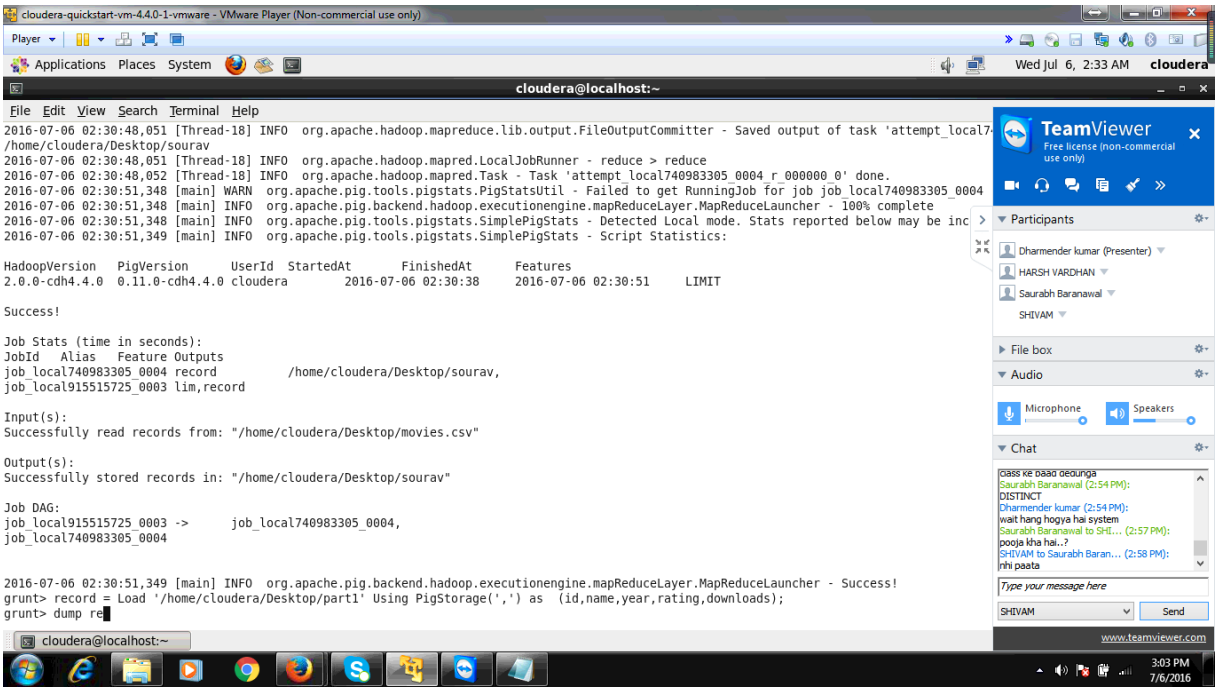

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 2:33 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
2016-07-06 02:30:48,051 [Thread-18] INFO org.apache.hadoop.mapreduce.lib.output.FileOutputCommitter - Saved output of task 'attempt_local7
/home/cloudera/Desktop/sourav
2016-07-06 02:30:48,051 [Thread-18] INFO org.apache.hadoop.mapred.LocalJobRunner - reduce > reduce
2016-07-06 02:30:48,052 [Thread-18] INFO org.apache.hadoop.mapred.Task - Task 'attempt_local740983305_0004_r_000000_0' done.
2016-07-06 02:30:51,348 [main] WARN org.apache.pig.tools.pigstats.PigStatsUtil - Failed to get RunningJob for job job_local740983305_0004
2016-07-06 02:30:51,348 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - 100% complete
2016-07-06 02:30:51,348 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Detected Local mode. Stats reported below may be inc
2016-07-06 02:30:51,349 [main] INFO org.apache.pig.tools.pigstats.SimplePigStats - Script Statistics:
HadoopVersion PigVersion UserId StartedAt FinishedAt Features
2.0.0-cdh4.4.0 0.11.0-cdh4.4.0 cloudera 2016-07-06 02:30:38 2016-07-06 02:30:51 LIMIT
Success!
Job Stats (time in seconds):
JobId Alias Feature Outputs
job_local740983305_0004 record /home/cloudera/Desktop/sourav,
job_local915515725_0003 lim,record
Input(s):
Successfully read records from: "/home/cloudera/Desktop/movies.csv"
Output(s):
Successfully stored records in: "/home/cloudera/Desktop/sourav"
Job DAG:
job_local915515725_0003 -> job_local740983305_0004,
job_local740983305_0004
2016-07-06 02:30:51,349 [main] INFO org.apache.pig.backend.hadoop.executionengine.mapReduceLayer.MapReduceLauncher - Success!
grunt> record = Load '/home/cloudera/Desktop/part1' Using PigStorage(',') as (id,name,year,rating,downloads);
grunt> dump re
TeamViewer
Free license (non-commercial use only)
Participants
Dharmender kumar (Presenter)
HARSH VARDHAN
Saurabh Baranawal
SHIVAM
File box
Audio
Microphone
Speakers
Chat
Saurabh Baranawal (2:54 PM):
DISTINCT
Dharmender kumar (2:54 PM):
wait hang hogya hai system
Saurabh Baranawal to SHI... (2:57 PM):
pooja kha hai..?
SHIVAM to Saurabh Baran... (2:58 PM):
nhi paata
Type your message here
SHIVAM
Send
www.teamviewer.com
cloudera@localhost:~
3:03 PM
7/6/2016

Figure-7.25

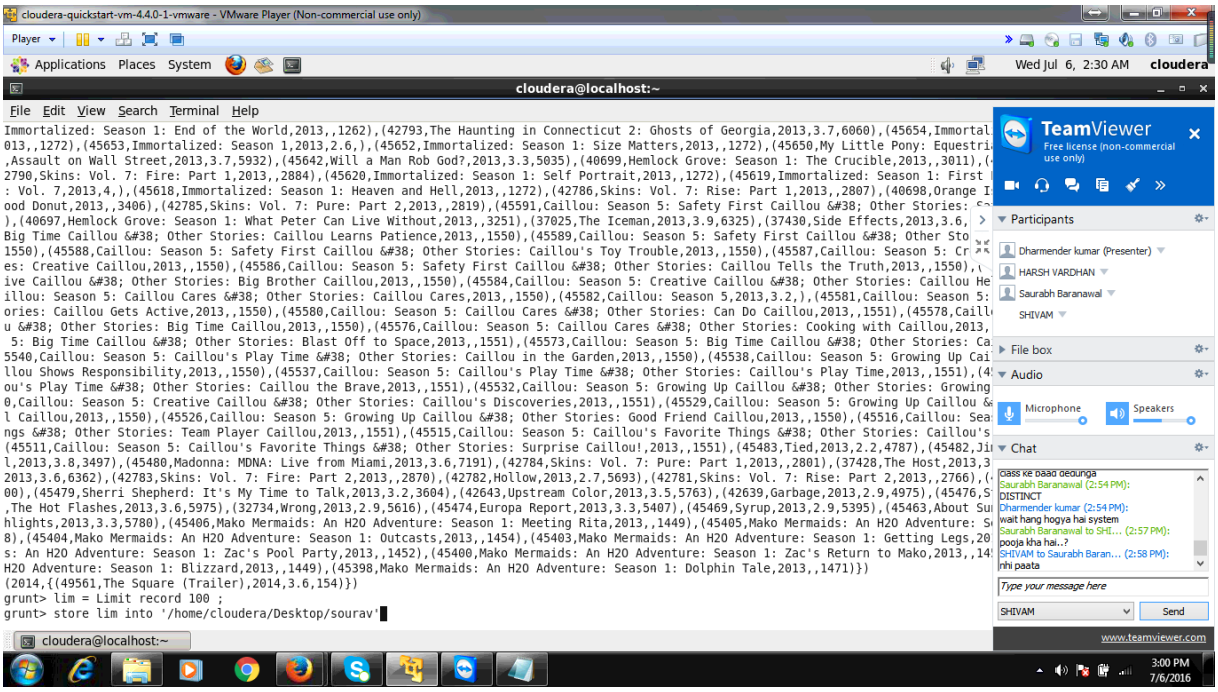

cloudera-quickstart-vm-4.4.0-1-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Wed Jul 6, 2:30 AM cloudera
cloudera@localhost:~
File Edit View Search Terminal Help
Immortalized: Season 1: End of the World,2013,,1262),(42793,The Haunting in Connecticut 2: Ghosts of Georgia,2013,3.7,6060),(45654,Immortal
013,,1272),(45653,Immortalized: Season 1,2013,2.6,),(45652,Immortalized: Season 1: Size Matters,2013,,1272),(45650,My Little Pony: Equestri
,Assault on Wall Street,2013,3.7,5932),(45642,Will a Man Rob God?,2013,3.3,5035),(40699,Hemlock Grove: Season 1: The Crucible,2013,,3011),(
2790,Skins: Vol. 7: Fire: Part 1,2013,,2884),(45620,Immortalized: Season 1: Self Portrait,2013,,1272),(45619,Immortalized: Season 1: First
: Vol. 7,2013,4,),(45618,Immortalized: Season 1: Heaven and Hell,2013,,1272),(42786,Skins: Vol. 7: Rise: Part 1,2013,,2807),(40698,Orange I
ood Donut,2013,,3406),(42785,Skins: Vol. 7: Pure: Part 2,2013,,2819),(45591,Caillou: Season 5: Safety First Caillou & Other Stories:
),(40697,Hemlock Grove: Season 1: What Peter Can Live Without,2013,,3251),(37025,The Iceman,2013,3.9,6325),(37430,Side Effects,2013,3.6,
Big Time Caillou & Other Stories: Caillou Learns Patience,2013,,1550),(45589,Caillou: Season 5: Safety First Caillou & Other Sto
1550),(45588,Caillou: Season 5: Safety First Caillou & Other Stories: Caillou's Toy Trouble,2013,,1550),(45587,Caillou: Season 5: Cr
es: Creative Caillou,2013,,1550),(45586,Caillou: Season 5: Safety First Caillou & Other Stories: Caillou Tells the Truth,2013,,1550),(
ive Caillou & Other Stories: Big Brother Caillou,2013,,1550),(45584,Caillou: Season 5: Creative Caillou & Other Stories: Caillou He
illou: Season 5: Caillou Cares & Other Stories: Caillou Cares,2013,,1550),(45582,Caillou: Season 5,2013,3.2,),(45581,Caillou: Season 5:
ories: Caillou Gets Active,2013,,1550),(45580,Caillou: Season 5: Caillou Cares & Other Stories: Can Do Caillou,2013,,1551),(45578,Caill
u & Other Stories: Big Time Caillou,2013,,1550),(45576,Caillou: Season 5: Caillou Cares & Other Stories: Cooking with Caillou,2013,
5: Big Time Caillou & Other Stories: Blast Off to Space,2013,,1551),(45573,Caillou: Season 5: Big Time Caillou & Other Stories: Ca
5540,Caillou: Season 5: Caillou's Play Time & Other Stories: Caillou in the Garden,2013,,1550),(45538,Caillou: Season 5: Growing Up Cai
llou Shows Responsibility,2013,,1550),(45537,Caillou: Season 5: Caillou's Play Time & Other Stories: Caillou's Play Time,2013,,1551),(4
ou's Play Time & Other Stories: Caillou the Brave,2013,,1551),(45532,Caillou: Season 5: Growing Up Caillou & Other Stories: Growing
0,Caillou: Season 5: Creative Caillou & Other Stories: Caillou's Discoveries,2013,,1551),(45529,Caillou: Season 5: Growing Up Caillou &
l Caillou,2013,,1550),(45526,Caillou: Season 5: Growing Up Caillou & Other Stories: Good Friend Caillou,2013,,1550),(45516,Caillou: Sea
ngs & Other Stories: Team Player Caillou,2013,,1551),(45515,Caillou: Season 5: Caillou's Favorite Things & Other Stories: Caillou's
(45511,Caillou: Season 5: Caillou's Favorite Things & Other Stories: Surprise Caillou!,2013,,1551),(45483,Tied,2013,2.2,4787),(45482,Ji
l,2013,3.8,3497),(45480,Madonna: MDNA: Live from Miami,2013,3.6,7191),(42784,Skins: Vol. 7: Pure: Part 1,2013,,2801),(37428,The Host,2013,3
2013,3.6,6362),(42783,Skins: Vol. 7: Fire: Part 2,2013,,2870),(42782,Hollow,2013,2.7,5693),(42781,Skins: Vol. 7: Rise: Part 2,2013,,2766),(
00),(45479,Sherri Shepherd: It's My Time to Talk,2013,3.2,3604),(42643,Upstream Color,2013,3.5,5763),(42639,Garbage,2013,2.9,4975),(45476,S
,The Hot Flashes,2013,3.6,5975),(32734,Wrong,2013,2.9,5616),(45474,Europa Report,2013,3.3,5407),(45469,Syrup,2013,2.9,5395),(45463,About Su
hlights,2013,3.3,5780),(45406,Mako Mermaids: An H2O Adventure: Season 1: Meeting Rita,2013,,1449),(45405,Mako Mermaids: An H2O Adventure: S
8),(45404,Mako Mermaids: An H2O Adventure: Season 1: Outcasts,2013,,1454),(45403,Mako Mermaids: An H2O Adventure: Season 1: Getting Legs,20
s: An H2O Adventure: Season 1: Zac's Pool Party,2013,,1452),(45400,Mako Mermaids: An H2O Adventure: Season 1: Zac's Return to Mako,2013,,14
H2O Adventure: Season 1: Blizzard,2013,,1449),(45398,Mako Mermaids: An H2O Adventure: Season 1: Dolphin Tale,2013,,1471)})
(2014,{(49561,The Square (Trailer),2014,3.6,154)})
grunt> lim = Limit record 100 ;
grunt> store lim into '/home/cloudera/Desktop/sourav'
cloudera@localhost:~
TeamViewer
Free license (non-commercial use only)
Participants
Dharmender kumar (Presenter)
HARSH VARDHAN
Saurabh Baranawal
SHIVAM
File box
Audio
Microphone
Speakers
Chat
class ke baad dedunga
Saurabh Baranawal (2:54 PM):
DISTINCT
Dharmender kumar (2:54 PM):
wait hang hogya hai system
Saurabh Baranawal to SHI... (2:57 PM):
pooja kha hai..?
SHIVAM to Saurabh Baran... (2:58 PM):
nhi paata
Type your message here
SHIVAM
Send
www.teamviewer.com
3:00 PM
7/6/2016

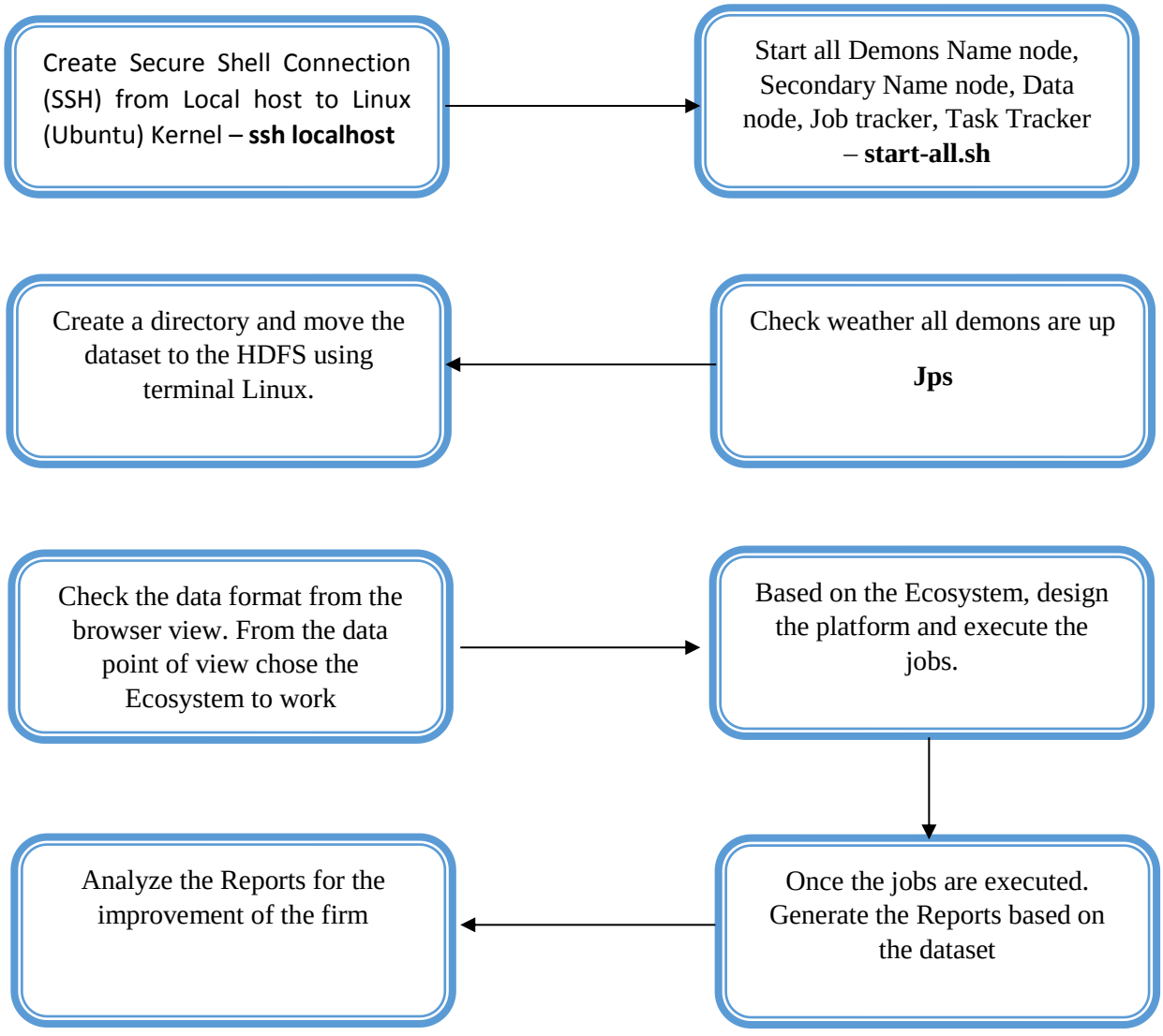
Figure 7.26

Chapter 8

Dependency Chart

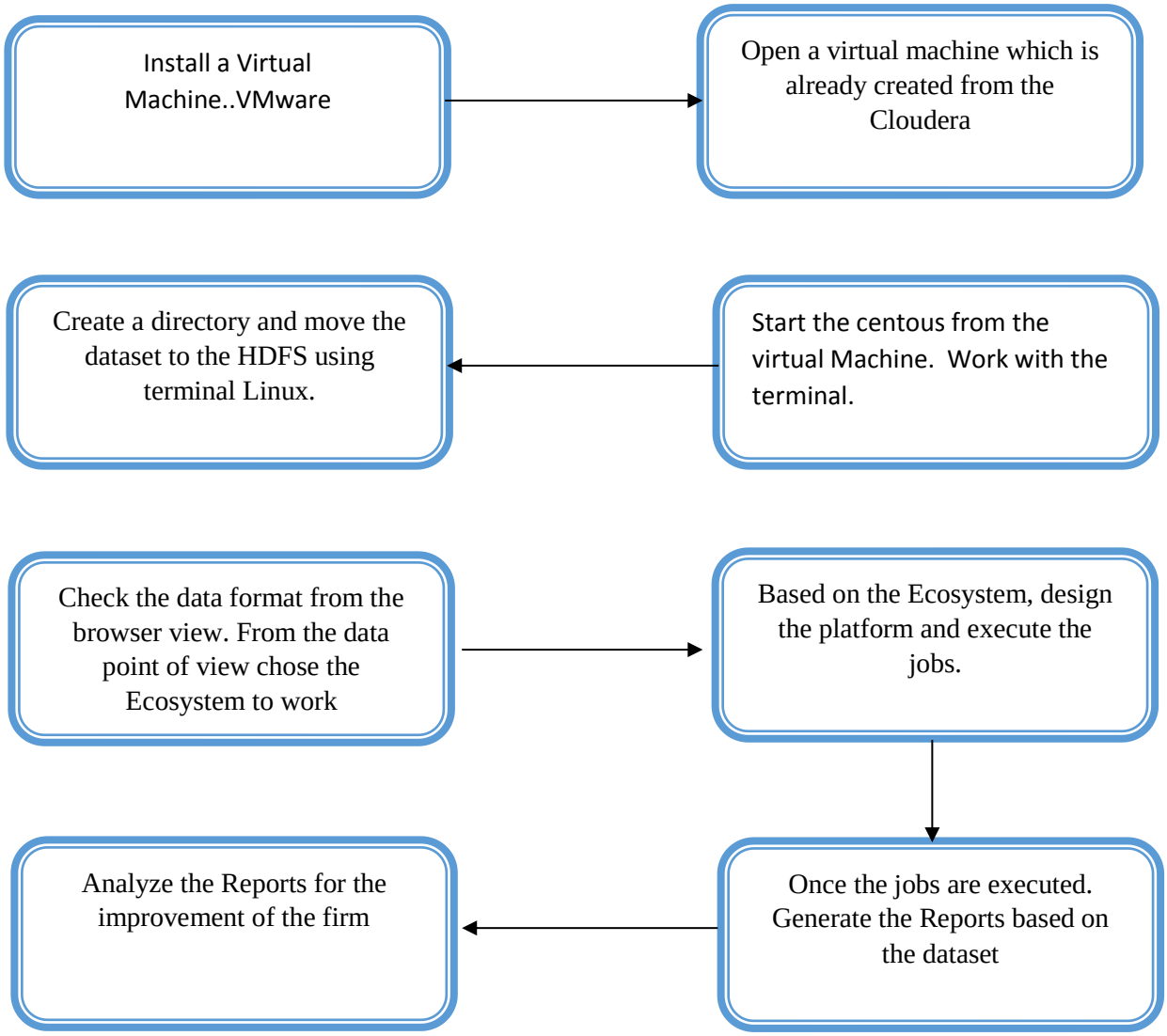
8.1. Business Flow:

8.1.1. Apache Hadoop Working Model-I:



Apache Hadoop working Model-I

8.2.2 Apache Hadoop Working Model-II:



Apache Hadoop Working Model-II

Chapter 9

Result and Discussion

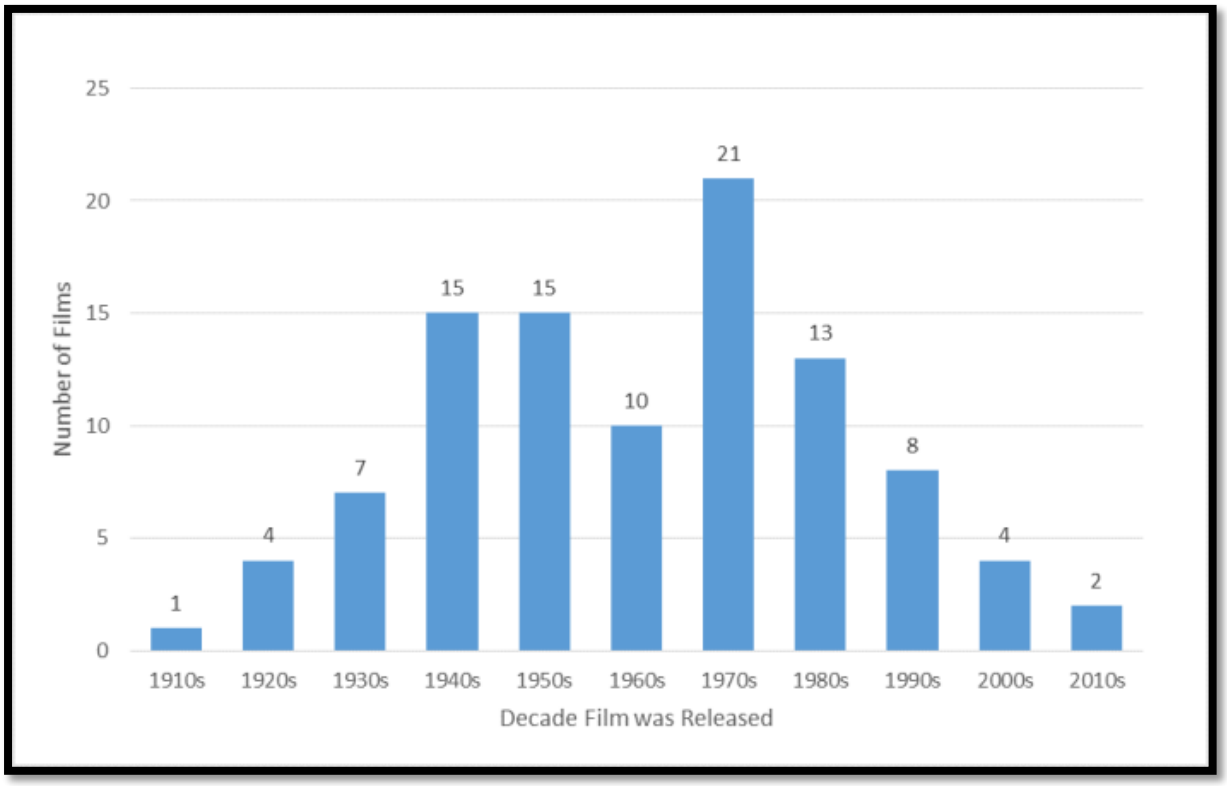


Figure 9.1.

The above figure depicts the variation in trends of movie releases in each decades. The plot is a Bar Graph plot that shows the number of movies that are released in the given decade.

From the plot it is clear that the trends in the movie releases is as follows

- Increases from 1910 to 1940s.
- Remains constant from 1940 to 1950s.
- Decreases from 1950 to 1960s
- Increases from 1960 to 1970s.
- And finally decreases from 1970 to 2010s.

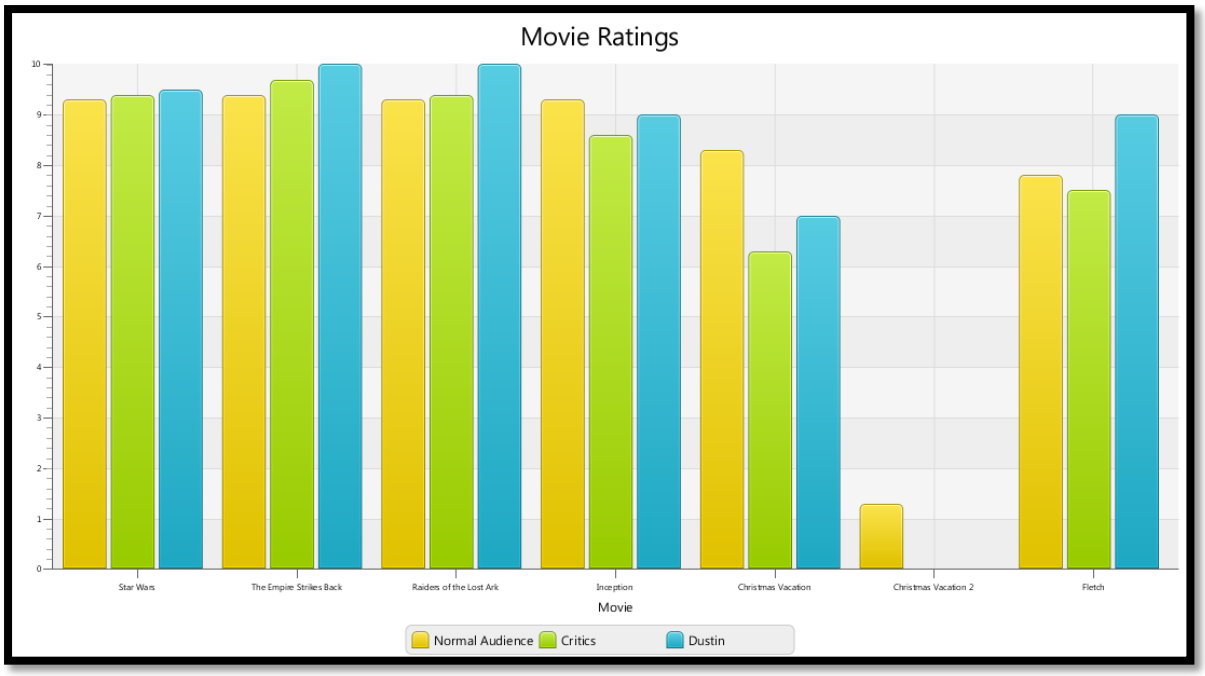


Figure 9.2.

The above bar graph gives a plot of ratings given by different kinds of audience for each movie.

The trends can be summarized as follows:

- "Stars wars" was mostly liked by Dustins.
- "Empire strikes back" was mostly liked by Dustins.
- "Raides of Lost Ark" was mostly liked by Dustins.
- "Inception" was most liked by Normal Audience.
- "Christmas Vacation" was most like by the normal audience.
- "Christmas Vacation 2" was most liked by normal people and was mostly disliked by other people.
- "Fletch" was again liked by the Dustins.

For the analysis of the data, the software "TIBCO SPOTFIRE" is used.

Chapter 10

Conclusion

Hadoop is trending technology in the market. Hadoop solves the big data problem more effectively and efficiently. More importantly Hadoop can analyze any kind of data. Analyzing the data based on Hadoop requires very less amount of time, and it reduces the production time which directly affects the economy of the organization.

Analyzing the dataset based on the Apache Hadoop will overcome all the issues caused by the traditional RDBMS and Master slave Architecture of Servers.

In our project we are trying to analyze Movie release dataset using Hadoop.

This analysis makes to analyze total number of movies, their ratings and their dates of release.

References

- **Hadoop:**
  https://hadoop.apache.org/
- **Java:**
  http://www.oracle.com/technetwork/java/javase/downloads/
- **Hive:**
  https://hive.apache.org/
- **Linux:**
  http://www.ubuntu.com/
  http://www.centos.org/