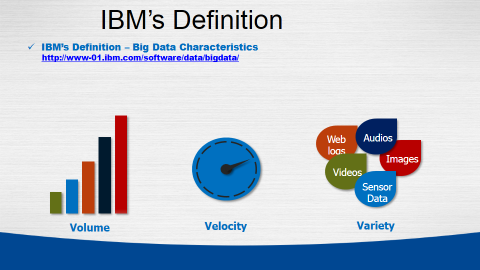
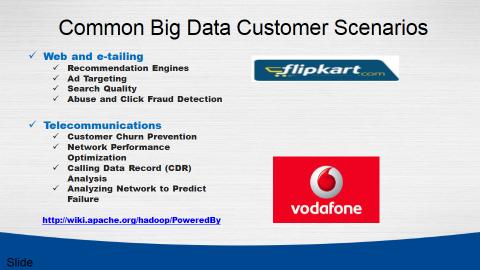
**PROJECT REPORT ON BUSINESS ANYLITCIS**

**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.**



* **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.



Big Data challenges

**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.

**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.

**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.

**Types of Data**

**There are Two types of Data**

1. **Unstructured data**

## Examples of 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.

# **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.

## Characteristics of Structured Data

Structured data first depends on creating a [data model](http://www.webopedia.com/TERM/D/data_modeling.html) – 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) .

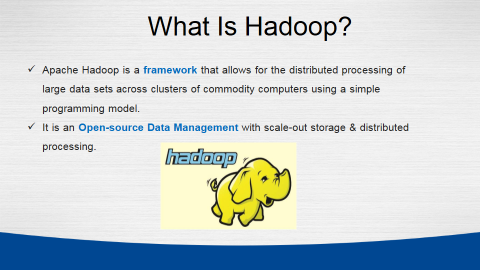
## Unstructured and Semi-Structured Data

[Unstructured data](http://www.webopedia.com/TERM/U/unstructured_data.html) 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.

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 .

### Big data tools

Software like [Hadoop](http://www.webopedia.com/TERM/H/hadoop.html) can process stores of both unstructured and structured data that are extremely large, very complex and changing rapidly.



**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.

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.

**When To Use Hadoop**

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.

### ****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.

### ****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.

I will not be showing the integration in this blog but will show them in the Hadoop Integration series. I am already excited about it and I hope you feel the same.

### ****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.

**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

## **When Not To Use Hadoop**

### ****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.

**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).

**2) 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.

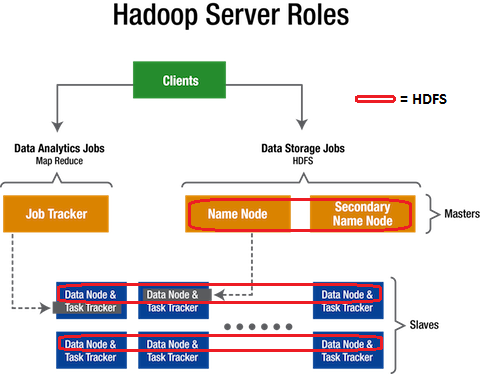
### ****3. Multiple Smaller Datasets****

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.

## **Architecture of Hadoop**

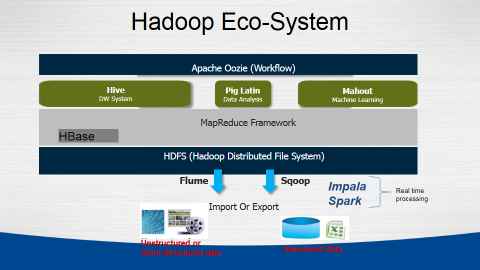
Below is a high-level architecture of multi-node Hadoop Cluster.

Here are few highlights of the Hadoop Architecture:



* Hadoop works in a master-worker / master-slave fashion.
* Hadoop has two core components: HDFS and MapReduce.
* **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.
* **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.
* The master contains the Namenode and Job Tracker components.
  + **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.
  + **Job Tracker** keeps track of the individual tasks/jobs assigned to each of the nodes and coordinates the exchange of information and results.
* Each Worker / Slave contains the Task Tracker and a Datanode components.
  + **Task Tracker** is responsible for running the task / computation assigned to it.
  + **Datanode** is responsible for holding the data.
* The computers present in the cluster can be present in any location and there is no dependency on the location of the physical server.

**HADOOP ECO SYSTEM**





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.

**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.



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

* recommendations, a.k.a. collective filtering
* classification, a.k.a categorization
* clustering
* frequent itemset mining, a.k.a parallel frequent pattern mining

## MapReduce

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

**HBAESE**

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.

* 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.**

## [**Sqoop**](http://sqoop.apache.org/)

**https://lh4.googleusercontent.com/KGJ467QzXrjbFNfulu_UDDkogngppSZvtKQTAqbt5rCNRD3dxcD2ktNApqjsH8H5Kc_JIySNEUKEy1V9eyEvGMuU7n9cuEVr13pxmCJIEJl6jJNQEl7R**

**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."**

**Steps To Installation Hadoop**

Windows

**Install Virtual machine**

**Windows OS**

**Run these commands on Ubuntu.**

**Upload Image CHD3 of Ubuntu**

**Upload Image CHD4 of Centos**

**Upload Image of Ubuntu**

|  |
| --- |
| **RUN THESE COMMANDS ON UBUNTU**  **Follow These Steps To install Hadoop On Ubuntu**  Step(1)**Update the repository:**  #Command: sudo apt-get **update**  Step (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 :  Step (3)  #Command:java -version  Step (4)Install openssh-server:  #Command: sudo apt-get install openssh-server  Step (5)Download and extract Hadoop:  #Command: wget http://archive.apache.org/dist/hadoop/core/hadoop-1.2.0/hadoop- 1.2.0.tar.gz  Command: tar -xvf hadoop-1.2.0.tar.gz  Step(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> |

|  |
| --- |
| Step(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>  Step(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:  Step(9)Check IP Address  #Command: ifconfig  Step(10)  #Command: sudo gedit /etc/hosts  Create a ssh key:  Step(11)  #Command: ssh-keygen -t rsa –P "" |

|  |
| --- |
| **Reboot The System** |

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

[**Sqoop**](http://sqoop.apache.org/)

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.

.



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.

**How Sqoop Works?**

The following image describes the workflow of Sqoop.



**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.

**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

|  |
| --- |
| **Step To Install and Create Table on Operating system** |

**Create Database**

**Install MySQL**

**Centos or Ubuntu**

**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 |

**Sqoop Import**

Sqoop is designed to import tables from a database into HDFS. The connect string is similar to a URL, and is communicated to Sqoop with the --connect argument. This describes the server and database to connect to; it may also specify the port.

**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.

**# 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.**

**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.

**# Write this way your command to Export t 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 “

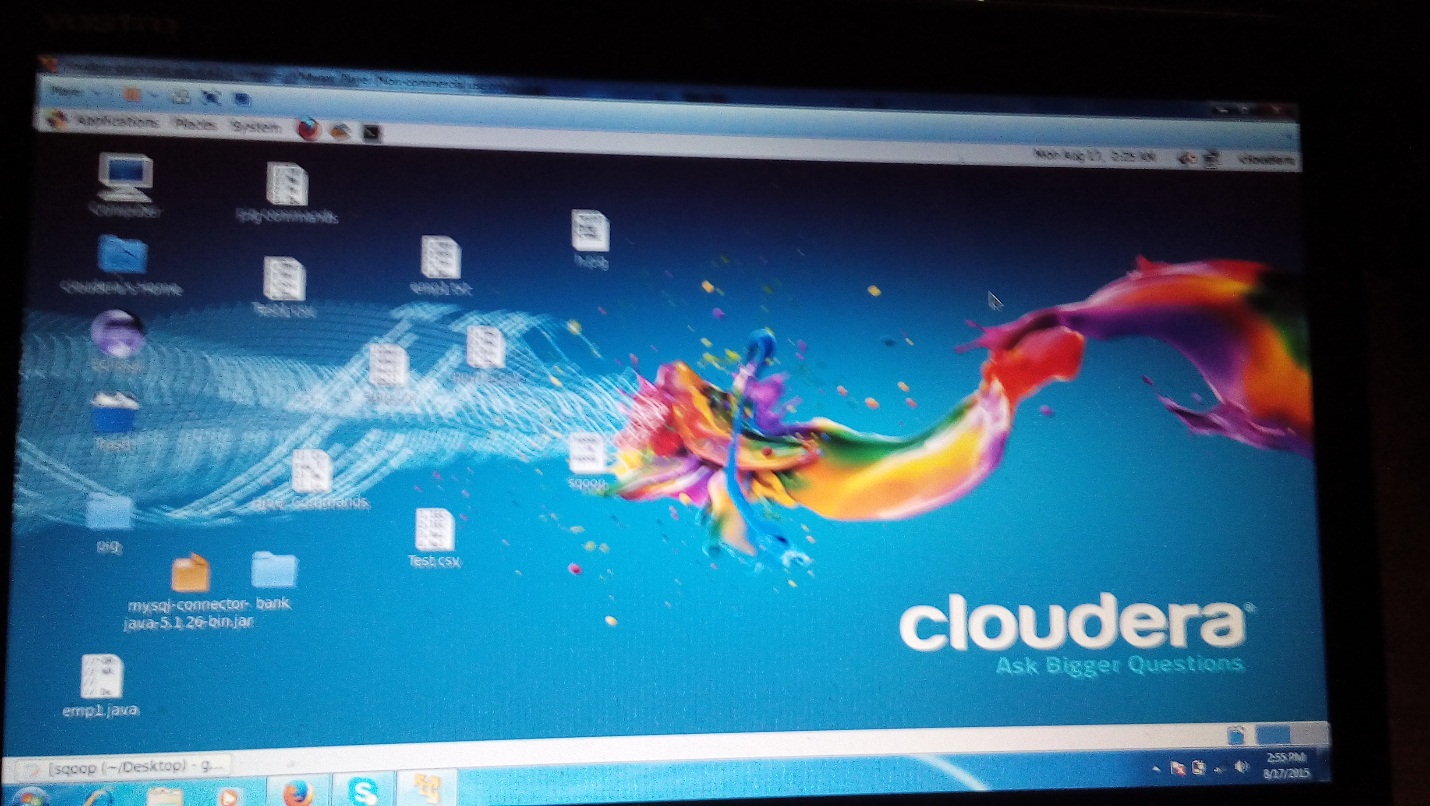
Or

$ sqoop export –-connect jdbc:mysq://localhost/Portnumber/databasename

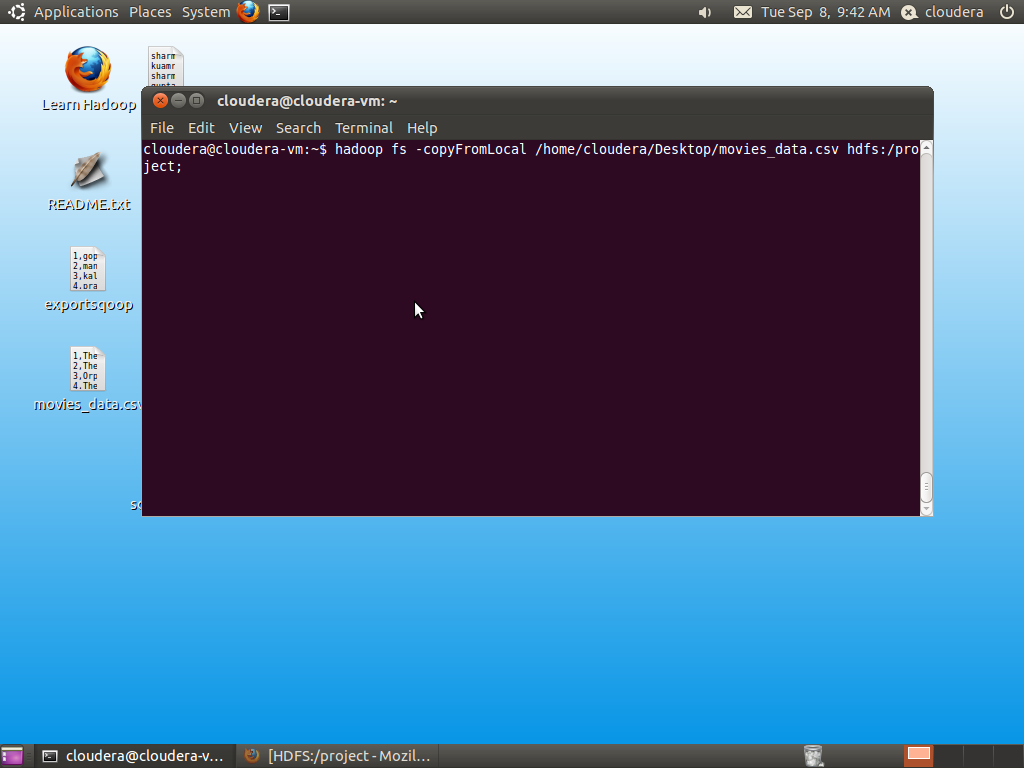
--username root -Password Password --table name

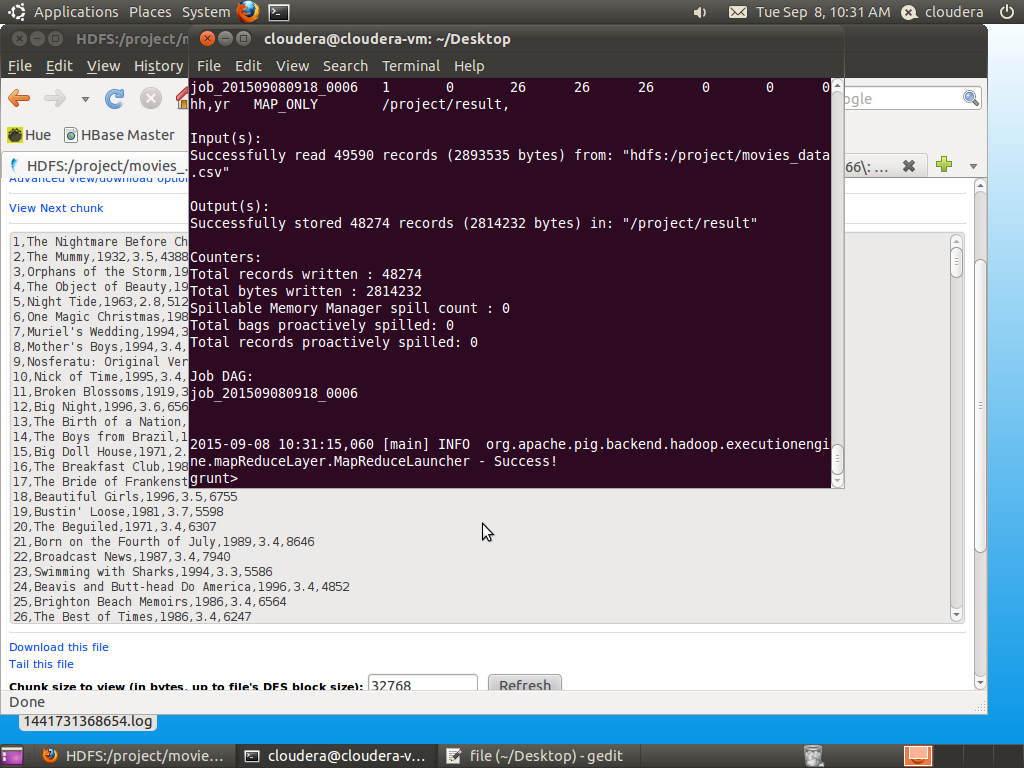
--export-dir “Path of table”

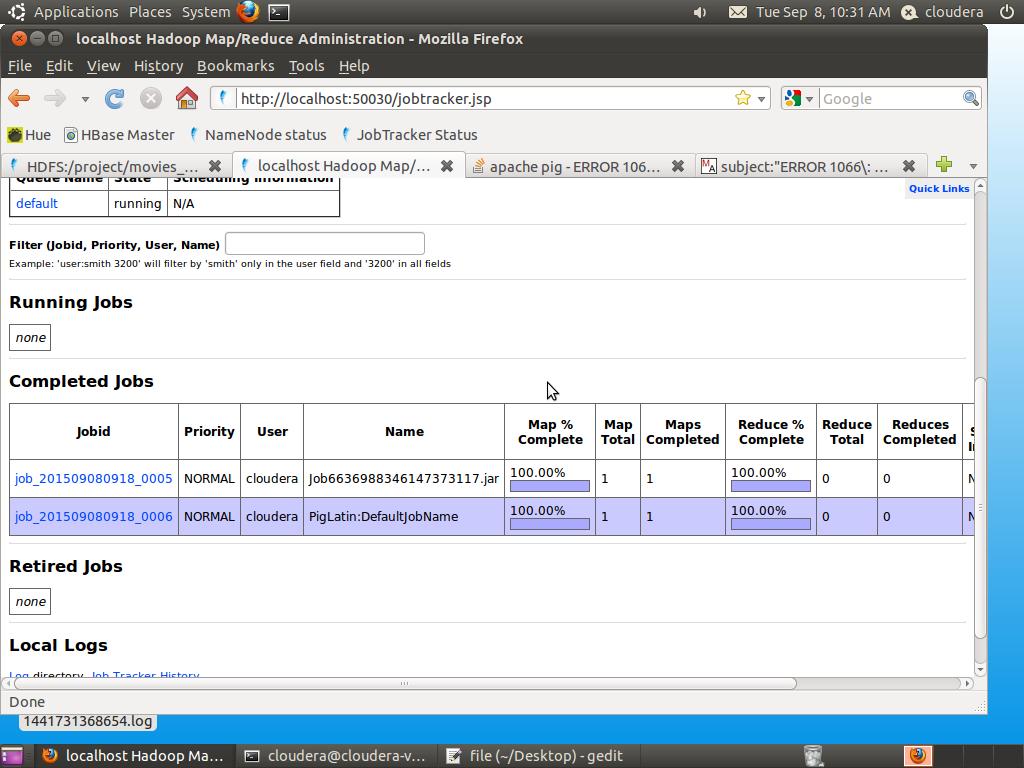
Open CHD4 Centos For Analyzing Big Data

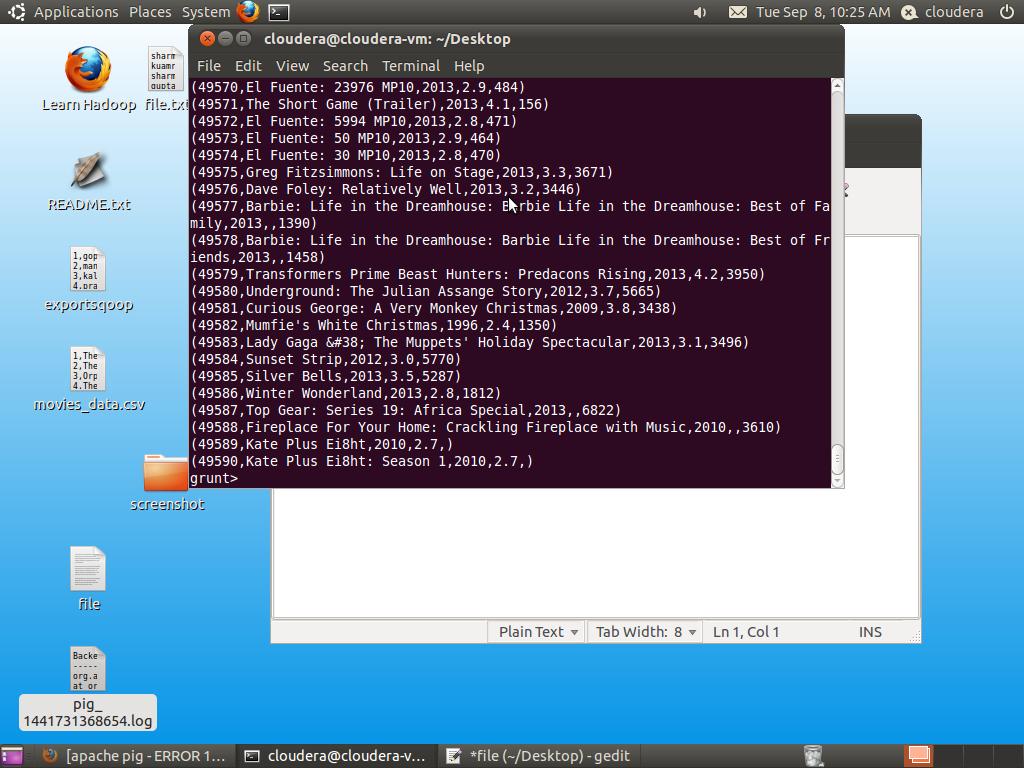


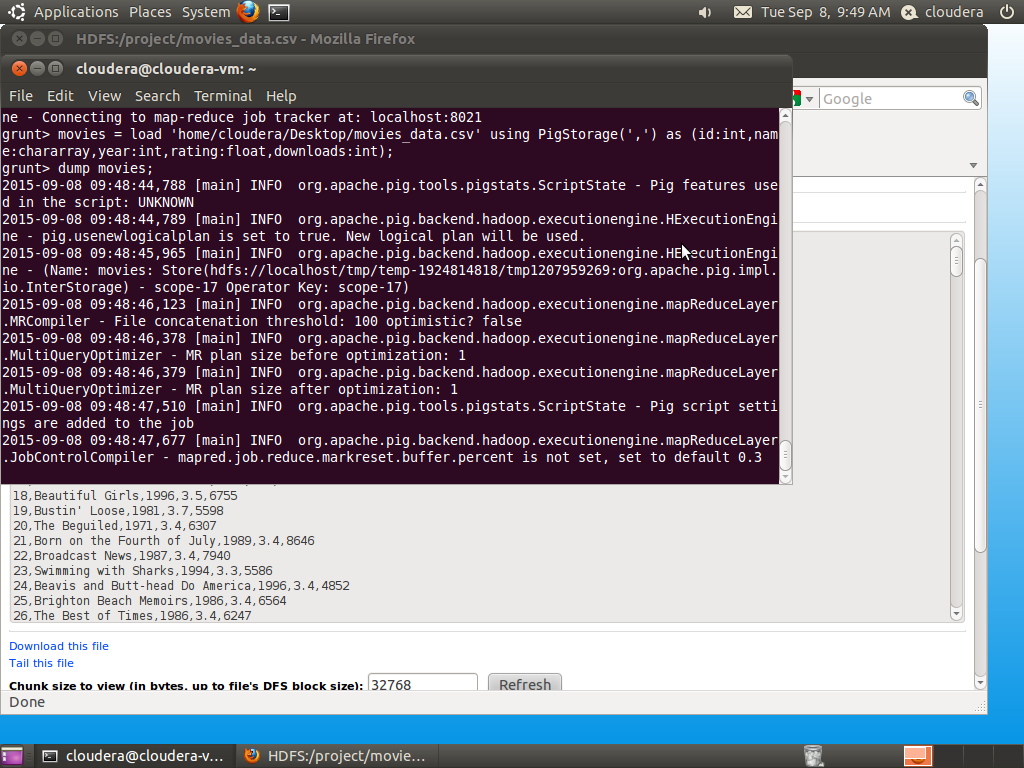
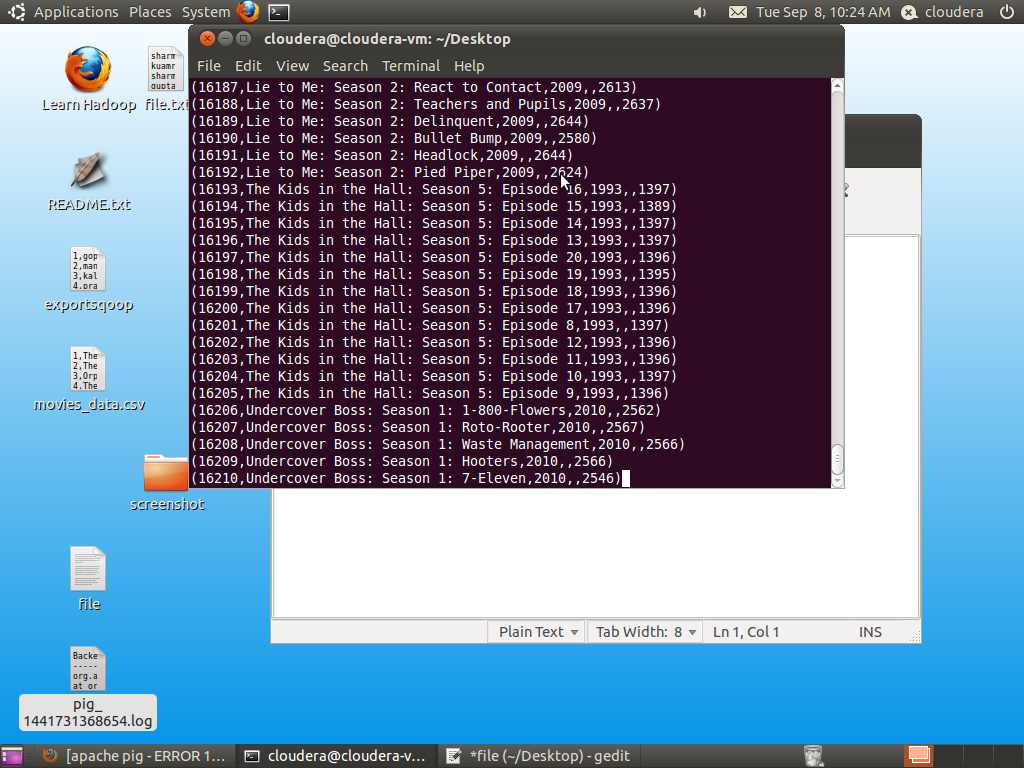
Copy the Input file into HDFS



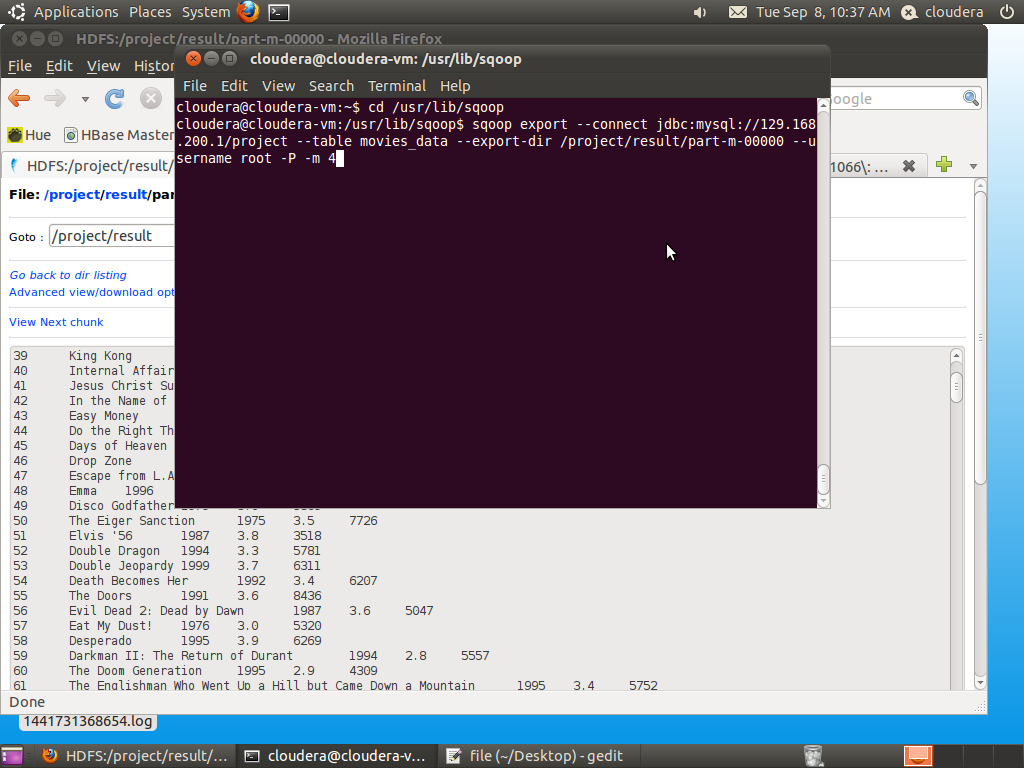
Copying Data into HDFS 

Resulted Data into The HDFS

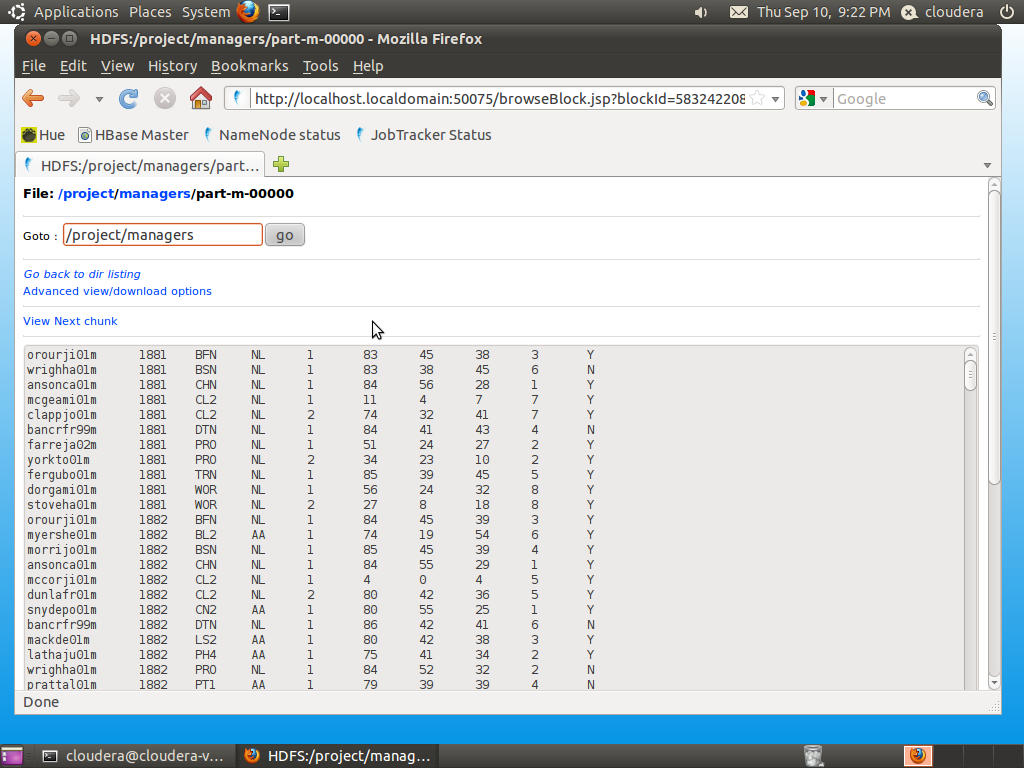
Create Schema In the Pig Latin and Filter through year 

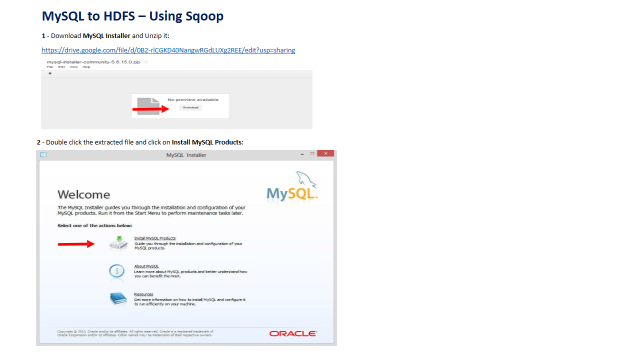
Show the Total Record of HDFS into Pig Schema

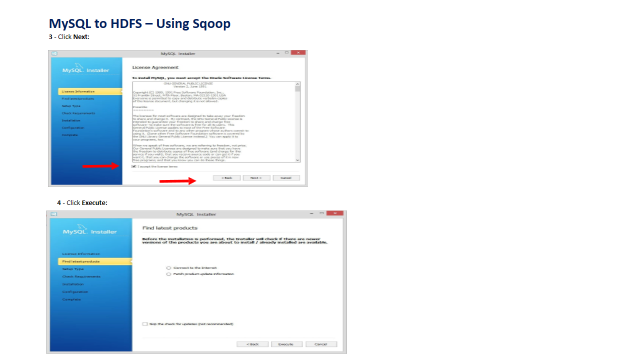
Filter the Big data of Baseball Match

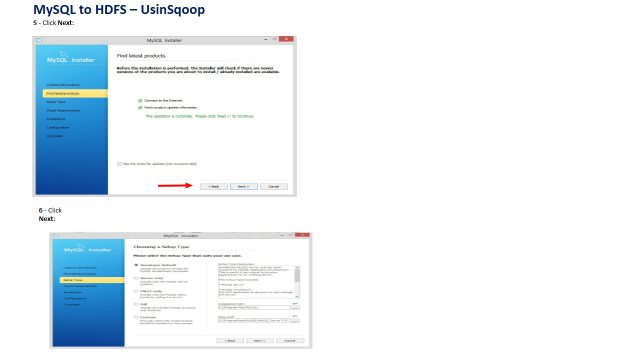


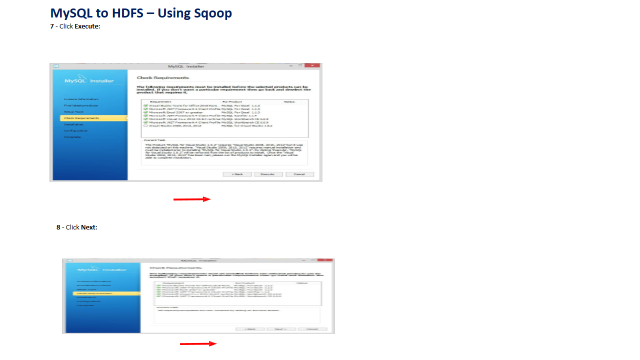
Output of Pig Stored into the HDFS Using Store Keyword

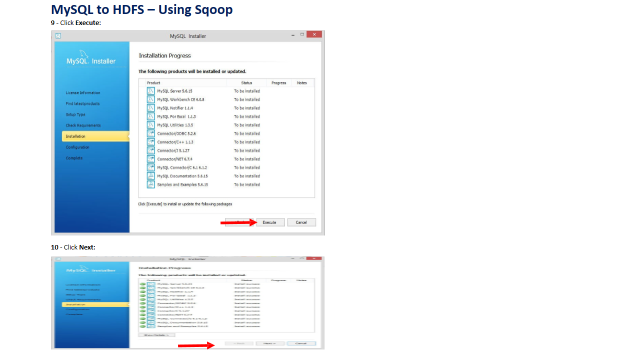


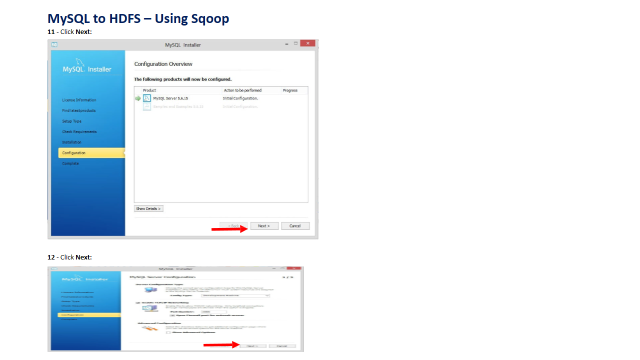
Install MySQL on Windows 

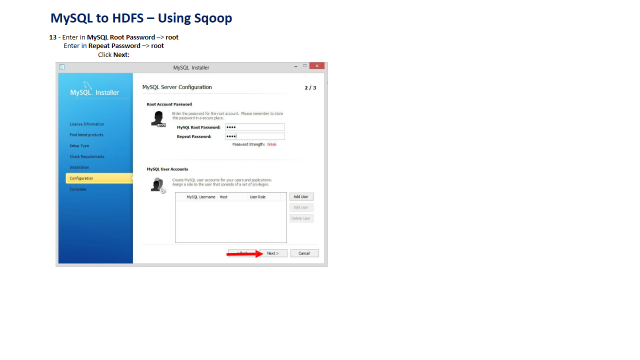


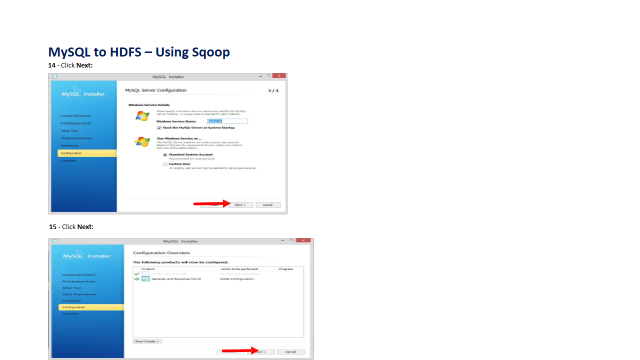
****

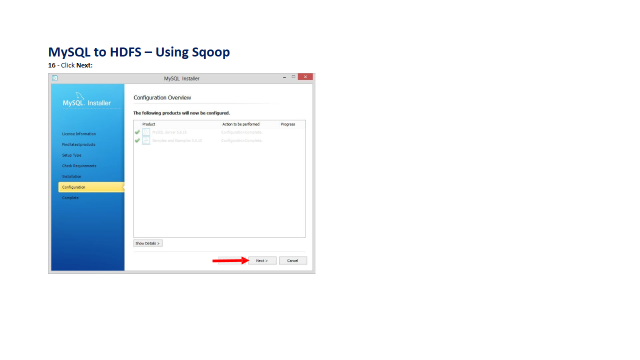
****

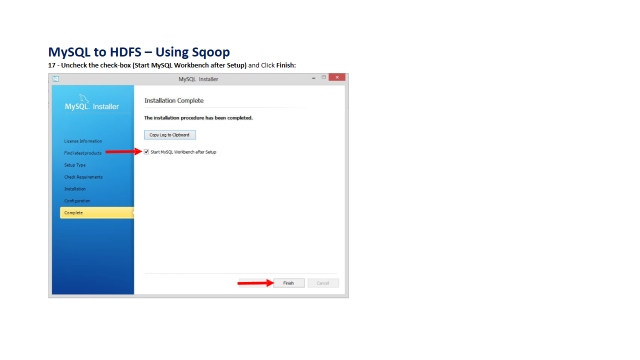
****

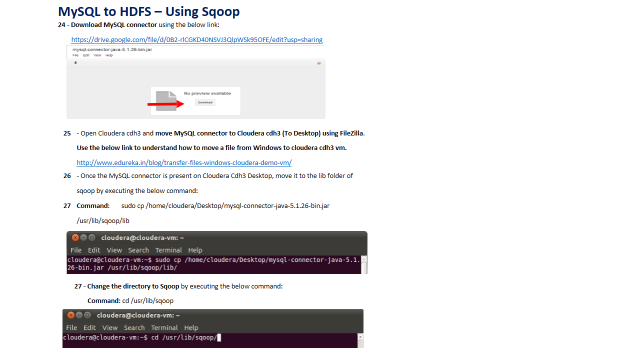
****

****

****

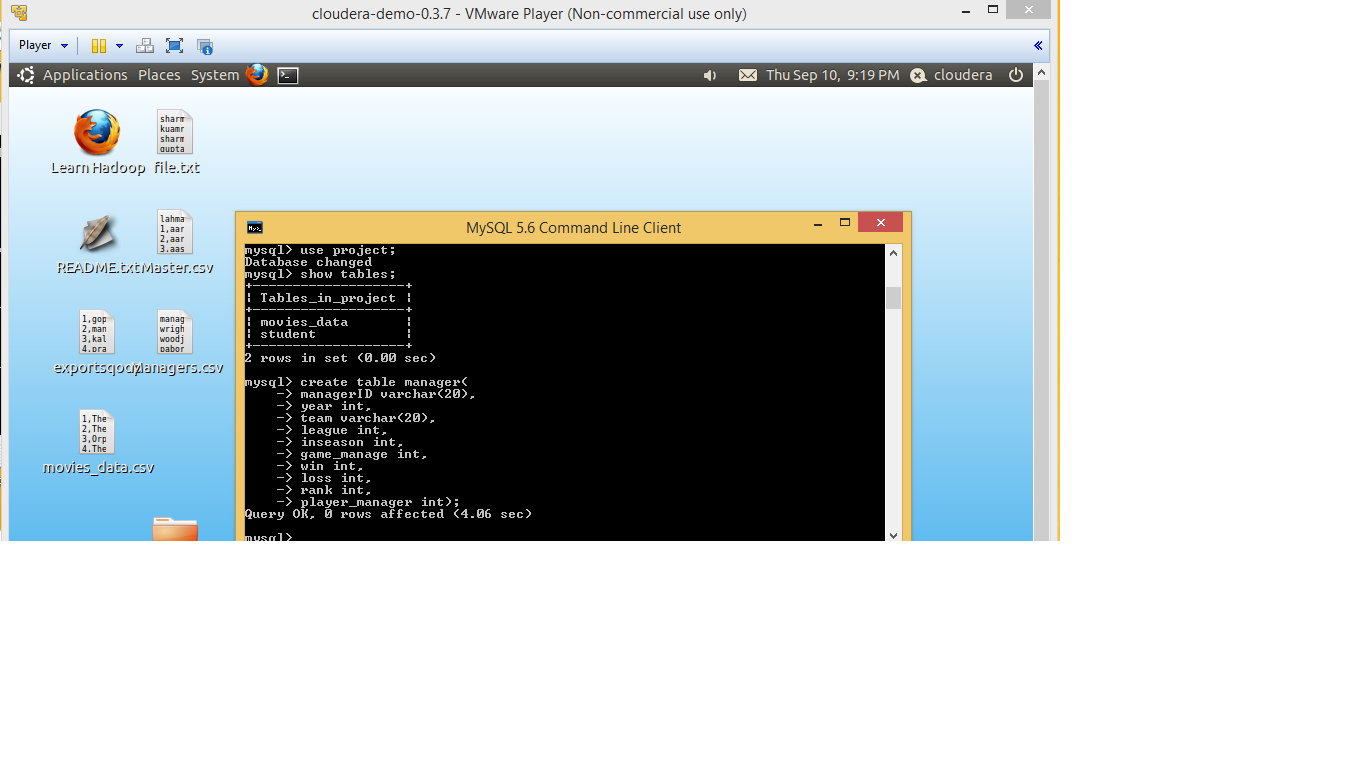
****

****

****

****

**Create table Schema for the Data of Baseball Match in the MySQL**

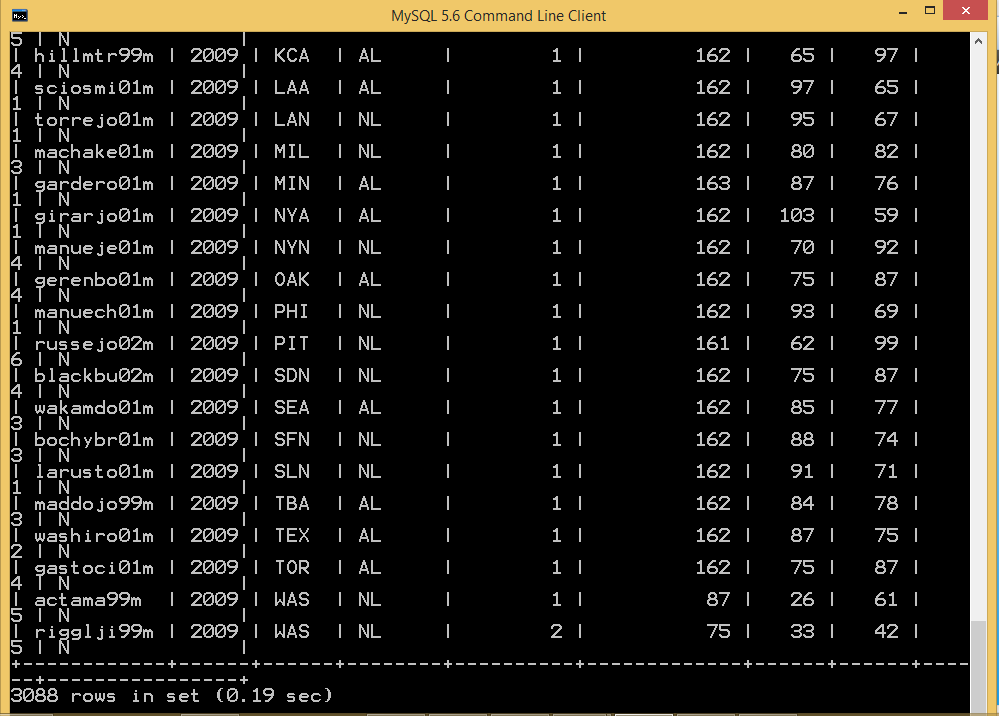
****

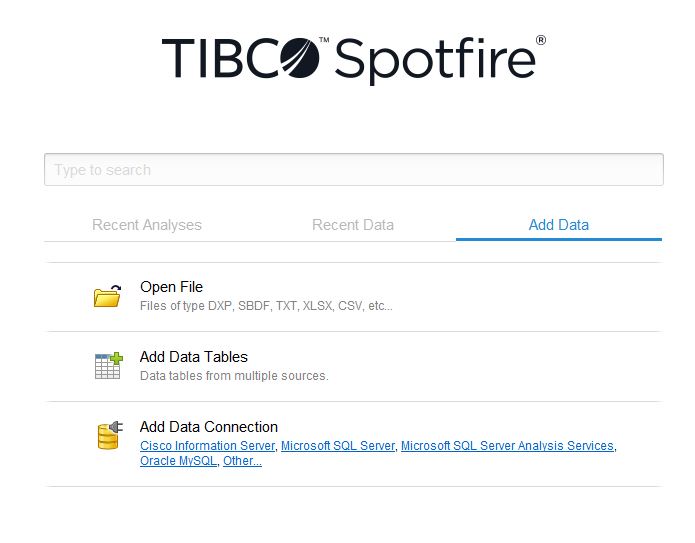
$ sqoop export –-connect jdbc:mysq://lPAddress/databasename

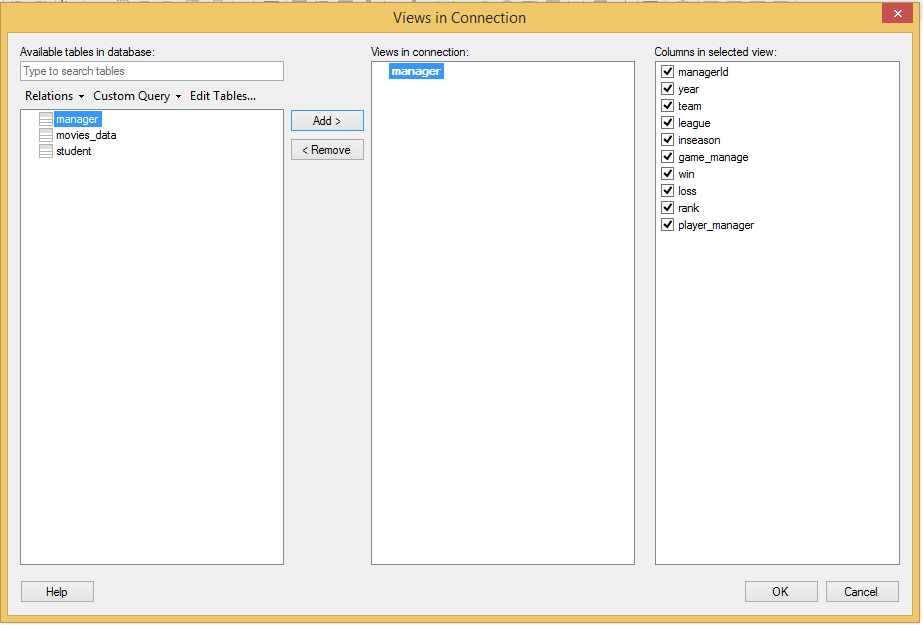
--username root -Password Password --table name

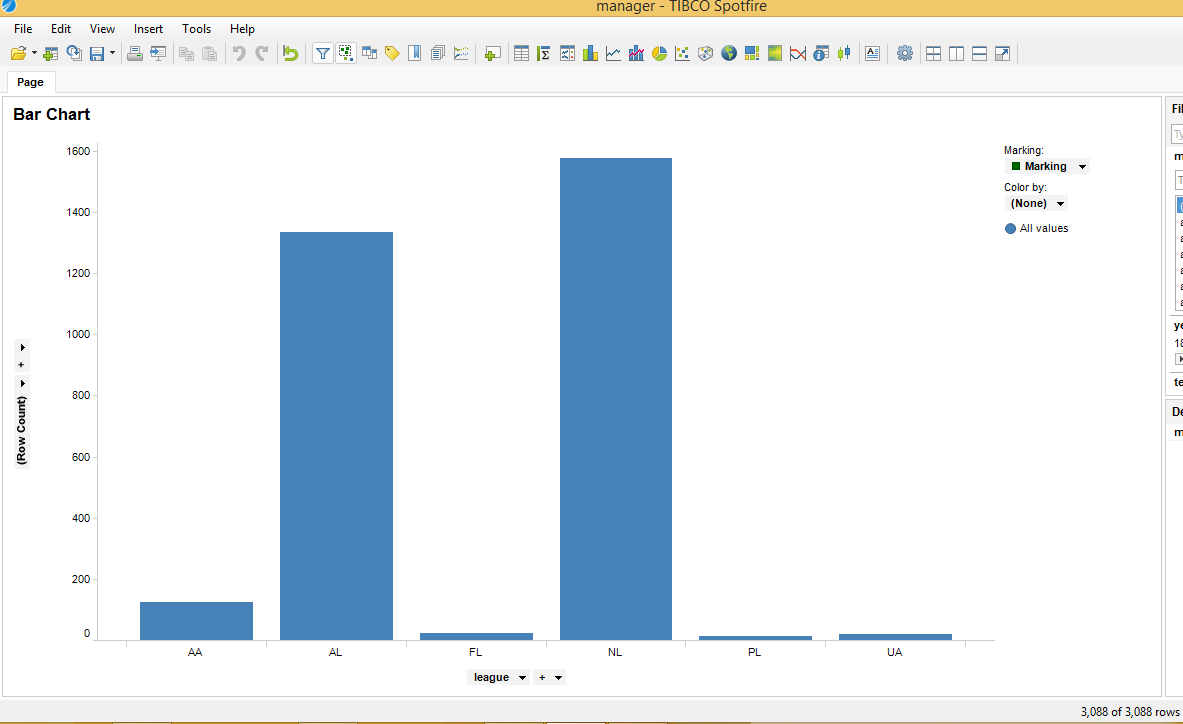
--export-dir “Path of table”

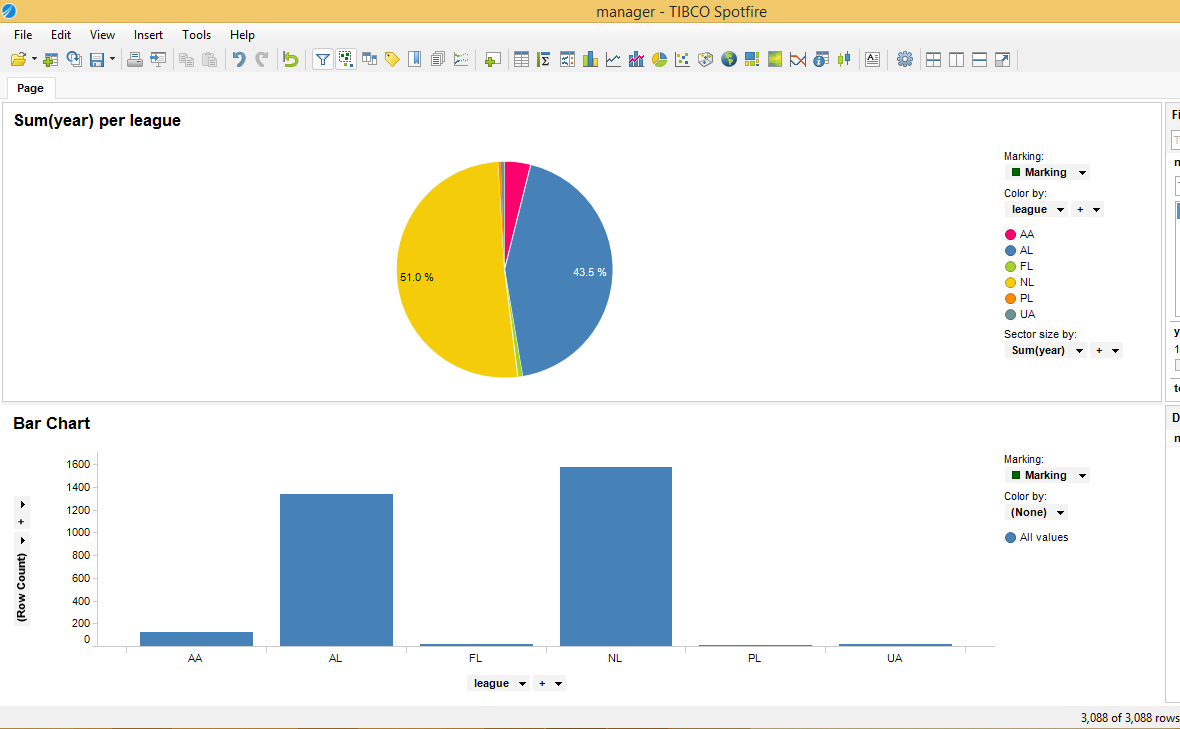
Using This Command the Analyzed Data of Pig Stored into MySQL

****

****

****

****

****