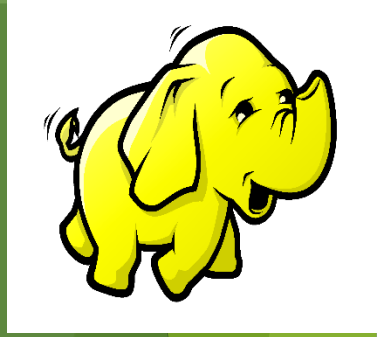
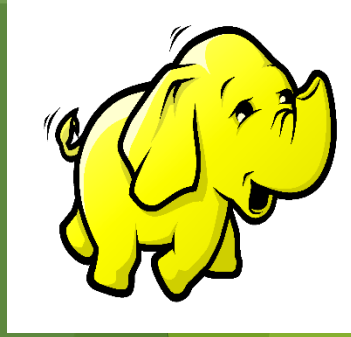


The Hadoop Ecosystem

By

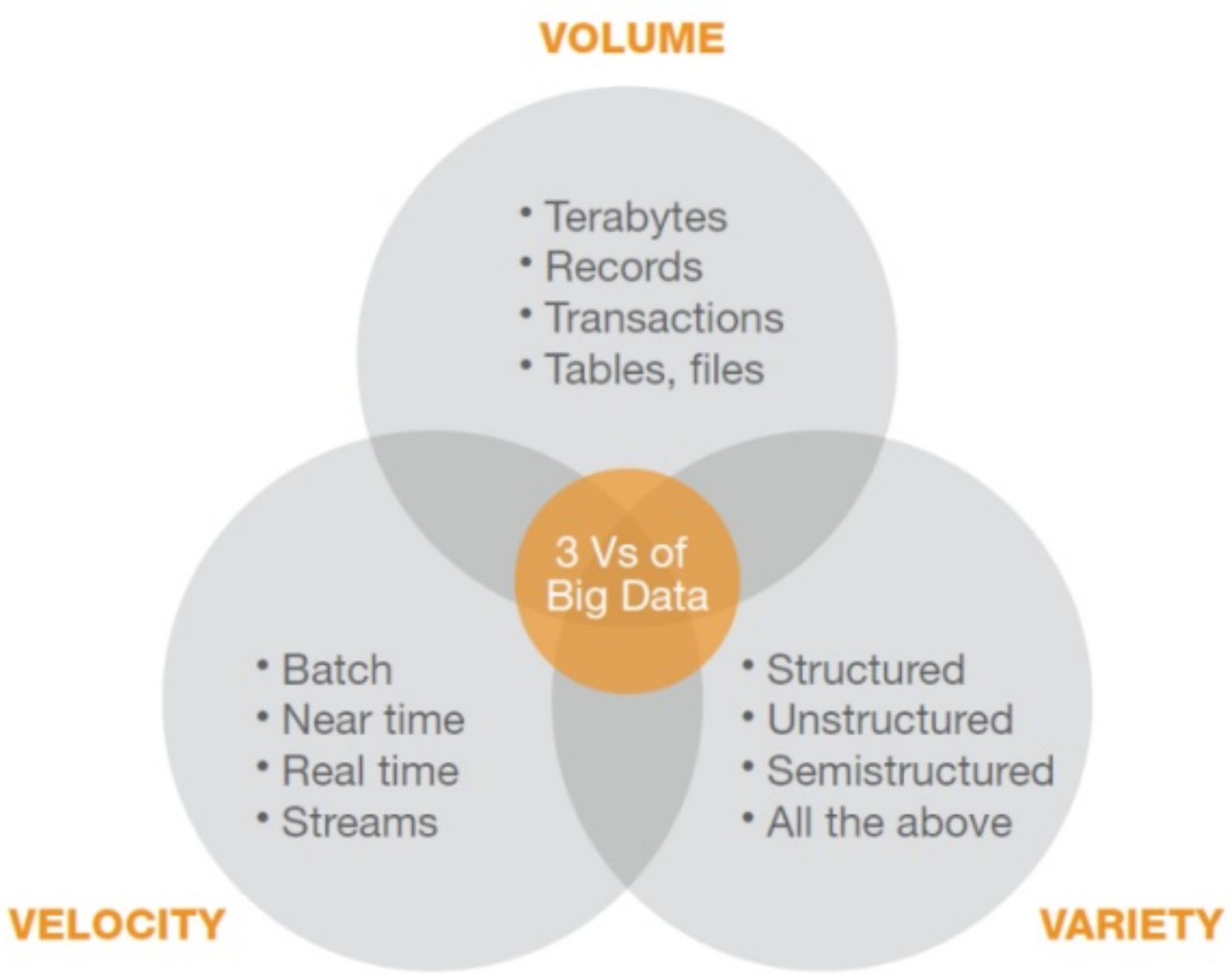
Pragya Singhvi

Overview

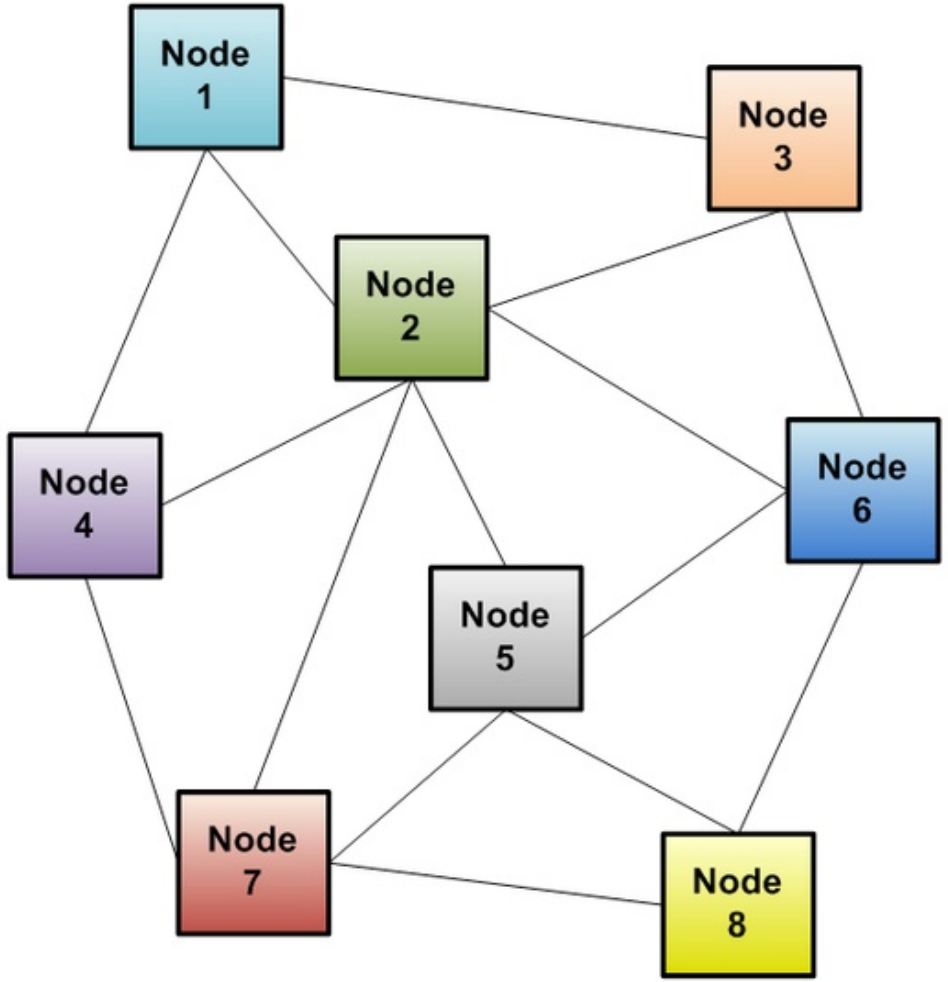


- Big Data Challenges
- Distributed system and challenges
- Hadoop Introduction
- History
- Who uses Hadoop
- The Hadoop Ecosystem
  - Hadoop core components
    - HDFS
    - Map Reduce
  - Other Hadoop ecosystem components
    - Hbase
    - Hive
    - Pig
    - Impala
    - Sqoop
    - Flume
    - Hue
    - Zookeeper
- Demo

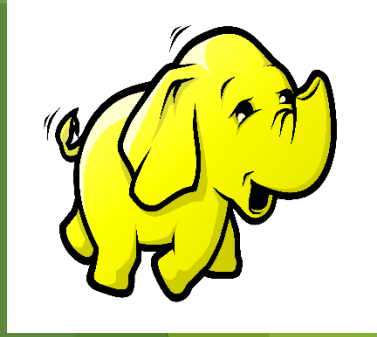
Big Data Challenges



Solution: Distributed system

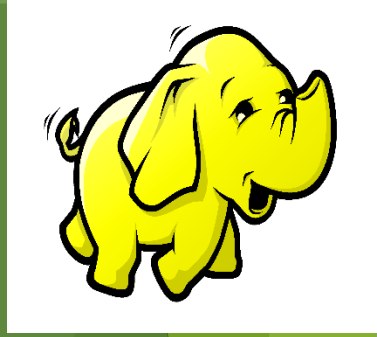


Distributed System Challenges



- Programming Complexity
- Finite bandwidth
- Partial failure
- The data bottleneck

New Approach to distributed computing

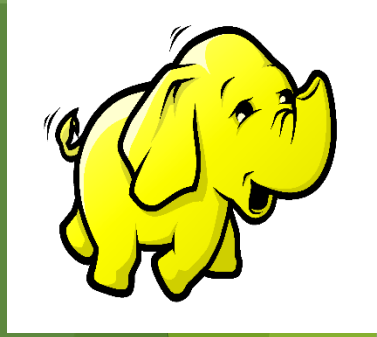


Hadoop:

A scalable fault-tolerant distributed system for data storage and processing

- Distribute data when the data is stored
- Process data where the data is
- Data is replicated

Hadoop Introduction

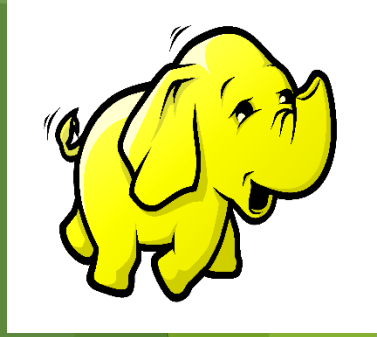


- Apache Hadoop is an open-source software framework for storage and large-scale processing of data-sets on clusters of commodity hardware.
- Some of the characteristics:
  - Open source
  - Distributed processing
  - Distributed storage
  - Scalable
  - Reliable
  - Fault-tolerant
  - Economical
  - Flexible

History

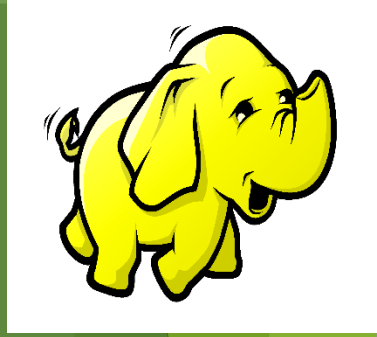
- Originally built as a Infrastructure for the “Nutch” project.
- Based on Google’s map reduce and google File System.
- Created by Doug Cutting in 2005 at Yahoo
- Named after his son’s toy yellow elephant.

Who uses Hadoop



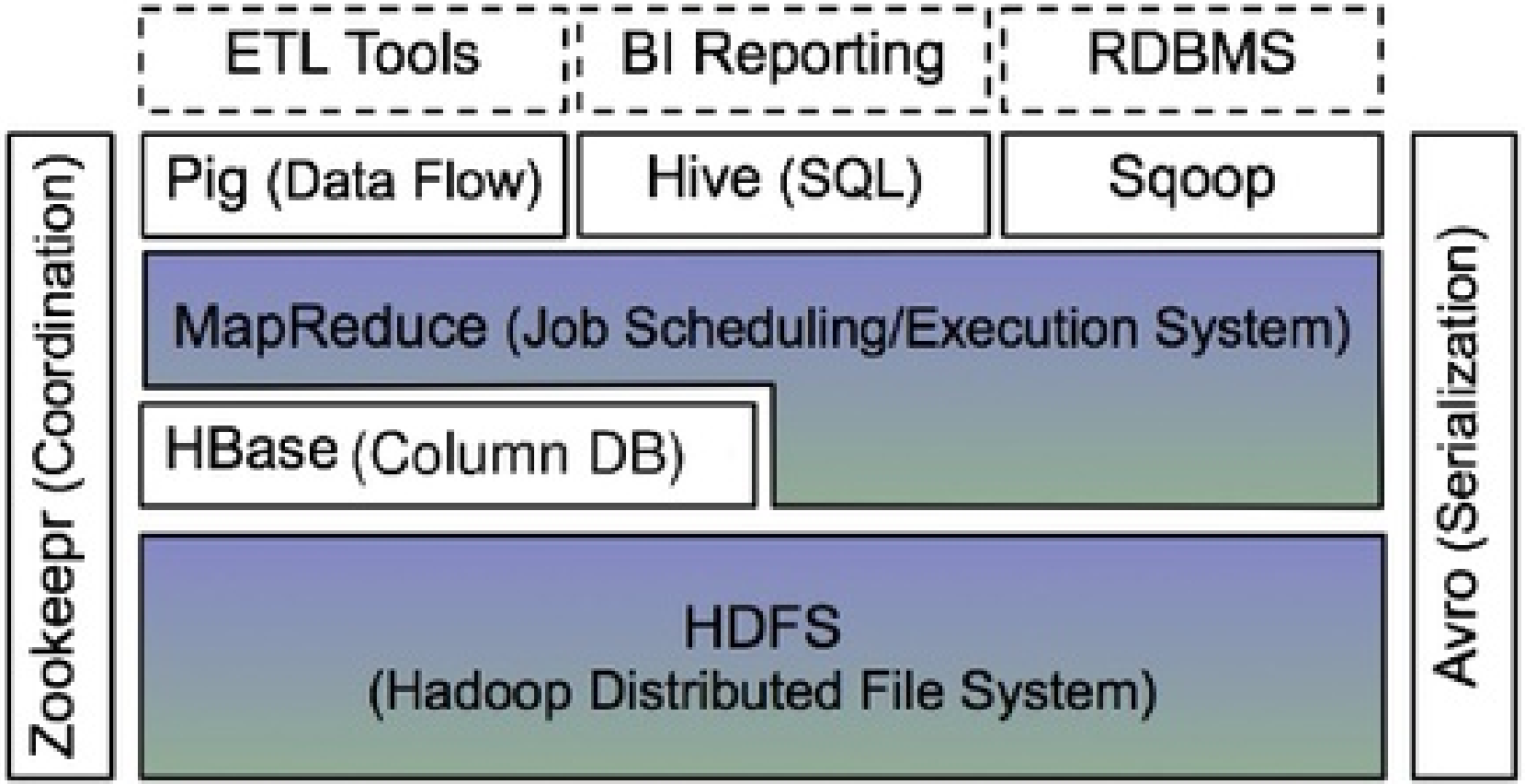
- http://wiki.apache.org/hadoop/PoweredBy
- http://wiki.apache.org/hadoop/Distributions%20and%20Commercial%20Support

The Hadoop Ecosystem



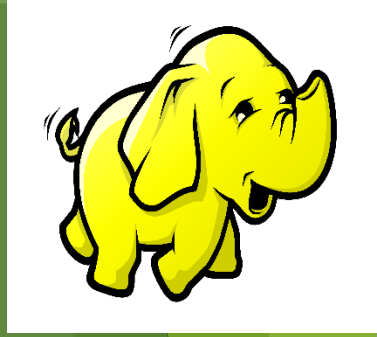
- http://hadoopecosystemtable.github.io/

The Hadoop Ecosystem



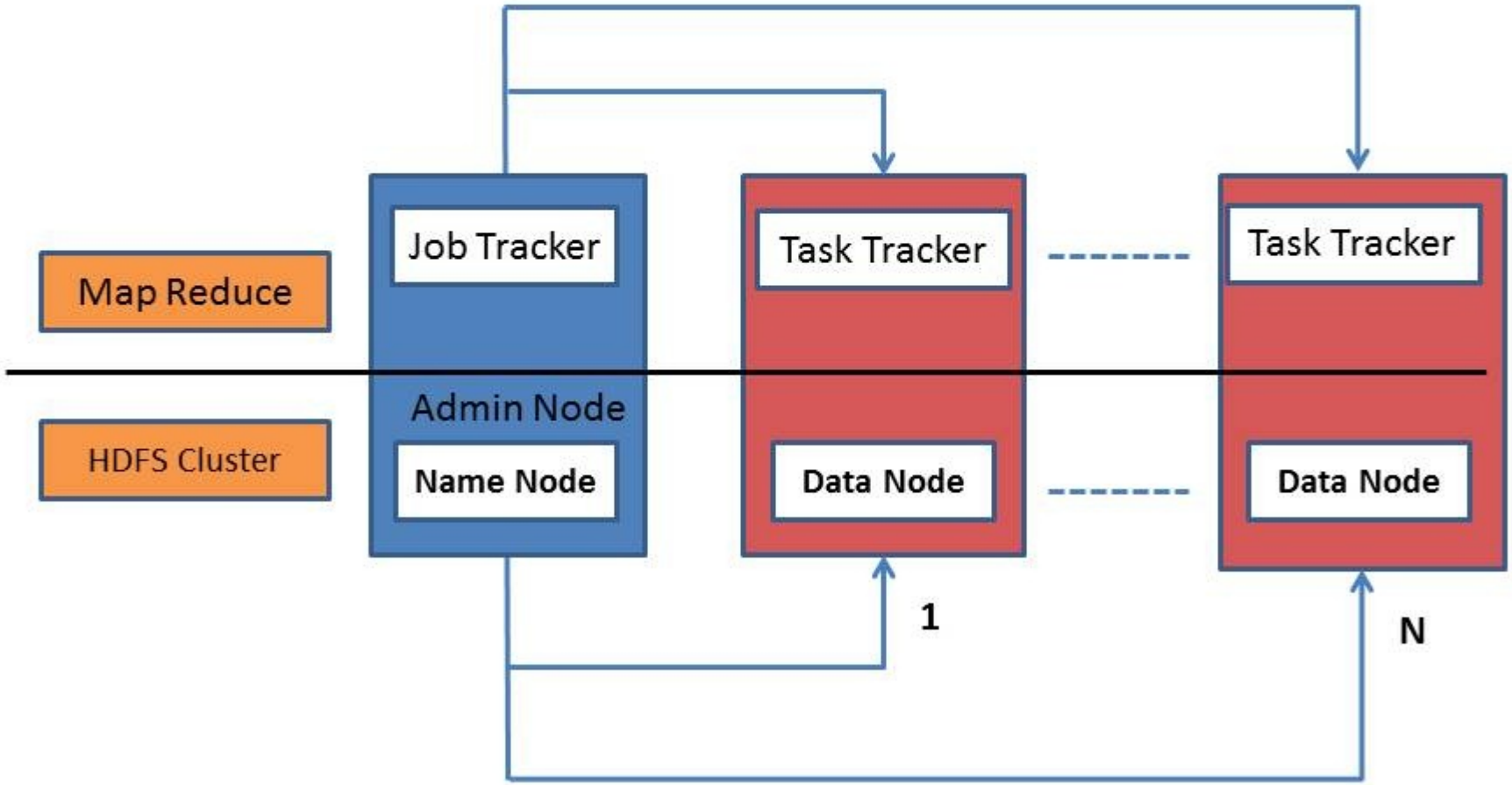
cloudera

Hadoop Core Components

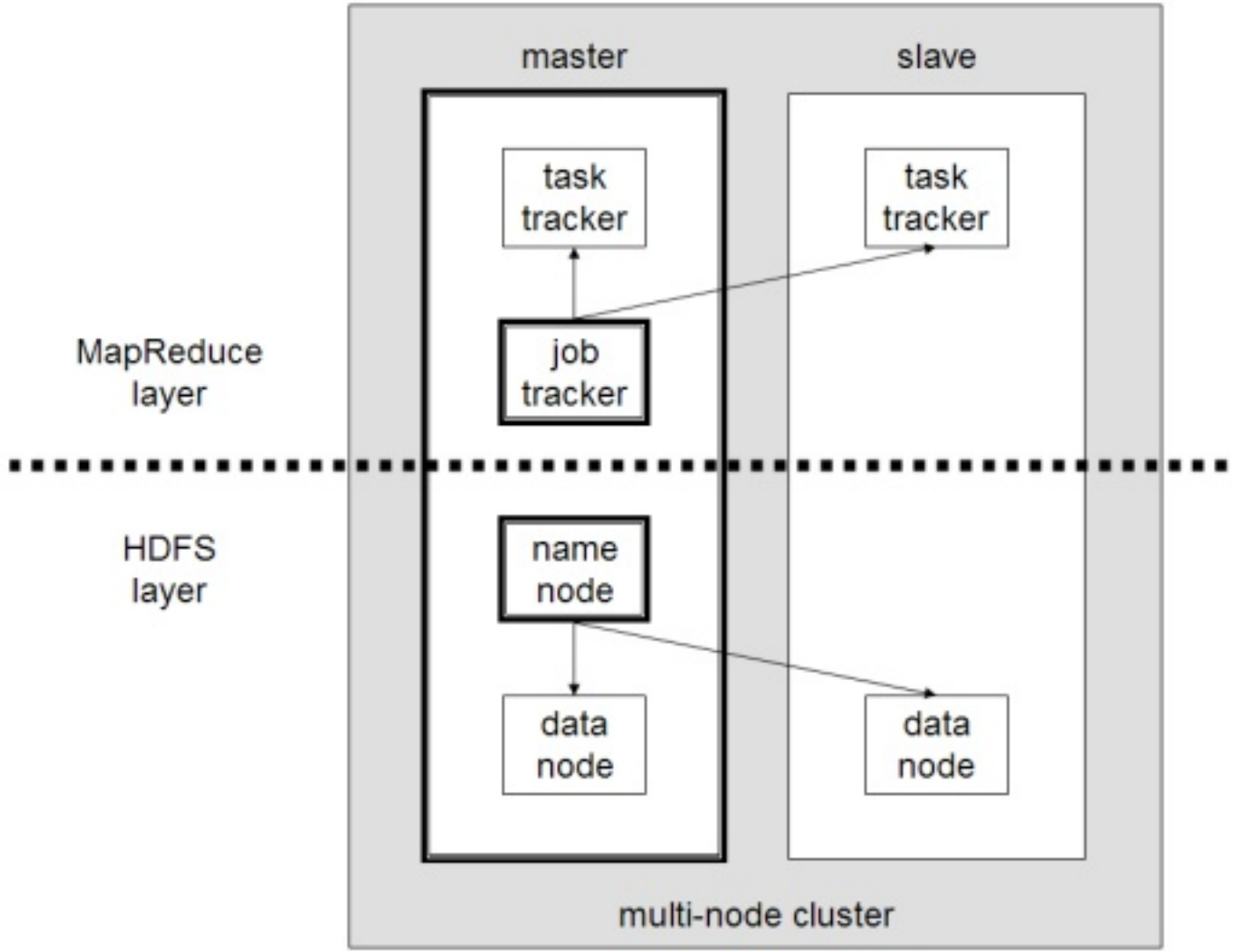


- HDFS – Hadoop Distributed File System (Storage)
- Map Reduce (Processing)

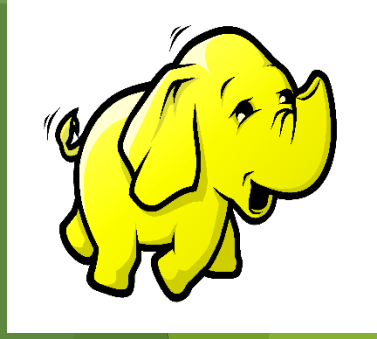
Hadoop Core Components



A multi-node Hadoop cluster

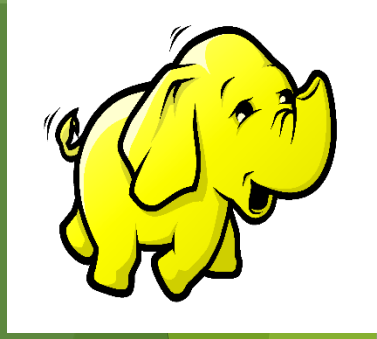


Nodes



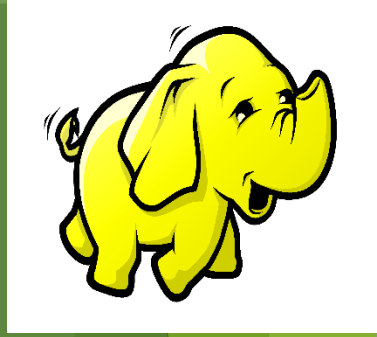
- NameNode:
  - Master of the system
  - Maintains and manages the blocks which are present on the DataNodes
- DataNodes:
  - Slaves which are deployed on each machine and provide the actual storage
  - Responsible for serving read and write requests for the clients
- Jobtracker:
  - takes care of all the job scheduling and assign tasks to Task Trackers.
- TaskTracker:
  - a node in the cluster that accepts tasks - Map, Reduce and Shuffle operations - from a jobtracker

HDFS



- Hadoop Distributed File System (HDFS) is designed to reliably store very large files across machines in a large cluster. It is inspired by the GoogleFileSystem.
- Distribute large data file into blocks
- Blocks are managed by different nodes in the cluster
- Each block is replicated on multiple nodes
- Name node stored metadata information about files and blocks

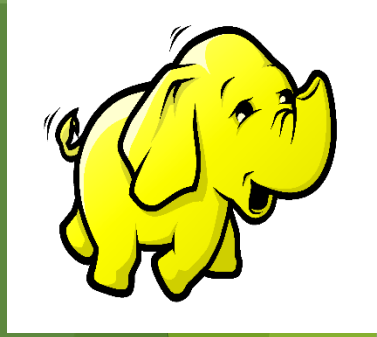
Map Reduce



- The Mapper:
  - Each block is processed in isolation by a map task called mapper
  - Map task runs on the node where the block is stored

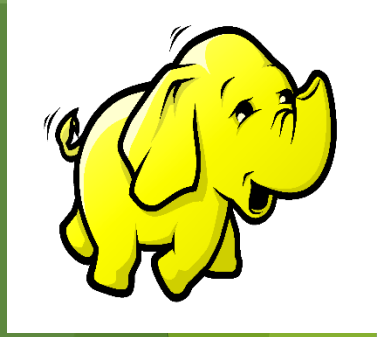
- The Reducer:
  - Consolidate result from different mappers
  - Produce final output

What makes Hadoop unique



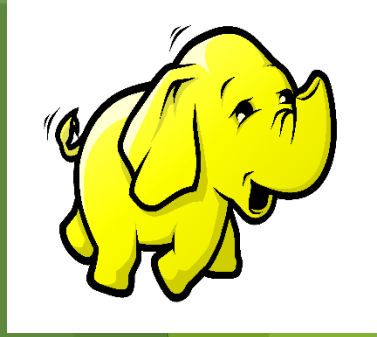
- Moving computation to data, instead of moving data to computation.
- Simplified programming model: allows user to quickly write and test
- Automatic distribution of data and work across machines

Other Hadoop components in Ecosystem



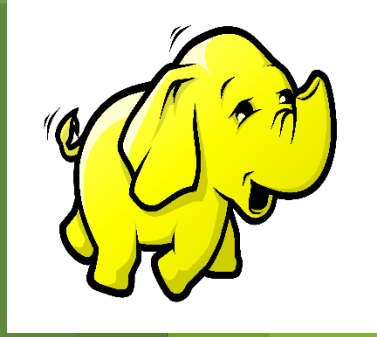
HBase	Hadoop database for random read/write access
Hive	SQL-like queries and tables on large datasets
Pig	Data flow language and compiler
Oozie	Workflow for interdependent Hadoop jobs
Sqoop	Integration of databases and data warehouses with Hadoop
Flume	Configurable streaming data collection
ZooKeeper	Coordination service for distributed applications

Hbase



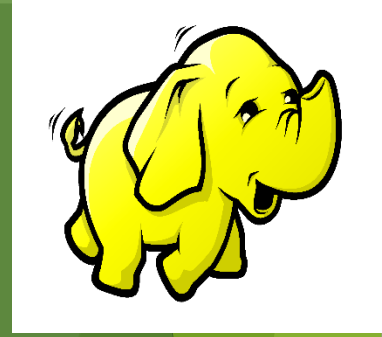
- HBase is an open source, non-relational, distributed database modeled after Google's BigTable.
- It runs on top of Hadoop and HDFS, providing BigTable-like capabilities for Hadoop.

Features of Hbase



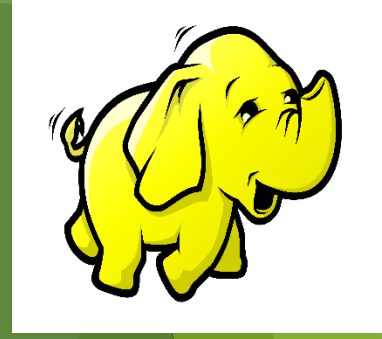
- Type of NoSql database
- Strongly consistent read and write
- Automatic sharding
- Automatic RegionServer failover
- Hadoop/HDFS Integration
- HBase supports massively parallelized processing via MapReduce for using HBase as both source and sink.
- HBase supports an easy to use Java API for programmatic access.
- HBase also supports Thrift and REST for non-Java front-ends.

Hbase in CAP theorem



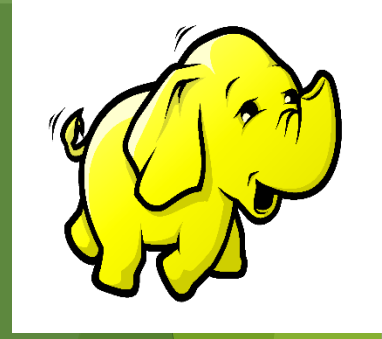
- Eric Brewer’s CAP theorem, HBase is a CP type system.

When to use Hbase



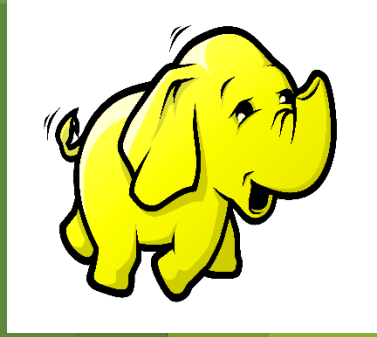
- When there is real big data: millions or billions of rows, in other way data can not store in a single node.
- When random read/write access to big data
- When require to do thousands of operations on big data
- When there is no need of extra features of RDMS like typed columns, secondary indexes, transactions, advanced query languages, etc.
- When there is enough hardware.

Difference between Hbase and HDFS



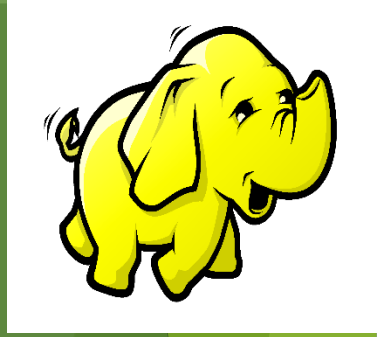
HDFS	Hbase
Good for storing large file	Built on top of HDFS. Good for hosting very large tables like billions of rows X millions of column
Write once. Append to files in some of recent versions but not commonly used	Read/write many
No random read/write	Random read/write
No individual record lookup rather read all data	Fast records lookup(update)

Hive



- An sql like interface to Hadoop.
- Data warehouse infrastructure built on top of Hadoop
- Provide data summarization, query and analysis
- Query execution via MapReduce
- Hive interpreter convert the query to Map reduce format.
- Open source project.
- Developed by Facebook
- Also used by Netflix, Cnet, Digg, eHarmony etc.

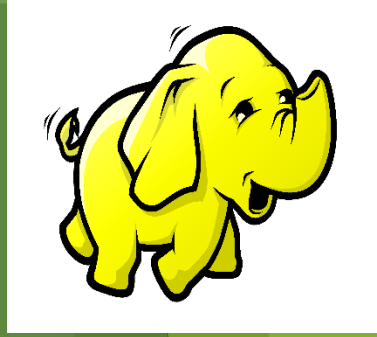
Hive



- HiveQL example:

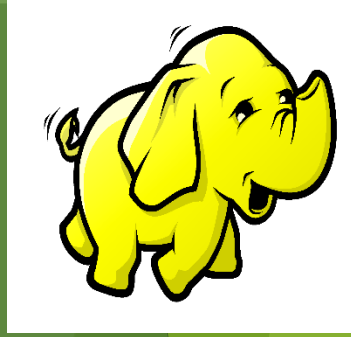
```
SELECT customerId, max(total_cost) from hive_purchases GROUP BY
customerId HAVING count(*) > 3;
```

Pig



- A scripting platform for processing and analyzing large data sets
- Apache Pig allows to write complex MapReduce programs using a simple scripting language.
- High level language: Pig Latin
- Pig Latin is data flow language.
- Pig translate Pig Latin script into MapReduce to execute within Hadoop.
- Open source project
- Developed by Yahoo

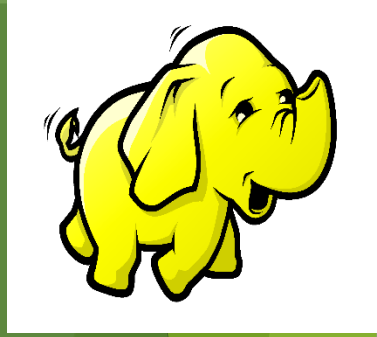
Pig



- Pig Latin example:

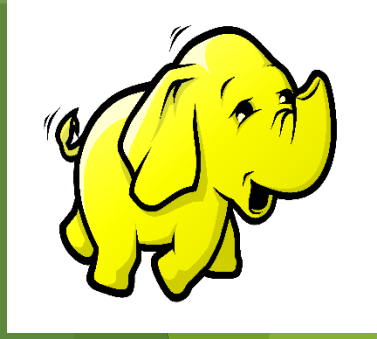
```
A = LOAD 'student' USING PigStorage() AS (name:chararray, age:int,
gpa:float);

X = FOREACH A GENERATE name,$2;

DUMP X;
```

Pig and Hive



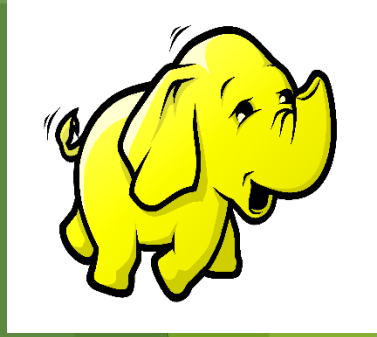
- Both requires compiler to generate Map reduce jobs
- Hence high latency queries when used for real time responses to ad-hoc queries
- Both are good for batch processing and ETL jobs
- Fault tolerant

Impala



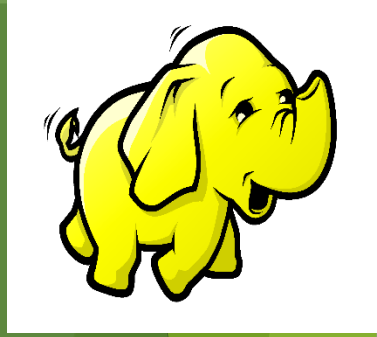
- Cloudera Impala is a query engine that runs on Apache Hadoop.
- Similar to HiveQL.
- Does not use Map reduce
- Optimized for low latency queries
- Open source apache project
- Developed by Cloudera
- Much faster than Hive or pig

Comparing Pig, Hive and Impala

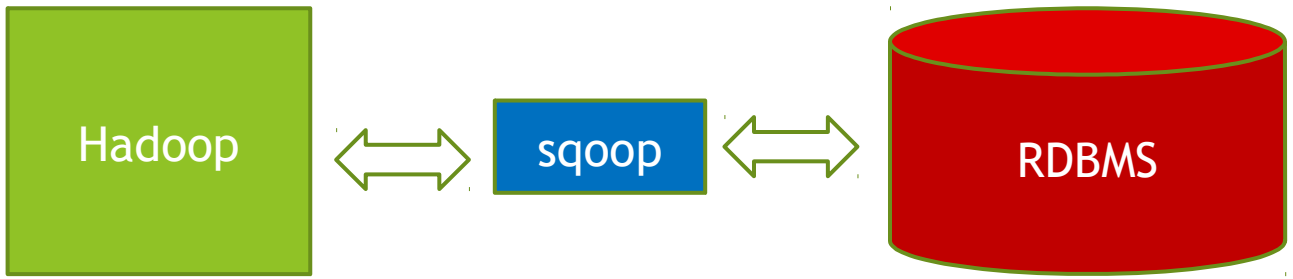


Description of Feature	Pig	Hive	Impala
SQL based query language	No	yes	yes
Schema	optional	required	required
Process data with external scripts	yes	yes	no
Extensible file format support	yes	yes	no
Query speed	slow	slow	fast
Accessible via ODBC/JDBC	no	yes	yes

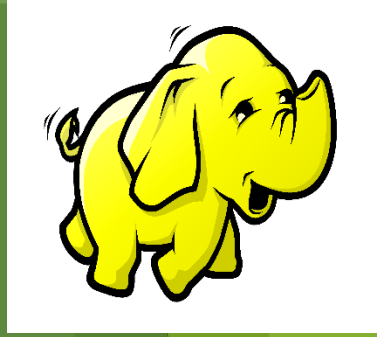
Sqoop



- Command-line interface for transforming data between relational database and Hadoop
- Support incremental imports
- Imports use to populate tables in Hadoop
- Exports use to put data from Hadoop into relational database such as SQL server



How Sqoop works



- The dataset being transferred is broken into small blocks.
- Map only job is launched.
- Individual mapper is responsible for transferring a block of the dataset.

How Sqoop works

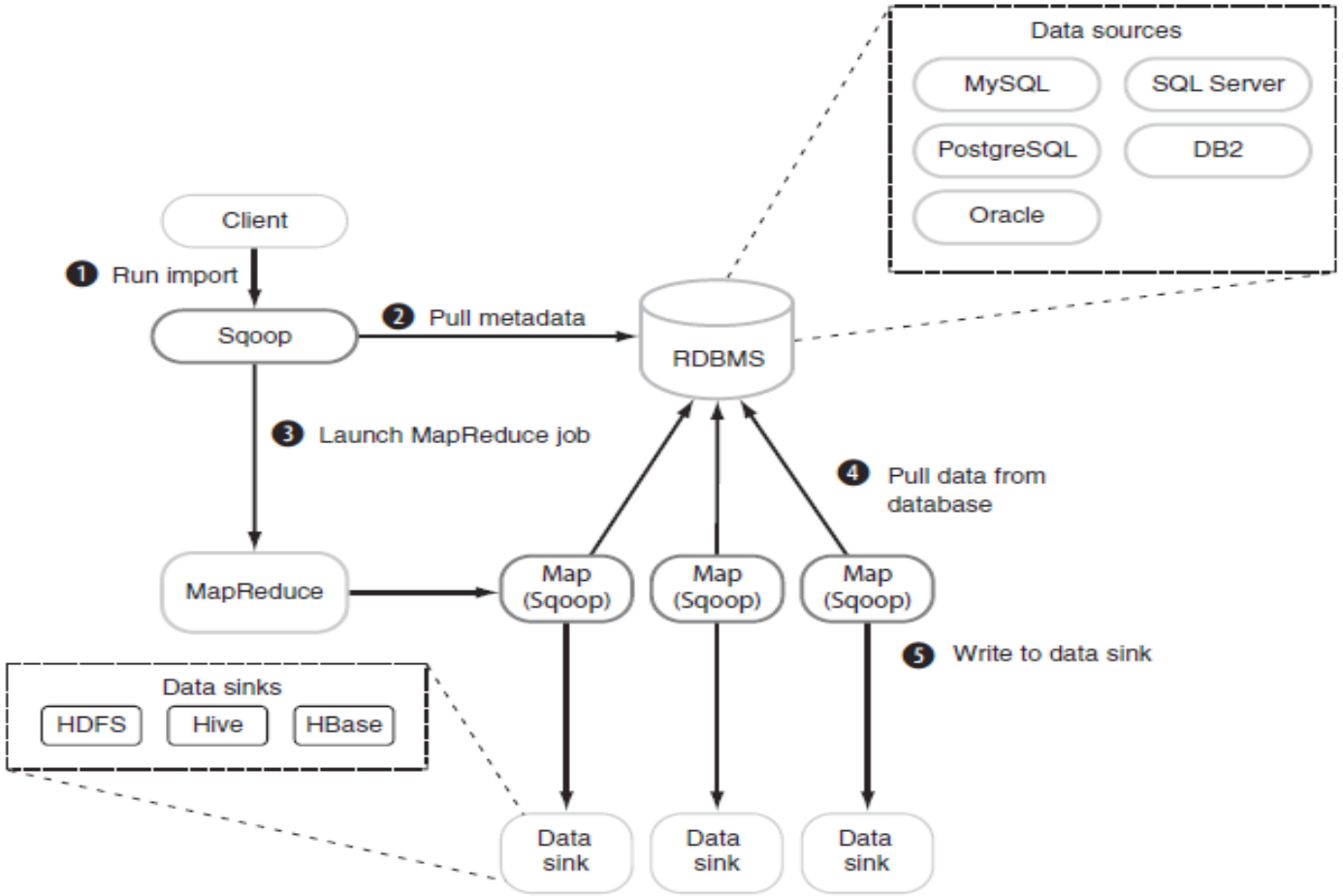
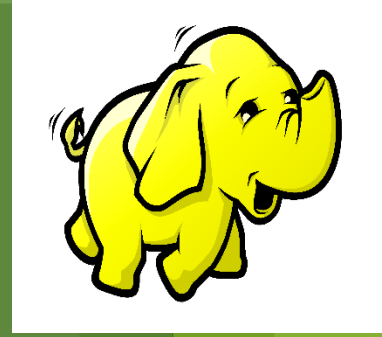


Figure 2.20 **Five-stage Sqoop import overview: connecting to the data source and using MapReduce to write to a data sink**

Flume



- Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).

How flume works

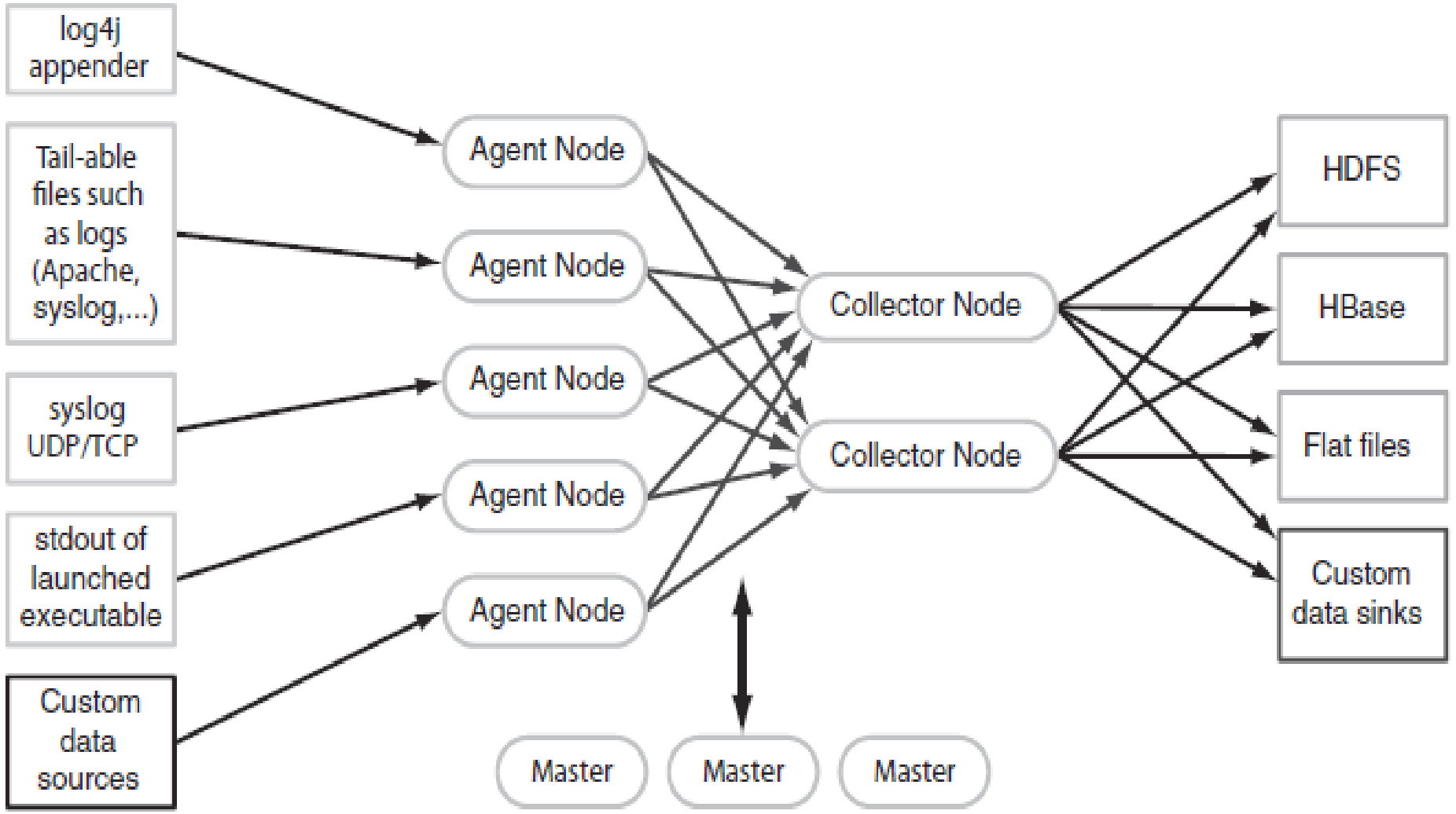
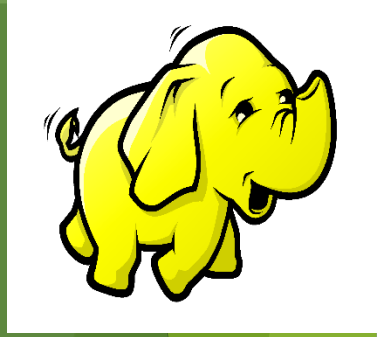


Figure 2.2 **Flume architecture for collecting streaming data**

How flume works

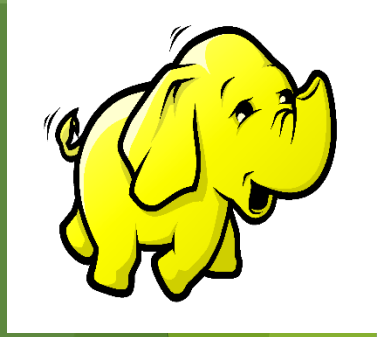


- Data flows like:

  Agent tier -> Collector tier -> Storage tier

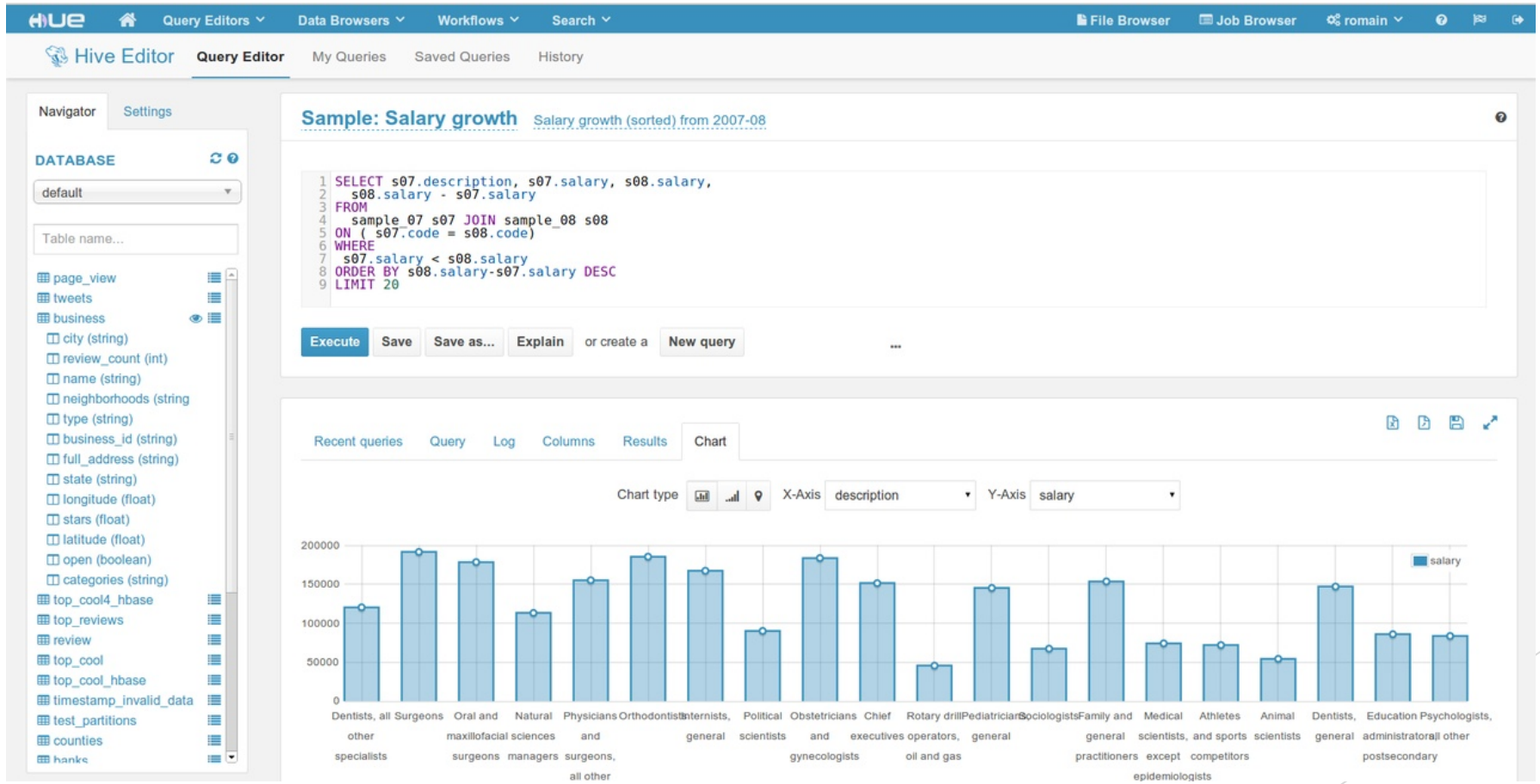
- **Agent nodes** are typically installed on the machines that generate the logs and are data's initial point of contact with Flume. They forward data to the next tier of ***collector nodes***, which aggregate the separate data flows and forward them to the final ***storage tier***.

Hue

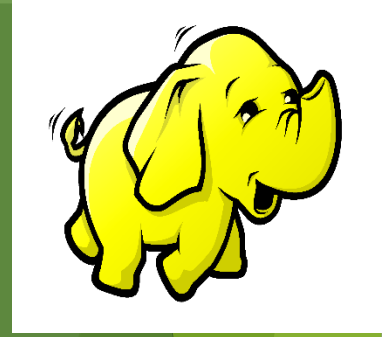


- Graphical front end to the cluster.
- Open source web interface.
- Makes Hadoop platform (HDFS, Map reduce, oozie, Hive, etc.) easy to use

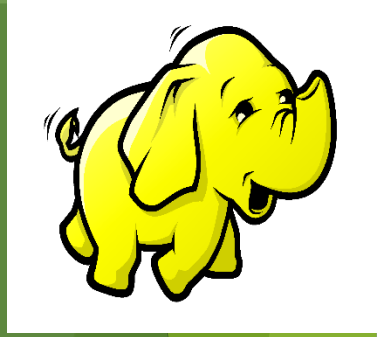
Hue


HUE
Query Editors
Data Browsers
Workflows
Search
File Browser
Job Browser
romain
Hive Editor
Query Editor
My Queries
Saved Queries
History
Navigator
Settings
DATABASE
default
Table name...
page_view
tweets
business
city (string)
review_count (int)
name (string)
neighborhoods (string
type (string)
business_id (string)
full_address (string)
state (string)
longitude (float)
stars (float)
latitude (float)
open (boolean)
categories (string)
top_cool4_hbase
top_reviews
review
top_cool
top_cool_hbase
timestamp_invalid_data
test_partitions
counties
banks
Sample: Salary growth
Salary growth (sorted) from 2007-08
SELECT s07.description, s07.salary, s08.salary,
s08.salary - s07.salary
FROM
sample_07 s07 JOIN sample_08 s08
ON (s07.code = s08.code)
WHERE
s07.salary < s08.salary
ORDER BY s08.salary-s07.salary DESC
LIMIT 20
Execute
Save
Save as...
Explain
or create a
New query
Recent queries
Query
Log
Columns
Results
Chart
Chart type
X-Axis
description
Y-Axis
salary
200000
150000
100000
50000
0
salary
Dentists, all other specialists
Surgeons
Oral and maxillofacial surgeons
Natural sciences managers
Physicians and surgeons, all other
Orthodontists
Internists, general
Political scientists
Obstetricians and gynecologists
Chief executives
Rotary drill operators, oil and gas
Pediatricians, general
Sociologists
Family and general practitioners
Medical scientists, except epidemiologists
Athletes and sports competitors
Animal scientists
Dentists, general
Education administrators, postsecondary
Psychologists, all other

Zookeeper



- Because coordinating distributed systems is a Zoo.
- ZooKeeper is a centralized service for maintaining configuration information, naming, providing distributed synchronization, and providing group services.

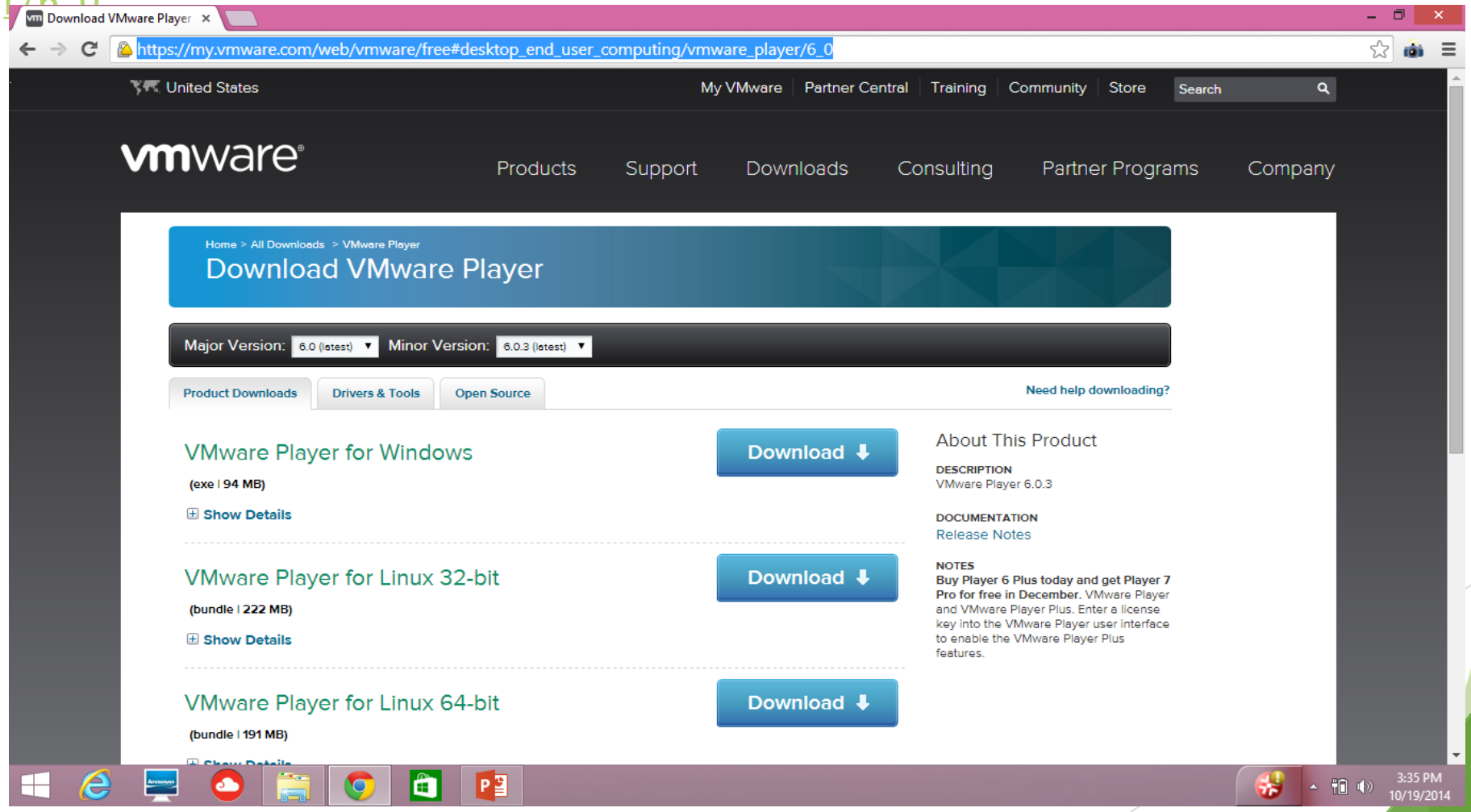


DEMO

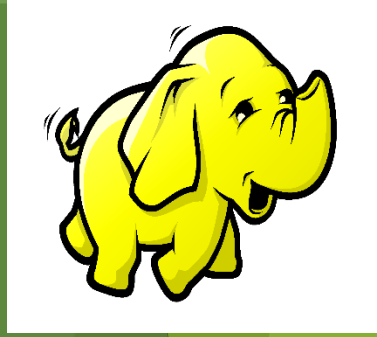
Hadoop Installation (CDH) for windows

- Download and install VM player

https://my.vmware.com/web/vmware/free#desktop_end_user_computing/vmware_player/6_0



Hadoop Installation (CDH) for windows

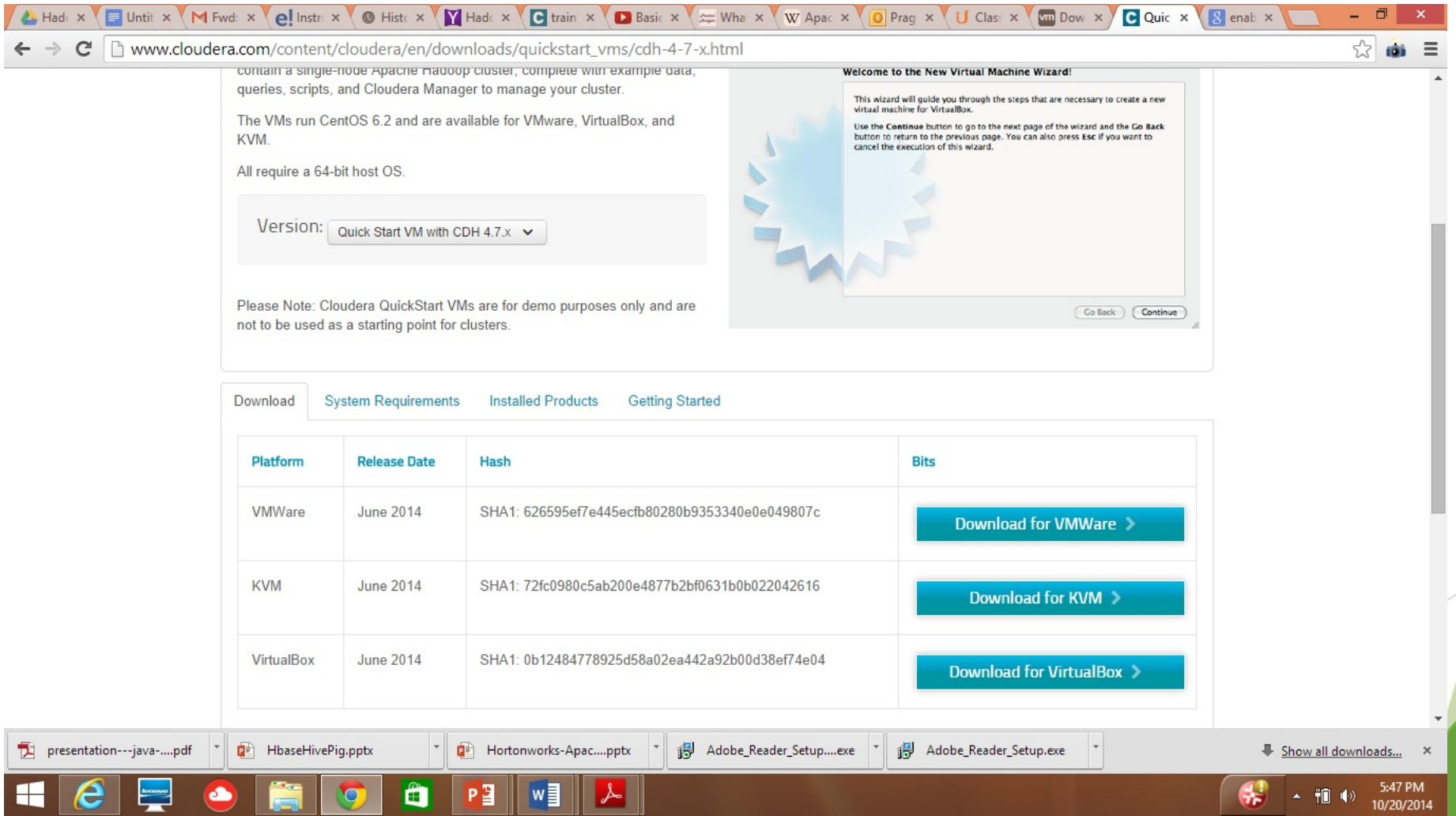


Make sure you have enabled virtualization in bios

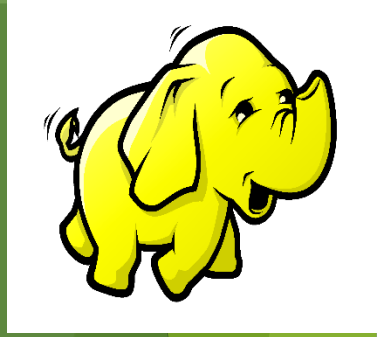
Hadoop Installation (CDH) for windows

- Download “Quick start VM with CDH” : Download for VMWare

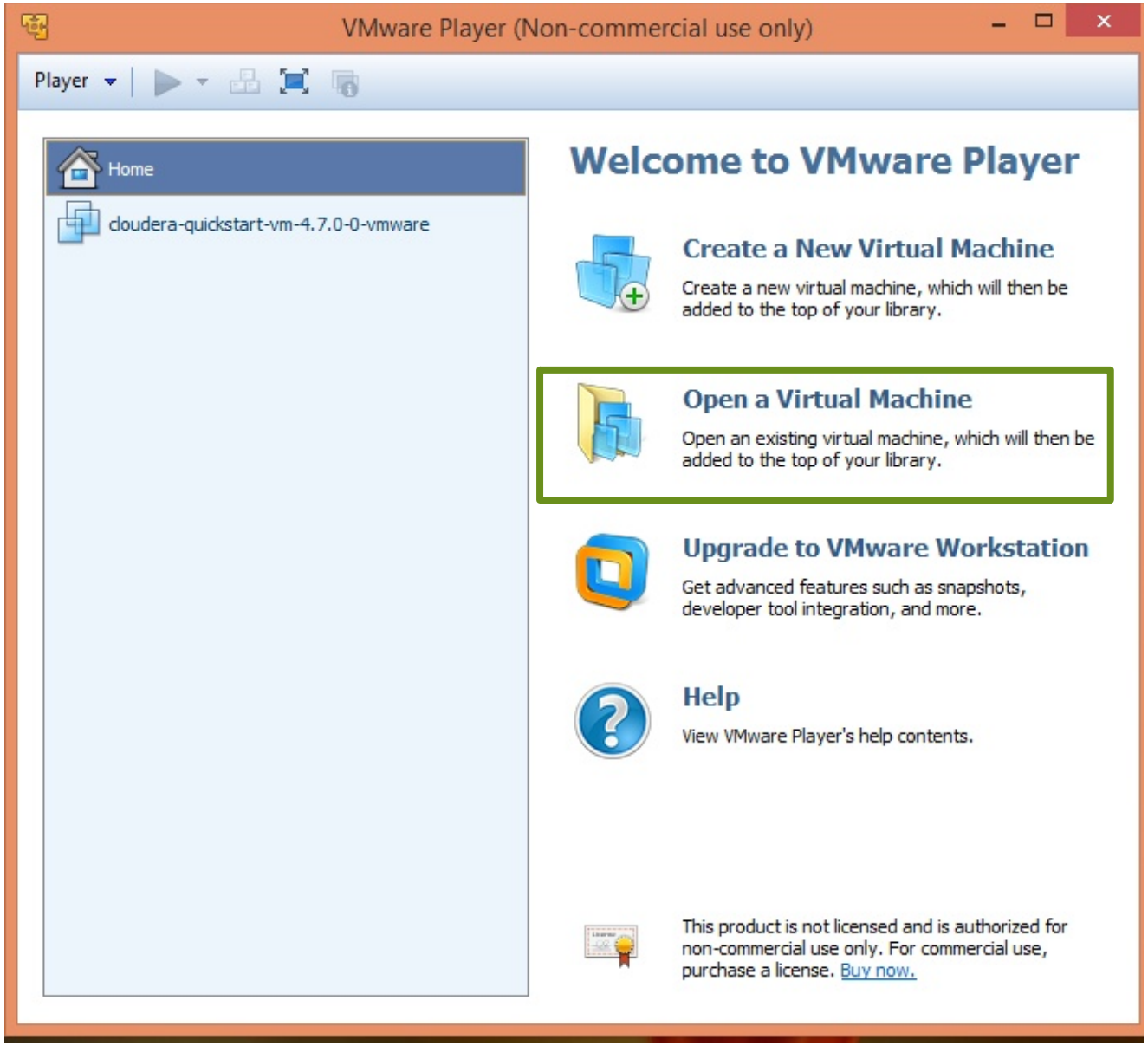
http://www.cloudera.com/content/cloudera/en/downloads/quickstart_vms/cdh-4-7-x.h

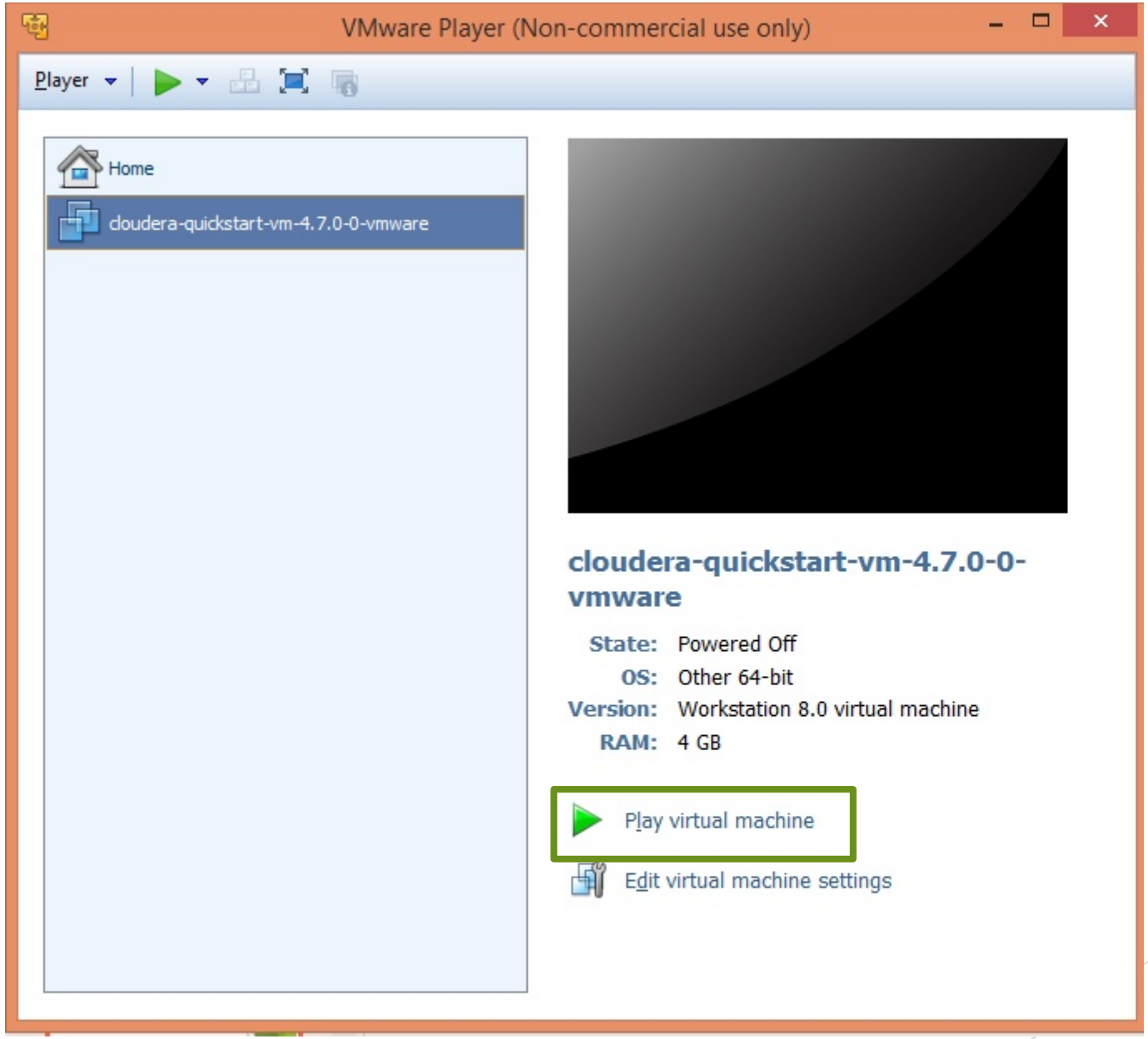


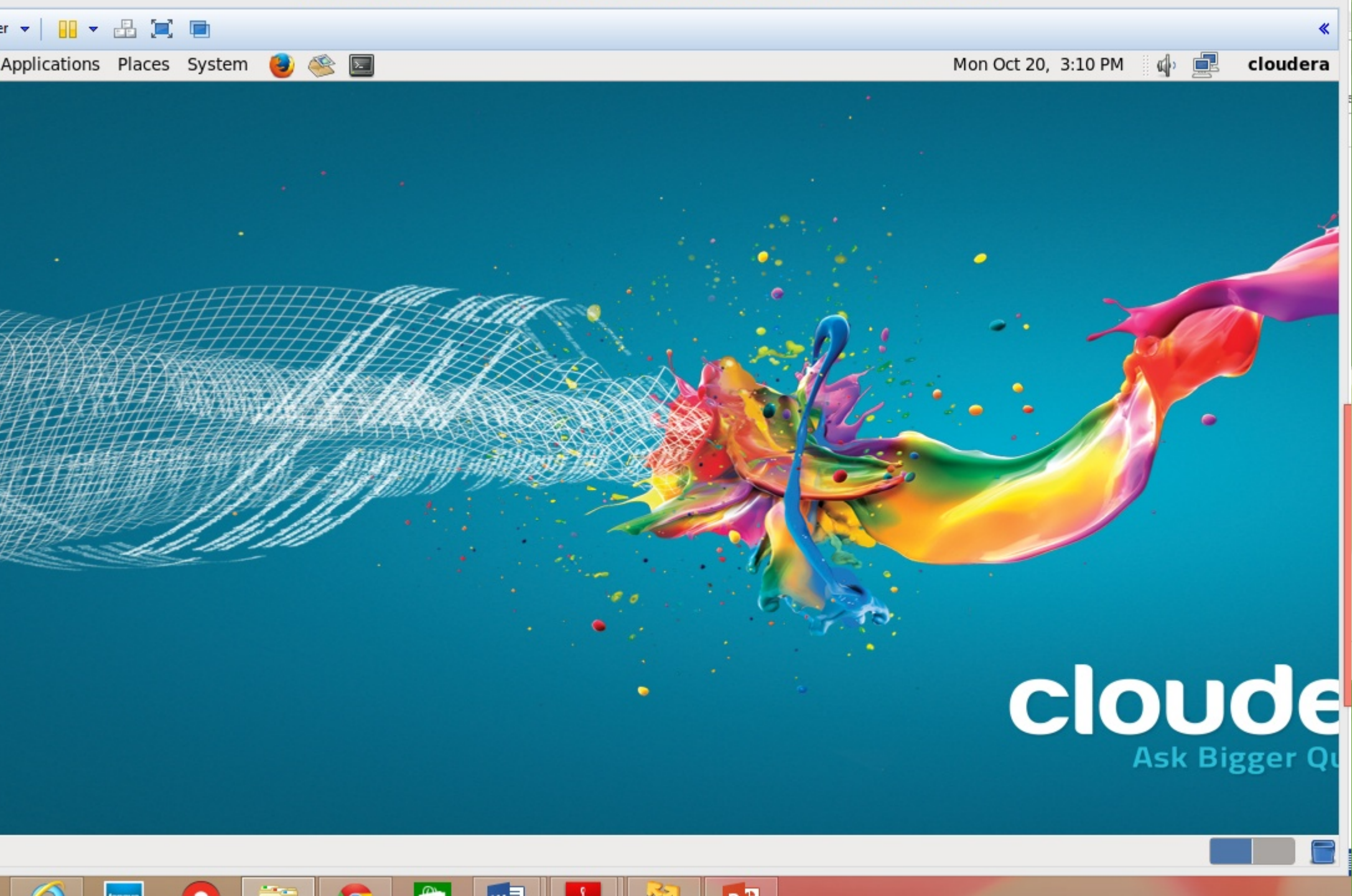
Hadoop Installation (CDH) for windows

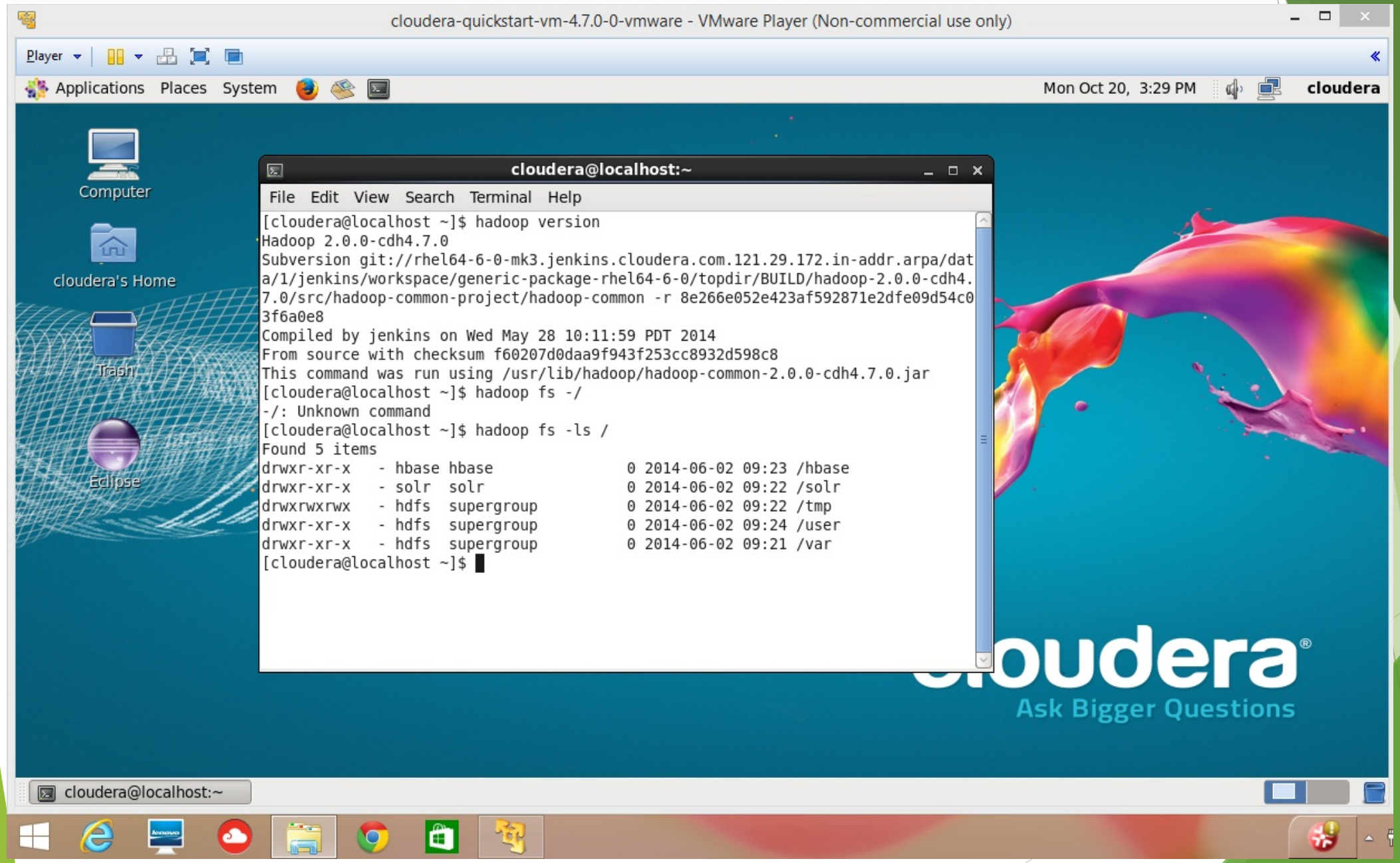


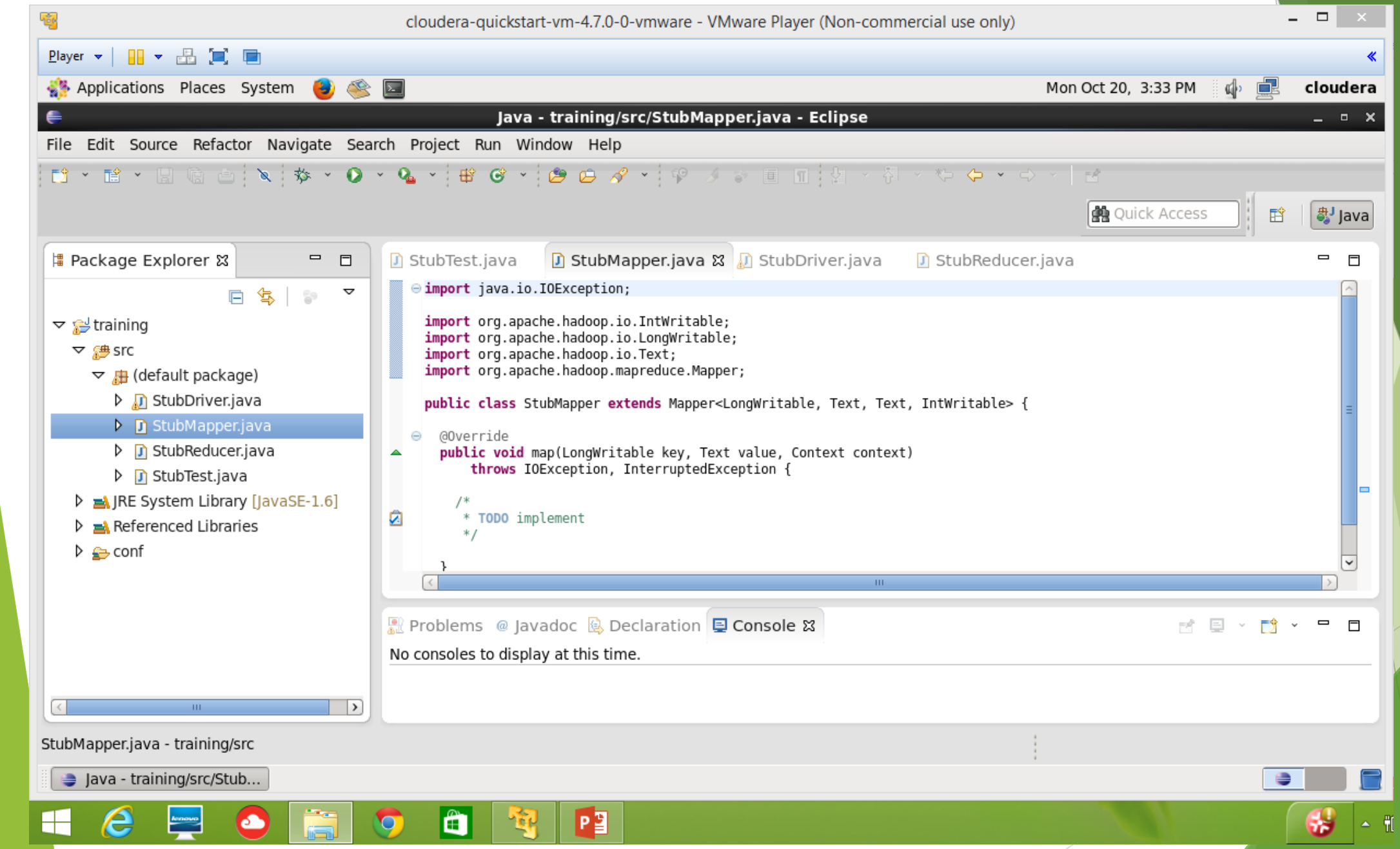
- Unzip “cloudera-quickstart-vm-4.7.0-0-vmware”
- Open CDH using VMPlayer:
  - Open VM Player
  - Click open a virtual machine
  - Select the file “cloudera-quickstart-vm-4.7.0-0-vmware” in the extracted directory of “cloudera-quickstart-vm-4.7.0-0-vmware”. Virtual machine will be added to your VM player.
  - Select this virtual machine and click play virtual machine.

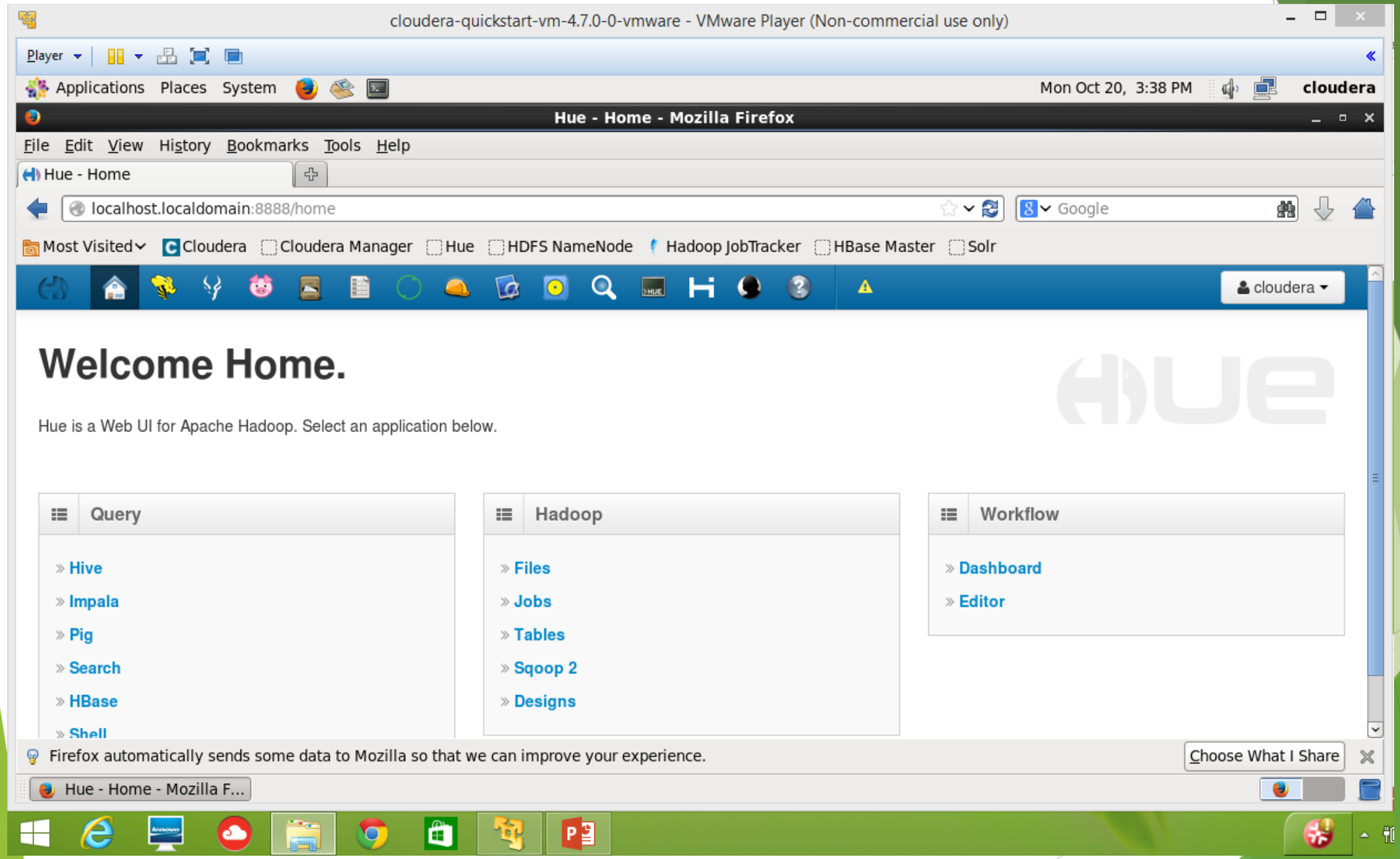

VMware Player (Non-commercial use only)
Player
Home
cloudera-quickstart-vm-4.7.0-0-vmware
Welcome to VMware Player
Create a New Virtual Machine
Create a new virtual machine, which will then be added to the top of your library.
Open a Virtual Machine
Open an existing virtual machine, which will then be added to the top of your library.
Upgrade to VMware Workstation
Get advanced features such as snapshots, developer tool integration, and more.
Help
View VMware Player's help contents.
This product is not licensed and is authorized for non-commercial use only. For commercial use, purchase a license. Buy now.

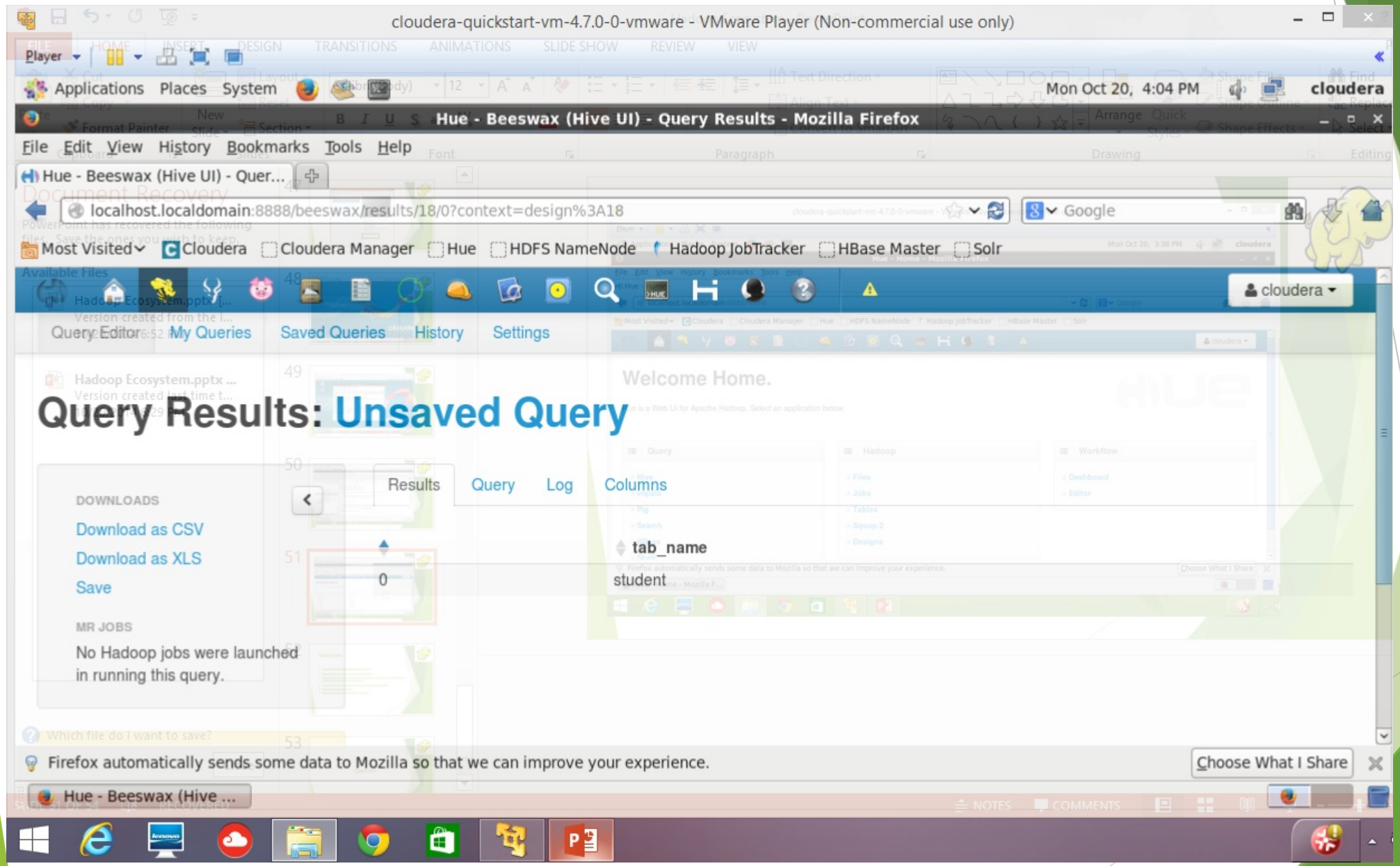

VMware Player (Non-commercial use only)
Player
Home
cloudera-quickstart-vm-4.7.0-0-vmware
cloudera-quickstart-vm-4.7.0-0-vmware
State: Powered Off
OS: Other 64-bit
Version: Workstation 8.0 virtual machine
RAM: 4 GB
Play virtual machine
Edit virtual machine settings


Applications
Places
System
Mon Oct 20, 3:10 PM
cloudera
cloude
Ask Bigger Qu

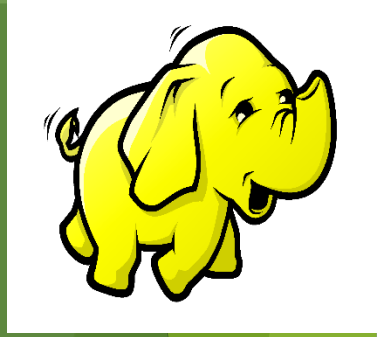

cloudera-quickstart-vm-4.7.0-0-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Mon Oct 20, 3:29 PM
cloudera
Computer
cloudera's Home
Trash
Eclipse
cloudera@localhost:~
File Edit View Search Terminal Help
[cloudera@localhost ~]$ hadoop version
Hadoop 2.0.0-cdh4.7.0
Subversion git://rhel64-6-0-mk3.jenkins.cloudera.com.121.29.172.in-addr.arpa/data/1/jenkins/workspace/generic-package-rhel64-6-0/topdir/BUILD/hadoop-2.0.0-cdh4.7.0/src/hadoop-common-project/hadoop-common -r 8e266e052e423af592871e2dfe09d54c03f6a0e8
Compiled by jenkins on Wed May 28 10:11:59 PDT 2014
From source with checksum f60207d0daa9f943f253cc8932d598c8
This command was run using /usr/lib/hadoop/hadoop-common-2.0.0-cdh4.7.0.jar
[cloudera@localhost ~]$ hadoop fs -/
-/: Unknown command
[cloudera@localhost ~]$ hadoop fs -ls /
Found 5 items
drwxr-xr-x - hbase hbase 0 2014-06-02 09:23 /hbase
drwxr-xr-x - solr solr 0 2014-06-02 09:22 /solr
drwxrwxrwx - hdfs supergroup 0 2014-06-02 09:22 /tmp
drwxr-xr-x - hdfs supergroup 0 2014-06-02 09:24 /user
drwxr-xr-x - hdfs supergroup 0 2014-06-02 09:21 /var
[cloudera@localhost ~]$
cloudera
Ask Bigger Questions
cloudera@localhost:~


cloudera-quickstart-vm-4.7.0-0-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Mon Oct 20, 3:33 PM
cloudera
Java - training/src/StubMapper.java - Eclipse
File Edit Source Refactor Navigate Search Project Run Window Help
Quick Access
Java
Package Explorer
training
src
(default package)
StubDriver.java
StubMapper.java
StubReducer.java
StubTest.java
JRE System Library [JavaSE-1.6]
Referenced Libraries
conf
StubTest.java
StubMapper.java
StubDriver.java
StubReducer.java
import java.io.IOException;
import org.apache.hadoop.io.IntWritable;
import org.apache.hadoop.io.LongWritable;
import org.apache.hadoop.io.Text;
import org.apache.hadoop.mapreduce.Mapper;
public class StubMapper extends Mapper<LongWritable, Text, Text, IntWritable> {
@Override
public void map(LongWritable key, Text value, Context context)
throws IOException, InterruptedException {
/*
* TODO implement
*/
}
Problems Javadoc Declaration Console
No consoles to display at this time.
StubMapper.java - training/src
Java - training/src/Stub...

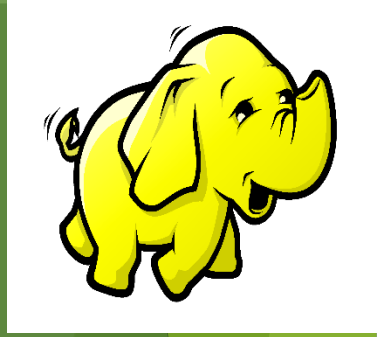

cloudera-quickstart-vm-4.7.0-0-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Mon Oct 20, 3:38 PM
cloudera
Hue - Home - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Hue - Home
localhost.localdomain:8888/home
Google
Most Visited
Cloudera
Cloudera Manager
Hue
HDFS NameNode
Hadoop JobTracker
HBase Master
Solr
cloudera
Welcome Home.
Hue is a Web UI for Apache Hadoop. Select an application below.
Query
Hive
Impala
Pig
Search
HBase
Shell
Hadoop
Files
Jobs
Tables
Sqoop 2
Designs
Workflow
Dashboard
Editor
Firefox automatically sends some data to Mozilla so that we can improve your experience.
Choose What I Share
Hue - Home - Mozilla F...


cloudera-quickstart-vm-4.7.0-0-vmware - VMware Player (Non-commercial use only)
Player
Applications Places System
Mon Oct 20, 4:04 PM
cloudera
Hue - Beeswax (Hive UI) - Query Results - Mozilla Firefox
File Edit View History Bookmarks Tools Help
Hue - Beeswax (Hive UI) - Quer...
localhost.localdomain:8888/beeswax/results/18/0?context=design%3A18
Google
Most Visited
Cloudera
Cloudera Manager
Hue
HDFS NameNode
Hadoop JobTracker
HBase Master
Solr
cloudera
Query Editor
My Queries
Saved Queries
History
Settings
Query Results: Unsaved Query
Results
Query
Log
Columns
DOWNLOADS
Download as CSV
Download as XLS
Save
MR JOBS
No Hadoop jobs were launched in running this query.
0
tab_name
student
Welcome Home.
Firefox automatically sends some data to Mozilla so that we can improve your experience.
Choose What I Share
Hue - Beeswax (Hive ...

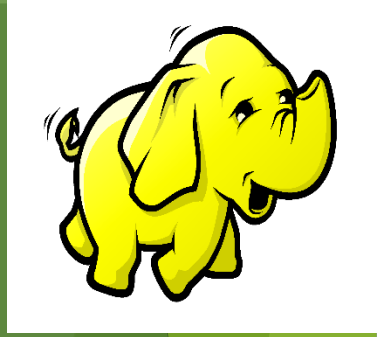
References



- http://training.cloudera.com/essentials.pdf
- http://en.wikipedia.org/wiki/Apache_Hadoop
- http://practicalanalytics.wordpress.com/2011/11/06/explaining-hadoop-to-management-whats-the-big-data-deal/
- https://developer.yahoo.com/hadoop/tutorial/module1.html
- http://hadoop.apache.org/
- http://wiki.apache.org/hadoop/FrontPage



Questions?



Thanks