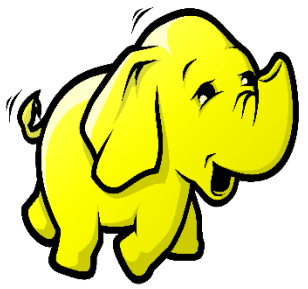


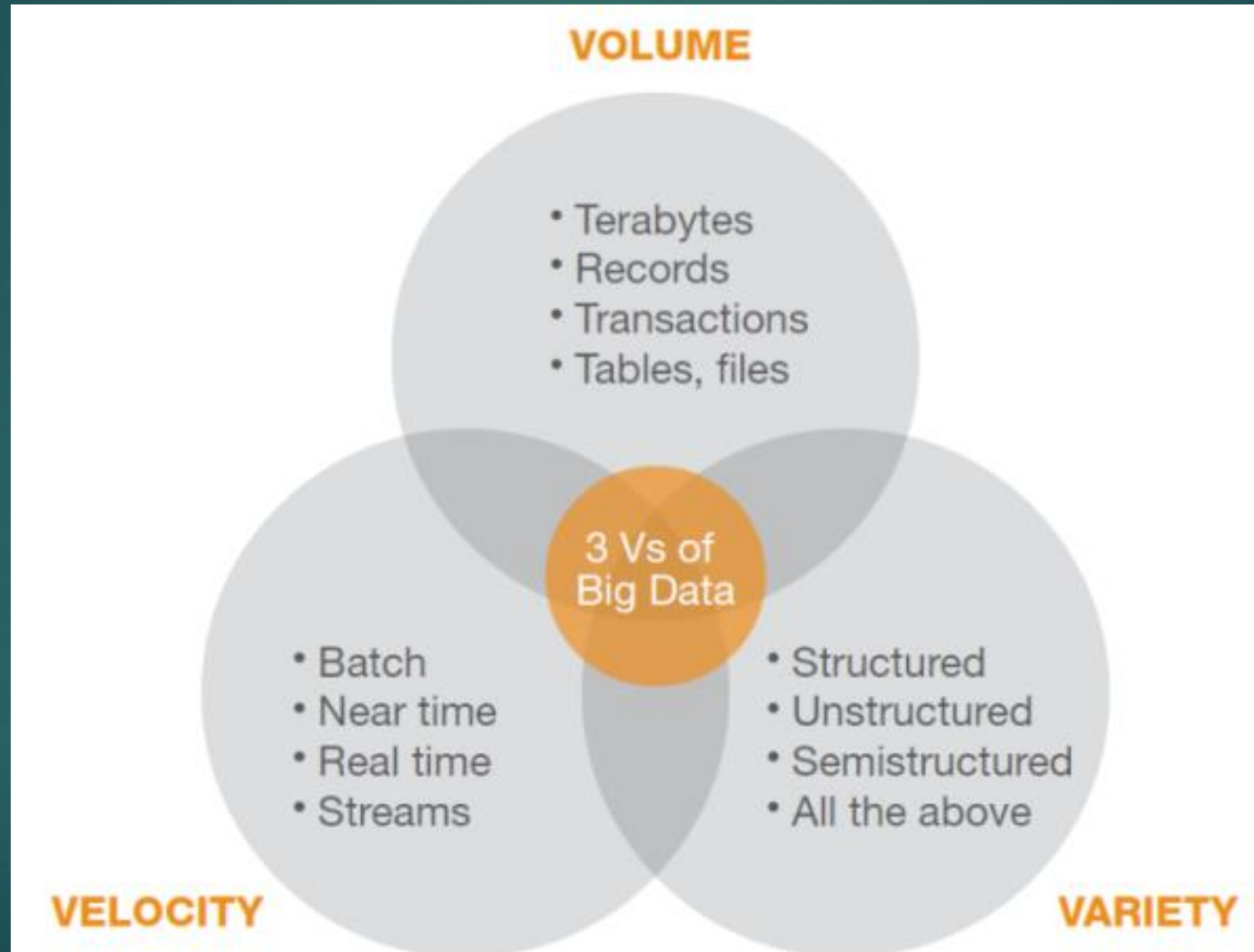
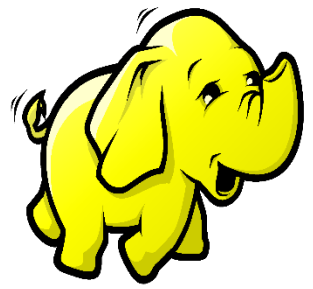
# The Hadoop Ecosystem

# Overview

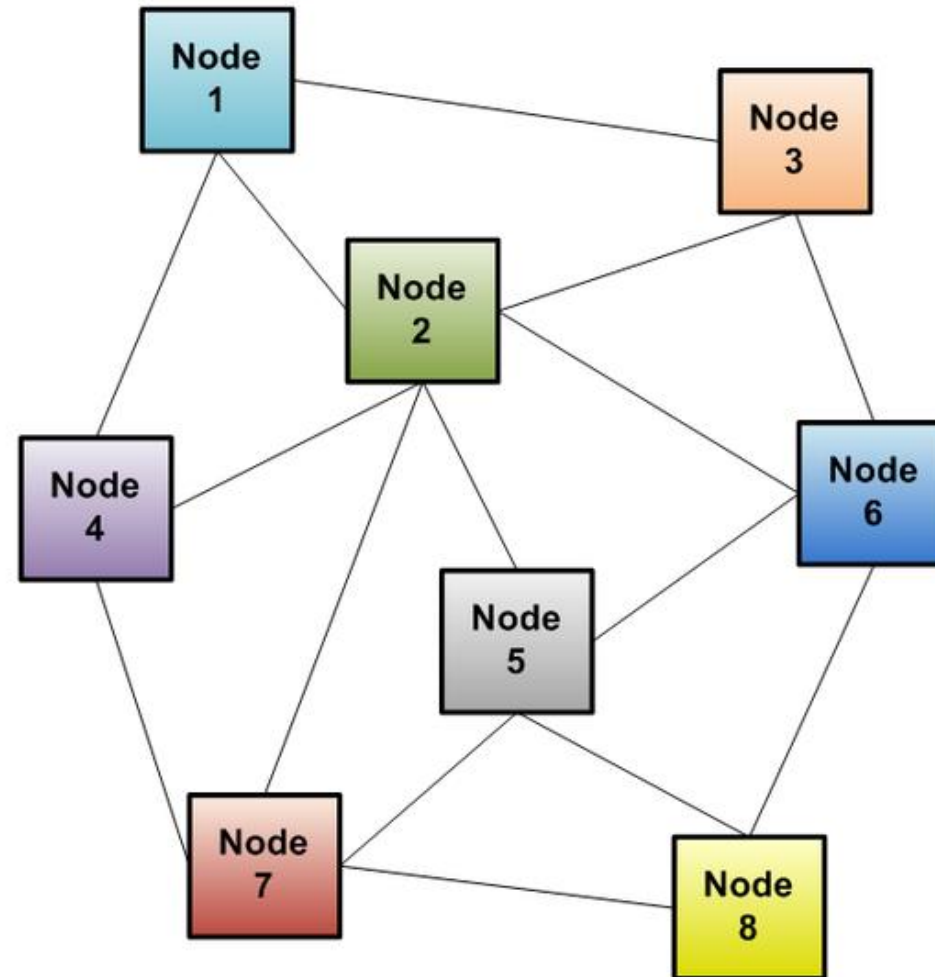
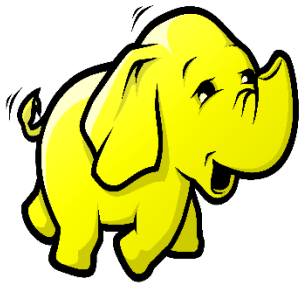
- ▶ Big Data Challenges
- ▶ Distributed system and challenges
- ▶ Hadoop Introduction
- ▶ History
- ▶ Who uses Hadoop
- ▶ The Hadoop Ecosystem
  - ❑ Hadoop core components
    - ❖ HDFS
    - ❖ Map Reduce
  - ❑ Other Hadoop ecosystem components
    - ❖ Hbase
    - ❖ Hive
    - ❖ Pig
    - ❖ Impala
  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



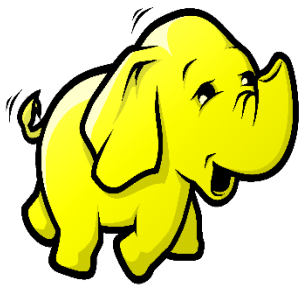
# Big Data Challenges



# Solution: Distributed system

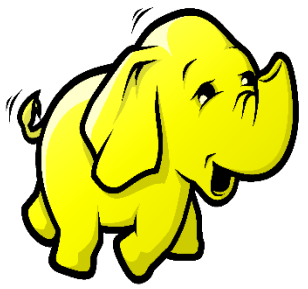


# Distributed System Challenges



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

# New Approach to distributed computing

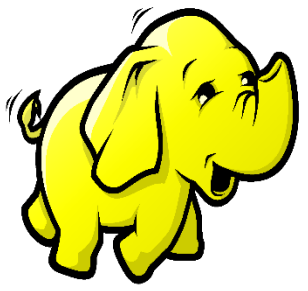


Hadoop:

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

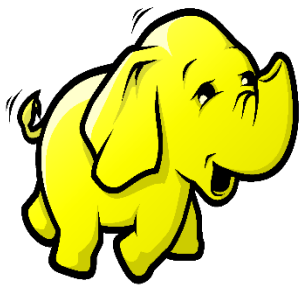
- ▶ Distribute data when the data is stored
- ▶ Process data where the data is
- ▶ Data is replicated

# Hadoop



- ▶ 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

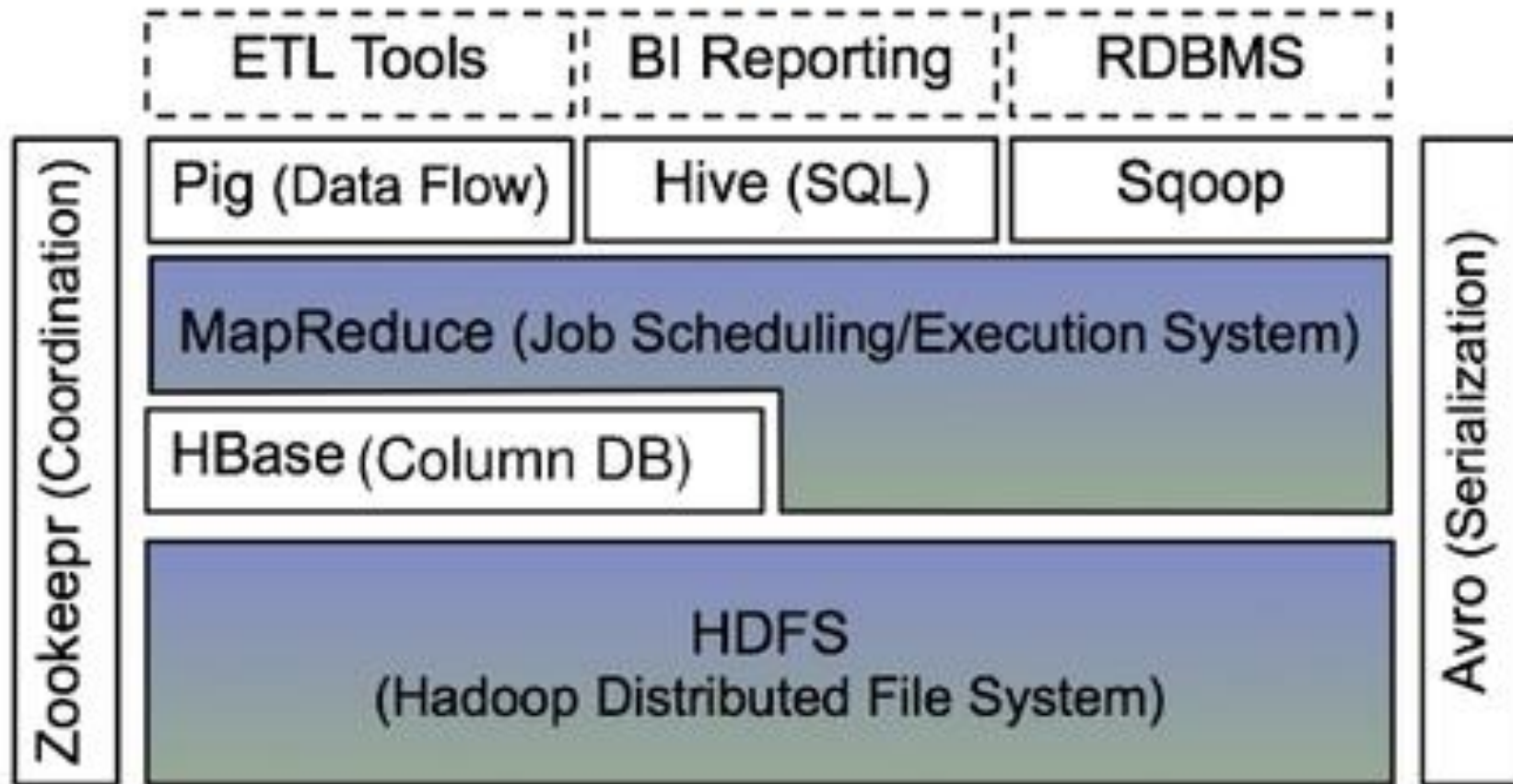
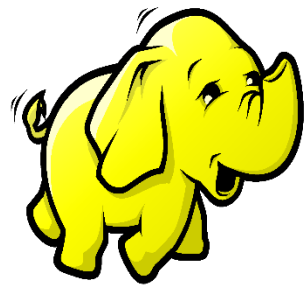
# The Hadoop Ecosystem



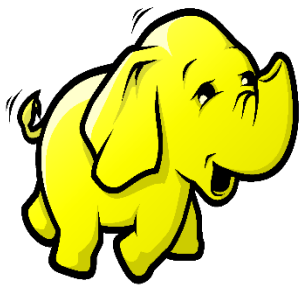
► <http://hadoopecosystemtable.github.io/>



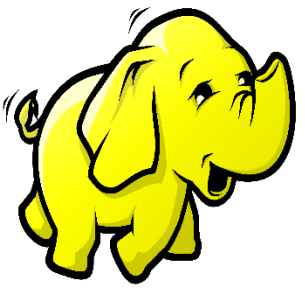
# The Hadoop Ecosystem



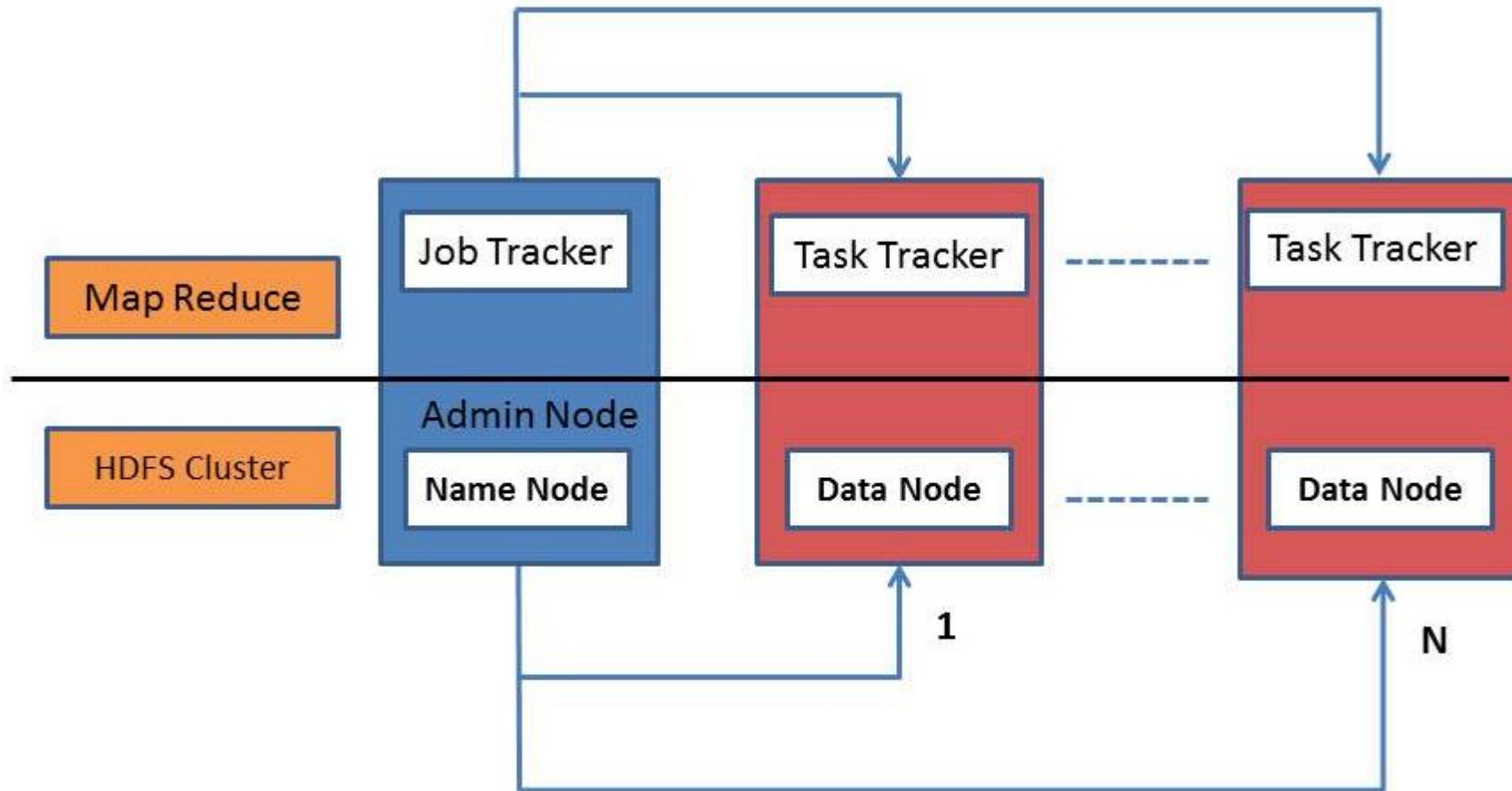
# Hadoop Core Components



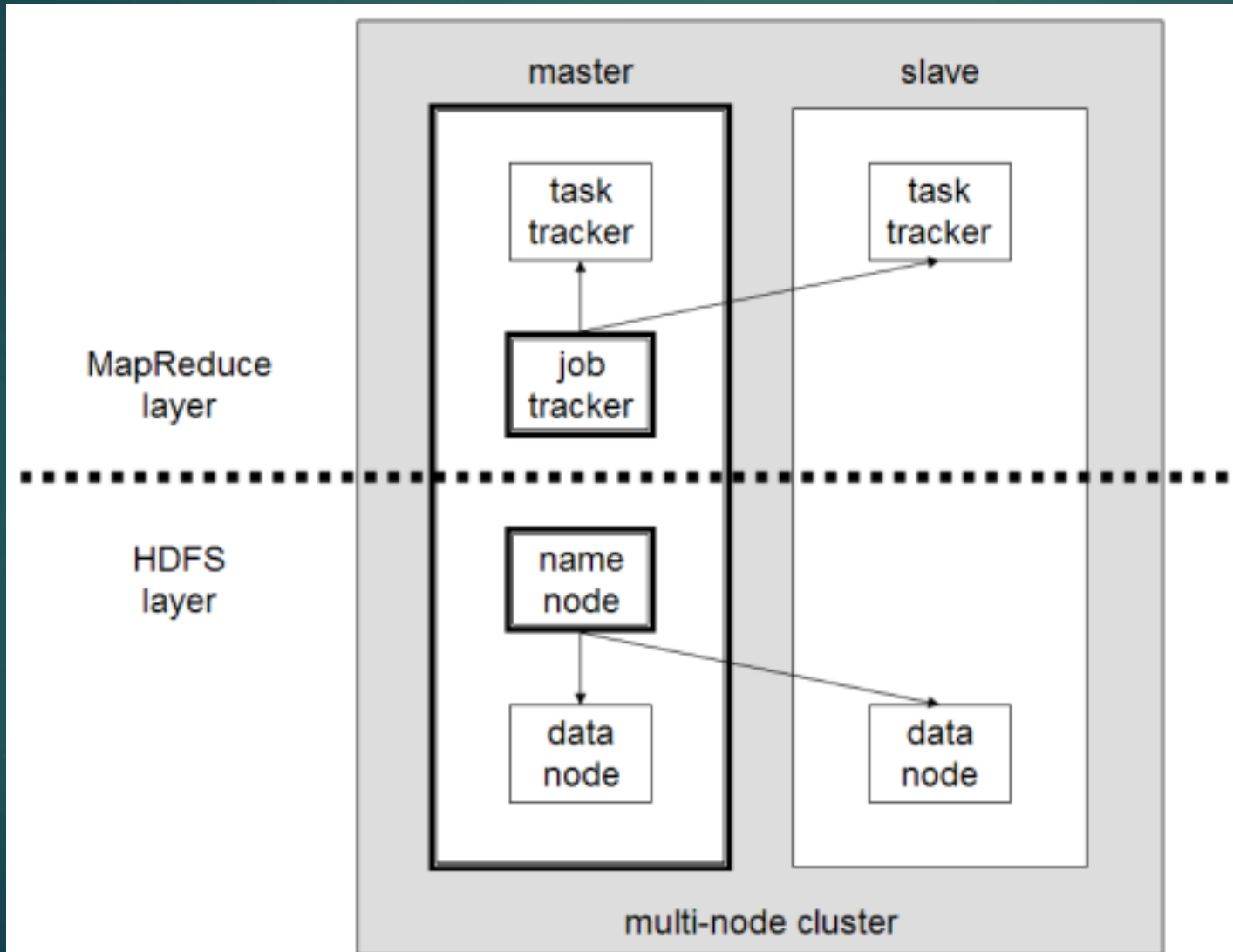
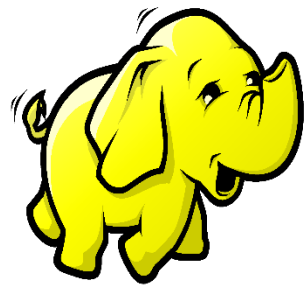
- ▶ HDFS – Hadoop Distributed File System (Storage)
- ▶ Map Reduce (Processing)



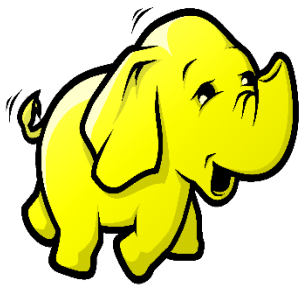
# Hadoop Core Components



# A multi-node Hadoop cluster

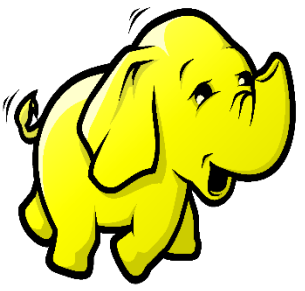


# Nodes



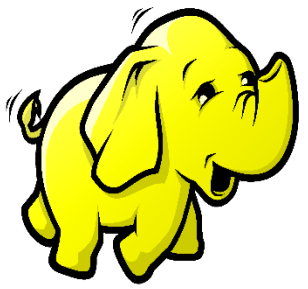
- ▶ **NameNode:**
  - ▶ Master of the system
  - ▶ Maintains and manages the blocks which are present on the DataNodes
- ▶ **DataNodes:**
  - ▶ Slaves which are deployed on each machine and provide the actual storage
  - ▶ Responsible for serving read and write requests for the clients
- ▶ **Jobtracker:**
  - ▶ 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:

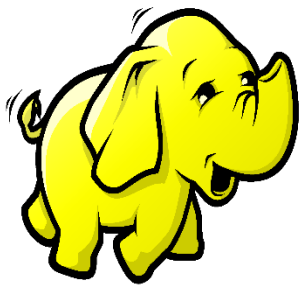
- ❑ Each block is processed in isolation by a map task called mapper
- ❑ Map task runs on the node where the block is stored

## ► The Reducer:

- ❑ Consolidate result from different mappers
- ❑ Produce final output



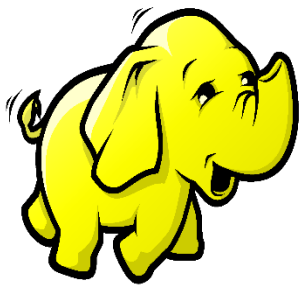
# What makes Hadoop unique



- ▶ 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

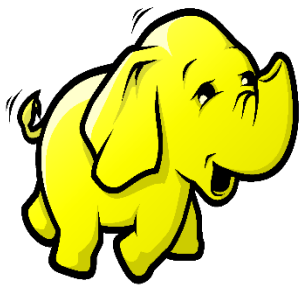


# Other Hadoop components in Ecosystem



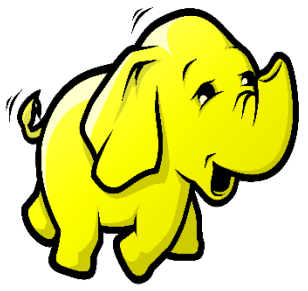
HBase	Hadoop database for random read/write access
Hive	SQL-like queries and tables on large datasets
Pig	Data flow language and compiler
Oozie	Workflow for interdependent Hadoop jobs
Sqoop	Integration of databases and data warehouses with Hadoop
Flume	Configurable streaming data collection
ZooKeeper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

# Hbase



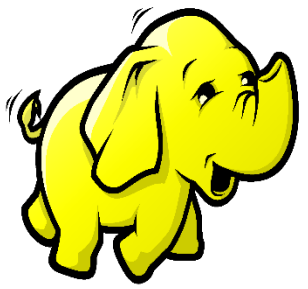
- ▶ HBase is an open source, non-relational, distributed database modeled after Google's BigTable.
- ▶ It runs on top of Hadoop and HDFS, providing BigTable-like capabilities for Hadoop.

# Features of Hbase



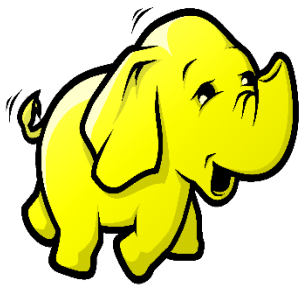
- ▶ Type of NoSql database
- ▶ Strongly consistent read and write
- ▶ Automatic sharding
- ▶ Automatic RegionServer failover
- ▶ Hadoop/HDFS Integration
- ▶ HBase supports massively parallelized processing via MapReduce for using HBase as both source and sink.
- ▶ HBase supports an easy to use Java API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

# Hbase in CAP theorem



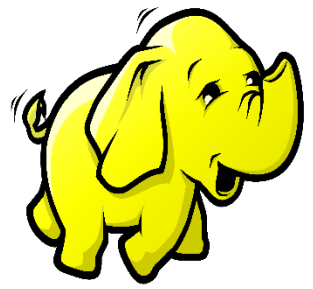
- ▶ Eric Brewer's CAP theorem, HBase is a CP type system.

# When to use Hbase



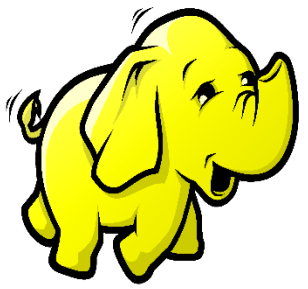
- ▶ When there is real big data: millions or billions of rows, in other way data can not store in a single node.
- ▶ When random read/write access to big data
- ▶ When require to do thousands of operations on big data
- ▶ When there is no need of extra features of RDMS like typed columns, secondary indexes,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

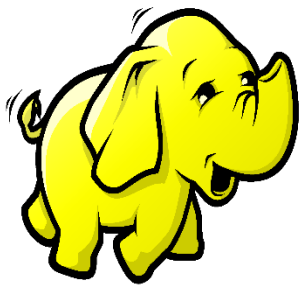
# Hive



- ▶ 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

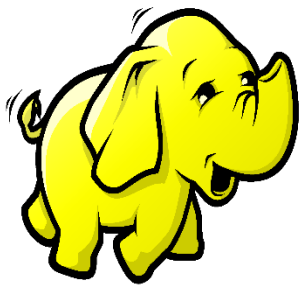


► HiveQL example:

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

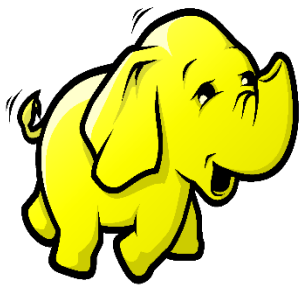


# Pig



- ▶ 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



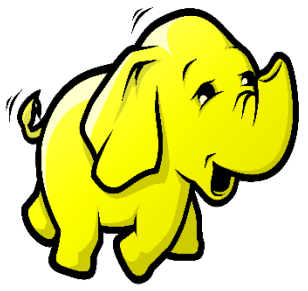
## ► Pig Latin example:

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

```
X = FOREACH A GENERATE name,$2;
```

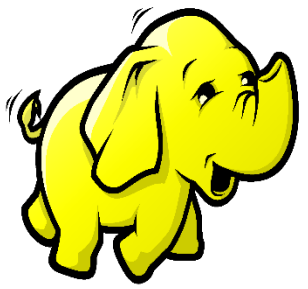
```
DUMP X;
```

# Pig and Hive



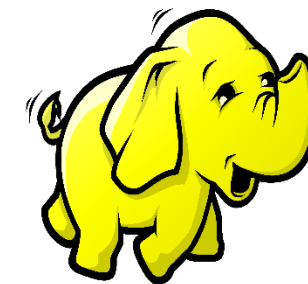
- ▶ Both requires compiler to generate Map reduce jobs
- ▶ Hence high latency queries when used for real time responses to ad-hoc queries
- ▶ Both are good for batch processing and ETL jobs
- ▶ Fault tolerant

# Impala



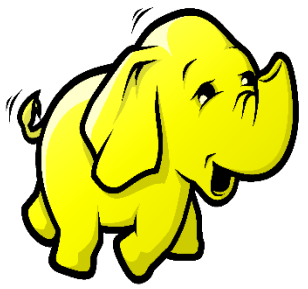
- ▶ Cloudera Impala is a query engine that runs on Apache Hadoop.
- ▶ Similar to HiveQL.
- ▶ Does not use Map reduce
- ▶ Optimized for low latency queries
- ▶ Open source apache project
- ▶ Developed by Cloudera
- ▶ Much faster than Hive or pig

# 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

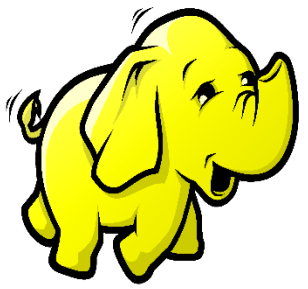
# Sqoop



- ▶ 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

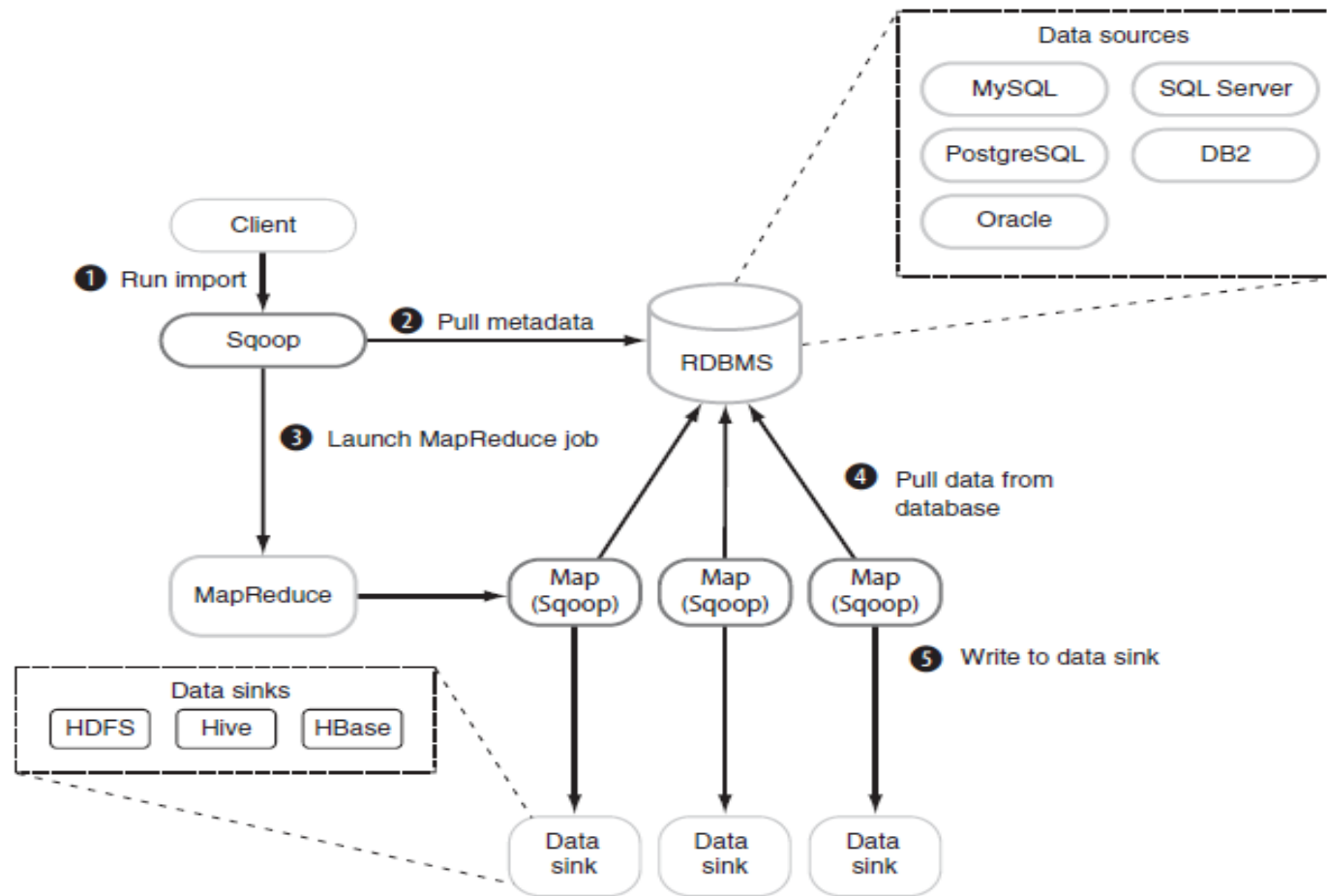
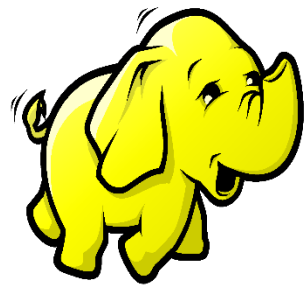


# How Sqoop works



- ▶ 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.

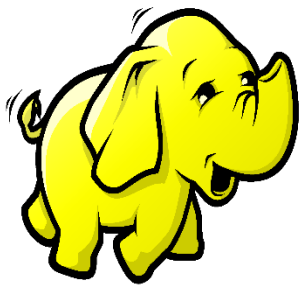
# How Sqoop works



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

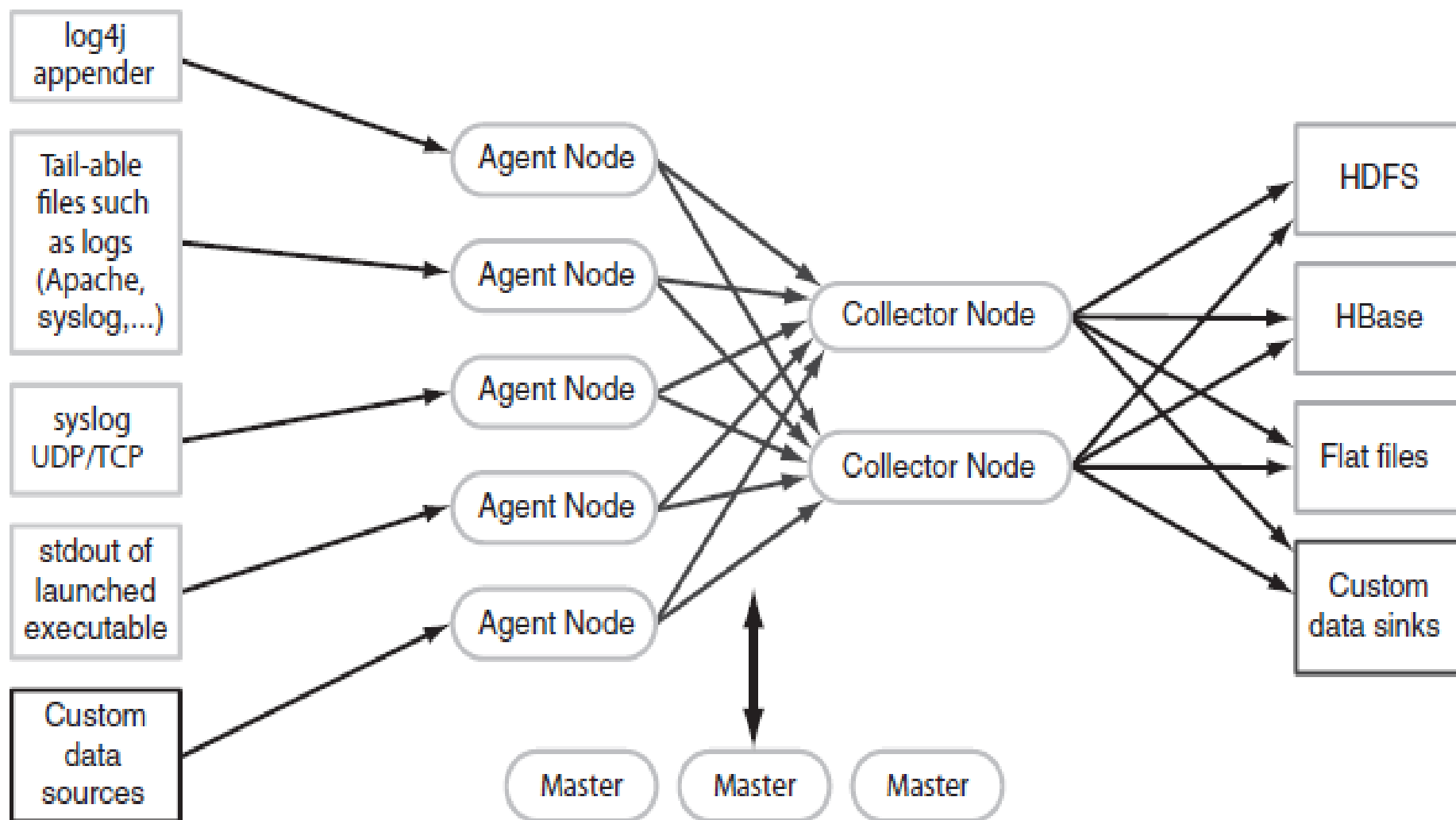
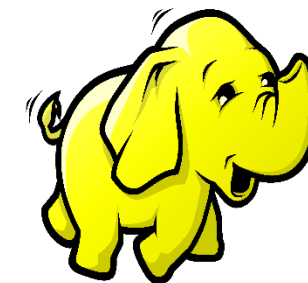


# Flume



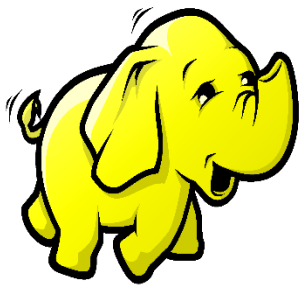
- ▶ Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS).

# How flume works



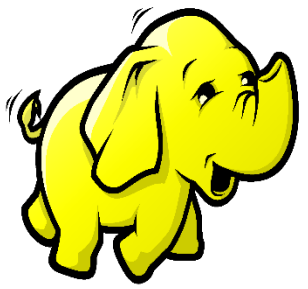
**Figure 2.2** Flume architecture for collecting streaming data

# How flume works



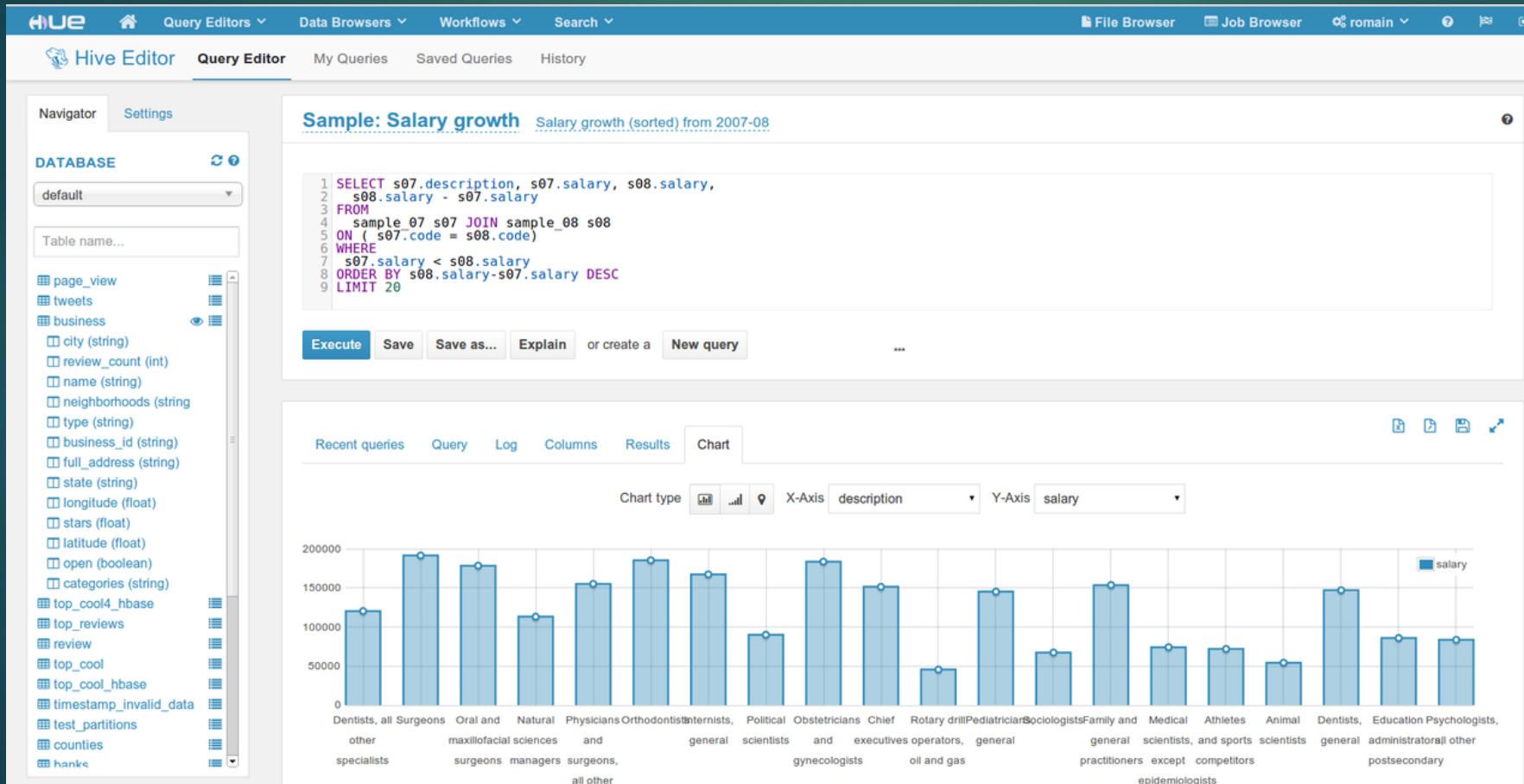
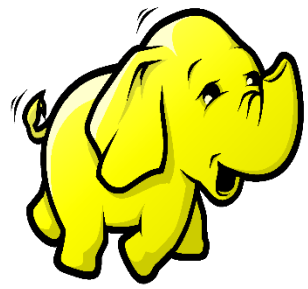
- ▶ Data flows like:  
Agent tier -> Collector tier -> Storage tier
- ▶ **Agent nodes** are typically installed on the machines that generate the logs and are data's initial point of contact with Flume. They forward data to the next tier of **collector nodes**, which aggregate the separate data flows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

# Hue

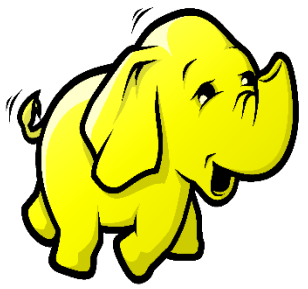


- ▶ Graphical front end to the cluster.
- ▶ Open source web interface.
- ▶ Makes Hadoop platform (HDFS, Map reduce, oozie, Hive, etc.) easy to use

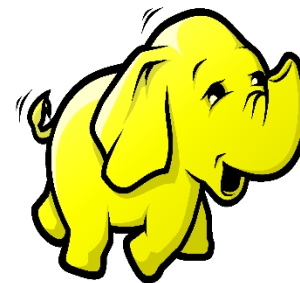
# Hue



# 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.

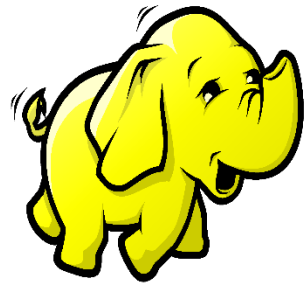


Lab

# Hadoop Installation (CDH ) for

▶ Download and install VM player

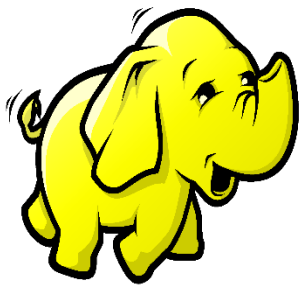
[https://my.vmware.com/web/vmware/free#desktop\\_end\\_user\\_computing/vmware\\_player/6\\_0](https://my.vmware.com/web/vmware/free#desktop_end_user_computing/vmware_player/6_0)



The screenshot shows the VMware Player download page in a web browser. The browser's address bar displays the URL: [https://my.vmware.com/web/vmware/free#desktop\\_end\\_user\\_computing/vmware\\_player/6\\_0](https://my.vmware.com/web/vmware/free#desktop_end_user_computing/vmware_player/6_0). The page features the VMware logo and navigation links for Products, Support, Downloads, Consulting, Partner Programs, and Company. A breadcrumb trail indicates the path: Home > All Downloads > VMware Player. The main heading is "Download VMware Player". Below this, there are dropdown menus for "Major Version: 6.0 (latest)" and "Minor Version: 6.0.3 (latest)". There are three tabs: "Product Downloads", "Drivers & Tools", and "Open Source". The "Product Downloads" tab is active, showing three download options: "VMware Player for Windows (exe | 94 MB)", "VMware Player for Linux 32-bit (bundle | 222 MB)", and "VMware Player for Linux 64-bit (bundle | 191 MB)". Each option has a "Download" button with a downward arrow. To the right of the download options, there is a section titled "About This Product" with sub-sections for "DESCRIPTION" (VMware Player 6.0.3), "DOCUMENTATION" (Release Notes), and "NOTES" (Buy Player 6 Plus today and get Player 7 Pro for free in December. VMware Player and VMware Player Plus. Enter a license key into the VMware Player user interface to enable the VMware Player Plus features.). The Windows taskbar is visible at the bottom of the screen, showing icons for various applications and the system clock indicating 3:35 PM on 10/19/2014.



# Hadoop Installation (CDH ) for windows

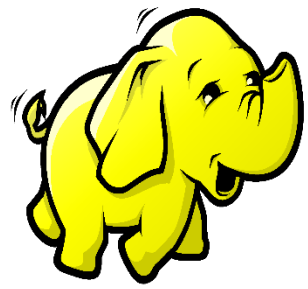


Make sure you have enabled  
virtualization in bios

# 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.html](http://www.cloudera.com/content/cloudera/en/downloads/quickstart_vms/cdh-4-7-x.html)



www.cloudera.com/content/cloudera/en/downloads/quickstart\_vms/cdh-4-7-x.html

contain a single-node Apache Hadoop cluster, complete with example data, queries, scripts, and Cloudera Manager to manage your cluster.

The VMs run CentOS 6.2 and are available for VMware, VirtualBox, and KVM.

All require a 64-bit host OS.

Version: Quick Start VM with CDH 4.7.x

Please Note: Cloudera QuickStart VMs are for demo purposes only and are not to be used as a starting point for clusters.

Welcome to the New Virtual Machine Wizard!

This wizard will guide you through the steps that are necessary to create a new virtual machine for VirtualBox.

Use the **Continue** button to go to the next page of the wizard and the **Go Back** button to return to the previous page. You can also press **Esc** if you want to cancel the execution of this wizard.

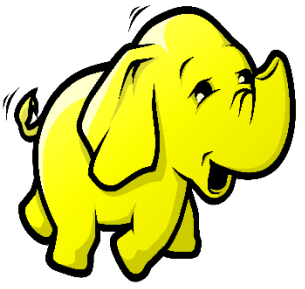
Go Back Continue

Download System Requirements Installed Products Getting Started

Platform	Release Date	Hash	Bits
VMWare	June 2014	SHA1: 626595ef7e445ecfb80280b9353340e0e049807c	<a href="#">Download for VMWare &gt;</a>
KVM	June 2014	SHA1: 72fc0980c5ab200e4877b2bf0631b0b022042616	<a href="#">Download for KVM &gt;</a>
VirtualBox	June 2014	SHA1: 0b12484778925d58a02ea442a92b00d38ef74e04	<a href="#">Download for VirtualBox &gt;</a>

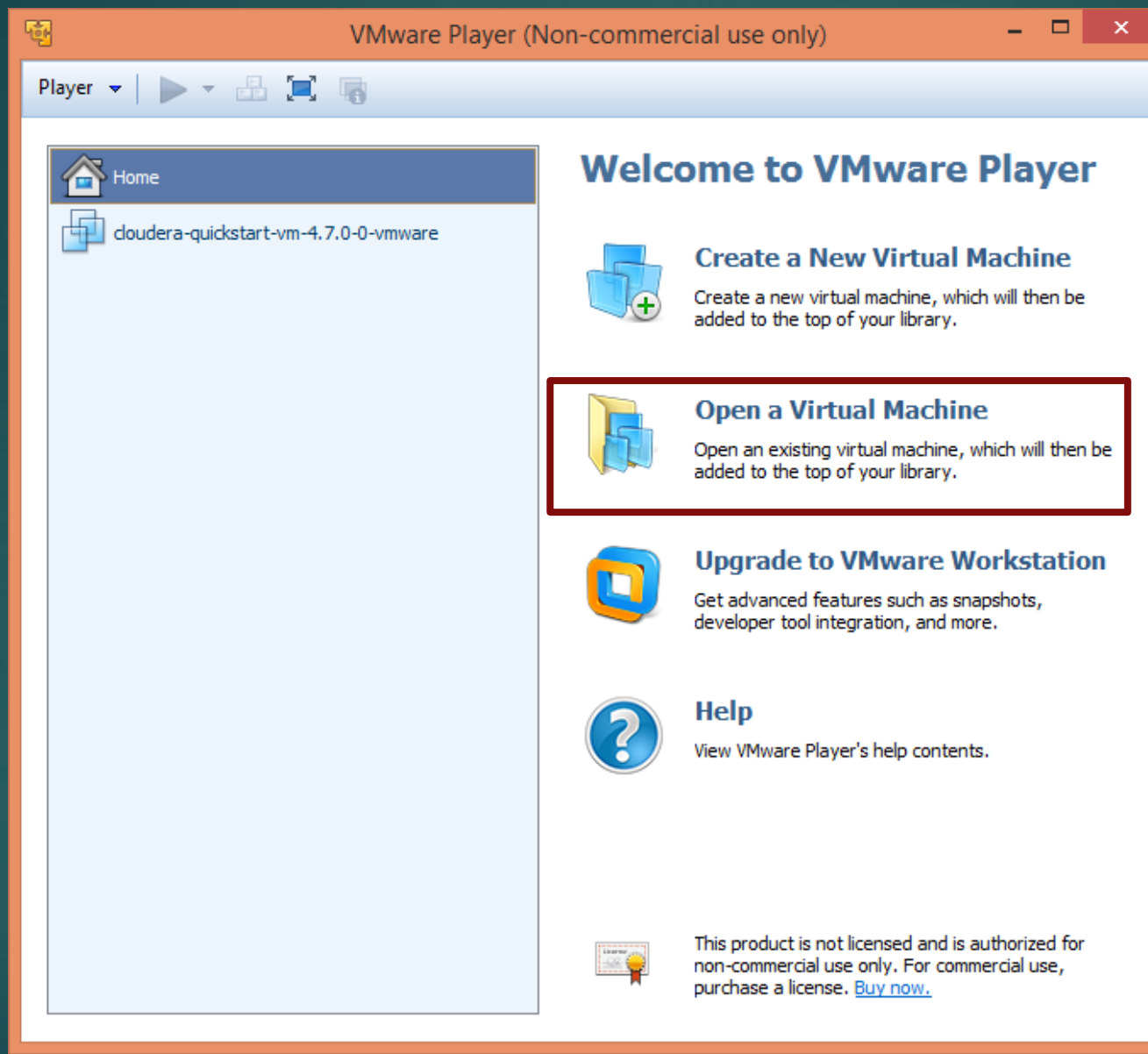
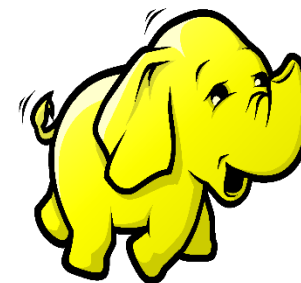
presentation---java-....pdf HbaseHivePig.pptx Hortonworks-Apac....pptx Adobe\_Reader\_Setup....exe Adobe\_Reader\_Setup.exe Show all downloads...

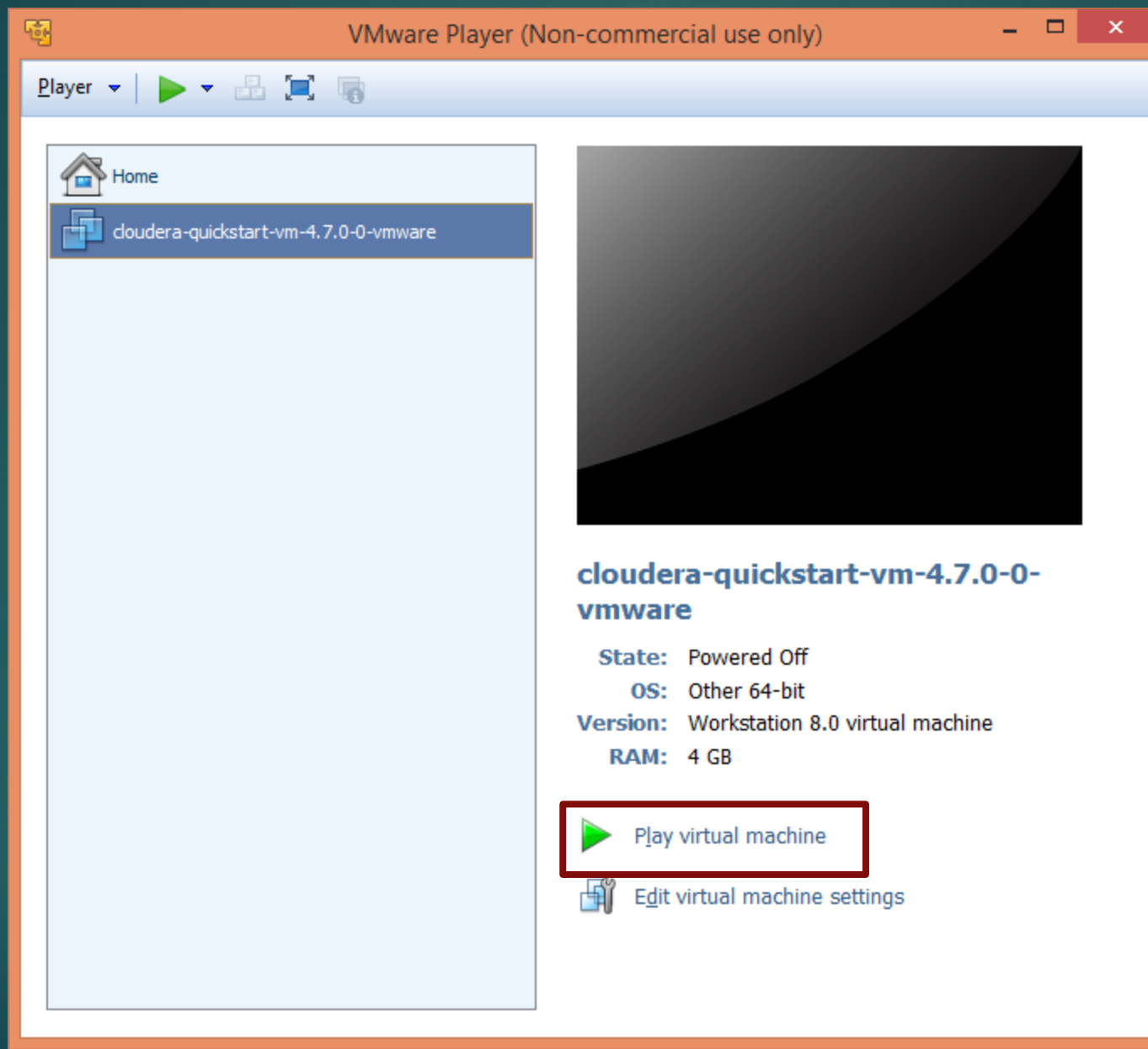
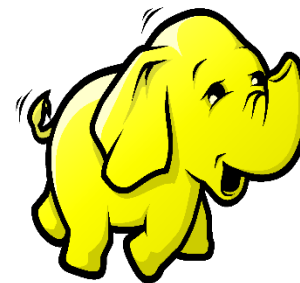
5:47 PM 10/20/2014



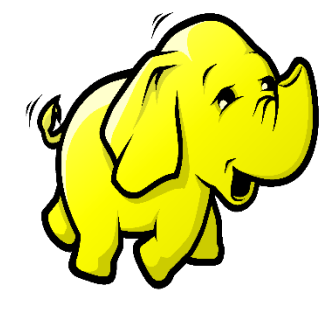
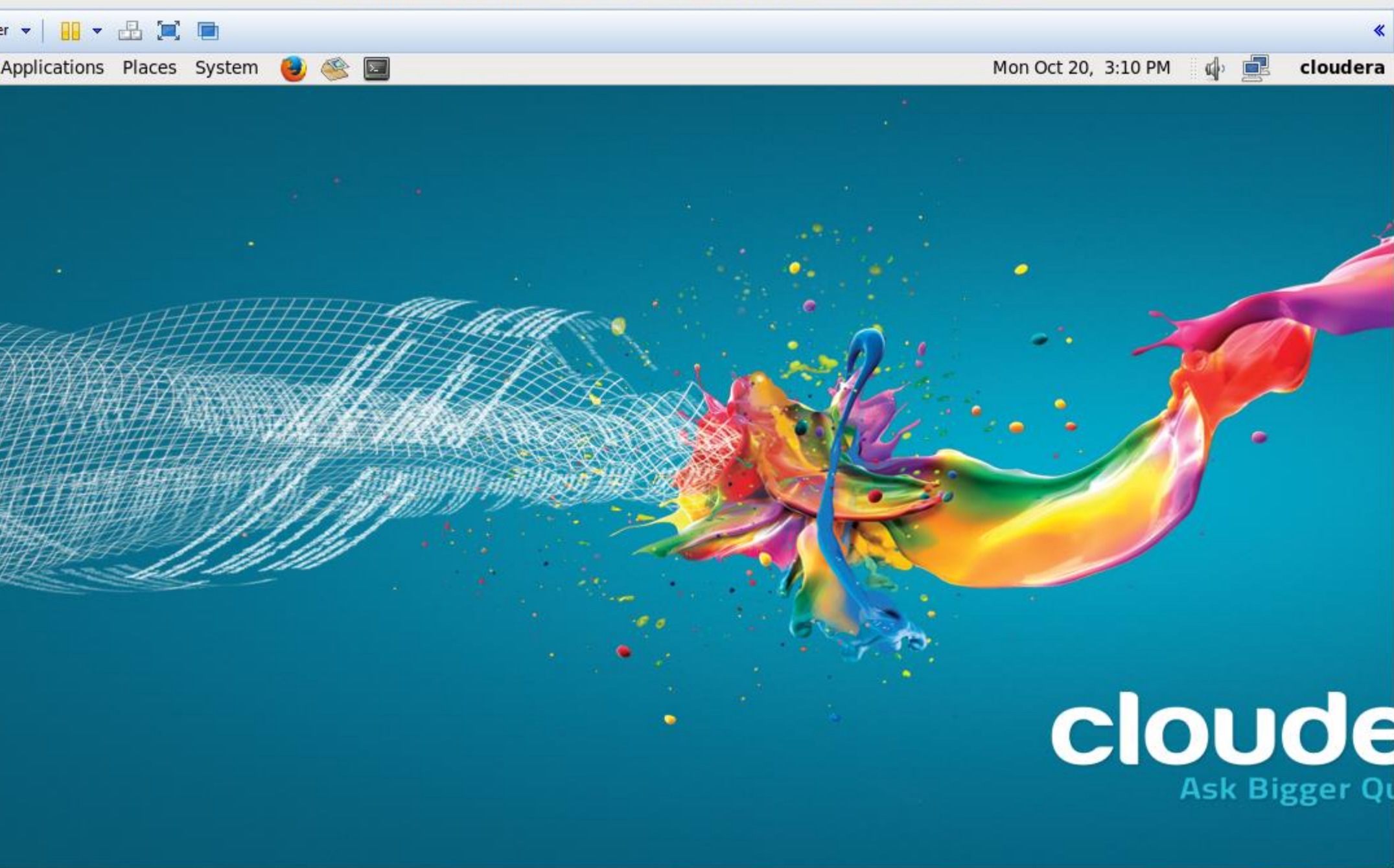
# Hadoop Installation (CDH ) for windows

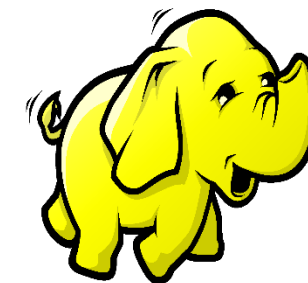
- ▶ Unzip “cloudera-quickstart-vm-4.7.0-0-vmware”
- ▶ Open CDH using VMPlayer:
  - ❑ Open VM Player
  - ❑ Click open a virtual machine
  - ❑ Select the file “cloudera-quickstart-vm-4.7.0-0-vmware” in the extracted directory of “cloudera-quickstart-vm-4.7.0-0-vmware”. Virtual machine will be added to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











cloudera-quickstart-vm-4.7.0-0-vmware - VMware Player (Non-commercial use only)

Player ▾ | [Icons] | Applications Places System [Icons] Mon Oct 20, 3:29 PM [Icons] 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:~

[Icons]



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

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
         */

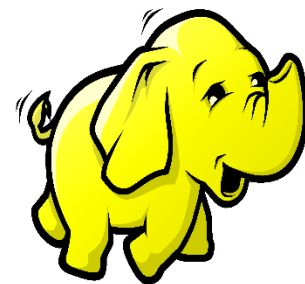
    }
}
```

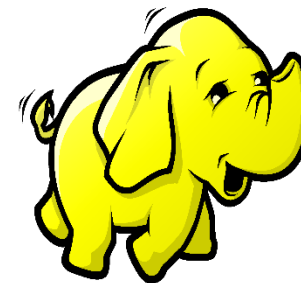
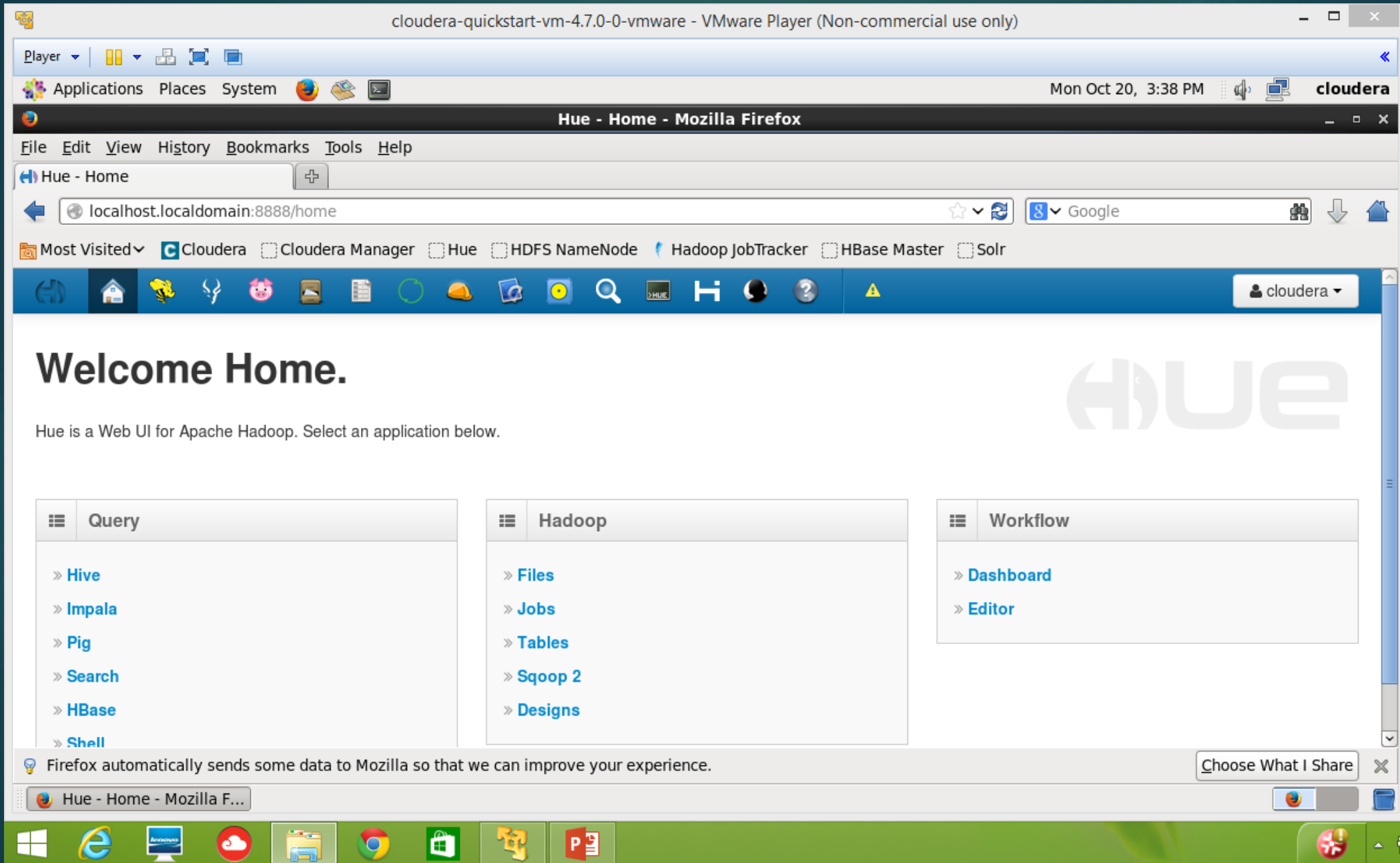
Problems @ Javadoc Declaration Console

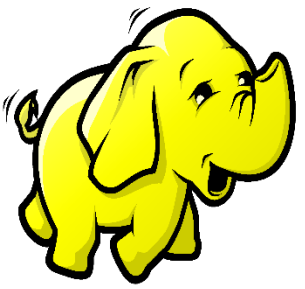
No consoles to display at this time.

StubMapper.java - training/src

Java - training/src/Stub...







cloudera-quickstart-vm-4.7.0-0-vmware - VMware Player (Non-commercial use only)

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

localhost.localdomain:8888/beeswax/results/18/0?context=design%3A18

Most Visited Cloudera Cloudera Manager Hue HDFS NameNode Hadoop JobTracker HBase Master Solr

Available Files

Query Editor: My Queries Saved Queries History Settings

Hadoop Ecosystem.pptx ...  
Version created from the l...

# Query Results: Unsaved Query

Downloads

- Download as CSV
- Download as XLS
- Save

MR JOBS

No Hadoop jobs were launched in running this query.

Welcome Home.

Query

Results Query Log Columns

tab\_name

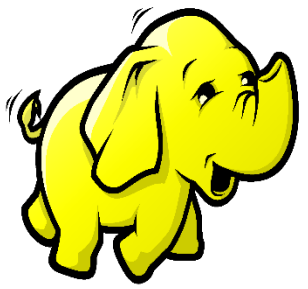
student

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://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>