Hadoop Setup

Jay Urbain, PhD

Setting Up Hadoop

- Hadoop binaries or vendor distribution
  - Use vendor distribution, check packages
- Hadoop version
  - Usually latest stable release ~2.7
- Location
  - Local cluster install, i.e., install from scratch
  - Free, but complicated unless standalone mode
- Local VM
  - Must install virtualization software, or Docker
- Cloud
  - Costs money to test
  - Need to remember to turn off
  - Great for episodic compute needs – IMHO what most companies should do.

Data Storage

- Local
  - File system (single)
- HDFS
  - pseudo or distributed
- Cloud
  - HDFS
  - S3 AWS/ BLOB Azure

- Note: Developing MapReduce algorithms are *complicated*. Keep things simple.

Libraries

- MapReduce
  - Version 1.0 or 2.0 (YARN)
- Other libraries
  - Hive and Pig are common, additional tools vary by vendor

Cloudera

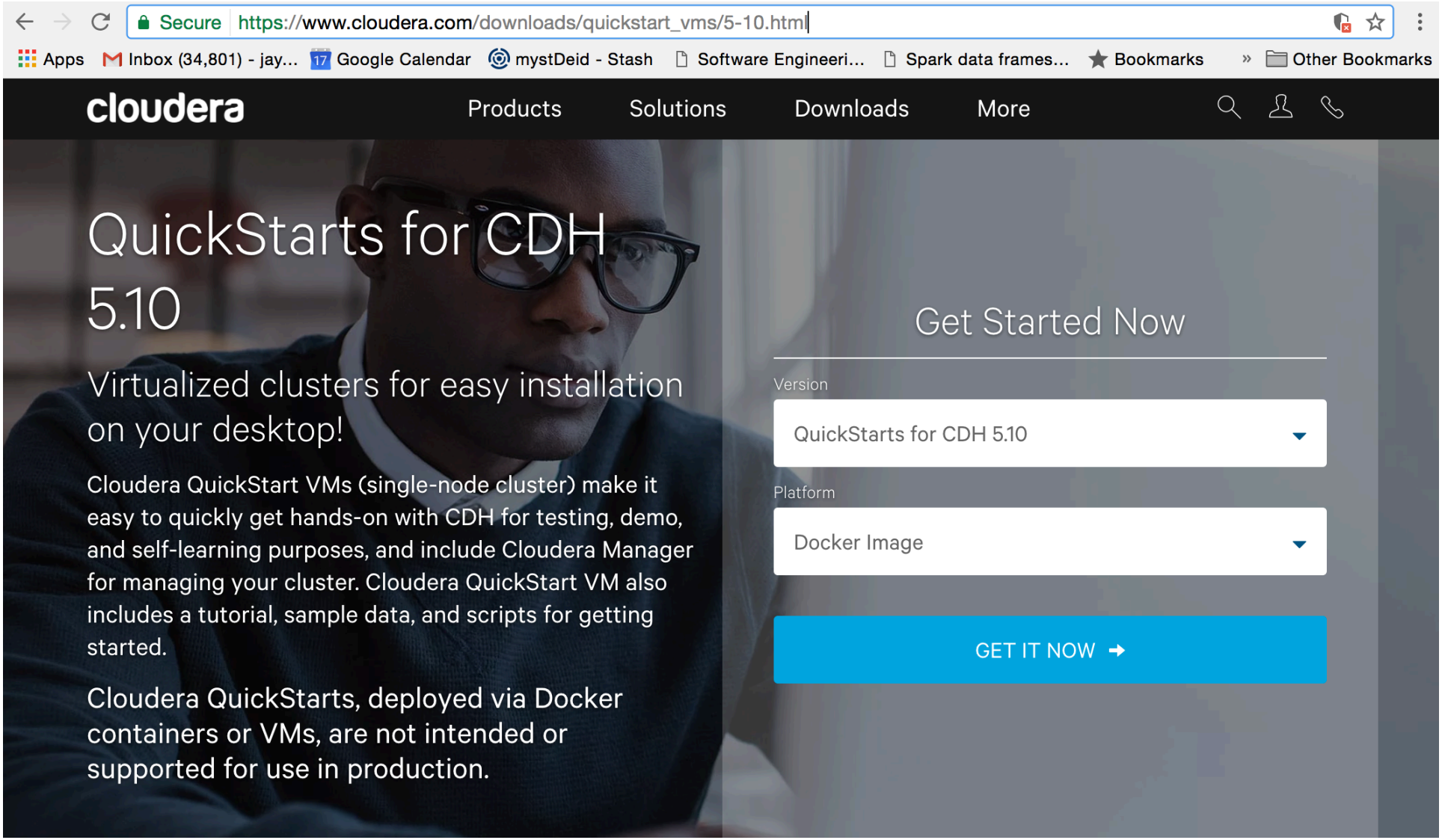
- OS + virtualization software
- Download Hadoop virtual machine image for that virtualization software
- Adding core libraries
  - Hive – SQL-like query, create batch (MapReduce) queries
  - Pig – ETL-like language
  - Impala – like Hive but real-time versus batch
  - Mahout – machine learning algorithms
  - Spark – in-memory processing of resilient distributed data set
  - Storm – complex event processing

MapReduce Programming Language

- Programming Languages
  - **Java**
  - Python (streaming)
  - C++ (pipes)
  - R

- Hadoop IDEs
  - **Eclipse (Java)**
  - Python (Sublime, PyCharm)
  - R Studio

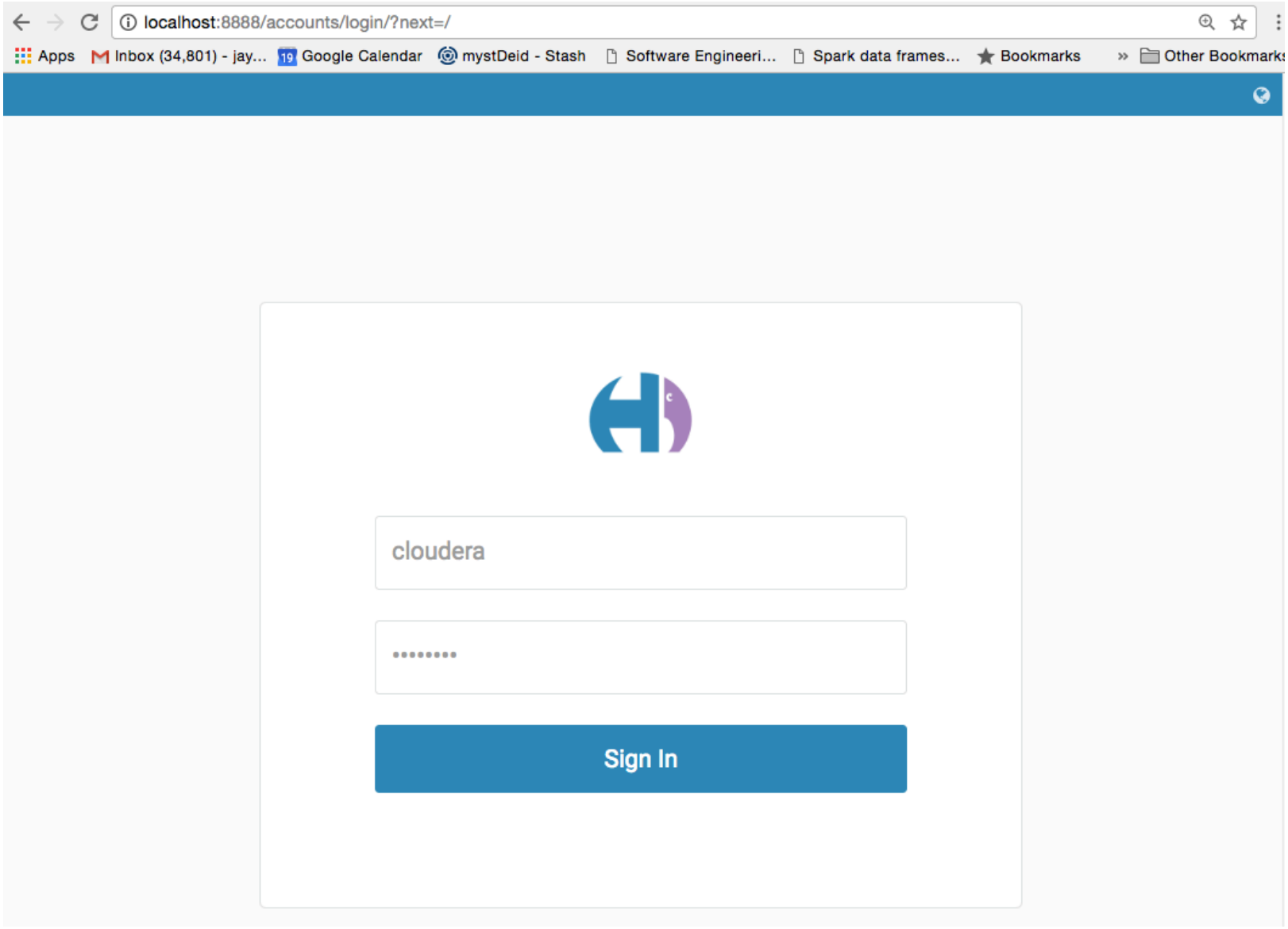
Cloudera Quickstart



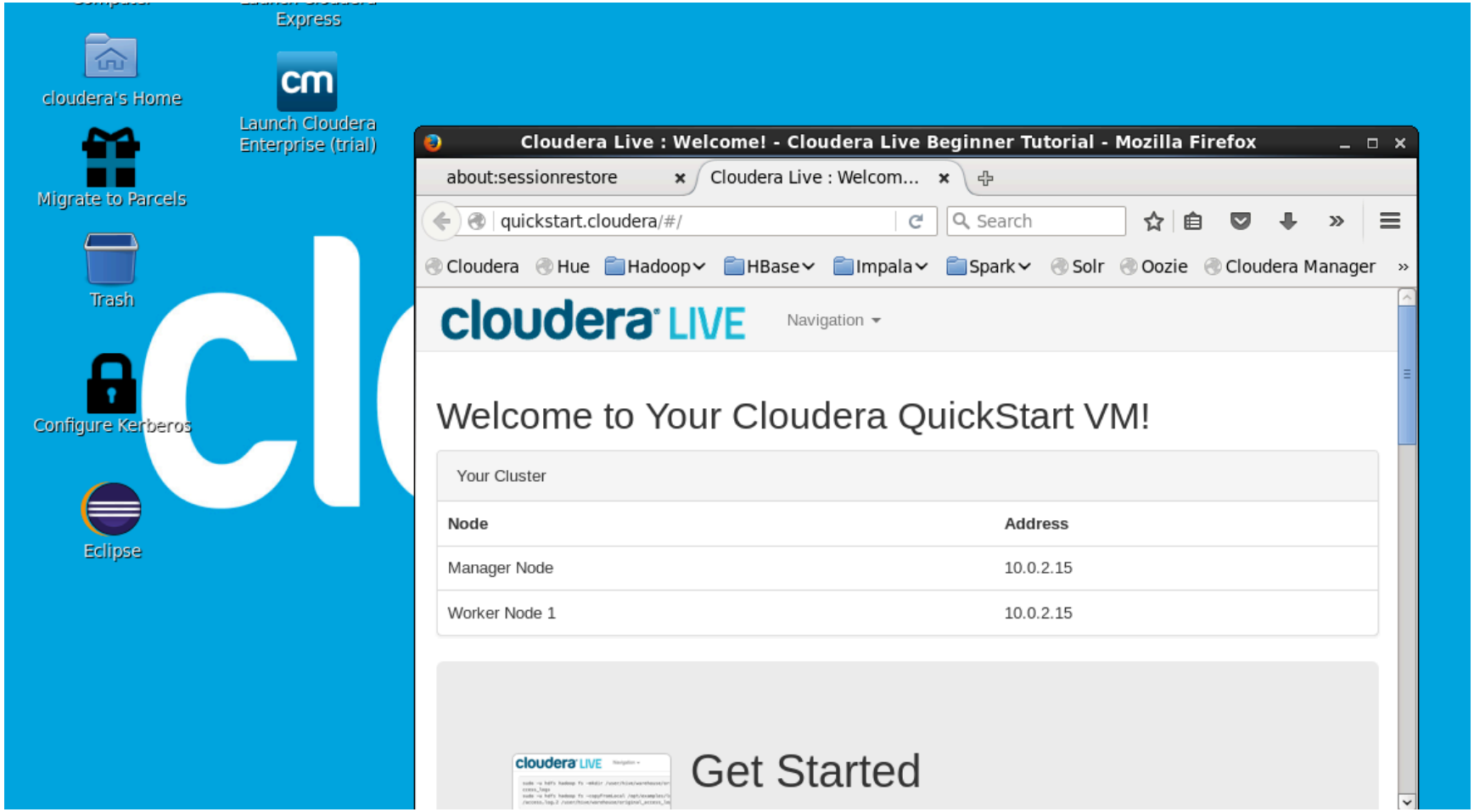
Hands on - Cloudera VM Walkthrough

- Which of the following Hadoop libraries gives you the ability to have SQL-like, or HQL query and creates batch MapReduce jobs with that query?
  - Impala
  - Mahout
  - Pig
  - Hive

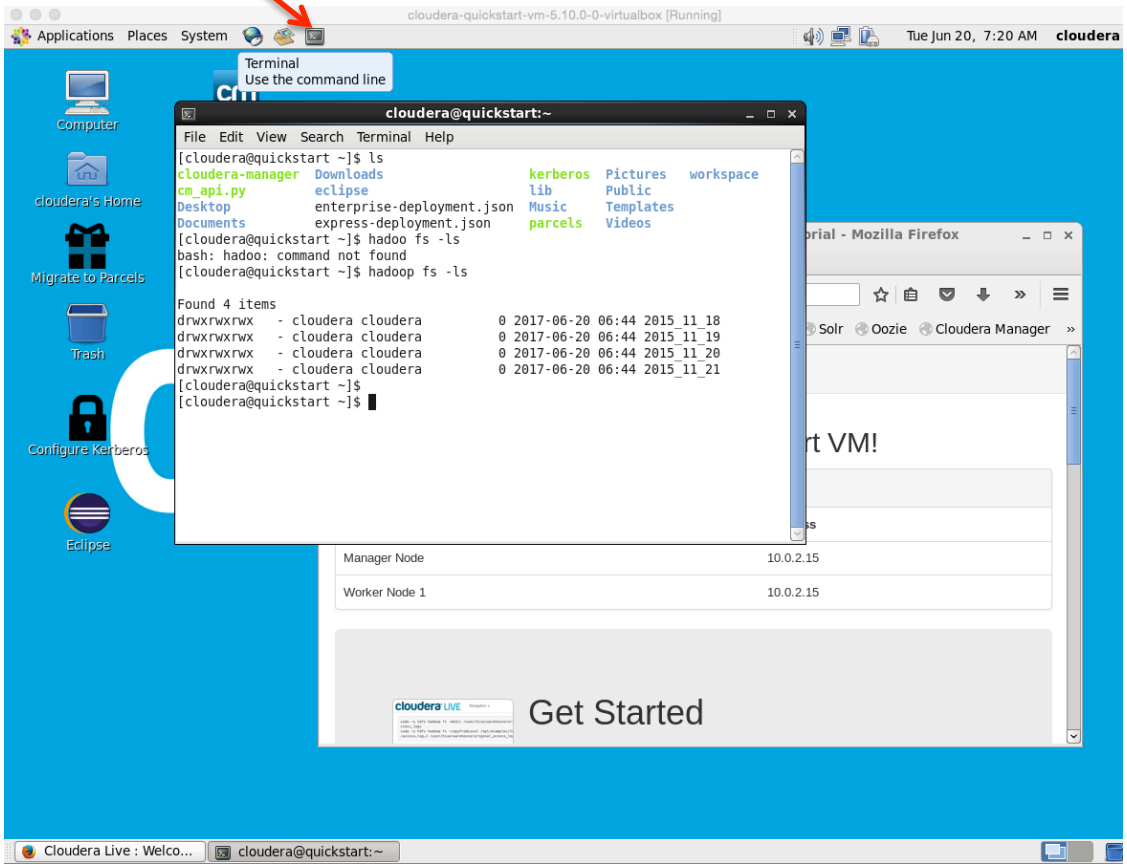
cloudera/cloudera



Startup Cloudera VM



Access the terminal

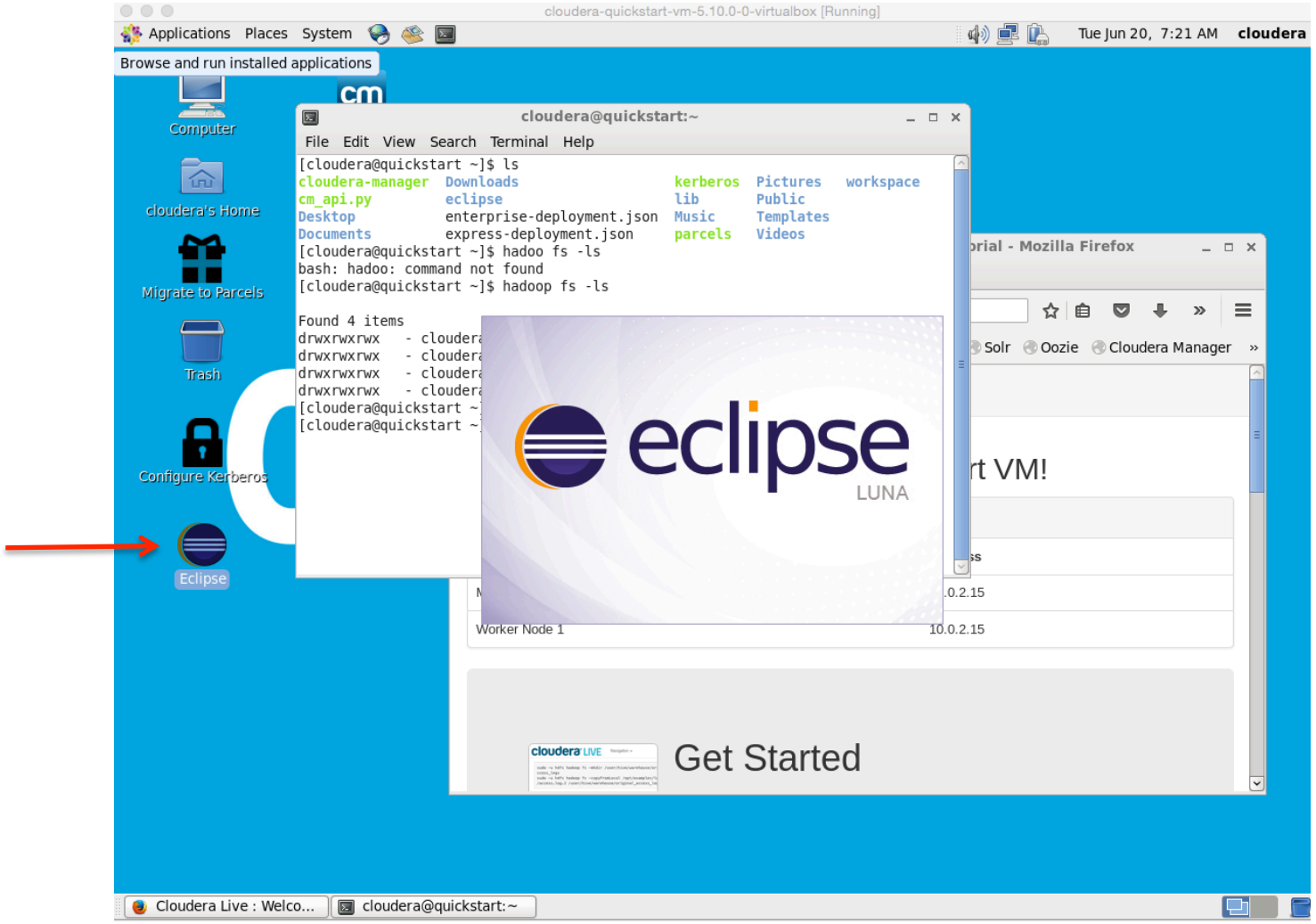


You have the **Auto capture keyboard** option turned on. This will cause the Virtual Machine to automatically **capture** the keyboard every time the VM window is activated and make it unavailable to other applications running on your host machine: when the keyboard is captured, all keystrokes (including system ones like Alt-Tab) will be directed to the VM.

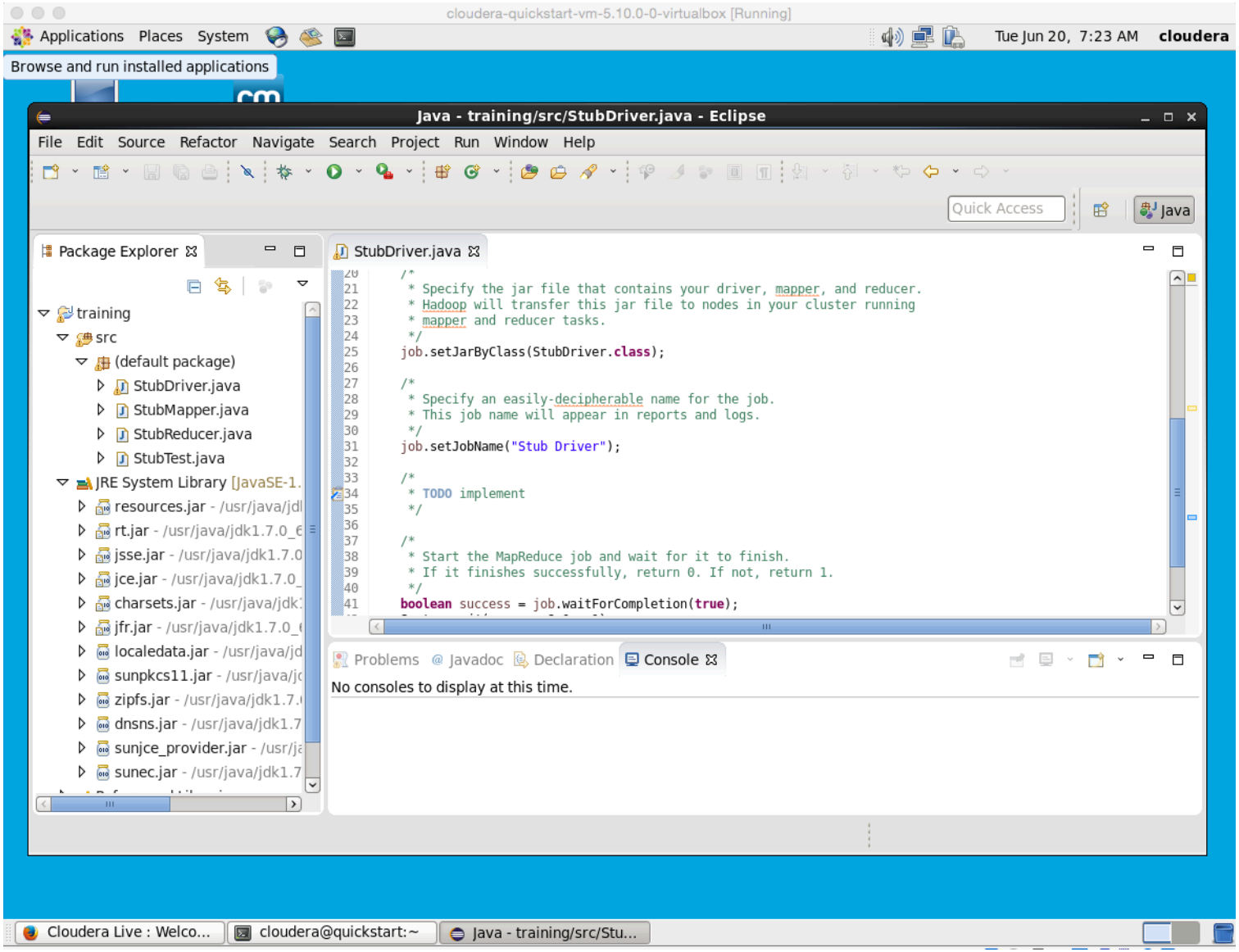
You can press the **host key** at any time to **uncapture** the keyboard and mouse (if it is captured) and return them to normal operation. The currently assigned host key is shown on the status bar at the bottom of the Virtual Machine window, next to the icon. This icon, together with the mouse icon placed nearby, indicate the current keyboard and mouse capture state.

The host key is currently defined as **Left ⌘**.

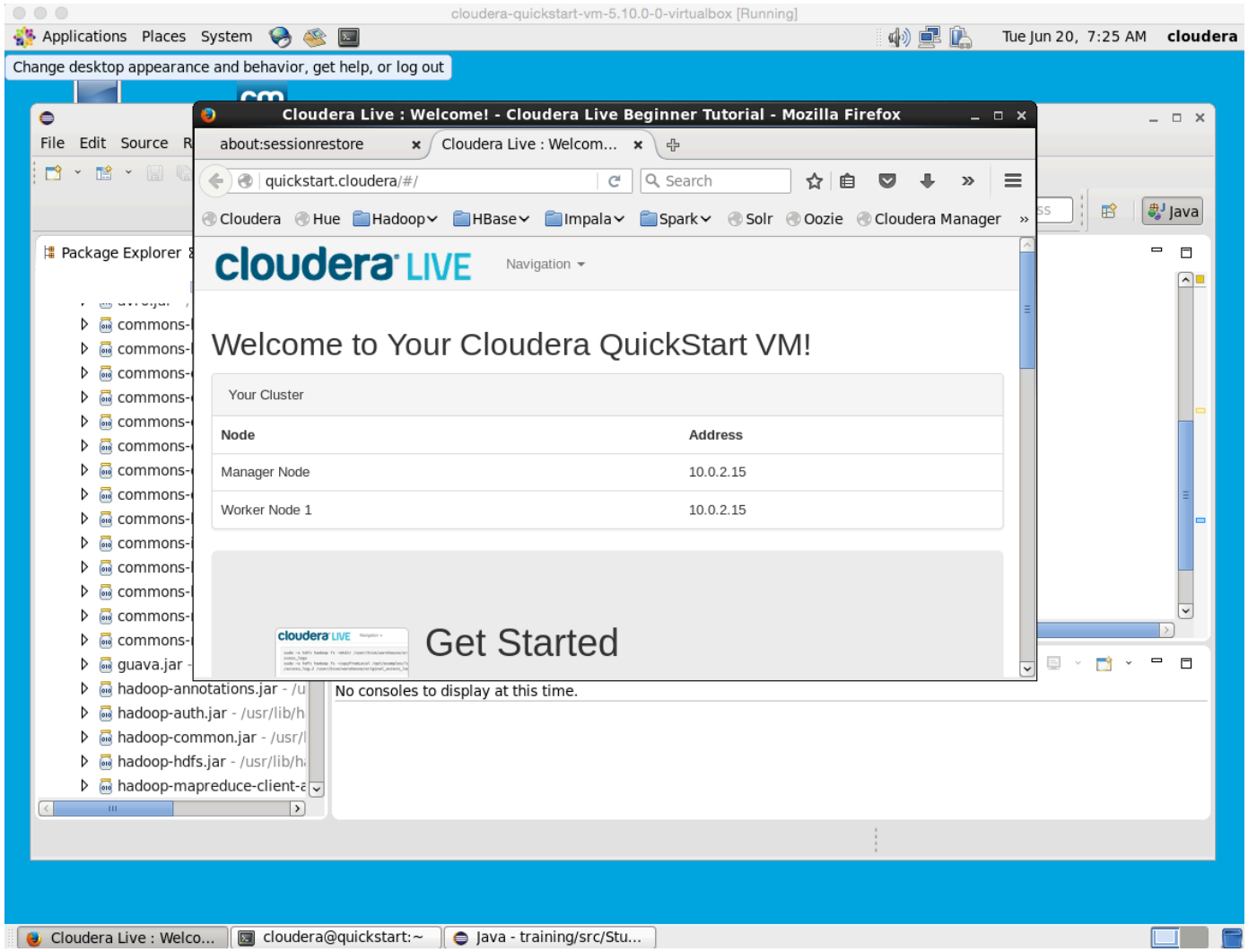
Eclipse


cloudera-quickstart-vm-5.10.0-0-virtualbox [Running]
Applications Places System
Tue Jun 20, 7:21 AM cloudera
Browse and run installed applications
Computer
cloudera's Home
Migrate to Parcels
Trash
Configure Kerberos
Eclipse
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$ ls
cloudera-manager Downloads kerberos Pictures workspace
cm_api.py eclipse lib Public
Desktop enterprise-deployment.json Music Templates
Documents express-deployment.json parcels Videos
[cloudera@quickstart ~]$ hadoo fs -ls
bash: hadoo: command not found
[cloudera@quickstart ~]$ hadoop fs -ls
Found 4 items
eclipse
LUNA
Solr
Oozie
Cloudera Manager
Worker Node 1
10.0.2.15
Get Started
Cloudera Live : Welco...
cloudera@quickstart:~

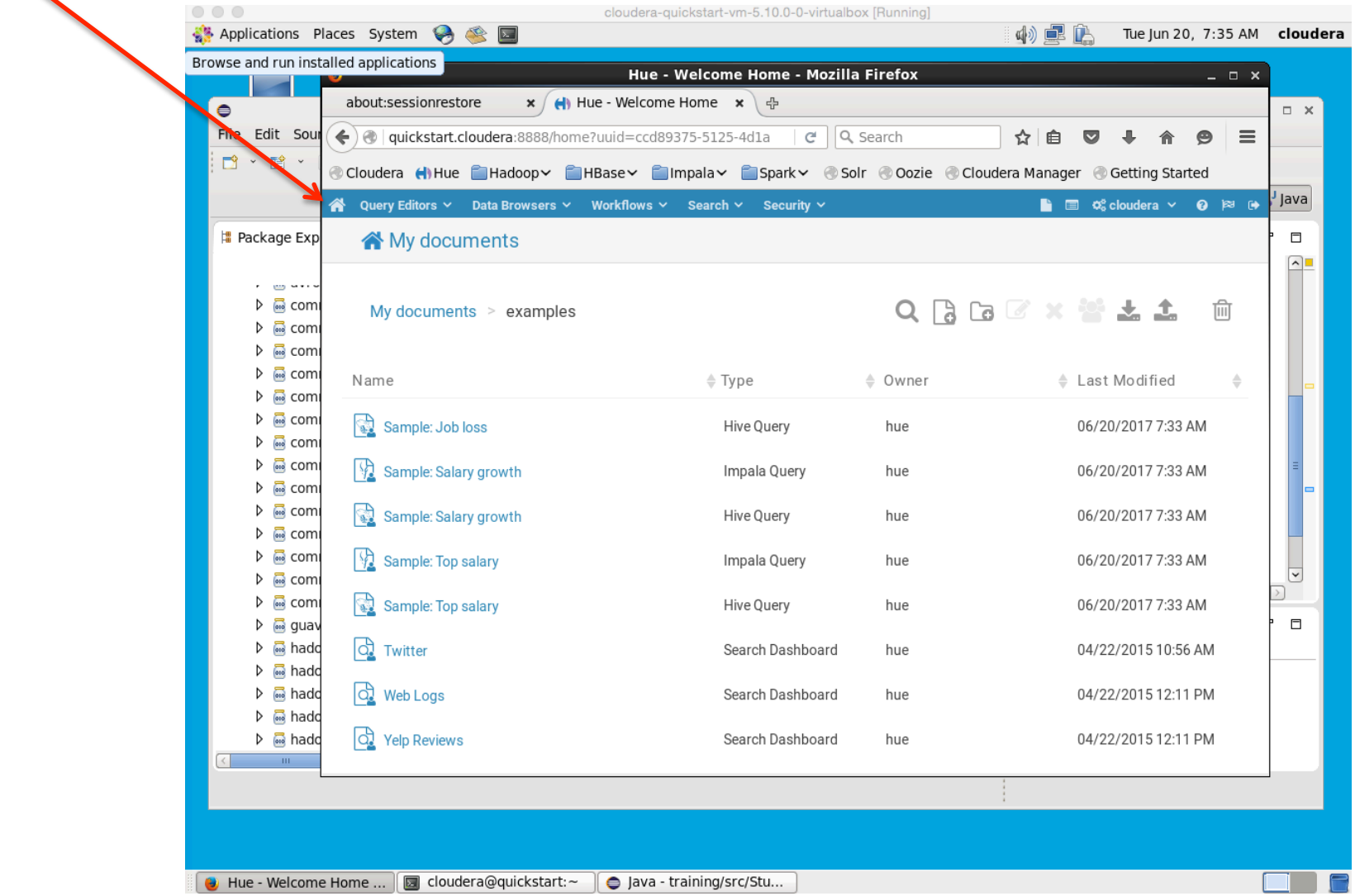
Stub Application and Libraries



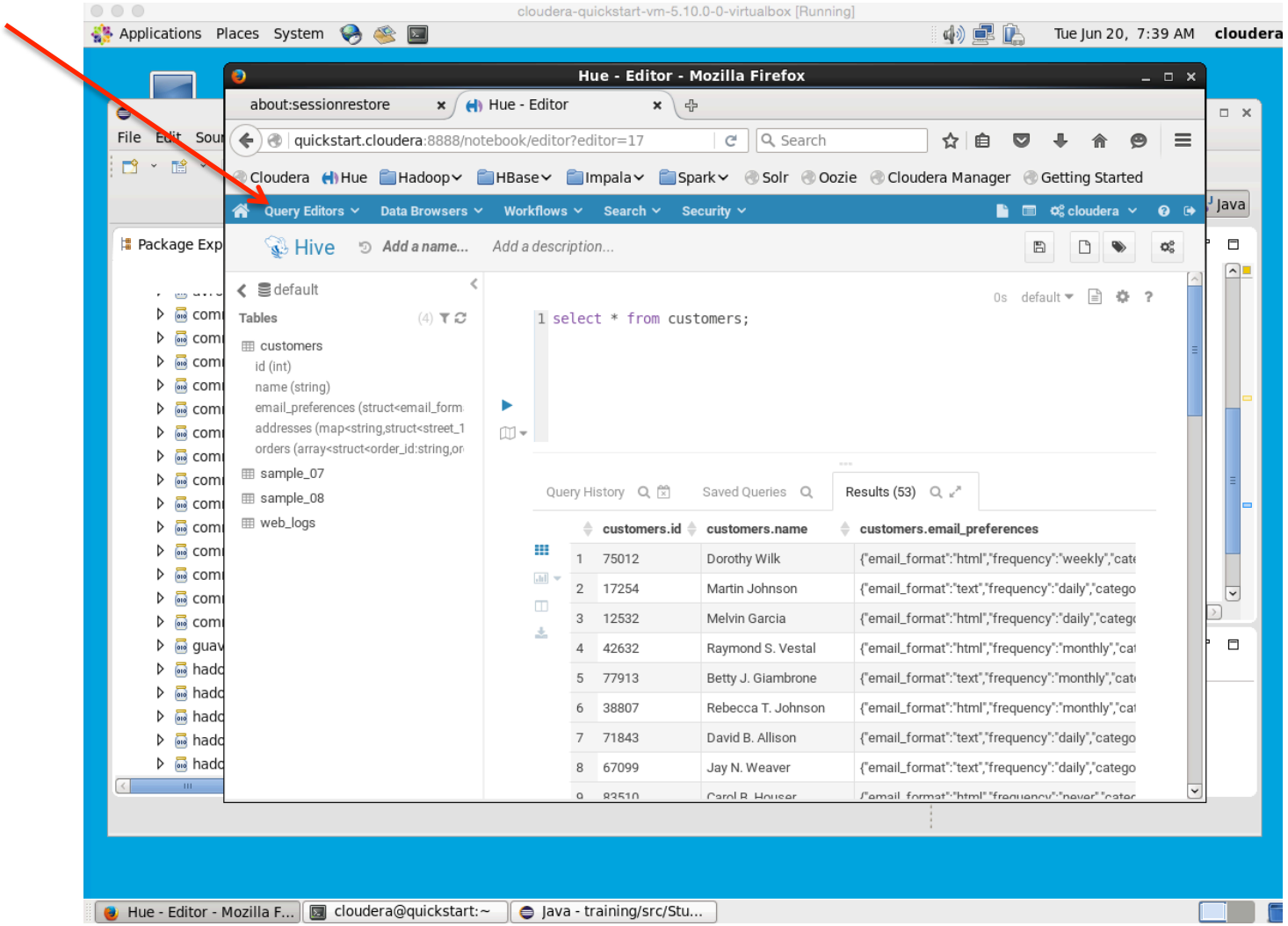
Firefox


cloudera-quickstart-vm-5.10.0-0-virtualbox [Running]
Applications Places System
Tue Jun 20, 7:25 AM cloudera
Change desktop appearance and behavior, get help, or log out
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
about:sessionrestore
Cloudera Live : Welcom...
quickstart.cloudera/#/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager
cloudera LIVE
Navigation
Welcome to Your Cloudera QuickStart VM!
Your Cluster
Node
Address
Manager Node
10.0.2.15
Worker Node 1
10.0.2.15
Get Started
File Edit Source
Package Explorer
Java
No consoles to display at this time.
Cloudera Live : Welco...
cloudera@quickstart:~
Java - training/src/Stu...

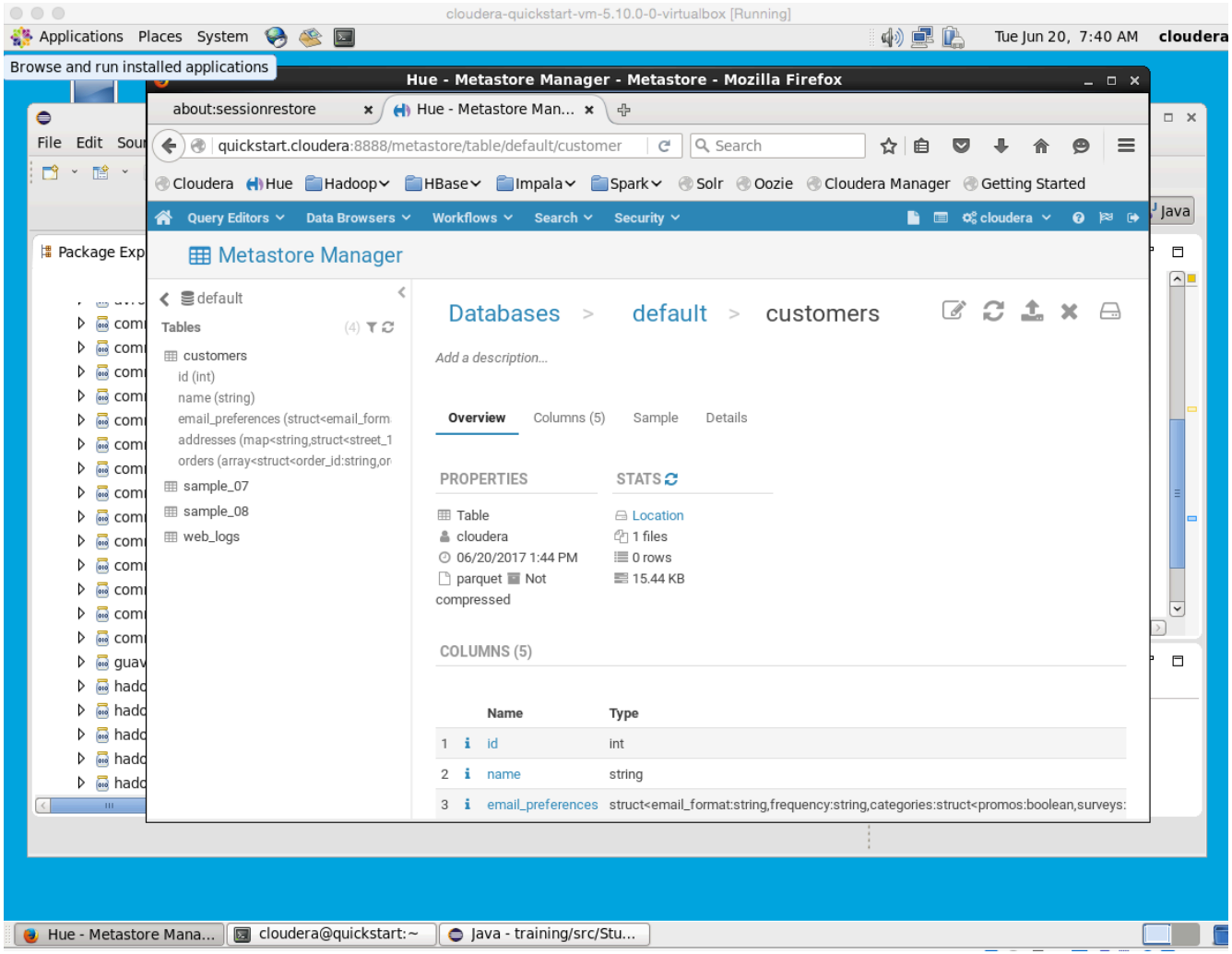
Applications via Hue Browser
Login/password: cloudera/cloudera



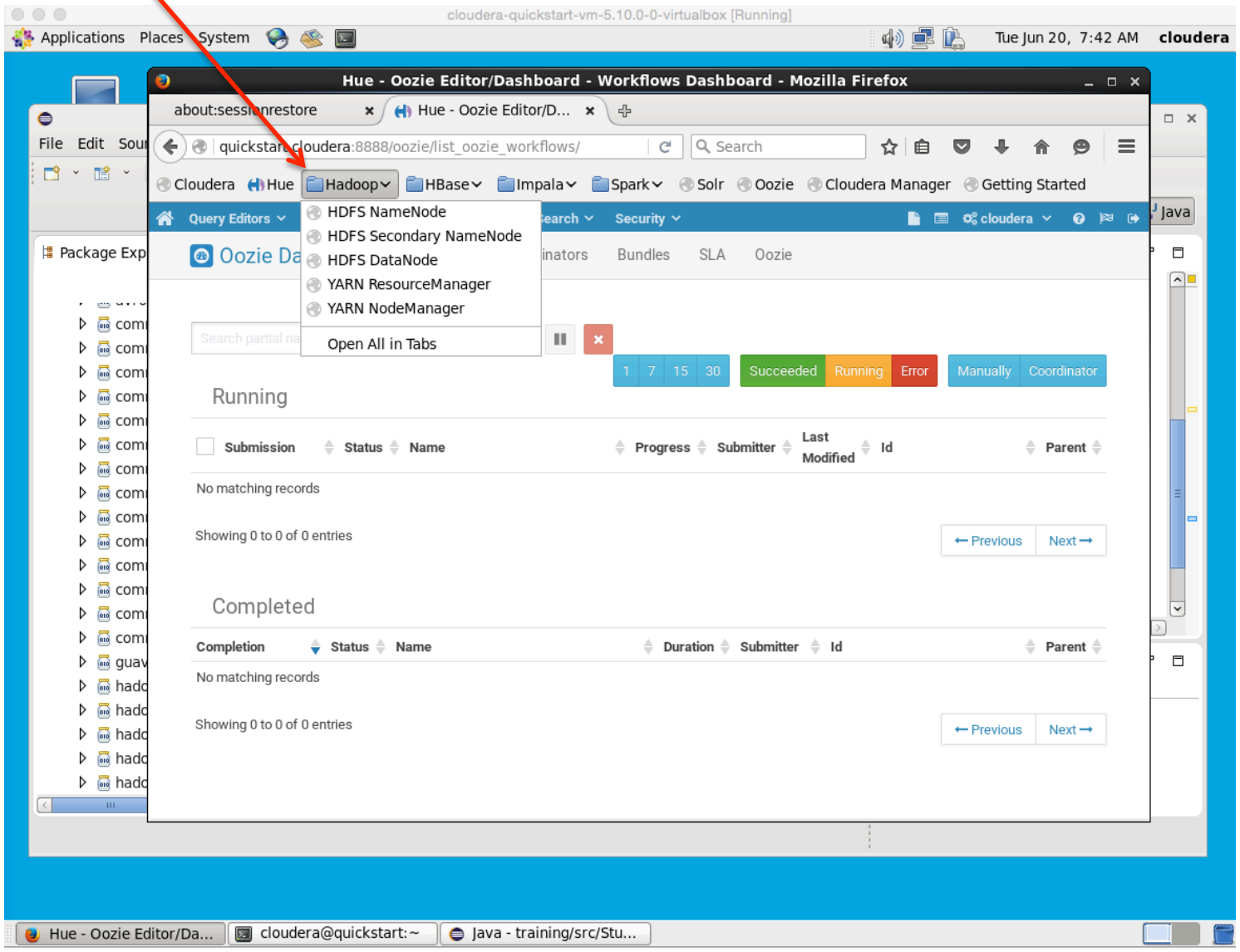
Example: Query Editors - Hive



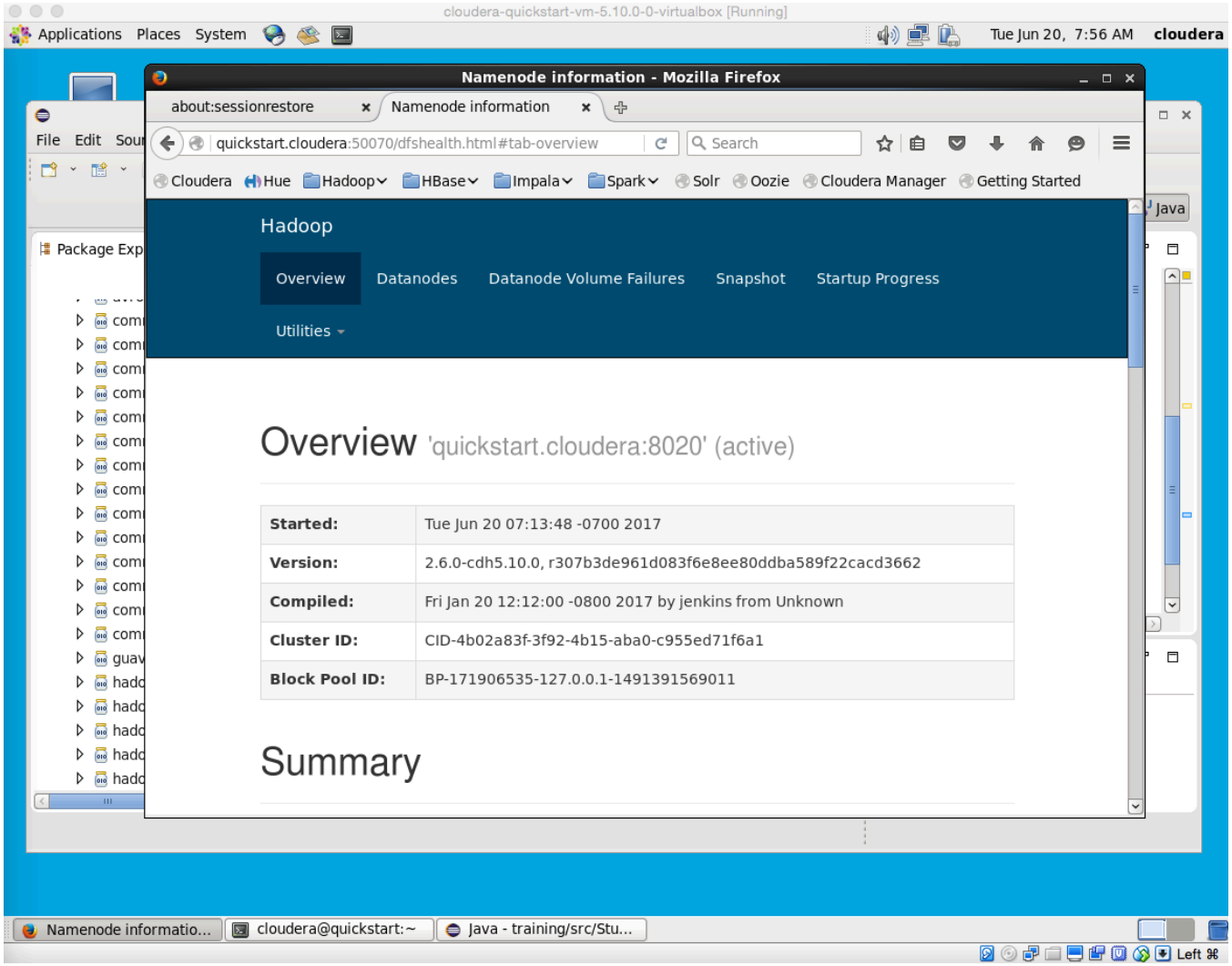
Metastore browser



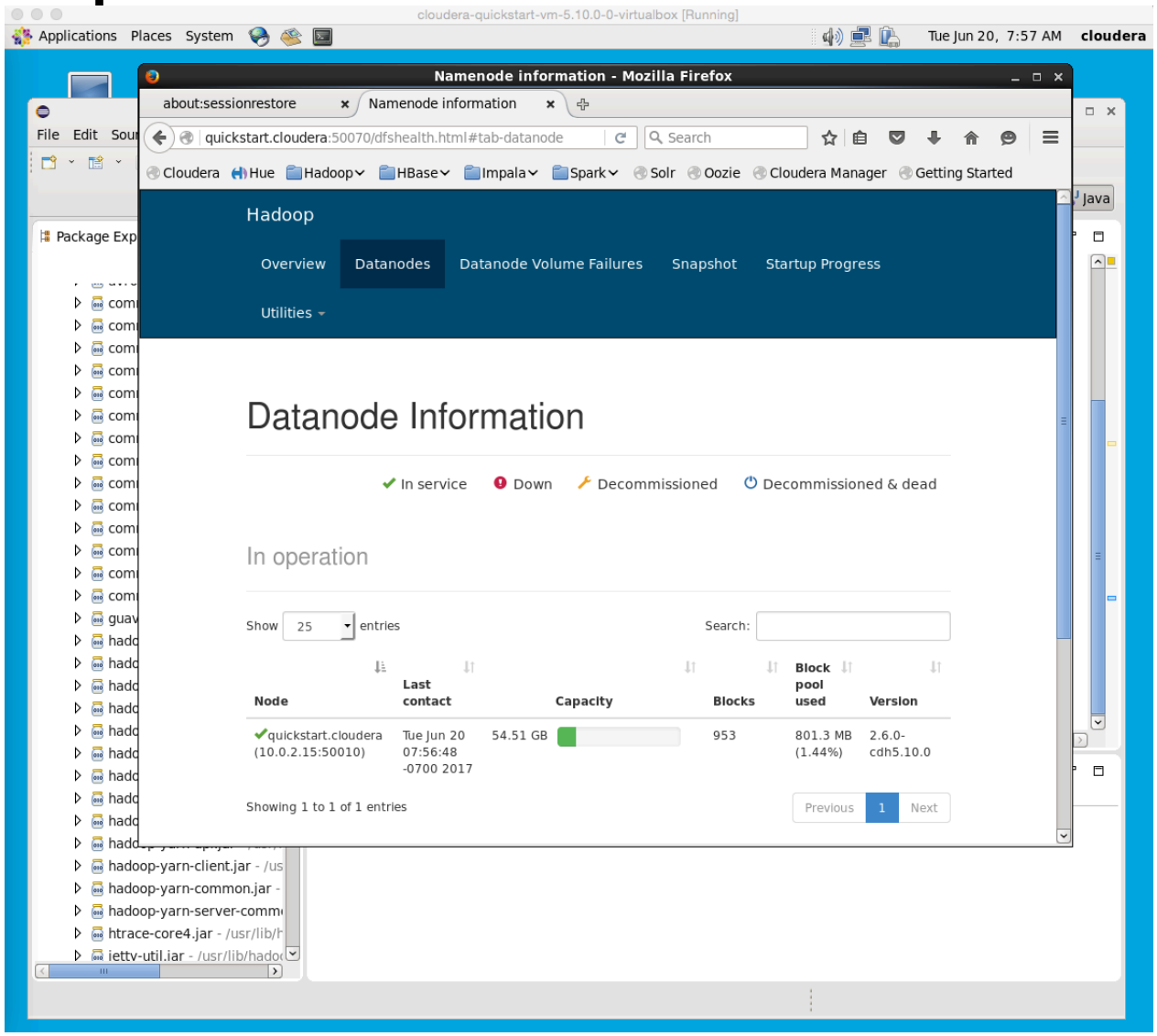
Hadoop Configuration



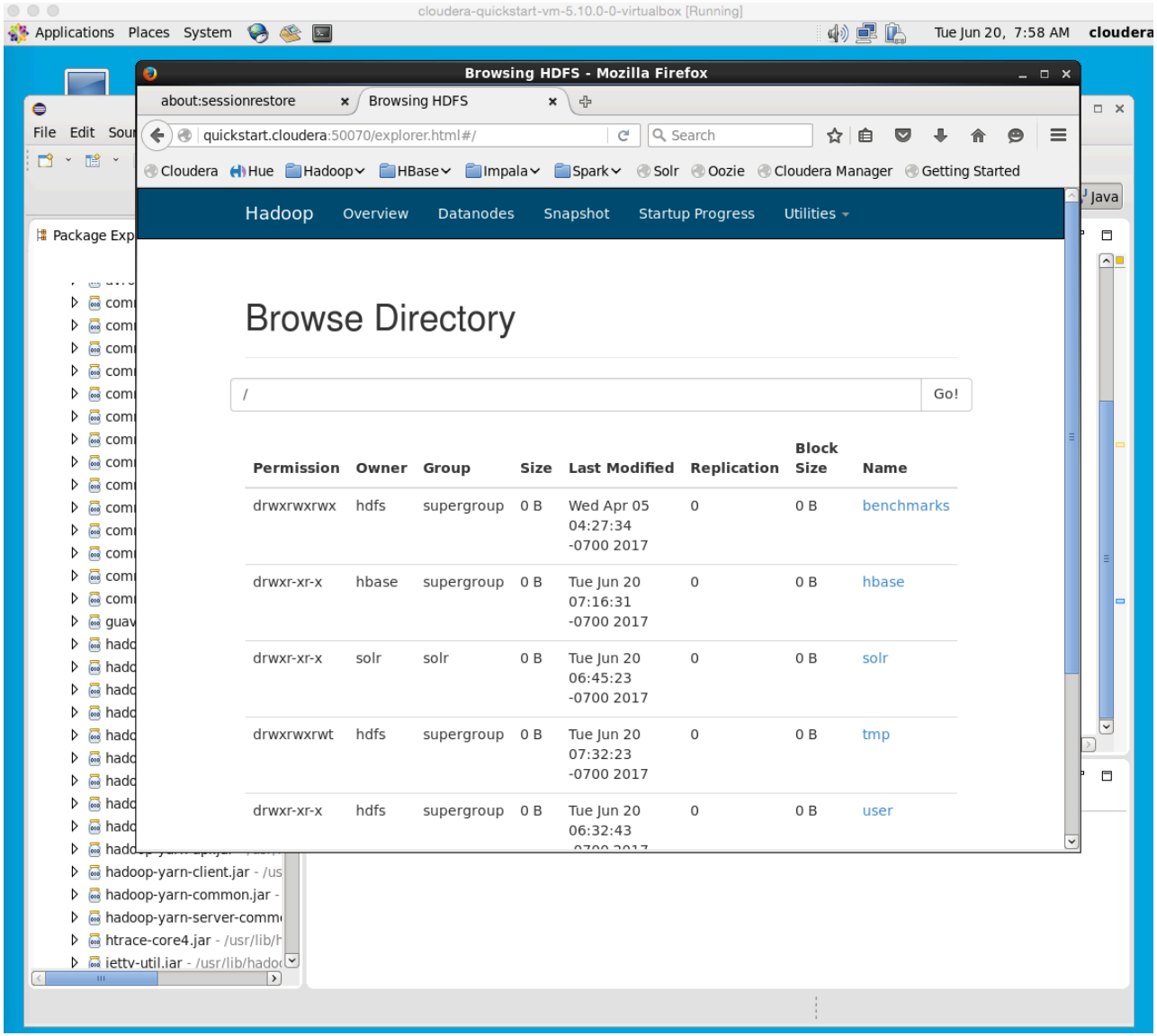
Hadoop Namenode



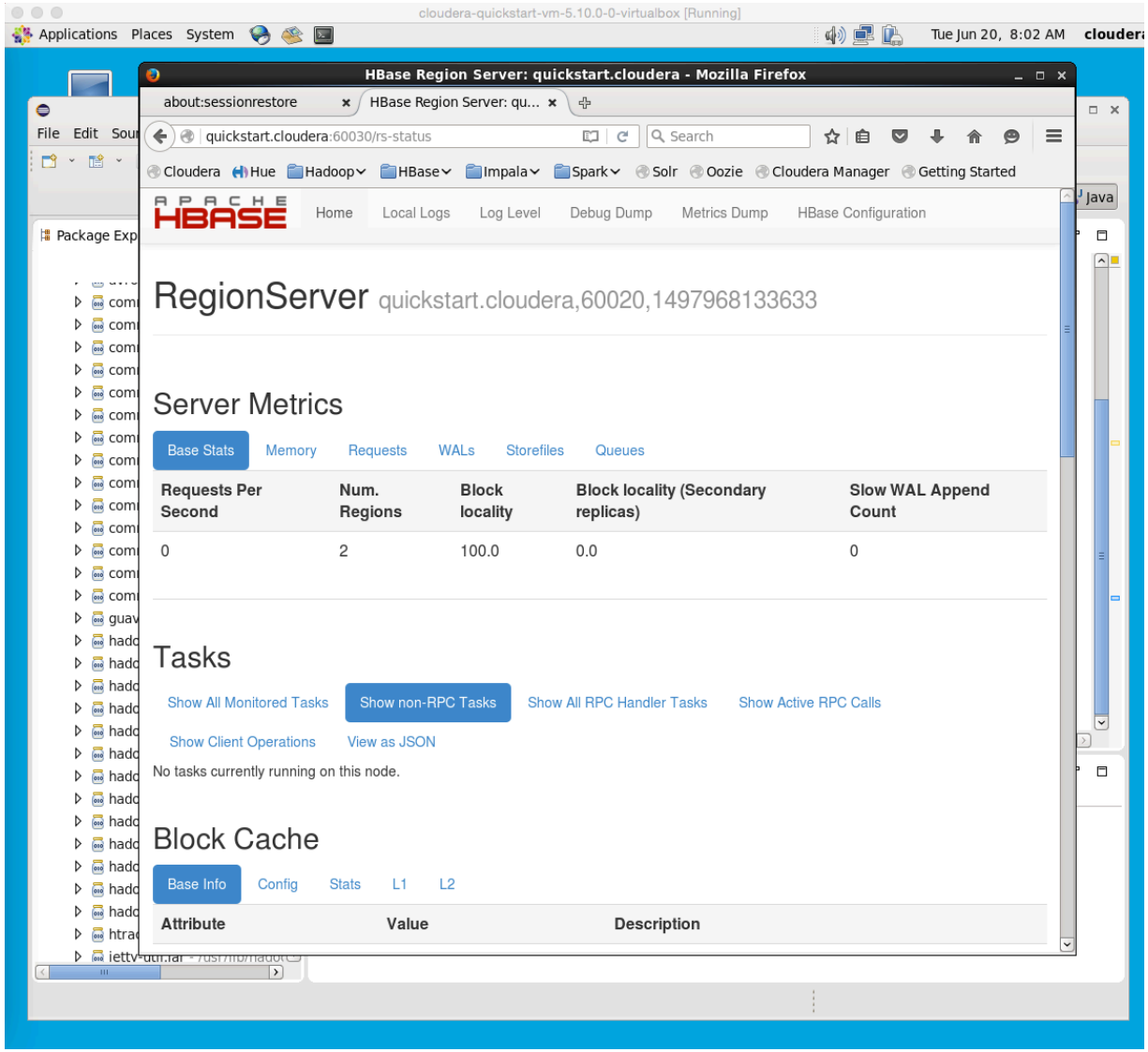
Datanode – single datanode due to pseudo distributed mode



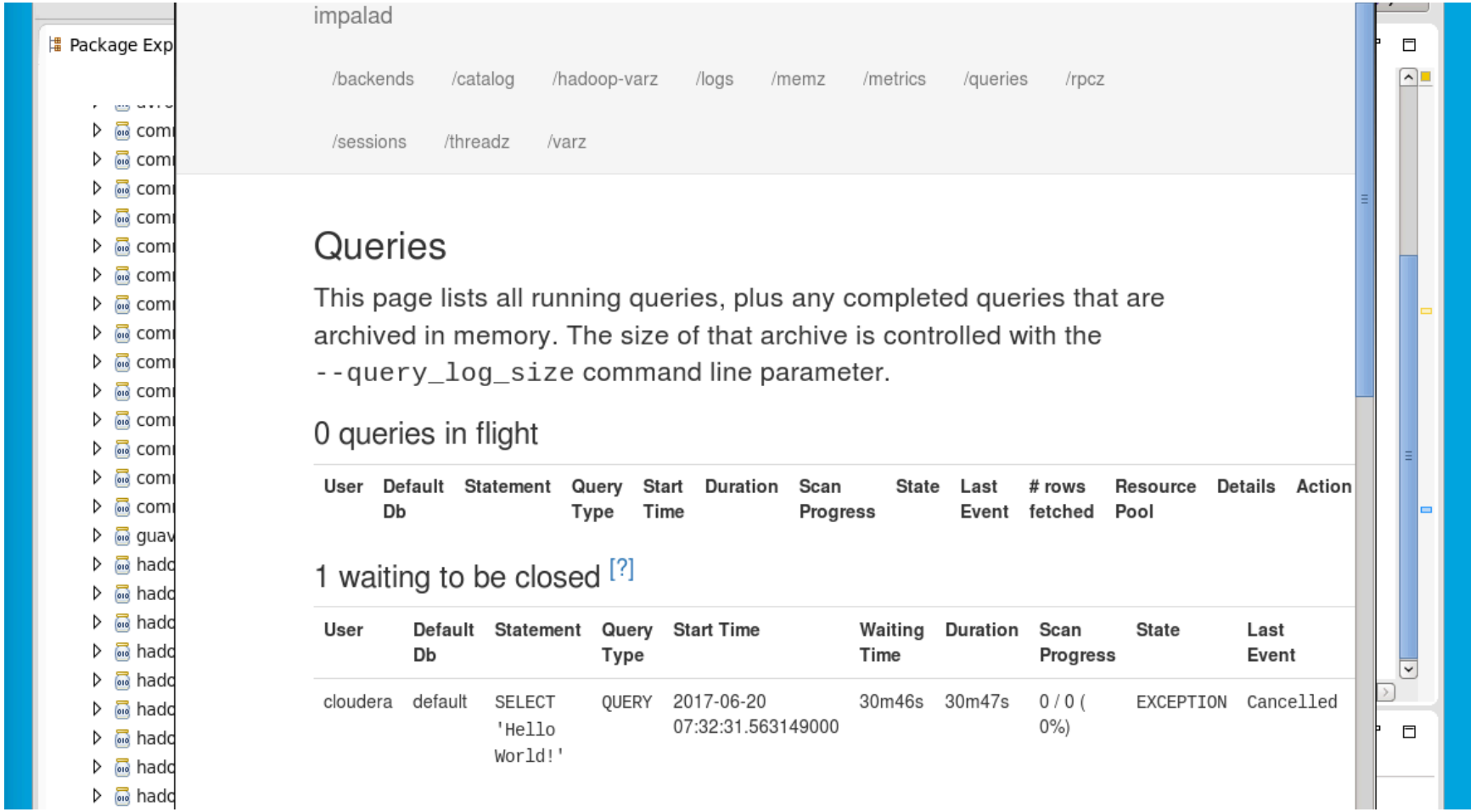
Datanode


cloudera-quickstart-vm-5.10.0-0-virtualbox [Running]
Applications Places System
Tue Jun 20, 7:58 AM cloudera
Browsing HDFS - Mozilla Firefox
about:sessionrestore
Browsing HDFS
quickstart.cloudera:50070/explorer.html#/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Hadoop Overview Datanodes Snapshot Startup Progress Utilities
Browse Directory
/
Go!
Permission Owner Group Size Last Modified Replication Block Size Name
drwxrwxrwx hdfs supergroup 0 B Wed Apr 05 04:27:34 -0700 2017 0 0 B benchmarks
drwxr-xr-x hbase supergroup 0 B Tue Jun 20 07:16:31 -0700 2017 0 0 B hbase
drwxr-xr-x solr solr 0 B Tue Jun 20 06:45:23 -0700 2017 0 0 B solr
drwxrwxrwt hdfs supergroup 0 B Tue Jun 20 07:32:23 -0700 2017 0 0 B tmp
drwxr-xr-x hdfs supergroup 0 B Tue Jun 20 06:32:43 0 0 B user
File Edit
Package Exp
Java
hadoop-yarn-client.jar
hadoop-yarn-common.jar
htrace-core4.jar
jetty-util.jar

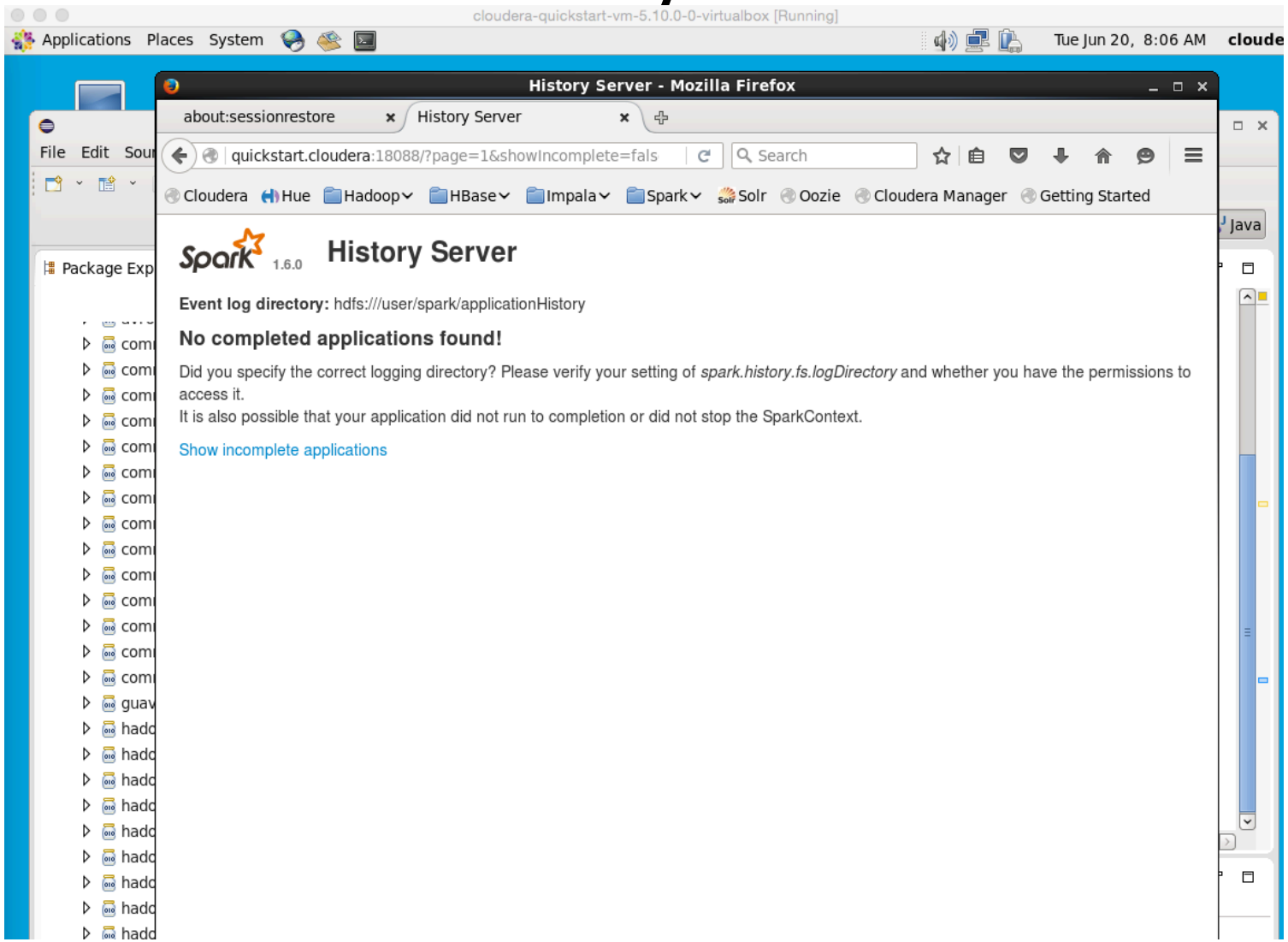
HBase


cloudera-quickstart-vm-5.10.0-0-virtualbox [Running]
Applications Places System
Tue Jun 20, 8:02 AM
HBase Region Server: quickstart.cloudera - Mozilla Firefox
about:sessionrestore
HBase Region Server: qu...
quickstart.cloudera:60030/rs-status
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
APACHE HBASE
Home Local Logs Log Level Debug Dump Metrics Dump HBase Configuration
RegionServer quickstart.cloudera,60020,1497968133633
Server Metrics
Base Stats Memory Requests WALs Storefiles Queues
Requests Per Second
Num. Regions
Block locality
Block locality (Secondary replicas)
Slow WAL Append Count
0
2
100.0
0.0
0
Tasks
Show All Monitored Tasks
Show non-RPC Tasks
Show All RPC Handler Tasks
Show Active RPC Calls
Show Client Operations
View as JSON
No tasks currently running on this node.
Block Cache
Base Info Config Stats L1 L2
Attribute
Value
Description

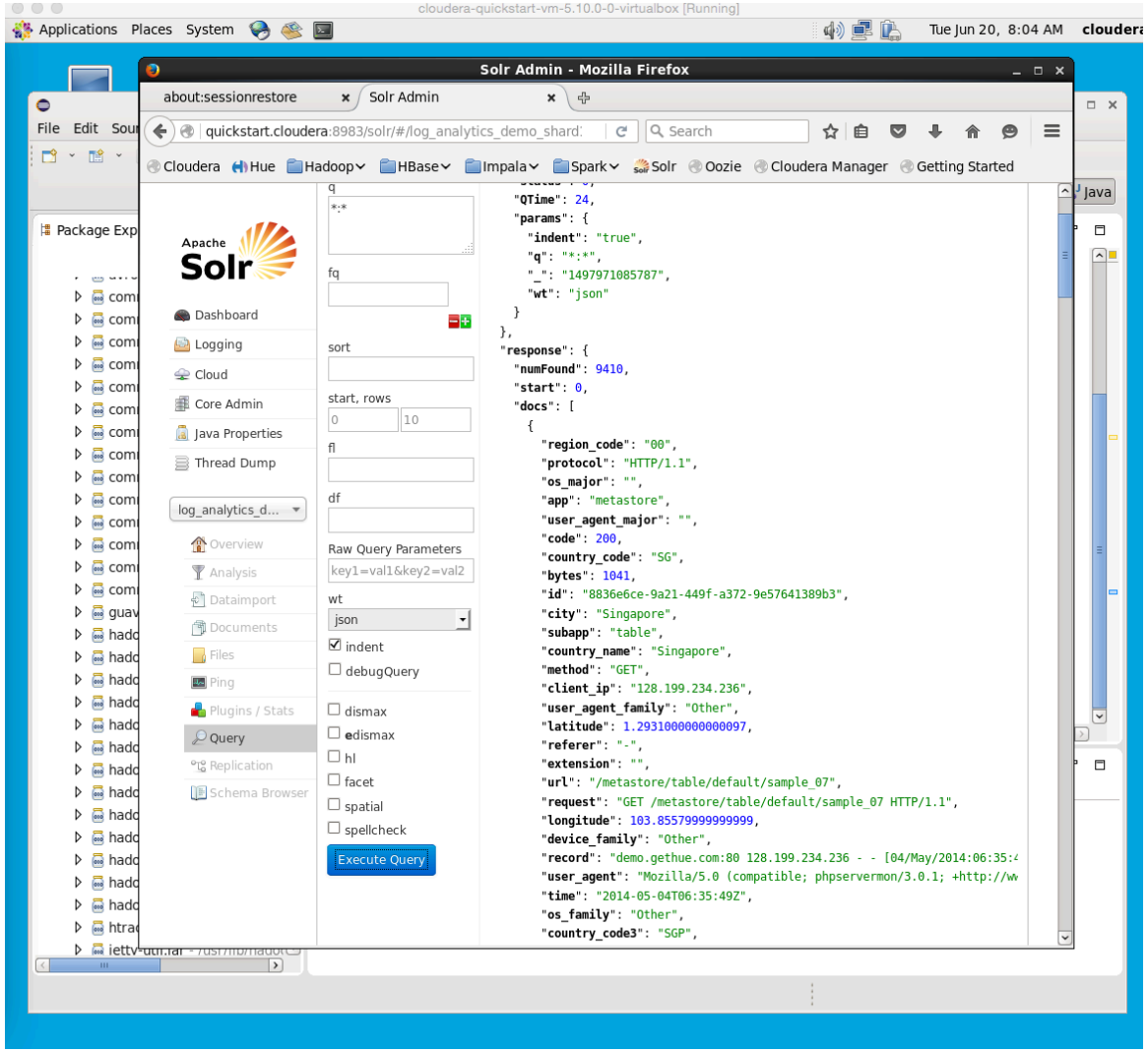
Cloudera Impala – in-memory database



Spark – in-memory processing & analytics



Solr Search


cloudera-quickstart-vm-5.10.0-0-virtualbox [Running]
Applications Places System
Tue Jun 20, 8:04 AM
Solr Admin - Mozilla Firefox
about:sessionrestore
Solr Admin
quickstart.cloudera:8983/solr/#/log_analytics_demo_shard
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Apache Solr
Dashboard
Logging
Cloud
Core Admin
Java Properties
Thread Dump
log_analytics_d...
Overview
Analysis
Dataimport
Documents
Files
Ping
Plugins / Stats
Query
Replication
Schema Browser
q
:
fq
sort
start, rows
0
10
fl
df
Raw Query Parameters
key1=val1&key2=val2
wt
json
indent
debugQuery
dismax
edismax
hl
facet
spatial
spellcheck
Execute Query
"QTime": 24,
"params": {
"indent": "true",
"q": "*:*",
"_": "1497971085787",
"wt": "json"
}
},
"response": {
"numFound": 9410,
"start": 0,
"docs": [
{
"region_code": "00",
"protocol": "HTTP/1.1",
"os_major": "",
"app": "metastore",
"user_agent_major": "",
"code": 200,
"country_code": "SG",
"bytes": 1041,
"id": "8836e6ce-9a21-449f-a372-9e57641389b3",
"city": "Singapore",
"subapp": "table",
"country_name": "Singapore",
"method": "GET",
"client_ip": "128.199.234.236",
"user_agent_family": "Other",
"latitude": 1.2931000000000097,
"referer": "-",
"extension": "",
"url": "/metastore/table/default/sample_07",
"request": "GET /metastore/table/default/sample_07 HTTP/1.1",
"longitude": 103.85579999999999,
"device_family": "Other",
"record": "demo.gethue.com:80 128.199.234.236 - - [04/May/2014:06:35:
"user_agent": "Mozilla/5.0 (compatible; phpservermon/3.0.1; +http://w
"time": "2014-05-04T06:35:49Z",
"os_family": "Other",
"country_code3": "SGP",

Review

- Which of the following Hadoop libraries gives you the ability to have SQL-like, or HQL query and creates batch MapReduce jobs with that query?
  - Postgress
  - Mahout
  - Hive
  - Pig
  - Impala

- Which IDE for Python
  - MS Visual Studio
  - Eclipse
  - Sublime
  - Rstudio
  - PyCharm
  - Spyder