

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
 - Free, but complicated unless standalone mode
- Local VM
 - Must install virtualization software
- Cloud
 - Costs money to test
 - Need to remember to turn off
 - Great for episodic compute needs

Data Storage

- Local
 - File system (single)
 - Note: Developing MapReduce algorithms are *complicated*. Keep things simple.
- HDFS – pseudo or distributed
- Cloud
 - HDFS
 - S3 AWS/ BLOB Azure

Libraries

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

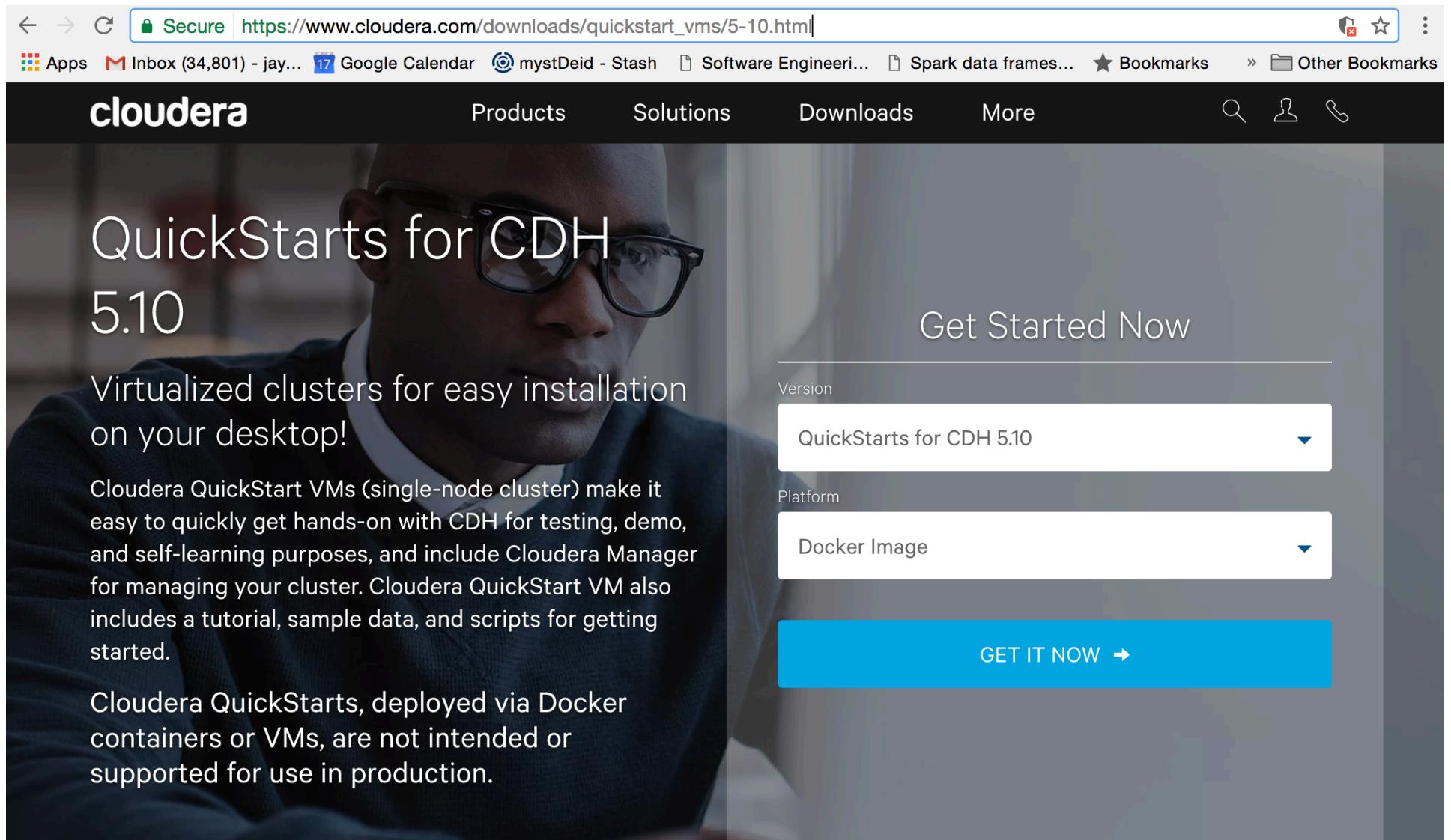
Cloudera

- 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



The screenshot shows a web browser displaying the Cloudera Quickstart landing page. The URL in the address bar is https://www.cloudera.com/downloads/quickstart_vms/5-10.html. The page features a dark background image of a person wearing glasses. On the left, there's a large heading "QuickStarts for CDH 5.10" and a subtext "Virtualized clusters for easy installation on your desktop!". Below this is a detailed description of what Cloudera QuickStart VMs are. On the right, there's a "Get Started Now" button with dropdown menus for "Version" (set to "QuickStarts for CDH 5.10") and "Platform" (set to "Docker Image"). A blue "GET IT NOW" button is at the bottom.

Secure | https://www.cloudera.com/downloads/quickstart_vms/5-10.html

Apps | Inbox (34,801) - jay... | Google Calendar | mystDeid - Stash | Software Engineeri... | Spark data frames... | Bookmarks | Other Bookmarks

cloudera Products Solutions Downloads More

QuickStarts for CDH 5.10

Virtualized clusters for easy installation on your desktop!

Cloudera QuickStart VMs (single-node cluster) make it easy to quickly get hands-on with CDH for testing, demo, and self-learning purposes, and include Cloudera Manager for managing your cluster. Cloudera QuickStart VM also includes a tutorial, sample data, and scripts for getting started.

Cloudera QuickStarts, deployed via Docker containers or VMs, are not intended or supported for use in production.

Get Started Now

Version: QuickStarts for CDH 5.10

Platform: Docker Image

GET IT NOW →

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

localhost:8888/accounts/login/?next=/

Apps Inbox (34,801) - jay... 19 Google Calendar mystDeid - Stash Software Engineeri... Spark data frames... Bookmarks Other Bookmarks

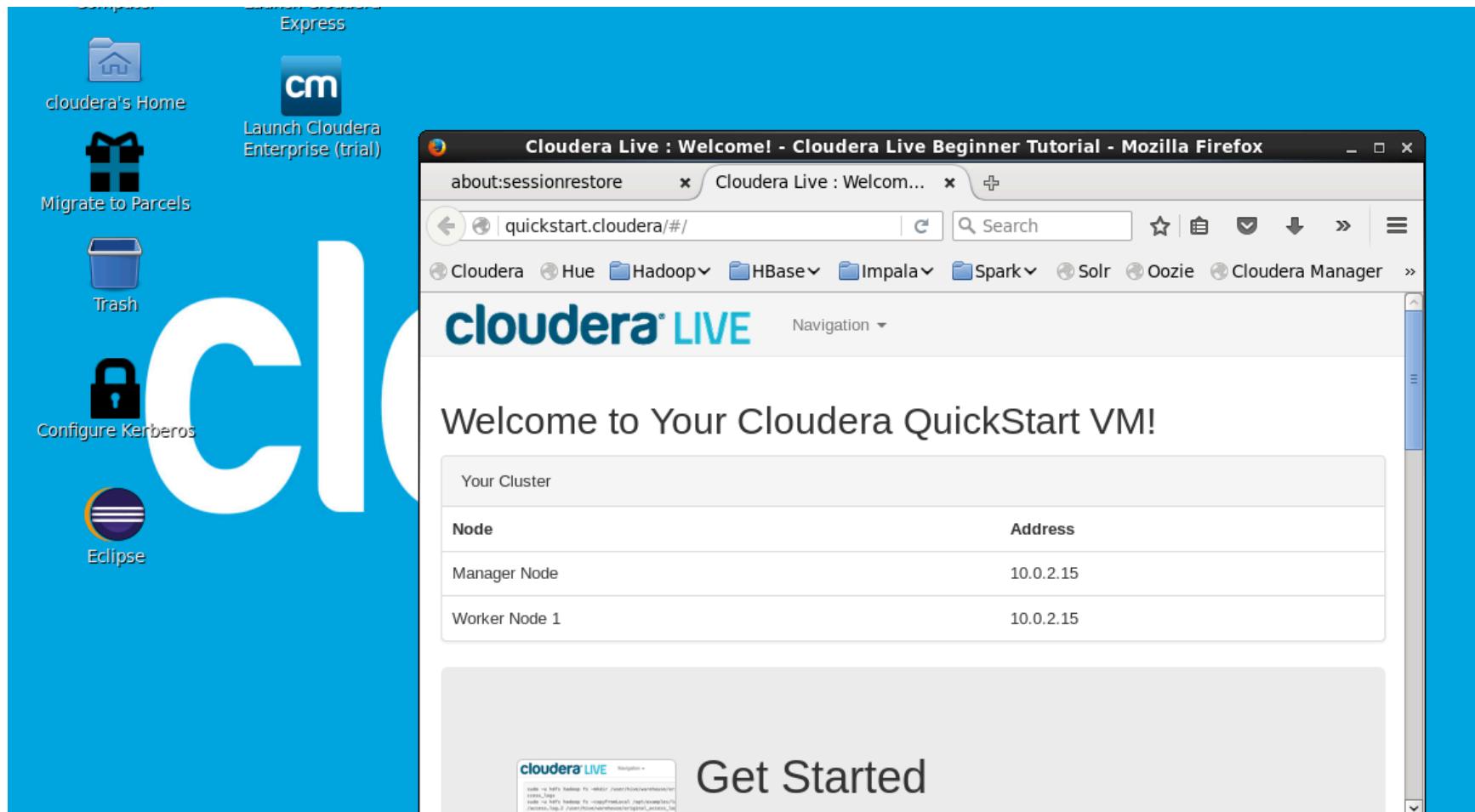


cloudera

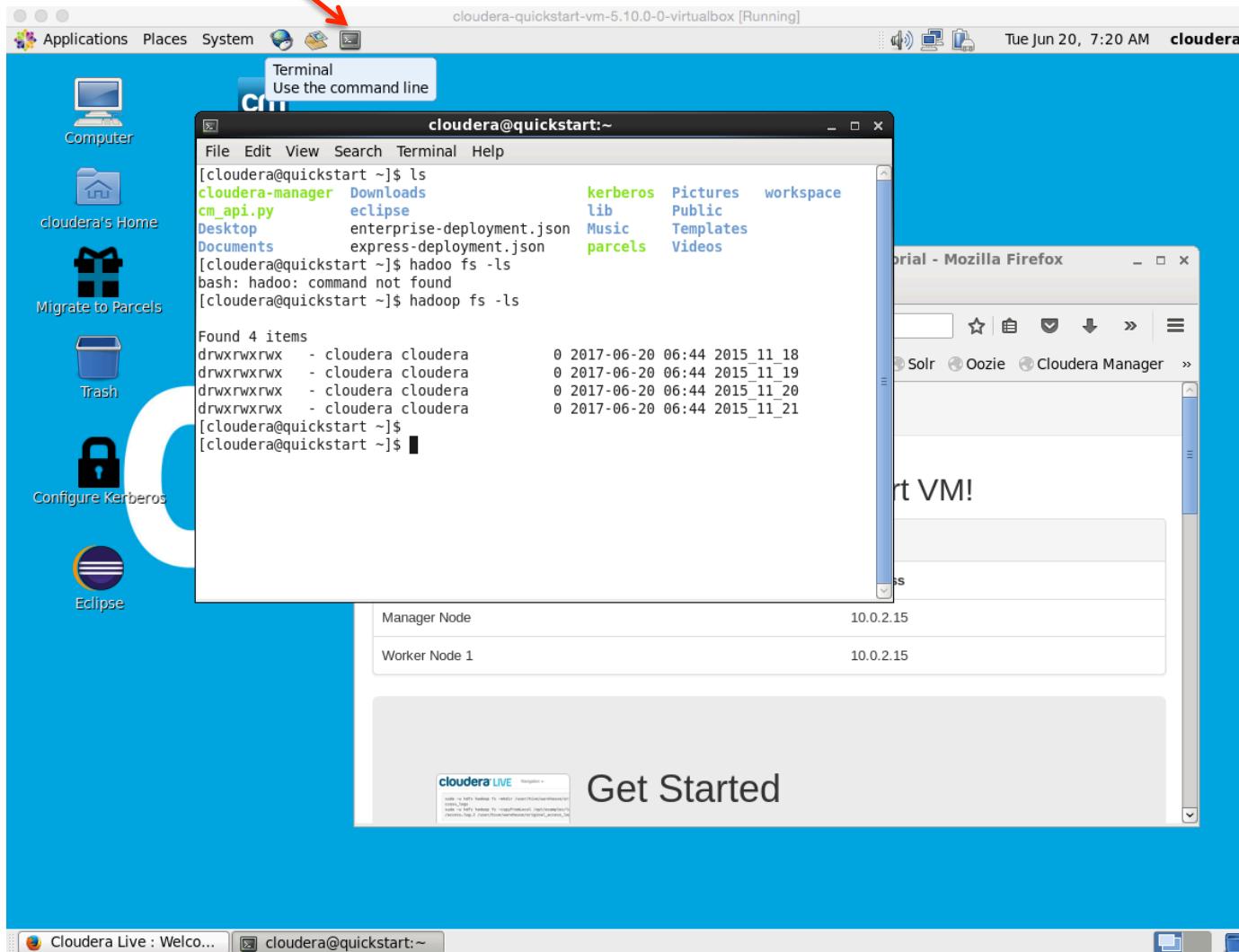
.....

Sign In

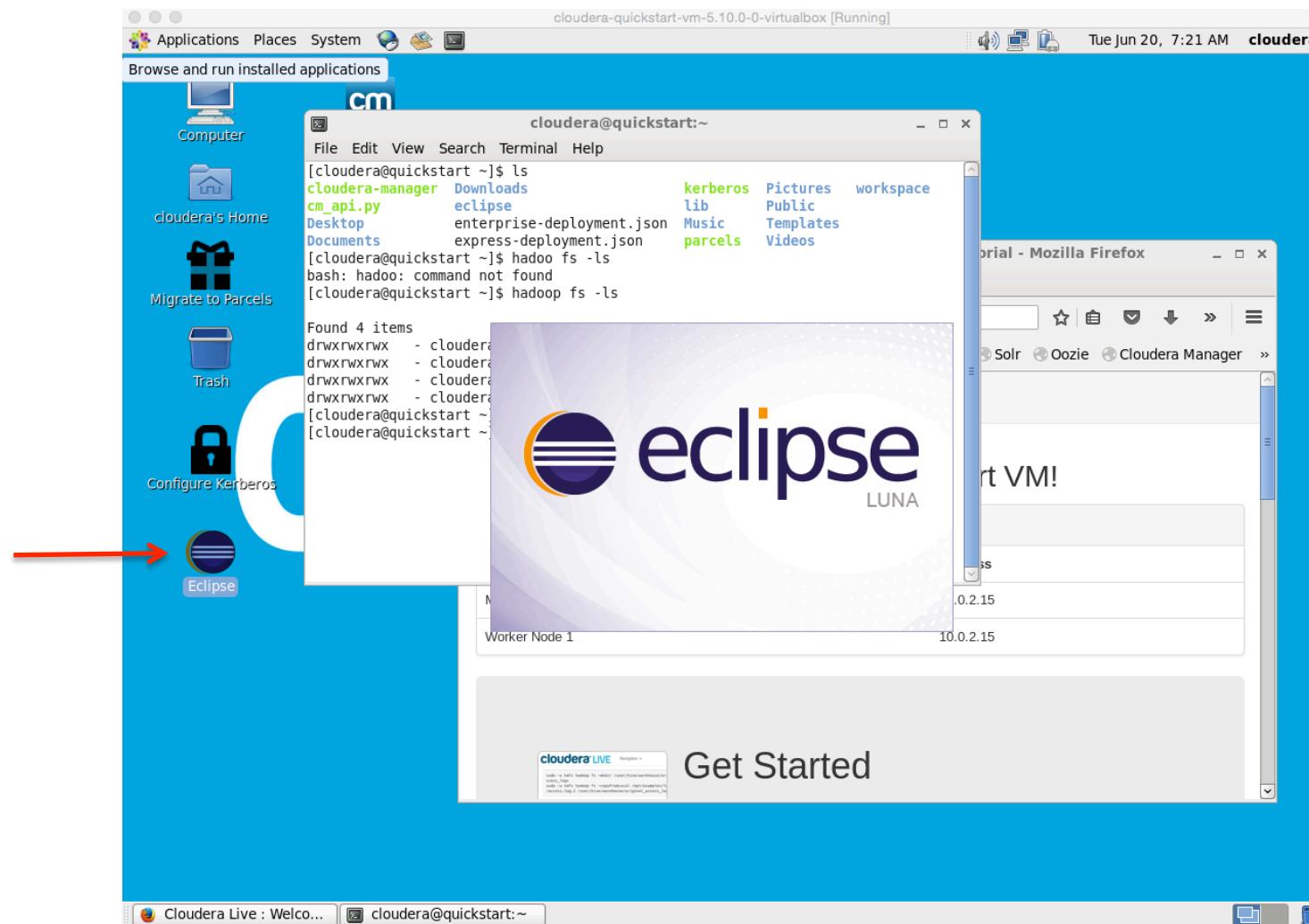
Startup Cloudera VM



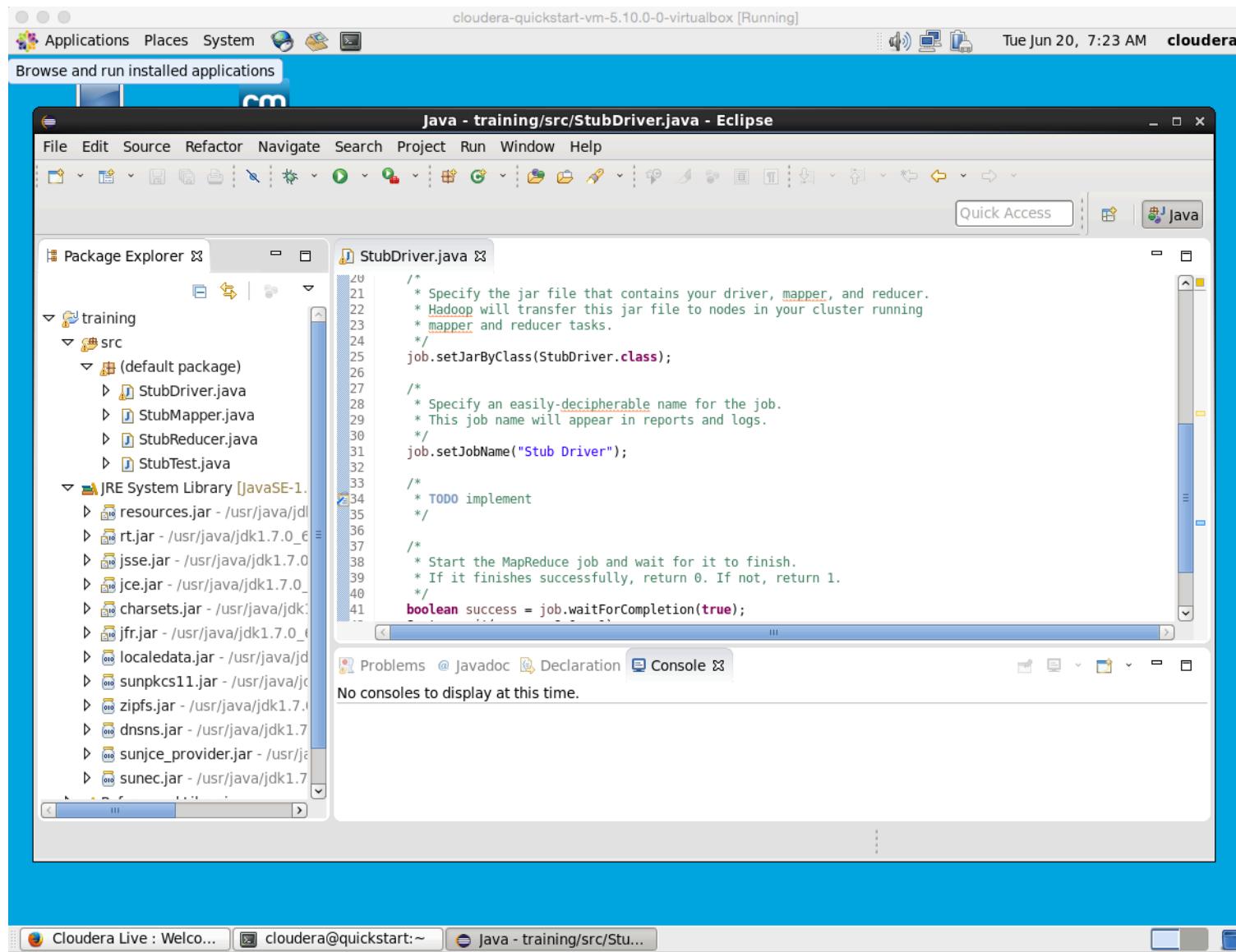
Access the terminal



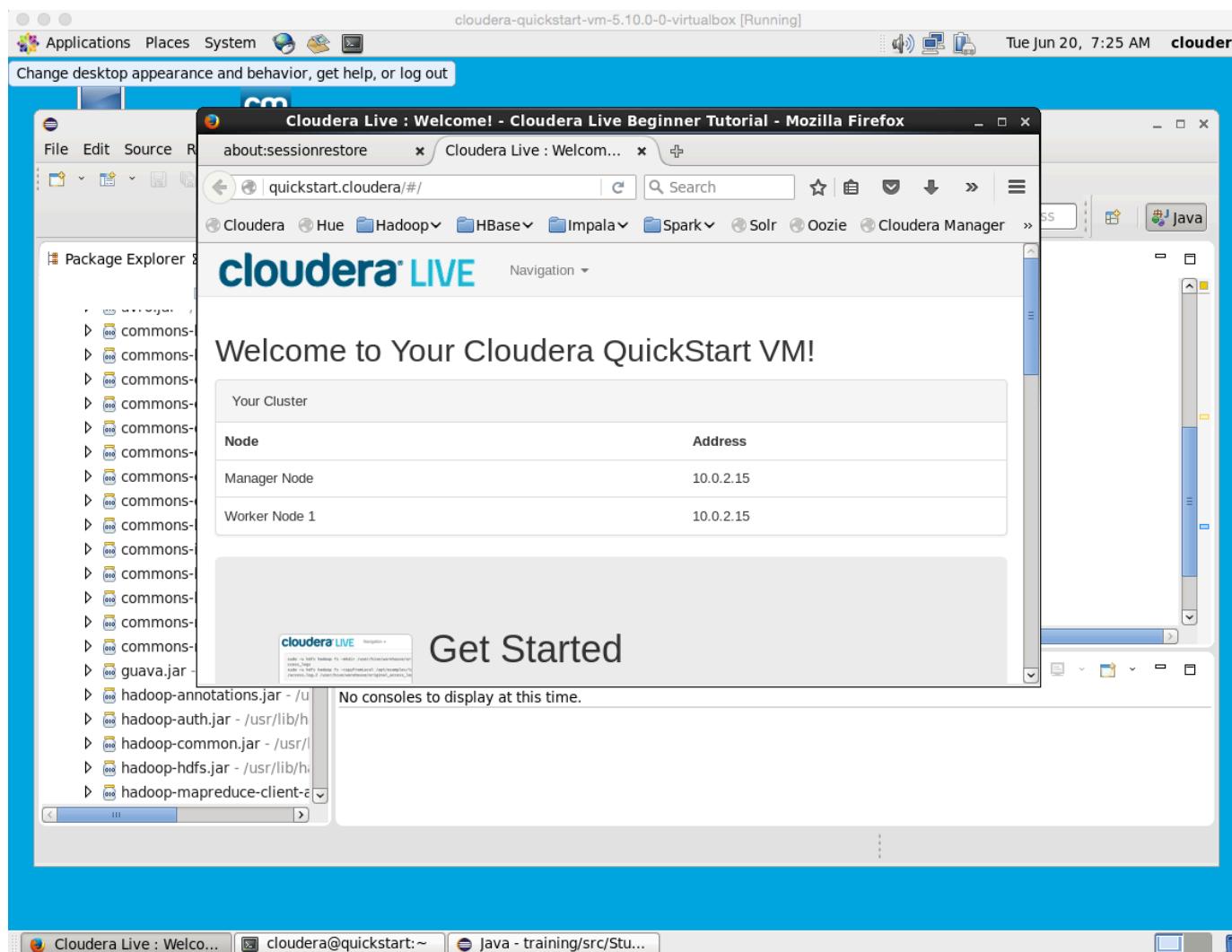
Eclipse



Stub Application and Libraries

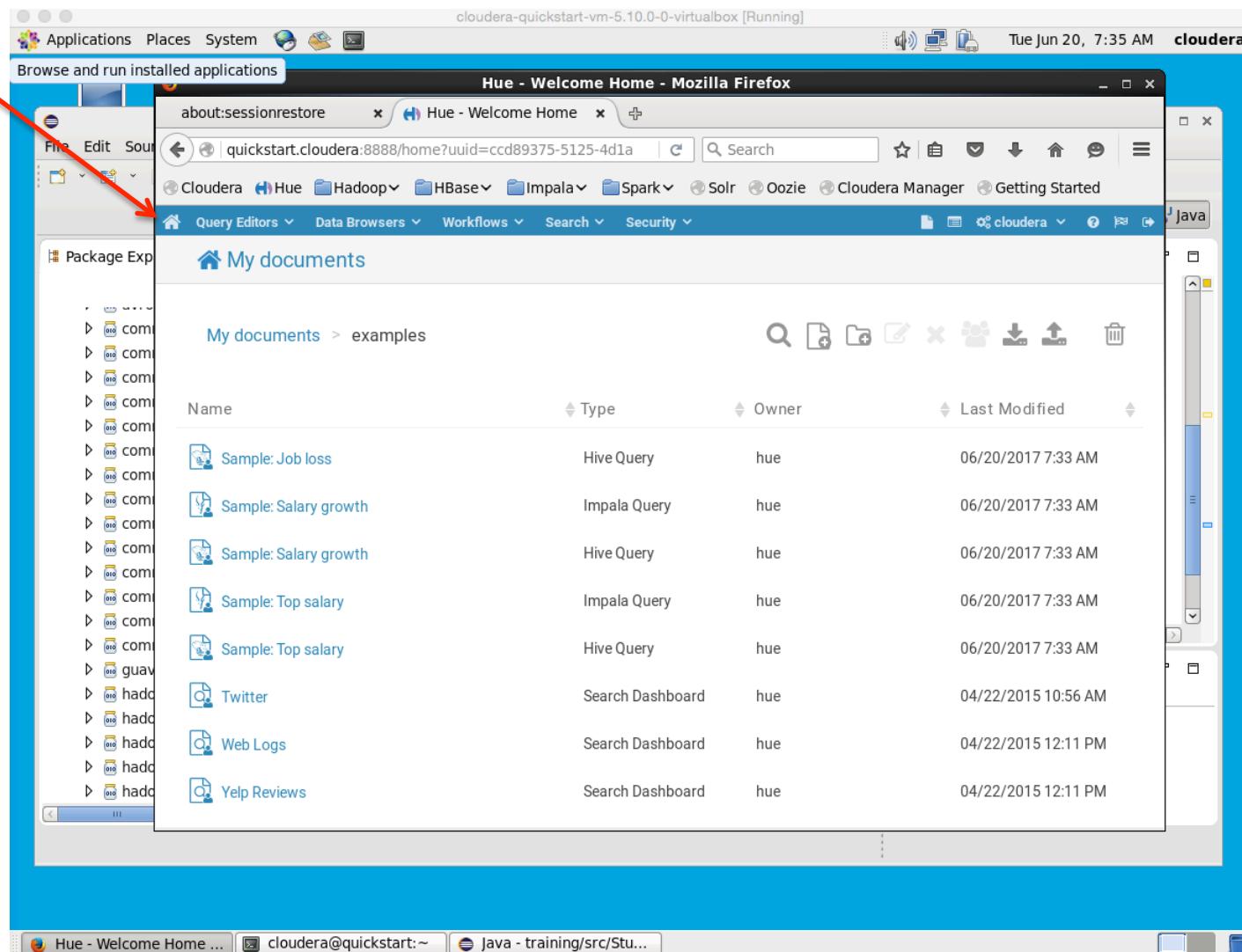


Firefox

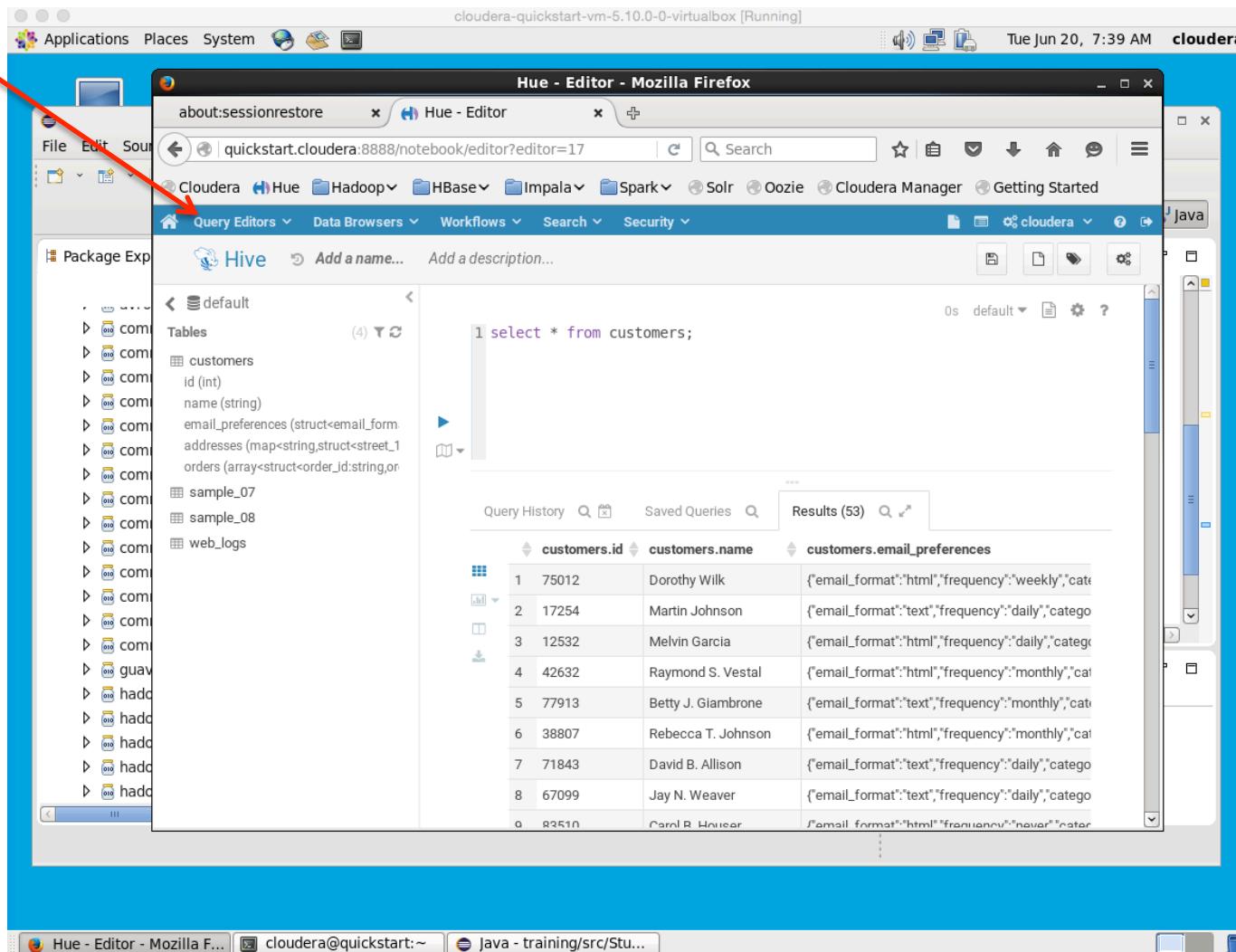


Applications via Hue Browser

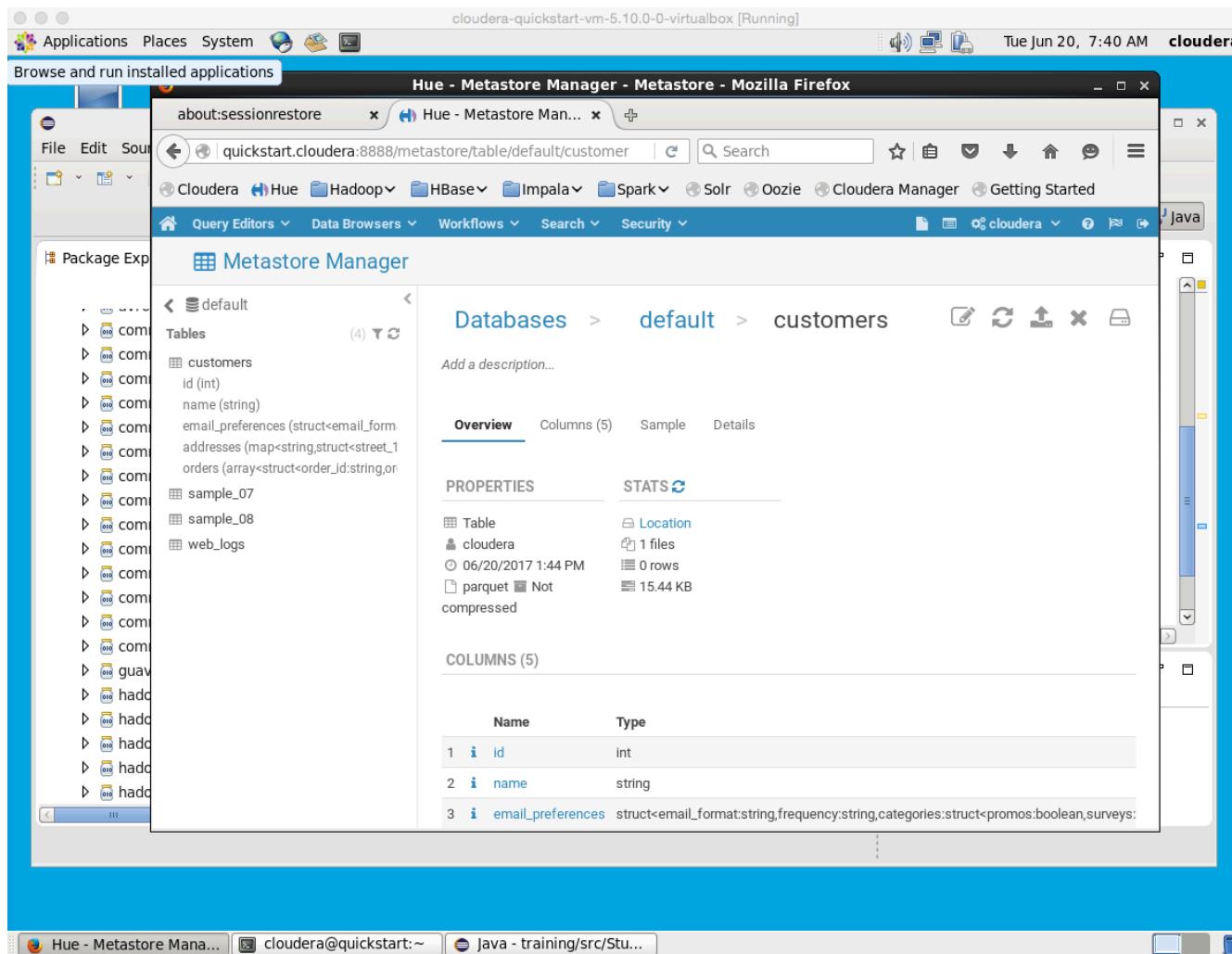
Login/password: cloudera/cloudera



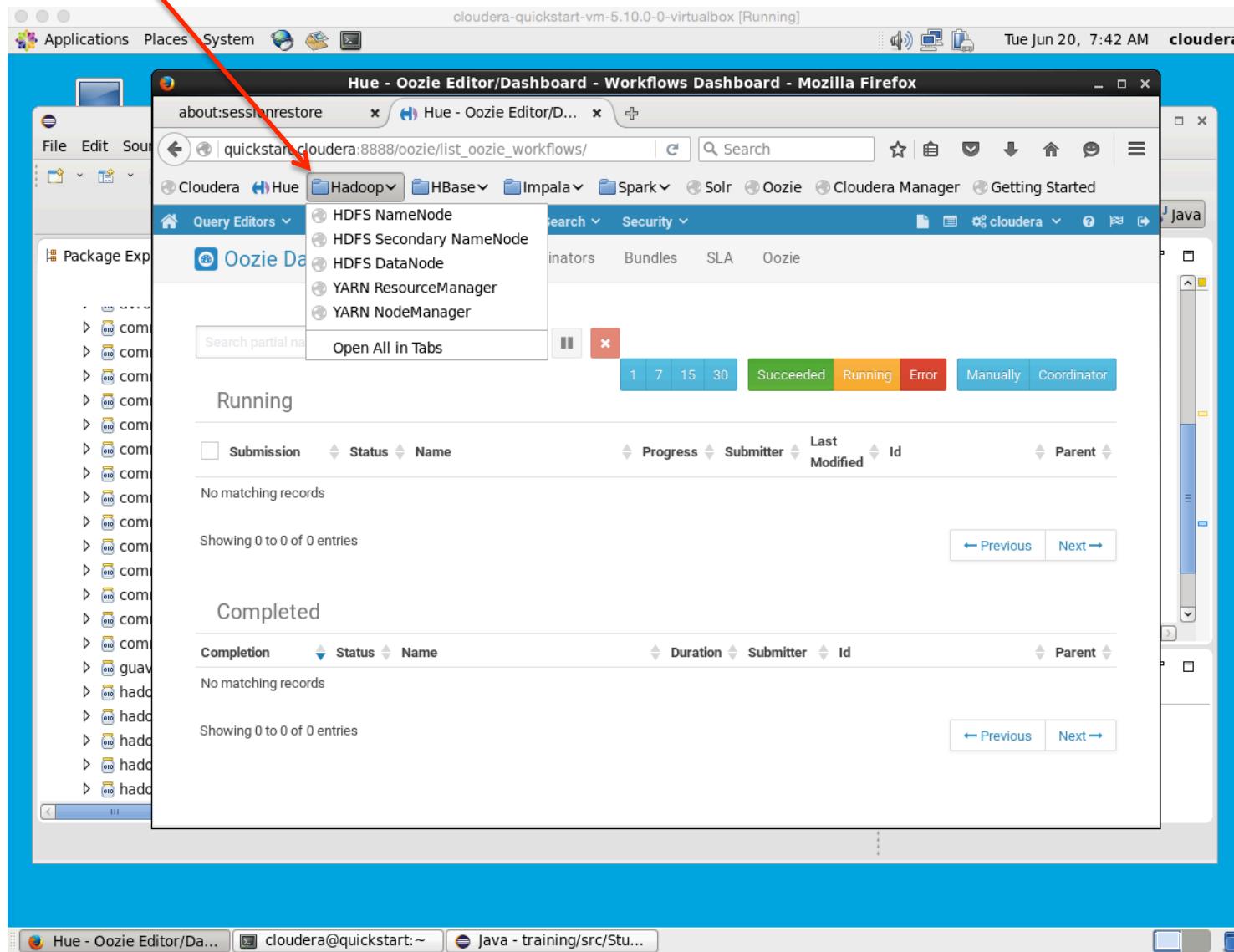
Example: Query Editors - Hive



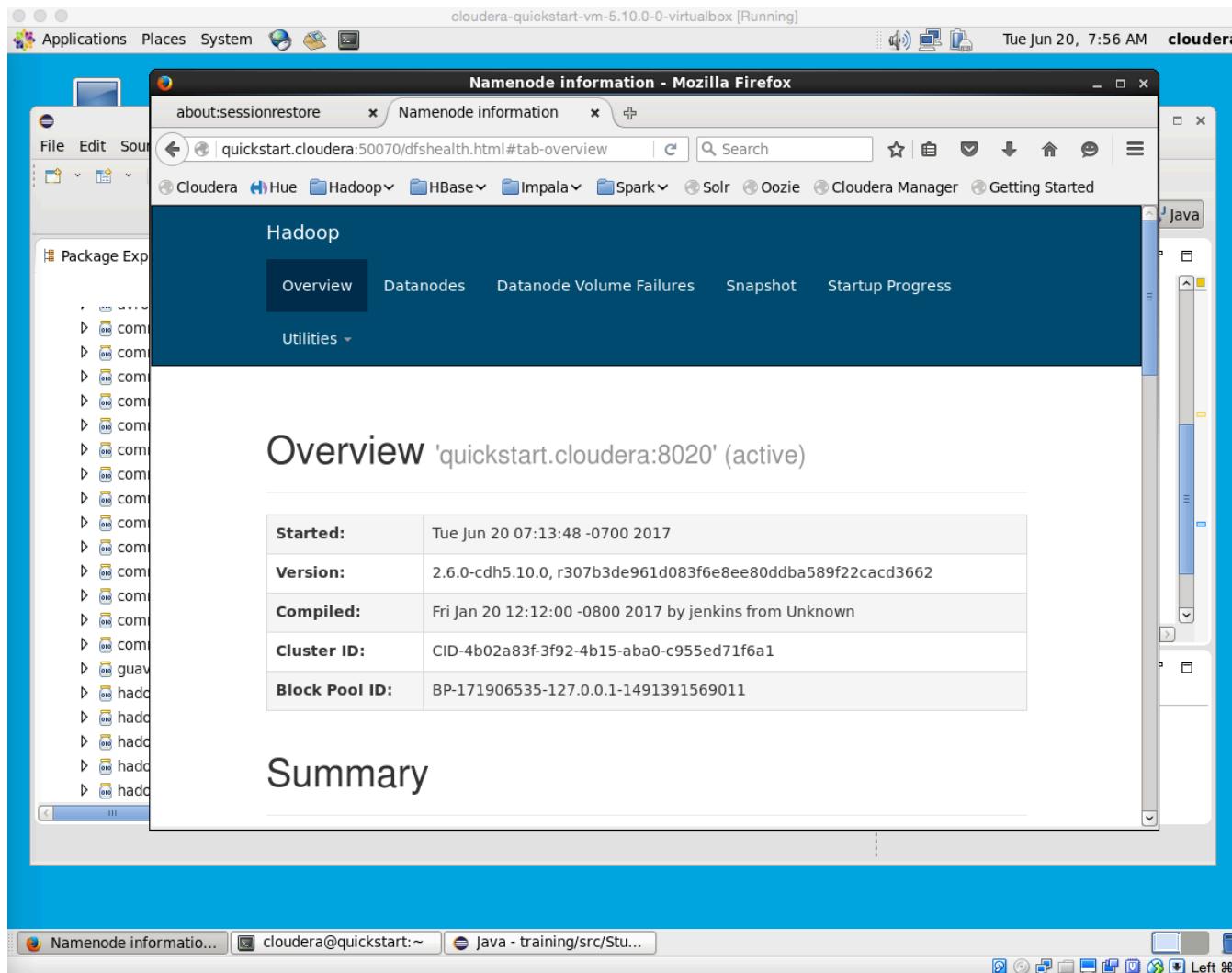
Metastore browser



Hadoop Configuration



Hadoop Namenode



Datanode – single datanode due to pseudo distributed mode

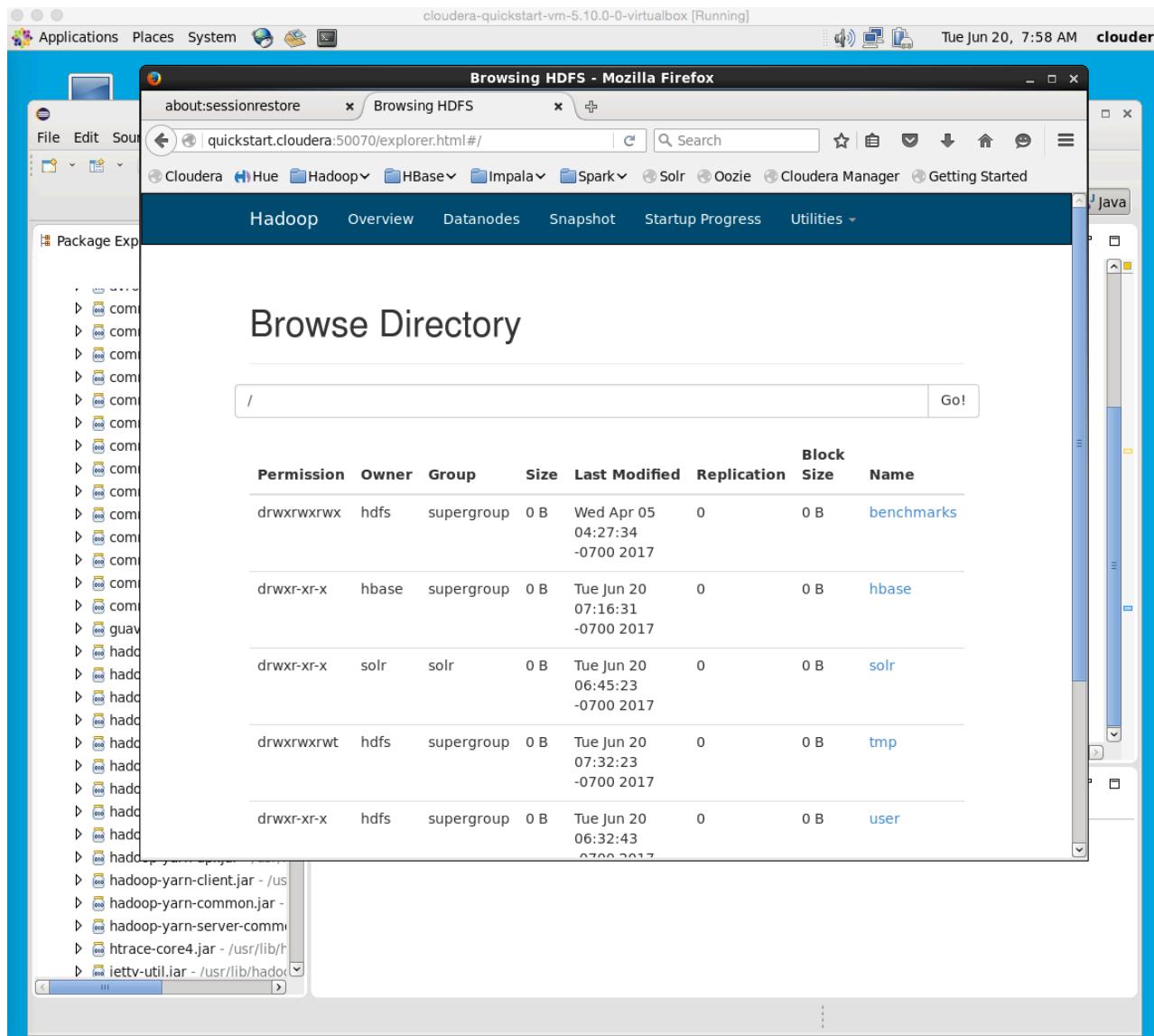
The screenshot shows a Firefox browser window titled "Namenode information - Mozilla Firefox" running on a Cloudera Quickstart VM. The URL is `quickstart.cloudera:50070/dfshealth.html#tab-datanode`. The page displays "Datanode Information" with a table listing a single datanode.

Node	Last contact	Capacity	Blocks	Block pool used	Version
quickstart.cloudera (10.0.2.15:50010)	Tue Jun 20 07:56:48 -0700 2017	54.51 GB	953	801.3 MB (1.44%)	2.6.0-cdh5.10.0

Below the table, it says "Showing 1 to 1 of 1 entries".

The browser interface includes a sidebar with a "Package Explorer" showing various Java packages like com, quickstart, hadoop, and org. A status bar at the bottom indicates "cloudera-quickstart-vm-5.10.0-0-virtualbox [Running]" and the date "Tue Jun 20, 7:57 AM".

Datanode



HBase

The screenshot shows a Mozilla Firefox browser window running on a Cloudera quickstart VM. The title bar indicates the session is running. The main content is the HBase Region Server status page.

RegionServer quickstart.cloudera,60020,1497968133633

Server Metrics

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

Attribute	Value	Description
Attribute	Value	Description

Cloudera Impala – in-memory database

The screenshot shows the Cloudera Impala web interface. On the left, there's a sidebar titled "Package Exp" with a tree view of various components like "com", "guav", "hadoop", etc. The main content area has a header "impalad" and a navigation bar with links: /backends, /catalog, /hadoop-varz, /logs, /memz, /metrics, /queries, /rpcz, /sessions, /threadz, and /varz.

Queries

This page lists all running queries, plus any completed queries that are archived in memory. The size of that archive is controlled with the --query_log_size command line parameter.

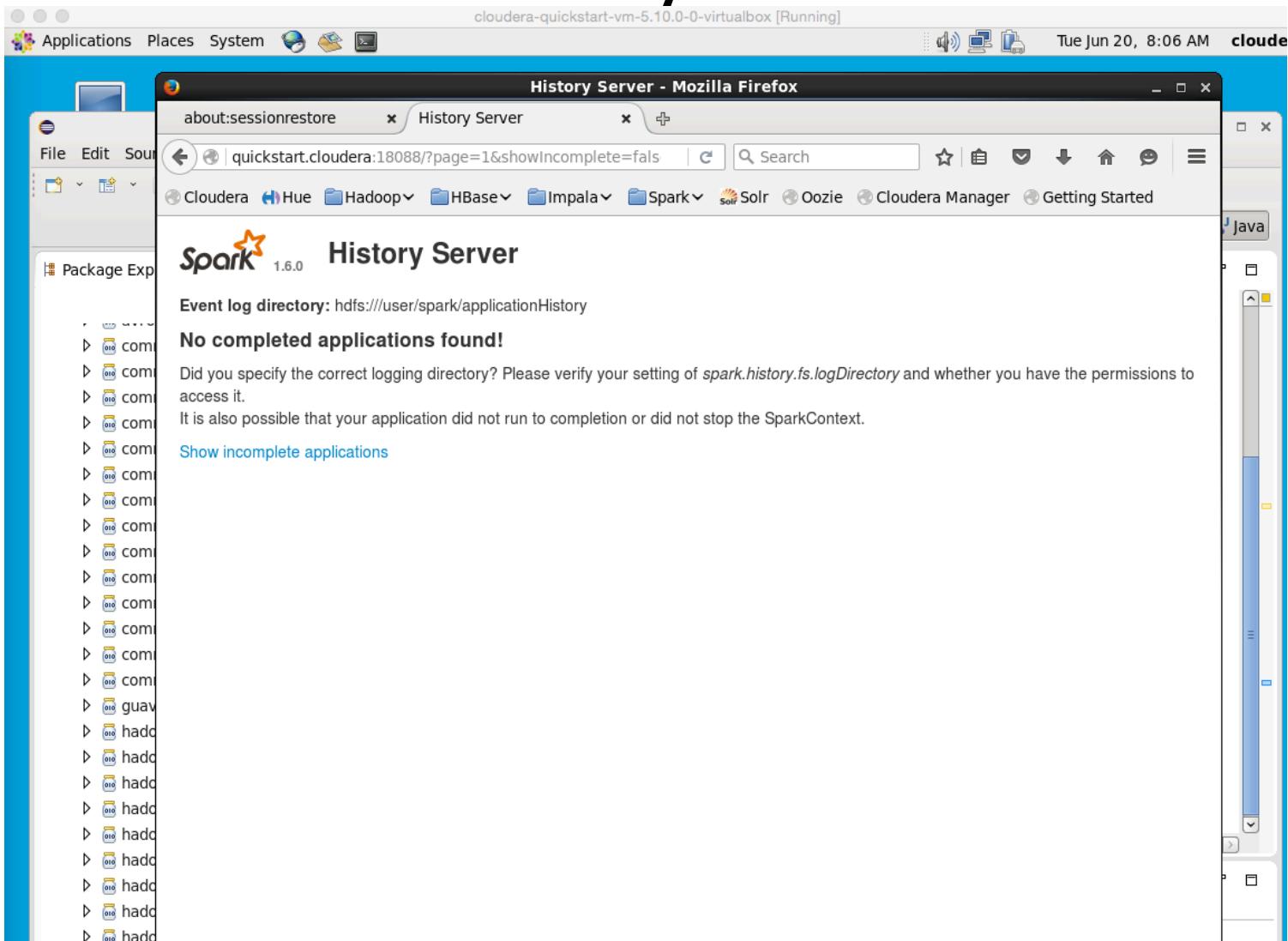
0 queries in flight

User	Default Db	Statement Type	Query	Start Time	Duration	Scan Progress	State	Last Event	# rows fetched	Resource Pool	Details	Action
cloudera	default	SELECT	QUERY	2017-06-20 07:32:31.563149000	30m46s	30m47s	0 / 0 (0%)	EXCEPTION	Cancelled			

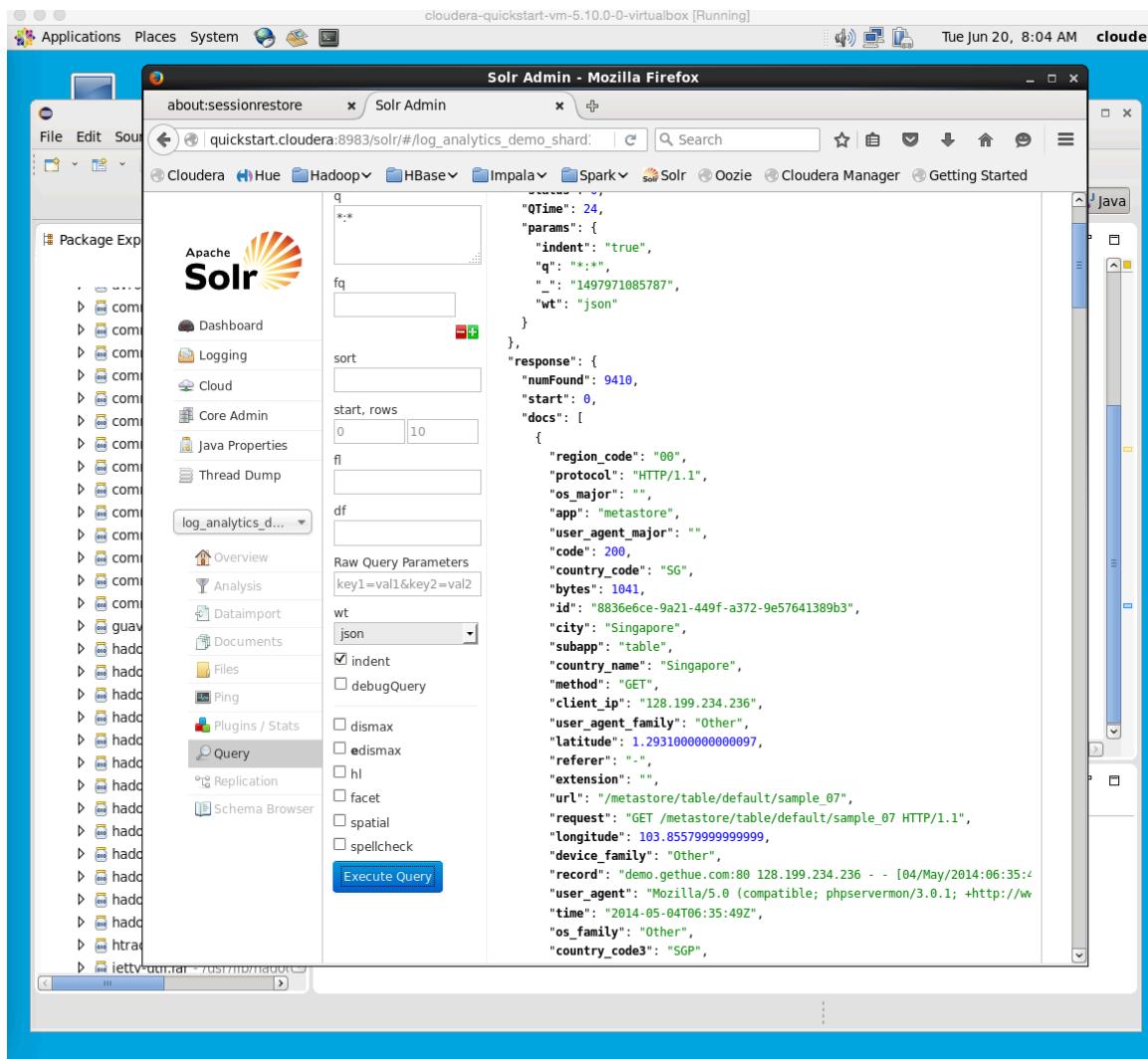
1 waiting to be closed [?]

User	Default Db	Statement Type	Query	Start Time	Waiting Time	Duration	Scan Progress	State	Last Event
cloudera	default	SELECT	QUERY	2017-06-20 07:32:31.563149000	30m46s	30m47s	0 / 0 (0%)	EXCEPTION	Cancelled

Spark – in-memory processing & analytics



Solr Search



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