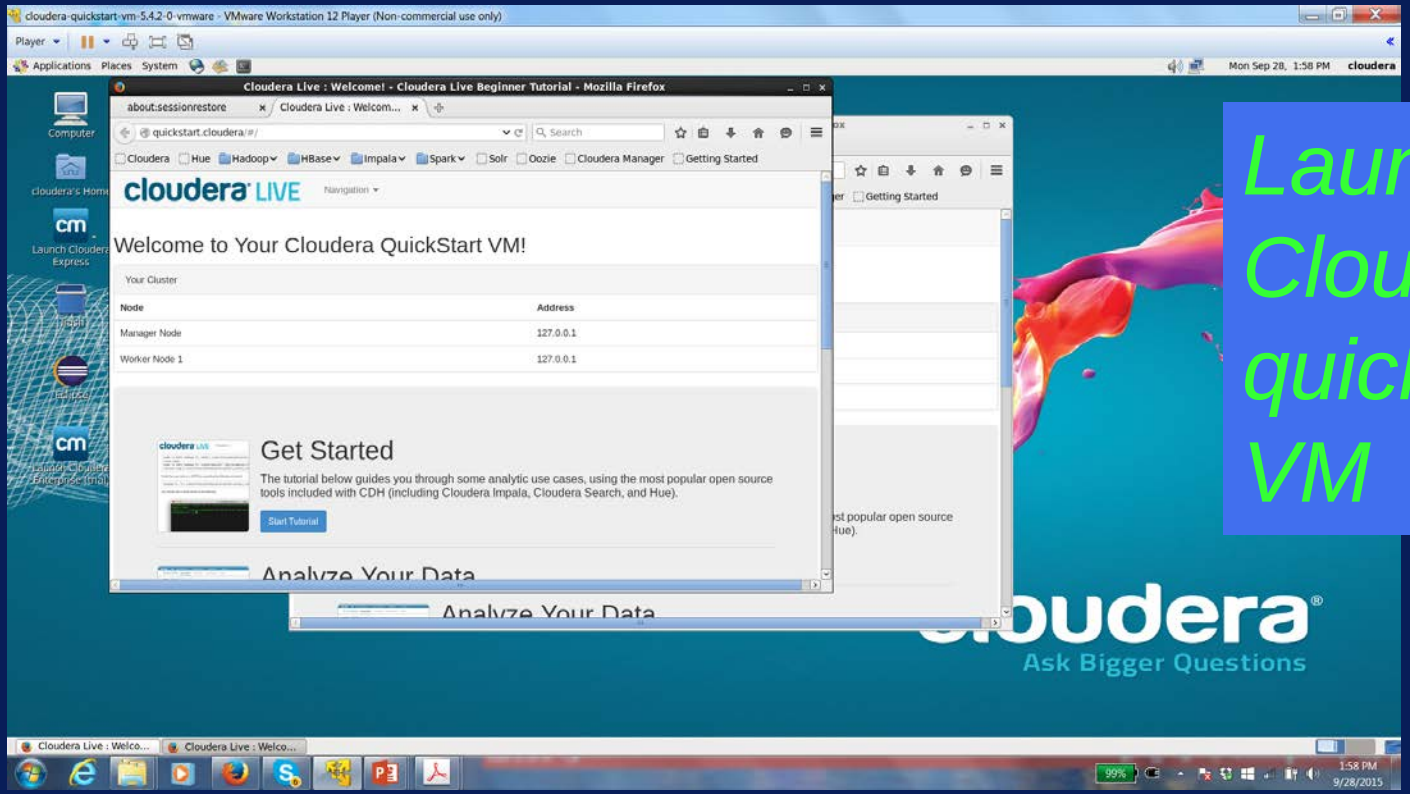
Introduction to Big Data

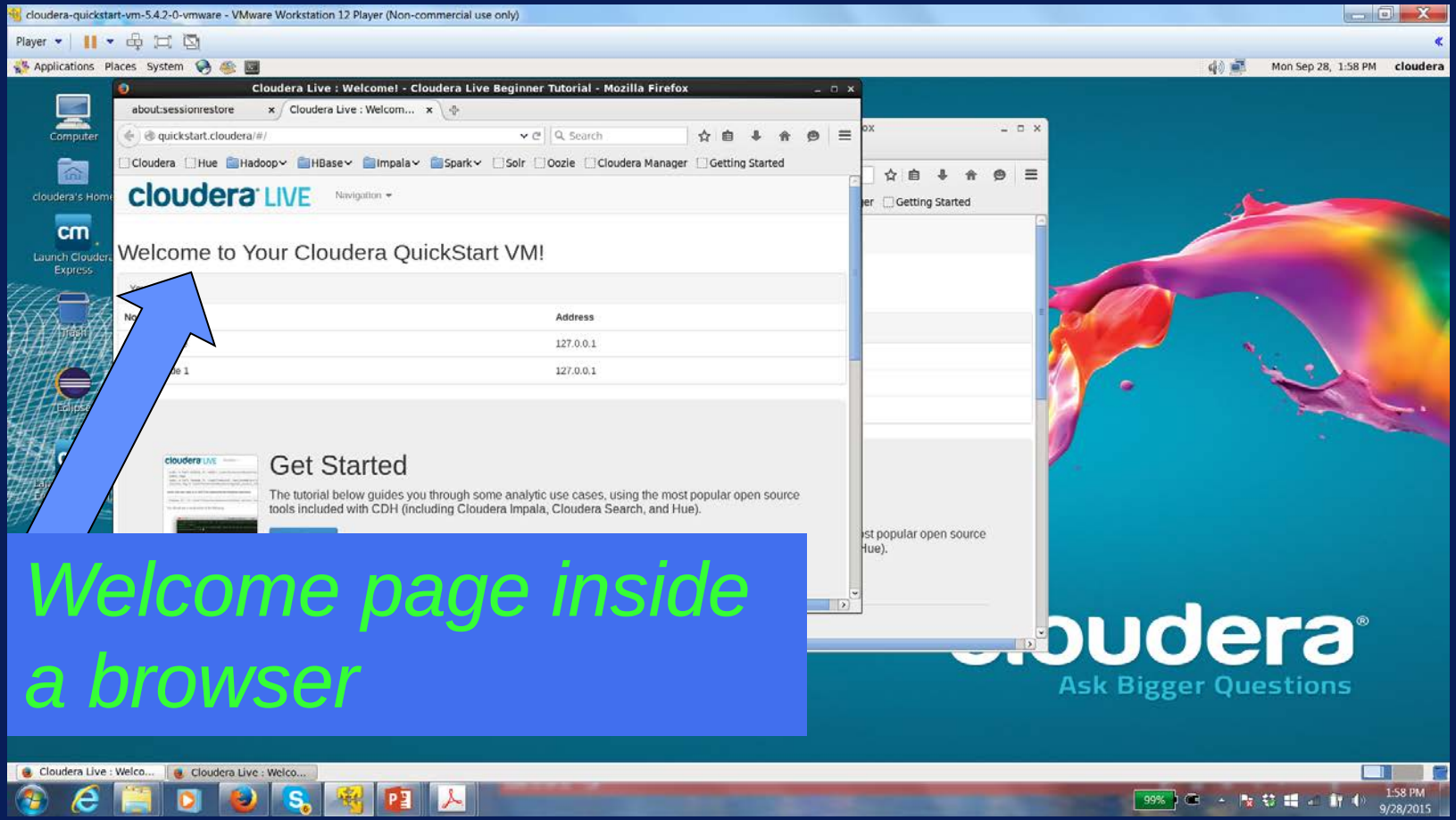
Lecture #1

Cloudera VM Tour

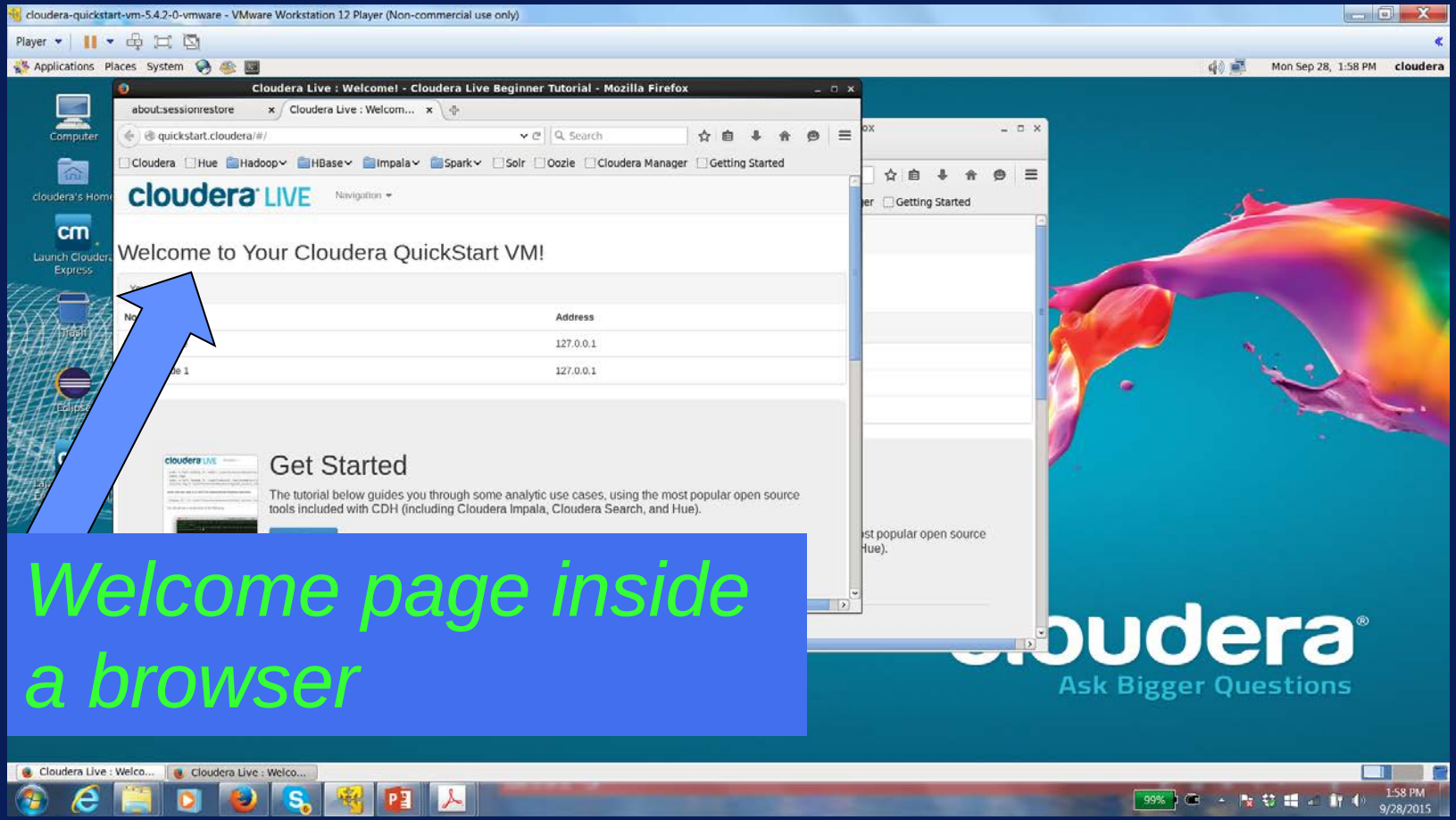


Launch the Cloudera quick start VM

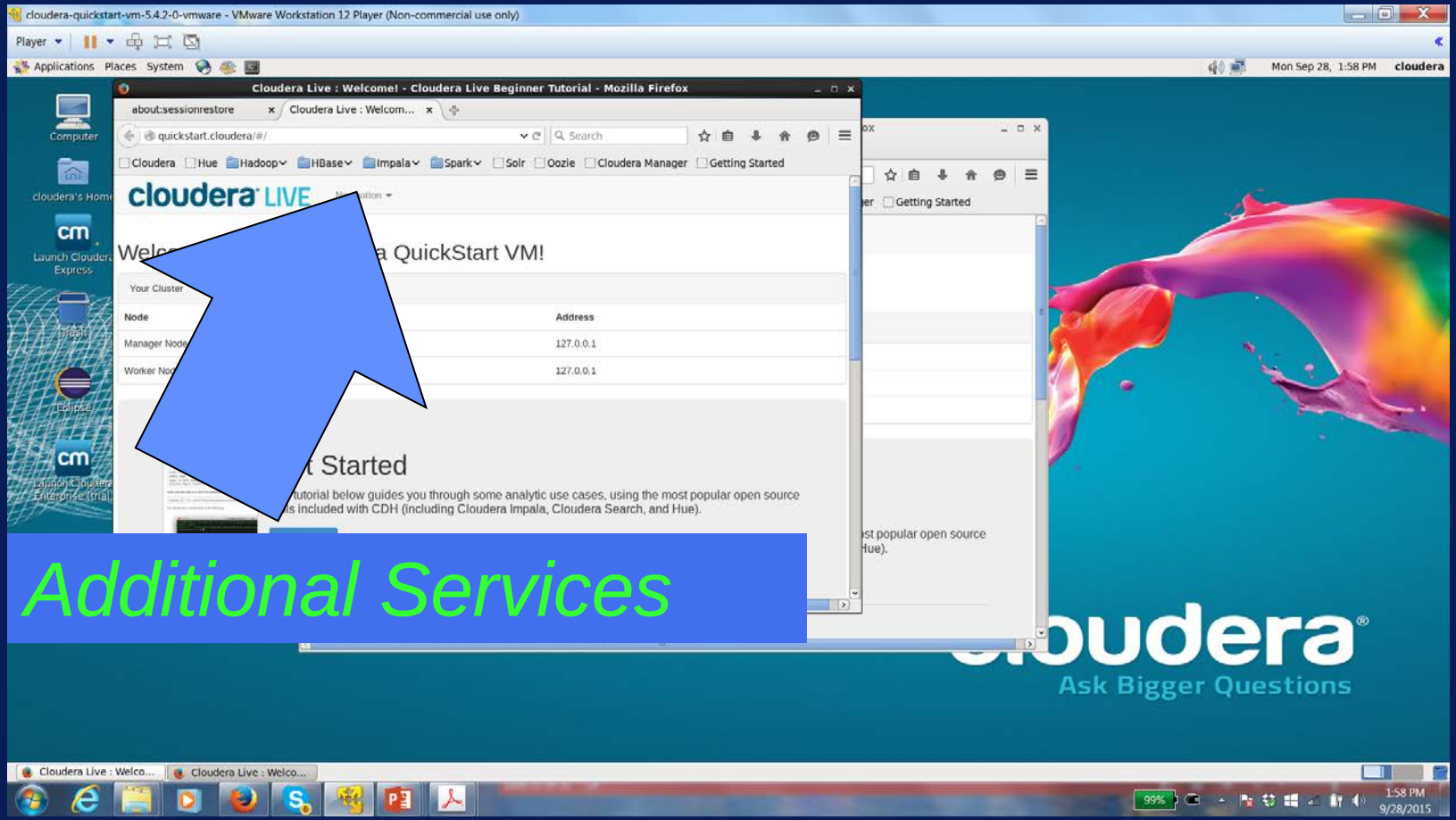
Inside the VM

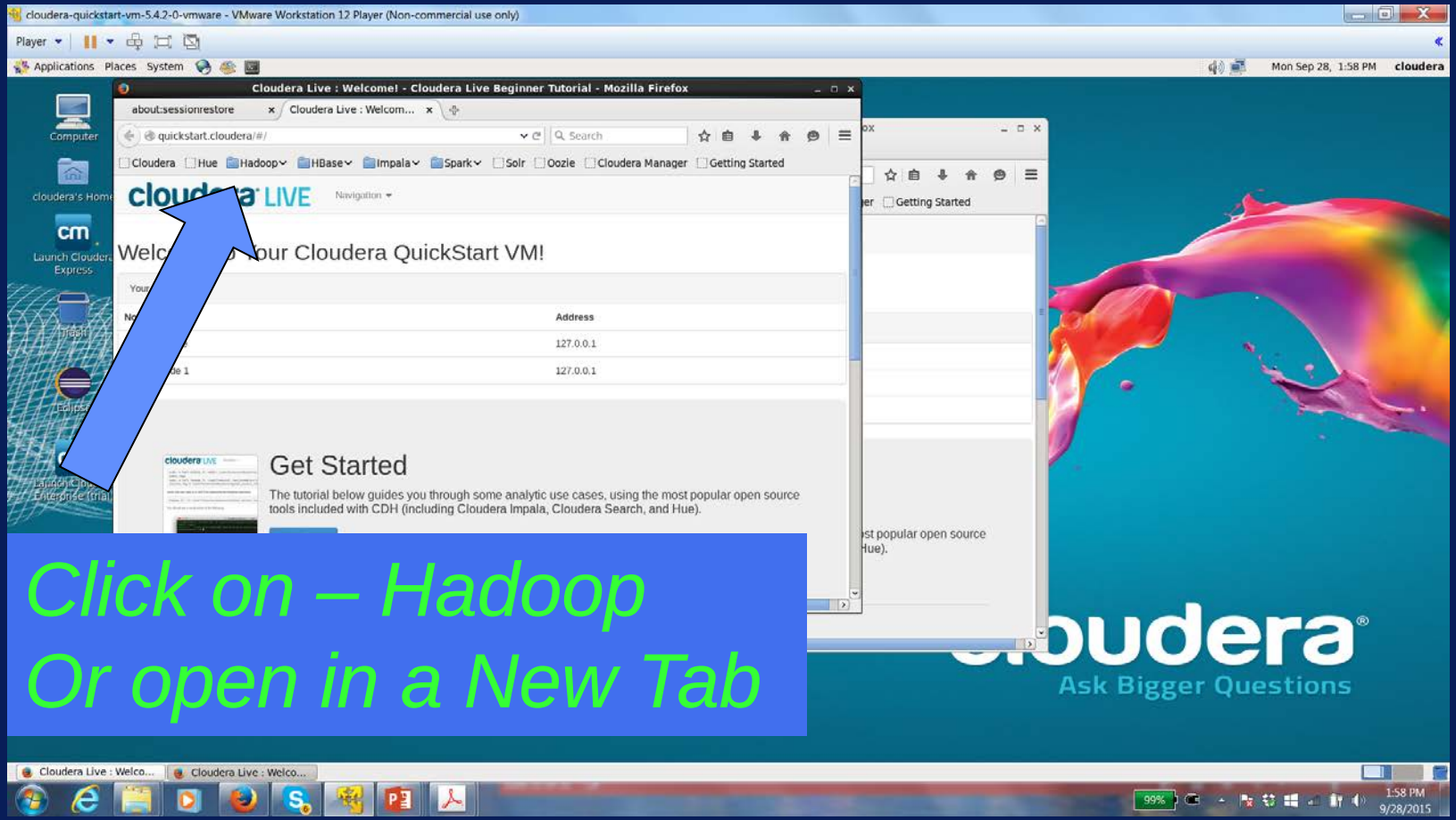


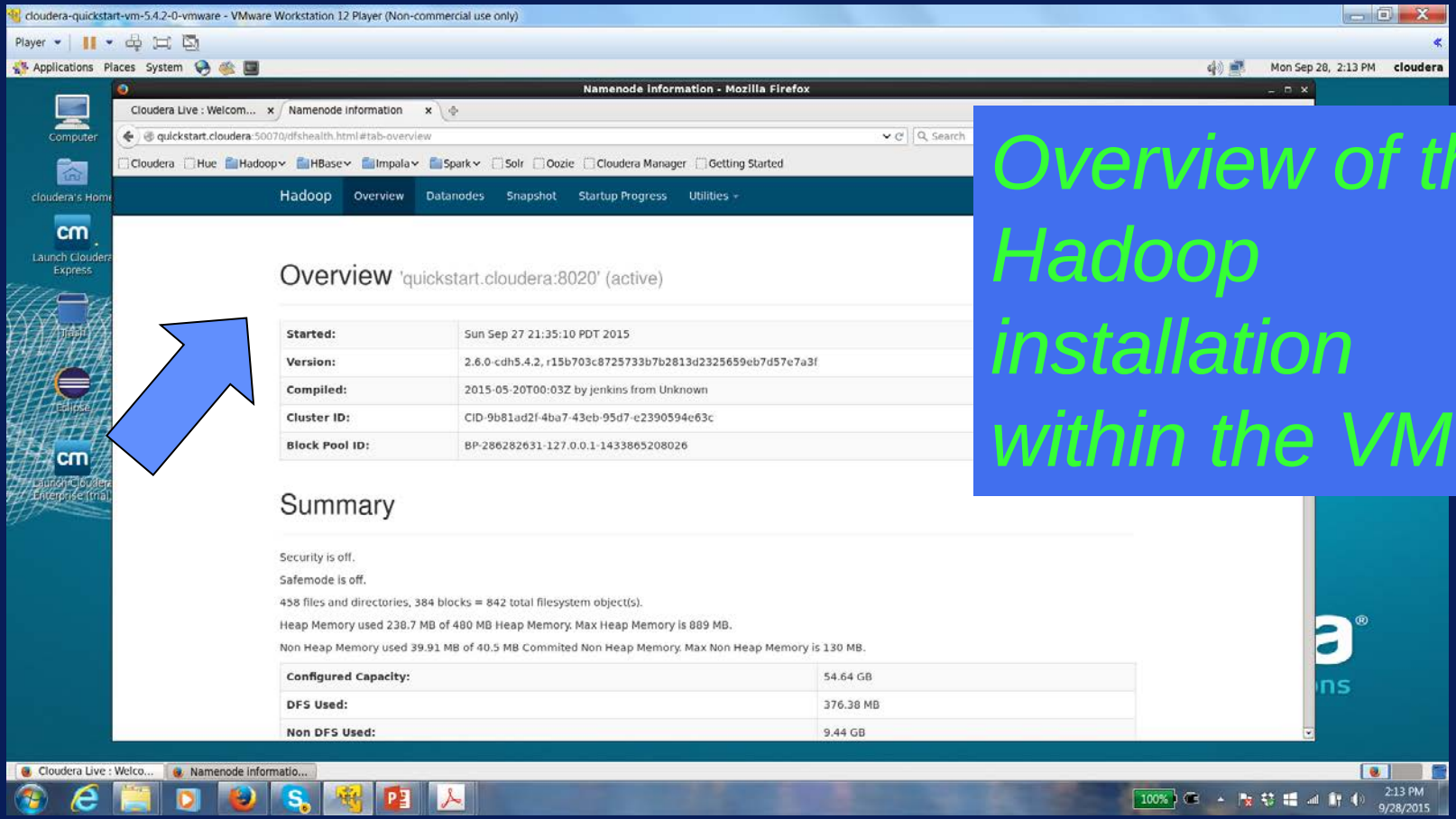
Welcome page inside a browser



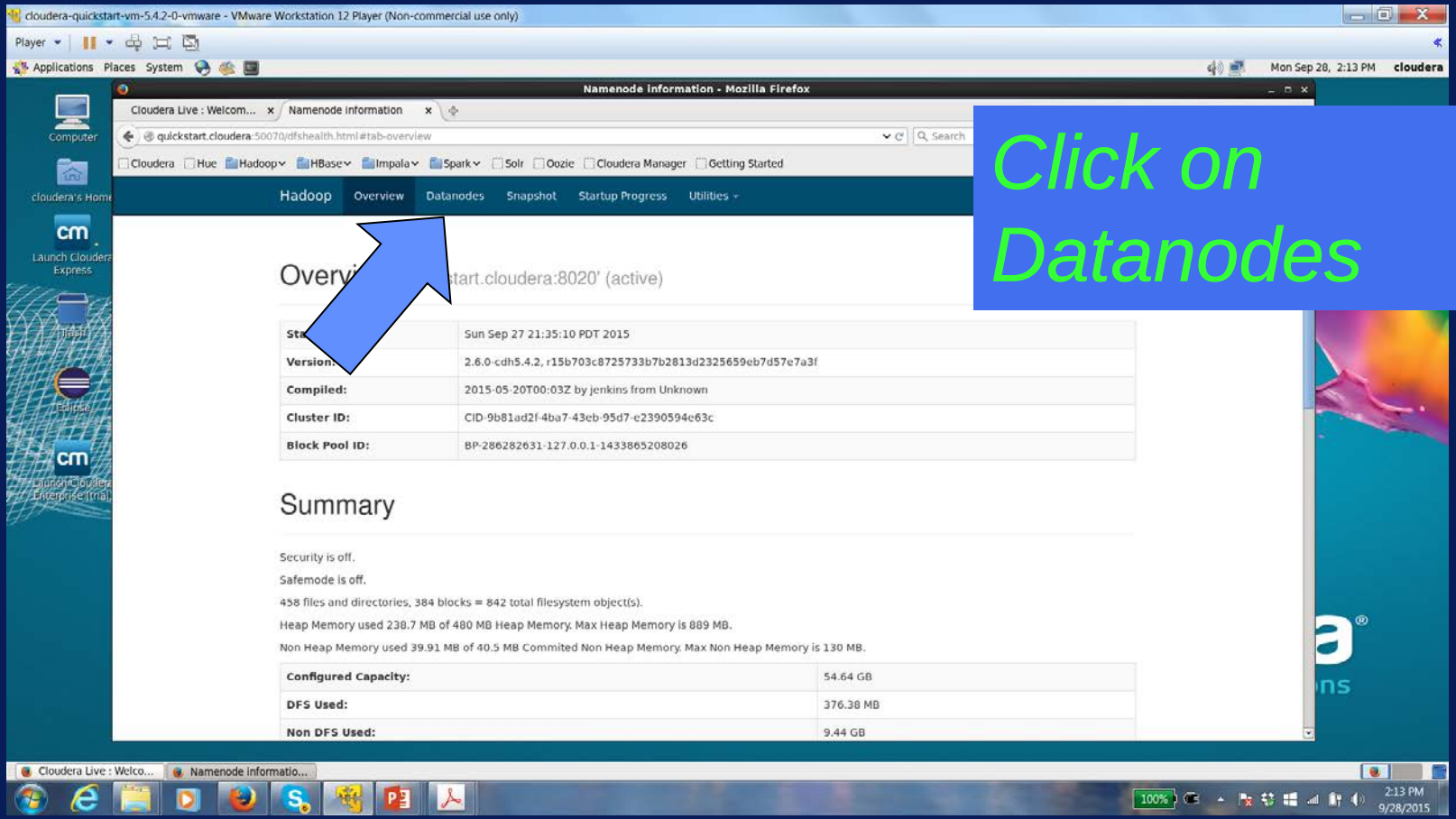
Welcome page inside a browser

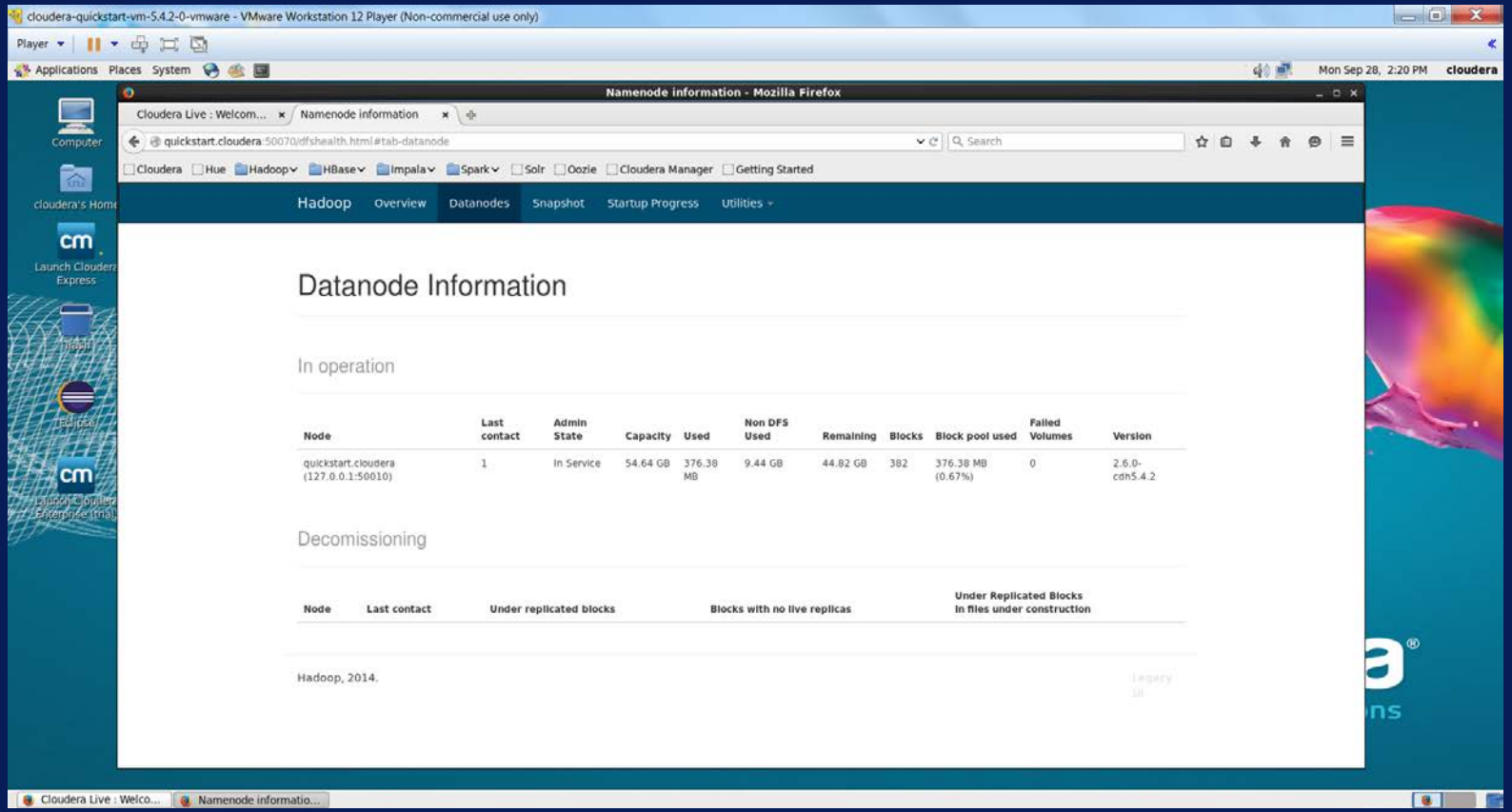

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 1:58 PM cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
about:sessionrestore
Cloudera Live : Welcom...
quickstart.cloudera/#/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
cloudera LIVE
a QuickStart VM!
Your Cluster
Node
Address
Manager Node
127.0.0.1
Worker Node
127.0.0.1
t Started
tutorial below guides you through some analytic use cases, using the most popular open source
s included with CDH (including Cloudera Impala, Cloudera Search, and Hue).
Computer
cloudera's Home
Launch Cloudera Express
Trash
Eclipse
Launch Cloudera Enterprise Trial
Additional Services
cloudera
Ask Bigger Questions
Cloudera Live : Welco...
Cloudera Live : Welco...
99%
1:58 PM
9/28/2015

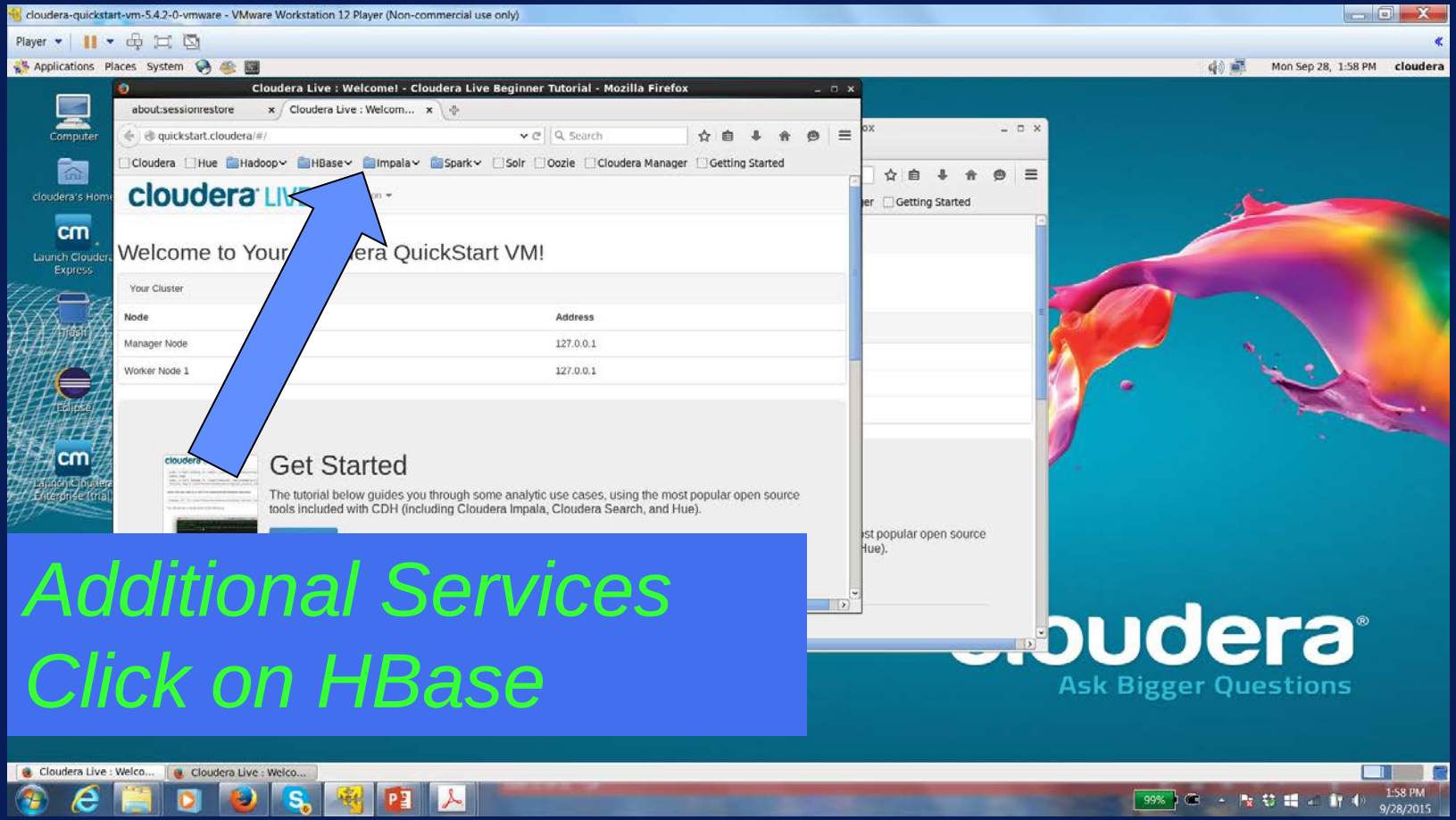

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 1:58 PM cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
about:sessionrestore
Cloudera Live : Welcom...
quickstart.cloudera/#/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
cloudera LIVE
Navigation
Address
127.0.0.1
127.0.0.1
Get Started
The tutorial below guides you through some analytic use cases, using the most popular open source tools included with CDH (including Cloudera Impala, Cloudera Search, and Hue).
Computer
cloudera's Home
Launch Cloudera Express
Trash
Launch Cloudera Enterprise (trial)
Click on – Hadoop
Or open in a New Tab
cloudera
Ask Bigger Questions
Cloudera Live : Welco...
Cloudera Live : Welco...
99%
1:58 PM
9/28/2015



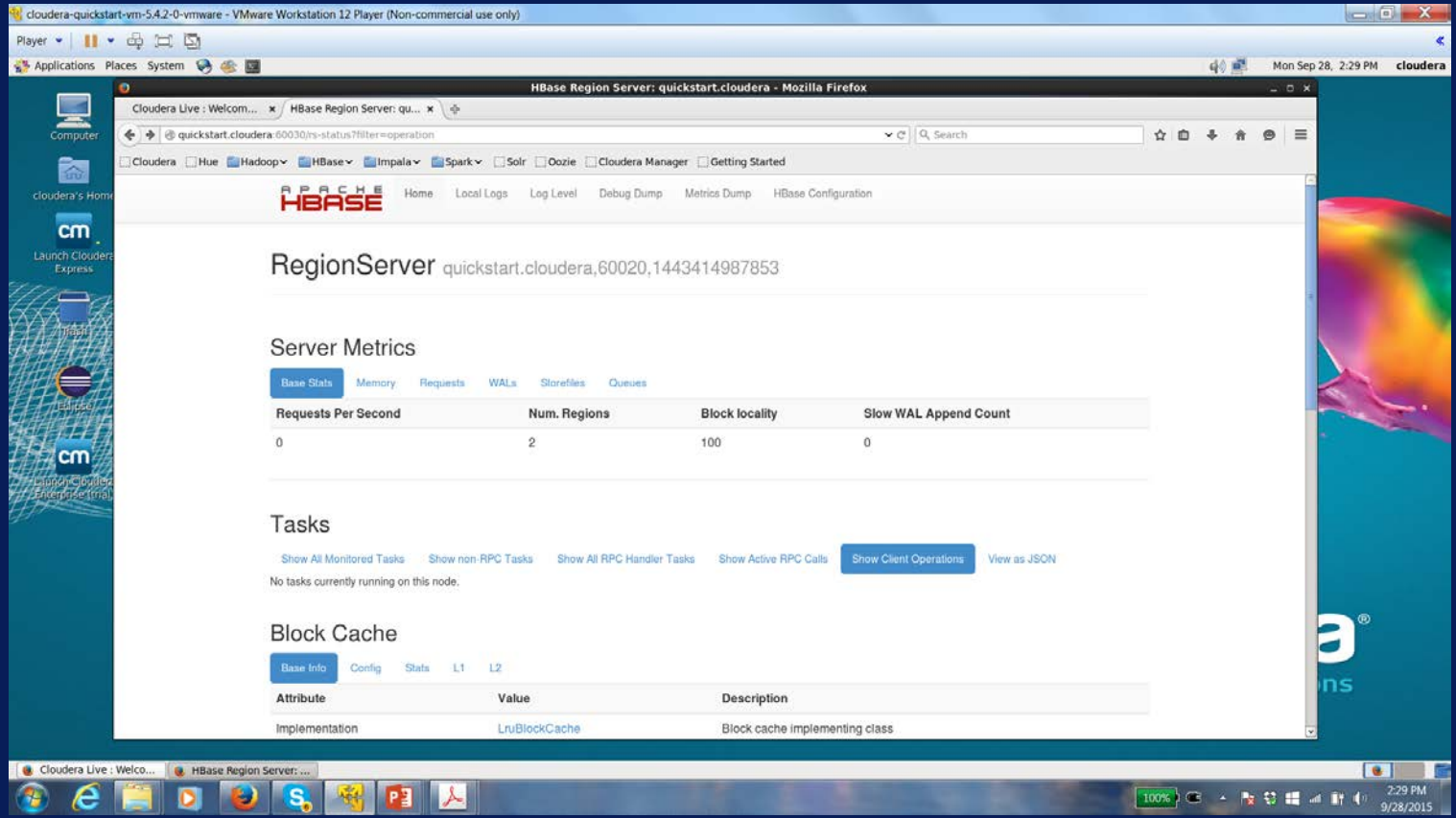
Overview of the Hadoop installation within the VM

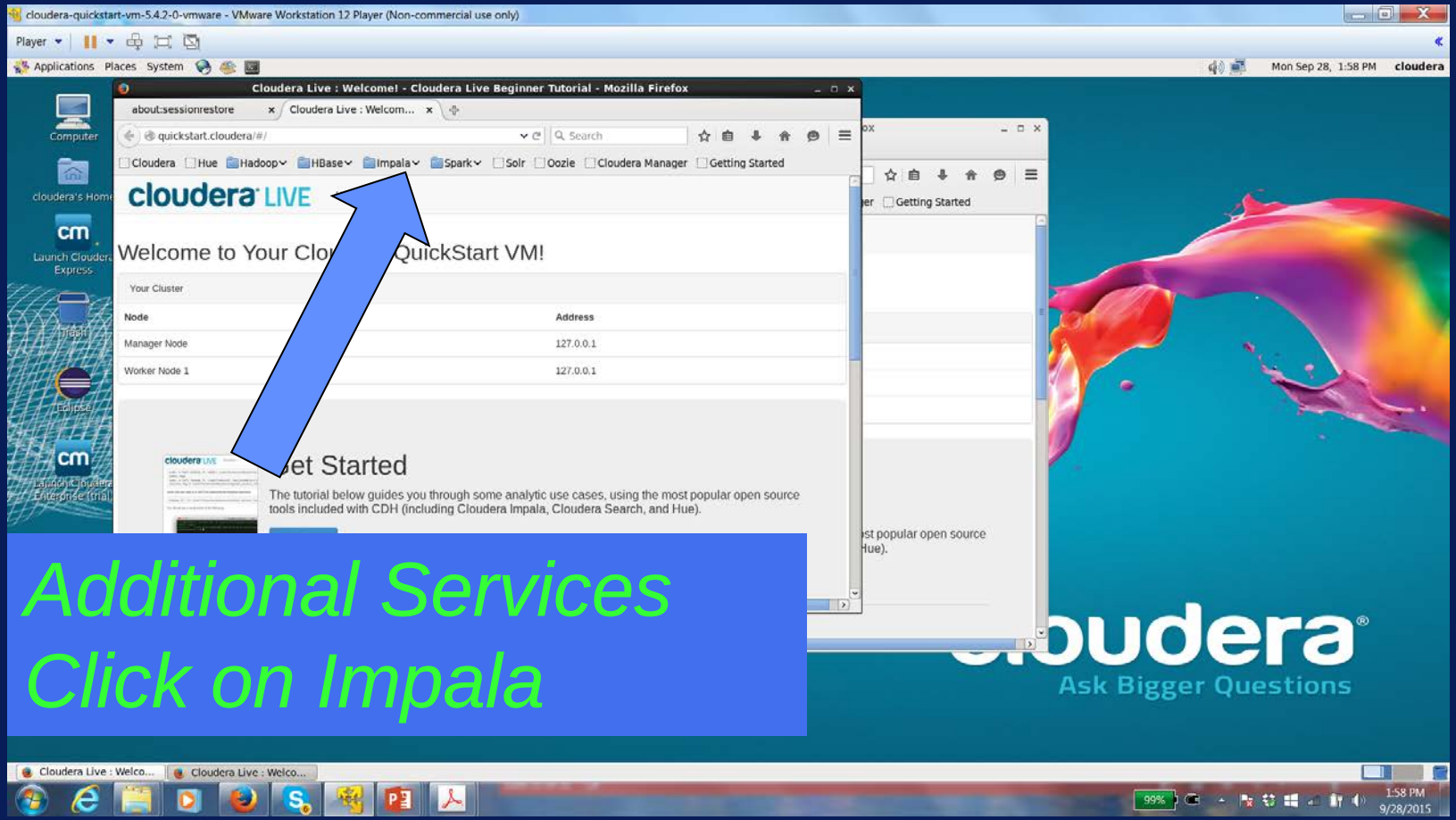

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 2:13 PM cloudera
Namenode Information - Mozilla Firefox
Cloudera Live : Welcom...
Namenode Information
quickstart.cloudera:50070/dfshealth.html#tab-overview
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Hadoop Overview Datanodes Snapshot Startup Progress Utilities
Sun Sep 27 21:35:10 PDT 2015
2.6.0-cdh5.4.2, r15b703c8725733b7b2813d2325659eb7d57e7a3f
2015-05-20T00:03Z by jenkins from Unknown
CID-9b81ad2f-4ba7-43eb-95d7-e2390594e63c
BP-286282631-127.0.0.1-1433865208026
Summary
Security is off.
Safemode is off.
458 files and directories, 384 blocks = 842 total filesystem object(s).
Heap Memory used 238.7 MB of 480 MB Heap Memory. Max Heap Memory is 889 MB.
Non Heap Memory used 39.91 MB of 40.5 MB Commited Non Heap Memory. Max Non Heap Memory is 130 MB.
Configured Capacity: 54.64 GB
DFS Used: 376.38 MB
Non DFS Used: 9.44 GB
Click on Datanodes
2:13 PM
9/28/2015

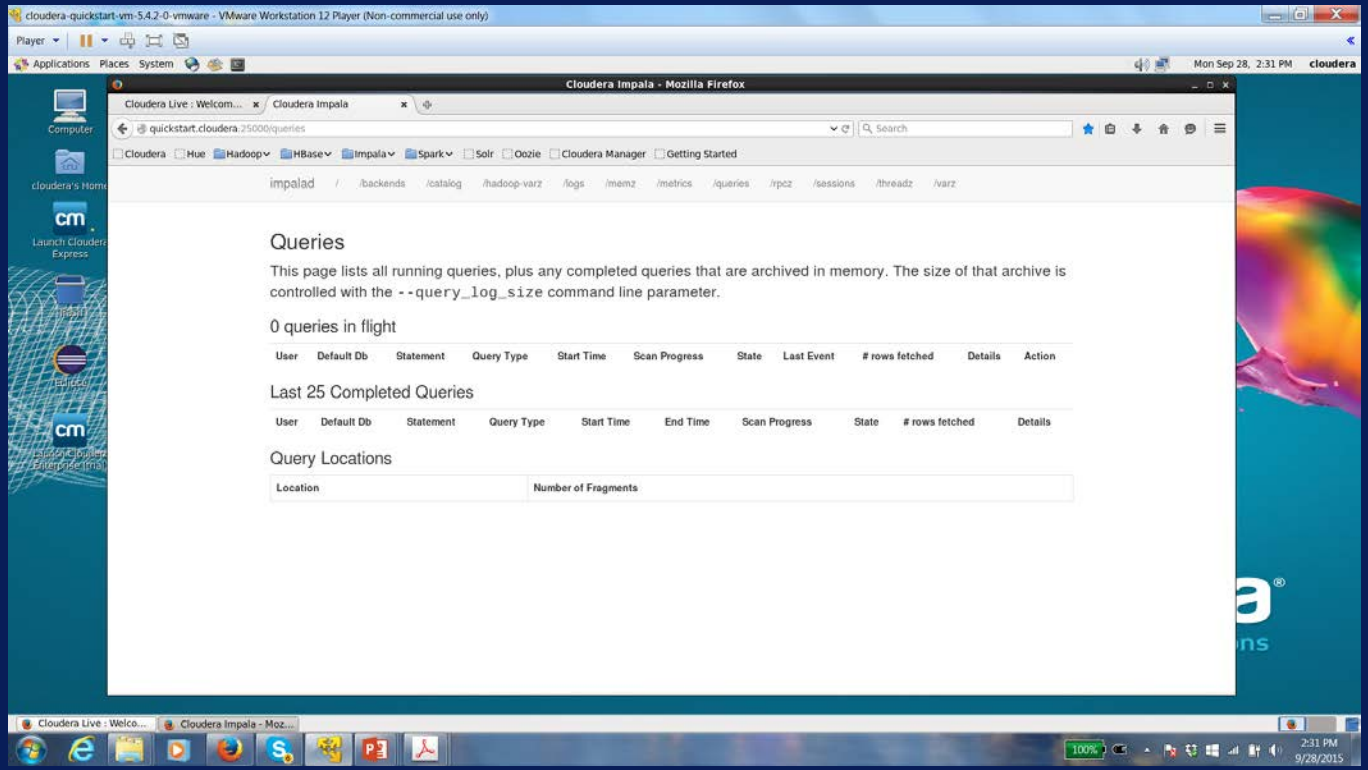

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 2:20 PM cloudera
Namenode information - Mozilla Firefox
Cloudera Live : Welcom...
Namenode information
quickstart.cloudera:50070/dfshealth.html#tab-datanode
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Hadoop Overview Datanodes Snapshot Startup Progress Utilities
Datanode Information
In operation
Node Last contact Admin State Capacity Used Non DFS Used Remaining Blocks Block pool used Failed Volumes Version
quickstart.cloudera (127.0.0.1:50010) 1 In Service 54.64 GB 376.38 MB 9.44 GB 44.82 GB 382 376.38 MB (0.67%) 0 2.6.0-cdh5.4.2
Decomissioning
Node Last contact Under replicated blocks Blocks with no live replicas Under Replicated Blocks In files under construction
Hadoop, 2014.
Computer
cloudera's Home
Launch Cloudera Express
Eclipse
Launch Cloudera Enterprise (trial)
Cloudera Live : Welco...
Namenode informatio...

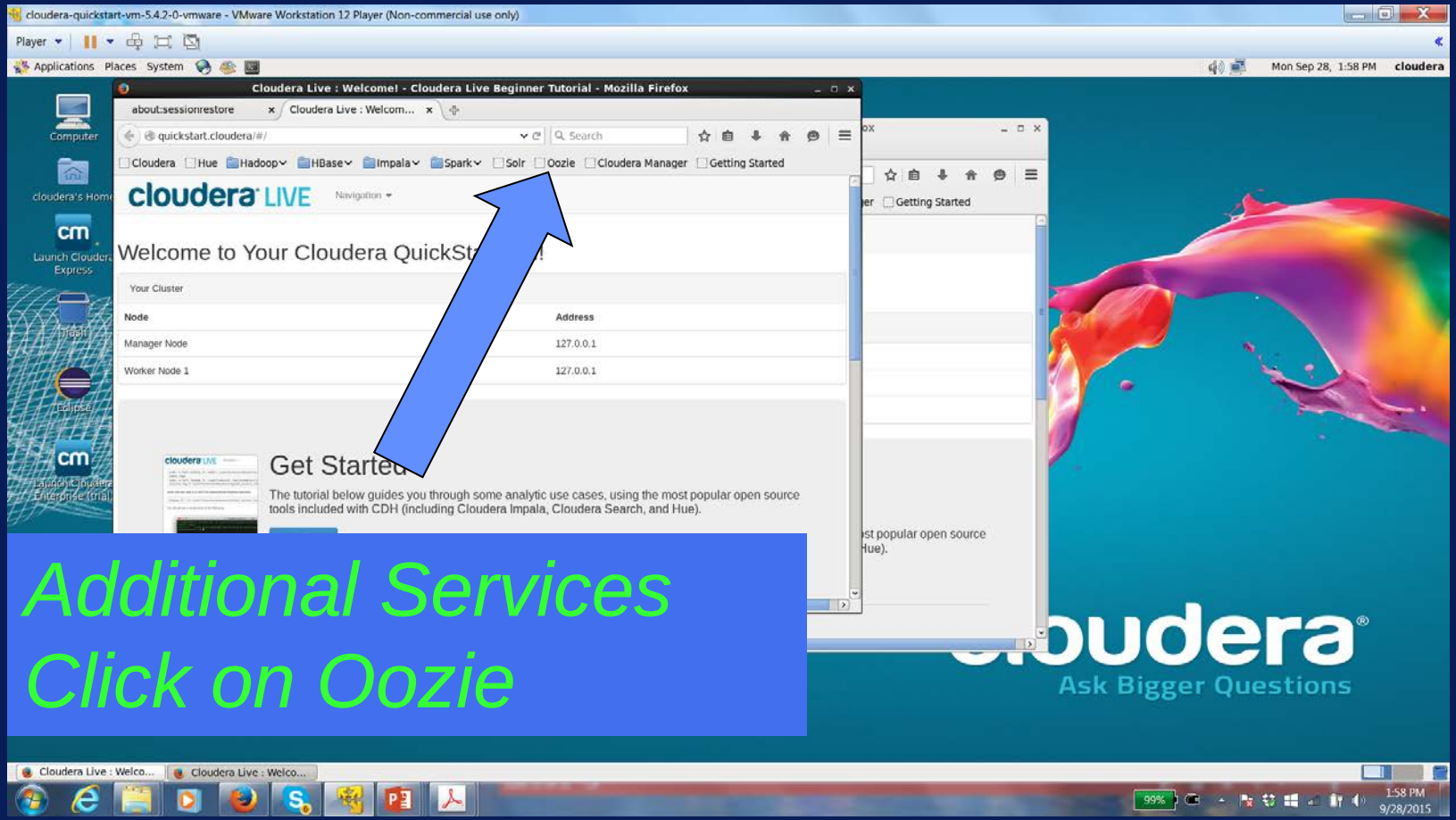


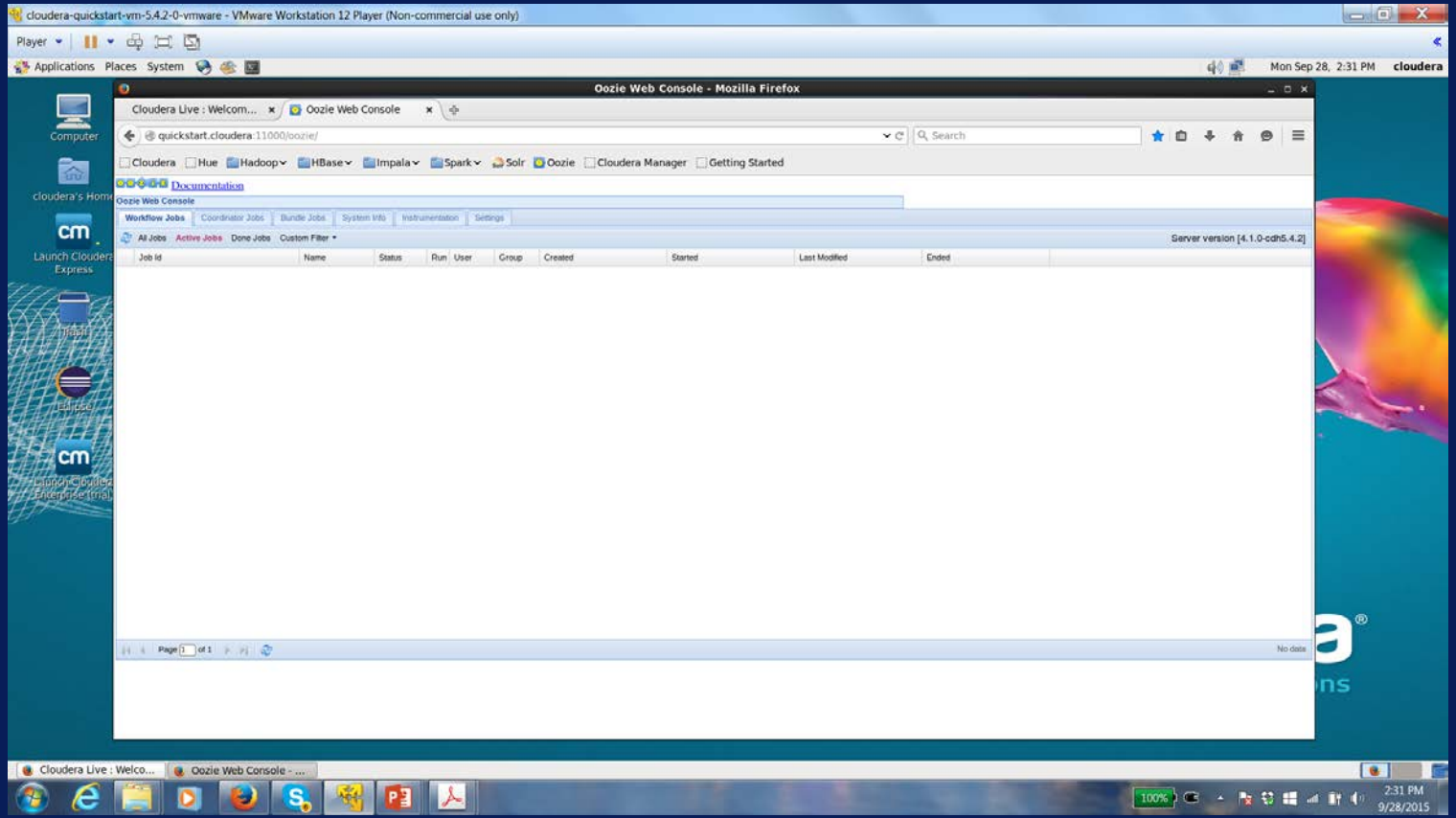
Additional Services
Click on HBase

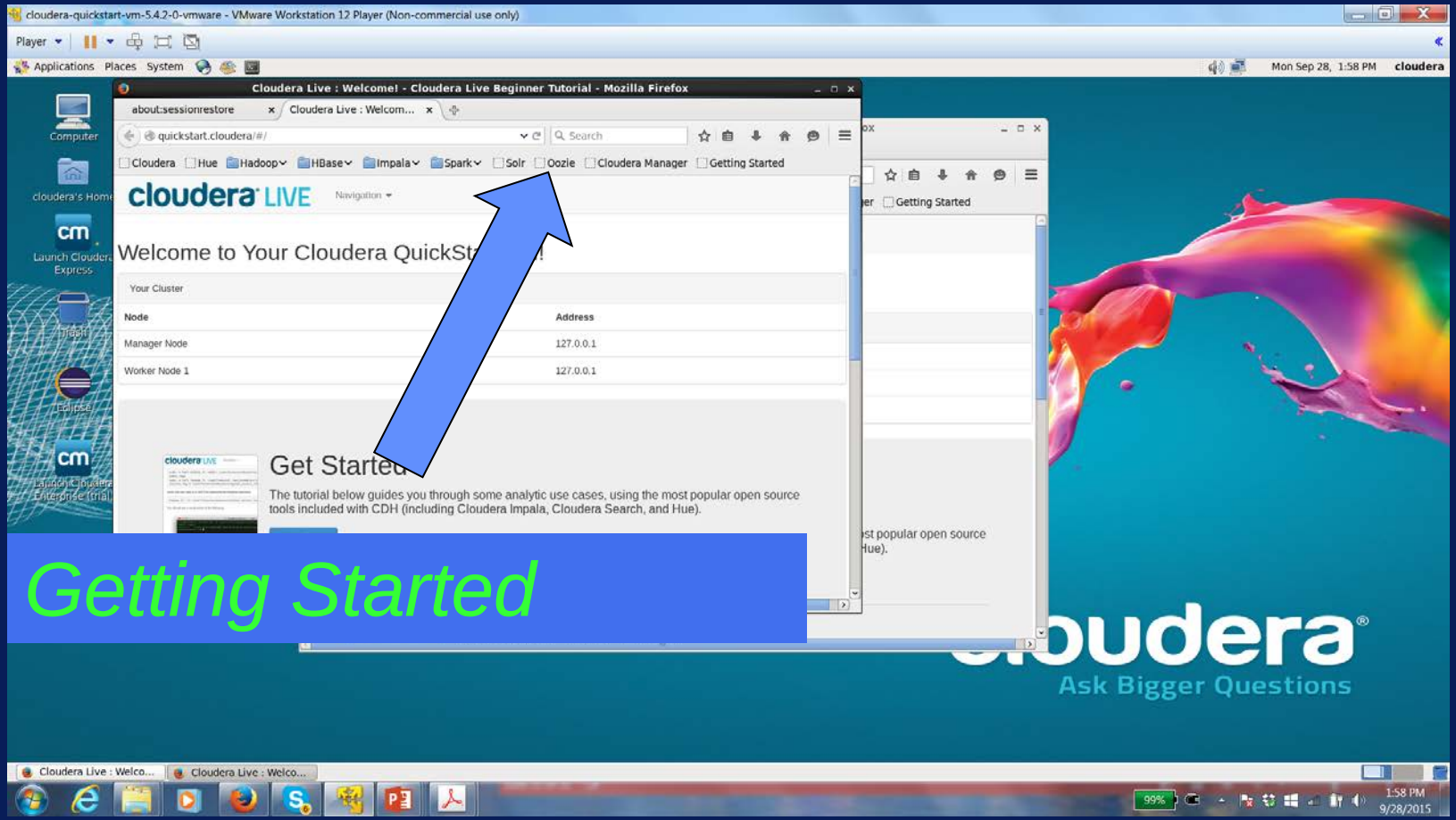

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 2:29 PM cloudera
HBase Region Server: quickstart.cloudera - Mozilla Firefox
Cloudera Live : Welcom...
HBase Region Server: qu...
quickstart.cloudera:60030/rs-status?filter=operation
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,60020,1443414987853
Server Metrics
Base Stats Memory Requests WALs Storefiles Queues
Requests Per Second
Num. Regions
Block locality
Slow WAL Append Count
0
2
100
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
Implementation
LruBlockCache
Block cache implementing class
Computer
cloudera's Home
Launch Cloudera Express
Trash
Eclipse
Launch Cloudera Enterprise Trial
Cloudera Live : Welco...
HBase Region Server: ...
100%
2:29 PM
9/28/2015


cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 1:58 PM cloudera
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 Getting Started
cloudera LIVE
Welcome to Your Clou QuickStart VM!
Your Cluster
Node
Address
Manager Node
127.0.0.1
Worker Node 1
127.0.0.1
et Started
The tutorial below guides you through some analytic use cases, using the most popular open source tools included with CDH (including Cloudera Impala, Cloudera Search, and Hue).
Computer
cloudera's Home
Launch Cloudera Express
Trash
Eclipse
Launch Cloudera Enterprise Trial
Additional Services
Click on Impala
Ask Bigger Questions
Cloudera Live : Welco...
Cloudera Live : Welco...
99%
1:58 PM
9/28/2015

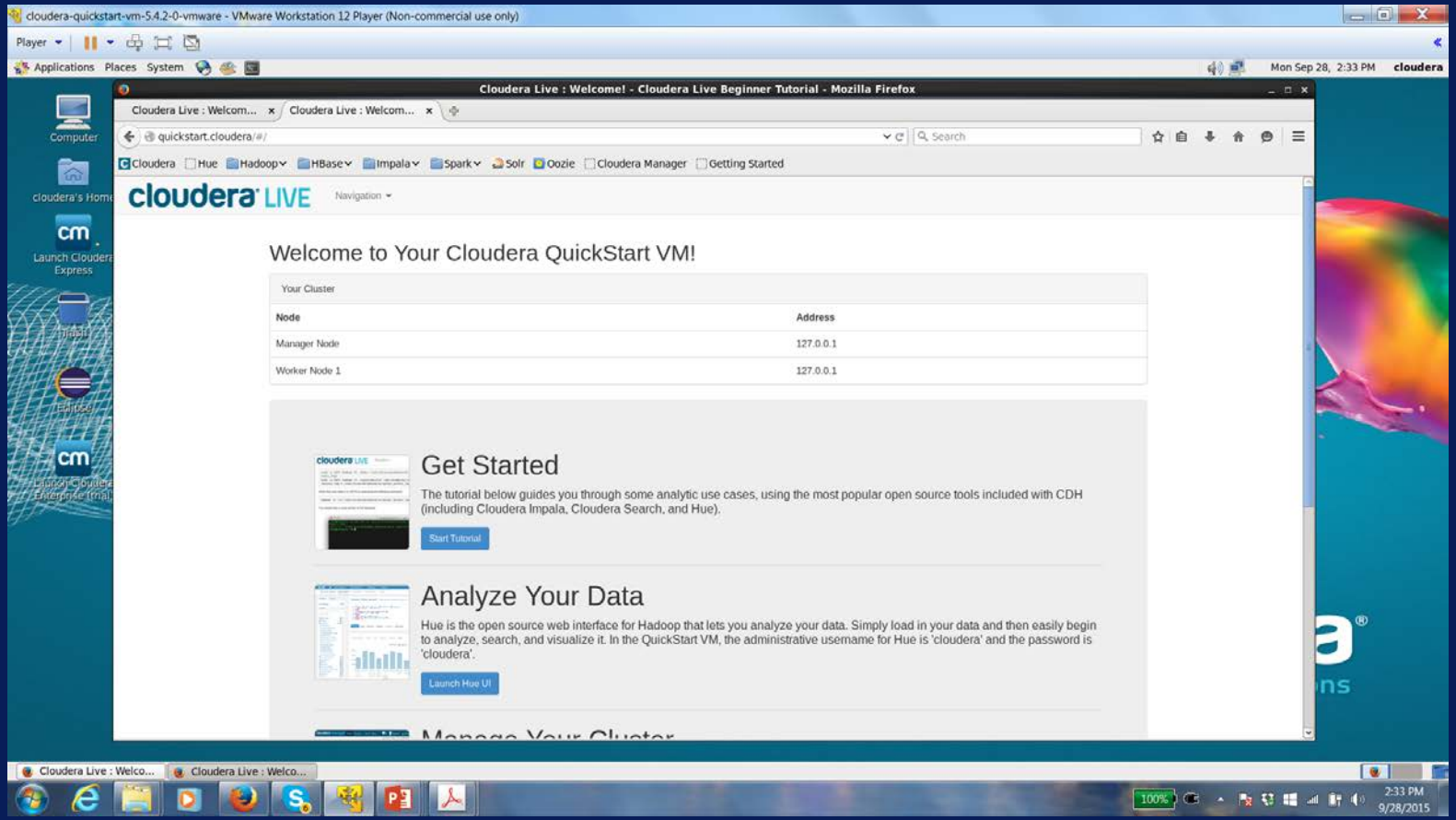

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 2:31 PM cloudera
Cloudera Impala - Mozilla Firefox
Cloudera Live : Welcom...
Cloudera Impala
quickstart.cloudera:25000/queries
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
impalad / /backends /catalog /hadoop-varz /logs /memz /metrics /queries /rpcz /sessions /threadz /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 Query Type Start Time Scan Progress State Last Event # rows fetched Details Action
Last 25 Completed Queries
User Default Db Statement Query Type Start Time End Time Scan Progress State # rows fetched Details
Query Locations
Location Number of Fragments
Computer
cloudera's Home
Launch Cloudera Express
Impala
Eclipse
Launch Cloudera Enterprise trial
Cloudera Live : Welco...
Cloudera Impala - Moz...
100%
2:31 PM
9/28/2015

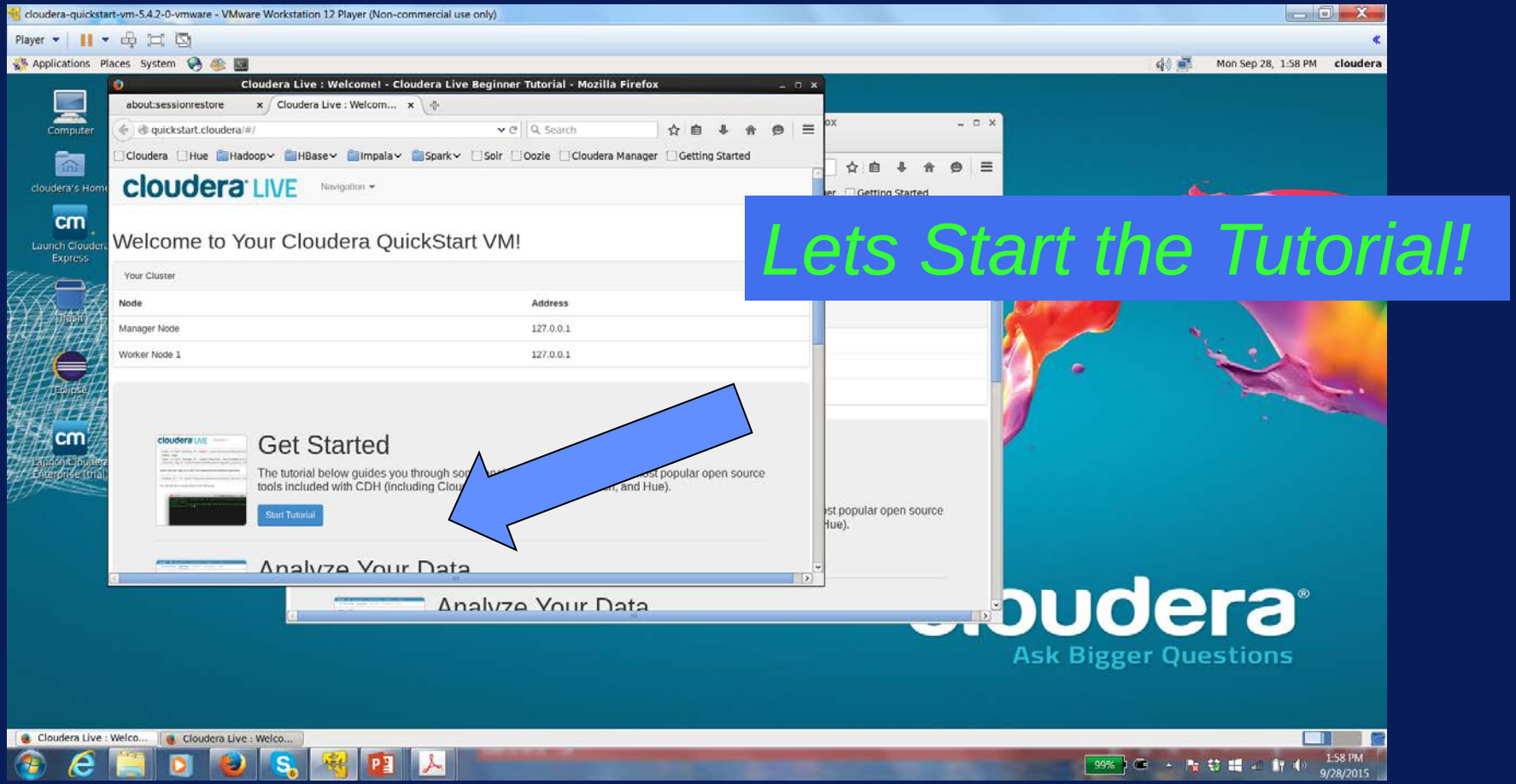

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 1:58 PM cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
about:sessionrestore
Cloudera Live : Welcom...
quickstart.cloudera/#/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
cloudera LIVE
Navigation
Welcome to Your Cloudera QuickSt
Your Cluster
Node
Address
Manager Node
127.0.0.1
Worker Node 1
127.0.0.1
Get Started
The tutorial below guides you through some analytic use cases, using the most popular open source tools included with CDH (including Cloudera Impala, Cloudera Search, and Hue).
Computer
cloudera's Home
Launch Cloudera Express
Trash
Eclipse
Launch Cloudera Enterprise trial
Additional Services
Click on Oozie
cloudera
Ask Bigger Questions
Cloudera Live : Welco...
Cloudera Live : Welco...
99%
1:58 PM
9/28/2015

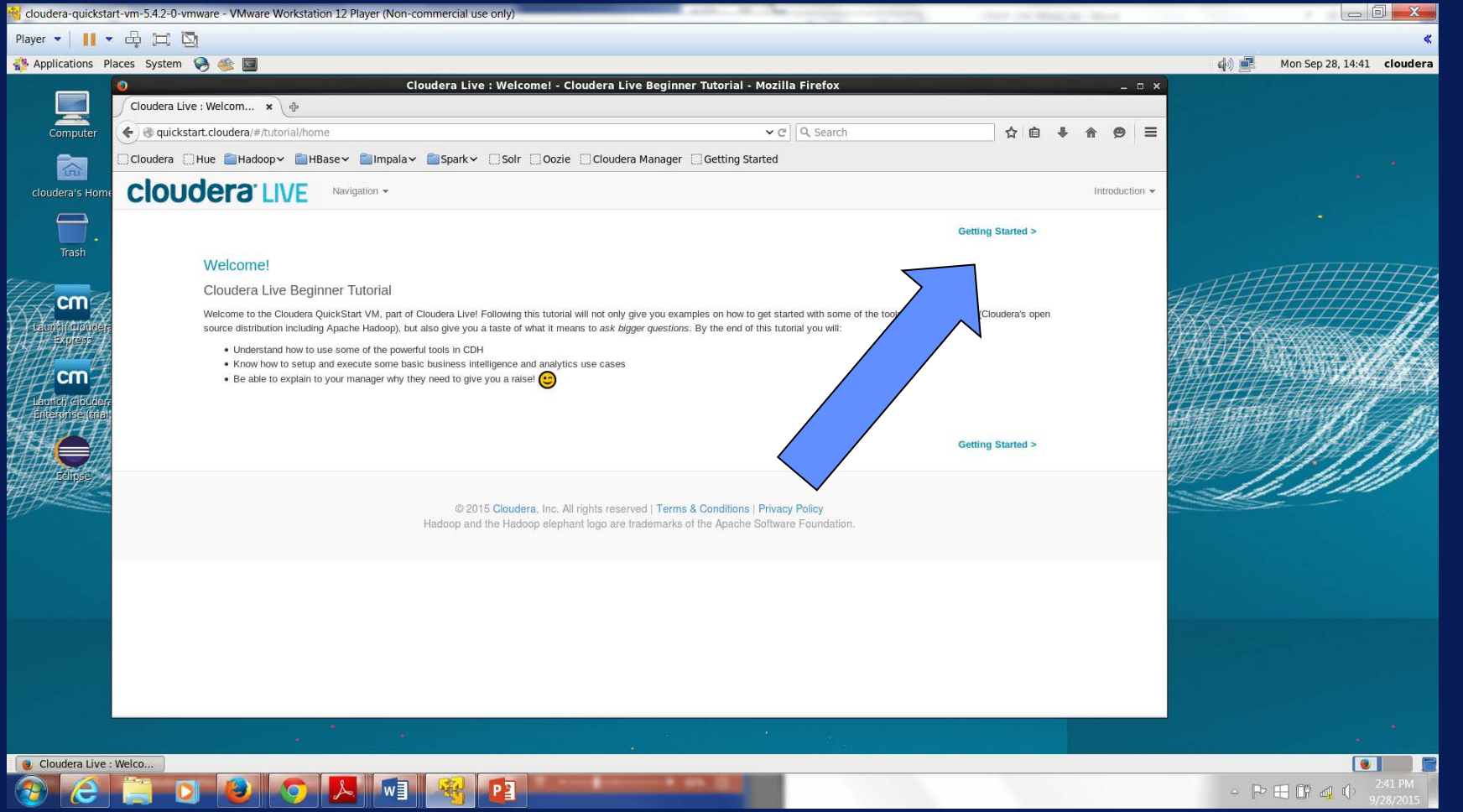

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 2:31 PM cloudera
Oozie Web Console - Mozilla Firefox
Cloudera Live : Welcom...
Oozie Web Console
quickstart.cloudera:11000/oozie/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Documentation
Oozie Web Console
Workflow Jobs Coordinator Jobs Bundle Jobs System Info Instrumentation Settings
All Jobs Active Jobs Done Jobs Custom Filter
Server version [4.1.0-cdh5.4.2]
Job Id Name Status Run User Group Created Started Last Modified Ended
Page 1 of 1
No data
Computer
cloudera's Home
Launch Cloudera Express
Trash
Eclipse
Launch Cloudera Enterprise (trial)
Cloudera Live : Welco...
Oozie Web Console - ...
100%
2:31 PM
9/28/2015

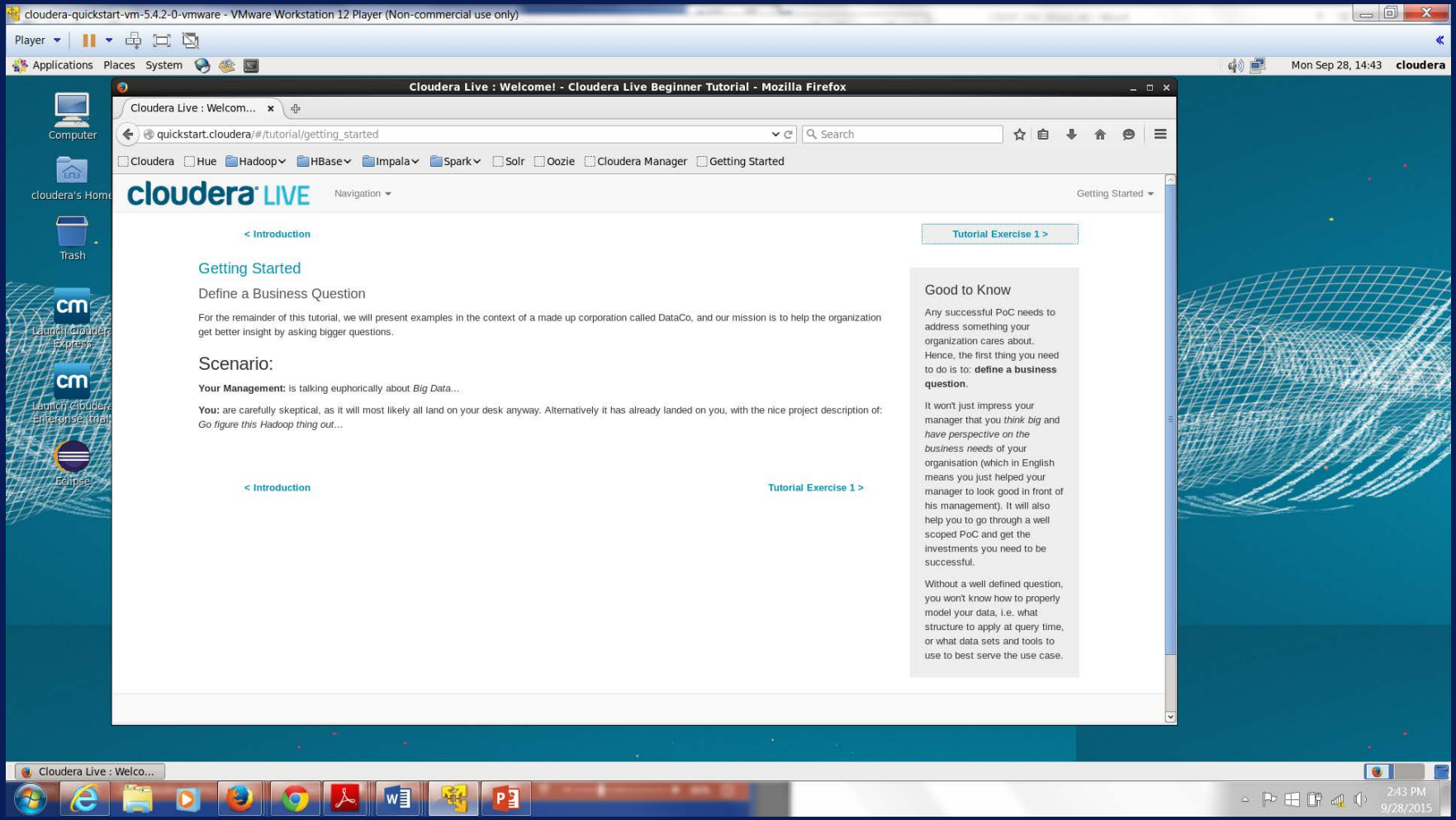


Getting Started


cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 2:33 PM cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
Cloudera Live : Welcom...
quickstart.cloudera/#/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
cloudera LIVE
Navigation
Welcome to Your Cloudera QuickStart VM!
Your Cluster
Node
Address
Manager Node
127.0.0.1
Worker Node 1
127.0.0.1
Get Started
The tutorial below guides you through some analytic use cases, using the most popular open source tools included with CDH (including Cloudera Impala, Cloudera Search, and Hue).
Start Tutorial
Analyze Your Data
Hue is the open source web interface for Hadoop that lets you analyze your data. Simply load in your data and then easily begin to analyze, search, and visualize it. In the QuickStart VM, the administrative username for Hue is 'cloudera' and the password is 'cloudera'.
Launch Hue UI
Computer
cloudera's Home
Launch Cloudera Express
Launch Cloudera Enterprise (trial)
Cloudera Live : Welco...
Cloudera Live : Welco...
100%
2:33 PM
9/28/2015


cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 1:58 PM cloudera
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 Getting Started
cloudera LIVE
Navigation
Welcome to Your Cloudera QuickStart VM!
Your Cluster
Node
Address
Manager Node
127.0.0.1
Worker Node 1
127.0.0.1
Get Started
Start Tutorial
Analyze Your Data
Lets Start the Tutorial!
Computer
cloudera's Home
Launch Cloudera Express
Trash
Eclipse
Launch Cloudera Enterprise Trial
Ask Bigger Questions
Cloudera Live : Welco...
Cloudera Live : Welco...
99%
1:58 PM
9/28/2015


cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 14:41 cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
quickstart.cloudera/#/tutorial/home
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
cloudera LIVE
Navigation
Introduction
Getting Started >
Welcome!
Cloudera Live Beginner Tutorial
Welcome to the Cloudera QuickStart VM, part of Cloudera Live! Following this tutorial will not only give you examples on how to get started with some of the tool (Cloudera's open source distribution including Apache Hadoop), but also give you a taste of what it means to ask bigger questions. By the end of this tutorial you will:
Understand how to use some of the powerful tools in CDH
Know how to setup and execute some basic business intelligence and analytics use cases
Be able to explain to your manager why they need to give you a raise!
Getting Started >
© 2015 Cloudera, Inc. All rights reserved | Terms & Conditions | Privacy Policy
Hadoop and the Hadoop elephant logo are trademarks of the Apache Software Foundation.
Computer
cloudera's Home
Trash
Launch Cloudera Express
Launch Cloudera Enterprise (trial)
Eclipse
Cloudera Live : Welco...
2:41 PM
9/28/2015


cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 14:43 cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
quickstart.cloudera/#/tutorial/getting_started
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
cloudera LIVE
Navigation
Getting Started
< Introduction
Tutorial Exercise 1 >
Getting Started
Define a Business Question
For the remainder of this tutorial, we will present examples in the context of a made up corporation called DataCo, and our mission is to help the organization get better insight by asking bigger questions.
Scenario:
Your Management: is talking euphorically about Big Data...
You: are carefully skeptical, as it will most likely all land on your desk anyway. Alternatively it has already landed on you, with the nice project description of: Go figure this Hadoop thing out...
< Introduction
Tutorial Exercise 1 >
Good to Know
Any successful PoC needs to address something your organization cares about. Hence, the first thing you need to do is to: define a business question.
It won't just impress your manager that you think big and have perspective on the business needs of your organisation (which in English means you just helped your manager to look good in front of his management). It will also help you to go through a well scoped PoC and get the investments you need to be successful.
Without a well defined question, you won't know how to properly model your data, i.e. what structure to apply at query time, or what data sets and tools to use to best serve the use case.
Computer
cloudera's Home
Trash
Launch Cloudera Express
Launch Cloudera Enterprise Trial
Eclipse
Cloudera Live : Welco...
2:43 PM
9/28/2015

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)

Player | Applications Places System | Mon Sep 28, 14:44 cloudera

Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox

Cloudera Live : Welcom...

quickstart.cloudera/#/tutorial/ingest_structured_data

Search

Cloudera | Hue | Hadoop | HBase | Impala | Spark | Solr | Oozie | Cloudera Manager | Getting Started

cloudera LIVE Navigation Tutorial Exercise 1

< Getting Started

Tutorial Exercise 2 >

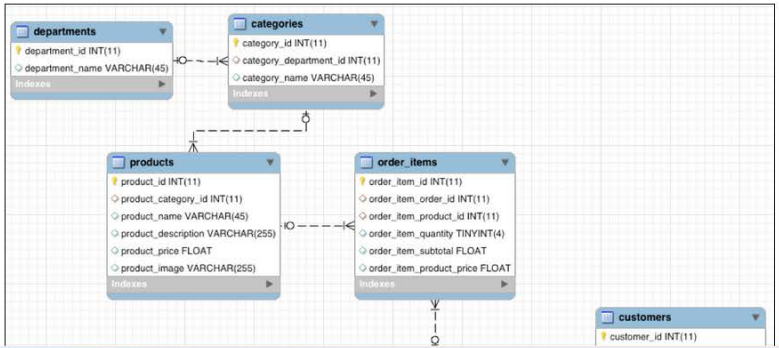
Tutorial Exercise 1

Ingest Structured Data

In this scenario, DataCo's business question is: *What products do our customers like to buy?* To answer this question, the first thought might be to look at the transaction data, which should indicate what customers actually do buy and like to buy, right?

This is probably something you can do in your regular RDBMS environment, but a benefit with Cloudera's platform is that you can do it at greater scale at lower cost, on the same system that you may also use for many other types of analysis.

What this exercise demonstrates is how to do exactly the same thing you already know how to do, but in CDH. Seamless integration is important when evaluating any new infrastructure. Hence, it's important to be able to do what you normally do, and not break any regular BI reports or workloads over the dataset you plan to migrate.



About Sqoop:

Apache Sqoop is a tool that uses MapReduce to transfer data between Hadoop clusters and relational databases very efficiently. It works by spawning tasks on multiple data nodes to download various portions of the data in parallel. When you're finished, each piece of data is replicated to ensure reliability, and spread out across the cluster to ensure you can process it in parallel on your cluster.

There are 2 versions of Sqoop included in Cloudera's platform. Sqoop 1 is a "thick client" and is what you use in this tutorial. The command you run will directly submit the MapReduce jobs to transfer the data. Sqoop 2 consists of a central server that submits the MapReduce jobs on behalf of clients, and a much lighter weight client that you use to connect to the server. The "Sqoop" you see in Cloudera Manager is the Sqoop 2 server, although Cloudera

Cloudera Live : Welco...

2:44 PM 9/28/2015

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)

Player | Applications Places System | Mon Sep 28, 14:44 cloudera

Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox

Cloudera Live : Welcom...

quickstart.cloudera/#/tutorial/ingest_structured_data

Search

Cloudera | Hue | Hadoop | HBase | Impala | Spark | Solr | Oozie | Cloudera Manager | Getting Started

cloudera LIVE Navigation Tutorial Exercise 1

< Getting Started

Tutorial Exercise 2 >

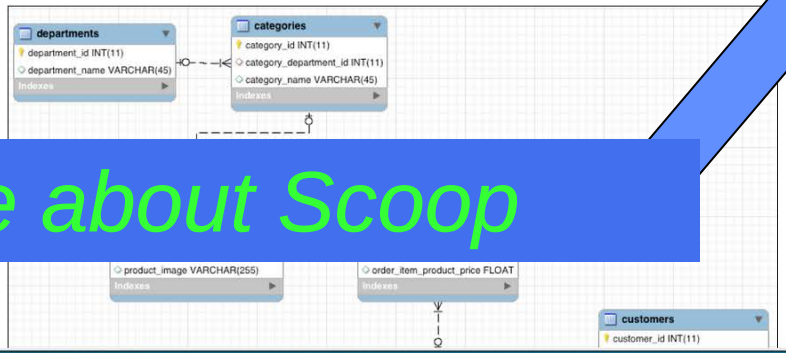
Tutorial Exercise 1

Ingest Structured Data

In this scenario, DataCo's business question is: *What products do our customers like to buy?* To answer this question, the first thought might be to look at the transaction data, which should indicate what customers actually do buy and like to buy, right?

This is probably something you can do in your regular RDBMS environment, but a benefit with Cloudera's platform is that you can do it at greater scale at lower cost, on the same system that you may also use for many other types of analysis.

What this exercise demonstrates is how to do exactly the same thing you already know how to do, but in CDH. Seamless integration is important when evaluating any new infrastructure. Hence, it's important to be able to do what you normally do, and not break any regular BI reports or workloads over the dataset you plan to migrate.



About Sqoop:

Apache Sqoop is a tool that uses MapReduce to transfer data between Hadoop clusters and relational databases very efficiently. It works by spawning tasks on multiple data nodes to download various portions of the data in parallel. When you're finished, each piece of data is replicated to ensure reliability, and spread out across the cluster to ensure you can process it in parallel on your cluster.

There are 2 versions of Sqoop included in Cloudera's platform. Sqoop 1 is a "thick client" and is what you use in this tutorial. The command you run will directly submit the MapReduce jobs to transfer the data. Sqoop 2 consists of a central server that submits the MapReduce jobs on behalf of clients, and a much lighter weight client that you use to connect to the server. The "Sqoop" you see in Cloudera Manager is the Sqoop 2 server, although Cloudera

More about Scoop

Cloudera Live : Welco...

2:44 PM 9/28/2015

Scroll down for more

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)

Player | Applications Places System

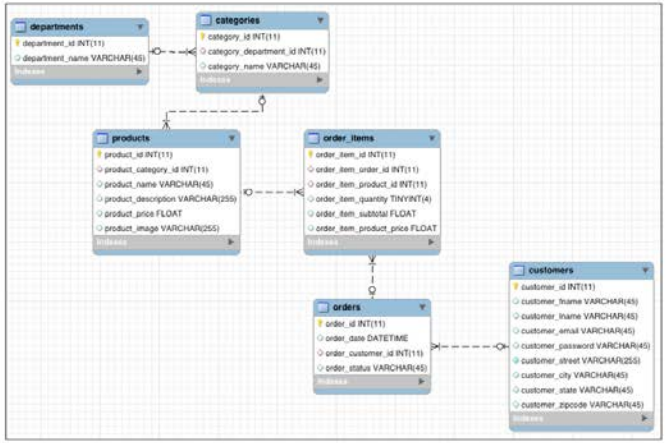
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox

Cloudera Live : Welcom...

quickstart.cloudera/#/tutorial/ingest_structured_data

Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started

cloudera LIVE Navigation



finished, each piece of data is replicated to ensure reliability, and spread out across the cluster to ensure you can process it in parallel on your cluster.

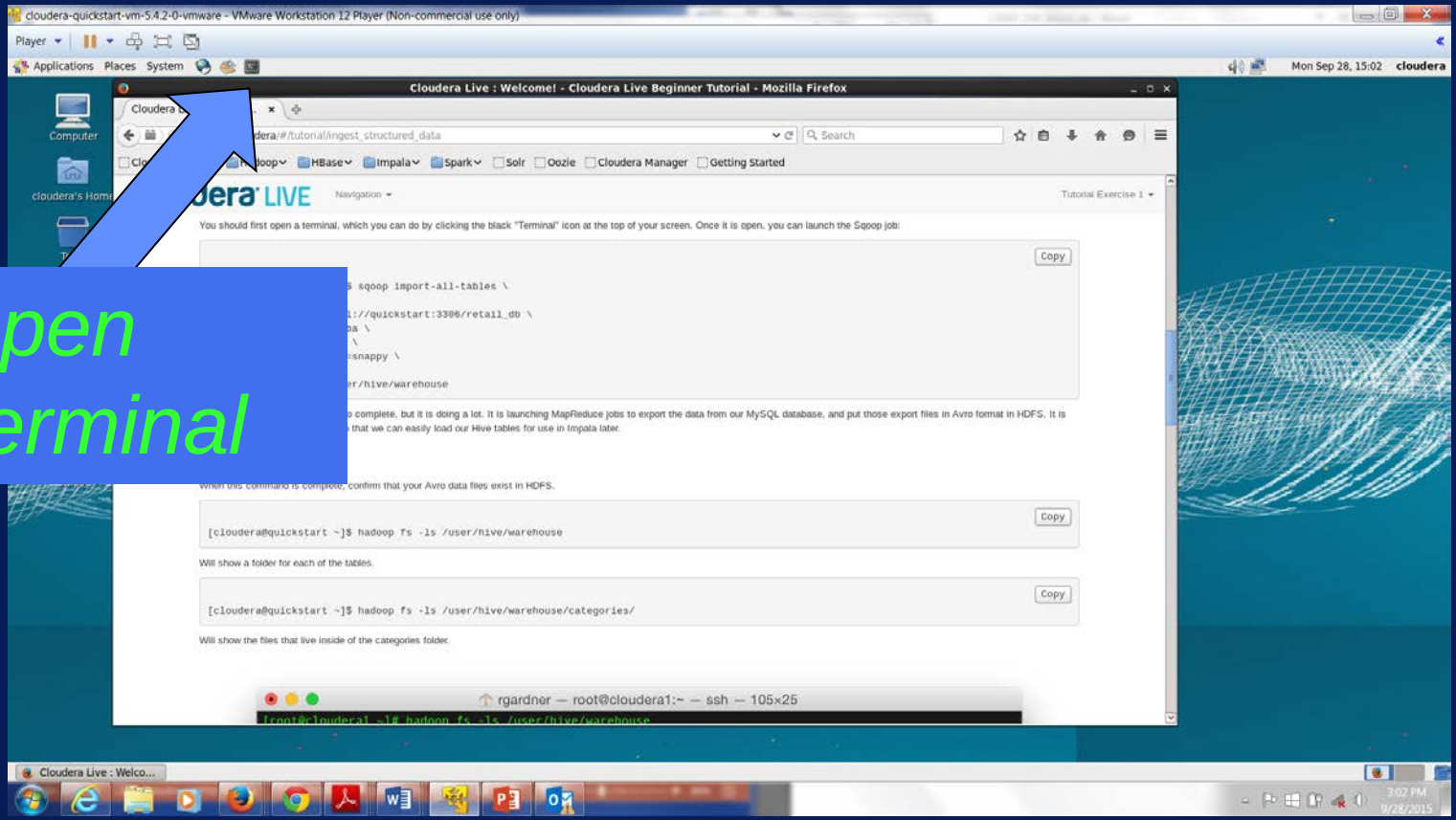
There are 2 versions of Sqoop included in Cloudera's platform. Sqoop 1 is a "thick client" and is what you use in this tutorial. The command you run will directly submit the MapReduce jobs to transfer the data. Sqoop 2 consists of a central server that submits the MapReduce jobs on behalf of clients, and a much lighter weight client that you use to connect to the server. The "Sqoop" you see in Cloudera Manager is the Sqoop 2 server, although Cloudera Manager will make sure that both the "sqoop" and "sqoop2" command are correctly configured on all your machines.

To analyze the transaction data in the new platform, we need to ingest it into the Hadoop Distributed File System (HDFS). We need to find a tool that easily transfers structured data from a RDBMS to HDFS, while preserving structure. That enables us to query the data, but not interfere with or break any regular workload on it.

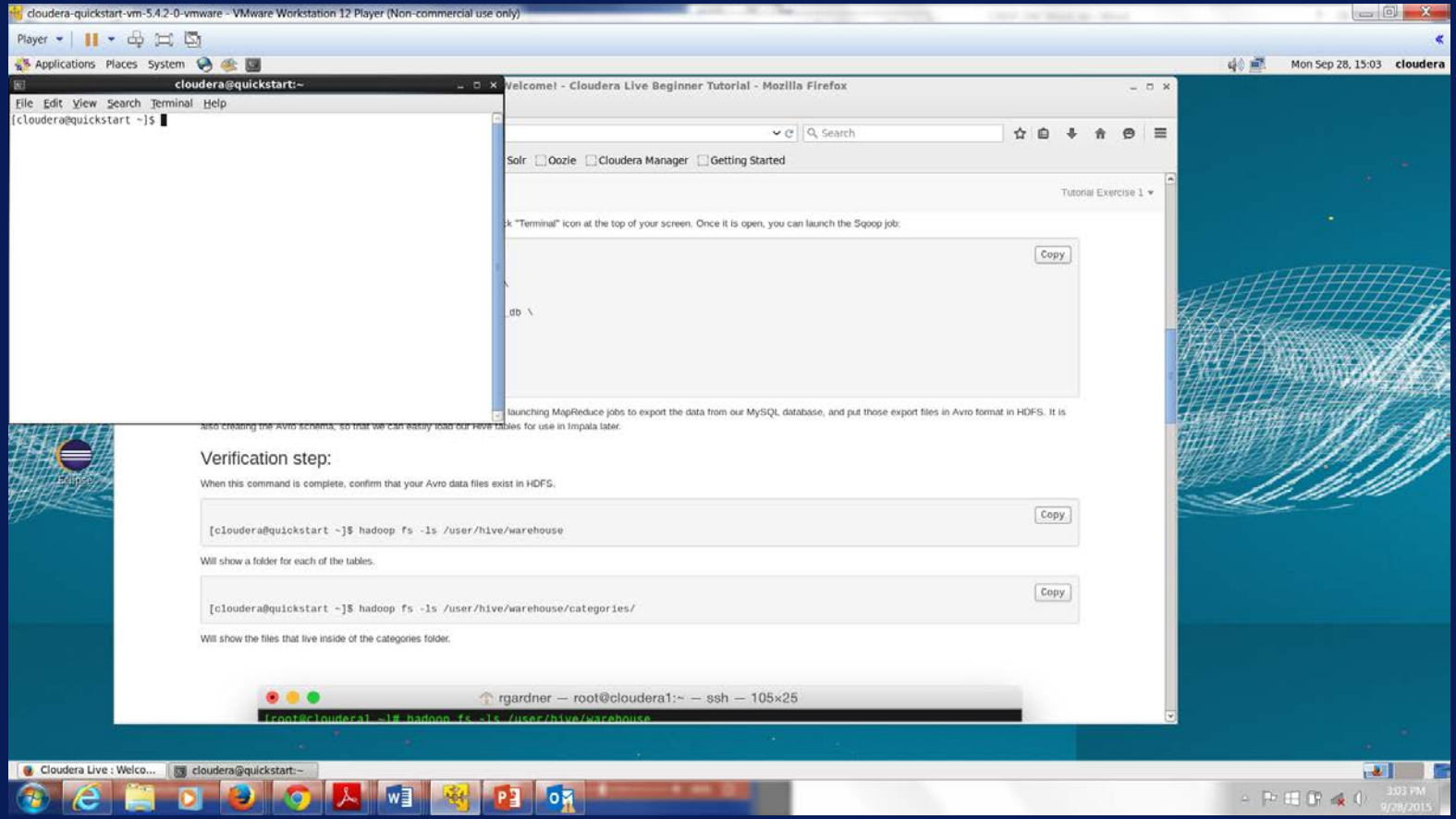
Apache Sqoop, which is part of CDH, is that tool. The nice thing about Sqoop is that we can automatically load our relational data from MySQL into HDFS, while preserving the structure.

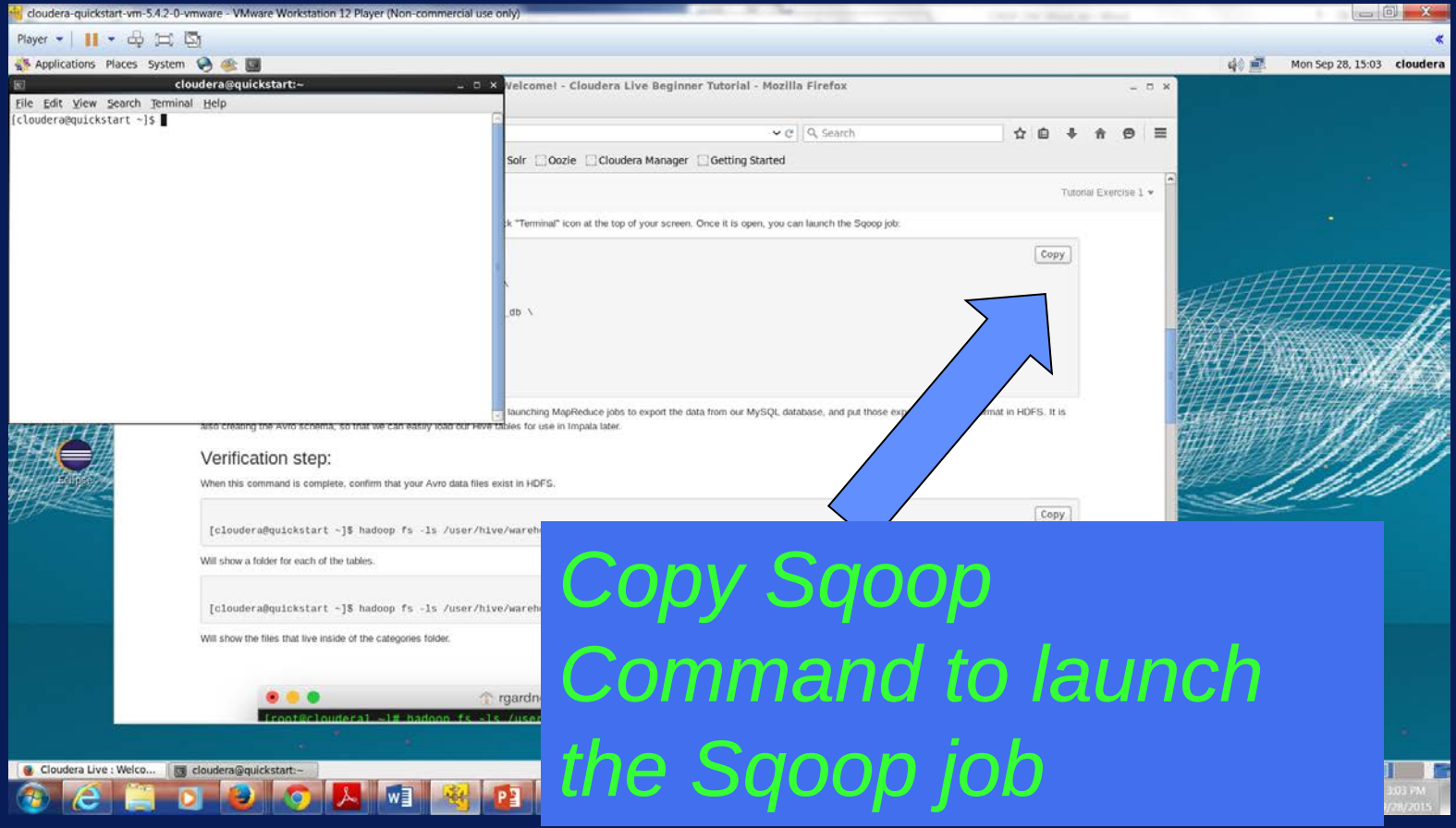
Cloudera Live : Welco...

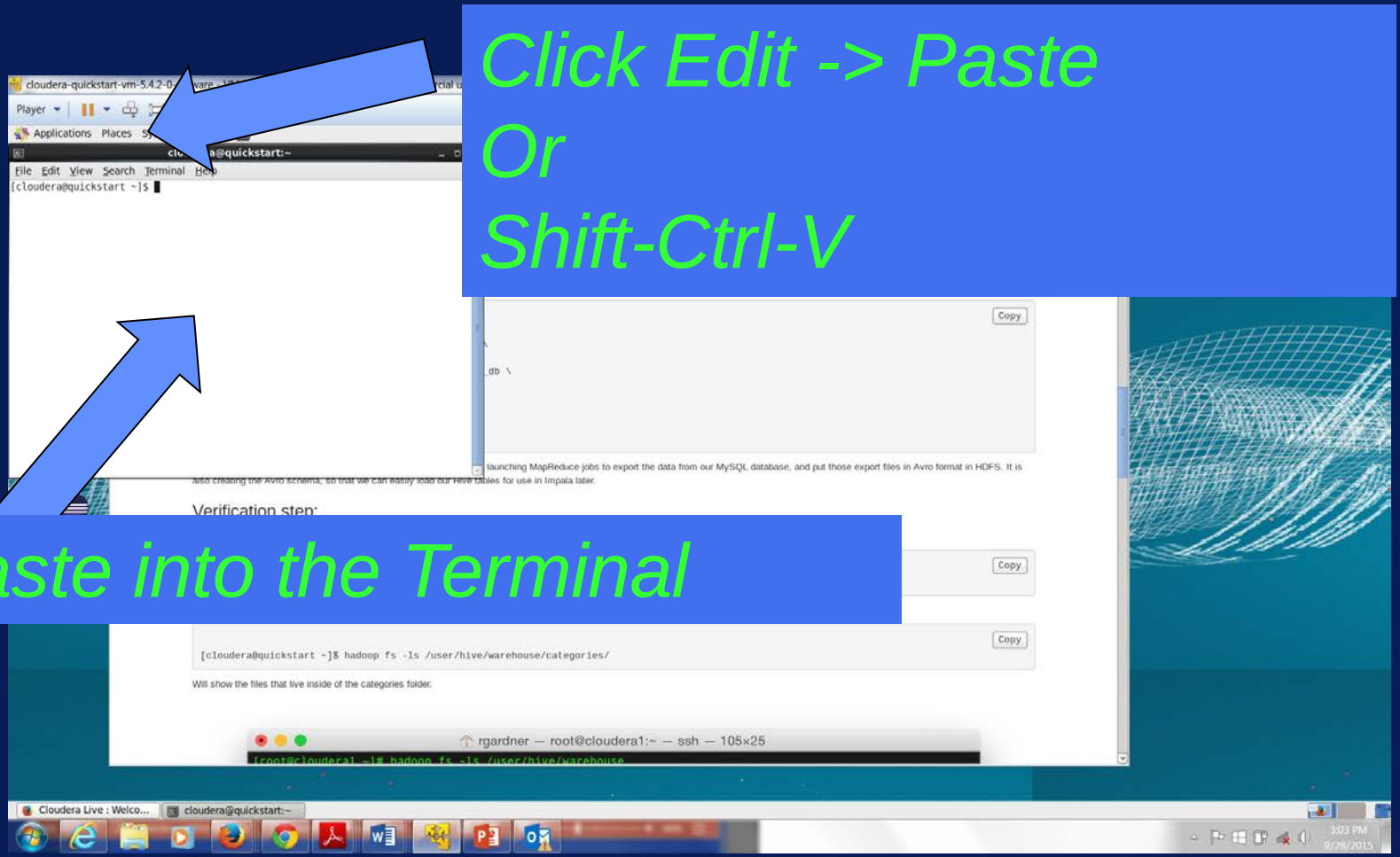
2:57 PM 9/28/2015

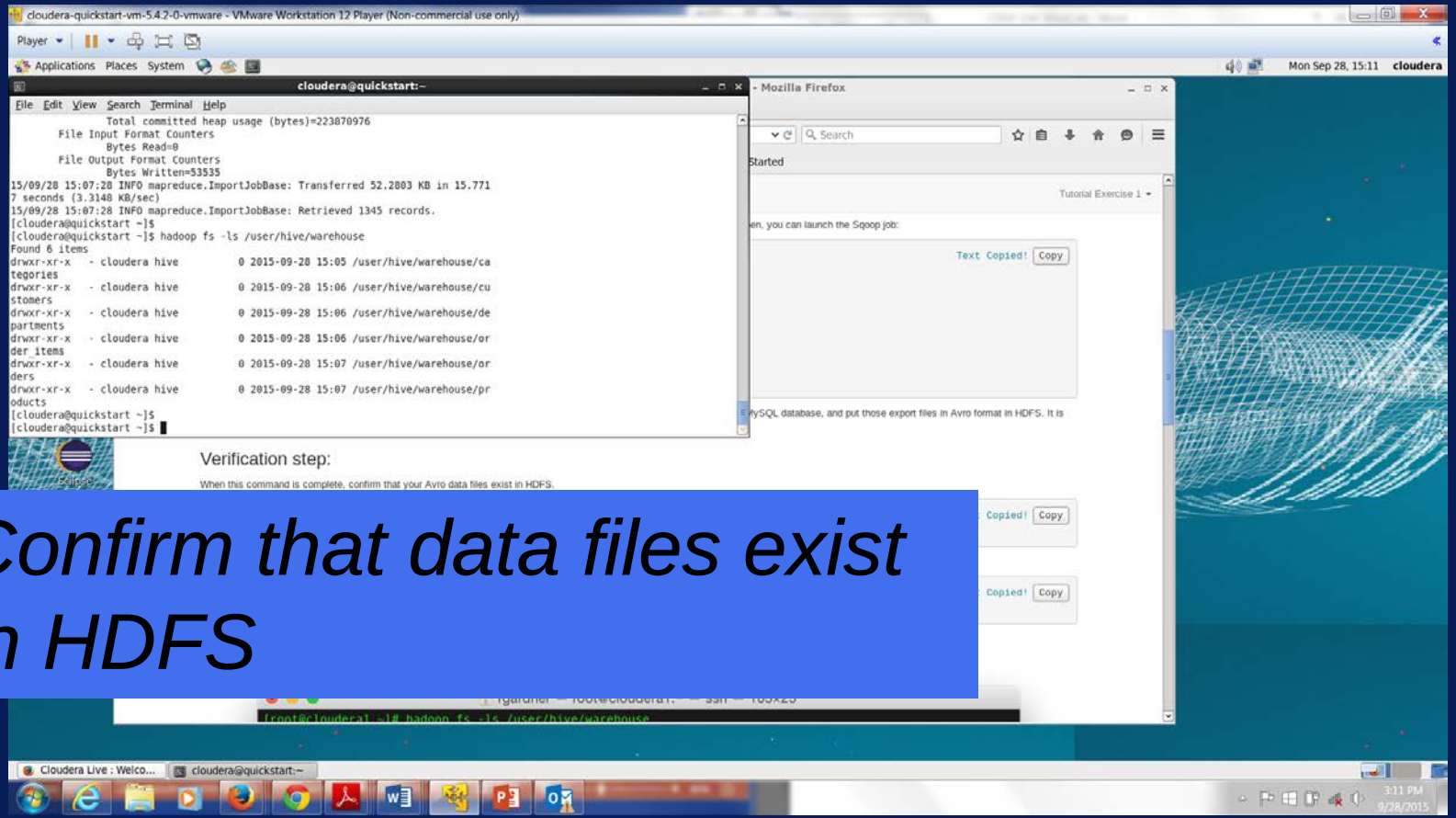


Open Terminal

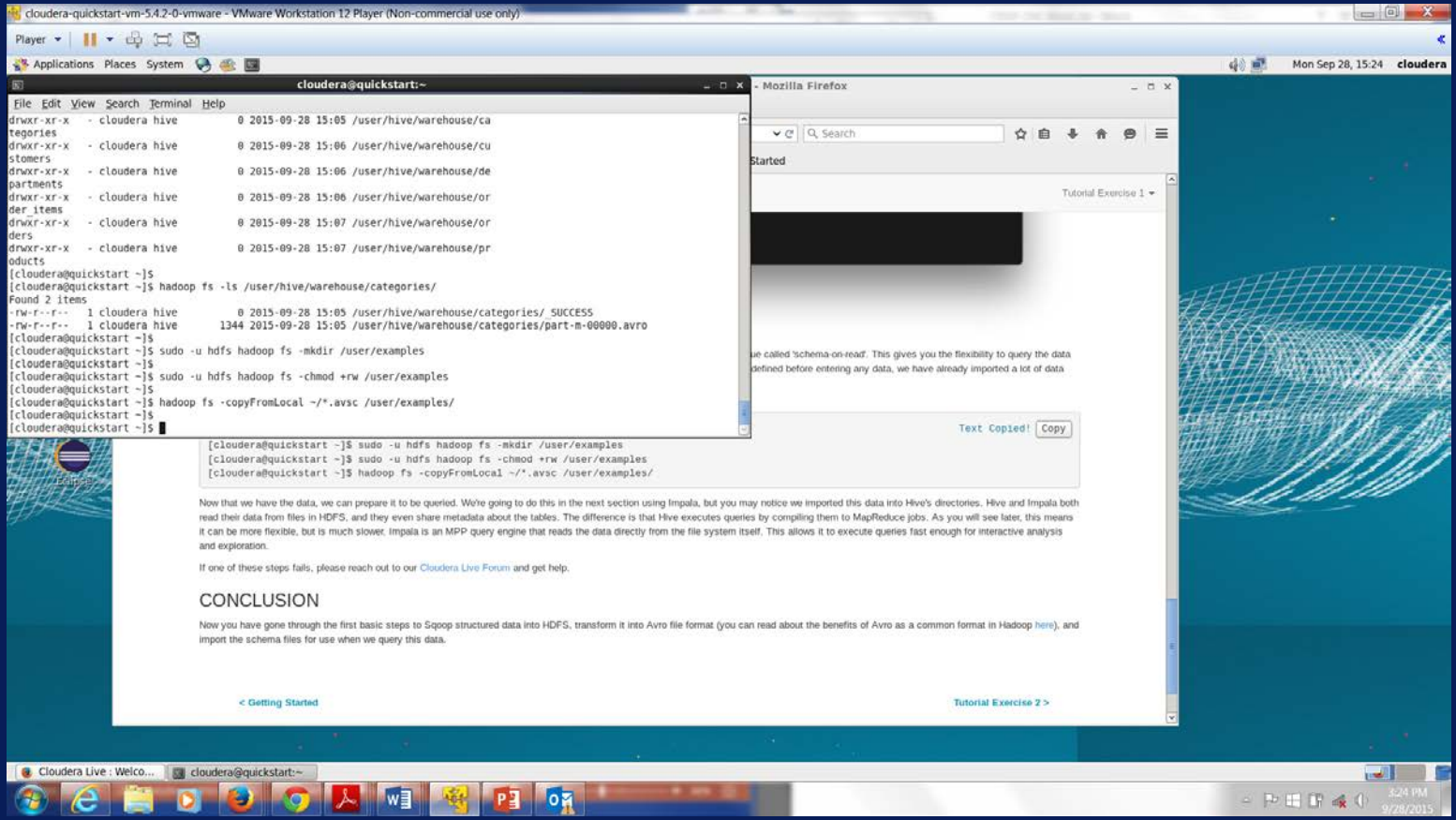

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 15:03 cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$
Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Search
Solr Oozie Cloudera Manager Getting Started
Tutorial Exercise 1
"Terminal" icon at the top of your screen. Once it is open, you can launch the Sqoop job:
Copy
launching MapReduce jobs to export the data from our MySQL database, and put those export files in Avro format in HDFS. It is
Verification step:
When this command is complete, confirm that your Avro data files exist in HDFS.
Copy
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse
Will show a folder for each of the tables.
Copy
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/categories/
Will show the files that live inside of the categories folder.
rgardner — root@cloudera1:~ — ssh — 105×25
Cloudera Live : Welco...
cloudera@quickstart:~

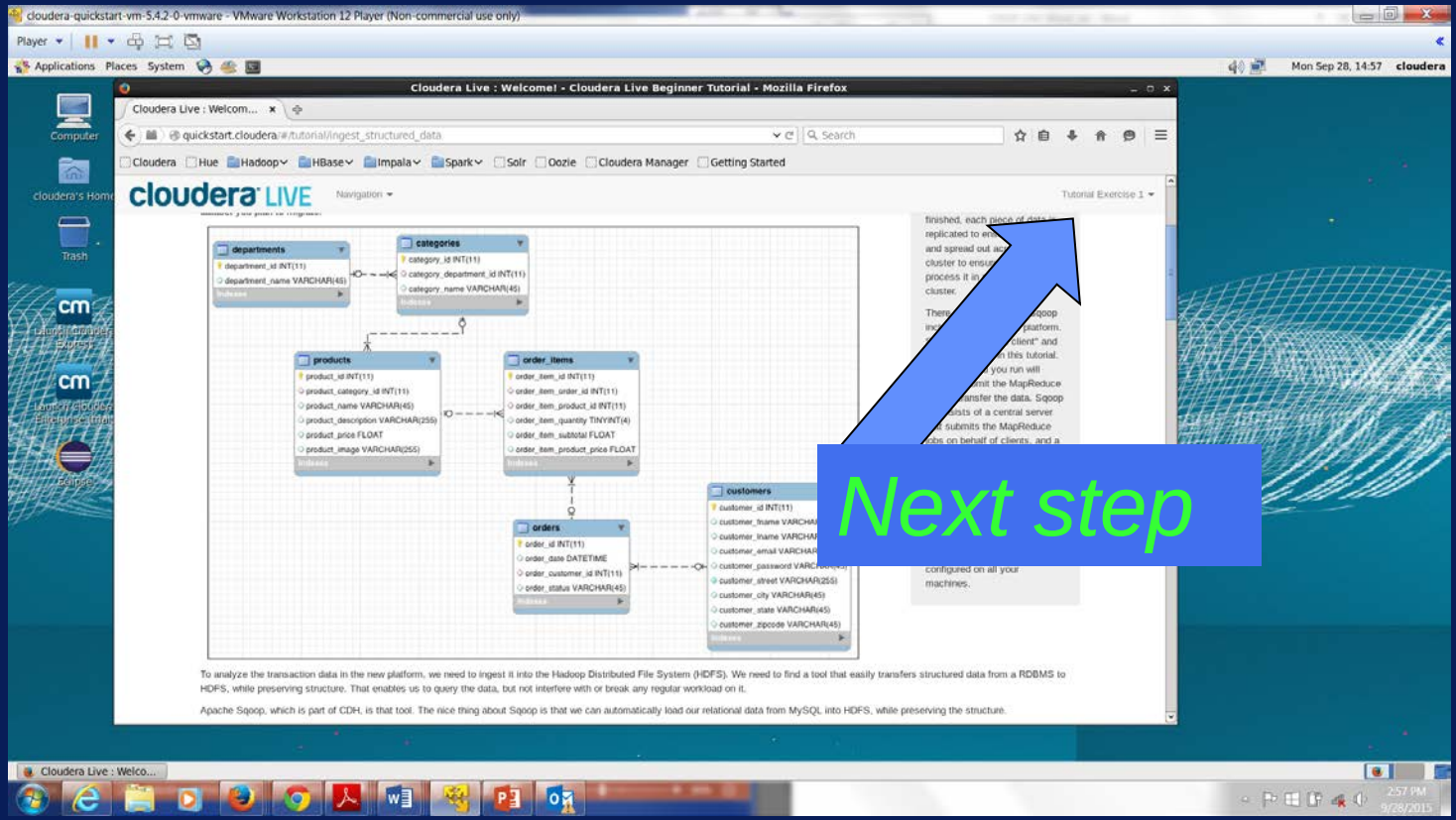

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 15:03 cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
[cloudera@quickstart ~]$
Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Search
Solr Oozie Cloudera Manager Getting Started
Tutorial Exercise 1
"Terminal" icon at the top of your screen. Once it is open, you can launch the Sqoop job:
Copy
db \
launching MapReduce jobs to export the data from our MySQL database, and put those exp
mat in HDFS. It is
also creating the Avro schema, so that we can easily load our Hive tables for use in Impala later.
Verification step:
When this command is complete, confirm that your Avro data files exist in HDFS.
Copy
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/wareh
Will show a folder for each of the tables.
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/wareh
Will show the files that live inside of the categories folder.
Cloudera Live : Welco...
cloudera@quickstart:~
Copy Sqoop Command to launch the Sqoop job

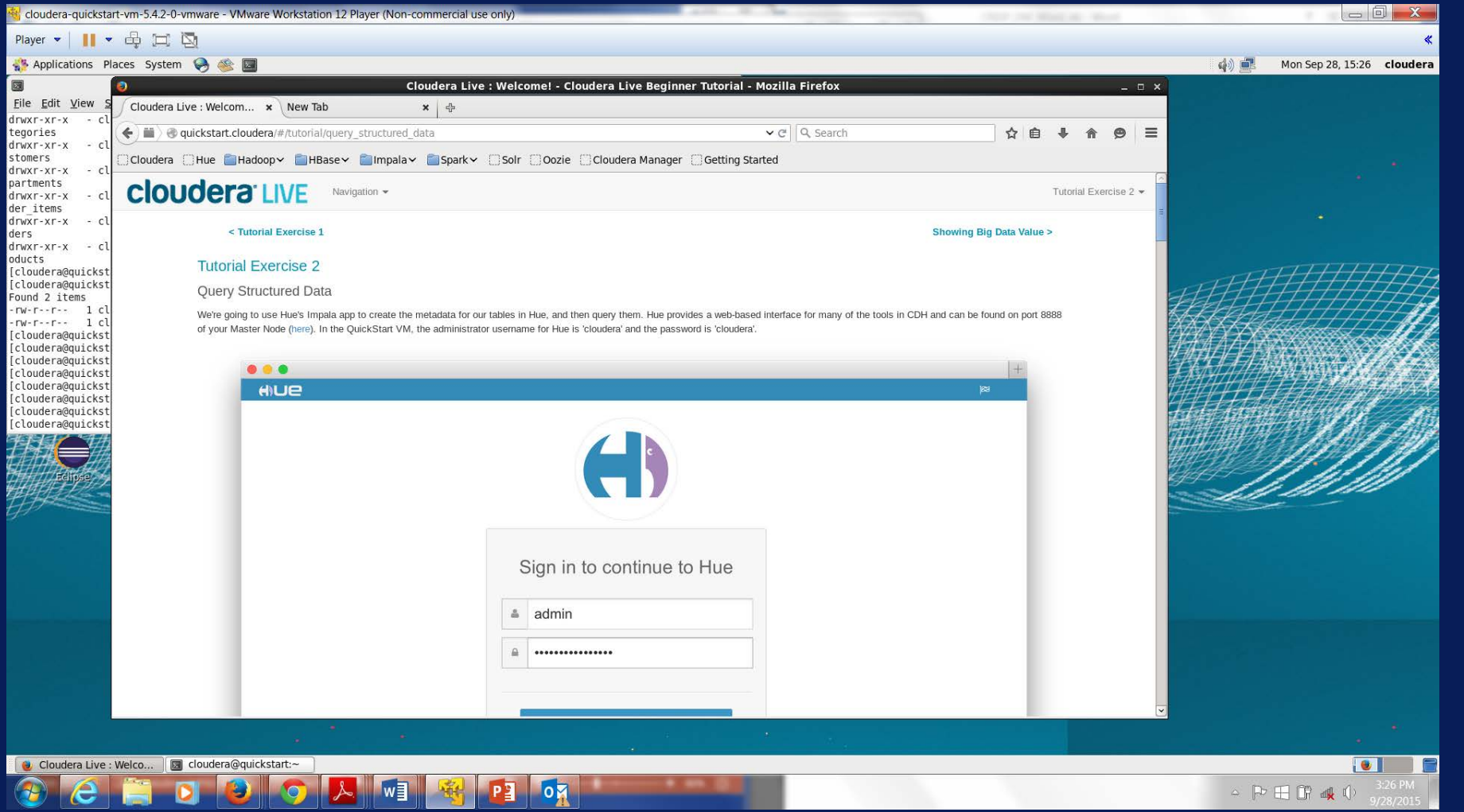

Click Edit -> Paste
Or
Shift-Ctrl-V
Paste into the Terminal
Player
Applications Places
File Edit View Search Terminal Help
[cloudera@quickstart ~]$
Copy
launching MapReduce jobs to export the data from our MySQL database, and put those export files in Avro format in HDFS. It is
Verification step:
Copy
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/categories/
Copy
Will show the files that live inside of the categories folder.
rgardner — root@cloudera1:~ — ssh — 105×25
Cloudera Live : Welco...
cloudera@quickstart:~

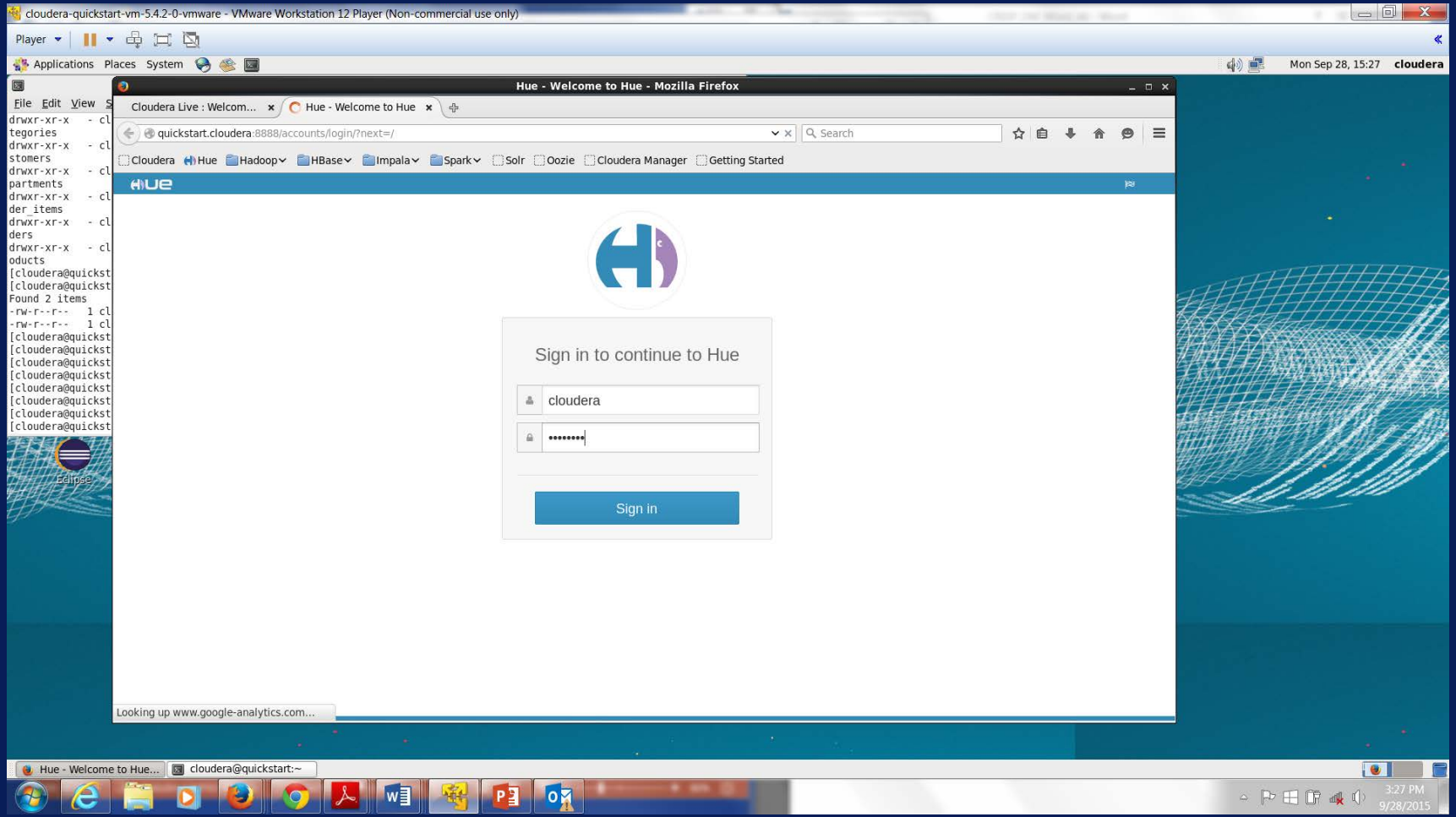


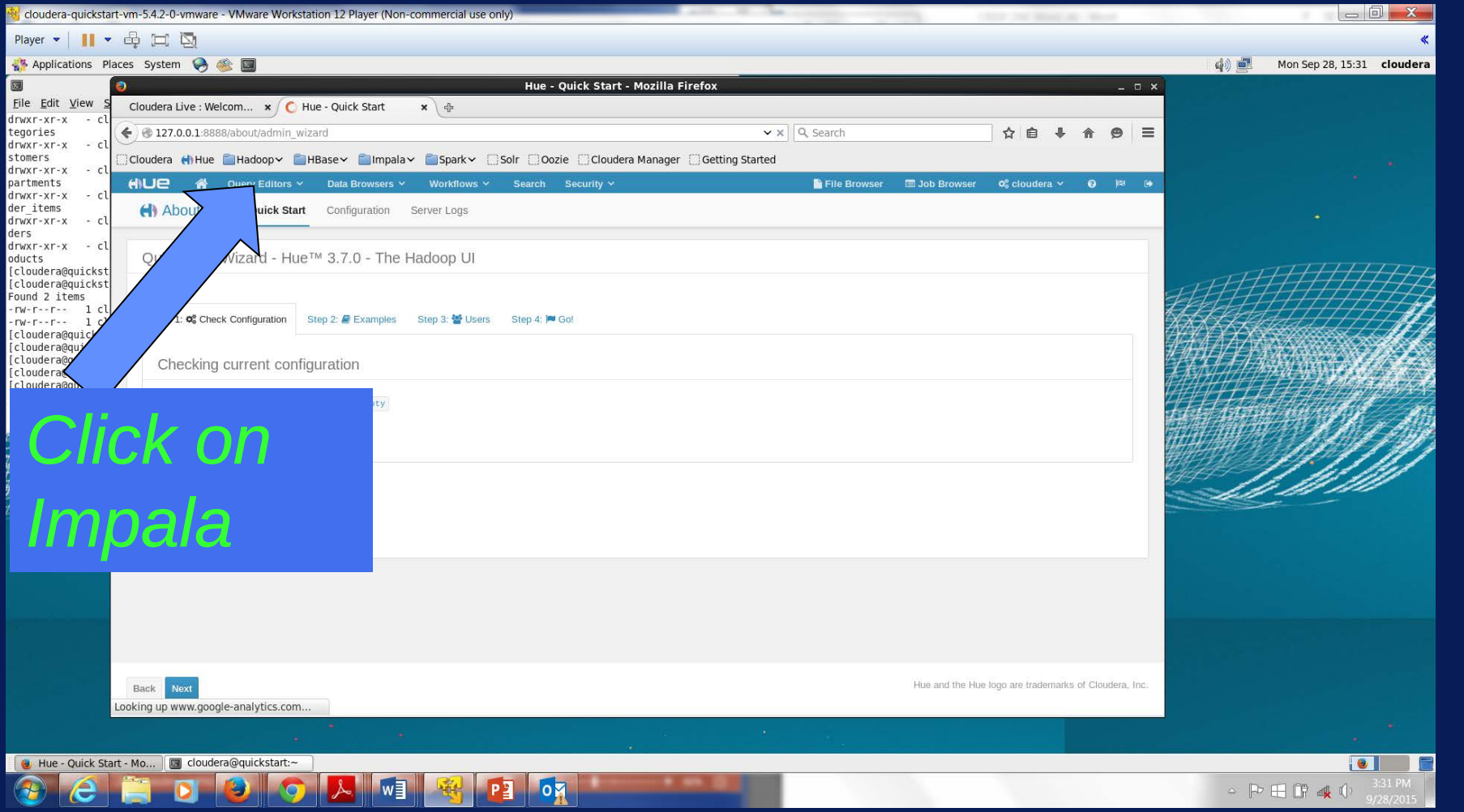
Confirm that data files exist in HDFS

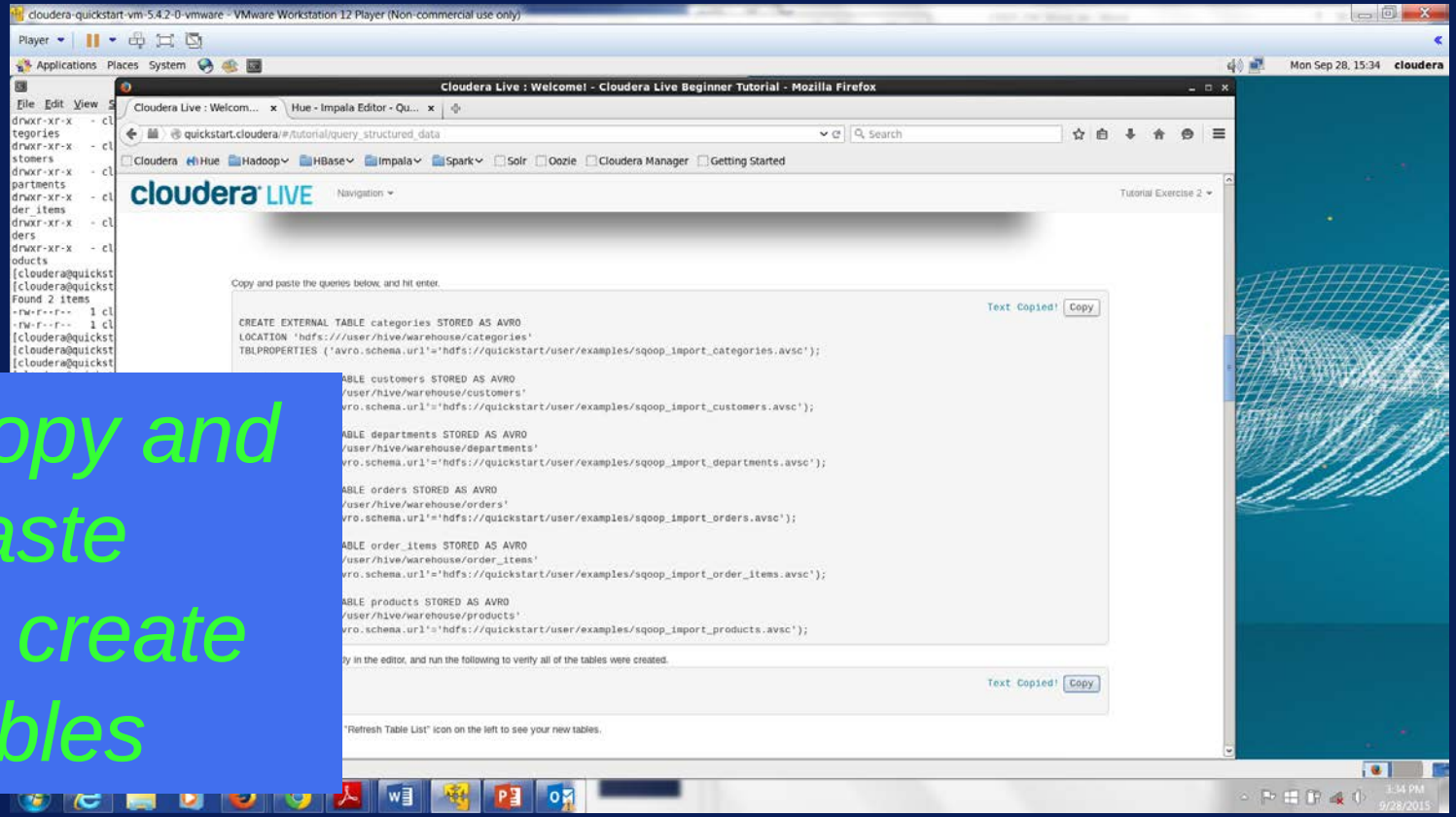

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 15:24 cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
drwxr-xr-x - cloudera hive 0 2015-09-28 15:05 /user/hive/warehouse/categories
drwxr-xr-x - cloudera hive 0 2015-09-28 15:06 /user/hive/warehouse/customers
drwxr-xr-x - cloudera hive 0 2015-09-28 15:06 /user/hive/warehouse/departments
drwxr-xr-x - cloudera hive 0 2015-09-28 15:06 /user/hive/warehouse/order_items
drwxr-xr-x - cloudera hive 0 2015-09-28 15:07 /user/hive/warehouse/orders
drwxr-xr-x - cloudera hive 0 2015-09-28 15:07 /user/hive/warehouse/products
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/categories/
Found 2 items
-rw-r--r-- 1 cloudera hive 0 2015-09-28 15:05 /user/hive/warehouse/categories/_SUCCESS
-rw-r--r-- 1 cloudera hive 1344 2015-09-28 15:05 /user/hive/warehouse/categories/part-m-00000.avro
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -mkdir /user/examples
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -chmod +rw /user/examples
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hadoop fs -copyFromLocal ~/*.avsc /user/examples/
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
Mozilla Firefox
Search
Started
Tutorial Exercise 1
called 'schema-on-read'. This gives you the flexibility to query the data
defined before entering any data, we have already imported a lot of data
Text Copied! Copy
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -mkdir /user/examples
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -chmod +rw /user/examples
[cloudera@quickstart ~]$ hadoop fs -copyFromLocal ~/*.avsc /user/examples/
Now that we have the data, we can prepare it to be queried. We're going to do this in the next section using Impala, but you may notice we imported this data into Hive's directories. Hive and Impala both read their data from files in HDFS, and they even share metadata about the tables. The difference is that Hive executes queries by compiling them to MapReduce jobs. As you will see later, this means it can be more flexible, but is much slower. Impala is an MPP query engine that reads the data directly from the file system itself. This allows it to execute queries fast enough for interactive analysis and exploration.
If one of these steps fails, please reach out to our Cloudera Live Forum and get help.
CONCLUSION
Now you have gone through the first basic steps to Sqoop structured data into HDFS, transform it into Avro file format (you can read about the benefits of Avro as a common format in Hadoop here), and import the schema files for use when we query this data.
< Getting Started
Tutorial Exercise 2 >
Cloudera Live : Welco...
cloudera@quickstart:~
3:24 PM
9/28/2015


cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 14:57 cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
quickstart.cloudera/#/tutorial/ingest_structured_data
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
cloudera LIVE
Navigation
Tutorial Exercise 1
departments
department_id INT(11)
department_name VARCHAR(45)
categories
category_id INT(11)
category_department_id INT(11)
category_name VARCHAR(45)
products
product_id INT(11)
product_category_id INT(11)
product_name VARCHAR(45)
product_description VARCHAR(255)
product_price FLOAT
product_image VARCHAR(255)
order_items
order_item_id INT(11)
order_item_order_id INT(11)
order_item_product_id INT(11)
order_item_quantity TINYINT(4)
order_item_subtotal FLOAT
order_item_product_price FLOAT
orders
order_id INT(11)
order_date DATETIME
order_customer_id INT(11)
order_status VARCHAR(45)
customers
customer_id INT(11)
customer_street VARCHAR(255)
customer_city VARCHAR(45)
customer_state VARCHAR(45)
customer_zipcode VARCHAR(45)
Next step
To analyze the transaction data in the new platform, we need to ingest it into the Hadoop Distributed File System (HDFS). We need to find a tool that easily transfers structured data from a RDBMS to HDFS, while preserving structure. That enables us to query the data, but not interfere with or break any regular workload on it.
Apache Sqoop, which is part of CDH, is that tool. The nice thing about Sqoop is that we can automatically load our relational data from MySQL into HDFS, while preserving the structure.
Cloudera Live : Welco...
2:57 PM
9/28/2015

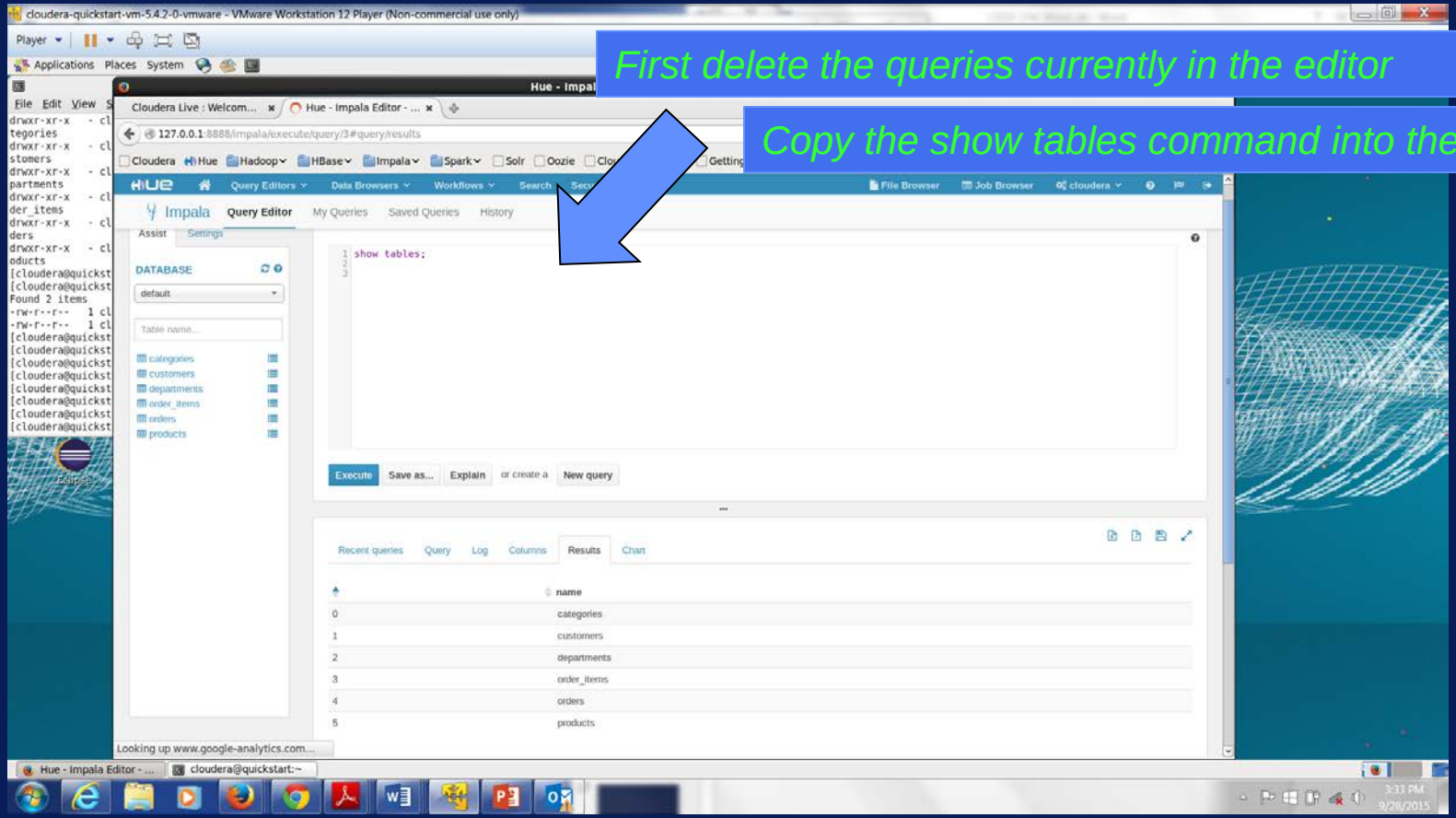

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 15:26 cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
New Tab
quickstart.cloudera/#/tutorial/query_structured_data
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
cloudera LIVE
Navigation
Tutorial Exercise 2
< Tutorial Exercise 1
Showing Big Data Value >
Tutorial Exercise 2
Query Structured Data
We're going to use Hue's Impala app to create the metadata for our tables in Hue, and then query them. Hue provides a web-based interface for many of the tools in CDH and can be found on port 8888 of your Master Node (here). In the QuickStart VM, the administrator username for Hue is 'cloudera' and the password is 'cloudera'.
HUE
Sign in to continue to Hue
admin
Eclipse
Cloudera Live : Welco...
cloudera@quickstart:~
3:26 PM
9/28/2015

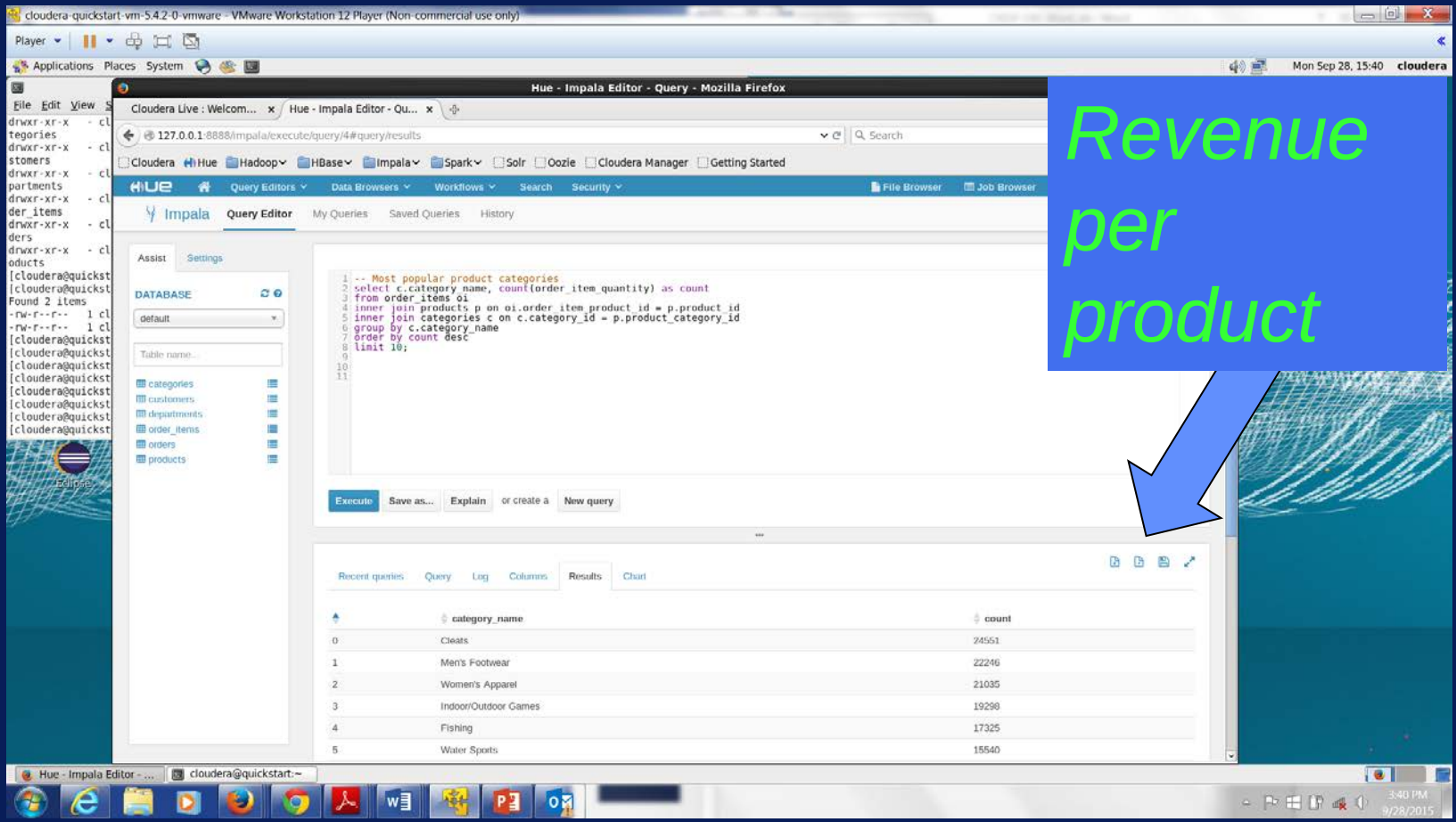

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 15:27 cloudera
Hue - Welcome to Hue - Mozilla Firefox
Cloudera Live : Welcom...
Hue - Welcome to Hue
quickstart.cloudera:8888/accounts/login/?next=/
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
HUE
Sign in to continue to Hue
cloudera
Sign in
Looking up www.google-analytics.com...
Eclipse
Hue - Welcome to Hue...
cloudera@quickstart:~
3:27 PM
9/28/2015

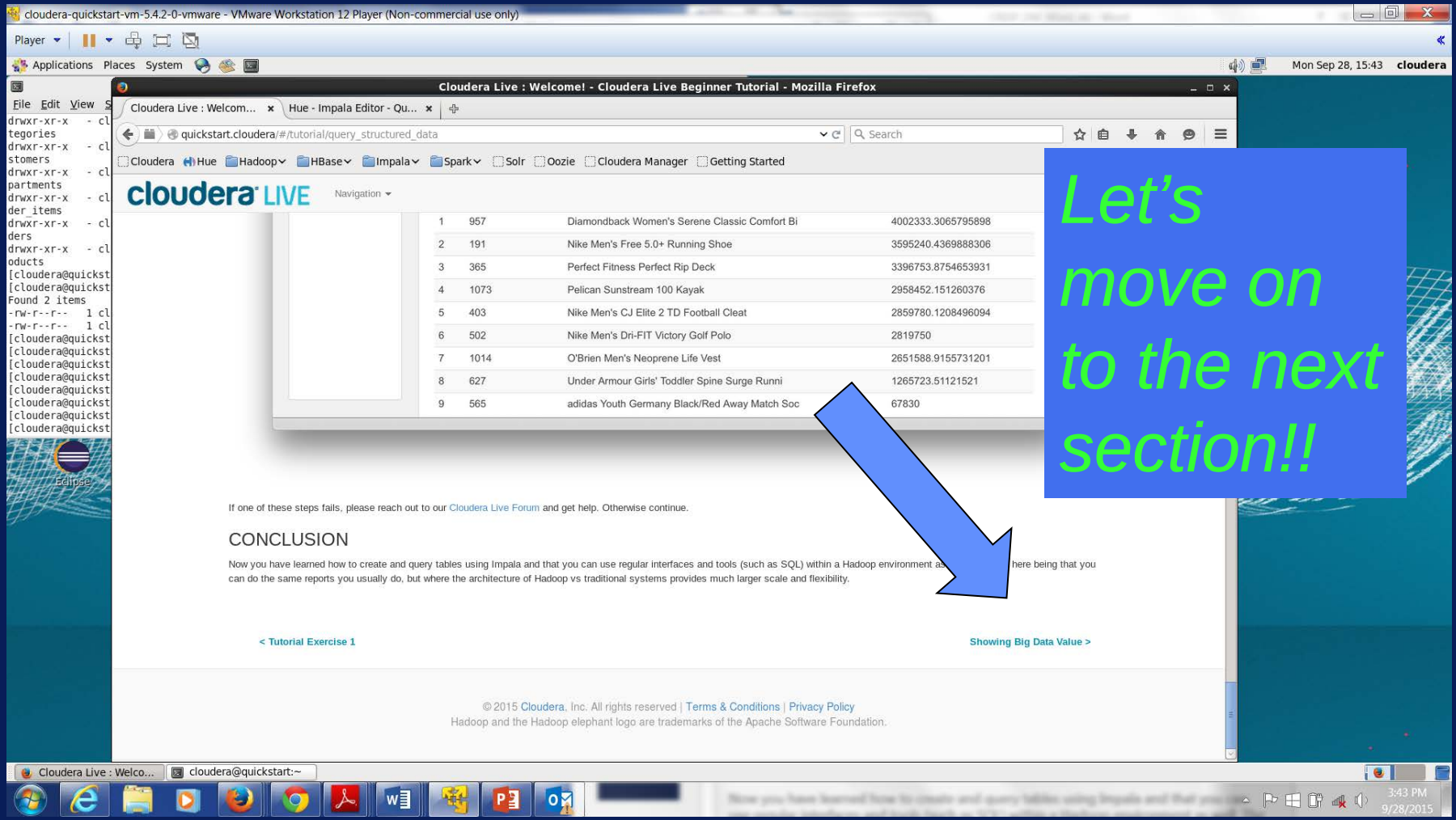

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 15:31 cloudera
Hue - Quick Start - Mozilla Firefox
Cloudera Live : Welcom...
Hue - Quick Start
127.0.0.1:8888/about/admin_wizard
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query Editors Data Browsers Workflows Search Security File Browser Job Browser cloudera
About
Quick Start Configuration Server Logs
Wizard - Hue™ 3.7.0 - The Hadoop UI
Step 1: Check Configuration Step 2: Examples Step 3: Users Step 4: Go!
Checking current configuration
Back Next
Hue and the Hue logo are trademarks of Cloudera, Inc.
Looking up www.google-analytics.com...
Click on Impala
Hue - Quick Start - Mo...
cloudera@quickstart:~
3:31 PM
9/28/2015

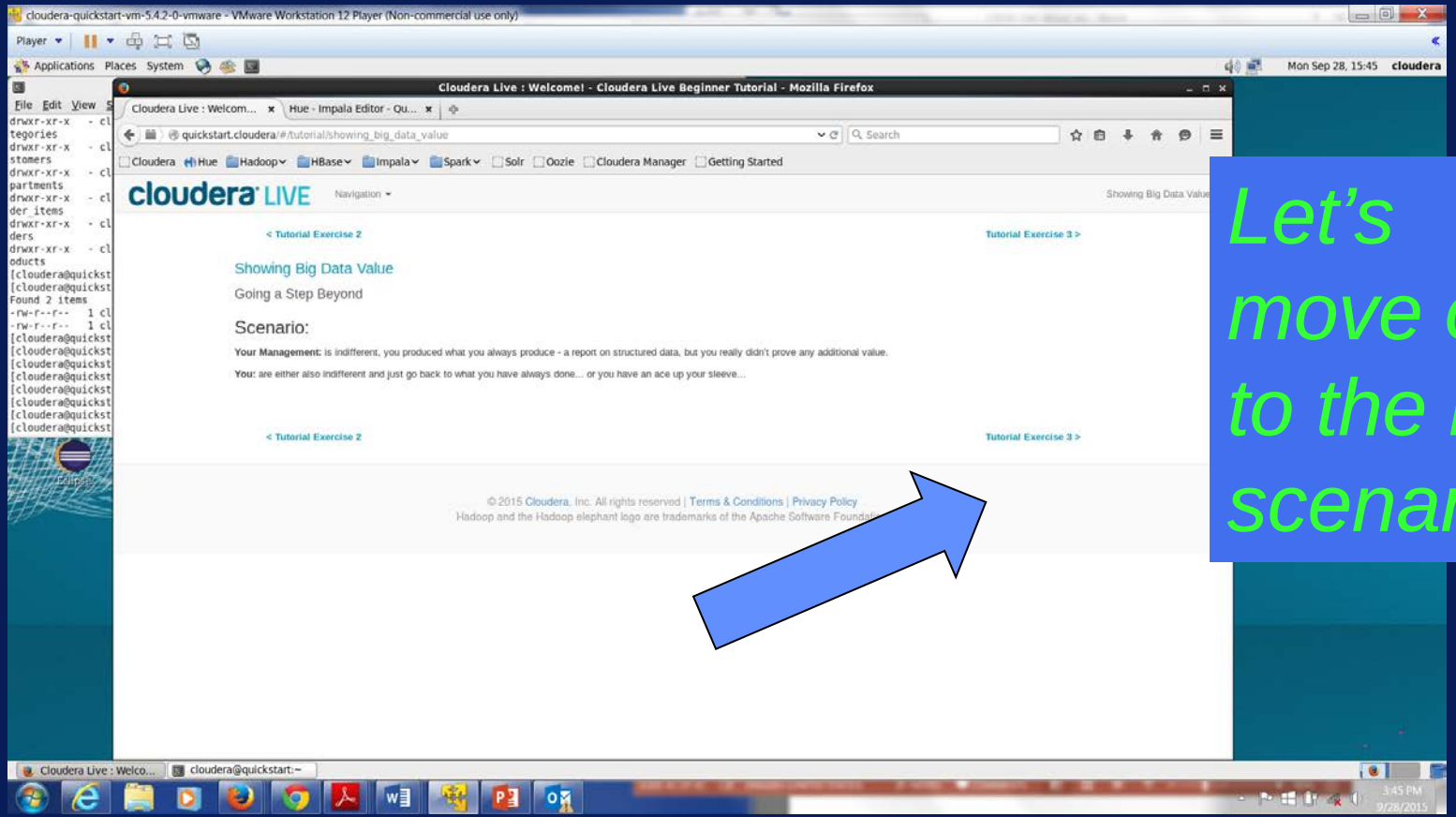


Copy and paste to create tables

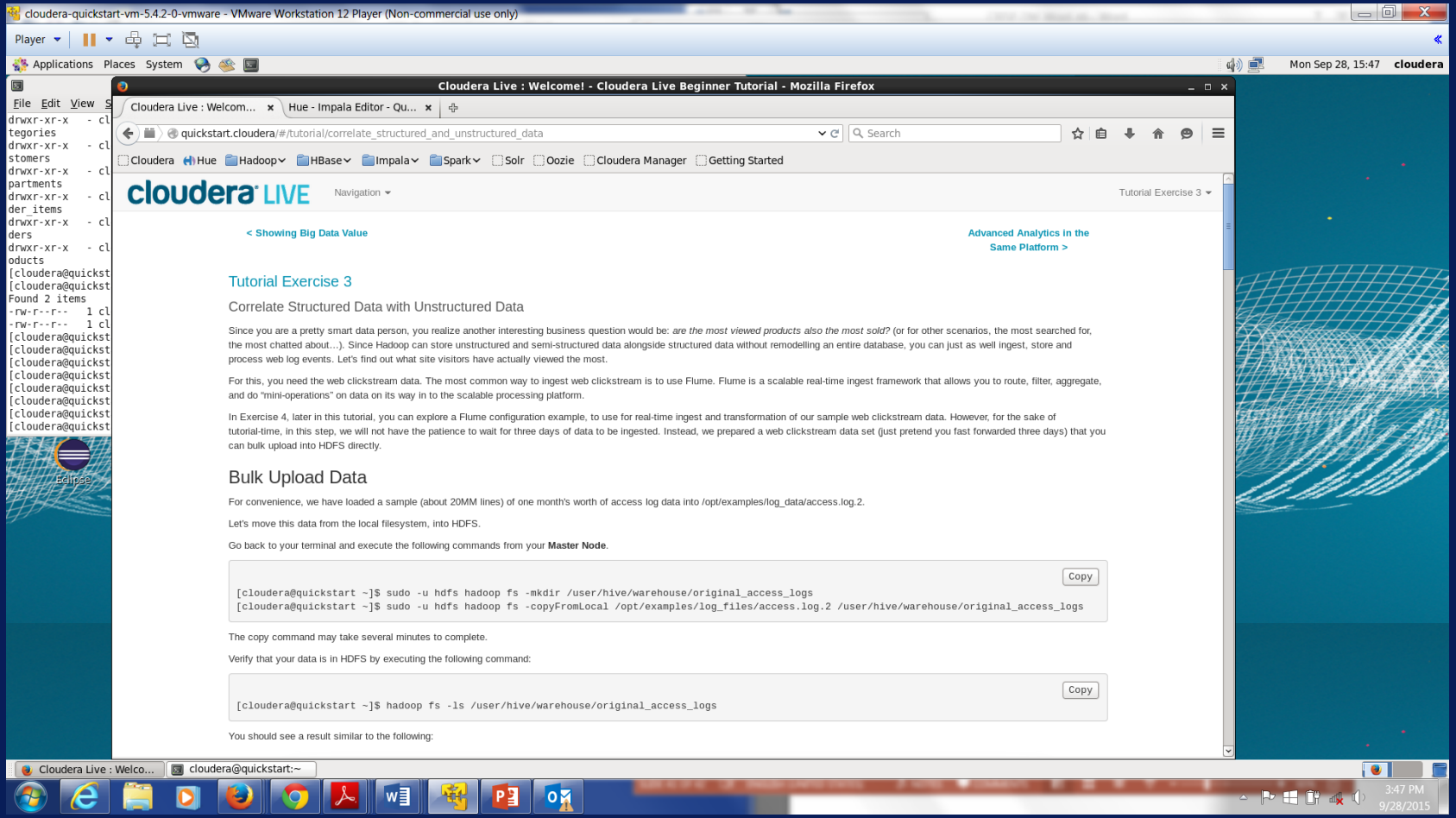

First delete the queries currently in the editor
Copy the show tables command into the window
cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Hue - Impala Editor - ...
127.0.0.1:8888/impala/execute/query/3#query/results
Impala
Query Editor
My Queries
Saved Queries
History
show tables;
Execute
Save as...
Explain
or create a
New query
Recent queries
Query
Log
Columns
Results
Chart
name
0 categories
1 customers
2 departments
3 order_items
4 orders
5 products

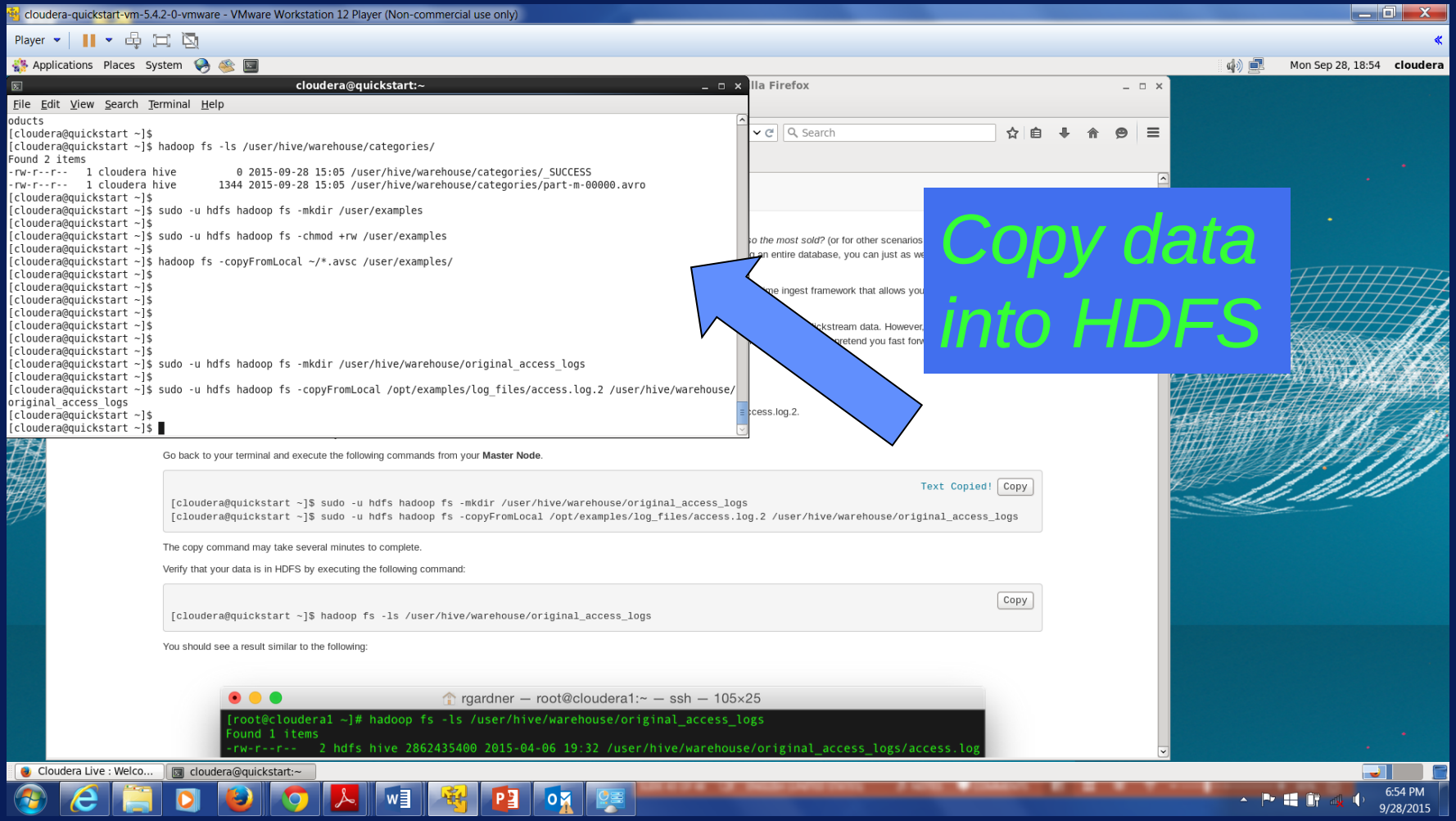

Revenue per product
Hue - Impala Editor - Query - Mozilla Firefox
-- Most popular product categories
select c.category_name, count(order_item_quantity) as count
from order_items oi
inner join products p on oi.order_item_product_id = p.product_id
inner join categories c on c.category_id = p.product_category_id
group by c.category_name
order by count desc
limit 10;
Execute
Save as...
Explain
or create a
New query
category_name
count
0 Cleats 24551
1 Men's Footwear 22246
2 Women's Apparel 21035
3 Indoor/Outdoor Games 19298
4 Fishing 17325
5 Water Sports 15540

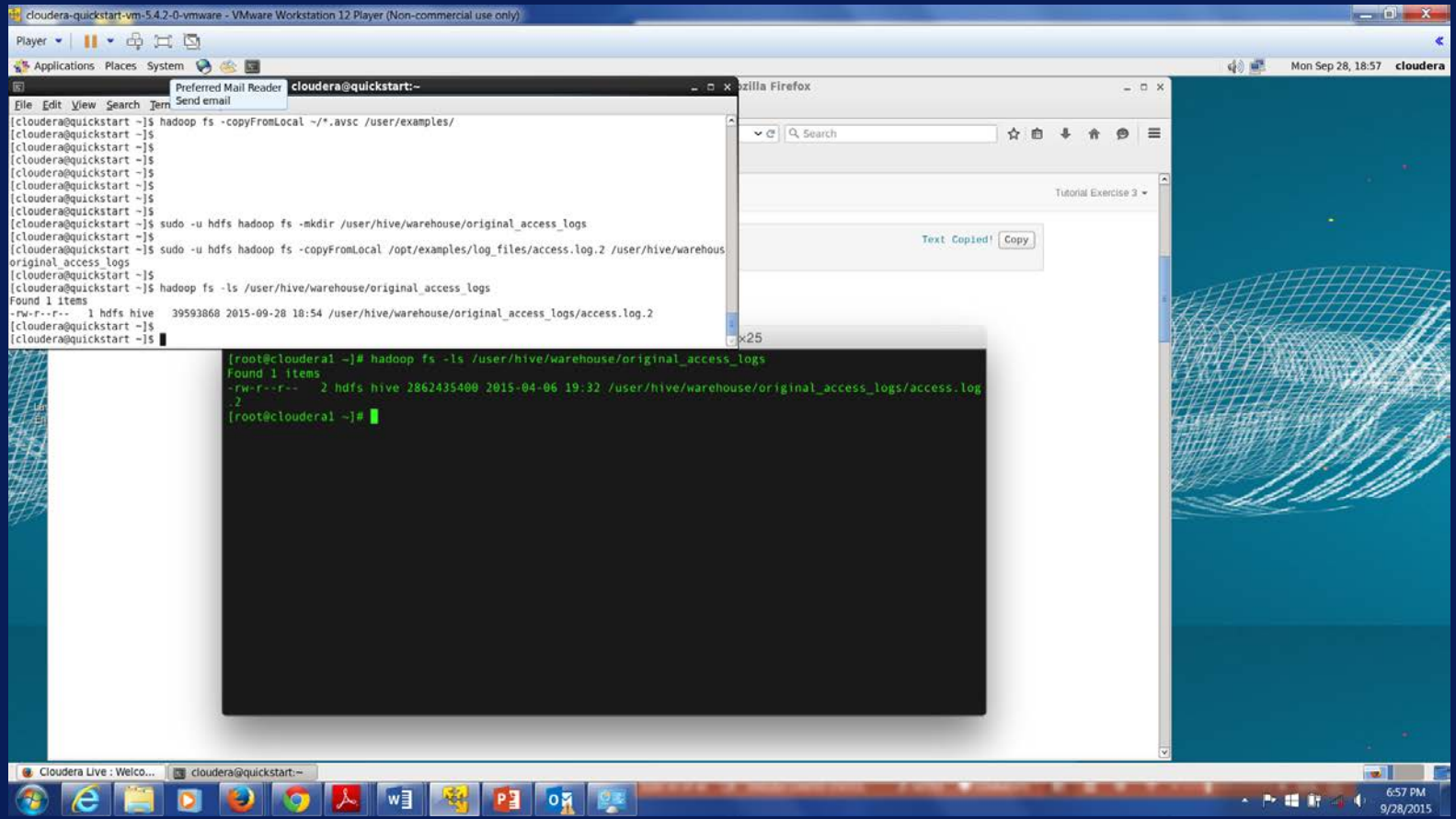

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 15:43 cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
Hue - Impala Editor - Qu...
quickstart.cloudera/#/tutorial/query_structured_data
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
cloudera LIVE
Navigation
1 957 Diamondback Women's Serene Classic Comfort Bi 4002333.3065795898
2 191 Nike Men's Free 5.0+ Running Shoe 3595240.4369888306
3 365 Perfect Fitness Perfect Rip Deck 3396753.8754653931
4 1073 Pelican Sunstream 100 Kayak 2958452.151260376
5 403 Nike Men's CJ Elite 2 TD Football Cleat 2859780.1208496094
6 502 Nike Men's Dri-FIT Victory Golf Polo 2819750
7 1014 O'Brien Men's Neoprene Life Vest 2651588.9155731201
8 627 Under Armour Girls' Toddler Spine Surge Runni 1265723.51121521
9 565 adidas Youth Germany Black/Red Away Match Soc 67830
Let's move on to the next section!!
If one of these steps fails, please reach out to our Cloudera Live Forum and get help. Otherwise continue.
CONCLUSION
Now you have learned how to create and query tables using Impala and that you can use regular interfaces and tools (such as SQL) within a Hadoop environment as
here being that you can do the same reports you usually do, but where the architecture of Hadoop vs traditional systems provides much larger scale and flexibility.
< Tutorial Exercise 1
Showing Big Data Value >
© 2015 Cloudera, Inc. All rights reserved | Terms & Conditions | Privacy Policy
Hadoop and the Hadoop elephant logo are trademarks of the Apache Software Foundation.
Cloudera Live : Welco...
cloudera@quickstart:~
3:43 PM
9/28/2015



Let's move on to the next scenario!!


cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 15:47 cloudera
Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox
Cloudera Live : Welcom...
Hue - Impala Editor - Qu...
quickstart.cloudera/#/tutorial/correlate_structured_and_unstructured_data
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
cloudera LIVE
Navigation
Tutorial Exercise 3
< Showing Big Data Value
Advanced Analytics in the Same Platform >
Tutorial Exercise 3
Correlate Structured Data with Unstructured Data
Since you are a pretty smart data person, you realize another interesting business question would be: are the most viewed products also the most sold? (or for other scenarios, the most searched for, the most chatted about…). Since Hadoop can store unstructured and semi-structured data alongside structured data without remodelling an entire database, you can just as well ingest, store and process web log events. Let's find out what site visitors have actually viewed the most.
For this, you need the web clickstream data. The most common way to ingest web clickstream is to use Flume. Flume is a scalable real-time ingest framework that allows you to route, filter, aggregate, and do "mini-operations" on data on its way in to the scalable processing platform.
In Exercise 4, later in this tutorial, you can explore a Flume configuration example, to use for real-time ingest and transformation of our sample web clickstream data. However, for the sake of tutorial-time, in this step, we will not have the patience to wait for three days of data to be ingested. Instead, we prepared a web clickstream data set (just pretend you fast forwarded three days) that you can bulk upload into HDFS directly.
Bulk Upload Data
For convenience, we have loaded a sample (about 20MM lines) of one month's worth of access log data into /opt/examples/log_data/access.log.2.
Let's move this data from the local filesystem, into HDFS.
Go back to your terminal and execute the following commands from your Master Node.
Copy
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -mkdir /user/hive/warehouse/original_access_logs
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -copyFromLocal /opt/examples/log_files/access.log.2 /user/hive/warehouse/original_access_logs
The copy command may take several minutes to complete.
Verify that your data is in HDFS by executing the following command:
Copy
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/original_access_logs
You should see a result similar to the following:
Eclipse
Cloudera Live : Welco...
cloudera@quickstart:~
3:47 PM
9/28/2015


cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 18:54 cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
oducts
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/categories/
Found 2 items
-rw-r--r--   1 cloudera hive   0 2015-09-28 15:05 /user/hive/warehouse/categories/_SUCCESS
-rw-r--r--   1 cloudera hive   1344 2015-09-28 15:05 /user/hive/warehouse/categories/part-m-00000.avro
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -mkdir /user/examples
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -chmod +rw /user/examples
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hadoop fs -copyFromLocal ~/*.avsc /user/examples/
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -mkdir /user/hive/warehouse/original_access_logs
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -copyFromLocal /opt/examples/log_files/access.log.2 /user/hive/warehouse/original_access_logs
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
Copy data into HDFS
Search
Go back to your terminal and execute the following commands from your Master Node.
Text Copied!
Copy
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -mkdir /user/hive/warehouse/original_access_logs
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -copyFromLocal /opt/examples/log_files/access.log.2 /user/hive/warehouse/original_access_logs
The copy command may take several minutes to complete.
Verify that your data is in HDFS by executing the following command:
Copy
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/original_access_logs
You should see a result similar to the following:
rgardner — root@cloudera1:~ — ssh — 105×25
[root@cloudera1 ~]# hadoop fs -ls /user/hive/warehouse/original_access_logs
Found 1 items
-rw-r--r--   2 hdfs hive 2862435400 2015-04-06 19:32 /user/hive/warehouse/original_access_logs/access.log
Cloudera Live : Welco...
cloudera@quickstart:~
6:54 PM
9/28/2015


cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 18:57 cloudera
Preferred Mail Reader
Send email
cloudera@quickstart:~
File Edit View Search Tern
[cloudera@quickstart ~]$ hadoop fs -copyFromLocal ~/*.avsc /user/examples/
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -mkdir /user/hive/warehouse/original_access_logs
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ sudo -u hdfs hadoop fs -copyFromLocal /opt/examples/log_files/access.log.2 /user/hive/warehous
original_access_logs
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ hadoop fs -ls /user/hive/warehouse/original_access_logs
Found 1 items
-rw-r--r--   1 hdfs hive   39593868 2015-09-28 18:54 /user/hive/warehouse/original_access_logs/access.log.2
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
ozilla Firefox
Search
Tutorial Exercise 3
Text Copied!
Copy
×25
[root@cloudera1 ~]# hadoop fs -ls /user/hive/warehouse/original_access_logs
Found 1 items
-rw-r--r--   2 hdfs hive 2862435400 2015-04-06 19:32 /user/hive/warehouse/original_access_logs/access.log
.2
[root@cloudera1 ~]#
Cloudera Live : Welco...
cloudera@quickstart:~
6:57 PM
9/28/2015

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)

Player

Applications Places System

Mon Sep 28, 19:04 cloudera

Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox

Cloudera Live : Welcom... | Hue - Impala Editor - Qu...

quickstart.cloudera/#/tutorial/correlate_structured_and_unstructured_data

Search

Cloudera | Hue | Hadoop | HBase | Impala | Spark | Solr | Oozie | Cloudera Manager | Getting Started

cloudera LIVE Navigation Tutorial Exercise 3

logs into individual fields using a regular expression. Second, you'll transfer the data from this intermediate table to one that does not require any special SerDe. Once the data is in this table, you can query it much faster and more interactively using Impala.

We'll query Hive using a command-line JDBC client for Hive called Beeline. You can invoke it from the terminal with the following:

Text Copied! Copy

```
[cloudera@quickstart ~]$ beeline -u jdbc:hive2://quickstart:10000/default -n admin -d org.apache.hive.jdbc.HiveDriver
```

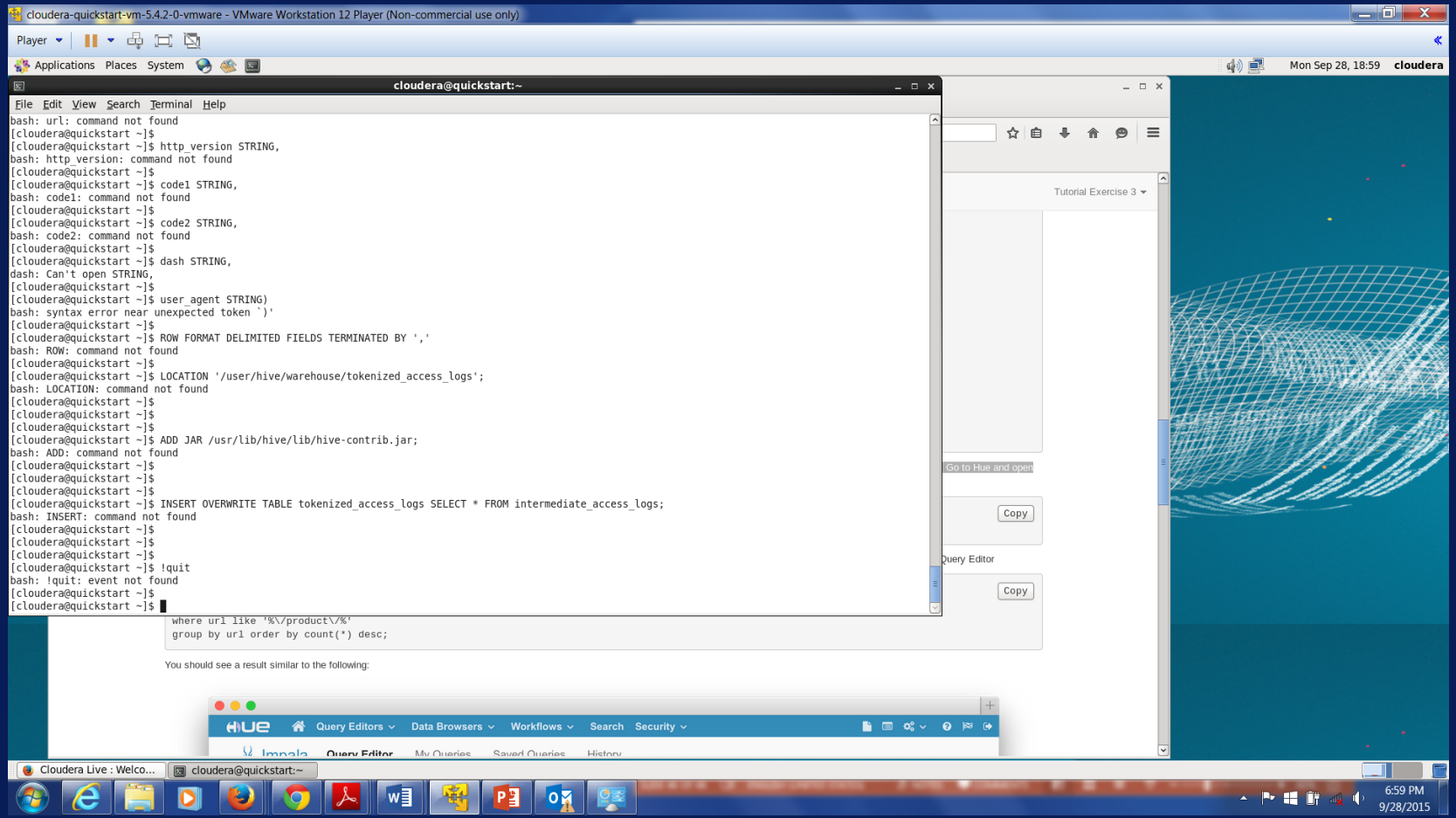
Once the Beeline shell is connected, run the following queries:

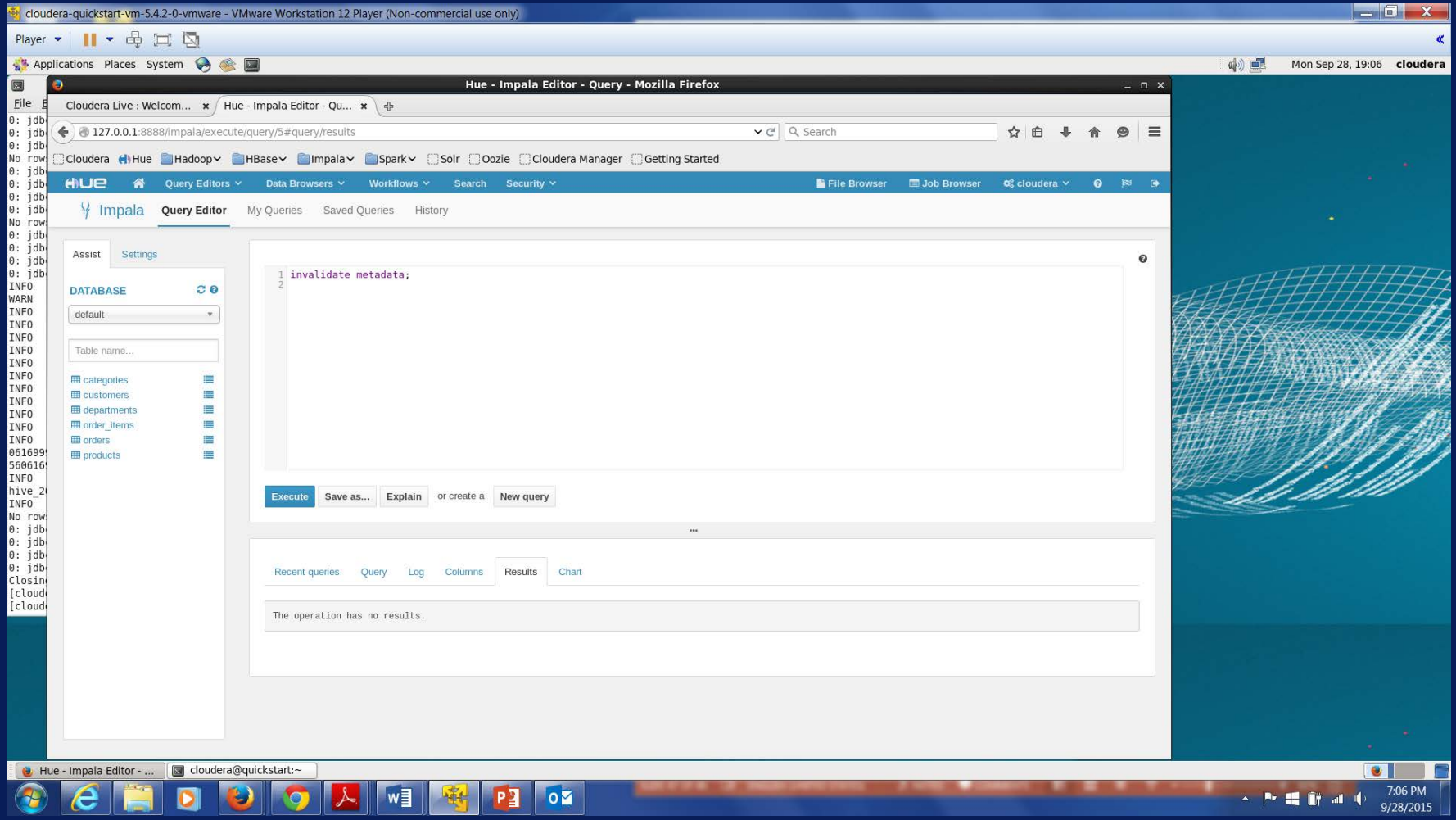
Text Copied! Copy

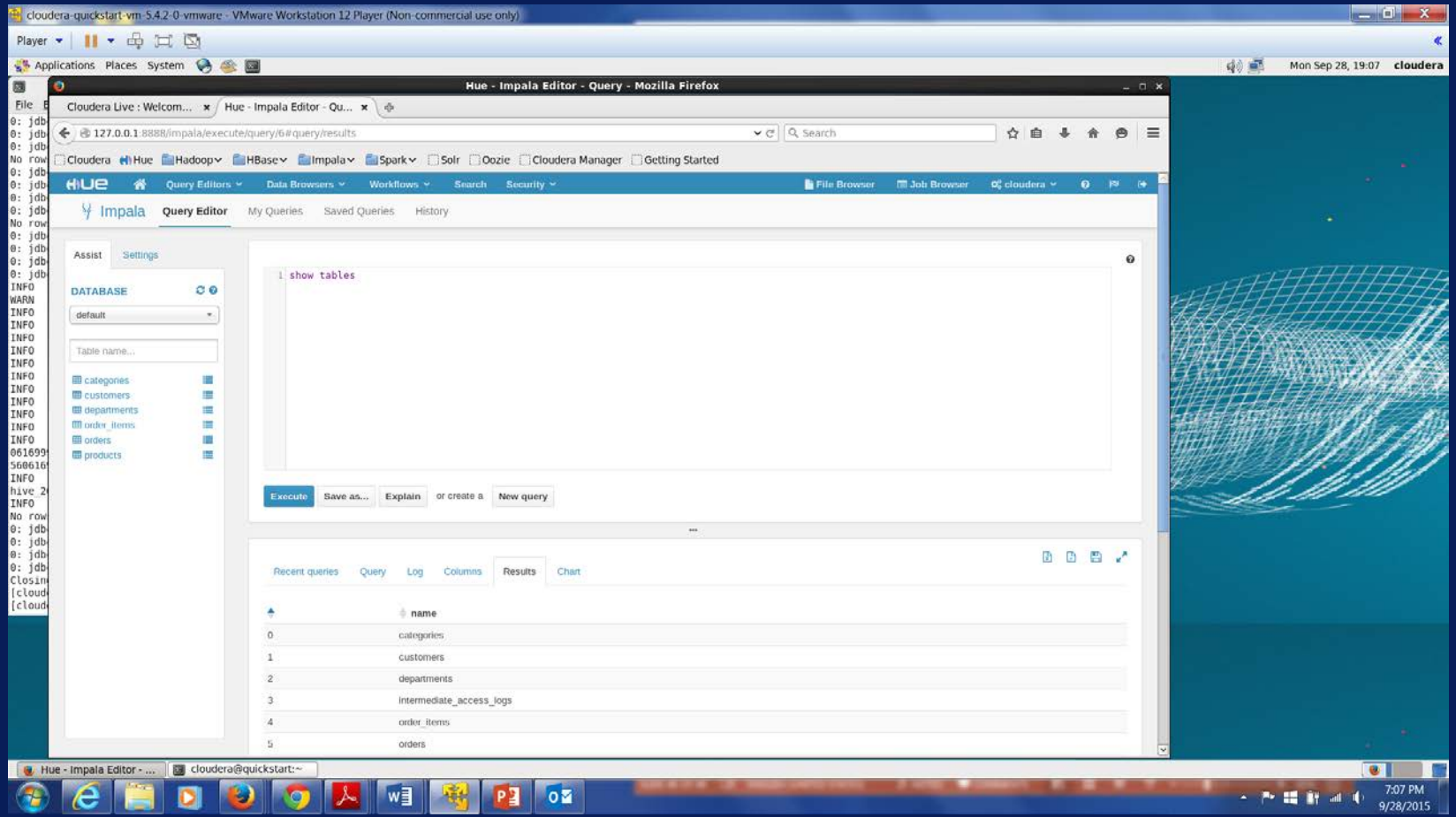
```
0: jdbc:hive2://quickstart:10000/default> CREATE EXTERNAL TABLE intermediate_access_logs (
 ip STRING,
 date STRING,
 method STRING,
 url STRING,
 http_version STRING,
 code1 STRING,
 code2 STRING,
 dash STRING,
 user_agent STRING)
 ROW FORMAT SERDE 'org.apache.hadoop.hive.contrib.serde2.RegexSerDe'
 WITH SERDEPROPERTIES (
 'input.regex' = '([^ ]*) - - \\[([^\\]]*)\\] "([^\ ]*) ([^\ ]*) ([^\ ]*)" (\\d*) (\\d*) "([^"]*)" "([^"]*)"',
 'output.format.string' = "%1$s %2$s %3$s %4$s %5$s %6$s %7$s %8$s %9$s"
 )
 LOCATION '/user/hive/warehouse/original_access_logs';

0: jdbc:hive2://quickstart:10000/default> CREATE EXTERNAL TABLE tokenized_access_logs (
 ip STRING,
 date STRING,
 method STRING,
 url STRING,
 http_version STRING,
 code1 STRING,
 code2 STRING,
 dash STRING,
 user_agent STRING)
 ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
```

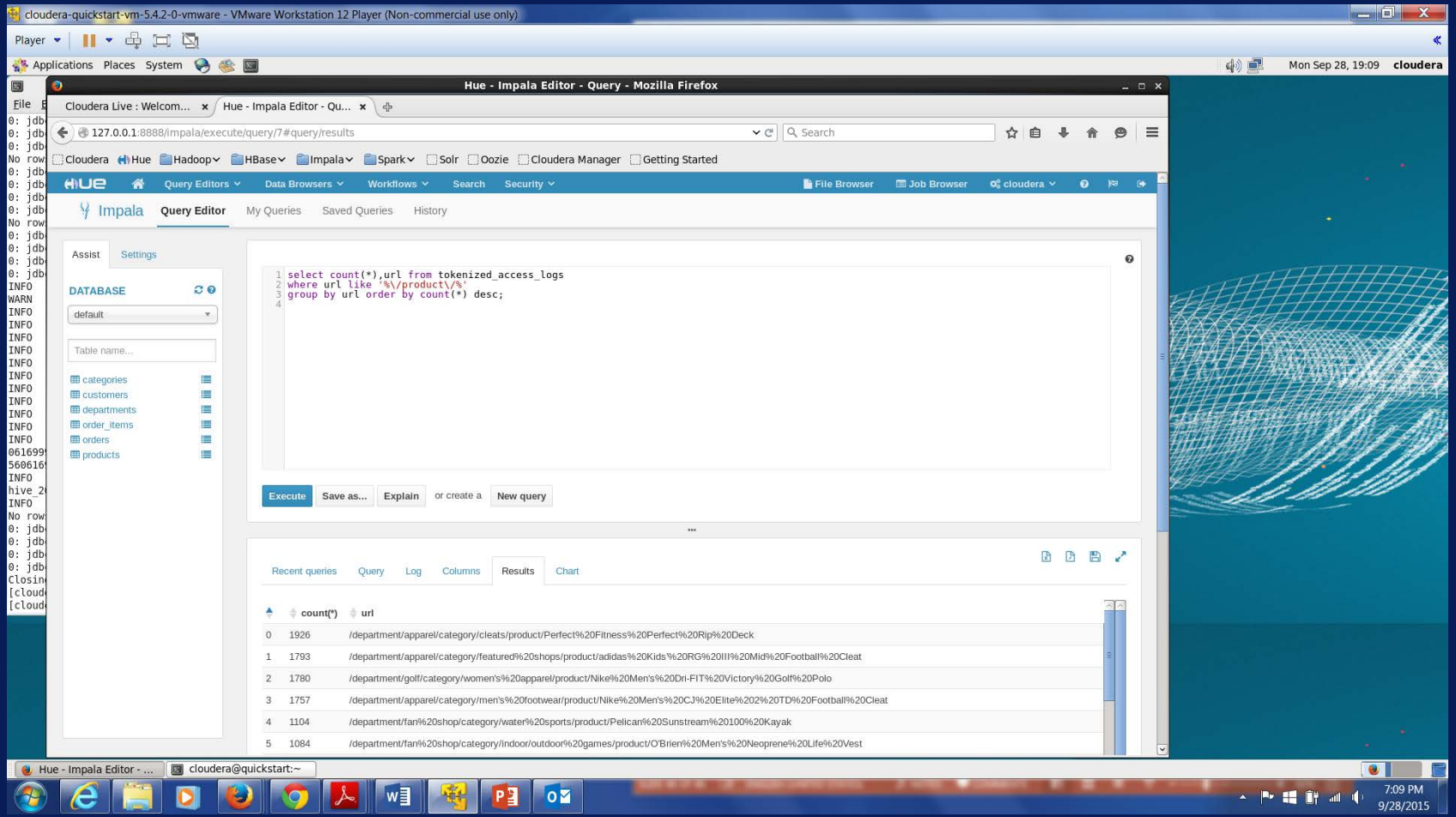
Cloudera Live : Welco... | cloudera@quickstart:~

7:04 PM 9/28/2015


cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 18:59 cloudera
cloudera@quickstart:~
File Edit View Search Terminal Help
bash: url: command not found
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ http_version STRING,
bash: http_version: command not found
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ code1 STRING,
bash: code1: command not found
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ code2 STRING,
bash: code2: command not found
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ dash STRING,
dash: Can't open STRING,
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ user_agent STRING)
bash: syntax error near unexpected token `)'
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
bash: ROW: command not found
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ LOCATION '/user/hive/warehouse/tokenized_access_logs';
bash: LOCATION: command not found
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ ADD JAR /usr/lib/hive/lib/hive-contrib.jar;
bash: ADD: command not found
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$ INSERT OVERWRITE TABLE tokenized_access_logs SELECT * FROM intermediate_access_logs;
bash: INSERT: command not found
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$!quit
bash: !quit: event not found
[cloudera@quickstart ~]$
[cloudera@quickstart ~]$
Tutorial Exercise 3
Go to Hue and open
Copy
Query Editor
Copy
where url like '%\/product\/%'
group by url order by count(*) desc;
You should see a result similar to the following:
Query Editors
Data Browsers
Workflows
Search
Security
Impala
Query Editor
My Queries
Saved Queries
History
Cloudera Live : Welco...
cloudera@quickstart:~
6:59 PM
9/28/2015


cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 19:06 cloudera
Hue - Impala Editor - Query - Mozilla Firefox
Cloudera Live : Welcom...
Hue - Impala Editor - Qu...
127.0.0.1:8888/impala/execute/query/5#query/results
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query Editors Data Browsers Workflows Search Security File Browser Job Browser cloudera
Impala Query Editor My Queries Saved Queries History
Assist Settings
DATABASE
default
Table name...
categories
customers
departments
order_items
orders
products
1 invalidate metadata;
2
Execute Save as... Explain or create a New query
Recent queries Query Log Columns Results Chart
The operation has no results.
Hue - Impala Editor - ...
cloudera@quickstart:~
7:06 PM
9/28/2015


cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 19:07 cloudera
Hue - Impala Editor - Query - Mozilla Firefox
Cloudera Live : Welcom...
Hue - Impala Editor - Qu...
127.0.0.1:8888/impala/execute/query/6#query/results
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query Editors Data Browsers Workflows Search Security File Browser Job Browser cloudera
Impala Query Editor My Queries Saved Queries History
Assist Settings
DATABASE
default
Table name...
categories
customers
departments
order_items
orders
products
1 show tables
Execute Save as... Explain or create a New query
Recent queries Query Log Columns Results Chart
name
0 categories
1 customers
2 departments
3 intermediate_access_logs
4 order_items
5 orders
Hue - Impala Editor - ...
cloudera@quickstart:~
7:07 PM
9/28/2015


cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)
Player
Applications Places System
Mon Sep 28, 19:09 cloudera
Hue - Impala Editor - Query - Mozilla Firefox
Cloudera Live : Welcom...
Hue - Impala Editor - Qu...
127.0.0.1:8888/impala/execute/query/7#query/results
Search
Cloudera Hue Hadoop HBase Impala Spark Solr Oozie Cloudera Manager Getting Started
Query Editors Data Browsers Workflows Search Security File Browser Job Browser cloudera
Impala Query Editor My Queries Saved Queries History
Assist Settings
DATABASE
default
Table name...
categories
customers
departments
order_items
orders
products
select count(*),url from tokenized_access_logs
where url like '%\/product\/%'
group by url order by count(*) desc;
Execute Save as... Explain or create a New query
Recent queries Query Log Columns Results Chart
count(*) url
0 1926 /department/apparel/category/cleats/product/Perfect%20Fitness%20Perfect%20Rip%20Deck
1 1793 /department/apparel/category/featured%20shops/product/adidas%20Kids'%20RG%20III%20Mid%20Football%20Cleat
2 1780 /department/golf/category/women's%20apparel/product/Nike%20Men's%20Dri-FIT%20Victory%20Golf%20Polo
3 1757 /department/apparel/category/men's%20footwear/product/Nike%20Men's%20CJ%20Elite%202%20TD%20Football%20Cleat
4 1104 /department/fan%20shop/category/water%20sports/product/Pelican%20Sunstream%20100%20Kayak
5 1084 /department/fan%20shop/category/indoor/outdoor%20games/product/O'Brien%20Men's%20Neoprene%20Life%20Vest
Hue - Impala Editor - ...
cloudera@quickstart:~
7:09 PM
9/28/2015

cloudera-quickstart-vm-5.4.2-0-vmware - VMware Workstation 12 Player (Non-commercial use only)

Player | Applications Places System | Mon Sep 28, 19:11 cloudera

Cloudera Live : Welcome! - Cloudera Live Beginner Tutorial - Mozilla Firefox

Cloudera Live : Welcom... | Hue - Impala Editor - Qu...

quickstart.cloudera/#/tutorial/correlate_structured_and_unstructured_data

Cloudera | Hue | Hadoop | HBase | Impala | Spark | Solr | Oozie | Cloudera Manager | Getting Started

cloudera LIVE Navigation Tutorial Exercise 3

By introspecting the results you quickly realize that this list contains many of the products on the most sold list from previous tutorial steps, but there is one product that did not show up in the previous result. There is one product that seems to be viewed a lot, but never purchased. Why?

Recent queries | Query | Log | Columns | Results | Chart

	product_id	product_name	revenue
0	1004	Field & Stream Sportsman 16 Gun Fire Safe	6637668.282318115
1	365	Perfect Fitness Perfect Rip Deck	4233794.3682899475
2	957	Diamondback Women's Serene Classic Comfort Bi	3946837.004547119
3	191	Nike Men's Free 5.0+ Running Shoe	3507549.2067337036
4	502	Nike Men's Dri-FIT Victory Golf Polo	3011600
5	1073	Pelican Sunstream 100 Kayak	2967851.6815185547
6	1014	O'Brien Men's Neoprene Life Vest	2765543.314743042
7	403	Nike Men's CJ Elite 2 TD Football Cleat	2763977.4868011475
8	627	Under Armour Girls' Toddler Spine Surge Runni	1214896.220287323
9	565	adidas Youth Germany Black/Red Away Match Soc	63490

Recent queries | Query | Log | Columns | Results | Chart

	count(*)	url	
0	248650	/department/apparel/category/featured%20shops/product/adidas%20Kids'%20RG%20III%20Mid%20Football%20Cleat	Missing???
1	248128	/department/apparel/category/cleats/product/Perfect%20Fitness%20Perfect%20Rip%20Deck	2nd
2	247914	/department/apparel/category/men's%20footwear/product/Nike%20Men's%20CJ%20Elite%202%20TD%20Football%20Cleat	8th
3	247456	/department/golf/category/women's%20apparel/product/Nike%20Men's%20Dri-FIT%20Victory%20Golf%20Polo	5th
4	149182	/department/fan%20shop/category/indoor/outdoor%20games/product/O'Brien%20Men's%20Neoprene%20Life%20Vest	7th
5	148995	/department/fan%20shop/category/fishing/product/Field%20&%20Stream%20Sportsman%2016%20Gun%20Fire%20Safe	1st!
6	148495	/department/fan%20shop/category/water%20sports/product/Pelican%20Sunstream%20100%20Kayak	6th
7	147921	/department/fan%20shop/category/camping%20&%20hiking/product/Diamondback%20Women's%20Serene%20Classic%20Comfort%20Bi	3rd
8	124969	/department/footwear/category/cardio%20equipment/product/Nike%20Men's%20Free%205.0+%20Running%20Shoe	4th
9	124555	/department/golf/category/shop%20by%20sport/product/Under%20Armour%20Girls'%20Toddler%20Spine%20Surge%20Runni	9th

Well, in our example with DataCo, once these odd findings are presented to your manager, it is immediately escalated. Eventually, someone figures out that on that view page, where most visitors

Cloudera Live : Welco... | cloudera@quickstart:~

7:11 PM 9/28/2015

Assignment