Data Analysis using Hadoop: Module 4, Lesson 6  
Hive CLI and Hive View Hands-On Lab

## Overview

In this lab, you will ssh into your provisioned HDInsight cluster ([Module 4 Lesson 3 Lab](https://github.com/MSFTImagine/computerscience/tree/master/Complimentary%20Course%20Content/Module4/Labs)) using PuTTY. You will then run Hive from both the Hive CLI (Command Line Interface) and Azure Ambari Hive View.

## Objectives

In this hands-on lab you will learn how to:

* Run Hive from the Command Line Interface
* Run Hive from Hive View
* Create Databases, managed and external tables

## Prerequisites

The following are required to complete this hands-on lab:

* A provisioned HDInsight cluster
* A web browser
* A Windows machine on which the [PuTTY](http://www.putty.org/) has been installed
* A Windows machine on which [WinsCp](https://winscp.net/eng/download.php) has been installed

Note: The Azure portal is continually improved and changed. The steps in this exercise reflect the user interface of the Microsoft Azure portal at the time of writing, but may not match the latest design of portal.

## Exercises

This hands-on lab includes the following exercises:

* Exercise 1: Setting up your data
* Exercise 2: Hive Command Line Interface
* Exercise 3: Run and debug the application

## Exercise 1: Setting Up the Data

In this exercise, you will learn how to bring data into your HDInsight cluster.

1. From Putty, sign in connect to the HDInsight cluster

(Refer to previous exercise 2 in [Module 4 Lesson 3 Lab](https://github.com/MSFTImagine/computerscience/tree/master/Complimentary%20Course%20Content/Module4/Labs)).

1. Create a directory in your cluster home directory.

From your Linux home directory, create a new directory. Into which you will copy the data that you will use throughout the exercises.

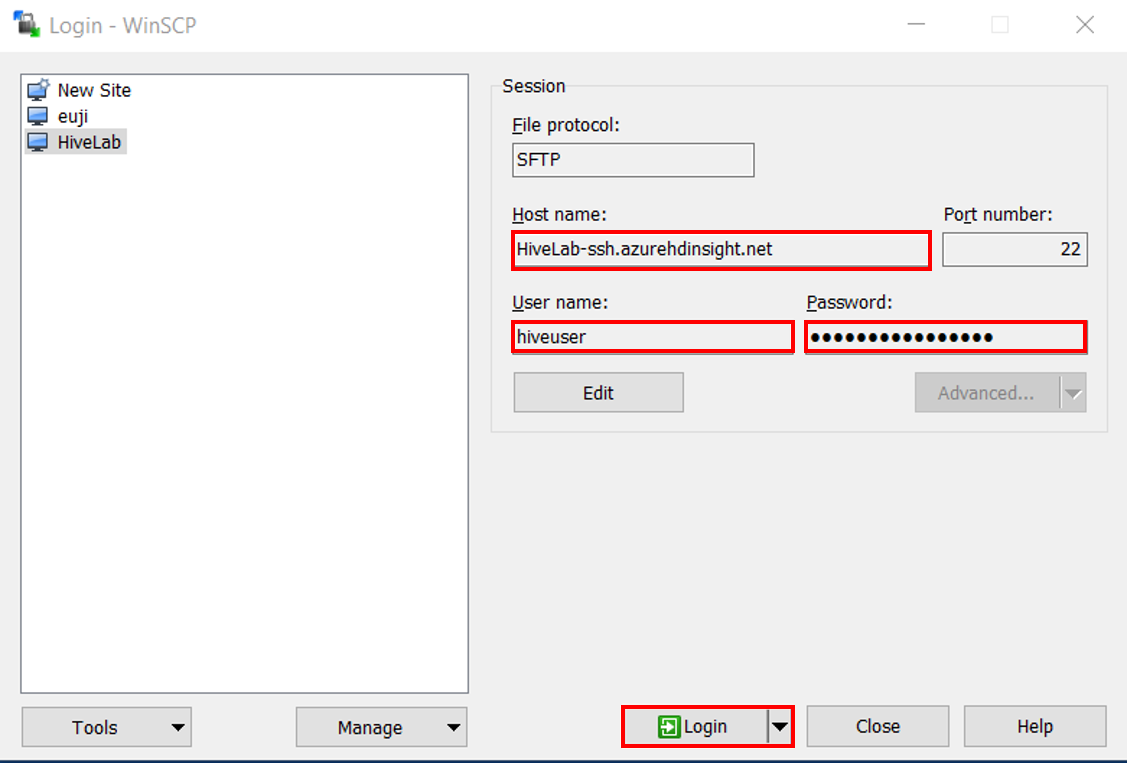
$> mkdir labdata

1. Import the sample data to your Windows computer

Download the sample data from [us-500.csv](https://docs.google.com/file/d/0BzwmWW0VpDH1NmJMT0Q1SkZIV0E/view?pref=2&pli=1) and save it to a new directory on your Windows computer

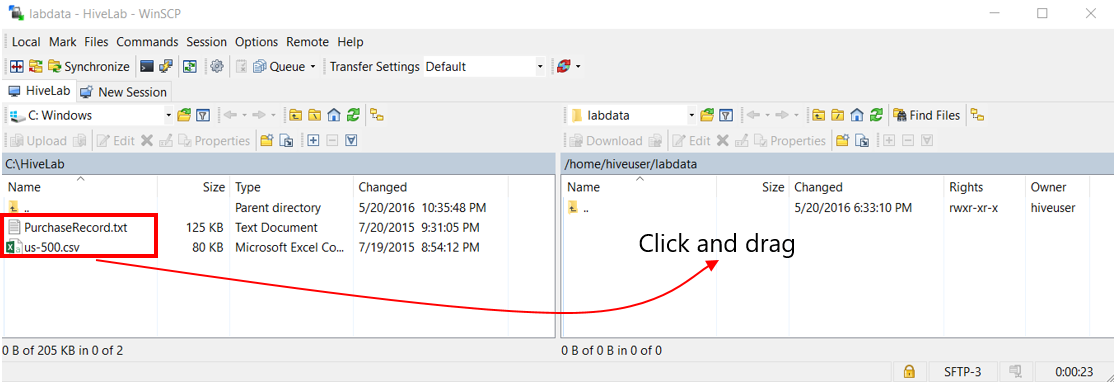
1. Connect WinsCP to the HDInsight Cluster

Start WinsCp and enter the host name, ssh user name and password. These parameters are the same as you entered when configuring PuTTY. Click login to connect.

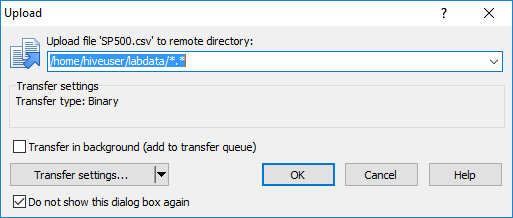


1. Copy/Move sample data to the HDInsight Cluster

On the right pane, select the labdata directory that you created in step 2 and double click. On the left pane, navigate to the folder containing your data in your Windows computer and click and drag sample data to the labdata directory on the right pane.



If it is your first time using WinsCp you will be prompted with this dialog:



Click OK.

## Exercise 2: Hive Command Line Interface

In this exercise we will explore some of basic commands available in the Hive Command Line Interface.

1. From your terminal in Putty, start the Hive Command Line Interface (CLI).

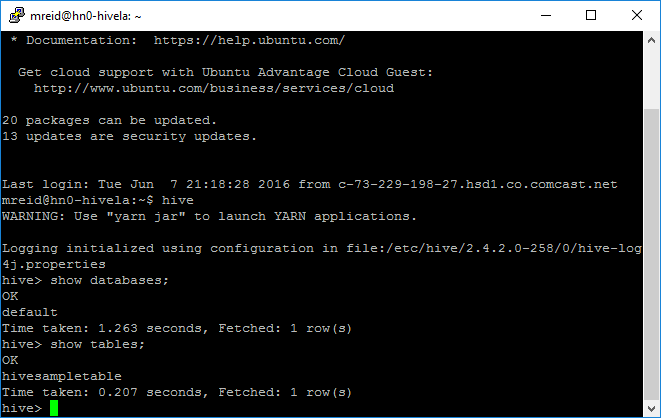
$> hive

1. From Hive CLI, check existing databases and tables in Hive

At this point, there will only be the default database. HDInsight includes a hivesampletable.

hive> show databases;

hive> show tables;



1. Review the table schema

Use the “describe” command to review the schema for this sample table.

hive> describe hivesampletable;

OK

clientid string

querytime string

market string

deviceplatform string

devicemake string

devicemodel string

state string

country string

querydwelltime double

sessionid bigint

sessionpagevieworder bigint

Time taken: 1.988 seconds, Fetched: 11 row(s)

1. Create a new database and start using it

You misspell hivelab and accidently create a new database named hivemab. Delete it, create a new one properly and start using hivelab.

hive> create database hivemab;

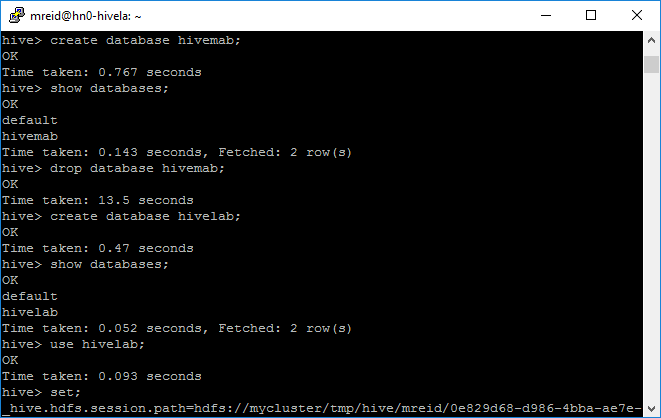
hive> show databases;

hive> drop database hivemab;

hive> create database hivelab;

hive> show databases;

hive> use hivelab;



1. Explore set variables in Hive CLI

Hive has many variables. ‘set’ prints all the variables in the hivevar (user-defined custom variables), hiveconf (Hive-specific configuration properties), system (configuration properties defined by Java) and env (Environmental variables defined by the shell) namespace. ‘set –v’ option prints in addition, all the properties defined by Hadoop. You can also use the set command to set property values.

hive> set;

hive> set -v;

hive> set env:HOME;

hive> set hive.cli.print.header=true;

hive> set hivevar:platform=Android;

1. Interact with the Operating System and HDFS

You can use ! to execute shell commands, however, neither pipes nor wild cards are supported. You can call HDFS commands by prefixing them with dfs. In Hive, a table is simply a directory in HDFS within the Hive warehouse directory. The contents of a table are all the files in that directory.

There is a subdirectory called “hivesampletable” in the /hive/warehouse directory. Within this directory, there is a file called HiveSampleData.txt which is a tab delimited text file representing the contents of the table.

hive> ! ls;

hive> ! date;

hive> dfs -ls /hive/warehouse/hivesampletable;

hive> dfs -cat /hive/warehouse/hivesampletable/HiveSampleData.txt;

1. Simple Query of hivesampletable

Try out a few simple HiveQL queries. First, go back to the default database where the hivesampletable exists. We first query all the columns in hivesampletable but limit the output to 50 rows. Next, we will query how long users dwelled on a device. Once again, we will limit the output to 50 rows. Finally, we will query the phone model that users looked at for Samsung phones.

hive> Use default;

hive> select \* from hivesampletable limit 50;

hive> select devicemake, querydwelltime from hivesampletable limit 50;

hive> select devicemodel from hivesampletable where devicemake='Samsung';

TRY IT:

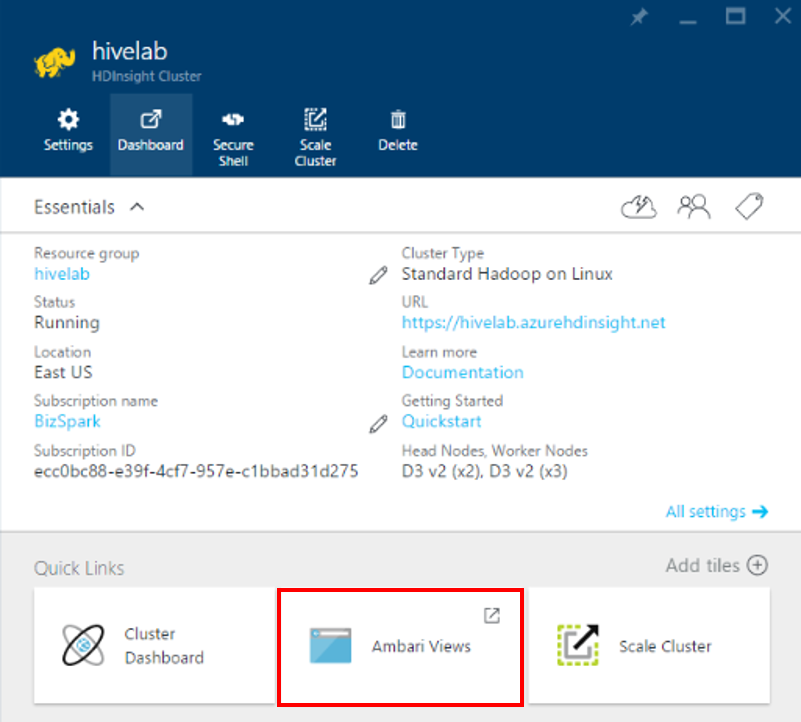
Query all items where the user dwelled on the item for more than 30 seconds.

How many items are there all together? Of these, how many were Samsung?

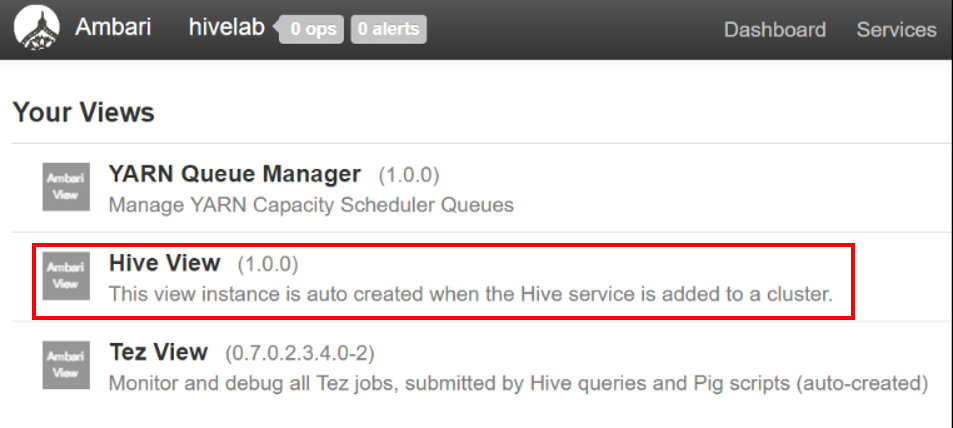
## Exercise 3: Hive View

The Hive Command Line Interface is very useful in certain situations but often, it is preferable to use a graphical interface to interact with Hive. In this exercise, we will explore the Hive View interface.

1. Select the hivelab HDInsight Cluster from the Azure portal



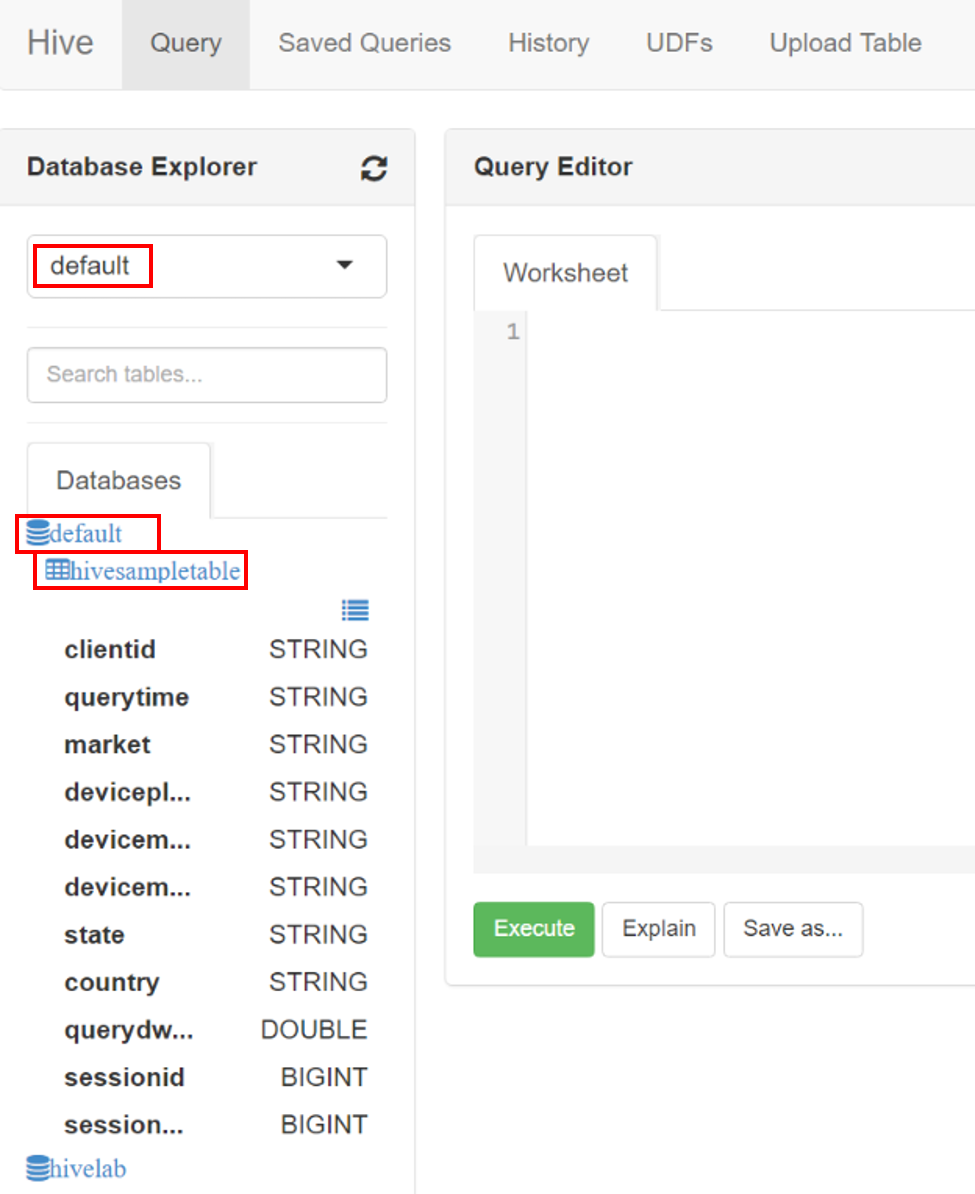
Once you have selected the hivelab HDInsight cluster, you will see the “Quick Links” section. From there, enter ‘Your Views’ by clicking on the table image in the navigation bar at the top of the window, then select Hive View.



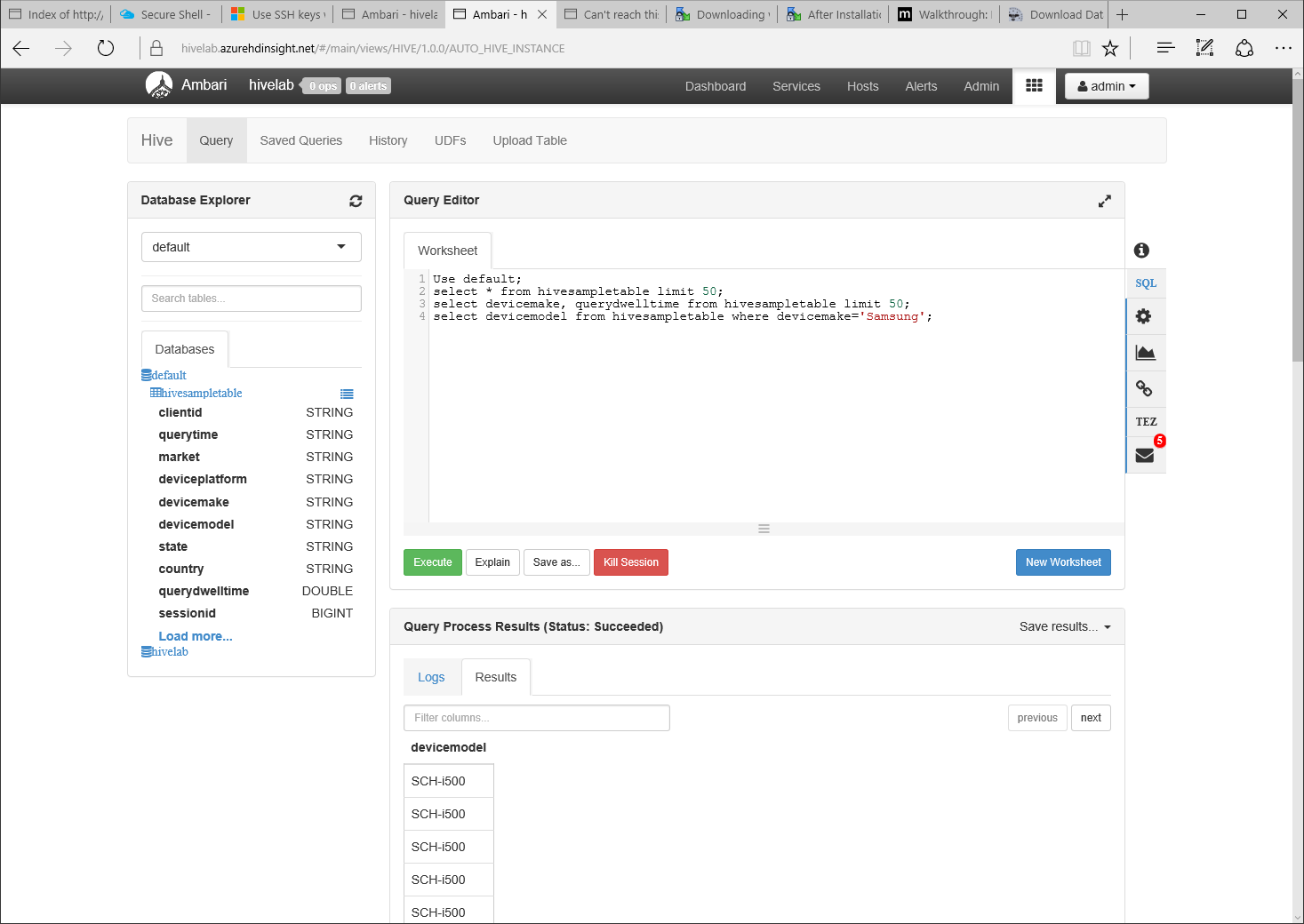
1. Databases, tables, and table schema from Hive View

Hive View shows all the databases that Hive is currently managing in the Database Explorer. Users can choose which database to use by clicking on the drop-down menu under Database Explorer and selecting the desired database.

In the Databases tab, you can drill down into the tables that are in a particular database. Here we see that the default database contains a single table – hivesampletable. By clicking on the table icon, the user can review the schema of the table.



1. Try all the commands we explored in the Hive Command Line Interface in the GUI.



TRY IT:

In the Query Editor to the right, try out all the commands that we explored in Exercise 2 under Hive CLI. Do they all work?

## Summary

In this hands-on lab, you learned how to:

* Move data into your HDInsight cluster using WinsCp
* Start Hive from the hive command line interface
* Create databases and issue queries against tables
* Set hive user-defined variables
* Start Hive from Hive View and explore databases and tables