Data Analysis using Hadoop: Module 4, Lesson 9  
HiveQL Hands-On Lab

## Overview

In the lab, you will practice using HiveQL to query data.

## Objectives

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

* Upload a table from Hive View
* Use WHERE clause
* Use GROUP BY clause
* Use Having clause
* Use Sort by clause
* Use aggregate and other functions

## Prerequisites

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

* A provisioned HDInsight cluster ([Module 4 Lesson 3 Lab](https://github.com/MSFTImagine/computerscience/tree/master/Complimentary%20Course%20Content/Module4/Labs))
* A web browser
* [PuTTY](http://www.putty.org/)
* [WinsCp](https://winscp.net/eng/download.php)
* Sample data from [Module 4 Lesson 6](https://github.com/MSFTImagine/computerscience/tree/master/Complimentary%20Course%20Content/Module4/Labs) Hive CLI and Hive View

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:

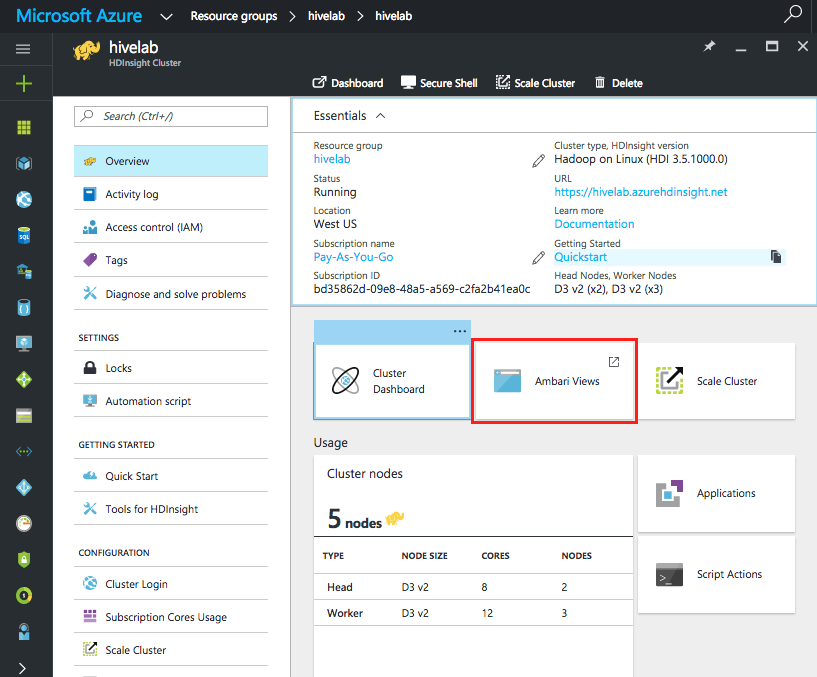
* Exercise 1: Upload a table from Hive View
* Exercise 2: Run queries from Query Editor

## Exercise 1: Upload a table from Hive View

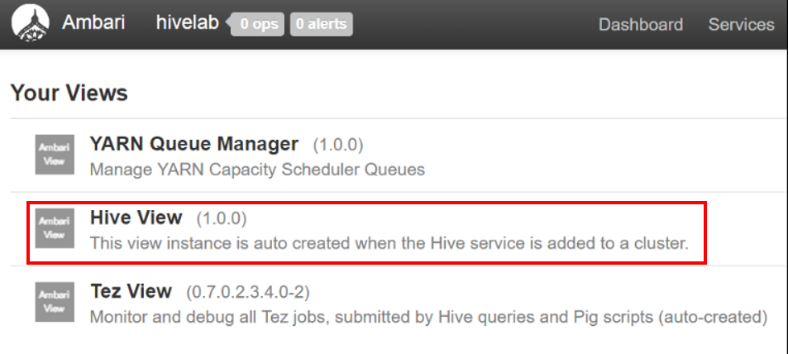
In this exercise, students will learn how to use Hive View to upload an external table.

1. Access the Hive View for the hivelab HDInsight cluster

Select the hivelab HDInsight Cluster from the Azure portal

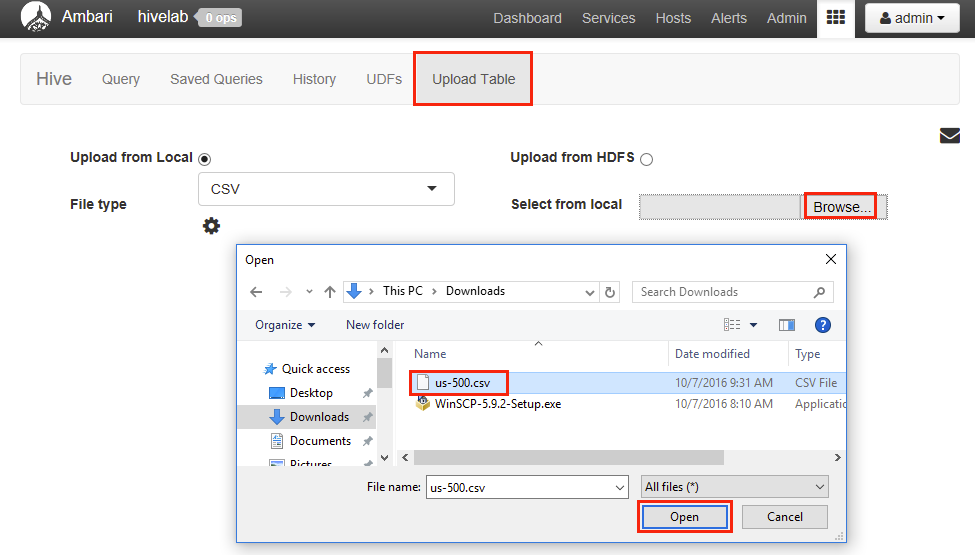


From the **Quick Links** section in the hivelab HDInsight cluster, click on **Ambari Views** and then select **Hive View**.



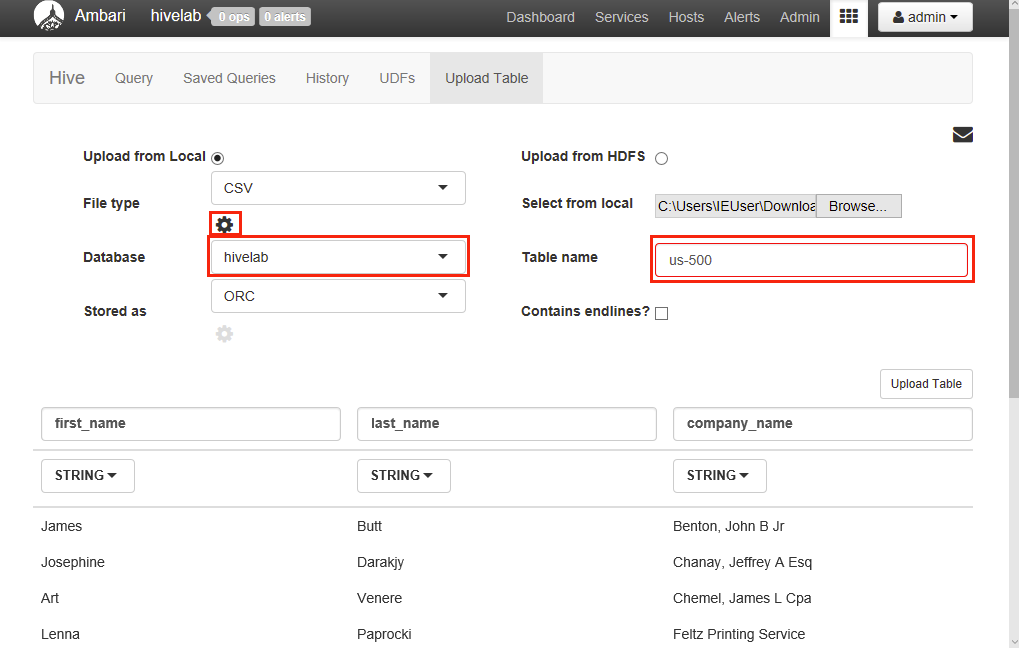
1. Select data to upload

From the top menu, click **Upload Table**, and the click **Choose File** to select the data file from which to upload the data. In a previous lab, students have downloaded a sample data set to their Windows computer. Select **us-500.csv**, a comma separated Excel table. Click **Open** to start uploading the file to Hive View



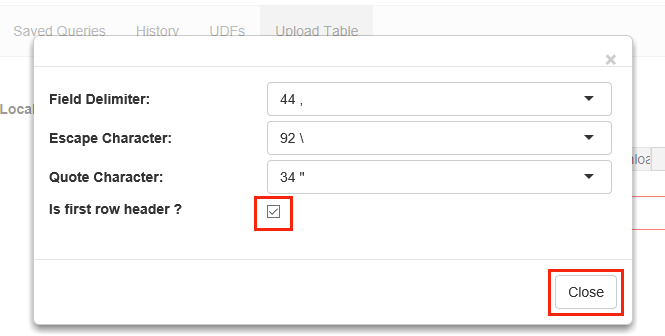
1. Setup the database and table column names

From the **Database** drop down menu, select **hivelab**. (We created this database in a previous lab. It is not necessary for the student to have successfully created this database to proceed. If hivelab is unavailable, either create it now or simply use the default database)



Select the “Settings” icon next to the File type field.

Make sure that the **Is First Row Header?** Checkbox is checked.



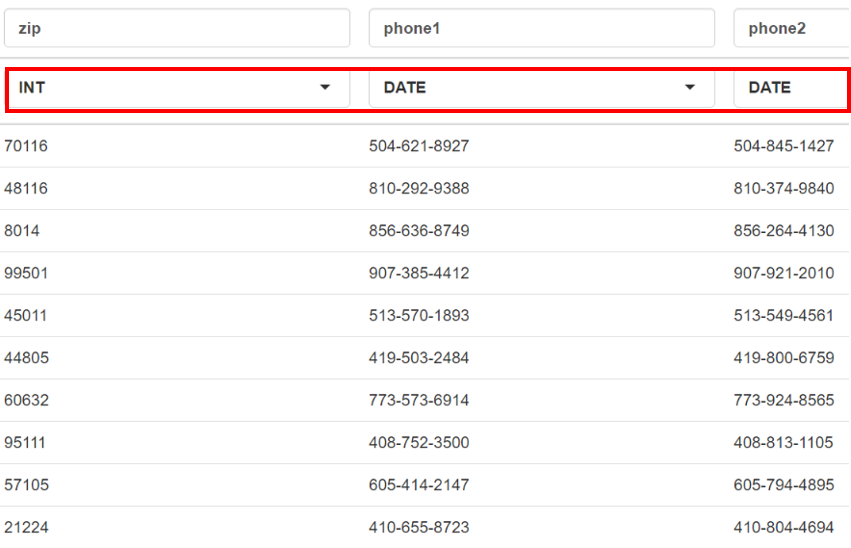
1. Review column data types and adjust as necessary

Review each of the column names and column data types. Scroll over to the right to view additional columns. Towards the end of the columns, there are 3 columns that have mismatched data types. Make the following changes:

**zip** 🡺 **STRING**

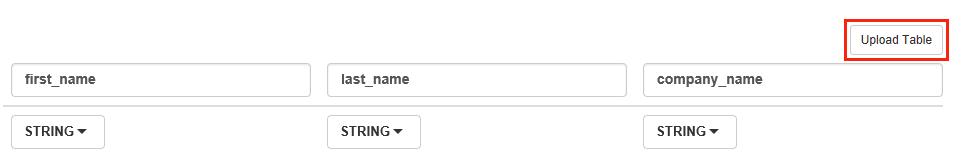
**phone1** 🡺 **STRING**

**phone2** 🡺 **STRING**



1. Click Upload Table

Once all the adjustments have been made, click on **Upload Table** on the right to begin uploading the data. .



Oops. What happened? There is an error stating that there is an error in the input. This is because Hive does not allow hyphens in its table names. Remove “**-**“ and try again.

## Exercise 2: Run queries from Query Editor

Now that the data has been uploaded to Hive, run some queries against this data.

1. Select the Database from **Database Explorer** under the **Query** tab.

Select the database into which the data was uploaded in Exercise 1. This will depend on whether the student saved it under the **default** database or the **hivelab** database.

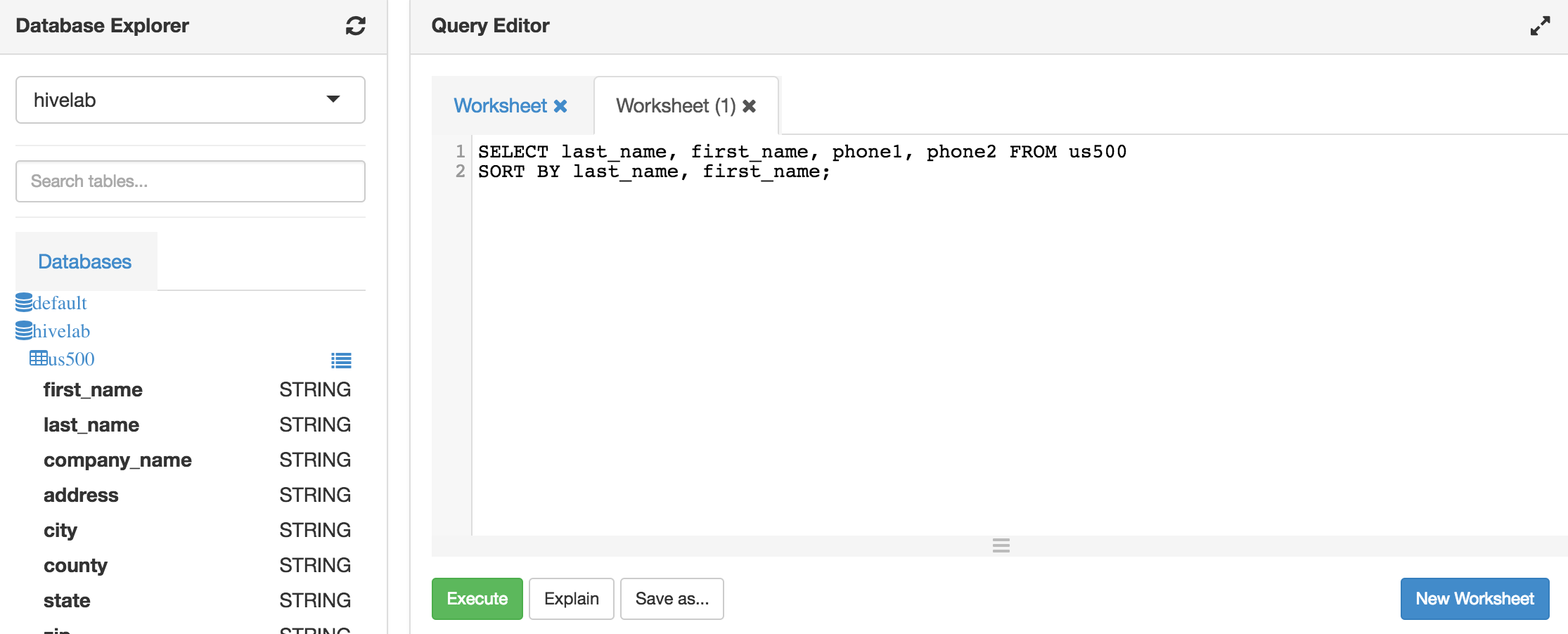
1. Create a phone catalog of the customer

Use the **New Worksheet** button at the bottom of the **Query Editor** to create a new worksheet. In the new worksheet, enter the following HiveQL statements:

SELECT last\_name, first\_name, phone1, phone2 FROM us500

SORT BY last\_name, first\_name;

Click **EXECUTE** to run the query.



TRY IT:

Create a phone catalog of customers from California only

Create a mailing list for the customers.

1. How many customers in each State

Use the **New Worksheet** button at the bottom of the **Query Editor** to create a new worksheet. In the new worksheet, enter the following HiveQL statements:

SELECT count(company\_name), state FROM us500

GROUP BY state;

Click **EXECUTE** to run the query. Something seems to be amiss with the data. The state field contains a lot of incorrect data. What happened?

1. How many customers have problem data?

Use the **New Worksheet** button at the bottom of the **Query Editor** to create a new worksheet. In the new worksheet, enter the following HiveQL statements:

The problem is that there are many company\_name fields that contain a comma in the middle of the field. This is causing Hive to interpret this as two separate fields. The CSV file accommodates the comma in the middle by placing the entire company\_name inside quotes(“).

Create and execute a query to filter out companies that has a quote (“) as the first character of its field. Hint: Use the function substr(field\_name, start\_index, length) to isolate the first character and compare with quote (“).

SELECT company\_name from us500

WHERE substr(company\_name,1,1) = '"';

Click **EXECUTE** to run the query.

It is very important to understand how data will be imported/loaded into Hive tables. Users must ensure that the data being loaded is in the proper format expected by Hive.

1. Which States in the US have clients most interested in new phones?

Change the database to **default** by selecting it in the **Database Explorer**.

Use the **New Worksheet** button at the bottom of the **Query Editor** to create a new worksheet. In the new worksheet, enter the following HiveQL statements:

SELECT avg(querydwelltime), state, country FROM hivesampledata

GROUP BY country, state

HAVING country = 'United States';

Click **EXECUTE** to run the query.

Customers who dwell on a query the longest seems to be a good logical choice in determining their level of interest in the phone and most likely customer to make a purchase.

The data needs to be grouped by country and state first and then from that group, we are only interested in customers from the United States. The average dwelling time of the customer is used as the indicator of level of interest.

TRY IT:

Which platform had the longest average query?

Which platform had the most number of queries?

How did number of queries in a platform vary depending on country?

## Summary

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

* Use Hive View to upload data to a table
* Use Hive View to create queries using WHERE, GROUP BY, SORT BY and HAVING clause
* Use aggregate functions in the query clause