Hands-On Exercise: Data Management

Files and Data Used in This Exercise

Exercise directory	$ADIR/exercises/data_mgmt
HDFS data directory	/analyst/dualcore/employees
Local data files	$ADIR/data/ratings_2012.txt $ADIR/data/ratings_2013.txt
Hive/Impala tables	customers

In this exercise, you will practice using several common techniques for creating and populating tables.
IMPORTANT: This exercise builds on previous ones.

Reviewing Existing Tables using the Table Browser

1. In Firefox, visit the Hue home page, click the menu toggle to the left of the Hue icon, and then click **Browsers > Tables**.

2. Make sure analyst database is selected.

3. Click the link for the customers table in the main panel to display the table overview and review the list of columns.

4. Click the **Sample** tab to view the first hundred rows of data.

Creating and Loading a Table using the Table Browser

Do the following to create and then load a table with product ratings data.

5. Before creating the table, review the files containing the product ratings data. The files are in $ADIR/analyst/data. You can use the head command in a terminal window to see the first few lines:

```
$ head $ADIR/data/ratings_2012.txt
$ head $ADIR/data/ratings_2013.txt
```

6. Copy the data files to the $ADIR/analyst/dualcore directory in HDFS. You may use either the Hue File Browser, or the hdfs command in the terminal window:

```
$ hdfs dfs -put $ADIR/data/ratings_2012.txt /analyst/dualcore/
$ hdfs dfs -put $ADIR/data/ratings_2013.txt /analyst/dualcore/
```

7. Return to the Table Browser in Hue. If the sample data is still shown, click on **analyst** in the breadcrumbs to return to the table list. Click on the **+ New** button in the upper right to start the table definition wizard.

8. The first wizard step is to pick whether to add the data at creation using a file, or to create an empty table so you can add the data later.

- With **Type** showing **file**, click the field next to **Path**. Navigate up the directory hierarchy to find the $ADIR/analyst/dualcore directory and choose ratings_2012.txt. (You will load the 2013 data later.)
- Check the information that has been added to the main panel, to verify that Hue is interpreting the data correctly. You should see **Format** options, including **Field Separator** (set to **^Tab()**), **Record Separator** (set to **New line**), and **Quote Character** (set to **Double Quote**). The **Has Header** box should *not* be checked. If any of these are set incorrectly, correct them. You should also see a **Preview** that shows the data separated into fields (labeled **field_1**, **field_2**, and so on).
- Click the **Next** button.

9. The next step is to set up the table specifications.
- Under **DESTINATION**, click in the **Name** field and change the supplied table name (based on the file name) to analyst.ratings. (This names the new table ratings and puts it in the analyst database.)
- Under **PROPERTIES**, choose the file format. **Format** should show **Text**. Correct it if needed.
- Click **Extras** to see the options provided there. The settings there are correct for this table, but note that this allows you to change your mind on some of the settings from the previous step. It also allows you to add a description for your table, and to set delimiters for Array, Map, and Struct fields. For this simple table, only the field terminator is relevant; collection and map delimiters are used for complex data in Hive and are not needed at this time.
- Scroll down to the **Fields** section. Notice that the field types are selected, however, you should check that these are correct. (For example, Hue typically chooses bigint for all integer fields, but perhaps you know that int or even tinyint is more appropriate). Use the following descriptions of the fields to add the field names and correct the types as needed:

Field Name	Field Type
posted	timestamp
cust_id	int
prod_id	int
rating	tinyint
message	string

- When you have added all the columns, click **Submit**. This will start a task to define the table in the metastore and create the warehouse directory in HDFS to store the data.
- When the task is complete, a **Task History** pop-up window will appear. Check that no error message is given, then click the **X** to close the pop-up window.

10. The new table ratings might appear in the analyst database in the Table Browser. Scroll down to confirm that the fields are correct and the data has been added to the table. If the table or the data does not appear, click the **Query** button to go to the Impala query editor and run the command:

```
REFRESH analyst.ratings;
```

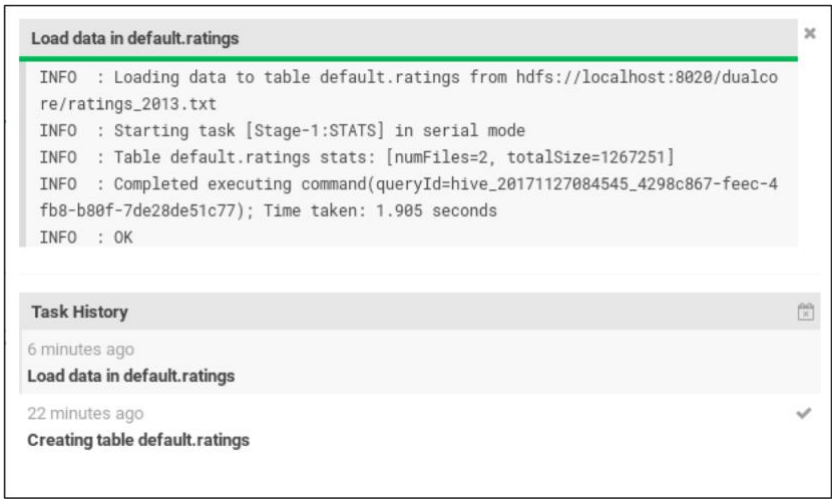
Return to the Table Browser and verify that the table is in the analyst database, with the data included. If it doesn't appear, click the **Refresh** button (↻) and choose **Clear cache**, then click **Refresh**. (This is just the cache for the display, not the Impala metadata cache). You might also need to refresh the browser window.

11. Try querying the data in the table. In Hue, click the **Query** button to switch to the Impala Query Editor. Check that the current database in the query editor window is set to analyst. Set it if needed.

Try counting the number of ratings:
SELECT COUNT(*) FROM ratings;
The total number of records should be 464.

12. You can also load data to an existing table. One way to do this is in Hue; another is to use the LOAD DATA INPATH command. Try doing it using Hue:

- Return to the Hue Table Browser, and click the link for the new ratings table in the main panel.
- Then click the **Import** button in the upper right corner.
- In the **Import Data** dialog, enter or browse to the HDFS location of the 2013 product ratings data file: $ADIR/analyst/dualcore/ratings_2013.txt. Be sure that the **Overwrite existing data** box is *not* checked, and then click **Submit**.
- The **Task History** pop-up window should appear again. Check that the data loaded without error. It should look similar to this:



Click the X to close the Task History window.

13. *Optional:* Verify that the 2013 data is shown alongside the 2012 data in the table's warehouse directory.

14. With the additional data, there are now 21,997 records. Try counting the records in the ratings table again.

- Use the command below, and note that the count is *not* 21,997.

SELECT COUNT(*) FROM ratings;

- After loading the new data, you need to refresh the metadata for this table, so the additional data can be accessed. (Refreshing is done automatically by Hue when you create a table, but not when you add data.) Enter and execute this command:

```
REFRESH ratings;
```

- Execute the count command again, and verify that 21,997 are included in the table.

Creating an External Table Using CREATE TABLE

The data for an employees table already exists in HDFS. In this section, you'll set up an external table so you can query this data. In the last section, you practiced creating a table using the Hue Table Browser; this time, use an Impala SQL statement. You may use either the Impala shell, or the Impala Query Editor in Hue.

15. Write and execute a CREATE TABLE statement to create an *external* table for the tab-delimited records in HDFS at $ADIR/analyst/dualcore/employees. The format is shown below:

Field Name	Field Type
emp_id	STRING
fname	STRING
lname	STRING
address	STRING
city	STRING
state	STRING
zipcode	STRING
job_title	STRING
email	STRING
active	STRING
salary	INT

16. Run the following query to verify that you have created the table correctly.

```
SELECT job_title, COUNT(*) AS num
FROM employees
GROUP BY job_title
ORDER BY num DESC LIMIT 3;
```

It should show that Sales Associate, Cashier, and Assistant Manager are the three most common job titles at Dualcore.

17. In a terminal window, execute the following command to import the suppliers table from MySQL as a new managed table in the analyst database:

```
$ sqoop import \
--connect jdbc:mysql://localhost/analyst_dualcore \
--username root --password cloudera \
--fields-terminated-by '\t' \
--warehouse-dir analist\dualcore \
--table suppliers
--hive-import
--create-hive-table
--hive-table analyst.suppliers
```

18. It is always a good idea to verify that data has been added as intended. Execute the following query to count the number of suppliers in Texas.

You may use either the Impala shell or the Hue Impala Query Editor.

```
INVALIDATE METADATA suppliers;
```

INVALIDATE METADATA tablename command from the analyst database, so Impala can find the new table.

then

```
SELECT COUNT(*) FROM suppliers WHERE state='TX';
```

The query should show that nine records match.

19. Use ALTER TABLE to rename the company column to name.

20. Use the DESCRIBE command on the suppliers table to verify the change.

21. Use ALTER TABLE to rename the entire table to vendors.

22. Although the ALTER TABLE command often requires that we make a corresponding change to the data in HDFS, renaming a table or column does not. You can verify this by running a query on the table using the new names, for example:

```
SELECT supp_id, name FROM vendors LIMIT 10;
```