Hands-On Exercise: Using Functions

Files and Data Used in This Exercise

Exercise directory	$ADIR/exercises/functions
Hive/Impala tables	orders, products

In this exercise, you will write queries using functions to analyze data in tables.

IMPORTANT: This exercise builds on previous ones.

Several analysis questions are described below, and you will need to run queries to answer them. You can use whichever tool you prefer—Hive or Impala—using whichever method you like best, including shell, script, or the Hue Query Editor.

There are several ways you could write these queries. An example solution for each is provided in the $ADIR/exercises/functions/solution/ directory.

Exploring the Dualcore Data

1. To begin, familiarize yourself with the orders table. You can look at the schema and get a sample of a few rows to see what the data looks like.
You'll also work with the products table. If you need to familiarize yourself with that table as well, do so now.

Busy Times

The order_date column in the orders table is a timestamp column that provides the time of the order, down to the second. You can use this to determine which are the busy times of day.

2. In the Impala documentation for date and time functions (https://docs.cloudera.com/documentation/enterprise/5-13-x/topics/impala_datetime_functions.html), find a function for extracting the hour from a timestamp type. Test the function by finding the distinct values for hours in the order_date column. (You should have only 0 to 23.)

3. Which are the three most active hours, and the three least active hours, for Dualcore orders?

Finding Profit

In the products table, the price column is what the customer pays for an item. The cost column gives the wholesale cost that Dualcore pays for the item. The *profit* on an item is the difference between what the customer pays and what Dualcore pays.

4. Find the five items that provide the largest profit. (They should all be servers.)

5. What items provide no profit, or actually cost Dualcore more than they charge for it? Use a filter to return only the rows for which this is the case. (Your query should return three rows.)

6. If either of the price or the cost column has a NULL value for a row, then that row would not be included in the previous queries. Do the following to check how complete this data is.

- First check either of the columns for NULL values using a null operator. Does the column have any NULL values?
- Now write a single query that will check both columns for NULL values. Are there any?

Averages

Answer the following about averages for different brands.

7. Write a query to find the average cost, average price, and average profit of all the items Dualcore carries. (The average profit is around $30.04.)

8. Modify the previous query to find the same averages for each brand *and* round the averages to the nearest penny (that is, the nearest integer). (Your query should return 47 rows.)

9. Modify the query to filter the results, so you only get those brands with an average price greater than $1000 (100000). Write down the largest average price for the next problem. (Your query should return only two rows.)

10. Find the items whose price is larger than the largest average price that you found for the previous problem. (Your query should return 20 rows.)

11. Which products are similar to those expensive ones but are also significantly cheaper?

12. How much does the price-to-profit ratio vary among items? If you get an odd result, check your query and think about what might give that result.

13. How much does it change within a brand?

14. What items have the largest price-to-profit ratio? (That is, they are relatively expensive for the amount of money they bring in to the company.)

15. Compare price-to-profit ratios for items similar to those. Purely from a price-to-profit perspective, would you recommend Dualcore stop carrying any particular items from a particular brand?

16. What other things might Dualcore need to consider before agreeing to stop carrying those items?