Apache Hadoop –

A course for undergraduates

Homework Labs, Lecture 13

Copyright © 2010-2014 Cloudera, Inc. All rights reserved.
Not to be reproduced without prior written consent.

Lab: Interactive Analysis with Impala

In this lab you will examine abandoned cart data using the tables created in the previous lab. You will use Impala to quickly determine how much lost revenue these abandoned carts represent and use several "what if" scenarios to determine whether Dualcore should offer free shipping to encourage customers to complete their purchases.

IMPORTANT: Since this lab builds on the previous one, it is important that you successfully complete the previous lab before starting this lab.

Step #1: Start the Impala Shell and Refresh the Cache

1. Issue the following commands to start Impala, then change to the directory for this lab:

```
$ sudo service impala-server start
$ sudo service impala-state-store start
$ cd $ADIR/exercises/interactive
```

2. First, start the Impala shell:

```
$ impala-shell
```

3. Since you created tables and modified data in Hive, Impala's cache of the metastore is outdated. You must refresh it before continuing by entering the following command in the Impala shell:

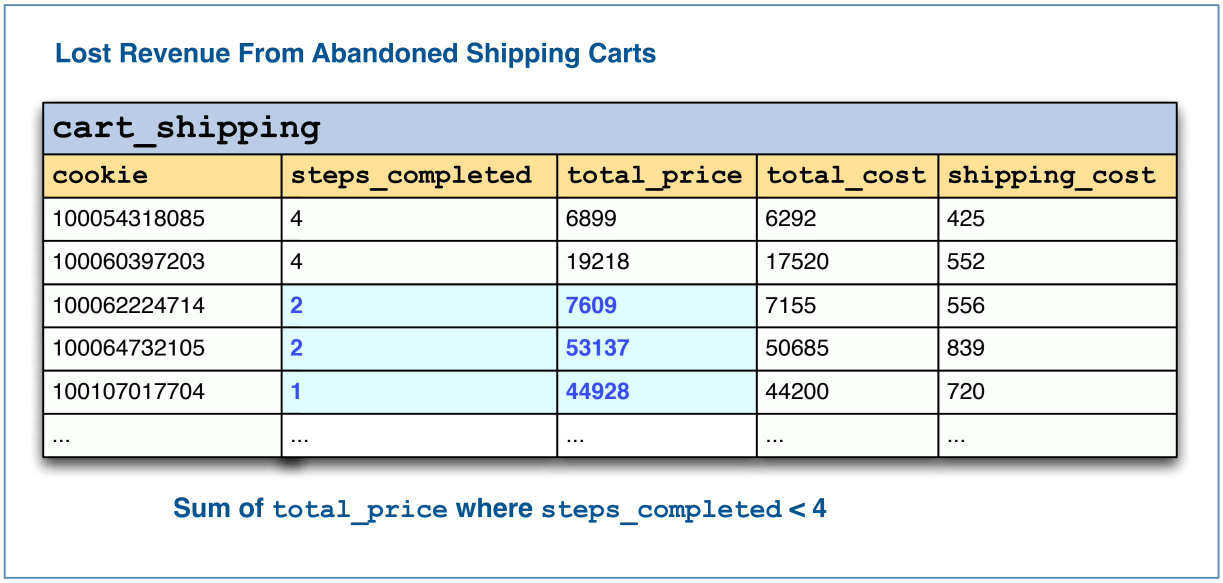
```
REFRESH;
```

Copyright © 2010-2014 Cloudera, Inc. All rights reserved.
Not to be reproduced without prior written consent.

Step #2: Calculate Lost Revenue

1. First, you'll calculate how much revenue the abandoned carts represent. Remember, there are four steps in the checkout process, so only records in the `cart_shipping` table with a `steps_completed` value of four represent a completed purchase:

```
SELECT SUM(total_price) AS lost_revenue
    FROM cart_shipping
    WHERE steps_completed < 4;
```



Lost Revenue From Abandoned Shipping Carts

cart_shipping

cookie	steps_completed	total_price	total_cost	shipping_cost
100054318085	4	6899	6292	425
100060397203	4	19218	17520	552
100062224714	2	7609	7155	556
100064732105	2	53137	50685	839
100107017704	1	44928	44200	720
...	...	...	...	...

Sum of `total_price` where `steps_completed` < 4

You should see that abandoned carts mean that Dualcore is potentially losing out on more than $2 million in revenue! Clearly it's worth the effort to do further analysis.

Note: The `total_price`, `total_cost`, and `shipping_cost` columns in the `cart_shipping` table contain the number of cents as integers. Be sure to divide results containing monetary amounts by 100 to get dollars and cents.

Copyright © 2010-2014 Cloudera, Inc. All rights reserved.
Not to be reproduced without prior written consent.

2. The number returned by the previous query is revenue, but what counts is profit. We calculate gross profit by subtracting the cost from the price. Write and execute a query similar to the one above, but which reports the total lost profit from abandoned carts.

3. How does this compare to the amount of profit Dualcore receives from customers who do complete the checkout process? Modify your previous query to consider only those records where `steps_completed = 4`, and then execute it in the Impala shell.

4. The previous two queries show the *total* profit for abandoned and completed orders, but these aren't directly comparable because there were different numbers of each. It might be the case that one is much more profitable than the other on a per-order basis. Write and execute a query that will calculate the *average* profit based on the number of steps completed during the checkout process.

Step #3: Calculate Cost/Profit for a Free Shipping Offer

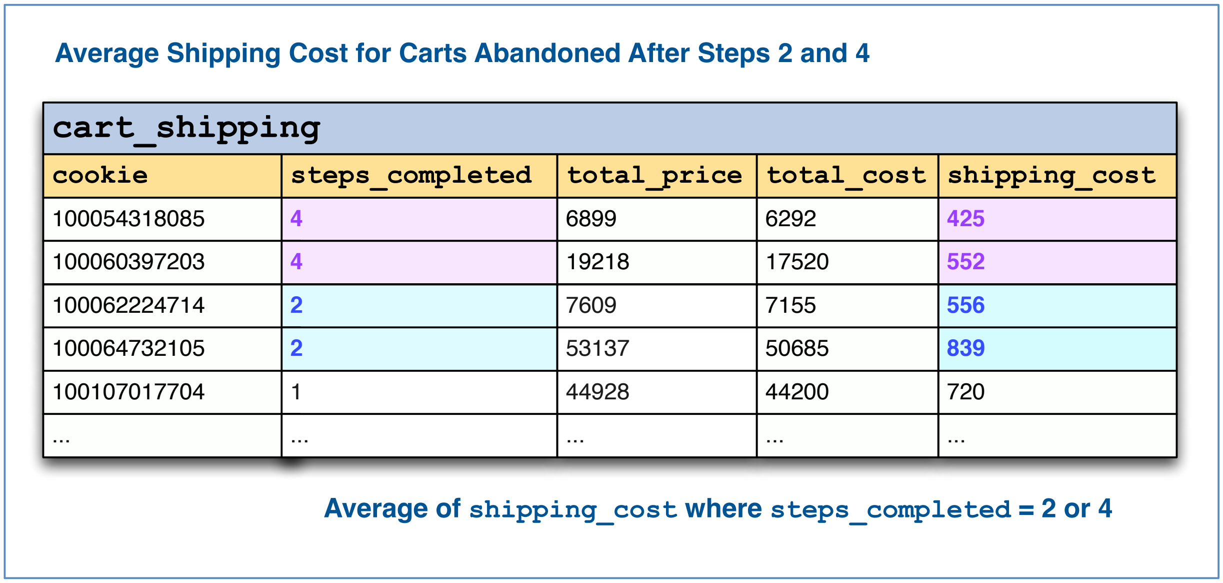
You have observed that most carts – and the most *profitable* carts – are abandoned at the point where the shipping cost is displayed to the customer. You will now run some queries to determine whether offering free shipping, on at least some orders, would actually bring in more revenue assuming this offer prompted more customers to finish the checkout process.

1. Run the following query to compare the average shipping cost for orders abandoned after the second step versus completed orders:

```
SELECT steps_completed, AVG(shipping_cost) AS ship_cost
   FROM cart_shipping
   WHERE steps_completed = 2 OR steps_completed = 4
```

Copyright © 2010-2014 Cloudera, Inc. All rights reserved.
Not to be reproduced without prior written consent.

```
    GROUP BY steps_completed;
```



Average Shipping Cost for Carts Abandoned After Steps 2 and 4

cart_shipping

cookie	steps_completed	total_price	total_cost	shipping_cost
100054318085	4	6899	6292	425
100060397203	4	19218	17520	552
100062224714	2	7609	7155	556
100064732105	2	53137	50685	839
100107017704	1	44928	44200	720
...	...	...	...	...

Average of shipping_cost where steps_completed = 2 or 4

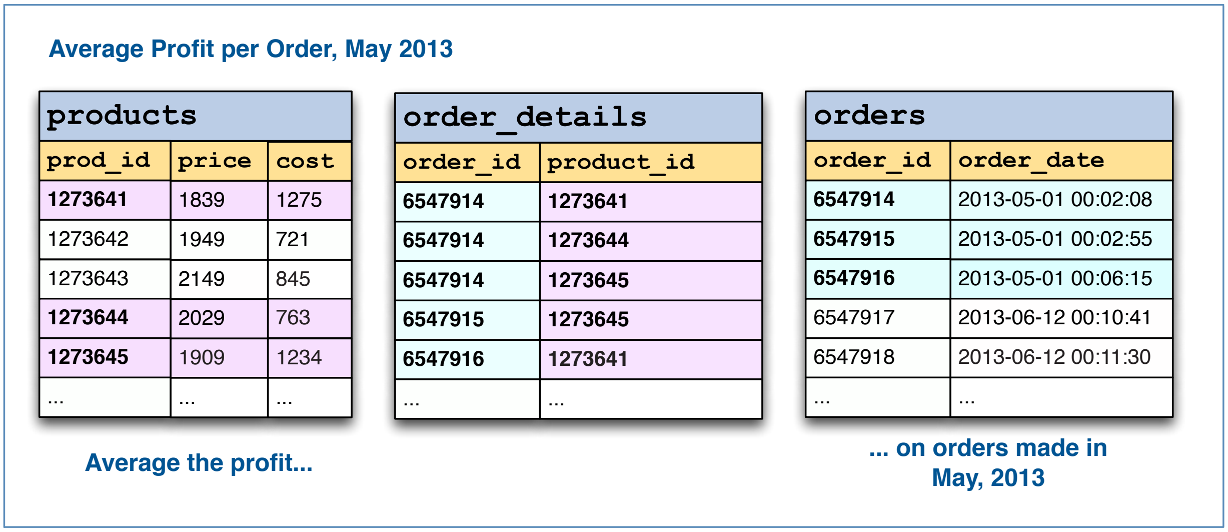
- You will see that the shipping cost of abandoned orders was almost 10% higher than for completed purchases. Offering free shipping, at least for some orders, might actually bring in more money than passing on the cost and risking abandoned orders.

2. Run the following query to determine the average profit per order over the entire month for the data you are analyzing in the log file. This will help you to determine whether Dualcore could absorb the cost of offering free shipping:

```
SELECT AVG(price - cost) AS profit
  FROM products p
  JOIN order_details d
    ON (d.prod_id = p.prod_id)
  JOIN orders o
    ON (d.order_id = o.order_id)
 WHERE YEAR(order_date) = 2013
```

Copyright © 2010-2014 Cloudera, Inc. All rights reserved.
Not to be reproduced without prior written consent.

```
      AND MONTH(order_date) = 05;
```



Average Profit per Order, May 2013

products

prod_id	price	cost
1273641	1839	1275
1273642	1949	721
1273643	2149	845
1273644	2029	763
1273645	1909	1234
...	...	...

order_details

order_id	product_id
6547914	**1273641**
6547914	**1273644**
6547914	**1273645**
6547915	**1273645**
6547916	**1273641**
...	...

orders

order_id	order_date
6547914	2013-05-01 00:02:08
6547915	2013-05-01 00:02:55
6547916	2013-05-01 00:06:15
6547917	2013-06-12 00:10:41
6547918	2013-06-12 00:11:30
...	...

Average the profit...

... on orders made in May, 2013

- You should see that the average profit for all orders during May was $7.80. An earlier query you ran showed that the average shipping cost was $8.83 for completed orders and $9.66 for abandoned orders, so clearly Dualcore would lose money by offering free shipping on all orders. However, it might still be worthwhile to offer free shipping on orders over a certain amount.

Copyright © 2010-2014 Cloudera, Inc. All rights reserved.
Not to be reproduced without prior written consent.

3. Run the following query, which is a slightly revised version of the previous one, to determine whether offering free shipping only on orders of $10 or more would be a good idea:

```
SELECT AVG(price - cost) AS profit
  FROM products p
  JOIN order_details d
    ON (d.prod_id = p.prod_id)
  JOIN orders o
    ON (d.order_id = o.order_id)
  WHERE YEAR(order_date) = 2013
        AND MONTH(order_date) = 05
        AND PRICE >= 1000;
```

   - You should see that the average profit on orders of $10 or more was $9.09, so absorbing the cost of shipping would leave very little profit.

4. Repeat the previous query, modifying it slightly each time to find the average profit on orders of at least $50, $100, and $500.

   - You should see that there is a huge spike in the amount of profit for orders of $500 or more (Dualcore makes $111.05 on average for these orders).

5. How much does shipping cost on average for orders totaling $500 or more? Write and run a query to find out

6. Since Dualcore won't know in advance who will abandon their cart, they would have to absorb the $12.28 average cost on *all* orders of at least $500. Would the extra money they might bring in from abandoned carts offset the added cost of free shipping for customers who would have completed their purchases anyway? Run the following query to see the total profit on completed purchases:

```
SELECT SUM(total_price - total_cost) AS total_profit
   FROM cart_shipping
```

Copyright © 2010-2014 Cloudera, Inc. All rights reserved.
Not to be reproduced without prior written consent.

```
   WHERE total_price >= 50000
     AND steps_completed = 4;
```

- After running this query, you should see that the total profit for completed orders is $107,582.97.

7. Now, run the following query to find the potential profit, after subtracting shipping costs, if all customers completed the checkout process:

```
SELECT gross_profit - total_shipping_cost AS
potential_profit
   FROM (SELECT
           SUM(total_price - total_cost) AS
gross_profit,
           SUM(shipping_cost) AS total_shipping_cost
         FROM cart_shipping
         WHERE total_price >= 50000) large_orders;
```

Since the result of $120,355.26 is greater than the current profit of $107,582.97 Dualcore currently earns from completed orders, it appears that they could earn nearly $13,000 more by offering free shipping for all orders of at least $500.

Congratulations! Your hard work analyzing a variety of data with Hadoop's tools has helped make Dualcore more profitable than ever.

This is the end of the lab.

Copyright © 2010-2014 Cloudera, Inc. All rights reserved.
Not to be reproduced without prior written consent.