Hortonworks

Apache Hive 0.13 Performance Benchmarks
Query Times in Hive 0.13 v. Hive 0.10

Results from the Stinger Initiative

June 2014

The Stinger Initiative

- **Apache Hive is the de facto standard for SQL-in-Hadoop**
- **The Stinger Initiative drove improved SQL semantics & performance**
- **Stinger Highlights:**
  - **13 months**
  - **145 developers**
  - **44 companies**
  - **3 Hive releases (0.11, 0.12 & 0.13)**
  - **392K lines of new Java code**

Benchmark Overview: Hive 10 v Hive 13

- **The TPC Benchmark™DS** is a decision support benchmark that models queries and data maintenance. It evaluates decision support systems that examine large volumes of data to answer real-world business questions.
- **Test:** 50 SQL queries on Hive 0.10 (RCFile) and Hive 0.13 (ORCFile) at scale of 30TB
- **Results: More than 100x speed up for 6 of 50 queries, with average acceleration of 52x across all queries**
- **Test Environment**
  - **Driven by the Hive Testbench**: https://github.com/cartershanklin/hive-testbench
  - **Nodes**: 20 nodes, 256 GB per node
  - **Drives:** 6x 4TB WDC WD4000FYYZ-0 drives per node
  - **Interconnect:** 10GB
  - **Processors**: 2x Intel(R) Xeon(R) CPU E5-2640 v2 @ 2.00GHz for total of 16 CPU cores per machine

Results for Interactive Queries

Queries #3, 7, 12, 15, 18,19, 26, 27, 42, 43, 52, 55, 82, 84, 91 & 96

Results for Interactive Queries

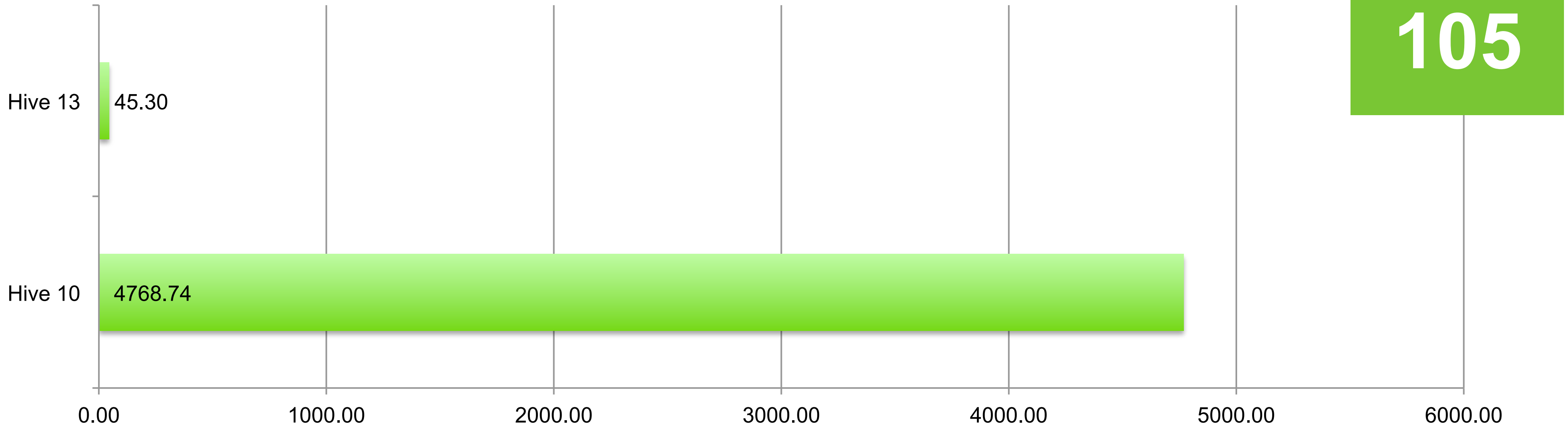
Interactive Queries: Star schema joins over single fact tables, which may involve advances SQL features such as windowing functions or rollups

Query #	Query Description	Hive 13	Hive 10	Change
55	For a given year, month and store manager calculate the total store sales of any combination all brands.	45.30	4,768.74	105X
27	For all items sold in stores located in six states during a given year, find summary statistics	99.05	9,988.27	101X
52	Report the total of extended sales price for all items of a specific brand in a specific year and month.	47.64	4,783.76	100X
42	For each item and a specific year and month calculate the sum of the extended sales price of store transactions.	51.16	4,681.42	92X
7	Compute the averages for promotional items sold in stores where promotion is not offered by mail or a special event.	111.89	9,795.81	88X
15	Report the total catalog sales for customers in selected geographical regions or who made large purchases.	129.43	9,299.68	72X
26	Computes averages for promotional items sold through the catalog channel…	79.14	4,718.24	60X
19	Select the top 10 revenue generating products bought by out of zip code customers for a given year…	106.94	5,668.60	53X
3	Report the total extended sales price per item brand of a specific manufacturer for all sales in a specific month.	127.99	5,433.11	42X
96	Count of sales from a named store to customers with a given number of dependents...	200.53	7,888.14	39X
91	Display total returns of catalog sales by call center and manager in a particular month…	51.61	1,460.19	28X
43	Report the sum of all sales from Sunday to Saturday for stores in a given data range by stores.	305.50	6,153.43	20X
82	Find customers who tend to spend more money (net-paid) on-line than in stores.	753.00	9,302.69	12X
84	List all customers living in a specified city, with an income between 2 values.	494.38	2,654.84	5X
12	Compute the revenue ratios across item classes.	163.42	NA	∞
18	Compute catalog sales in a given year by customers meeting certain characteristics.	162.55	NA	∞

All times in seconds

Query 55

For a given year, month and store manager calculate the total store sales of any combination all brands.



X Factor
105

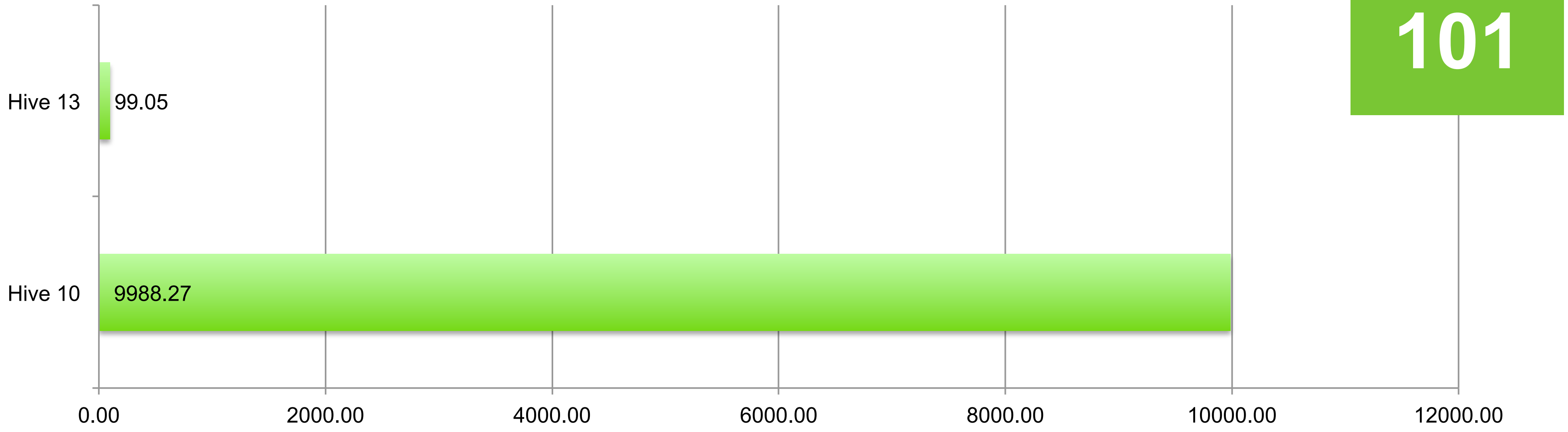
All Values in Seconds

Query 27

For all items sold in stores located in six states during a given year, find the average quantity, average list price, average list sales price, average coupon amount for a given gender, marital status, education and customer demographic.

X Factor

101

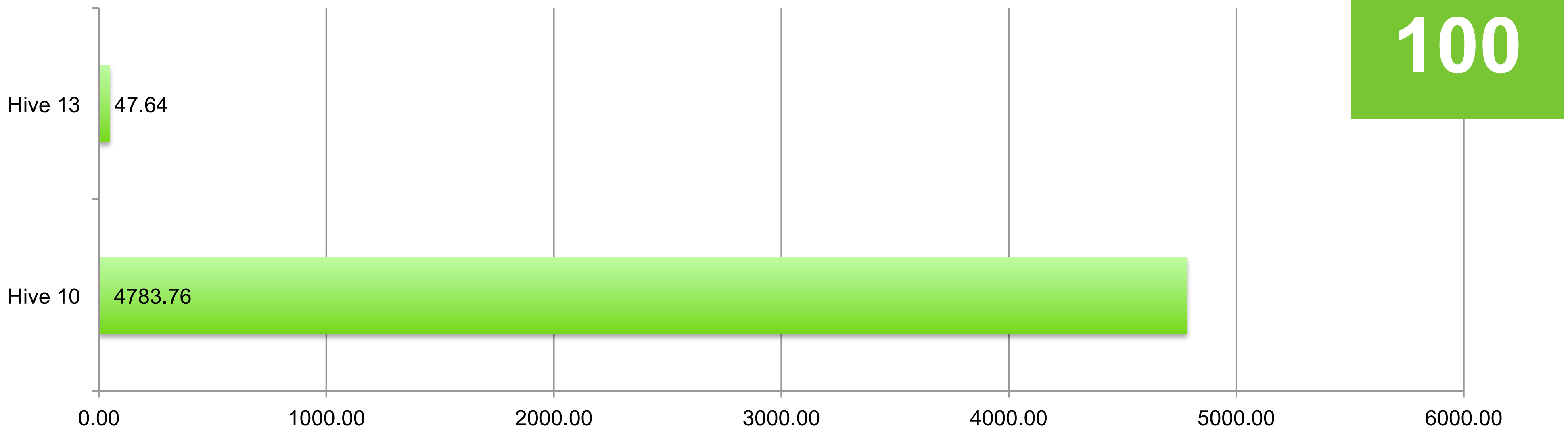


All Values in Seconds

Hortonworks

Query 52

Report the total of extended sales price for all items of a specific brand in a specific year and month.

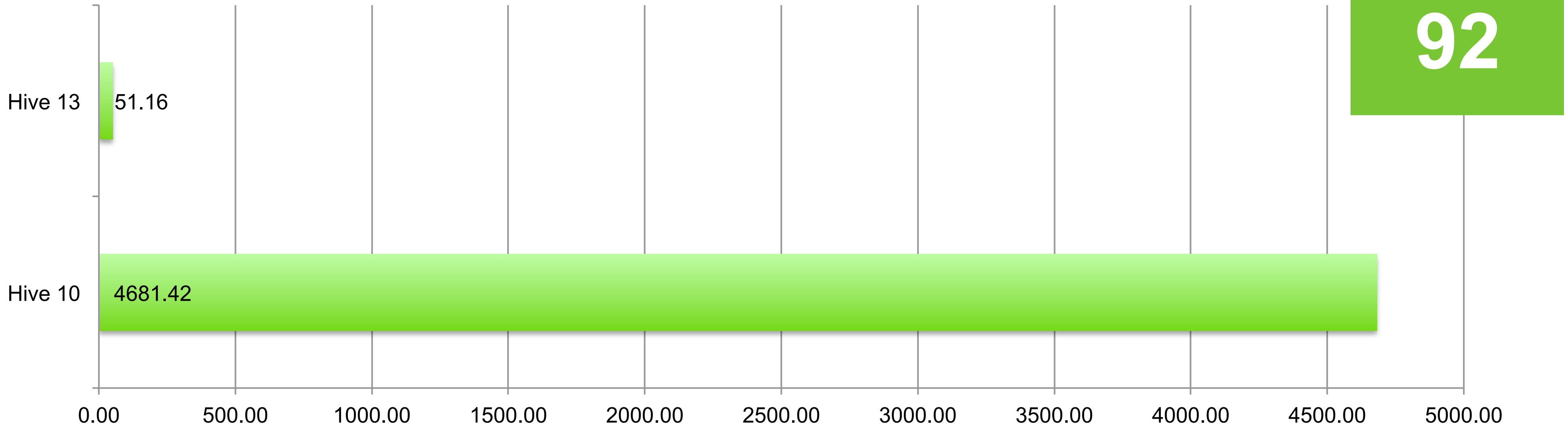


X Factor
100

All Values in Seconds

Query 42

For each item and a specific year and month calculate the sum of the extended sales price of store transactions.



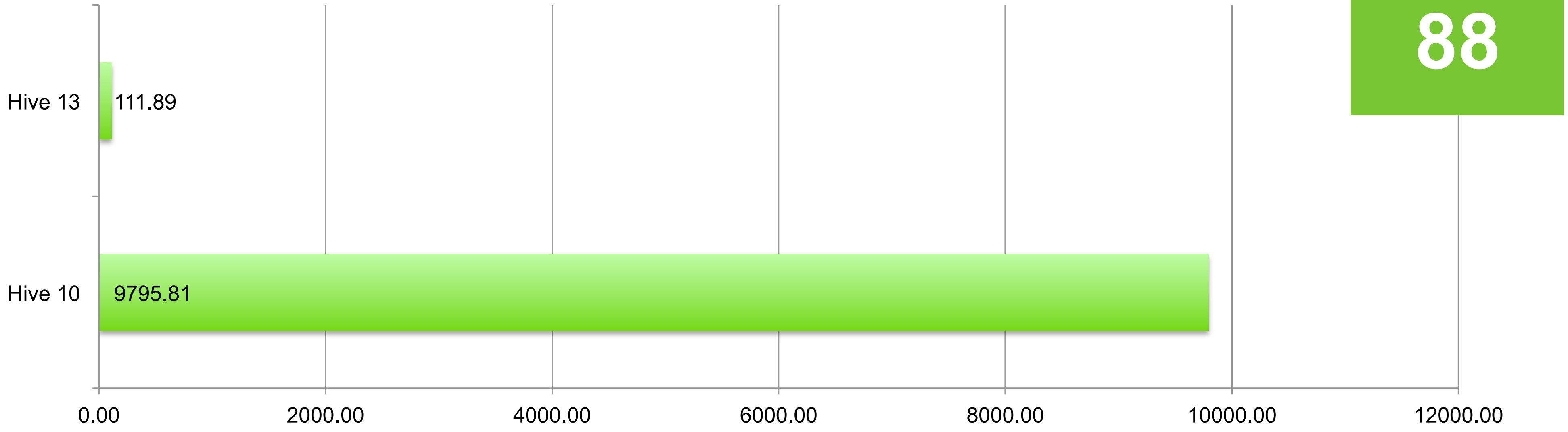
X Factor

92

All Values in Seconds

Query 7

Compute the average quantity, list price, discount, and sales price for promotional items sold in stores where the promotion is not offered by mail or a special event. Restrict the results to a specific gender, marital and educational status.

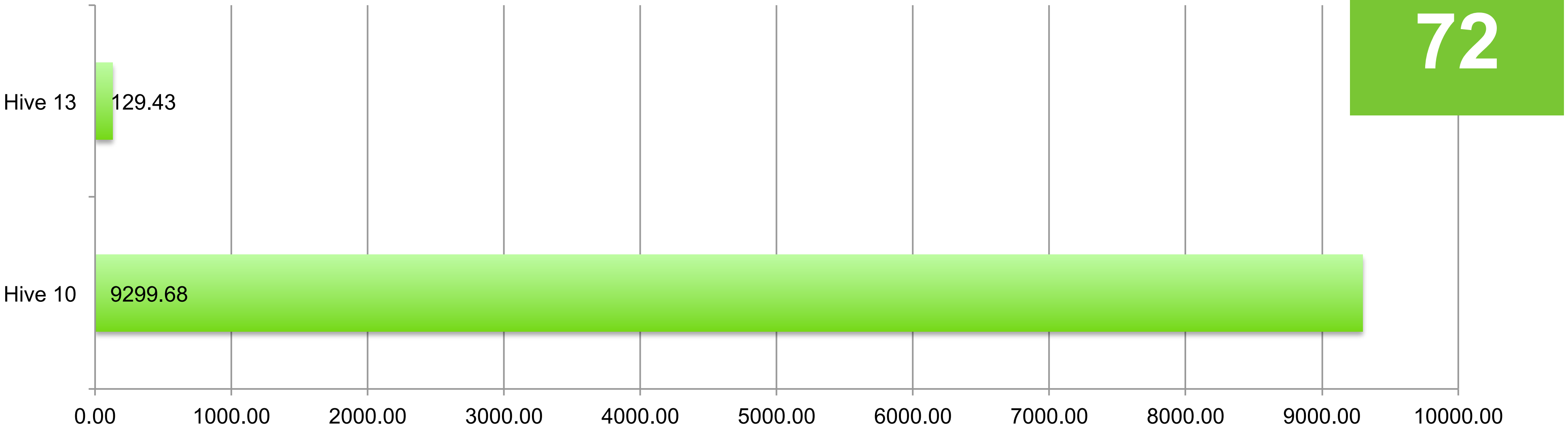


X Factor
88

All Values in Seconds

Query 15

Report the total catalog sales for customers in selected geographical regions or who made large purchases for a given year and quarter.

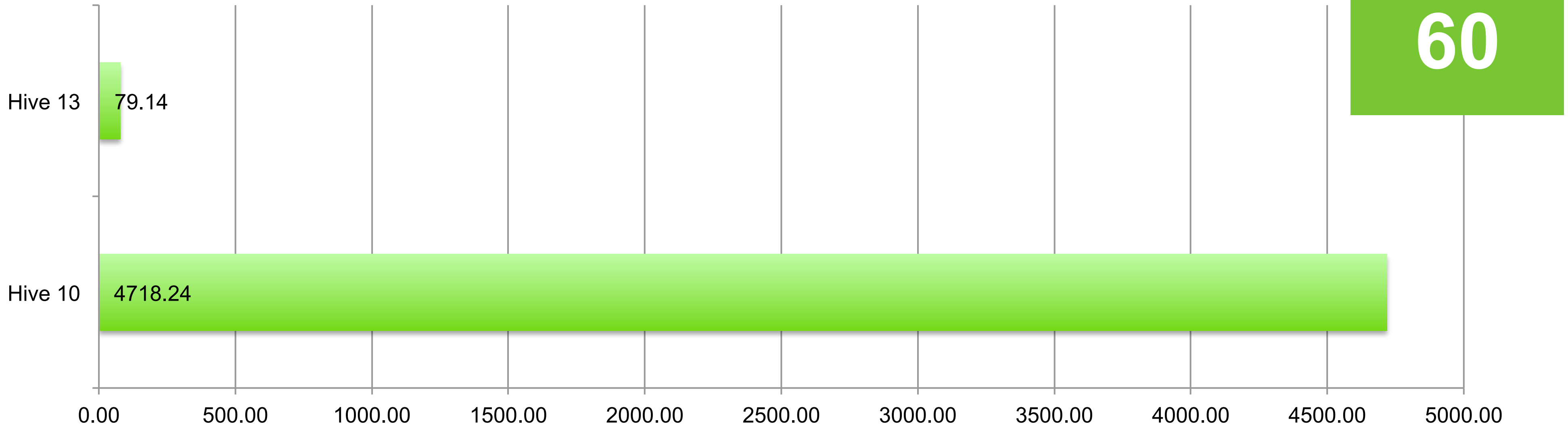


X Factor
72

All Values in Seconds

Query 26

Computes the average quantity, list price, discount, sales price for promotional items sold through the catalog channel where the promotion was not offered by mail or in an event for given gender, marital status and educational status.



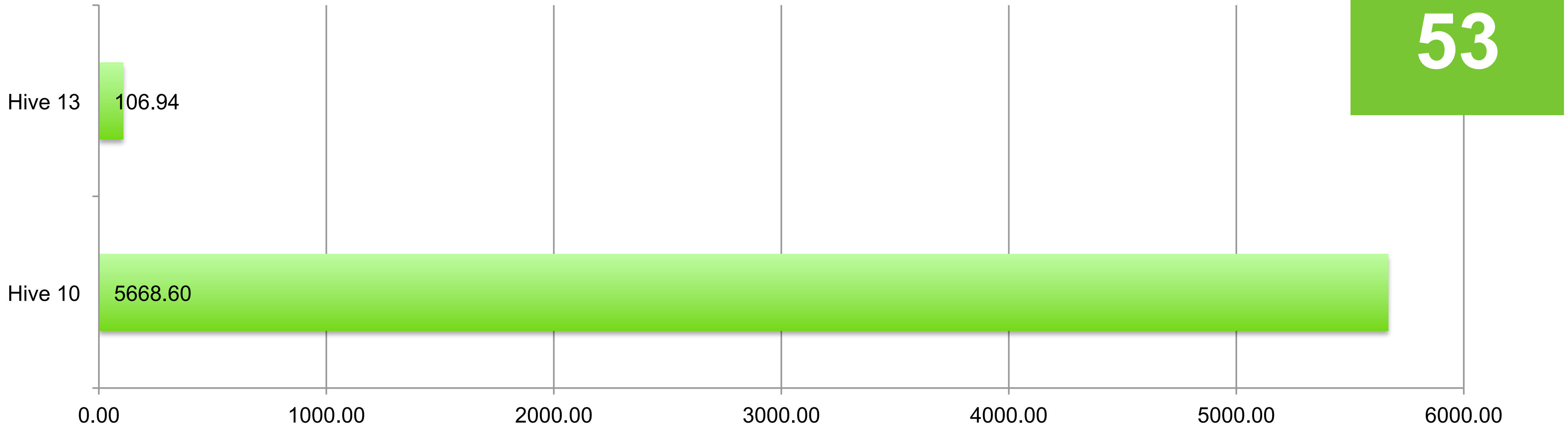
X Factor

60

All Values in Seconds

Query 19

Select the top 10 revenue generating products bought by out of zip code customers for a given year, month and manager.



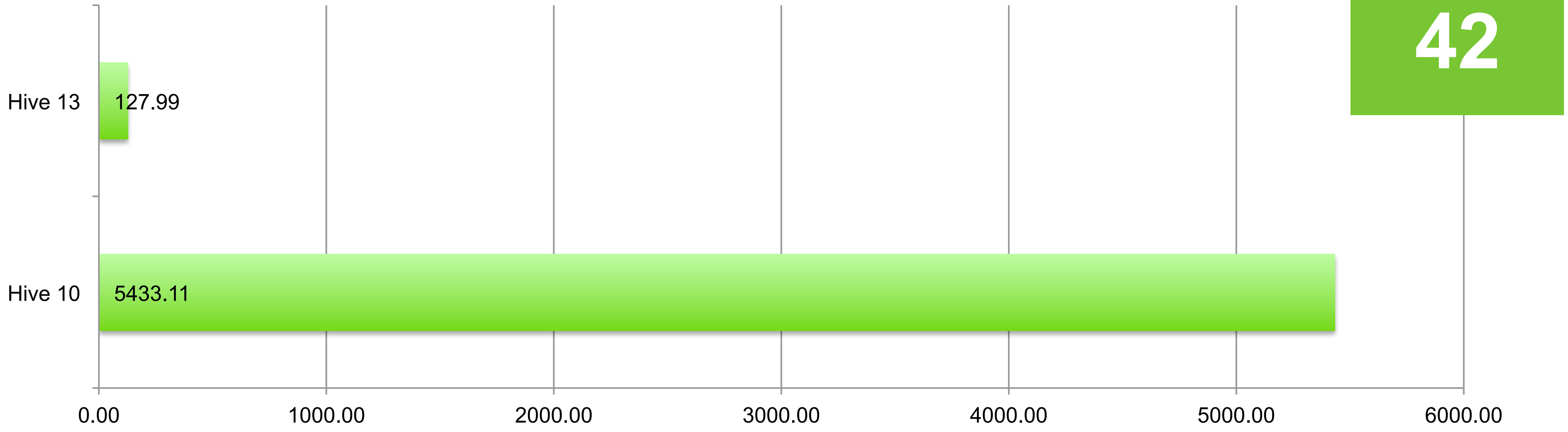
X Factor

53

All Values in Seconds

Query 3

Report the total extended sales price per item brand of a specific manufacturer for all sales in a specific month of the year.



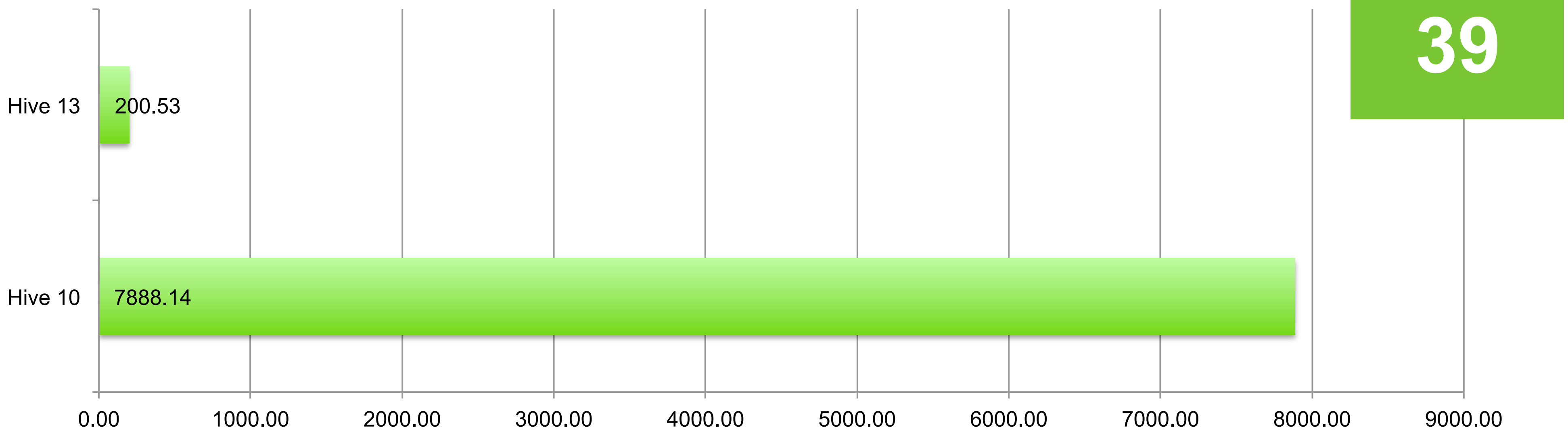
X Factor

42

All Values in Seconds

Query 96

Compute a count of sales from a named store to customers with a given number of dependents made in a specified half hour period of the day.



X Factor
39

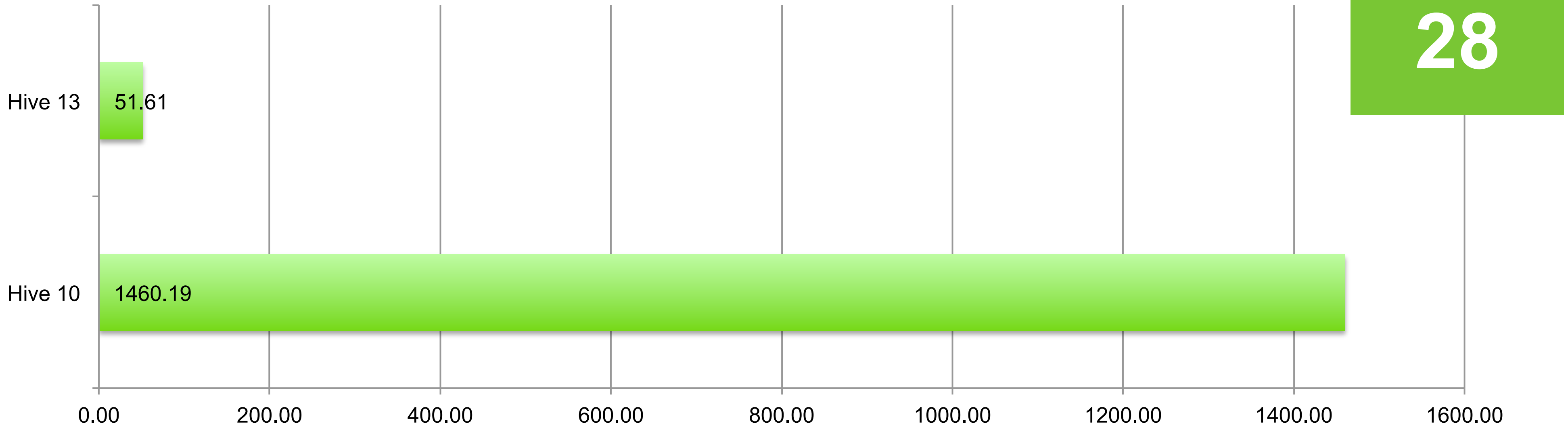
All Values in Seconds

Query 91

Display total returns of catalog sales by call center and manager in a particular month for male customers of unknown education or female customers with advanced degrees with a specified buy potential and from a particular time zone.

X Factor

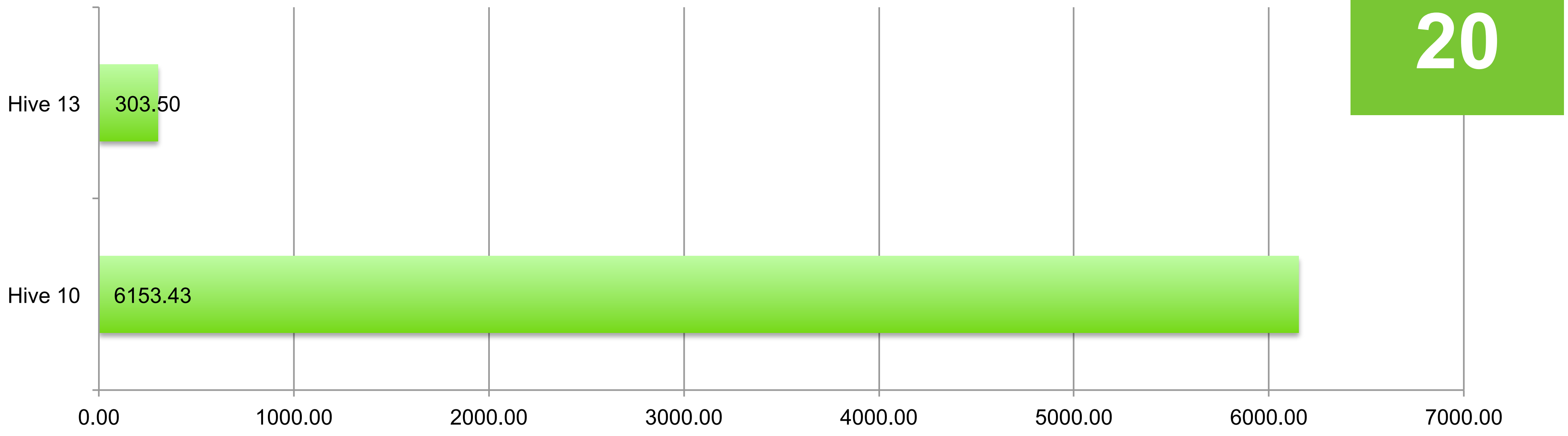
28



All Values in Seconds

Query 43

Report the sum of all sales from Sunday to Saturday for stores in a given data range by stores.



X Factor

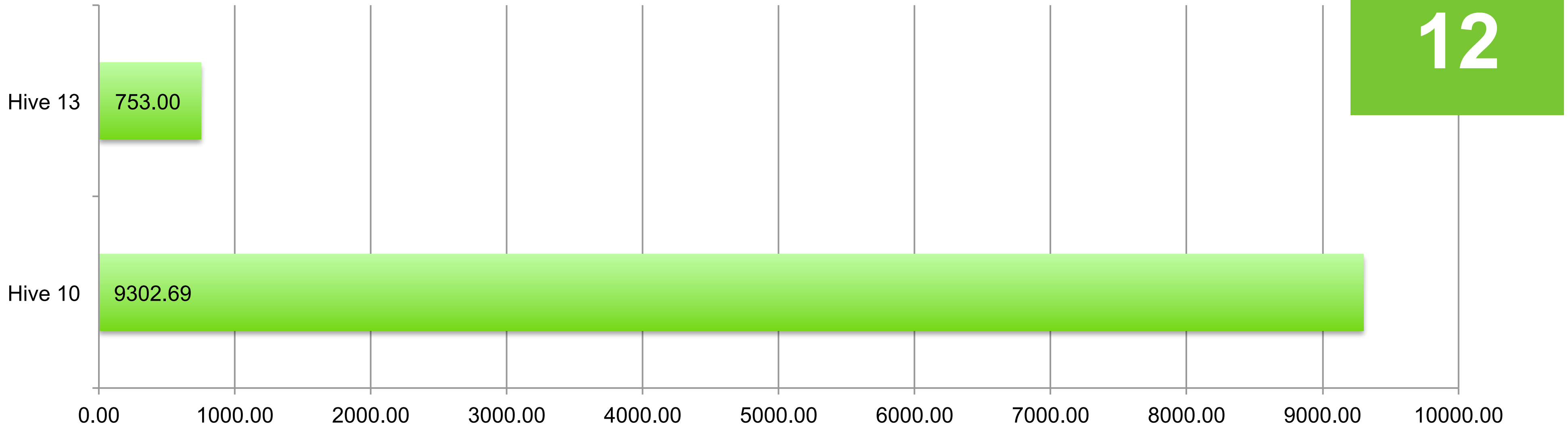
20

All Values in Seconds

Hortonworks

Query 82

Find customers who tend to spend more money (net-paid) on-line than in stores.



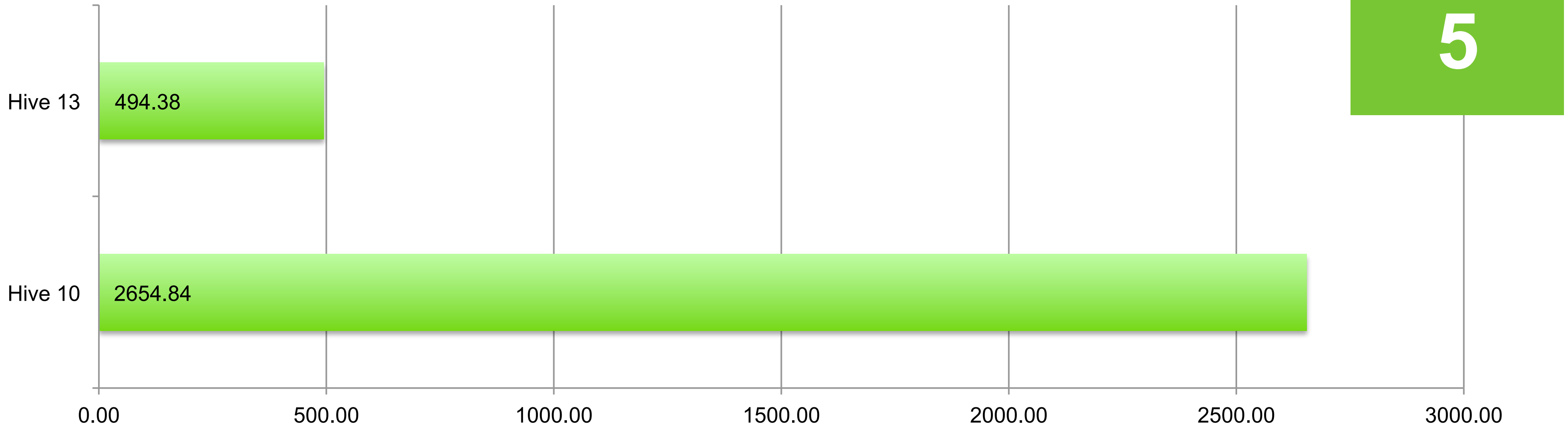
X Factor

12

All Values in Seconds

Query 84

List all customers living in a specified city, with an income between 2 values.



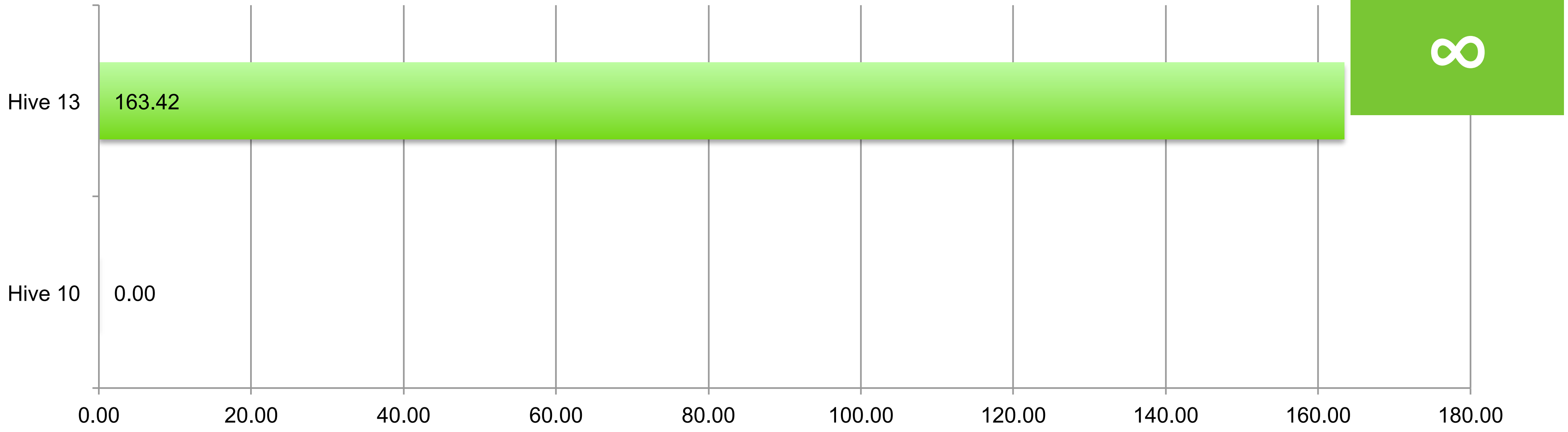
X Factor

5

All Values in Seconds

Query 12

Compute the revenue ratios across item classes: For each item in a list of given categories, during a 30 day time period, sold through the web channel compute the ratio of sales of that item to the sum of all of the sales in that item's class.



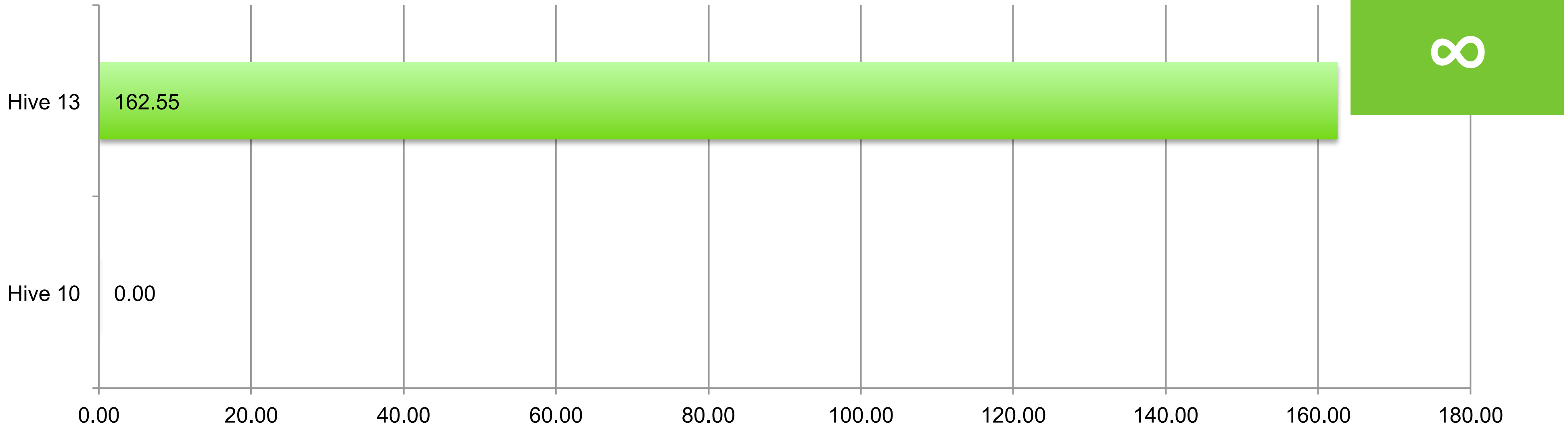
X Factor

∞

All Values in Seconds

Query 18

Compute, for each county, the average quantity, list price, coupon amount, sales price, net profit, age, and number of dependents for all items purchased through catalog sales in a given year by customers who were born in a given list of six months and living in a given list of seven states and who also belong to a given gender and education demographic.



X Factor

∞

All Values in Seconds

Results for Deep Reporting Queries

Queries #13, 17, 20, 21, 25, 28, 32, 40, 45, 46, 48, 49, 50, 58, 66, 68, 76, 79, 85, 87, 88, 89, 90, 92, 93, 94, 95 & 97

Results for Deep Reporting Queries

Deep Reporting: Complex queries involving multiple fact tables or large intermediate datasets

Query #	Query Description	Hive 13	Hive 10	Change
58	Retrieve the items generating the highest revenue…	217.68	34,620.94	159X
40	Compute the impact of an item price change on the sales by computing the total sales for items in a 30 day period…	231.04	19,434.19	84X
68	Compute the per customer extended sales price, extended list price and extended tax for "out of town" shoppers…	85.51	7,091.34	83X
66	Compute web and catalog sales and profits by warehouse	619.73	40,677.91	66X
95	Produce a count of web sales and total shipping cost and net profit in a given 60 day period…	334.37	20,473.84	61X
93	For a given merchandise return reason, report on customers' total cost of purchases minus the cost of returned items.	3,670.04	200,501.26	55X
21	For all items whose price was changed on a given date, compute the percentage change in inventory…	29.00	1,393.71	48X
46	Compute the per-customer coupon amount and net profit of all "out of town" customers buying from stores…	171.84	8,236.25	48X
88	How many items do we sell between pacific times of a day in certain stores to a certain type of customer?	1,767.31	72,721.59	41X
32	Compute the total discounted amount for a particular manufacturer in a particular 90 day period for catalog sales…	151.89	6,103.18	39X
17	Analyze, for each state, all items that were sold in stores in a particular quarter and returned in the next three quarters…	300.94	11,578.61	38X
94	Produce a count of web sales and total shipping cost and net profit in a given 60 day period…	181.77	5,859.67	32X
79	Compute the per customer coupon amount and net profit of Monday shoppers	272.71	8,568.70	31X
76	Computes the average quantity, list price, discount, sales price for promotional items sold through the web channel…	257.89	7,346.07	28X
92	Compute the total discount on web sales of items from a given manufacturer over a particular 90 day period…	860.82	9,768.99	11X
97	Generate counts of promotional sales and total sales, and their ratio from the web channel…	1,178.89	10,802.96	9X

All times in seconds

Results for Deep Reporting Queries (cont)

Deep Reporting: Complex queries involving multiple fact tables or large intermediate datasets

Query #	Query Description	Hive 13	Hive 10	Change
87	Count how many customers have ordered on the same day items on the web and the catalog…	1,672.35	12,308.54	7X
13	Calculate the average sales quantity, average sales price, average wholesale cost, total wholesale cost for store sales…	8,528.86	46,260.90	5X
50	For each store count the number of items in a specified month that were returned after 30, 60, 90, 120 and >120 days…	3,125.99	8,063.41	3X
20	Compute the total revenue and the ratio of total revenue to revenue by item class for specified item categories…	77.59	NA	∞
25	Get all items that were sold in stores in a particular month and year and returned in the next three quarters…	318.87	NA	∞
28	Calculate the average list price, number of non empty (null) list prices and number of distinct list prices…	2,227.67	NA	∞
45	Report the total web sales for customers in specific zip codes, cities, counties or states, or specific items…	112.09	NA	∞
48	Calculate the total sales by different types of customers…	1,813.69	NA	∞
49	Report the top 10 worst return ratios (sales to returns) of all items for each channel by quantity and currency…	559.75	NA	∞
85	For all web return reasons calculate the average sales, average refunded cash and average return fee…	500.67	NA	∞
89	All month and combination of item categories, classes and brands that have had monthly sales larger than 0.1 percent…	164.49	NA	∞
90	The ratio between the number of items sold over the internet in the morning versus items sold in the evening…	131.18	NA	∞

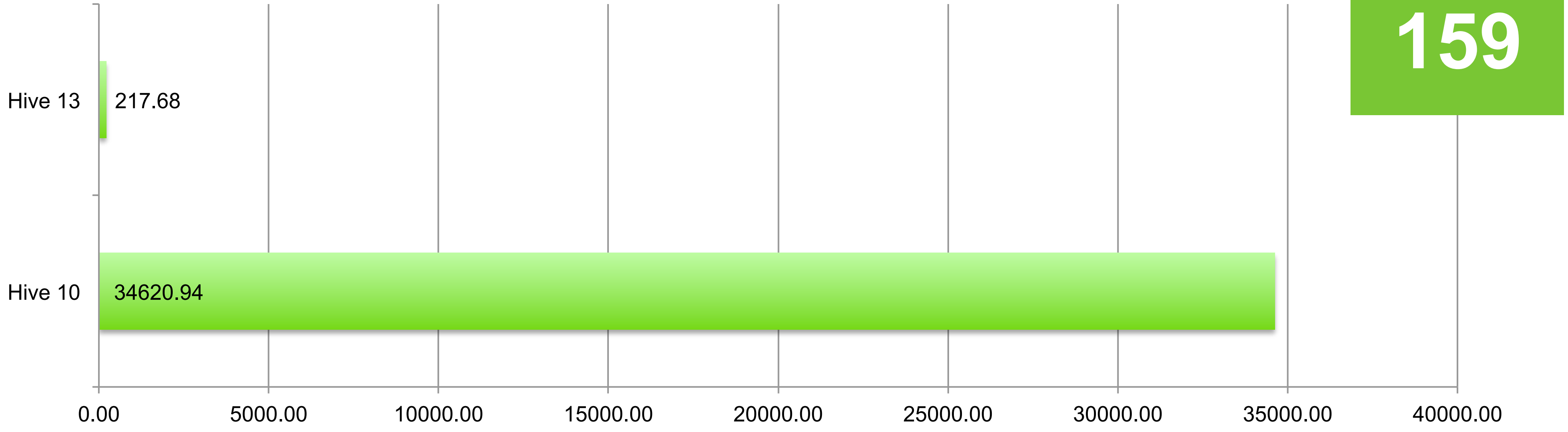
All times in seconds

Query 58

Retrieve the items generating the highest revenue and which had a revenue that was approximately equivalent across all of store, catalog and web within the week ending a given date.

X Factor

159

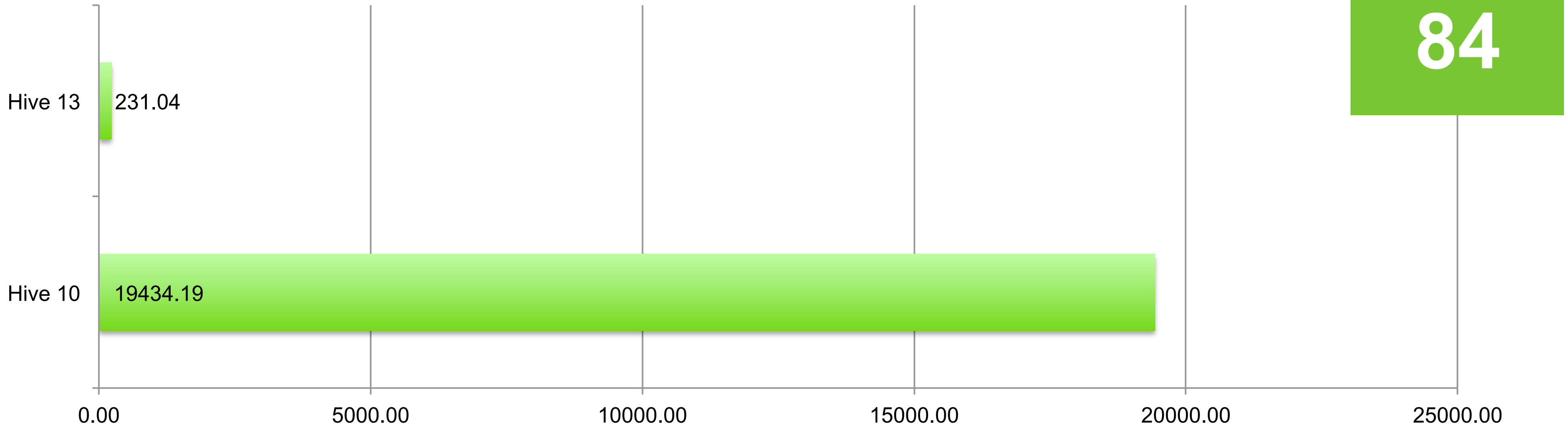


All Values in Seconds

Hortonworks

Query 40

Compute the impact of an item price change on the sales by computing the total sales for items in a 30 day period before and after the price change. Group the items by location of warehouse where they were delivered from.



X Factor

84

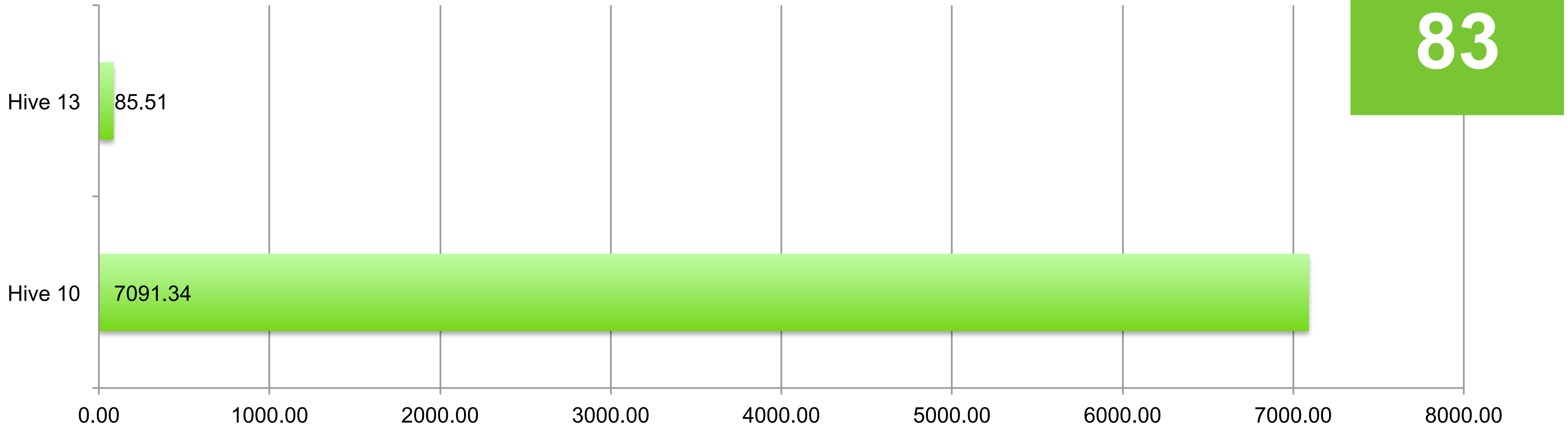
All Values in Seconds

Query 68

Compute the per customer extended sales price, extended list price and extended tax for "out of town" shoppers buying from stores located in two cities in the first two days of each month of three consecutive years. Only consider customers with specific dependent and vehicle counts.

X Factor

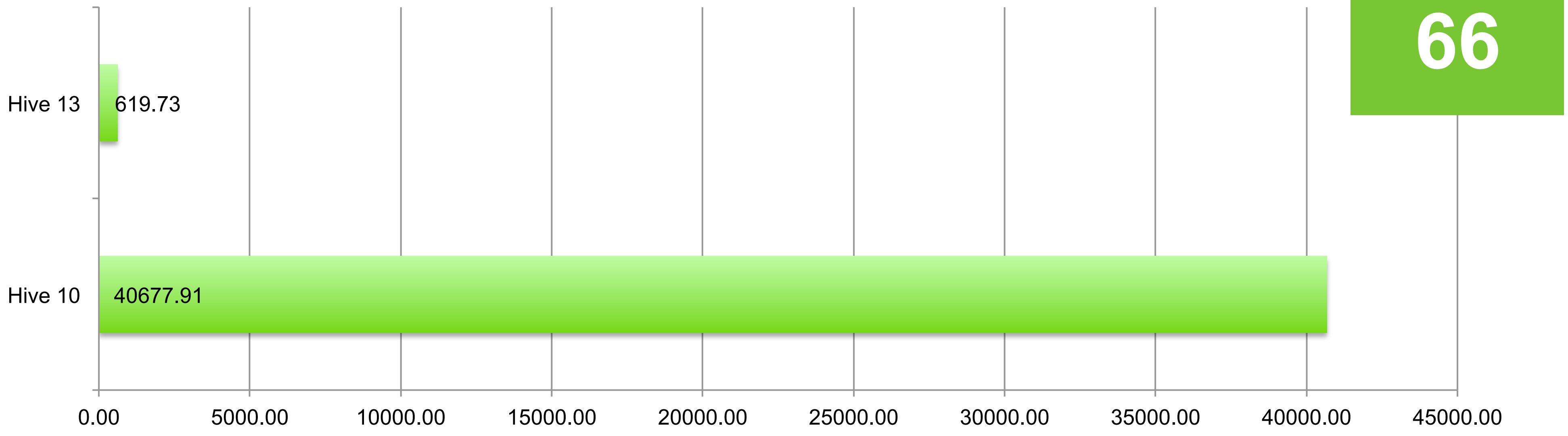
83



All Values in Seconds

Query 66

Compute web and catalog sales and profits by warehouse. Report results by month for a given year during a given 8-hour period.



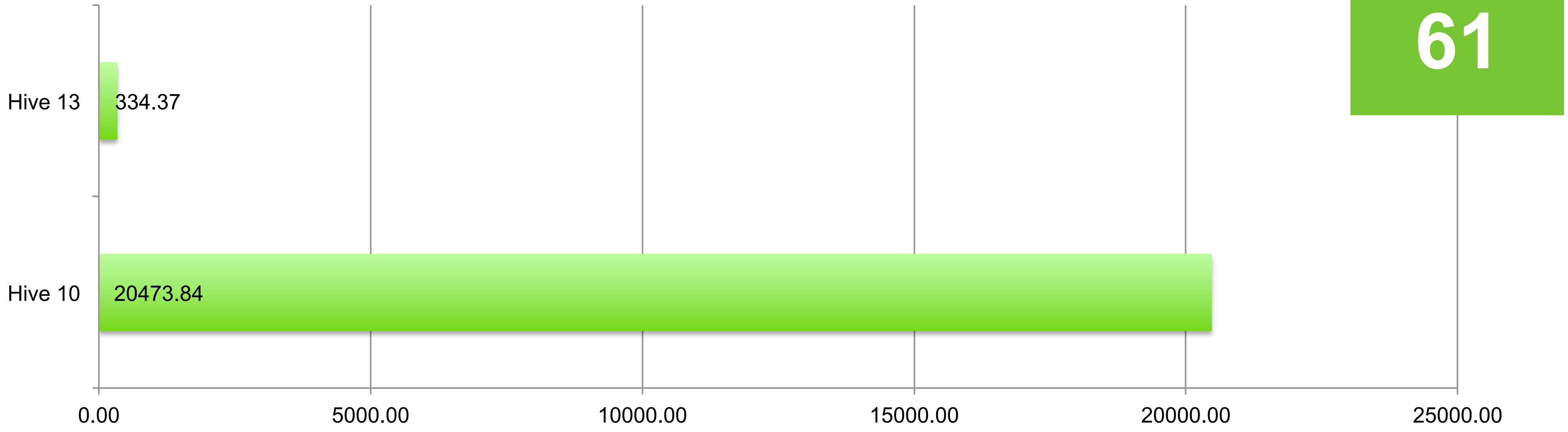
X Factor

66

All Values in Seconds

Query 95

Produce a count of web sales and total shipping cost and net profit in a given 60 day period to customers in a given state from a named web site for returned orders shipped from more than one warehouse.



X Factor

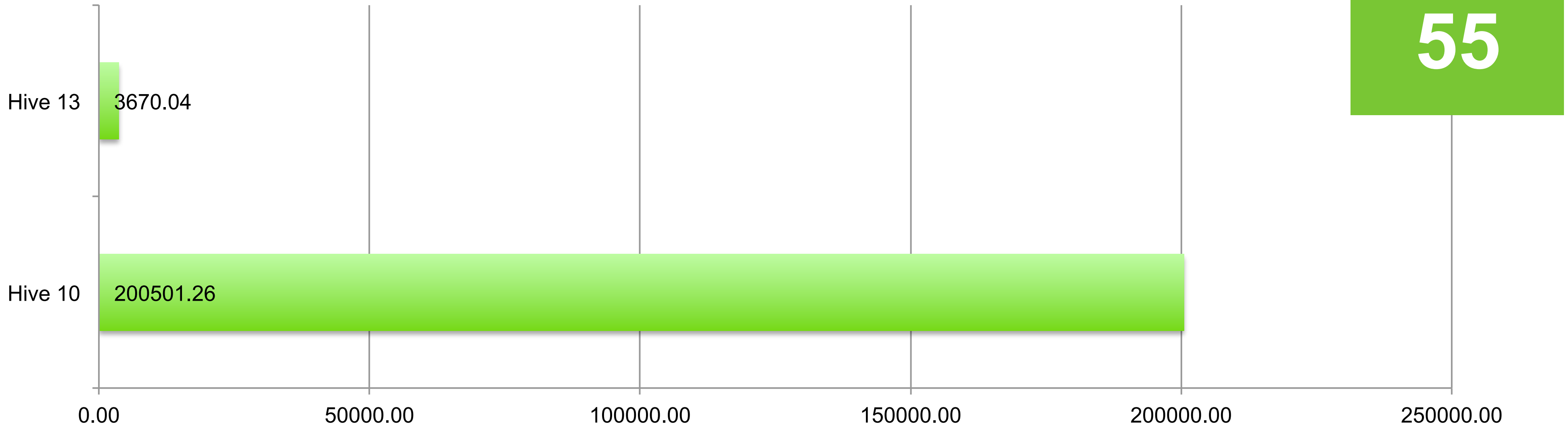
61

All Values in Seconds

Hortonworks

Query 93

For a given merchandise return reason, report on customers’ total cost of purchases minus the cost of returned items. Limit the output to the 100 customers with the highest value of total purchases.



X Factor

55

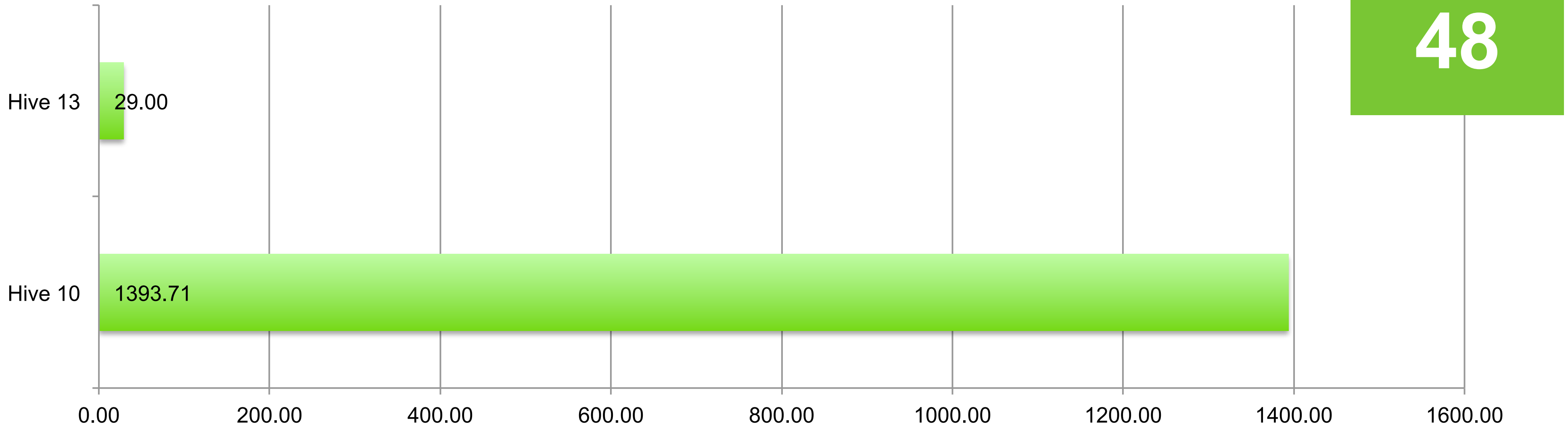
All Values in Seconds

Query 21

For all items whose price was changed on a given date, compute the percentage change in inventory between the 30-day period BEFORE the price change and the 30-day period AFTER the change. Group this information by warehouse.

X Factor

48



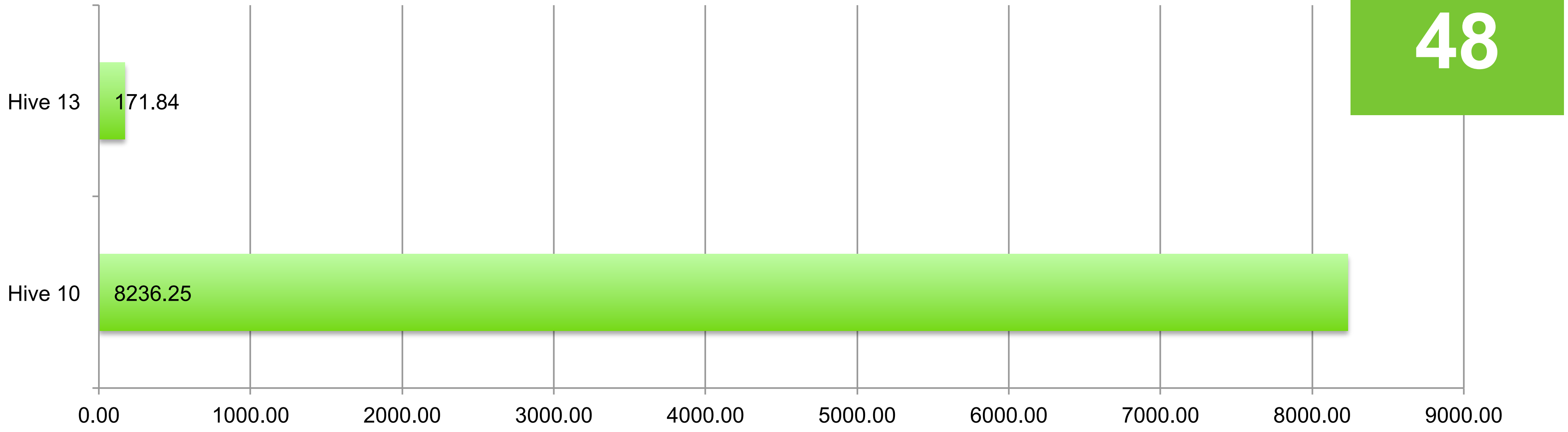
All Values in Seconds

Query 46

Compute the per-customer coupon amount and net profit of all "out of town" customers buying from stores located in 5 cities on weekends in three consecutive years. The customers need to fit the profile of having a specific dependent count and vehicle count. For all these customers print the city they lived in at the time of purchase, the city in which the store is located, the coupon amount and net profit.

X Factor

48

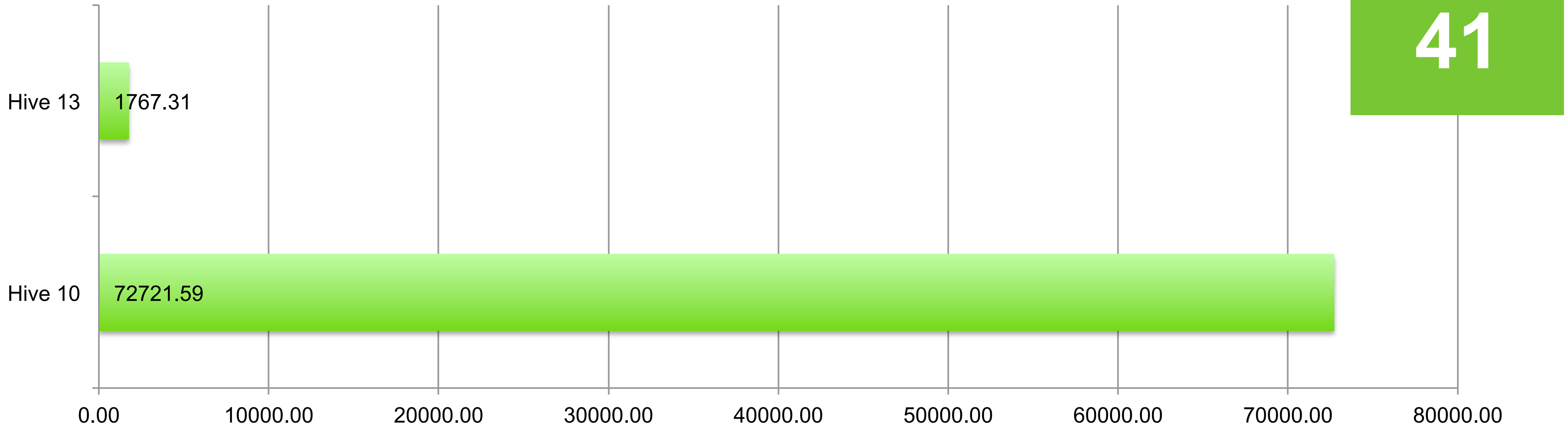


All Values in Seconds

Hortonworks

Query 88

How many items do we sell between pacific times of a day in certain stores to customers with one dependent count and 2 or less vehicles registered or 2 dependents with 4 or fewer vehicles registered or 3 dependents and five or less vehicles registered. In one row break the counts into sells from 8:30 to 9, 9 to 9:30, 9:30 to 10 ... 12 to 12:30



X Factor

41

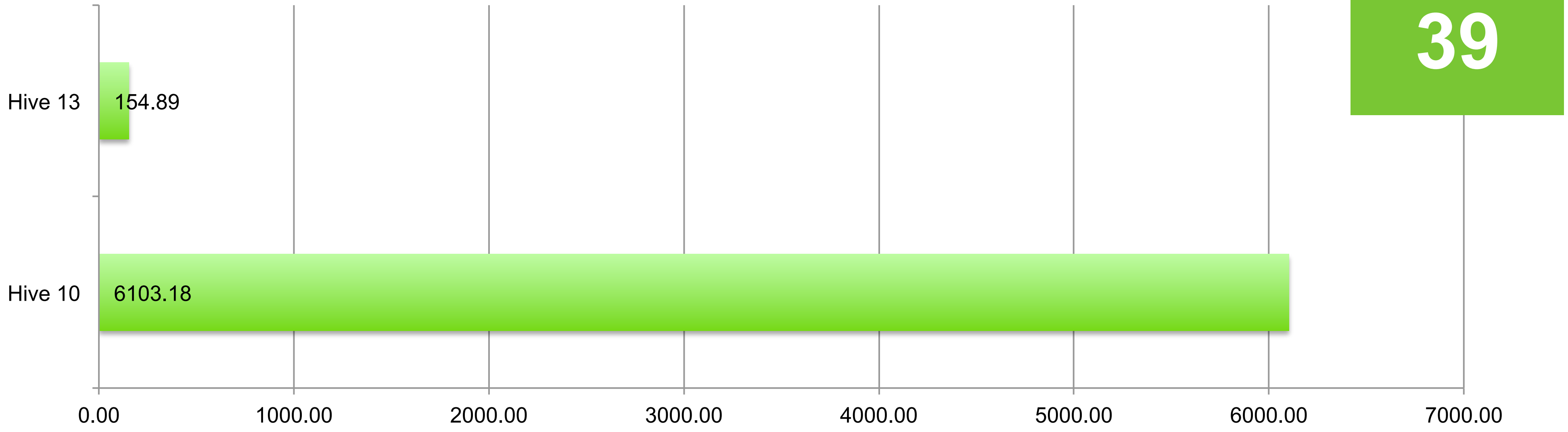
All Values in Seconds

Query 32

Compute the total discounted amount for a particular manufacturer in a particular 90 day period for catalog sales whose discounts exceeded the average discount by at least 30%.

X Factor

39



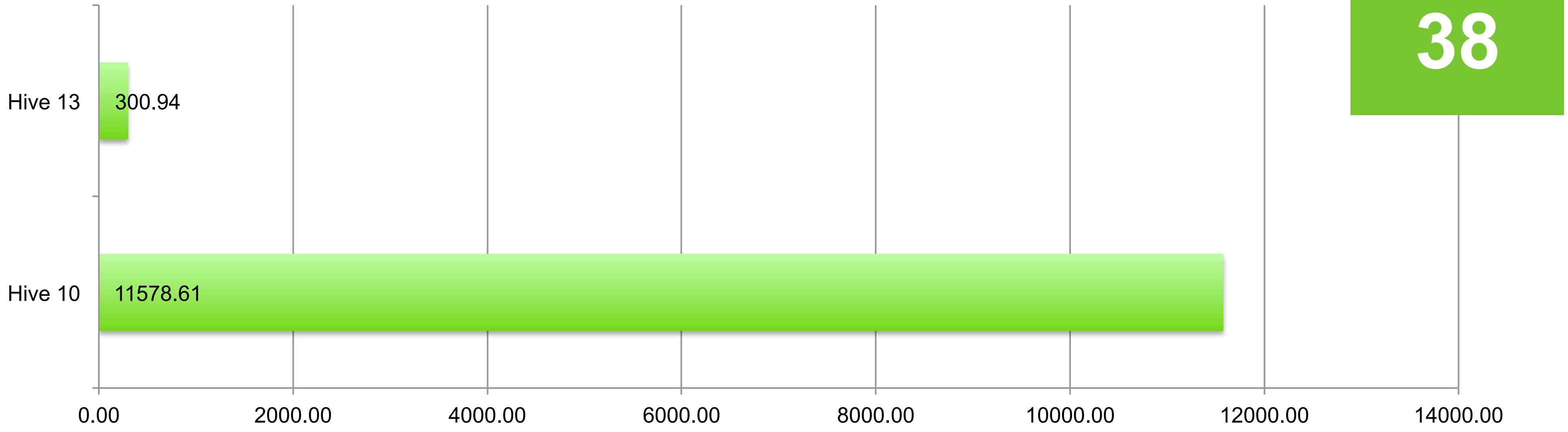
All Values in Seconds

Query 17

Analyze, for each state, all items that were sold in stores in a particular quarter and returned in the next three quarters and then re-purchased by the customer through the catalog channel in the three following quarters.

X Factor

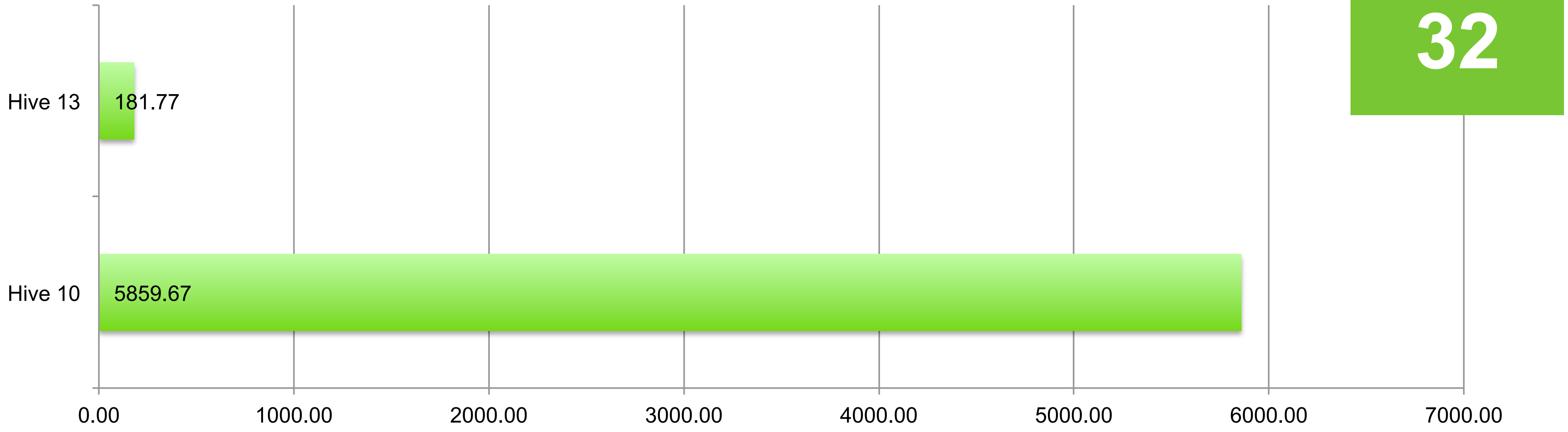
38



All Values in Seconds

Query 94

Produce a count of web sales and total shipping cost and net profit in a given 60 day period to customers in a given state from a named web site for non returned orders shipped from more than one warehouse.



X Factor

32

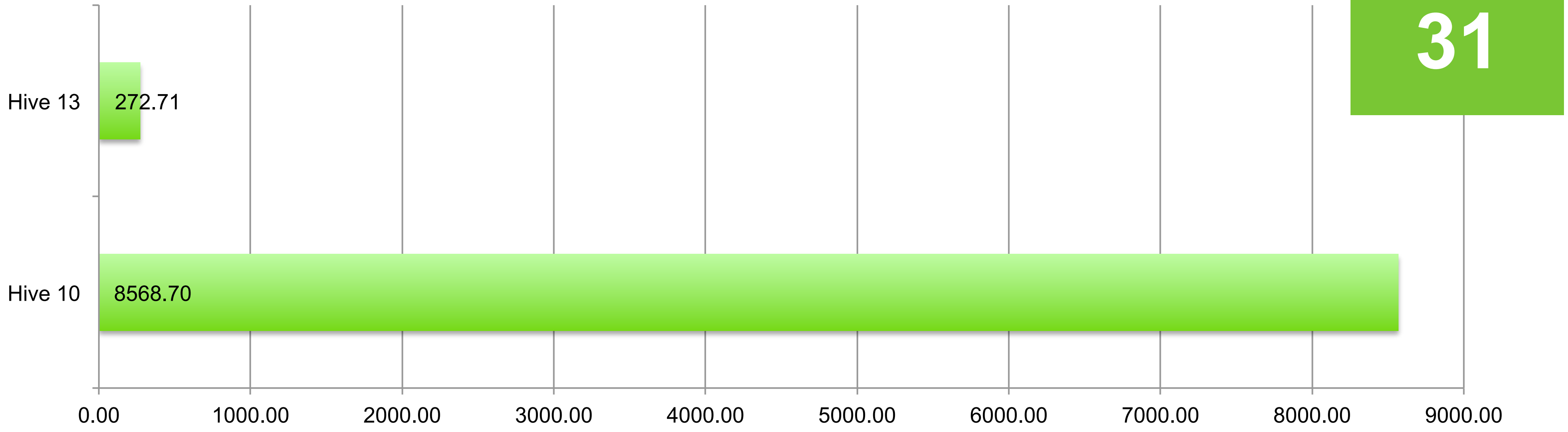
All Values in Seconds

Query 79

Compute the per customer coupon amount and net profit of Monday shoppers. Only purchases of three consecutive years made on Mondays in large stores by customers with a certain dependent count and with a large vehicle count are considered.

X Factor

31



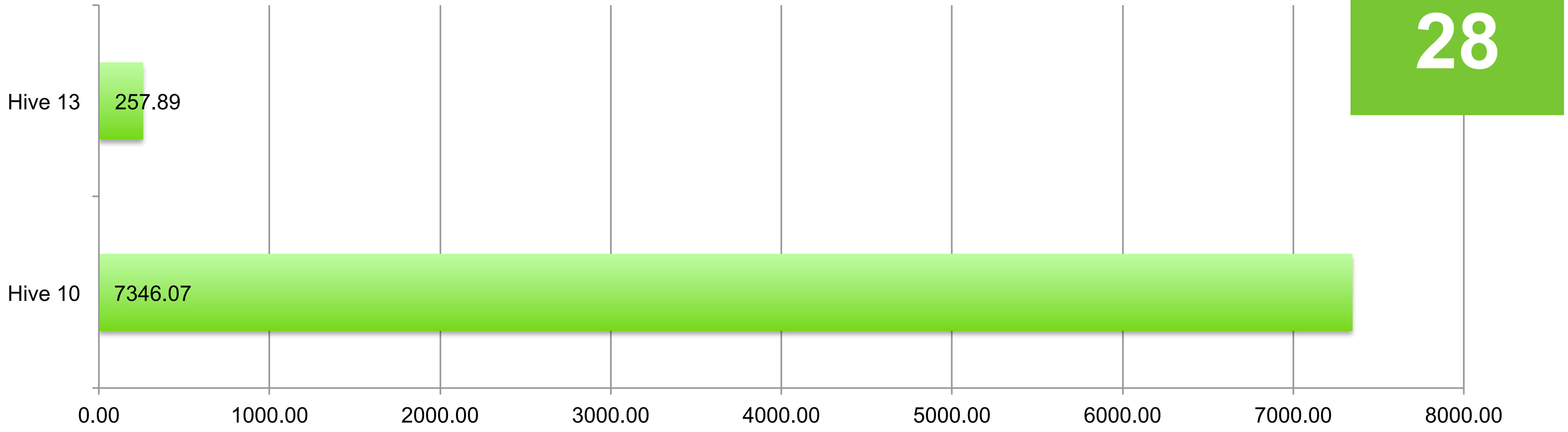
All Values in Seconds

Query 76

Computes the average quantity, list price, discount, sales price for promotional items sold through the web channel where the promotion is not offered by mail or in an event for given gender, marital status and educational status.

X Factor

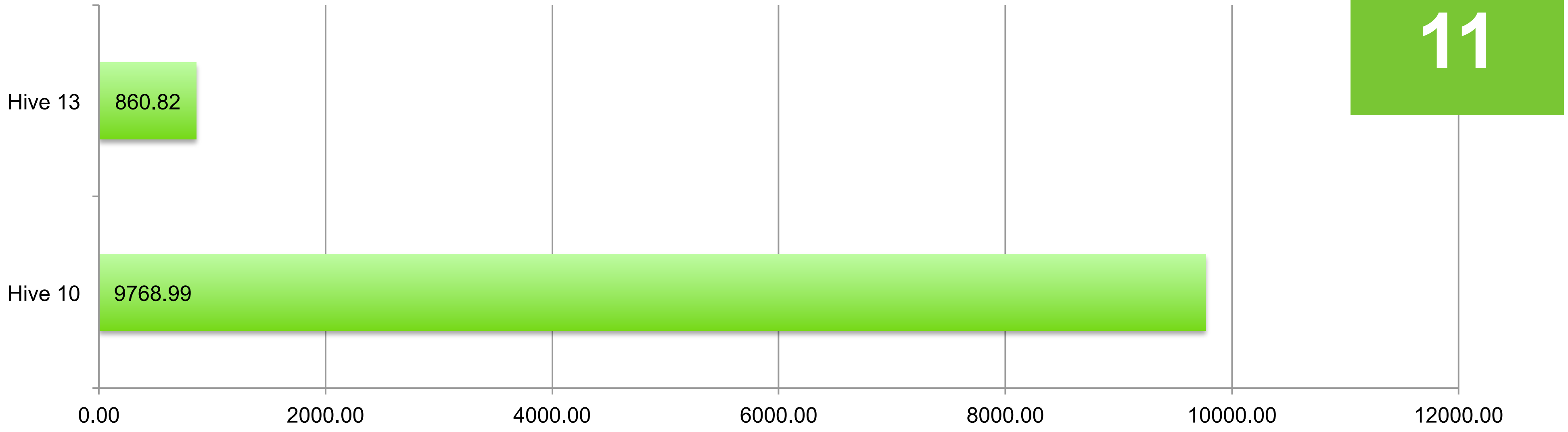
28



All Values in Seconds

Query 92

Compute the total discount on web sales of items from a given manufacturer over a particular 90 day period for sales whose discount exceeded 30% over the average discount of items from that manufacturer in that period of time.

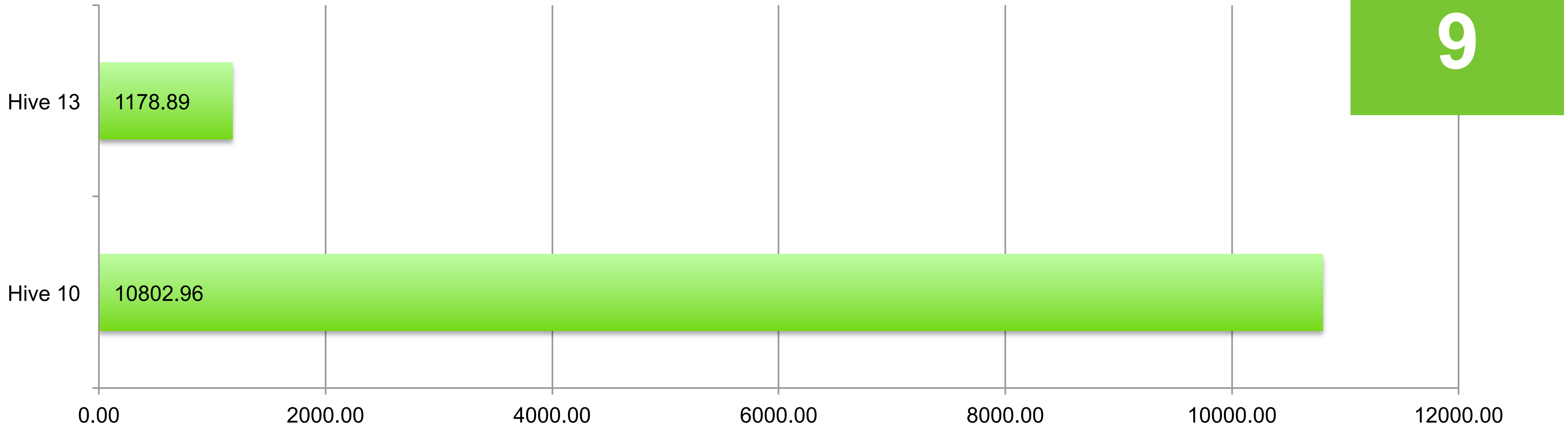


X Factor
11

All Values in Seconds

Query 97

Generate counts of promotional sales and total sales, and their ratio from the web channel for a particular item category and month to customers in a given time zone.



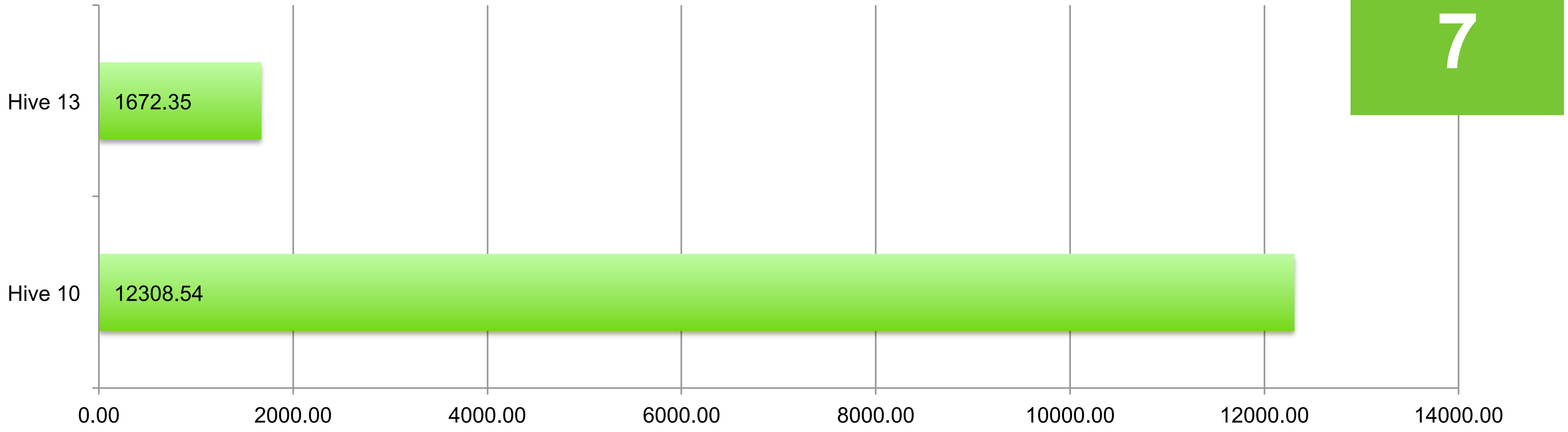
X Factor

9

All Values in Seconds

Query 87

Count how many customers have ordered on the same day items on the web and the catalog and on the same day have bought items in a store.



X Factor

7

All Values in Seconds

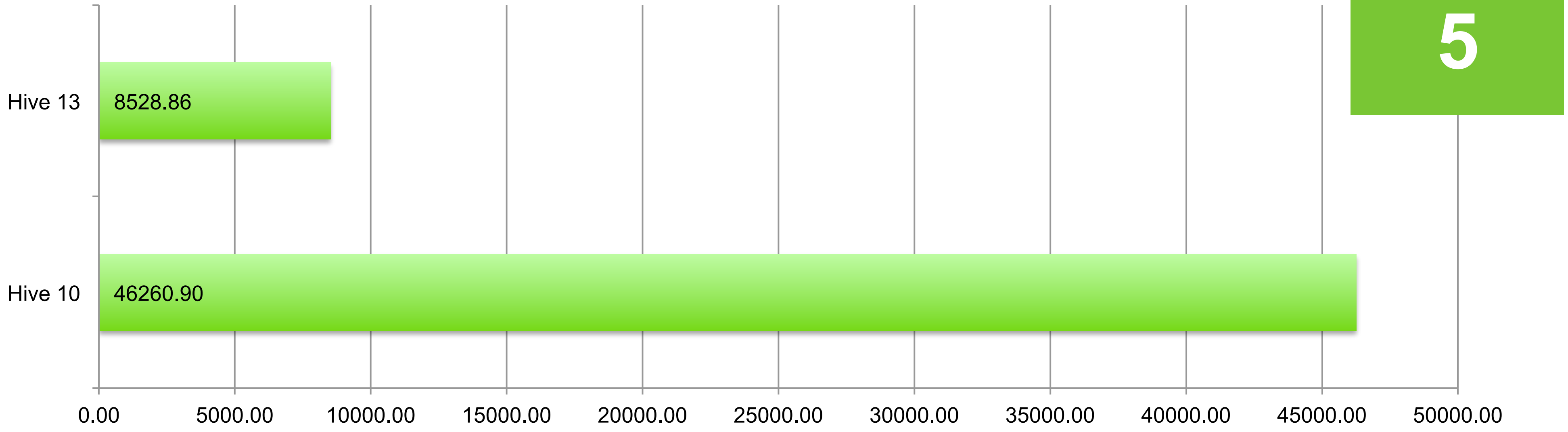
Hortonworks

Query 13

Calculate the average sales quantity, average sales price, average wholesale cost, total wholesale cost for store sales of different customer types (e.g., based on marital status, education status) including their household demographics, sales price and different combinations of state and sales profit for a given year.

X Factor

5

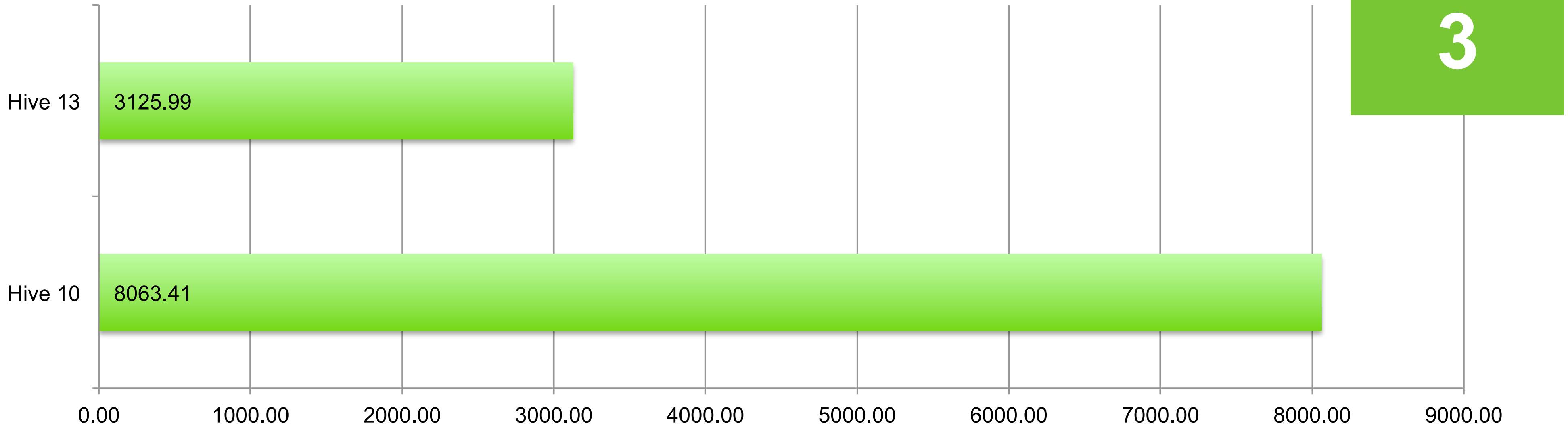


All Values in Seconds

Hortonworks

Query 50

For each store count the number of items in a specified month that were returned after 30, 60, 90, 120 and more than 120 days from the day of purchase.



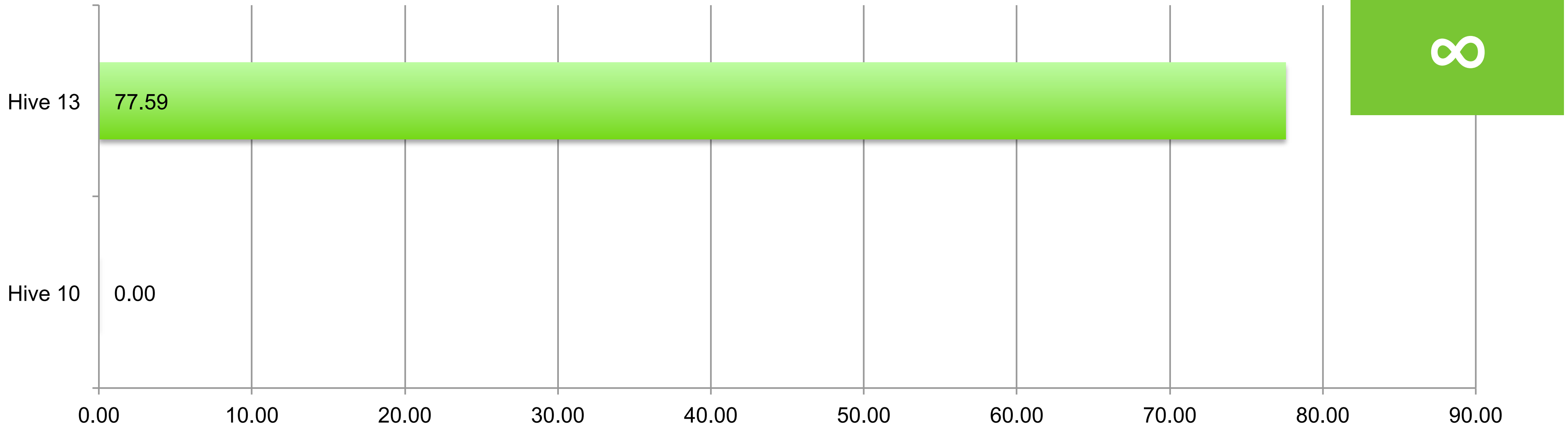
X Factor

3

All Values in Seconds

Query 20

Compute the total revenue and the ratio of total revenue to revenue by item class for specified item categories and time periods.



X Factor

∞

All Values in Seconds

Hortonworks

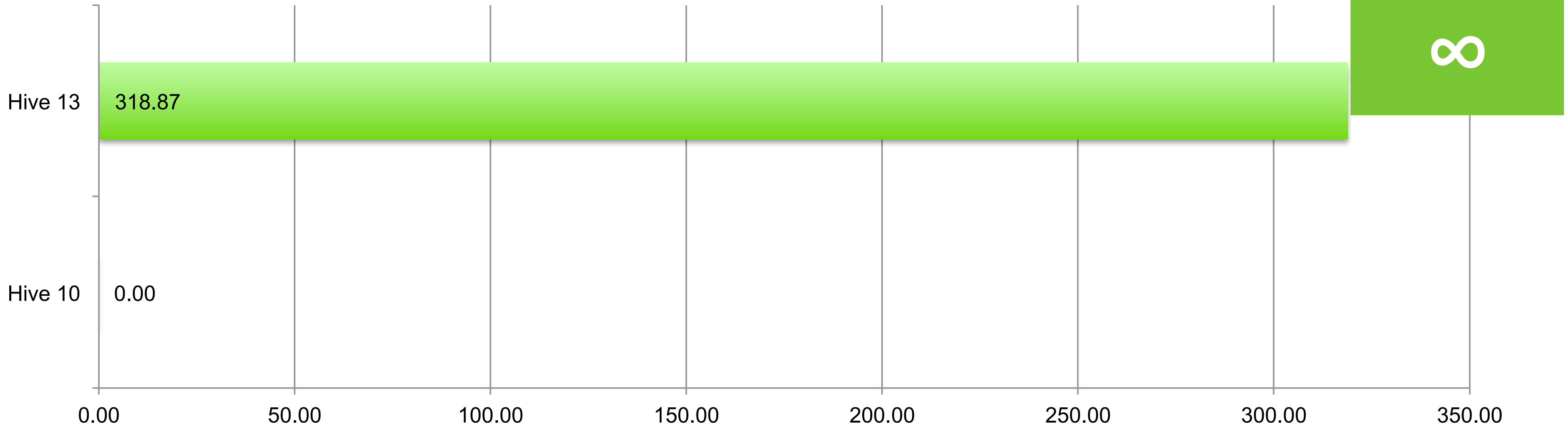
Query 25

Get all items that were sold in stores in a particular month and year AND returned in the next three quarters AND re-purchased by the customer through the catalog channel in the six following months.

For these items, compute the sum of net profit of store sales, net loss of store loss and net profit of catalog. Group this information by item and store.

X Factor

∞



All Values in Seconds

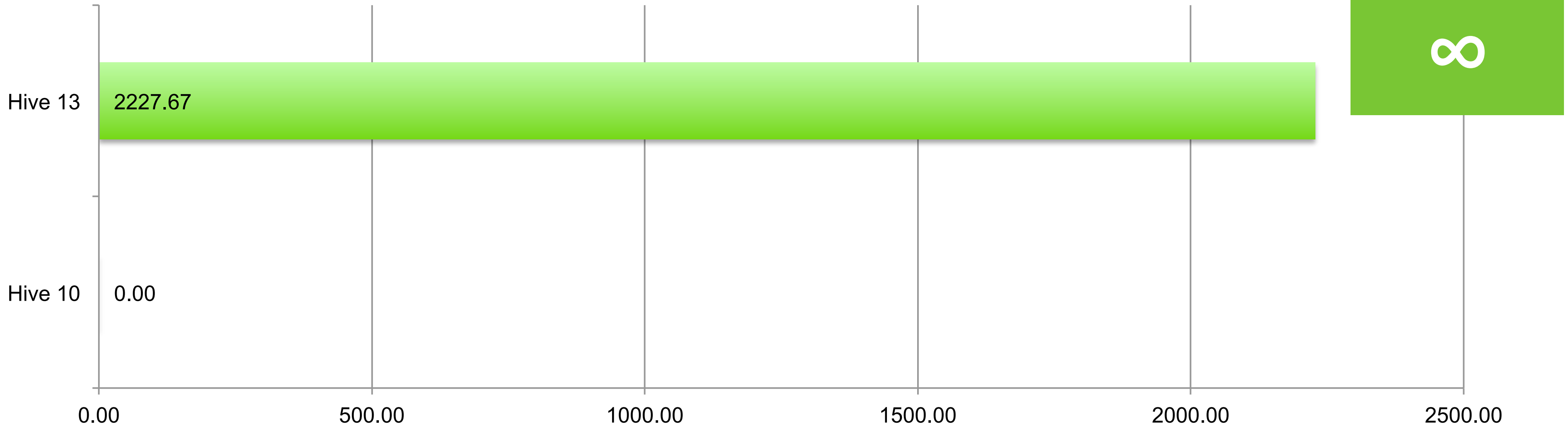
Hortonworks

Query 28

Calculate the average list price, number of non empty (null) list prices and number of distinct list prices of six different sales buckets of the store sales channel. Each bucket is defined by a range of distinct items and information about list price, coupon amount and wholesale cost.

X Factor

∞

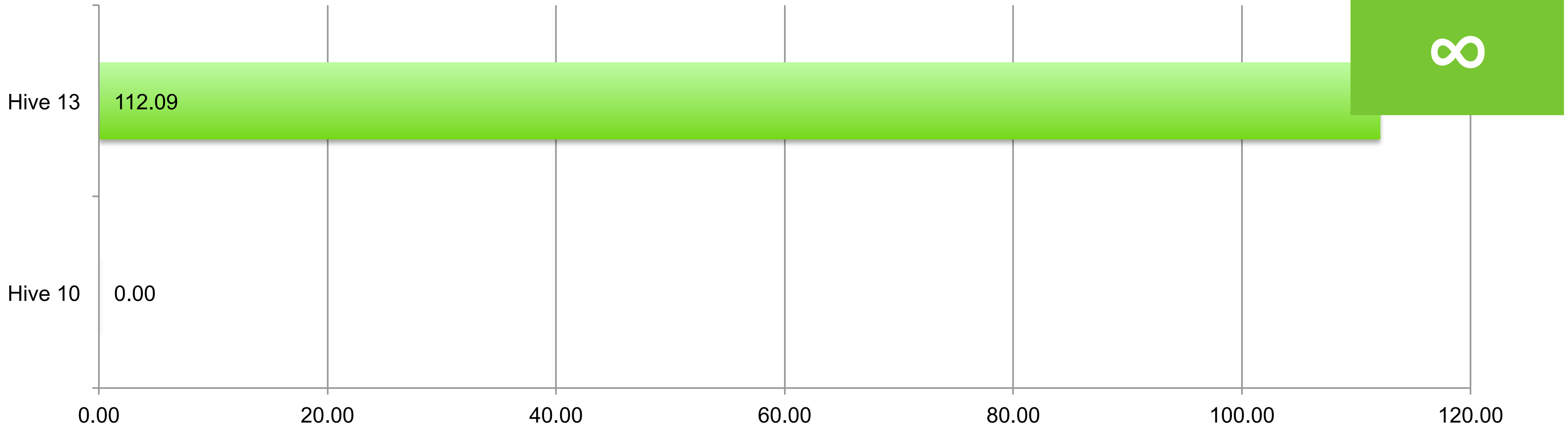


All Values in Seconds

Hortonworks

Query 45

Report the total web sales for customers in specific zip codes, cities, counties or states, or specific items for a given year and quarter.



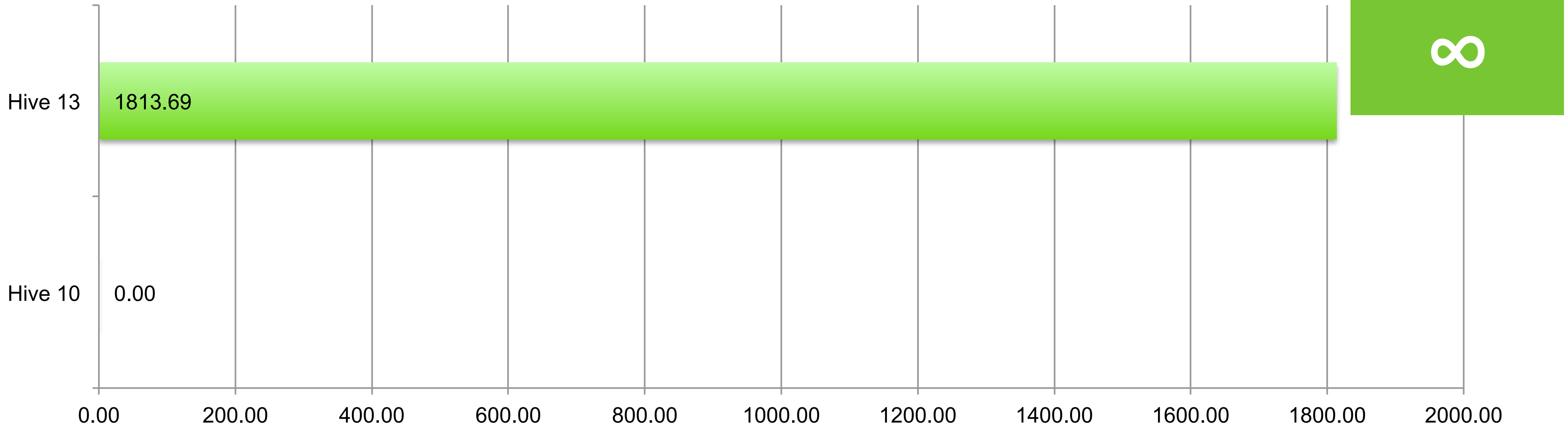
X Factor

∞

All Values in Seconds

Query 48

Calculate the total sales by different types of customers (e.g., based on marital status, education status), sales price and different combinations of state and sales profit.



X Factor

∞

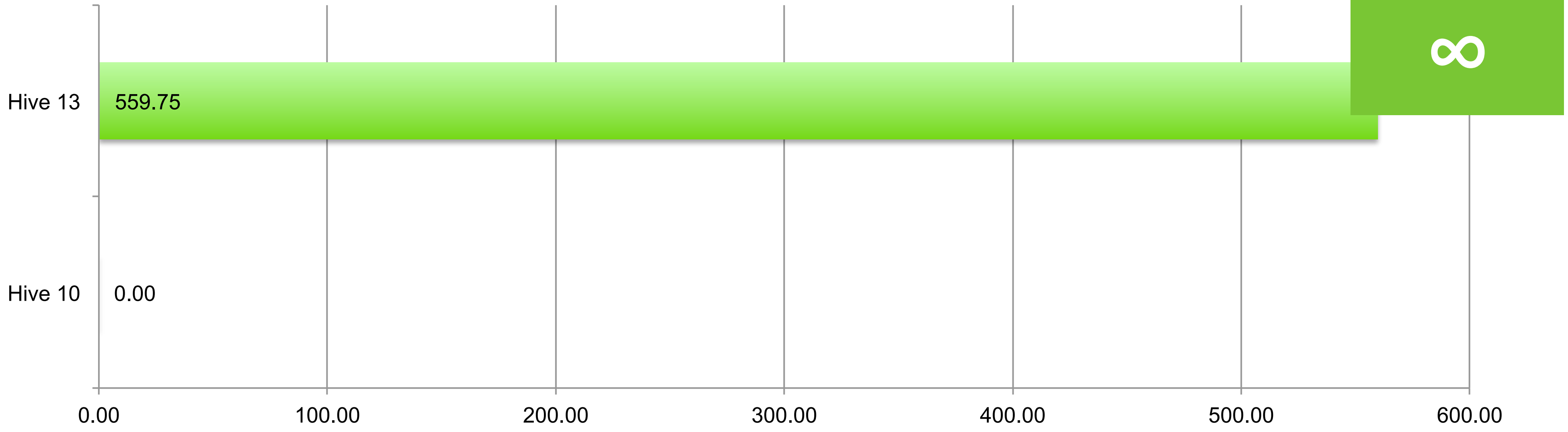
All Values in Seconds

Query 49

Report the top 10 worst return ratios (sales to returns) of all items for each channel by quantity and currency sorted by ratio. Quantity ratio is defined as total number of sales to total number of returns. Currency ratio is defined as sum of return amount to sum of net paid.

X Factor

∞



All Values in Seconds

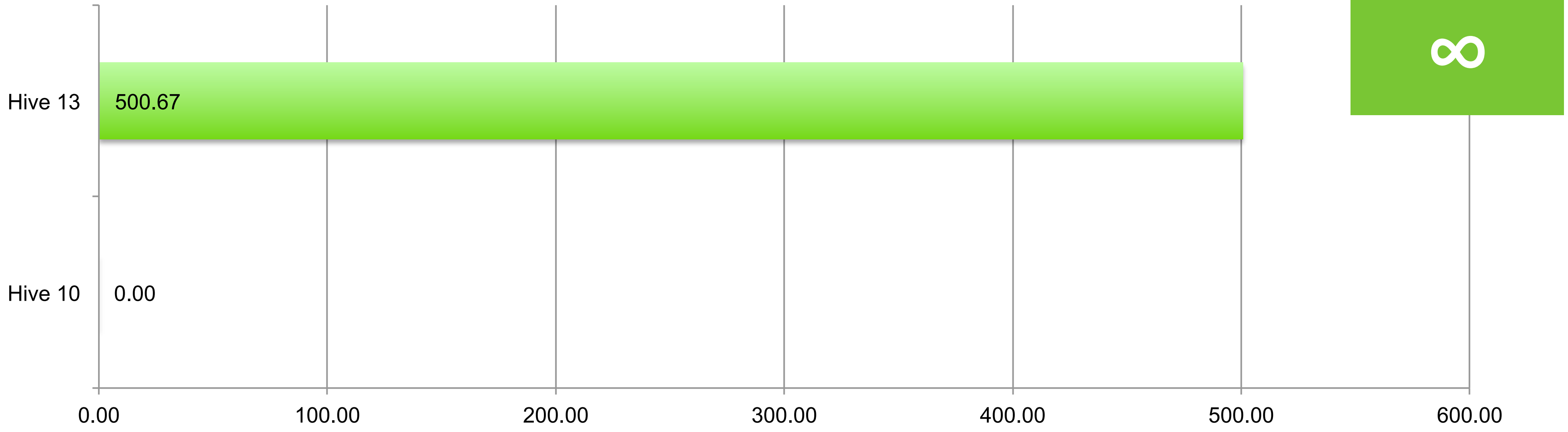
Hortonworks

Query 85

For all web return reason calculate the average sales, average refunded cash and average return fee by different combinations of customer and sales types (e.g., based on marital status, education status, state and sales profit).

X Factor

∞

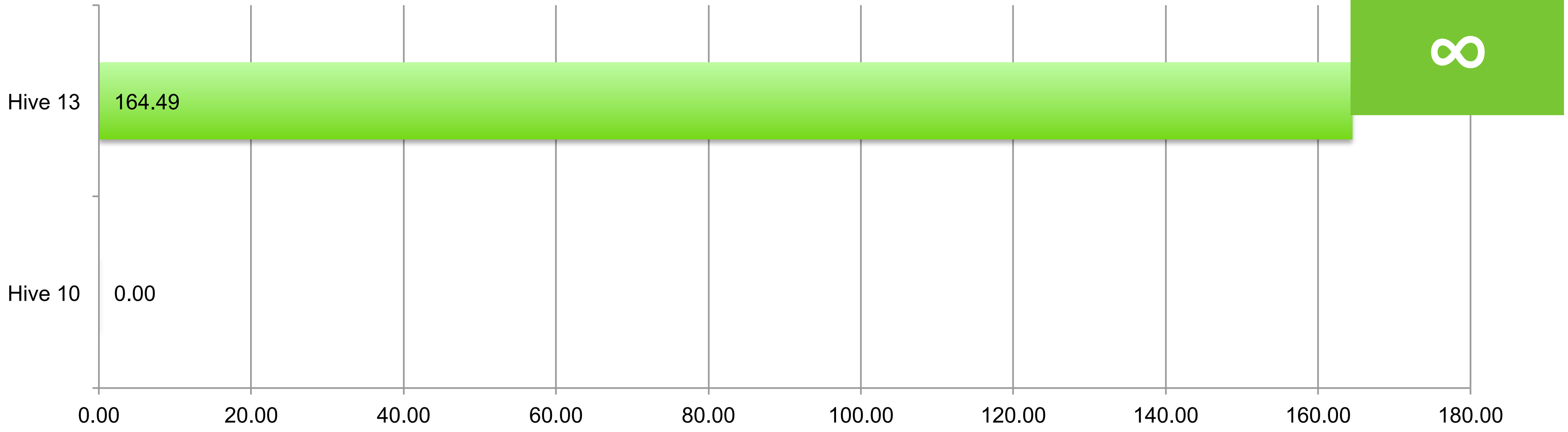


All Values in Seconds

Hortonworks

Query 89

Within a year list all month and combination of item categories, classes and brands that have had monthly sales larger than 0.1 percent of the total yearly sales.



X Factor

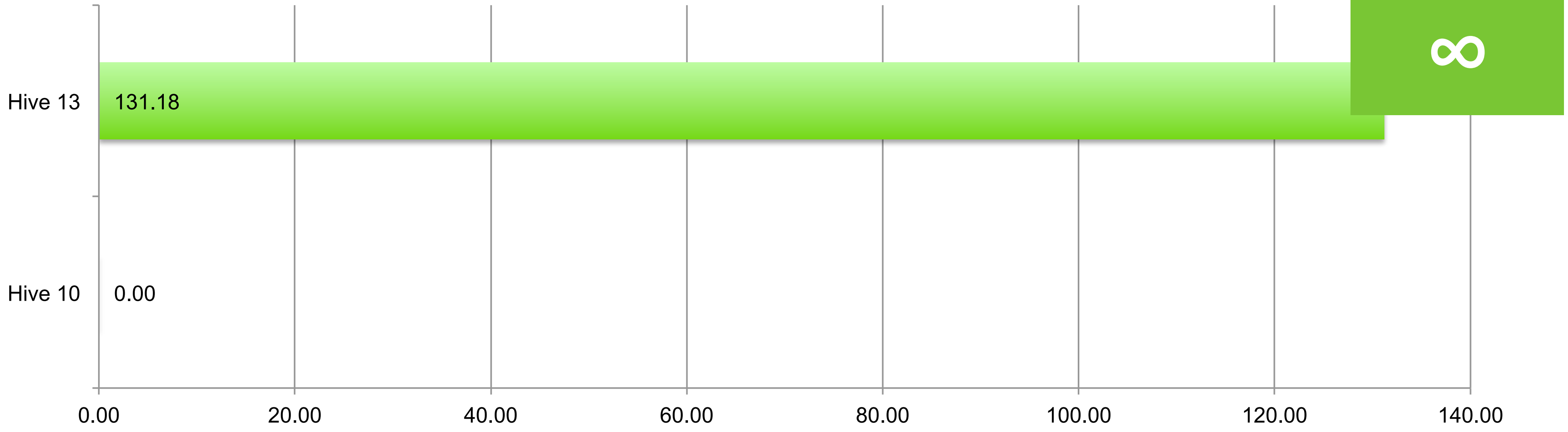
∞

All Values in Seconds

Hortonworks

Query 90

What is the ratio between the number of items sold over the internet in the morning (8 to 9am) to the number of items sold in the evening (7 to 8pm) of customers with a specified number of dependents. Consider only websites with a high amount of content.



X Factor

∞

All Values in Seconds

Results for Data Mining Queries

Queries #34, 39, 64, 71, 73 & 98

Results for Data Mining Queries

Deep Mining: Queries that return large amounts of data for further processing by other tools

Query #	Query Description	Hive 13	Hive 10	Change
73	Count the number of customers with specific buy potentials and whose dependent count to vehicle count ratio is…	48.98	5,866.06	120X
71	Select the top 10 revenue generating products, sold during breakfast or dinner time for one month…	91.61	10,498.57	115X
34	Display all customers with specific buy potentials and whose dependent count to vehicle count ratio is larger than 1.2…	125.72	6,745.30	54X
39	Query with multiple, related iterations…	111.74	2,452.08	22X
64	Find those stores that sold more cross-sales items from one year to another	6,821.24	34,289.66	5X
98	Report on items sold in a given 30 day period, belonging to the specified category.	1,085.06	NA	∞

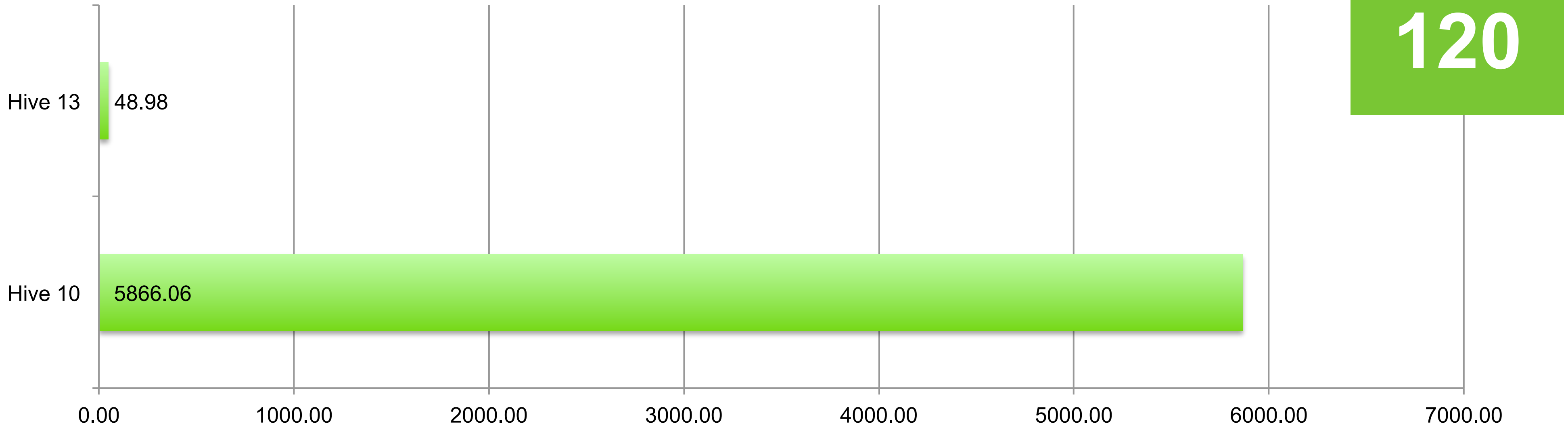
All times in seconds

Query 73

Count the number of customers with specific buy potentials and whose dependent count to vehicle count ratio is larger than 1 and who in three consecutive years bought in stores located in 4 counties between 1 and 5 items in one purchase. Only purchases in the first 2 days of the months are considered.

X Factor

120



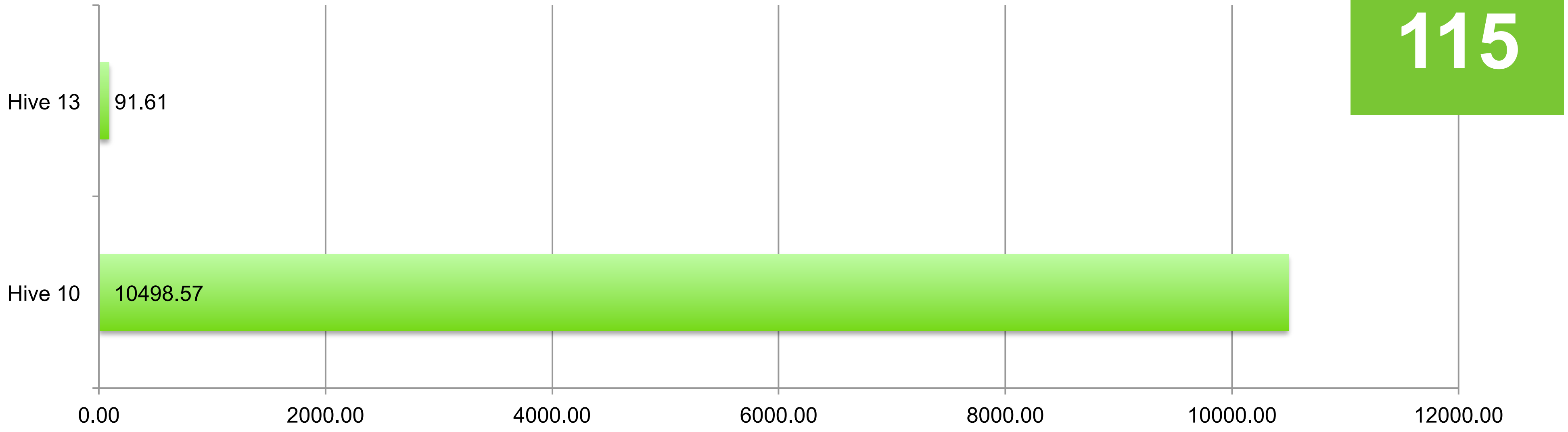
All Values in Seconds

Query 71

Select the top 10 revenue generating products, sold during breakfast or dinner time for one month managed by a given manager across all three sales channels.

X Factor

115



All Values in Seconds

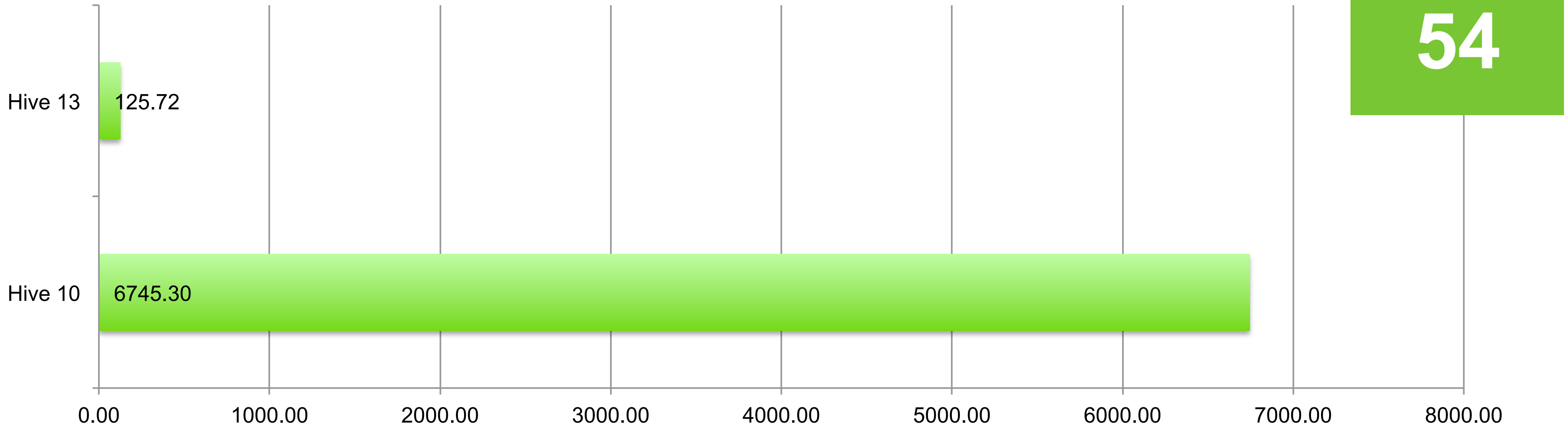
Hortonworks

Query 34

Display all customers with specific buy potentials and whose dependent count to vehicle count ratio is larger than 1.2, who in three consecutive years made purchases with between 15 and 20 items in the beginning or the end of each month in stores located in 8 counties.

X Factor

54



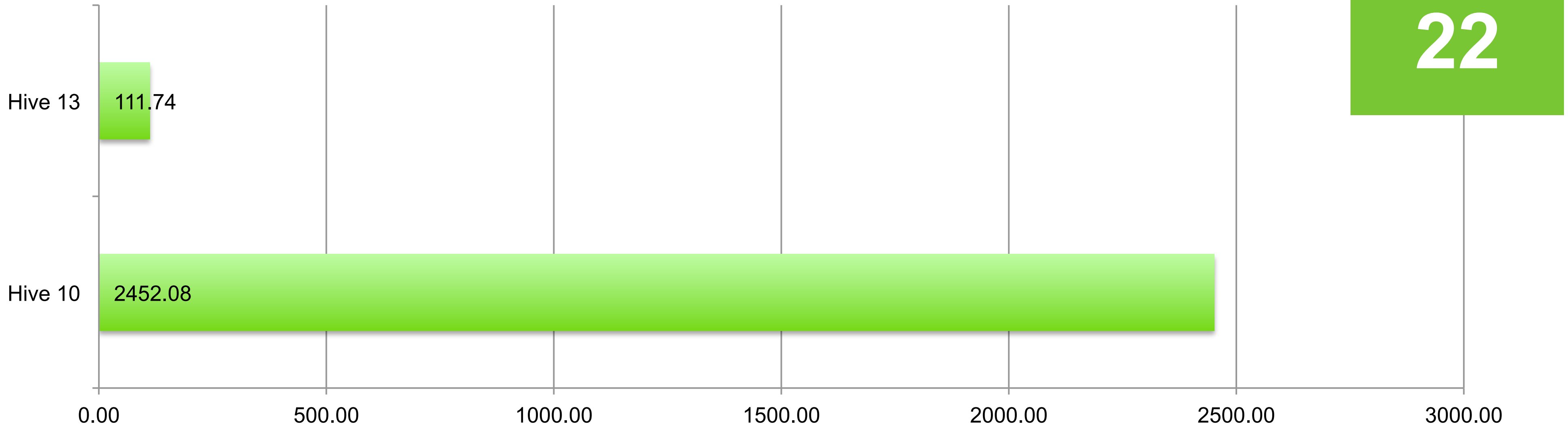
All Values in Seconds

Hortonworks

Query 39

This query contains multiple, related iterations:

- Iteration 1: Calculate the coefficient of variation and mean of every item and warehouse of two consecutive months
- Iteration 2: Find items that had a coefficient of variation in the first months of 1.5 or large



X Factor

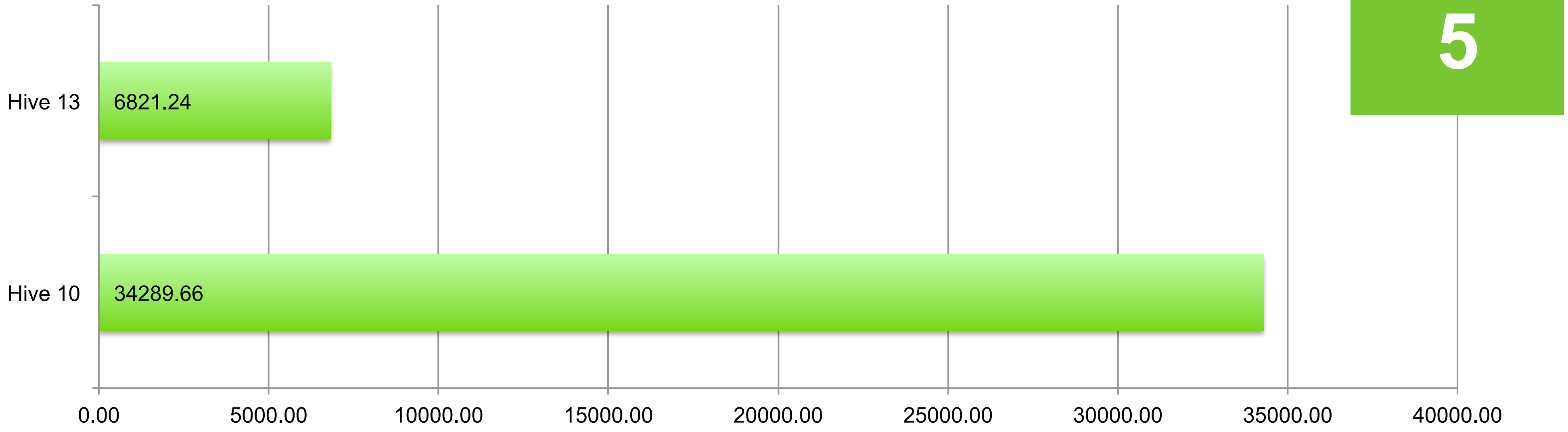
22

All Values in Seconds

Hortonworks

Query 64

Find those stores that sold more cross-sales items from one year to another. Cross-sale items are items that are sold over the Internet, by catalog and in store.



X Factor

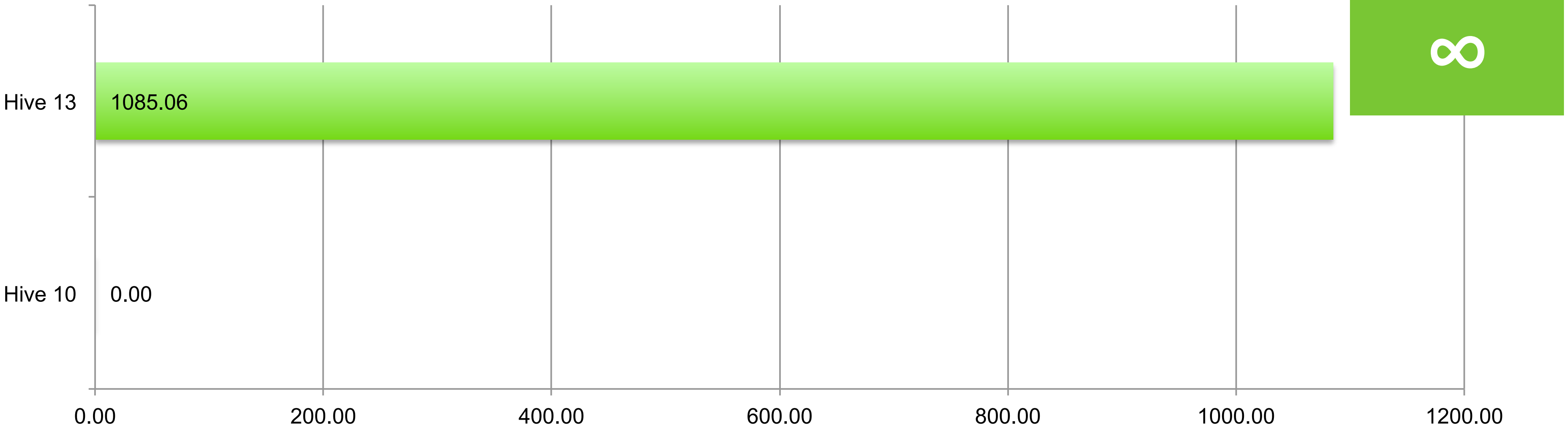
5

All Values in Seconds

Hortonworks

Query 98

Report on items sold in a given 30 day period, belonging to the specified category.



X Factor

∞

All Values in Seconds