CS346 Advanced Databases Report

Valentin Kodderitzsch (1931737)

1 Individual Reflection

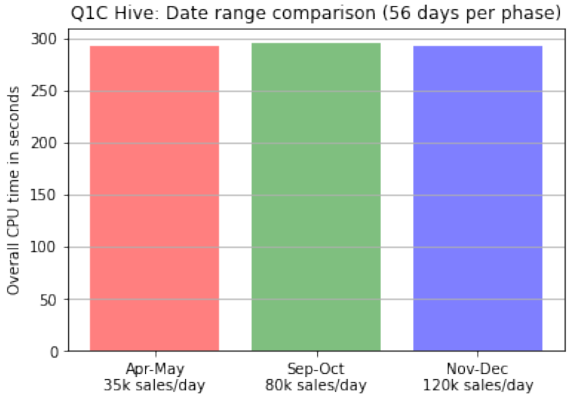
My first contribution to this group project was my SQL experience. We were working as a team and collectively we were able to solve most queries before the end of labs. However, I quickly realised that query 1 B is a non-standard SQL query. After some googling we found a solution using partitions. Here my past SQL experience proved useful as I have worked for a year in industry for a company that maintained SQL code in almost every product they sold. During said placement experience I have been told that non-standard queries should be avoided. Reason being that non-standard queries which use partitions or recursion border the limits of the SQL paradigm and make the code less readable. In scenarios where an aggregated output can only be achieved using non-standard SQL other data manipulation paradigms should be used. My initial thought got confirmed at the end of our project when I profiled the Hive versus MapReduce implementations for query 1 B.

My second contribution to this group project was the detailed exploratory analysis (EDA) I provided. To this end, I learned about the multitude of techniques proposed by Turkey [4, 3] such as outliers, the interquartile range and min, max statistics. I also learned about the coefficient of variation [1] which is an interesting statistic to compare the measure of dispersion between multiple attributes of different scales. To automate my EDA pipeline I developed my own summary statistics function using Python. A sample output of said function is shown in figure 1b.

Due to my EDA I was able to help my peers during their development of the MapReduce implementations. Concrete examples include discussion about null values and how they are stored, the total number of records or discussions about date ranges.

My final contribution to this project was deciding on a profiling strategy, running all the tests and analysing the collected test data. After having read the section about selectivy estimators in the databases textbook by Elmasri et al. [2] I was convinced that choosing a different date range would affect the running times of the queries. Especially, as the sold date attribute had a non-uniform, cyclical, three step distribution. However, I was proven wrong as shown in figure 1a and more details can be found in the report.

In the end, my main learning from this project is that Hive is superior to MapReduce for the vast majority of analytical queries. The only exception are non-standard queries which are meant to be executed as compiled queries. Additionally, I have seen that any Hadoop architecture imposes a start up time overhead on queries which made me realise that said paradigm only makes sense for data which a normal relational database cannot handle.



(a) Testing assumed selectivity of attributes

```
====== NAME: ss_net_paid (115,203,420) ==================
Null values count and %: 5,181,664 || 4.50%

Duplicates count: 114,178,206
Uniques count: 1,025,214

Min and max: 0.00 || 19,744.00

Mean and median (50%): 1,721.10 || 865.04

Std and CV: 2,182.28 || 1.2680

----- Q1, Q3 and IQR:  236.04 || 2,364.12 ||| 2,128.08
Gaus: Q1, Q3 and skew: 258.97 || 3,183.23 ||| 2.1422

Outliers count below and above: 0 || 7,662,532
```

(b) Custom EDA method

Figure 1: Learnings

References

[1] Hervé Abdi. "Coefficient of variation". In: *Encyclopedia of research design* 1.5 (2010).
[2] R Elmasri et al. *Fundamentals of Database Systems*. 7th ed. Springer, 2015. ISBN: 978-0133970777.
[3] Georgy Shevlyakov et al. "Robust versions of the Tukey boxplot with their application to detection of outliers". In: *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE. 2013, pp. 6506–6510.
[4] John W Tukey et al. *Exploratory data analysis*. Vol. 2. Reading, MA, 1977.