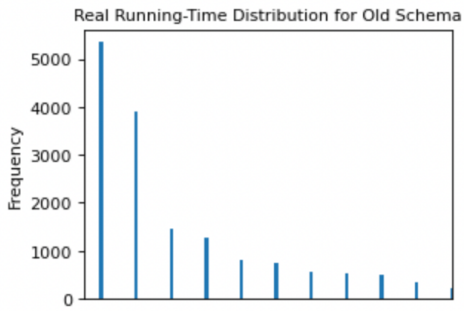
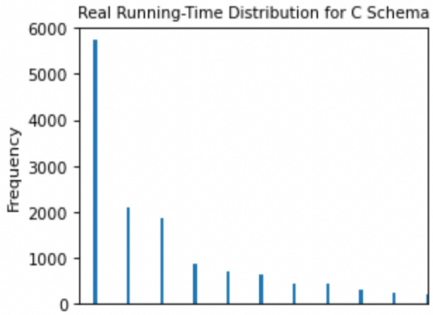
Delphi EPI4 Analysis

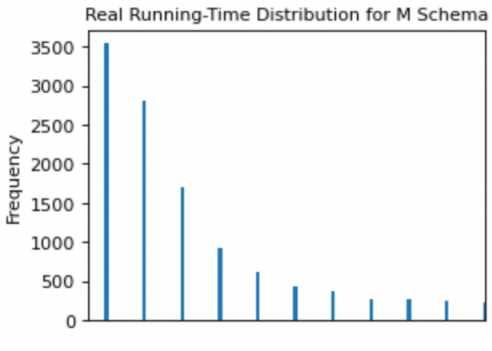
The goal of this analysis is to communicate the run-time differences between the three schemas that the data presented for the implementation of SQL queries. The analyses were conducted using a Jupyter notebook, which is currently available on the Github repository. The analyses used various libraries – Pandas, NumPy, Matplotlib, Statistics, SciPy, Warnings.

The analyses initially discovered that the results did not match one to one with the query data file. To avoid improper matching and analyses, the unevenness in the rows was controlled for the three results files for the three schemas. This step was initially done to import the geographical, time, publication, and signal axes values and query type for the three results files. After aggregating the relevant data and correctly matching the required data points, the analyses was performed.





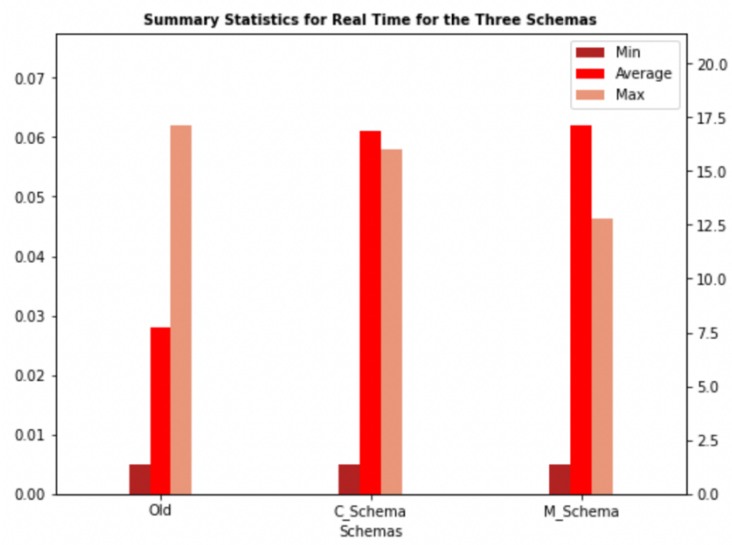
After accounting for the unevenness in the files, the first step posed the answer to check the distribution of the real running-time for the three results files. This step was considered to implore any significant changes that existed between the results of three results files. As we can see from the three plots, we can deduce that the three files have the same distribution for the real running time.



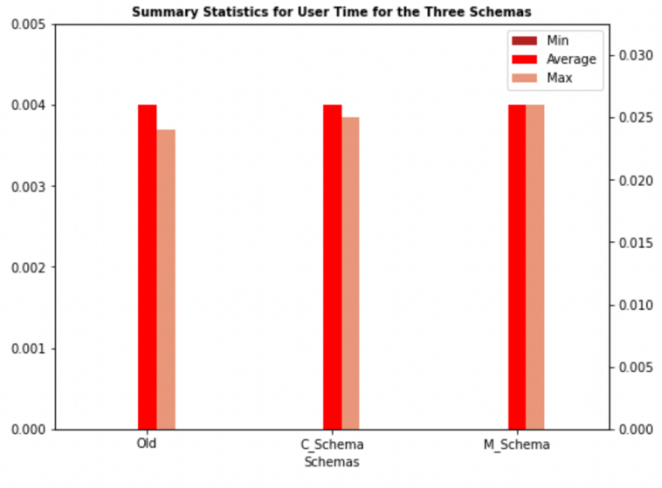
The three plots show that most of the run-times were very close to zero, implying the quick run-time for most of the queries that are presented in the query file. This directly implies the presence of few complex queries in the query file. The similar distribution meant that the next step would be to compare the summary statistics for the different run-times – real, user, and kernel – for the three schemas. This analysis would provide an overall idea of the run-times of the three schemas.

After evaluating the summary statistics, we discovered that the average real running time for old schema is 0.028 unix utility, but the worst-case and the best-case run-time for old schema are 17.121 unix utility and 0.005 unix utility, respectively. The number for the average real running time for C schema is 0.061 unix utility, including the worst-case and the best-case run-time for C schema are 15.997 unix utility and 0.005 unix utility, respectively. For the M schema, the average real running time is 0.062 unix utility, the worst-case is 12.788 unix utility, and the best-case run-time is 0.005 unix utility. For the user running time, we discovered that the average running time for old schema is 0.004 unix utility, with the worst case and the best-case run-time being 0.024 unix utility and 0.0 unix utility, respectively. On the other hand, the average user running time for C schema is 0.004 unix utility, along with the worst case and the best-case run-time for C schema being 0.025 unix utility and 0.0 unix utility, respectively. To conclude, the average user running time for M schema is 0.004 unix utility, but the worst case and the best-case run-time for M schema are 0.026 unix utility and 0.0 unix utility, respectively. We did not perform any analysis on the rows as the results were expected to be the same because the same queries were run for the three schemas.

The above-mentioned summary statistics show that there are some differences in the summary statistics when compared between the three schemas. Therefore, we intended to compare the running time for real and user. The analyses concluded the exclusion of kernel running time as the correlation matrix showed an extremely high (-0.9) negative correlation between kernel running time and user running time. Therefore, the analyses were narrowed down to comparisons strictly for real and user running time for the three schemas.

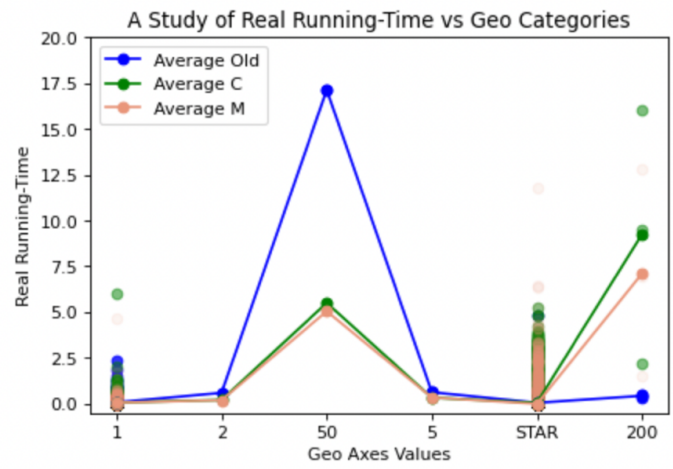


The results for real time conclude that the old schema on an average performed better compared to the new schemas, however, for certain queries it had a worse worst performance compared to the worst performance of the M schema (it was similar to the worst performance of C schema). This implied that for certain queries, the old schema could take much longer compared to M schema. M schema had the best max run-time, implying that it never took extremely long. This helps us also conclude that C schema overall did not outperform the old schema and M schema in any factor of measure for the real run-time.



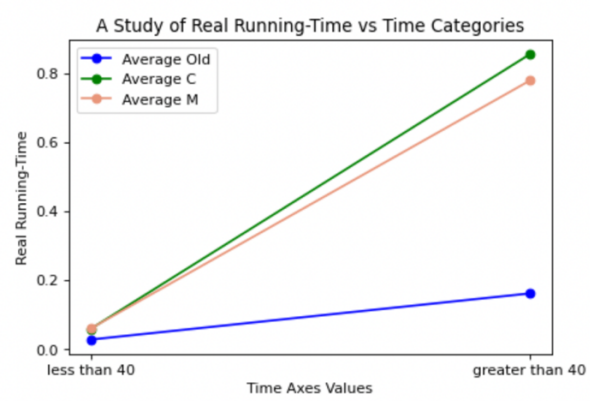
The plot on the left for the user time did not show significant differences; the plot only indicates that the old schema somewhat worked well with the user time for the worst performances.

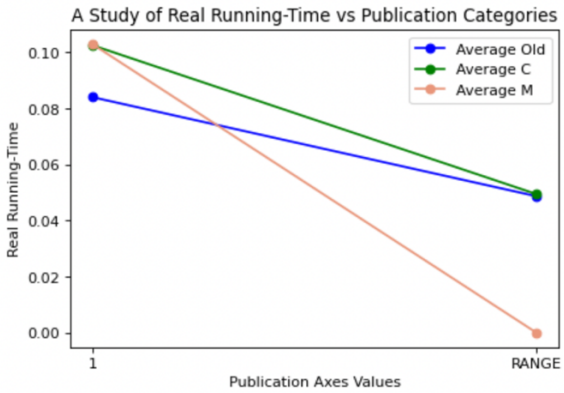
The analyses conducted has only considered factors individually for the running time for the three schemas. The results show that on an average the old schema had the best real running-time. The M schema had the best worst running-time compared to the other schemas. The user running time does not fluctuate significantly for the three schemas. However, it would be important to evaluate and compare if there could be potentially significant differences in the real running-time when we compare across different axes.



We had restricted our analyses to real running time as we had discovered not too much difference for the user running time. The last few plots helped us understand how the three schemas differed in their mean running time for the different values that were used for the four different axes. We did not consider the signal axis for analysis as the signal axis only had one value of 1. Therefore, the analyses were conducted on the remaining three axes. We discovered that the schema M had the best running time for all geo values except 200. This is contrary to what we had discovered that the old schema was generally the best. For the queries that had data collected for time value less than 40, we discovered that the old schema performed the best in terms of the running time.

Finally, the publication axis showed mixed results, where M was the best for a range of publication values and the old schema was the best performer for the publication axis values as 1 for the four different categories - lag, latest, asof, and issue.





Therefore, to conclude, the analyses considered only the real time as the user time produced similar results for the three schemas and the kernel time was highly negatively correlated with user time. This changed the direction of the analyses towards real running time, which was further brought down to multiple axes values. The conducted analyses overall shows that the old schema outperformed the two new schemas in terms of real running-time. The advantage of the M schema was its worst-case running time being 33% less than the worst-case running time for both the old schema and C schema. C schema did not really show any advantages over the other two and the further detailed analyses based on the axes solidifies the argument. Finally, the M schema performs better than the old schema for geo and publication as opposed to the old schema that outperforms the other two schema for time axis.