# **High Performance Computing and Big Data**

# Declaration

No part of this project has been submitted in support of an application for any other degree or qualification at this or any other institute of learning. Apart from those parts of the project containing citations to the work of others, this project is my own unaided work. This work has been carried out in accordance with the Manchester Metropolitan University research ethics procedures.

Signed: Mason Bowers

Date: 24/05/2024

# Section 1: Big Data Assignment

H1: “Rides in 2014 starting from Baylis Road, Waterloo Station were shorter compared to other stations."

H0: “No statistically significant Difference in ride time between Baylis Road and other Stations."

This will be tested by a one-tailed T-Test with a confidence interval of 95%.

### *Feature Analysis:*

Quantitative Features: Duration, Rental ID, Start Date, End Date

Qualitative Features: Start Station, End Station

Rows: 11481596

Columns: 9

### *Features Related to Hypothesis:*

Duration is the time that the bike was rented for. It is measured in seconds, although the granularity is 1 minute. This means that the smallest difference possible between 2 duration measurements is 60 seconds.

Start Station ID is the unique ID number corresponding to the Station that the ride began at.

Start Station Name is the Name of the Station that the ride began at. Directly related to the station ID.

Start Date is the date and time at which the ride started, given in the format “DD/MM/YY 00:00”

Before any other treatment of data is done, all other columns will be removed from the Dataframe to save processing time.

### *Removal of Outliers:*

The dataset contains many values for duration which are simply erroneous (such as duration < 0) or values that are not feasible for actual use of the bikes (values > 12 hours). The maximum value of duration even states that a user rented a bike for 30 days.

A first step is to remove any rows with null values in any field, followed by removing any rows where Duration < 0, as this is clearly an anomalous result.

The method of dealing with outliers will be to only work within 1.5 IQR’s. This method is commonly used to remove outliers and will make sure that any calculations made, and conclusions drawn are valid.

This calculation removed 672554 entries and reduced the Standard Deviation from 12967 to 519.

### *Feature Engineering:*

To allow separating the DataFrame by month, the month must be extracted from the start date column. This is done by creating a new column with just the date, and then a column containing just the month. This is done using the Split() function.

### *Testing for Statistical Significance:*

Now the data must be split into two groups, rides starting from Baylis Road and rides starting elsewhere. These groups are then split into months, and the mean is calculated. The result is two lists, each containing 12 values (The mean value of each month).

It must now be checked that each list is normally distributed before any tests for Statistical Significance can be selected. For this, a Shapiro-Wilkes test is used, and returns p-values of 0.62 and 0.60 for Baylis and other rides respectively.

A one-tailed T-test can now be done to test the hypothesis.

The T-Test returned a p-value of 1.81410918e-07 which is less than our alpha value of 0.05, and so we reject the null hypothesis and conclude that rides in 2014 starting at Baylis Road were shorter compared to other stations.

# Section 2: Big Data and the Global Environment

Big Data has the potential to impact many facets of society, one of which being our carbon footprint. As we move ever closer to ‘smart cities’ dominated by ICT infrastructure, local governments’ lack of usage of big data becomes more and more apparent, with their reliance on outside sources for their data collection, preventing any forward movement in this regard (Giest, 2017). Despite this, interest in Big Data has increased dramatically (Batty, 2013) and the private sector has begun to use Big Data to work towards improving their environmental impact. BT, for instance, has established research programs working on using Big Data. (Keeso, 2014).

Ecology is also moving towards the use of Big Data, with institutions such as the National Center for Ecological Analysis and Synthesis rising into the top 1% of ecology institutions worldwide in terms of scientific impact (Hampton, 2013). This shows that Big Data has already revolutionized this scientific field.

A novel proposition in this domain is the usage of Big Data to predict transport carbon emissions. Transport emissions are the leading source of greenhouse gas emissions. This project considers modelling carbon emissions in a grid based on factors such as time of day and weather, achieving a prediction accuracy of over 90%. This research could be extremely useful when considering city planning and road networks (Lu et al, 2017)

Legal acts are already in place which govern big data, such as the EU General Data Protection Regulation, which governs not only direct identifiers but also indirect identifiers, such as IP addresses. Also, in the case of opt-in data sharing, the data subjects have often given permission for their data to be gathered for a different reason than the big data utilization of it, meaning that in the case of big data, anonymization should be pursued (Gruschka, 2018). Ethical use of Big Data must also be considered, with people often being unaware of the data they generate that is being collected, and often not being aware of where data they are shown has originated from. This leads to a power imbalance; online consumers see causation implied by big data and do not know the true underlying reason for the trend, whereas the data collector does. Anonymization of data is another ethical concern, as data cannot be stripped of all features relating to a certain group without stripping it of its value (Zwitter, 2014).

To conclude, the wide array of data gathering and prediction techniques employed today could have huge positive societal impact, but it is also likely that issues will arise with how the data is being used. That is to say, the data must be used responsibly and impartially, else it can easily lead to giving more societal influence to those who collect and use it. Although acts are in place to protect individuals in a dataset, more work must be done to ensure the protection of our entire society’s wellbeing, in an age where misrepresentation of data can cause catastrophic effect.

### *References*

Giest, S. (2017) ‘Big data analytics for mitigating carbon emissions in Smart Cities: Opportunities and challenges’, *European Planning Studies*, 25(6), pp. 941–957. doi:10.1080/09654313.2017.1294149.

Batty, M. (2013) ‘Big Data, Smart Cities and city planning’, *Dialogues in Human Geography*, 3(3), pp. 274–279. doi:10.1177/2043820613513390.

Keeso, A. (2014) Big Data and Environmental Sustainability: A conversation starter. Oxford: Smith School of Enterprise and the Environment, University of Oxford.

Dubey, R. et al. (2019) “Can big data and predictive analytics improve social and environmental sustainability?,” Technological Forecasting & Social Change, 144, pp. 534–545. Available at: <https://doi.org/10.1016/j.techfore.2017.06.020>.

Hampton, S.E. et al. (2013) “Big data and the future of ecology,” Frontiers in Ecology and the Environment, 11(3), pp. 156–162.

Lu, X. et al. (2017) ‘Predicting transportation carbon emission with Urban Big Data’, IEEE Transactions on Sustainable Computing, 2(4), pp. 333–344. doi:10.1109/tsusc.2017.2728805.

Zwitter, A. (2014) ‘Big Data Ethics’, *Big Data &amp; Society*, 1(2), p. 205395171455925. doi:10.1177/2053951714559253.

Gruschka, N. *et al.* (2018) ‘Privacy issues and data protection in Big Data: A case study analysis under GDPR’, *2018 IEEE International Conference on Big Data (Big Data)* [Preprint]. doi:10.1109/bigdata.2018.8622621.

# Section 3: High Performance Computing

### *Chapter 1: Compiling And Running the Implementations*

CPU used: Intel(R) Core(TM) i9-10900K CPU @ 3.70GHz with 10 cores.

Serial Implementation:

‘gcc md.c -lm’ is used to compile the serial implementation. ‘-lm’ is needed for access to math libraries used in the code.

./a.out runs the code, which will model the 20000 gas particles over 10 timesteps.

The Serial Implementation took 108.709 seconds to initialize and complete all 10 timesteps. This time was taken from the time given by the solution in the terminal. The result was a center of mass equal to (-0.09509,-0.16562,49.64602).

|  |  |
| --- | --- |
| Num of Cores | Time Taken (seconds) |
| N/A | 108.709 |

MPI Implementation:

‘mpicc md-MPI.c -lm’ is used to compile the MPI implementation, with ‘mpirun -np 1 ./a.out’ being used to run the code on 1 core, while ‘mpirun -np 10 ./a.out’ would be used to run the code on 10 cores. This implementation was run between 1 and 10 cores, with some being omitted for reasons that will be discussed in the next chapter. Time was taken from the time given in the Terminal, measured by the solution.

Result: Centre of mass = (-0.09509, -0.16562,49.64602)

|  |  |
| --- | --- |
| Num of Cores | Time Taken (seconds) |
| 1 | 107.834 s |
| 2 | 55.1179 s |
| 3 | ABORTED |
| 4 | 27.0961 s |
| 5 | 23.4573 s |
| 6 | ABORTED |
| 7 | ABORTED |
| 8 | 13.482 s |
| 9 | ABORTED |
| 10 | 12.172 s |

A graph with purple line

Description automatically generated

For clarity, Aborted Processes will appear as 0 time taken.

OpenMP Implementation:

‘gcc -fopenmp md-openMP.c -lm’ is used to compile the OpenMP implementation. ‘export OMP\_NUM\_THREADS=1’ would be used to set the number of cores used to 1. the code can then be run with ‘./a.out’.  This implementation was run between 1 and 10 cores, resulting in Centre of mass = (-0.09509, -0.16562,49.64602). Time was taken from the time given in the Terminal, measured by the solution.

|  |  |
| --- | --- |
| Num of Cores | Time Taken (seconds) |
| 1 | 111.964 s |
| 2 | 54.4214 s |
| 3 | 38.5908 s |
| 4 | 28.4821 s |
| 5 | 22.1399 s |
| 6 | 18.4422 s |
| 7 | 15.7552 s |
| 8 | 14.0551 s |
| 9 | 13.0952 s |
| 10 | 12.1159 s |

A graph with purple line

Description automatically generatedA graph with a purple line

Description automatically generated

### *Chapter 2: Discussing different Implementations*

The OpenMP implementation achieved the fastest score of 12.1159 when run on 10 cores, with Serial being the slowest as it cannot be run on multiple cores.

The MPI solution makes use of MPIGather to retrieve all data across the different cores, which is suited more to multiple clusters each with their own memory, rather than a single PC. This will lead to having greater overhead costs than the OpenMP implementation. MPIGather also requires Number of Particles/ Num of cores to result in an integer, meaning some core counts cannot be used.

The OpenMP solution does not need to retrieve information due to all processes sharing memory, although this may lead to issues with some larger-scale operations as a single system may not have enough memory.

Conclusion

Due to MPI needing Number of Particles/ Num of cores to result in an integer, in many cases, the OpenMP solution will be more appropriate, such as when only one machine is being used, and the machine has enough storage space to handle the simulation. When separate clusters are being used for the simulation, the OpenMP solution would not be suitable due to its lack of data retrieval from the other clusters, and so the MPI solution would be more suitable. One could expect the MPI solution to have better scaling past the point of these tests, due to its ability to utilize multiple clusters and so, access higher performance computers. OpenMP was likely only faster than MPI in these tests due to the overheads of MPI programming and the time taken to gather data from all nodes.

# Appendix

OneDrive Directory containing the Big Data.ipnyb code can be found at: [HPC&BD\_20047539\_Mason\_Bowers](https://stummuac-my.sharepoint.com/:f:/g/personal/20047539_stu_mmu_ac_uk/Ep_9W-soX81DkwVL-BawgUMBLTy8y_pPiSKaeDmVUvB8qQ?e=itaijZ)

It is accessible to Dr Seun Ajao and Dr Michael Bane.