CSE 603:HOMEWORK ASSIGNMENT-3

BY,
SREE HARSHA KONDURI
UBITNAME: sreehars
PERSON #: 50060926

COMPARE THE PERFORMANCE ON INCREASE IN VARIABLES AND OBSERVATIONS KEEPING NODES CONSTANT

THE BELOW GRAPH IS TO DENOTE CONSTANT OBSERVATION SIZE OF 10000, INCREASING VARIABLE SIZE FROM 10 to 50 AND CONSTANT NODES=12.

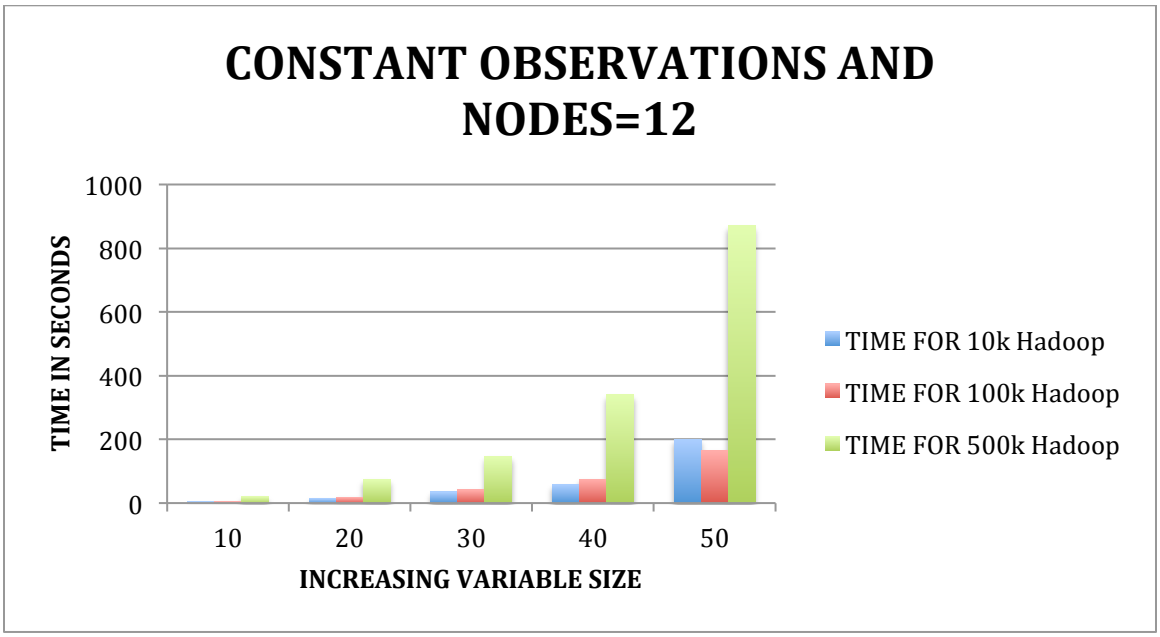

Figure a Compare performance of Hadoop with Constant observations and Nodes

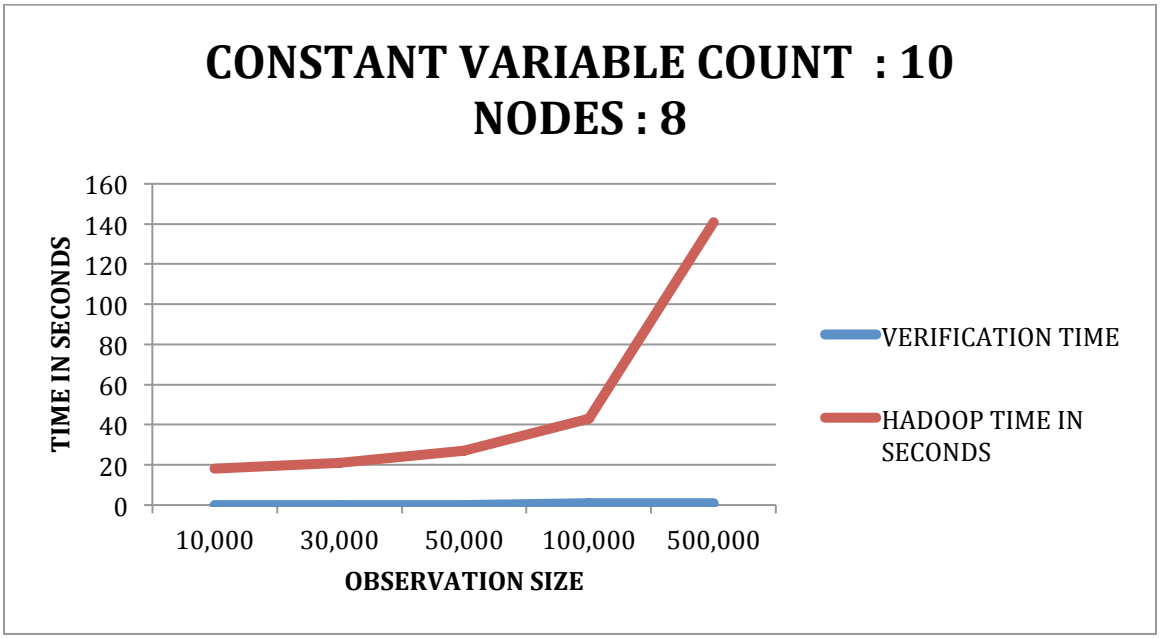

Figure b : Compare the Performance of Hadoop with Constant Variables Constant Nodes and Increasing Observations

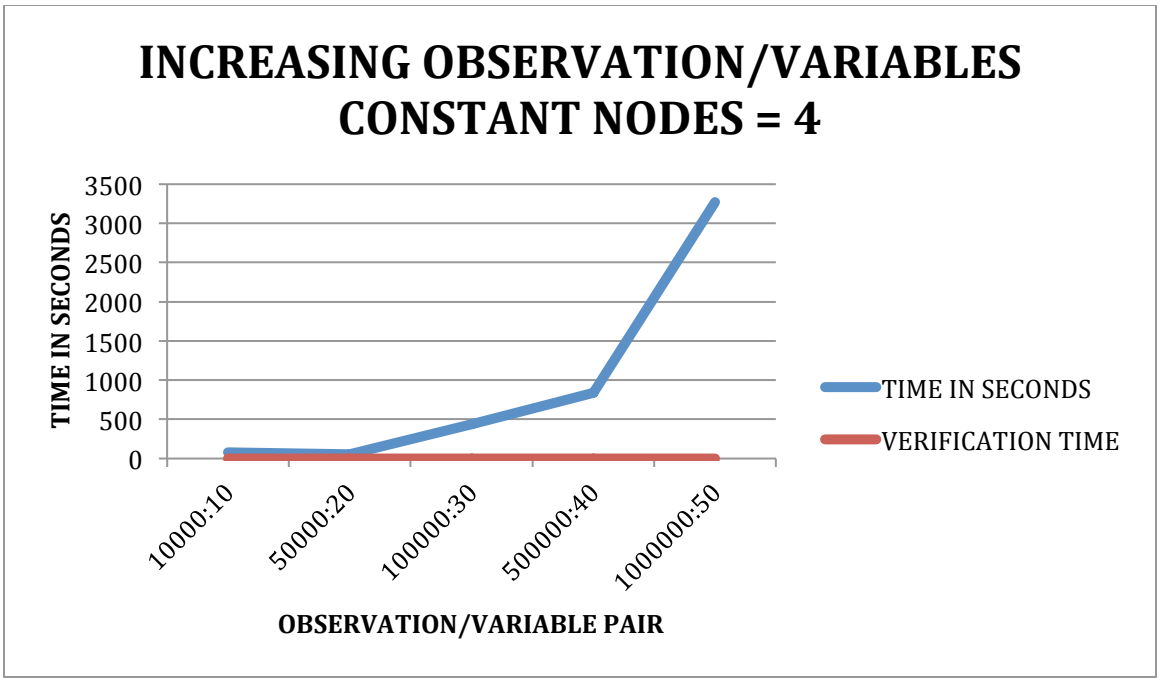

Figure c : Comparing Performance of Increasing Observations and Increasing Variables with Netezza Analytics Library

Please Note that since we cannot perform verification for variables more than 30 I have used 0 for the particular field.

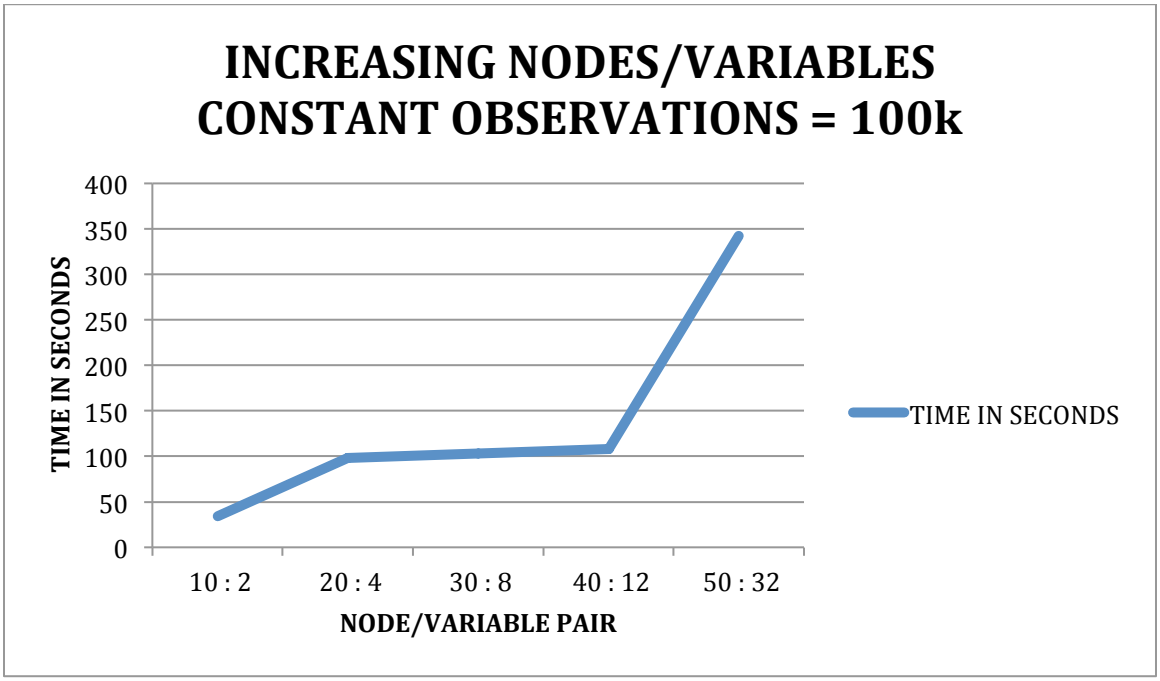

Figure d : Comparing Performance of Increasing Observations and Increasing Variables

Graph Depicting Time For Constant Nodes = 4 and Varying Observations and Variables.

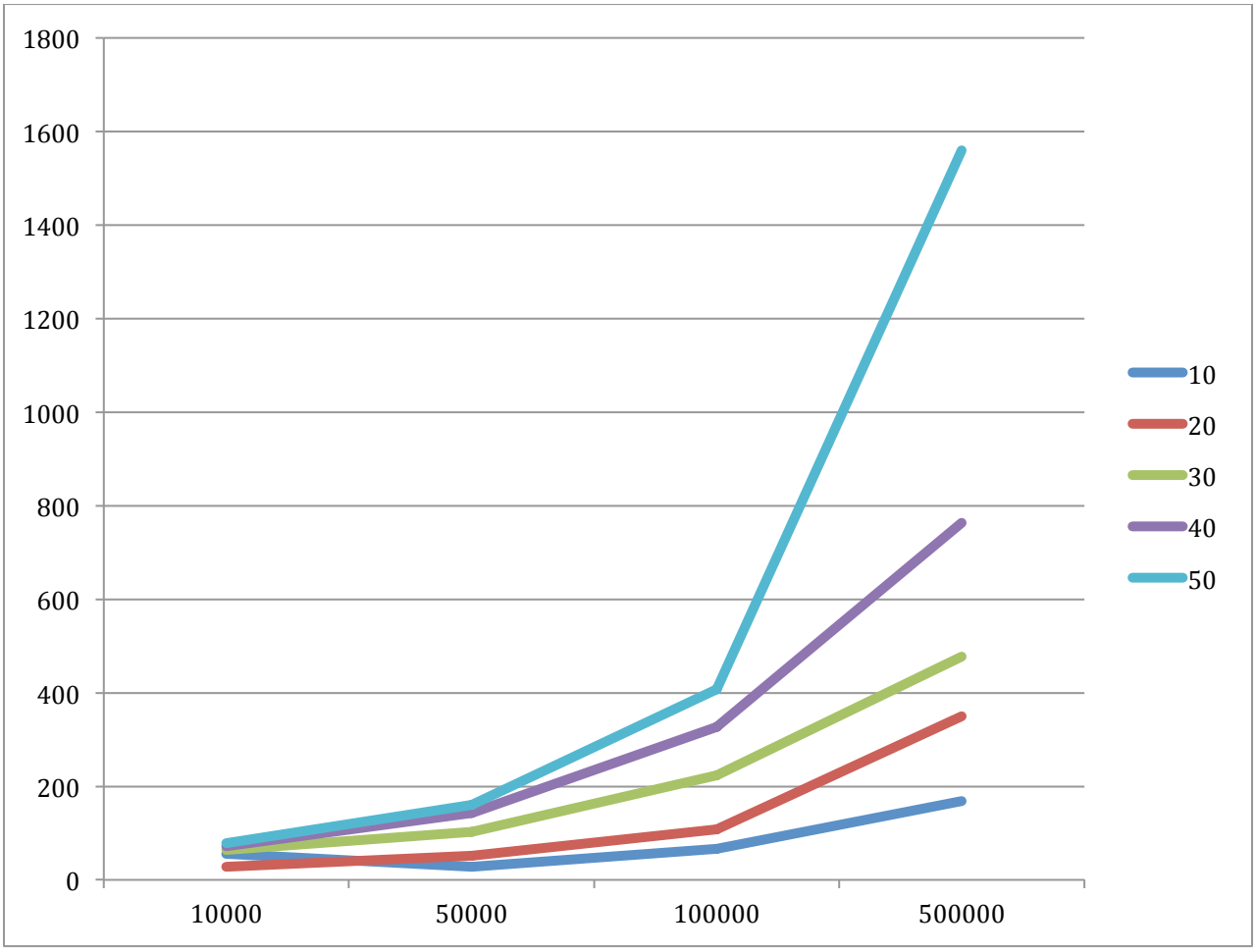

Figure e : Constant Nodes – 4 X:AXIS = Observations, Legend : Variables and Y:AXIS = time

IMPACT IN VARIATION OF PROBLEM SIZE ON PERFORMANCE

Hadoop behaves strangely for small data sets. The time for calculating correlation for observations of size 10000 and 10 variables is larger than then that for observations of 10000 and 20 variables. This can be seen In figure(c). It shows that the system is not proper for smaller data sets.
As the observations are increased with increasing variables and constant nodes we see a linear increase in time after some constant variables ‘n’. This can be associated to the fact that calculating more observations on the same nodes takes more time.

PERFORMANCE BOTTLENECKS OF THE IMPLEMENTATION

One bottleneck is while comparing the time for nz..corr_matrix_agg the number of variables is limited to 31 which prevents us from comparing too many variables.
Also Hadoop has non-computational overhead for calculating correlation for constant observations upto some point.
This is because a constant amount of time is taken to allocate mappers and reducers in hadoop and beyond allocation the jobs run in parallel. So for small data sets the performance of Hadoop is very poor.

COMPARE WITH NETEZZA ANALYTICS FUNCTION

Figure a, b, c, d depict the performance of the NETEZZA ANALYTICS LIBRARY with corresponding configurations. I observed that the Netezza Library function was orders of 10 faster than Hadoop.

REPORT TOP 10 CORRELATION PAIRS

10,000 / 50 variables

Correlation	-----	X	Y
0.0356021562752619		6	28
0.03517046649998187		35	42
0.035125947059918165		5	26
0.03073452467523025		2	7
0.028133319555499327		24	40
0.02712792752347848		0	4
0.024794248543615963		16	47
0.024064369851381218		7	38
0.024046003608758852		42	45
0.023814046522602314		28	34

100,000 / 40 variables

Correlation	------	X	Y
0.009867978254817462		30	37
0.009340066737905987		31	35
0.0085627104081431		0	19
0.00825293120375324		1	8
0.007889178455324403		15	36
0.007570915906834985		25	34
0.007254453330714038		11	18
0.007163871717687819		6	22
0.0068750851384874094		5	33
0.006742968206071961		4	30

10,000 / 30 variables
Correlation ------ X Y
0.02678552365834398 1 3
0.02552061545547558 13 29
0.02537851213276721 5 21
0.025033423446684588 22 29
0.023492397552317915 1 28
0.022016955095549868 16 21
0.021923761231217035 5 18
0.02190896458794473 10 27
0.021422484913052188 5 19
0.021235474487862935 2 6

10,000 / 20 variables
Correlation ------ X Y

0.03061507611465416 6 12
0.025302586440828178 12 14
0.023415085356778574 0 5
0.022671285856259684 14 18
0.020382187363784617 2 12
0.02008648381461696 7 17
0.019553462430603092 0 2
0.0186478376649955 8 9
0.01776975442698409 1 13
0.016928976285959223 13 14

10,000 / 10 variables
Correlation ------ X Y

0.018920782425998293 4 9
0.014307428027397292 5 6
0.012852657804846054 1 3
0.01062332579726599 8 9
0.010602786304902884 0 9
0.009587657108939952 4 8
0.00920348300624569 1 9
0.00715006380991591 1 5
0.006855102998199534 0 5
0.006360215655702815 7 9

CONCLUSION:

I have observed that the time taken for Hadoop is very large due to the overhead incurred due to allocating and deallocating nodes. Upto around 100,000 Hadoop runs slower than sequential Correlation of Matricies. After that Hadoop slows grows faster than the sequential correlation but is still less than both our Implementation of Correlation in Netezza and the Netezza Analytics Library.