International Journal of Control Theory and Applications

ISSN : 0974-5572

Volume 10 • Number 4 • 2017

Performance and Data Analysis on Multi-core System for a Big Data R Application

Chetna Dabas[1], Aniket Dabas[2] and J. P. Gupta[3]

[1] *Computer Science and Engineering and Information Technology, Jaypee Institute of Information Technology, Noida, India*
E-mail: chetna.dabas@jiit.ac.in
[2] *Intelligent Networks, Ericsson India Global Services Pvt. Ltd., Noida, India*
[3] *Ex Chancellor, Lingya's University*

Abstract: In the present era of Multi-Core technology, due to the evolutions of multi core designs, it is getting complex to estimate and conceptualize the intensity of difficulty the Data Science and Big Data Communities are about to encounter. Further, Big Data Analytics is still being explored on Multi-Core systems. This research paper aims to address the above stated issue on a multi-core platform. Performance comparison of one such big data application namely LinkedIn Connections application in R language on a multi-core system has been carried out in this research work. This research paper presents the results and analysis of the proposed big data application (in Ri386 3.2.1) along with the performance analysis and results of the proposed big data application by varying the number of cores on the multi-core machine. The comparison analysis is carried out primarily in terms of CPU usage and memory usage apart from eleven other parameters which plays a crucial role in evaluating the system –application performance. This kind of experimentation and analysis may play a vital role in driving the technology and market trends of an organization or company.

Keywords: Big Data; Multi-core; Analysis; Performance

I. INTRODUCTION

Big data analytics equips organizations and data scientists to play a vital role in data architecture, data acquisition, data analysis and archiving. This technology also enables to analyze a mix of unstructured, semi-structured and structured data in the tunnel of mining precious business information and insights. On the other hand, single core era has almost come to an end and it's all about multi-core processing now a days.

Multi-core processing equips a user with the capability to execute processes in a very less span of time as compared with the serial versions of the processors. If any application makes use of both the technologies for a more lucrative purpose in terms of analysis or performance, then issues also gets vigorous [1-3].

To start with the basics of the proposed work, firstly it is of crucial importance to understand the Big Data Analysis model and the multi-core architecture before presenting the proposed work. The next section throes light on the same.

II. SETTING THE STAGE: BIG DATA ANALYSIS MODEL AND MULTI-CORE ARCHITECHTURE

The prime challenges posed by Big Data Analysis are its high dimensionality, computational complexity, visualization and real and distributed computation apart from many others. The prime challenges associated with multi-core systems are allocation of work to different processors or cores in order to achieve an optimized throughput. If there is a big data application it has to undergo the challenges posed by both the technologies [1-6].

(A) Data Analysis Model

The data analysis in engineering is primarily composed of accessing of data, transforming it for some specific purpose, visualizing it (especially the big data) for analyzing it for meaningful purposes, modeling it and finally analyzing the data.

Apart from the flow that exists between various components of the Data Analysis Model presented in Fig 1, there is a consistent independent feedback mechanism which exists back and forth amongst two subsequent modules [1, 2, 3].

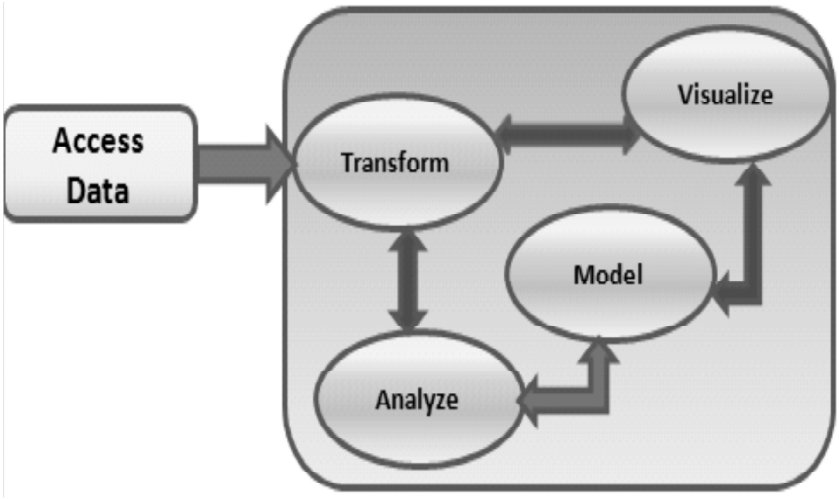

Figure 1: Data Analysis Model

This bears a critical problem of interdependence between different modules in the analysis phase of big data catering to any application. The data set utilized in the proposed work underwent through these phases as a part of analysis that was carried out consisting up of data frames consisting up of 95 observations of 59 variables.

(B) Multi-core Architecture

In layman terms, the Multi-core architecture is multiple cores burnt on a single chip. In technical terms and at an abstract level, it consists up of four layers namely the bus interface at the lowest level, shared memory working and interacting on its top, Individual memories (as many as the number of cores) associated with the shared memory on still higher level. The bus interface directly communicates with the off-chip components.

Since, the experimentation in the proposed work consumed 4 logical cores on the multi-core system; four-core architecture is displayed in Fig. 2.

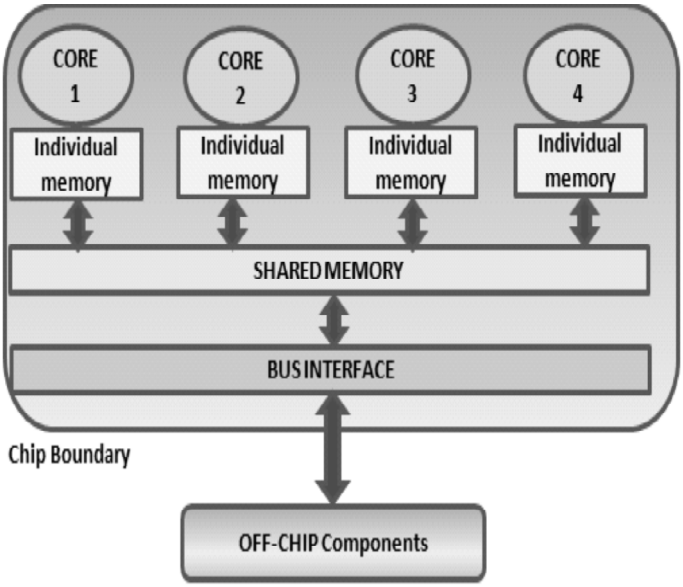

Figure 2: Multi-core Architecture

(C) Multi-core System Performance Metrics

The performance of a multi-core system is evaluated by the two basic laws namely Amdahl's Law and Gustafson's law as described below:

- Amdahl's law is widely accepted and use in evaluating multi-core processor performance analysis. Amdahl's law states the expected speedup of an algorithm via parallelization in relationship with the section of the algorithm which is serial versus parallel. If the proportion of parallel to serial execution is higher, then the chances of speedup are higher as the number of processor (core) increases [1]. The speed up is computed by:

$$\frac{1}{\left((1-f)+\left(\frac{f}{n}\right)\right)} \tag{1}$$

where: f: fraction of the program that is infinitely parallelizable and n is the number of processor or cores

Here, the LinkedIn R application has been designed to work on parallel data sets of the LinkedIn connection s and the application.

- Gustafson's law is stated for massively parallel architectures for large data sets. The speedup is represented by:

$$\frac{s'+(p'x\,n)}{1} \tag{2}$$

where: s' is the serial time spent on the parallel system, p' is the parallel time spent on the parallel system and n is the number of processors [4-6].

The next section presents the experimental work as a part of the proposed research work.

III. EXPERIMENTAL WORK

In the proposed work, data and performance analysis of a LinkedIn R application has been designed and executed on Ri386 version 3.2.1 and a multi-core platform with four logical cores. The data analysis has been carried out with data frame of 95 observations on 59 variables. Factors were utilized for factor label information for annotation. Logistics were also part of the composed and complex data set apart from other parameters. The performance analysis metrics were primarily the CPU usage and memory usage apart from eleven other parameters stated during the course of the proposed work.

(A) Experimental Environment Specifications

The execution of the proposed work was performed on R(i386 3.2.1), 32-bit and windows 7 professional (32-bit) . The experimental work in the proposed research was carried out on Intel Core i3 2100CPU @3.10 GHz system. The installed memory on this multi-core system was 2 GB out of which only 1.88GB was usable. The graphs were plotted while executing the LinkedIn application in R and the comparison experimental work based on 13 parameters was plotted in Microsoft Excel on the multi-core system with above mentioned specifications. Several R packages were installed in order to execute the desired proposed application on R.

chetna.csv file was used to extract the LinkedIn connections and their detailed description consisting up of 94 observations of 59 variables and for performing analysis on the application script designed in R. This file contained the LinkedIn connections data from the profile of the author. The data frame of the csv file of the R application consisted up 59 variables like First.Name, Company, Job.Title, Mobile.Phone, E.mail.Address etc. The factors of the data frame were consumed for the factor label information for annotation while execution. Various R-scripts were designed as a part of the proposed R LinkedIn connections application and executed.

The Results and Analysis of the proposed work is presented in the next section ahead.

IV. RESULT AND ANALYSIS

This section is comprised up of two prime subsections corresponding to R Application execution and analysis and Multi-core performance analysis for R LinkedIn Application. Up-next is the subsection on R LinkedIn Application Execution and Analysis.

(A) R LinkedIn Application Execution and Analysis

As a part of the experimentation results, a snapshot of output in the form of how many known connections are presently recruited by which companies is depicted in Fig. 1.

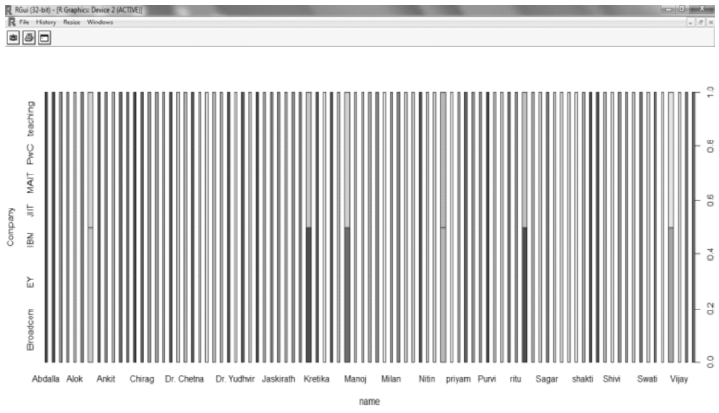

Figure 3: Snapshot of output: Employee Name (connection) Vs Company name (in LinkedIn profile)

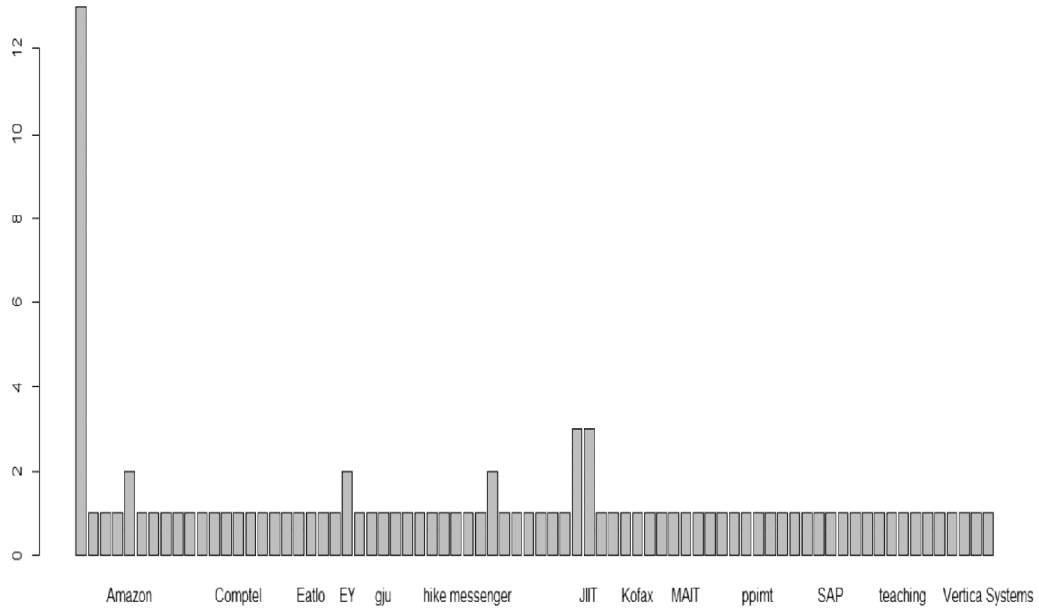

Figure 4: Snapshot of output: Company Vs Number of connections in (LinkedIn profile)

Fig. 3 depicts which connections (people) are hired by which company or organization.

As per this analysis, amongst the live LinkedIn connections of the author, 13 connections were recruited by Jaypee Institute of Information Technology, 3 by JIIT (alias Jaypee Institute of Information Technology), 3 by Amazon, 2 by EY, 2 by IBM and 70 by Others(other companies or organizations) as depicted in Fig 4 .

Further, as a part of analysis it was observed that designation 13 connections were at Assistant Professor Designation, 4 software engineers, 2 associate professors, 2 consultants, 2 directors and 69 others.

Fig. 5 presented ahead, is a snapshot of Line Plot in R for each person (connection) on LinkedIn profile Vs Job Titles which depicts how many connections (persons) are on the same designation presently.

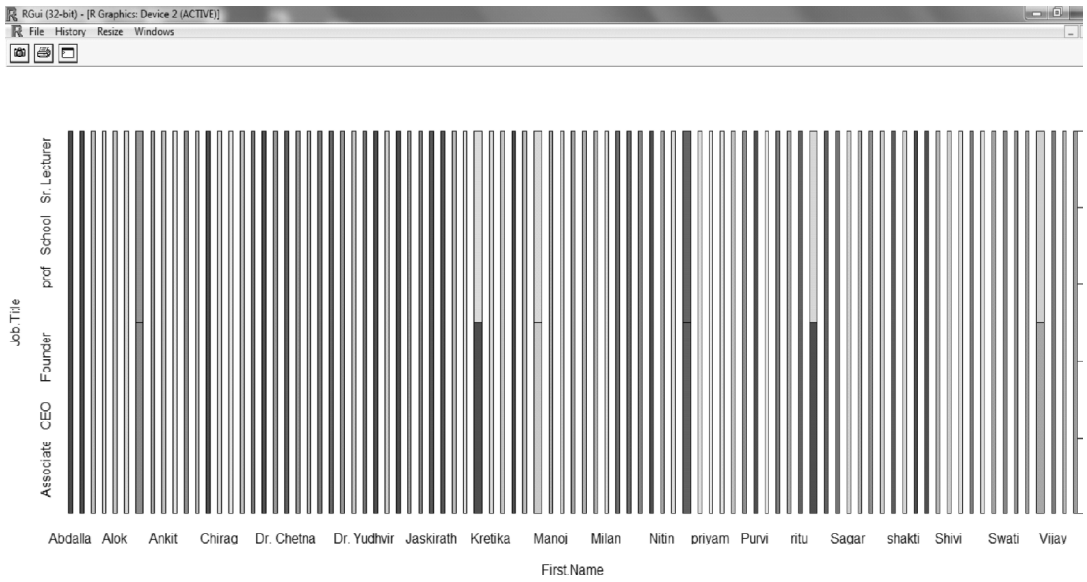

Figure 5: Snapshot of Line Plot in R for each person (connection/Employee) on LinkedIn profile Vs Job Titles

The result plots from Fig.3, Fig. 4 and Fig. 5 represents a snapshot of the data part, since there were 95 observations of 59 variables, the entire plot was too huge to be plotted on one screen. This kind of study and analysis may play a vital role in driving the influencing trends of an organization or company.

Next subsection takes up the journey of the R LinkedIn application on multi-core system in terms of performance analysis.

(B) Multi-core Performance Analysis for R LinkedIn Application

The Multi-core performance analysis for the R LinkedIn Application was carried out in five phases as described in the subsections below.

1) Case 1: The system performance when neither the Ri386 3.2.1 was launched on the multi-core system under consideration and hence nor the LinkedIn R application was started is Case 1.

The performance graph in this scenario in terms of parameters like available physical memory (both paged and non-paged), CPU utilization, number of processes, number of threads, number of handles etc is presented.

Here the processor affinity was initialed as CPU0, CPU1, CPU2 and CPU3. Apart from other observations that will be discussed in the last subsection of this section, it was observed that the CPU usage turned out to be 0%. The snapshot of the performance of this scenario is presented in Fig. 6.

1) Case 2: Fig. 7 depicts Case 2. Here Ri386 3.2.1 was up and running on the multi-core system with above presented technical specifications. But, the LinkedIn R application was not yet started. Here, the logical cores chosen on the multi-core machine were CPU0 and CPU 1 only. The performance graph in Case 2 as well is represented in terms of similar parameters as in Case 1 above. In this case, the CPU usage turned out to be 1% only.

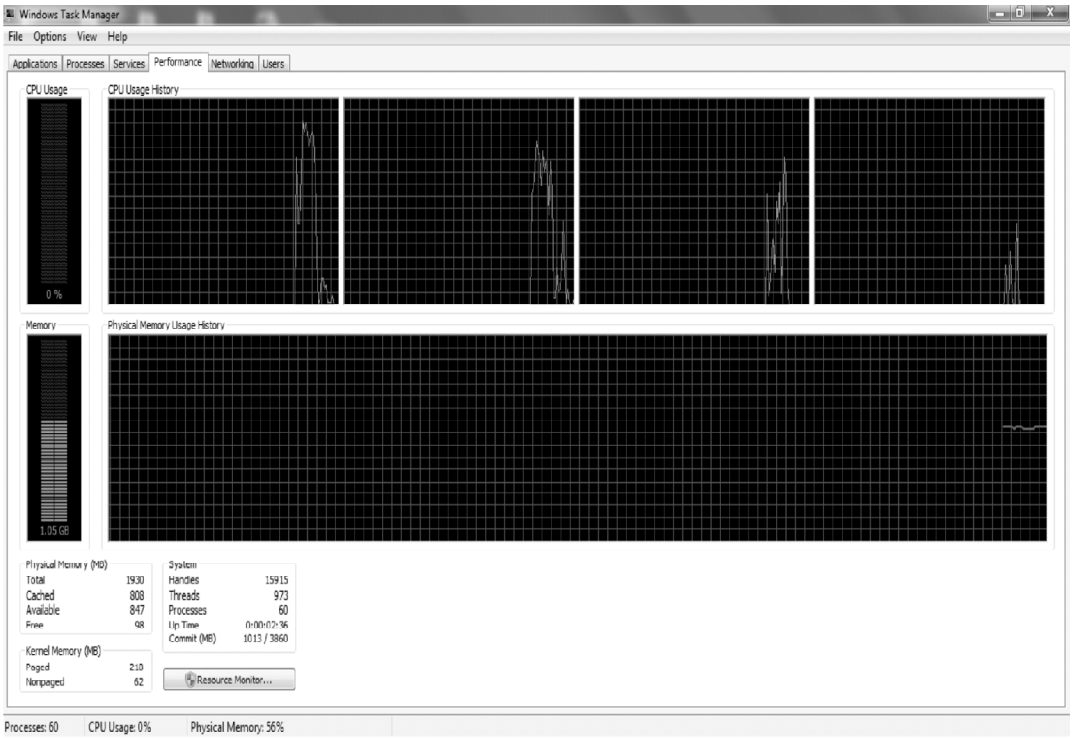

Figure 6: (Case1) Snapshot of system performance without R i386 3.2.1(CPU0, CPU1, CPU2, CPU3)

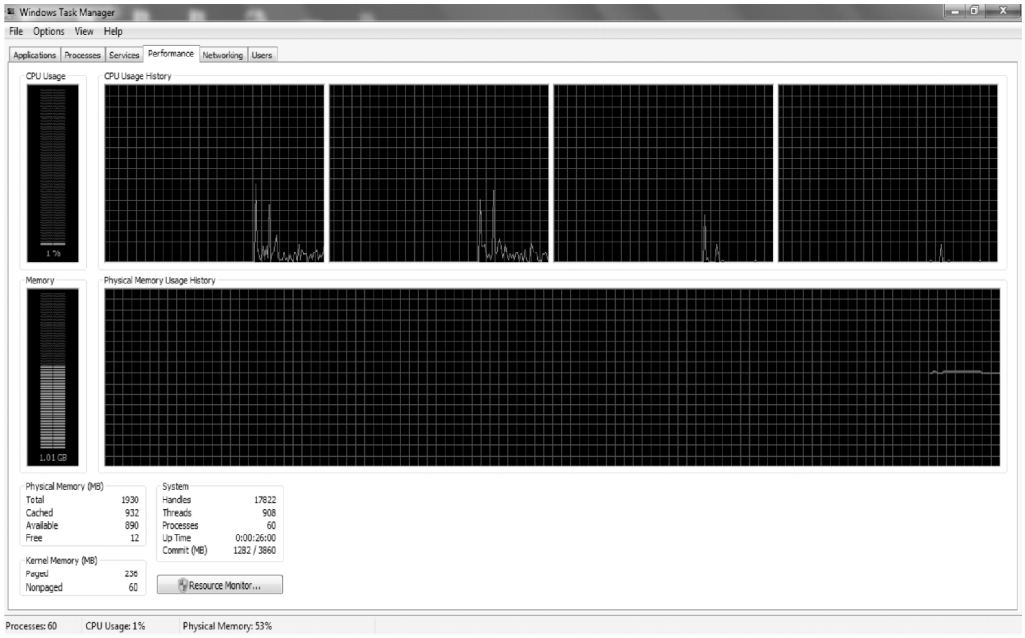

Figure 7: (Case2) Snapshot of system performance while running R i386 3.2.1 (CPU0 and CPU1)

2) Case 3: This Case is depicted in Fig. 8 and represents the situation when both the Ri386 3.2.1 and the LinkedIn R application were up and running. The snapshot of Case 3 with CPU0 and CPU1 (only) with the similar performance parameters as mentioned in the above two cases is presented in Fig. 8. Again, the CPU usage was 1% only.

3) Case 4: Case 4 utilized all four logical cores of the multi-core system under consideration i.e CPU0, CPU1, CPU2 and CPU3. Here, both the Ri386 3.2.1 and LinkedIn R application were started and running.

Apart from other observations that will be discussed in the last subsection of this section, it was observed that the CPU usage rose up to 11%. The snapshot of the performance is captured and depicted in Fig. 9. Here the processor affinity was set to CPU0, CPU1, CPU2 and CPU3.

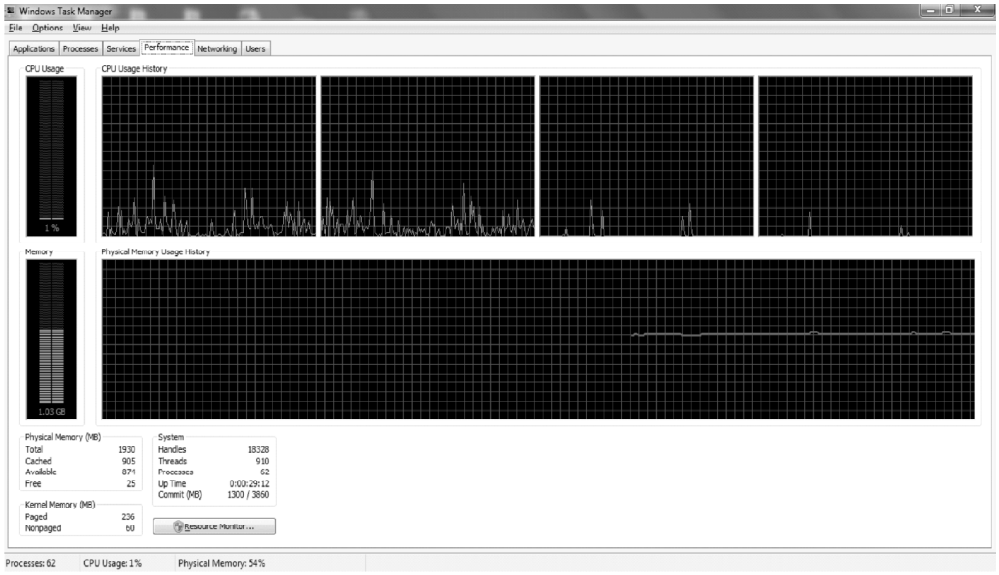

Figure 8: (Case 3) Snapshot of system performance while LinkedIn R Application running (CPU0 and CPU1)

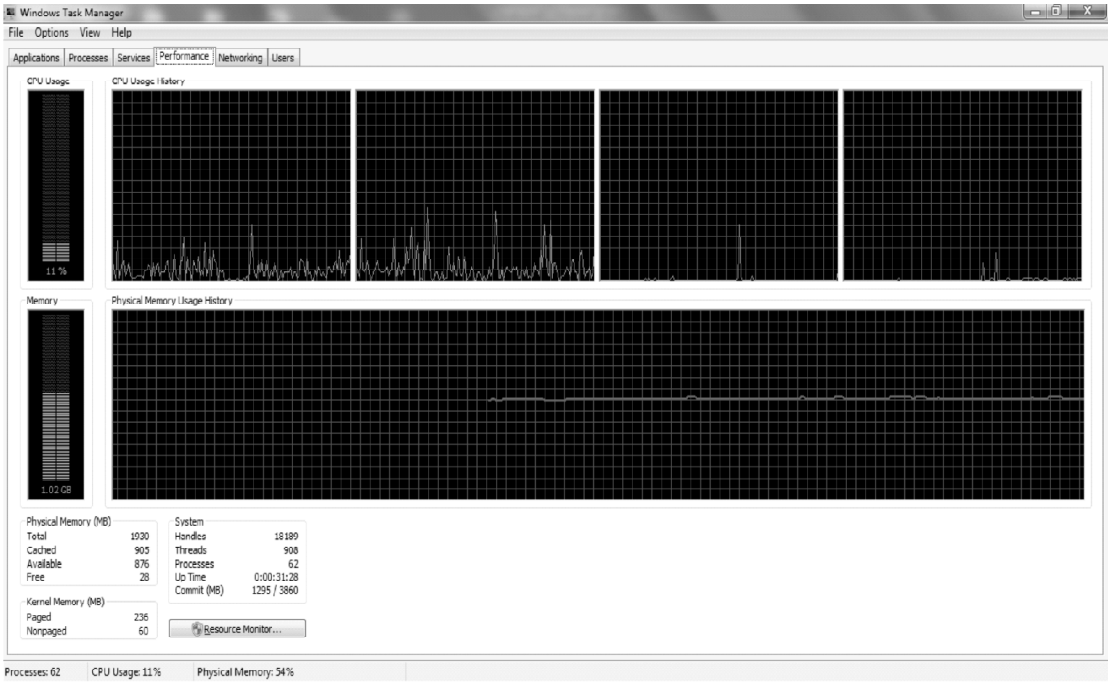

Figure 9: (Case 4)Snapshot of system performance while LinkedIn R Application running (CPU0, CPU1, CPU2, CPU3)

Next subsection is related to Case 5 where apart from the Ri386 version 3.2.1 and the LinkedIn R application, other applications were also launched in order to seek a comparison.

4) Case 5: Fig. 10 represents Case 5, with a snapshot of CPU Performance captured along with RGUI (32-bit), Facebook, NIVIDIA Visual Profiler, Matlab (version R2013a), MS-Word , Calculator, Paint, Internet Explorer, chetna.csv file are up and running on the multi-core system under work.

Figure 10: (Case 5): Snapshot of CPU Performance [RGUI(32-bit), Matlab (version R2013a), NIVIDIA Visual Profiler, Facebook, MS-word, Calculator, Paint, internet explorer, chetna.csv file] [CPU0,CPU1,CPU2 and CPU3]

All four logical cores were assigned tasks in this scenario. It was observed as a part of this study that the CPU usage went up to 14 % and free memory becomes zero.

Apart from these individual scenario observations, overall performance comparison and analysis of the multi-core system while running and executing the LinkedIn R application was performed as a part of the proposed research work. The same has been discussed in the upcoming subsection of this work.

(C) Overall Performance Comparisons and Analysis

The LinkedIn R application was executed and the results were tabulated in Table I, Table II and Table III. Table I is filled with performance data corresponding to all five cases as considered in the above subsection.

The number of cores on the multi-core system under consideration wee varied in order to analyze the performance better. Results tabulated in Table I reveal that the execution and performance of LinkedIn R application (with data frames of 95 observations on 59 variables) gives an improved CPU usage when all four cores of the multi-core system are on work and CPU usage is a little less when only two cores are being used.

Table I
Performance Analysis for LinkedIn R Application on Multi-core System CPU Usage

Cases	*No of Cores*	*CPU Usage (%)*
Case1	4	0
Case 2	2	1
Case 3	2	1
Case 4	4	11
Case 5	4	14

Results tabulated in Table II reveals the memory usage by the LinkedIn R application on the multi-core system by varying the number of cores. Here, it was observed as a part of this study that the execution and performance of the LinkedIn R application (with data frames of 95 observations on 59 variables) gives an improved memory usage when more cores are used as compared with less number of cores being used to perform the same task.

Table II
Performance Analysis for LinkedIn R Application on Multi-core System Physical Memory Usage

Cases	*No of Cores*	*Physical Memory Usage (%)*
Case1	4	56
Case 2	2	53
Case 3	2	54
Case 4	4	54
Case 5	4	75

Table III is composed of results tabulated for all the five cases discussed in the above few subsections based on technical performance parameters like number of cores (logical), CPU usage, memory usage, handles, threads, processes, paged kernel memory, non-paged kernel memory, physical memory, cached physical memory, free memory and physical memory usage.

Table III
Performance Analysis Table for LinkedIn R Application on Multi-core with various parameters

S. No.	*Cases*	*Cores (Logical)*	*CPU Usage (%)*	*Memory Usage (GB)*	*Handles*	*Threads*	*Processes*	*Paged Kernel memory (MB)*	*Non-paged Kernel memory (MB)*	*Physical Memory (Total) (MB)*	*Cached Physical Memory*	*Available Physical Memory*	*Free Physical Memory*	*Physical Memory Usage (%)*
1	Case 1	4	0	1.05	15915	973	60	218	62	1930	808	847	98	56
2	Case 2	2	1	1.02	17822	908	60	236	60	1930	932	890	12	53
3	Case 3	2	1	1.03	18328	910	62	236	60	1930	905	874	25	54
4	Case 4	4	11	1.02	18189	908	62	236	60	1930	905	876	28	54
5	Case 5	4	14	1.42	22589	1078	72	247	61	1930	487	465	0	75

Here, it was observed as a part the proposed work that the free physical memory kept on decreasing as there were more applications and utility software was getting running and executed. The free physical memory went from 98(MB) to 0(MB) from execution of Case 1 to Case 0. Also, the cached physical memory was 808 while in

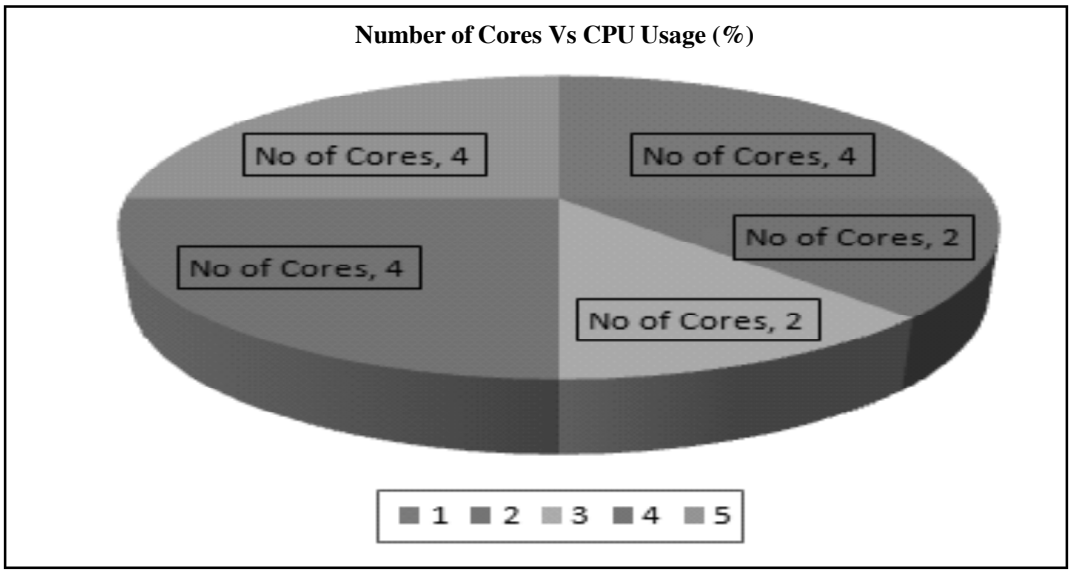

Figure 11: Plot of Number of Cores Vs CPU Usage

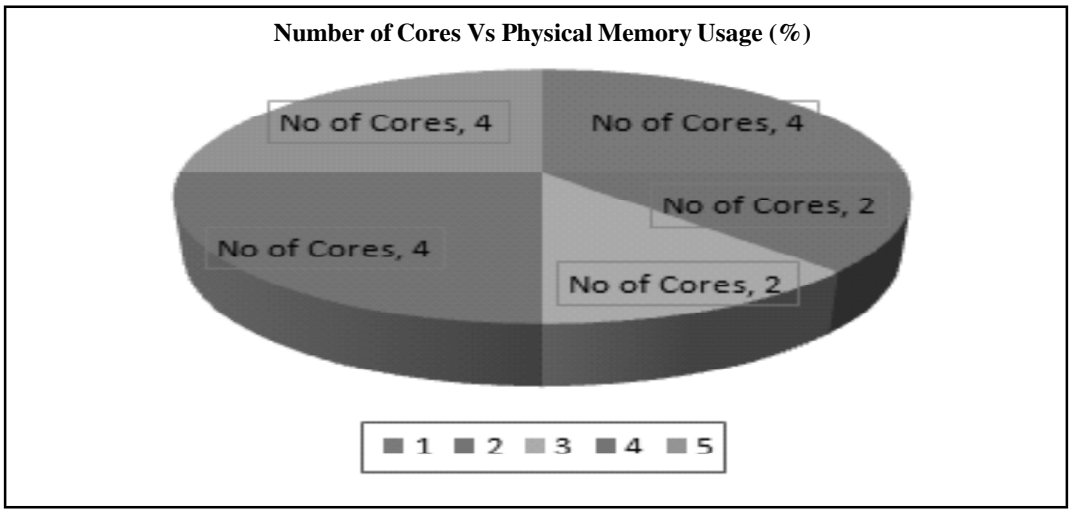

Figure 12: Plot of Number of Cores Vs Memory Usage

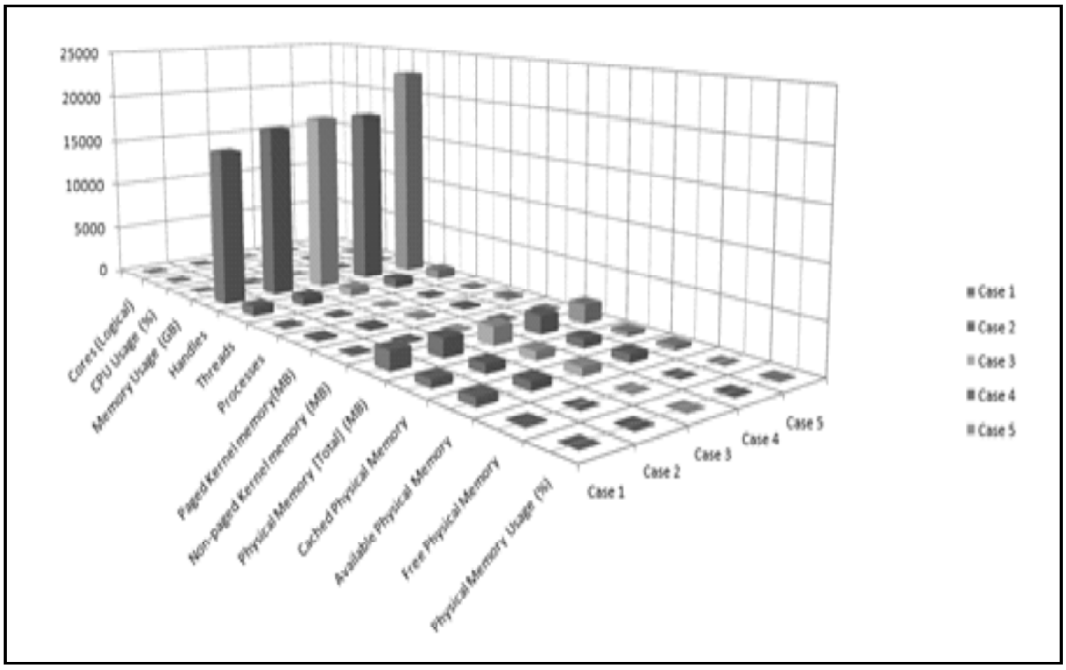

Figure 13: Performance Analysis Plot for Linked in R Application on Multi-core System containing all parameters

Case 1 where as it was 487 in Case 5. The CPU usage took off at 0% and was 14% in Case 5, 11% in Case 4. It was also noted that the number of threads and processes were increased in number while executing the LinkedIn R application on multi-core machine. The physical memory usage was also increased in due course as the work progressed from Case 1 to Case 4 and then towards Case 5.

Fig 11 presents a 3D plot of number of cores of the multi-core system versus CPU usage for all five cases as per the results tabulated in Table I. Fig. 12 depicts a 3D plot of the number of cores of the multi-core system versus Memory usage corresponding to all five case discussed above, in accordance with the results tabulated in Table II.

Since a picture is worth thousand words, the results tabulated in Table III corresponding to all five cases discussed have been transformed into a 3D bar plot in Fig.13 where the results are crystal clear.

CONCLUSIONS

In this paper namely, Data and Performance Analysis of Big Data R application on multi-core system, the data and performance analysis of a LinkedIn R application has been carried out for 95 observations of 59 variables in the data frame consisting up of factor label information for annotation apart from other components.

The performance analysis for the LinkedIn R application has been performed on a multi-core system with four logical cores. The number of cores has been varied during the analysis portion of the work. The technical parameters consumed for this study are primarily CPU usage and Memory usage apart from 11 other parameters in order to evaluate the performance. Results are tabulated and plotted corresponding to the same. It was concluded that the CPU usage was improved for the LinkedIn R application when the number of cores used were more as compared to its performance when the number of cores used were less, which is a good sign. The physical memory usage was also high with the execution of the LinkedIn R application. The proposed LinkedIn R application may be further be utilized for making more intense analysis of the big data, or better purposes and igniting new industry trends.

REFERENCES

[1] Jacques Bughin. Big data, Big Bang? Journal of Big Data, Springer (2016) 3: 2.

[2] Chun-Wei Tsai, Chin-Feng lai,Han-Cheih Chao and Athanasios V.Vasilakos, Big data Analytics: A Survey, Journal of Big Data(2015) 2: 21.

[3] Chetna Dabas, Big Data Analytics for Exploratory Social Network Analysis,International Journal of Information Technology and Management, Inderscience [In Press] [Listed in Forthcomming Articles Available: *http://www.inderscience.com/info/ingeneral/forthcoming.php?jcode=ijitm*]

[4] Gustafson, J.L., "Reevaluating Amdahl's Law," Comm. ACM, May 1988, pp 532-533.

[5] M. J. Flynn, "Some Computer Organizations and Their Effectiveness," IEEE Transactions on Computers, Vol. C-21, No. 9, September 1972, pp. 948-960.

[6] Rauber, Thomas, Rünger, Gudula, Parallel Programmingfor Multicore and Cluster Systems(2013), ISBN: 978-3-642-37801-0, DOI:10.1007/978-3-642-37801-0.