International Journal of Information Technology and Business Management
29th June 2013. Vol. 14 No.1
© 2012 JITBM & ARF. All rights reserved

ISSN 2304-0777 www.jitbm.com

# THE EFFECT OF THE DATA STORAGE TECHNIQUES ON THE ANALYSIS PERFORMANCE IN THE DATA MINING STUDIES: A SAMPLE APPLICATION WITH WEKA

**Erkan Özhan**[1] **, Erdem Ucar**[2]

Corlu Vocational School, Namik Kemal University, Turkey[1]
Department of Computer Engineering, Trakya University, Turkey[2]
erkanozhan@gmail.com[1] , erdemucar@trakya.edu.tr

## ABSTRACT

*In this study, the data storage techniques which are necessary to perform the data mining applications have been anaysed. In the data mining applications, the data are processed through some stages until the analysis stage. The proper performance of these stages affects both the accuracy of the analysis and the performance. Also, the analyses should be carried out the with the appropriate hardware. The effective use of the physical memory and the processor capacity bears a lot of importance in terms of analysis period and cost. In the study, database and file environment, which are two types of the data, which are to be processed through the data mining applications, have been examined. Some analyses have been carried out on a sample data cluster for this operation. It has been searched whether there is a connection between the size of the data cluster and storage environments through a gradual increase in the data analysed. For this purpose, a cluster composed of 25 million data has been put into analysis. Firstly, an analysis for two types storage environments has been performed by increasing these data with the use of Navie Bayes algorithm. The results obtained have been grouped as physical memory and the CPU usage. Secondly, the performance assessment has been examined in relation to time through the record of the analysis periods . The data obtained from the analyses in the database and the file environments have been indicated and evaluated in the final chapter of our study.*

**Keywords:** Data mining, hardware, database, file, performance.

## I. INTRODUCTION

The data applications aim to examine if there is a connection between the data which cannot be related with each other at first sight. The connections which human brain has difficulty in making between the data can be achieved through the data mining applications. The data mining carries the meaning of the analysis of the data of high quantity stored in a computer. Besides this, the mining involves the analyses which benefit from the statistics and artificial intelligence techniques which are generally employed in the large-scaled data sets(Olson & Delen, 2008). It is a demanding and challenging process to find and extract the meaningful data especially from the big data clusters. Another thing that should not be forgotten during this process is the hardware necessity. It is of great importance to determine the minimum hardware needs when working with the big data clusters.

Even though today's computers' capacity and operation power increase continuously, they still fall short in many applications. The need for physical memory and processor limits data mining applications.

The most important point affecting a computer's speed is CPU(Central Processing Unit). The other pieces of the hardware affect the performance; however, the capacity of the CPU proves the most dominant component(Ozguler, 2007). The memory is an important resource which needs being administered carefully. In the present time, though an average home computer has hundred times a bigger memory than the most sophisticated computer at the beginning of the 1960s, today's programs require memory(Tanenbaum, 1992).

It is possible to define the database concept in various ways. They can be defined as the whole of interrelated data brought together to solve a problem, or a collection of useful data organized in a particular way (Celikol, 2007). The database is a data storage technique frequently used in today's enterprises. At present, they are frequently employed for the purpose of finance, marketing, reporting, data collecting and etc. In this study, MySQL has been used as the database administration software. MySQL is a software which is distributed free of charge with open source code.

Relational database management system (RDBMS) is a device which is commonly used in enterprises, research and education fields, and internet search engines, and it is an important factor to have a good database to administer the information sources and to have access to them(Dubois, 2009).

There are a great number of applications available for the data mining. In this study, Weka software, which was developed by Waikato University and is a widely used, has been employed. It is a software which is distributed free of charge with General Public License (GNU). In the Weka interface exists a lot of algorithms to do analysis and Naive Bayes is the algorithm chosen for this study. Classification is conducted using the calculus of probabilities in Naive Bayes algorithm. In this algorithm, the qualifications are independent from each other(Tuncer, 2010).

As for the functional dimension, hardware and software are connected to each other. The softwares reach the results via hardware. The software with the required hardware is believed to have provided the primary condition to produce accurate and rapid outcomes. Therefore, the hardware pieces, especially the physical memory and processors should be sufficient and used effectively.

## II. LITERATURE SURVEY

There are invaluable studies in the data mining techniques in which the data analysis is carried out. The authors(Bradford, 1998) put effort into measuring the performance of the CPU cache memory on 8 different sets of data. In the study carried out on two different types of cache memory, they evaluated the results in the form of kB. Yet, the data analysis has been conducted only in a CPU-based form.

The authors(Tiwari, Jha, & Yadav, 2012) also have examined the algorithms within Weka software and they stated the results obtained. They, later, tested the algorithm's classification success; however, they did not perform a hardware-based test.

In an another study, the researchers(Liu, Pisharath, Liao, Memik, Choudhary, & Dubey, 2004) conducted 8 different performance analyses, 6 of which are classifiers, with the use of a software called MineBech. In these analyses, the algorithm success was tried to be measured, and they obtained important results. Still, a hardware-based test was not carried out.

In another study, the authors(Li, et al., 2007) analysed the data clusters different from each other and they employed a simulation software named SoftSDV. They performed analyses with the help of the simulations on the CPU cache memory, and they seized results of great importance. They, still, did not carried out an analysis related to the data storage techniques.

## III. METHODOLOGY

In this study, mainly two different analyses were conducted. These analyses were classified as data transfer, processing performance, processor's performance, and operation duration and the results were recorded. The analyses were conducted on the same computer. Firstly, 25 million data, in the form of 1.5 million lines and 17 columns, were transferred to MySQL database. On the other hand, all data were divided into 6 files in a way that they would make 250 thousand data in 17 columns with the condition that their content would stay the same.

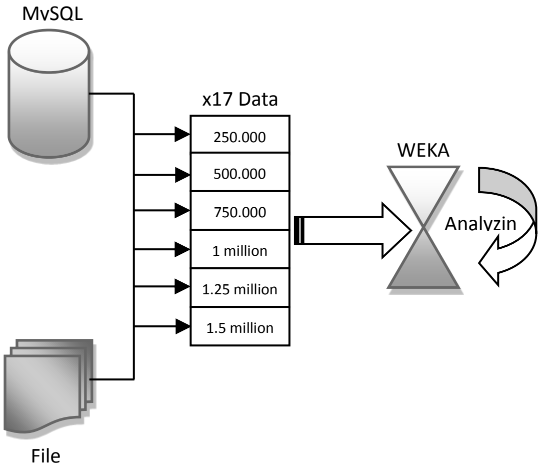


**Fig.1** The stages of the performance analysis

As shown in Figure 1, the tests, firstly, were done for 250 thousand data. At this level, 250 thousand data were imported with the Open DB (Database) application in the Weka software, and the quantity of the physical memory used was measured. The program was shut down and restarted and this time the data were transferred from the file to the Weka software with the use of Open File application. The physical memory used was measured again. This operation was repeated with the increase rate of 250 thousand data until 1.5 million data were obtained .

After the data's import values were determined, the processing stage, the second stage came forward. At this stage, 250 thousand data taken from the database were processed with NavieBayes algorithm. After this process was over, the quantity of the physical memory spent was recorded. Following that, the same data were transferred to the Weka software and were analysed with the NaiveBayes software again and the quantity of the physical memory used during the data analysis duration was recorded. The quantities of the memory spent was recorded as GB and this cycle was repeated until 1.5 data were obtained.

After the analysis of the data, the 3rd stage was realized. At this stage, after the data were taken from the database and the files, the processing time was measured. Firstly, the processing time of the 250 thousand data after it was taken from the database was measured. Then, the processing time of the same data after it was taken a single file was measured and recorded.

After analysis durations were measured, the 4th stage was initialized to determine the CPU use rates. At this stage, the average CPU use was measured while the data were being analysed. Firstly, 250 thousand data were taken from the database and transferred to the Weka software, and when the analysis with the NaiveBayes algorithm was started, The CPU use began to be monitored. The same operation was repeated for the data taken from the file. The assets obtained at the end of the analysis were recorded. During the process until the end of the analysis, the average use of the CPU was determined through the Source Watch software developed by Microsoft.

## IV. RESULTS AND DISCUSSIONS

The first of the results obtained from the tests is the quantity of the physical memory used during the data transfer. These results are shown in Figure 2.

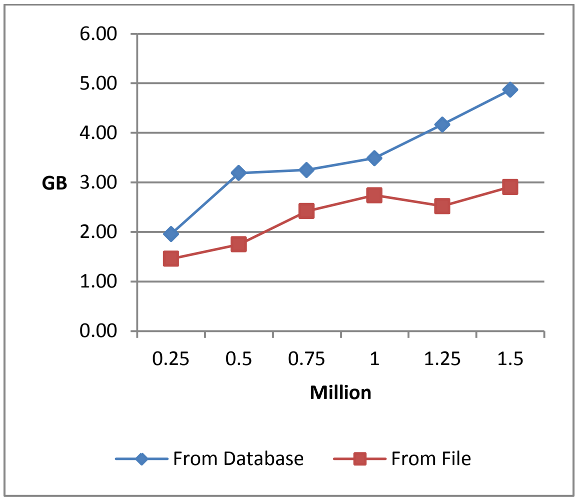


**Fig.2** The quantity of the physical memory used for the data transfer

As seen in Figure 2, the quantity of the memory used during the data transfer from the database is higher than the quantity of the memory used during the data transfer from the file.

Another result obtained from this study is the physical memory levels spent during the data processing stage. At the stage of data processing, NaiveBayes algorithm was used in each test and the quantity of physical memory used was measured when the "analysed completed" message was seen. This situation is shown in Figure 3.

As can be seen from the Figure 3, the quantities used for the analysis are nearly the same for the two types of data storage techniques. Although the quantities of the memory used for the data transfer are different, this situation shows that the quantities of the memory required at the end of the analysis are the same.

The 3rd result obtained from the study is whether two types of data storage techniques affect the time passed for the data analysis. The results are shown in Figure 4.

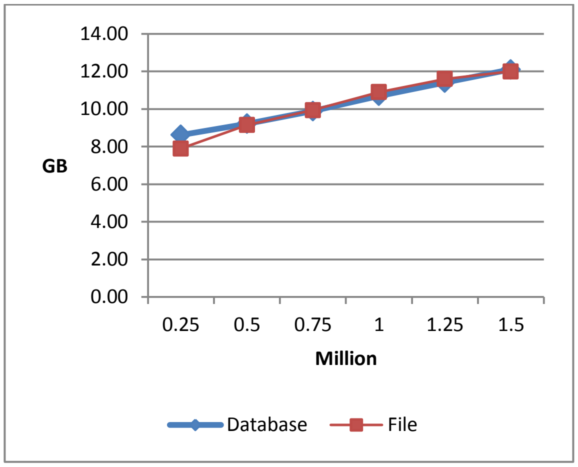


**Fig.3** The physical memory values used at the end of the analysis

As shown in Figure 4, the time passed for the data analysis is nearly the same for the two types of data storage techniques. The fact that the data storage technique is different did not affect the time required for the data analysis.

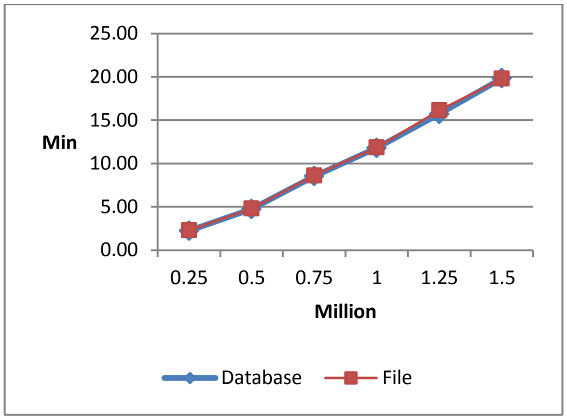
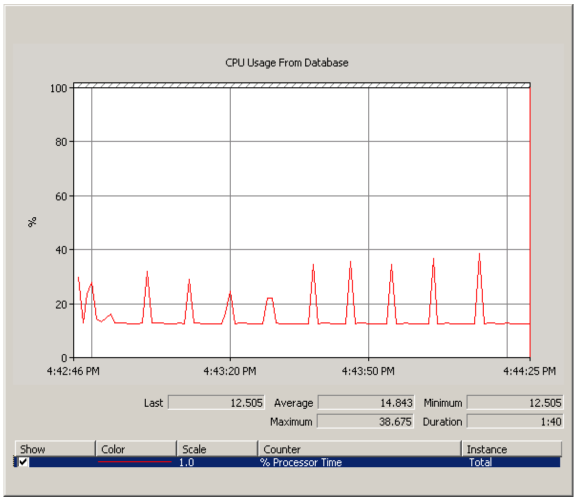


**Fig.4** Analysis durations according to the data storage techniques

Another factor examined in the study is the CPU use rate. The CPU use rate values used by the two types of the data storage techniques in the analysis of the total 1. 5 million data with the increase rate of 250 thousand data are shown in Figure 5 and 6.

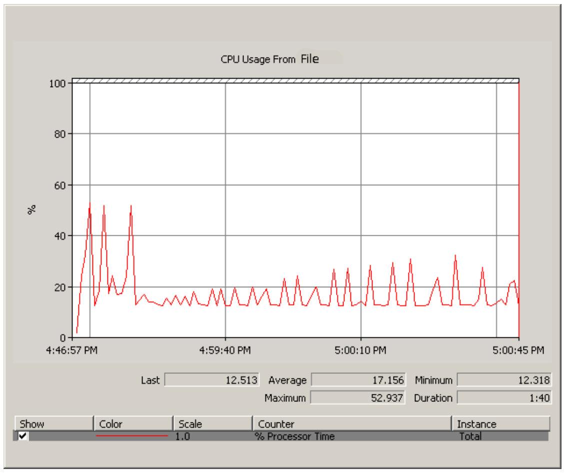


Last 12.505 Average 14.843 Minimum 12.505
Maximum 38.675 Duration 1:40

| Show | Color | Scale | Counter | Instance |
|---|---|---|---|---|
| ☑ | — | 1.0 | % Processor Time | Total |

**Fig.5** CPU values used in the analysis of the data taken from the database

As shown in Figure 5, the average CPU use for the analysis of the data taken from the database is 14.843 %. During the analysis, the maximum rate value is 38.673 % and the minimum value is 12,505 %.



Last 12.513 Average 17.156 Minimum 12.318
Maximum 52.937 Duration 1:40

| Show | Color | Scale | Counter | Instance |
|---|---|---|---|---|
| ☑ | — | 1.0 | % Processor Time | Total |

**Fig.6** CPU values used in the analysis of the data taken from the file

As can be seen from the Figure 6, the average CPU use rate is 17,156 % in the analysis of the data taken from a single file. During the analysis, the maximum rate value is 52,937 %, and the minimum is 12,318 %. When these rates are compared with the rates of the CPU used for the database, it can be seen that there is no difference except for the maximum CPU use rate. However, the CPU use rate average is so close to each other. On the other hand, this situation is not reflected on the analysis as a factor changing the calculation time as shown in Figure 4.

As a result, when the data taken from the different data storage techniques were analysed through Weka application employing different data mining techniques, it was found that the data did not differ except for the data transfer stage. When the data transferred from the database exceed the physical memory threshold, the researchers can solve this problem at the data transfer stage by retransferring the data to the program after transferring the data from the database to a file. Yet, the physical memory overflow may occur during the analysis. The most important advantage of the database use is that it gives the opportunity to analyse the data in segments. In the future, the researchers can reconduct this analysis for the algorithms of different types. To be able to take large quantities of data in a single file, firstly transferring them to a database and then dividing them into segments can prove a solution.

## REFERENCES

[1] Bradford, J. P. (1998). Performance and Memory-Access Characterization of Data Mining Applications. *Workshop Held In Conjunction With The 31st Annual International Symposium On Microarchitecture*, (pp. 91-100).

[2] Celikol, S. (2007). *SQL Veri Tabani Alt Programi* (1 ed.). Trabzon, Turkey: ABP Publications.

[3] Dubois, P. (2009). *MySQL Developer's Library* (4 ed.). Upper Saddle River, NJ, USA: Pearson Education, Inc.

[4] Li, W., Li, E., Jaleel, A., Shan, J., Chen, Y., Wang, Q., et al. (2007). Understanding the Memory Performance of Data-Mining Workloads on Small, Medium, and Large-Scale CMPs Using Hardware-Software Co-simulation. *Performance Analysis of Systems & Software ISPASS 2007 IEEE International Symposium* , (pp. 35-43). San Jose, CA.

[5] Liu, Y., Pisharath, J., Liao, W.-k., Memik, G., Choudhary, A., & Dubey, P. (2004). Performance Evaluation And Characterization Of Scalable Data Mining Algorithms. *Proceedings of IASTED.*

[6] Olson, D. L., & Delen, D. (2008). *Advanced Data Mining Techniques* (1 b.). Berlin, Germany: Springer-Verlag.

[7] Ozguler, M. (2007). *Bilgisayar Donanimi(Computer Hardware)* (9 ed.). Trabzon, Turkey: ABP Publications.

[8] Tanenbaum, A. S. (1992). *Modern Operation Systems* (1 ed.). Amsterdam, Netherlands: Prentice-Hall International Inc.

[9] Tiwari, M., Jha, M. B., & Yadav, O. (2012). Performance analysis of Data Mining algorithms in Weka. *IOSR Journal of Computer Engineering*, (pp. 32-41). India.

[10] Tuncer, T. (2010). *Bilgisayar Aglari Icin Saldiri Tespit Sistemi Tasarimlari ve EPGA Ortaminda Gerceklestirilmesi.* Doktora Tezi, Fırat University, Engineering Faculity, Elazig.