Journal of Engineering and Applied Sciences 13 (1): 137-149, 2018
ISSN: 1816-949X

Mathematical Model for the Research of Systems with Massively Parallel Processing Based on Big Data

[1]Serbin Vassiliy, [1]Moldagulova Aiman, [1]Duisebekova Kulanda, [1]Satybaldiyeva Ryskhan,
[1]Orazbekov Sayatbek, [1]Tursynkulova Akbota, [1]Alimzhanova Laura, [1]Zhuanyshev Ilyas,
[1]Rakhmetulayeva Sabina and [2]Bektemyssova Gulnara
[1]Departament of Information Systems,
[2]Departament of International Information Technology,
International Information Technology University, Almaty, Kazakhstan

Abstract: The right choice of technologies for managing big data banks will solve almost all key tasks of banks. The goal of the study is to make a comparative performance test and crash-test for Greenplum, Netezza, Exadata and Oracle systems on OS AIX based on big data SAS campaign management, also, formalization this processes by mathematical modeling. The 14 SAS campaigns of a large Kazakhstani bank were selected as big data. In the testing took into account the time and load scaling.

Key words: Cluster, supercomputer, SAS campaign, mathematical model, stress testing, process

INTRODUCTION

A comparative study of big data processing performance (Hashem *et al*., 2015) on different software and hardware systems (Watson *et al*., 2015; Loghin *et al*., 2015) is essential for high-performance computing, both in terms of optimising the load on existing machines and the new platform procurement policy (Mihovska *et al*., 2015). The 14 major marketing SAS campaigns of one of the Kazakhstani Bank were selected as research samples. This selection allows a more adequate assessment of computer systems capabilities. Additionally with the help of specialized tools, we analyzed the characteristics of the studied systems of massively parallel architectures such as Greenplum, Netezza, Exadata and Oracle systems.

Big data is able to tackle almost all of the key tasks of the banks such as customer acquisition, improving the quality of services assessment of borrowers, combating fraud and etc. By increasing the speed and quality of reporting, developing depth of analysis big data technologies help banks to meet the requirements of financial regulators.

The data collected in banks are unstructured by their nature. Unstructured data is the fastest growing type of data generated today. Experts estimate that 80-90% of the data in any organization is unstructured (Anonymous, 2014; Gantz and Reinsel, 2012). Big data paradigm allows to solve the problem of unstructured data processing. Useful information is retrieved by overlaying the raw data with framework. That way, it is possible to make some sense out of unstructured data. Hadoop and similar tools are used to create structure (by using key value pairs) where there is no structure.

Data volumes are growing faster than the rate at which the methods used to work with that data can improve. At the moment there is no such technological product that can cope with the mentioned task. However, the emergence of big data issues markedly accelerated the trend at which they are researched.

The objective of this research is to compare performance tests of greenplum, netezza, exadata and Oracle systems on OS AIX based on big data SAS campaign management.

Related works on nosql data warehousing: Big data is important in banking industry due to operations in the highly competitive environment and large data arrays which are mostly unstructured.

According to Stonebraker *et al.* (2014), many enterprises are confronted with the problem of integration of diverse data structures such as spreadsheets, web sources, XML, traditional DBMSs. In this regard, the Apache Hadoop data management platform is recognized as a leader in big data area and it is uniquely equipped to handle the volume, variety and velocity of unstructured data generated within many businesses.

Fan and Bifet (2013) concluded that large amount of useful data is getting lost because most of the time new data is not tied to files and unstructured.

Corresponding Author: Serbin Vassiliy, Departament of Information Systems, International Information Technology University, Almaty, Kazakhstan

According to Doan *et al.* (2009), managing unstructured data is now an increasingly challenging task to many domains. Ordonez (2013) stated that analyzing the integration of large amounts of relational and unstructured data together is very challenging and leads to a new research area of big data analysis.

Big data has become synonymous with NoSQL databases. However, NoSQL has several drawbacks, for example, it does not support SQL, transactions, reports and other additional characteristics (Han *et al*., 2011).

Every day enormous measures of information (e.g., tests) is produced by researchers and analysts (e.g., high-vitality physical science, science, engineering, bioinformatics and etc.), however separating useful information using RDBMS is almost impossible (Cuzzocrea *et al*., 2013). Here, the challenge is in retrieving useful information from it (Bakshi, 2012).

Although, leading organizations progressively perceive the imperativeness of leveraging their information as a vital asset and are familiar with the term of "big data", a lot of people are still not sure about the costs and profits from new projects involving big data (Gopalkrishnan *et al*., 2012). The regular monitoring of information flow by the system prompts big data issues that are brought by the volume, variety and velocity properties of big data. The learning of the system attributes requires applicable mathematical modelling strategies (Suthaharan, 2014).

Despite, the fact that the analysis of significant information archives has been extensively studied recently, the question of extending the real models and calculations to specific multidimensional information configuration is still open (Cuzzocrea *et al*., 2012).

Although, data storage capability has increased significantly and there are many available data centres around the world it is still very hard to capture and store big data efficiently and make, it easily accessible (Mohamed *et al*., 2008).

Hadoop and HDFS are the dominant platforms for big data analysis at large web companies. Hadoop MapReduce programming model allows data analyses to be expressed much more easily, written and debugged much more quickly. To give a better understanding of when to use MapReduce, it is paramount to differentiate it with other information transforming frameworks. It has been noticed that MapReduce frameworks do not make traditional information handling more obsolete (Pearson and Silva, 2014).

There has been a noticeable tendency for applications to grow in number that are also able to handle huge volumes of data. For instance, Google's MapReduce and its open-source equal Hadoop are powerful tools for building such applications (Kim and Shim, 2012).

Massively Parallel Processing (MPP) (Lugowski *et al*., 2015) is a class of parallel computing system architectures. A distinctive feature of the architecture is that memory is physically divided. The system is built from the individual nodes (node), comprising a processor a local memory bank, communication processors and network adapters, sometimes-hard drives and other input-output devices. Only processors of the same node have access to the random access memory bank of the unit. The nodes are connected by special communication channels. The user can define a logical number of the processor to which he or she is connected and organize the exchange of messages with other processors. Two modes of operation for the operating system are used on the machines of massively parallel architecture.

In one mode, the complete operating system runs only on the management machine (front-end) and each node runs a heavily abridged version of the operating system that supports the operation of the branch of the parallel application allocated to it. In the second mode, each module operates full, most often a UNIX-like system which is installed separately.

Solutions from large companies: EMC greenplum, IBM netezza and Oracle exadata were selected as the objective of the study (Pricket, 2012; Burns, 2011).

Greenplum Software a company engaged in the development of a DBMS for data warehouses. The company specializes in the enterprise data cloud solutions for large-scale data warehousing and analytical systems. Greenplum Database DBMS is based on a modified PostgreSQL database with Massively Parallel Processing (MPP). Greenplum implemented MapReduce functionality and column-oriented organization of tables in their database as part of the so-called polymorphic data storage technology (No *et al*., 2000).

Netezza is an American company, the developer of software and hardware data storage-relational database server clusters, providing massively parallel processing. A distinctive feature of all Netezza complexes is the use of programmable gate arrays in the data processing nodes, providing compression and filtering of data and thus, the allowing to reduce the costs of storage and input-output operations during the execution of data selection queries. The company was founded in 2000; in 2010, it was absorbed by IBM Corporation and in 2011 was fully integrated into the corporation, hardware and software systems are being released under the name IBM PureData for analytics, since, 2012 (Martin *et al*., 2013).

Exadata line of software and hardware systems that was being released commercially by the Oracle Corporation from 2008 to mid-2009 based on Hewlett-Packard server hardware, later based on the hardware from the absorbed Sun Microsystems. The complex is a cluster of database management servers, based on Oracle RAC technology, delivered as pre-assembled telecommunication closets of 42 unit dimensions filled with servers, node storage and switches, InfiniBand or Ethernet (Hussain *et al*., 2013).

Competitors also noted that being focused on OLTP and OLAP-processing simultaneously, makes the systems less effective for analytical processing on which similar solutions from Teradata and Netezza are concentrated in particular, non-optimality of usage of the approach with the symmetrical access from all servers to all storage nodes (symmetrical parallelism) as opposed to the complete separation of data between nodes in competing analytical systems with massively parallel processing is noted (Halstead *et al*., 2015).

Formal representation of the multi-dimensional schema: Construction of mathematical model. Currently, analysts are extremely relevant in all areas of business. This study looks at how to get current information from the accumulated large volumes of raw information. The use of advanced analytical technologies, the ability to extract the necessary knowledge from big data, integrate them into operational processes and insert, convert all this into operational management decisions. The article uses big data, the business analysis of innovations as a competitive advantage and the construction of a mathematical model for decision-making.

In addition, the problems associated with data management such as collection, storage, structuring and classification, using Hadoop, MapReduce methods in the study, it was decided to develop a mathematical model for determining the preferences of customers for the specific choice of a particular company. It should be noted that regardless of their volume and quality, the data is not very useful if they are not retained in such an environment and format which give them access and the most important thing is to analyze them. Big data does not provide success and progress. In order to benefit from the data, it is necessary to analyze them and perform some action based on the results of the analysis. Hadoop and MapReduce systems do not automatically interpret data collected from various data sources, so, we attempted to create mathematical models for further analysis and decision making on the basis of test experimental data.

The use of scoring models to provide accelerated and extended analysis makes it possible to quickly make decisions on the choice of the desired product and gives good results for processing large raw data. For the analysis and forecast of the statistical data, it is necessary to construct a mathematical model that reflects the relationship between the four solutions in order to increase performance of big data processing (Greenplum, Netezza, Exadata and Oracle (AIX) Optimized).

Sufficient amount of information is one of the basic prerequisites for the model. It is assumed that in the second-tier banks of Kazakhstan there is a sufficient or even an excessive amount of information on the bank's customers, since, all banks are offering many products for a long time, at least 5 years. Accordingly, there is enough accumulated data for the analysis. Conversely task is to process and analyze that large volumes of information.

In this study, there are developed mathematical models that can be used to develop a software for large data set, that can be used for further decisions. Covered methods: linear regression, logistic regression, neural networks.

Building a model using the method of multiple linear regression. In this research, for processing data and decision making methods, we used linear regression, multiple linear regression and neural networks. The initial assumptions:

- Errors are not correlated
- An error has a normal distribution

Thus, our task is to find the weights w such that it will satisfy the condition:

$$\min_w \sum_i^n \varepsilon_i^2 \tag{1}$$

The solution of the method of multiple linear regression obtained by the least squares method. Decision algorithm represented as:

$$\sum_{i=1}^{N_D}\left(1-\sum_{j=0}^{p} w_j x_{ij}\right)^2 + \sum_{i=N_D+1}^{N_D+N_{ND}}\left(\sum_{j=0}^{p} w_j x_{ij}\right)^2 \tag{2}$$

where, $x_{i0} = 1$, if i = 1, 2, ..., n. In matrix form:

$$\begin{pmatrix} 1_D & X_D \\ 1_{ND} & X_{ND} \end{pmatrix}\begin{pmatrix} w_0 \\ w^T \end{pmatrix} = \begin{pmatrix} 1_D \\ 0 \end{pmatrix} \tag{3}$$

Or:

$$Xw^T = y^T \quad (4)$$

Here:

$$X = \begin{pmatrix} 1_D & X_D \\ 1_{ND} & X_{ND} \end{pmatrix}$$

is the matrix of size N_D*(p+1):

$$X_D = \begin{pmatrix} x_{11} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{ND1} & \cdots & X_{NDp} \end{pmatrix}$$

is the matrix of size nN_D*p:

$$X_D = \begin{pmatrix} x_{ND+11} & \cdots & x_{ND+1p} \\ \vdots & \ddots & \vdots \\ x_{ND+nND1} & \cdots & X_{ND+nNDp} \end{pmatrix}$$

is the matrix of size nN_D*p. And $y^T = ({}^{1D}{}_0)$. The matrix form of minimization task (1) represented as:

$$\min\left(Xw^T - y^T\right)^T\left(Xw^T - y^T\right) \quad (5)$$

Now, to solve this problem (Eq. 5), we find the derivative and equate it to 0:

$$X^T X^T\left(Xw^T - y^T\right) = 0 \quad (6)$$

Or:

$$X^T X w^T = X^T y^T \quad (7)$$

As the result we get the system of linear equations in matrix form:

$$\begin{pmatrix} 1_D & 1_{ND} \\ X_D & X_{ND} \end{pmatrix}\begin{pmatrix} 1_D & X_D \\ 1_{ND} & X_{ND} \end{pmatrix}\begin{pmatrix} w_0 \\ w^T \end{pmatrix} = \begin{pmatrix} 1_D & 1_{ND} \\ X_D & X_{ND} \end{pmatrix}\begin{pmatrix} 1_D \\ 0 \end{pmatrix} \quad (8)$$

And:

$$\begin{pmatrix} n & n_D\mu_D + n_{ND}\mu_{ND} \\ n_D\mu_D^T + n_{ND}\mu_{ND}^T & X_D^T X_D + X_{ND}^T X_{ND} \end{pmatrix}\begin{pmatrix} w_0 \\ w^T \end{pmatrix} = \begin{pmatrix} n_D \\ n_D\mu_D^T \end{pmatrix} \quad (9)$$

where μ_D and μ_{ND} determine the average variable vectors for overdue and not overdue loans, respectively. Assuming that the learning sample is relevant to all the assumptions of linear regression and all assumptions are met, then, we get:

$$\begin{aligned} &X_D^T X_D + X_{ND}^T X_{ND} = nE\left[X_i X_j\right] = \\ &Cov\left(X_i X_j\right) + n_D\mu_D\mu_D^T + n_{ND}\mu_{ND}\mu_{ND}^T \end{aligned} \quad (10)$$

Let covariance function C mean learning sample. Now, we obtain:

$$X_D^T X_D + X_{ND}^T X_{ND} = nC + n_D\mu_D\mu_D^T + n_{ND}\mu_{ND}\mu_D^T \quad (11)$$

Using Eq. 11 in the system (Eq. 9) we obtain:

$$nw_0 + \left(n_D\mu_D + n_{ND}\mu_{ND}\right)w^T = n_D \quad (12)$$

$$\begin{pmatrix} n_D\mu_D^T + \\ n_{ND}\mu_{ND}^T \end{pmatrix}w_0 + \begin{pmatrix} nC + n_D\mu_D\mu_D^T + \\ n_{ND}\mu_{ND}\mu_D^T \end{pmatrix}w^T = n_D\mu_D^T \quad (13)$$

Then, substitute the first Eq. 12 in the second and we get:

$$\begin{aligned} &\left(\frac{\left(n_D\mu_D^T + n_{ND}\mu_{ND}^T\right)\left(n_D - \left(\mu_D + n_{ND}\mu_{ND}\right)w^T\right)}{n}\right) + \\ &nCw^T + \left(n_D\mu_D\mu_D^T + n_{ND}\mu_{ND}\mu_D^T\right)w^T = n_D\mu_D^T \end{aligned}$$

That is:

$$\left(\frac{n_D n_{ND}}{n}\right)\left(\mu_D - \mu_{ND}\right)w^T + nCw^T = \left(\frac{n_D n_{ND}}{n}\right)\left(\mu_D - \mu_{ND}\right)^T$$

$$Cw^T = a\left(\mu_D - \mu_{ND}\right)^T \quad (14)$$

where, α is constant. The resulting dependence (Eq. 14) gives the desired optimal weight vector:

$$w = \left(w_0, w_1, \ldots, w_p\right)$$

Solving a task using regression model. E(Y|x), Y-dependent variable; x-factor or explanatory variable; $p(x_i) = E(Y|x_i)$. Probability:

$$p\left(x_i\right) = G\left(x_i, w\right) = \frac{e^{w_0 + w_1 x_{i1} + w_2 x_{i2} + \ldots, w_p x_{ip}}}{1 + e^{w_0 + w_1 x_{i1} + w_2 x_{i2} + \ldots, w_p x_{ip}}} = \frac{e^{x_i w}}{1 + e^{x_i w}} \quad (15)$$

$$In\left(\frac{p\left(x_i\right)}{1 - p\left(x_i\right)}\right) = w_0 + w_1 x_{i1} + w_2 x_{i2} + , \ldots, + w_p x_{ip} + e_i \quad (16)$$

$$L\left(w\right) = \prod_{i=1}^{n} p\left(x_i\right)^{yi}\left(1 - p\left(x_i\right)\right)^{1-yi} \quad (17)$$

$$l\left(w\right) = In\left(L\left(w\right)\right) = \sum_{i=1}^{n}\left\{\begin{matrix} y_i In\left(p\left(x_i\right)\right) + \\ \left(1 - y_i\right)In\left(1 - p\left(x_i\right)\right) \end{matrix}\right\} \quad (18)$$

$$\frac{dl(w)}{dw_0} = \sum_{i=1}^{n}(y_i - p(x_i)) = 0 \quad (19)$$

$$\frac{dl(w)}{dw_j} = \sum_{i=1}^{n}(y_i - p(x_i))x_{ij} = 0, j = 1, 2, \ldots, p \quad (20)$$

Solving a task using neural networks: To build a model with neural networks, back-propagation algorithm was used. In this research, neural network with back-propagation algorithm composed of several layers of neurons, each neuron of layer i is connected to each neuron of layer i+1:

$$E(w) = \frac{1}{2}\sum_{j=1}^{p}(y_j - d_j)^2 \quad (21)$$

Where:
y_j = The value of jth neural network output
d_j = The target value of jth output
p = The number of neurons in the output layer

The weights vary with each iteration by the equation:

$$\Delta w_{ij} = -\mu\frac{\partial E}{\partial w_{ij}} \quad (22)$$

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial y_j}\times\frac{\partial y_j}{\partial S_j}\times\frac{\partial S_j}{\partial w_{ij}} \quad (23)$$

Where:
μ = The parameter which defines the speed of learning
y_j = The value of jth neural network output
S_j-a = The weighted sum of the input signals, defined by the formula:

$$S_j = \sum_{i=1}^{n} w_i x_{ij} \quad (24)$$

Here:

$$\frac{\partial S_j}{\partial w_{ij}} = x_i \quad (25)$$

Where:
x_i = The value of the ith input of neuron. The definition of the first factor of the formula
k- = The number of neurons in layer (n+1)

Assuming that:

$$\frac{\partial E}{\partial y_i} = \sum_k \frac{\partial E}{\partial y_k}\times\frac{\partial y_k}{\partial S_k}\times\frac{\partial S_k}{\partial y_j} = \sum_k \frac{\partial E}{\partial y_k}\times\frac{\partial y_k}{\partial S_k}\times w_{jk}^{(n+1)} \quad (26)$$

Next, determine the recursive formula for determining the nth layer, if it is known the next (n+1)-layer:

$$d_j^{(n)} = \frac{\partial E}{\partial y_j}\times\frac{\partial y_j}{\partial S_j} \quad (27)$$

The last layer is determined as follows:

$$d_j^{(n)} = \left[\sum_k \delta_k^{(n+1)}\times w_{jk}^{(n+1)}\right]\times\frac{\partial y_j}{\partial S_j} \quad (28)$$

Equation 22 defined as:

$$\Delta w_{ij}^{(n)} = -\mu d_j^{(N)}\times x_i^n \quad (29)$$

$$w_{ij}^{(n)}(t) = w_{ij}^{(n)}(t\text{-}1) + \Delta w_{ij}^{(n)}(t) \quad (30)$$

A mathematical model of the problem 1 and 2 for N = 4 and the sample solution:

$$P = w_0 + w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 \quad (31)$$

This model is solved by least squares method. LSM is one of the basic methods of regression analysis to estimate the unknown parameters of regression models for the sample data.

If the system of equations has a solution, then sum of least squares is equal to zero and exact solutions of equations can be found analytically or for example by the various numerical methods of optimization. If the system is over determined that is the number of independent equations is larger than the number of unknown variables, the system has no exact solution and the method of least squares allows you to find a certain "optimal" vector x in the sense of maximum closeness of the vectors P or as close as possible of deviations vector e to zero (closeness is understood in the sense of Euclidean distance):

$$J = \sum_{i=1}^{N}(P_i - P_i)^2 \rightarrow \min \quad (32)$$

Where:
N = The dimension
P = The target function
x_1-x_4 = The factors, explanatory variables
w_0-w_4 = The unknown coefficients

Introducing the concept of a residual function:

$$J = \sum_{i=1}^{N}\left(\left(\begin{matrix} w_0 + w_1x_{1i} + w_2x_{2i} + \\ w_3x_{3i} + w_4x_{4i}\end{matrix}\right) - \breve{P}_i\right)^2 = \sum_{i=1}^{N}\left(P_i - \breve{P}_i\right)^2 \rightarrow \min \quad (33)$$

To ensure the minimum of the expression (Eq. 33), it is necessary that the partial derivative of this expression with respect to all the parameters w_0-w_4 are equal to zero:

$$\frac{\partial J}{\partial w_0}=0;\frac{\partial J}{\partial w_1}=0;\frac{\partial J}{\partial w_2}=0;\frac{\partial J}{\partial w_3}=0;\frac{\partial J}{\partial w_4}=0$$

$$\frac{\partial J}{\partial w_0}=2\sum_{i=1}^{N}\left(\begin{matrix}w_0+w_1x_{1i}+w_2x_{2i}+\\ w_3x_{3i}+w_4x_{4i}-\breve{P}_i\end{matrix}\right)=0$$

$$\frac{\partial J}{\partial w_1}=2\sum_{i=1}^{N}\left(\begin{matrix}w_0+w_1x_{1i}+w_2x_{2i}+\\ w_3x_{3i}+w_4x_{4i}-\breve{P}_i\end{matrix}\right)\times x_{1i}=0$$

$$\frac{\partial J}{\partial w_2}=2\sum_{i=1}^{N}\left(\begin{matrix}w_0+w_1x_{1i}+w_2x_{2i}+\\ w_3x_{3i}+w_4x_{4i}-\breve{P}_i\end{matrix}\right)\times x_{2i}=0$$

$$\frac{\partial J}{\partial w_3}=2\sum_{i=1}^{N}\left(\begin{matrix}w_0+w_1x_{1i}+w_2x_{2i}+\\ w_3x_{3i}+w_4x_{4i}-\breve{P}_i\end{matrix}\right)\times x_{3i}=0$$

$$\frac{\partial J}{\partial w_4}=2\sum_{i=1}^{N}\left(\begin{matrix}w_0+w_1x_{1i}+w_2x_{2i}+\\ w_3x_{3i}+w_4x_{4i}-\breve{P}_i\end{matrix}\right)\times x_{4i}=0$$

Then, the system of linear algebraic equations is as follows:

$$\sum_{i=1}^{N}x_{1i}+\sum_{i=1}^{N}x_{2i}+\sum_{i=1}^{N}x_{3i}+\sum_{i=1}^{N}x_{4i}=\sum_{i=1}^{N}\breve{P}_i$$

$$\sum_{i=1}^{N}x_{1i}+\sum_{i=1}^{N}x_{1i}\times x_{1i}+\sum_{i=1}^{N}x_{2i}\times x_{1i}+\sum_{i=1}^{N}x_{3i}\times x_{1i}+\sum_{i=1}^{N}x_{4i}\times x_{1i}=\sum_{i=1}^{N}\breve{P}_i$$

$$\sum_{i=1}^{N}x_{2i}+\sum_{i=1}^{N}x_{2i}\times x_{1i}+\sum_{i=1}^{N}x_{2i}\times x_{2i}+\sum_{i=1}^{N}x_{3i}\times x_{2i}+\sum_{i=1}^{N}x_{4i}\times x_{2i}=\sum_{i=1}^{N}\breve{P}_i\times x_{2i}$$

$$\sum_{i=1}^{N}x_{3i}+\sum_{i=1}^{N}x_{1i}\times x_{3i}+\sum_{i=1}^{N}x_{2i}\times x_{3i}+\sum_{i=1}^{N}x_{3i}\times x_{3i}+\sum_{i=1}^{N}x_{4i}\times x_{3i}=\sum_{i=1}^{N}\breve{P}_i\times x_{3i}$$

$$\sum_{i=1}^{N}x_{4i}+\sum_{i=1}^{N}x_{1i}\times x_{4i}+\sum_{i=1}^{N}x_{2i}\times x_{4i}+\sum_{i=1}^{N}x_{3i}\times x_{4i}+\sum_{i=1}^{N}x_{4i}\times x_{4i}=\sum_{i=1}^{N}\breve{P}_i\times x_{4i}$$

First of all, we note that for the linear models LSM estimates are linear as it follows from the above given equation. For unbiasedness of LSM estimates, necessary and sufficient conditions for a major regression analysis:

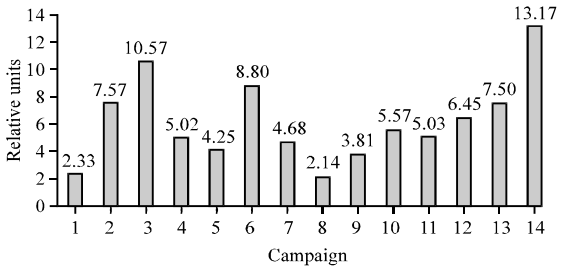

Fig. 1: Duration of performance SAS-campaigns

conditional factors on the expectation of the random error should be zero. This condition is particularly satisfied if:

- The mathematical expectation of the random error is zero
- Factors and random errors-independent random variables

The first condition can be considered fulfilled always for models with a constant. Since, the constant assumes a non-zero expectation error (so, the model with constant generally preferred).

The second condition-the condition of exogenous factors-fundamental. If this property is not satisfied, it can be assumed that almost any evaluation will be very poor: they are not wealthy (that is, even a very large amount of data does not provide a qualitative assessment in this case). In the classical case stronger assumption of determinism factors should be made as opposed to random error which automatically means that the condition exogenous.

Experiments: During current research, a comparative performance test and load test for Greenplum, Netezza, Exadata and Oracle systems on OS AIX based on big data SAS campaign management were conducted of one of the largest Kazakhstan banks.

Performance test was designed to measure the maximum performance with a given set as campaign of circumstances and settings.

Measurement of the duration of SAS-campaigns was conducted before the experiment in accordance with Fig. 1. The testing yielded following data in accordance with Table 1.

The settings of systems of each product are illustrated in Table 2. Technical solutions and optimization were done for all DBMS on the side of SAS (Serbin *et al*., 2016).

Manual processing of formulas was completed to improve descent into database queries. A number of

Table 1: Time of campaigns in relative units

Name campaign	Current oracle	Netezza	Greenplum	Exadata	Oracle (AIX) optimized
1	2.33	0.21	0.23	0.15	0.85
2	7.57	2.43	2.51	0.67	2.33
3	10.57	3.37	2.34	1.03	3.79
4	5.02	2.87	2.47	1.34	2.75
5	4.25	2.87	2.19	1.23	2.70
6	8.80	3.85	2.52	1.52	3.73
7	4.68	2.35	1.33	1.75	2.03
8	2.14	1.05	0.52	0.36	0.78
9	3.81	2.47	1.10	0.71	1.71
10	5.57	2.02	1.10	0.76	1.62
11	5.03	2.43	1.53	0.97	1.93
12	6.45	2.73	1.39	1.14	2.79
13	7.50	3.46	1.79	1.24	2.82
14	13.17	3.41	2.44	2.22	4.42
Averege	6.21	2.54	1.68	1.08	2.45

Table 2: Specifications

Variables	Oracle current	Netezza	Greenplum	Exadata	Oracle (AIX) Optimized
Configuration database	24 core Power 7 (3.4 GHz) 160 GB RAM 3000 MB/sec SAN	NZ1000-3 (4 SPU-only 24 CPU+FPGA)	4 server segment to 16 CPU cores each	XZ-2 Half Rack 4 database nodes to 16 CPU cores storage node 7-12 CPU cores	24 core Power 7 (3.4 GHz) 160 GB RAM 2300 MB/sec San
SAS compute server configuration	Wirth. (VMWare) 12 vCPU~300 MB/sec SAN Storage SAS 9.2	16 physical cores CPU 1 GB/sec DAS storage SAS 9.3	16 physical cores CPU 1 GB/sec DAS storage SAS 9.3	16 physical cores CPU 1 GB/sec DAS storage SAS 9.3	16 physical cores CPU 1 GB/sec DAS storage SAS 9.3
Configuring SAS Mid-Tier Server	Wirth. (VMWare) 4 vCPU 12 GB RAM	Wirth. (VMWare) 4 vCPU 12 GB RAM	Wirth. (VMWare) 4 vCPU 12 GB RAM	Wirth. (VMWare) 4 vCPU 12 GB RAM	Wirth. (VMWare) 4 vCPU 12 GB RAM
Network interface SAS-DB	Unknown	10 GB	10 GB	10 GB	To 1 GB
Cost, mill. $	0.5-1	2.5-3.5	1-2.5	4-5	0.5-1

formulas used in campaigns have been changed so that queries could use them to descend into the database. Since, the tested Version SAS CM 5.41 allows to edit the formula "globally" in other words, system-widely. Testing are run sets of test's company or packets.

First packet ("odd") 7 test campaigns runs in parallel (N1, 3, 5, 7, ..., from list). Campaigns are run at the same time (with minimal intervals, the entire run should keep within 1 minute) in the execute mode preselected option "test" (test upload). Then, the expected completion of all 7 campaigns.

The second pack ("even") the remained 7 test campaigns launch by parallel (campaign No. 2, 4, 6, 8, ..., from the list). The process run campaigns as in the first round.

The results of every test are the time of execute of every company by separately that defined by log "SAS marketing automation core" by type of records as "executed all communications for (name of company) in 20 min 20.30 sec".

It is assumed that the current release of Oracle acceleration is equal to 1.00, the relative acceleration at a netezza, greenplum, exadata and Oracle (AIX) was optimized in accordance with Fig. 2. Percentage of acceleration to "optimize" Oracle AIX is shown in Table 3.

The schedules shown above from first "pack" second "pack" is rather similar. "Average" acceleration of packs of campaigns, we count on a equation:

Table 3: Percentage of acceleration to "optimize" Oracle AIX: (negative acceleration (%), positive-deceleration (%))

Name campaign	Current Oracle (%)	Netezza (%)	Green plum (%)	Exadata (%)	Oracle (AIX) optimized
1	175.0	-75.2	-72.4	-82.1	-
2	225.2	4.4	8.0	-71.0	-
3	179.3	-11.1	-38.1	-72.7	-
4	82.9	4.6	-10.2	-51.1	-
5	57.8	6.5	-18.8	-54.4	-
6	135.9	3.2	-32.5	-59.3	-
7	130.2	15.6	-34.4	-14.1	-
8	174.6	34.9	-33.3	-53.3	-
9	122.8	44.2	-35.6	-58.7	-
10	244.4	24.9	-32.1	-53.1	-
11	160.4	25.9	-20.6	-49.7	-
12	131.0	-2.1	-50.1	-59.1	-
13	166.2	22.9	-36.5	-55.9	-
14	197.9	-23.0	-44.9	-49.8	-
Outcome	0.0	3.8	-31.4	-55.9	-

$$A_{avg} = \frac{S-T}{S} \quad (34)$$

Where:

S = The sum of all duration of calculation of campaigns from a pack on the tested contour

T = The sum of all times of campaigns from a pack on the "optimized" Oracle AIX (Mukazhanova and Serbin, 2016)

Summary comparative results are presented on Table 4. The less negative number is higher acceleration.

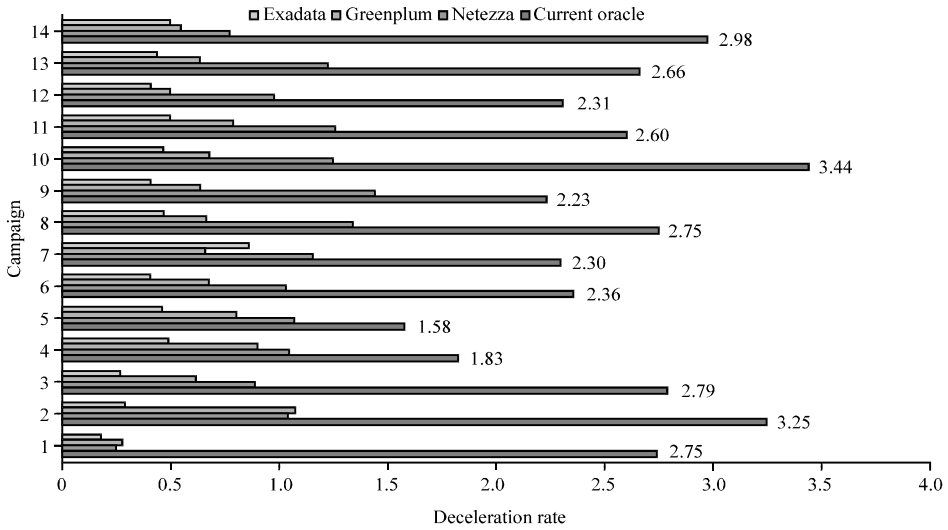

Fig. 2: Relation of time to "optimize" Oracle AIX (time/time) (<1-acceleration, >1-deceleration)

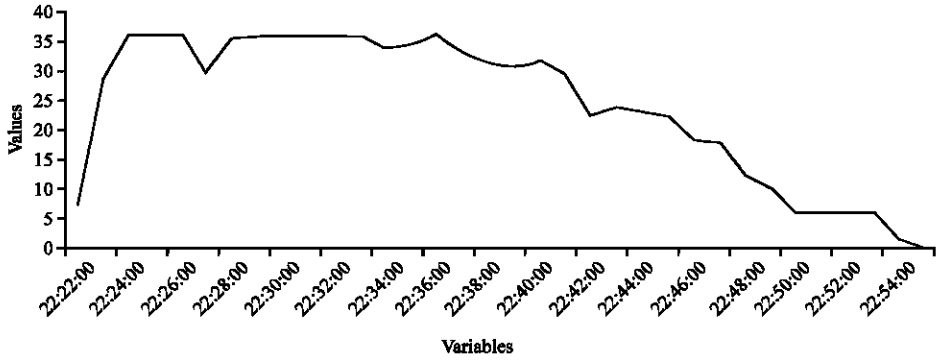

Fig. 3: Graphic running task Netezza

Table 4: Summary comparative results

Variables	Netezza (%)	Greenplum (%)	Exadata (%)	Oracle (%)
Pack 1 (odd)	-1.2	-25.1	-57.6	138.0
Pack 2 (even)	9.3	-38.6	-54.0	171.8

Positive percent-delay: Based on the research of productivity of big data processing which was done on three data systems and on optimized Oracle, the following results were obtained. Based on the runtime performance of campaigns, it could be noticed that Exadata accelerates processing of campaigns by 55.9% compared to the current solution. Greenplum accelerates processing of campaigns by 31.4%. Netezza did not show any acceleration.

Load testing was designed in order to determine a software's behavior under both normal and anticipated peak load conditions. It helps to identify the maximum operating capacity of an application as well as any bottlenecks and determine which element is causing degradation.

Production schedules on the systems used in the analysis are given. Schedules contain identical temporary scales across that at an opportunity to compare schedules with each other. Statistics at tests I gathered on servers with 30 sec intervals.

Schedule of running task "Task" is the settlement task started in system: either SQL inquiry or the stored process of SAS (Fig. 3-5). The schedule of netezza CPU aggregate are constructed on data of Excel of the utility of monitoring. The green schedule is the maximum load on one of knots (Fig. 6).

The schedule of greenplum segment node are constructed with use of means of NMON Analyser on the basis of the data collected by the utility of nmon on one of segment knots (Fig. 7).

The schedule of exadata storage sell node are constructed with use of means of NMON analyser on the basis of the data collected by the utility of nmon on one of Storage of servers and on one of Database of servers (Fan and Bifet, 2013) (Fig. 8).

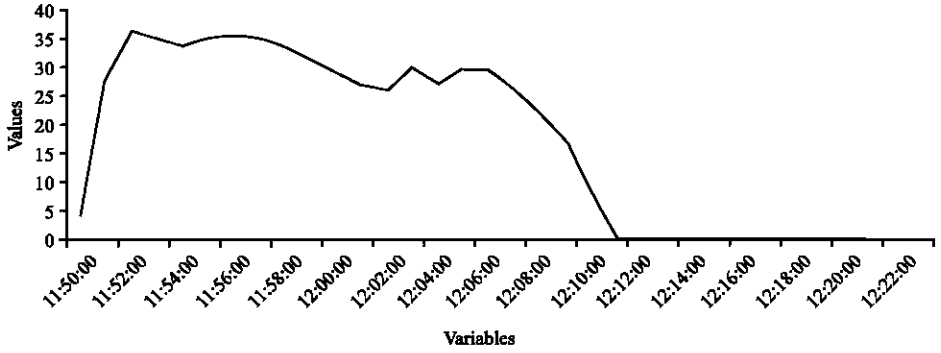

Fig. 4: Graphic running task Greenplum

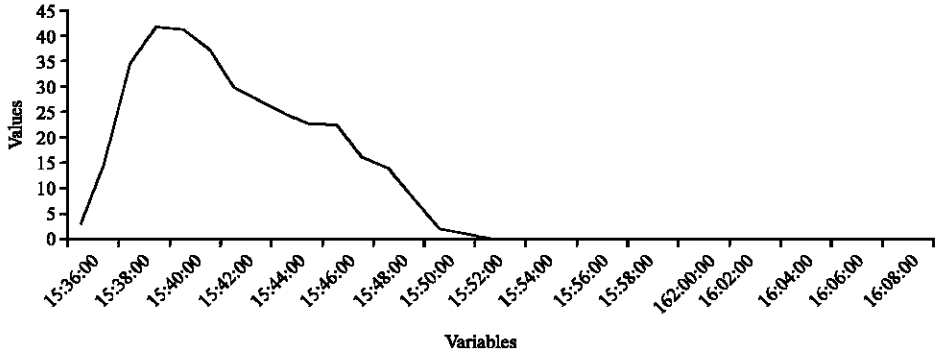

Fig. 5: Graphic running task Exadata

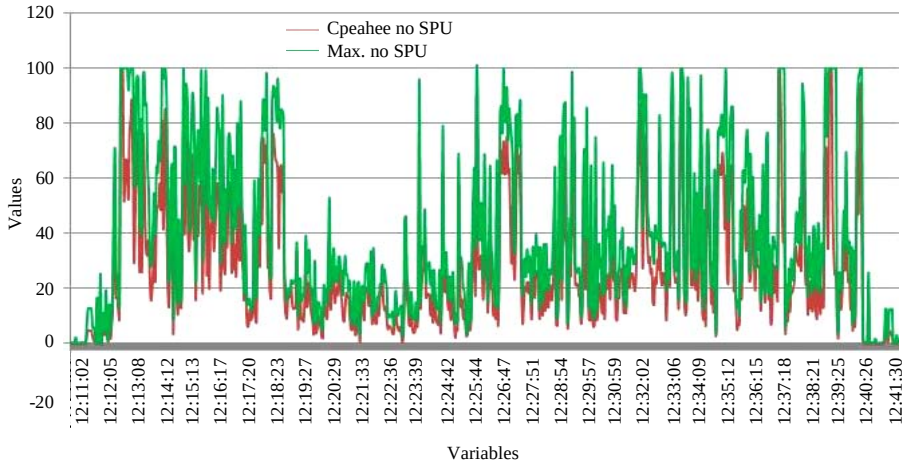

Fig. 6: Netezza CPU aggregate

Oracle production load of CPU and Disks constructed by means of the utility of nmon from the AIX server on which the DB settles down (Fig. 9).

Stress testing on SAS meta/mid-tier server: It should be noted that all tests have shown SasCM-Prod-Meta server overload by CPU (all 4 cores are disposed) in the first 3-4 min from the moment of launching campaign. This server is responsible for the campaign launch, information card structure recording and generation of tasks. This server does not work with data. Server load depends on the number and complexity (by nodes) of running campaigns, complexity of information card. Exadata overload is especially clearly noticeable, the tasks that are

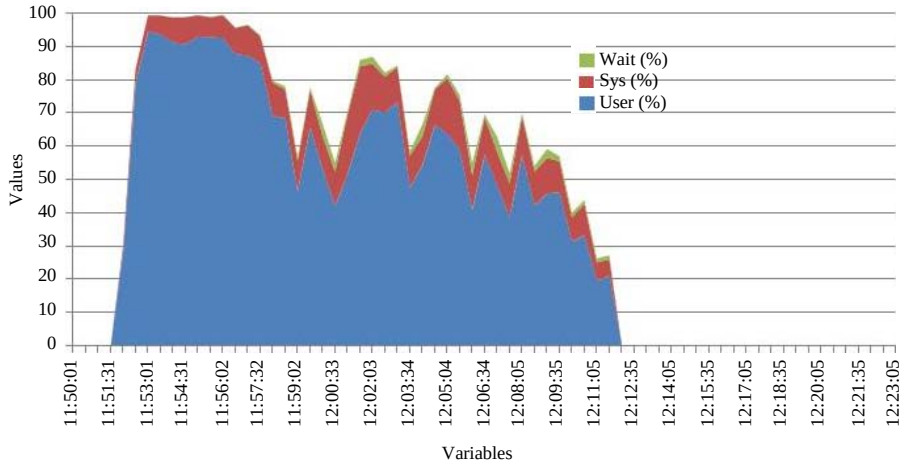

Fig. 7: Greenplum segment node

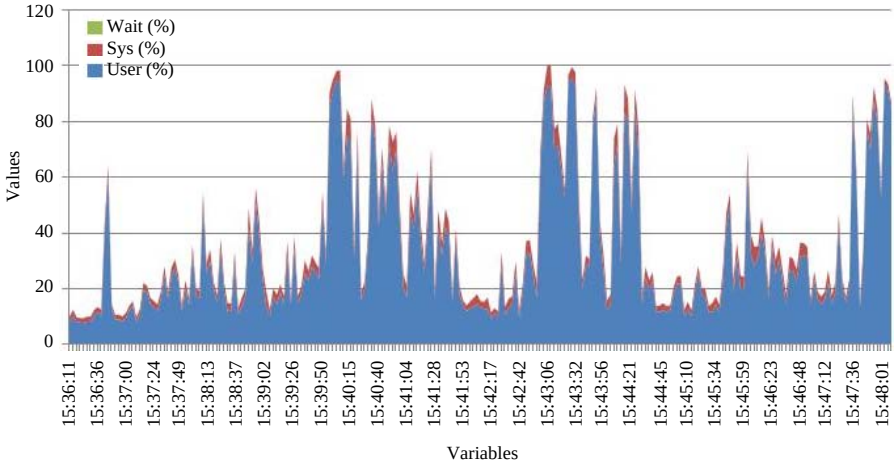

Fig. 8: Exadata storage sell node

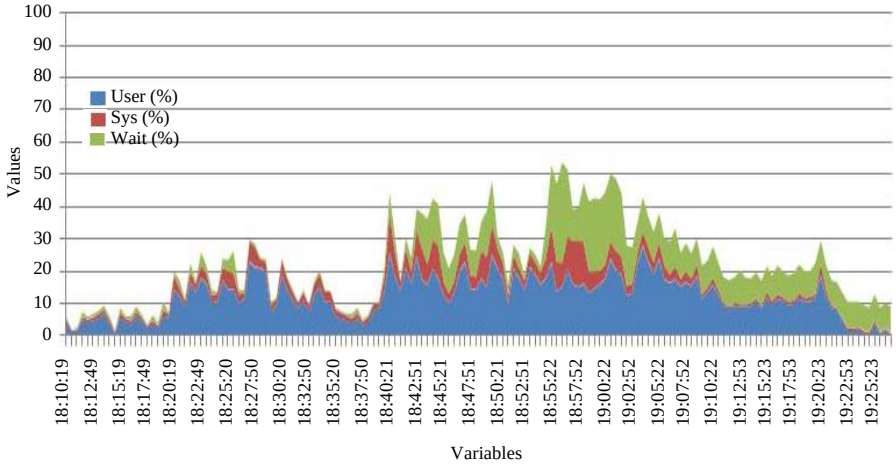

Fig. 9: Oracle production load of CPU and disks

Table 5: Analysis of bottlenecks

Systems	Netezza	Greenplum	Oracle exadata
Estimated bottlenecks	70-80%: CPU disks 20%: transmission of data ODBC interface restrictions	50%: CPU server segment 50%: calculations in SAS (architecture of SAS-data transfer, unreleased requests)	SAP Architecture (data transfer from/to the database "thin table", "unreleased" requirements. Partially storage SAS compute server
Time scaling (possibility of accelerating the same load by increasing the database configuration)	Yes (by strengthening the CPU of the disk subsystem by increasing the number of CPU	Yes (limited) increasing the number of segmented database servers. Also by reducing the number of "unreleased" queries after migration to SAS 9.4 or by manual editing of nodes in the campaigns	No because the emphasis is no longer in the database limited acceleration is possible due to manual editing of nodes in the campaigns, the elimination of "not started" (automatically it is not possible)
Scaling the load (possibility to take a greater load without increasing the operating time by increasing the database configuration)	Limited (when the load grows, limit of ODBC will increase)	Yes (there is no reason to wait for problems with scaling)	Yes based on the chart of loads, 2-3 times increase in the load does not require an expansion of the database resources, there is a margin

formed in the first minutes are executed faster and Mid-Tier Server is forced to generate new tasks immediately after completing the initial volume.

It should be noted that the situation of simultaneous campaign opening/launch (at 5 sec intervals) is not typical for the industrial operation of a system.

Greenplum (data caching in memory (OS file cache)): Stress testing schedules in the segment server Greenplum show that performance increases due to data caching in Random Access Memory (RAM). All showcase occupies about 200 GB, the server, provided for testing the total amount of Random Access Memory (RAM) has 768 GB (4 segment servers, 192 GB in each).

During the tests, all showcases were cached in Random Access Memory (RAM). The graphs show that there is no reading from the disk on the segment node.

In this task, greenplum database factually functions as "In-memory" solution. But it is important to note that the Linux file cache is not controlled and when the active volume of data exceeds storage capacity in the system, this effect can quickly degrade. Analysis of bottlenecks of the system is illustrated in Table 5.

It should be noted that the system for testing was presented in different configurations (Table 2). Exadata was represented to scale Half Rack (1/2 of the full version of the stand) while Netezza and Greenplum to scale corresponded to 1/4 of the rack.

The Oracle database (AIX) on the SSD-drives was not connected to the network 10 Gb/sec which could affect the test results but performance analysis showed that the basic speed limit has full CPU load on the database server. The network does not have a decisive influence. The reliability of the results is confirmed by experiments on real systems deployed with real data.

CONCLUSION

The idea of the specialized analytical car has captured many leading players of the IT industry today-for a solution of the problem of big data in particular a possibility of the solution of analytical tasks in real time, traditional infrastructures are no good any more. New devices are specially designed under needs of business analytics and combine the best qualities of modern equipment rooms and program architecture to exclude emergence of any bottlenecks on the way of huge data flows and to provide necessary flexibility, scalability and reliability.

Due to collection of large amount data sets and their unstructured nature banks require distributed approach of computation. The most popular solutions of MPP are EMC Greenplum, IBM Netezza and Oracle exadata. Greenplum is based on PostgresSQL RDBMS and implements MapReduce functionality. Netezza is the complex of data processing nodes that provides compression and filter data during processing. Exadata is the cluster of database management servers based on Oracle technology and distinguished from other solutions by the approach with symmetrical access from all servers.

On the basis of the performance study for processing of large data, the following results were obtained: Exadata gave the best results of system testing. Based on the run-time campaigns in relative units it can be seen that Exadata was accelerates processing of campaigns by 55.9% compared to the current version of oracle. Greenplum speeds up the processing of campaigns by 31.4%.

Thus, the acceleration of processing large data on green plum in accordance with the equipment is two times higher than the current version of Oracle, on Exadata-4 times the acceleration.

In conclusion, the analysis conducted for solutions for increasing processing performance of large data with the Oracle with the option database real application cluster exadata, may be identified as the most powerful system in comparison with analogues MPP-solutions.

On the basis of experiments to identify the best system performance with massively parallel processing of big data architecture, there was built the mathematical model of prediction. This model can be used to identify a trend in performance of other big data processing.

REFERENCES

Anonymous, 2014. Reliable thermal paper supplier?. Tele-Paper (M) Sdn Bhd, Shah Alam, Malaysia. http://www.webopedia.com/quick_ref/just-how-much-data-is-out-there.html.

Bakshi, K., 2012. Considerations for big data: Architecture and approach. Proceedings of the IEEE Conference on Aerospace Conference, March 3-10, 2012, IEEE, Herndon, Virginia, ISBN:978-1-4577-0556-4, pp: 1-7.

Burns, R., 2011. Exadata is still oracle. Teradata, Islamabad, Lahore, Pakistan.

Cuzzocrea, A., A. Papadimitriou, D. Katsaros and Y. Manolopoulos, 2012. Edge betweenness centrality: A novel algorithm for QoS-based topology control over wireless sensor networks. J. Network Comput. Appl., 35: 1210-1217.

Cuzzocrea, A., D. Sacca and J.D. Ullman, 2013. Big data: A research agenda. Proceedings of the 17th International Symposium on Database Engineering and Applications, October 09-13, 2013, ACM, New York, USA., ISBN: 978-1-4503-2025-2, pp: 198-203.

Doan, A., J.F. Naughton, R. Ramakrishnan, A. Baid and X. Chai *et al*., 2009. Information extraction challenges in managing unstructured data. ACM SIGMOD Rec., 37: 14-20.

Fan, W. and A. Bifet, 2013. Mining big data: Current status and forecast to the future. ACM. SIGKDD. Explor. Newsl., 14: 1-5.

Gantz, J. and D. Reinsel, 2012. The digital universe in 2020: Big data, bigger digital shadows and biggest grow the in the far east. IDC December 2012. http://www.emc.com/collateral/analyst-reports/idc-the-digital-universe-in-2020.pdf.

Gopalkrishnan, V., D. Steier, H. Lewis and J. Guszcza, 2012. Big data, big business: Bridging the gap. Proceedings of the 1st International Workshop on Big Data Streams and Heterogeneous Source Mining: Algorithms Systems Programming Models and Applications, August 12, 2012, ACM, Beijing, China, ISBN:978-1-4503-1547-0, pp: 7-11.

Halstead, R.J., I. Absalyamov, W.A. Najjar and V.J. Tsotras, 2015. FPGA-based multithreading for In-memory hash joins. Proceedings of the 7th Biennial International Conference on Innovative Data Systems Research (CIDR '15), January 4-7, 2015, CIDR, Asilomar, California, USA., pp: 1-9.

Han, J., E. Haihong, G. Le and J. Du, 2011. Survey on NoSQL database. Proceedings of the 6th International Conference on Pervasive Computing and Applications, October 26-28, 2011, Port Elizabeth, pp: 363-366.

Hashem, I.A.T., I. Yaqoob, N.B. Anuar, S. Mokhtar, A. Gani and S.U. Khan, 2015. The rise of big data on cloud computing: Review and open research issues. Inform. Syst., 47: 98-115.

Hussain, S.J., T. Farooq, R. Shamsudeen and K. Yu, 2013. Overview of Oracle RAC. In: Expert Oracle RAC 12c, Shamsudeen, R., S.H. Jaffar, Y. Kai and F. Tariq (Eds.). Apress, Berkeley, California, USA., ISBN:978-1-4302-5044-9, pp: 1-28.

Kim, Y. and K. Shim, 2012. Parallel top-k similarity join algorithms using MapReduce. Proceedings of the 2012 IEEE 28th International Conference on Data Engineering (ICDE), April 1-5, 2012, IEEE, Washington, DC, USA., ISBN:978-0-7695-4747-3, pp: 510-521.

Loghin, D., B.M. Tudor, H. Zhang, B.C. Ooi and Y.M. Teo, 2015. A performance study of big data on small nodes. Proc. VLDB. Endowment, 8: 762-773.

Lugowski, A., S. Kamil, A. Buluc, S. Williams and E. Duriakova *et al*., 2015. Parallel processing of filtered queries in attributed semantic graphs. J. Parallel Distrib. Comput., 79: 115-131.

Martin, D., O. Koeth, J. Kern and I. Ivanova, 2013. Near real-time analytics with IBM DB2 analytics accelerator. Proceedings of the 16th International Conference on Extending Database Technology, March 18-22, 2013, ACM, New York, USA., ISBN:978-1-4503-1597-5, pp: 579-588.

Mihovska, A., S. Kyriazakos and J. Soldatos, 2015. Resource Management Support of Big Data Procurement. In: Resource Management in Future Internet, Poulkov, V. and P. Ramjee (Eds.). River Publisher, New York, USA., ISBN:978-93102-44-6, pp: 19-35.

Mohamed, N., J. Al-Jaroodi and I. Jawhar, 2008. Middleware for robotics: A survey. Proceedings of the 2008 IEEE International Conference on Robotics, Automation and Mechatronics, September 21-24, 2008, IEEE, Chengdu, China, ISBN:978-1-4244-1675-2, pp: 736-742.

Mukazhanova, M.N. and V.V. Serbin, 2016. Research of load testing of systems with a massive parallel architecture on the big data. Proceedings of the 2nd International Conference on Information Technologies in Science and Industry 2016, May 19, 2016, International IT University Press, Almaty, Kazakhstan, pp: 79-82.

No, J., R. Thakur and A. Choudhary, 2000. Integrating parallel file I/O and database support for high-performance scientific data management. Proceedings of the ACM/IEEE 2000 International Conference on Supercomputing, November 4-10, 2000, IEEE, Dallas, Texas, USA., pp: 57-57.

Ordonez, C., 2013. Can we analyze big data inside a DBMS?. Proceedings of the 16th International Workshop on Data Warehousing and OLAP, October 28-28, 2013, ACM, New York, USA., ISBN:978-1-4503-2412-0, pp: 85-92.

Pearson, S.S. and Y.N. Silva, 2014. Index-Based RS Similarity Joins. In: Similarity Search and Applications, Traina, A.J.M., C. Traina and R.L.F. Cordeiro (Eds.). Springer, Cham Municipality, Switzerland, ISBN:978-3-319-11987-8, pp: 106-112.

Pricket, M.T., 2012. Oracle cranks up the flash with exadata X3 systems. Situation Publishing, London, England, UK. https://www.theregister.co.uk/2012/10/01/oracle_exadata_x3_systems/

Serbin V., K. Duysebekova and A. Altaybek, 2016. Research performance systems with massively parallel processing to big data. Eurasian Union Sci., 1: 118-122.

Stonebraker, M., A. Pavlo, R. Taft and M.L. Brodie, 2014. Enterprise database applications and the cloud: A difficult road ahead. Proceedings of the 2014 IEEE International Conference on Cloud Engineering (IC2E), March 11-14, 2014, IEEE, Boston, Massachusetts, USA., ISBN:978-1-4799-3766-0, pp: 1-6.

Suthaharan, S., 2014. Big data classification: Problems and challenges in network intrusion prediction with machine learning. ACM. SIGMETRICS. Perform. Eval. Rev., 41: 70-73.

Watson, R.N., J. Woodruff, P.G. Neumann, S.W. Moore and J. Anderson *et al*., 2015. Cheri: A hybrid capability-system architecture for scalable software compartmentalization. Proceedings of the 2015 IEEE International Symposium on Security and Privacy (SP), May 17-21, 2015, IEEE, San Jose, California, USA., ISBN:978-1-4673-6949-7, pp: 20-37.