Development and Robust Testing of Big Data Analysis System based on Optimized Java Technique and Parallel Oracle Database

Haichun Zuo
GUANGZHOU NANYANG COLLEGE, Guangzhou 510925, China
*Corresponding author(E-mail: haichun1234@163.com)

Abstract
With the rapid development of artificial intelligence and data control science, the demand for big data analysis with programming language and database servers is growing. To deal with the hard challenge, in this research paper, we conduct research on the development and robust testing of big data analysis system based on the optimized Java technique and parallel Oracle database. The large relational database Oracle has been widely used in all walks of life and has gradually become the important database platform of the enterprise information construction with the expansion of Oracle database. We finalize our research in the aspects. (1) We combine the clustering method to present the new data mining algorithm. Clustering analysis method is based on the data sets in data mining association between measures of the automatic which is divided into the several clusters make between data points within the same cluster as similar as possible. (2) The optimization of Oracle database server. Distributed application architecture of Oracle database has two kinds of distributed database system and advanced replication system, we combine the advances of the parallel feature to enhance the traditional database structure for big data. (3) We optimize the basic Java organization and programming habit to enhance the systematic performance of the data processing for Java with re-organized data structure to reduce the time-consuming. In the experimental part, we compare our algorithm with other state-of-the-art algorithms numerically and visually, the result indicates that our method performs better. In the final part, we summarize our method and propose the prospect for the coming research.

Key words: Big Data Analysis, Data Control System, Robust Testing, Optimized Java, Parallel Oracle Database

1. INTRODUCTION

Big data platform for data management system of the main challenges can be summarized as the generation of large amount of data, data, retrieving and updating speed and data variety three aspects. Big data in the word "big" is often used to describe the space in daily life. Therefore, when it comes to big data, people's first impression is the amount of data must be very big, which require large data problem must satisfy the characteristics of the data volume is very big(Barr, 2015). This is actually a mistake understanding of big data. From the perspective of the scientific research, three aspects of data can describe the characteristics of the data as a three dimension(Ceccato, 2016; Kristo, 2013). Any dimension of the demand over the processing capacity of the existing system, such problems worth studying is the big data problem. Theoretically, primary steps of dealing with big data could be separated into following classifications. (1) The data collection step. The big data acquisition is the basis of the entire process, with the development of Internet technology and applications and the popularity of a variety of terminal equipment, make the data producer range is more and more big, the output of data is also more and more, the link between the data is becoming more and more complex, it is also the embodiment of the "big" in the big data, so we need to improve data collection speed and accuracy. (2) Data processing and integration. Data processing and integration mainly on the large amounts of data were collected in the previous step, the appropriate pre-treatment, including the formatting, de-noising, and further integrated storage(Guo, 2016). Because the data collection procedures to collect various data, the data structure is not unified, not conducive to after data analysis, and some data belongs to the invalid data, will need to be removed, otherwise it will affect the accuracy and reliability of the data analysis, therefore, need to be unified data format and remove invalid data. Often design some filter to complete the task. (3) Data analysis and understanding. After completed the data acquisition and processing, to analyze the data, because after the data analysis can reflect all the important value of big data(Hwang, 2015). Object is on the analysis of the data after the step of data processing and integration of unified format data, need according to the application requirements and the value direction of the data needed for the original sample data further processing and analysis. Existing data analysis usually refers to the use of the data warehouse and data mining tools for the centralized storage of data analysis, data analysis service and the difference between traditional data analysis is its object oriented is not data, but data services. (4) Data instructions. Data interpretation is explanation for big data analysis and presentation, in data processing, the interpretation of the data results step is the user directly faces with the achievements of big data analysis steps for future analysis and operation(Bellec, 2015; Deng, 2016).
Figure 1. The Basic Tasks and Applications of Big Data Technology

As shown in the figure one, the basic applications of the big data could vary from user to user. The large relational database Oracle has been widely used in all walks of life and has gradually become the important database platform of enterprise information construction, but with the expansion of Oracle database. The increase of database users, database performance problem more and more prominent, therefore, it is necessary to adjust the Oracle database performance and optimize, under the condition of meet the requirements, system performance to achieve the best and minimum system cost. Stored procedure is an important object in the database (Lee, 2015). Oracle database allows stored procedure as a database object is generally stored in the database to perform once, compiled code stored in the cache memory, when using direct call again and no compiler can perform again. Use the stored procedures, therefore, we can greatly improve the speed of data storage and query, still can reduce the complexity of the whole system, improve the reusability, security, scalability and maintainability and enhance the performance of the whole application system. The traditional method of large database to its robustness of data mining structure modeling using data gathered scheduling method based on genetic algorithm, there is a big path loss, the robustness of data mining. According to the above problem, this paper proposes the improved detecting outliers based on local genetic evolution of large database type of the cross data mining model, through mathematical model building and data mining algorithm is improved and raising of large flow media cross model of data mining in the database performance to improve the robustness of database access (Kanewala, 2013).

This paper is organized as the following parts. In the section two, we theoretically analyze the state-of-the-art data mining and big data analysis algorithms to serve as the foundation for the system. In the section three, we introduce basic features and optimization technique for the Oracle database server to enhance its robustness when facing with large-scale data stream. In the fourth section, we build up our system with the review and introduction of the Java and corresponding techniques. In the last two sections, we summary down work with simulation on large-scale database. Moreover, prospect is given in the final parts.

2. THE PROPOSED METHODOLOGY

2.1 The Pattern Recognition Algorithms for Data Mining

Data mining algorithm is to create the mechanism of data mining model. In order to create the model, algorithm will be the first to analyze a set of data and find specific patterns and trends. Algorithm uses results of this analysis to define the mining parameters of the model, the parameters used in the whole data set to extract the feasible mode and detailed statistics. In general, the traditional data mining algorithm could be categorized as the follows.

1) Bayesian algorithm. Bayesian algorithm is a kind of classification algorithm. To spill the modeling with the premise of the algorithm on the assumption that the input attributes unrelated down calculation input columns and the conditional probability between the predictable columns. Compared with other algorithms, the algorithm for computation task is small, so we can quickly generate mining model. To find out the relationship between the input columns and columns can be predicted can use this algorithm to the initial data detection and then according to results of the algorithm when using other computational cost is larger, more precise algorithm
to create other mining model. In the following formula, we define the basic process of the Bayesian algorithm that is theoretically proposed by the research group (Rudebeck, 2014).

\[ P(Z^k | X^i, m) = \prod_{i=1}^k P(z_i^k | X^i, m) = \prod_{i=1}^k P(z_i^k | x_i^k, m) \] (1)

(2) The decision tree. Decision tree is the kind of classification and regression algorithms. Used for discrete and continuous attributes in predictive modeling. The basic principle of decision tree is recursively split into subsets on the data, so that each subset contains target variable similar state. The target variable is predictable attributes. The primary information entropy is defined below where denotes the feature distribution.

\[ I(S) = I(p) = - \sum_{i=1}^n p_i \log_2 p_i \] (2)

(3) Clustering analysis method. Clustering analysis method is based on the data sets in data mining association between measures of the automatic is divided into the several clusters, make between data points within the same cluster as similar as possible, as far as possible between different clusters of data points. The discovery of outlier data can often make people find some real, but unexpectedly knowledge based on the research of the data from the group and found that the abnormal behavior and pattern, has very important significance. Generate data points corresponding to subspace and attribute weight vector from the group to stray subspace the weights of attributes in the assigned by higher, further put forward the concepts of subspace weighted distance. Based on the thought of the density to detect outliers, analysis and calculation of data objects subspace to impact factor from the group, determine whether to outliers. Algorithm can effectively adapt to detect outliers in high-dimensional data and we show the distance metric that is theoretically proposed by the research group in the literature (Tse, 2016).

\[ D(X, X') = \|W^T X - W^T X'\|^2 = \|W^T (X - X')\|^2 \] (3)

Based on the prior data mining algorithms, we will propose our own method in the later parts. Rough set theory is an extension of the set theory it mainly studies under the condition of incomplete information and incomplete data mining technology and many of the traditional data mining technology which is only for precise set and does not apply to become warped rough set, but in reality the rough set is a common phenomenon. Such a record of the decision table reduction handles the means. The object of the class with the same regularity, but such objects is not enough contracted, in order to get a more simplified decision rules need to value reduction, and these rules is the reduction of decision rules. In following formula 4, we define the decision rules that will play as the guidance of the decision.

\[ r_g: \text{des}(X) \rightarrow \text{dex}(Y) \] (4)

Where the \( \text{des}(X) \) represents the rules of the thing before, the \( \text{dex}(Y) \) represents the after a rule, they express the causal relationship. The so-called decision rule reduction is to use the existing decision-making logic, elimination algorithm necessary conditions of each decision rule. Wavelet transform is years from many different fields such as mathematics and signal processing of an integrated approach. In general, the wavelet transform is the kind of data, functions or operators into different frequency components, and then in to the study of the components that match the resolution instead of scale. Therefore, wavelet transform was used to interested in many of object is to provide practical and useful mathematical expressions.

Similarity search in data mining are presented: interested in model, the data set based on similarity measure to find similar pattern. Wavelet several different methods was applied to the similar search. First, the wavelet transform to the wavelet domain, the original data can only rely on wavelet coefficients to achieve dimension reduction and similarity search can be in wavelet domain and high efficiency. The sample mean and standard deviation are shown.

\[ x_i = \frac{1}{k} \sum_{i=1}^k x_i, \quad s_j = \sqrt{\frac{1}{k-1} \sum_{i=1}^k (x_i - x_j)^2} \] (5)

Because many decision support application test characteristics, so users in most cases do not need to be completely accurate query response and more hope the fast approximate query response, we define the search feature in the formula 6.

\[ \delta = \sqrt{\frac{1}{k-1} \sum_{i=1}^k (\log u_i - \chi)^2} \] (6)

In this formula, the \( k \) represents the total number of sample input attribute, the \( \chi \) represents the standard difference, the \( u_i \) denotes the input attribute sample. For large spatial database, the wavelet clustering algorithm is a clustering of the multi-resolution analysis. Spatial data object can be expressed in n-dimensional feature space, and the numerical character of space object can also be used feature vectors. By classifying data space
grid can not only reduce the number of data objects, and the error is very small. From a signal processing perspective, if the feature space collection of the basic objects is regarded as mundane visa number, the high frequency part of signal corresponds to a region, in the area, the distribution of the object has a fast change, and n visa number corresponds to the low frequency part of the feature space of the area of high range, in this region, the object is focused and the use wavelet transform to decompose the signal into different frequency sub-band. In the figure 2, we demonstrate the visual illustration of the data partition and classification (Wang, 2014).

2.2. The Optimization of Oracle Database Server

There are many factors affecting the Oracle database, but there's no absolute evaluation index, evaluation of Oracle database optimization effect. The survey found that refer to many factors, such as the system database of various performance indicators before and after the optimization, one of the important evaluation standard is the running efficiency of the SQL statements but in many cases, when the root cause of the concern of many users and appear in the programmer write SQL statements. Distributed application architecture of Oracle database has two kinds of distributed database system and advanced replication system. Advanced replication system based on the distributed database system, the use of the Oracle advanced replication technology in each distribution site asked update and the synchronize data. Oracle distributed database system allows the application from the local or remote database operation data. Application need to connect to a local database can transparently access remote database data, without having to know the physical location and type of the remote database. Database link is database basic object of distributed application architecture, which defines the target database and the connection of users with the distributed database system and the advanced replication system through the database link communications between sites. In the figure 3, we illustrate the basic components of the Oracle server which needs for optimization(Yang, 2014).

To optimize the server, we should consider the following issues. (1) The library structure optimization plays an important role in the whole system optimization process and it is also the most basic part, mainly includes two parts of optimization, logical structure and physical structure optimization. One of the most effective ways to improve the retrieval efficiency is to create indexes and it is an effective to improving the efficiency of the whole system several times or more. But it should be clear, not index the more the better, should be reasonable to set the position and number of the index to ensure the optimal index Settings. If often need to modify the data column on build lots of index which will cause the loss of the performance of the system and storage space of a lot of waste also could cause more serious consequences. Database allows the processing can be divided into the database server and client application processing with shared data
management by computer processing has database management system, and focuses on the application of workstation run data interpretation and display data. (2) The optimization of the program structure design. Architecture of the system is in the first need to consider the problems in the system development, according to the basic needs of the system to choose the appropriate architecture is more than important. The system structure is expanding constantly, complexity of system performance are also improved, the application of environment relative to the defined pattern before running can be a lot of complex network environment, it can be used for the user and the database. Server limited system resources cannot provide enough data please request and lead to system disorder and even system crash. A stored procedure is used to complete a specific task, if a task logical change, simply change the stored procedure without changing the client code, which can reduce the cost of software maintenance. (3) The optimization of SQL statements. The optimization of SQL statements relative hardware to improve the operation more simple and fast achieve the optimal effect which will be a good SQL statements compared with inferior SQL statement, it would be easy to contrast, we found that the good SQL statements can reduce the number of requests, it can reduce network bandwidth utilization. When dealing with the amount of data has increased dramatically, the influence of the superiority of the SQL statement effect more apparent. As previously, in many cases, problems haven't find out the real reason, after will take various methods to optimize system, it not only waste time and energy, the most important thing is to cannot fundamentally solve the problem and the optimization of SQL statements should be focused on and should be checked early which because of the relatively easier to solve. In the following code fragment, we illustrate the server optimization procedures.

![Oracle](image1.png)

**Figure 3. Basic Components of the Oracle Database Server**

![Oracle](image2.png)

**Figure 4. Oracle Server Optimization Steps and Procedures**

2.3 The Basic Features of Java and the Optimization Framework

In many traditional languages, the program was done as part of the boot process is loaded immediately. Then the initialization, then the program began to run. The initialization process of languages must be carefully controlled to ensure that the static initialization order not cause trouble. Java is cross-platform and object-oriented programming language for the distributed computing environment. Object-oriented is a natural extension of the real world model, any entity in the real world can be as object, through the message interaction
between objects, if the traditional procedural programming language is a process as the center algorithm as the drive, object-oriented programming languages on the object as the center based on message drive, all of object oriented programming languages support three concepts: encapsulation, polymorphism and inheritance, encapsulation of the Java language is stronger, because the Java without a global variable, the main function, most of the members in Java objects, Java offers users a series of classes, the Java class hierarchy, a subclass can inherit properties and methods of the parent and it is different from other object-oriented programming language is the Java only supports the single inheritance. Java mainly by the Java virtual machine level in object code implementation platform neutrality, the JVM is an abstract machine, it adhere to the specific operating system, itself has the virtual machine instructions and has its own stack such as a set of registers. We summarize the Java features as the follows.

- **Binding compiler.** JAXB binding compiler is the core of the JAXB processing model and its function is to convert or binding source XML schema set of JAXB content class. Run the JAXB binding compiler, with XML Schema as input and generates some Java derived classes, map the source of the XML Schema constraints to the Schema derived classes and JAXB reference implementation together, can make the customer application of the XML contents of Java content tree. The Java content tree composed the XML content, and according to the Schema defined in the constraint condition to test the effectiveness of the content tree.

- **Schema derived classes.** They generated by JAXB shrink translation. Some of these special class varies because of input Schema, JAXB requires developers to different Schema for different data object definition. For example, the chapter books, online order system oriented database or the employees are individually defined Schema. At compile time, because of the input Schema, JAXB compiler to generate the Schema of a derived class is also different.

- **Binding framework implementation.** JAXB binding framework implementation is the runtime API and it will provide a Java application of XML content solutions group, marshalling and test the effectiveness of the interface. Binding framework from the XML Schema source file and the binding declaration of the generated classes and interfaces, plus a few JAXB utility classes.

- **XML output file.** The XML output file output that is grouped into an XML file of XML content. In the JAXB, the marshalling content involves to parses an XML tree object, and written in an XML file, i.e. the XML, file an accurate representation and it is effective for the basic source schema. JAXB XML data can be composed of XML documents to handle content of SAX or the DOM node. In JAXB process support for XML input file according to the source Schema defined in the constraint conditions of validity check. It is optional, however, the inspection process such as your through other means that input file is valid, for performance reasons, the step during operation may choose to skip the efficacy but whenever, in before or during the efficacy is very important, because have assumed that the generated XML marshalling period. File and the source Schema effective.

- **Java application is using JAXB binding framework XML data, test and modify the content of the Java objects, marshalling to return the content of the Java web client application of XML data. Typically, JAXB binding framework packaging in a larger Java application, it provides a graphical user interface, XML transformation, data processing and other necessary functions for later processing steps.**

Because of the characteristics of Java, we can under the condition of without recompiling the other code, compile only need to modify the unit, and the modified files compiled. Class files in the path of the Java, wait until the next time the Java virtual machine reactivated, this logical because Java application will load the new changes of class files, also did update their function which is of the dynamic characteristics of Java.

### 2.4. The Finalized Big Data Analysis System

In the following figure 6, we illustrate the overview flowchart of the proposed big data analysis system. In the system, Bilinear data coordination problem of the process constraint model is a kind of special form of nonlinear model, the model of nonlinear term is the product of two variables, such as industrial group in the process of balance equation and the energy balance equation. Double linear process of the robust data coordinate calculation is relatively complex, if we can reduce complexity, the development of suitable for the actual application of the simple algorithm, bilinear process can improve the robust data coordination technology of actual use. In this paper, by using a two-step two bilinear data coordination problem can be converted to linear data coordination problems, and then use penalty function optimization method, the virtual measurement equation and the equivalent power method, the double linear process of robust data coordination problem into the square estimation problem, it's very easy, at the same time, in the process of solving considering the upper and lower bounds of the variables in order to reduce error. The location of streaming data scheduling transformation strategy is defined as the follows.
\[ x = \sum_{i=1}^{N} s_i \psi_i = \psi_s, \quad \psi = [\psi_1, \psi_2, \ldots, \psi_N] \]  

If the significant error in the process, has no bad resistance due to the least squares method, using the least squares objective function to get coordinate value of easy to be contaminated by gross errors and biases. The optimized goal could be summarized as follows.

\[ \rho_{\text{Optimized}}(x) = \frac{1}{2} (x - \bar{x})^T \sum^{-1} (x - \bar{x}) \]  

Robust data coordination does not need to be based on the core ideal statistical distribution properties of the error, so not easily polluted by the significant error. Robust data coordination purpose is to reduce the influence of gross errors, when significant error in the process of the robust function of significant error measurements give the small power this prevents the other measured values by the pollution, so the robust data coordination with resistance. Equivalent weight method can make the calculation becomes very easy, so it is widely used in the data regression in robust estimation. The corresponding equation could be defined as follows.

\[ \int_{\text{Robustness}} \rho(v) \cdot f(t) \cdot dl = \min \]  

Robust data coordination problem for the bilinear process, due to the core flow rate and the composition variables are independent measurement, unrelated. So using two-step, along the flow direction search first minimize the corresponding objective function values, and then along with the direction search component variables and the corresponding objective function values. Therefore, it is able to double the linear robust data coordination problem into the two linear robust data coordination problems.

3. EXPERIMENT

In this section, we conduct the experimental simulation on the proposed methodology. Our experiment is executed through following steps. (1) Data selection. Required data extracted from the database and its related properties. (2) Data preprocessing. The data produced in the data selection stage, according to the need for reprocessing to ensure the data integrity and consistency, loss and distortion and noise data uses data smoothing techniques for processing. (3) The large amount of data in the database samples contains the knowledge through online learning data compression in weight between the nodes. Fuzzy theory can play an important role in knowledge and the rule acquisition. Human language and thought are fuzzy, fuzzy thinking forms and a variety of language expression, perfect and efficient characteristics. There is a lot of knowledge in daily work and the study is fuzzy, the fuzzy knowledge in the control and decision making often play a huge role. From the table of the testing, we simulate our method in different datasets and obtain satisfactory result. Moreover, later, we conduct more corresponding experiment with the comparison of the other state-of-the-art methods. Through the illustrated result, our system outperforms.
<table>
<thead>
<tr>
<th>Dataset</th>
<th>MI</th>
<th>CDI-1</th>
<th>AI</th>
<th>CDI-2</th>
<th>SMOTE</th>
<th>WDeNN</th>
<th>LMMNN</th>
<th>DNet</th>
<th>CDPPD</th>
<th>TIBDEDT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Appetency</td>
<td>0.03</td>
<td>0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.03</td>
<td>0.32</td>
<td>0.30</td>
<td>0.30</td>
</tr>
<tr>
<td>Churn</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Upselling</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Ada-agnostic</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Ada-prior</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Sylva-agnostic</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Sylva-prior</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Brazil/Rouses</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Marketing</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Backache</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>BioMed</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Schino</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Flix</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Rto</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Ted</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Tr23</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Balances</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Cleveland</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Croc</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Credit</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Eccit</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Corazon</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Heart</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Hepatitis</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Hungarian</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Ionosphere</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Lymphs</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Forns</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
<tr>
<td>Primary</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
<td>0.15</td>
</tr>
</tbody>
</table>

**Friedman** α = 0.011 (Bonferroni-corrected)
4. CONCLUSION

In this research paper, we conduct research on the development and robust testing of big data analysis system based on the optimized Java technique and parallel Oracle database. Distributed database is the combination of the computer network and the database products, compared to centralized database with high scalability and reliability it can achieve resource sharing, improve the utilization rate of the data, and can realize remote control of distributed data. Oracle is with the characteristics of distributed database are in the Oracle database itself will melt of the advantages of distributed database. Big data has a large amount of data and complex data structure, data is to produce low speed, the density of the data value and these features added to the difficulty of the big data for effective analysis, large data analysis has become one of the core content of exploration data development, therefore, we must be for big data analysis of researches in connotation and extension. Our research combines the Java and database server with the optimized data mining algorithm to finalize the big data analysis challenge. The experiment indicate that our methodology is robust and effective which means the lower time-consuming. In the future, we plan to combine the cloud computing technique to extend our current down research to the cloud computing system to enhance the algorithm performance on the bigger data scope which will be meaningful.

REFERENCES


