УДК 519.256

АНАЛІЗ ВЕЛИКОГО ОБСЯГУ ДАНИХ ПРО СТАН ВИСОКОТЕХНОЛОГІЧНОГО ОБЛАДНАННЯ

АНАЛИЗ БОЛЬШОГО ОБЪЕМА ДАННЫХ О СОСТОЯНИИ ВЫСОКОТЕХНОЛОГИЧЕСКОГО ОБОРУДОВАНИЯ

ANALYSIS OF A LARGE VOLUME OF DATA ON THE STATE OF HIGH-TECH EQUIPMENT

Д.С. ШИБАЕВ, В.В. ВЫЧУЖАНИН, докт.техн.наук,
Н.О. ШИБАЕВА, канд.техн.наук
Одесский национальный морской университет, Украина

Ідеологічною основою дослідження є аналіз даних, що отримуються в результаті роботи великої кількості високотехнічного обладнання. Дані розподіляються по базах даних, в залежності від різних характеристик. Складність подальшої обробки, залежить від обсягу інформації, яку необхідно проаналізувати, а також архітектурного типу зберігання даних.

Використання технології data mining дозволяє істотно поліпшити показники аналізу інформації з наступною системою короткострокового пошуку за значенням. Використання такої технології дозволить підвищити ефективність роботи архівів суднових показників за весь час експлуатації судна. Сама технологія аналізу даних не досконала і потребує постійної модифікації, для збільшення власної ефективності.

Додавання сучасної архітектури перебору даних у базах, дозволить збільшити ефективність аналізу даних, що складаються з великої кількості показників стану судна та його обладнання. Однією з таких архітектур − є Map Reduce.

Ключові слова: *аналіз даних, суднові показники, бази даних, пошукові алгоритми, великі дані.*

Идеологической основой исследования является анализ данных, получаемых в результате работы большого количества высокотехнического оборудования. Данные распределяются по базам данных, в зависимости от различных характеристик. Сложность последующей обработки, зависит от объема информации, которую необходимо проанализировать, а также архитектурного типа хранения данных.

Использование технологии data mining позволяет существенно улучшить показатели анализа информации с последующей системой краткосрочного поиска по значению. Использование такой технологии позволит повысить эффективность работы архивов судовых показателей за все время эксплуатации судна.

Сама технология анализа данных недоскональная и требует постоянной модификации, для увеличения собственной эффективности.

Добавление современной архитектуры перебора данных в базах, позволит увеличить эффективность анализа данных, состоящих из большого количества показателей состояния судна и его оборудования. Одной из таких архитектур − есть Map Reduce.

Ключевые слова: *анализ данных, судовые показатели, базы данных, поисковые алгоритмы, большие данные.*

The ideological basis of the study is to analyze the data obtained in the result of a large number of high-tech equipment. The data is distributed in databases, depending on various characteristics. The complexity of the subsequent processing depends on the amount of information you need to perform, as well as architectural type of data storage.

The use of data mining technology allows to significantly improve the analysis of information and subsequent short-term search value. The use of this technology will improve the efficiency of the archives of marine indicators for all time of operation of the vessel. The technology of data analysis is not thorough and requires permanent modification to increase their own efficiency.

The addition of modern architecture through data in the databases, will allow to increase efficiency of data analysis, consisting of a large number of indicators of the condition of the vessel and its equipment. One of these architectures is Map-Reduce.

Keywords: *data analysis, ship performance, databases, search algorithms, big data.*

Introduction. Modern swimming facilities are equipped with a large number of various special systems. The main task of such systems is obtaining readings from operating equipment and transmission to the user terminals.

However, due to the growth in the number of shipboard systems, which utilize for its operation a modern digital algorithms, as well as taking into account the increasing automation of modern ships, increasing the number of systems of transfer of the readings over a period of time which contributes to the development and improvement of monitoring systems and processing of such information.

Analysis. Management system (SU) of the ship is a distributed (decentralized) network SU, composed of individual subsystems and hardware complexes for various purposes, the United working on a data exchange system (OXOD) included in the communication system.

From the point of view of processes in the communication system, its model contains several interacting layers.

At the base lies the transport subsystem, on which the layer of the network operating system that organizes the applications and user provisioning. On top of the operating system is the application layer. In particular, because of the special role of data warehouses and database management systems

(DBMS), this class of system applications is segregated into a separate network layer.

At the next level are working the system services that use a DBMS as a tool to search for and provide information in a comfortable for decision form to the officials of SU, and also perform some processing of information (directory service, e-mail, the system of collective work).

A centralized system of collection and storage of information, are not intended for a fast output of information received with regard to dynamic add new information.

To solve this problem, many SU are formed based on the separation of the data keyed into a database for relational and non-relational. Such a mixed architecture the base complicates the process of information processing and output of results at user request.

One of the methodologies, which helps to solve tasks of different classes of search patterns and the interpretation of the results is the methodology of data mining Data Mining. It is used to detect and explore patterns in arrays of semi-structured information and building models describing the behavior of complex systems [1].

A characteristic feature of the data analysis methods Data Mining is the use of various algorithms for finding patterns in the data.

Expansion of the set of data-mining models in various algorithmic nature can be productive in the class of problems where not accurately work classical methods: statistical, analytical or deterministic.

Each stage of the research data, we can build a finite number of hypotheses that can be confirmed or not be confirmed subsequently.

The more constructed models and descriptions are close to the hypotheses, the more we have the right to assume the accuracy of the result. Of course, in the study of real data, any conclusions you can do with a finite degree of accuracy.

When analyzing data using the methodology of Data Mining, we can build descriptive and predictive models based on multiple algorithms.

The obtained results can move us in finding hidden patterns in the data.

At the same time, then often the situation arises when future values calculated by one model differ from those values found for the other model. This means that at least one of the models gives only approximate values of the target variable.

With a groundswell of information in the world and the need to handle it in a reasonable time faced the issue of vertical scalability is the growth of processor speed has stayed at 3.5 GHz, the speed of reading from the disk also grows quiet pace, plus the price powerful servers are always more total price a few simple servers. In this situation, the ordinary relational database, even clustered on the disk array, is not able to solve the problem of speed, scalability and throughput.

The only way out − horizontal scaling where multiple independent servers connected by the fast network and each owns/handles only part of the data and/or only a portion of read requests-update. In this architecture, to increase the capacity of the storage (capacity, response time, throughput) it is only necessary to add a new server to the cluster. Procedures sharding, replication, failover (the result will be obtained even if one or more servers stopped responding), data redistribution when adding nodes involved herself NoSQL database [2].

The purpose of the study. An important factor arising in the development of systems analysis marine systems is their storage. As storage system it is necessary to use a combined structure composed of several DBMS is able to place the information received on the network.

For this, it is important to develop a data warehouse consisting of a certain number of DBMS.

The concept of data warehouse (DWH) is the idea of separating the data used for operational processing and for the solution of analysis tasks.

It allows to use data structures that meet the requirements of their storage, including use in OLTP systems and systems analysis.

This separation allows to optimize the structure of the operational data store (operational databases, files, spreadsheets, etc.) to perform operations of insertion, modification, deletion and searching, and data structures used for the analysis (for analytical queries) [3].

In order for the update operation and the reading was effective, NoSQL databases, you must use a structure with random access, such as B-trees.

So, if we abandon the arbitrary update data, and process the entire set sequentially, it is possible to achieve significant performance gains.

All of these factors, generalize the formation of such a search architecture like Hadoop.

The standard used in the Hadoop architecture is the use of MapReduce tasks (Fig.1).

At the beginning of calculations the input data set is divided into several subsets. Each subset is processed on a separate node of the cluster.

Map task on each node receives at input a set of pairs key-value and returns another set of pairs.

Next, all pairs are grouped by key, sorted, and fed to the input of the reduce task, which creates the final result or input for another map task.

This achievement consists of performing three steps:

– The Map phase: the processors and disks on each node are busy processing their data.

– The Reduce phase: the results obtained in the first stage, are transferred over the network and aggregated.

– In the third stage the results are stored in the file system (visible on the chart jump record in HDFS).

Data processing in a MapReduce architecture is a complex technical process, consisting of a large set of sequentially executed operations:

– Run application: transfer of the application code into the main (master) and slave nodes (workers).

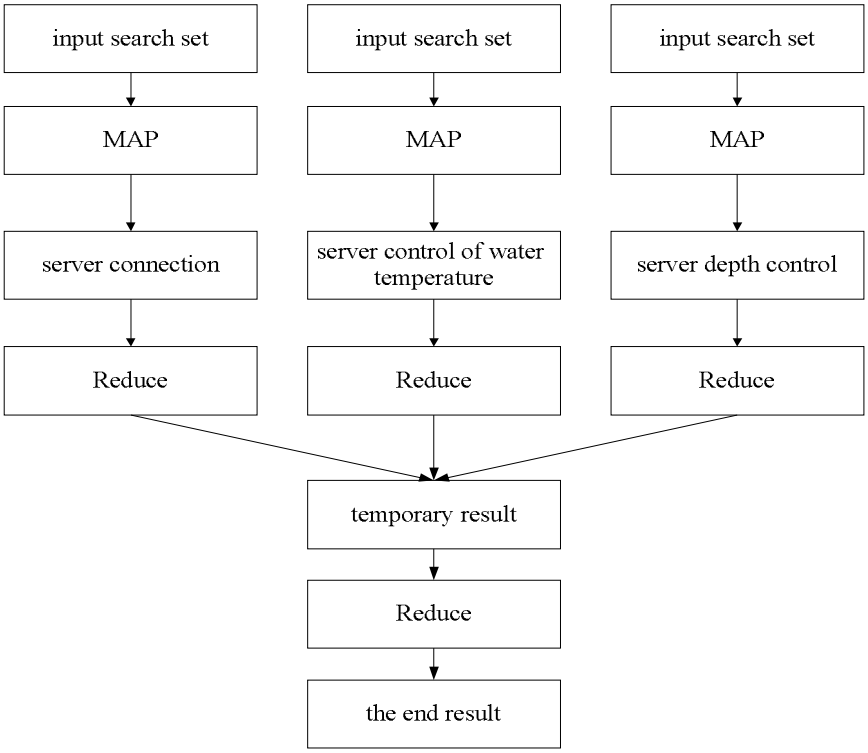

Fig. 1. MapReduce архитектура

– Master assigns specific tasks (Map or Reduce) and distributes portions of the input data to compute nodes (workers).

– Map-nodes assigned reading input and start processing.

– Map-nodes locally keep intermediate results: each node stores the result on local disks.

– Reducer nodes read intermediate data from Map and Reduce nodes perform data processing.

– Reduce-sites store the final results in output file, usually in HDFS.

One of the concepts that make it possible to establish a unified data storage system is a system of support and decision-making (DSS). In DSS, these two types of data are called respectively the operational data sources (OID) and the data store.

The basis of the method of processing and storage is factor analysis, facilitating the identification of possible used information, depending on the circumstances of its application [4].

The results of the study. The formal outcome of the first stage of the use of factor analysis – obtaining the mixing matrix and on its basis – the correlation matrix.

The confusion matrix is a table, which records the measurement results of the observable variables: in the columns of the matrix (number of variables) presented to the evaluation subjects (or one subject) of each variable; the matrix rows are different observations for each variable. If the task of the researcher is to construct a factor space for one subject, it is necessary to provide a plurality of such observations.

In that case, when the construction of the group factor space, it is enough to obtain one rating from each Respondent. For subsequent calculation according to this correlation matrix with a fairly reliable correlation coefficients should provide the necessary number of observations, i.e. the number of rows in the matrix of confusion.

This allows you to use statistical evidence of information search, starting from a similar search situations arising in the process of using a software system.

Conclusion. Develop a software solution and method of analysing large amounts of information in databases as relational and not relational, will speed up the process of iterating through the data and search information from ship systems monitoring equipment.

REFERENCES

1. *Арский Ю.М. Принципы конструирования интеллектуальных систем. / Ю.М. Арский, В.К. Финн // Информационные технологии и вычислительные системы. − № 4. − М., 2008. – С. 4-37.*
2. *Булычев А.В. Технологии интеллектуального анализа Data Mining и их использование при решении задач логистической оптимизации. / А.В. Булычев, В.Б. Бритков // Труды 51-й научной конференции МФТИ «Современные проблемы фундаментальных и прикладных наук»: Часть VII. Управление и прикладная математика. Т. 3. − М.: МФТИ, 2008. − 138 с.*
3. *Барсегян А.А. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. / А.А. Барсегян, М.С. Куприянов, В.В. Степаненко, И.И. Холод // БХВ-Петербург, 2008. – 173 с.*
4. *Березин Ф.А. Уравнение Шредингера / Ф.А. Березин, М.А. Шубин. −М.: Изд-во МГУ, 1983. – 295 с.*

Стаття надійшла до редакції 25.09.2017