УДК 004.65

Informatics and Mathematical Methods in Simulation
Vol. 8 (2018), No. 1, pp. 78-85

ОПТИМИЗАЦИЯ МЕТОДОВ ПРОГНОЗИРОВАНИЯ, ОБРАБОТКИ И АНАЛИЗА ИНФОРМАЦИИ В РАЗНОСТУКТУРНЫХ ХРАНИЛИЩАХ ДАННЫХ

Д.С. Шибаев, В.В. Вычужанин, Н.О. Шибаева, Н.Д. Рудниченко

Одесский национальный морской университет,
ул. Мечникова, 34, Одесса, 65029, Украина; e-mail: denshibaev@outlook.com

В статье предложен метод оптимизации обработки большого объема данных, предназначенный для диагностирования состояния оборудования сложных технических систем. Такое диагностирование позволяет повысить эффективность анализа неисправностей в технических системах с целью их скорейшего устранения. Это достигается за счет прогнозирования отказа критически уязвимых систем, способных вывести из эксплуатации транспортное средство, либо привести к последствиям, угрожающим жизни людей. Метод реализован в информационной системе, основанной на наборе алгоритмических согласованностей, направленных на идентификацию информации в высоконагруженных сетях и ее обработку в моменты передачи от сложных технических систем к центральному хранилищу данных. Функционирование информационной системы основывается на применении гибридных алгоритмов поиска и сортировки данных, концепции Data Mining, а также алгоритмов для работы с большими объемами данных. Методы отбора и распределения данных выполняются на уровне контроля показаний датчиков технической системы и распределяются до момента записи в хранилище данных. Это позволяет идентифицировать информацию, необходимую для построения прогноза отказа критически уязвимых технических систем. С целью увеличения скорости обработки данных, применяется язык разработки C/C++, который позволяет взаимодействовать информационной системе с датчиками контроля показаниями. В качестве хранилища данных используются реляционная и нереляционная базы данных, позволяющие сохранять большие массивы информации и обеспечивать постоянных доступ к данным. Разработанная информационная система может быть использована в транспортных средствах с используемой системой контроля показаний в реальном времени.

Ключевые слова: анализ данных, большие данные, компьютерные системы, базы данных, модульное программное обеспечение, Big Data, Data Mining, коллекции данных, гибридные хранилища данных

Введение

Надежное функционирование современных, например, судовых сложных технических систем (СТС) [1-4] часто основывается на использовании удаленных программных решений по прогнозированию их работоспособности [5,6].

Для решения подобных задач используются клиент-серверные архитектуры, на основе сервера баз данных (БД), применяемых для хранения полученных данных, а также серверные анализаторы данных. Использование такой архитектуры позволяет выявлять неисправности в работе технической системы на ранних этапах ее выхода из строя, а также уменьшить вероятность отказа технической системы при несвоевременном ее обслуживании. Работоспособность клиент-серверного хранилища данных зависит от модели и структуры хранения данных. Использование реляционной модели относится к классической нотации для разработки клиент-серверных

приложений, ориентированных на структурное хранение данных. К основным преимуществам реляционных БД следует отнести:

– наличие языка манипулирования и описания данных (sql);

– разработанную систему оптимизации запросов, способную выполнять декомпозицию запросов, обеспечивая параллельное выполнение таких работ на кластерных архитектурах;

– использование механизма контроля, включающего атомарность, согласованность, долговечность и т.д.;

– присутствие общесистемных ресурсов, обеспечивающих работу реляционной архитектуры повсеместно.

Работа реляционной архитектуры использует единую структуру передачи транзакционного запроса от пользователя с дальнейшей его обработкой системой управления базами данных (СУБД), а также модели хранения данных, применяемых в различных условиях. Главными составляющими, обеспечивающими дальнейшую работу пользователей с данными, при помощи которых выполняется проверка на права доступа, организацию, структуру и прочее, являются базы метаданных. Физическое расположение данных также обеспечивает база метаданных, формирующая схему распределения и контроля изменений в хранилище данных.

В связи с усложнением задач диагностирования состояния СТС, связанных с генерацией большего объема данных для оценки текущего состояния объектов диагностирования, использование классических реляционных хранилищ привело к существенному повышению требований к кластерным архитектурам. В результате возросли сложности задач, решаемых реляционными БД. Существующие проблемы в подобных системах стали проявляться более значимо, что привело к усугублению недостатков использования подобной архитектуры. Одной из весомых проблем стала потеря соответствия данных. Ее проявление выражается в том, что реляционные БД не позволяют хранить агрегаты в явном виде, что существенным образом влияет на распознание таблиц при их постоянном росте. В связи с этим усложняется связывание таблиц при выполнении запроса и, как следствие, осложняется формирование агрегата. Второй весомой проблемой является масштабирование данных. Это связано с использованием дорогостоящих аппаратно-программных комплексов, необходимых для работы параллельных систем БД и дальнейшей поддержкой отказоустойчивости. Чем сложнее формируются запросы в БД, тем больше падает производительность системы, связанная с межмашинным обменом данными в кластерной архитектуре.

Немаловажной составляющей, связанной с использованием реляционной архитектуры БД, является сложность модификации или реорганизации соответствующих таблиц при необходимости. Это достаточно весомый недостаток архитектуры, так как он не позволяет в кратчайшие сроки исправить ошибку, которая могла быть совершена на начальных этапах. Полное изменение всей схемы данных, а также модификация таблиц – это существенная проблема, которая может возникнуть.

Одним из решений проблемы реляционных БД является альтернативная технология, позволяющая хранить данные нереляционными методами, получившая название «NoSQL». Согласно [7], существует две причины, при которых использование нереляционной архитектуры является актуальным:

– эффективность разработки приложений (использование БД в качестве основного хранилища данных является стандартным решением, однако много времени уходит на отображение данных из структур, хранящихся в БД; использование нереляционной архитектуры может обеспечить уменьшение объемов кода, упростить последующую отладку приложения, а также обеспечить общее взаимодействие БД и программного решения);

– работа с большими объемами данных (использование реляционных архитектур является достаточно дорогостоящим и сложным решением в случаях, когда объемы

хранимых данных велики, что связано с архитектурными особенностями реляционных БД, которые изначально планировались для стационарного использования, а позже были доработаны до серверных реалий, однако использование кластерных серверов является сложным техническим условием, при котором реляционная архитектура не эффективное решение, как от него требуют разработчики) [8].

Нереляционные БД изначально созданы для использования на кластерных решениях. От этого повышается эффективность их использования, общее масштабирование, а также возможность работы с большими объемами удаленной информации. В связи с этим требуется разработать единое решение, направленное на повышение эффективности использования методов диагностирования СТС, а также обеспечить возможность удаленного прогнозирования их технического состояния.

С ростом использования технологических решений, основанных на архитектуре «Big Data», возникли противоречия, связанные с необходимостью хранения больших объемов неструктурированных данных с целью дальнейшего их структурирования методами разработки схемы БД. В результате развитие реляционных архитектур приостановлено из-за сложности добавления новых решений, а также изменившейся логики работы таких систем. Сегодня не используются прямые запросы в БД. Произошли глобальные изменения в пользовательских интерфейсах и моделях их использования, растет популярность веб-архитектур, требующих более сложные и гибкие решения в области хранения и работы с данными.

Целевой задачей Data Mining является осуществления поисковых операций, применяемых к законам поведения и общего функционирования исследуемой системы, состоящей из числовых данных. Обязательное условие, возникающее при использовании методологий Data Mining, заключается в трактовке полученных закономерностей, сформированных в результате определения практической полезности данных. В связи с этим возникает необходимость в разработке метода оптимизации большого объема данных, позволяющего распределять показания состояния СТС. Метод должен основываться на использовании разноструктурных хранилищ данных, а также работать в высоконагруженной сети [9].

Целью работы является разработка метода оптимизации обработки большого объема данных, получаемых вследствие диагностирования СТС, позволяющего обеспечить сбор и обработку больших объемов информации в высоконагруженных сетях с учетом реального времени.

Основная часть

Разрабатываемый метод основан на использовании набора алгоритмических согласованностей. Они направлены на идентификацию информации в высоконагруженных сетях и ее обработку в моменты передачи от СТС к центральному хранилищу данных. Такие решения образуют единый комплекс, являющийся информационной системой, способной функционировать как дополнительная составляющая для оценки и прогнозирования состояния СТС.

Одним из требований к информационной системе, ориентированной на анализ данных, является своевременное обеспечение аналитической информацией, необходимой для принятия решения. Все собранные данные требуют дополнительной операции обработки – очистки. При использовании очистки может происходить процесс удаления выбросов (нехарактерных и ошибочных значений), обработка отсутствующих значений параметров, выполнение численного преобразования и т.д. На рис. 1 представлен алгоритм, реализующий разработанный метод анализа большого объема данных. Использование подобного решения позволяет собирать аналитические данные и выполнять предварительную их оценку за счет использования алгоритмов интеллектуального анализа данных. Разработанное решение применяется для контроля

состояния СТС. Для предотвращения трудностей, возникающих вследствие непрерывных изменений данных и проблемы контроля за СТС со стороны аналитика, используется технология Data Mining, направленная на сбор данных с рабочего окружения технической системы, с возможностью динамического масштабирования. В целях повышения эффективности системы анализа, применены алгоритмы самостоятельного обучения, обеспечивающие системе автономность от внешнего воздействия.

Использование алгоритмов, способных обучаться в процессе своей работы, активно применяют при решении задач классификации с обучением. Классификация с обучением подразумевает следующие действия:

– подготовка данных (имеющийся набор объектов с известными метками классов разбивается на две части: обучающую выборку и тестовую выборку);

– обучение модели (параметры модели классификации подбираются на основе обучающей выборки таким образом, чтобы добиться наилучшего соответствия между предсказанными и фактическими метками классов);

– тестирование модели (полученная в результате обучения модель проверяется на достоверность, для чего вычисляется процент недостоверных результатов классификации объектов из тестовой выборки).

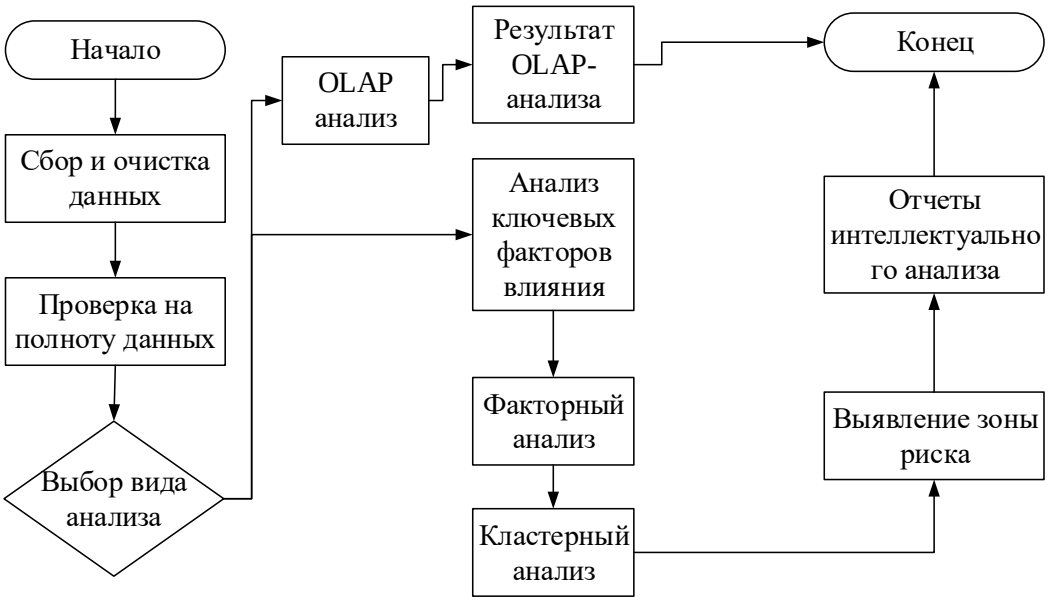

Рис. 1. Алгоритм, реализующий разработанный метод анализа данных в высоконагруженных сетях СТС

Целевой составляющей разработанной программной системы является извлечение структурированной информации из разно-структурированного хранилища данных. Такая задача является актуальной из-за повышения уменьшения работоспособности СТС при их эксплуатации в различных условиях, а также изменения конструктивной сложности таких систем. На практике применяются хранилища показаний работоспособности технической системы, имеющей строго структурированную структуру. Это обусловлено проблемой оперирования только структурированной информацией, которую используют такие методы анализа, как OLAP или Data Mining. Наиболее остро проблема извлечения информации возникает при работе с неструктурированными данными, имеющими текстовую форму, содержащими большие объемы информации и имеющие полезную форму.

Поисковый процесс включает в себя стадию поиска информационных ресурсов, являющихся релевантными задаче, с последующим извлечением из них исходных

коллекций данных. При этом ресурсы, принимающие участие в поиске, могут содержать в себе как структурированные, так и слабоструктурированные наборы данных. На следующем этапе данные пропускаются через средства анализа. Немаловажным является использование механизмов, способствующих увеличению скорости аналитической обработки. На рис. 2 представлены этапы, входящие в основу извлечения сущностей, которые формируются в разноструктурированных коллекциях данных и используются для дальнейшего анализа получаемой информации, необходимой для решения задачи классификации информации. В кластерных решениях, применяемых для обработки большого объема данных, использование коллекций является неотъемлемой составляющей при работе алгоритмов Hadoop.

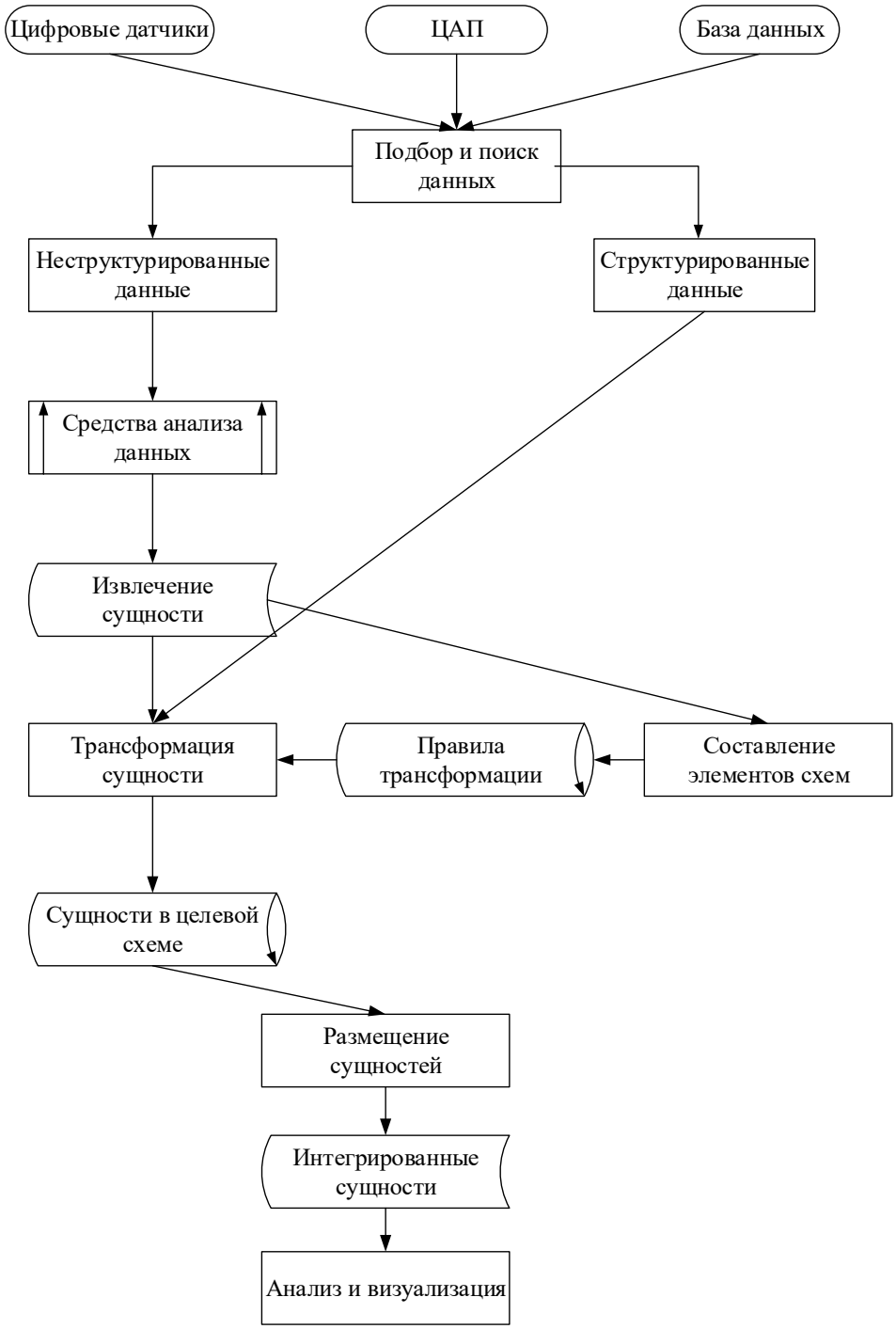

Рис. 2. Схема внутреннего анализа сущностей данных

По окончанию процесса анализа, наступает процесс интеграции собранных коллекций, онтологий и извлеченных сущностей в общую интегрированную коллекцию. Первым этапом работы такой схемы является использования сопоставления элементов исходных схем и целевой схемы, в которой применяются различные методы автоматизации. Все эти правила применяются для трансформации данных, использующих более тонкие методы интеграции. Одним из таких методов является метод расширения сущностей, позволяющий установить сходство сущностей между собой. По окончанию использования методов расширения наступает процесс слияния данных, содержащий информацию об одной сущности реального состояния СТС, сформированной из различных источников данных.

Процесс аналитической обработки, используемый для работы репрезентативной выборки, формируется из двух частей: построение модели и использование построенной модели для новых наборов данных. Процесс, строящий модель, является достаточно ресурсоемкой задачей, которая может изменяться в зависимости от применяемого алгоритма и включает в себя кеширование, сканирование, расчет множества вспомогательных параметров и т.д.

Использование уже готовых моделей к новым данным требует намного меньше ресурсов в связи с упрощенной работой с вычислением простых функций. Из этого следует, что использование небольших множеств данных может давать существенный результат при вычислении разноструктурных наборов данных и влиять на временные интервалы, выделяемые на расчет и анализ информации в базах данных. В качестве практического метода, используемого для построения и получения результатов от репрезентативных выборок, применяют сэмплинг. Его главной особенностью является сохранение скорости аналитической обработки данных без потери качества результирующего анализа.

Выводы

В результате анализа современных методов обработки разноструктурированной информации в высоконагруженных сетях разработан метод оптимизации обработки большого объема данных, решающий проблему диагностирования состояния оборудования в СТС. Основой метода является информационная система, способная извлекать и анализировать потоки информации в высоконагруженных сетях передачи данных и использовать полученную информацию для дальнейшего прогнозирования состояния СТС.

Работоспособность информационной системы основывается на применении гибридных алгоритмов поиска и сортировки данных, концепции Data Mining, а также математических алгоритмов для работы с большими объемами данных.

Список литературы

1. Вычужанин, В.В. Повышение эффективности эксплуатации судовой системы комфортного кондиционирования воздуха при переменных нагрузках. Монография / В.В. Вычужанин. — Одесса: ОНМУ, 2009. — 206 с.
2. Вычужанин, В.В. Математические модели нестационарных режимов воздухообработки в центральной СКВ / В.В. Вычужанин // Вісник Одеського національного морського університету, збірник наукових праць, 2007. — № 23. — С.172-185.
3. Рудниченко, Н.Д. Оценки структурного и функционального рисков сложных технических систем / Н.Д. Рудниченко, В.В Вычужанин // Восточно-Европейский журнал передовых технологий, Інформаційні технології. Системи управління, 2014. — Том 1, № 2 (67). — С. 18-22.

4. Vychuzhanin, V. Devising a method for the estimation and prediction of technical condition of ship complex systems / V. Vychuzhanin, N. Rudnichenko, V. Boyko, N. Shibaeva, S. Konovalov // Восточно-Европейский журнал передовых технологий, 2016. — 6 (9). — Pp. 4-11.
5. Вычужанин, В.В. Информационное обеспечение мониторинга и диагностирования технического состояния судовых энергоустановок / В.В.Вычужанин // Вісник одеського національного морського університету, збірник наукових праць, 2012. — № 35. — С. 111-124.
6. Шибаева, Н.О. Информационное обеспечение дистанционной оценки рисков сложных технических систем / Н.О. Шибаева, В.В. Вычужанин // Информатика и математические методы в моделировании, 2016. — Том 6, № 2. — С. 133-141.
7. Кудрявцев, К.Я. Методы повышения скорости поиска информации в базах данных / К.Я. Кудрявцев, А.Е. Коротков // LAP Lambert Academic Publishing, 2012. — 84 с.
8. Samet, H. The Quadtree and Related Hierarchical Data Structures / H. Samet. // ACM Comput. Surv, 1984. — Pp. 187-260.
9. Arasu, A. Efficient exact set-similarity joins / A. Arasu, V. Ganti, R. Kaushik // Proceedings of the 32nd international conference on Very large data bases. VLDB '06. VLDB Endowment, 2006. — Pp. 918-929.

ОПТИМІЗАЦІЯ МЕТОДІВ ПРОГНОЗУВАННЯ, ОБРОБКИ ТА АНАЛІЗУ ІНФОРМАЦІЇ В РІЗНОСТУКТУРНИХ СХОВИЩАХ ДАНИХ

Д.С. Шибаєв, В.В. Вичужанін, Н.О. Шибаєва, Н.Д. Рудніченко

Одеський національний морський університет,
вул, Мечнікова, 34, Одеса, 65029, Україна; e-mail: denshibaev@outlook.com

У статті запропоновано метод оптимізації обробки великого обсягу даних, призначений для діагностування стану устаткування складних технічних систем. Таке діагностування дозволяє підвищити ефективність аналізу несправностей в технічних системах з метою їх якнайшвидшого усунення. Це досягається за рахунок прогнозування відмови критично уразливих систем, здатних вивести з експлуатації транспортний засіб, або призвести до наслідків, що загрожують життю людей. Метод реалізований в інформаційній системі, заснованої на наборі алгоритмічних узгодженостей, спрямованих на ідентифікацію інформації в високонавантажених мережах та її обробку в моменти передачі від складних технічних систем до центрального сховища даних. Функціонування інформаційної системи ґрунтується на застосуванні гібридних алгоритмів пошуку та сортування даних, концепції Data Mining, а також алгоритмів для роботи з великими обсягами даних. Методи відбору і розподілу даних виконуються на рівні контролю показань датчиків технічної системи і розподіляються до моменту запису в сховище даних. Це дозволяє ідентифікувати інформацію, необхідну для побудови прогнозу відмови критично вразливих технічних систем. З метою збільшення швидкості обробки даних, застосовується мова розробки C/C++, яка дозволяє взаємодіяти інформаційній системі з датчиками контролю показників. В якості сховища даних використовуються реляційне та нереляційне сховище даних, що дозволяє зберігати великі масиви інформації і забезпечувати постійний доступ до даних. Розроблена інформаційна система може бути використана в транспортних засобах з системою контролю показників в реальному часі.

Ключові слова: аналіз даних, великі дані, комп'ютерні системи, бази даних, модульне програмне забезпечення, Big Data, Data Mining, колекції даних, гібридні сховища даних

OPTIMIZATION OF PREDICTION METHODS, PROCESSING AND ANALYSIS OF INFORMATION IN DIFFERENT STRUCTURE DATA

D.S. Shibaiev, V.V. Vychuzhanin, N.O. Shibaieva, N.D. Rudnichenko

Odesa National Maritime University,
34, Mechnikova Str., Odesa, 65029, Ukraine; e-mail: denshibaev@outlook.com

In article the method of optimization of processing of a large volume of the data intended for diagnostics of a condition of the equipment of difficult technical systems is offered. Such diagnostics allows to increase the efficiency of the analysis of faults in technical systems with a view to their early elimination. This is achieved by predicting the failure of critically vulnerable systems capable of decommissioning the vehicle, or lead to consequences that threaten people's lives. The method is implemented in an information system based on a set of algorithmic consistency aimed at identifying information in high-load networks and its processing at the time of transmission from complex technical systems to the Central data warehouse. The functioning of the information system is based on the use of hybrid algorithms for searching and sorting data, the concept of Data Mining, as well as algorithms for working with large amounts of data. The methods of data selection and distribution are carried out at the level of sensor readings control of the technical system and distributed until recording in the data warehouse. This allows to identify the information necessary for the prediction of failure of critically vulnerable technical systems. In order to increase the speed of data processing, the C\C++ development language is used, which allows the information system to interact with the reading control sensors. As a data warehouse uses the relational and not relational database, allowing to save large amounts of data and provide constant access to data. The developed information system can be used in vehicles with the system used to monitor the readings in real time.

Keywords: data analysis, big data, computer systems, databases, modular software, Big Data, Data Mining, data collections, hybrid data stores