УДК 004.65

Informatics and Mathematical Methods in Simulation
Vol. 7 (2017), No. 4, pp. 318-324

ОПТИМИЗАЦИЯ ОТБОРА И АНАЛИЗА ИНФОРМАЦИИ В РАЗНОСТРУКТУРНЫХ 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 работе предложена методология обработки большого объема информации, целью которой является уменьшение времени на прогнозирование состояния сложных технических систем, уменьшение затрат на передачу данных через спутниковые линии связи и возможность использования модульного программного продукта в качестве практического решения проблемы передачи большого объема информации через высоконагруженные спутниковые линии связи.
Использование предложенной методологии позволит сократить затраты, возникающие при передаче данных через спутниковые линии связи, а также повысить эффективность обработки большого объема информации о состоянии оборудования сложных технических систем.

Ключевые слова: анализ данных, большие объемы данных, компьютерные системы, базы данных, модульное программное обеспечение, SCADA-системы, передача данных, протоколы передачи данных, VSAT

Введение

Современные алгоритмические решения, направленные на работу с динамически-расширяемыми массивами информации, представляют собой сложные информационно-технические системы, разделенные на автономные подсистемы, основной задачей которых является реплицирование информации в центральные узлы обработки и управления данными [1,2]. Такие решения позволяют модульно наращивать хранилища данных в реальном времени, с целью дистанционной проверки сохраняемой информации удаленными аналитиками.

Сохранность целостности данных, обеспечение качественной масштабируемости системы, а также быстрый доступ к информации является актуальной задачей, решение которой позволит улучшить систему обработки больших объемов данных из сложных технических систем (СТС) [3,4].

СТС оснащается автоматизированными системами, выполняющими анализ работы их оборудования в реальном времени. Сбор данных с таких систем выполняет сервер контроля и обработки информации. Использование централизованных комплексов, интегрированных в СТС, обеспечит:

- контроль технологических процессов при взаимодействии оборудования;
- анализ состояния подсистем СТС;
- оптимизацию обслуживания систем;
- уменьшение нагрузки на обслуживающий персонал.

Известные сложные компьютерные системы (СКС) разделяются на категории, зависящие от области их применения. Одним из направлений в диагностировании с использованием СКС является анализ работоспособности СТС на транспорте [5]. Такая СКС состоит из активного сервера или пары серверов, а также линии передачи данных. Датчики параметров подсистем СТС выполняют передачу данных в строго

обозначенные периоды времени, которые могут быть заданы индивидуально для каждого типа оборудования СТС.

Накопление информации, полученной в результате работы СКС, позволяет контролировать качество и точность работы СТС, а также выполнять прогнозирование технического состояния оборудования систем. При этом возникают проблема хранения и сортировки информации, их анализа и контроля, а также сложности в накоплении информации для использования ее в различных аналитических целях [6].

Одним из оптимальных решений по хранению и накоплению информации является организация сервера баз данных, способного принимать информацию от оборудования СТС. Такая система позволяет хранить большое количество информации на протяжении длительного времени и быть доступной техническому персоналу в любое время. Для этого используют дисковые массивы большого объема, защищенные от внешнего воздействия и позволяющие добиться целостности данных.

Современные СКС, используемые на водном транспорте, представляют собой сочетания модульных интегрированных комплексов и вспомогательных контролеров управления. Такие комплексные решения являются SCADA-системами. Однако в СКС возникают проблемы, связанные со сбором и обработкой информации. Это связано с большими массивами информации, формируемыми и передаваемыми по внутренней локальной сети транспортного средства и требующими детального анализа для своевременного выявления неисправностей, анализа потребления топливных и горюче-смазочных материалов СТС [7,8].

Работа SCADA-систем основывается на промежуточном считывании информации с датчиков параметров СТС, занесением показаний в базу данных (БД) с дальнейшим архивированием показаний для анализа и прогнозирования технического состояния систем.

В связи с этим возникает необходимость создания единого решения, способного гарантировать качественную запись данных с датчиков контроля параметров подсистем СТС в централизованную БД. Такое решение должно обладать повышенной отказоустойчивостью, гарантировать проверку целостности данных, а также обеспечивать простое масштабирование системы.

Одним из классических решений такой задачи является использование реляционных БД, обладающих набором возможностей, способных улучшить качество процесса записи информации. Это позволяет:

- упростить проектирование модели данных;
- упорядочить структуру таблиц;
- сохранить целостность данных;
- контролировать дублирование данных.

Такие БД отказоустойчивы и обеспечивают хранение больших объемов информации без дополнительного контроля. Использование их в качестве центрального хранилища СКС улучшит показатели сохранности информации, а также гарантирует целостность данных при передаче информации от мест контроля технического состояния оборудования до записи информации в реляционную БД. Результатом использования реляционной БД в составе SCADA-системы (рис. 1) станет возможность сохранения [9]. При этом возникает проблема с передачей данных в удаленные центры обработки информации, которые обладают достаточной вычислительной мощностью, необходимой для работы с большими объемами накопленной информации за время эксплуатации транспортного средства. Передача большого объема данных, хранимых в реляционных БД, требует сохранения структуры таблиц данных для удаленных аналитических центров, а также необходимость в высокоскоростной линии передачи данных.

Из анализа систем сбора и хранения информации о состоянии СТС следует, что проблемой современных СКС, работающих по архитектурному принципу SCADA,

является невозможность обеспечения быстрой передачи данных для их обработки. Причина заключается в монолитности используемых архитектур хранения информации и большой нагрузке на каналы передачи данных. Это делает невозможным анализировать данные о техническом состоянии оборудования СТС в режиме реального времени [10-12].

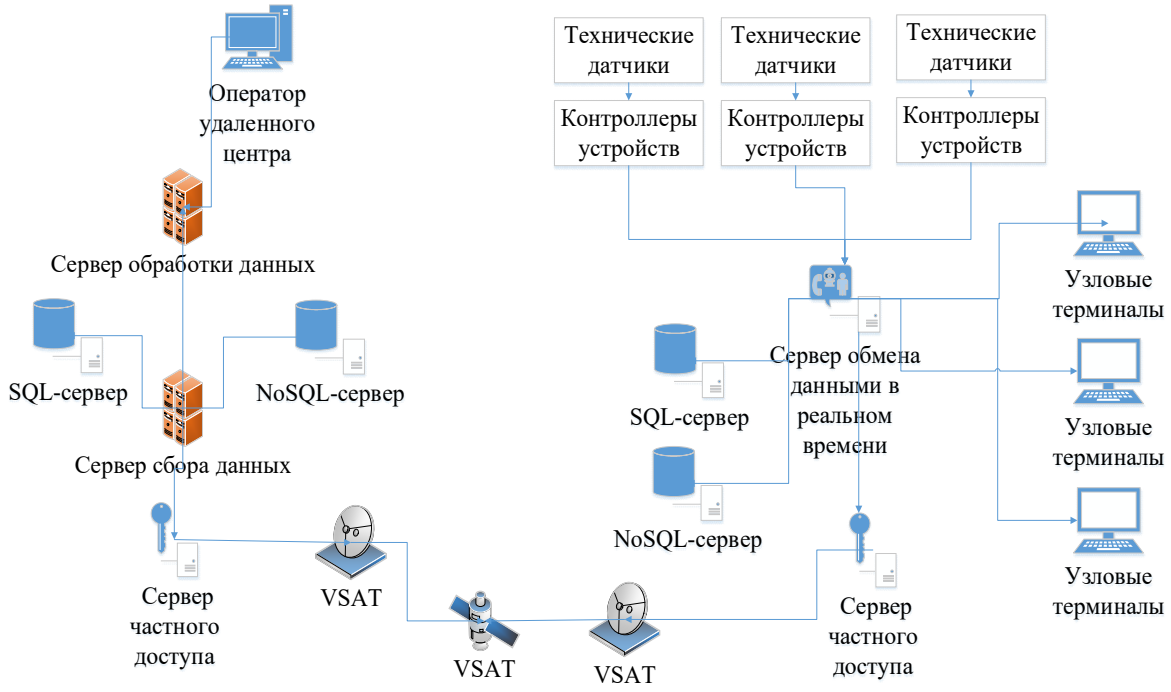

Рис. 1. Схема распределения хранения и анализа данных в SCADA СКС

Целью работы является разработка методологии передачи и обработки большого объема информации в реальном времени, включающей поиск, отбор и анализ данных, необходимых для прогнозирования технического состояния оборудования СТС.

Основная часть

Первоначальным этапом разработки решения по передаче и обработке большого объема информации, является формализация задач с использованием известного программного решения. Основные решаемые задачи:

– разработка алгоритма перехвата и промежуточного анализа потока передаваемой информации технического состояния СТС к серверу хранения результатов;

– подсчет исходных метаданных в пакете;

– разработка алгоритма распределения пакетов согласно считанным метаданным;

– разработка алгоритма архивирования данных согласно нормам передачи через спутниковые линии;

– построение модели передачи данных, основанной на минимизации размера передаваемого пакета;

– разработка алгоритма проведения анализа данных, необходимых для построения прогноза состояния СТС;

– прогнозирование состояния СТС, основанное на полученных данных из SCADA-систем.

В качестве программной реализации задач, сформированных при разработке передачи и обработки большого объема информации, необходимо создать модульное

программное решение, допускающее его использование как части функционального модуля SCADA-системы.

Начальным этапом реализации модульного программного решения является формулирование программных требований, выполняемых средствами разработки программных продуктов. Для этого используются инструменты проектирования, в состав которых входят CASE-средства. Было выполнено построение диаграммы вариантов использования, отражающей программное взаимодействие подсистем СТС в модульном программном решении (рис. 2).

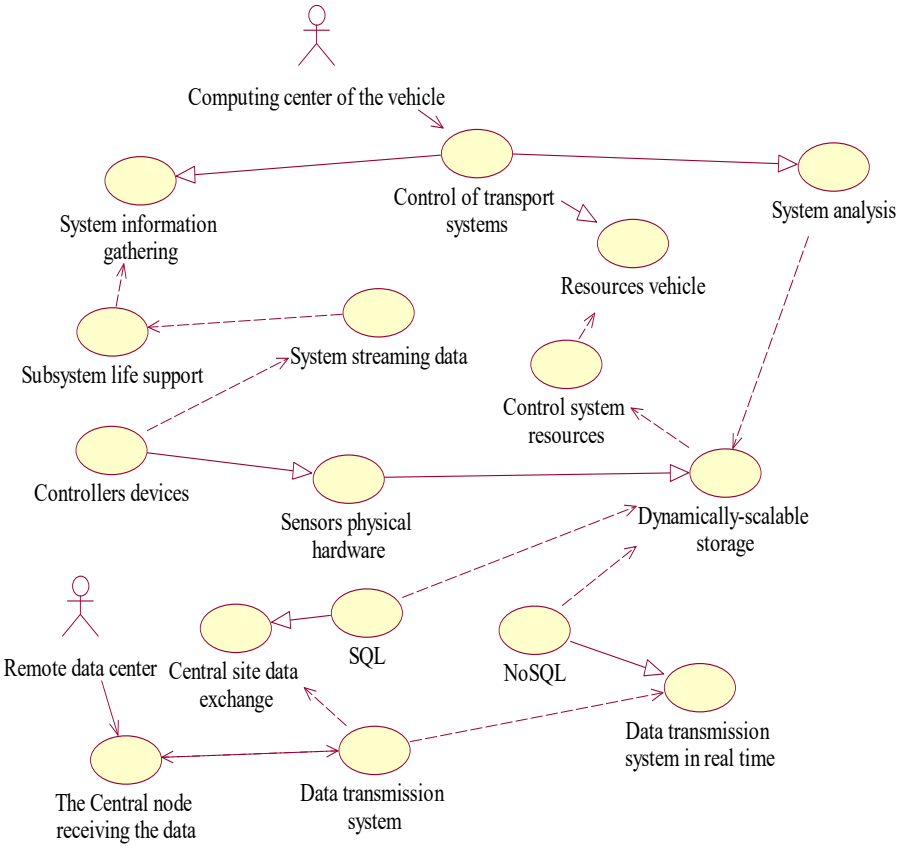

Рис. 2. Диаграмма взаимодействия подсистем СТС в модульном программном решении

Под модульностью программного решения подразумевается использование отдельных функциональных частей, осуществляющих работу с потоками информации, передаваемой по сетям, считывание и идентификацию метаданных пакетов информации, использование метода распределения информации согласно полученным метаданным и т.д. Такая комплексная реализация создаст объемное, монолитное решение, которое сложно интегрировать в SCADA-систему из-за разнообразности использования технологического оборудования СТС и средств контроля за их работоспособностью.

Для определения последовательности использования модулей программного решения был разработан обобщенный алгоритм обработки большого объема информации в реальном времени (рис. 3).

Для данных, которые являются приоритетно важными, необходимо предоставить отдельное хранилище, построенное на нереляционной архитектуре (NoSQL). Такое хранилище позволит добиться максимальной скорости записи информации, а также минимизировать сжатый архив информации для дальнейшей передачи по спутниковым каналам. В качестве используемого типа NoSQL хранилища предлагается использовать «bigtable-подобную» базу данных, использующую семейства колонок в качестве

хранилища информации. Такая архитектура БД похожа на классическую реляционную базу данных, однако самым большим отличием является представление колонок в виде разреженной матрицы, строки и столбцы которой используются в качестве ключей. Использование «bigtable-подобной» БД подходит для взаимодействия с большим объемом информации, что делает модульную систему достаточно сложным средством обработки больших массивов данных.

Рис. 3. Обобщенный алгоритм работы модульного программного обеспечения

В качестве способа распределения метаданных предлагается использовать прямую классификацию, позволяющую выполнять сортировку по типам датчиков сбора информации. Это позволит увеличить точность отбора данных, а также сфокусировать прогноз состояния СТС на конкретном оборудовании.

Для минимализации передаваемых данных был выбран алгоритм - деревья Хаффмана, который является одним из самых подходящих алгоритмов сжатия с контролем целостности данных. Он основывается на использовании кодов Хаффмана, при помощи которых можно сократить названия и системные маркировки датчиков, что уменьшит размер передаваемого массива.

В качестве спутниковой системы передачи данных рассматривается архитектура, основывающаяся на VSAT, способная передавать данные в реальном времени на скорости 4 мбит/сек. Такое решение анализирует данные, полученные от контроллеров датчиков физических устройств СТС, и использует такую информацию в качестве входных данных для построения различных прогнозов.

Выводы

Из анализа устройства СКС, входящих в состав SCADA архитектур, можно сделать вывод о необходимости разработки методологии обработки большого объема информации. Результатом является реализация модульного программного решения, способного улучшить итоговый процесс прогнозирования состояния СТС. Такое решение позволит внедрить NoSQL в средства оценки технического состояния систем, а также использовать улучшенные средства обработки больших массивов информации.

Список литературы

1. Кормен, Т. Алгоритмы: построение и анализ / Т. Кормен, Ч. Лейзерсон, Р. Ривест. — М: МЦНМО, 2002. — 960 с.
2. Кнут, Д.Э. Искусство программирования. Т.3. – Сортировка и поиск / Д. Кнут. — М.: Изд. Дом «Вильямс», 2010. — 788 с.
3. Вычужанин, В.В. Повышение эффективности эксплуатации судовой системы комфортного кондиционирования воздуха при переменных нагрузках / В.В. Вычужанин. — М: Одесса, ОНМУ, 2009. — 206 с.
4. Vychuzhanin, V. Assessment of risks structurally and functionally complex technical systems / V. Vychuzhanin, N. Rudnichenko. // Eastern-European Journal of Enterprise Technologies. — 2014. — Т. 1, № 2 (67). — С. 18–22.
5. 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. // Eastern-European Journal of Enterprise Technologies. — 2016. — Т. 6, № 9 (84). — С. 4–11
6. Finkel, R A. Quad trees a data structure for retrieval on composite keys / R.A. Finkel, J.L. Bentley. // Acta Informatica. — 1974. — Pp. 1–9.
7. Андреев, Е.Б. SCADA-системы: взгляд изнутри / Е.Б. Андреев, Н.А. Куцевич, О.В. Синенко. — М.: РТСофт, 2004. — 176 с.
8. Прошин, Д.И. Проблемы выбора инструментальных средств построения SCADA-систем / Д.И. Прошин, Л.В. Гурьянов // Информатизация и Системы Управления в Промышленности. — 2010. — № 1 (25). — С. 21–25.
9. Ефимов, И.П. SCADA-система TraceMode / И.П. Ефимов, Д.А. Солуянов. — Ульяновск: УлГТУ, 2010. — 158 с.
10. Кудрявцев, К.Я. Методы повышения скорости поиска информации в базах данных / К.Я. Кудрявцев, А.Е. Коротков. // LAP Lambert Academic Publishing, 2012. — 84 с.
11. Samet, H. The Quadtree and Related Hierarchical Data Structures / H. Samet. // ACM Comput. Surv. — 1984. — Pp. 187–260.

12. 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 ВІДБОРУ ТА АНАЛІЗУ ІНФОРМАЦІЇ В РІЗНОСТРУКТУРНИХ СХОВИЩАХ ДАНИХ

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

Одеський національний морський університет,
вул, Мечнікова, 34, Одеса, 65029, Україна, denshibaev@outlook.com

В роботі запропоновано методологію обробки великого обсягу інформації, результатом якої є зменшення часу на прогнозування стану складних технічних систем, зменшення витрат на передачу даних через супутникові лінії зв'язку і можливість використання модульного програмного продукту в якості практичного вирішення проблеми передачі великого обсягу інформації через високонавантажені супутникові лінії зв'язку. Використання запропонованої методології дозволить скоротити витрати, що виникають при передачі даних через супутникові лінії зв'язку, а також підвищити ефективність обробки великого обсягу інформації, отриманої в результаті передачі масиву даних від контролерів станів обладнання складних технічних систем.

Ключові слова: аналіз даних, великі об’єми даних, комп'ютерні системи, бази даних, модульне програмне забезпечення SCADA-системи, передача даних, протоколи передачі даних, VSAT

OPTIMIZATION OF SELECTION AND ANALYSIS OF INFORMATION IN DATA WAREHOUSES RESTRUCTURING

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

Odesa National Maritime University,
34, Mechnikova Str., Odesa, 65029, Ukraine, denshibaev@outlook.com

Proposed methodology processing of a large volume of information, the result of which is to reduce the time to forecast the state of complex technical systems, reducing costs for data transmission via satellite links and the ability to use modular software product as a practical solution to the problem of delivering large amounts of high-load information via satellite links. Using the proposed methodologies will reduce the costs incurred when transmitting data via satellite links, as well as to improve the efficiency of processing a large amount of information obtained as a result of transfer of array data from controller condition of the equipment complex technical systems.

Keywords: data mining, big data, computer systems, databases, modular software, SCADA, data transmission, data transfer protocols, VSAT