УДК 004.62:004.75

**ПОДХОДЫ К** **ХРАНЕНИЮ И ОБРАБОТКЕ БОЛЬШИХ ОБЪЕМОВ СТРУКТУРИРОВАННЫХ ДАННЫХ**

**Быченков А.К., Матчин В.Т.**

Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет информационных технологий, радиотехники и электроники» (МИРЭА), 119454, Россия, г. Москва, проспект Вернадского, 78, e-mail: bychenkov.a.k@edu.mirea.ru, matchin@mirea.ru

В эпоху цифровизации и информационного бума данные стали ключевым ресурсом для принятия решений и получения ценной информации. Современные организации сталкиваются с огромными объемами структурированных данных, и эффективное их хранение является ключевой задачей. С каждым днем в связи с ростом объемов, разнообразием и скоростью поступления данных из различных источников сложность этой задачи возрастает. В статье будут рассмотрены подходы к хранению и обработке больших объемов структурированных данных.

Ключевые слова: большие данные, хранение структурированных данных, озеро данных

**APPROACHES TO STORING AND PROCESSING LARGE VOLUMES OF STRUCTURED DATA**

**Bychenkov A.K, Matchin V.T.**

*Federal State Educational Institution of Higher Education “Moscow State University of Information Technologies, Radio Engineerng and Electronics” (MIREA), 119454, Russia, Moscow, Vernadscogo avenue, 78 e-mail: bychenkov.a.k@edu.mirea.ru, matchin@mirea.ru*

**In the era of digitalization and the information boom, data has become a key resource for decision-making and obtaining valuable information. Modern organizations face huge amounts of structured data, and storing it efficiently is a key task. Every day, due to the growing volume, variety and speed of data receipt from various sources, the complexity of this task increases. The article will consider approaches to storing and processing large amounts of structured data.**

Key words: big data, structured data storage, data lake

В современном мире существуют различные подходы к хранению и обработке больших объемов структурированных данных. Но до определенного момента, практически единственным ответом на вопрос «как хранить и обрабатывать данные?» являлось использования реляционных СУБД. Это может быть PostgreSQL, MySQL, Oracle, SQL Server.

Способы хранения данных в СУБД можно разделить на 2 категории: ориентация на строки и ориентация на столбцы. Изменяя способ хранения данных на жестком диске компьютера, можно повлиять на производительность базы данных и ее оптимизацию для транзакционных или аналитических рабочих нагрузок.

Данные, которые необходимо хранить могут измеряться в терабайтах и петабайтах, их необходимо хранить максимально эффективно. Для этого могут применяться различные алгоритмы сжатия данных: zstd, lz4, zlib. Каждый алгоритм позволяет сконфигурировать уровень сжатия. Чем выше уровень сжатия, тем дольше оно происходит. Фактически сжатие данных позволяет улучшить показатели производительности ввода/вывода и уменьшить размер данных за счет ресурсов процессора. Поэтому всегда нужно соблюдать баланс между доступными ресурсами процессора и требуемой скоростью сжатия. Есть два типа сжатия в СУБД:

1. Сжатие данных на уровне таблицы — применяется ко всей таблице. Доступно для таблиц как со строковой, так и с колоночной ориентацией данных.
2. Сжатие данных на уровне столбца — применяется к отдельному столбцу. Позволяет использовать различные алгоритмы сжатия для разных столбцов одной таблицы. Этот тип сжатия доступен только для таблиц с колоночной ориентацией данных.

Другим параметром хранения данных является надежность, исключающая минимизирующая риск потери. Надежность определяется устойчивостью системы к отказам оборудования. Для этого экземпляры СУБД запускают на множестве компьютеров – в кластере, настраивают репликацию данных. В данном случае используемая СУБД должна поддерживать репликацию. Реализация репликации базы данных сопряжена с рядом проблем, например:

1. Согласованность данных: обеспечение согласованности данных между репликами может оказаться сложной задачей, особенно в сценариях, где данные часто меняются. Синхронизация и распространение изменений требует сложных алгоритмов и методов, позволяющих избежать конфликтов и поддерживать согласованность.
2. Задержка репликации: поскольку репликация часто предполагает распространение изменений на удаленные реплики, между основной базой данных и репликами может возникнуть небольшая задержка. Минимизация задержки репликации может иметь высокий приоритет, особенно в сценариях, где критически важны данные в реальном времени. Использование синхронной репликации базы данных может помочь свести задержки репликации к минимуму.
3. Задержка сети: на репликацию данных в географически распределенных местоположениях может влиять задержка в сети. Обеспечение эффективной передачи данных и минимизация задержек становится критически важным для поддержания быстродействующей и надежной системы репликации.
4. Масштабируемость: по мере увеличения объема данных и количества транзакций обеспечение масштабируемости системы репликации становится непростой задачей. Масштабирование для удовлетворения растущей рабочей нагрузки, например, во время праздничной распродажи, при сохранении производительности и оперативности — это тонкий баланс.
5. Решение конфликта: когда одновременные обновления происходят как в основной базе данных, так и в реплике, необходим механизм, позволяющий определить, какое обновление преобладает. Выбор неправильной стратегии может привести к потере данных или повреждению информации.

Данные необходимо хранить надежноДля обеспечения отказоустойчивости, по

Список литературы *(источники располагаются по алфавиту)*

1. Костогрызов А.И., Степанов П.В. Инновационное управление качеством и рисками в жизненном цикле систем формационных систем / - М.: Изд-во ВПК, 2008. - 404 с.

2. ITILv3.Глоссарий терминов и определений, ITIL® V3 Glossary Russian Translation v0.92, 30 Apr 2009.

3. Бирюков А.Н. Лекции о процессах управления информационными технологиями / М.: Интуит, 2010.- 216 с.

References (*порядок источников соответствует русскому варианту)*

1. Kostogryzov AI Stepanov PV Innovative quality management and risk management in the life cycle of systems, information systems / - M .: Publishing House of the military-industrial complex, 2008. - 404 p.

2. ITILv3.Glossary terms and definitions, ITIL® V3 Glossary Russian Translation v0.92, 30 Apr 2009.

3. Biryukov A.N. Lectures on management processes of information technology / M .: Intuit, 2010.- 216 p.