УДК 004.65

СОВРЕМЕННЫЕ ПЛАТФОРМЫ ДЛЯ РЕАЛИЗАЦИИ БОЛЬШИХ ДАННЫХ

М. Приведенцев, студент IV курса направления «Бизнес-информатика» Саранского кооперативного института (филиала) автономной некоммерческой образовательной организации высшего образования Центросоюза Российской Федерации «Российский университет кооперации»

Научный руководитель: **С. С. Голяев**, кандидат педагогических наук, доцент, заведующий кафедрой информационных технологий и математики Саранского кооперативного института (филиала) автономной некоммерческой образовательной организации высшего образования Центросоюза Российской Федерации «Российский университет кооперации»

Рассматриваются проблемы выбора и использования средств работы с большими данными.

Ключевые слова: СУБД, большие данные, СУБД «NoSQL».

По мере роста применения средств работы с большими данными к СУБД предъявляются все более высокие требования с точки зрения производительности и масштабируемости. На протяжении последних лет реляционные СУБД играли ключевую роль во многих областях деятельности, однако современным приложениям нужна функциональность, не свойственная этим системам, в том числе возможность изменения схем, а также поддержка многообразия типов и моделей данных. Помимо этого, у новым СУБД должна быть возможность элегантно, экономично и автоматически масштабироваться.

Элегантность – возможность добавления узлов по мере роста объемов данных для сохранения гарантированной скорости выполнения запросов.

Автоматизм – способность сбалансировано распределять данные по мере добавления узлов.

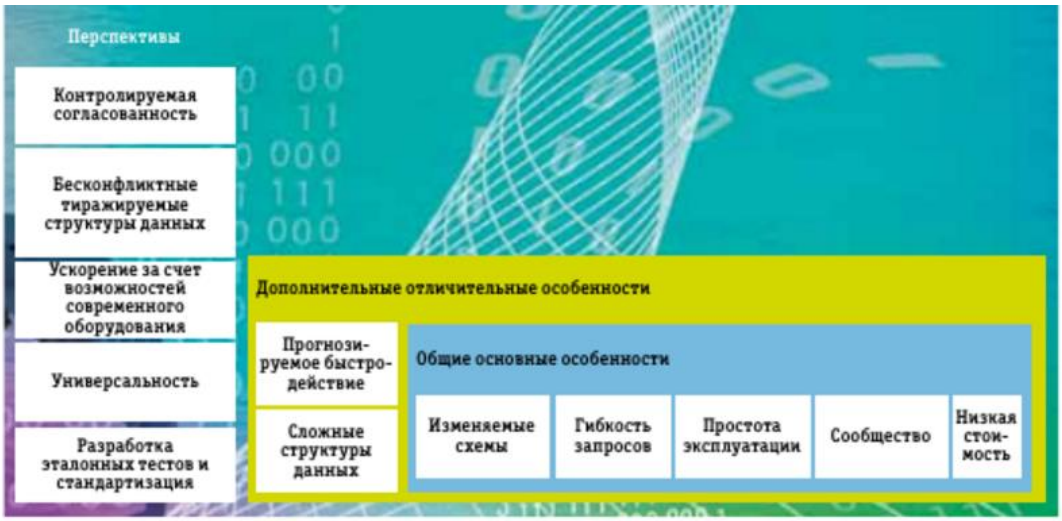
А благодаря *экономичности* затраты на развертывание должны снижаться по мере совершенствования аппаратного обеспечения. Иначе говоря, экономия, достигнутая за счет снижения затрат на вычисления и хранение, распространяется на общую стоимость внедрения и эксплуатации СУБД.

Все три названные характеристики присущи СУБД NoSQL, и хотя их название нацелено на противопоставление базам SQL, к нереляционным могут относиться и системы, поддерживающие как SQL, так и другие способы опроса. Системы NoSQL были созданы, чтобы расширить, а не заменить функционал реляционных СУБД.

По сути, многие свойства NoSQL-систем не новы и не уникальны. Главное отличие в том, что в системах NoSQL акцент делается на другом наборе функций.

Системы NoSQL бывают операционными и аналитическими. О первых говорят редко, и их особенности известны хуже, хотя системы NoSQL стремительно развиваются и функциональные границы между ними стираются. Начали появляться системы, поддерживающие многие модели данных, и эта тенденция сохраниться, причем ожесточенная конкуренция на рынке СУБД позволяет предположить, что разработчики продуктов NoSQL будут расширять круг их применений, заполняя пустые ниши.

Из рисунка можно понять, какие характеристики привлекают заказчиков в системах NoSQL. Главные общие черты для большинства систем NoSQL и их характеристик обычно отмечают изменяемые схемы данных, гибкость запросов, простоту эксплуатации, наличие сообщества и низкую стоимость.



Текущие характеристики и прогнозируемые направления развития систем NoSQL

Изменяемые схемы. Во многих современных приложениях схема данных часто меняется. Управление подобными данными с помощью традиционных реляционных СУБД, не предполагающих изменения однажды созданной схемы, невозможно: в достаточно сложной среде даже для простого изменения схемы, например для добавления одного столбца, может потребоваться неделя. В системах NoSQL этой проблемы нет: в них часто используется документная модель данных, когда база представляет собой коллекцию документов, или модель «ключ-значение», в которой данные представлены в виде соответствующих пар. В таких системах можно вначале загрузить данные, а уже потом определять и переопределять схемы. Гибкие модели данных позволяют, держать в одной и той же коллекции две записи с переменным числом описательных атрибутов. Процессы управления данными протекают проще, а времени на внедрение новой функциональности требуется гораздо меньше.

С использованием гибких схем тесно связана еще одна особенность – возможность применения гибких методов поиска по базе запросов, составленных в свободной форме на основе регулярных выражений, и поиска по ключевым словам. Гибкость запросов особенно полезна, когда система NoSQL служит одновременно и как репозиторий метаданных для описания гетерогенных срезов данных, и как средство обнаружения этих срезов. Такой сценарий применения преобладает в организациях, в которых данные хранятся разрозненно.

Идеальным вариантом было бы существование единственной оптимальной схемы данных для всей информации предприятия, которую использовали бы все приложения. Но этой идиллии не суждено воплотиться в жизнь, по крайней мере пока не будет предложен простой способ обнаружения связей между разобщенными гетерогенными данными. Сегодня эти связи устанавливаются путем создания метаданных, описывающих основные данные и сохраняемых в системе NoSQL. Гибкие механизмы опроса позволяют выполнять поиск в метаданных по ключевым словам. Гибкие механизмы

опроса и схемы позволяют анализировать квазиструктурированные срезы данных и проводить предварительный анализ. К системам NoSQL, которым присуща такая гибкость, относятся документные хранилища MongoDB, CouchDB и MarkLogic.

Большинство систем NoSQL открыты, что дает дополнительные стимулы для совместного планирования дальнейшего развития продукта. Вокруг систем NoSQL также нередко образуется активное сообщество пользователей.

Росту сообщества способствует простота развертывания и тестирования многих систем NoSQL, которые разрабатываются без расчета на обязательное наличие в организации администратора базы данных. Простота использования – большой плюс по сравнению с тяготами освоения некоторых традиционных реляционных СУБД. Усилия по повышению продуктивности труда разработчиков приложений, вкладываемые создателями NoSQL-систем, сполна окупаются – небольшие проекты нередко разрастаются в крупные, занимая постоянное место в экосистеме предприятия. В подобных случаях разработчики приложений, некогда впервые попробовавшие поработать с той или иной системой NoSQL, становятся ее проводниками в своей организации. Именно поэтому создатели систем NoSQL стараются обеспечить максимальную простоту начала использования.

Некоторые системы NoSQL работают только в облаке, но нередко предприятиям нужна гибридная модель, при которой часть сервиса действует локально. Реализация системы NoSQL в такой форме – задача сложная, поскольку нужно достичь баланса между общей стоимостью развертывания и обеспечением требуемого быстродействия, приватности и безопасности. Необходимо, чтобы общая стоимость владения системой NoSQL при фиксированном объеме данных уменьшалась пропорционально совершенствованию оборудования. Универсальность Некоторые специалисты по СУБД считают, что универсальность недостижима и для каждого класса приложений необходимо разрабатывать свои движки по обработке данных. Но внедрять слишком много платформ данных в пределах одного предприятия

непрактично: наличие отдельной платформы для различных документных хранилищ, приложений, работающих с ключами и значениями, поточными данными, а также для поддержки разных видов аналитической обработки (графы, реляционная модель и т. д.) усложнит администрирование, при этом вырастет общая стоимость развертывания, снизятся темпы внедрения новых приложений и сервисов, работающих с данными.

Направления дальнейшей работы над системами NoSQL можно предположить исходя из их основных особенностей, не забывая о повышении производительности и универсальности. Контролируемая согласованность Ранние системы NoSQL обычно не отвечали стандартным требованиям к транзакционным системам – ACID (atomicity, consistency, isolation, durability – атомарность, согласованность, изолированность, долговечность), применяя вместо этого модель BASE (basic availability, soft state, eventual consistency – базовая доступность, негарантированное сохранение состояния и возможная согласованность). Эта модель предъявляет менее строгие требования к согласованности, допуская временное расхождение копий одних и тех же данных, вследствие чего в определенных ситуациях увеличивается доступность распределенной системы.

Одно из интересных направлений дальнейших исследований состоит в том, чтобы помочь разработчикам приложений разобраться с последствиями выбора различных уровней согласованности для разных сценариев. Стоит также изучить влияние различных вариантов согласованности на быстродействие в условиях многоцентровых сред с разными нагрузками – например, с большим количеством операций записи либо чтения. Бесконфликтное тиражирование структур. В средах, допускающих конкурентный доступ к сложным структурам данных, нужны механизмы управления такими структурами, особенно когда для отказоустойчивости применяется тиражирование, в том числе в системах, распределенных между несколькими ЦОД с большой задержкой передачи между центрами.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Джигнеш Пател. Операционные СУБД NoSQL: сегодня и завтра. Журнал «Открытые системы СУБД (№3 2016).