УДК 004.65

Михасик Е. И.

ИСПОЛЬЗОВАНИЕ NOSQL-ТЕХНОЛОГИЙ НА ПРИМЕРЕ APACHE CASSANDRA

Белорусский национальный технический университет,
г. Минск, Республика Беларусь
Научный руководитель: ст. преподаватель Астапчик Н. И.

Каждый год объем информации, которая окружает человека, растет в геометрической прогрессии. Чтобы ее сохранить для будущих поколений используются базы данных, которые позволяют пользователю с помощью запросов получать необходимые данные по интересующей теме.

Последнее время среди баз данных наибольшей популярностью пользуются реляционные базы данных – это совокупность взаимосвязанных таблиц, каждая из которых содержит информацию об объектах определенного типа. Строка таблицы содержит данные об одном объекте (например, товаре, клиенте), а столбцы таблицы описывают различные характеристики этих объектов – атрибутов (например, наименование, код товара, сведения о клиенте). Записи, т. е. строки таблицы, имеют одинаковую структуру – они состоят из полей, хранящих атрибуты объекта.

Несмотря на понятную структуру, работая с реляционными базами данных можно столкнуться с целым набором проблем:

- реляционные базы плохо масштабируются;
- проектирование крупных баз данных с множеством компонентов требует значительных усилий и усложняет понимание самой структуры базы данных;
- эволюция схемы данных занимает долгие часы, так как в реляционных базах данных упор делается именно на ее структуру (таблицы).

Именно рост информации, которую необходимо уместить в базе данных, и решение вышеотмеченных проблем привели к появлению революционной идеи создания NoSQL баз данных.

В основе данного подхода лежит теорема CAP, которая содержит в себе три базовых свойства и только два из них можно получить единовременно:

- согласованность данных (Consistency) – все данные должны быть полными и непротиворечивыми;
- доступность (Availability) – максимально возможная скорость ответа сервера на запрос для записи и чтения;
- устойчивость к разделению (Partition tolerance) – в случае разделения системы на несколько частей каждая из них, если она доступна, должна быть в состоянии работать автономно, отдавая корректный отклик и предоставляя свои данные.

Основная проблема при работе с NoSQL-технологией – это отсутствие четкой схемы данных и на первый взгляд может оставлять чувство хаоса, однако, все данные, которые заносятся в базу имеют именованные поля и в них хранятся данные определенного типа.

Самым ярким представителем NoSQL-технологии является база данных Apache Cassandra, первая версия которой была разработана для работы Facebook еще в 2008 году.

В Cassandra приложение работает с пространством ключей, что соответствует понятию схемы базы данных в реляционной модели. В этом пространстве ключей могут находиться несколько колоночных семейств, что соответствует понятию реляционной таблицы. В свою очередь, колоночные семейства содержат колонки, которые объединяются при помощи ключа в записи.

Колонка состоит из трех частей: имени, метки времени и значения. Колонки в пределах записи упорядочены.

В отличие от реляционной базы данных, никаких ограничений на то, чтобы записи содержали колонки с такими же именами, как и в других записях – нет. Также в последних версиях Cassandra появилась возможность выполнять запросы определения и изменения данных при помощи языка SQL,

а также создавать вторичные индексы. Конкретное значение, хранимое в Cassandra идентифицируется:

– пространством ключей – это привязка к приложению (предметной области). Позволяет на одном кластере размещать данные разных приложений;

– колоночным семейством – это привязка к запросу;

– ключом – это привязка к узлу кластера. От ключа зависит на какие узлы попадут сохранённые колонки;

- именем колонки – это привязка к атрибуту в записи. Позволяет в одной записи хранить несколько значений.

Данные распределяются по узлам кластера двумя способами.

Первый: от ключа берется хэш и делится на диапазоны.

Второй распределяет ключи по узлам по порядку. Для каждого пространства ключей задается уровень репликации – на какое количество узлов должны быть записаны данные. Чем больше уровень, тем безопаснее, но дольше запись.

Узлы кластера равноправны, и клиент может соединяться с любым – каждый узел умеет определять реплики, на которых лежат нужные данные, и запрашивать их. Узел, который обрабатывает запрос от клиента, называется координатором.

Схема работы NoSQL-технологии позволяет хранить информацию на разных носителях, однако необходимо, чтобы к этим носителям был постоянный доступ.

Таким образом, NoSQL-технологии обладают большими возможностями хранения, лучше поддаются масштабированию и обладают простой формулировкой запросов, однако, они очень привязаны к конкретной СУБД и необходимо особое внимание уделить проектированию модели данных.