ОБРАБОТКА БОЛЬШИХ ОБЪЕМОВ ДАННЫХ

Казаков Олег Эдуардович

магистрант, ФГБОУ ВО «Восточно-Сибирский государственный университет технологий и управления», РФ, г. Улан-Удэ

Аннотация. Работа посвящена методам обработки больших объемов данных. В рамках статьи описаны принципы работы с большими данными, а также методы, позволяющие повысить скорость обработки данных.

Ключевые слова: большие данные, обработка данных, обработка больших объемов данных, модель MapReduce.

Эксплуатация систем, для работы которых необходима база данных (БД), подразумевает что в базу постоянно будут записываться различные данные. Со временем объем БД, необходимой для работоспособности системы, настолько увеличится, что возникнет проблема со скоростью обработки данных.

Работа с базами данных большого объема отличается от работы с обычными базами и требует особого подхода т.к. обрабатывать большие объемы данных, используя только метод увеличения мощности оборудования является далеко не самым практичным.

Принципы работы с большими данными

1. **Горизонтальная масштабируемость**. В связи с тем, что рост объемов данных предсказать довольно сложно – любая система, которая предназначена для обработки таких данных, должна быть расширяемой. То есть должна быть возможность беспрепятственно увеличивать производительность оборудования в случае необходимости.

2. **Отказоустойчивость**. Из принципа горизонтальной масштабируемости следует, что машин в кластере может быть много. В свою очередь это означает, что в процессе эксплуатации некоторые из них гарантированно будут выходить из строя. Поэтому система должна учитывать возможность таких сбоев и уметь справляться с ними без каких-либо серьезных последствий.

3. **Локальность данных.** В больших распределённых системах данные распределены по большому количеству машин. Необходимо по возможности хранить и обрабатывать данные на одном и том же сервере т.к. расходы на передачу данных могут превысить расходы на саму обработку. Поэтому данный принцип является одним из важнейших.

На сегодняшний день существует множество методов обработки больших объемов данных, которые позволяют повысить масштабируемость и не требуют постоянного обновления оборудования.

Возможности СУБД

Современные базы данных имеют обширный функционал, использование которого позволяет значительно увеличить скорость обработки данных:

Предварительный обсчет данных. Сведения, которые планируется использовать для анализа можно предварительно обсчитать в свободное время, тем самым подготовив их для дальнейшей обработки.

Кэширование таблиц. Для сокращения количества обращений к дисковой памяти наиболее часто используемые в процессе анализа данные можно кэшировать в оперативную память.

Разбиение таблиц на разделы и табличные пространства. Разбиение таблиц на разделы можно осуществить таким образом, чтобы при обращении к данным было минимальное количество операций с дисками. Также можно размещать на отдельных дисках данные, индексы и вспомогательные таблицы (т.к. невозможно одновременно считывать и записывать данные на один и тот же диск). Это позволит СУБД считывать и записывать данные параллельно.

Кроме того, повысить скорость считывания информации из базы данных можно следующими способами: индексирование, построение планов запросов, параллельная обработка SQL запросов и т.п.

Использование нескольких моделей

Известно, что чем более простые механизмы анализа используются, тем быстрее данные анализируются. Основная идея следующая: не тратить время на обработку того, что можно не анализировать.

Следуя данному принципу в процессе обработки часть данных постепенно отсеивается. Вначале используют наиболее простые алгоритмы т.к. объем данных очень велик. Затем по мере отсеивания данных используют все более сложные алгоритмы до тех пор, пока не будут обработаны все исходные данные. В результате общее время, необходимое для обработки всех данных, уменьшается на порядки.

Параллельная обработка

Исходные данные можно разбить на несколько групп по общим признакам. Например, можно выделить группы клиентов, товаров. Затем вместо построения одной сложной модели для всех данных, можно для каждой группы построить одну более простую модель.

Таким образом благодаря описанному подходу значительно повышается скорость анализа и снижаются требования к памяти т.к., во-первых, данные обрабатываются не целиком, а по частям, во-вторых используются более простые алгоритмы, в-третьих при таком подходе обработку данных можно осуществлять параллельно для нескольких групп данных.

Репрезентативные выборки

Основная идея данного метода заключается в том, чтобы не строить одну сложную модель на весь объем исходных данных. Вместо этого подготавливается некоторое подмножество данных, для которого будет построена модель. Затем построенная модель обрабатывает оставшийся набор данных. Таким образом экономится значительная часть ресурсов т.к. применение готовой модели к новым данным значительно эффективней.

Примеры существующих моделей

A/B testing – методика, в которой контрольная выборка поочередно сравнивается с другими. Тем

самым удается выявить оптимальную комбинацию показателей.

Regression – набор статистических методов для выявления закономерности между изменением зависимой переменной и одной или несколькими независимыми. Часто применяется для прогнозирования и предсказаний. Используется в data mining.

Simulation – моделирование поведения сложных систем часто используется для прогнозирования, предсказания и проработки различных сценариев при планировании.

MapReduce – модель предложена компанией Google. С помощью данной модели можно эффективно решать ряд задач. Например, когда необходимо подсчитать сколько раз каждое слово встретилось в тексте. Кроме того, в данной модели соблюдены все описанные принципы работы с большими данными.

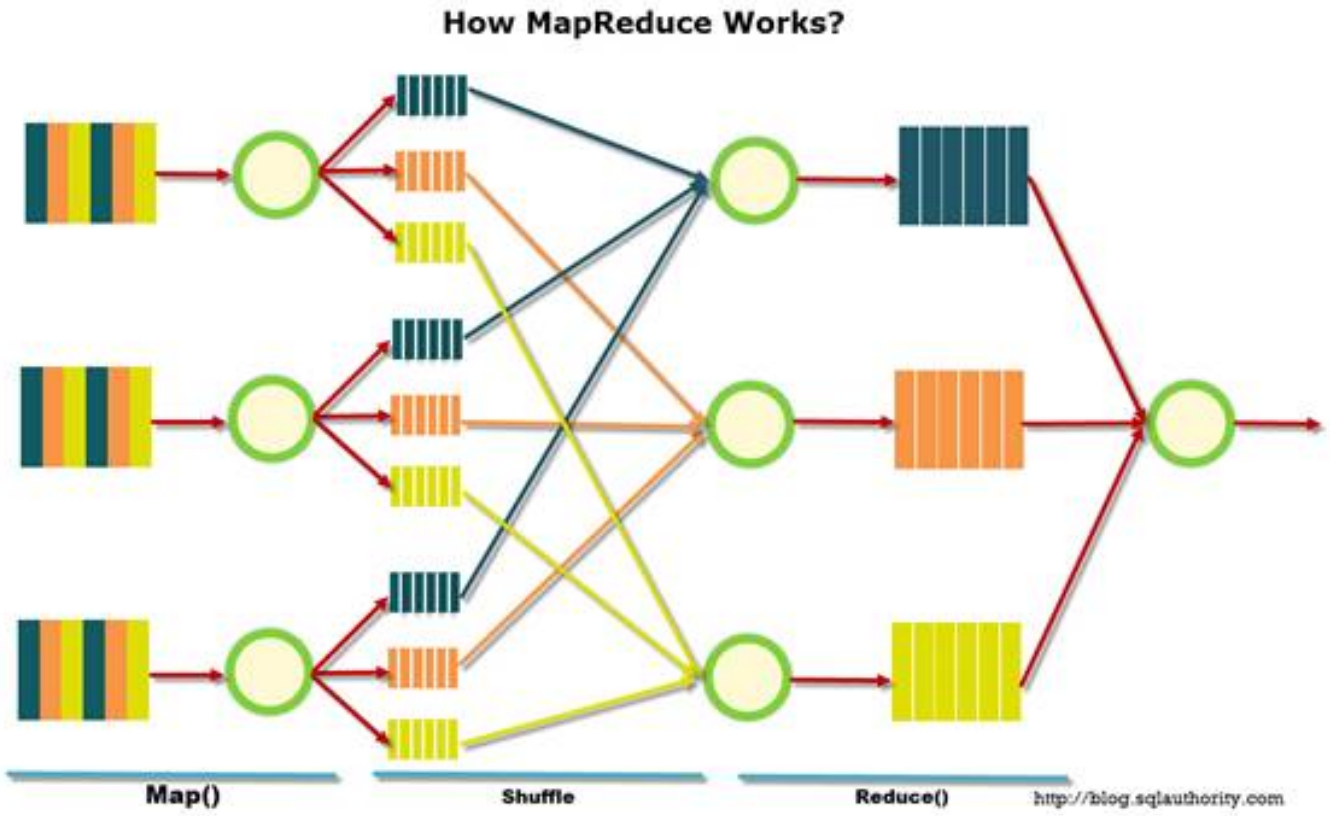

Рисунок 1. Схема работы MapReduce

Заключение

Обработка больших объемов данных – это задача, актуальность которой с каждым днем только растет. В связи с этим интерес к инструментам сбора, обработки, управления и анализа больших объемов данных проявляют едва ли не все ведущие ИТ-компании.

Описанные в статье методы позволяют оптимизировать работу с базой данных и повысить эффективность обработки больших объемов данных, что позволяет решать поставленную задачу с приемлемой скоростью. Однако это только малая часть существующих на сегодняшний день методов для обработки больших объемов данных.

Список литературы:

1. Билл Фрэнкс. Революция в аналитике: Пер. с англ. / Билл Фрэнкс. – Альпина Паблишер, 2018 – 75 с.

2. Анализ больших объемов данных [Электронный ресурс]. – Режим доступа: https://basegroup.ru, свободный.

3. Big Data – Buzz Words[Электронный ресурс]. – Режим доступа: https://blog.sqlauthority.com, свободный.