

УДК 004.65

ОПЫТ ВНЕДРЕНИЯ КОЛОНОЧНОЙ БАЗЫ ДАННЫХ SLICKHOUSE В ПРОЦЕСС ПЕРВИЧНОЙ ОБРАБОТКИ ДАННЫХ

Еременко М.И. (Университет ИТМО)

Научный руководитель – к. ф.-м. н., доцент Старовойтов А.А.
(Университет ИТМО)

Аннотация. В докладе рассматривается опыт практического внедрения колоночной базы данных SlickHouse в процесс расчета показателей активности работы пользователей с одним из продуктов ООО “Дневник ру”.

Введение. В аналитических системах, где подавляющее большинство SQL операции имеют тип “чтение” для большого объема данных, традиционные СУБД имеют проблемы с производительностью. Для таблиц на несколько десятков тысяч записей уже необходимо создавать индексы, чтобы получать приемлемую производительность. Данная проблема может быть решена путем использования колоночных СУБД, которые позволяют увеличить скорость выполнения SQL запросов в несколько раз, а также уменьшить занимаемое данными место на диске.

Основная часть. Для аналитики данные можно запрашивать у приложения либо в уже агрегированном состоянии, либо нет. При запросе агрегированных данных у системы, которая занимается их анализом, не возникает потребности в дополнительных расчетах. Но, если у разработчиков аналитической системы нет прямого доступа к СУБД, в которой хранятся данные, то при изменении условий расчета показателей или добавлении новых, необходимо обращаться к администратору этой базы данных, что в свою очередь может замедлять процесс разработки нового функционала.

Использование не агрегированных данных снимает нагрузки с администратора базы данных, данный подход может быть реализован множеством способом от прямого подключения приложения к необходимой базе данных, до использования полных бэкапов нужной базы данных.

В работе рассматривается вариант перехода на запрос не агрегированных данных и их расчет посредством колоночной базы данных. Был произведен анализ нескольких колоночных баз данных основываясь на потребностях приложения “Мониторинг активности Дневник.ру” компании ООО “Дневник ру”, по средствам которого было принято решение использовать базу данных SlickHouse, разрабатываемую компанией Яндекс. Однако данное решение имеет ряд ограничений, такие как плохая поддержка оператора JOIN и отсутствие серверных скриптов, но плюсы в виде открытого исходного кода, поддержки и развитии продукта со стороны Яндекс, а также высоких показателей производительности, перевешивают ограничения. Так как первый релиз SlickHouse был выпущен в 2016 году, то данная СУБД является достаточно молодой и находится в процесс постоянных доработок, из-за чего можно говорить, что ограничения, представленные выше, могут быть исправлены в последующих релизах.

Также, в работе рассматривается архитектура разработанного решения по внедрению SlickHouse в процесс расчета данных.

Выводы. При сборке и расчете статистики работы приложения возникают проблемы с получением актуальных данных для расчетов. В работе рассмотрено решение, используемое в одном из продуктов ООО “Дневник ру”, которое заключается в использовании колоночной базы данных в процессе расчета показателей, используемых для аналитики данных. На рынке

СУБД представлено обширное количество колоночных баз данных, для решения проблемы рассматриваемой в данной работе была выбрана колоночная база данных ClickHouse путем анализа и сравнения ее с конкурентами.

Еременко М.И. (автор)

Подпись

Старовойтов А.А. (научный руководитель)

Подпись