

Исследование производительности различных систем управления базами данных в контексте аналитических запросов

Гониченко Н.И. (ИТМО), Верзаков А.Ю. (ИТМО), Харитонов А.Е. (ИТМО)

Научный руководитель – Харитонов А.Е. (ИТМО)

Введение. Системы управления базами данных (СУБД) позволяют эффективно хранить и обрабатывать большие объёмы данных. Помимо хранения данных важным пунктом работы с ними является аналитика. Аналитика данных позволяет компаниям понять закономерность тех или иных процессов, выявить наиболее частые запросы пользователей в следствие чего правильно распределить ресурсы. Но аналитические запросы имеют характерные особенности в следствие чего им стоит уделять отдельное внимание. В данном докладе будет проведено исследование производительности различных СУБД в контексте аналитических запросов.

Основная часть. Целью аналитических запросов является выявление дополнительной информации из существующих данных [1]. То есть они направлены на извлечение большого объема данных и последующую агрегацию, что влечёт большую нагрузку на СУБД. Эта нагрузка не является целевой, потому что не влияет на работоспособность доменной модели. Следовательно, имеет смысл по возможности разделять основную СУБД и СУБД для аналитики. Из этого следует, что изначально нужен какой-то объём данных.

В данном исследовании изначально будет заполнена СУБД PostgreSQL. Схема будет отражать доменную модель приложения для авиакомпании. Затем будет выполнен процесс по извлечению этих данных, их преобразованию и загрузке в другие СУБД. Трансформация обусловлена тем, что для выполнения аналитических запросов обычно используется объединение данных из нескольких таблиц, что и влечёт большие потери по производительности. В аналитических СУБД данные, как правило, денормализованы, что позволяет эффективно решать задачи аналитики. В качестве аналитических СУБД упор будет сделан на колоночные СУБД поскольку основная рабочая нагрузка состоит из непредвиденных запросов, затрагивающих большие объёмы исторических данных. При использовании архитектуры с хранением данных по столбцам СУБД приходится считывать только атрибуты, требуемые для обработки заданного запроса [2]. Запросы будут выполняться на одной физической машине, чтобы результаты можно было считать адекватными.

Выводы. Проведено исследование производительности различных СУБД в контексте аналитических запросов. Результаты проведенного исследования помогут выбрать СУБД для задачи аналитики данных.

Список использованных источников:

1. П.А.Курапов, А.Ф. Мелик-Адамян. Оптимизация аналитических запросов в гетерогенных системах – 2021.
2. Майкл Стоунбрейкер, Угур Кетинтемел, перевод: Сергей Кузнецов. «Один размер пригоден для всех»: идея, время которой пришло и ушло - 2005