

МОДЕРНИЗАЦИЯ АРХИТЕКТУРЫ СИСТЕМЫ ОБРАБОТКИ И ВИЗУАЛИЗАЦИИ ДАННЫХ МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ

Автор – Д.К. Орлова

(Университет ИТМО, Санкт-Петербург)

Научный руководитель – А.В. Гаврилов

(Университет ИТМО, Санкт-Петербург)

Визуализация данных — это представление информации в удобной для анализа графической форме. Системы визуализации активно используются в медицинских, социологических и прочих исследованиях. Использование различного рода графиков и схем позволяет быстрее и эффективнее анализировать полученную в ходе исследований информацию. В ходе исследований приходится работать с большим объемом данных. Проблема эффективного получения, хранения и обработки такого объема данных является актуальным вопросом. От реализации данных операций зависит доверие к системе (данные не должны быть потеряны ни при каких обстоятельствах, ведь каждая порция данных важна в исследовании, так как может влиять на результат) и удовлетворенность пользователей системой (система должна показывать приемлемое время отклика, то есть данные для визуализации должны быть получены как можно быстрее).

Рассматриваемая система обработки и визуализации данных состоит из следующих компонентов: модуля получения измененных данных из разрозненных файловых хранилищ, модуля обработки данных и модуля визуализации. Система загружает данные из файлов в единую базу данных для дальнейшей их обработки и визуализации. Для ускорения визуализации в системе применяется система кэширования, в которой все данные одного направления исследования хранятся вместе, при этом обновление данных в кеше происходит только после их вытеснения и загрузки заново. Измененные данные загружаются из файлов по расписанию, соответственно задержка получения данных пользователями при активной работе с данными в лучшем случае может достигать пары часов.

С течением времени изменились требования к объему данных и частоте их обновления. Обновление данных производится все чаще и текущая реализация системы не позволяет получать их вовремя, а для сохранения производительности приходится увеличивать количество оперативной памяти сервера, при этом масштабируемость системы ограничена.

Целью данной работы является модернизация архитектуры системы получения, хранения и обработки информации данных исследований для их последующей визуализации, которая бы решала поставленные выше проблемы. Для обеспечения расширяемости системы она должна быть горизонтально масштабируемой, для этого было решено изменить систему кэширования существующей системы и сделать ее распределенной; для увеличения показателей ее средней работ использовать 2-х уровневый кеш, устройство которого рассматривается в работе. Первый уровень кэширования представляет собой разработанный автором распределенный кеш данных в куче Java-машин, позволяющий получать данные датасетов в распределенной системе в кратчайшее время. Для решения проблемы большой задержки обновления данных в системе визуализации, было решено отслеживать изменения в исходных файлах постоянно, а не по расписанию и использовать поток сообщений, который бы получал измененные данные из исходных файлов, затем отправлял их в единое хранилище и в систему для дальнейшей обработки. Для того, чтобы обновление данных в системе визуализации не было замедлено необходимостью ожидания вытеснения данных из кэша, было решено реализовать систему кэширования с поддержкой инкрементных обновлений.

В результате проведенного анализа способов работы с большими объемами данных была разработана концепция архитектуры системы получения, хранения и обработки

информации данных исследований для их последующей визуализации. На основе разработанной концепции был реализован прототип, который показал, что разработанная архитектура сильно уменьшает задержку получения обновленных данных, увеличивает время отклика системы за счет усовершенствования системы кэширования и позволяет горизонтально расширять системы за счет подключения новых узлов обработки и кэширования данных.