УДК 004.415.2

ОБРАБОТКА ДАННЫХ ИЗ РАЗНОРОДНЫХ ИСТОЧНИКОВ ДЛЯ СИСТЕМЫ АНАЛИЗА МЕДИЦИНСКИХ ИССЛЕДОВАНИЙ

Орлова Д.К. (Университет ИТМО, Санкт-Петербург)
Научный руководитель – Гаврилов А.В.
(Университет ИТМО, Санкт-Петербург)

Целью данной работы является унификация форматов данных медицинских исследований из разнородных источников и уменьшение задержки их получения для последующего анализа. Система будет решать проблемы объединения данных из различных источников разных форматов в единое хранилище с унифицированным форматом, эффективного анализа данных и обеспечение актуальности данных в системе.

Медицинские исследования содержат большое количество разнородных данных, которые часто хранятся в разных источниках – файлах различных форматов, реляционных и NoSQL базах данных. Для успеха исследования необходимо анализировать весь имеющийся набор данных анализов и жизненных показателей пациентов за все время наблюдения. Для получения более полной картины эффективности препаратов и влияния на организм человека лекарственных веществ данные разных исследований рассматривают в совокупности. Для возможности проведения анализа как одного, так как группы исследований необходимо преобразовывать данные из разных источников к одному виду. Для обеспечения высокой скорости анализа необходимо организовать хранение данных исследований в централизованном хранилище и предусмотреть эффективную систему кэширования, чтобы преобразования и извлечения данных не происходили каждый раз, как они понадобились для анализа. Актуальность и полнота данных в системе анализа медицинских данных является одним из важнейших аспектов, так как отсутствие даже небольшой части показателей может приводить к неправильным выводам исследователей и дополнительной работе исследователей, следовательно система должна быть надежной и отказоустойчивой, чтобы из-за сбоев в системе не возникало ситуаций, когда порция данных из файлов не была записана в центральное хранилище.

Для решения вышепоставленных проблем было решено разработать архитектуру системы обработки данных медицинских исследований, которая будет состоять из системы ETL (системы извлечения, преобразования и загрузки данных), центрального хранилища данных и системы эффективной обработки данных. ETL система использует механизм потоковой обработки данных для уменьшения задержек в получении новых данных и изменении уже имеющихся. Система предоставляет возможность вручную задавать соответствие между данными из источника и поддерживаемым стандартом данных. Система запоминает выборы пользователей (стандартизированное поле и соответствующее название колонки, заданное пользователем). В случае настройки соответствия для нового файла будет происходить поиск полей из файла в уже имеющихся настройках и пользователю будут предлагаться на выбор возможные варианты. Для определения обновленных, удаленных или новых данных было решено использовать механизм хеширования. Ключевые поля записей таблиц определяются отдельным потоком на основе уникальности имеющихся в базе полей. Для каждой строки с данными вычисляется хэш код по ключевым полям и хэш код по всем полям, на основе которых происходит определение является ли порция данных новой или измененной. Для увеличения показателей обработки данных из центрального хранилища было решено использовать 2-х уровневый кеш, устройство которого рассматривается в работе. 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. Разработанная архитектура является распределенной,

что позволяет производить горизонтальное масштабирование системы в случае возрастания объема данных и обеспечивает высокую надежность системы.

В результате проведенного анализа существующих подходов и инструментов для реализации ETL систем и способов работы с большими объемами данных была разработана концепция архитектуры системы получения, хранения и обработки информации данных исследований из разнородных источников. Для тестирования системы был разработан прототип, удовлетворяющий разработанной архитектуре. Тестирование прототипа показало, что система успешно производит приведение данных из различных источников к стандартизированному виду, своевременно обнаруживает измененные данные. Задержка получения данных из источников является допустимой, а потерь данных не происходит в связи с их дублированием на нескольких узлах системы.