Data Warehouse VS Data Lake

13.04.2023

Data Warehouse VS Data Lake

Хранилище данных (англ. Data Warehouse, DWH) – удобное решение для предприятий и организаций, принципы работы которого мы решили осветить в нашей сегодняшней статье. Опираясь на собственный опыт в построении хранилищ данных для финансовых организаций мы так же постараемся максимально доступно представить все выгоды использования DWH, а так же проведем сравнение с его «конкурентом» – облачным хранилищем.

Хранилище данных представляет собой предметно-ориентированную информационную базу данных, которые обрабатываются и хранятся в едином аппаратно-программном комплексе, обеспечивающем быстрый доступ к оперативной и исторической информации, многомерный анализ данных, получение прогнозов и статистики в разрезах согласованной нормативно-справочной информации. Строится на базе систем управления данными и систем поддержки принятия решений. Поступающие в хранилище данные, как правило, доступны только для чтения.

Зачастую традиционная архитектура хранилищ имеет трехуровневую структуру, состоящую из следующих уровней:

• Нижний уровень, содержащий сервер базы данных. Используется

для извлечения данных из множества различных источников.
• Средний уровень, содержащий сервер OLAP. Преобразует данные в структуру, лучше подходящую для анализа и сложных запросов.
• Верхний уровень – уровень клиента. Здесь содержатся инструменты, используемые для высокоуровневого анализа данных, создания отчетов.

Приведем список выгод от внедрения корпоративного хранилища данных:

1. Появление информационной системы хранения данных с использованием единой справочной информации;
2. Возможности проведения всестороннего анализа бизнеса;
3. Возможности проведения анализа с использованием исторических данных;
4. Возможность соединения и анализа информации, ранее хранившихся в разных информационных системах;
5. Возможность анализа и скрещивания разных по роду данных;
6. Появление основы для более качественного расчета себестоимости услуг.

Однако, компании все чаще переходят на облачные хранилища данных, вместо традиционных локальных систем. Облачное хранилище данных (англ. Data Lake) – это модель облачных вычислений, предусматривающих хранение данных в Интернете с помощью поставщика облачных вычислительных ресурсов, который предоставляет хранилище данных как сервис и обеспечивает управление им. Здесь имеет место ряд отличий от традиционных хранилищ: облачные хранилища данных быстрее и дешевле настроить и масштабировать, они могут выполнять сложные аналитические запросы гораздо быстрее благодаря использованию массовой параллельной обработке. А также нет необходимости покупать физическое оборудование.

Но, не смотря на то, что облачные хранилища данных – это большой шаг вперед по сравнению с традиционным подходом к архитектуре, пользователи по-прежнему сталкиваются с рядом проблем при их настройке:

• Загрузка данных в облачные хранилища нетривиальна, а для крупномасштабных конвейеров данных требуется настройка, тестирование и поддержка процесса ETL. Эта часть процесса обычно выполняется сторонними инструментами;
• Чтобы не допустить снижения производительности запросов обновления, вставки и удаления должны выполняться осторожно;

• Трудно иметь дело с полуструктурированными данными. Их необходимо нормализовать их в формате реляционной базы данных, что требует автоматизации больших потоков данных;
• Как правило, в облачных хранилищах данных не поддерживаются вложенные структуры;
• Оптимизация кластера: для достижения оптимальной работы необходимо постоянно пересматривать и при необходимости дополнительно настраивать конфигурацию;
• Необходимость оптимизации запросов в связи с тем, что пользовательские запросы могут не соответствовать передовым методам и, следовательно, будут выполняться намного дольше;
• Несмотря на то, что поставщики хранилищ данных предоставляют множество возможностей для резервного копирования данных, этот процесс требует мониторинга и пристального внимания.