Авито

Анализ больших данных компании Avito на базе аналитической платформы Vertica

Обзор

Компания Avito, крупнейший в России сайт объявлений, стремится собирать и сохранять всю информацию, имеющую отношение к ее бизнесу. Большие данные для нее — реальность, в условиях которой необходимо добиваться успеха. Сотрудникам Avito изо дня в день приходится иметь дело как с гигантскими объемами, так и с бесконечными потоками самых разнообразных и сложных данных.

Задача

В настоящее время данные поступают из 26 различных систем, включая основной интернет-сайт и другие веб-проекты (например, Domofond.ru), системы CRM, программные решения для управления баннерами, рассылки писем и SMS, а также компоненты мобильных приложений и многое другое. Наглядным примером больших данных, поступающих в корпоративную сеть с очень высокой скоростью, может служить кликстрим — поток событий, отражающий все действия пользователей на сайтах: клики, переходы на другие страницы и прочие действия. В настоящее время фиксируется почти 1 млрд действий в сутки, иногда до 2–3 млн действий в минуту. Наиболее разрозненными являются данные из бэк-офисных систем компании, в которых хранятся объявления клиентов: их структура стремительно усложняется и становится многоуровневой. Накапливаются и внешние данные — из систем контекстной рекламы «Яндекс.Директ», Google Adwords, «Рубикон», ряда других источников маркетинговых данных, из социальных сетей, а также из Европейского центробанка о курсах всех европейских валют.

«В компании приживается подход, при котором топ-менеджеры, прежде чем принимать решения, внимательно изучают данные, проверяя различные гипотезы»

ИВАН ГУЗ
BI-директор
Avito

Использовавшиеся до 2013 года инструменты, функционирующие на базе PostgreSQL и Microsoft Excel, не позволяли ни собирать, ни обрабатывать данные кликстрима. С их помощью удавалось только подготовить достаточно простую отчетность.

В середине 2013 года, вскоре после того как Avito заняла лидирующую позицию в своем рыночном сегменте, поглотив двух конкурентов — проекты OLX.ru и Slando.ru, в компании было создано подразделение бизнес-аналитики. Именно на него руководство Avito возложило задачу монетизации данных компании на основе глубокого анализа поступающей информации.

Avito

Краткий обзор:

- **Отрасль:**
  Онлайн торговля
- **Регион:**
  Москва, Россия
- **Задача:**
  Монетизация данных: разностороннее использование имеющейся информации для развития различных направлений бизнеса и получения дополнительной выгоды
- **Решение:**
  СУБД Vertica
- **Результаты:**
  + Создана единая корпоративная аналитическая система на основе BI-платформы.
  + Превращение данных в актив, позволяющий компании успешно зарабатывать деньги.
  + Получение максимальной выгоды от использования имеющихся данных.
  + Переход к активному применению данных для принятия различных решений и снижения бизнес-рисков.

Решение

Расширяемая BI-платформа

Чтобы обеспечить монетизацию данных и превратить их в актив, позволяющий успешно зарабатывать деньги, подразделение BI решило для начала сконцентрировать усилия на оптимизации моделей ценообразования для платных услуг, баннерной рекламе, A/B-тестировании пользовательских интерфейсов сайта, автоматической модерации объявлений и подготовке финансовой отчетности. «Для выполнения намеченных планов требовался мощный инструмент, позволявший не только собирать, хранить и обрабатывать большие объемы сложных данных, но и наращивать возможности, которые могут потребоваться для решения будущих задач», — рассказывает Иван Гуз, BI-директор компании Avito.

Ключевым требованием к аналитической платформе стала практически неограниченная масштабируемость — как по объемам и скорости сбора, хранения и обработки данных, так и по их сложности. «На момент выбора платформы невозможно было спрогнозировать ни количество систем, данные из которых придется собирать (поначалу использовались всего две), ни объемы этих данных, ни скорость их поступления, ни конкретные аналитические задачи, ни методы их решения», — поясняет Николай Голов, архитектор корпоративного хранилища данных компании Avito. Причем речь шла о полноценной масштабируемости BI-платформы не только по производительности и объемам данных, но и по аналитическому охвату стремительно усложняющихся данных. Еще один важный параметр выбора — поддержка SQL-запросов, а третье ключевое требование — возможность масштабирования на основе стандартных серверов архитектуры x86, приобретенных у разных вендоров.

Выбор платформы осуществлялся на основе тестирования, проведенного силами специалистов Avito. В качестве тестовых брались реальные данные компании —требовалось удостовериться, что платформа сможет их быстро загружать.

На финальной стадии остались три претендента: платформы Vertica, Greenplum, Oracle Exadata. В решении Oracle специалистов Avito не устроили возможности использования стандартных серверов и масштабирования на их основе; Greenplum показала себя слабее Vertica по скорости загрузки данных. «Мы знали, что наш проект находится на переднем крае технологических инноваций, поэтому одним из критериев выбора поставщика аналитической платформы была возможность обратиться к коллегам из других российских компаний в случае возникновения сложных проблем. К тому времени Vertica уже использовалась в компании Yota, и в конечном итоге мы отдали предпочтение именно этому продукту», — вспоминает Николай Голов, архитектор корпоративного хранилища данных компании Avito.

Внедрение аналитической платформы осуществлялось собственными силами. К моменту начала реализации проекта вся серверная инфраструктура компании размещалась в Швеции, серверы приобретались по лизинговой схеме у владельца центра обработки данных. Платформа Vertica была развернута на кластере из трех серверов, еще два использовались для извлечения, преобразования и загрузки данных (ETL). Интеграция с другими системами не составила труда. Для интеграции с кликстримом была использована схема с промежуточным кэшированием данных, описывающих историю событий за последние три-четыре дня в СУБД MongoDB.

Основные сложности возникли, когда потребовалось расширить аналитическую систему. После того как объем данных превысил определенный уровень, пришлось отказаться от прежних подходов к работе с ними и найти новые, более эффективные способы их обработки. Кроме того, пришлось столкнуться со сложностями при расширении кластера, на котором располагалась Vertica. В Avito считают, что все эти затруднения возникли из-за недостатка опыта на момент внедрения платформы.

Преимущества

Данные приносят деньги

Аналитическая система на базе платформы Vertica стала неотъемлемой частью бизнес-модели Avito, без которой достижение успеха уже невозможно. Глава подразделения бизнес-аналитики Иван Гуз с середины 2016 года входит в группу высших руководителей компании, которая принимает все ключевые решения, касающиеся развития бизнеса Avito.

В аналитическую систему стекается вся поступающая в компанию информация. Система Vertica располагается на кластере из 14 серверов в одном из московских ЦОД (это требуется по закону, поскольку Avito является оператором персональных данных); еще три сервера находятся в «холодном» резерве, один — для процедур ETL, также восемь серверов MongoDB работают в составе кластера, обеспечивающего кэширование событий кликстрима. В скором времени число серверов для Vertica планируется увеличить в два раза — до 28, также предполагается в кластер MongoDB добавить еще один дополнительный сервер и установить второй резервный сервер для ETL.

Работа по расширению возможностей системы постоянно продолжается: по мере появления в Avito функциональных блоков, поддерживающих новые направления бизнеса компании, они интегрируются с Vertica, а кроме того, добавляются инструменты для анализа новых данных. Имеющиеся инструменты позволяют использовать самые разные способы анализа данных, в том числе новейшие, такие как глубинное обучение и компьютерное зрение.

«На момент выбора платформы невозможно было спрогнозировать ни количество источников данных, ни их объемы, ни скорость поступления информации, ни задачи, которые придется решать, ни методы их решения»

НИКОЛАЙ ГОЛОВ
архитектор корпоративного хранилища данных
Avito

www.microfocus.com

Основными пользователями системы являются сотрудники Avito, работающие в подразделении BI: из трех десятков специалистов шестеро занимаются обслуживанием и развитием системы на базе Vertica, остальные — анализом данных для различных прикладных задач. В их числе оптимизация ценообразования платных услуг с учетом их категории, географии пользователей и прочих параметров, модерация (выявление дубликатов, нарушение категорий, борьба с мошенническими объявлениями), оптимизация баннерной рекламы (включая проведение CTR-аукционов), организация таргетированных рассылок электронной почты и SMS, аналитика в области CRM. Отчеты для инвесторов с итогами и прогнозами по финансовым показателям, трафику и другим величинам тоже готовятся при помощи Vertica.

Аналитическими возможностями Vertica пользуется и высшее руководство компании. «У нас приживается подход, при котором топ-менеджеры, прежде чем принимать решения, внимательно изучают данные, проверяя различные гипотезы», — поясняет Иван Гуз, BI-директор Avito. Исследованием данных занимаются аналитики, работающие в ключевых бизнес-подразделениях компании. В скором времени планируется предоставить доступ к Vertica любому сотруднику Avito, если в ходе выполнения служебных задач ему полезно анализировать те или иные данные.

Подразделение BI в своей работе придерживается концепции Data Lab, основа которой — поиск новых идей и подходов по применению данных и аналитики с ощутимой монетизируемой пользой для бизнеса. «Мы поняли, например, что алгоритмы поиска мошеннических объявлений нужно постоянно совершенствовать, чтобы добиваться большей эффективности», — отмечает Николай Голов, архитектор корпоративного хранилища данных компании Avito. Сегодня все новые бизнес-проекты Avito реализуются при участии подразделения BI, совместно с ним разрабатываются всевозможные способы работы с данными — от интеграции новых функциональных модулей с аналитической системой на базе Vertica до анализа и использования данных. Все новые данные, которые будут появляться в компании, также будут стекаться в аналитическую систему — это касается и новых функциональных модулей, и новых внешних источников данных, которые Avito решит задействовать.

Многие задачи из области BI решаются путем организации конкурсов по анализу открытых данных, в которых Avito предлагает принять участие внешним специалистам, выбирая затем наиболее перспективные идеи и подходы. Зачастую авторов наиболее интересных идей компания зачисляет в свой штат.

«Компания Avito проделала колоссальную работу не только по внедрению и освоению аналитической платформы Vertica, но и по формированию культуры использования данных и BI-инструментов для решения самых разных бизнес-задач, — комментирует достигнутые результаты Евгений Степанов, руководитель направления Big Data Platform компании Micro Focus в России. — В проектах этого заказчика проявились наиболее яркие преимущества, которыми обладает Vertica. Avito блестяще использует возможности этой платформы, демонстрируя мировой уровень работы с Большими Данными».

VERTICA | MICRO FOCUS