

Аналитическая платформа «Полимати́ка»

В начале июня 2016 г. компания «Полимати́ка Рус» выпустила новую версию своего продукта — отчетно-аналитическую платформу «Полимати́ка 5.0». В интервью Роман Раевский — генеральный директор «Полимати́ка Рус» и главный конструктор системы «Полимати́ка» рассказывает, как компании удалось достичь обработки миллиардов записей за секунды.



Роман Раевский — генеральный директор «Полимати́ка Рус».

SN. Как давно ваши решения присутствуют на рынке?

Р.Р. Мы уже более 12 лет занимаемся разработкой инструментов для бизнес-аналитики и интеллектуальной обработки информации. Непосредственно работа над аналитической платформой «Полимати́ка» ведется, начиная с 2010 года, при этом основные принципы были заложены еще в 2004 году.

Текущее поколение системы ориентировано, в первую очередь, на аналитическую обработку больших массивов исходных данных с использованием методов машинного обучения и алгоритмов интеллектуального анализа информации.

SN. В июне компания IBS представила новые АПК «СКАЛА», а также специализированное решение «СКАЛАСР/Аналитика», которое было разработано совместно с вашей компанией. Почему компания IBS выбрала вас для своего проекта?

Р.Р. Компанией IBS был протестирован ряд аналитических инструментов, включая решения западных производителей. По результатам тестов «Полимати́ка» показала лучшие результаты по скорости и качеству аналитической обработки информации. Если быть точнее, то «Полимати́ка» вообще оказалась единственной системой, которая способна обрабатывать постоянно нарастающие объемы данных за приемлемое время и выдавать корректный результат.

SN. В чем уникальность вашей разработки?

Р.Р. Сейчас на рынке очень большое количество продуктов, которые позиционируют себя как BI-системы. Несмотря на это, все эти продукты можно условно разделить на два класса решений — интерактивные визуализаторы и сложные BI-решения с аналитикой.

Первые позволяют легко визуализировать имеющиеся данные в определенном разрезе и выполнять над ними простые операции такие, как фильтрация или сортировка. Но они не поддерживают сложных

аналитических операций и глубинного исследования исходных данных.

Второй класс решений использует выборки или агрегацию исходных данных, сложные методы моделирования и последующее наложение рассчитанной модели на исходный массив данных. Такой подход приводит к потере исходной детализации и, как следствие, пропуску аномалий и не всегда точным результатам анализа. Помимо этого, такие решения требуют серьезных затрат для внедрения и имеют ограниченный круг пользователей.

Наше решение является универсальным — мы используем лучшее из каждого класса решений. Система не использует агрегацию данных, моделирование, выборки. В «Полимати́ке» пользователь всегда работает с исходными данными и может в любой момент «провалиться» до нужного уровня детализации и представить результат в виде интерактивной диаграммы или графика, что расширяет возможности анализа, повышает точность работы системы и, как результат, улучшает качество принятых решений.

SN. Вы позиционируете систему как полностью российскую разработку. Вы не используете модули западных компаний?

Р.Р. Мы не используем в системе сторонних модулей или open-source библиотек. Платформа «Полимати́ка» разрабатывалась с нуля и является на 100% российской разработкой.

Большинство продуктов, особенно сложных BI-решений, имеют довольно длинную историю и ориентированы на работу в семействе ОС Windows. Из-за этого исторически как десктопные, так и серверные компоненты этих решений, имеют ограничения на объем обрабатываемых данных. Мы изначально были ориентированы на использование Unix- и Linux-операционных систем, что позволило избежать ограничений на уровне операционной системы.

SN. Какие еще особенности системы вы можете выделить?

Р.Р. «Полимати́ка» использует собственный OLAP-сервер, поддерживающий технологию онлайн-обработки данных в оперативной памяти — OLAP-In-Memory. К слову, наш OLAP-сервер является десятым OLAP-сервером в мире. Аналитический сервер «Полимати́ки» сочетает в себе OLAP-обработку и алгоритмы DataMining, связанные друг с другом.

Кроме того, при вычислениях мы используем гибридную технологию CPU+GPU, то есть ускорения вычислений на графических процессорах GPU-acceleration. При этом система автоматически определяет, какие ресурсы необходимо задействовать для той или иной операции. При работе с небольшими объемами записей

вычисления выполняются на центральном процессоре, а при обработке уже больших объемов часть вычислений переносится на графический ускоритель. Это позволяет сохранять комфортное время отклика системы при возрастающих объемах данных.

SN. Есть какие-то ограничения по объемам данных? Каковы пределы масштабируемости системы?

Р.Р. Все данные мы измеряем в записях, а не в байтах. Ограничения по масштабируемости системы зависят от используемого аппаратного обеспечения.

Например, видеокарта NvidiaTesla K80 с 24 Гбайт памяти способна за один цикл обработать 4 млрд строк. Таким образом, пределы масштабирования определяются числом требующихся GPU для обработки требуемого объема строк за один цикл. Если нужно сортировать, скажем, 7 млрд строк, то для этого нужно использовать две NvidiaTesla K80 и т.д. Подчеркиваю, что это необходимо для обеспечения времени отклика системы на уровне пары секунд — без видеокарт система будет работать — только медленнее.

Если оперировать конкретными цифрами, то логическое ограничение по количеству записей в системе — 2^{64} , или более миллиарда строк. Для понимания: объем данных в 2 млрд строк генерирует сеть гипермаркетов за 2 года работы, а 6 млрд строк — это, примерно, объем данных всего федерального казначейства.

SN. Какие источники данных поддерживает система? Откуда вы можете брать данные?

Р.Р. Система поддерживает интеграцию со всеми популярными базами данных. Для получения данных используется собственный модуль импорта, что позволяет нам как разработчикам системы легко добавлять поддержку новых источников данных при необходимости.

Загрузить данные в систему можно с помощью простого SQL-запроса. Если говорить о базах данных класса NoSQL, то с ними система работает на более низком уровне, например, REST API или прямого обмена информацией между C++ библиотеками.

«Полимати́ка» также поддерживает импорт данных из текстовых файлов, CSV- и Excel-файлов, баз данных 1С и прочих популярных форматов.

SN. Каким образом происходит связь с источником? Где хранятся данные, с которыми работает система?

Р.Р. Мы используем собственное хранилище данных, собственный контейнер OLAP-кубов, собственное расчетное ядро и свой серверный слой. При первом подключении к источнику необходимые данные загружаются на сервер «Полимати́ки». Добавление новых данных в систему про-

исходит согласно установленному расписанию обновления, при этом инкрементная подгрузка идет со скоростью порядка сотни тысяч новых записей в секунду.

Дополнительно система обеспечивает сжатие данных в 12+ раз в сравнении с объемом данных в источнике. Например, если размер таблиц в БД источника составляет 12 Гбайт, то размер OLAP-куба, сформированный из этих таблиц без потерь детализации, не будет превышать 1 Гб.

SN. Один из ключевых трендов — это аналитика/анализ информации, относящейся к объекту/цели из разных источников (неструктурированных данных) и выдача результатов в реальном времени. Насколько «Полиматики» готова к этому?

Р.Р. Ни одна из существующих систем не даст вам возможность работы в реальном времени, всегда будут задержки, связанные с процессами передачи, загрузки и обработки данных. Если говорить об аналитике в «псевдореальном» времени, когда задержки не превышают нескольких секунд, то в случае «Полиматики» все зависит от расписания времени обновления данных.

На практике у большинства клиентов аналитика «в реальном времени» подразумевает задержки до минуты между событием и фактом его анализа, то есть сравнения со всей историей. Теоретически это время можно сократить до 10 секунд, что решит задачи подавляющего большинства аналитиков.

SN. Какие еще результаты тестирования платформы вы можете представить?

Р.Р. Мы проводили тестирование совместно с компанией IBS. На объеме данных

более 2 млрд строк время на выполнение OLAP-операций составляет не более 3 секунд — и это при обработке всего массива данных. Сложные операции Data Mining, например, кластеризация и прогнозирование — от 17 секунд при обработке всего массива данных.

SN. Учитывая названные возможности системы, насколько она сложна в использовании?

Р.Р. С самого начала мы ориентировались на скорость и простоту работы в системе при использовании сложных аналитических инструментов. Вся работа пользователя в «Полиматики» происходит в окне браузера с помощью мыши.

Если взять, например, операцию кластеризации, то для этого необходимо выделить факт, по которому будет происходить расчет, и нажать кнопку «кластеризация» — система сама предложит оптимальное число итоговых кластеров и покажет результат работы. При этом пользователю не нужно знать, как это работает. Он всего лишь должен представлять, какие объекты он хочет подвергнуть анализу, и какие их характеристики его интересуют. Остальное происходит автоматически.

SN. Раз зашла речь о пользователях, то хотелось бы знать: поддерживается ли в системе совместная работа? Поддерживается ли самостоятельное исследование? Как это происходит?

Р.Р. В системе реализована система профилей (мы их называем ситуационными панелями), которая позволяет, опять же, за пару кликов отправить ссылку на профиль коллегам и поделиться результатами аналитической работы. С помощью про-

филей пользователь может загрузить необходимый экран и провести дополнительное исследование интересующих показателей.

То есть ситуационная панель — это начальная точка работы пользователя. Кто-то ограничится ее просмотром, а кому-то потребуется просмотреть все детали, историю, тонкости связи различных информационных объектов.

SN. Проекты? Внедрения?

Р.Р. У нас есть завершённые проекты в розничной торговле, банковской сфере, нефтегазовой отрасли. Мы активно сотрудничаем с государственными заказчиками, в настоящий момент система проходит тестирование в нескольких крупных государственных организациях. В ближайших планах компании — выйти на такие сегменты рынка, как телекоммуникационная отрасль и сфера электронной коммерции.

SN. Каковы ваши планы по развитию?

Р.Р. В настоящий момент мы только завершили работу над пятой версией системы. И в ближайшее время мы будем проводить ее тестирование совместно с нашими партнерами. Планы по развитию системы разработаны на несколько лет вперед. Например, до конца года мы планируем добавить поддержку сценариев работы системы, когда пользователь может сохранить последовательность действий и автоматизировать последующие процессы обработки данных. В наших планах также — реализация сценариев загрузки и взаимодействия кубов.■

Check Point: платформа R80

Апрель 2016 г. — Компания Check Point® Software Technologies представила платформу управления безопасностью нового поколения — Check Point R80 — плод более чем трехлетних научных исследований и разработок. В основе платформы лежат технологии, дающие возможность консолидировать все составляющие их среды безопасности в одном месте и создавать условия для улучшенной автоматизации задач по защите.

Распространение персональных устройств и взаимозависимость цифровой мировой экономики — глобальные тенденции, которые позволяют улучшить выполнение наших рабочих задач. Вместе с тем они поставили перед бизнесом задачи по борьбе с постоянно эволюционирующими угрозами. Новая платформа R80 помогает удовлетворить эту потребность, предлагая унифицированный подход к безопасности, который позволит консолидировать все аспекты защиты, делая их более эффективными и адаптивными к внедрению мощных решений по безопасности. Это достигается тремя ключевыми способами:

— **внедрение принципа «одна консоль — одна политика».** Новая платформа R80 обеспечивает постоянную защиту всей инфраструктуры с единой консоли управления периметром, филиалами ЦОДа и

облачными загрузками. Это позволяет упорядочить процессы и получить улучшенную видимость администрирования политик, в том числе дать нескольким администраторам возможность работать с одной и той же политикой без конфликта системы управления;

— **повышение уровня безопасности в соответствии с бизнес-процессами и сетевыми архитектурами.** Технологии платформы R80 помогают командам по безопасности идти в ногу с изменяющимися политиками за счет сегментации на управляемые секции и приведения безопасности в соответствие с ИТ-процессами. Они также позволяют повысить уровень самообслуживания в области безопасности для владельцев бизнеса и освободить ключевых сотрудников для решения стратегических ИБ-задач;

— **интеграция управления угрозами для общего видения рисков во всей сети.** С платформой R80 у команд безопасности будет полная видимость состояния защиты, они смогут быстрее реагировать на новые атаки. Мониторинг, регистрация, отчеты и сравнительный анализ событий будут отображаться на единой, наглядной и настраиваемой панели управления. Это позволит командам по безопасности не только быстро найти и разрешить проблему безопасности, но также создавать специализированные отчеты для всех — от ключевых руководителей бизнеса до технологических партнеров.

Компания Check Point лидер “по полноте видения” последнего исследования

Gartner “Magic Quadrant for Enterprise Network Firewalls, May, 2016” (<http://www.checkpoint.com/resources/gartner-leader-enterprise-network-firewall/index.html>, прим. ред.).

Fujitsu: сервер PRIMERGY нового поколения

Июнь 2016 г. — Компания Fujitsu представила новое поколение флагманских стоечных серверов PRIMERGY. Четырехsocketные x86-серверы PRIMERGY RX4770 M3 оснащены ОП DDR 4 и процессорами Intel Xeon E 7-4800/8800 v4. Устанавливаемый в стойку сервер оптимизирован для поддержки критически важных для бизнеса приложений, требовательных к вычислительным ресурсам. Он ориентирован для задач, требующих обработки больших объемов данных, например, для БД SAP HANA, обрабатываемых в ОП, или бизнес-аналитики в режиме реального времени. Новое поколение серверов PRIMERGY обладает максимально высокой устойчивостью к сбоям, гарантированной расширенными критически важными функциями обеспечения надежности, готовности и безопасности, а также улучшенными показателями консолидации и управляемости.

Новые серверы PRIMERGY установили новый мировой рекорд по результатам эталонного теста SPECint_rate_base_2006 для вычислений общего назначения (<https://www.spec.org/cpu2006/results/>).