ДЕСЯТЬ ГЛАВНЫХ

ТЕНДЕНЦИЙ 2017 ГОДА

В ОБЛАСТИ БОЛЬШИХ ДАННЫХ

Каждый год в компании Tableau проводится обсуждение направлений развития отрасли. По его результатам составляется список главных тенденций следующего года в области больших данных. Вот наши прогнозы на 2017 год.

Десять главных тенденций 2017 года в области работы с большими данными

2016 год стал переломным для больших данных: еще больше организаций научились хранить и обрабатывать данные самого разного формата и объема и получать на их основе ценную информацию. В 2017 году продолжится распространение систем, поддерживающих большие объемы структурированных и неструктурированных данных. Рынку потребуются платформы, помогающие обеспечить безопасность и надежность хранения и администрирования больших данных и при этом позволяющие конечным пользователям их анализировать. Эти системы будут совершенствоваться для эксплуатации в ИТ-инфраструктурах компаний на основе корпоративных стандартов.

Большие данные становятся быстрыми и доступными благодаря новым возможностям ускорения Hadoop

Разумеется, можно выполнять машинное обучение или проводить анализ тональности текста с помощью Hadoop, но при этом сразу возникает вопрос: насколько быстро работает интерактивный SQL? Ведь SQL — это проводник для бизнес-пользователей, которые хотят применять данные Hadoop для быстрого отображения показателей эффективности на унифицированных информационных панелях и для разведочного анализа данных.

Потребность в скорости способствовала распространению более быстрых баз данных, таких как Exasol и MemSQL, систем хранения на базе Hadoop, например Kudu, и технологий ускоренного выполнения запросов. Использование специализированных SQL-систем для Hadoop (Apache Impala, Hive LLAP, Presto, Phoenix и Drill) и решений OLAP для Hadoop (AtScale, Jethro Data и Kyvos Insights) в сочетании с этими средствами ускорения запросов еще больше размывает границу между традиционными хранилищами и решениями для работы с большими данными.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ: AtScale BI on Hadoop benchmark Q4 2016 («Компания AtScale: сравнительное тестирование решений для бизнес-аналитики на основе Hadoop, 4-й квартал 2016 года»)

2

Теперь большие данные — это не только Hadoop: специализированные решения для Hadoop теряют актуальность

В последние годы на волне популярности больших данных появилось несколько технологий, призванных удовлетворить потребность в аналитике на основе Hadoop. Однако корпоративные компании со сложной разнородной ИТ-инфраструктурой больше не хотят развертывать отдельную систему бизнес-аналитики только для одного источника данных (Hadoop). Ведь ответы на их вопросы могут быть в любом из множества источников, включая системы учета, облачные хранилища, а также структурированные и неструктурированные данные из массивов Hadoop и других хранилищ. (К слову, даже реляционные базы данных оснащаются механизмами для работы с большими данными. Например, SQL Server 2016 с недавнего времени поддерживает формат JSON.)

В 2017 году потребители будут требовать аналитику по всем видам данных. Платформы, не зависящие от типа и источника данных, будут набирать популярность, а те, что были созданы специально для Hadoop и не допускают более широкого применения, окажутся не у дел. Уход с рынка Platfora лишь первая ласточка.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ: Uncommon sense: The big data warehouse («Неочевидное вероятное: хранилище больших данных»)

3

Организации стремятся к максимальной эффективности при создании озер данных

Озеро данных подобно искусственному водохранилищу. Сначала строится плотина (кластер), а затем хранилище наполняется водой (данными). Когда озеро наполнится, воду (данные) можно использовать в различных целях: для производства электроэнергии, для купания или как источник питьевой воды (для прогнозной аналитики, машинного обучения, обеспечения информационной безопасности и т. д.).

До сегодняшнего дня наполнение озера было самоцелью. Но в 2017 году эта ситуация изменится, поскольку бизнес стал более критично относиться к Hadoop. Организации будут требовать от озер данных воспроизводимости и гибкости для получения быстрых ответов на вопросы. Компании будут более тщательно просчитывать окупаемость инвестиций в персонал, данные и инфраструктуру. Эта тенденция усилит взаимодействие бизнеса и ИТ, а платформы самообслуживания получат более широкое признание как средство эффективного использования больших данных.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ: Maximizing data value with a data lake («Повышение эффективности использования данных с помощью озера данных»)

4

Более зрелые архитектурные решения отходят от универсальных платформ

Технология Hadoop — это уже не просто средство пакетной обработки, используемое для работы с данными лишь в научных целях. Она превратилась в многоцелевое решение для специализированного анализа. В некоторых компаниях она даже заменила хранилища данных при создании текущих отчетов о регулярных операциях.

В 2017 году организации будут стремиться использовать специализированные архитектурные решения для удовлетворения этих разнообразных потребностей. Прежде чем принимать решение о выборе стратегии работы с данными, они будут исследовать множество факторов: профили пользователей, вопросы, объем, частоту обращений, скорость передачи данных и уровень агрегации. Архитектуры, отвечающие современным требованиям, будут создаваться исходя из конкретных потребностей. Они будут сочетать в себе лучшие инструменты для самостоятельной подготовки данных, основные механизмы Hadoop и аналитические решения для конечных пользователей таким образом, чтобы архитектуру можно было адаптировать к изменениям потребностей бизнеса. Гибкость архитектуры в конечном счете станет решающим аргументом при выборе технологий.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ: The cold/warm/hot framework and how it applies to your Hadoop strategy («Платформы с разной оперативностью доступа к данным: влияние на стратегию использования Hadoop»)

5

Главный стимул для инвестиций в большие данные — многообразие, а не объем или скорость

Компания Gartner определяет большие данные с помощью трех категорий: большой объем, высокая скорость и многообразие. Все эти три показателя продолжают расти, но именно скорость становится решающим стимулом для инвестиций в большие данные, согласно результатам недавнего исследования, проведенного компанией New Vantage Partners. Эта тенденция сохранится, так как организации стремятся к интеграции большего количества источников и уделяют основное внимание разнообразию больших данных. Форматов данных становится все больше: неструктурированные данные в формате JSON, вложенные разновидности в других базах данных (реляционных и NoSQL), структурированные данные (Avro, Parquet, XML). В этих условиях решающее значение приобретают модули объединения (коннекторы). В 2017 году аналитические платформы будут оцениваться исходя из возможности прямого подключения к принципиально различным источникам в режиме реального времени.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ: Variety, not volume, is driving big data initiatives («Применение больших данных определяется их разнообразием, а не объемом»)

6

Spark и машинное обучение повышают доступность больших данных

Система Apache Spark, изначально входившая в комплекс решений Hadoop, становится наиболее привлекательной платформой для работы с большими данными в крупных компаниях. В рамках опроса, проведенного среди архитекторов данных, ИТ-руководителей и бизнес-аналитиков, около 70 % респондентов предпочли Spark стандартному решению MapReduce, ориентированному на пакетную обработку и неспособному взаимодействовать с интерактивными приложениями и производить обработку потоковых данных в режиме реального времени.

Эти возможности сложных вычислений на основе больших данных привели к росту популярности платформ, включающих в себя машинное обучение на базе большого объема вычислений, искусственный интеллект и алгоритмы на графах. В частности, Microsoft Azure ML получила распространение благодаря доступности даже для начинающих пользователей и простоте интеграции с существующими платформами Microsoft. Доступность машинного обучения для широкого круга пользователей приведет к появлению новых моделей и приложений, которые будут создавать петабайты новых данных. По мере развития машинного обучения и усложнения систем в центре внимания окажутся поставщики средств самообслуживания, которые должны будут обеспечить доступность этих данных для конечного пользователя.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ: Why you should use Spark for machine learning («Почему стоит использовать Spark для машинного обучения»)

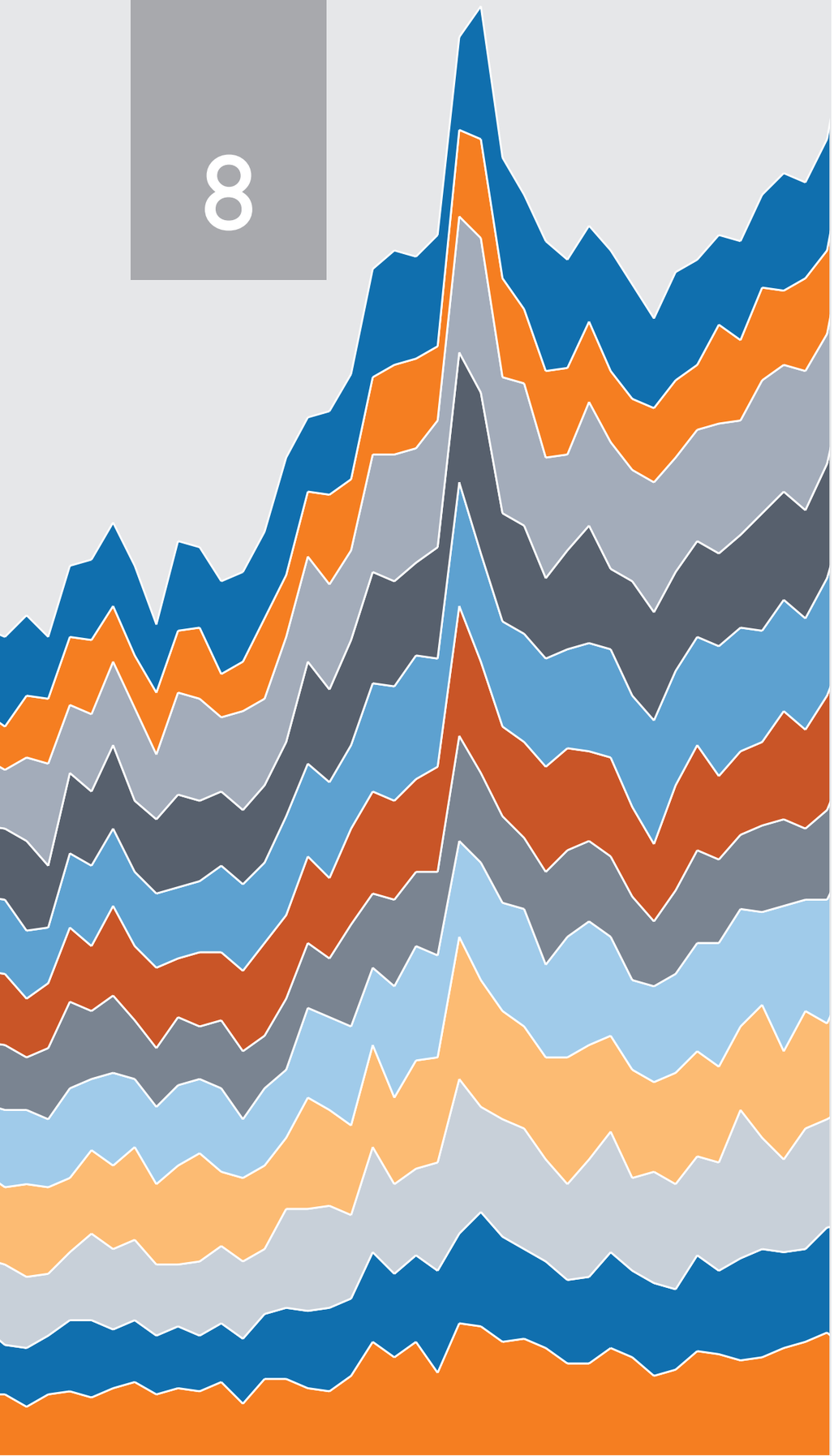
7

Объединение Интернета вещей, облачных технологий и больших данных создает новые возможности для самостоятельной аналитики

По всей видимости, в 2017 году у каждого объекта будет датчик, отправляющий информацию в центр обработки данных. Интернет вещей создает огромные массивы структурированных и неструктурированных данных, все большая часть которых развертывается в облачных службах. Эти данные часто неоднородны и располагаются во множестве реляционных и нереляционных систем: от кластеров Hadoop до баз данных NoSQL. В то время как инновации в области хранения и управляемых услуг ускорили сбор данных, доступ к ним и получение полезной информации все еще представляют собой сложную задачу для конечного пользователя. В связи с этим растет спрос на аналитические инструменты, которые позволяют свободно объединять множество разнообразных источников облачных данных. Эти инструменты дают организациям возможность изучать и визуализировать любые виды данных, где бы они ни хранились, и помогают определить скрытые преимущества инвестиций в Интернет вещей.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ: Tableau on solving IoT's last-mile challenge («Завершающая стадия реализации Интернета вещей: концепция Tableau»)

8



Самостоятельная подготовка данных становится обычным делом — пользователи берут большие данные в свои руки

Одна из самых актуальных задач — это повышение доступности данных Hadoop для бизнес-пользователей. Развитие платформ самостоятельной аналитики помогло продвинуться в ее решении. Но корпоративные пользователи требуют ускорения и упрощения подготовки данных к анализу, особенно при работе с данными различного типа и формата.

Гибкие средства самостоятельной подготовки обеспечивают не только предварительную обработку данных Hadoop в исходном хранилище, но и создание моментальных снимков баз данных для повышения скорости и удобства анализа. Множество инновационных решений в этой области было предложено компаниями, специализирующимися на подготовке больших данных для обработки конечными пользователями, например Alteryx, Trifacta и Paxata. Эти инструменты упрощают задачу для компаний, отстающих или серьезно запаздывающих на пути внедрения Hadoop, и в 2017 году их популярность будет расти.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ: Why self-service prep is a killer app for big data («Определяющая роль средств самостоятельной подготовки в системах больших данных»)

9

Большие данные выходят на передний план: Hadoop становится корпоративным стандартом

Системы Hadoop продолжают набирать популярность в качестве основной составляющей корпоративной ИТ-инфраструктуры. В 2017 году повысится объем инвестиций в средства обеспечения безопасности и управления для корпоративных информационных систем. Решение Apache Sentry обеспечивает детализированную проверку подлинности на основе ролей для доступа к данным и метаданным, хранящимся в кластере Hadoop. Решение Apache Atlas, разработанное для управления данными, позволяет компаниям внедрять единую классификацию данных в пределах инфраструктуры. Решение Apache Ranger предоставляет возможность централизованного управления безопасностью систем Hadoop.

В этих условиях корпоративные заказчики начинают требовать аналогичных возможностей от имеющихся у них платформ РСУБД. Эти возможности становятся основными для новейших технологий в области больших данных, разрушая еще одно препятствие на пути к распространению корпоративных систем больших данных.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ: The phases of Hadoop maturity: Where exactly is it going? («Стадии зрелости: в каком направлении развивается Hadoop»)

10

Распространение каталогов метаданных помогает пользователям находить большие данные, представляющие интерес для анализа

Долгое время компании отбрасывали значительное количество данных, поскольку не успевали их обрабатывать. Hadoop позволяет им обрабатывать большие объемы данных, однако способ их организации, как правило, представляет сложности для поиска.

Каталоги метаданных позволяют обнаружить подходящие данные и извлечь из них ценную информацию с помощью средств самообслуживания. На данный момент в этой нише представлены такие компании, как Alation и Waterline, которые используют технологии машинного обучения для автоматизации поиска в массивах Hadoop. Они создают каталоги файлов с использованием тегов, выявляют связи между ресурсами данных и даже предлагают варианты поисковых запросов в пользовательском интерфейсе. Это помогает потребителям и хранителям данных сократить время, необходимое на проверку, поиск и корректный запрос данных. В 2017 году вслед за ростом популярности самостоятельной аналитики возрастет интерес к возможностям самостоятельного обнаружения данных и спрос на соответствующие решения.

ДОПОЛНИТЕЛЬНАЯ ИНФОРМАЦИЯ: Data catalogs as a strategic requirement for data lakes («Каталоги данных как обязательный элемент озер данных»)

О компании Tableau

Внедрение визуализации данных в розничные системы и процессы гораздо проще, чем вы думаете.

Компания Tableau Software помогает людям найти и проанализировать данные, независимо от их объема и количества источников. Решения Tableau позволяют быстро и удобно подключать, объединять, визуализировать данные и делиться информационными панелями на любом устройстве, от ПК до iPad. Создавайте и публикуйте информационные панели с возможностью автоматического обновления и делитесь ими с коллегами, партнерами и клиентами — без специальных навыков программирования. Установите бесплатную пробную версию.

TABLEAU.COM/TRIAL