Аналитика в реальном времени – смена парадигмы бизнеса

Направления бизнес-анализа и BI[1] (business intelligence) в последние годы показывают самые стабильные темпы роста. С появлением технологий in-memory акцент во многих BI-проектах сместился к построению аналитических систем, способных функционировать практически в реальном времени. О технологиях и решениях на базе разработок компании SAP, позволяющих реализовывать такие системы, рассказывает директор компании "Терн" Екатерина Лозовая.

Лозовая Екатерина Александровна – директор ЗАО "Терн".

Введение

Направление решений "вычисления in-memory", позволяющих переносить в оперативную память БД (частично или полностью) и одновременно выполнять в ОП OLTP- (online transaction processing) и OLAP-приложения (oline application processing), стало развиваться с 2010 г. и сегодня число подобных предложений в различной архитектурной реализации на рынке уже около десяти.

Пионером и одним из лидеров этого направления является компания SAP с решением HANA. Отличительной особенностью решений па базе платформы SAP HANA является высокая интегрированность БД для OLAP- и OLTP-приложений, что в значительной степени упрощает администрирование и управление подобными системами. После ряда анонсов, сделанных SAP за последние месяцы, решения этого класса получили новый импульс развития, благодаря появлению нового функционала и упрощению процедуры миграции на новую архитектуру уже развернутых приложений.

Компания "Терн" считает платформу SAP HANA одной из приоритетных для продвижения в своих проектах, связанных с аналитикой в реальном времени.

1) Приложения BI, в основном, связаны с репортингом. Приложения бизнес-анализа позволяют исследовать бизнес-процессы и отвечать на вопросы "что будет происходить", "что будет происходить при изменении каких-либо условий" и т.п., *прим. ред.*

Требования бизнеса

Объем данных, с которыми работают компании, удваивается каждый год. Помимо этого, бизнес становится гораздо более динамичным, требуя управления бизнес-процессами, максимально приближенного к реальному времени с мгновенным предоставлением заинтересованным лицам всей необходимой информации.

Вычисления в оперативной памяти и мобильные технологии открывают новые пути для трансформации бизнеса и получения конкурентных преимуществ. За счет их использования у организаций появляются новые возможности, в частности, более быстрое реагирование на изменение спроса и предоставление персональных предложений для удовлетворения потребностей клиентов. Сотрудники смогут работать эффективнее, поскольку будут получать представление о состоянии дел в реальном времени. Раскрыв весь потенциал имеющихся данных, руководители смогут принимать более взвешенные решения, оперативно реагировать на изменение рыночной ситуации и ускорять важные процессы, например планирование.

Компания получает полную свободу в поиске новых возможностей и переходу к таким моделям ведения бизнеса, которые невозможно было использовать ранее.

Проблемы перехода к ведению бизнеса в реальном времени

Чтобы перейти к ведению бизнеса в реальном времени, требуется непрерывно контролировать основные бизнес-процессы (финансовый учет, продажи и производство), получать данные из новых источников, например из социальных сетей, от мобильных приложений или датчиков оборудования. Кроме того, для принятия взвешенных бизнес-решений необходимо сразу анализировать эти данные и использовать передовые методы моделирования, в частности прогнозное моделирование. И, наконец, требуется обеспечить сотрудникам доступ в реальном времени к информации с помощью любого устройства, чтобы в критической ситуации они могли бы действовать без промедления.

Однако, сменить парадигму ведения бизнеса не так просто, и здесь можно выделить несколько проблем. *Во-первых*, это многочисленность БД, с которыми приходится иметь дело при анализе бизнес-информации, вследствие чего возникают многоэтапность и длительность обработки данных при принятии решений.

Для подготовки стандартных отчетов необходимо использовать процедуры ETL, обеспечивающие передачу данных из транзакционной системы в аналитическую и занимающие часто непозволи-

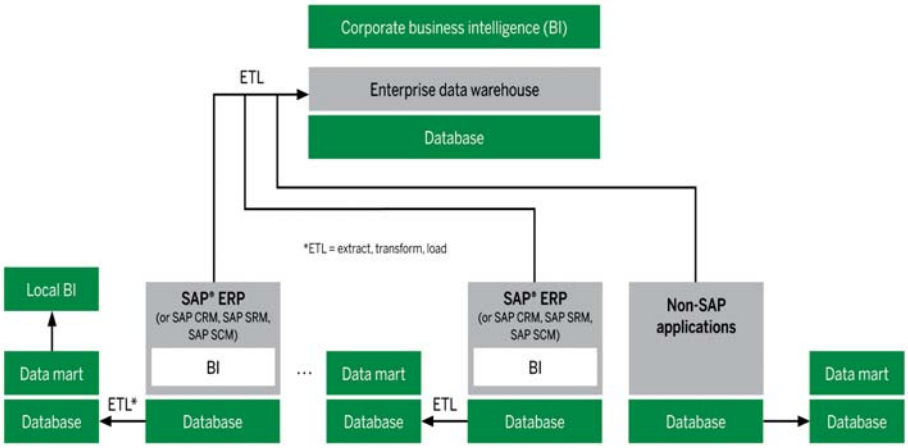

Рис. 1. В современных корпоративных системах данные для OLTP- и OLAP-приложений обычно хранятся в разных БД.

Табл. 1. Сравнение различных уровней памяти для процессора Intel Nehalem.

Тип памяти	Размер	Задержка
L1 cache	64 KB	~4 cycles [2 ns]
L2 cache	256 KB	~10 cycles [5 ns]
L3 cache (shared)	8 MB	35–40+ cycles [20 ns]
Main memory	GBs up to terabytes	100–400 cycles
Solid state memory	GBs up to terabytes	5,000 cycles
Disk	Up to petabytes	1,000,000 cycles

тельно много времени. И хотя в последнее время ETL-процедуры были в значительной степени оптимизированы, их продолжительность остается за пределами реального времени.

При использовании существующей технологии управления базами данных компании вынуждены постоянно идти на компромисс. Так, оптимизация корпоративной системы обработки данных для формирования отчетности, обеспечивающая, с одной стороны, охват и глубину данных, а с другой – скорость и простоту их обработки, попросту невозможна. Для создания отчета, позволяющего провести комплексный и глубокий анализ, необходимо заложить время на обработку огромных объемов данных – оно измеряется часами и даже днями. Если же компания стремится ускорить и упростить формирование отчетности, то ей придется сократить объем информации и уменьшить количество параметров, используемых для формирования отчета. Ни в том, ни в другом случаях получить обновление в реальном времени невозможно: в обычной ИТ-среде придется дожидаться окончания пакетной обработки заданий, выполняемой ночью. Все это усложняет ИТ-ландшафт, замедляет выполнение процессов и порой ограничивает доступ пользователей к нужной информации в нужное время.

Сложившаяся ситуация отражена на рис. 1. Крупные компании, как правило, имеют множество ERP-систем (enterprise resource planning), каждая из которых имеет собственную БД для операционных данных. Данные для аналитики через ETL-процедуры консолидируются в отдельном корпоративном хранилище данных (data warehouse) и доступны для бизнес-пользователей через BI-решения. Для анализа самых последних данных (которые обычно не синхронизованы с транзакционными/операционными данными) используются дополнительные витрины данных и локальные BI-клиенты.

Во-вторых, для традиционных ИТ-архитектур характерна необходимость переноса данных с уровня хранения (как правило это внешние СХД, в основном, построенные на жестких дисках) на уровень сервера СУБД. Вследствие существенной разницы в производительности на этих уровнях, это стало одним из самых узких мест, затрудняющих переход к ведению бизнеса в реальном времени (табл. 1).

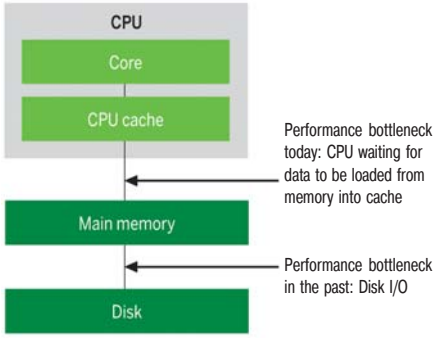

Рис. 2. Бурное развитие технологий привело к тому, что сдерживающим фактором производительности в современных OLAP-системах стала пропускная способность между основной памятью и кэшем ЦПУ.

Решению данной проблемы во многом мешала дороговизна чипов памяти DIMM-уровня. Однако в последние годы ситуация изменилась. Основная память перестала быть ограничивающим ресурсом. В 2012 г. стали доступны серверы с более чем 2 Тбайт RAM. Одновременно существенно возросло число ядер на сервер. В 2012 г. их число составляло 80, а в ближайшем будущем прогнозируется рост до 128. Это привело к тому, что сдерживающим фактором в производительности уже стало место между оперативной памятью и кэшем ЦПУ (рис. 2).

Платформа SAP HANA

Такое бурное развитие технологий явилось предпосылкой создания платформы SAP HANA – программно-аппаратного комплекса, позволившего перенести хранение (части или полностью) БД для OLTP- и OLAP-приложений непосредственно в оперативную память. В настоящее время со стороны аппаратной составляющей она поддержана сертифицированными серверными решениями от HP, IBM, Fujitsu, Hitachi, Cisco, Dell, NEC, Huawei. По заявлениям представителей SAP, все сертифицированные решения имеют примерно одинаковую производительность при развертывании на них SAP HANA.

Одновременно SAP HANA является законченной и полной системой управления базой данных (DBMS) со стандартным интерфейсом SQL, транзакционной изоляцией и восстановлением (ACID – atomicity, consistency, isolation, durability) и с высокой доступностью.

SAP HANA поддерживает SQL92, а приложения SAP, использующие Open SQL, могут выполняться на SAP HANA платформе без каких-либо изменений. Дополнительные функциональные возможности, типа freestyle-поиск, реализованы в SAP HANA как расширения SQL.

Аналитические и специальные интерфейсы

В дополнение к SQL, SAP HANA непосредственно поддерживает BI-клиентов, используя многомерные выражения (MDX – multidimensional expressions) для таких продуктов как Microsoft Excel и BI-сервисы (BICS – Business Intelligence Consumer Services) через внутренний интерфейс для решений SAP BusinessObjects™.

Построчное и поколоночное хранение БД

Одна из главных оптимизационных задач, решаемых SAP HANA, – достижение высоких коэффициентов попадания данных на уровни кэширования ЦПУ. Это достигается сжатием данных и адаптацией способа хранения данных в БД для каждой задачи. Например, когда обработка выполняется построчно и большинство обрабатываемых полей находятся внутри строки, используется построчная запись (как правило, это OLTP-обработка, *прим. ред.*). Если вычисления выполняются на одном

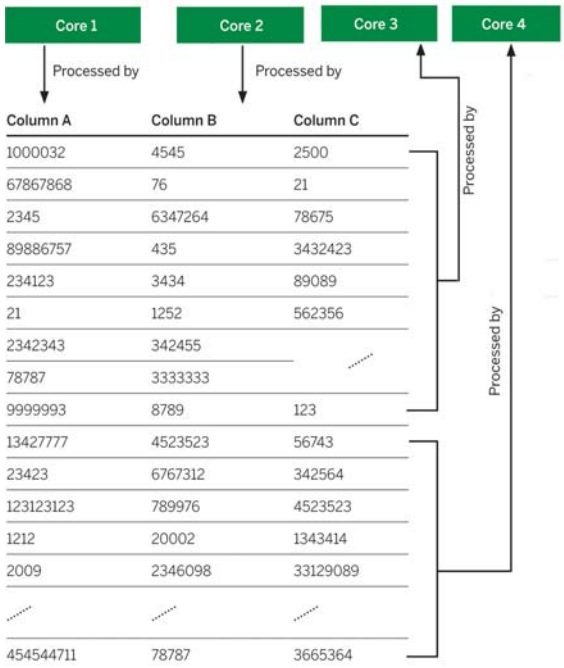

Рис. 3. При поколоночном хранении операции по отдельным столбцам, а также части данных самих столбцов могут быть переданы на обработку отдельным ядрам.

или нескольких столбцах, используется поколоночное хранение, при котором каждый столбец представляется отдельным сжатым блоком данных (в основном, для OLAP-приложений, *прим. ред.*). Как показывают оценки, проведенные SAP, при операциях записи (OLTP-запросы) только около 30% таблиц требуют построчной записи, оставшиеся 70% хорошо выполняют запись и при поколоночном хранении, тем самым, обеспечивая высокие показатели компрессии данных – порядка 20% и, соответственно, существенно упрощая миграцию БД в ОП.

При поколоночной записи существенно упрощается процесс распараллеливания вычислений по различным ядрам. Если необходимо выполнить поиск по множеству столбцов или агрегацию нескольких столбцов, то такие операции могут быть назначены различным ядрам процессора. Более того, при выполнении операции в рамках одной колонки, колонка может быть разбита на несколько партиций, каждая из которых передается на выполнение отдельному ядру (рис. 3).

SAP Business Suite на платформе SAP HANA – единая платформа для аналитики и транзакционных приложений

SAP Business Suite – комплекс решений, разработанных для совместной работы и предназначенных для повышения эффективности бизнеса. Это полностью интегрированный программный пакет решений, поддерживающий выполнение основных бизнес-операций с помощью лидирующих на рынке приложений, процессов и технологий.

Теперь базовые решения SAP – "Управление взаимоотношениями с клиентами" (SAP CRM), "Управление ресурсами предприятия" (SAP ERP), "Управление логистической цепочкой" (SAP SCM) и "Управление взаимоотношениями с поставщиками" (SAP SRM)[2] – могут исполь-

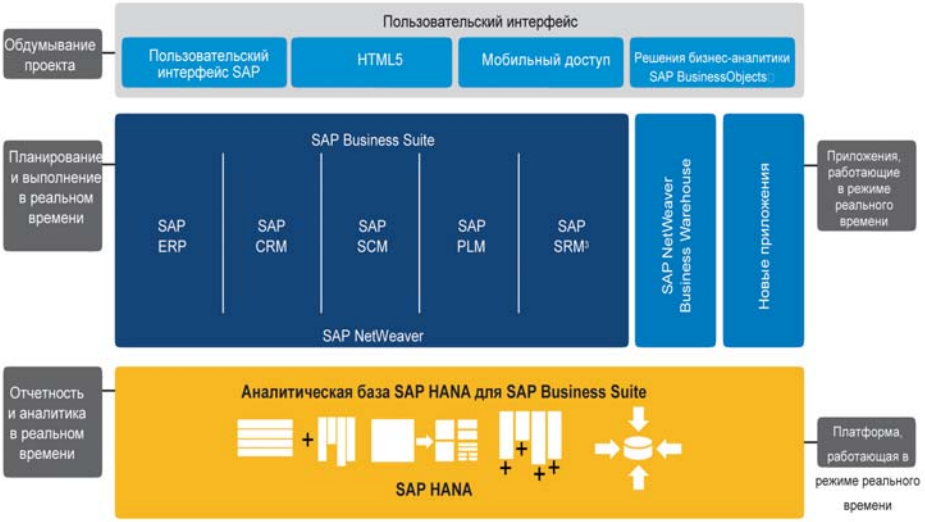

Рис. 4. Открытая платформа для внедрения инноваций без нарушения текущей работы компании.

зовать SAP HANA в качестве главной базы данных. Это позволяет обрабатывать в памяти огромные объемы информации в реальном времени вне зависимости от того, генерируется ли эта информация внутренними приложениями или импортируется из внешних источников, в частности, из социальных сетей. Кроме того, можно применять расширенные средства аналитики, например прогнозную аналитику, ко всем транзакционным данным, чтобы принимать верные решения на основе фактов. При этом снижается сложность и исключается дублирование данных и систем.

Большое количество бизнес-сценариев (для маркетингового анализа, финансового закрытия, управления дебиторской задолженностью, планирования материальных ресурсов, анализа покупательского поведения и потребительских предпочтений и т. д.) и наиболее часто используемых форм для операционной отчетности и аналитики были оптимизированы с целью создания максимальной ценности для клиентов.

Впервые компании получают возможность обрабатывать транзакции и выполнять аналитику в реальном времени с использованием единого источника достоверных данных. Это открывает пути для внедрения интеллектуальных инноваций, ускорения процессов и упрощения взаимосвязей в бизнесе.

Компания "Терн" приняла деятельное участие в реализации крупного проекта в ОАО "Сургутнефтегаз", и хотелось бы привести слова начальника управления информационных технологий ОАО "Сургутнефтегаз" Рината Гимранова: "Бизнес сегодня требует от технологий возможности свести к минимуму временной разрыв между анализом информации, принятием решения и его выполнением. Перевод SAP Business Suite на платформу SAP HANA – это не только упрощение бизнес-коммуникаций и передачи информации в режиме реального времени, но и способ сделать стратегию компании максимально близкой и доступной всем сотрудникам".

Платформа SAP HANA обеспечивает открытость, необходимую для внедрения инноваций без нарушения текущей деятельности компании. Кроме того, она позволяет использовать решения нового поколения, работающие в реальном времени и изначально построенные на данной платформе (рис. 4).

Новые возможности для пользователей

Пользователи теперь могут работать в реальном времени с унифицированным представлением информации в рамках нужного контекста. Интерфейс, ориентированный на обычных пользователей, разработан на основе последних достижений инженерной мысли, обеспечивает простой и удобный способ взаимодействия с приложениями и не требует долгого освоения. Руководители и сотрудники получают возможность совместной работы в режиме реального времени и доступ к данным любого типа с помощью любого устройства.

Упрощение ИТ-инфраструктуры

SAP Business Suite на платформе SAP HANA позволяет упростить ИТ-инфраструктуру, а объединение процессов аналитики и транзакций снижает совокупную стоимость владения. SAP HANA предоставляет уникальную возможность эффективного управления транзакционными и аналитическими потоками, благодаря чему значительно упрощается ИТ-ландшафт. Если компания применяет интегрированный сценарий, то для решений SAP Business Suite используется один экземпляр платформы SAP HANA в качестве основной базы данных. Поскольку процессы обработки транзакций и аналитики обращаются к общей схеме базы данных, то необходимость в репликации данных отпадает.

SAP Business Suite работает на платформе SAP HANA, что позволяет получать данные с любым уровнем детализации. Возможно также проведение прогнозного анализа, обработка структурированных и неструктурированных данных и управление различными ресурсоемкими операциями в режиме реального времени. По мере развития моделей ведения бизнеса сложные логические операции можно будет вывести на уровень базы данных.

Интеллектуальные бизнес-инновации

Динамичные компании привлекают клиентов за счет применения интеллектуальных моделей ведения бизнеса и процессов. Комплекс решений SAP Business Suite на платформе SAP HANA поможет задействовать бизнес-инновации путем переосмысления бизнес-процессов и изобретения более интеллектуальных моделей, чем раньше. Например, можно получить более точное представление о потребительских сегментах и данные в новых форматах, включая данные от датчиков оборудования или из социальных сетей. Для этого можно воспользоваться инструментами прогнозной аналитики, которые помогут сменить модели монетизации, чтобы перестать зависеть от продукции и перейти на модели предоставления услуг, основанные на предпочтениях клиентов.

Среди возможностей, предоставляемых SAP Business Suite на платформе SAP HANA, можно отметить следующие:

- улучшение положения на рынке посредством создания новых бизнес-моделей на основе данных;
- принятие более обоснованных решений за счет использования инструментов прогнозного анализа и моделирования по "большим данным" для минимизации рисков;
- трансформирование бизнес-процессов с помощью встроенной в транзакции аналитики для увеличения валового дохода.

Интеграция транзакционных и аналитических приложений

В рамках платформы SAP HANA аналитические приложения могут непосредственно работать с данными операционных приложений, используя один источник, но разные модели данных. При этом аналитические приложения могут не терять своей эффективности в производительности, например за счет того, как уже отмечалось, что таблицы для записи данных в транзакционных приложениях могут храниться в поколоночном виде.

Там, где требуется выполнение процедур ETL, например, в целях обогащения данных дополнительной информацией или в целях устранения ненужной или избыточной информации, используются технологии ETL или Replication Server, работающие в реальном времени.

Необходимо заметить, что при миграции на платформу SAP HANA способ хранения таблиц задается автоматически, хотя можно использовать и ручное управление.

Производительность SAP HANA

Вопрос производительности при переходе к анализу данных в реальном времени – ключевой. Платформа SAP HANA в данном случае выполняет основную роль. Высокая производительность обработки запросов на ее базе достигается за счет трех факторов:

- перемещения хранения БД в RAM;
- обеспечения близкой к линейной масштабируемости по производительности при увеличении числа ядер;
- высокого уровня попадания данных в кэш ЦПУ за счет внутренней оптимизации алгоритмов обработки запросов.

2) По состоянию на январь 2013 г. решение SAP «Управление взаимоотношениями с поставщиками» не поставляется с SAP HANA. В будущем планируется интеграция.

3) SAP HANA™ for Next-Generation Business Applications and Real-Time Analytics. Explore and Analyze Vast Quantities of Data from Virtually Any Source at the Speed of Thought, SAP, ноябрь 2012.

4) "HANA Performance: Efficient Speed and Scale-out for Real-time BI", SAP, ноябрь 2012.

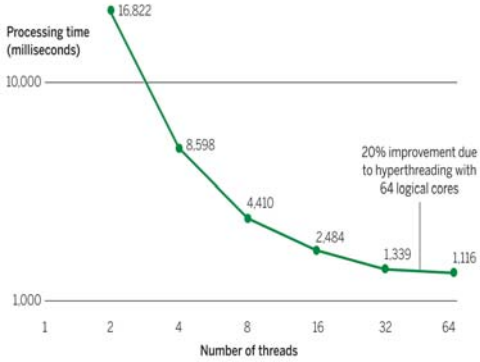

Рис. 5. Выполнение теста Joining TPC-H на наборе данных (120 млн записей) на платформе SAP HANA™ на базе серверов 4S Nehalem-EX (2.26 GHz) с 64 логическими ядрами.

Результаты измерения производительности SAP HANA, проведенные компанией SAP, проиллюстрируем на двух примерах[3,4].

Пример 1. Масштабирование производительности SAP HANA при добавлении ядер на операции Joining TPC-H[3].

Данный тест выполнен SAP совместно с Intel на базе серверов 4S Nehalem-EX (2.26 GHz) и на наборе данных, содержащем 120 млн записей.

На рис. 5 представлены тестирования производительности при масштабировании SAP HANA, выполненные SAP совместно с Intel. Время обработки с 16,8 сек на базе двух ядер улучшается до 1,4 сек с использованием 32 ядер.

Увеличение производительности достигается SAP HANA за счет внутренней оптимизации выполнения SQL-инструкций, позволяющей выполнять обработку параллельно на множестве ядер и узлов в составе SAP HANA.

Пример 2. Масштабирование производительности SAP HANA в зависимости от числа потоков на реальных запросах для SAP ERP системы[4].

Тестовые данные представляли собой одну большую таблицу, содержащую 61 колонку и 1200 млрд записей, с несколькими таблицами меньшего размера. Общий объем некомпрессионных данных составлял 1 Пбайт. При этом какие-либо настройки не выполнялись, а индексация таблиц не проводилась. Данные представляли собой информацию по продажам и были взяты из реальной SAP ERP системы.

Конфигурация тестовой системы представляла собой кластер из 100 IBM X5 серверов с общим объемом RAM – 100 Тбайт. 95 узлов были сконфигурированы как узлы SAP HANA и 5 узлов были выделены для поддержания отказоустойчивости. Каждый узел содержал: 4 процессора с 10 ядрами каждый и 2 гиперпотока (всего 40 ядер и 80 гипер-потоков на узел/сервер); 1 Тбайт RAM; 3,3 Тбайт дисковой памяти.

Данные были распределены между 95 узлами на основе Customer_ID. При компрессии 1 Пбайт исходных данных сжимался до 49,2 Тбайт (за счет перевода таблиц из построчного хранения в поколоночное, что делается автоматически при миграции данных из внешних источников для приложений аналитики), в результате, занимая только 517 Гбайт RAM (или половину возможного) на каждом узле.

Пакет запросов состоял 18 различных SQL, который включал 10 базовых запросов и их вариации по временным интервалам (месяц, квартал и т.д.) и охватывал диапазон BI-запросов от уровня департамента до корпоративного уровня: основные отчеты, итеративные запросы (drill downs), ранжирование, погодовой анализ.

При тестировании измерялось как время обработки запроса, так и число обрабатываемых запросов за один час. Сначала запросы были выполнены в одном потоке, затем в множественных параллельных потоках – 10, 20, 30, 40, 50 и 60. При этом при 50 потоках среднее время обработки запроса составило 1, 6 сек, что только в 3,4 раза больше, чем в однопоточном режиме.

Максимальное параллельное число обслуживаемых пользователей может быть получено делением максимальной производительности обрабатываемых запросов в час (112 602) на среднее время обдумывания запроса пользователем (это примерно 3 мин или 20 запросов в час). В результате получаем 5 630 пользователей, которые могут работать в реальном времени (максимальное время обработки наиболее сложных запросов не превышает 11,5 сек) на данной конфигурации системы с заданным типом BI-запросов.

Заключение

Рассмотренные выше возможности SAP HANA, безусловно, производят впечатление, однако, всегда надо помнить о том, что ни один даже самый мощный инструментарий не дает положительного эффекта без привлечения грамотной команды специалистов, способных профессионально разобраться в сути проблем и правильно применить инновационные разработки. Как правильно отмечается в ряде аналитических исследований, успех ИТ-проекта на 70% зависит от людей и лишь на 30% – от применяемых технологий. Хочу, в свою очередь, пожелать всем читателям Storage News исключительно успешных проектов и пригласить для профессионального общения на традиционную ежегодную BI-конференцию компании "Терн", которая пройдет 21 мая в Москве.