SAP HANA – платформа для бизнеса в реальном времени

Вводная лекция

1. Хранение и обработка данных в стиле SAP: BW или HANA

SAP SE — немецкая компания, производитель программного обеспечения для организаций.

Компания SAP была создана пятью бывшими сотрудниками IBM.

Наименование SAP составлено на основе первых букв полного названия: «Systeme, Anwendungen und Produkte in der Datenverarbeitung» / «Systems, Applications and Products in Data Processing».

Объемы данных растут, а сроки на принятие управленческого решения уменьшаются. Необходимо построить аналитическую систему, чтобы быстро получать агрегированные данные из различных ИТ-приложений.

Если за основу для создания такой аналитической системы взять технологии компании SAP, решения которой в корпоративном сегменте для многих отраслей стали стандартом де-факто, то с точки зрения скорости получения данных, соотношения цена/качество и общей стоимости владения можно назвать два варианта: SAP BW и HANA.

Один из них уже есть в практике достаточно многих российских предприятий. Он связан с построением хранилища данных на основе платформы SAP Business Warehouse (BW) и внедрением поверх него инструментов Business Intelligence. Опыт подобных проектов, в том числе реализованных компанией EPAM Systems, доказывает, что таким образом можно серьезно сократить сроки получения необходимой информации.

Закупки, продажи, производство, логистика, обслуживание клиентов – каждая из этих сфер деятельности современного предприятия ежеминутно генерирует поток данных. Они распределяются по широкому спектру различных информационных систем – это учетные приложения, промышленные ERP-системы, логистические и биллинговые решения и т.д.

С архитектурной точки зрения многие из них, – к примеру, ERP-системы или банковские автоматизированные системы, – работают на **реляционных базах** данных, в основе которых лежат принципы онлайновой транзакционной обработки (**OLTP**). Подобная структура позволяет обеспечить высокую производительность при регистрации данных. Каждый автоматизированный бизнес-процесс оставляет свой «след» в одном или нескольких приложениях компании, и объемы накопленной информации постоянно растут.

Для разгрузки учетных систем и высокой скорости получения аналитических данных часть информации можно переместить в другую структуру – **хранилище данных**. Оно спроектировано уже по совершенно иным архитектурным принципам – для онлайновой аналитической обработки (**OLAP**). Формирование аналитических данных в разных разрезах, drill-down и slice and dice обеспечиваются надстройкой над хранилищем данных в виде BI-инструментов (рис.1). Помимо ERP-системы, к хранилищу можно подключить все необходимые информационные активы организации, превратив его в полноценное корпоративное хранилище

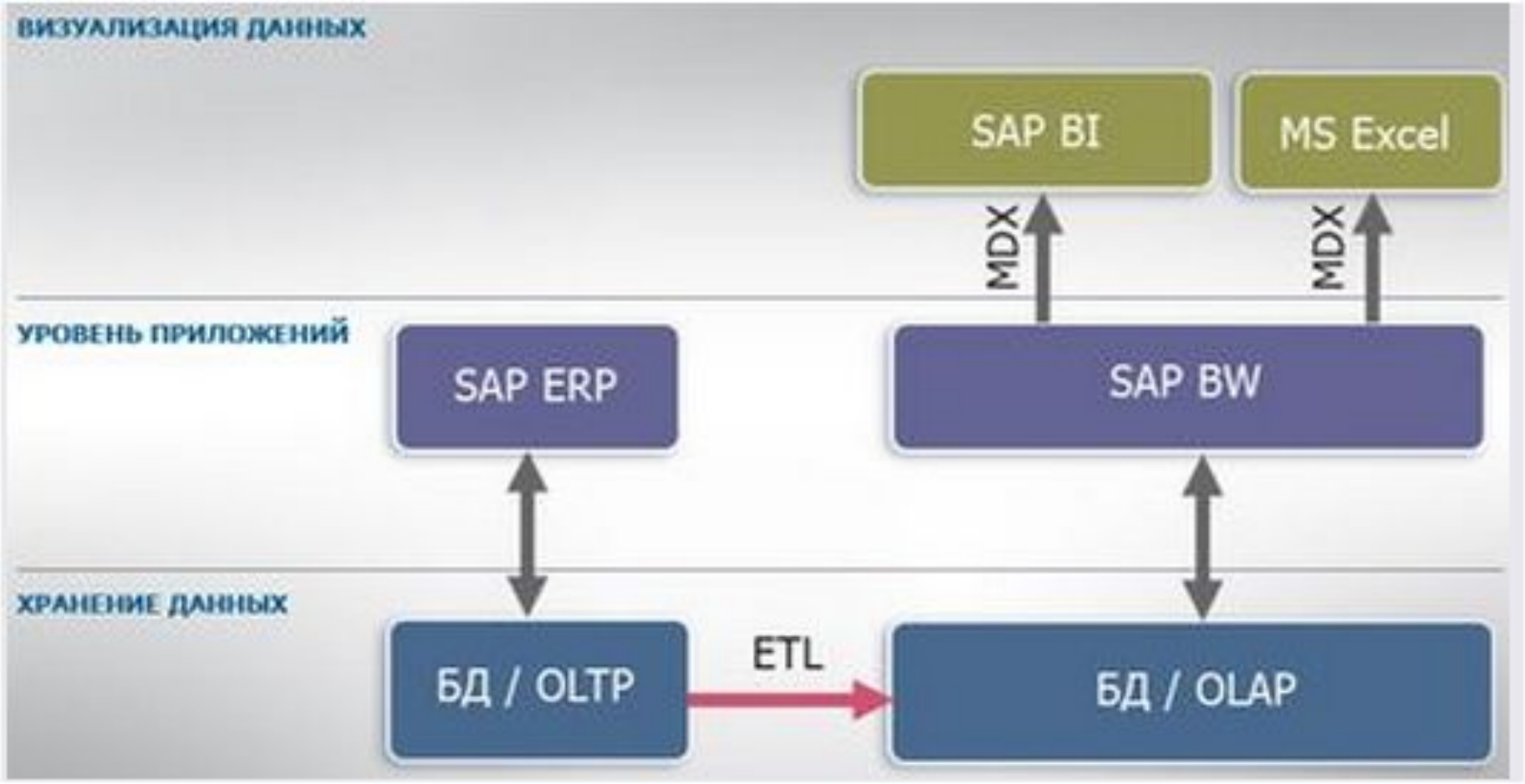


Рис.1 Взаимодействие систем при использовании хранилища данных

Вариант на основе SAP BW

Хранилища данных сейчас используются достаточно широко, но на определенном этапе некоторые компании попадают в ситуацию, когда возможности имеющихся систем уже не позволяют получать качественную аналитическую информацию и необходимую скорость для принятия управленческих решений.

Проблема заключается в том, что для переноса данных в классическое хранилище используется процесс **экстракции, трансформации и загрузки** (**ETL**). При очень больших массивах данных его выполнение потребует достаточно длительного времени, так что начать работать с актуальной аналитикой через считанные минуты после внесения изменений в учетную систему бизнес-пользователи не смогут. Кроме того, возникают вопросы и относительно полноты аналитических данных.

Нередко в хранилище загружаются только агрегированные данные, без возможности их рассмотрения на более детальном уровне. Другой вид - информация ограничена строго определенным временным горизонтом, без возможности увидеть всю историю конкретного показателя. В большинстве случаев причины таких ограничений чисто технические и связаны с тем, что традиционные базы данных уже не могут с приемлемой скоростью обработать накопленные объемы информации. Как следствие, пользователи при выполнении своих BI-запросов

Для многих бизнес-задач тех скорости и качества, которые может обеспечить связка «**хранилище данных – BI-инструменты**», бывает достаточно. В качестве примера можно привести процессы формирования обязательной отчетности, в некоторых случаях - бюджетирования и т. д.

Однако есть задачи, для которых требуется получать информацию практически в режиме реального времени.

В случае с производственными предприятиями примерами могут служить анализ загрузки производственных мощностей, анализ эксплуатационных характеристик оборудования предприятия и простоев при внеплановых ремонтах, анализ движения технико-материальных ценностей, состояние склада.

Для компаний ритейл-сектора и сферы услуг это, к примеру, моментальный анализ рентабельности различных сегментов бизнеса, перерасчет цен в условиях быстро меняющейся ситуации на рынке, оперативное планирование загрузки персонала в офисах и торговых отделениях и т. д.

Для финансовых организаций – анализ и управление в реальном времени потоком денежных средств и управление ликвидностью, анализ открытых валютных позиций в банках и др. В этом случае скорость, которую обеспечивает применение хранилища данных, может оказаться слишком низкой, что приведет к росту затрат или потерям компании из-

Технологичный авангард: вариант на основе SAP HANA

Для выполнения задач, где от руководителей требуется более быстрая реакция на ситуацию на предприятии или в бизнесе в целом, SAP предлагает использовать платформу SAP HANA (также в сочетании с BI-средствами). В ее основе лежит использование построенной на принципах in-memory гибридной базы данных. Это дает возможность сохранять информацию в базе данных как в традиционной построчной модели, так и в поколоночной. Поколоночное хранение обеспечивает высокую скорость агрегирования показателей и использование внутренней компрессии данных, что также положительно влияет на потребление доступной памяти. Встроенный OLAP-процессор агрегирует большие объемы данных на лету, без необходимости построения, заполнения, хранения и использования промежуточных агрегатов. При этом важно отметить, что есть возможность детализировать полученную аналитическую информацию до уровня исходных данных. Кроме того, при работе платформы максимально используются возможности современных процессоров для распараллеливания операций по обработке данных. В результате удается быстро получать нужную аналитическую информацию.

Данные для анализа в SAP HANA также могут поступать из базы данных, находящейся под управлением ERP-системы. Отличие от использования хранилища на основе SAP BW – отсутствие процессов ETL. Загрузка информации осуществляется с помощью технологии репликации данных, которая позволяет переносить все изменения, которые произошли внутри ERP-системы, в базу данных внутри SAP HANA в режиме, максимально приближенном к real-time. В результате пользователям не нужно ожидать очередного (как правило, ежесуточного) срабатывания ETL-процесса или работать с устаревшими данными. Результаты всех изменений данных, происходящих в ERP, оперативно доступны через привычные интерфейсы бизнес-аналитики SAP BI (рис. 2).

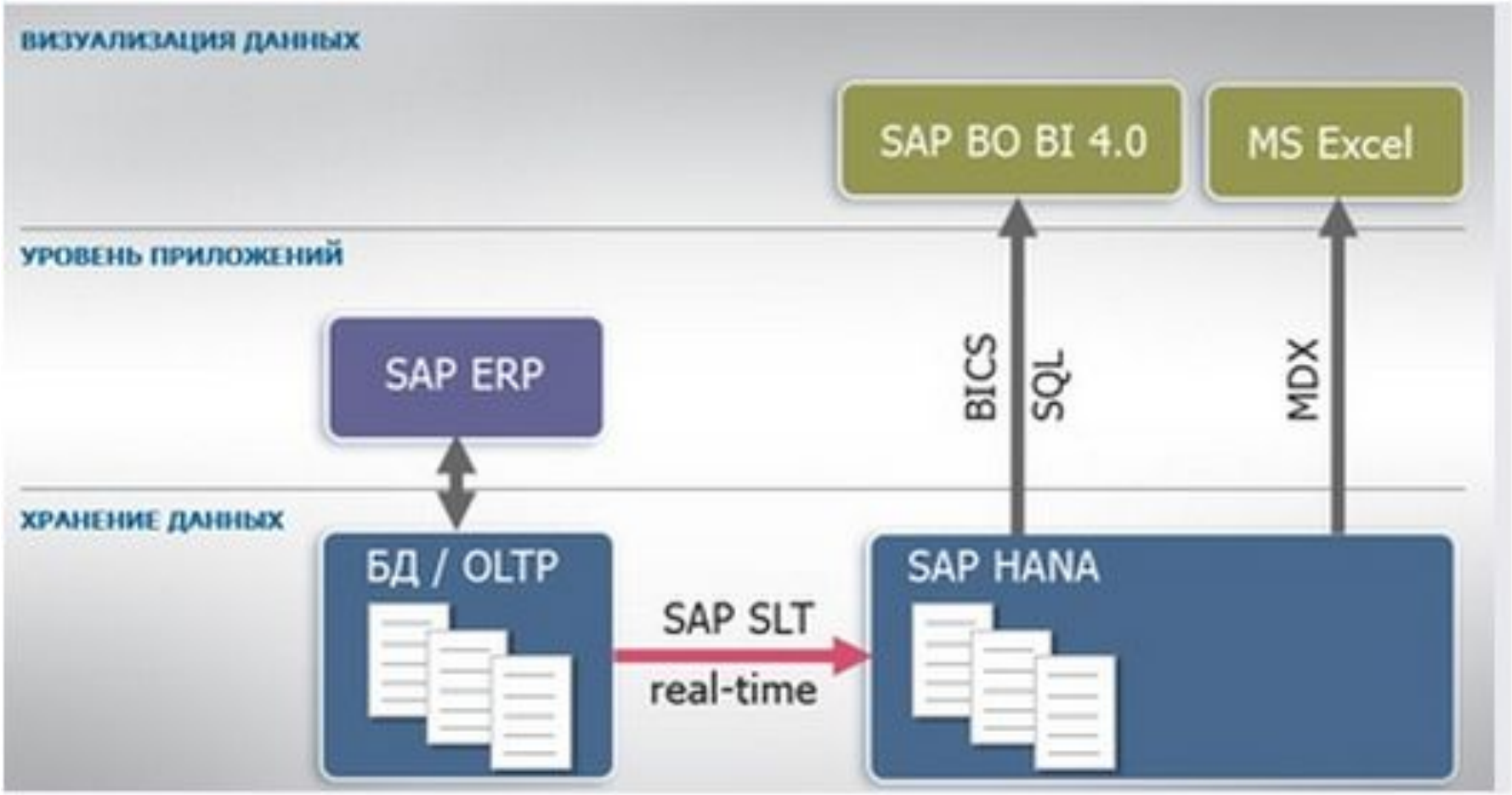


Рис.2 Взаимодействие систем при обработке данных SAP ERP

Решение на основе комплекса SAP HANA

Как и в случае с традиционным хранилищем данных, в качестве источников информации SAP HANA может использовать не только ERP-систему, но и другие приложения. В этом случае для загрузки данных используются средства SAP BusinessObjects Data Services (рис.3).

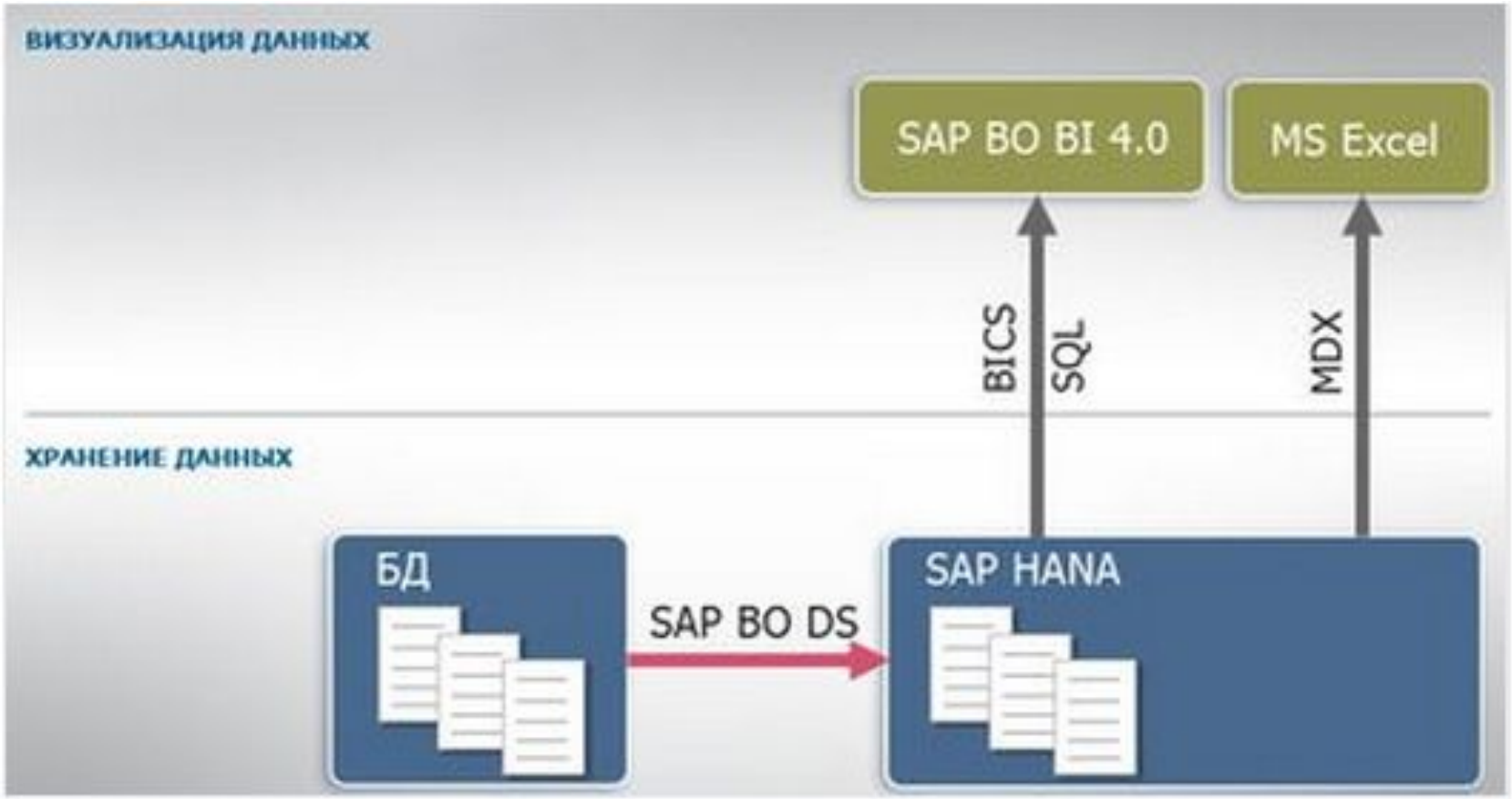


Рис.3 Взаимодействие систем при обработке данных

Если используются данные из различных источников на основе комплекса SAP HANA.

В такой конфигурации скорость обработки запросов даже при анализе больших объемов данных остается высокой, хотя о работе в режиме реального времени (с точки зрения актуальности данных) речь уже не идет.

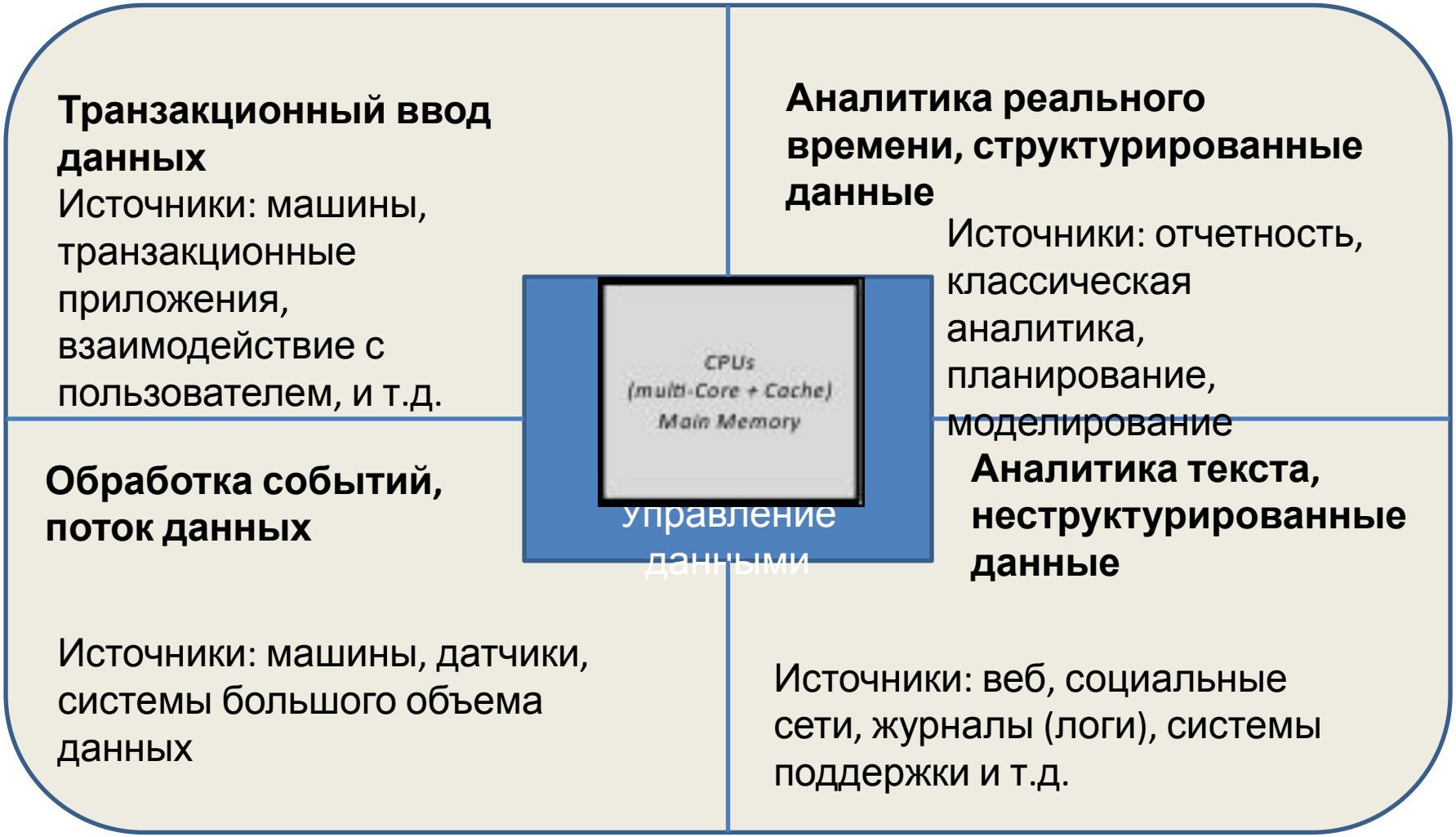
Сейчас есть проекты, в которых SAP HANA выступает в качестве базы данных и для SAP BW. Такой подход помогает сохранить выстроенную инфраструктуру компании, сделанные ранее инвестиции и значительно повысить скорость работы SAP BW. Кроме того, у компании появляется возможность постепенно технологически развивать текущий ландшафт, но при этом существенно экономить финансовые средства.

Важная особенность платформы - функциональность Calculation Engine. Она позволяет выполнять ресурсоемкие операции над данными непосредственно в оперативной памяти и значительно сокращать обмен информацией между базой данных и приложениями. В результате SAP HANA может использоваться и как платформа для разработки in-memory-приложений, которые будут полностью работать в оперативной памяти. Для определения закономерностей, прогнозирования ситуации и других видов предиктивного анализа Calculation Engine может использовать собственную библиотеку функций статистической обработки данных. Также для этих целей имеется возможность задействовать всю мощь и широкий спектр библиотек открытого языка R. Тем самым аналитика переходит на новый уровень – от констатации фактов к выявлению закономерностей и предвидению ситуации.

В отличие от хранилищ данных, в состав SAP HANA входит не только программное обеспечение, но и аппаратная часть – комплекс blade-серверов на базе архитектуры Intel Nehalem-EX CPU. Такой подход также позволяет экономить ресурсы на интеграцию комплекса в уже существующий на предприятии ландшафт.

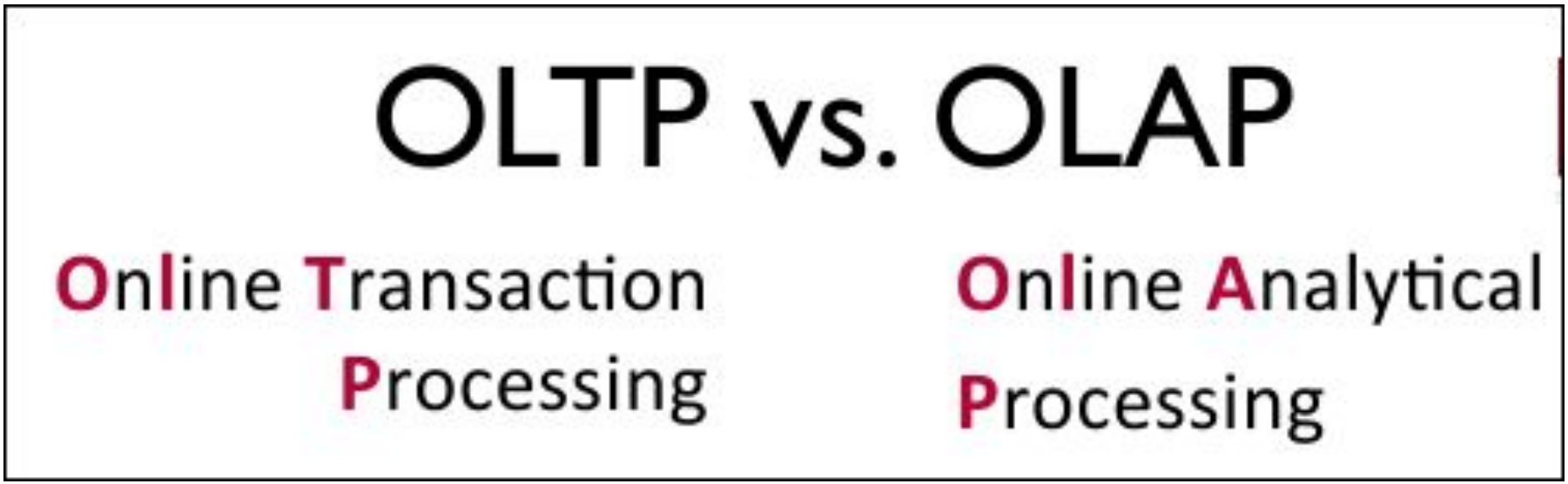
2. Новые требования к системе управления базой данных предприятия

Проблема: разнообразные приложения, создающие данные



Таким образом, требование первое - система управления базой данных предприятия должна быть в состоянии обрабатывать данные, поступающие из различных типов источников.

OLTP и/или OLAP – альтернатива или объединение



Далее, система управления данными предприятия должна быть в состоянии обрабатывать транзакционные и аналитические типы запросов, которые отличаются в нескольких измерениях. Типичные запросы для **оперативной обработки транзакций (OLTP)** – это создание заказов на продажу, счетов-фактур, бухгалтерских данных, выборка заказа для одного клиента, или отображение основных данных клиента.

Online Analytical Processing (OLAP) состоит из аналитических запросов. Типичные запросы OLAP-стиля - напоминания (напоминание об оплате), перекрестные продажи (продажи дополнительных продуктов или услуг клиенту), оперативная отчетность, или анализ тенденций на основе истории.

Так как всегда считалось, что эти типы запросов значительно отличаются, было принято разделение системы управления данными на две отдельные системы обработки OLTP и OLAP запросов. В литературе утверждается, что OLTP нагрузка при записи интенсивней, в то время как OLAP-нагрузки появляются только при чтении, и что две рабочие нагрузки полагаются на "Противоположные законы физики баз данных".

Тем не менее, исследования в текущих корпоративных систем показали, что это утверждение не соответствует действительности. Основное различие между системами, которые обрабатывают эти типы запросов, в том, что OLTP системы обрабатывают больше запросов с одним объектом выборки или запросов, которые из большого объема данных возвращают всего несколько объектов, в то время как системы OLAP агрегируют лишь несколько столбцов таблицы, но для большого количества объектов.

Для синхронизации аналитической системы с транзакционной системой (системами), необходим многотиражный ETL (Extract-Transform-Load) процесс.

Процесс ETL занимает много времени и является относительно сложным, потому что все соответствующие изменения должны быть извлечены из внешнего источника или источников, если их несколько, данные преобразуются в формат, необходимый для аналитики, и загружаются в целевую базу данных.

Недостатки разделения OLAP и OLTP

Несмотря на то, что разделение базы данных на две системы позволяет рабочей нагрузке специфичным образом оптимизироваться в обеих системах, оно приводит к целому ряду недостатков:

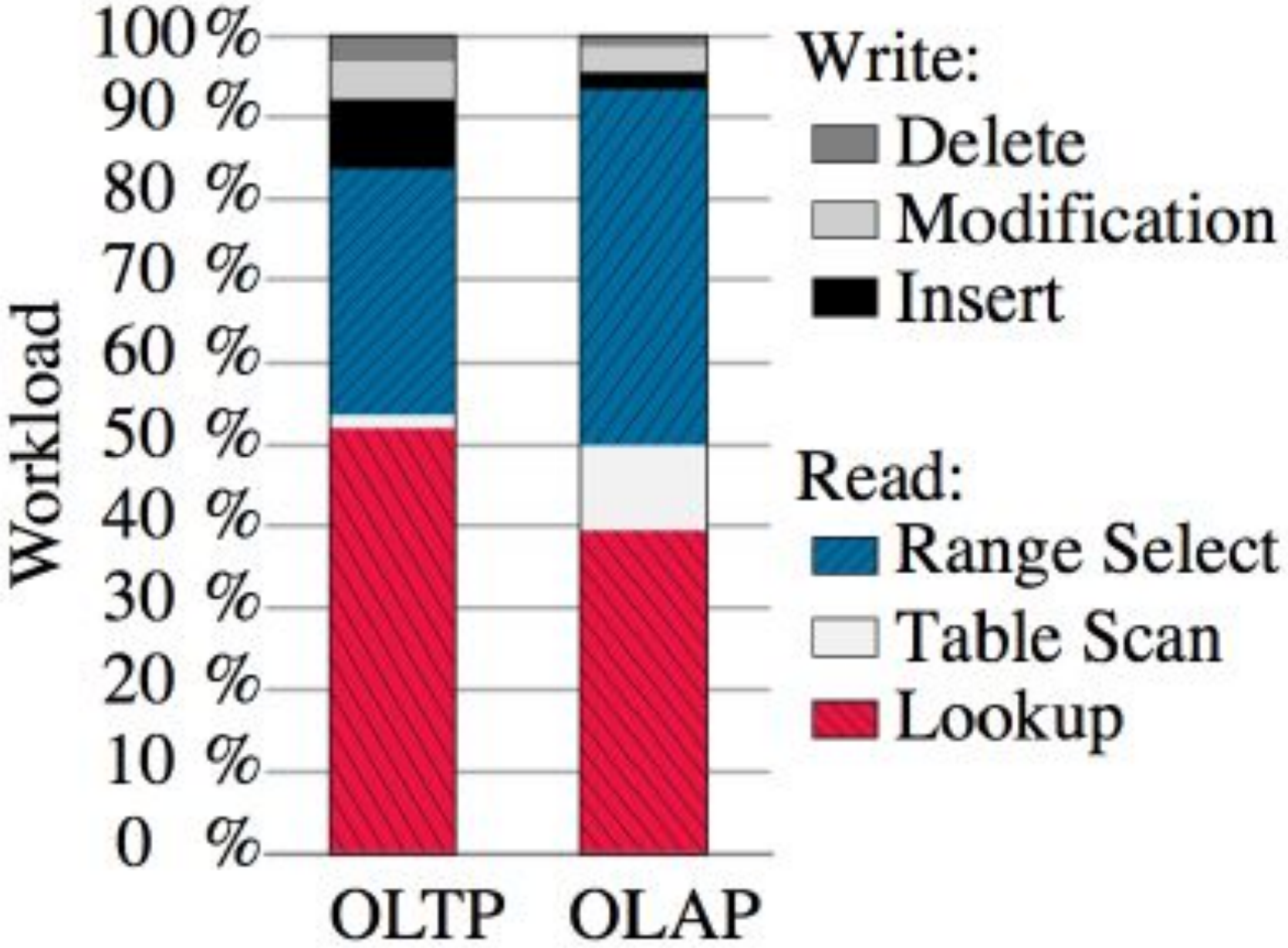
- Система OLAP не имеет последних (актуальных) данных, так как процесс ETL вводит задержку. Задержка может варьироваться от нескольких минут до нескольких часов или даже дней. Следовательно, многие решения должны опираться на устаревшие данные, а не использовать новейшую информацию.
- Для достижения приемлемой производительности, системы OLAP работают с предопределенными материальными агрегатами, что уменьшает гибкость запросов пользователя.
- Избыточность данных является высокой. Аналогичная информация хранится в обеих системах, просто различно оптимизирована.
- Схемы, используемые OLTP и OLAP системами, различны, что вносит сложность для приложений, использующих их обе, и сложно для процесса ETL синхронизации данных между системами.

OLTP против OLAP: миф различного шаблона доступа

При обработке транзакций часто принимают, что доли чтения и записи равны, в то время как на самом деле в аналитической обработке доминирует больше чтение и варьируются запросы. Тем не менее, анализ рабочей нагрузки из нескольких систем реального заказчика показывает, что OLTP и OLAP системы не так и отличаются, как ожидалось в классических корпоративных системах. Как показано на рисунке, OLTP процессы в системе имеют более 80% запросов на чтение. Менее 10% от фактического объема - запросы на изменение существующих данных, например, обновление и удаление. Системы OLAP обрабатывают еще большее количество запросов на чтение, которые и составляют около 95% рабочей нагрузки.

Обновления в транзакционной нагрузке представляют особый интерес. Анализ обновлений в различных сферах промышленности отличаются (показано на рисунке).

Это подтверждает, что количество обновлений в OLTP системах является достаточно низким и варьируется по отраслям. В проанализированных высокотехнологичных компаниях, пики «частота обновления» около 12%, это означает, что около 88% из всех сохраненных в базе данных транзакций никогда не обновляются. В других секторах исследование показало, что возможны даже более низкие проценты обновления, например, менее 1% в банковском и дискретном производстве.


100 %
90 %
80 %
70 %
60 %
50 %
40 %
30 %
20 %
10 %
0 %
Workload
OLTP
OLAP
Write:
Delete
Modification
Insert
Read:
Range Select
Table Scan
Lookup

Предприятие с разреженными данными

Анализируя данные предприятия в стандартном программном обеспечении, специальные характеристики данных. Самое интересное, что большинство таблиц очень широки (по характеристикам объектов – сущностей) и содержат сотни столбцов. Тем не менее, многие атрибуты такого анализа не используются вообще: 55% из всех столбцов не используются в среднем компаниями. Это связано с тем, что стандартное программное обеспечение должно поддерживать работу многих потоков в различных отраслях промышленности и стран, однако одна компания не использует все из них. Кроме того, во многих колонках NULL или значения по умолчанию являются доминирующими, так что энтропия (содержание, разнообразие информации) в этих столбцах является очень низким (около нуля).

Но даже специфичные столбцы, которые используются в компании, часто имеют низкую мощность значений, т.е. существует очень мало различных значений. Часто в связи с тем, что эти модели данных в реальном мире, и каждая компания имеет лишь ограниченное количество продуктов, которые могут быть проданы, для ограниченного числа клиентов, с помощью ограниченного числа сот|