Проектирование баз данных, ч.2

ФИО преподавателя: Смирнов Михаил Вячеславович

e-mail: smirnov.mirea@gmail.com

Лекция 10

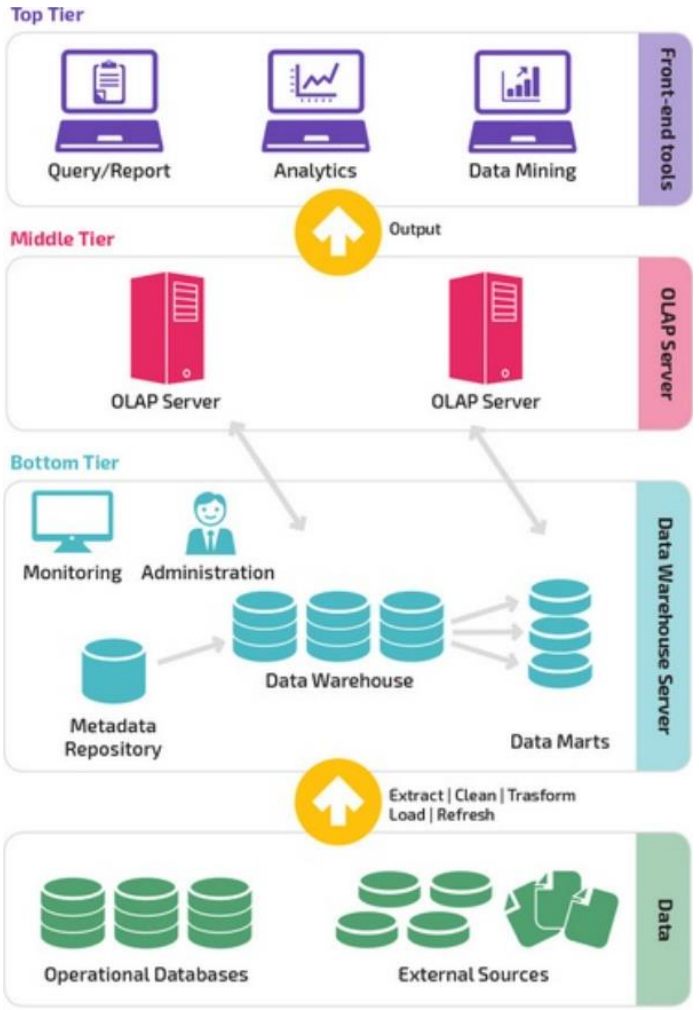
Основы использования хранилищ данных

Классическая трехуровневая архитектура хранилища данных

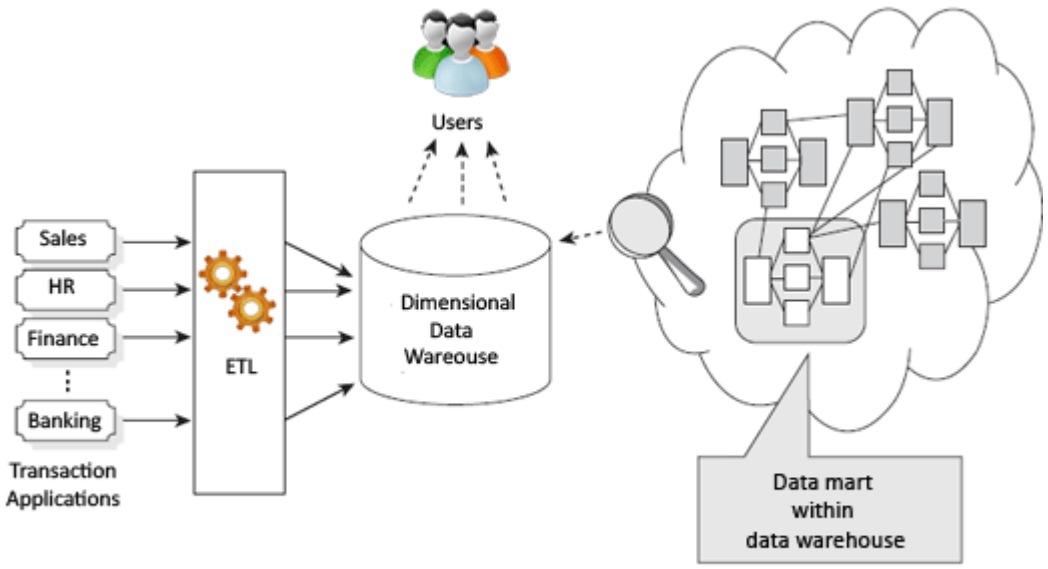
Нижний уровень: сервер базы данных, используемый для извлечения данных из множества различных источников.

Средний уровень: средний уровень содержит сервер OLAP, который преобразует данные в структуру, лучше подходящую для анализа и сложных запросов.

Верхний уровень: уровень клиента. Инструменты, используемые для высокоуровневого анализа данных, создания отчетов и анализа данных.

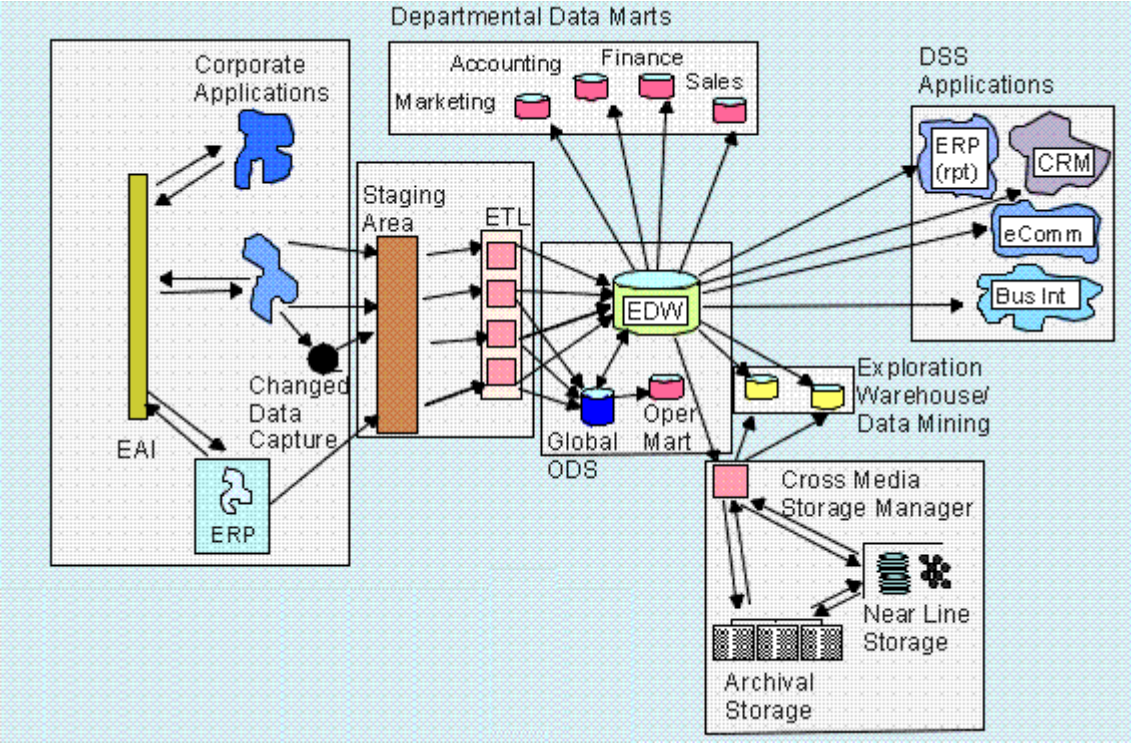


Модель хранилища мистера Кимбалла (Kimball)



- Transaction applications – системы или сервисы, созданные для обработки бизнес транзакций. Результатом обработки являются данные, сохраненные в РМД или в файлах.
- ETL – сервисы приведения данных к стандартному виду. Расшифровывается как достать, изменить, загрузить.
- A dimensional data warehouse. Содержит корпоративные данные в виде, наиболее удобном для гранулирования. Используется модель измерений. Как правило имеет форму звездочки. Аналитические системы способны извлекать данные напрямую из модели.
- Datamart. Витрина данных. Существует за пределами хранилища и представляет собой логический концепт.

Модель хранилища мистера Инмона (Inmon)



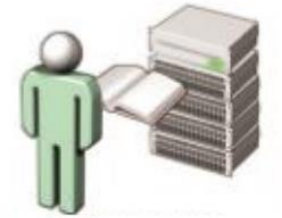
- Corporate Applications. В данной модели называются также системами источника, и предоставляют данные в хранилище.
- ETL Processes. Также называются data services. После обработки данных в рамках установленного формата, они загружаются в корпоративное хранилище. Процессы ETL могут выполняться в режиме batch или в режиме real-time.
- Enterprise data warehouse. Центральный элемент архитектуры, объединенный репозиторий атомарных данных, находящихся в третьей нормальной форме (см. реляционную модель).
- Data marts. Выбирают данные из хранилища с помощью инструментов агрегации. Аналитические приложения обращаются не к хранилищу, а к заранее приготовленным витринам данных.

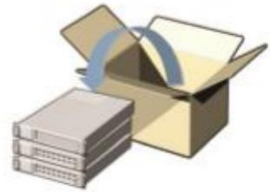
Критерии выбора Kmb-Inm.

Характеристика	Kmb	Inm
Уровень бизнес-решений	Тактический	Стратегический
Требования по интеграции данных	Отдельные бизнес-требования	Корпоративные бизнес-требования
Структура данных	Простые количественные данные, KPI	Многослойные данные, качественные данные
Стабильность данных	Источники редко меняются	Источники часто меняются
Сложность исполнения	Маленькая команда проекта	Большая команда проекта
Время на развертывание	Срочно требуется хранилище	Есть существенное время на проектирование и релиз
Затраты	Сравнительно низкие затраты	Высокие затраты со старта

Варианты имплементации хранилищ данных


Сборка на заказ


Эталонные архитектуры


Готовые решения по ХД

Хранилища можно построить с использованием типового инструментария многопользовательских реляционных баз данных (Oracle, postgreSQL, MS SQL Server).

Также, может быть использовано специальное программное обеспечение для построения хранилищ данных (Autonomous Data Warehouse, IBM Netezza Performance Server).
Такие продукты, помимо софта могут также комплектоваться и железом.

Типовой состав участников проекта реализации хранилища данных

менеджер проекта

архитектор решения

инженер данных

АД

специалист по инфраструктуре

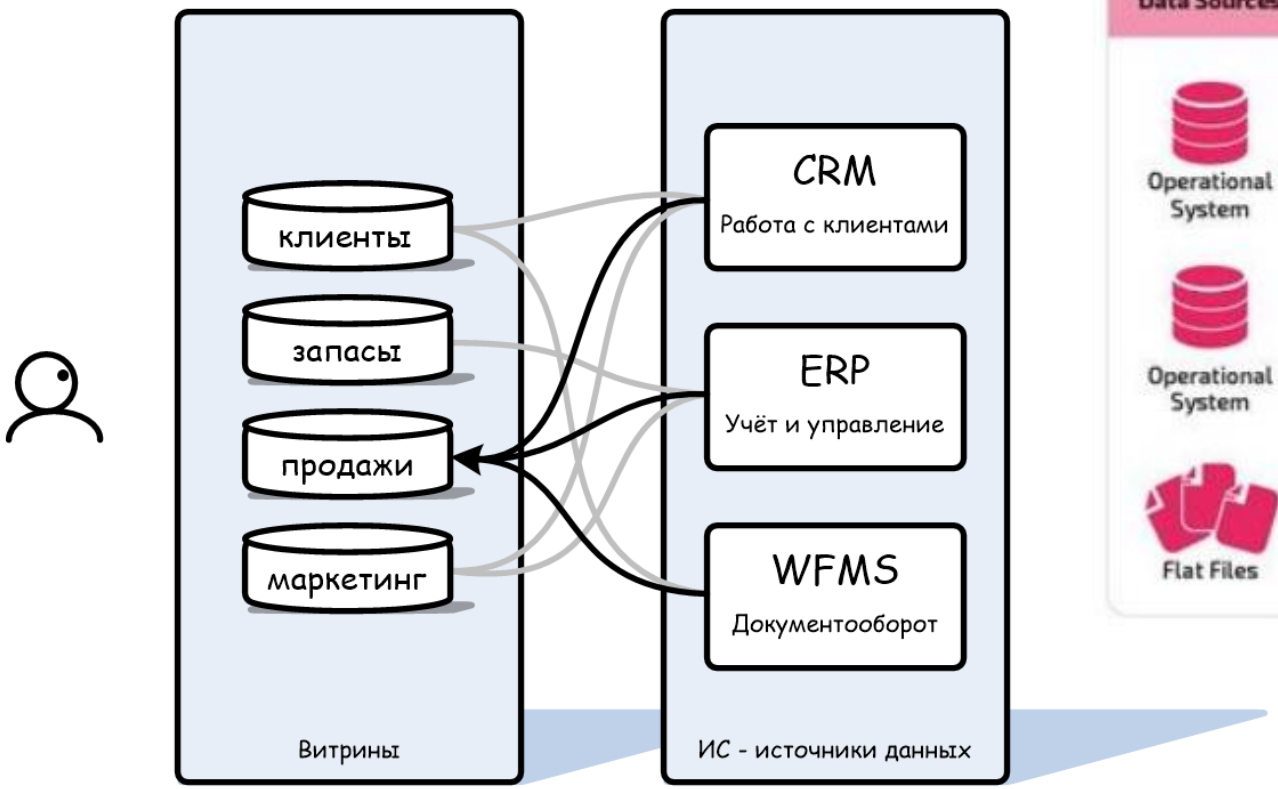
разработчик ETL процедур

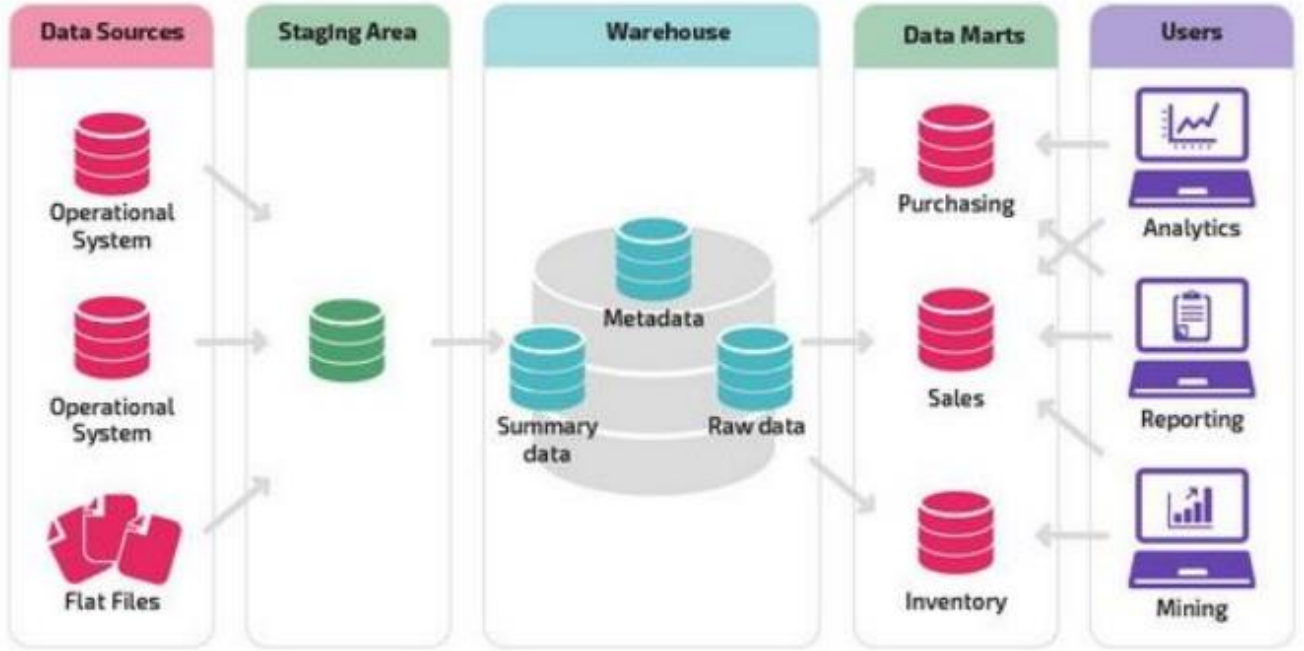
бизнес-аналитик

тестировщик

специалисты по обслуживанию данных

Место витрин данных (data mart) в процессе аналитики данных





Почему витрины, а не простые запросы и представления?

- Изоляция данных. Обновление данных, например загрузка справочника, может быть растянуто во времени, при этом до окончания загрузки текущая версия справочника должна быть полностью доступна с исключением «грязного чтения» загружаемой версии.
- Гарантии атомарности операций при обновлении данных. В случае сбоев и ошибок загрузки данных витрина остаётся в состоянии, которое предшествовало сбойному процессу. Другими словами, или данные обновляются полностью, или не обновляются вовсе, не оставляя следов сбойных операций.
- Системная темпоральность. Мало какая реляционная СУБД имеет функцию системной темпоральности «из коробки». Ведение системного времени и версионирование записей по системному времени позволяет сравнивать состояние данных витрины между двумя разными моментами времени или проводить «расследование», основываясь на данных, которые были в витрине в определенный момент в прошлом.
- Эффективное выполнение различных видов запросов: сравнительно редких и тяжелых аналитических запросов, затрагивающих большой объем данных (OLAP-нагрузка), и множества одновременных простых запросов (OLTP-нагрузка). Как правило, СУБД заточены на какой-то один вариант нагрузки: OLAP или OLTP.

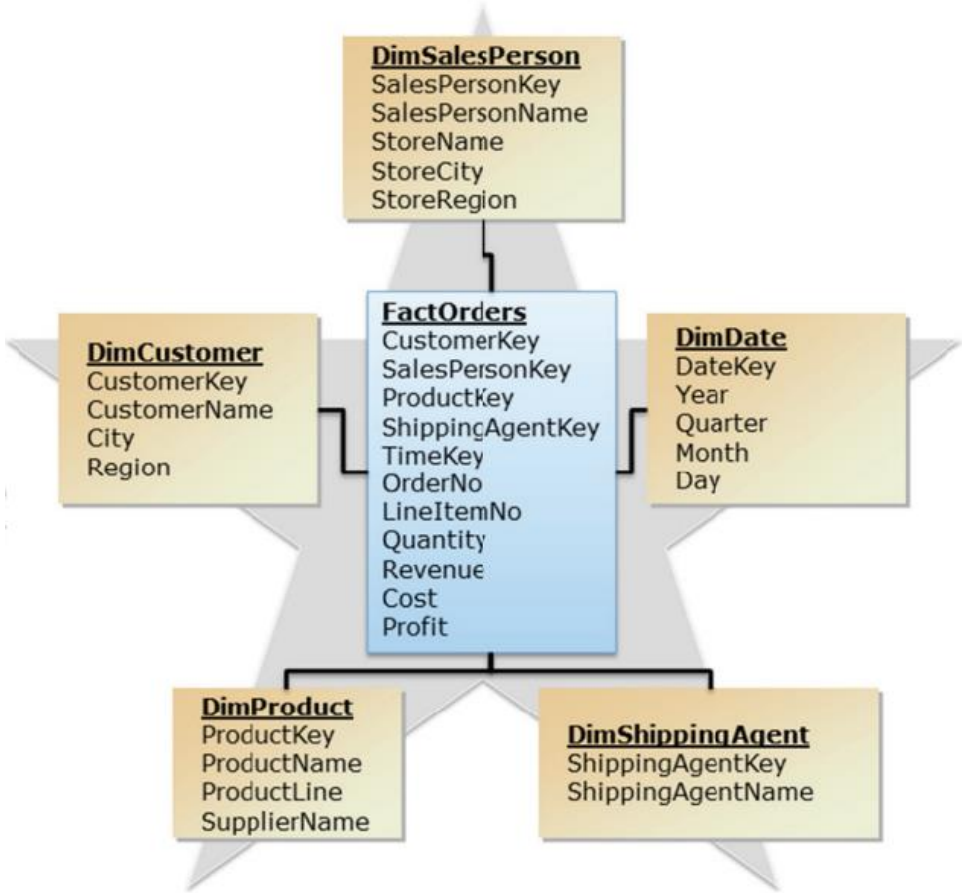
ETL и хранилище данных

1. Извлечение данных из пула источников данных. Данные хранятся во временной промежуточной базе данных.
2. Выполняются операции преобразования, чтобы структурировать и преобразовать данные в подходящую форму для целевой системы хранилища данных.
3. Структурированные данные загружаются в хранилище и готовы к анализу.

ELT и HDFS

1. Данные сразу же загружаются после извлечения из исходных пулов данных. Промежуточная база данных отсутствует, что означает, что данные немедленно загружаются в единый централизованный репозиторий.
2. Данные преобразуются в системе хранилища данных для использования с инструментами бизнес-аналитики и аналитики.

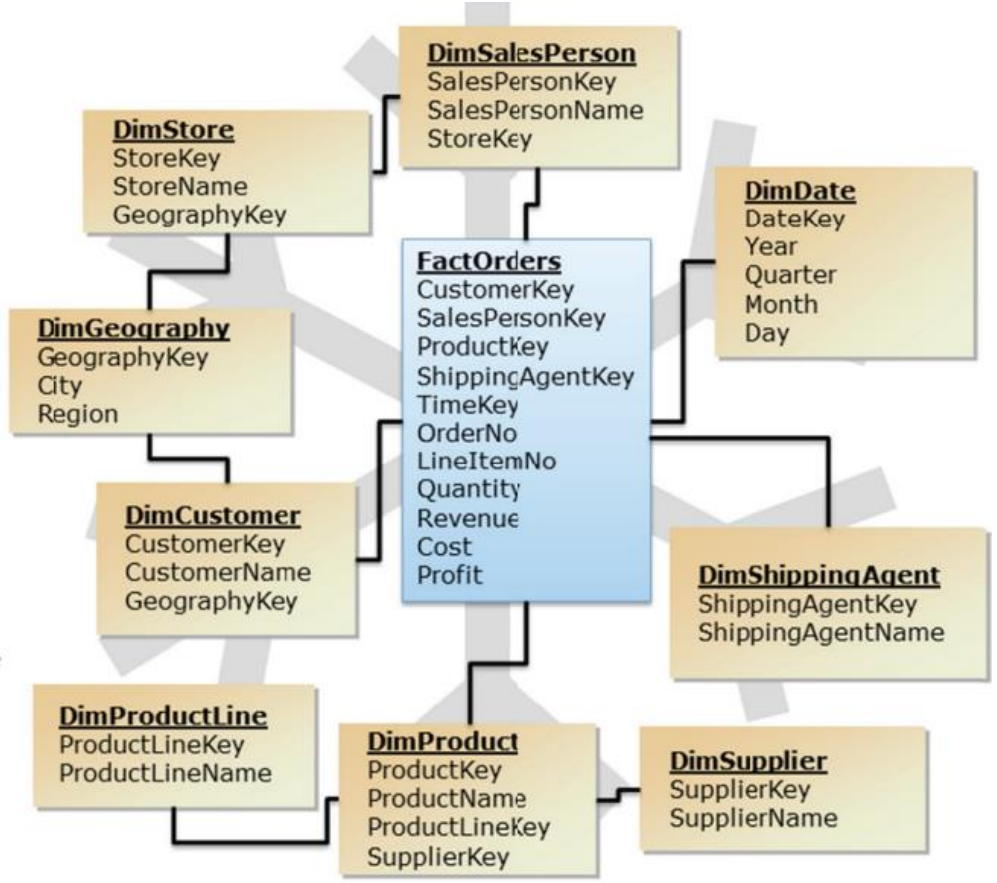
Логическая схема хранилища звезда (денормализация)



1. Сгруппированные измерения располагаются в таблицах измерений.
2. Сгруппированные показатели в таблице фактов.
3. Объединение таблицы фактов и таблиц измерений с использованием внешних ключей.

4. Синяя – таблица фактов. Красные – таблицы измерений.

Логическая схема хранилища звезда (нормализация)



Нормализованные таблицы измерений.

Применяются, когда:

- измерение может быть разделено между несколькими измерениями
- - имеется иерархия и таблица измерений содержит маленький набор данных, который может часто обновляться
- - имеются множественные таблицы фактов с различной грануляцией

Чтение на дом

- https://www.oracle.com/ru/database/what-is-a-data-warehouse/#link1
- https://habr.com/ru/post/650237/ - статьи и ссылка на практикум по витринам данных
- https://coderlessons.com/tutorials/bolshie-dannye-i-analitika/teoriia-khraneniia-dannykh/10-oltp-protiv-olap - про OLTP и OLAP
- https://zen.yandex.ru/media/id/5aef49c279885e47d5eb6199/o-bazah-dannyh-prosto-razlichiia-olap-i-oltp-5bf5557e5184cc00a99028ff - про OLTP и OLAP очень просто
- Английский Кренке, стр. 573-587.

Спасибо за внимание!