



Обработка статистической информации. Хранилища данных.

Информационное обеспечение предметно-ориентированных ИС.

Проблемы бизнеса



1. Важная для бизнеса информация распределена среди разнообразных, зачастую разнородных систем.
2. Поиск информации, требуемой для принятия бизнес-решений это очень затратное по времени мероприятие, к тому же с высокой вероятностью недостоверности.
3. Очень сложно ответить на фундаментальные вопросы бизнеса.



Фундаментальные вопросы бизнеса (Forbes)

В чем ценность моего продукта
(предложения)?

Насколько масштабируется
мой бизнес?

Какая оптимальная (максимальная)
цена у моего товара?

Будет ли спрос на мой товар?
Какой его объем?

Как продавать товар?

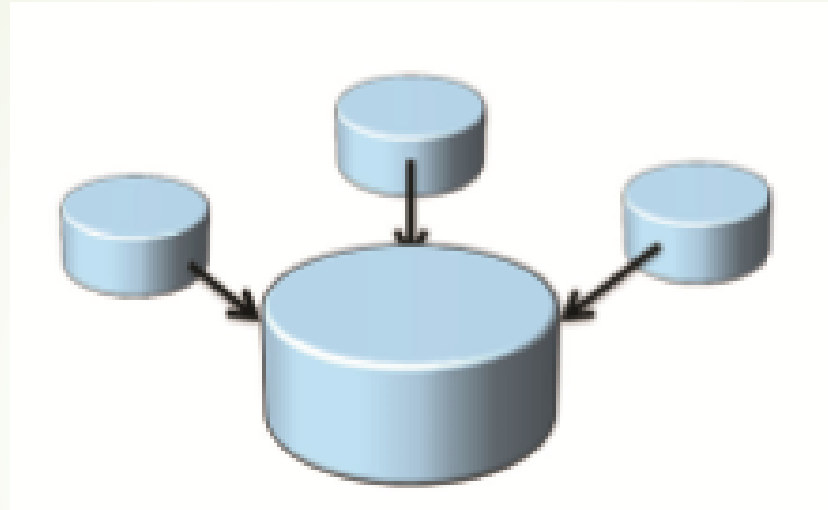
Какой капитал нужен для...?

Прогноз по операционной окупаемости...

Прогноз по инвестиционной окупаемости

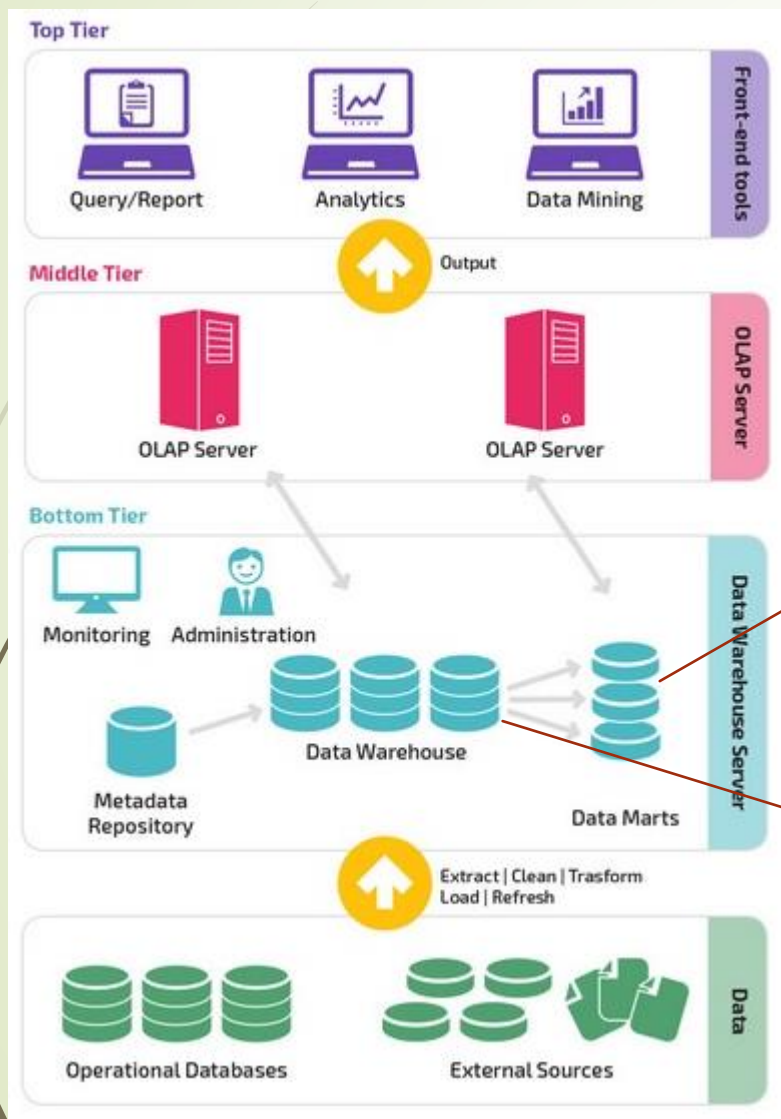
Конечная цель (достижимость?)

Что такое хранилище данных?



1. Централизованное хранилище бизнес информации, необходимой для отчетности и анализа.
2. Обычно, хранилище данных:
 - содержит большие массивы архивных (исторических данных);
 - оптимизировано под запросы к данным (а не на добавление-изменение, как РМД);
 - пополняется новой бизнес информацией через определенные интервалы времени;
 - представляет собой основу для решения корпоративных задач.

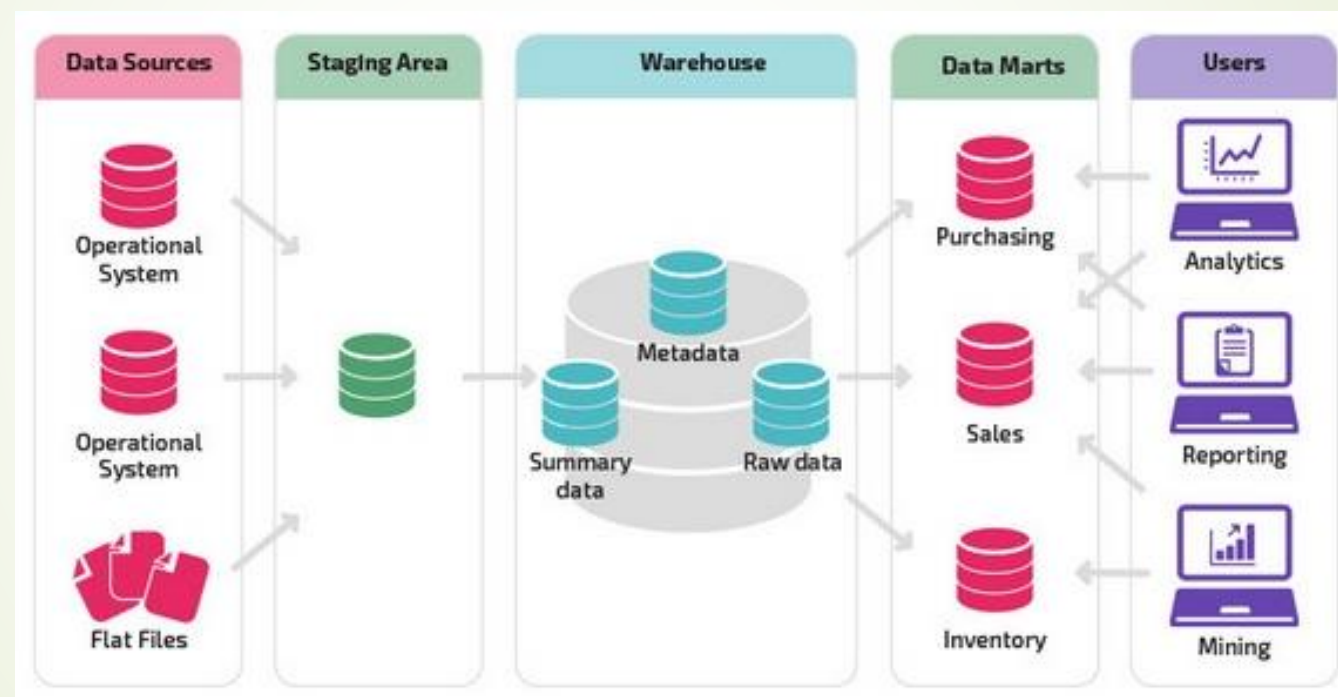
Классическая трехуровневая архитектура хранилища данных



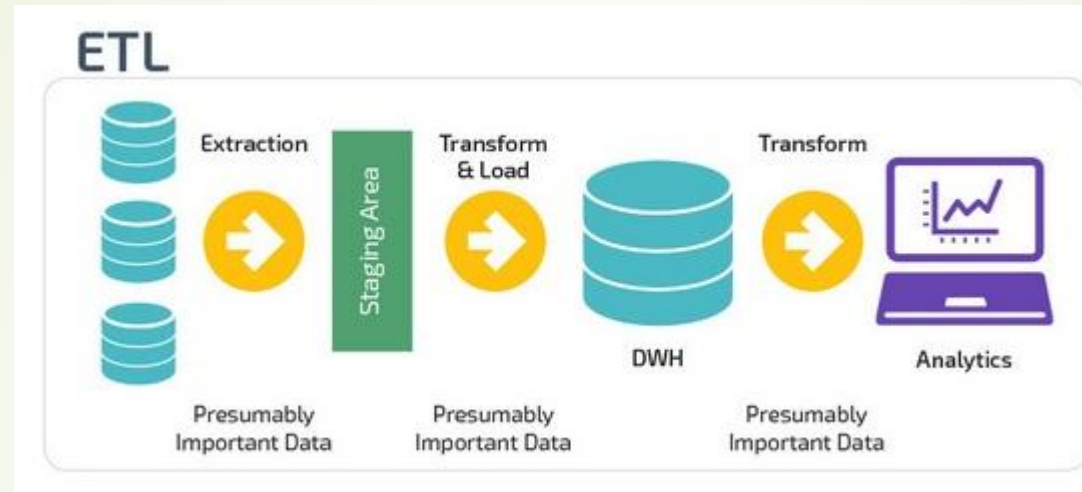
Kimball's WH: Хранилище данных - это просто **сочетание различных витрин данных**, которые облегчают отчетность и анализ, "снизу-вверх"

Inmon's WH: хранилище данных является централизованным хранилищем всех корпоративных данных, "сверху-вниз"

Подробнее про витринки



ETL (extract, transform, load)



1. Сначала извлекают данные 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.

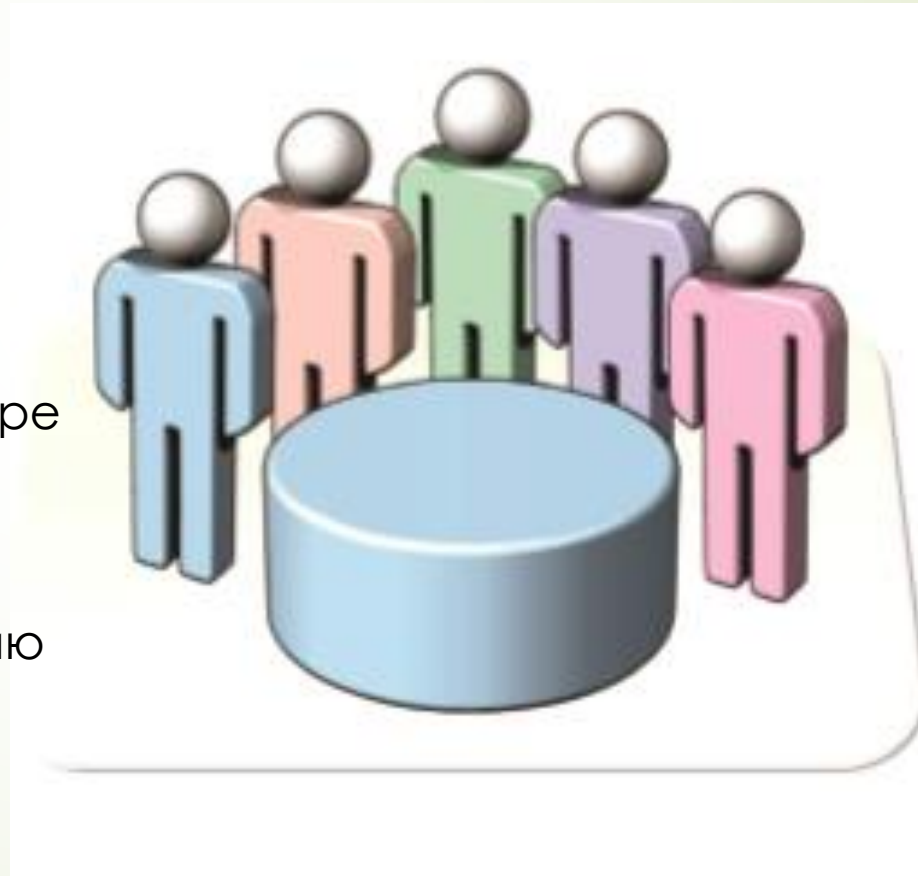
ELT (extract, load, transform)



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.

Участники проекта реализации хранилища данных

- менеджер проекта
- архитектор решения
- инженер данных
- АД
- специалист по инфраструктуре
- разработчик ETL процедур
- бизнес-аналитик
- тестировщик
- специалисты по обслуживанию данных



Варианты имплементации



Сборка на заказ



Эталонные архитектуры



Готовые решения по ХД



Логическая схема хранилищ данных

1. Введение в моделирование измерений данных
2. Схемы “звезда” (“star”)
3. Соображения на тему таблиц измерений
4. Соображения на тему таблиц фактов
5. Схемы “Снежинка” (“snowflake”)
6. Временные измерения данных

Моделирование измерений данных

Бизнес вопросы сфокусированы на показателях, агрегированных из бизнес измерений.

Показатели – это факты о бизнесе.

Измерения – это места, где могут быть агрегированы показатели.

Примеры измерений:

- доход от продаж, по продавцам
- прибыль, по продуктам
- объем заказов, по продуктам
- стоимость, по продуктам
- доход от продаж, по покупателям
- прибыль, по регионам
- доход от продаж, за отчетный квартал



Схемы “Звезда”

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



Синяя – таблица фактов.
Красные – таблицы измерений.

Соображения относительно таблиц измерений

Денормализация:

- таблицы измерений обычно “широкие”
- дублирующие значения обычно предпочтительны для соединений

Ключи:

- создаются новые суррогатные ключи для идентификации каждого столбца
- наибольшую производительность обеспечивают простые ключи типа integer
- содержат бизнес-ключи



Подстраиваемые изменения:

Измерения могут использоваться многими таблицами фактов.

Соображения относительно таблиц фактов

Грануляция

- использование самого нижнего уровня детализации, который относится ко всем измерениям
- создание нескольких таблиц фактов в случае, когда требуются несколько грануляций

Ключи:

- первичный ключ композитный, включает в себя внешние ключи таблиц измерений

Параметры:

- аддитивные, могут быть агрегированы во всех измерениях
- неаддитивные
- полуаддитивные

Дегенеративные измерения.

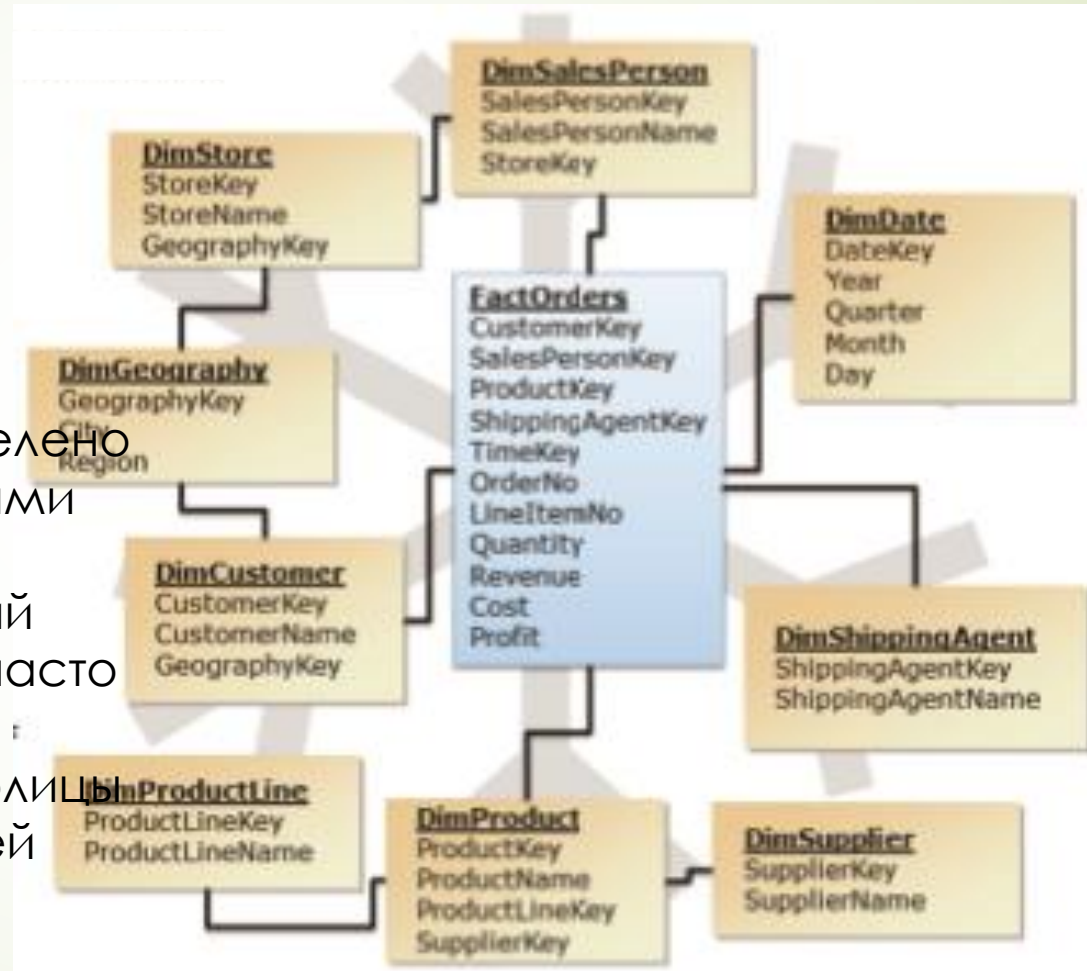


Схемы “Снежинка”

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



Временные измерения

Реализует последовательную гранулярность для временного анализа и отчетов.

Содержит в себе временную иерархию: Год-Квартал-Месяц-День

Может содержать специфические бизнес-атрибуты (праздники, налоговые периоды и т.д.)

Может быть организована (заполнена) различными способами:

- SQL скрипты с функцией DATETIME
- функции Excel
- автогенерация средствами BI

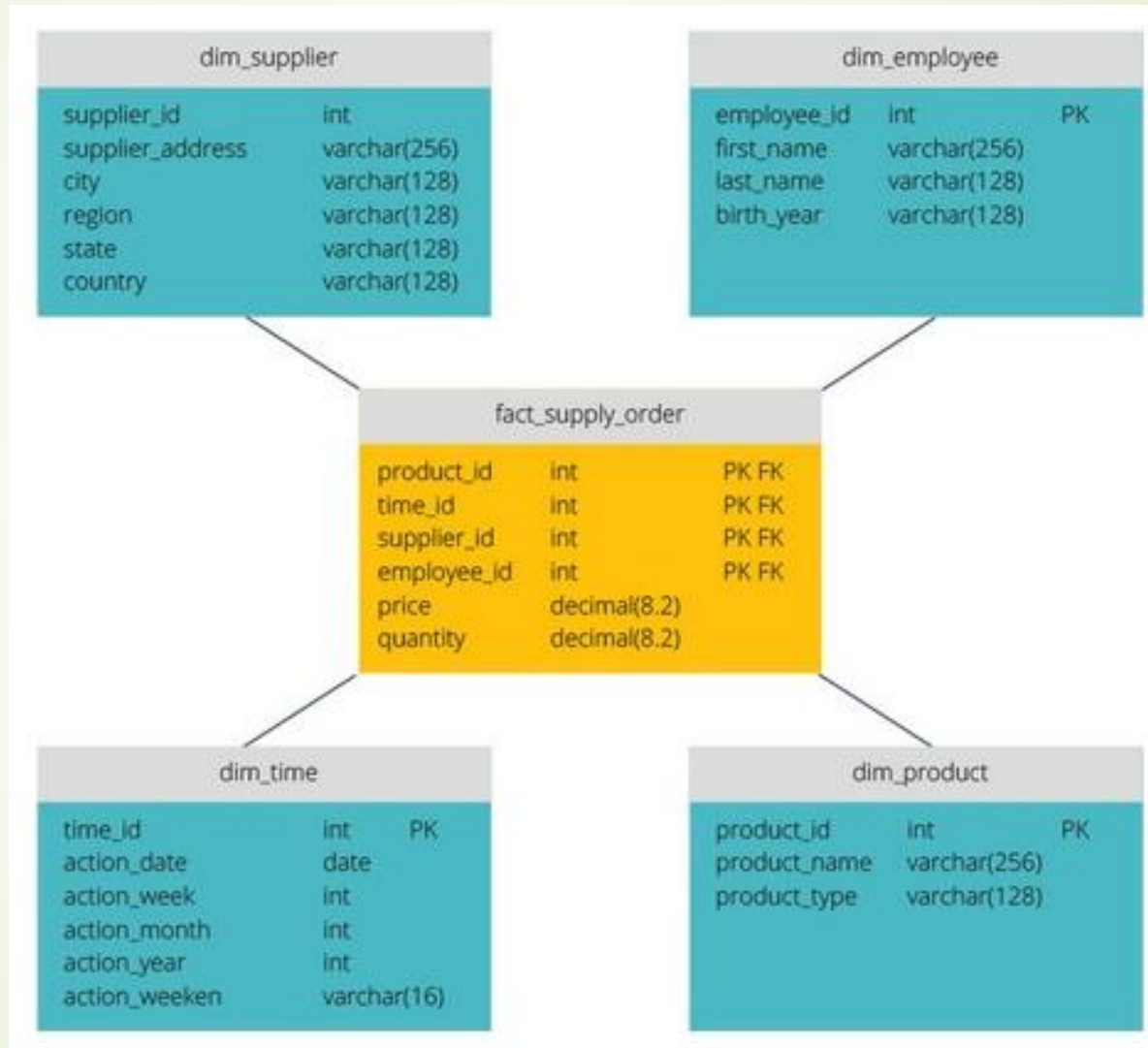




Задание: построить физическую
модель данных из приведенных схем
хранилищ данных



Тип: “Звезда”.



Тип: “Снежинка”.

