Средства интеграции и аналитики данных

Александр Дольник
alexander.dolnik@gmail.com

Цель лекции: сформировать общее представление о сборе и обработке больших объёмов данных

“Мы сами этим занимаемся, но я не понимаю зачем это нужно”

Специалист Колл-Центра

“Кто владеет информацией - тот владеет миром”

придумал Натан Ротшильд,
но употребил в своей речи
Уинстон Черчиль

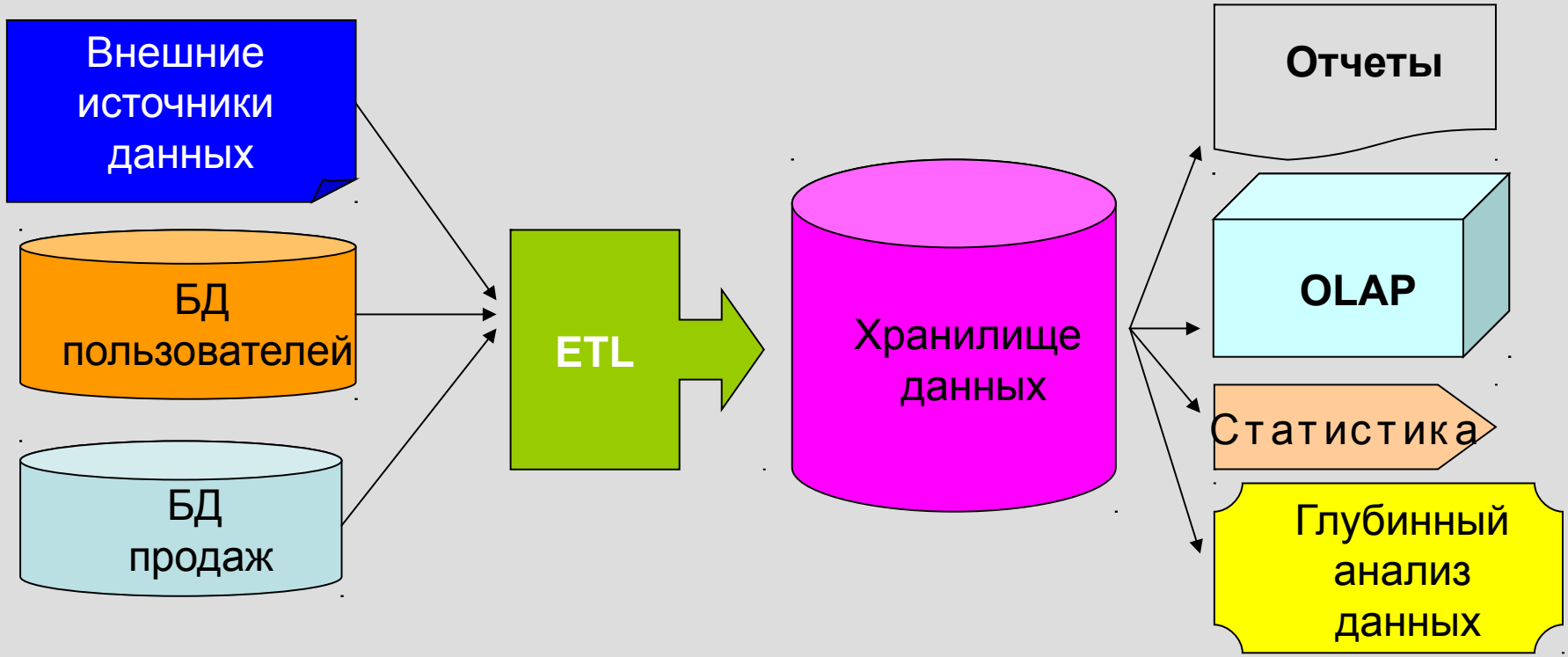
Задачи

- Обсудить проблемы и способы их решения в сфере интеграции данных
- Основы разработки больших систем по хранению и обработки данных. Понятия об этапах проектирования
- Побывать на фронте - передовой анализа: аналитические запросы и отчёты

Обзор

- Интеграция данных
  - Основные этапы и задачи интеграции данных
  - Пример среды Kettle
- Структуры хранения данных
  - Метаданные
  - Хранилища данных
  - Дизайн хранилищ данных
- Загрузка больших объёмов данных
  - Особенности выполнения запросов по загрузке данных
- На передовой анализа данных
  - Иерархии и рекурсивные функции
  - Аналитические запросы

Архитектура “Хранилищ данных”



Операционные БД

- Ежедневные операции
- Текущие данные

Хранилища данных

- Анализ и поддержка принятия решения
- Исторические данные
- Интегрированные данные из различных источников

ИНТЕГРАЦИЯ ДАННЫХ

Интеграция данных

- Определение. Что это такое?
- Какие задачи выполняет?
- Какие системы по интеграции данных известны? (альтернативы)

- Хороший процесс
  - Есть ли универсальное решение?
  - Визуализация и интерактивность

Интеграция данных

- Определение. Что это такое?
- Какие задачи выполняет?
- Какие системы по интеграции данных известны? (альтернативы)
  - Скрипты;
  - Приложения;
  - Сервисы, в том числе и web services
- Хороший процесс
  - Есть ли универсальное решение?
  - Визуализация и интерактивность

Пример.
Обогащение данными из сети

Ежедневная загрузка данных с сайта
http://www.atsenergo.ru/results/rsv/oes/index.htm

Час	Объем покупки, мВт.ч	Объем продажи, мВт.ч	Индекс рвц на покупку ээ, руб./МВт.ч.	Max индекс рвц, руб./МВт.ч	...
0	27 995.670	29 545.896	788.62	1056.68	...
1	27 769.408	29 090.217	792.49	1051.97	...
2	27 828.475	29 049.130	792.20	978.75	...
3	28 194.453	29 354.610	799.85	914.89	...
4	28 922.995	29 845.665	804.63	929.06	...
...	...	...	...	...	...

Типичные проблемы процессов интеграции

- Понятность и документированность
- Расширяемость и масштабируемость
  - Изменяющиеся требования;
  - Резкий рост обрабатываемых данных
- Устойчивость к сбоям и контроль
  - Обработка ошибок;
  - Логирование процесса;
  - Возможность приостанавливать и запускать с места остановки

К примеру. Устойчивость к сбоям...

- Временная недоступность сервера в момент загрузки
  - Ошибки HTTP: 400, 404, 500 и другие
- Изменение порядка столбцов и их переименование
  - Разработчики таблицы решили поменять местами колонки;
  - Может быть связано с добавлением новых характеристик;
  - Удалена по каким-либо причинам одна из характеристик
- Изменение единицы измерения
  - мВт.ч -----> кВт.ч
- Изменение формата представляемых данных
  - Разделитель запятая -----> разделитель точка

Проектирование процесса

- Четко разделить на три стадии:
  - сбор данных из внешних источников
  - преобразование и валидация данных
  - загрузка данных в базу и логирование
- Описать используемые в процессе метаданные
- Оценить вероятность ошибки и её критичность
- Продумать механизмы реакции на ошибки и расчитать требуемые ресурсы с учётом обслуживания процесса в будущем

Создание процесса

Используйте специализированные инструменты с графической средой разработки

- Предположим, данные с веб сайта заложен в wget
- Создадим процесс с использованием Kettle

Видео с использованием Kettle

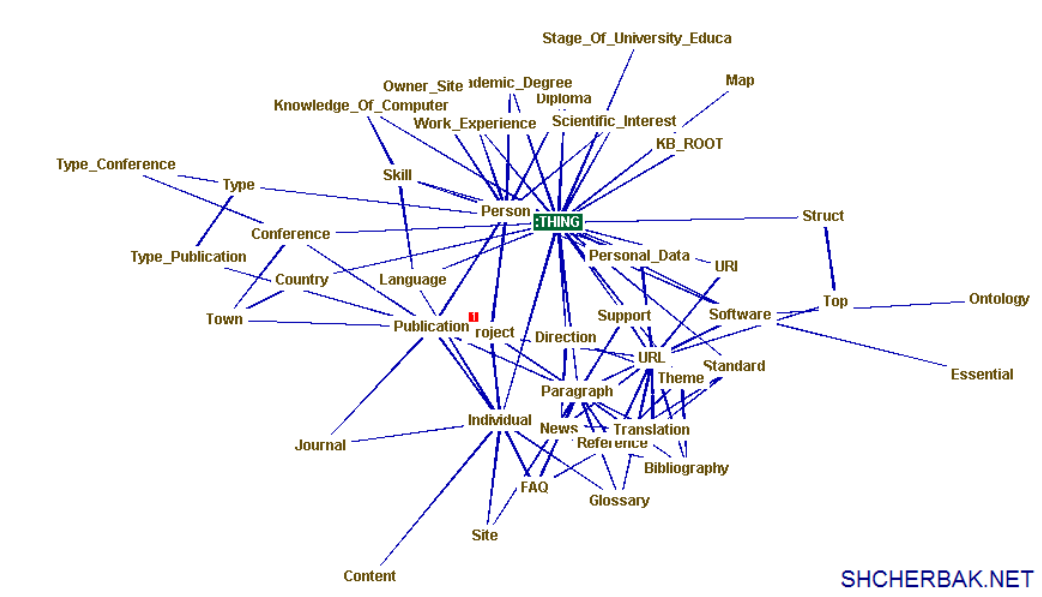
СТРУКТУРЫ ХРАНЕНИЯ ДАННЫХ

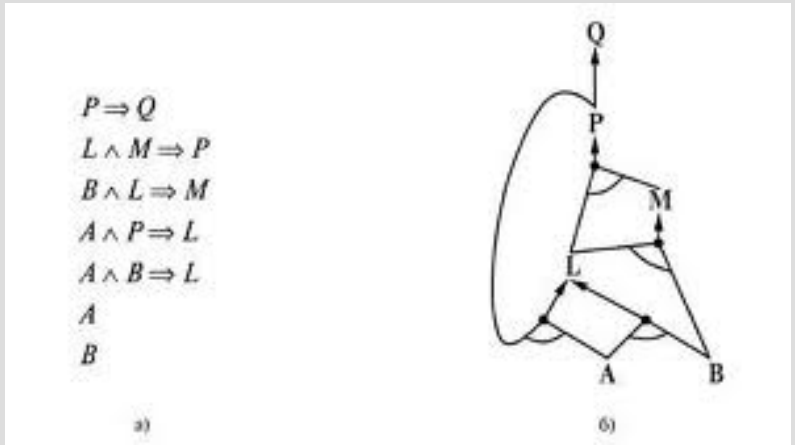
Метаданные

- Типы данных, простые ограничения типа max, min, формат...
- Проверки и зависимости между значений:
  - среднее между максимальным и минимальным
  - проверка на корректность формирования email
- Функциональные зависимости

Онтологии

- Определение
  - ключевые понятия +
  - связи +
  - аксиомы для логического вывода и доказательства теорем :)





Эксперимент по онтологиям

- Какие данные представлены в таблице?

?	?????????	????????????	????????????
1	**КНР**	1 357 150 000	30 марта 2013
2	**Индия**	1 233 221 000	30 марта 2013
3	**США**	315 370 000	30 марта 2013
4	**Индонезия**	245 000 000	1 января 2013

Эксперимент по онтологиям

- Какие данные представлены в таблице?

?	?????????	????????????	????????????
1	**КНР**	1 357 150 000	30 марта 2013
2	**Индия**	1 233 221 000	30 марта 2013
3	**США**	315 370 000	30 марта 2013
4	**Индонезия**	245 000 000	1 января 2013

Правильный ответ – население по странам мира

- Можно ли сделать программу для автоматического аннотирования таблиц?

Автоматическое аннотирование

- Поискать в сети, но такой таблицы может не быть
- Поиск фактов в сети из текста и вывод на основе статистики
- Выводы на основе онтологий

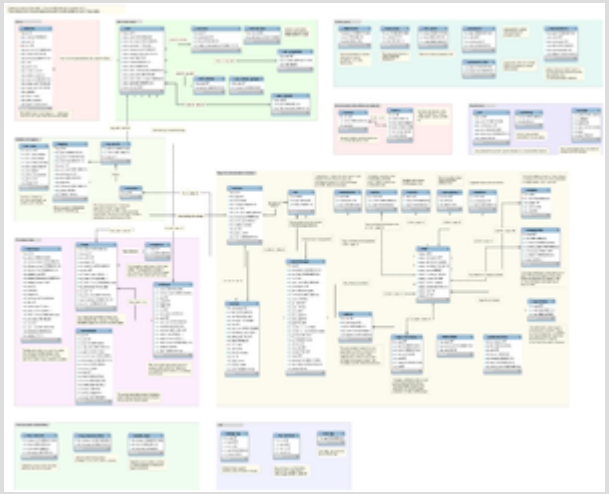
Petros Venetis, Alon Halevy, Jayant Madhavan, Marius Paşca, Warren Shen, Fei Wu, Gengxin Miao, and Chung Wu. 2011. Recovering semantics of tables on the web. Proc. VLDB Endow. 4, 9 (June 2011), 528-538.

Хранилища данных

- Онтологии являются связующим звеном между внешними источниками и хранилищами данных
- Хранилище данных – специализированная БД без транзакционной активности для нужд бизнес-анализа
  - Интегрированость
  - Некорректируемость
  - Зависимость от времени
  - Удобство в использовании аналитиками

Дизайн хранилищ данных

- Вариант 1. Как учили в теории БД – нормализация



- Вариант 2. Большая таблица фактов со многочисленными измерениями

Модели хранилищ данных

- Использование многоразмерной модели данных
- Известно как схема типа звезда или снежинка
- Много снега – снежный ком, много звезд – галактика
- Необходимо различать факты (facts), значения (measures), измерения (dimensions) и иерархии (hierarchies)

Схема снежинка

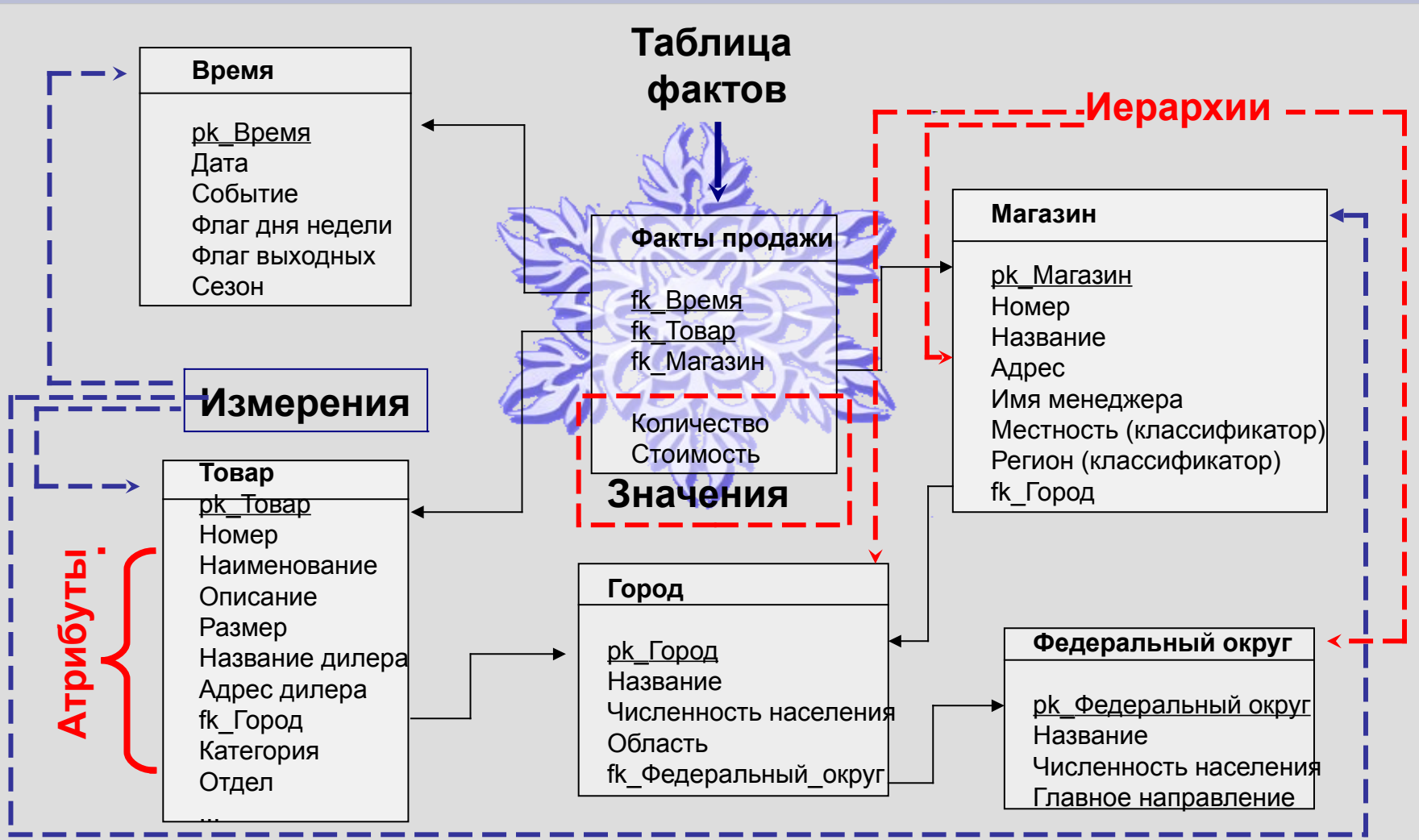
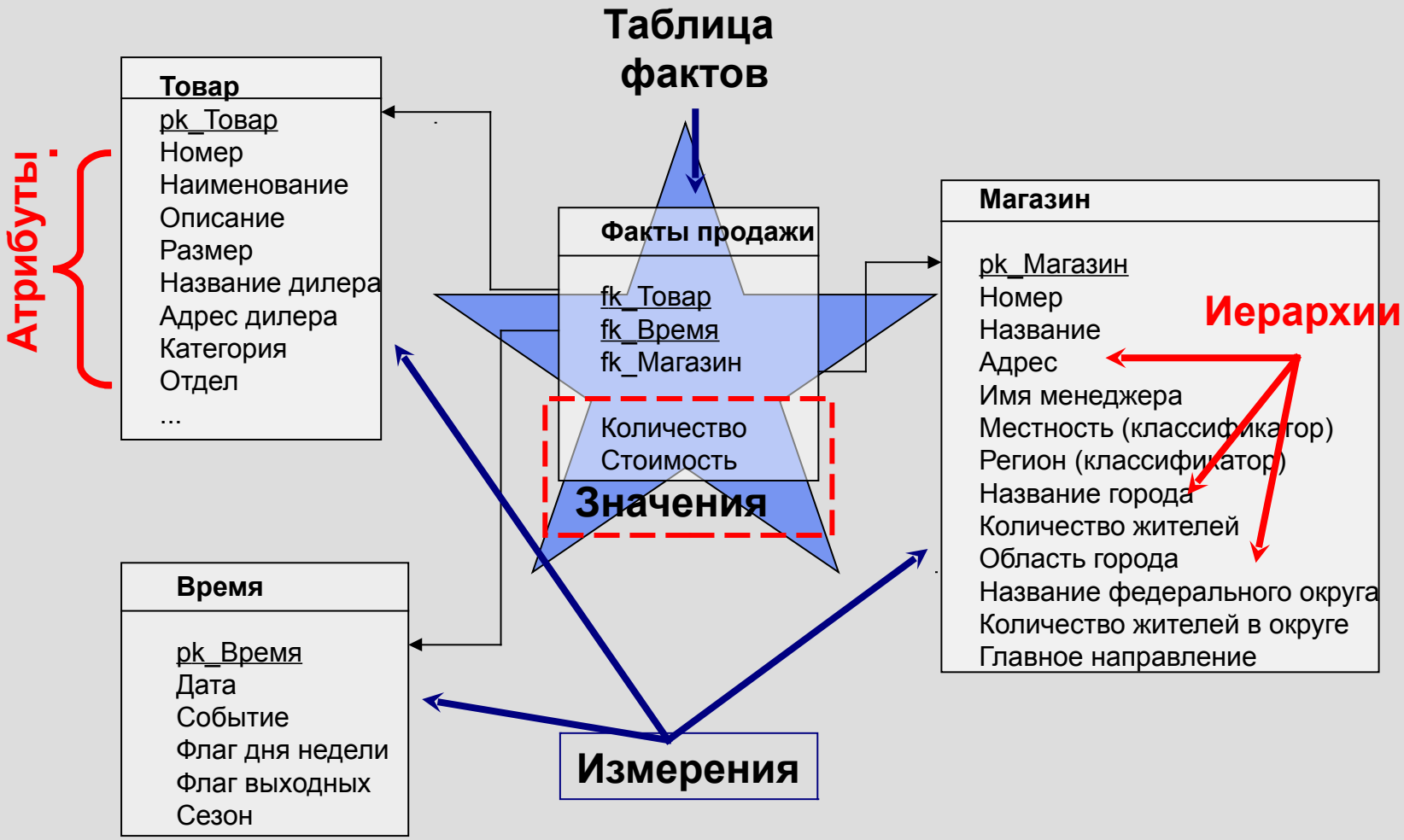


Схема звезда



ЗАГРУЗКА ДАННЫХ

Загрузка

Цель: быстрая загрузка в ХД

Загрузка Δ намного быстрее, чем всего подряд

- SQL подобное обновление медленное
  - Огромное перекрытие (оптимизация, блокировки, и так далее) для каждого вызова SQL
  - Блочная загрузка намного быстрее
- Индексы сильно замедляют процесс загрузки
  - Удалить и построить заново после загрузки
  - Можно сделать для части индекса
- Параллелизм
  - Измерения могут быть загружены одновременно
  - Таблицы фактов могут быть загружены одновременно

НА ПЕРЕДОВОЙ АНАЛИЗА ДАННЫХ

Три фронта анализа данных

- *On-Line Analytical Processing* (OLAP) - технологии интерактивной аналитической обработки данных для поддержки принятия решений
  - основан на мультиразмерном представлении данных;
  - Должен быть интуитивно понятен пользователю (drag-and-drop) в плане манипулирования данными
    - Динамическая навигация по иерархиям
    - Автоматическое агрегирование по значениям
    - Изменение позиций колонок и строк (pivot)
    - Включать сложные формулы
    - ….
- *Reporting* - построение сводных отчетов
- *Data mining* - технологии глубинной аналитической обработки данных

Типовые OLAP операции

- Две операции используют иерархии:
  - Roll-up – преобразует детализированные значения в агрегированные данные

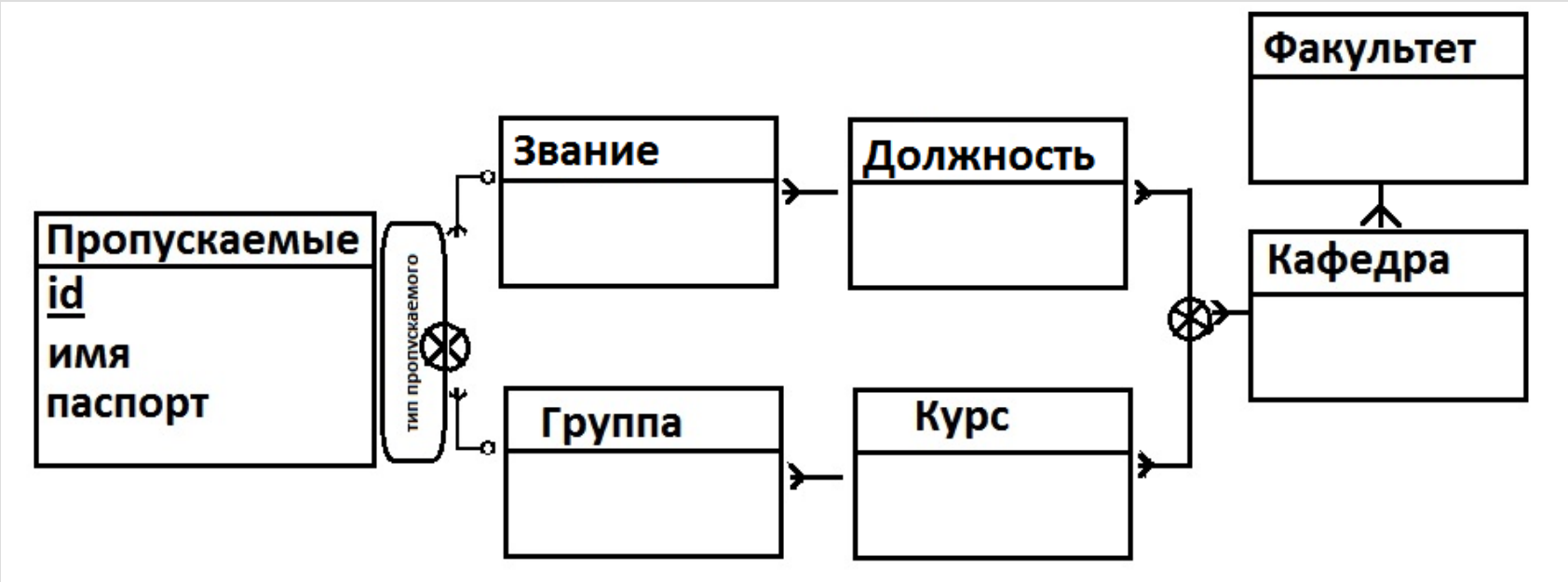
  - Drill-down/Up – преобразует агрегированные значения в более детализированные
- Операция Pivot
- Slice and dice

Операции для построения сводных отчетов

- Иерархии и работа с ними
- Аналитические функции
- Пользовательские агрегирующие функции
- Материализованные представления

Типы иерархий - 1

- Простые иерархии (деревья) 1- ко многим
  - Сбалансированные иерархии (расстояние до листьев – константа)
  - Несбалансированные иерархии – рекурсивные иерархии [будут разобраны далее]
  - Обобщённые иерархии
  - Неполные иерархии (пропущены штаты)



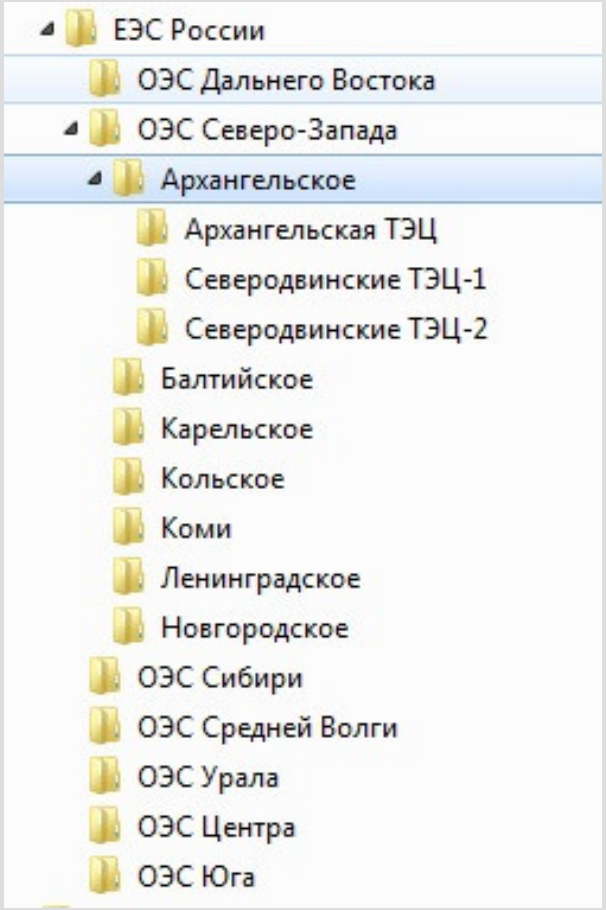
Типы иерархий – 2

- Нестрогие иерархии (многие ко многим)
- Альтернативные иерархии (время, неделя)
- Параллельные иерархии
  - Независимые (разные)
  - Зависимые (уровень для разной аналики общий, переиспользование | офис и зона действия)

Рекурсивные запросы

id	title	pid
1	*ЕЭС России*	null
2	*ОЭС Северо-Запада*	1
3	*ОЭС Дальнего Востока*	1

Как узнать уровень вложенности элемента?



Рекурсивные запросы - решение

Как узнать уровень вложенности элемента?

```
WITH RECURSIVE
 Rec (id, pid, title)
 AS (
   SELECT id, pid, title FROM test_table
   UNION ALL
   SELECT Rec.id, test_table.pid, Rec.title
    FROM Rec, test_table
    WHERE Rec.pid = test_table.id
   )
 SELECT  title, id, count(*) as level FROM Rec
 GROUP BY title, id;
```

id	title	pid
1	*ЕЭС России*	null
2	*ОЭС Северо-Запада*	1
3	*ОЭС Дальнего Востока*	1

Проблема построения анализа с использованием языка SQL

- Иногда для анализа данных не хватает стандартного SQL
- Пользовательские функции и хранимые процедуры – не декларативны и синтаксис не естественный для языка SQL
- Хранимые процедуры плохо поддаются ускорению

Аналитические функции. Достоинства

Для решения проблемы Oracle предложила аналитические функции

- Лаконичную и простую формулировку.
- Снижают нагрузки на сеть.
- Перенос вычислений на сервер.
- Лучшую эффективность обработки запросов.

Аналитические функции. Пример

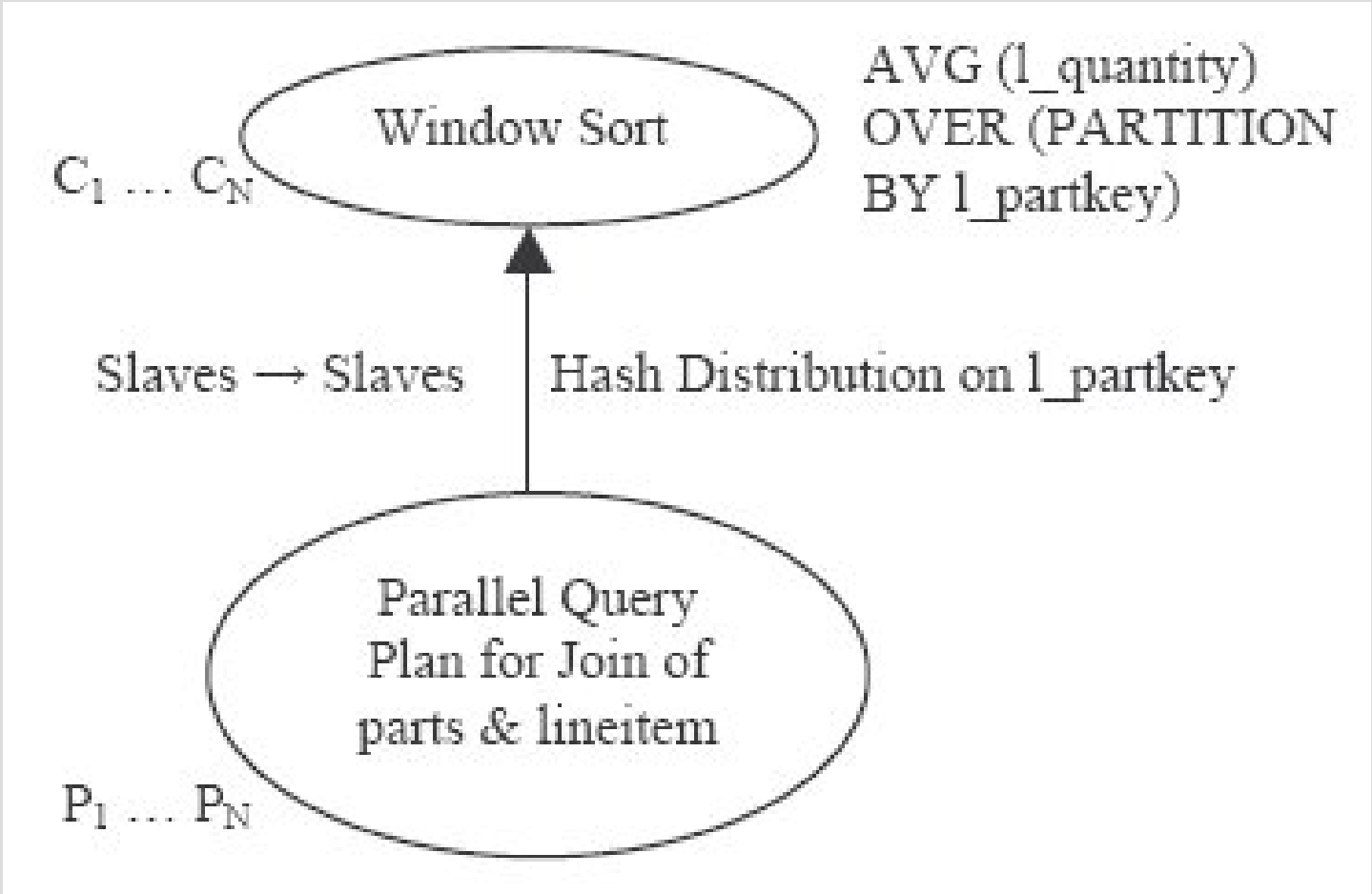
```
SELECT ename, deptno, job,
        SUM(sal) OVER (PARTITION BY deptno, job)
sum_sal
FROM emp;
```

- Групповая операция происходит над группой в PARTITION BY
- Один проход (если смотреть план)

Аналитические функции. Разновидности

- Функции ранжирования
- Оконные функции, позволяющие вычислять различные агрегаты (групповые операции), функции подсчета долей, статистические функции LAG/LEAD с запаздывающим/опережающим аргументом
- Итоговые функции по группе элементов
- Статистические функции (линейная регрессия и т. д.)

Маштабируемость аналитических функций



ДОМАШНЕЕ ЗАДАНИЕ. T1

A
1
7
5
100
23
1000

A	B
1	5
5	7
7	23
23	100
100	1000

1. Построить такое решение
2. Оценить его сложность выполнения
3. Построить оптимальное решение (и доказать, что оно оптимально)

ДОМАШНЕЕ ЗАДАНИЕ. Т2

Сделать на SQL MVIEW mWorkDay без использования вставки данных для которого можно было бы выполнить следующую команду и получить все рабочие дни в указанном диапазоне:

```
select *
from mv_WorkDay
where
    date between '01.01.2000' and '01.01.2100';
```

- Органичения по задаче – нужны даты от 2000 года до 2100

ДОМАШНЕЕ ЗАДАНИЕ. Т3

- Пусть есть таблица с данными о погоде (числовая): дата, время, значения температуры, облачность, давление (P0), направление ветра (СВ, Ю,), скорость ветра с сервиса gismeteo
- Файл можно скачать из архива погоды на rp5 или запросив у меня/Димы
- Известно, что формат несколько менялся – победит тот, кто сможет написать такой скрипт, который будет устойчив к изменениям (изменения скрыты)