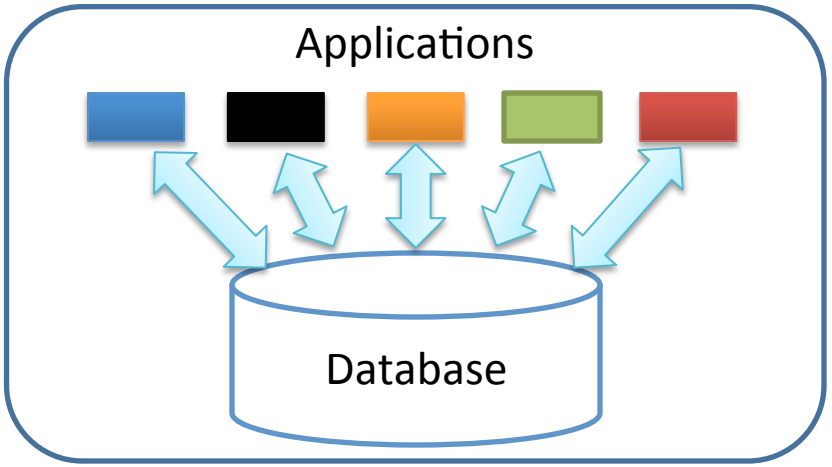
Управление данными: новые задачи и современные подходы

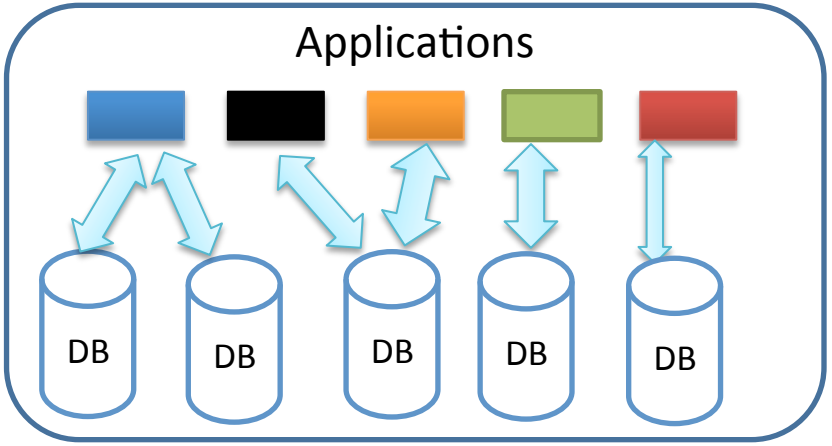
Б.А. Новиков
Санкт-Петербургский государственный университет
b.Novikov@spbu.ru

Управление данными

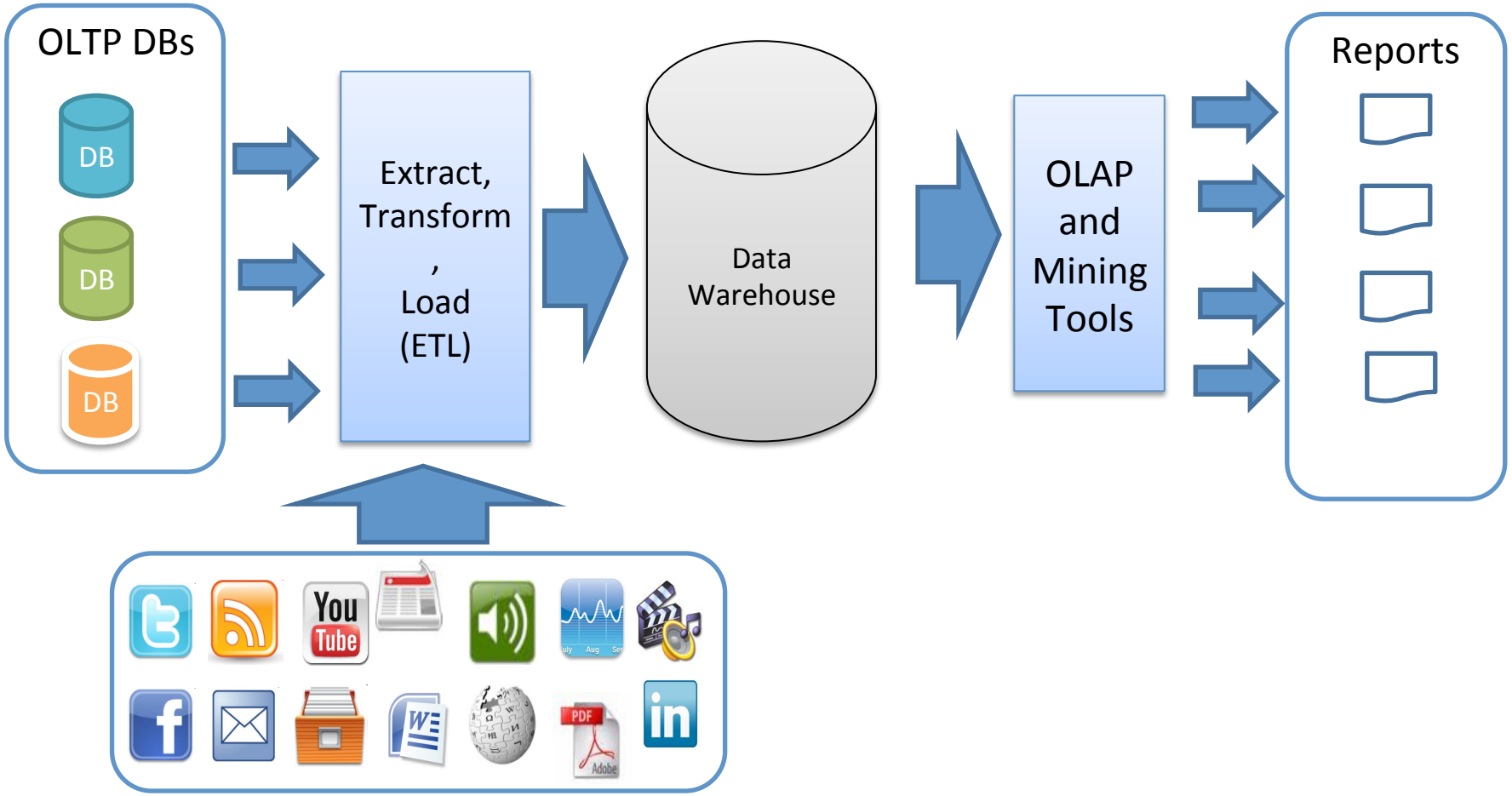
- Идеальная картина мира из учебников по СУБД: общие данные для всех приложений.
- **На практике этого никогда не было**



- В лучшем случае БД используется небольшим числом приложений
- Приложение может работать с несколькими БД



Аналитическая обработка



Большие данные

- Новое V-свойство каждые несколько месяцев:
  - Volume
  - Variety
  - Velocity
  - Veracity
  - Value
  - ?

- *Big Data: It's Not Just running the Analytics on the cloud*

- H. V. Jagadish

Большие данные: Объем

- Масштабирование не помогает:
  - Jim Gray (2004)
    - Возможности хранения данных превышают возможности их обработки, и со временем этотразрыв увеличивается
  - Anastasia Ailamaki (2014)
    - Количество вновь создаваемых данных превышает возможности их хранения, и разрыв со временем увеличивается

Большие данные: Многообразие

- Структурированность
  - Структурированные
  - Слабоструктурированыне
  - неструктурированные
- Динамика
  - Мало изменяющиеся
  - Интенсивно меняющиеся
  - Потоки данных
- Надежность и достоверность

Большие данные: скорость

- Jim Gray (2004)
  - Возможности хранения данных превышают возможности их обработки, и со временем этотразрыв увеличивается

- Ценность реузльтатов анализа падает со временем
- Ограничения реального времени
- Необходимость приближенной обработки

Большие данные: достоверность

- Качество данных:
  - Полнота
  - Согласованность
  - Целостность
- Как измерить качестводанных?
- Как оценить качество реузльтатов анализа?

- Становились ли ящеры плоскими?

Большие данные: ценность

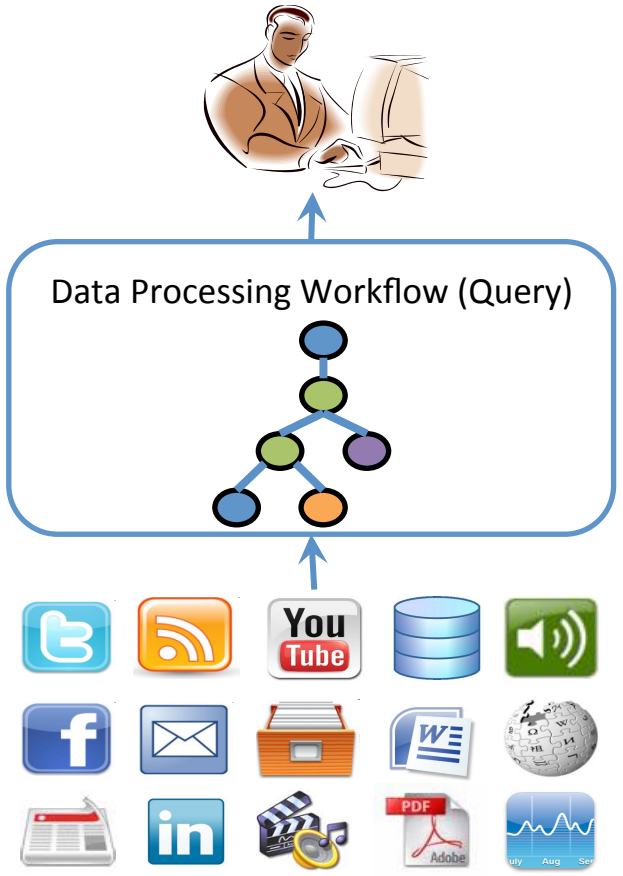
- Трудно добиться, но легко измерить

Декларативные спецификации

Высокоуровневое описание результата, а не процесса вычислений

Языки запросов обеспечивают:

- Высокая выразительность
- Компактность спецификаций
- Возможна эффективная реализация
- Возможности оптимизации



Модель данных

- Недостаточность реляционной модели данных
  - Динамика ?
  - Подобие ?
  - Идентификация ?
- Объектные расшинерия
  - Побочные ээфекты ?

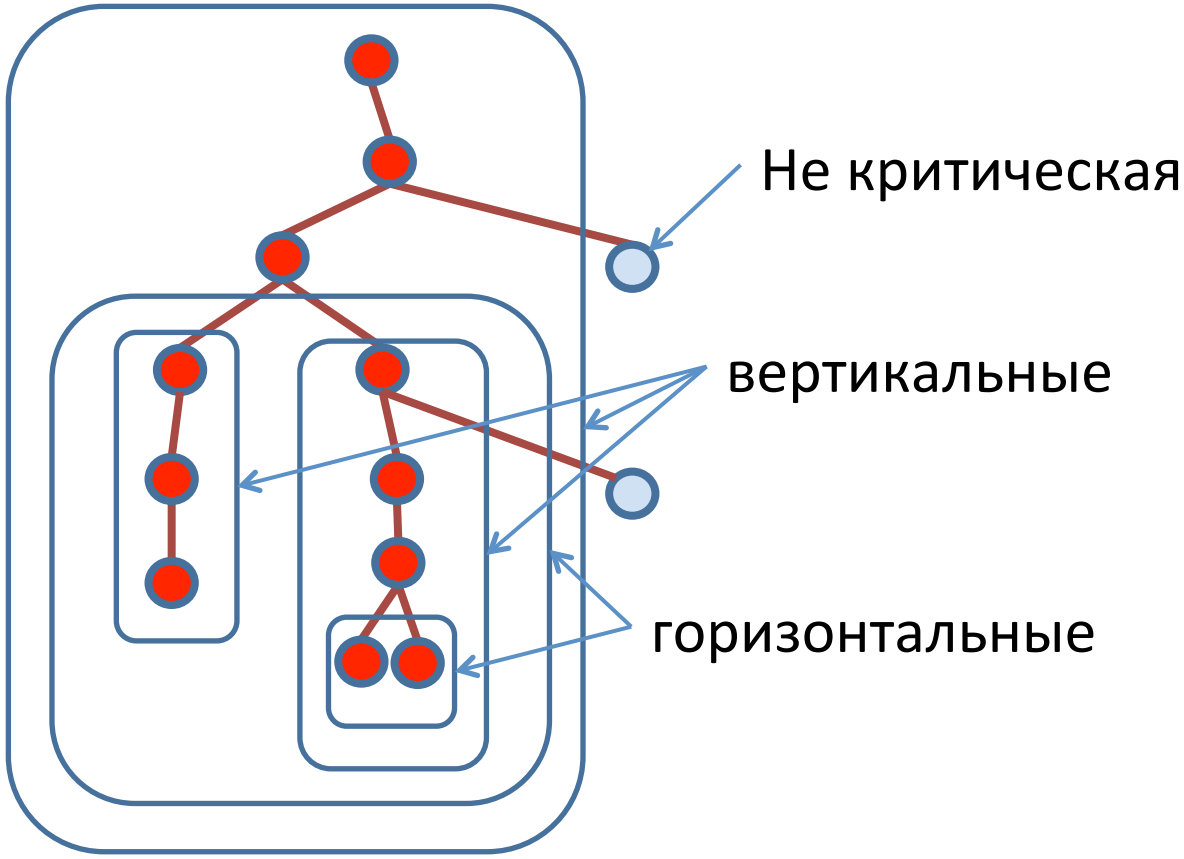
Приближенное выполнение

- Алгоритмы, допускающие приближенное вычисление
- Совместимо ли с языком запросов?
- Баланс между качеством и используемыми ресурсами

Выполнение сценариев обрабоки в реальном времени

- Реальное время: предсказуемость времени отклика
- Задача распределения ресурсов: получить наилучшее достижимое качество при ограничениях на используемые ресурсы

Распределение ресурсов



Оптимизация сценариев

- Необходимо превратить набор операций в алгебру
- Классы операций
  - Фильтры (возможно, map)
  - Агрегирования (возможно, reduce)
  - Соединения (взаимосвязи между объектами)
- Критерии оптимизации: стоимость, качество, надежность, …

Другие подходы

- noSQL = nonsenseQL
- Распределенные системы
- Строки или колонки?
- Алгебры на основе поиска по ключевым словами
- Data spaces

Проблемы оперативной образотки

- За последние 30 лет допустимое время ответа выросло с 0.25 сек до 2.5 сек., при увеличении частоты процессоров в 10000 раз.
- Почему?
- Очень большое количество очень мелких обращений к системе хранения (не имеет значения, SQL или нет)

Благодарности

- В работе участвоут сотрудники, аспиранты и студенты кафедры информационно-аналитических систем
- Спасибо!