Одиннадцатая независимая научно-практическая конференция «Разработка ПО 2015»

22 - 24 октября, Москва

2015 CEE-SEC(R) Разработка ПО

Обработка «умных данных»

Дмитрий Бугайченко

@mail.ru group

OK.ru – мифы:

OK.ru – реальность:

OK.ru – немного о размере

- 200 000 000 пользователей, 12 000 000 000 дружб, 10 000 000 сообществ…
- 8000 серверов по всему миру
- 1 терабит трафика в секунду
- 6 терабайт новых данных доступных для анализа в сутки
- …

100 гигабайт это:

- 50 000 электронных книг
- 10 000 музыкальных композиций
- 5 000 фоток с современного смартфона
- 10 HD фильмов
- 1 сезон любимого сериала в хорошем качестве

100 гигабайт, что выберете вы?

vs.

Как данные работают на нас

- 100 гигабайт
  - Каталог и рекомендации музыки: +300% к активности
  - Похожие видео: +100% просмотров в разделе похожих
- 400 гигабайт
  - Рекомендации сообществ: +50% кликов на витрине сообществ
- 10 терабайт
  - Сортировка «ленты»: +20% к «классам» из ленты

Как заставить данные работать на вас?

Собираем данные

Запись в СУБД

- ☑ Надежность
- ☑ Контроль схемы
- ☑ Простота анализ
- ☐ Скорость
- ☐ Гибкость
- ☐ Простота сбора

Агрегация логов

- ☑ Скорость
- ☑ Простота сбора
- ☑ Гибкость
- ☐ Простота анализа
- ☐ Потери при сборе
- ☐ Отсутствие схемы

Apache Kafka

- ☑ Скорость
- ☑ Простота сбора
- ☑ Гибкость
- ☑ Потоковый анализ
- ☑ Управляемые потери
- ☑ Управляемая схема

Сохраняем данные

Текстовые форматы

- Простой текст, csv
  - ☑Эволюция
  - ☐Скорость
  - ☐Объем
- JSON
  - ☑Эволюция
  - ☑Скорость
  - ☐Объем

Бинарные форматы

- SequenceFile
  - ☐Эволюция
  - ☑Скорость
  - ☑Объем
- Apache Avro
  - ☑Эволюция
  - ☑Скорость
  - ☑Объем

Apache Parquet

- **Эволюция** – колоночный формат, схема в заголовке
- **Скорость** – поднимает с диска только нужные колонки, базовые индексы
- **Объем** – сжатие перечислимых типов из коробки!

Исследуем данные

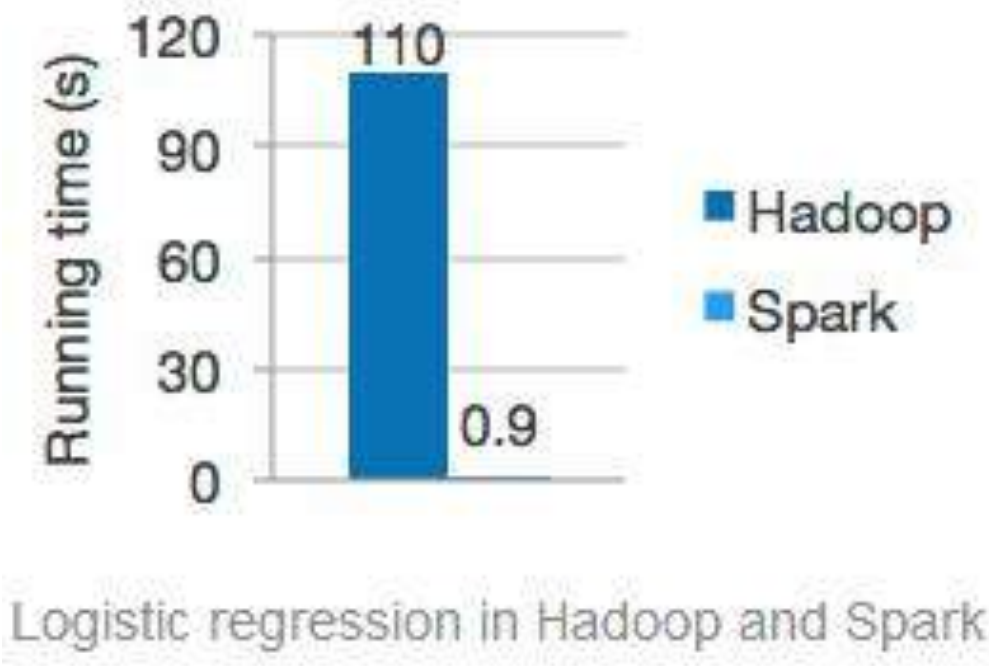
Зачем?

- Поиск закономерностей
- Оценка эффективности
- Базис для принятия стратегических решений
- Идеи для новых фич
- Идеи улучшения существующих фич

Как?

- Сэмплирование
- Статистический анализ
- Корреляционный анализ
- Визуализация
- Машинное обучение
- Блендинг
- …

Чем?


Running time (s)
120
90
60
30
0
110
0.9
Hadoop
Spark
Logistic regression in Hadoop and Spark

Шуруп забитый молотком держит лучше, чем гвоздь закрученный отверткой

Наполняем ящик исследователя

- Apache Pig, Hive
  - ETL
  - Простые агрегаты и статистика
  - Блендинг
  - Сэмплирование
- Apache Tez
  - Каскадные агрегаты
  - Многоуровневый блендинг
- Apache Spark
  - Ad-hook запросы
  - Простые ML модели
- Python, R
  - Статистика
  - Продвинутые ML модели
  - Базовая визуализация
- Tableau
  - Визуализация

Наполняем ящик исследователя

- Apache Pig, Hive
  - ETL
  - Простые агрегаты и статистика
  - Блендинг
  - Сэмплирование
- Apache Tez
  - Каскадные аггрегаты
  - Многоуровневый блендинг

>1Tb

- Apache Spark
  - Ad-hoc запросы
  - Простые ML модели
- Python, R
  - Статистика
  - Продвинутые ML модели
  - Базовая визуализация
- Tableau
  - Визуализация

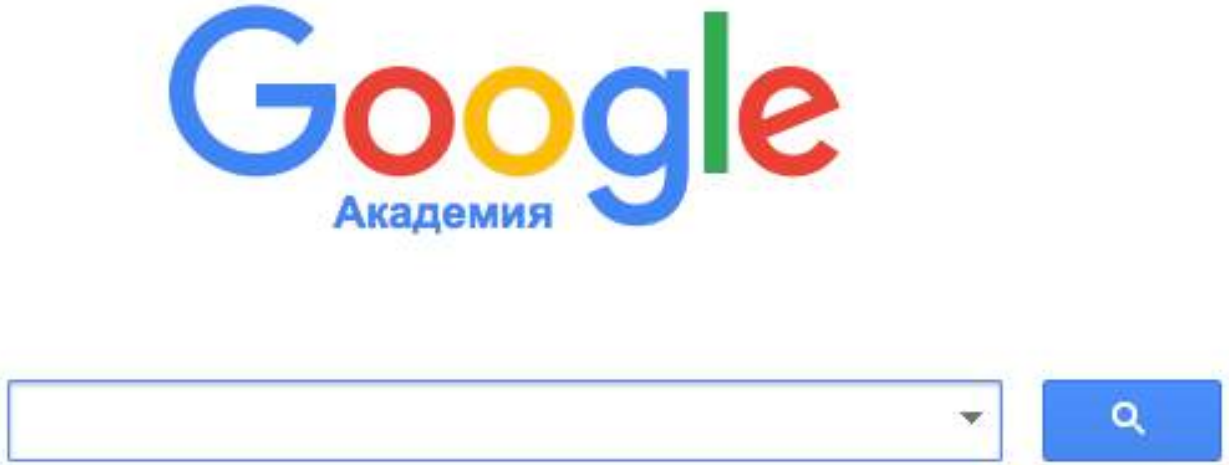
10Gb-1Tb

<2Gb

Строим модель

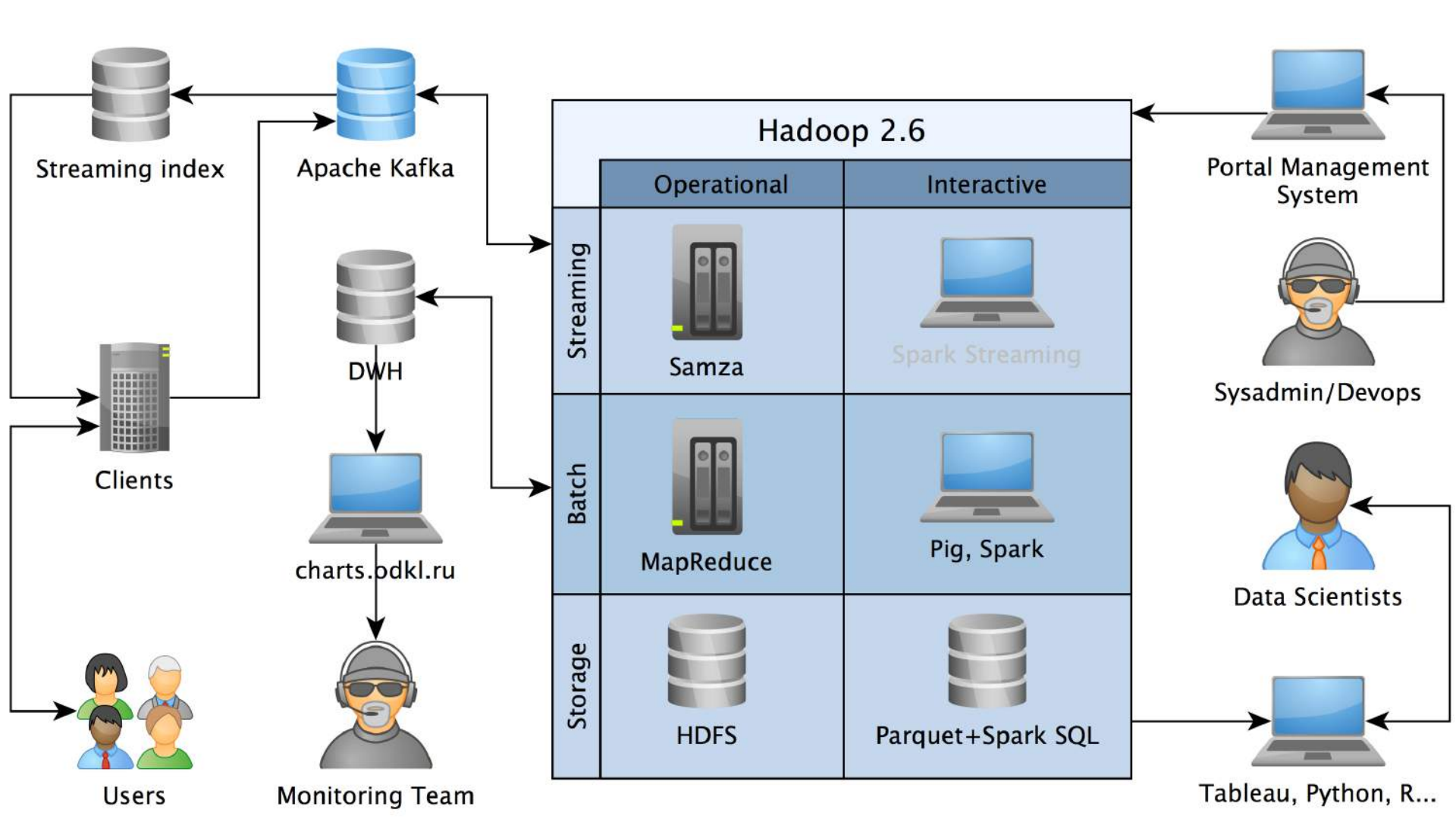
- Модель предсказывает реакцию пользователя на действия системы
- Предсказание модели используется для принятия решений
- **Самое важное**: четкие критерии оценки качества в оффлайне и онлайне

Главный инструмент построения модели



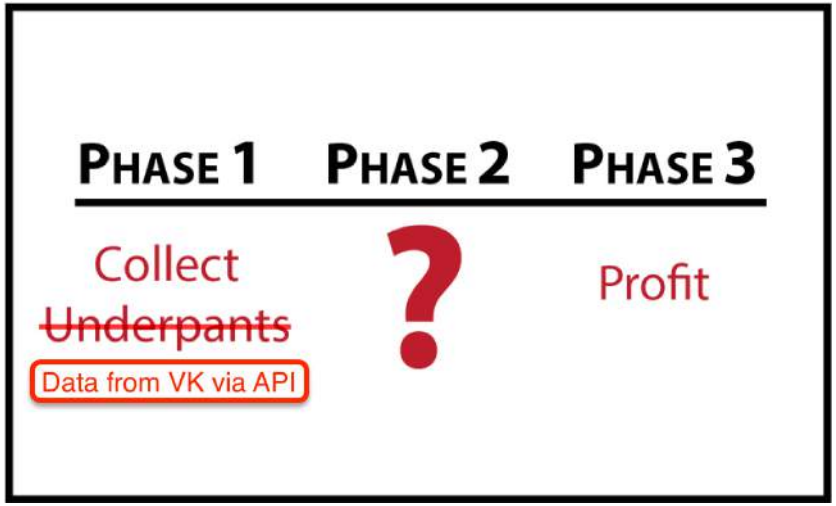
Внедрение модели

- Анализ эффективности
  - Собрать данные для расчета KPI
  - Инфраструктура для A/B теста
  - Выбор аудитории для A/B теста
- Интегрируем в прод
  - Максимально простой модуль расчета
  - Подготовка данных заранее
- Масштабируем и ускоряем
  - Переписывание стандартных алгоритмов
  - Переход на стриминг
- Автоматизируем
  - Интеграция с шедулером/ стримингом
- **Мониторим!**


Streaming index
Apache Kafka
Hadoop 2.6
Operational
Interactive
Streaming
Batch
Storage
Samza
Spark Streaming
MapReduce
Pig, Spark
HDFS
Parquet+Spark SQL
Portal Management System
Sysadmin/Devops
Data Scientists
Tableau, Python, R...
DWH
Clients
charts.odkl.ru
Users
Monitoring Team

Немного философии*

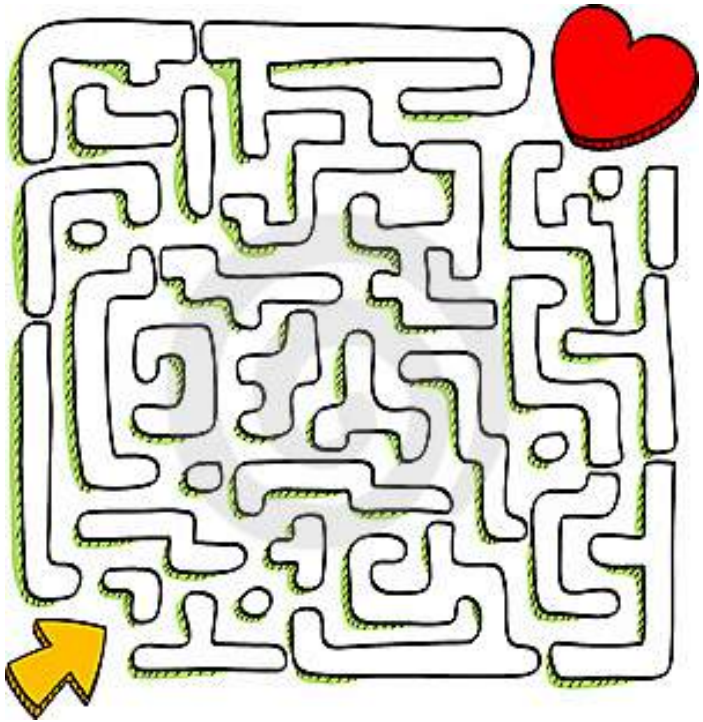
Data Science in lab



Data Science in industry

* Личное мнение автора

Data Science at OK.ru



Спасибо за внимание!