XII международная конференция

CEE-SECR / РАЗРАБОТКА ПО

28 - 29 октября, Москва

Семантическое ядро рунета – высоконагруженная content-based рекомендательная система реального времени на базе Amazon Kinesis/Lucene

Александр Сербул

ООО “1С-Битрикс”

О чем поговорим?

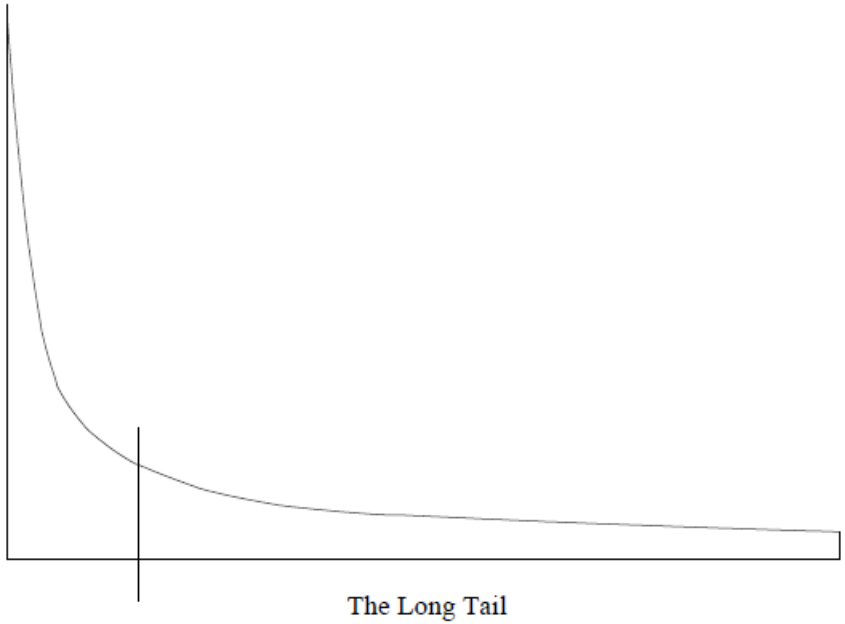
- Типы и алгоритмы рекомендательных систем – кратко
- Как мы собираем данные о действиях клиентов
- Архитектура нашей рекомендательной системы
- Статистика использования
- Советы как писать свои рекомендательные системы

Типы и алгоритмы рекомендательных систем

- Релевантный контент – «угадываем мысли»
- Релевантный поиск
- Предлагаем то, что клиенту нужно как раз сейчас
- Увеличение лояльности, конверсии

Объем продаж товаров

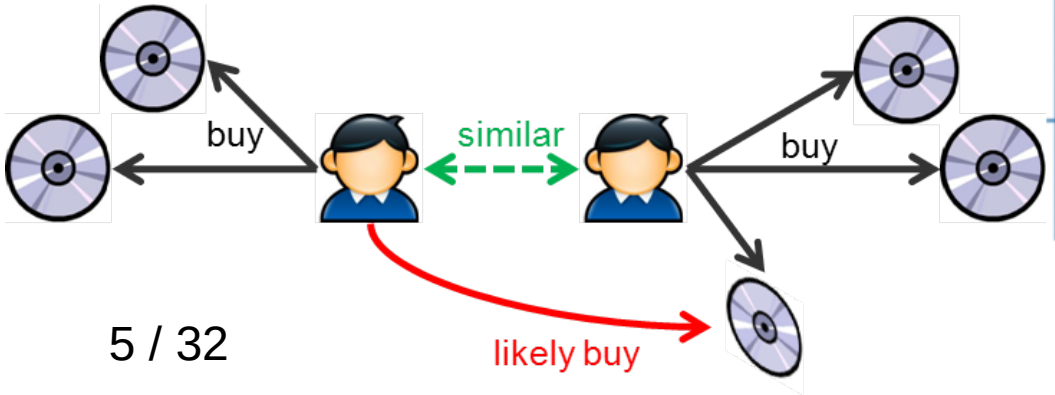
- Best-sellers
- Топ-продаж…
- С этим товаром покупают
- Персональные рекомендации



«Mining of Massive Datasets», 9.1.2: Leskovec, Rajaraman, Ullman (Stanford University)

Коллаборативная фильтрация

- Предложи Товары/Услуги, которые есть у твоих друзей (User-User)
- Предложи к твоим Товарам другие связанные с ними Товары (Item-Item): «сухарики к пиву»



	M_1	M_2	M_3	M_4	M_5	M_6
U_1	✓	✓	✓	✓		
U_2	✓		✓	✓	✓	
U_3		✓				✓

Возможности коллаборативной фильтрации

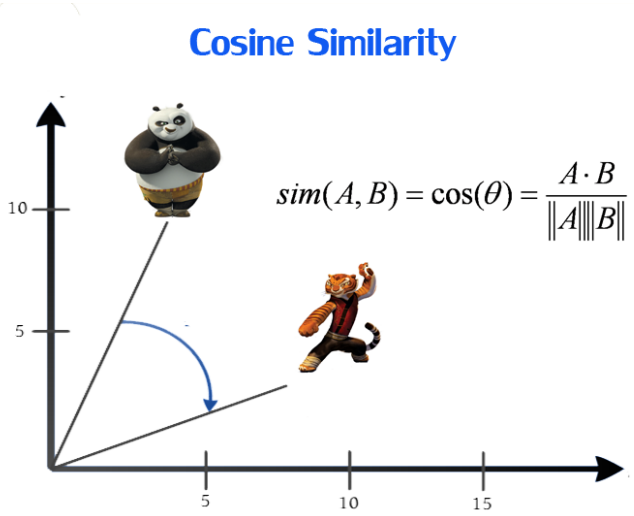
- Персональная рекомендация (рекомендуем посмотреть эти Товары)
- С этим Товаром покупают/смотрят/… (глобальная)
- Топ Товаров на сайте

Apache Spark MLlib (als), Apache Mahout (Taste) + неделька

Объем данных

Объем модели, требования к «железу»

Content-based рекомендации



- Купил пластиковые окна – теперь их предлагают на всех сайтах и смартфоне.
- Купил Toyota, ищу шины, предлагают шины к Toyota
- Поисковый «движок»: Sphinx, Lucene (Solr)
- «Обвязка» для данных
- Хранение профиля Клиента
- Реализация: неделька. Риски – объем данных, языки.

Тюнинг рекомендательных систем

- Рекомендовать постоянно «возобновляемые» Товары (молоко, носки, …)
- Рекомендовать фильм/телевизор – один раз до покупки
- Учет пола, возраста, размера, …

Как собирать данные от клиентов?

- Как собирать?
- Куда собирать?
- Как обрабатывать?
- Потоковые алгоритмы...

Технологии - RabbitMQ

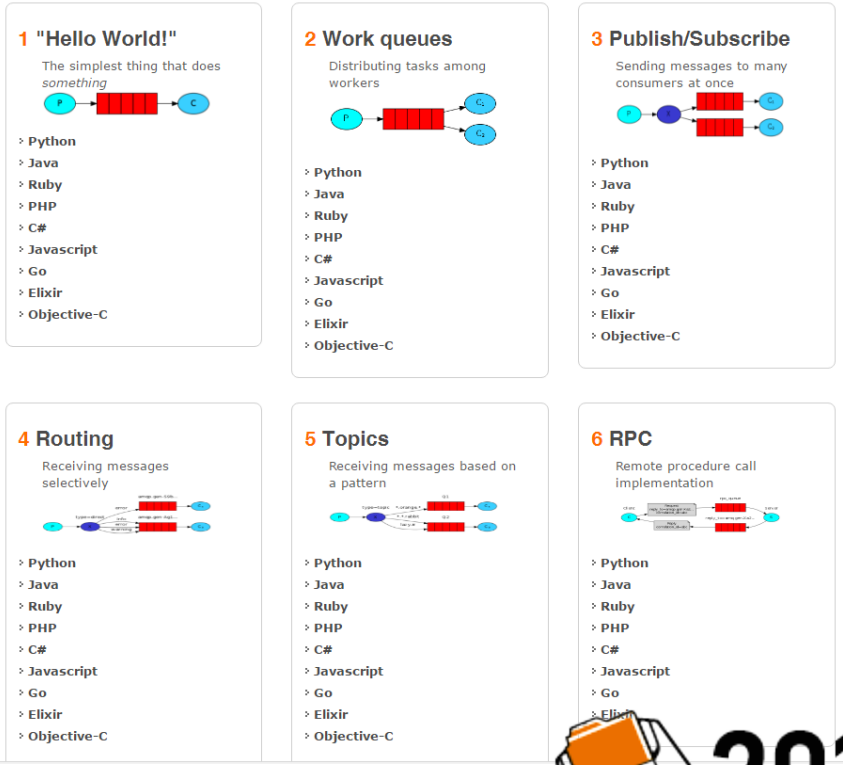
http://www.rabbitmq.com

1.Очереди сообщений

на все вкусы

2.AMQP

3.Erlang



Технологии - Apache Kafka

http://kafka.apache.org/

1.“LinkedIn”

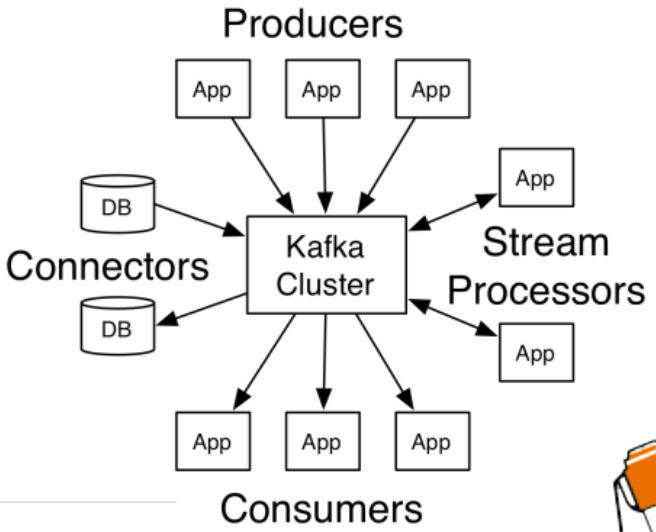
2.Не совсем очередь

3.Совсем не очередь!

4.Клиентское приложение

«держит» курсор потока

5.Scala



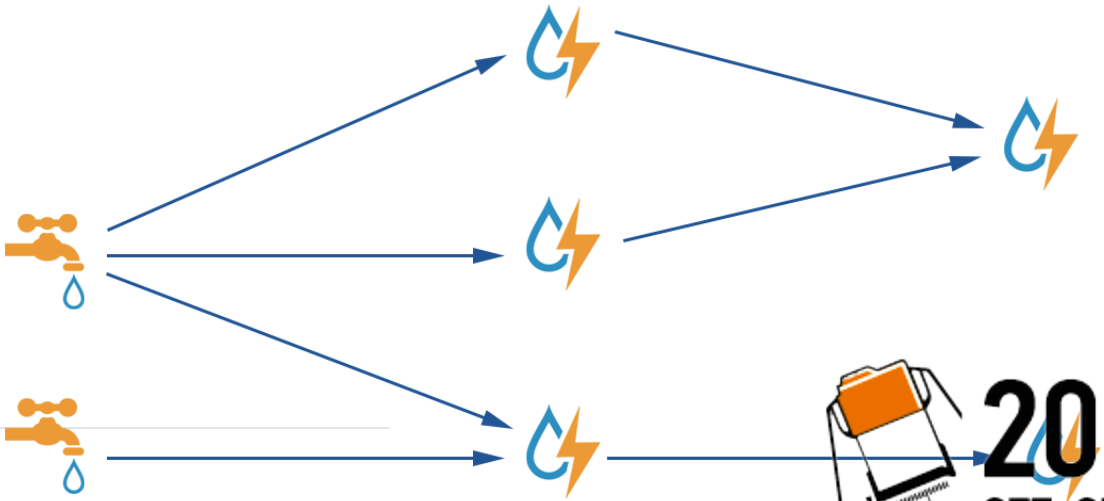
Технологии - Apache Storm

http://storm.apache.org

1.Task parallel

2.Удобные, гибкие workflow

3.Clojure/JVM



Технологии - Pinba

http://pinba.org

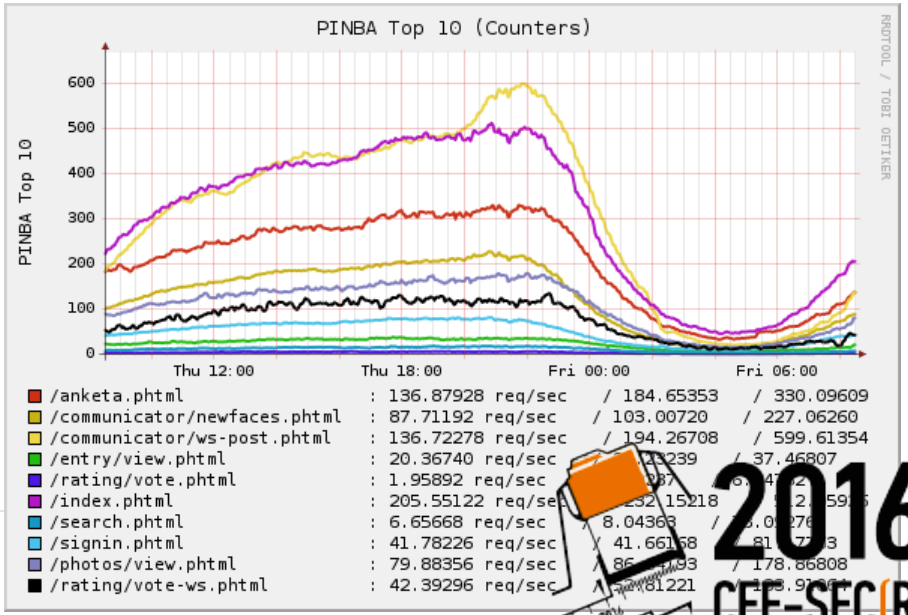
Агрегация внутри собств.

движка в MySQL

1.Интеграция с PHP

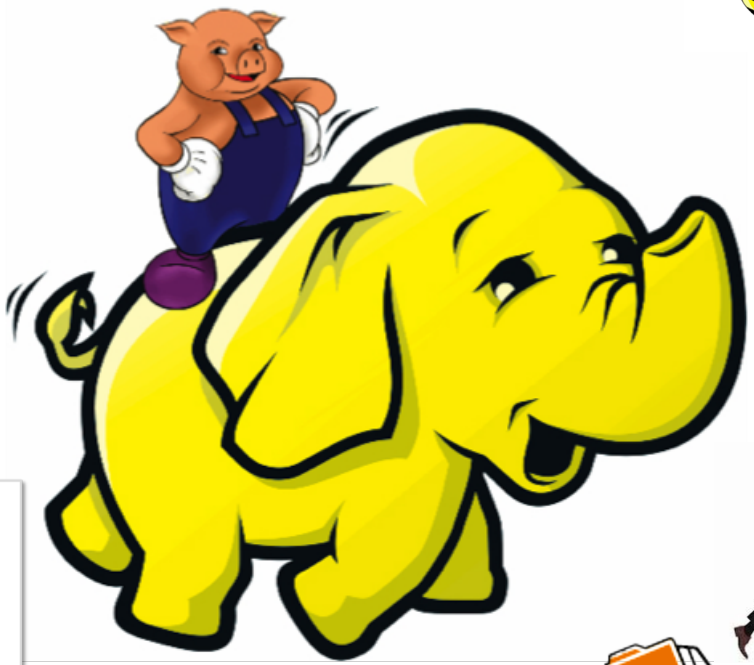
2.Быстро, удобно

3.Badoo.com



Apache Hadoop

- Платформа:
- - вычисления (MapReduce)
- - файловая система (HDFS)
- - “SQL-запросы” по данным

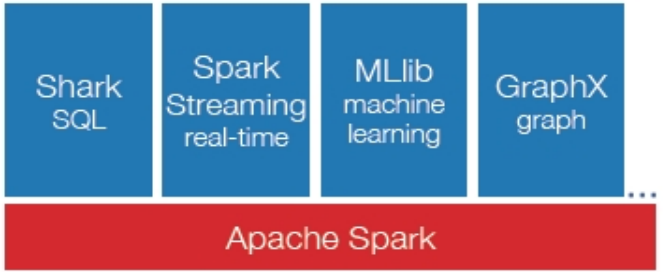


Apache Spark

- Скорость!
- Работа в памяти
- Кэширование в памяти
- Простота развертывания

Spark
Lightning-Fast Cluster Computing

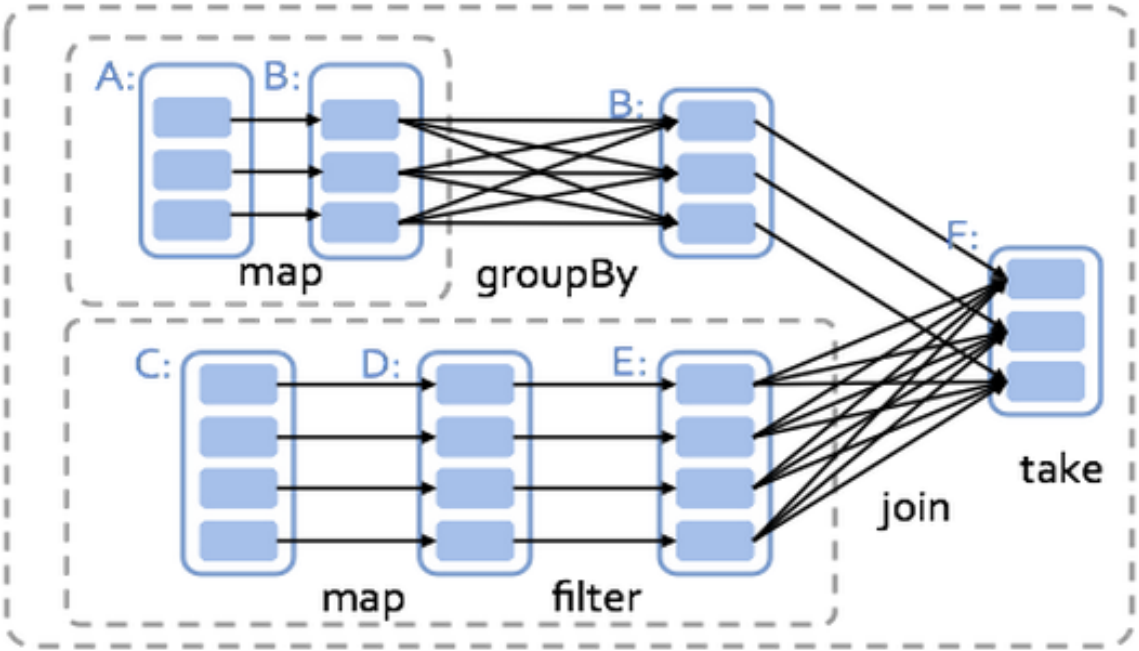
Sophisticated: can run today's most advanced algorithms



Парадигма MapReduce

«Mining of Massive Datasets»: Leskovec, Rajaraman, Ullman (Stanford University)

Apache Spark



«Online» алгоритмы, они другие!

- Кластеризация
- Уникальные элементы
- Агрегация
- Ограничения по памяти
- Это – уже не SQL ;-)

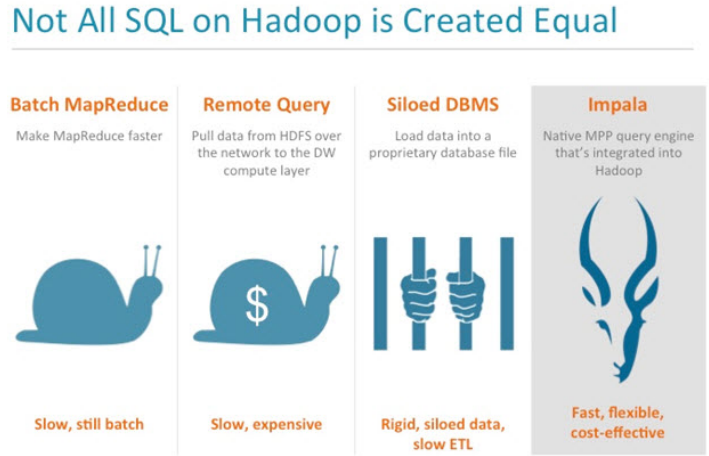
«Mining of Massive Datasets»: Leskovec, Rajaraman, Ullman (Stanford University)

Война систем хранения

- SQL на MapReduce: Hive, Pig, Spark SQL
- SQL на MPP (massive parallel processing):

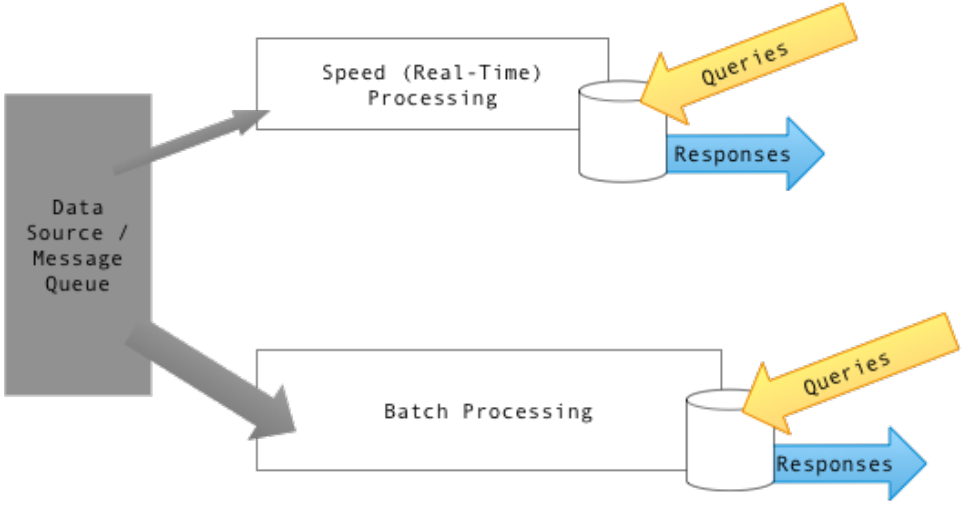
Impala, Presto, Amazon RedShift, Vertica

- NoSQL: Cassandra, Hbase, Amazon DynamoDB
- Классика: MySQL, MS SQL, Oracle, …



Lambda-архитектура

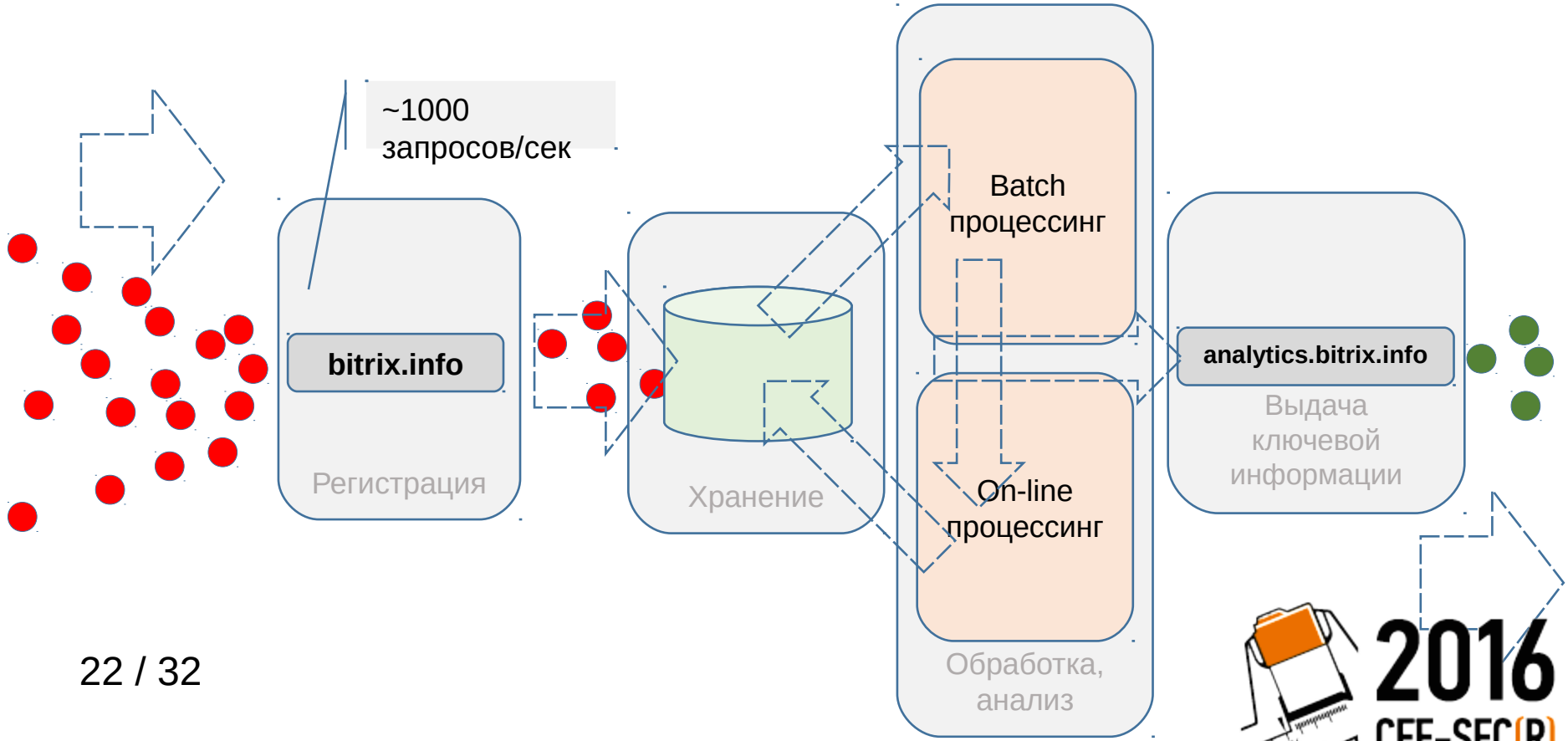
- Людей заваливает данными…
- С большой скоростью ☺



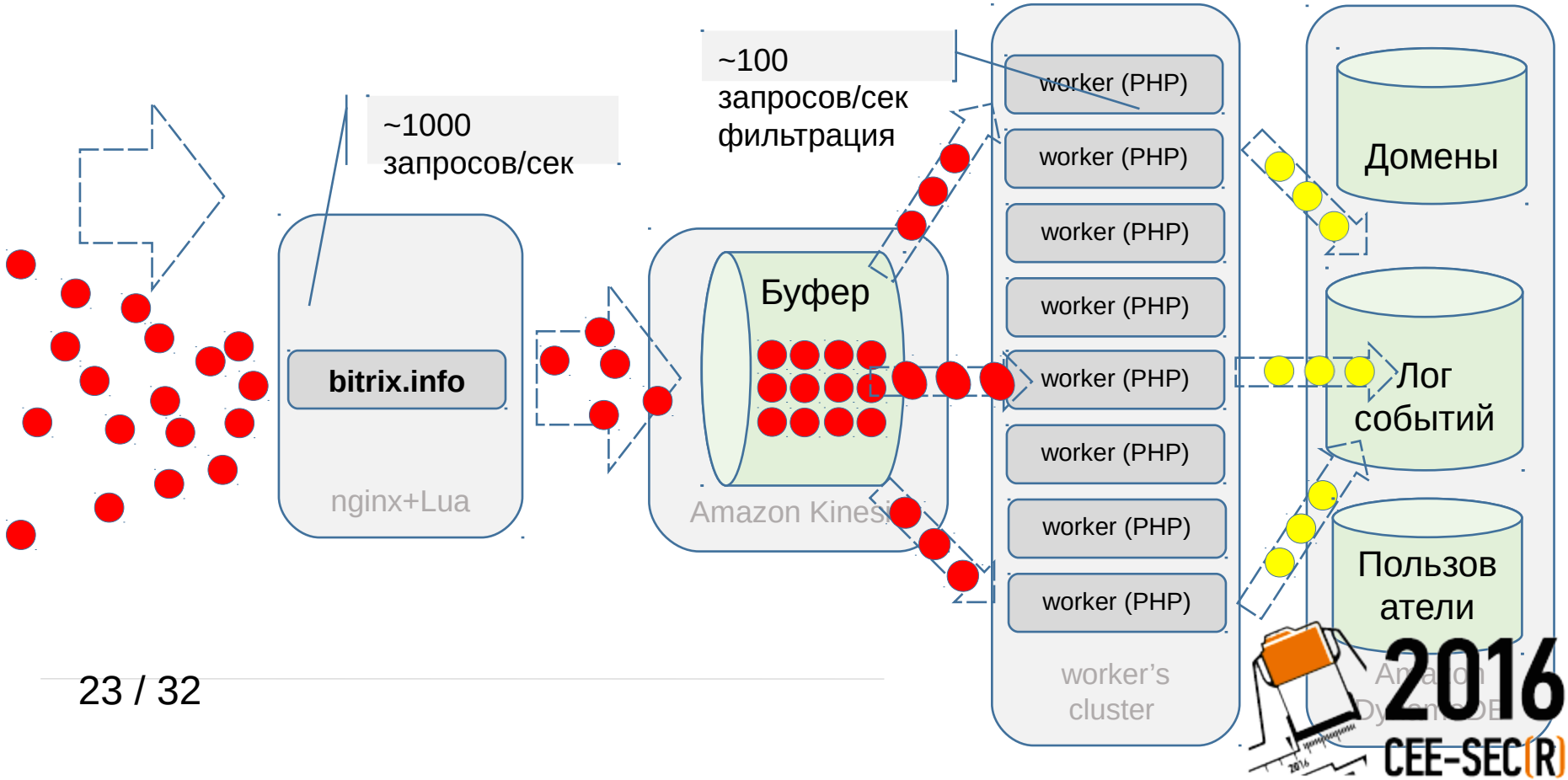
BigData – «под капотом». Виды событий.

- Кука Пользователя
- Хэш лицензии
- Домен
- ID товара
- Название Товара
- Категории Товара
- ID рекомендации
- ряд других

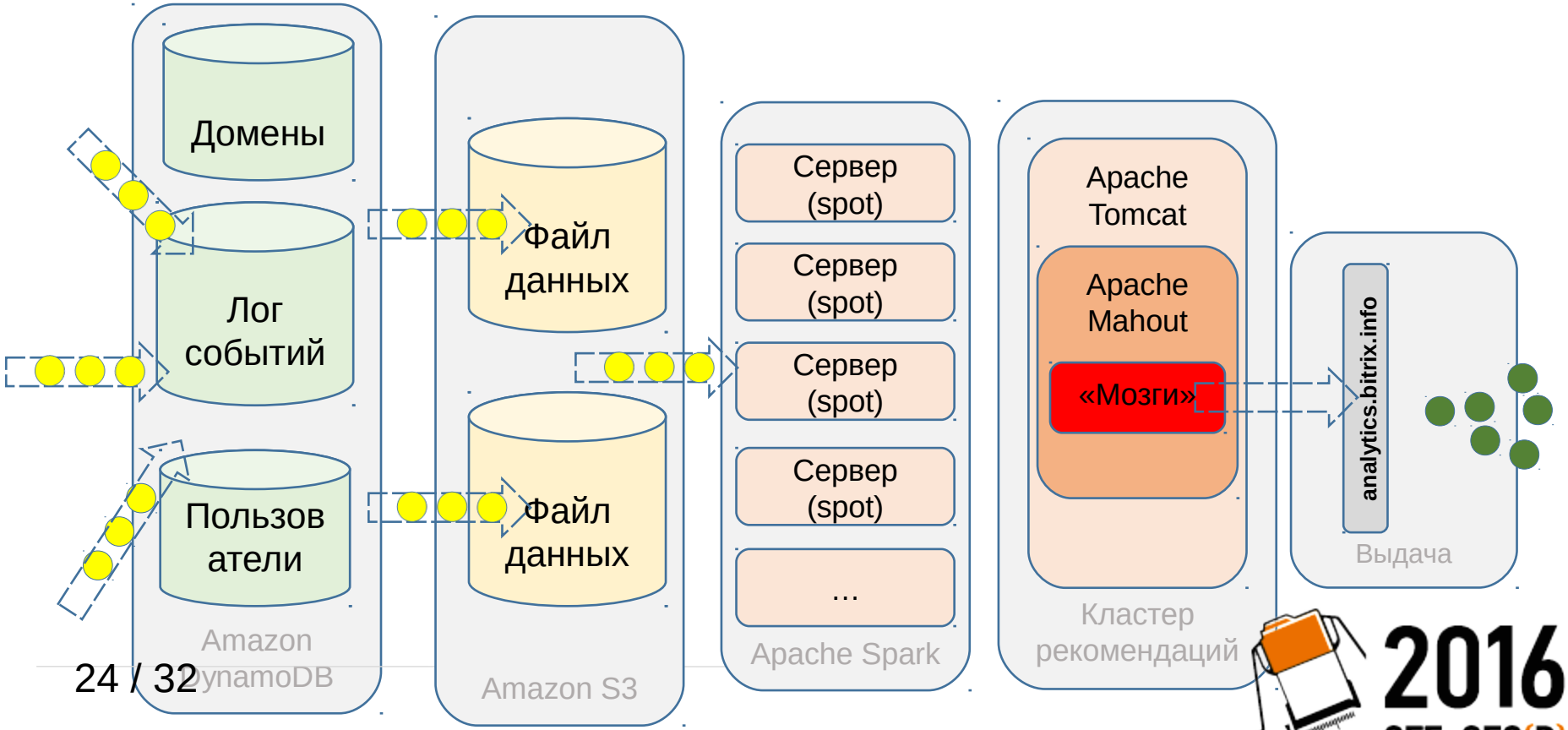
BigData – «под капотом». С птичьего полета.



BigData – «под капотом». Регистрация событий.



BigData – «под капотом». Обработка, анализ, выдача.



Apache Lucene

- Doug Cutting: Nutch, Hadoop (Yahoo!)… сейчас в Cloudera
- Lucene: Solr, ElasticSearch
- Lucene: грамотная многопоточность из коробки

elastic

2016 CEE-SEC(R)

Apache Lucene: +/-

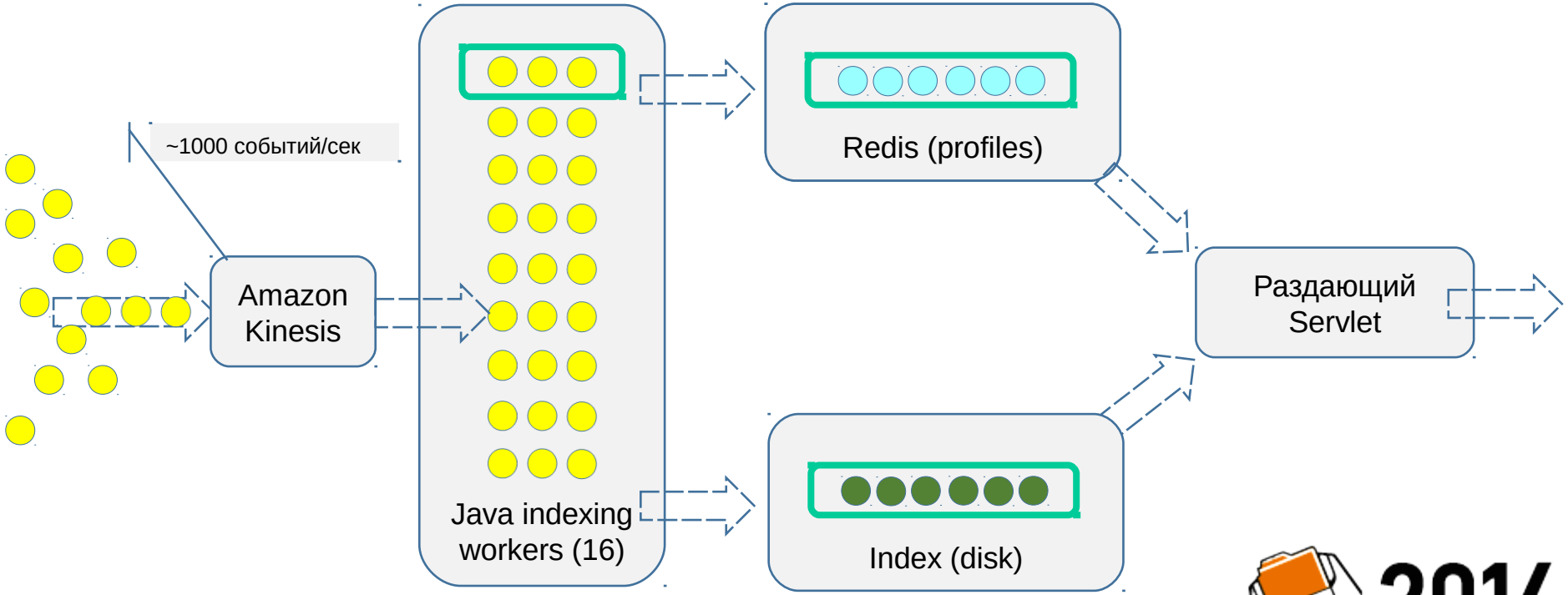
- (-) Нет нормальной поддержки русского языка
- (-) Нет русской морфологии
- (-) Документация иногда оставляет желать лучшего
- (-) Нет решения для 100% онлайн индексации

- (+) Компактный индекс (гигабайты)
- (+) Лаконичное API
- (+) Транзакционность
- (+) Thread-safety

Redis

- Профиль Пользователя: десятки тэгов
- Стемминг Портера
- Высокочастотные слова
- Алгоритмы вытеснения тэгов
- Куда можно развивать… (word2vec, glove, синонимы ...)

Архитектура content-based рекомендаций



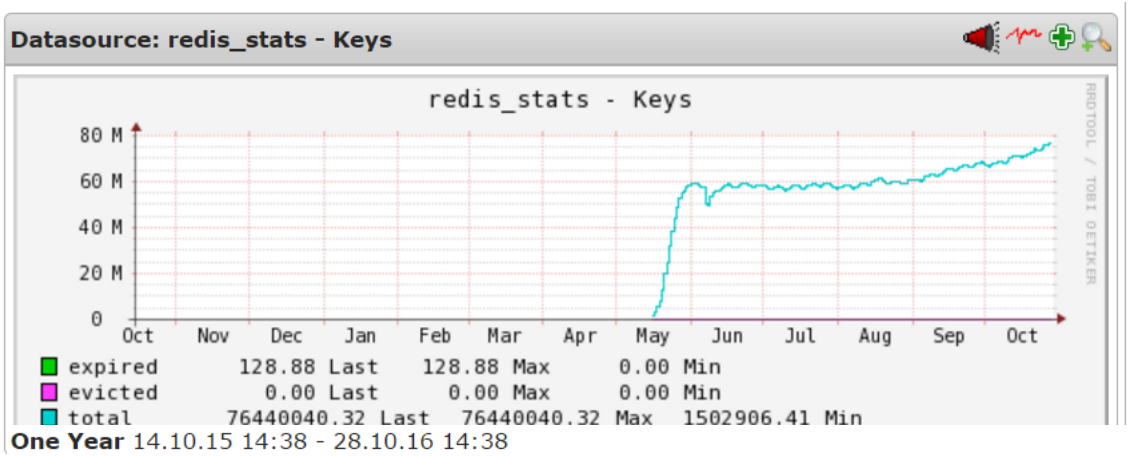
Процессы

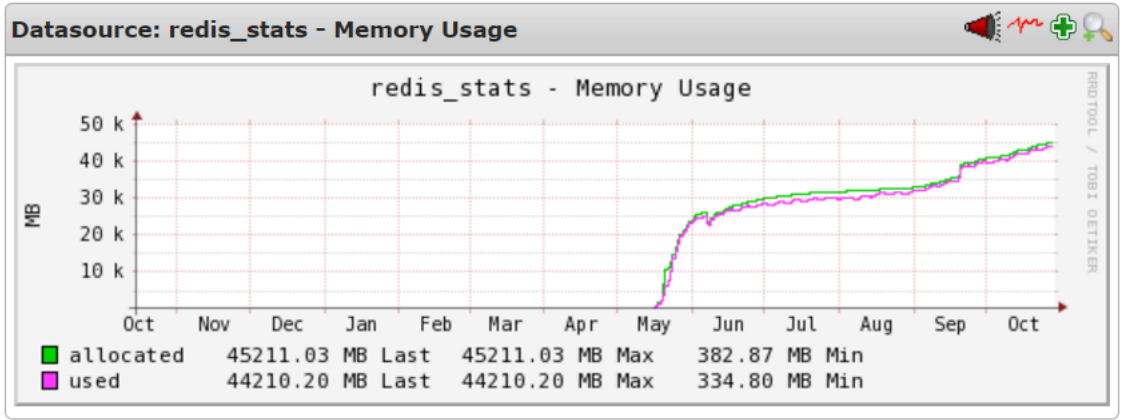
- Многопоточный индексатор, java/lucene
- Amazon Kinesis – как буффер
- Индекс в папке на диске, вытеснение
- Как реализован “онлайн”
- Раздающий Servlet

Цифры и данные

- “Потребители”: десятки тысяч интернет-магазинов
- “Поставщики”: все сайты на Битрикс, больше 100к
- Тэги Профиля: название страницы, h1
- Индекс Товаров: название, краткое описание, разделы
- Индекс: гигабайты, сотни файлов в папке

Цифры и данные





2016 CEE-SEC(R)

Спасибо за внимание!
Вопросы?

Александр Сербул
@AlexSerbul
serbul@1c-bitrix.ru