Одиннадцатая независимая научно-практическая конференция «Разработка ПО 2015»

22 - 24 октября, Москва

2015 CEE-SEC(R) РазработкаПО

Облачный сервис персональных рекомендаций для 20 000 интернет-магазинов: секреты, алгоритмы, технологии

Александр Сербул

ООО «1С-Битрикс»

1С·БИТРИКС

О чем поговорим…

- Рекомендательные сервисы – суть
- Снаружи: популярные алгоритмы и техники реализации
- А когда много данных…
- Изнутри: как устроен наш облачный сервис «1С-Битрикс BigData»
- Куда двигаться дальше

1С·БИТРИКС

Персональные рекомендации – зачем?

- Предсказать мысли, желания клиента
- Если клиент готов – соблазнить его, привязать к себе
- Не спамить клиента мусором, не раздражать
- Соблазнять клиента – регулярно (рассылки, push)

1) Релевантность, 2) Разумность, 3) Вовремя, 4) Не пережать

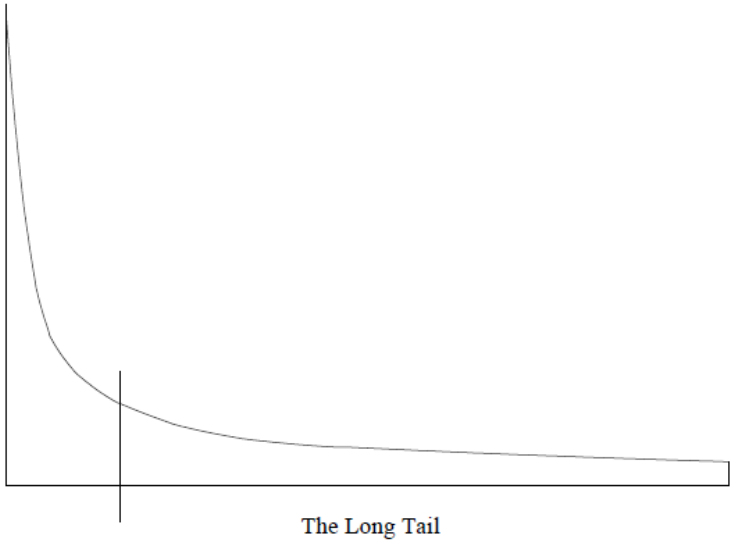
Нас уже прослушивают:

Windows 10, Android, …

1С·БИТРИКС

Как соблазнять?

- Не персональные «крючки»:
- Топ продаж (best sellers)
- С этим Товаром покупают (аксессуары)
- С этим Товаром смотрят
- Другие смотрят сейчас
- Скидка на очень популярный товар



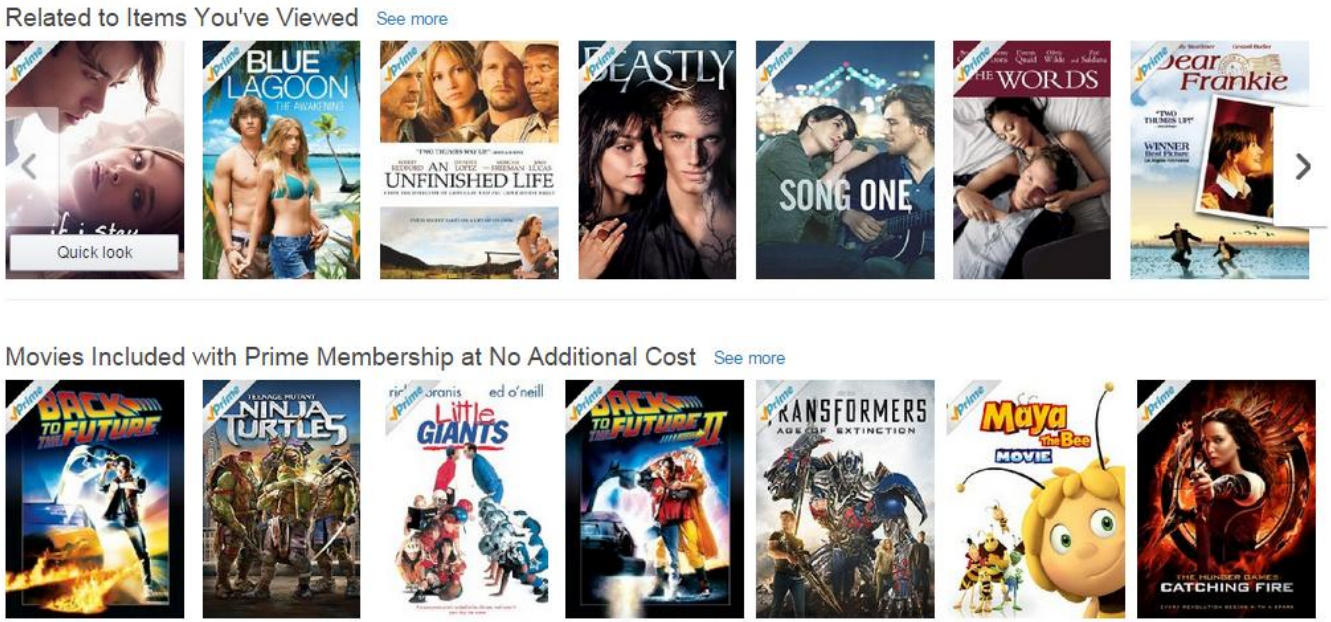
«Mining of Massive Datasets», 9.1.2: Leskovec, Rajaraman, Ullman (Stanford University)

Небольшой набор товаров. Хвост. Спам – для некоторых.

1С·БИТРИКС

Amazon.com

- Персональные, не персональные



1С·БИТРИКС

Amazon.com

- Не персональные?!

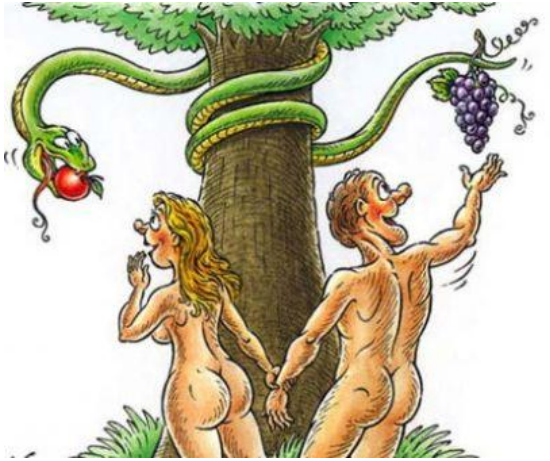
1С·БИТРИКС

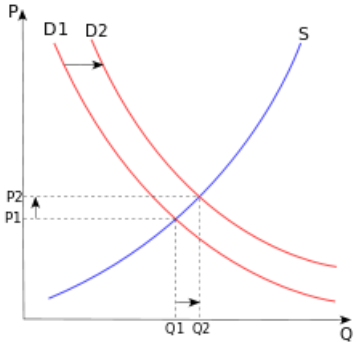
Как соблазнять?

- Персональные «крючки»:

Рекомендуем именно вам в данный момент:

- Купить, посмотреть
- Люди, похожие на вас («близкие по духу»)
- «Хорошая» скидка, «хорошая» цена
- Полезный контент
- Релевантный поиск

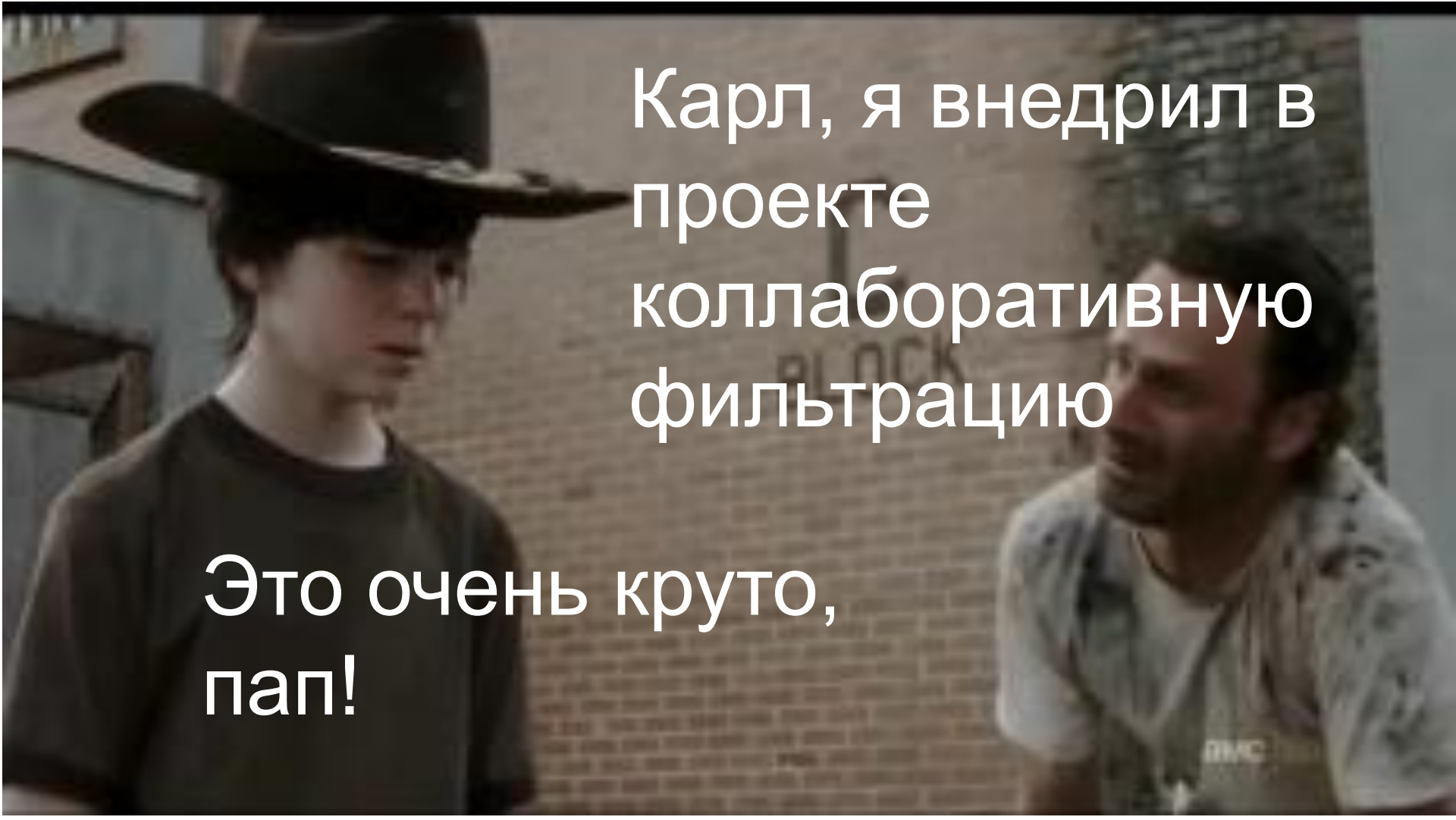




1С·БИТРИКС

С целью персональных рекомендаций – понятно. Теперь сухая конкретика и код.

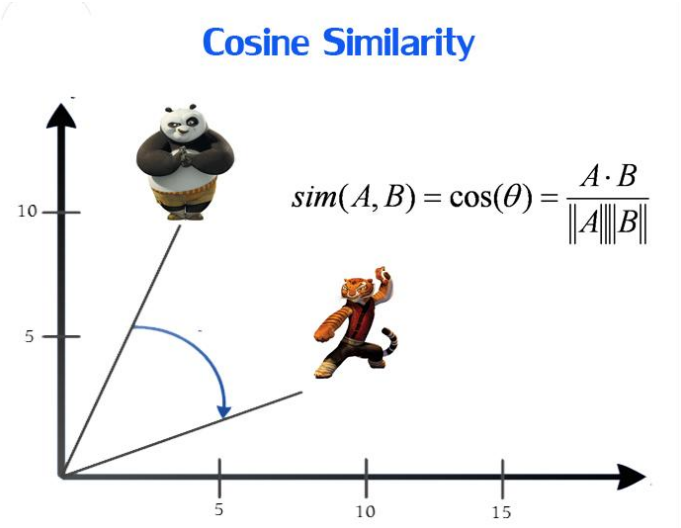
1С·БИТРИКС


Карл, я внедрил в проекте коллаборативную фильтрацию
Это очень круто, пап!


Но я так и не понял,
как и почему она
работает.
СОВСЕМ!!!

Content-based рекомендации

- Купил пластиковые окна – теперь их предлагают на всех сайтах и смартфоне, в Windows 10 и во сне.
- Купил Toyota, ищу шины, предлагают шины к Toyota вверху списка
- Vector space model, tf/idf
- word2vec



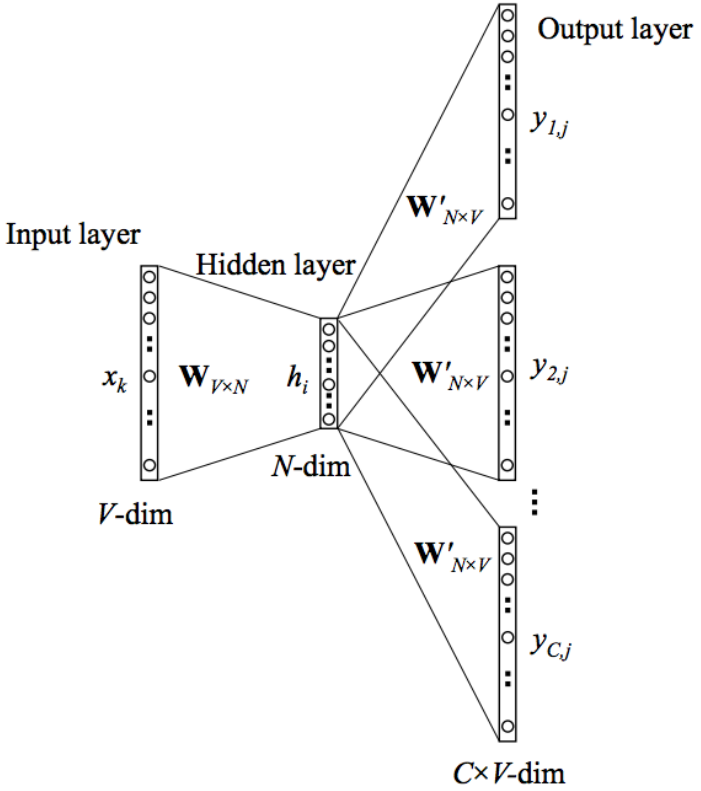
	eveyrthing	interesting	learning	lerning	like	Machien	machine	not	predicts	problems	solving	sure	What
1	0	1	0	0	1	0	0	0	0	1	1	0	0
2	0	0	1	0	0	0	1	0	0	0	0	0	1
3	0	0	0	0	0	0	0	1	0	0	0	1	0
4	1	0	0	1	0	1	0	0	1	0	0	0	0

1С·БИТРИКС

word2vec, SVD/PCA

- Сжимаем размерность
- «Склеиваем» синонимы
- Skip-gram
- Continuous bag of words (CBOW)

- «Похож» на матричную факторизацию

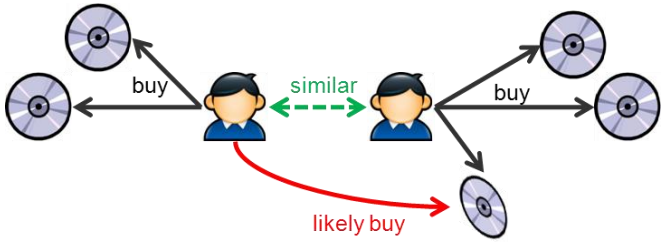


1С·БИТРИКС

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»

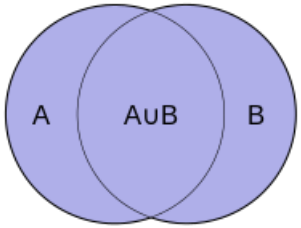
	M_1	M_2	M_3	M_4	M_5	M_6
U_1	✓	✓	✓	✓		
U_2	✓		✓	✓	✓	
U_3		✓				✓



1С·БИТРИКС

Коллаборативная фильтрация - алгоритмы

- User-User: поиск похожих «в лоб» (kNN), k-d tree, LSH
- Item-Item: Amazon, работает гораздо быстрее
- Item-Item «плюшки» - с этим Товаром покупают
- Mahout Taste (матрица в памяти)
- Spark MLLib (ALS)



$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}.$$

Товары в моем профиле

Их связи с другими Товарами

Взвешенное среднее для предсказания моих невыраженных интересов

1С·БИТРИКС

Коллаборативная фильтрация – сжатие Товаров

- «Единый» каталог
- Склеить дубликаты
- Передать «смысл» между Товарами
- Улучшить качество персональных рекомендаций
- Семантическое сжатие размерности, аналог матричной факторизации
- Скорость
- Ранжирование результатов

1С·БИТРИКС

Minhash

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

- Min-wise independent permutations locality sensitive hashing scheme
- Снижаем размерность
- Совместима с LSH (следующий слайд)

Pr[hmin(A) = hmin(B)] = J(A,B)

- Размер сигнатуры: 50-500

simhash

Row	S_1	S_2	S_3	S_4	$x+1 \mod 5$	$3x+1 \mod 5$
0	1	0	0	1	1	1
1	0	0	1	0	2	4
2	0	1	0	1	3	2
3	1	0	1	1	4	0
4	0	0	1	0	0	3

Text shingling

- Shingle – «черепица»
- Устойчивость к вариантам, опечаткам

«Штаны красные махровые в полоску»

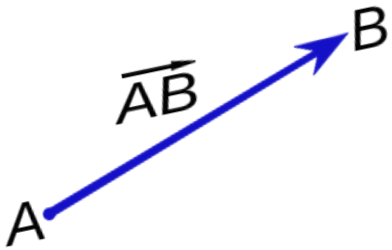
{«штан», «таны», «аны », «ны к», «ы кра», «крас», …}

«Красные полосатые штаны»

1С·БИТРИКС

Векторизация описания Товара

- Текст: «Штаны красные махровые в полоску»
- Вектор «bag of words»: [0,0,0,1,0,…0,1,0] – ~ 10000 - 1000000 элементов (kernel hack)
- Minhash-сигнатура после shingling:
- [1243,823,-324,12312,…] – 100-500 элементов, совместима с LSH



1С·БИТРИКС

Locality-Sensitive Hashing (LSH)

- Вероятностный метод снижения размерности
- Использовали для minhashed-векторов
- Banding:

b – корзины, r – элементов в корзине.

P{ “Векторы совпадут хотя-бы в одной корзине” }:

An approximation to the threshold is $(1/b)^{1/r}$

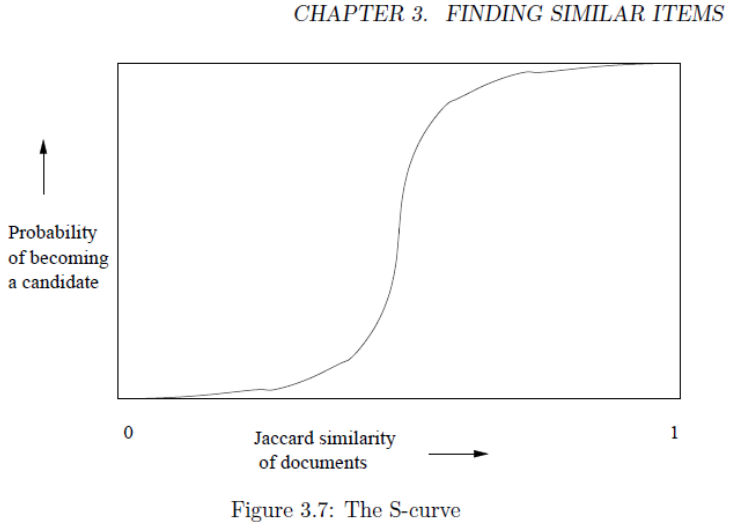
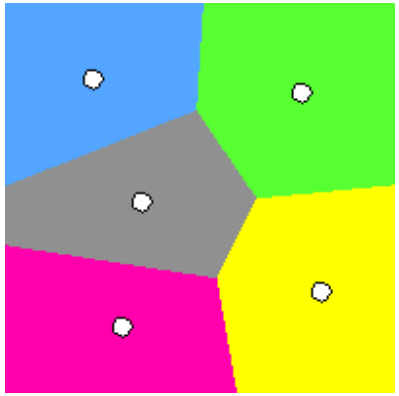


Figure 3.7: The S-curve

1С·БИТРИКС

Кластеризация каталога



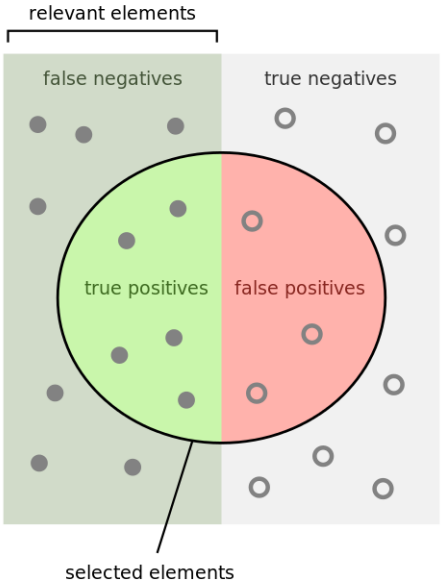
- Apache Spark
- 2-3 часа, 8 spot-серверов
- 10-20 млн. Товаров => 1 млн. кластеров
- Адекватные по смыслу кластера
- Персональные рекомендации - стали в разы «лучше»
- DynamoDB – хранение кластроидов

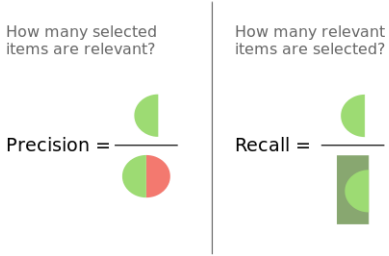
1С·БИТРИКС

application_1437483006918_0346	hadoop	MahoutModelsBuilderClustered	SPARK	default	Sun, 06 Sep 2015 12:00:12 GMT	Sun, 06 Sep 2015 14:39:44 GMT	FINISHED	SUCCEEDED	History

Измерение качества персональных рекомендаций

- Recall, precision
- Предсказываем на «старой» модели
- Смотрим фактические значения профиля – на текущей модели
- Считаем recall

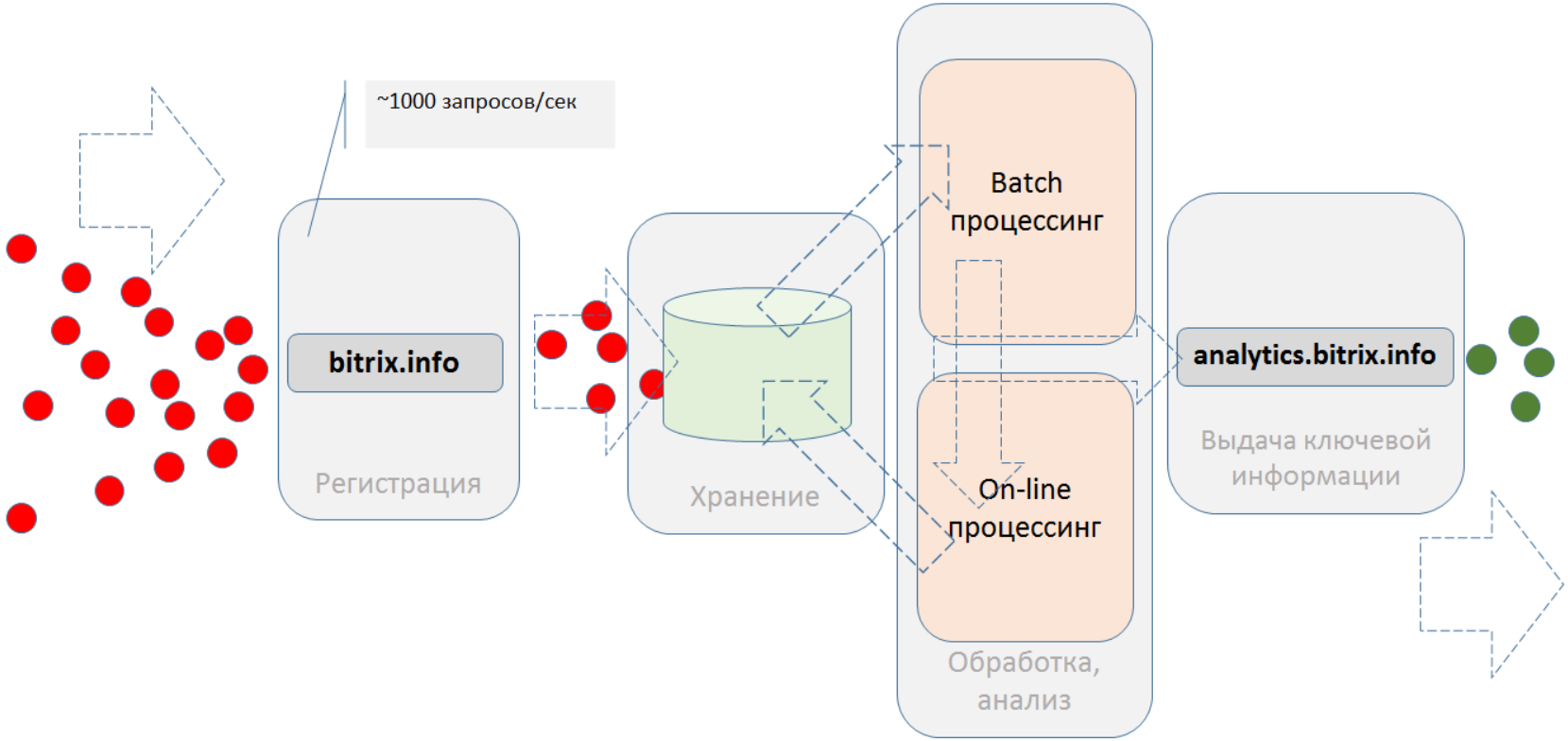




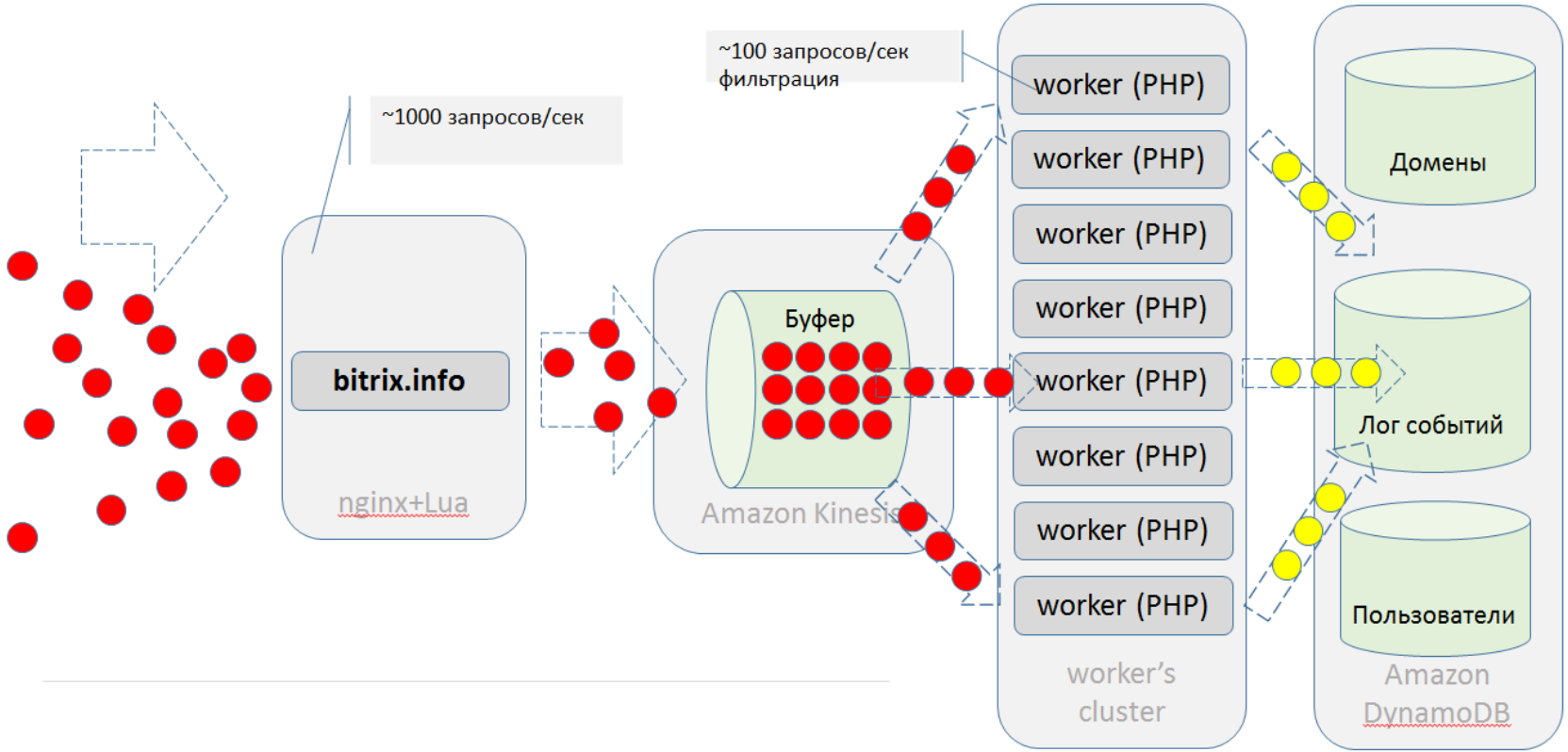
Архитектура нашего облачного сервиса

1С·БИТРИКС

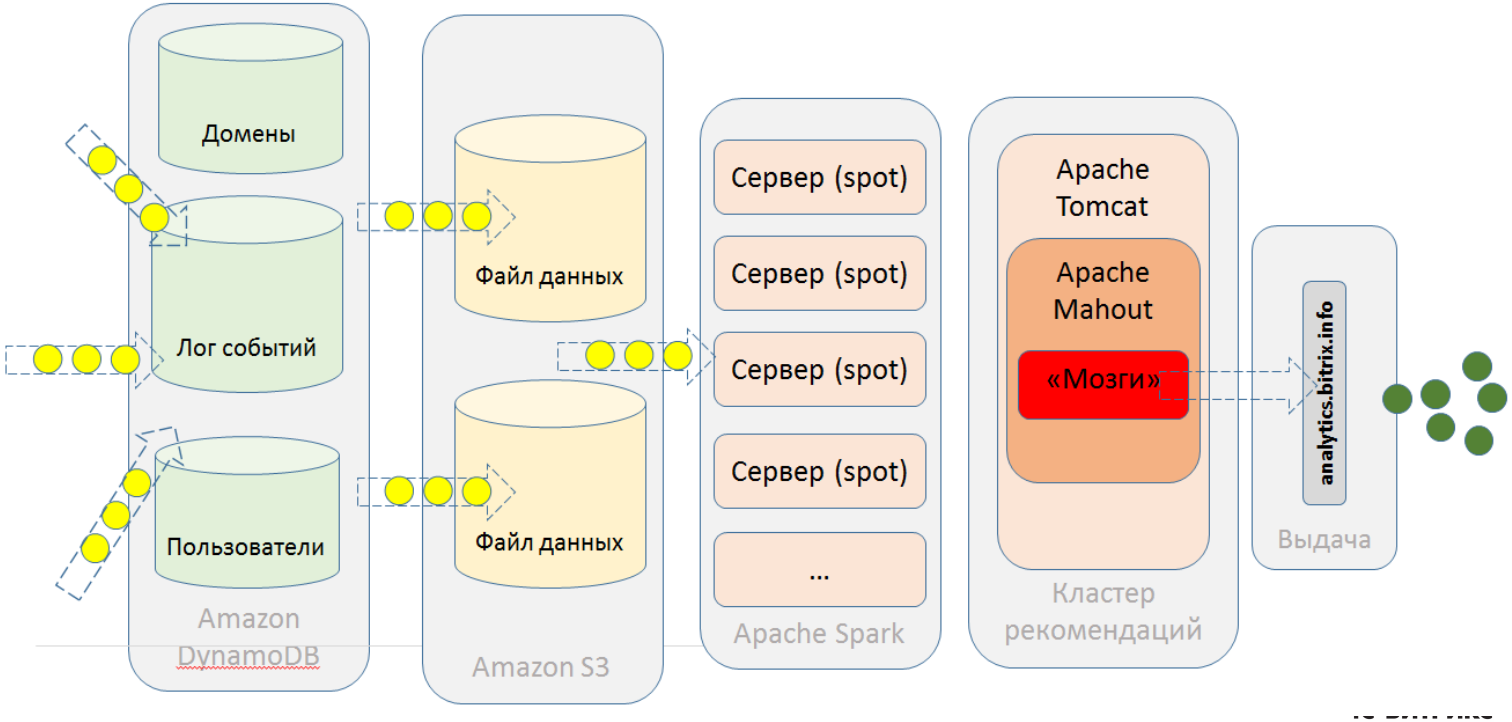
Сервис «1С-Битрикс: BigData» - общий вид



Сервис “1С-Битрикс: BigData”



Сервис “1С-Битрикс: BigData”



Цифры кратко

- Тысячи запросов в секунду к сервису
- ~20 тысяч интернет-магазинов
- Ощутимый рост конверсии – до 50-80%, зависит от размера магазина
- Активное использование «С этим Товаром покупают»!?
- 1 сервер рекомендаций (70G ОЗУ) + небольшой кластер Spark
- Обсчитываем событий: > 855 миллионов
- Уникальных посетителей: > 332 миллиона

1С·БИТРИКС

API. Персональная рекомендация

- *https://analytics.bitrix.info/crecoms/v1_0/recoms.php?op=recommend&uid=#кука#&count=3&aid=#хэш_лицензии#*
- op=recommend
- uid – кука Пользователя
- aid – хэш от Лицензии
- count – число рекомендаций

{
"id":"24aace52dc0284950bcff7b7f1b7a7f0de66aca9",
"items":["1651384","1652041","1651556"]
}

API. Похожие Товары на данный

- *https://analytics.bitrix.info/crecoms/v1_0/recoms.php?op=simitems&aid=#хэш_лицензии#&eid=#id_товара#&count=3&type=combined&uid=#кука#*
- op=simitems
- uid – кука Пользователя
- aid – хэш от Лицензии
- eid – ID Товара
- type - view|order|combined
- count – размер выдачи

API. Топ Товаров на сайте

- *https://analytics.bitrix.info/crecoms/v1_0/recoms.php?op=sim_domain_items&aid=#хэш_лицензии#&domain=#домен#&count=50&type=combined&uid=#кука#*
- op=sim_domain_items
- uid – кука Пользователя
- aid – хэш от Лицензии
- domain – домен сайта
- type - view|order|combined
- count – размер выдачи

Куда развиваться

- Пол, возраст, ценовая категория клиента – машинное обучение
- Разные виды товаров: возобновляемые, невозобновляемые
- Цена товара
- Внутренние циклы (готов покупать), модели Маркова
- Классификация групп лояльности, кластерный анализ
- Релевантный поиск

1С·БИТРИКС

Спасибо за внимание!
Вопросы?

Александр Сербул

serbul@1c-bitrix.ru

AlexSerbul

1С·БИТРИКС