2. Overfitting [Электронный ресурс] – Режим доступа: https://en.wikipedia.org/wiki/Overfitting
3. Linear regression [Электронный ресурс] – Режим доступа: https://en.wikipedia.org/wiki/Linear_regression
4. Support-vector machine [Электронный ресурс] – Режим доступа: https://en.wikipedia.org/wiki/Support-vector_machine

АНАЛИЗ ЭФФЕКТИВНОСТИ РЕКЛАМЫ КАК ETL ПРОЦЕСС

Харитонов Н.В., Хоронеко М.П., Медунецкий М.А.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Стержанов М.В. – к.т.н.,доцент

В данной работе рассматривается механизм оценки эффективности интернет рекламы при помощи стека технологий Big Data. Речь пойдет о ETL процессе (реализованном на планировщике задач Luigi) с помощью которого мы собираем, храним и обрабатываем большие объемы данных. В качестве результата будет представлена архитектурная схема платформы, осуществляющей данный процесс, и пример возвращаемых данных - еженедельный отчет, включающий в себя статистические данные рекламной кампании. Эволюция вычислительных систем обусловлена развитием задач бизнеса, а именно задач сбора данных, хранения и обработки полученных результатов. Конкурирующими показателями для данных систем становятся такие характеристики как скорость работы, отказоустойчивость, безопасность и защищенность от внешних воздействий. Глобальная информатизация привела к тому, что централизованные информационные системы становятся достаточно уязвимыми. Не менее уязвимым становится бизнес, работа которого зависит от качественной работы информационных систем.

В настоящее время наблюдается устойчивый рост интереса к практическому применению технологий Big Data в сфере маркетинга. Среди решаемых проблем можно выделить задачи повышения конкурентоспособности, создания новых услуг, совершенствования управления взаимоотношениями с клиентами. В результате развития Интернет, социальных сетей, и иных сетевых сервисов непрерывно растут потребности в информационных продуктах и услугах. Чтобы предлагать клиентам такие услуги, предприятиям приходится анализировать большие объемы данных из различных источников. Поэтому накопленная информация становится стратегически важным активом, от эффективности управления которым существенно зависят результаты деятельности предприятий.

В рамках данной работы мы представляем платформу для анализа эффективности интернет рекламы с использованием стека технологий Big Data.

Big Data [1] – обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами, появившимися в конце 2000-х годов и альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence.

Нами предлагается использование следующих параметров для оценки эффективности интернет-рекламы: количество просмотров; количество переходов; количество нажатий мыши по рекламной области; местоположения; время просмотра. Мы логируем все доступные акты показа рекламы, агрегируем их, проводим обработку, а затем создаем отчет за заданный период времени (неделя, месяц, год).

Отчет включает в себя: данные о рекламодателе; рассмотренный период; количество показов; количество переходов; отношение переходов к показам; количество уникальных местоположений; количество уникальных пользователей; количество уникальных устройств и т. д.

Для более детального описания, представим архитектуру платформы в виде диаграммы на рисунке 1:

Представление архитектуры и стека используемых технологий

Архитектура системы представляет собой ETL[2] процесс, организованный на кластере компьютеров с использованием технологий Hadoop и HDFS. Рассмотрим каждый из шагов данного процесса.

Extract. На этом шаге мы собираем данные из различных источников (массива социального Веба, логов действий пользователей, корпоративных баз данных, разнообразных датчиков, внешнего набора данных и т.д.), а затем передаем их на следующий этап для выполнения преобразований. В данном случае для извлечения данных нами используются RTB[3] аукционы и подход “1x1 pixel”. Каждый показ рекламы записывается в HDFS[4] базу данных (см. hdfs://logs на Рис. 1). Каждая запись

содержит в себе такие поля как “время показа”, “идентификатор рекламодателя”, “идентификатор устройства”, “географическая широта”, “географическая долгота” и т.д.

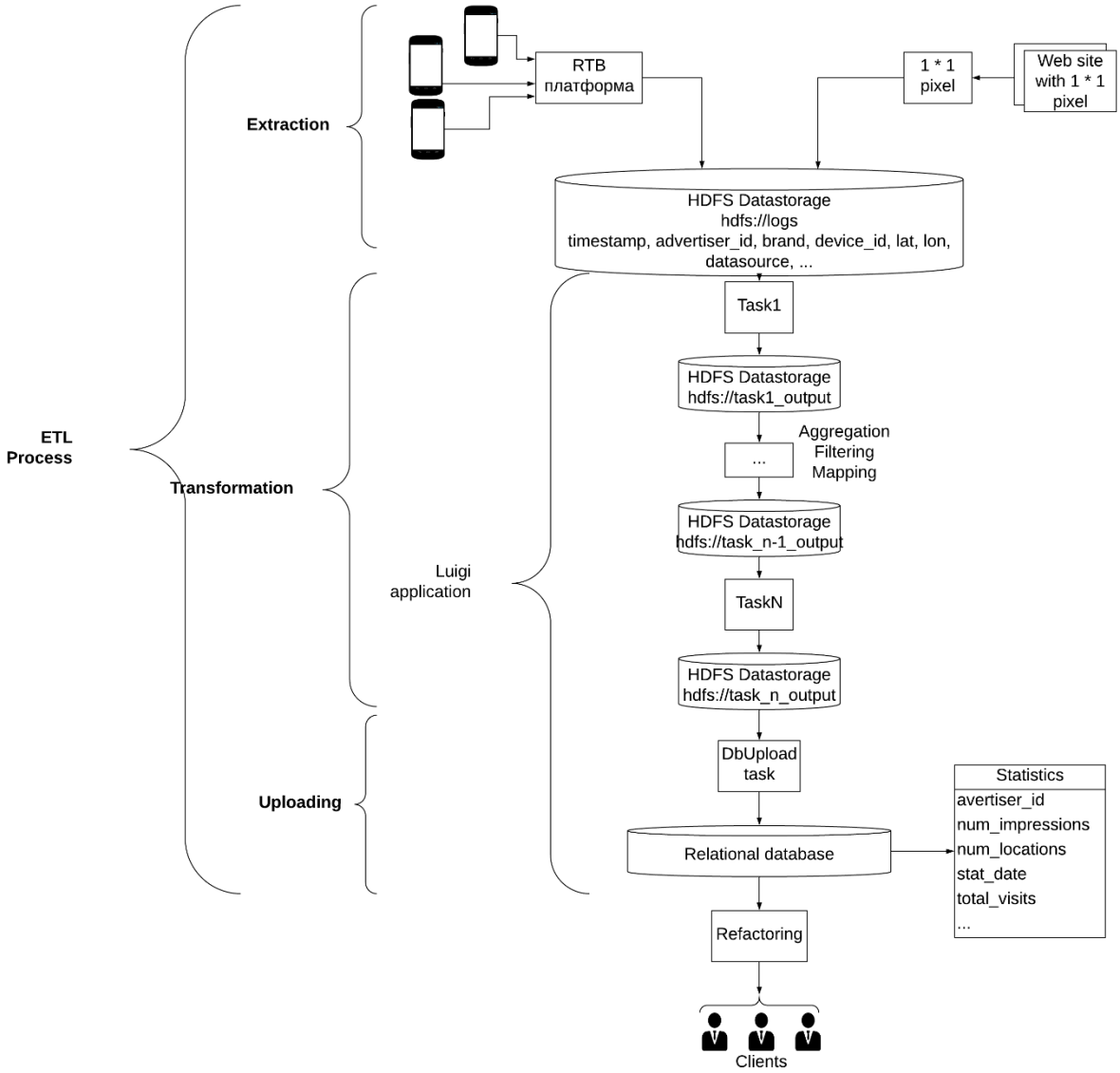

Рис 1. Представление архитектуры и стека используемых технологий

Transform. Данный шаг является «сердцем» любого ETL. Здесь требуется ясное понимание бизнес задач, а также наличия базовых знаний в предметной области. В нашей системе на данном шаге производится фильтрация, группировка и агрегирование извлеченных данных и вычисление информации, которая будет представлена в отчете. Ежеминутно мы получаем сотни тысячи строк логов (см. Рис. 2) от RTB платформы и REST сервисов (1 * 1 pixel).

```
1 df = spark.read.parquet("hdfs://logs/dt_log=2019-01-15")
2 df.count()/(24*60)
3 Out[1]: 460148
```

Рис 2. Расчет количества поступающих в минуту строк логов

Для их преобразования используется фреймворк для распределенной обработки данных Apache Spark [5] (в частности PySpark) и планировщик задач Luigi[6]. Задачи образуют граф зависимостей, а фреймворк разрешает зависимости, отслеживает граф выполнения, управляет запуском задач, обрабатывает ошибки с возможностью перезапуска нужных задач, распределяет ресурсы рабочих процессов с возможностью параллельной работы независимых частей графа задач.

Load. Наконец, мы загружаем обработанные данные в базу MySQL. Полученный набор данных используется для генерации отчетов (пример отчета на Рис. 3), которые затем отправляются рекламодателям.

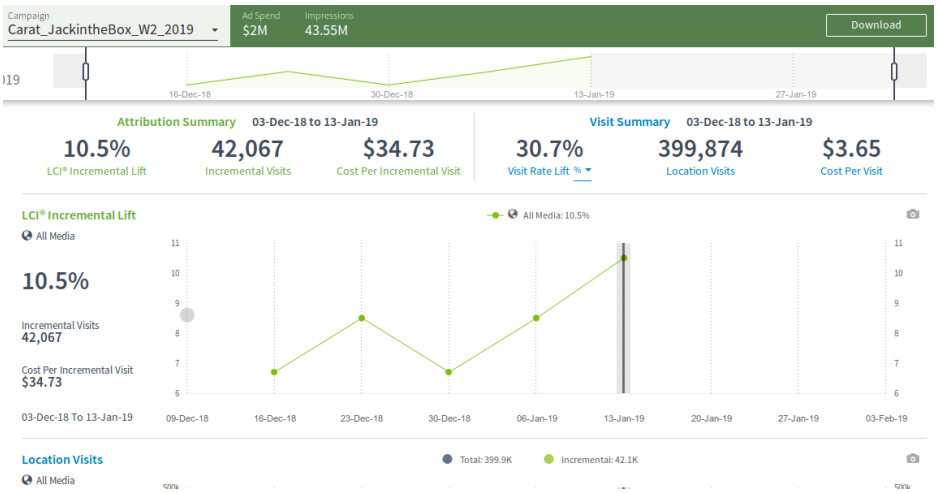

Рис 3. Пример отчета о рекламной кампании

Вывод

Данная система обрабатывает более полумиллиона событий в минуту, обрабатывает их и выдает отчет в форме, удобной для клиентов. В результате клиент может оценить эффективность рекламной кампании. Планируется перевод ключевых модулей с языка Python на Scala, что позволит ускорить работу системы и увеличить отказоустойчивость.

Список использованных источников:

1. Breur, Tom. "Statistical Power Analysis and the contemporary "crisis" in social sciences". *Journal of Marketing Analytics*, July 2016 – 65 с.
2. Loshin, D. Business Intelligence - 2nd / Loshin, D. - Morgan Kaufmann, 2012. - 400 с.
3. Stange. M. Real-Time Advertising. Business & Information Systems Engineering / M. Stange, B. Funk. - Fakultät WirtschaftswissenschaftenLeuphana University LüneburgLüneburgGermany, 2014. - 308 с.
4. White, T. Hadoop: The Definitive Guide, Fourth Edition / T. White. – Gravenstein Highway North, Sebastopol: O'Reilly Media, 2015. – 728 с.
5. Karau, H. Learning Spark / h. Karau, A. Konwinski, P. Wendell, M. Zaharia. – Gravenstein Highway North, Sebastopol: O'Reilly Media, 2015. – 255 с.
6. GitHub [Электронный ресурс] / Luigi. - Режим доступа : https://github.com/spotify/luigi. - Дата доступа : 23.12.2018.

ГИБРИДНАЯ СТРАТЕГИЯ ОТОБРАЖЕНИЯ ЛОГИЧЕСКИХ АДРЕСОВ В ФИЗИЧЕСКИЕ В НАКОПИТЕЛЯХ ИНФОРМАЦИИ НА ОСНОВЕ ФЛЭШ-ПАМЯТИ

Шиманский И.Д.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Заливако С.С. – кандидат техн. наук

Статья посвящена проблеме отображения логических адресов в физические в накопителях информации на основе NAND-технологии флэш-памяти. Выделяются и описываются характерные особенности основных стратегий отображения. Особое внимание уделено гибридной стратегии отображения. Даётся сравнение всех рассмотренных стратегий. В качестве ключевого свойства для сравнения используется объём динамической памяти, требуемый для хранения таблиц отображения. Представлены достоинства и недостатки гибридной стратегии относительно остальных, а также вектор работы для более детального изучения её характеристик.

В данной работе в качестве примера устройства используется Samsung K9XXG08UXM [1]. Необходимые характеристики приведены в таблице 1. Размер логической страницы, используемой файловой системой, будем считать равной 4КБ. Стоит подчеркнуть, что в данном примере предполагается, что размеры физических и логических страниц равны.