УДК 659.113.7

ИСПОЛЬЗОВАНИЕ ТЕХНОЛОГИИ BIG DATA ДЛЯ АНАЛИЗА ЭФФЕКТИВНОСТИ РЕКЛАМЫ

М.А. Медунецкий
Студент БГУИР

Н.В. Харитонов
Студент БГУИР

М.П. Хоронеко
Студент БГУИР

М.В. Стержанов
Доцент кафедры информатики БГУИР

Белорусский государственный университет информатики и радиоэлектроники, Республика Беларусь
E-mail: maxim.horoneko@gmail.com., mmed.log@gmail.com, nikita.kharitonov99@gmail.com. sterjanov@bsuir.by.

М.А. Медунецкий
Победитель множества олимпиад по программированию, в том числе полуфинала международной студенческой олимпиады (ACM ICPC). В 2016 году поступил на факультет КСиС Белорусского Государственного Университета Информатики и Радиоэлектроники.

Н.В. Харитонов
В июне 2016 года закончил светлогорскую гимназию, где принимал активное участие в олимпиадном движении. В школьное время неоднократно становился победителем областной и призером республиканской олимпиады по физике.

М.П. Хоронеко
В 2016 году закончил светлогорскую гимназию. Принимал активное участие в олимпиадном движении. В школьное время неоднократно становился победителем олимпиад по математике, программированию и астрономии.

М.В. Стержанов
Родился в 1984 году в Минске. В 2001 поступил в БГУИР на специальность Информатика.

Аннотация. 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.

Ключевые слова: Большие данные, реклама, анализ

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 ин-

формация 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 нам акты показа рекламы, затем агрегируем их, проводим их обработку, а затем создаем отчет как конечный результат за определенный 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

Можно заметить, что вся архитектура представляет собой ETL процесс, который организован на кластере компьютеров с использованием технологий Hadoop и HDFS.**ETL - Extract, Transform, Load**

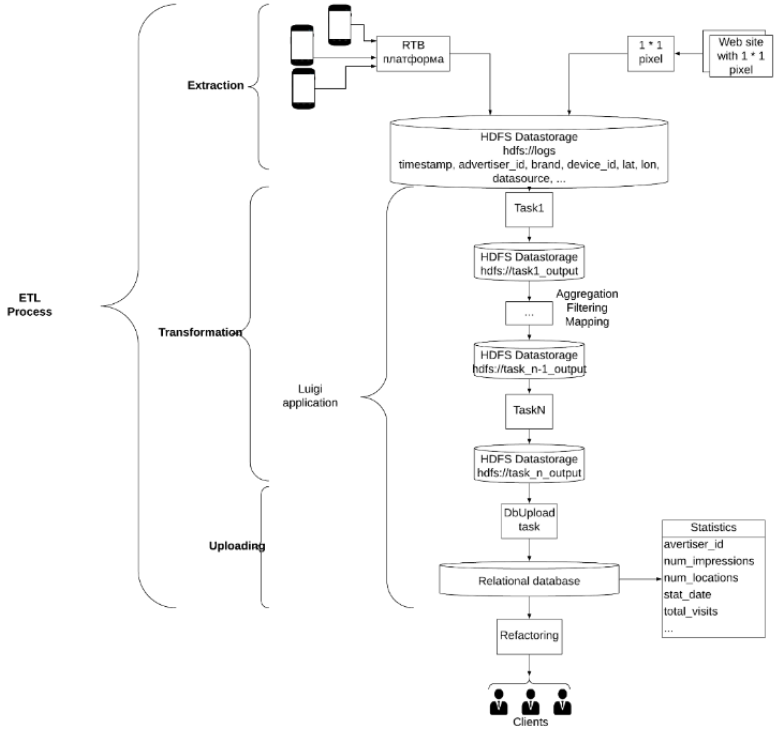

Рисунок 1. Представление архитектуры и стека используемых технологий

Extract, Transform и Load [2] — это три концептуально важных шага, определяющих устройство большинства современных конвейеров данных. На сегодняшний день это базовая модель того, как сырые данные сделать готовыми для анализа.

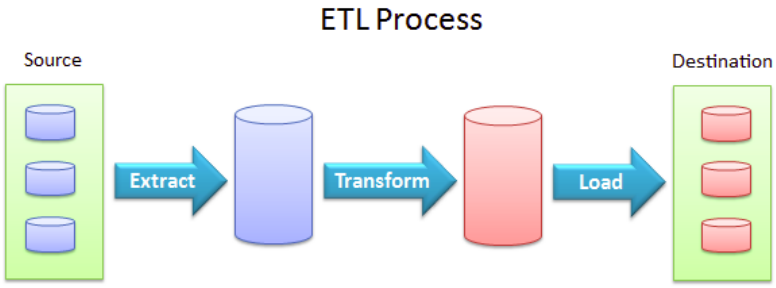

Рисунок 2. Схема базового ETL процесса

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. (рисунок 3)

```
In [16]: df.printSchema()
root
 |-- ts: timestamp (nullable = true)
 |-- bucket: integer (nullable = true)
 |-- aicode: integer (nullable = true)
 |-- advertiser: string (nullable = true)
 |-- campaign: string (nullable = true)
 |-- brand: string (nullable = true)
 |-- iol: string (nullable = true)
 |-- ad: string (nullable = true)
 |-- aud: string (nullable = true)
 |-- site: string (nullable = true)
 |-- did: string (nullable = true)
 |-- didt: string (nullable = true)
 |-- latitude: double (nullable = true)
 |-- longitude: double (nullable = true)
 |-- loc_uncertainty: double (nullable = true)
 |-- ref: string (nullable = true)
 |-- custom1: string (nullable = true)
 |-- custom2: string (nullable = true)
 |-- rdr_url: string (nullable = true)
 |-- agency: string (nullable = true)
 |-- inv_type: string (nullable = true)
 |-- channel: string (nullable = true)
 |-- publisher: string (nullable = true)
 |-- creative: string (nullable = true)
 |-- ip: string (nullable = true)
 |-- plcmnt: string (nullable = true)
 |-- xd_loc: string (nullable = true)
 |-- http_status: integer (nullable = true)
 |-- server_name: string (nullable = true)
 |-- server_addr: string (nullable = true)
 |-- server_port: integer (nullable = true)
 |-- api_key: string (nullable = true)
 |-- remote_addr: string (nullable = true)
 |-- x_forwarded_for: string (nullable = true)
 |-- jwat: string (nullable = true)
 |-- referer: string (nullable = true)
 |-- ua: string (nullable = true)
 |-- ndat: string (nullable = true)
 |-- reserved10: string (nullable = true)
 |-- reserved11: string (nullable = true)
 |-- reserved12: string (nullable = true)
 |-- advertiser_id: integer (nullable = true)
 |-- campaign_id: integer (nullable = true)
 |-- mgrs1: long (nullable = true)
 |-- mgrs10: long (nullable = true)
 |-- mgrs100: long (nullable = true)
 |-- mgrs1000: long (nullable = true)
 |-- mgrs10000: long (nullable = true)
 |-- mgrs100000: integer (nullable = true)
 |-- ds: string (nullable = true)
 |-- b: integer (nullable = true)
 |-- recovered: integer (nullable = true)
 |-- dt: string (nullable = true)
```

Рисунок 3. Входные данные (поля логов)

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 (рисунок 4) 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
```

Рисунок 4. 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]. Они отлично подходят для наших целей, так как задача PySparkTask из модуля luigi.contrib.spark позволяет нам использовать все лучшие стороны Spark'а в нашем приложении. Задачи образуют граф зависимостей (рисунок 5), 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.

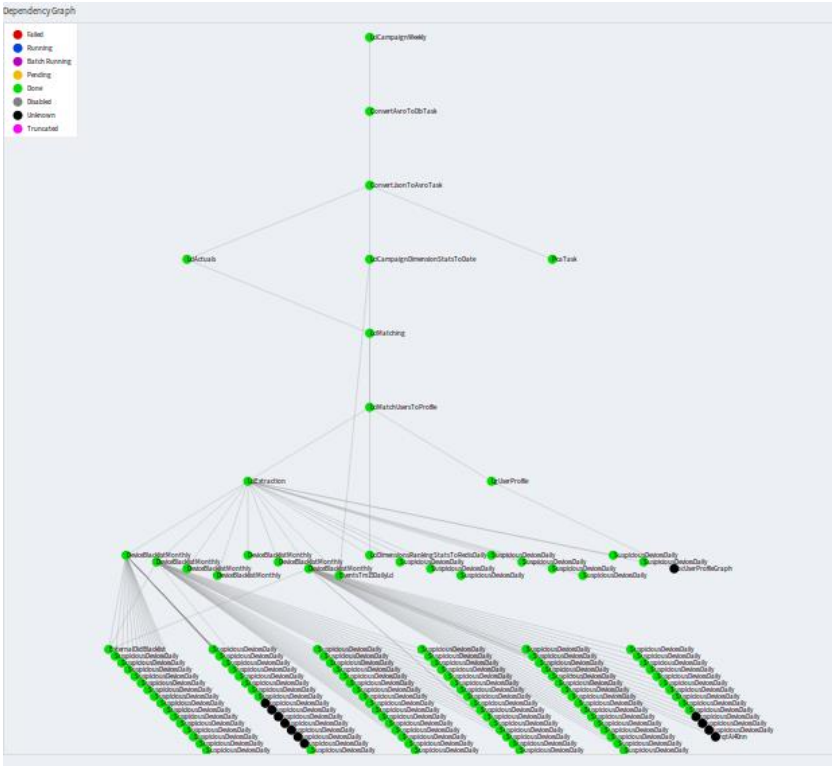

Рисунок 5. Граф зависимости Luigi задач

Load. Наконец, мы загружаем обработанные данные в базу MySQL. Полученный набор данных используется для генерации отчетов (пример отчета на Рис. 6), которые затем отправляются рекламодателям.

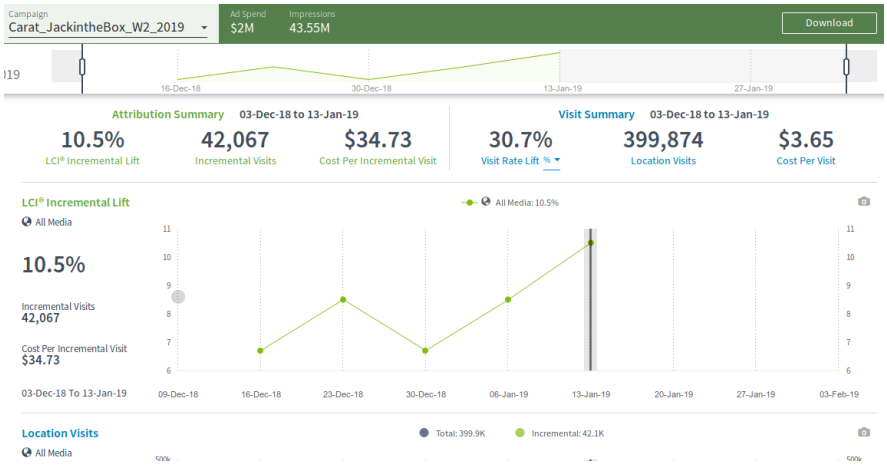

Рисунок 6. Пример отчета о рекламной кампании

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.

Литература

[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.

USING THE BIG DATA TECHNOLOGY STACK FOR ANALYSIS OF THE ADVERTISEMENT EFFICIENCY

M. MEDUNETSKI
Student of BSUIR

N. HARITONOV
Student of BSUIR

M. HORONEKO
Student of BSUIR

M. STERJANOV
Associate professor of the Informatics Department of the BSUIR

Abstract. This paper discusses the mechanism for evaluating the effectiveness of Internet advertising using the Big Data technology stack. It's about the ETL process implemented on the Luigi Task Scheduler, with which we collect, store and process large amounts of data. As a result, the architectural scheme of the platform implementing this process is presented. Also, an example of the data returned is a weekly report that includes the statistics of an advertising campaign is provided.

Keywords: big data, advertisement, analysis