An Artificial Intelligence Enabled Data Analytics Platform for Digital Advertisement

Naz Albayrak[1], Aydeniz Özdemir[1], and Engin Zeydan[2]

[1]Messaging and Advertising Department, Türk Telekomunikasyon A.S, Istanbul, Turkey, 34889.
Email: {naz.albayrak, aydeniz.ozdemir}@turktelekom.com.tr

[2] Centre Technologic de Telecomunicacions de Catalunya, Castelldefels, Barcelona, Spain, 08860.
Email: engin.zeydan@cttc.cat

***Abstract*—To offer the most suitable product to the end-users in digital marketing, end-user behaviour needs to be observed in *real time* while browsing through websites. This enables to profile and define the end-user behaviour with more accurate success rates. However, this analysis requires the development of special methods for processing large data. Within the scope of this demo, we demonstrate a big data analytics platform to increase the effectiveness of digital advertising. During experiments, we demonstrate via the dashboard interface the successful mappings of end-user's various features including URL-category report (Cookie ID - Subscriber ID, URLs- Classified Interactive Advertising Bureau (IAB) category) analysis, non-functional resorts (end-user behaviour report, category score report, device usage report), the most visited websites and new category labeling for classification purposes.**

***Index Terms*—artificial intelligence, web categorization, user profiling, big data.**

I. Introduction

Digital advertising, also called Internet advertising ("Internet marketing") is used to leverage businesses with Internet technologies to deliver promotional advertisements to end-users. Digital advertising includes promotional advertisements and messages delivered through email, social media websites, online advertising on search engines, banner ads on mobile or web sites. The main reason for advertisers to turn to digital advertisement is that it can offer targeting far beyond traditional methods. For example, an airline company has the opportunity to display ads on Facebook and sell tickets to users of a telecommunications company by combining the data of content provider and the carrier. A simple example is as follows: assume that the following targeted criteria is needed for web advertisement to appropriate users: (i) received a minimum of one signal in the last one year in the United Arab Emirates, (ii) Income segment is high, (iii) Smart device users. If various data sets can be combined appropriately for campaign analysis, a significant portion of subscription ads can be displayed to targeted users matched in Facebook.

When browsing on any website, end-user bahaviour needs to be monitored in "real time" so that the most suitable product can be offered to an end-user. As a result, it will be possible to profile and segment the end-user with a more accurate precision. Major telecommunication providers have millions of broadband users. Targeted marketing of telecommunication providers can be done by monitoring and profiling the traffic belonging to their end-users. Those end-users should also be willing to allow their broadband data to be processed within an *advertisement analytical platform* deployed inside an infrastructure provider. These issues can be addressed in the specific regulation rules of the country. The proposed *advertisement analytical platform* in this demo is a platform with a powerful infrastructure in which this big data can be processed, categorized and the results can be reported via dashboard. As a result of this analysis, advertisers can use those cookies to collect anonymous information about a websites visitors. Finally, this information is used to create profiles containing specific details about an end-user to display relevant ads. There are basically three types of cookies: (i) **1st party cookies:** Data belongs directly to a telecommunication company (CRM, website visitors, etc.), (ii) **2nd party cookies:** 1st party data belonging to the company that the telecommunication company collaborates with. At this point, two (or more) companies can share data with each other at any time., (iii) **3rd party cookies:** Data anyone can access and is sold via data market place or data exchange systems.

Google and Facebook duopoly are currently dominating the digital advertisement market. However, Mobile Network Operators (MNOs) are also interested in investing in digital ad business [1]. For example, Verizon has acquired Yahoo and AOL and launched Oath digital ad platform [1]. Telefonica has launched the fourth platform in order to increase their added-value media or internet services. AT&T has acquired AppNexus, DirecTV and Time Warnet aiming to become a TV and content company. AT&T intends to combine its traditional telecom data with the appnexus's media data. Deutsche Telekom has a single platform called Emetriq, a subsidiary targeted to create highly targeted digital ads [2]. In the literature there exists also works related to web site analysis and classification models [3], [4]. The authors in [3] provide an overview of the web analysis process for web sites. In the paper [4], a classification model is used to identify relevant documents for a user from the web, specifically for the web user profiling problem.

In this demo, by analyzing the packet source (using Deep

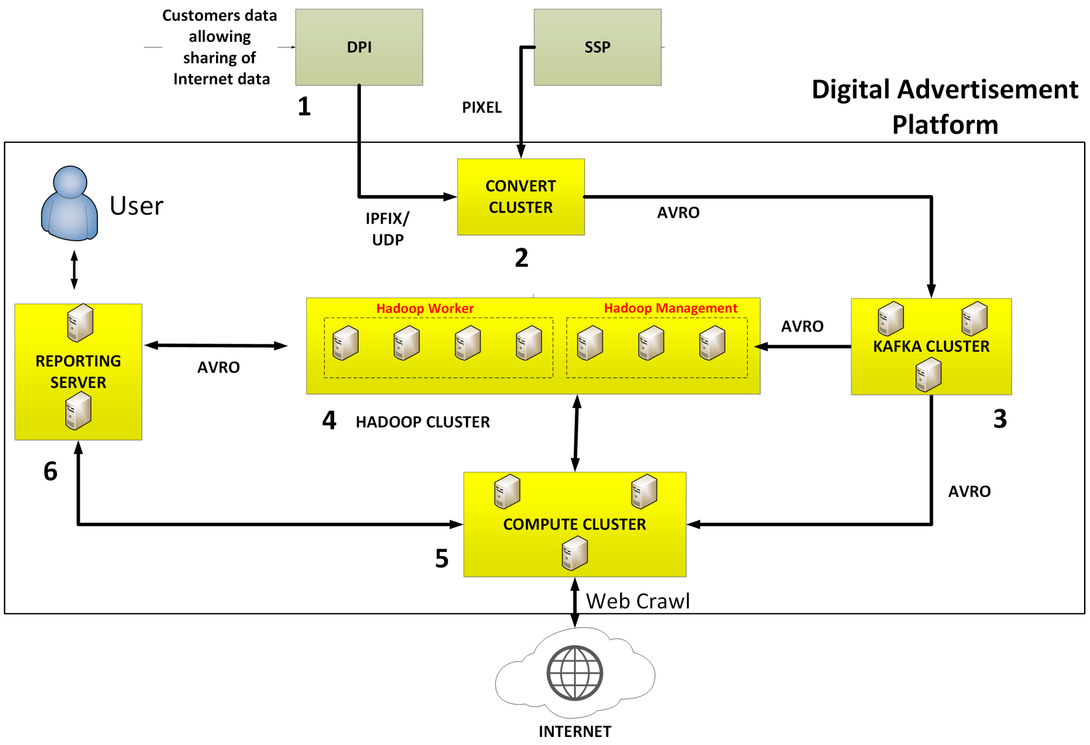


Fig. 1: General architecture of the digital advertisement platform.

Packet Inspection (DPI)) data, Internet usage habits of test end-users are analyzed. Internet usage is classified on an industry basis so that the market can be addressed in the most accurate way. Therefore, depending on test end-users' visiting website, the customer group is labelled with the corresponding category, e.g. "real estate" or "automobiles" sector. This classification allows companies that want to advertise to these customer groups to show instant advertisements. Classification of the websites can be done according to the Interactive Advertising Bureau (IAB) content taxonomy [5]. It is also possible to expand these main categories with subcategories customized for MNOs' needs.

II. Artificial Intelligence Supported Data Analytics Platform

In the proposed *advertising analytics platform*, not only DPI data of the Internet usage patterns of MNOs' end-users, but also the data from the Supply Side Platform (SSP) is used. SSPs publisher's video, native, or mobile inventory are platforms that allow advertisers to buy ad space from the inventory via software. Therefore, a publisher integrated with the SSP will have the opportunity to meet with many advertisers. By matching the "Bid Request & Response" data provided by the Supply Side Platform with the IP Address and UA (User Agent - Device & Browser Information) fields, the user's Cookie-ID information is obtained. Thus, customer segmentation with cookie-ID is becoming available for digital ad serving. All these data are available as reports from the analytics platform.

Dataset Description: The data fields in the **SSP data** are as follows[1]: TimeStamp (TS)$^{(*)}$, UserUniqueID (Uuid), FullUrl$^{(*)}$, UserIP and UserPort$^{(*)}$ (source and destination)$^{(*)}$, ISP, City, Country, Languages (browser), DeviceType (and DeviceModel, DeviceMaker), Duration (on website) and UserAgent$^{(*)}$ (agent with user's browser data). The **DPI data (IPfix)** consists of the following fields: IPv4Address and TransportPort$^{(*)}$ (source and destination), datetime$^{(*)}$, flowStartSeconds, flowEndSeconds, HttpContentType, HttpRequestMethod (GET or POST), HttpResponseStatus, HttpUrl$^{(*)}$, SubscriberIdentifier, IncomingOctets (arriving data volume), HttpUserAgent (device category), HttpContentType (for content filtering only html purposes), Service (http or https), HttpReferer (used to improve hit rate statistics).

A. Demonstration Architecture

Fig. 1 gives a general architecture and corresponding components of the proposed *digital advertisement platform* used in our demonstration. All data belonging to end-users (that allow the processing of their Internet usage data) flows instantly into the *advertising analytics platform* as given in Fig. 1. In first step, SSP and DPI data marked as (1) are collected by *convert cluster* marked as (2) from their respective providers. SSP data is collected with SFTP and in batch form. DPI data is streaming and is collected with Ipfix Netflow v9 (UDP). Inside *convert cluster* servers, streaming Ipfix and Pixel data from DPI and SSP servers respectively are combined and the

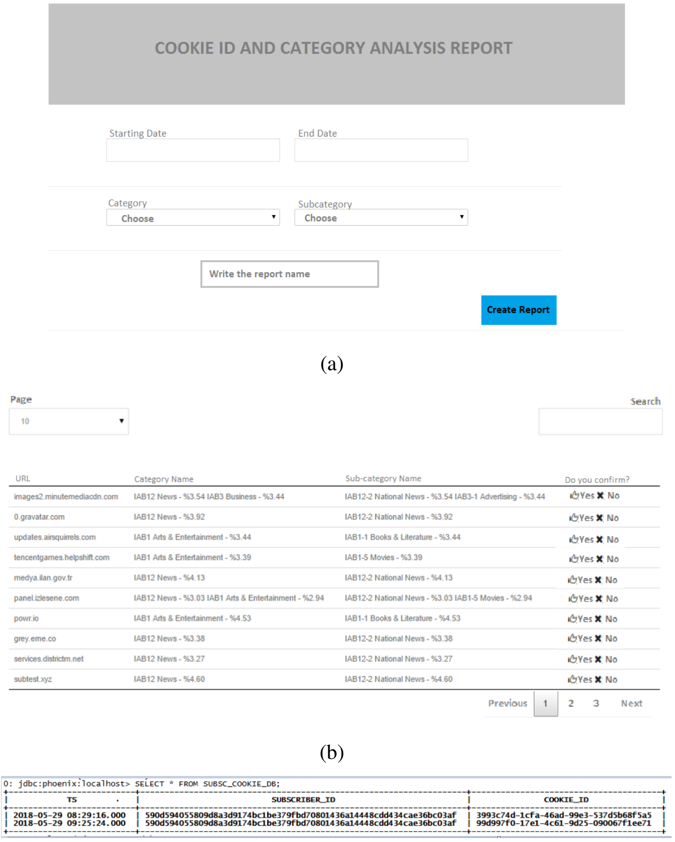
[1] The field marked with (*) are used for mapping of SSP and DPI datasets together.

corresponding data (Ipfix and Pixel) are converted into AVRO data format [6]. Later, it is sent to the Kafka servers to be consumed by Kafka consumers. *Kafka cluster* marked as (3) consists of three servers that provide communication between applications. The data available in *Kafka cluster* are distributed to related applications, i.e. *Hadoop and compute clusters*. The data is held on the Kafka server for up to 24 hours for categorization purposes. Hadoop cluster marked as (4) consists of four Hadoop workers and three Hadoop management servers. *Hadoop cluster* holds the HBASE database and is interacting with *reporting server* and *computer cluster*. Regarding the data storage policy, once the user information has been cleared, the information of the most visited sites is kept in HBASE for statistical use without holding personal information of website visits. This data is parsed so that the user and URL cannot be matched again after processing in Kafka. *Compute cluster* marked as (5) has also three servers and is used to do the crawling of incoming URLs and the category-based classification. Full analytics code improvements have been made in order to categorize the system by eliminating the parametric values in the full URL. There are two methods to determine which category each site visited by the user belongs to: First, for global web sites, the category information of companies that provide pre-identification is used. Second, for local web sites, the categories are determined by text mining. For the second method, Naive Bayes algorithm is used to decompose Web sites into categories. Full URL address is in the data with the http extension collected at the *digital advertisement platform*. Http data accounts for approximately 25% of data traffic. 75% of data is with https extension and these data are categorized based on domain knowledge.

In addition, end-user segments are created to indicate the Internet habits of these end-users by using the data such as how frequently end-users enter the web sites in these categories, and how long they stay in the web sites in this category. *Reporting servers* marked as (6) consists of two servers where we run the interface for dashboard.

B. Analysis of Demonstration Results

We have collected two months of data (January-February 2018) for SSP and 1 week of data (between 22-29 January 2018) for DPI for demonstration purposes. Each DPI observation consists of 300 million records per day. At peak hours of traffic, streaming DPI traffic consists of one million packets per second. Out of these, 5321 URLs are determined to be unique in the first given month. We also have approximately 5205 unique usable URLs in the second month of data. We have used 2000 of these data for training purposes. For testing purposes, remaining 3321 of URLs and the accumulated second months' 5205 URLs are used. Using Naive Bayes, URL - IAB category classification accuracy of 70% is achieved. The start-end dates, category and subcategories are selected from the interface shown in Fig. 2(a). Different analysis results including URL-category-cookie ID-subscriber ID report (cookie ID and subscriber ID analysis (an example in Fig. 2(c)), URLs and their classified IAB categories (an example in Fig. 2(b))), non-functional resorts (end-user behaviour report, category score report, device usage report), the most visited websites and new category labeling for classification can be done with the dashboard interface of Fig. 2.



(a)

URL	Category Name	Sub-category Name	Do you confirm?
images2.minutemediacdn.com	IAB12 News - %3.54 IAB3 Business - %3.44	IAB12-2 National News - %3.54 IAB3-1 Advertising - %3.44	Yes ✖ No
0.gravatar.com	IAB12 News - %3.92	IAB12-2 National News - %3.92	Yes ✖ No
updates.airsquirrels.com	IAB1 Arts & Entertainment - %3.44	IAB1-1 Books & Literature - %3.44	Yes ✖ No
tencentgames.helpshift.com	IAB1 Arts & Entertainment - %3.39	IAB1-5 Movies - %3.39	Yes ✖ No
medya.ilan.gov.tr	IAB12 News - %4.13	IAB12-2 National News - %4.13	Yes ✖ No
panel.izlesene.com	IAB12 News - %3.03 IAB1 Arts & Entertainment - %2.94	IAB12-2 National News - %3.03 IAB1-5 Movies - %2.94	Yes ✖ No
powr.io	IAB1 Arts & Entertainment - %4.53	IAB1-1 Books & Literature - %4.53	Yes ✖ No
grey.eme.co	IAB12 News - %3.38	IAB12-2 National News - %3.38	Yes ✖ No
services.districtm.net	IAB12 News - %3.27	IAB12-2 National News - %3.27	Yes ✖ No
subtest.xyz	IAB12 News - %4.60	IAB12-2 National News - %4.60	Yes ✖ No

(b)

```
0: jdbc:phoenix:localhost> SELECT * FROM SUBSC_COOKIE_DB;
+-------------------------+------------------------------------------------------------------+--------------------------------------+
|           TS            |                          SUBSCRIBER_ID                           |              COOKIE_ID               |
+-------------------------+------------------------------------------------------------------+--------------------------------------+
| 2018-05-29 08:29:16.000 | 590d594055809d8a3d9174bc1be379fbd70801436a14448cdd434cae36bc03af | 3993c74d-1cfa-46ad-99e3-537d5b68f5a5 |
| 2018-05-29 09:25:24.000 | 590d594055809d8a3d9174bc1be379fbd70801436a14448cdd434cae36bc03af | 99d997f0-17e1-4c61-9d25-090067f1ee71 |
+-------------------------+------------------------------------------------------------------+--------------------------------------+
```

(c)

Fig. 2: (a) Dashboard for Analysis Query (b) Analysis Report Results for URL versus IAB category (c) Cookie-ID versus Subscriber ID.

Acknowledgements

This work was partially funded by Spanish MINECO grant TEC2017-88373-R (5G-REFINE) and by Generalitat de Catalunya grant 2017 SGR 1195.

References

[1] "Oath ad platform." https://www.oath.com/, 2018. [Online; accessed 21-September-2018].

[2] "emetriq GmbH." https://www.emetriq.com/, 2018. [Online; accessed 21-September-2018].

[3] D. Booth and B. J. Jansen, "A review of methodologies for analyzing websites," in *Web technologies: Concepts, methodologies, tools, and applications*, pp. 145–166, IGI Global, 2010.

[4] J. Tang, L. Yao, D. Zhang, and J. Zhang, "A combination approach to web user profiling," *ACM Transactions on Knowledge Discovery from Data (TKDD)*, vol. 5, no. 1, p. 2, 2010.

[5] "Interactive Advertising Bureau (IAB)." https://www.iab.com/, 2018. [Online; accessed 21-September-2018].

[6] "Apache Avro." https://avro.apache.org/docs/1.2.0/, 2017. [Online; accessed 04-Dec.-2018].