GOCE DELCEV UNIVERSITY - STIP
FACULTY OF COMPUTER SCIENCE

ISSN 2545-4803 on line
DOI: 10.46763/BJAMI

BALKAN JOURNAL OF APPLIED MATHEMATICS AND INFORMATICS (BJAMI)

BALKAN JOURNAL
OF APPLIED MATHEMATICS AND INFORMATICS

YEAR 2021

VOLUME IV, Number 1

GOCE DELCEV UNIVERSITY - STIP, REPUBLIC OF NORTH MACEDONIA
FACULTY OF COMPUTER SCIENCE

ISSN 2545-4803 on line

BALKAN JOURNAL OF APPLIED MATHEMATICS AND INFORMATICS

BALKAN JOURNAL
OF APPLIED MATHEMATICS AND INFORMATICS

(BJAMI)

YEAR 2021 VOLUME IV, Number 1

AIMS AND SCOPE:
BJAMI publishes original research articles in the areas of applied mathematics and informatics.

Topics:
1. Computer science;
2. Computer and software engineering;
3. Information technology;
4. Computer security;
5. Electrical engineering;
6. Telecommunication;
7. Mathematics and its applications;
8. Articles of interdisciplinary of computer and information sciences with education, economics, environmental, health, and engineering.

Address of the editorial office
Goce Delcev University – Stip
Faculty of philology
Krste Misirkov 10-A
PO box 201, 2000 Štip,
Republic of North Macedonia

BALKAN JOURNAL
OF APPLIED MATHEMATICS AND INFORMATICS (BJAMI), Vol 3

ISSN 2545-4803 on line
Vol. 4, No. 1, Year 2021

EDITORIAL BOARD

CONTENT

DOI: 10.46763/BJAMI21310059h
UDC: 004.89:519.856]:005.92

HOW DATA ENGINEERING AND BIG DATA ANALYTICS CAN CONTRIBUTE TO INTRODUCING INTELLIGENCE IN BUSINESS: A CASE STUDY[1]

Olga Petan, Ljubinka Sandjakoska, Atanas Hristov

Abstract. Nowadays, in the era of big data, data engineering becomes one of the key features in providing competitive advantages to companies. A lot of approaches are introduced into data analytics in order to improve the businesses. This paper presents one possible way of introducing intelligence. As a case study, a comprehensive analysis of an e-commerce site is done. The analysis includes user segmentation, analysis of the conversion funnel, product analysis, and ad campaigns analysis. The data engineering process results in important findings: documents and dashboards. The findings give additional value for employees and for automation of business processes to make them more intelligent.

Keywords. data engineering, big data analytics, business intelligence.

1. Introduction

Nowadays, in the era of big data, data engineering becomes one of the keys to providing competitive advantages to companies. The advantages are highly correlated with the consolidation of various digital marketing and e-commerce tools that are able to do advanced analyses. Effective and efficient data analysis is based on a proper data architecture that will ensure correct, reliable, consistent, safe, and accessible data. All of these data features should be present in the quantitative and qualitative analysis of the data engineering process. The main goal of the data engineering process is to source, transform and analyze data from each subsystem. It is the core module between the data sources, their creation and storage, and the data analysis (Figure 1).

Designing and managing data flows that integrate information from various sources into a common pool typically involves implementing data pipelines based on the ETL (Extract, Transform, and Load) model (Figure 2).

In order to obtain novel business insights from the data analysis, ETL models should provide information clarity, completeness, quality, and velocity. Different data tools are used in this process: Apache Hadoop, Apache Spark, Apache Kafka, Apache Cassandra, SQL and NoSQL (relational and non-relational databases), etc.

A lot of approaches are introduced into data analytics so as to improve the businesses. The approaches fall into one of the categories: descriptive, diagnostic, predictive, and prescriptive. Unlike businesses, the research community does not pay much attention to developing data engineering and data analytics tools. The new approaches result from the practical work of the companies and the experience that lead to following the clients' requirements.

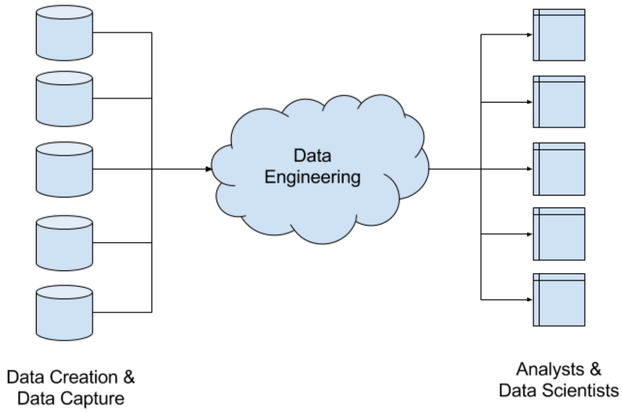

Figure 1. *Data engineering between data creation and data analysis [22]*

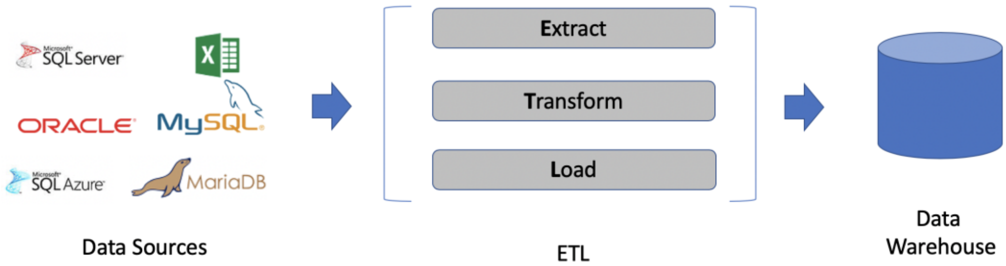

Figure 2. *ETL process overview [21]*

The paper is organized as follows. After the first introductory section there is the section that presents a case study, where a description of the data sources, e-commerce analysis issues, data architecture, and data visualization is included. At the end, the concluding remarks are given, followed by future work directions.

2. Case study

As a case study, a comprehensive analysis of an e-commerce site is done. The analysis includes user segmentation, analysis of the conversion funnel, product analysis, and ad campaigns analysis. Although all concepts discussed in this paper can be implemented in any cloud service provider, Microsoft Azure is the provider of choice for this case study.

The main goal of *user segmentation* is to answer questions such as: who our buyers are, who does not buy from the site, who should be targeted with some specific ad campaign, etc. *Analysis of the conversion funnel* is important in order to be aware of a possible bottleneck that can cause dropping out of potential users. Moreover, it is important to know what the business does right to reinforce those actions. The obtained data from the *product analysis* actually answers the questions: which products are the best sellers, which do not sell? Are people buying for themselves, or are they buying gifts? Is it possible to boost the sales of a certain product by creating ad campaigns or by offering discounts to specific user segments? Additionally, the goal is to investigate other products that people tend to buy when buying a specific product. *Ad campaigns analysis* is a challenging data analytics task because the competitive advantage of the company depends on it. Once user segments are known and the products that need to be boosted, or tailored, ad campaigns to boost sales or help visitors convert can be created. Since these are paid ads, and there is a risk of targeting wrong user segments, ads' performance needs to be additionally analyzed. Answers to the questions: How much return on investment does each ad campaign bring? Are the current actions and analyses working? What needs to change? have a lot of significance in the complete data engineering process.

The data engineering process aims to present important findings in two ways:

(1) as **documents** in a central repository - so that employees in various departments of the company can access the information they need to do their work, and

(2) as **dashboards**, so they can visually explore and make conclusions from the analyses. Hence, it is obvious that a central repository is needed to store aggregated data, results and findings, and documents.

The first step is aggregating data from Facebook Ads, Google Ads, Google Analytics, Shopify, and Klaviyo. The aggregation aims to capture user behavior data, including how the user came to the e-commerce site - either through an ad or by searching, whether they made a purchase or add products to the cart but did not make a purchase, whether they responded to an email campaign etc. The attitude herein is that unless time is taken to design and implement a comprehensive data architecture, which connects all tools a business uses in its operations, it will be impossible to draw useful and usable conclusions from the data. Disability to rely on correct data limits the implementation of any advanced intelligent technique. From this, we can conclude that data engineering is the most important phase of any data-driven project.

2.1. **Data Sources**. The e-commerce site considered in this paper uses the following digital tools:

- Shopify, to build the site, create product catalogs, create shopping carts and transaction functionalities, gather buyers' information, and generally facilitate shopping;
- Facebook Ads and Google Ads, to create ad campaigns;
- Klaviyo, to create email campaigns;
- Google Analytics, to track user behavior on the site and to measure site performance.

All these tools have APIs, and data can be accessed programmatically, bypassing tools' GUI. However, finding ways to join these different data sets is more challenging than it seems at first.

2.1.1. **Dataset.** The aggregated dataset is constructed by joining together datasets extracted from the tools mentioned above. All datasets contain information for the period between January 1st 2019, and January 1st 2021.

First, we extract the product dataset from **Shopify.** This dataset contains information about which product is sold during the annotated period, how many units were sold, how much revenue that product generated, the average price of the product, whether there were refunds for that product, etc. Dataset columns: **Product, Transaction ID, Product Revenue, Unique Purchases, Quantity, Average Price, Product Refund Amount, Cart-to-Detail Rate, Buy-to-Detail Rate.**

Next, we extract the ad campaigns datasets from **Google Ads** and **Facebook Ads**. These datasets contain information about the ad campaigns from each source (Google or Facebook), how many transactions were carried out by one buyer, what the average order was, etc. Dataset columns: **Source / Medium, Transaction ID, Users, Sessions, Revenue, Transactions, Average Order, Ecommerce Conversion Rate, Per Session Value.**

Next, we extract the ad campaigns dataset from **Klaviyo**. This dataset contains information about email ad campaigns, how much revenue each campaign brought from how many transactions, etc. Dataset columns: **Source / Medium, Transaction ID, Users, Sessions, Revenue, Transactions, Average Order, Ecommerce Conversion Rate, Per Session Value.**

We can also extract auxiliary datasets for all three datasets mentioned above. These datasets include information such as the number of users, number of new users, what the bounce rate for a session was (how many people landed on a page and left without interacting), the average session duration, number of transactions, and revenue. Auxiliary dataset columns: **Users, New Users, Sessions, Bounce Rate, Pages / Session, Avg. Session Duration, Ecommerce Conversion Rate, Transactions, Revenue.**

Finally, we extract the user demographics and user behavior datasets from Google Analytics. These datasets contain information about users' age, gender, whether they are new visitors on the site or they have browsed the site before, etc. Dataset columns: **Age, Gender, User Type, Sessions, Bounce Rate, Pages / Session, Avg. Session Duration, Transactions, Revenue, Ecommerce Conversion Rate.**

2.2. **E-commerce analyses issues**. Since there are five APIs that need to be called, a data pipeline that executes scripts that access and transform them should be created. The goal is to join five data sets into a single data set. Looking at the APIs' documentation, it is obvious that only Shopify and Klaviyo track user IDs directly. This means that there is a need for specific functionality to be developed, a functionality that enables creating a user ID for the rest of the tools: Facebook Ads, Google Analytics, and Google Ads.

However, data privacy concerns need to be taken into consideration when the user ID functionality is created. For example, Google specifically states that the business should let visitors know through a pop-up that their activity is being monitored, and they need to give the business consent. The business also needs to provide an opt-out functionality and not store user IDs if the visitors disallow it. If the business fails to comply with Google's guidelines, the business' account might be terminated. See Figure 3.

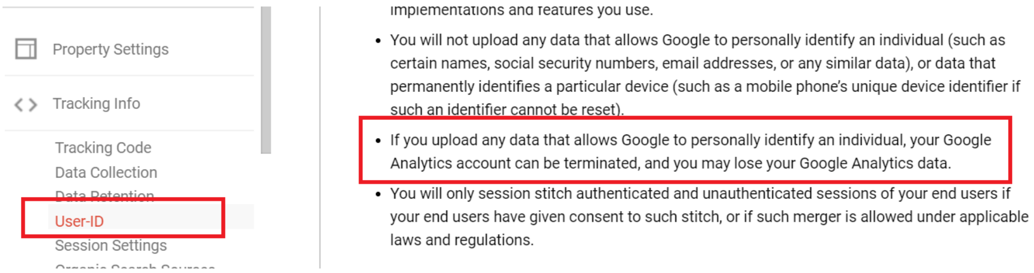

Figure 3. *Google Analytics Requirements*

Google Ads also needs to reference a user ID variable in order to differentiate users. The way to do this is to add the variable in the Re-marketing Tag code. See Figure 4.

Facebook, on the other hand, does provide the possibility of embedding a user ID in the general tracking code, Facebook Pixel. The only thing that needs to be done is to pass the user ID variable in the tracking function. See Figure 5.

The proper way to implement Google Analytics and Google Ads tracking code is to use Google Tag Manager. Furthermore, the Facebook Pixel code can also be implemented using Google Tag Manager. This way, all tracking code for all tools is compiled in one place, and the business can have a clear and consolidated picture of all the events and goals created to track user behavior. The user ID is a global variable put in the Google Tag Manager's data layer and then called in the tracking code for specific tools (Figures 3 through 5). See Figure 6.

```
<script>
gtag('event', 'page_view', {
'send_to': 'AW-CONVERSION_ID',
'user_id': 'replace with value',
'value': 'replace with value',
'items': [{
'id': 'replace with value',
'google_business_vertical': 'retail'
}]
});
</script>
```

Figure 4. *Google Ads User Id*

Use the following code to set the User ID:

```
fbq('init', '{pixel-id}', {uid: '{user-id}'});
```

Figure 5. *Facebook Pixel User Id*

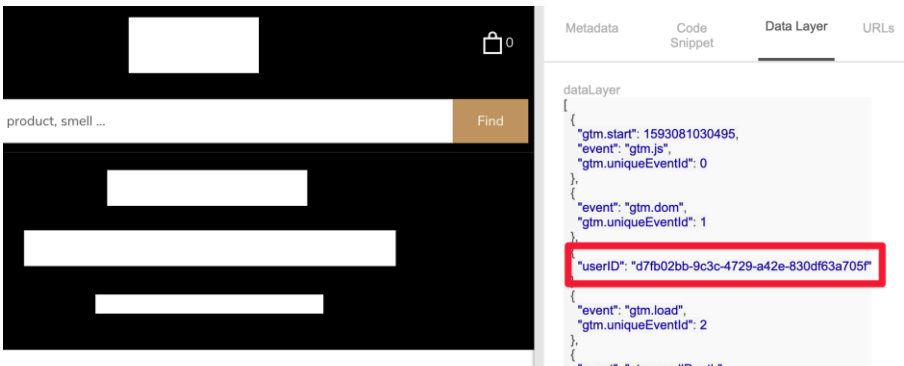

Figure 6. *User Id in the Data Layer*

Why is all this needed if Shopify and Klaviyo already track user IDs? Is there no other way data sets can be joined together? Granted, Shopify communicates with all the other tools mentioned here. However, since Shopify only gathers data of the people who completed their transaction, the later analyses would only be analyzing buyers and would fail to differentiate non-buyers. As expected, if the goal is to increase ROI (Rate of Investment) on ad campaigns, a clear differentiation between who the buyers are and why they buy, on the other hand, is necessary. Also, the information about why people do not buy is needed, so proper action can be taken to convert no-buyers. Specifically, Shopify and Google Analytics can be joined together through the Transaction Id column. Google Analytics has a default segment called Converters that uses this Transaction Id column, but, as stated previously, users cannot be easily differentiated from sessions. See Figure 7.

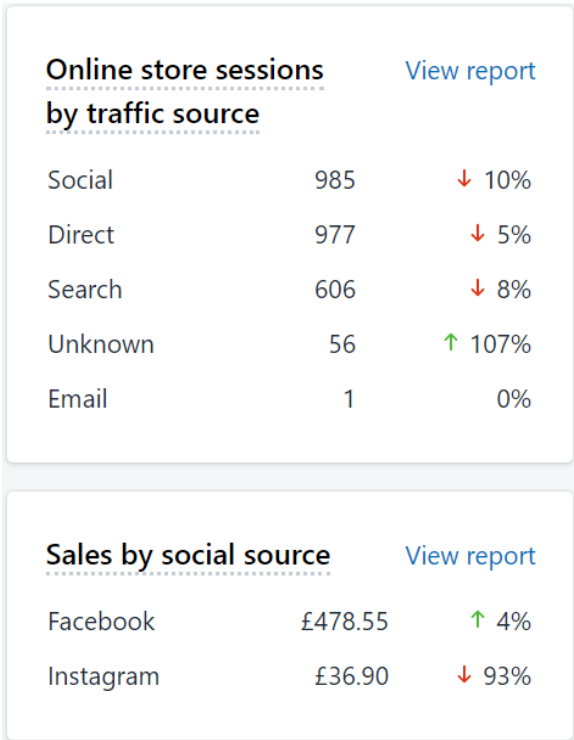

Figure 7. *Shopify Connected*

2.3. **Data architecture**. Once the justification behind creating a global user ID variable that will be embedded in the tracking code of all previously mentioned tools is established, the decision of what needs to be done to aggregate all the data sets and where to store the aggregated data should be made.

First of all, a data pipeline to orchestrate the ETL processes (Extract, Transform, Load) should be built. Developing the mentioned data pipeline is outside of the scope of this paper, but generally, business stakeholders together with data analysts define key performance indicators, pose questions they would like the analyses to answer, and write a document with everything that needs to be taken into consideration. Based on this document, scripts that will call a tool's API and pull specific data are created. In Microsoft Azure, these scripts can be written in U-SQL. The Data Lake Analytics jobs run U-SQL scripts. See Figure 8.

```
@product_mart =
    SELECT isreparented,
                        statecode,
                        productid,
                        statuscode,
                        createdon,
                        productnumber,
                        producttypecode,
                        name,
                        productstructure,
                        isstockitem,
                        modifiedby,
                        modifiedon,
                        createdby,
                        parentproductid,
                        quantitydecimal,
                        validfromdate,
                        validtodate
    FROM @product
        WHERE statuscode == 0;

OUTPUT @product_mart
TO @mart + "product.csv"
USING Outputters.Text( outputHeader: true, delimiter : '|', rowDelimiter : "~");
```

Figure 8. *U-SQL script*

Once all scripts are written, the next phase is creating a data lake that will store the data set. A data lake is a specific choice because the business needs to store both structured and unstructured data. That means all the documents that must be shared with different employees across the business need to be considered. Since data warehouses are mainly used for structured data, this data storage option is discarded for this particular use case. Microsoft Azure offers the Data Lake service for this purpose.

With the dedicated storage for the data, the scripts that pull the data and store it in the dedicated storage, and the jobs that run these scripts, the only requirement left is to automate these processes, which can be done with a data pipeline. The Azure service for creating data pipelines is Data Factory. See Figure 9 and Figure 10.

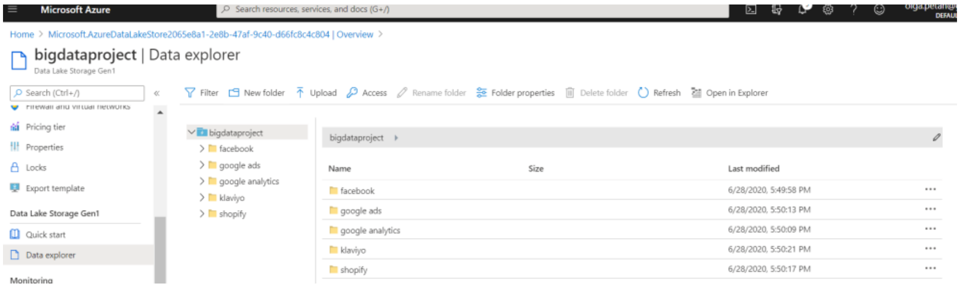

Figure 9. *Data Lake*

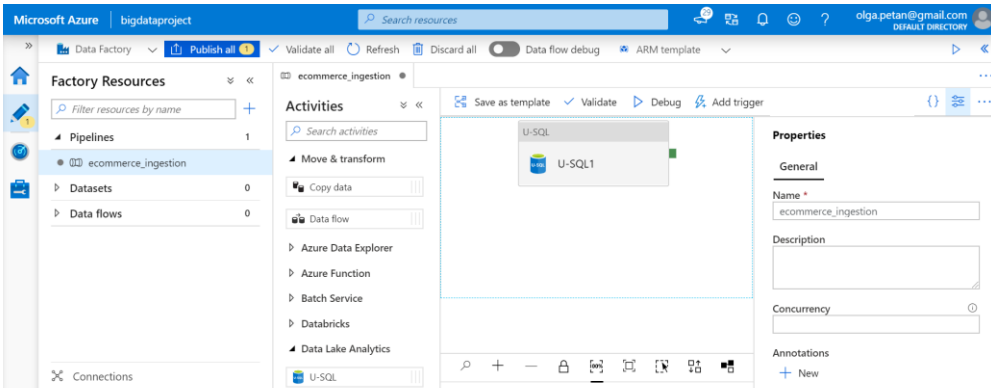

Figure 10. *Data Factory*

With these procedures, the analyses can be automated using the scripts run by the pipeline. This ensures that the data is correct, clean, organized, safe, reliable, and accessible when it is needed, and the next phase of exploring the data can begin.

2.3. **Data visualization**. One often overlooked aspect of any data analysis is data storytelling. In its essence, data storytelling is the process of translating numbers back to words. That allows the business stakeholders to know what has been found in the data, what those findings mean for the business, how they can use that information in their departments, etc. Usually, the non-technical staff wants to see metrics like sales by period, sales per product, sales per store. The end goal of the processes described in this paper is to provide useful information for the business's teams. Data storytelling is the second most important process, after data engineering, to which special attention should be paid.

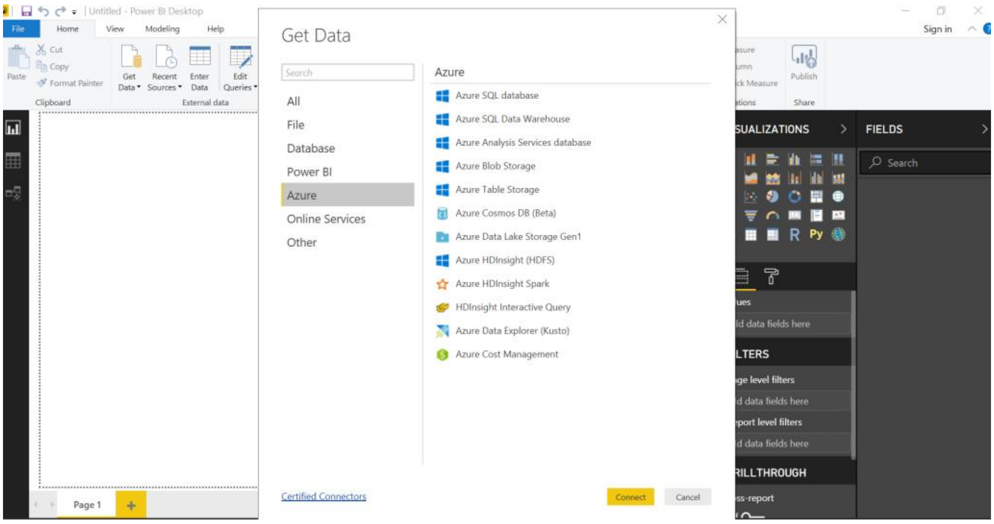

Figure 11. *Power BI*

Power BI is a useful data visualization tool to create dashboards. Power BI directly communicates with the Data Lake previously created. Moreover, an ETL process can also be created to specify exactly the data that needs to be fed in a specific dashboard. This helps immensely from a data loading perspective and speed, since not all data is loaded at once, and the client can quickly use the dashboard. See Figure 11.

3. Conclusion

The work described in this paper is the first step in developing a comprehensive data architecture that joins together data from several digital tools. That allows a business to do advanced analyses on their products and customers. Keeping in mind that a business's goal is to increase the return of investment on their ad spent, the data engineering processes described here enable the business to assign a unique, anonymous user-id to a single user and track that user across the conversion funnel. By tracking users through the conversion funnel, the business can identify bottlenecks in the different conversion steps and remove them so that more website visitors can buy more products more easily. Moreover, by dis-aggregating session data from tools like Google Analytics into user data, the business can more precisely classify users into user personas and create tailored ad campaigns, which will increase the odds that the ad campaign is successful.

The data engineering process results with important findings: documents and dashboards. A consolidated, central repository of all the data that a business stores will ensure that all the data is used for analyses, all findings, dashboards, and other documents are available and visible to all team members who need to access them so that collaboration between departments is facilitated. Thus, the findings give additional value to the employees and the automation of the business processes.

In this paper, using a case study, we try to depict the influence of data engineering on the business. Introducing intelligence in business based on data analytics improves the complete business process, ensuring a competitive advantage for the company. The main contribution of this paper is in the area of data engineering and big data analytics, with a special focus on the tools for developing an effective business model and enchant of the business processes.

4. Future work

Since the dynamics of the working environment is complex, a lot of aspects in improving business should be considered. The development of tools that are based on artificial intelligence is more than a trend. If we want to achieve a competitive advantage, we should create and use smart engines to make our business more "intelligent". One way is to use deep learning as an advanced machine learning concept. Our future work is focused on developing deep learning neural networks to predict the features of customer behavior.

References

[1] *Cielen D., Meysman A. D. B., and Ali M.:* (2016). Introducing Data Science: Manning Publishing, book
[2] *Foulds I.:* (2018). Learn Azure in a Month of Lunches: Manning Publishing, book
[3] *Nuckolls L. R.:* (2020). Azure Storage, Streaming, and Batch Analytics: Manning Publishing, book
[4] *Marz N., and Warren J.:* (2015). Big Data: Manning Publishing, book
[5] *Kleppmann M.:* (2017). Designing Data-Intensive Applications: The Big Ideas Behind Reliable, Scalable, and Maintainable Systems: O'Reilly Media, book
[6] *Siegel E.*: (2013). Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die: Wiley, book
[7] *Alhlou F., Asif S., and Fettman E.:* (2016). Google Analytics Breakthrough: From Zero to Business Impact: Wiley, book
[8] *Weber J.:* (2015). Practical Google Analytics and Google Tag Manager for Developers: Apress, book
[9] *Facebook's API.*: https://developers.facebook.com/docs/analytics/send_data/user_properties/
[10] *Google Analytics API.*: https://developers.google.com/analytics/devguides/reporting/core/v4
[11] *Google Ads API.*: https://developers.google.com/google-ads/api/docs/start
[12] *Shopify API.*: https://developers.shopify.com/
[13] *Klaviyo API.*: https://www.klaviyo.com/docs/api/metrics
[14] *Google Analytics User ID.*: https://www.clickinsight.ca/blog/implement-user-id-tracking
[15] Google Analytics User ID.: https://www.simoahava.com/analytics/improve-data-collection-with-four-custom-dimensions/#5-user-id
[16] *Google Ads User ID.*: https://support.google.com/google-ads/answer/9199250?hl=en
[17] *Azure Data Lake.:* https://azure.microsoft.com/en-us/solutions/data-lake/
[18] *Azure Data Factory.*: https://docs.microsoft.com/en-us/azure/data-factory/
[19] *Azure Data Lake Analytics.:* https://docs.microsoft.com/en-us/azure/data-lake-analytics/
[20] *U-SQL.:* https://docs.microsoft.com/en-us/azure/data-lake-analytics/data-lake-analytics-u-sql-get-started
[21] *ETL as a service.*: http://www.datapower.vn/etl-as-a-service.html
[22] *What Is Data Engineering?:* https://www.dremio.com/data-lake/data-engineering/
[23] *P. Ghavare and P. Ahire,* (2018), Big Data Classification of Users Navigation and Behavior Using Web Server Logs,Fourth International Conference on Computing Communication Control and Automation (ICCUBEA) , pp. 1-6, doi: 10.1109/ICCUBEA.2018.8697606.
[24] *S. Subbalakshmi and C. Prabhu,* (2018), Protagonist of Big Data and Predictive Analytics using data analytics, International Conference on Computational Techniques, Electronics and Mechanical Systems (CTEMS), pp. 276-279, doi: 10.1109/CTEMS.2018.8769141.
[25] *L. C. Boldt et al*., (2016) Forecasting Nike's sales using Facebook data, IEEE International Conference on Big Data (Big Data), pp. 2447-2456, doi: 10.1109/BigData.2016.7840881.
[26] *M. Cirlugea, P. Faragó and S. Hintea*, (2020), Statistical Study of Small Business Customers using Facebook Ads and Google Analytics," 43rd International Conference on Telecommunications and Signal Processing (TSP) pp. 212-215, doi: 10.1109/TSP49548.2020.9163447.

Olga Petan
University of Information Science and Technology,
Faculty of Applied IT, Machine Intelligence and Robotics,
Partizanska bb, Ohrid, Macedonia
Email address: olga.petan@isvma.uist.edu.mk

Ljubinka Sandjakoska
University of Information Science and Technology,
Faculty of Computer Science and Engineering,
Partizanska bb, Ohrid, Macedonia
E-mail address: ljubinka.gjergjeska@uist.edu.mk

Atanas Hristov
University of Information Science and Technology,
Faculty of Information and Communication Sciences,
Partizanska bb, Ohrid, Macedonia
E-mail address: atanas.hristov@uist.edu.mk