Custom application programming interface data extractor applied to the Klarna e-commerce dataset

Anas El Attaoui[1], Alaeddine Boukhalfa[2], Sara Rhouas[1], Norelislam El Hami[1]
[1]Science and Engineering Laboratory, National School of Applied Sciences, Ibn Tofail University, Kenitra, Morocco
[2]Systems Engineering Laboratory (LGS), Sultan Moulay Slimane University, Beni Mellal, Morocco

Article Info

Article history:

Received Mar 16, 2023
Revised May 15, 2023
Accepted May 19, 2023

Keywords:

Big data
Business intelligence
Data analysis
E-commerce sales
Non structured data
Webpage dataset

ABSTRACT

The use of smart technologies, the internet of things (IoT), social media, and others produce a billion or more pieces of data in different formats. Big data has risen to become the most sought-after field in computer science. The e-commerce evolved significantly and continued to flow until now and even after the pandemic. So, big data technologies helped with the development and approach to collecting, storing, processing, and extracting the data in this field. This paper proposes an application programming interface (API) data extractor tool applied to a collection of e-commerce public websites named "Klarna dataset" to extract its data, and an analysis of the results. The study of e-commerce sales has given results matching universal e-commerce sales tendencies. The peak of the number of e-commerce transactions and sales was between 2018-2019. Thus, the highest e-commerce sales price was in the United States for "luxury" or "fancy" products, and the highest sales in Europe were in Frankfurt, Germany, for hardware and gaming material.

Corresponding Author:

Anas El Attaoui
Science and Engineering Laboratory, National School of Applied Sciences, Ibn Tofail University
Hay Rahma Sector A N 345, Sale, Morocco
Email: anaselatt045@gmail.com

1. INTRODUCTION

The term "big data" is used to refer to vast amounts of diverse data types, and it is a rapidly growing field. The explosion of big data is driven by an abundance of sources, such as smartphones and the internet, and it is utilized in a wide range of applications [1]. To manage this large volume of data, new tools and strategies are being developed, such as powerful computer systems, advanced software, and sophisticated algorithms. These tools and strategies enable businesses and industries, such as e-commerce, to make better use of the data and improve the services they offer. The field of big data is constantly evolving and has become an essential component in many industries. It is crucial to stay updated with the newest methods of handling and analyzing data in order to gain valuable insights and stay ahead of the competition [2]–[5].

Undoubtedly, with the rapid advancement of technology, e-commerce has grown exponentially, making it possible for businesses of all sizes to establish an online presence. This has led to a variety of product data being available on the web in various forms, such as customer reviews, product rankings, images, and more. As a result, many businesses have been able to start and improve their e-commerce operations without the need for a physical storefront. This has not only made it easier for businesses to reach customers but also increased the competition in the e-commerce market, leading to a wider variety of products and services available to consumers. Furthermore, the availability of this product data has enabled businesses to make more informed decisions and improve the customer experience [6], [7]. This data also plays a crucial role in benchmarking and studying competition, which can assist in making important business decisions, whether it be launching a new product or developing an existing one. We are currently in an era where billions of users purchase a wide range of products from online shopping websites, and this number is increasing rapidly day by day. As a result, analyzing big data has become an essential aspect of e-commerce. It allows businesses to gain a deeper understanding of customer behavior and preferences, as well as market trends, which can help them make more informed decisions and stay ahead of the competition. It also enables businesses to identify areas for improvement, optimize their marketing strategies and customer service, and ultimately increase their revenue. Overall, big data analysis is a crucial tool for businesses to stay competitive in the fast-paced and ever-evolving e-commerce market [8]–[10].

In this research paper, we work with e-commerce data from the Klarna product page dataset, which is a dataset of openly accessible pages that correspond to products sold online on multiple e-commerce sites between 2018 and 2019. On each page, analysts defined some key points of interest, manually labeling them. These labels are presented in the hypertext markup language (HTML) code. The purpose of this study is to give an in-depth analysis of Klarna's e-commerce sales using an innovative method and algorithm known as the application programming interface (API) data extractor. The major goal of this program is to extract and parse data from the Klarna e-commerce dataset so that it can be easily used for further analysis. One of the key findings of the study was that the number of e-commerce transactions and sales was important in the period between 2018 and 2019. This trend is in line with what has been widely reported across the industry, as more and more people turned to online shopping as a way to avoid going to physical stores and to make it easier to buy products during sales or happy days to gift someone.

2. PROPOSED METHOD

Klarna product page dataset is a public collection of online product pages that were gathered from various e-commerce websites. This dataset includes offline snapshots of 51,701 product pages sourced from 8,175 different merchants across 8 different countries, including the United States, Great Britain, Sweden, the Netherlands, Finland, Norway, Germany, and Austria. The data collection period was between 2018 and 2019. Each page in the dataset has been labeled by analysts with 5 key pieces of information, including the product price, image, name, add-to-cart button, and cart button (if available). These labels are included in the HTML code of each page as an attribute called «Klarna-ai-label» and can take on one of four possible values: price, name, picture, and add to cart [11]. The data gathered, stored, and processed in big data can come from a variety of fields and be generated by a multitude of heterogeneous data sources, resulting in a massive amount of structured, unstructured and multimodal data [12], [13]. Processing large amounts of digital data from many channels necessitates the use of specialized information technology (IT) solutions. There are dozens of them, and most of them are based on the open-source concept, HADOOP, and others for example [14]–[16].

From structured data, we can obtain a relational database where the tables are made up of columns that represent the different fields or attributes of the data. The relational model is a way of organizing and structuring data in a database, where data is organized into tables and relationships are established between the tables using foreign keys [17]–[19]. Cardinality in a relational database refers to the number of unique values that can exist in a particular column or relationship. The three types of cardinalities are one-to-one, one-to-many, and many-to-many. In a one-to-one relationship, each value in one column corresponds to a unique value in another column. In a one-to-many relationship, one value in a column can correspond to multiple values in another column. In a many-to-many relationship, multiple values in one column can correspond to multiple values in another column [20]–[22]. Additionally, the structured data format allows for more efficient and effective data analysis and can be used as input for business intelligence models [23]–[25].

In this study, we proposed a new approach to extracting and structuring the data from the Klarna dataset. Specifically, we have developed a custom API data extractor, which is designed to retrieve non-structured data from the dataset and transform it into a structured format. The API can extract the data from the dataset, identify key elements and attributes, and then organize the data into a structured format that is more easily analyzed and understood. This approach allows us to extract valuable insights and information from the Klarna dataset that would otherwise be difficult or impossible to access.

The Klarna dataset is structured in the form of a tree, with the top-most level being the root labeled "Klarna." Within this root level, there are multiple folders, each containing additional nested folders. Each e-commerce sale transaction within the dataset as seen in the Figure 1 is represented by two files one in JSON (JavaScript object notation) and the others in MHTML (multipurpose internet mail extensions (MIME) encapsulation of aggregate HTML documents): "page_metadata.json" and "source.mhtml" the "source.mhtml" file contains information about the transaction, such as the price, product_name, its image, and add_to_card button. The "page_metadata.Json" file contains further information like the country, currency, date, and many more. Together, these two files provide a comprehensive picture of each e-commerce sale transaction in the dataset.

2.1. Algorithm

The first step in extracting data from the Klarna dataset involves reading through all the folders within the tree structure. This includes traversing through any subfolders and subdirectories that may be present. Once all the folders have been read, the algorithm searches each folder for specific files containing data, such as “page metadata.json” and “source.mhtml,” and saves them in memory. These two files are referred to as “data files” in the remainder of this study. The algorithm for extracting data files from folders tree as you can see in the Figure 2.

After extracting the data files, our next step is to retrieve all the information contained within them. The algorithm in Figure 3 begins by iterating through each data file, reading, and analyzing its contents. As it reads through the file, it looks for specific patterns or markers that indicate the presence of relevant information, such as “Klarna-ai-label” that is in “source.mhtml” files and JSON objects in “page_metadata.json”. Once all the information has been collected, we will save it to an output file called “database.json”. This file will be in the JSON format, which is a lightweight data-interchange format that is simple for people to read and write while also being simple for computers to process and generate. The “database.json” file will be used to store all the information we have collected, so that it can be easily accessed and used for further analysis or processing. The API data extractor is a tool that enables the extraction of data from the Klarna dataset, which is non-structured in its raw form. The extractor converts this data into a structured format, making it more usable and accessible for analysis. Noting that the first and second steps are referred to by step 1 and step 2, the full algorithm for its functioning is as follows in Figure 4.

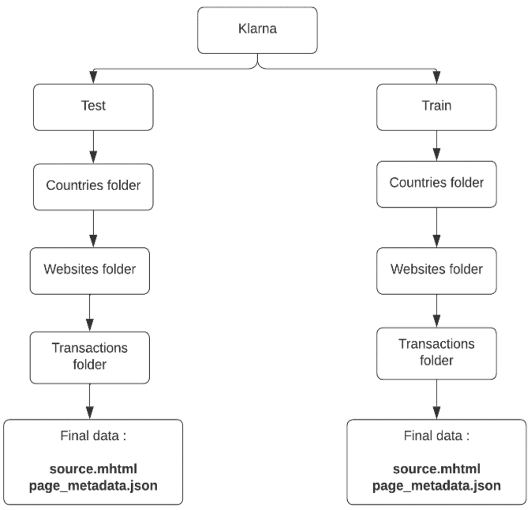

Figure 1. Klarna e-commerce dataset structure

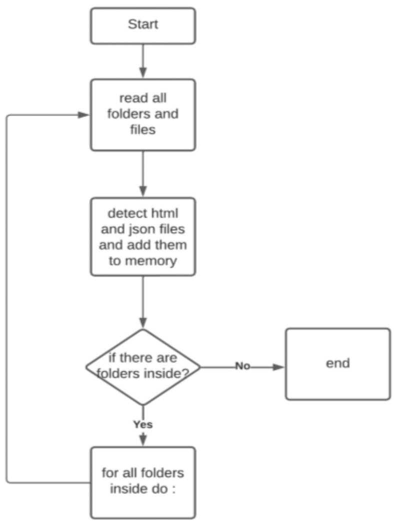

Figure 2. Extracting data files from folders tree algorithm

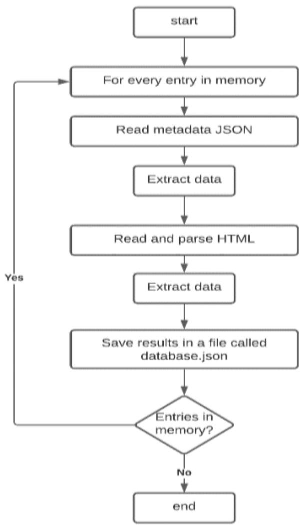

Figure 3. Extracting data from data files algorithm

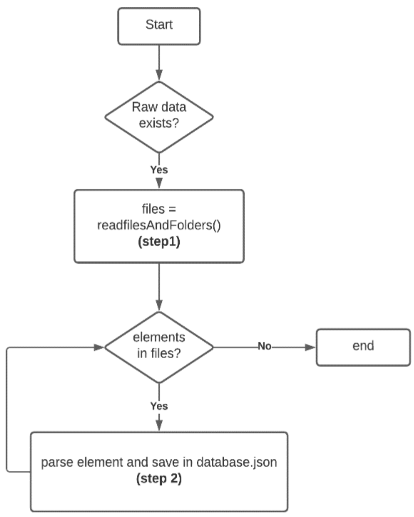

Figure 4. API data extractor full algorithm (step1+step 2)

3. RESULTS AND DISCUSSION

Previously, we used the API data extractor to extract data from the Klarna dataset. This data was then structured and organized in a way that made it easy to understand and use. Once the data was structured, it was stored in a file called "database.json." The Table 1 is a representation of all the data that was retrieved using the API data extractor and stored in the "database.json" file.

In this section, we will take the "database.json" file as input for a business intelligence IT tool, and we will use this tool to conduct a thorough analysis of the data contained within the file. By using a business intelligence tool, we will be able to manipulate and visualize the data, making it much easier to understand and interpret quickly and easily. Overall, using the "database.json" file as an input on a business intelligence IT tool will be a powerful resource for uncovering insights and understanding the data.

Table 1. Klarna e-commerce dataset extracted data

Data type	Source file
Country Currency Date City Region Region_code Country_code Continent_code In_eu Utc_offset Website Country_calling_code Timezone Languages User_agent	page_metadata.json
Page_type Price Product_name Picture Add_to_cart	source.mhtml

3.1. Data transformation

Before utilizing the "database.json" file, we performed several modifications to it, such as specifying the data type for each column and rearranging the columns in a specific order. It's worth noting that the data pertains to eight different countries. As a result, the prices for the products are expressed in various currencies. To facilitate comparisons, we converted all of the prices to the same currency United States dollar (USD) by joining our database with an international exchange rate table that contains exchange rates of August 17, 2022, at 13:40 UTC (coordinated universal time).

3.2. Data modeling

For organization and clarity, we divided our dataset into four separate tables. One table, known as the "fact table" that contains information about e-commerce sales. The remaining three tables, referred to as "dimension tables" that include information about all of the actors that can potentially influence these sales. Additionally, we created a dimension table specifically for dates, which allows for easy filtering and analysis of the sales based on the date they occurred. As a summary, we have at the end five tables, one as a fact table called "Klarna e-commerce sales", and four dimensions called "territories", "websites", "currencies", and "dates" as you can see in Figure 5. This decomposition of the dataset into multiple tables allows for more efficient querying, analysis, and understanding of the data.

3.3. Data visualization

Data visualization is the act of successfully communicating information and insights from data by utilizing graphical representations such as charts and graphs. It is a strong tool for making large data sets easier to comprehend. To thoroughly analyze our data, we will utilize a variety of different data visualization techniques to represent the information in a clear and easy-to-understand manner. By using four graphs, we will be able to effectively communicate different aspects of the data, highlight important trends, and identify patterns that might not be immediately obvious from looking at raw numbers.

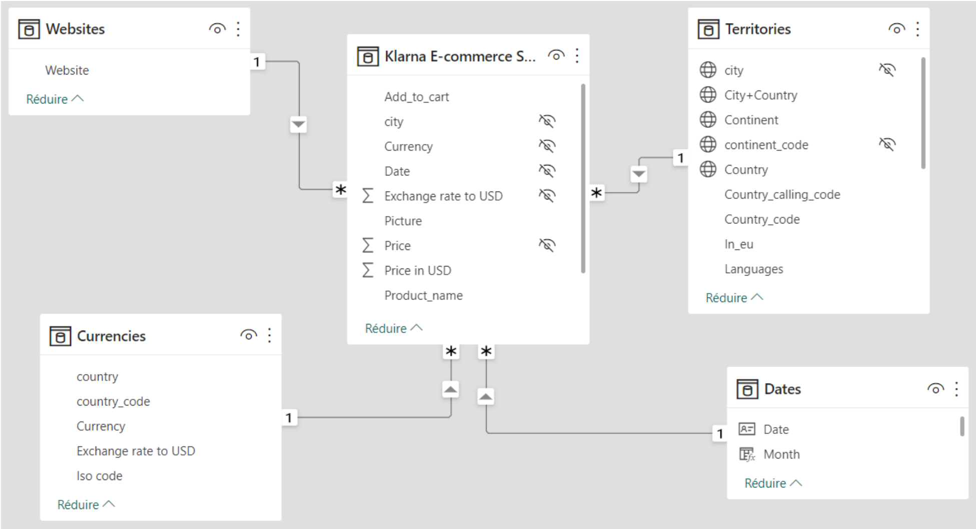

Figure 5. The conceptual model of Klarna e-commerce dataset

3.3.1. Moving average of number of transactions per date

The Figure 6 is a representation of the moving average of the number of transactions per date over 14 days. The moving average is a statistical method that is used to smooth out fluctuations in data by calculating the average of a set of data over a specific period. In this case, it is used to show the overall trend in the number of transactions per date over 14 days. The graph is plotted with the number of transactions per date on the y-axis and the date on the x-axis. The line representing the moving average is plotted in blue and is intended to show the overall trend in the data. The three red points on the graph represent the values of the number of transactions for the dates December 21, 2018, February 20, 2019, and July 10, 2019, respectively. These red points are intended to demonstrate how the number of transactions on these specific dates compares to the overall trend represented by the moving average. There are upward and downward trends in the data, with a significant increase beginning in November 2018. This increase can be seen to peak on December 21 of 2018, which is likely due to the holiday season as people tend to do more shopping online to fulfill their needs for christmas. Moreover, the increase in online shopping during this period can also be due to the convenience of online shopping, as well as the availability of more diverse products at competitive prices. Online retailers may have also run holiday promotions and sales that encouraged customers to make purchases.

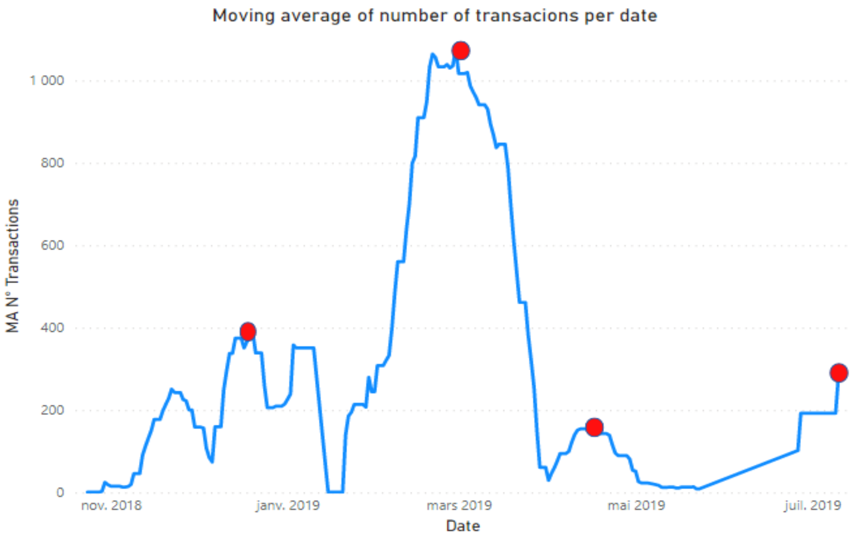

Figure 6. Moving average of number of transactions per date

After the peak in December, the number of transactions decreases, but then there is a sudden increase between February and March, hitting a global optimum in March of 2019. This increase in activity can be explained by the variety of events that occur during these two months, which can greatly affect e-commerce sales. For example, valentine's day on February 14th and International women's day on March 8th led to an increase in the sales of gifts, flowers, and expensive merchandise that symbolize love, affection, and care. Additionally, spring break, which varies by location and school, can also affect sales of travel packages and vacation-related items. At the end of spring break in early May, transaction activity typically slows down before picking up again during the summer. This trend could be attributed to the summer sales season, as people tend to have more disposable income during this time of year and may be inclined to make purchases for vacations, travel, and other needs. Overall, this graph provides a clear representation of the trend in the number of transactions over a period, and it is evident that certain events and factors have a significant impact on the data.

3.3.2. Moving average of total sales in USD per date

The Figure 7 is a representation of the moving average of total sales per date over 14 days. We have observed that the optimal outcome for the total sales is the same as that for the number of transactions. This correlation means that as the number of transactions increases, there is a corresponding increase in the overall sales figures.

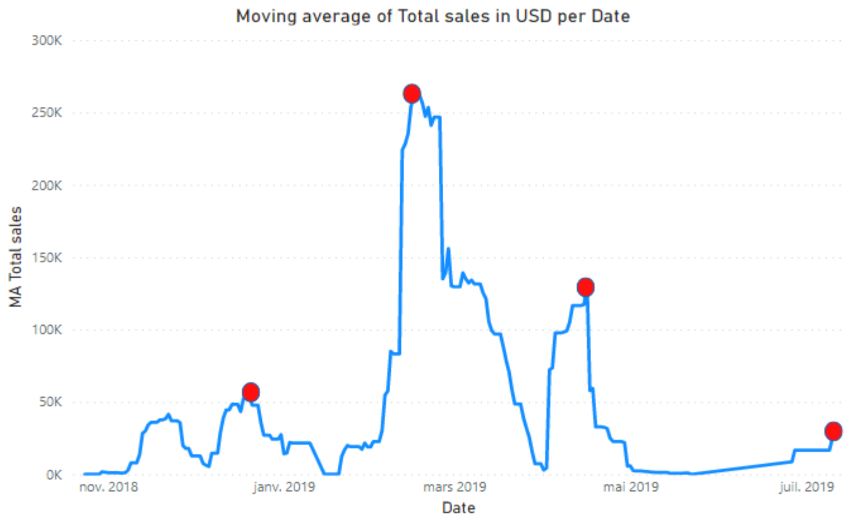

Figure 7. Moving average of total sales in USD per date

3.3.3. Total sales by country tree

The Figure 8 displays the distribution of total sales across various countries, cities, dates, websites, and product categories. Our analysis reveals that most of the total sales were generated in Chicago, United States, on February 11, 2019, on a well-known website, with the product category being "luxury" or "fancy." This date is close to valentine's day, a day known for gift-giving and romantic gestures, which confirms our initial hypothesis from the analysis of the first graph regarding the number of transactions per date. It is worth mentioning that for confidentiality reasons, the specific website and product names have been removed from the graph, and replaced by website 1, website 2, website 3..., and product 1, product 2, product 3..., respectively.

3.3.4. Total sales by continent tree

The Figure 9 present the total sales for each country included in our dataset, we can see the top three countries in terms of e-commerce sales that generated the highest percentage of total sales among all the countries in the dataset are the United States the top performer, followed by Germany and Finland. This data shows that e-commerce is a thriving industry in these countries, and it's worth noting that these countries have a relatively high purchasing power. This information can be useful for businesses looking to expand their e-commerce operations, as these countries may be key markets to target.

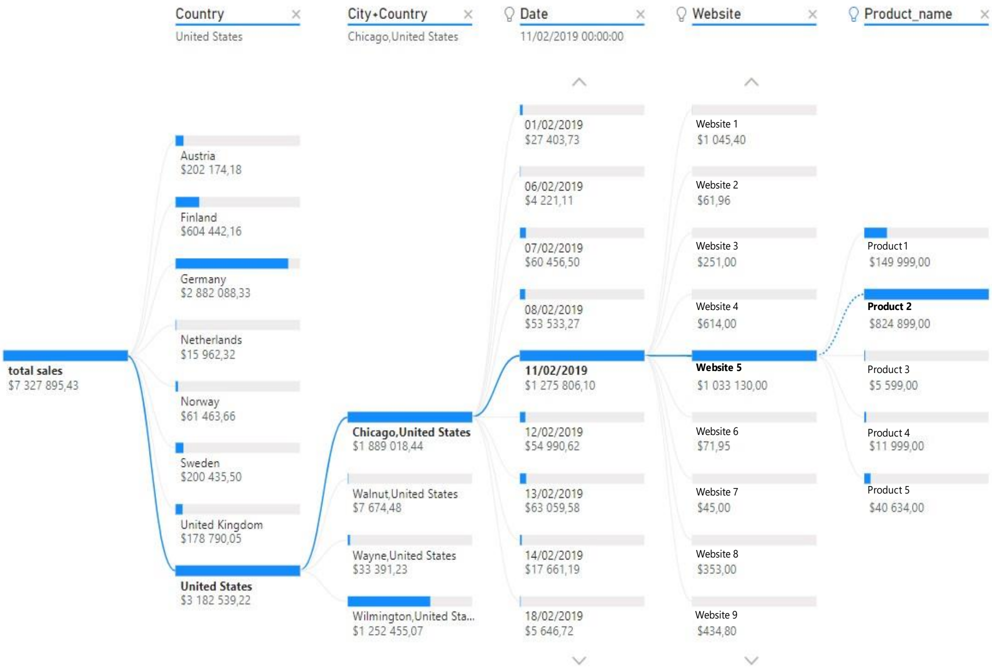

Figure 8. Total sales by country tree

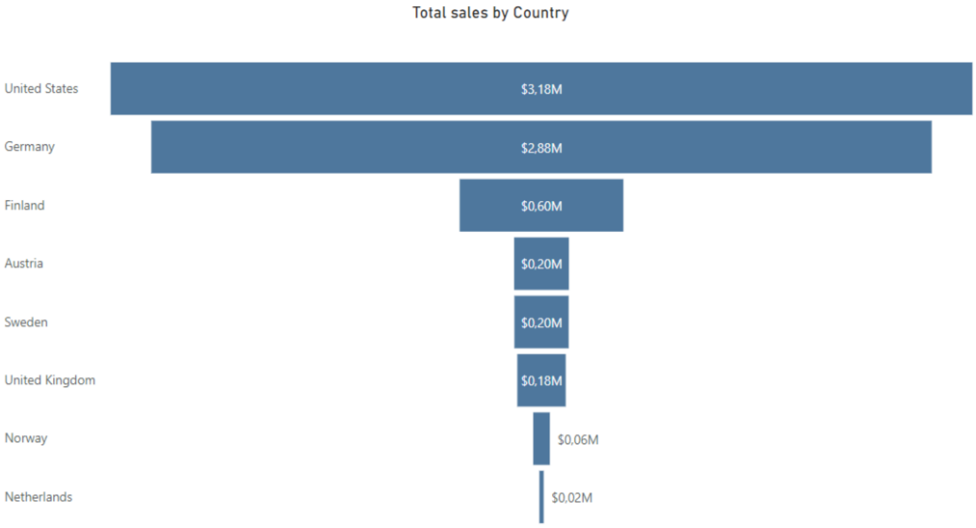

Figure 9. Total sales by country

4. CONCLUSION

Processing large amounts of digital data from various sources necessitates the use of specialized IT tools of which there are many options. In this study, we utilized an API data extractor to process a Klarna e-commerce dataset and conducted an analysis of the data. The results obtained were found to correspond with real-world events and global trends, which lends credibility to the method of data extraction, transformation, and loading as well as the method used. This API data extractor is a powerful tool that allows us to gather and process data from different sources, such as e-commerce websites. This makes it possible for us to extract and analyze data in a structured and efficient manner, which is crucial when dealing with large amounts of data. Furthermore, the data analysis provides insights that can be used to improve e-commerce strategies, track customer behavior, and identify market trends, which can be beneficial for businesses and researchers alike. Overall, the use of specialized IT tools and the method employed in this study demonstrate the effectiveness of data analysis in understanding real-world events and global trends.

REFERENCES

[1] Y. Talaoui, M. Kohtamäki, M. Ranta, and S. Paroutis, "Recovering the divide: a review of the big data analytics-strategy relationship," *Long Range Planning*, vol. 56, no. 2, p. 102290, 2023, doi: 10.1016/j.lrp.2022.102290.

[2] N. Deepa *et al.*, "A survey on blockchain for big data: approaches, opportunities, and future directions," *Future Generation Computer Systems*, vol. 131, pp. 209–226, 2022, doi: 10.1016/j.future.2022.01.017.

[3] W. A. Günther, M. H. R. Mehrizi, M. Huysman, and F. Feldberg, "Debating big data: a literature review on realizing value from big data," *Journal of Strategic Information Systems*, vol. 26, no. 3, pp. 191–209, 2017, doi: 10.1016/j.jsis.2017.07.003.

[4] K. Manley, C. Nyelele, and B. N. Egoh, "A review of machine learning and big data applications in addressing ecosystem service research gaps," *Ecosystem Services*, vol. 57, p. 101478, 2022, doi: 10.1016/j.ecoser.2022.101478.

[5] J. O'Reilly, "Big data," *Network Storage*. Elsevier, pp. 127–149, 2017, doi: 10.1016/b978-0-12-803863-5.00008-x.

[6] T. Zheng, "Research on the design of rural tourism e-commerce system based on big data technology," *Procedia Computer Science*, vol. 208, pp. 414–419, 2022, doi: 10.1016/j.procs.2022.10.058.

[7] Z. Zhu, S. Wang, F. Wang, and Z. Tu, "Recommendation networks of homogeneous products on an E-commerce platform: measurement and competition effects," *Expert Systems with Applications*, vol. 201, p. 117128, 2022, doi: 10.1016/j.eswa.2022.117128.

[8] Y. Jing and P. Yang, "Research on the risk prevention of cross-border e-commerce logistics in China by applying big data technology," *Procedia Computer Science*, vol. 208, pp. 407–413, 2022, doi: 10.1016/j.procs.2022.10.057.

[9] G. Ma, J. Ma, H. Li, Y. Wang, Z. Wang, and B. Zhang, "Customer behavior in purchasing energy-saving products: big data analytics from online reviews of e-commerce," *Energy Policy*, vol. 165, p. 112960, 2022, doi: 10.1016/j.enpol.2022.112960.

[10] P. A. Riyaz and S. M. Varghese, "A scalable product recommendations using collaborative filtering in hadoop for bigdata," *Procedia Technology*, vol. 24, pp. 1393–1399, 2016, doi: 10.1016/j.protcy.2016.05.159.

[11] A. Hotti, R. Risuleo, S. Magureanu, A. Moradi, and J. Lagergren, "Graph neural networks for nomination and representation learning of web elements," *Neural Networks*, 2021, [Online]. Available: https://arxiv.org/abs/2111.02168v3.

[12] S. T. Faraj Al-Janabi and M. A. S. Rasheed, "Public-key cryptography enabled kerberos authentication," *4th International Conference on Developments in eSystems Engineering, DeSE 2011*. IEEE, pp. 209–214, 2011, doi: 10.1109/DeSE.2011.16.

[13] P. Koupil, J. Bártík, and I. Holubová, "MM-evocat: a tool for modelling and evolution management of multi-model data," *International Conference on Information and Knowledge Management, Proceedings, ACM 2022*, pp. 4892–4896, 2022, doi: 10.1145/3511808.3557180.

[14] J. S. Kumar, B. K. Raghavendra, S. Raghavendra, and Meenakshi, "Performance evaluation of map-reduce jar pig hive and spark with machine learning using big data," *International Journal of Electrical and Computer Engineering*, vol. 10, no. 4, pp. 3811–3818, 2020, doi: 10.11591/ijece.v10i4.pp3811-3818.

[15] B. E. B. Semlali and C. El-Amrani, "Big data and remote sensing: a new software of ingestion," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 2, pp. 1521–1530, 2021, doi: 10.11591/ijece.v11i2.pp1521-1530.

[16] M. A. Rahman, J. Hossen, A. Sultana, A. Al-Mamun, and N. A. A. Aziz, "A smart method for spark using neural network for big data," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 3, pp. 2525–2534, 2021, doi: 10.11591/ijece.v11i3.pp2525-2534.

[17] S. Shetty, B. D. Rao, and S. Prabhu, "Growth of relational model: Interdependence and complementary to big data," *International Journal of Electrical and Computer Engineering*, vol. 11, no. 2, pp. 1780–1795, 2021, doi: 10.11591/ijece.v11i2.pp1780-1795.

[18] C. J. F. Candel, D. Sevilla Ruiz, and J. J. García-Molina, "A unified metamodel for NoSQL and relational databases," *Information Systems*, vol. 104, p. 101898, 2022, doi: 10.1016/j.is.2021.101898.

[19] B. Bender, C. Bertheau, T. Körppen, H. Lauppe, and N. Gronau, "A proposal for future data organization in enterprise systems-an analysis of established database approaches," *Information Systems and e-Business Management*, vol. 20, no. 3, pp. 441–494, 2022, doi: 10.1007/s10257-022-00555-6.

[20] V. C. Storey, "Understanding semantic relationships," *The VLDB Journal*, vol. 2, no. 4, pp. 455–488, 1993, doi: 10.1007/BF01263048.

[21] V. Leis, A. Gubichev, A. Mirchev, P. Boncz, A. Kemper, and T. Neumann, "How good are query optimizers, really?," *Proceedings of the VLDB Endowment*, vol. 9, no. 3, pp. 204–215, 2016, doi: 10.14778/2850583.2850594.

[22] H. Yu, M. Davis, C. S. Wilson, and F. T. H. Cole, "Object-relational data modelling for informetric databases," *Journal of Informetrics*, vol. 2, no. 3, pp. 240–251, 2008, doi: 10.1016/j.joi.2008.06.001.

[23] H. Chen, R. H. L. Chiang, and V. C. Storey, "Business intelligence and analytics: from big data to big impact," *MIS Quarterly: Management Information Systems*, vol. 36, no. 4, pp. 1165–1188, 2012, doi: 10.2307/41703503.

[24] C. Shao, Y. Yang, S. Juneja, and T. GSeetharam, "IoT data visualization for business intelligence in corporate finance," *Information Processing and Management*, vol. 59, no. 1, p. 102736, 2022, doi: 10.1016/j.ipm.2021.102736.

[25] J. Wang, A. H. Omar, F. M. Alotaibi, Y. I. Daradkeh, and S. A. Althubiti, "Business intelligence ability to enhance organizational performance and performance evaluation capabilities by improving data mining systems for competitive advantage," *Information Processing and Management*, vol. 59, no. 6, p. 103075, 2022, doi: 10.1016/j.ipm.2022.103075.

BIOGRAPHIES OF AUTHORS

Anas El Attaoui He is a Ph.D. student in computer science from National school of applied science, Ibn Tofail University, Morocco. He received an engineer degree (2019) in industrial engineering from National school of applied science, Ibn Tofail University, Morocco. His research interests reside in the fields of big data and business intelligence. He can be contacted at email: anaselatt045@gmail.com.

Alaeddine Boukhalfa is a professor of computer science at the National School of applied sciences in Khouribga, Sultan Moulay Slimane University Beni Mellal, Morocco. He received an engineer degree in computer science from National school of applied science, Abdelmalek Essaâdi University, Tetouan, Morocco. He got his Ph.D. degree in computer sciences at national school of applied sciences, Ibn Tofail University Morocco. His research interests include big data and deep learning. He can be contacted at email: alaeddine.boukhalfa@gmail.com.

Sara Rhouas currently, she is a Ph.D. student in computer science from National school of applied science, Ibn Tofail University, Morocco. She received an engineer degree (2019) in industrial engineering from National school of applied science, Ibn Tofail University, Morocco. Her research interests reside in the fields of big data, interoperability and artificial intelligence. She can be contacted at email: rhouas.sara@gmail.com.

Norelislam El Hami is a professor of computer science at the national school of applied sciences in Kenitra, Ibn Tofail university, Morocco. He obtained a diploma of state engineer in 2000 from the polytechnic Faculty of MONS (FPMS) in Belgium. The author's areas of specialization were computer and telecommunications. In June 2012, he earned a Ph.D. in computer science from the national institute of applied sciences of rouen (INSA) in France, and Ph.D. in applied mathematics and computer science from Mohammed V-Agdal University, EMI, Rabat-Morocco. He can be contacted at email: norelislam@outlook.com.