Avrupa Bilim ve Teknoloji Dergisi
Sayı 21, S. 680-689, Ocak 2021
© Telif hakkı EJOSAT'a aittir
Derleme Makalesi

www.ejosat.com ISSN:2148-2683

European Journal of Science and Technology
No. 21, pp. 680-689, January 2021
Copyright © 2021 EJOSAT
Review Article

Big Data Companies and Open Source Movement

Necmi Gürsakal[1*], Sevda Gürsakal[2], Sadullah Çelik[3]

[1] Fenerbahçe University, Faculty of Engineering, Departmant of Industrial Engineering, İstanbul, Turkey, (ORCID: 0000-0002-7909-3734), negursakal@gmail.com
[2] Bursa Uludağ University, Faculty of Economics and Administrative Sciences, Departmant of Econometrics, Bursa, Turkey, (ORCID: 0000-0002-1324-3648), sdalgic@uludag.edu.tr
[3*] Aydın Adnan Menderes University, Nazilli Faculty of Economics and Administrative Sciences, Departmant of Econometrics, Aydın, Turkey, (ORCID: 0000-0001-5468-475X), ssadullah.celik@gmail.com

(First received 5 November 2020 and in final form 31 January 2021)

(**DOI:** 10.31590/ejosat.822219)

ATIF/REFERENCE: Gürsakal, N., Gürsakal, S. & Çelik, S. (2021). Big Data Companies and Open Source Movement. *European Journal of Science and Technology*, (21), 680-689.

Abstract

The purpose of this study is to discuss the misuse of open source software by big data companies for other reasons. Developments in information and communication technologies in recent years have increased the use of Big Data and open source software. Open source software such as R, Python, Hadoop, Spark, MapReduce are developed by many people and these are used in many technologies such as Big Data, Data Science, Artificial Intelligence, Internet of Things and Blockchain. Open source software is also of great importance in terms of approaches that add value to Big Data such as machine learning and deep learning. The source code of these software is open to everyone and everyone can contribute and use them for free for their desired purpose. Today, many big data companies such as Apple, Amazon, Google, Facebook, Microsoft, Samsung, Yahoo and Qualcomm are working hard to accelerate machine learning and develop hardware suitable for software. Also, big data companies have started to share their open source software information by establishing the TODO Group. Unfortunately, the open source movement aimed at sharing, devotion; has begun to turn into a romantic effort that serves big data companies in the face of organized movements. The open source software movement whose aim is to provide free, reliable and quality software to everyone; Used by big data companies for profit (other than the purposes of the movement). On the other hand, the open source software movement is of great importance in terms of the rapid spread of information, the use and sharing of the produced codes by everyone. Big data companies first use the movement for software development and then make this software for a fee. Microsoft has done this in the NodeXL program, which is used for visualizing networks.

Keywords: Big Data, Big Data Companies, Open Source Software.

Büyük Veri Şirketleri ve Açık Kaynak Hareketi

Öz

Bu çalışmanın amacı, açık kaynak kodlu yazılımların büyük veri şirketleri tarafından amaçları dışında kötüye kullanılabileceğini tartışmaktır. Son yıllarda bilişim ve iletişim teknolojilerinde yaşanan gelişmeler Büyük Veri ve açık kaynak kodlu yazılımların kullanımını artırmıştır. R, Python, Hadoop, Spark, MapReduce gibi açık kaynak kodlu yazılımlar çok sayıda kişi tarafından geliştirilmekte ve bunlar Büyük Veri, Veri Bilimi, Yapay Zeka, Nesnelerin İnterneti ve Blok Zincir gibi birçok teknolojide kullanılmaktadır. Makine öğrenmesi ve derin öğrenme gibi Büyük Veri'ye değer katan yaklaşımlar açısından da, açık kaynak kodlu yazılımların önemi büyüktür. Bu yazılımların kaynak kodları herkese açıktır ve bunlara herkes katkıda bulunup istediği amaç doğrultusunda ücretsiz kullanabilir. Bugün Apple, Amazon, Google, Facebook, Microsoft, Samsung, Yahoo ve Qualcomm gibi birçok büyük veri şirketi, makine öğrenmesini hızlandırmak ve yazılıma uygun donanım geliştirmek için yoğun çalışmalar yapmaktadır. Ayrıca büyük veri şirketleri, TODO Group'u kurarak açık kaynak kodlu yazılım bilgilerini birbirleriyle paylaşmaya başlamışlardır. Ne yazık

[*] Corresponding Author: ssadullah.celik@gmail.com

ki, paylaşımı, özveriyi amaçlayan açık kaynak hareketi; büyük veri şirketlerinin organize hareketleri karşısında onlara hizmet eden romantik bir çabaya dönüşmeye başlamıştır. Amacı ücretsiz, güvenilir ve kaliteli yazılımı herkese sunmak olan açık kaynak kodlu yazılım hareketi; büyük veri şirketleri tarafından (hareketin amaçları dışında) kâr amacıyla kullanılmaktadır. Diğer taraftan, açık kaynak kodlu yazılım hareketi bilginin hızla yayılımı, üretilen kodların herkes tarafından kullanımı ve paylaşılması açısından da büyük öneme sahiptir. Büyük veri şirketleri hareketi önce yazılım geliştirme amacıyla kullanmakta, daha sonra ise bu yazılımı ücretli hale getirmektedirler. Microsoft, ağların görselleştirilmesinde kullanılan NodeXL programında bunu yapmıştır.

Anahtar Kelimeler: Büyük Veri, Büyük Veri Şirketleri, Açık Kaynak Kodlu Yazılım.

1. Introduction

> "Copyleft-All Rights Reversed"[†]
> Richard Stallman

In the early 2000s, three phenomena drew attention: the rapid spread of open source software, significant capital investments in open source projects, and a new organizational structure in which these projects were managed (Tirole & Lerner, 2000). In the process of processing raw data, when the amount of this data constantly increased and reached a size that could not be processed or stored on commercial computers, the digital world began to seek solutions in terms of software and hardware. The processing of data in distributed systems with open source software such as Hadoop[‡], cloud computing technology and other technological developments brought out the Big Data phenomenon:

The tools available for the problems of Big Data such as volume, velocity and variety have been developed in recent years. Generally, these technologies are not expensive enough to hamper startups, and most software is open source software. Hadoop, the most commonly used framework, combines commercial machines with open source software and provides tools for data analysis by distributing incoming data streams to cheap disks (McAfee et al., 2012).

As we approach the end of the second decade of the 2000s, there are many opposing views on Big Data in societies, from the fact that this development will be a panacea to all kinds of problems, to the issue that the big brother will constantly watch and observe us. The Big Data phenomenon is largely about analyzing unstructured data. This phenomenon, which we have mentioned, has brought diversity in data types with it, and as seen in Figure 1, the uncertainty especially in the data types of the sensors has increased.

Also, the fact that data becomes dynamic from a static state and begins to be streaming data is another important development at this point. When we remember the saying "Needs are the mother of discovery", we think that all these developments may have brought to the agenda "do it yourself" type open source solutions. In short, the incredible change in the amount and type of data may have supported different solutions in the development of software for processing this data.

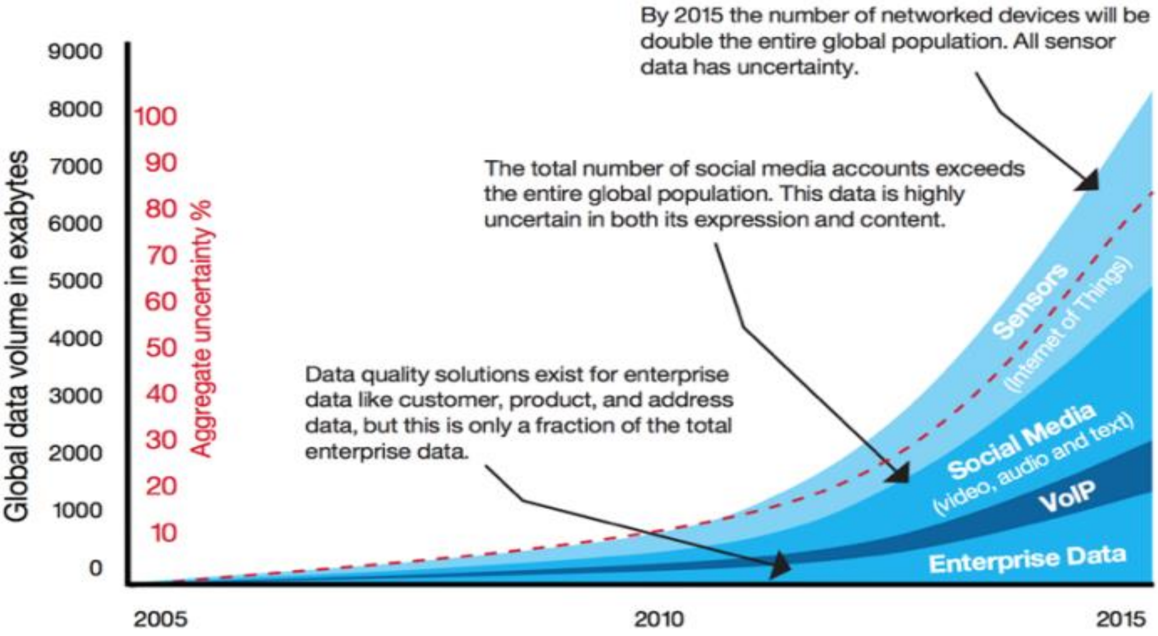

Figure 1. Exponential increase in global open source data over the last 10 years (Dash line describes uncertainty in the type of data) (IBM, 2013).

If we want to understand what is happening in the digital world and to have a solid place for ourselves in the digital world as a country, we need to examine the developments in today's developed countries and especially in the USA. The first point to be understood here is that the popularity of Big Data phenomenon is gradually giving way to machine learning[§]. Today, the Big Data

[†] The pseudonym of the general public license (General Public License) is transformed into a word game "Copyright", meaning copyright, to "Copyleft", saying "all rights are not reserved on the contrary."

[‡] For open source projects related to Hadoop only, see. http://www.learnquest.com/e-learning-samps/LQFTF/April/Hadoop/U1P2.pdf

[§] When we look at the searches for "Big Data" and "Machine Learning" with Google Trends, it is seen that MachineLearning caught "Big Data" in early May 2017 worldwide; In the USA, we see that "Machine Learning" surpasses "Big Data" in October 2016 (Piatetsky, 2017).

boom has been left behind and almost all Silicon Valley is focused on machine learning and driverless car production. Today companies like Apple, Qualcomm, Samsung and Google; They are focused on developing hardware suitable for software to speed up machine learning.

No matter how important the technical details of Big Data are, analytical approaches such as machine learning and deep learning are important in terms of adding value to Big Data. Open source[**] software is of great importance in terms of both Big Data and approaches that add value to Big Data such as machine learning and deep learning.

The fact that people started to share their data on social media without any charge has both improved the Big Data phenomenon and increased the demand for new software for such data. Data is increasing enormously and new software is needed for storing, processing and analyzing this data as the costs of storing data are decreasing rapidly. This demand was met by those who made open source software, who was mostly voluntary. Especially in the 2000s, open source software started to take an important place in the eyes of big data companies such as Facebook, Amazon, Google, Apple and Microsoft.

There are three types of licenses in software technologies:

- *Licensed software:* In this case, the software product is produced and controlled by a software company. The source code of the program is not available for the persons or companies that the company producing the software licensed. The issue of annual maintenance and upgrade costs of the program is determined by an agreement between the seller and the purchaser. Databases of Oracle, IBM or Teradata or Windows programs sold by Microsoft can be considered as licensed software.
- *Open source software:* The source code of the software is freely available. However, in this case, companies can turn the software into money by selling value-adding components such as management tools or support services. Examples of this are Cloudera and other companies marketing for Hadoop and MongoDB company for MongoDB. Open source software: "It guarantees full access to the source code, the right to run the program for any purpose without restriction, the right to change the source code, the right to distribute original and modified software, the right to know open source rights" (Carillo and Okoli, 2008). Besides, another point to be added here is that licenses are free, but it should be aware that operating a program is a costly process that requires repair, maintenance and expertise when necessary.
- *Cloud services:* Cloud services are data and software services offered by companies from data centers to customers on the Internet. Payments can be based on subscription and term of use. Examples of this are Google App Engine and Amazon Elastic MapReduce, which provide cloud computing services.

Today, digital infrastructure is largely provided by open source software. Especially in recent years, big data companies such as Google, Amazon, Facebook and Microsoft, which have determined the future of digital technology, attach importance to open source software and sometimes make their software open source. However, just as physical infrastructure has maintenance problems, digital infrastructure also has maintenance problems.

2. Big Data

"Big Data" is a term often used by software and hardware companies to increase their sales. This term has impressive business potential and is an extremely important technological trend. Although Big Data has emerged with the introduction of the internet into our lives, it has just begun to be noticed by people. Every movement or click we make on the internet during the day causes data formation. Considering that millions of people do the same movement during the day, a huge data stack is produced every second. On the other hand, this is not only limited to social media, but terabytes of data are also produced every second in many areas, and this enormous data is recorded in the databases of organizations (Çelik, 2018: 48-49). When organizations bring together this data available to them, they know more about their customers than ever before. For example, public health officials need more detailed information to better make policy decisions regarding the management of increasingly scarce resources. The ability to gain insight from Big Data will undoubtedly be of immense socio-economic significance (Cavoukian & Jonas, 2012).

"Big Data refers to datasets whose size is beyond the ability of typical database software tools to capture, store, manage, and analyze" (Manyika et al., 2011). To better understand Big Data, five basic properties that define it, namely 5V (Volume, Velocity, Variety, Value, and Veracity) should be well known (Figure 2).

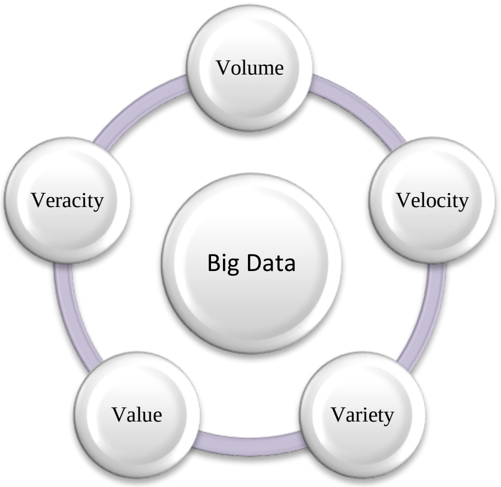

Figure 2: Big Data Characteristics

Volume: refers to the amount of data and creates difficulties in processing large amounts of data with traditional computation or techniques. The volume of the data is a problem for Big Data (Zuech et al., 2015).

[**] Source code is a set of computer-implemented instructions for a program to accomplish its purpose. The source code is then converted into machine code by a tool called a compiler.

Velocity: refers to the speed at which data is processed, and it poses a major problem for Big Data when the speed of data moves too fast to be processed by conventional computing or techniques (Zuech et al., 2015).

Variety: refers to the complexity of the data (Zuech et al., 2015). Big Data consists of high dimensional, many sources or many different data structures (structured, semi-structured or unstructured). Since these different data structures contain complex problems, some problems may arise during the analysis process.

Value: Indicates whether a value-added product that can be used by a business has been obtained as a result of Big Data analysis. If a correlation, estimation, statistical result or a value added product is not obtained as a result of the analysis, this data is no different from information garbage.

Veracity: It refers to the reliability or usefulness of the results obtained from data analysis. In short, it expresses whether the data is received from an accurate and reliable source. As the number of data sources and types increases, it becomes harder to maintain trust in Big Data Analytics (Najafabadi et al., 2015).

Big data is not about the data itself, but about new use cases and new insights. Big data analytics examines huge, detailed datasets to reveal hidden patterns, unknown correlations, market trends, customer preferences, and new business insights. Because it can only store bulk data, people can now ask questions previously impossible with a traditional data warehouse. Imagine for a moment looking at the picture of the Mona Lisa and seeing only large pixels. This is the view you get from customers in a data warehouse. To get a detailed view of your customers, you need to store fine, detailed, nanoscale data on these customers and use big data analytics such as data mining or machine learning to see the detailed portrait (Talend, 2020).

The size of big data and the processing power required to analyze it, the open source community and computers have led to major advances in many analytical fields such as image recognition, speech recognition, medical diagnosis and gaming. While it makes statistical data easier to understand and interpret, machine learning includes algorithmic methods that enable machines to solve problems without specific computer programming. Therefore, machine learning pioneers predictive modeling tasks (Li et al., 2018; Alanazi et al., 2017). That is, machine learning is a discipline in computer science where machines (i.e. computers) are programmed to process input data, learn or train the basic model, and finally elaborate insights into new data. Thus, machine learning is essentially a certain class of artificial intelligence that includes learning mechanisms (such as a deep neural network) that can be properly trained in finding complex patterns within Big Data and transforming inputs with large data sets. Machine learning can analyze large-scale data from different environments and combine them to make some predictions (Rajkomar et al., 2019).

Deep learning and Big Data are the two newest trends in the rapidly growing digital world. Deep learning machine learning and pattern recognition are also actively used. Deep learning has achieved great success in many application areas such as speech recognition, image recognition and natural language processing. Today, Big Data contains great opportunities and transformative potential for various industries. On the other hand, there are unprecedented challenges to using data and information. As data grows, deep learning begins to play an important role in delivering analytical solutions based on Big Data prediction (Chen and Lin). Deep learning algorithms are used to exploit the predictive power of Big Data in fields such as search engines, medicine, and astronomy. Deep learning and Big Data are jointly recognized as "two major trends that are key drivers for the future growth of the American economy." (Crego et al., 2013). Unlike most traditional learning techniques, deep learning is controlled in deep architectures to automatically learn hierarchical representations. or unsupervised learning methods (Yu & Deng, 2010). Deep architectures can capture statistical input patterns that are often initiated more complex, hierarchically, to achieve adaptability to new areas compared to traditional learning methods (Arel et al., 2010).

How have the initially negative views of big data companies towards the open source movement, which gained a different momentum with the Big Data phenomenon, changed? Our study, which will deal with these and similar questions, aims to examine how the big data companies we mentioned look at the open source movement and how these companies interact with the open source movement. Following an introduction to the article, the history of the open source movement will be briefly summarized, and we will try to take a look at the interaction between the big data companies and the open source movement these days. The study is complemented by an evaluation.

3. A Brief History of the Open Source Movement

Richard Stallman, born in 1953, began hacking regularly at the university in 1971 (Williams, 2002: 49). Stallman said, “In my beliefs, freedom is more important than technological progress. I always prefer a less advanced free program to a more advanced but nonfree program. I wouldn't give up my freedom for something like this. My rule is that if I cannot share, I will take it ”(Williams, 2002: 116).

While working at the MIT Artificial Intelligence Laboratory in 1980, Richard Stallman was having trouble with printers that malfunctioned and left the job without completing the desired output. To solve this problem, Stallman wrote a few lines of code, asking the printer to notify those who had the job that the job was finished or that the printer was unable to complete the output (malfunction). However, such a change in the source code of the Xerox laser-printer was not possible for free or for money.

It is possible to take the issue back to the past, but the open source movement first started in 1984, when Richard Stallman, an MIT researcher and formerly a hacker, started a movement called the "free software movement" and started the free operating system GNU and profit It started with the founding of the non-purpose Free Software Foundation. Richard Stallman also introduced a type of license called “General Public License” (GPL) provided that copying, modifying, upgrading, modifying the source code is under the same license as the original code (Friedman, 2007: 106).

Later, Linus Torvalds, a 21 year old student at the University of Helsinki in 1991, developed an operating system called Linux as open source software to compete with Microsoft's operating system, Windows. The development of the GNU/Linux operating system under the same license as other people's improvements also contributed to this movement. Developed by 12 thousand people, Linux today has 18 million lines of source code, and millions of people worldwide use Linux in their Android smartphones, scientific studies and data centers (Athens, 2016).

Linus Torvalds wrote, "This business has been developing since April and is nearing completion. I would like to know what features people want in the software. I like the suggestions, but I cannot promise to implement them :-)" (Athens, 2016). Torvalds wasn't a social person, “I didn't start Linux as a collaboration project, I started it for myself. I wanted results, but I also enjoyed programming. I had no intention of using the open source code method while making the project public. I just wanted to comment on my work "(Asay, 2016). He didn't really like other people, but he liked other people who were interested in his project. According to Torvalds, open source was a great way to work together and produce code (Asay, 2016). According to him, all good programmers did programming because it was fun, not to earn money or to be overly praised by the public (BrainyQuote, 2020).

In fact, the open source movement meant much more than what words mean. Linux has proven to the world that working together, volunteering, and a management style based on merit can produce powerful products (Kepes, 2013). So how has the open data movement developed? It is a commonly held view that hacking is at the root of both the open source movement and the open data movement. According to the hacking culture: Computers can create a better world, everyone should be able to freely access computer technology and information, every hacker should be evaluated according to the code they write, and merit should be at the forefront. However, with all of this, it is debatable whether anyone who develops code is doing it for fun, like Linus Torvalds.

Although a software developer like Torvalds was later rewarded with stocks by companies, employees in these projects work for free and are not rewarded with stocks at the end of the job. The question of why software developers working in open source projects do this job for free has been asked in the literature and the answers to this question have been sought. Although there are various answers such as interest, entertainment, and learning curiosity, the most logical of these are the expectation of career advancement in the future and ego pleasure. Some authors call this "signaling incentive" (Tirole and Lerner, 2000: 15).

Meeting at the Free Software Summit organized by Tim O'Reilly, the founder of O'Reilly Media, which has made significant contributions to the popularization of terms such as Web 2.0 and open source, open source advocates established an organization called Open Source Initiative (OSI) in 1998. The Free Software Foundation and Open Planet Initiative (OPI), founded by Richard Stallman, are today seen as the main schools of the free and open source software movement.

After the early 2000s, the interactions between the open source software movement, Big Data and open data have been increasing. At the leading edge of this tsunami, the combination of innovative business and technology trends promises us a smarter future based on open source software and cross-organizational collaboration duo called Open Data Science. Open Data Science is such a move that it uses open source tools to make data, analytics, and computing a connected ecosystem that works together (Chambers et al.2017).

There is an important point here. Being an open data pro and a Big Data pro is about transformations of knowledge and power. However, these are opposed to each other. On the one hand, the actions of the open data proponents focus on the use and distribution of data, on the other hand, they relate to Big Data, which is the quantification of social life. Big Data, on the other hand, is an approach that uses all kinds of digital chips people leave behind without realizing it. In short, it is as if someone wants to freely use and distribute data, while carrying wood to the quarry of Big Data (Baack, 2015).

4. Open Source Code Applications

Open source software can be big, important, and transformational. But since we live in a for-profit capitalist society, this does not kill traditional software vendors, licensed software is still winning (Kepes, 2013). The interest in open source software is explained by the "network effect". The more consumers use a particular product, the more valuable it becomes to them. The utility function increases with the increase in the number of people using this product. The emerging externalities are due to the network effect, which is the most important factor for end users of particular software, and this factor affects people's use of this software (Celińska, 2016).

OSS projects are carried out by hundreds or even thousands of volunteers. Jobs are not given, people take the jobs they want, and there is no detailed project plan. The following quote can give an idea about how open source projects are run:

> “The Apache Group (AG) is the informal organization of the people in the core group responsible for developing the Apache HTTP Server Project, consisting entirely of volunteers. None of these developers need to spend large periods on the project in a consistent or planned manner, the process is maintained in a decentralized manner and with asynchronous communications. AG uses an e-mail list to communicate with each person, and a minimal voting system is used to solve problems ”(Mockus et al., 2000).

Let's briefly talk about some open source applications:

4.1. GitHub

Source control management, as the name suggests, is about the tools to manage the source code of any project, making it easy to do business together and efficiently. Source control management provides tools for various code developers to collaborate and contribute. Developed by Linus Torvalds in 2005, open source GitHub is a project management system established to enable different people we know or do not know to support and participate in a project that we are currently under construction or are developing. GitHub is a social code sharing company and hosts 31 million open source projects developed by 12 million people (Anthes, 2016).

In recent years Microsoft's control system Source Depot could not support the size of Windows. Microsoft has become friendly to Linux, almost saying 'I love it'. As a result, Microsoft announced that they adapted to GitHub developed by Linux (Paul, 2017).

4.2. Open Source Applications in Mobile Communication

The android operating system, first Android Inc. It was developed by a company called and acquired by Google in 2005. This system based on Linux is partially open source. iOS is Apple's mobile operating system. Android has 87% of the market and iOS has 12% (Moontechnolabs, 2018). In fact, both Google and Apple are profiting from this business. Apple sells expensive phones and has developed country markets such as the USA, Europe, and Australia. Google is dominant in the Asian and African mobile operating system market. In this way, Google enables more people to reach web services and search engines (Weinberger, 2016).

4.3. Open Source Applications in Cloud Computing

50% of the data centers in the world are located in the USA and China (Cimpanu, 2017). An increasing number of people do it instead of storing information on their computers, but on servers, they can access via the Internet. Millions of people carry their data and work to sites owned by Google (Johnson, 2008).

The interesting thing about cloud computing is that we redefine everything we did before as cloud computing. The computer world is under the influence of even more fashion than the fashion that women are interested in (Johnson, 2008). According to Richard Stallman, cloud computing is a trap and this trap will force people to buy more licensed systems, costing them more over time. “This stupidity is worse than stupidity; it's a marketing hype ”(Johnson, 2008). Hackers' goal is to change the world through software. They despised bad software, academic bureaucracy, and selfish behavior (Williams, 2002: 46).

Even stranger, they began to compete on an open source basis, still run by Amazon, Microsoft, and Google. No, it is not open code in the sense of "exporting kernel codes". “Although cloud services are based on open source code, they are generally licensed,” says Doug Cutting[††]. If you ask me, "in general" means "always" (Asay, 2017).

Open source software, specific to Azure, Microsoft speaks the language of love, deciding it should appeal to developers (Asay, 2017). Although all major cloud operators appear to compete on an open source basis on the surface, this is not the case in reality (Asay, 2017). Since the cloud business is not just a code business, what happens is those giant companies such as Amazon, Microsoft and Google use the "open source movement" for marketing purposes to different degrees.

4.4. Open Source Applications in Data Science: R and Python

R and Python, two open source software languages, are among the languages that are frequently used today, especially in data science applications. R is an implementation of the S language developed at Bell Labs and was put into use in 1995 after six years of preparation by Ross Ihaka and Robert Gentleman. Python, inspired by languages such as C, Modulo-3 and ABC, was developed by Guido Von Rossum in 1991. Both languages have a large number of packages and libraries that can be used in applications. Users close to statistics prefer R, and those close to engineering prefer Python. The design and development of R is managed by the R-core Group and the R Foundation. For Python, the Python Software Foundation does the same task. While R focuses on data analysis, statistics, and graphical models, Python is more concerned with code readability and efficiency (Tutorial, 2020).

5. How Do Big Data Companies View the Open Source Movement?

The example of Microsoft is interesting in understanding the value of the open source and free software today. In the early 2000s, big software companies started to perceive free software as a threat. Steve Ballmer, the former CEO of Microsoft, described Linux as cancer, referring to its free license in 2001 (Smith, 2001). Big firms used to say, "Who can trust these long-haired anarchist guys?" they were asking. Those who developed the programs said, "Why should I give the program I have written with great effort to these guys?" they said. However, things have changed over time. The same Steve Ballmer said to Microsoft CEO Satya Nadella, 15 years later, this time in December 2016, when Microsoft announced that SQL servers would run on Linux, ”(Kawamoto, 2016).

The rapid development of technology in the cyber field has made legal regulations in this field far behind. The Internet environment, which reminds us of the days of the Wild West attack, first experiences technological developments, but legal regulations cannot keep up with these developments. This phenomenon affects almost all of the events and relationships in cyber space. The open source movement can also be interpreted as filling the gap in cyberspace.

Derek Weeks, vice president of software security firm Sonatype, says that "interest in open source software has grown in recent years to meet the growing demand for new technology." For this reason, system developers produce more code in a shorter time. Researchers at Sonatype estimate that 80% to 90% of every modern application developed consists of open source components. Although open source development is so effective, it can also pose serious security problems. In a recent survey, 25% of developers in the government and industry say their organizations have experienced a security breach of over 70% compared to 2014. Emile Monette, a cyber supply chain expert at the Cyber Security and Infrastructure Security Agency, says "Most open source developers focus more on functionality than security and do not keep track of the latest vulnerabilities and updates." Therefore, it is difficult for agencies to know if the software they are purchasing has an open source vulnerability overlooked by the vendor. According to Weeks; no one can write great code, and all code everywhere is likely to have a vulnerability, whether it's an open source component or from scratch (Corrigan, 2019).

When we want to share a photo on Facebook, an infrastructure that stores and publishes this photo as data does this on Facebook. But does Facebook do all these things, or does Facebook compile and compile the work done by some volunteers to get them done? The correct answer is the second, and Facebook does not even thank these volunteers (Eghbal, 2016: 19).

In 2015, Google opened the source code for TensorFlow, a set of tools developed for deep learning applications. "He hoped that a community of students and hobbyists would be created and these people would contribute," he told Google. By the end of the 2000s, big data companies had begun to learn how to benefit from open source software. For example, some modules of NodeXL, which was developed as an open source for drawing and analysis of networks, started to be sold by Microsoft for money.

[††] Author of Hadoop codes.

In 2016, the official federal source code policy was determined and published for the first time in the USA (Scott, 2016). With this policy, it was aimed to ensure compliance of the public sector with open source applications. Some authors see this as an important triumph of the open source movement (Finley, 2016). Today, open source software is used almost everywhere, from our phones to our cars. On the other hand, large companies are increasingly interested in open source software. Walmart has implemented a cloud computing management system, the Exxon Mobil developer toolkit, and these are all open source software. London Stock Exchange Group, JP Morgan and Wells Fargo support Hyperledger (Finley, 2016).

Craig Mundie, Vice President of Microsoft, described the sharing of source code with the public domain 15 years ago in 2002 as "unhealthy". At that time, “History has shown us that although such models have a place; this model did not reveal successful software that is widely accessible to consumers, easy to use, powerful and in the mass market ”(Charny, 2002). The media, accustomed to hearing these from people at the highest levels of Microsoft, naturally described this as "the last step in Microsoft's protracted public relations campaign, which was conducted within the framework of the war with the open source code movement (Charny, 2002).

Someone who developed an open source software in the eyes of big data companies was, after all, a resume who applied for the job and spoke money no matter what the market. According to the information obtained from a survey, the percentage of companies using open source software in their operations almost doubled from 42% to 78% from 2010 to 2015, and the percentage of those who contributed to open source projects is% in 2014-2015. It increased from 50 to 64% (Anthes, 2016). Twitter works with thousands of Linux servers. It took and used the 13 million-line Linux kernel[‡‡] for Google Android, the world's largest open source project.

Since 1999, IBM has been supporting open source Linux operating systems and spending important financial, technical and marketing resources for the growth of Linux technology for the development of Linux technology. IBM's interest in Linux stems from its being a good operating system as well as providing an attractive total cost for IBM customers. For example, IBM spent one billion dollars on the development of FOSS (Free Open Source Software) and they did its public relations campaign with the slogan "Peace, Love and Linux" (Söderberg, 2008: 5). 10% of productive code developers working in FOSS projects write 72% of the codes (Söderberg, 2008: 28). IBM is also included in the Apache community, which develops open source Internet technologies, and contributes regularly and effectively to this community. Professor William Scherlis of Carnegie Mellon University says, “A hierarchical structure that is generally controlled in large and successful projects such as Apache and Eclipse; There are detailed ownership and governance structures in The Apache Foundation and The Eclipse Foundation (Anthes, 2016).

Software development today is a complex business. Communities that develop open source code can have large and complex relationships. There are two groups in the open source software business. The first of these is the "community" that will write open source software, the second is the "companies" that are interested in the results of this project and are in competition. The community and companies, in other words, the stakeholders of the project, form an ecosystem consisting of units in interaction that will produce software and services for a common market. It is clear that, for both the community and companies, it is important how cooperation will work and how the business will align with the policies and expectations of stakeholders (Gonzalez-Barahona et al., 2012). Projects are often communicated with non-formal open forums, topic audiences, and e-mailing lists, and the stakeholders of the project (Linåker et al., 2016).

In a study that quantitatively analyzed the Apache Hadoop ecosystem, stakeholders of this project were classified and ranked as follows (Linåker et al., 2016):

- *Infrastructure provider:* Sells infrastructure based on Apache Hadoop (Wandisco).
- *Platform user:* Uses Apache Hadoop in data storage and processing (Baidu, Xiaomi, eBay, Twitter, Yahoo).
- *Product provider:* Sells Apache Hadoop packaged products (Cloudera, Hortonworks, Huawei).
- *Product supporter:* Provides support for Apache Hadoop without a product vendor (Intel).
- *Service provider:* Sells related services to Apache Hadoop (Altiscale, Microsoft, NTT Data).

Again in the same study, the cooperation between stakeholders was examined with the measures of network science and it was found that the network has weak connections, low density; It has been determined that the most important forces that sustain the business in the Apache Hadoop ecosystem are the product suppliers (Cloudera, Hortonworks, Huawei) (Linåker et al., 2016).

Although open source software is useful, it also contains some risks. Open source software is available in numerous online repositories. Therefore, developers are not able to know the quality or security level. When companies do not invest in open source software, they run into a number of risks. As a result, they have to pay extra to eliminate these risks.

The risks posed by open source software are:

- Most software developers work with the idea that they can access open source software for free and use it unlimitedly. But this is wrong. All open source software works subject to their own licenses and some restrictions. Technology developed using GPL, LGPL, AGPL, CDDL, MPL and Open SSL can result in the whole project being open source license. The developer of this software will be known as the owner of the software. But its developer will not be known as a single person. This software will then become "de facto open source software". This concept is known as "copyleft" license (Escrowlondon, 2018).
- Open source software has no legal obligations to security and there may not be a community that can help you maintain security. Software developers are usually not security experts. For this reason, software developers

[‡‡] A kernel is a piece of software that provides a layer between hardware and software on a computer.

may not have enough knowledge about how to implement best practices (Infocyte, 2019).

- As of 2019, there are more than 200 license types. Most of these licenses are not compatible with each other. As the number of components increases, it becomes difficult to track and compare all license terms (Infocyte, 2019).
- If a company is good at buying, using open source code in technology could negatively affect it. If the buyer breaches open source licensing obligations, the code may fall into the hands of the public and the company's competitors (Escrowlondon, 2018).
- When open source code is open, hackers and malicious users will be able to access this code and find vulnerabilities within the open source code (Escrowlondon, 2018).
- If software developers instead of integrating all components; If they copy and paste sections of code from open source software, they can inadvertently increase some risk. When this is done, it will be impossible to trace the code for security (Infocyte, 2019).

6. Conclusions and Recommendations

Among the approaches that add value to big data, open source software is of great importance. In the long run, this software will support developments in subjects such as the Internet of Things (IoT) and Industry 4.0. Quantum computers and cloud computing services to be provided on this subject will also be discussed with open source software more and more every day. For example, a company called D-Wave opened Qbsolv, its basic quantum computing software, as open source software in early 2017 (Jackson, 2018). Once a die-hard foe of open source, Microsoft's love for open source has outstripped even Google today (Asay, 2017).

In October 2015, Amazon said Elastic copied a free software tool that they use to search and analyze data and sell them as a paid service. Within a year, Amazon started making more money from Elasticsearch developed by Elastic, making it easier for people to use the tool with its other offerings. That's why Elastic added premium features to Elasticsearch and limited what companies like Amazon can do. But Amazon still copied most of these features and continued to use them for free. Elastic Amazon sued in federal court in California for infringing its trademark. "Misleading consumers," Amazon said. Amazon denied this was anything wrong. The trial continues as of 2019 "(Wakabayashi, 2019). In addition, Amazon uses Amazon Web Services to copy and integrate software developed by other big data companies. Amazon combines discounts by embedding competing offers to make their products cheaper and gives their service an advantage. While these actions allow customers to buy more from Amazon, software developers may not gain any financial gain. On the other hand, despite all these negativities, small companies have no choice but to work with Amazon. Given Amazon's wide reach with customers, startups often accept some restrictions on promoting their products. In addition, small companies voluntarily share customer and product information with Amazon and pay Amazon a certain fee to sell via Amazon Web Services (Wakabayashi, 2019).

"According to GitHub, the top ten projects on its platform have thousands of individual contributors… but mostly these are run by large tech companies. Is this a problem for technology?" (Swanner, 2019) It is undoubtedly a problem. In the meantime, let us remind you that Microsoft owns GitHub. Swift, the software language developed by Apple and made open source in 2015, is growing rapidly. But few realize that Apple has patented features for this language. Swift is indeed open and free, but features of the language are patented… (Swanner, 2019).

Let us return to Richard Stallman's sentence quoted at the very beginning of our article: "All rights are not reserved". When we examine the examples of organized movements supported by the power of big data companies, unfortunately, this sentence is doomed to remain as a sentence smelling romance. Big data companies know very well how the efforts of those who develop open source software with great good will to monetize their products, and they abuse their goodwill efforts.

Today, the power of big data companies is accepted by everyone. These companies are very valuable, highly organized and perhaps the best managed companies in the world. The love of open source software of these companies we mentioned continues increasingly and the work has come to the software to be produced for quantum computers. We do not know if it is possible to say that this is a different kind of exploitation. However, we can at least qualify this relationship as "abuse". When we are asked to predict the future, we can say the following: We can't think that the unorganized, well-intentioned and sharing software developers who take part in the open source code movement against FAMA (Facebook, Amazon, Microsoft, Alphabet) have the slightest chance.

As a result, whether the open source movement will continue to deliver innovation for years to come is still a controversial issue. Given the current speed, the open source movement has, it is probably not wise to bet against it. However, the challenges customers face in integrating open source projects and dealing with the increased complexity cannot be easily set aside. On the other hand, there seems to be no alternative for companies considering using Big Data Technology to gain a competitive advantage. Because open source software should be used regardless of the complexity. "Open source has increasingly become a very important part of technology at Bloomberg," said Gideon Mann, Bloomberg's head of data science, to Datanami last year. "Nowadays it is not possible to stay competitive without using open source, so there are a lot of open sources we use" (Woodie, 2018). There is a virtuous cycle at the heart of the success of open source. As open source software gets better, it attracts more users, which makes the product even better over time (Woodie, 2018). In essence, being open is a fundamental vision of how the world should work. However, in a world working with more and more software and cloud services, the question faced by businesses everywhere is now, "Should we be open or closed?" so far it has become clear in most places, but "How clear should we be?" (Robinson, 2019) of course, that is another question that needs to be discussed.

References

Alanazi, H. O., Abdullah, A. H., & Qureshi, K. N. (2017). A Critical Review for Developing Accurate and Dynamic Predictive Models Using Machine Learning Methods in Medicine and Health Care. *Journal of Medical Systems*, *41*(4), 69.

Anthes, G. (2016). Open Source Software No Longer Optional. *Communications of the ACM,* 59(8), 15-17.

Arel, I., Rose, D. C., & Karnowski, T. P. (2010). Deep Machine Learning-a New Frontier in Artificial Intelligence Research [research frontier]. *IEEE Computational Intelligence Magazine*, *5*(4), 13-18.

Asay, M. (2016). Why Linux Creator Linus Torvalds Doesn't Really Care About Open Source. 2016, February 22. Retrieved from https://www.techrepublic.com/article/linux-creator-linus-torvalds-doesnt-really-care-about-open-source/. 12.03.2020.

Asay, M. (2017, March 30). BIG Open-Source Love Microsoft and Google? You Still Won't Catch AWS. Retrieved from https://www.theregister.co.uk/2017/03/30/doe_open_sourcey_ness_in_cloud_matter/. 10.03.2020.

Baack, S. (2015). Datafication and Empowerment: How the Open Data Movement Rearticulates Notions of Democracy, Participation, and Journalism. *Big Data & Society*, *2*(2), 2053951715594634.

BrainyQuote (2020). Linus Torvalds Quotes. Retrieved from https://www.brainyquote.com/quotes/linus_torvalds_137861. 12.03.2020.

Carillo, K., & Okoli, C. (2008). The Open Source Movement: a revolution in software development. Journal of Computer Information Systems, 49(2), 1-9.

Celińska, D. (2016). Why Do Users Choose Open Source Software? Analysis of the Network Effect. *Informatyka Ekonomiczna*, *39*(1), 9-22.

Chambers, M., Doig, C., & Stokes-Rees, I. (2017). Breaking Data Science Open. O'Reilly Media, Incorporated.

Charny, B. (2002, January 2). Microsoft Raps Open Source Approach. Retrieved from https://www.cnet.com/news/microsoft-raps-open-source-approach/. 12.05.2020.

Chen, X. W., & Lin, X. (2014). Big Data Deep Learning: Challenges and Perspectives. *IEEE access*, *2*, 514-525.

Cimpanu, C. (2017, June 2). Hadoop Servers Expose over 5 Petabytes of Data. Retrieved from https://www.bleepingcomputer.com/news/security/hadoop-servers-expose-over-5-petabytes-of-data/. 12.05.2020.

Corrigan, J. (2019). How do agencies make sure the crowdsourced code that underlies nearly every piece of tech on the market is safe to use? Retrieved from https://www.nextgov.com/cybersecurity/2019/05/inside-governments-open-source-software-conundrum/157186/. 25.12.2020.

Crego, E., Munoz, G. & Islam, F. (2013, July 26). Big Data and Deep Learning: Big Deals or Big Delusions? Retrieved from https://www.huffpost.com/entry/big-data-and-deep-learnin_b_3325352. 01.09.2020.

Çelik, S. (2018). Büyük Veri. *Gece Kitaplığı. Ankara*. ISBN: 978-605-288-811-7.

Eghbal, N. (2016). *Roads and Bridges: The Unseen Labor Behind Our Digital Infrastructure*. Ford Foundation.

Escrowlondon (2018). Top risks in using open source code in software development. Retrieved from https://www.escrowlondon.com/news/top-3-risks-using-open-source-code/

Finley, K. (2016, August 11). Open Source Won. So, Now What? Retrieved from https://www.wired.com/2016/08/open-source-won-now/. 12.05.2020.

Friedman, T. L. (2007). The World is Flat 3.0: A Brief History of the Twenty-first Century/Thomas L. Friedman. *NY.: Picador*.

Gonzalez-Barahona, J. M., Izquierdo-Cortazar, D., Maffulli, S., & Robles, G. (2012). Using Software Analytics to Understand How Companies Interact in Free Software Communities.

IBM (2013). IBM White Paper. Big Data for the Intelligence Community. Retrieved from https://pdfs.semanticscholar.org/6bff/f82a993eab399a84dd82081977a8e1fcba57.pdf. 12.05.2020.

Infocyte (2019). 7 Risks Posed by Open-Source Software and How to Defend Yourself https://www.infocyte.com/blog/2019/06/18/7-risks-posed-by-open-source-software-and-how-to-defend-yourself/

Jackson, M. (2018, December 19). Quantum Computing Progress Will Speed up Thanks to Open Sourcing. Retrieved from https://singularityhub.com/2017/01/28/quantum-computing-progress-will-speed-up-thanks-to-open-sourcing/. 20.06.2020.

Jani, K. (2016). The Promise and Prejudice of Big Data in Intelligence Community. *arXiv preprint arXiv:1610.08629*.

Johnson, B. (2008). Cloud Computing is a Trap, Warns GNU Founder Richard Stallman. *The Guardian*, *29*.

Kawamoto, D. (2016, March 11). Ballmer: Linux no Longer a Cancer. Retrieved from https://www.informationweek.com/software/ballmer-linux-no-longer-a-cancer--/d/d-id/1324661. 13.07.2020.

Kepes, B. (2013). Open Source is Good and All, But Proprietary is Still Winning.

Li, Y., Wu, F. X., & Ngom, A. (2018). A Review on Machine Learning Principles for Multi-view Biological Data Integration. *Briefings in Bioinformatics*, *19*(2), 325-340.

Linåker, J., Rempel, P., Regnell, B., & Mäder, P. (2016, March). How Firms Adapt and Interact in Open Source Ecosystems: Analyzing Stakeholder Influence and Collaboration Patterns. In *International Working Conference on Requirements Engineering: Foundation for Software Quality* (pp. 63-81). Springer, Cham.

Manyika, J., et. al. (2011). Big data: The Next Frontier for Innovation, Competition, and Productivity. McKinsey Global Institute. Retrieved from http://www.mckinsey.com/Insights/MGI/Research/Technology_and_Innovation/Big_data_The_next_frontier_for_innovation. 10.10.2020.

McAfee, A., Brynjolfsson, E., Davenport, T. H., Patil, D. J., & Barton, D. (2012). Big Data: The Management Revolution. *Harvard Business Review*, *90*(10), 60-68.

Mockus, A., Fielding, R. T., & Herbsleb, J. (2000, June). A Case Study of Open Source Software Development: The Apache Server. In Proceedings of the 22nd International Conference on Software Engineering (pp. 263-272).

Moontechnolabs (2018, September 14). Apple vs Android - A Comparative Study 2017. Retrieved from https://www.moontechnolabs.com/apple-vs-android-comparative-study-2017/. 22.07.2020.

Najafabadi, M.M., Villanustre, F., Khoshgoftaar, T.M. et al. (2015). Deep Learning Applications and Challenges in Big Data Analytics. Journal of Big Data 2. https://doi.org/10.1186/s40537-014-0007-7.

Paul, J. (2017). Microsoft is Now Using Linus Torvalds' Open Source Tool for Windows Development. Retrieved from https://itsfoss.com/microsoft-using-git/. 20.07.2020.

Piatetsky, G. (2017). KDnuggets, Machine Learning Overtaking Big Data?. Retrieved from https://www.kdnuggets.com/2017/05/machine-learning-overtaking-big-data.html. 22.07.2020.

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine Learning in Medicine. *New England Journal of Medicine*, *380*(14), 1347-1358.

Robinson, S. (2019, January 30). Open Source Technology Will Influence the Future of Cloud. Retrieved from https://www.businessinsider.com/sc/open-source-technology-future-of-cloud-2019-1. 06.09.2020.

Robles, G., González-Barahona, J. M., Izquierdo-Cortazar, D., & Herraiz, I. (2009). Tools for the Study of the Usual Data Sources Found in Libre Software Projects. *International Journal of Open Source Software and Processes (IJOSSP)*, *1*(1), 24-45.

Scott, T. (2016). (United States Chief Information Officer), Rung E. Anne United States Chief Acquisition Officer, “M-16-21 Memorandum for the Heads of Departments and Agencies”, https://sourcecode.cio.gov/. 12.02.2020.

Smith, JT. (2001, June 1). Microsoft's Ballmer: Linux is a Cancer. 1 June 2001. Retrieved from https://www.linux.com/news/microsofts-ballmer-linux-cancer. 25.02.2020.

Söderberg, J. (2008). Hacking Capitalism: The Free and Open Source Software Movement.

Swanner, N. (2019, August 8). Big Tech Controls Many Major Open Source Projects. Is that a Problem? Retrieved from https://insights.dice.com/2019/08/05/open-source-google-microsoft-apple-github/. 12.04.2020.

Talend (2020). What is Big Data? [Free guide & definition]. (2020, July 22). Retrieved from https://www.talend.com/resources/future-big-data/. 05.09.2020.

Tirole, J., & Lerner, J. (2000). The Simple Economics of Open Source.

Tutorial (2020). Choosing Python or R for Data Analysis? An Infographic. Retrieved from https://www.datacamp.com/community/tutorials/r-or-python-for-data-analysis#gs.G2t5njE. 15.06.2020.

Wakabayashi, D. (2019, December 15). Prime Leverage: How Amazon Wields Power in the Technology World. Retrieved from https://www.nytimes.com/2019/12/15/technology/amazon-aws-cloud-competition.html. 17.05.2020.

Weinberger, M. (2016, December 21). The Whole 'Mac vs. PC' Thing is so Over, and 'Android vs. iPhone' is Close Behind. Retrieved from https://www.businessinsider.com/apple-mac-vs-microsoft-windows-pc-is-over-2016-12. 18.06.2020.

Williams, S. (2002). Free as in Freedom (2.0)-Richard Stallman and the Free Software Revolution. *Boston: The Free Software Foundation*.

Woodie, A. (2018). Weighing Open Source's Worth for the Future of Big Data. (2018, February 26). Retrieved from https://www.datanami.com/2018/02/26/weighing-open-sources-worth-future-big-data/

Yu, D., & Deng, L. (2010). Deep Learning and its Applications to Signal and Information Processing [exploratory dsp]. *IEEE Signal Processing Magazine*, *28*(1), 145-154.

Zuech, R., Khoshgoftaar, T.M. & Wald, R. Intrusion Detection and Big Heterogeneous Data: a Survey. Journal of Big Data 2, 3 (2015). https://doi.org/10.1186/s40537-015-0013-4.