Electronic Letters on Science & Engineering 16(2) (2020)
Available online at dergipark.org.tr/else

A Hadoop Application for Urban Computing in Smart City

Sevcan TURAN[1*], Halit ÖZTEKİN[2]

[1]Çanakkale Onsekiz Mart University, Çan Vocational School; sevcanturan@comu.edu.tr; 0000-0003-4278-7406
[2]Sakarya University of Applied Sciences; halitoztekin@subu.edu.tr; 0000-0001-8598-4763

Abstract: Data analytics, which is the process of obtaining meaningful information from the rapidly occurring big data, has taken its place among popular topics. According to smart transportation systems, which are part of smart cities, big data is also generated in the field of transportation. Within the scope of this study, a data analytics application will be carried out with the average speed and vehicle count data obtained by the sensor systems installed on the roads for the City Pulse Smart City project which is accessible through the internet. Using the Hadoop open-source framework software, which is popular in data analytics, graphics and information will be obtained for interpretation about the flow of traffic on the road.

Keywords: Data Analytics, Big Data, Hadoop, Intelligent Transportation Systems

Akıllı Şehirde Kentsel Bilgi İşlem için Hadoop Uygulaması

Özet: Hızla oluşan büyük veriden anlamlı bilgi elde etme süreci olan veri analitiği, popüler konular arasında yerini almıştır. Akıllı şehirlerin bir parçası olan akıllı ulaşım sistemleriyle birlikte ulaşım alanında da büyük veri üretilmektedir. Bu çalışma kapsamında, internet üzerinden erişilebilen City Pulse Smart City projesi için yollara kurulan sensör sistemleri ile elde edilen ortalama hız ve araç sayısı verileriyle veri analizi uygulaması yapılacaktır. Veri analitiğinde popüler olan Hadoop açık kaynak çerçeve yazılımı kullanılarak, yoldaki trafik akışının yorumlanması için grafikler ve bilgiler elde edilecektir.

Anahtar Kelimeler: Veri Analitiği, Hadoop, Akıllı Ulaşım Sistemleri

Reference to this paper should be made as follows (bu makaleye aşağıdaki şekilde atıfta bulunulmalı):
Sevcan, T., Halit, Ö., 'An Hadoop Application for Urban Computing in Smart City', Elec Lett Sci Eng , vol. 16(2) , (2020), 193-201

1. Introduction

With the spread of the Internet, everyone and everything has become able to produce information at any moment and big data has entered our lives. Big data; In addition to the structural data produced in the traditional way, it is defined as the data that contains different structures and contains content such as text, image, and video that do not have common features to be kept with traditional database systems [1-3]. According to the statistics published on Statista's web page, it is estimated that big data, which had a volume of 2 (two) zetabytes in 2010, would be 175 zetabays in 2025 [4].

Obtaining information that is meaningful from the rapidly formed big data, which will guide the business world and guide the planning for the future is called data analytics [2]. For example, information shared in social media, instant messages, etc. By monitoring the users' preferences, product presentations are made accordingly, the data in the field of health can be determined and the outbreaks that can occur and the measures to be taken against these outbreaks can be

determined and plans for the future in the electrical infrastructure can be determined by the data collected from the smart grid systems [5].

Tools are being developed to perform data analytics operations. One of them is open source Apache Hadoop framework software, and the framework mainly includes Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop MapReduce, Hadoop YARN [6]. Hadoop accessibility (can work on cloud systems or large computer clusters), robustness (can work on computer clusters, tolerate any computer error), scalability (allows development by creating new nodes in computer clusters by tracking the growth of data) and simplicity (parallel coding It is preferred by both academic researchers and industry because of its features [7]. For example, Çiftçi and Ertuğrul have proposed a Hadoop MapReduce based search engine service for food and additive information search applications to be run on mobile devices. In search engine service, the data is taken from websites and recorded on Hadoop HDFS and analyzed by Mapping / Reduce operations. When the user performs a query with his mobile application, the information is returned in the ready databases, if not, or the results are processed by searching the websites and the answer is returned to the user [8]. Yılmazel has run algorithms such as mean, median, peak value, variance, standard deviation, maximum / minimum, skewness, and quarters, which are frequently used in statistics science, on big data. The accuracy of the results produced by the system was checked against the results obtained with the R programming language and it was reported that the same results were obtained. The effect of adding new nodes to the system on the speed of the algorithm has been examined and it has been observed that adding a new node reduces the processing time [9].

One of the application areas of data analytics is smart transportation systems. Intelligent transportation systems: highlights areas such as driving safety, preventing congestion, eliminating accident risks, saving time by finding the ideal route, and protecting the environment with less energy consumption [10]. To reach these goals, data is tried to be collected from all movements in traffic. With the increase in internet applications of objects, instantly generated data in the field of transportation has increased. Information about the big data reaching the level of Petabayt in the field of transportation; from smart cards, sensors, social media, mobile electronic system integration cameras, transportation vehicles, mobile phones etc. can be obtained [3]. Data analytics in smart transportation systems, according to Li Zhu et al. It can be used to gather, and process data generated by different systems, to control and manage traffic instantly, to detect accidents and emergencies instantly and to take necessary steps [5].

Studies on data analytics in smart transportation systems are coming to the fore. Ying et al. In their work, they addressed the issue of finding the best route for trucks to transport wastes from construction and demolition processes. In this study, depending on the risk ratio to be taken into consideration, the best route alternatives are offered to provide city security and reduce the accident rate [11]. In their work, Andy and his colleagues have developed software called BIGSEA, which consists of many applications, to provide safe and comfortable transportation to the city. This software uses big data and data analytics; it offers end users the most ideal route, city administrators, bus usage trends to manage the public transportation system, transportation information for passengers who will use public transportation, detection of potential traffic jams, and mood analysis on traffic or possible accidents through comments on social media [12]. In their study, Alfredo et al. Developed a data analytics infrastructure to provide solutions to the problem of transport of truck containers of logistics companies in smart cities. Their aim is to minimize unloaded journeys with high costs. For this, an optimization study has been made with the genetic algorithm by making use of the information obtained from previous transports [13]. In their studies, Darwish and Bakar presented a fog-based real-time data analytics approach for vehicles connected to the Internet. Due to the slowness of cloud computing, considering the fact that decision-making processes against traffic instantaneous events may be delayed, the

architecture of making fast real-time transactions with fog computing has been described [10]. In their study, Robert and his colleagues conducted an analysis of the traffic situation in China, about ways to obtain big data and how data analytics can be used in transportation systems [14].

In this study, using the traffic data in the City Pulse Smart City [15] open data set, an example of data analytics application with Hadoop distributed file system will be realized. Our goal is to graphically express the speed distributions of vehicles passing through a point.

The next steps of the study; In the second part, the basic components of Hadoop architecture, in the third part, our application of data analytics will be in the form of a result / discussion in the fourth part.

2. Hadoop Components

Hadoop is a decentralized system with master / slave structure, and clusters can be created from servers with different features, and different tasks and powers can be given to servers [7]. Thus, lower cost servers are used together instead of high cost servers, and it is possible to create a low cost system for data analytics. For example, Aysan and Özbilgin created cloud with single card computers with Unix operating system on them and used the Hadoop system. They showed that Hadoop can be used with single card computers that have low cost, small physical structure, and low energy consumption [16]. Hadoop has three types of setups, local, fake distributed and fully distributed. Yavuz et al. Reported in detail the operations to be performed for the fully distributed installation and operation of Hadoop basic components using three servers running Debian 5.0.3 operating system [17].

2.1 Hadoop Distributed File System (HDFS)

HDFS is the file system that enables large files to be stored in a manner suitable for distributed data processing systems [7]. Thanks to HDFS, low-performance servers are run together to create a virtual disk and store and process big data.

In HDFS, the files are stored by dividing them into 128 MB blocks by default, and the size of the blocks can be changed depending on the desired file size depending on the intended purpose [18]. Backups of the files are created to prevent any data loss; the default number of backups is 3 [18]. As can be seen in Figure 1, file operations in HDFS consist of two basic parts: NameNode and DataNode. NameNode is the main unit (master) of the system and is responsible for keeping file information such as the name of the files, number of copies, where the files are saved, and file operations such as saving, deleting and creating backups of the files. There is one NameNode in each server cluster. DataNode, on the other hand, are servers that are slaves and where data is stored on their physical disks and file operations are performed. When there is a change in the information of the files they store, DataNodes send their new status to NameNode and update the information [7].

2.2 Hadoop MapReduce

MapReduce is used for parallel information processing on a distributed server system. The information given in the login file is sent to the DataNodes in the server cluster. In each server, the data is first transformed into <key, value> pairs in the “map (matching)” stage and filtered and the number of matching maps is obtained. From this output, the values belonging to the same key are combined and lists are created and exchanged between servers. As a result of the relocation, each server performs a "reduce" on the <key, value> pairs sent to it, and results from the values with the same key [6-7].

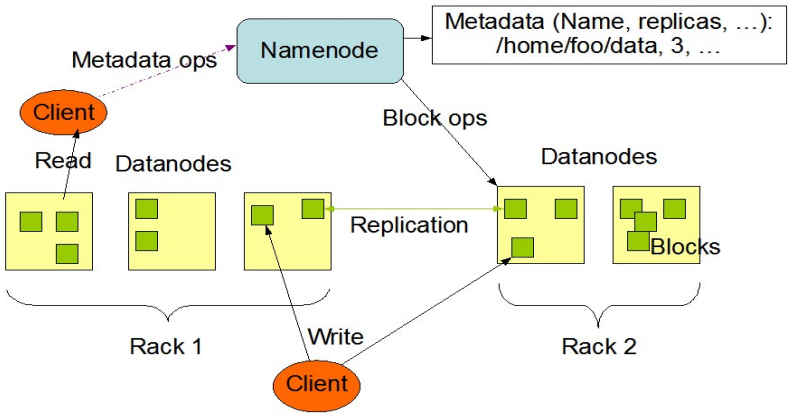

Figure 1. HDFS Architecture

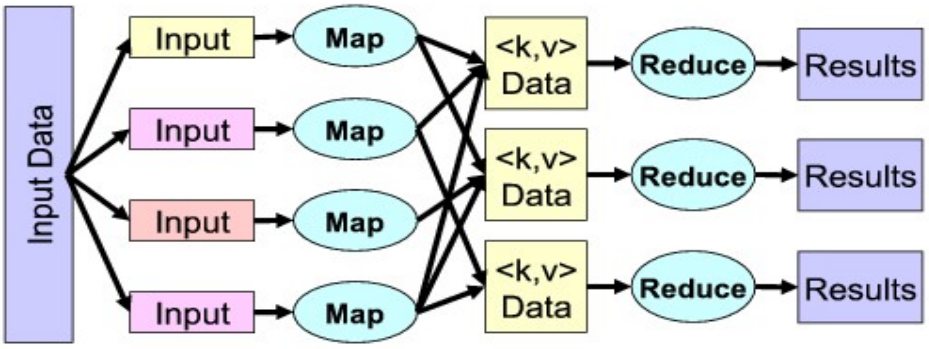

Figure 2. Map-Reduce Architecture

2.3. Hadoop YARN

It is the component responsible for managing resources and business processes. The yarn: There is a basic resource manager that manages the use of resources in the system by all client processes, and node managers that perform monitoring operations of the node's CPU, memory, and network usage in each node. AppMaster is created for each request, and the transactions of the clients are followed. MapReduce works based on YARN [19].

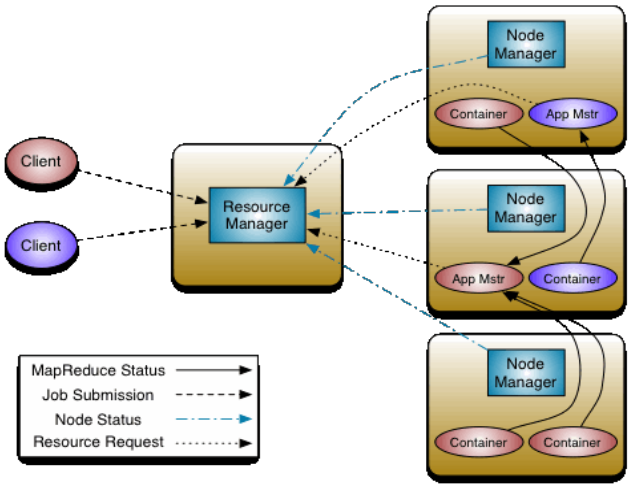

Figure 3. YARN Architecture[19]

3. Case Study

In our study, traffic information data from City Pulse Smart City open data set was used. As traffic data; The vehicles passing between the coordinate point (56.22579478256016, 10.116589665412903) on the city of Arhusvej, Hinnerup, Denmark (56.23172069428216, 10.104986076057457) and 1030 meters away from the coordinate point; The file in which the average speed, average measurement time, time of recording of the data, and the number of vehicles passed was recorded in the timeframe of February 2 - June 9, 2014. There are 32076 records in the file.

In our study, Hadoop 3.1.3 version was used as a single cluster-single node (Standalone). Our node computer has Windows 10 Professional installed, with 16 GB Ram and Intel® Core ™ i7-8750H 2.21 GHz processor. Our application was developed using the Eclipse editor in Java language. The data set has been saved to the Hadoop HDFS distributed file system. On the data set, there are Hadoop MapReduce operations, how many vehicles pass on the basis of the speed limits we set on the traffic data, the average speed and the number of vehicles at the time of the day. The results obtained as a result of matching-reduction processes are presented as pie and column graphs.

During Hadoop MapReduce operations on the data set, the speed is divided into 5 groups as 0-39, 40-59, 60-79, 80-99, and more than 100. During pairing, the speed range was taken as the key and the number of vehicles was taken as the value. During the reduction, the total number of vehicles was found for vehicles within the same speed range. The algorithm is given the below.

Algortihm 1. Speed Range- Vehicle Count

BEGIN

1- Create job
2- Set job input, output folder
3- Set job mapper, reducer properties
4- Start Job
   a. Map key, value pair

While mapper has new line from file
i) Split value from mapper
ii) Get velocity (V) and vehicle count (C) from value
iii) ***If*** *V>=0 and V <40 then Set <key,value> = (0-39 , C)*
else if *V >=40 and V<60 then Set <key,value> = (40-59 , C)*
else if *V>=60 and V <80 then Set <key,value> = (60-79 , C)*
else if V>=80 and V<100 then Set <key,value> = (80-99 , C)*
else *Set <key,value> = (100-500, C)*

   b. Shuffle key, value pair
   c. Reduce key, value pair

While reducer has new Key-Value Pair
While value has more element
i) Read value as y
ii) Sum += y
Set <key, value> = (key from reducer, sum)

   d. Save reduced result to HDFS
5- Visualize results

END

Table 1 lists the results obtained as a result of Mapping-Reduction processes. The results are shown graphically in Figure 4. Then, the mapping was done by taking the time information in the data as the key, the speed and the number of vehicles as the value.

Table 1. The Total Number of Vehicles by Speed Range

Travel Speed(kmph)	No. of speed observations	Rate (%)
0-39	10781	7,81
40-59	83976	60,83
60-79	42358	30,68
80-99	883	0,64
100+	51	0,04

As can be seen in Figure 4, there is generally smooth traffic on the road. 91.51% of the vehicles can move at an average speed of 40 to 79 km / h. There are not many vehicles traveling over 80 km / h speed. In general, speed limits are complied with. Looking at the bar graph and Table 2, it is seen that the hours that pass most of the vehicles on the road are at 5, 6 and 7 in the morning, and the average speeds during these hours are lower than other hours. It may be thought that traffic planners should make a decision by paying attention to these hours. In addition, it is seen that the average speeds are the highest in the hours with the lowest number of vehicles and it can be concluded that these hours should be paid attention to control the speed limits.

When our speed range - number of vehicles application runs on Hadoop, the statistical information about the time elapsed for the thread and the processed data are as in Table 2. As it can be seen in Table 2, there are 32075 records in our file, 32075 key-value pairs were created during matching, 5 key-value pairs were obtained from these key-value pairs during the reduction phase.

Table 2. The Average Speed Range- Total Number of Vehicles Matching / Reduction Operation Statistics

File System Counters
FILE: Number of bytes read=67, FILE: Number of bytes written=438455
FILE: Number of read operations=0, FILE: Number of large read operations=0
FILE: Number of write operations=0
HDFS: Number of bytes read=1746653, HDFS: Number of bytes written=56
HDFS: Number of read operations=8, HDFS: Number of large read operations=0
HDFS: Number of write operations=2
Job Counters
Launched map tasks=1, Launched reduce tasks=1, Data-local map tasks=1
Map-Reduce Framework
Map input records=32075, Map output records=32075, Map output bytes=320138
Map output materialized bytes=67, Input split bytes=90, Combine input records=***32075***
Combine output records=***5***, Reduce input groups=***5***, Reduce shuffle bytes=67
Reduce input records=***5***, Reduce output records=***5***, Spilled Records=10, Shuffled Maps =1
Failed Shuffles=0, Merged Map outputs=1

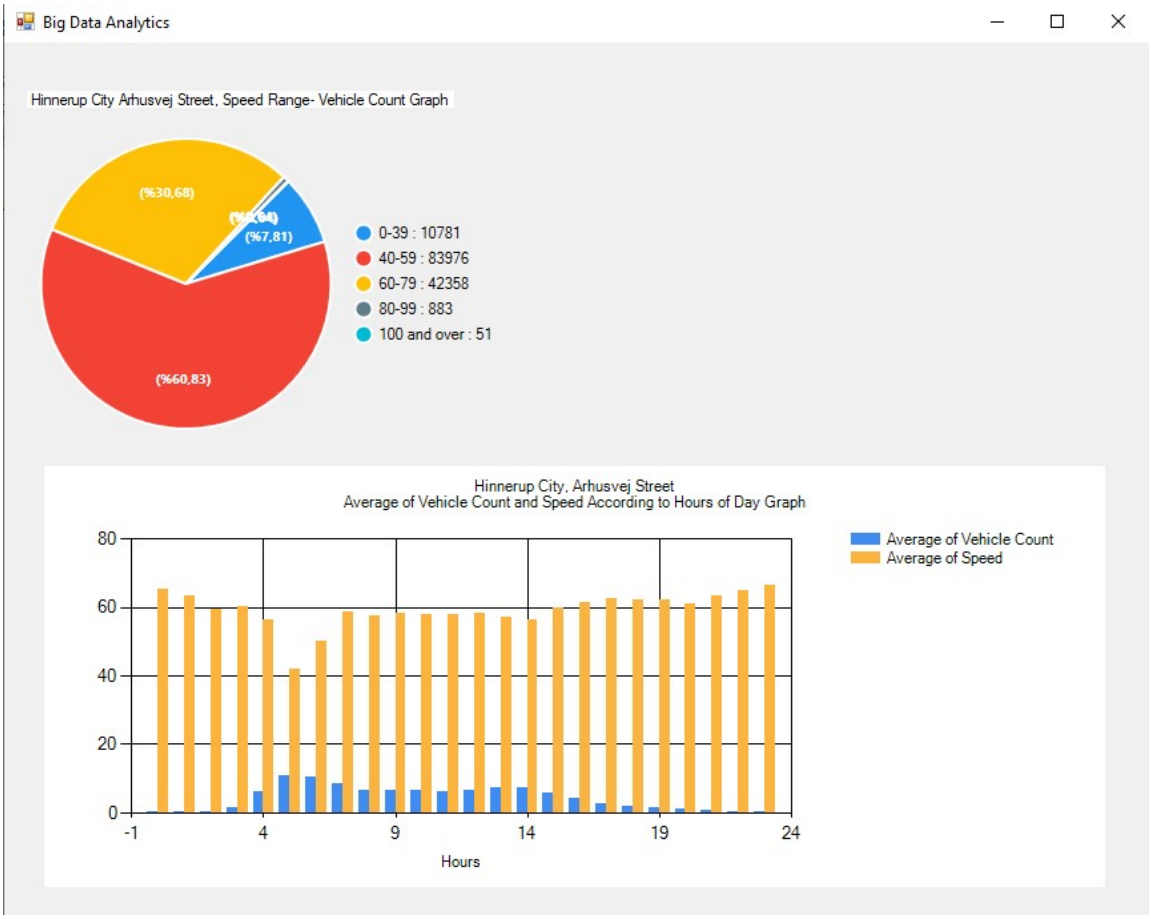

Figure 4. Graphical Display of Data Obtained from Map-Reduce

4. Discussion and Conclusions

Big data; business world, management decisions of countries etc. It is an important resource used to give direction. Data analytics applications come to the fore in the processing of big data. One of the areas where big data begins to form is the transportation systems in the cities. Studies are carried out to facilitate the control of transportation systems of smart cities. In this study, we have implemented a sample matching-reduction application with a single node on the Windows operating system with Hadoop using the City Pulse Smart City open data set. With our application, we obtained how many vehicles passed through the speed ranges we determined from the street we selected. Using this information, it can be found how many vehicles do not meet speed limits. Also, with our application, we obtained how many vehicles passed in the speed ranges we determined from the street we selected and at what hours the vehicle density was higher. Using this information, it can be found how many vehicles do not comply with the speed limits and at what hours traffic jams may occur.

Our next work may be to develop an application to monitor and manage real transportation information and traffic conditions in our country by installing the Hadoop system on distributed nodes.

Acknowledgement

The authors would like to thank the Open Data Aarhus (ODAA) initiative and CityPulse EU FP7 Project for publishing open datasets.

Compliance with Ethical Standards

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

Conflict of Interest

Authors have no conflict of interest.

References

[1] Aktan, E., "Büyük Veri: Uygulama Alanları, Analitiği ve Güvenlik Boyutu", *Bilgi Yönetimi(2018)*, 1-22.

[2] Cuzzocrea, A., Song, I.-Y., & Davis, K., "Analytics over Large-Scale Multidimensional Data:The Big Data Revolution", *DOLAP '11: Proceedings of the ACM 14th international workshop on Data Warehousing and OLAP(2011),* 101–104.

[3] Chen, H., Chiang, R., & Storey, V., "Business Intelligence and Analytics: From Big Data to Big Impact", *Business Intelligence Research*, Vol. 36, No. 4 , 1165-1188.

[4] O'Dea, S., "Volume of data/information created worldwide from 2010 to 2025", https://www.statista.com/statistics/871513/worldwide-data-created/ [Accessed: 28-Feb-2020].

[5] Zhu, L., Yu, F. R., Wang, Y., Ning, B.,& Tang, T., "Big Data Analytics in Intelligent Transportation Systems: A Survey", *IEEE Transactions on Intelligent Transportation System(2019)*, vol 20, 383-398.

[6] Apache Hadoop, Apache Software Foundation: https://hadoop.apache.org/ [Accessed: 03-Apr-2020].

[7] Lam, C., "Hadoop in Action", *Stamford: Manning Publications Co(2011)*.

[8] Çitçi M. A., Çelik E. D., "Hadoop ve Mapreduce Teknolojisi aracılığıyla Gıda-tabanlı Mobil Uygulamaları için bir Arama Hizmeti", *Cumhuriyet Üniversitesi Fen Fakültesi Fen Bilimleri Dergisi(2017)*, 38: 79-94.

[9] Yılmazel Ö., "Hadoop Üzerinde Ölçeklenebilir Betimleyici İstatistik Uygulamaları", *Nicel Bilimler Dergisi(2019)*. 1:43-58.

[10] Darwish T. S. J., Bakar K. A., "Fog Based Intelligent Transportation Big Data Analytics in The Internet of Vehicles Environment: Motivations, Architecture, Challenges, and Critical Issues", *IEEE Access, Access, IEEE(2018)*, 6:15679-15701.

[11] Qiu, Y., Zhao, X., Zhang, X., "Optimal Routing for Safe Construction and Demolition Waste Transportation:A CVaR Criterion and Big Data Analytics Approach", *Technical Gazette(2019),* 26, 1128-1135.

[12] Alic, A.S., Almeida J., Aloisio, G., "BIGSEA: A Big Data analytics platform for public transportation", *Future Generation Computer Systems(2019)*, 243-269.

[13] Cuzzocrea, A., Nolich, M., Ukovich, W., "A Big-Data-Analytics Framework for Supporting Logistics Problems in Smart-City Environments", *Procedia Computer Science(2019)*, 159:2589–2597.

[14] Biuk-Aghai, R.P., Kou, W. T., Fong, S., "Big Data Analytics for Transportation: Problems and Prospects for its Application in China", *IEEE Region 10 Symposium (TENSYMP)(2019)*, 173-178.

[15] Kolozali, S., Bermudez-Edo, M., Puschmann, D., Ganz, F. and Barnaghi, P., "A Knowledge-Based Approach for Real-Time IoT Data Stream Annotation and Processing", *2014 IEEE International Conference on Internet of Things (iThings), and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social*

Computing (CPSCom), Taipei. 215-222.

[16] Aysan L., Özbilgin İ.G., "Tek Kart Bilgisayarlar ile Bulut Oluşturarak MapReduce İşlemleri Denemesi", *Bilişim Teknolojileri Dergisi(2015),* 8:179-191.

[17] Yavuz G., Aytekin S., Akçay M., "Apache Hadoop ve Dağıtık Sistemler Üzerindeki Rolü", *Dumlupınar Üniversitesi Fen Bilimleri Enstitüsü Dergisi(2012)*, 27: 43-54.

[18] HDFS Default. Apache Hadoop: https://hadoop.apache.org/docs/r2.4.1/hadoop-project-dist/hadoop-hdfs/hdfs-default.xml. [Accessed: 25-Mar-2020].

[19] YARN. Apache Hadoop: https://hadoop.apache.org/docs/stable/hadoop-yarn/hadoop-yarn-site/YARN.html. [Accessed: 26-Mar-2020].