European Journal of Technique

journal homepage: https://dergipark.org.tr/en/pub/ejt

Vol.13, No.2, 2023

Research Article

Detection of Shadow IT Incidents for Centralized IT Management in Enterprises using Statistical and Machine Learning Algorithms

Mücahit Kutsal[1*], Bihter Daş[2], Ziya Aşkar[3], Ali Necdet Güvercin[4], Resul Daş[5]

[1*] University of Gdańsk, Institute of Theoretical Physics and Astrophysics, 80-308, Gdańsk, Poland. (m.kutsal.423@studms.ug.edu.pl).
[2] Firat University, Software Engineering Department, 23119, Elazig, Turkey. (bihterdas@firat.edu.tr).
[3] Arçelik A.Ş. Karaagac Caddesi 2-6, Sutluce Beyoğlu 34445 Istanbul, Turkiye. (ziya.askar@arcelik.com).
[4] Arçelik A.Ş. Karaagac Caddesi 2-6, Sutluce Beyoğlu 34445 Istanbul, Turkiye. (alinecdet.guvercin@arcelik.com).
[5] Firat University, Software Engineering Department, 23119, Elazig, Turkey. (rdas@firat.edu.tr).

ARTICLE INFO

Received: Oct., 30. 2023
Revised: Nov., 26 2023
Accepted: Nov, 27. 2023

Keywords:
Shadow IT
Centralized IT Management
Enterprise IT
IT Monitoring and Control

Corresponding author: *Mücahit Kutsal*

ISSN: *2536-5010* | e-ISSN: *2536-5134*

DOI: https://doi.org/10.36222/ejt.1382461

ABSTRACT

Software as a Service (SaaS) is a software service where software solutions are offered to users via the internet, usually subscription-based or sometimes opened to access by selling a license key, distributed over the cloud, and updates are automatically delivered to users because they are distributed over the cloud. The number of SaaS provider companies is increasing day by day, and with this increase, unauthorized purchase of SaaS applications has become a problem for corporate-sized companies. Without the company's approval, SaaS software and hardware used by employees increase Shadow IT which means there is a potential risk of security breaches, data loss, and compliance issues as the IT department is unaware of the usage and unable to monitor and control the systems effectively. In this study, in order to avoid the problems that may be caused by Shadow IT, unauthorized SaaS applications in Arçelik Global have been detected by utilizing statistical and machine learning approaches. In the experiment, Interquartile Range, K-Means and Stabilization algorithms were used for the detection of unauthorized SaaS applications. Using all three algorithms, low, medium and high-risk shadow IT detection was made for Arçelik company. We see that the proposed stabilization approach explores unauthorized SaaS applications much more distinctively than the other two algorithms. The proposed approach can be used in the future to detect unauthorized software from other companies.

1. INTRODUCTION

Today, free, flexible, cloud-based information technology (IT) applications and services that are easily accessible over the internet are increasing rapidly [1]. In order to work more efficiently, people can use these IT applications or services from home, at work, or while traveling on personal laptops, tablets, and smartphones. While the use of these IT technologies provides an advantage in accelerating the digital transformation, on the other hand, it causes the employees to turn to IT offers without the approval of the organization they work for, in short, to increase the shadow IT [2,3]. Shadow IT is the usage of information technology systems, devices, software, applications, and services without the open approval of the IT department. The fact that software as a service (SaaS) applications are easily accessible over the internet and can offer different solutions for different problems is very attractive to users [4-6]. This situation causes different departments to purchase SaaS applications without the knowledge of the IT department, in other words the emergence of Shadow IT. Shadow IT plays a critical role in both security and financial investment in companies. Employees' use of some features such as file sharing, storage, and collaboration may result in the leakage of sensitive data [7]. The increase in the use of mobile devices in business environments, the use of wearable devices, that is, bring your own (BYO) devices, causes new forms of devices to enter businesses and all these devices to offer different operating system types. BYO devices create an increasingly heterogeneous and difficult space for IT to manage. This trend often referred to as "shadow IT", creates a significant security risk for companies [8,9]. In addition, IT has little visibility into the implications of corporate data in cloud environments. As a result, it becomes very difficult to manage BYO devices that contain both company-provided applications and personally owned applications [10,11]. With the increasing use of shadow IT, it becomes more difficult for organizations to adjust to legal or contractual IT regulations. The reliability and accuracy of big data analytics are weakening because shadow IT increases in unknown corporate data sources. All these difficulties add great importance to the studies on shadow IT detection. Another concept, IT consumerization, refers to the transportation of software and hardware products designed for personal use to the organization and their use for business purposes. Companies can provide consumption IT or enterprise IT to their employees [12,13]. Enterprise IT is a concept that includes enterprise services and support, as well as their strategy, management, budgets, and policy. Figure 1 outlines the distinction between shadow IT and current concepts [14,15].

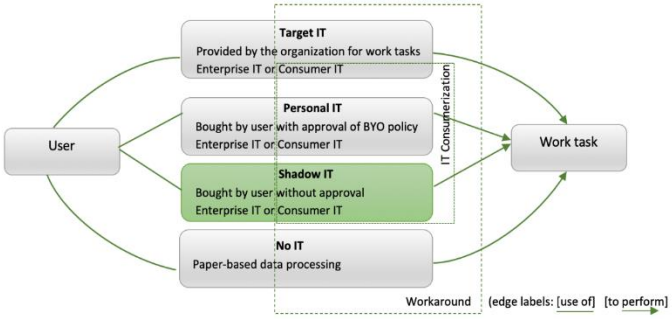

Figure 1. Shadow IT and closely related concepts [1]

1.1. Current Problems with Shadow IT and Motivation.

In order to keep security at the highest level in a company, the IT department must be aware of every software used in the company. In this way, it can be decided whether the software used in the company is safe or not. If there is a lack of control over the software used in an organization, this will cause the organization to experience different security vulnerabilities. One of the main problems that Shadow IT can cause is the loss of important internal data. An unauthorized SaaS application used can take data that should not go out of the organization by backing up/storing it without permission, and if this confidential data is in the hands of malicious people, a cyber security breach may occur within the organization. Therefore, an unauthorized SaaS application will cause a legal violation. An unauthorized SaaS application may also be in conflict with an agreement that the organization has to comply with. In order not to be exposed to all these problems, the IT department should have full authority over the SaaS applications used in the organization. Another problem that we may encounter within the organization as a result of Shadow IT is the updates for SaaS applications because SaaS applications are centrally distributed on the cloud, users do not have a choice whether to accept updates to their SaaS applications. Therefore, the IT department of the organization should be prepared to avoid errors, legal incompatibilities, and security vulnerabilities that may be caused by a new update to a SaaS application, and for this, the IT department should be aware of all the SaaS applications that the organization uses. In addition, there may be updates on the pricing of the SaaS applications used over time, and as the institution becomes dependent on this SaaS application, it has to comply with the new pricing, which means a financial risk for the institution. Finally, another problem created by Shadow IT is efficiency. A SaaS application may lose its efficiency over time, and put its customers in trouble, or a SaaS application in use may lose its ability to meet the demands of the organization. In order for the organization to cope with all these problems, shadow IT detection must be done correctly. In this study, unauthorized SaaS software currently used in Arçelik Global company in Turkey has been identified, with a suggested approach to avoid the problems that may be caused by Shadow IT in a company. In addition, in the future, a detection system has been developed in the company for SaaS applications that may be used without permission.

1.2. Main Contributions

The main contributions of this paper are outlined below:

- Shadow IT detection was performed using statistical and machine learning algorithms on firewall log data in a corporate-sized company.
- A novel shadow IT detection system was developed that works on the data flowing over the firewall.
- The number of research articles on shadow IT detection in the literature is quite low. This study will shed light on future academic studies for shadow IT detection.

1.3. Novelty.

Although the importance of Shadow IT in the business world has increased recently, it still does not receive the necessary attention in the academic world. Even if there are articles in which the concept of shadow IT is mentioned, there are very little experiments on shadow IT detection. As far as we know, there is no research paper in the literature that distinguishes shadow IT detection as low risk, medium risk and high risk. In this study, a new approach for detection of shadow IT was proposed and the performance of this approach was compared with existing methods and divided according to risk groups. Therefore, we believe that this study will contribute to the scientific world in detecting shadow IT.

The remainder of this paper is organized as follows. Section 2 presents related works and general information about K-Means and Interquartile algorithms used for shadow IT detection. Section 3 includes details of the materials and implementation. Section 4 presents the experimental results and discussion. Section 5 presents the conclusion.

2. RELATED WORK

This section examines the state of the art on the topic of shadow IT such as detection of shadow IT, and review papers. Although the importance of Shadow IT has increased a lot in the business world, there are not many scientific studies on this subject. There are hardly any articles on scanning Shadow IT. As far as we know, our study will be the first study to detect Shadow IT. Therefore, in the literature review section, the most considerable studies in which the term Shadow IT is used are mentioned. In the studies that have been cited as [16,17], it is mentioned that Shadow IT increases the security vulnerability in companies and damages the organizational data flow. In [18], 129 best shadow applications were classified as excel macro, cloud solutions, software, ERP, business intelligence systems, hardware, websites. In [19], the authors also explained other system concepts such as rogue IT and shadow, workaround and feral, which are closely related to this concept, apart from shadow IT, and stated the differences between these concepts and shadow IT. In [20,21] studies, the authors interpret shadow IT as an insider threat because a non-malicious employee is installing unapproved software. They also stated that it was caused by the company's employees not complying with the information security policies. In the studies referenced as [22,23], the authors argued that Shadow IT systems are more efficient and effective in practice than the existing official and standard systems used in companies. In [24,25], the authors discussed the concept of shadow IT in determining the relationship between business and IT. They also explored the role of social media software and self-made macros in Excel or Access software for shadow IT. In [26-28], the difficulties experienced in detecting shadow IT with increasing new technologies were mentioned. These studies stated that there are very few academic publications on shadow IT detection, which is largely due to the difficulty of accessing the data and the fact that they are unregistered. In [29], the authors did a study on the relationship of shadow IT with security. An evaluation model for shadow ITs in a company is presented in [30]. A sample shadow assessment document is presented based on several weighted and evaluated criteria. In [31], the authors conducted a review article on shadow IT.

The acquisition and preprocessing of log files from servers are of paramount importance in the realm of cybersecurity, system performance monitoring, error analysis, and overall information security management. Log files are text documents that meticulously record events in a system, and they are employed for monitoring activities, detecting errors, identifying cybersecurity

threats, and defending against security breaches. Initially, the preprocessing stage of log files obtained from servers involves transforming this data into meaningful and usable information [32, 33]. This stage encompasses steps such as cleaning up unnecessary information, organizing log records, and standardizing them into a consistent format. This process enables log data to be more effectively analyzed and interpreted. The cleaning and analysis of log files form the cornerstone of an organization's information security strategies. For instance, analyses performed on log files to detect abnormal activities can identify potential security breaches in advance, allowing for preventive measures to be taken. Moreover, the utilization of log files for monitoring system performance and error analysis facilitates the swift identification and resolution of issues. Therefore, the routine collection of log files from servers plays a critical role in maintaining a secure network and ensuring the healthy operation of information systems [34, 35].

2.1. k-Means Algorithm.

The K-Means algorithm is a clustering algorithm that tries to divide the dataset into groups it belongs to and is an unsupervised learning [36]. It makes elements within the same group as similar as possible, but tries to keep the clusters apart. The k value in K-Means determines the number of clusters. The algorithm has a simple working logic. Assigns data points to a cluster so that the sum of the squared distance between the data points and the cluster's centroid is minimal. This is how it calculates that the data belongs to that cluster. While calculating the centroid of the clusters, it takes the arithmetic average of all data points belonging to that cluster. The fact that the data in the clusters is similar to each other is determined by the fact that the data has little variation [37,38]. Since the K-Means algorithm uses distance-based measurement to determine the similarity between data points, normalization is recommended to determine which cluster the data belongs to. Because of K-Means is an iterative algorithm and centers are randomly started, using different centroids could provide a solution when K-Means stays at a local optimum and not converge to the global optimum.

2.2. Interquartile Range Algorithm.

Quartile range, which is widely used in statistics, is the rule of dividing an ordered data set into four equal parts, each consisting of quarter data. In this rule, the middle half (50%) of the data is represented as medians, while the third quartile (25%) and first quartile (25%) are represented as quartiles. The interquartile range (IQR) is calculated by subtracting the first quartile from the third quartile. The interquartile range indicates the range where most of the data values are found. The interquartile range is also expressed by Equation 1.

$$IQR = Q3 - Q1 \quad (1)$$

In this equation; Q1 is a number between the smallest number and the median of the data set. Q2 is the median of the data. Q3 is a number between the median and the highest value of the data set. The quartiles gap is preferred as a statistical measure of spread since it is not affected by extremely small or extremely large extremes in the sorted data. If there is a very extreme outlier in the dataset, the quartile range may be preferred. While IQR is used to identify outliers in a data set, it indicates where most of the data is. It also shows the central trend of the data [39, 40]. Figure 2 shows the graphical representation of the interquartile range.

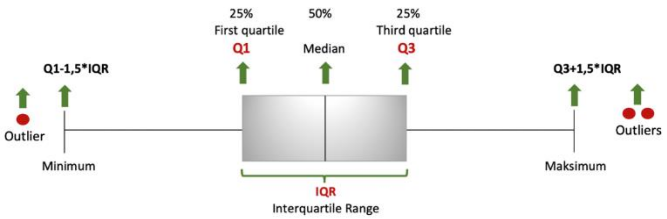

Figure 2. The graphical representation of IQR

3. MATERIALS AND METHODS

In this section, the steps of the application carried out for shadow IT detection to detect unauthorized SaaS software within Arçelik Global company are explained in detail. The Shadow IT detection system consists of 3 main parts: (i) data set collection (ii) data preprocessing (iii) data analysis. Figure 3 shows the flowchart of the proposed shadow IT detection approach.

Algorithm 1. The algorithm of the shadow IT detection model

Procedure: The shadow IT detection model
Input: Log data stored by IBM Qradar software **Output:** Low, medium, high risk SaaS applications
Step 1: Collecting log data **Step 2:** Extracting unnecessary columns from data stored in Qradar (Data refining) **Step 3:** Extracting the matching records by filtering the data **Step 4:** Converting the data to json format in Qradar and sending it to an API written in python. **Step 5:** Saving the data coming to Rest API directly to a collection in Cosmos DB before analysis as a daily-report collection **Step 6:** Processing of incoming data to Rest API at pre-process stage - data transformation for analysis - min-max normalization for K-Means algorithm - calculating date from timestamp value **Step 7:** Detection of low, medium, high risk anomalies by K-Means, IQR and Stabilization algorithms **Step 8:** Saving the detected anomalies to a collection in Azure Cosmos DB as anomalies collection

3.1. Data Collection.

The log data used as a data source in this study were stored with the IBM Qradar software, which was created with the Fortinet Firewall solution. These data are taboos and raw log data that are formed as a result of requests made out of Arçelik company. In terms of volume, terabytes of data are generated monthly. Table1 shows a small and anonymized example of the used tabular log data.

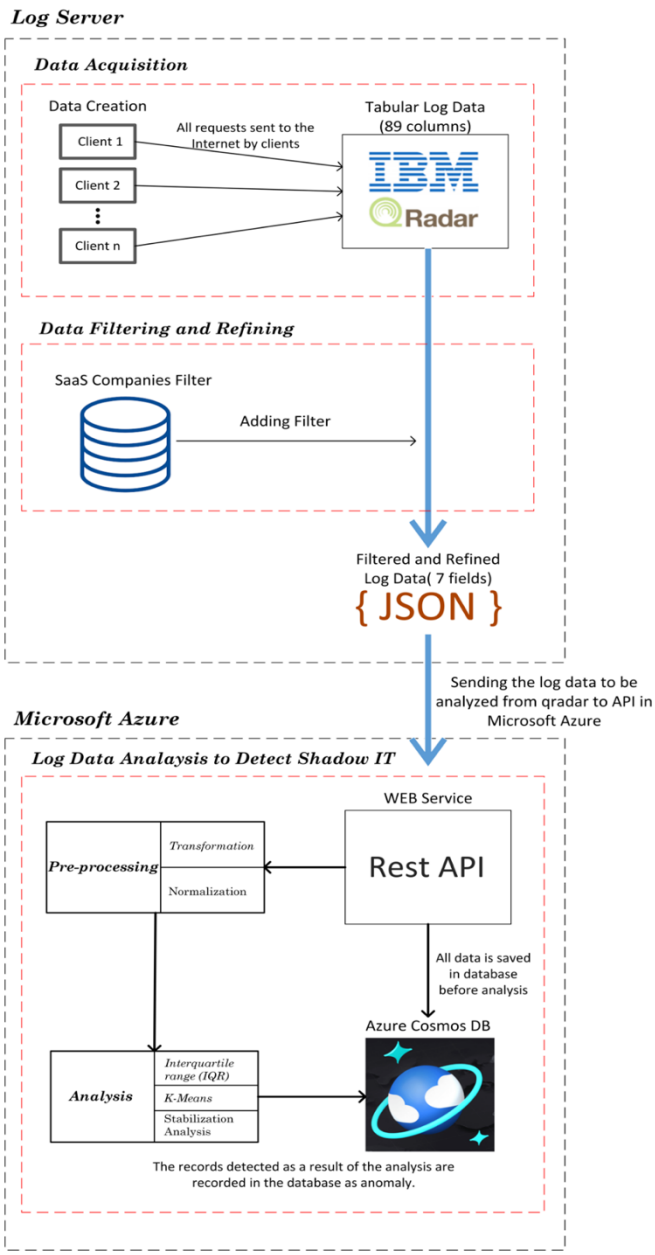

Figure 3. The shadow IT detection model

Table I
An example of tabular log data.

	Hostname	Request URL	Request Time	Source IP
1	e.g.1.com	/login	16410	10.92.*
2	e.g.2.com	/homepage	164114	10.134.*
3	e.g.3.com	/list/1	165064	10.86.*
4	e.g.4.com	/services/add	165134	10.134.*
5	e.g.5.com	/register	164624	10.92.*
6	e.g.6.com	/login	1647106	10.86.*
7	e.g.7.com	/livecall/s2	164736	10.17.*

Our log data kept in IBM Qradar software is in tabular format and is stored in a table called events. There are 89 columns in this table. The most important columns that can be used in the application are extracted in the Data refining section and are also used in the analysis section.

3.1. 1. Data Filtering and Refining

In this subsection, filtering and refining are performed on the monthly terabytes of data stored in the log server. In this study, we did not deal with all the data generated by Arçelik employees. We only used the requests sent by the employees of Arçelik company to the SaaS applications. Therefore, a filter was applied to extract the needed data, and after the filtering was completed, only the log data of the requests sent to the SaaS applications by Arçelik employees remained, and then the columns needed for this log data were extracted. Table 2 shows the event matrix after filtering and refining. The event matrix is the number of requests made by Arçelik employees to websites belonging to SaaS applications via their browsers. On the event matrix, each column corresponds to a SaaS company, and each row corresponds to a day. For example, a request made by any employee of Arçelik company to xyzsaas.com company on 01.01.2022 via browser increases the number in the relevant cell on the event matrix by one.

Table II
An example of event matrix.

Date	e.g.1.com	e.g.2.com	e.g.3.com	e.g.4.com
01.01. 2022	5	14	23	16
02.01.2022	16	25	34	47
03.01.2022	68	77	86	75
04.01.2022	91	99	83	3
05.01.2022	98	89	98	0
06.01.2022	53	62	30	41
07.01.2022	56	65	74	83
08.01.2022	87	96	11	7
09.01.2022	30	49	42	51

Then, we converted the data into json format and sent it to the web service running in the cloud environment. Figure 4 shows the json object that belongs to the requests sent by three Arçelik employees to the examplesaas.com application on a daily basis.


```
{
  "Start Time (Minimum)": 1655889723818,
  "hostname": "examplesaas.com",
  "hitcount": 96,
  "request_url": "examplesaas.com/examplepath",
  "source": [
    {
      "ip": "10.[redacted].60",
      "log_source": "ARCELIK_SM_FW_[redacted]",
      "hitcount": 23
    },
    {
      "ip": "10.[redacted].116",
      "log_source": "ARCELIK_SM_FW_[redacted]",
      "hitcount": 17
    },
    {
      "ip": "10.[redacted].173",
      "log_source": "ARCELIK_LogY_FW_[redacted]",
      "hitcount": 56
    }
  ]
}
```


Figure 4. Json Ocject Example for Requests of the Three Arcelik Employees.

3.2. Preprocessing

The data preprocessing is a crucial step in the data analysis pipeline, involving the cleaning and transformation of raw data into a format that is suitable for analysis. It is an essential phase that directly impacts the quality and effectiveness of subsequent data modeling or analytical processes. Preprocessing sub-stage performs the transformation and normalization of the data. Therefore, these processes make the data analyzable. The key steps involved in data preprocessing include:

3.2.1 Data Transformation

In this preprocessing subsection, a data transformation takes place. The Start time (Minimum) field in the data we obtained in json format represents a data in timestamp format. Date is calculated from this timestamp data. Date data is needed because we analyze according to days in the analysis phase. No other type of data transformation is needed in this part.

3.2.2 Normalization

Normalization is to treat the data in a single order in cases where the difference between the data is too great. By applying the normalization process, the number group is multiplied by a fixed number without disturbing the ratio between them, and by adding a fixed number, the numbers are taken into a certain range. Although there are different normalization methods, min-max normalization is a well-known and used method. In this method, the largest and smallest values in a group of data are considered. All other data is normalized to be 0 for the smallest value and 1 for the maximum value, and all data is spread over this 0-1 range. We applied min-max normalization to the data in the K-Means algorithm. Normalization is not performed in other algorithms.

3.3 Data Analysis

In this subsection, the use of 3 different algorithms for shadow IT detection is explained. k-Means and IQR algorithms are previously known algorithms used in data clustering in the literature. Here, these algorithms are not used for clustering, but for the detection of SaaS applications that can be purchased without the knowledge of the IT department. The other algorithm is an algorithm that we proposed and we call stabilization. The Stabilization algorithm is an algorithm that is used to detect the SaaS applications currently used in Arçelik Global, the constraints and steps are developed by us, and we derive from the Interquartile Range algorithm.

3.3.1 Analysis with k-means

We have previously mentioned that the K-Means algorithm is a distance-based unsupervised machine learning algorithm that tries to find k number of clusters in a data set, and that these clusters are as separate from each other, but the elements are tried to be kept as close as possible. In this study, we used the K-Means algorithm as an anomaly detection algorithm, not as a clustering algorithm. In our Shadow IT detection system, our cluster number is one and the anomalies we are trying to detect are events that are far from this cluster on the event matrix. Before an event occurred on the event matrix, we brought all the events back 30 days in the relevant column where that event would occur, and ran the K-Means algorithm on them. Since we made anomaly analysis separately for each column on the event matrix, this algorithm was applied separately for each column. Our aim is to determine a major cluster in each column on the event matrix and to determine the outlier values that are far from this cluster. In order to evaluate an event as an anomaly, the exact midpoint of the cluster is determined by the K-Means algorithm. Then, the Euclidean distance of all events is checked for this middle value. The Euclidean distance between the points p and q is given in Equation.

$$d(p, q)= \sqrt{p^2+q^2} \quad (2)$$

The following conditions were checked to rank the detected anomalies as low, medium and high.

i. The Euclidean distance is between 0.75-0.85, then low.

ii. The Euclidean distance is between 0.85-0.90, medium.

iii. The Euclidean distance is between 0.90-1, then high.

3.3.2 Analysis with IQR

The interquartile rule is a statistical algorithm that measures the distribution and variability of the data by dividing the dataset into quarters. In the IQR algorithm, first the median of the data set is taken and the data is divided into three parts as first quartile (Q1), second quartile (Q2) and third quartile (Q3). The value of Q2 is directly equal to the median itself. Then, the Q1 value was subtracted from the Q3 value to obtain the IQR value. We added the IQR value to the Q3 value to obtain the upper threshold value. The formula we used to obtain the upper threshold value is shown in Equation 3.

$$\text{Upper threshold} = Q3 + IQR*1.5 \quad (3)$$

In order to evaluate an event as an anomaly, we evaluated the values above the upper threshold as anomaly in the analyzed data set. In addition, the following conditions were checked in order to rank the anomalies detected here as low, medium and high.

i. The values > upper threshold then low risk.

ii. The values > (upper threshold) *2 then medium risk.

iii. The values > (upper threshold) *4 then high risk.

For this algorithm, we performed a 15-day backward analysis on the event matrix. We ran the IQR algorithm for each column on the event matrix.

3.3.2 Analysis with Our Approach

This proposed approach, unlike the other two algorithms, tries to detect continuity, not exactly an outlier anomaly on the data set. The basis of this proposed approach, which we named "Stabilization", is based on the IQR algorithm. For this algorithm, not the upper limit, but the lower limit and an additional tolerance value are used. In the proposed approach, unlike the algorithm above, for determining the lower limit, the IQR value is not added to the Q3 value, instead it is subtracted from the Q1 value. The formula used for the lower bound is shown in Equation 4.

$$\text{Lower threshold} = Q3 + IQR*1 \quad (4)$$

Here, in order to decide whether a new event will be marked as an anomaly, it is checked whether the lower limit has been exceeded. If a stable request has been sent to the said SaaS application from within the Arçelik company in the last 30 days, we have marked it as an anomaly. We used the tolerance value so that the Arçelik company is closed on holidays, holidays and similar special days and no requests are made to prevent this anomaly. As a result of such an analysis, we decided that if the number of data below the lower limit determined on the 30-day dataset is more than the

tolerance value, there is an anomaly in the dataset. In order to rank the anomalies detected here as low, medium and high, first the tolerance value was set as 10 and the following conditions were checked in turn.

i. High risk if the number of data below the lower limit is less than *((tolerance)/ 2^2)*
ii. If the number of data below the lower limit is less than *((tolerance)/ 2^1)*, medium risk.
iii. If the number of data below the lower limit is less than *((tolerance)/ 2^0)*, that is, less than the direct tolerance, low risk.

4. EXPERIMENTAL RESULTS AND DISCUSSION

The proposed approach for detecting Shadow IT is an algorithm based on the interquartile rule statistical algorithm, the rules, and constraints of which were developed by us. With our approach, which we call stabilization, the software used by Arçelik employees without the permission of the IT department has been identified. As in other algorithms, low, middle, and high-risk shadow IT detections were made. Table 3 shows the detection results of low, medium, and high-risk shadow ITs with three algorithms for Arçelik company.

Table III
Some of the detected shadow IT incidents

Domain	Algorithms	Count	Date	Importance
E.g.1	Stabilization	915	22-05-13	High
E.g.2	Stabilization	114	2022-05-13	High
E.g.3	Stabilization	112	2022-05-13	High
E.g.4	k-means	1340	2022-05-12	Low
E.g.5	k-means	5565	2022-05-13	High
E.g.6	k-means	2126	2022-05-13	Middle
E.g.7	IQR, Stabilization	119	2022-05-11	High
E.g.8	Stabilization	935	2022-05-09	High
E.g.9	Stabilization	132	2022-05-07	Middle
E.g.10.com	k-means	128	2022-05-13	High
E.g.11.com	IQR	81	2022-05-07	High

When K-Means and IQR algorithms were first to run to detect unauthorized SaaS software in Arcelik, it was seen that with these two algorithms, SaaS applications that did not only exist before but came into use suddenly could be detected. However, we needed to detect not only newly purchased applications but also previously purchased applications that have been in use for a long time. Based on this need, we developed the stabilization algorithm. After adding this algorithm to the system, the shadow IT detection approach we developed gained the ability to detect not only unauthorized software to be purchased after the day it was actively used, but also purchased and regularly used the software before the day it was used. According to the experimental results obtained, we could say that the stabilization approach detects shadow ITs much more than the other two algorithms and is a much more useful algorithm for the company.

After detecting the shadow IT detections at Arçelik as low, medium, and high risk, one of the biggest problems we encountered was Cross-Origin Resource Sharing (CORS). CORS is simply a web page making a request to another web page, that is, to a domain belonging to that application. For example, when an employee requests the example1.com SaaS application, it may be requesting example2.com as well, regardless of the application that is actually making the request in the background. This request is stored as a firewall log. This scenario means the following for shadow IT: Company employee actually buys and uses the "example1" application, but because the "example1" application uses the services of the "example2" application in the background, it is stored in firewall logs as if the employee also purchased and used the "example2" application. Therefore, it is possible that the "example2" application will be detected as shadow IT in Arçelik company by the shadow IT detection system we have developed. To avoid this, when the system detects an anomaly, we created a dictionary of the most common keywords in SaaS applications that we know to determine whether it is a background service used by an application or a purchased SaaS application. We looked at how many times the values in this dictionary were passed in the path of the requests. Figure 5 shows the most frequent keywords and their numbers for the "vimeo.com" web request.

Figure 5. Dictionary matches in request path.

In addition, by calculating how many times the subdomain values of the requested application are passed, we have created a different insight in order to determine whether the data we have is really a shadow IT. Figure 6 shows the number of subdomains in the requested web address.

Figure 6. Subdomain counts.

In order to comprehensively enrich our study and develop a more robust analysis, we conducted a review of similar research efforts. Our investigation revealed a limited body of literature addressing the identification of shadow IT incidents, with most studies primarily focused on defining the concept of shadow IT. Notably, the works of Silic et al. [7] and Rentrop et al. [30] provided valuable insights into guiding our approach for this study.

5. CONCLUSION

In our study, we discussed the detection of Shadow IT that may exist within an institution with statistical and machine learning methods. Throughout the study, we applied one statistical, one machine learning algorithm and a new approach based on statistic to the event matrix obtained from the firewall logs of the company enabling the detection of SaaS applications currently used in the organization. In addition, we developed a detection system to identify SaaS applications that may be purchased without the knowledge of the IT department in the future. Our study is ready to use in the future to determine the Shadow IT situations within different institutions and organizations. Based on this paper, in the future, when it is desired to detect Shadow IT within an institution,

it will be sufficient to apply the detection algorithms to this event matrix by establishing the data pipeline in which the relevant institution creates the event matrix.

Thanks to this end-to-end shadow-IT detection system we have developed, we have presented an effective solution in terms of manageable IT. Although the work we have presented is a general solution that can be applied by different institutions in the future, researchers who will apply to this study can contact the authors in the future as well. Based on this study, different shadow IT detection systems can be developed in the future using different algorithms (e.g. deep learning time series analysis). Within the scope of our study, we did not conduct a study on the actions that can be taken within the organization after a shadow IT incident is detected, but after a shadow IT incident is detected, actions such as legalizing this SaaS software within the organization or completely removing it from the processes within the organization should be taken, so studies can be carried out on the actions that can be taken against shadow IT incidents detected in the future.

ACKNOWLEDGEMENT

This work was supported by the Arçelik A.S. Digital Transformation, AI and Big Data Research and Development Center. The authors want to thank Emre Bahadır Sever (e-mail: bahadir.sever@arcelik.com) for his contribution and support.

REFERENCES

[1] Haag, S.; Eckhardt, A. Shadow IT. *Bus Inf Syst Eng*. 2017, vol. 59, no. 6, pp. 469–473, doi: 10.1007/s12599-017-0497-x.

[2] György A.; Cleven A.; Uebernickel F.; Brenner W. Exploring the shadows: IT governance approaches to user-driven innovation. *In: Proceedings of the 20th European Conference on Information Systems.* 2012, Barcelona.

[3] Segal M. Dealing with the realities of shadow IT. *In: Datacenter J*. http://www.datacenterjournal.com/dealing-realities-shadow/. Accessed 22 Nov. 2016.

[4] Brancheau J.C; Brown, C. The management of end-user computing: Status and Directions. *ACM Computing Surveys*, 1993, vol. 25, no. 4, pp. 437–482.

[5] Klotz, S.; Kopper, A.; Westner, M., Strahringer, S. Causing factors,outcomes, and governace of Shadow IT and business-managed IT: a systematic literature review. *International Journal of Information Systems and Project Management.* vol.7, no.1, 2019.

[6] Rentrop, C., Zimmermann, S. Shadow IT - Management and Control of Unofficial IT," *ICDS 2012: The Sixth International Conference on Digital Society*, Proceedings pp. 98-102.

[7] Silic M.; Back, A. Shadow IT – A view from behind the curtain. *Computers & Security*, vol. 45, pp. 274–283, Sep. 2014, doi: 10.1016/j.cose.2014.06.007.

[8] Allen, D.; Burton, F.G.;Smith, S.D.;Wood, D.A. Shadow IT Use, Outcome Effects, and Subjective Performance Evaluation. *Rochester, NY*, Jun. 27, 2017. doi: 10.2139/ssrn.2993443

[9] Alojairi, A. The Dynamics of IT Workaround Practices- A Theoretical Concept and an Empirical Assessment. *International Journal of Advanced Computer Science and Applications*, 2017, 8(7), 527-534.https://doi. org/10.14569/IJACSA.2017.080773.

[10] Behrens, S. Shadow Systems: The Good, the Bad and the Ugly. *Communications of the ACM*, 2009, 52(2), 124-129. https://doi.org/10.1145/1461928.1461960.

[11] Behrens, S.;Sedera, W. Why Do Shadow Systems Exist after an ERP Implementation? Lessons from a Case Study. *Proceedings of the 8th Pacific Asia Conference on Information Systems*, 2004, 1713-1726.

[12] Burnett, M. M.; Scaffidi, C. End-User Development. In Soegaard, M. and Friis, R. (Eds.), The Encyclopedia of Human-Computer Interaction. Aarhus: *The Interaction Design Foundation*, 2013.

[13] Chua, C. E. H.; Storey, V. C.; Chen, L. Central IT or Shadow IT? Factors Shaping Users' Decision to Go Rogue with IT. *Proceedings of the 35th International Conference on Information Systems*, 2014, 1-14. Atlanta: The Association for Information Systems.

[14] Haag, S.; Eckhardt, A. Normalizing the Shadows- The Role of Symbolic Models for Individuals 'Shadow IT Usage. *ICIS 2014*, 2014, 1-13.

[15] D. A. Aziz, "Webserver based smart monitoring system using ESP8266 node MCU module," *International Journal of Scientific & Engineering Research, vol. 9*, pp. 801-808, 2018.

[16] Strong, D.M.; Volkoff O. A roadmap for enterprise system implementation. *Computer*, 37 (6) (2004), pp. 22-29.

[17] Oliver, D.; Romm, C.T. ERP systems in universities: rationale advanced for their adoption Idea Group Publishing, *Hershey, PA* (2002).

[18] Chefjec, T. Resultats De L'Enquete Sur Le Phenomene du Shadow IT http://chejfec.com/2012/12/18/resultats-complets-de-lenquete-shadow-it/ (2012) Retrieved on March 2014.

[19] Rentrop, C.; van Laak, O.; Mevius M. Schatten-IT: ein Thema für die interne Revision Revisionspraxis–Journal für Revisoren, Wirtschaftsprüfer, *IT-Sicherheits und Datenschutz beauftragte (2)* (2011), pp. 68-76.

[20] Warkentin, M. Willison, R. Behavioral and policy issues in information systems security: the insider threat. *Eur Journal Inform System*, 18 (2) (2009), p. 101.

[21] Puhakainen, P. Siponen, M. Improving employees' compliance through information systems security training: an action research study MIS Q, 34 (4) (2010).

[22] Behrens, S.; Sedera W. Why do shadow systems exist after an ERP implementation? Lessons from a case study. *8th Pacific Asia conference on information systems.* Shanghai, China; 2004.

[23] Harley, B. Wright, C.; Hall, R.; Dery K. Management reactions to technological change the example of enterprise resource planning. *J Appl Behav Sci*, 42 (1) (2006), pp. 58-75.

[24] Jones, D.; Behrens, S; Jamieson, K.; Tansley, E. The rise and fall of a shadow system: lessons for enterprise system implementation ACIS, Hobart, Tasmania (2004).

[25] Sherman, R. Shedding light on data shadow systems Inform Manage Online (29 April, 2004), p. 1002617-1.

[26] Haag, S.; Eckhardt, A. Justifying Shadow IT Usage, *PACIS 2015 Proceedings*. 241.https://aisel.aisnet.org/pacis2015/241.

[27] Behrens S. Shadow systems: the good, the bad and the ugly Commun ACM, 52 (2) (2009), pp. 124-129.

[28] Mahmood, M.A.; Siponen, M.; Straub, D.; Rao, H.R.; Raghu, T. Moving toward black hat research in information systems security: an editorial introduction to the special issue MIS Q, 34 (3) (2010), pp. 431-433.

[29] Silic, M.;Back, A. Information security and open source dual use security software: trust paradox open source software: quality verification. *Springer (2013)*, pp. 194-206.

[30] Rentrop, C.; Zimmermann, S. Shadow IT evaluation model. *In 2012 Federated Conference on Computer Science and Information Systems (FedCSIS)*, Sep. 2012, pp. 1023–1027.

[31] Raković, L.; Sakal, M.; Matković, P.; Marić, M. Shadow IT – Systematic Literature Review. *Information Technology and Control*, vol. 49, no. 1, Art. no. 1, Mar. 2020, doi: 10.5755/j01.itc.49.1.23801.

[32] Das, R., Turkoglu, I. "Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method", Elsevier, Expert Systems with Applications, c. 36, sy 3, Part 2, ss. 6635-6644, Nis. 2009, doi: 10.1016/j.eswa.2008.08.067.

[33] Daş R., Turkoglu, I. "Extraction of Interesting patterns through association rule mining for improvement of website usability", IU-JEEE, c. 9, sy 2, ss. 1037-1046, Oca. 2009.

[34] Das, R., M. Z. Gündüz, "Analysis of cyber-attacks in IoT-based critical infrastructures", International Journal of Information Security Science, c. 8, sy 4, ss. 122-133, Ara. 2019.

[35] Das, R., Turkoglu, I, Poyraz, M., "Analyzing of system errors for increasing a web server performance by using web usage mining", Istanbul University - Journal of Electrical & Electronics Engineering, c. 7, sy 2, ss. 379-386, Ocak 2012.

[36] Minh, H.L. Sang-To, T.; Abdel Wahab, M.; Cuong-Le, T. A new metaheuristic optimization based on K-means clustering algorithm and its application to structural damage identification. *Knowledge-Based Systems*, vol. 251, p. 109189, Sep. 2022, doi: 10.1016/j.knosys.2022.109189.

[37] Abernathy, A.; Celebi, M.E. The incremental online k-means clustering algorithm and its application to color quantization. Expert Systems with Applications, vol. 207, p. 117927, Nov. 2022, doi: 10.1016/j.eswa.2022.117927.

[38] Li, Y.; Chu, X.; Tian, D.; Feng, F.; Mu, W. Customer segmentation using K-means clustering and the adaptive particle swarm optimization algorithm. *Applied Soft Computing*, vol. 113, p. 107924, Dec. 2021, doi: 10.1016/j.asoc.2021.107924.

[39] Cho, I.; Park, S.; Kim, J. A fire risk assessment method for high-capacity battery packs using interquartile range filter. *Journal of Energy Storage*, vol. 50, p. 104663, Jun. 2022, doi: 10.1016/j.est.2022.104663.

[40] Selvaraj, E. Collier, J.D.; Culver, E.; Brady, J.M.; Bailey, A.; Pavlides, M. THU460 - Temporal increase in interquartile range iron-corrected T1 in high-risk patients with large-duct primary sclerosing cholangitis. *Journal of Hepatology*, vol. 77, p. S322, Jul. 2022, doi: 10.1016/S0168-8278(22)01009-1.

BIOGRAPHIES

Mucahit Kutsal received his B.Sc. in Software Engineering from Firat University in 2023 and he is currently studying his M.Sc in Quantum Information Technologies at University of Gdansk. He studied at Czestochowa University of Technology for 1 academic year as part of an exchange program during his Bachelor's degree. During his Bachelor's degree, he was involved in various scientific studies and took part in various projects in the information technology industry. His current research areas include geometric deep learning, drug discovery, cryo-EM image processing, bioinformatics, cloud computing and quantum computation.

Bihter Das graduated B.S. and M.S. degrees from the Department of Computer Science at the Firat University in 2004 and 2007 respectively. Then she received Ph.D. degree at the Department of Software Engineering at the same university in 2018. She also worked between September 2017 and June 2018 as a visiting scholar at the Department of Computing Science at the University of Alberta, Edmonton, Canada. Her current research areas include data science, big data, data analytics, bioinformatic, digital signal processing, genome data analysis.

Ziya Askar received the B.Sc. degree from the Department of Computer Engineering at Istanbul Technical University in 1998. He has more than 25 years of experience in information technology and currently works as a Director of Enterprise Architecture at Arçelik Global and Emerging technologies. His current research areas include enterprise architecture, process digitalization, machine learning, functional programming, domain driven design in enterprise organizations.

Ali Necdet Guvercin received the B.Sc. degree from the Department of Computer Engineering at Çankaya University in 2009 and the M.Sc. degree from the Department of Computer Engineering at Karabük University in 2019. He is currently a Ph.D. student in the Department of Computer Engineering at Kocaeli University. He has more than 10 years of experience in information technology and currently works as a Senior Enterprise Architect at Arçelik Global. His current research areas include intelligent automation, robotic automation, artificial intelligence, machine learning, and autonomous systems.

Resul Das is a full professor in the Department of Software Engineering, Technology Faculty, Firat University, where he has been a faculty member since 2011. He graduated with B.Sc. and M.Sc. degrees from the Department of Computer Science at Firat University in 1999 and 2002 respectively. Then he completed his Ph.D. degree at the Department of Electrical-Electronics Engineering at the same university in 2008. He served as both a lecturer and network administrator at the Department of Informatics at Firat University from 2000 to 2011. In addition, he has been the CCNA and CCNP instructor and the coordinator of the Cisco Networking Academy Program since 2002 at this university. He worked between September 2017 and June 2018 as a visiting professor at the Department of Computing Science at the University of Alberta, Edmonton, Canada supported by the TÜBİTAK-BIDEB 2219 Post-Doctoral Fellowship. He has many journal papers and international conference proceedings. he served as Associate Editor for the Journal of IEEE Access and the Turkish Journal of Electrical Engineering and Computer Science from 2018 to 2021. He entered the 2% of the "World's Most Influential Scientists" list published by Stanford University researchers in 2020, 2021, 2022. His current research areas include computer networks and security, cyber-security, software design methods, software testing, IoT/M2M applications, graph visualization, knowledge discovery, and data fusion.