Deep Science Publishing
https://doi.org/10.70593/978-93-49307-51-3_3

DeepScience
Open Access Books

Chapter 3

Design and methodology for IAM data collection and machine learning model experimentation

Esther Chinwe Eze

University of North Texas, United States

3.0 Project Design

This chapter talks about the design approach of the project. It also explains how the IAM training dataset used for the project was collected, and the cleaning and processing techniques.

3.1 Research Architecture and Methodology

The methodology involved in the design and implementation of this project involved three (3) major steps- the installation and configuration of an IAM server (WSO2) which aimed to depict an enterprise environment to provide a real-life dataset for training. The second step after the server setup was to collect real-life data from the server in the form of logs and perform data processing on the collected data. The final step was the selection of a suitable supervised Machine learning algorithm and providing an experiment based on it. In the server, 15 users were created by the super admin and each user was assigned roles with different privileges as shown in Figure 7 below. The experiment was based on users performing normal activities for normal logs and brute force attacks for malicious logs.

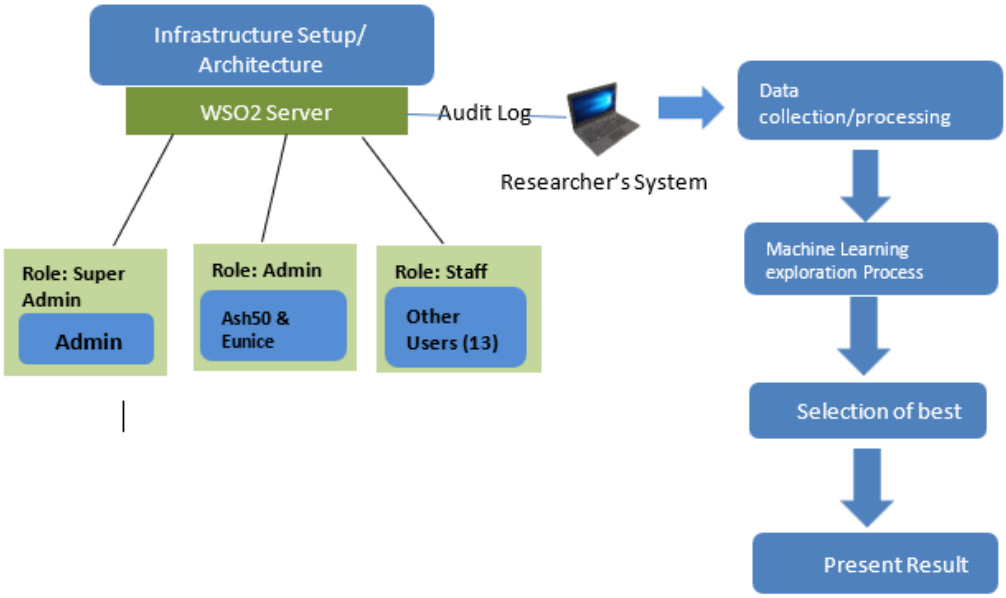

Figure 7: Research Methodology

The initial data obtained from the server before the transformation was in the form of an audit log. An audit log is used to track the sequence of actions that affects a particular task carried out on the server (WSO, 2021).

3.2 Environment/ Testbed Setup

This sub-section describes and explains the environmental setup and simulation of the testbed. The setup went thus:

- First, the WSO2 identity server 5.11.0 which is the latest release was downloaded from the official website.

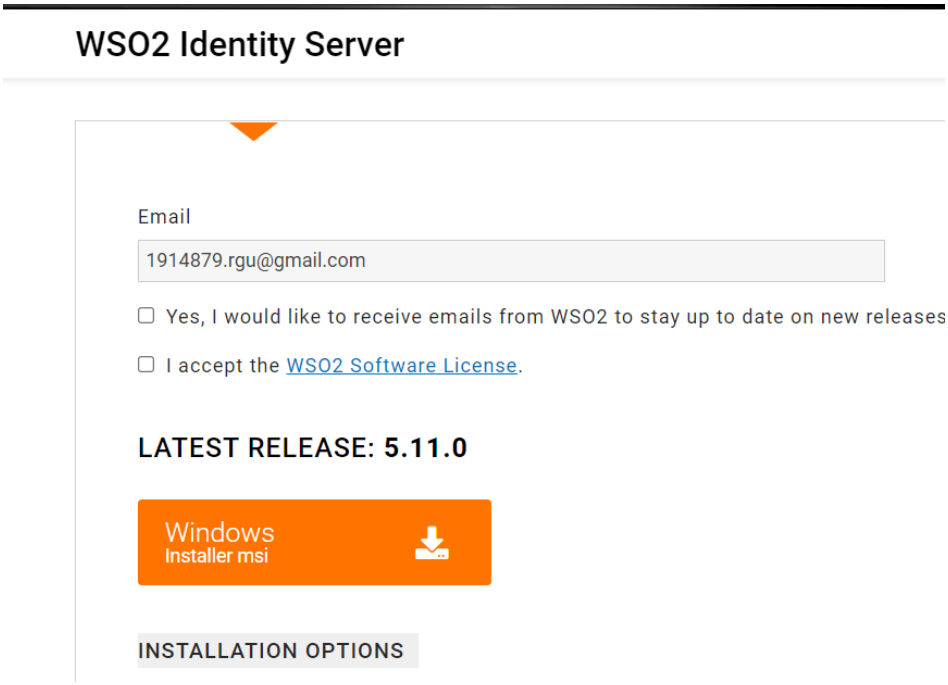

Figure 8: Download WSO2 Server

- After the download, the installation was done. The server was installed on the researcher's local machine as opposed to the initial plan. The initial plan was for the server to be installed on a Virtual Machine (VM) but the installation was not successful hence the reason it was installed on the Local machine of the researcher. This method restricted the attacks carried out especially malicious attacks like privilege escalation which couldn't be done directly on the local machine of the researcher for security reasons.
- After installation, the server is started on the command prompt and the server is left to load successfully. This process allows the necessary programs needed by the server to load properly.
- Once the server is loaded properly, the carbon URL is copied and pasted into a browser which then loads the dashboard of the server.
- The admin then logs into the server and configures it.

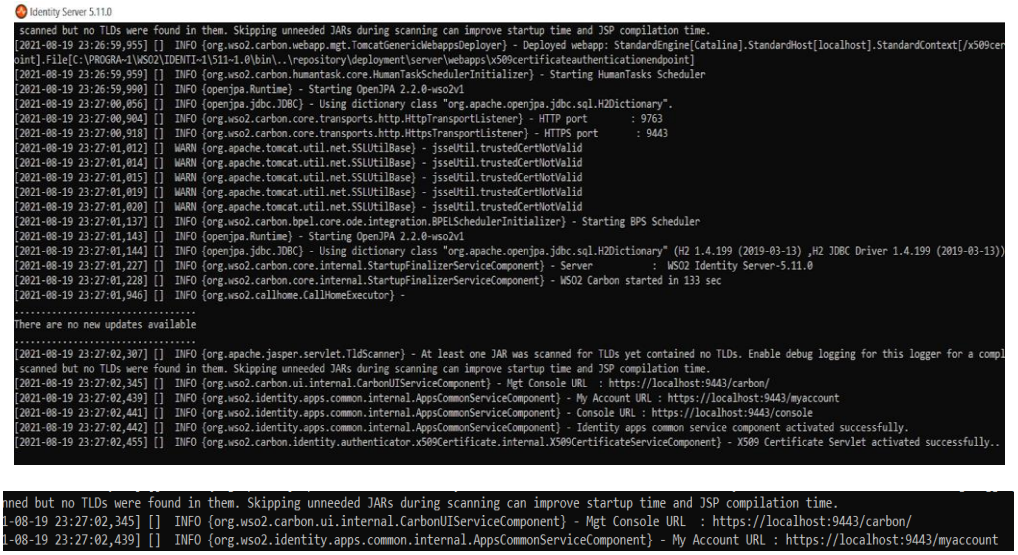

Figure 9: Start Server in Command Prompt

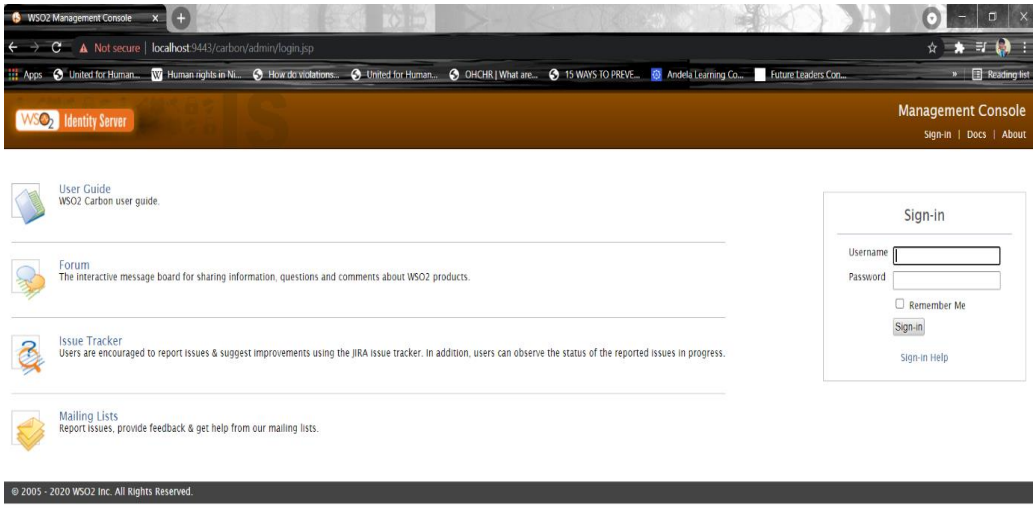

Figure 10: Carbon URL & WSO2 Identity Server Dashboard

As seen in Figure 10 above the WSO2 server is a simulated environment with lots of functionalities and features that include user stores, single sign-on, provisioning, service providers, identity providers, claims, etc. Exploring all features of the server seemed like an impossible task given the time frame of the project and for this reason, the project only focused on the User Management Architecture (user authentication and privilege aspect) of the identity server.

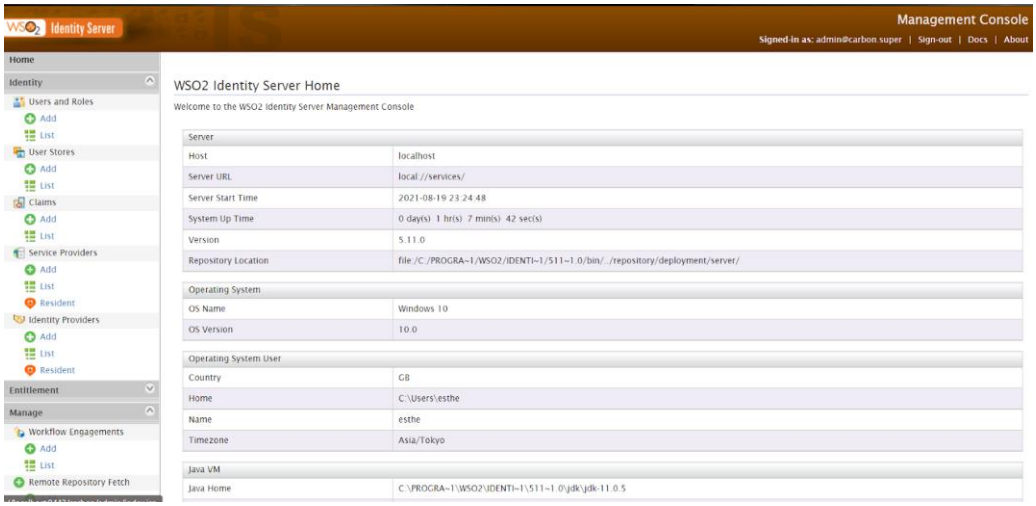

Figure 11: Server Home

3.2.1 Simulation of Benign Activities

The benign activities followed the user management functionality. The main concepts of user management are users, permission, roles, and user stores.

- **Users:** Users are consumers who interact with your organizational applications, databases, and other systems (WSO, 2021).
- **Permissions:** Permission is a delegation of authority or a right that is assigned to a user or a group of users to act on a system (WSO, 2021).
- **User roles:** A user role is a grouping of permissions. In addition to assigning individual permissions to users, admins can create user roles and assign those roles to users (WSO, 2021).

Scenario

Esty store is a fashion company that has 15 staff and a super Admin. Each of these staff has a particular or similar designation. Depending on their designation they all have different permission and access levels. The super admin creates the users and assigns two staff (Ash50 and Eunice) Admin roles, while the remaining users have staff roles. The scenario is otherwise known as a role-based access control which is an approach used to restrict access to authorized users based on their role (WSO, 2021).
The scenario explained in Figure 12 below is how the benign data is collected based on the activities of users according to their roles and privileges.

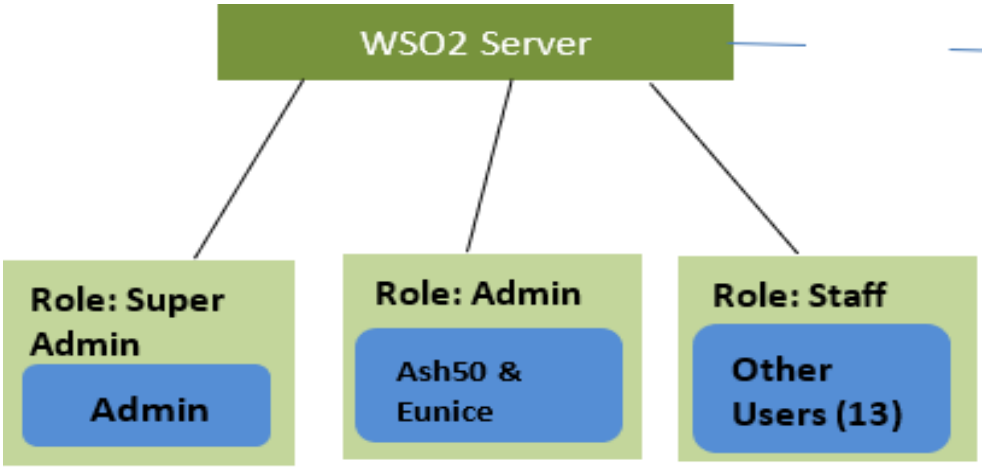

Figure 12: Benign Data Collection

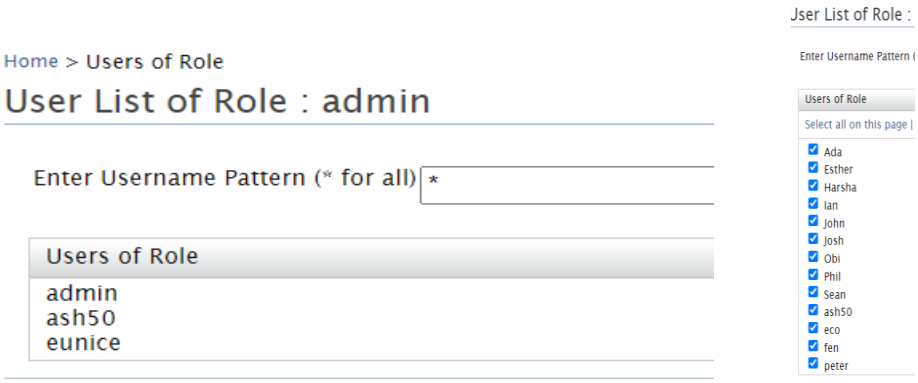

Figure 13: List of Admins & List of other users who are not Admins

The super Admin (default) has unrestricted control while the assigned Admin (Ash50 & Eunice) has restricted control.

Examples of Admin Activities:

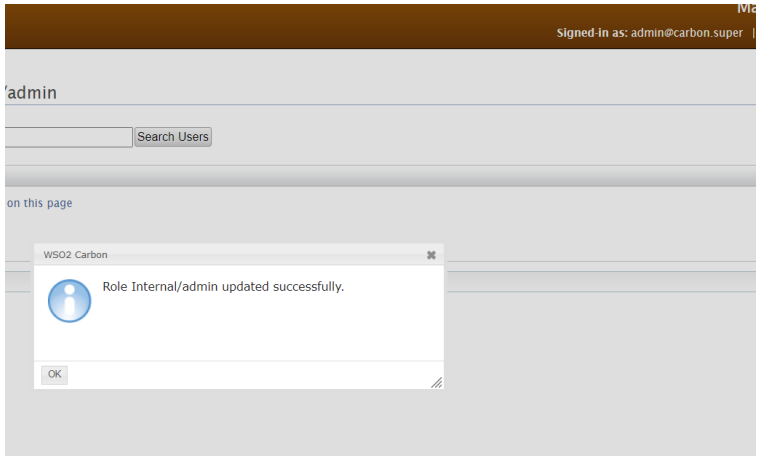

Figure 14: Admins assign roles

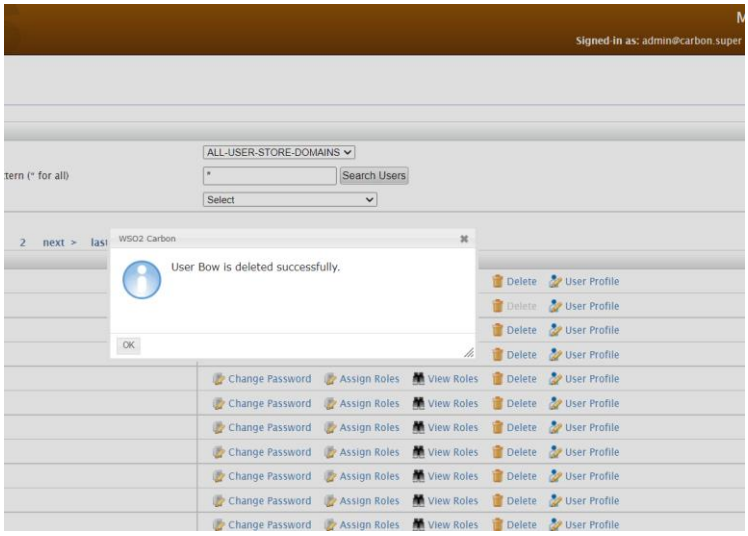

Figure 15: Admins can delete users

3.2.2 Simulation of Malicious Data

The approach used to collect malicious data is similar to the benign activities carried out. The researcher intended to carry out privilege escalation attacks as well as brute force attacks, however, the operation was not carried out as intended due to the failed attempt to install the server on a VM. The intended purpose could not be performed on the researcher's local system for security reasons. However, the researcher resolved to perform a manual brute-force attack on the server. A brute force attack involves guessing and combining usernames and passwords. BFA is a simple method of attack but highly successful.

Figure 16: Brute Force attack on Admin

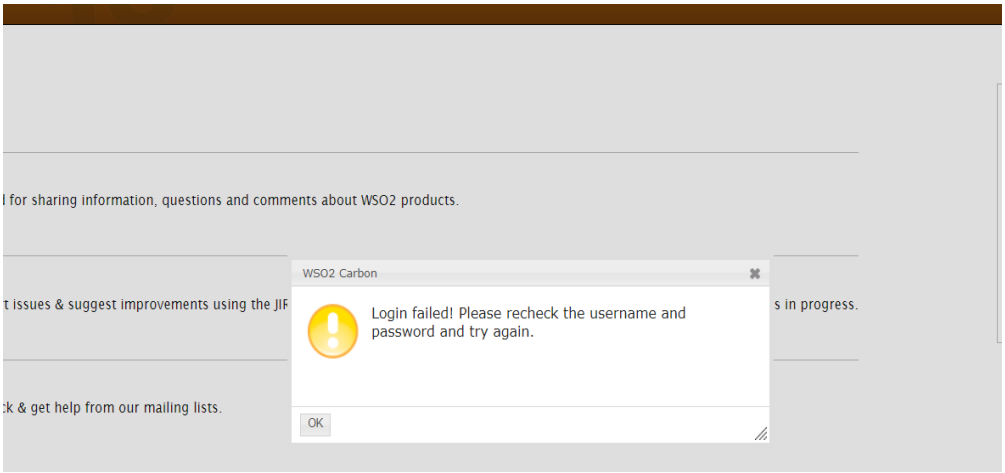

Figure 17: Failed Attempt

As shown in Figures 16 & 17, an unauthorized user attempts to log in as an Admin to access Admin privilege. The result is a failed attempt. The process of performing normal and malicious activities for data collection took one week.

3.3 Data Collection

The major step that was carried out to collect data was setting up a testbed that mimics normal and malicious activities. The process involved the installation and configuration of an IAM server (WSO2) which aimed to depict an enterprise environment to provide a real-life dataset for training.

The IAM dataset used for this project is multivariate data. The dataset was obtained as an unstructured audit log from the installed and configured WSO2 server – an open-source identity and access management server. The IAM features mimicked a real-world enterprise environment. The log captured all activities carried out on the WSO2 server and was transformed into a .csv format for training.

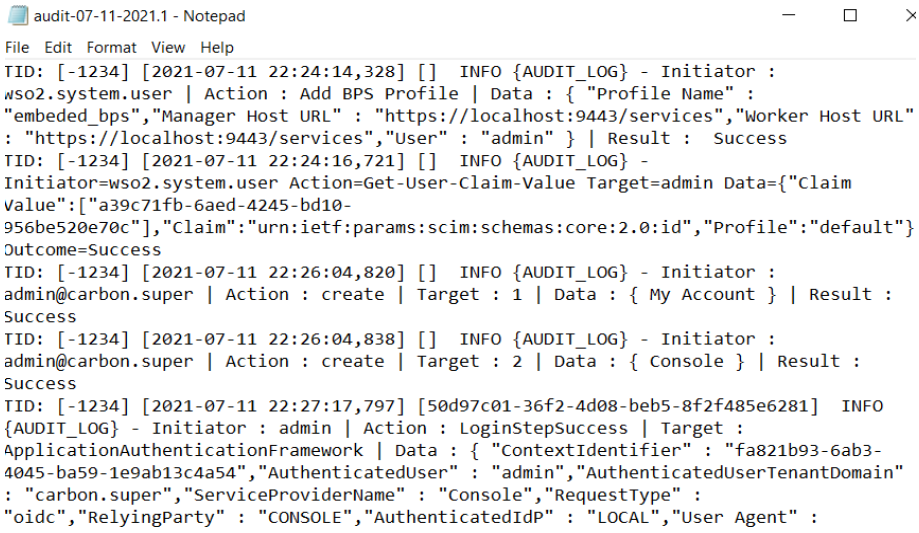

Figure 18: Data in unstructured form (Audit log from WSO2 server)

	A	B	C	D	E	F	G	H	I	J
1	User	TimeStamp	overflow	TimestampDelta	Action	overflow2	overflow3	ActionDelta	Password	Outcome
2	Admin	44388	0.934768519	0	Create	9347685185	9347685185	0	rqntDi16p(	Success
3	Admin	44388	0.934768519	0	Create	9347685185	9347685185	0	9hEkFp	Success
4	Admin	44388	0.935613426	0.000844907	Login	9356134259	9347685185	8449074	DpnYsH	Success
5	Admin	44388	0.935625	1.16E-05	Login	935625	9347685185	9346749560	r4rfjsWtB(	Success
6	Admin	44388	0.935648148	2.31E-05	Get-User-Claim-Values	9356481481	9347685185	8796296	5AlS76Xi4(	Success
7	Admin	44388	0.939108796	0.003460648	Login	9391087963	9347685185	43402778	OzFGNuz1	Success
8	Admin	44388	0.942592593	0.003483796	Login	9425925926	9347685185	78240741	90y6TfrOx	Success
9	Admin	44388	0.946076389	0.003483796	Login	9460763889	9347685185	113078704	v4pRdWT5	Success
10	Admin	44388	0.952604167	0.006527778	Logout	9526041667	9347685185	178356482	wUSO0B	Success
11	Admin	44392	0.92625	0.026354167	Login	92625	92625	0	sduYFpu5q	Success
12	Admin	44392	0.927210648	0.000960648	Get-User-List	9272106482	92625	9272013857	gEMoEY	Success
13	Admin	44392	0.928622685	0.001412037	Create	9286226852	92625	9286134227	YUuVsCiX9	Success
14	Admin	44392	0.928634259	1.16E-05	Get-Roles-of-User	9286342593	92625	9286249968	sokwked6l	Success
15	Admin	44392	0.932858796	0.004224537	Update	9328587963	92625	9328495338	MSSSfAD	Success
16	Admin	44392	0.933020833	0.000162037	Update	9330208333	92625	9330115708	3s4yoXmV	Success
17	Admin	44392	0.933252315	0.000231481	Update	9332523148	92625	9332430523	QsQLikOw	Success
18	Admin	44392	0.9365625	0.003310185	Get-Roles-of-User	9365625	92625	9273000	qL0UlFbA(	Success
19	Admin	44392	0.9365625	0	Get-User-List	9365625	92625	9273000	5Rr2V9Z	Success
20	Admin	44392	0.9365625	0	Add User	9365625	92625	9273000	6IvMDvEx	Success
21	Admin	44392	0.936898148	0.000335648	Add User	9368981481	92625	9368888856	sX29iHVGL	Success
22	Admin	44392	0.936898148	0	Get-User-List	9368981481	92625	9368888856	vdBnyy2v6	Success
23	Admin	44392	0.937175926	0.000277778	Get-Roles-of-User	9371759259	92625	9371666634	peTnodygz	Success
24	Admin	44392	0.938738426	0.0015625	Update-Roles-of-User	9387384259	92625	9387291634	CNxwwsLE	Success

Figure 19: Data in Structured form (.csv format)

3.4 Dataset Description

For this research, two IAM datasets were used. The first is called “ben_IAM” which was generated as an audit log, consisting of normal activities carried out on the server. The second dataset is called “mal_IAM” which was also generated in the same way as the first but consists of a series of malicious activities. The malicious activity was focused on brute force attacks. Both datasets were extracted from raw log files that contained a series of variables. Relevant features were later extracted to generate both datasets. Each dataset consisted of 10 variables (columns) and 1001 observations (rows).

No	Features	Description	Data type
1.	User	The name of the user or logger performing an activity.	Character (Chr)
2.	Timestamp	The time and date of the occurrence of an event or activity.	Integer (int)
3.	Overflow 1	The sequence of the Timestamp	Numeric (num)
4.	Timestamp Delta	The difference in the timestamp of each activity.	Numeric (num)
5.	Action	The type of event or action performed by a user.	Character (Chr)

6.	Overflow (Login)	The login time	Numeric (num)
7.	Overflow (Logout)	The Logout time	Numeric (num)
8.	Action Delta	The intervals between login and logout	Numeric (num)
9.	Password	Passwords	Character (Chr)
10.	Outcome	The outcome or result of every activity performed by the users	Character (Chr)

Table 3: Description of Features and Data Types

3.4.1 Data cleaning and preparation

The process of data cleaning is very crucial. It involves the process of converting raw data to logical data that can be used for training purposes. Raw data can be difficult to work with unless it has been transformed (pre-processed). Oftentimes, raw data may lack headers, contain wrong character encoding, or wrong data types, hence, the importance of data cleaning. The preparation process went thus:

- The datasets were obtained in a text format (.txt) and were transformed to a comma-separated value format (.csv) and saved in Excel spreadsheets.
- The ben_IAM dataset had several variables but was cleaned to 10 variables.
- Afterward, the “read.csv” function in R was used to read the IAM datasets and stored them in a data frame for use.
- The “read.csv” function of the readr library in R is used to read or load the dataset into R.

```
```{r}
#import datasets
ben_IAM <- read.csv("C:/Users/esthe/Desktop/MSC PROJECT/WSO2/ben_IAM.csv")
mal_IAM <- read.csv("C:/Users/esthe/Desktop/MSC PROJECT/WSO2/mal_IAM.csv")

tail(ben_IAM)
```

Figure 20: Loading the Dataset

3.4.2 Missing Values

The datasets did not contain any missing values. The sum() function in R was used to check for missing values and check the completeness of data. Since no missing value was found, it eliminated the need for data imputation (an approach used to replace
```

missing values with a reasonable guess about what the missing values would have been if not missing)

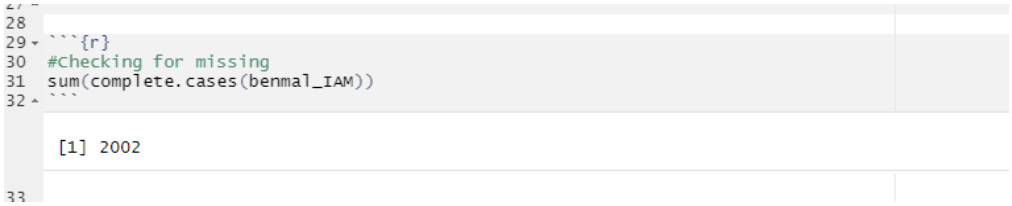

Figure 21: Checking for missing Values

3.5 Choice of Development Tools

In experiments, a testbed environment had to be set up and for this reason, an IAM server (WSO2) was installed and configured. In addition, three other tools were employed to carry out the Machine Learning experimental process – R studio which is a Data science software, the Weka GUI machine learning software, and MS Excel. The choice of these tools is due to their flexibility in performing unique tasks and ease of use.

3.5.1 Wso2 Server

The WSO2 is a 100% open-source IAM technology that offers an enterprise platform for integrating Application Programming Interfaces (APIs), Applications, and web services locally and across the internet. It provides a solution for end-to-end API management in the cloud, on-prem, or hybrid environment. The increase in the number of consumers and users has made managing services/ microservices such as security, and access control difficult to handle. WSO2 offers a solution to this problem hence making it an important technology in the aspect of Identity and access management (WSO, 2021).

3.5.2 R Studio

R studio is an Integrated Development Environment (IDE) for R, a programming language used for statistical computing and graphics. The researcher chose to use R for the following reasons:

- **Scalability** – R is very powerful and flexible in processing large amounts of data.
- **Integration** – it can read various data formats such as CSV, text, etc. It can scrape data directly from websites.

- **Open Source** – it is free and does not require a license or subscription. This also means that it is constantly improved by a community of users and developers.

3.5.3 WEKA (Waikato Environment for Knowledge Analysis)

Weka is an open-source GUI-based Machine Learning algorithm tool that contains a collection of tools for data preparation, classification, regression, clustering, association rules mining, data mining tasks, and visualization. Reasons for using Weka:

- **Ease of Use** – The Weka application served as an alternative tool for the completion of the project. It was very easy to use and learn since the researcher had no prior knowledge of the tool.
- Weka has a large selection of machine-learning algorithms to choose from for classification and regression problems (Jason, 2016).
- Easy to configure each machine learning algorithm which saves time.
- It also gives accurate results.
- Great tool for ranking and feature selection.