המחלקה למערכות מידע ניהוליות

פרוייקט גמר תשף 2020

זיהוי פעולות חשודות (אנומליות) ברשת הביתית

NETWORK TRAFFIC ANOMALY DETECTOR

מסמך זה מהווה חלק מדרישות לקבלת תואר ראשון במערכות מידע BSC

מכללת אחווה

שמות המבצעים: אלמוג אליאס, קובי תורג'מן וכנרת כהן.

שם המנחה: ד"ר יצחק רוזן.

**הצהרה**

הפרויקט נעשה בהנחיית ד"ר צחי רוזן

המכללה האקדמית אחווה- מחלקת מערכות מידע ניהוליות

מכיוון שהפרויקט הנו בלמידת מכונה ושימוש בתחום איתור התקפות בתעבורת הרשת על בסיס אנומליה, הוא חדשני יחסית, דרוש מחקר רב על מנת לבצע את הפרויקט בנוסף למימוש עצמו.

הפרויקט מציג את עבודתנו האישית ומהווה חלק מקבלת תואר ראשון במערכות מידע ניהוליות.

# תיאור הארגון

## 1.1.1. רקע כללי

רשת מקומית LAN מאפשרת למחשב אחד המחובר לרשת להשתמש במשאבים ממחשבים אחרים. שימושים נוספים הם הרצת יישומים ממחשב אחד על מחשב אחר.

חסרונותיה של העבודה ברשת הם סיכון לפרטיות המידע וריבוי מקורות לתוכנות זדוניות. המחשבים שברשת יכולים להתחבר ביניהם בכבלים או בקשר אלחוטי ויש לוודא שאין פעילות חריגה על הרשת, ושרק אנשים מורשים יוכלו להשתמש במחשב שכבר מחובר אל הרשת. מעבר לזאת יש להגביל את הגישה של מי שכן מורשה כך שתוכנם של מסמכים יהיה גלוי רק לבעלי סמכות רלוונטית עבור תוכנו של כל מסמך או תוכנה.

## 1.1.2. לקוח

לקוחות המערכת הם משתמשי הבית החכם. הלקוח מגדיר לנו מה לעשות ובהתאם לזה נבנה כנדרש. המערכת תשמש בעיקר את דיירי הבית ומשתמשי הרשת המקומית LAN. הגדרנו את קובי להיות המייצג של הלקוח, אשר יהיה מסוגל להגדיר את הדרישות של הלקוח.

## 1.1.3. מומחי היישום

מומחי היישום הינם בעלי הבית אשר יטמיעו את המערכת.

ההחלטה לגבי הפיתוחים תהיה על ידי צוות היגוי שתורכב מראש צוות האפיון, ומומחה אבטחת מידע.

## 1.1.4. משתמשים

אחראי הרשת המקומית – בעל הבית.

# 1.2. מטרות ויעדי הארגון

## 1.2.1. מטרת הארגון

הטכנולוגיה חשופה לשלל פרצות מידע, ויש צורך תמידי לשכלל ולהתמודד עם באגים, אשר הולכים והופכים למתוחכמים יותר .ישנם שיקולים שחייבים להילקח בחשבון כשעוסקים בחיבור מחשבים לרשת. הנחה היא שאי אפשר לעבוד ללא תקשורת בין מחשבים ולכן המטרה היא לזהות פעילות חריגה עויינת של תוקפים למניעת התפשטות ברשת המקומית.

## 1.2.2. עקרי ענף המערכת

הענף בו עוסק הפרויקט הינו תחום הסייבר על ידי שיטות ואלגוריתמים של למידת מכונה.  
קהל היעד של הפרויקט הנו:

1. משתמשי רשת מקומית
2. משתמשי הרשת ביתית
3. משתמשי הבית חכם

## 1.2.3. הצגת מטרות המערכת

הגדרת הפרויקט היא בניית תשתית שבאמצעותה אפשר לחקור תופעה אנומלית.

לדמות התקפות באמצעות הסימולטור ולהגדיר את ההתנהגות שלהן באמצעות אלגוריתם של למידת מכונה.

גבולות המערכת-

1. הקמת מעבדה (חומרה) – סקלביליות
2. סימולטור- מדמה סביבה רגילה וגם סביבה שמדמה התקפות.
3. אוספים את התנועות ברשת - גשש שאוסף את התנועות ברשת בזמן אמת ומתעד אותם - ושליפה של ערכים להגדרת מודל לאנליזה עליו נעבוד.
4. הקמת DB – 2 שכבות: שכבת RAW DATA, שכבת עיבוד המידע.
5. אנליזה של הנתונים באמצעות אלגוריתם של למידת מכונה.
6. ויזואליזציה - הצגת הנתונים

## תיאור הבעיה

כמות המידע שגוברת עם הזמן והתממשקות מול רכיבים נוספים גורמת לסיכון גבוה יותר לתקיפת סייבר בעולם וברשת הפנימית בפרט.

כיום ישנם חברות קיימות אשר משלבות מוצר חומרתי (נתב) או חוצץ בין הרשתות כדי לזהות אנומליות אך לא בהכרח כל התעבורה עוברת דרך הנתב.

אף על פי כן, מוצרים אלו מאתרים איומים והתקפות בתעבורה ע"י חיפוש תבניות ספציפיות או ע"י הכרת דפוס הפעולה של הקובץ הזדוני, כך שלכל מתקפה קיימת תבנית ייחודית עבורה.

כאשר נחשף איום חדש יוצרים תבנית תואמת לאיום ולפי התבנית המערכת מזהה את האיום לאחר מכן.

ההנחה היא שהתוקף כבר נמצא ברשת הביתית וע"י למידת וחקירת הרשת הוא מזהה מטרות תקיפה.

**הבעיה** היא זיהוי פעילות בשלב  Active Reconnaissance של תוקף וזמן התגובה לאיום זה מהרגע שבו הוא מזוהה עד לבניית התבנית עבורו, כיוון שכבר נוצרה פגיעה במערכת.

היום בעולם המודרני ישנה עלייה בכמות המידע העוברת ברשת ובעקבות כך נוצר ריבוי בסכנות וסיכוי גבוה יותר לפריצות ע"י תוקפים פוטנציאלים.

מדובר על **מערכת מידע מורכבת** שלא ניתנת לחיזוי מתקפות סייבר עתידיות.

## סקירת ספרות

הוחלט להתבסס על הספר "Anomaly Detection Principles and Algorithms.pdf"

אנומליה

בכריית מידע, זיהוי אנומליות Anomaly Detection (לעיתים מכונה גם גילוי אנומליות או זיהוי חריגים) מתייחס לאיתור תצפיות או אירועים אשר אינם תואמים לדפוס הצפוי או ליתר התצפיות בבסיס הנתונים.

לרוב, התצפיות החריגות מעידות על בעיה. אנומליה בתעבורת רשת תקשורת עשויה להעיד על מתקפת סייבר על רשת התקשורת.

ניתן לחלק את הטכניקות לזיהוי אנומליות לשלושה סוגים:

טכניקות לא מונחות (unsupervised), אשר מניחות כי רוב התצפיות בבסיס הנתונים מייצגות מקרים נורמליים או תקינים. לשם כך ניתן, למשל, להשתמש בטכניקות של ניתוח אשכולות כדי לאפיין את התצפיות הנורמליות. תצפית שאינה שייכת לאף אשכול מוגדרת כאנומליה.

טכניקות מונחות (supervised), אשר בהן התצפיות בבסיס הנתונים סווגו מראש לתצפיות "נורמליות" או לתצפיות "לא נורמליות". במקרה כזה, ניתן להשתמש בשיטות של למידה חישובית לאימון מסַווגים, המאפשרים סיווג של מקרים חדשים שאותם לא פגשנו בתהליך הלמידה.

טכניקות מונחות למחצה (semi-supervised), אשר בונות מודל המייצג את דפוס ההתנהגות הרגיל מתוך אוסף של תצפיות נורמליות. לאחר מכן בוחנים באמצעות המודל את הנראות של תצפיות חדשה.

כאשר מיושמים אלגוריתם לאיתור חריגות, יש לקחת בחשבון שלושה מקרים אפשריים:

1. Correct Detection: חריגות שנמצאו בנתונים אכן תואמות בדיוק את החריגות בתהליך.

2. False Positive: התהליך ממשיך להיות תקין, אך נצפים ערכי נתונים לא צפויים, למשל בגלל רעש מערכת פנימי.

3. False Negative: התהליך הופך להיות חריג, אך ההשלכות אינן רשומות כנתונים הלא תקינים, למשל, בגלל שהאות החריגה חזקה באופן לא מדויק לעומת הרעש במערכת.

מה זה נורמה?

באנליזה מתמטית, נורמה היא פונקציה ממשית המוגדרת על מרחב וקטורי, ומתאימה לכל וקטור ערך ממשי, באופן שמתמלאים מספר תנאים. תנאים אלו מבוססים על התכונות היסודיות של האורך המוכר במרחב האוקלידי. מרחב וקטורי שמוגדרת עליו נורמה נקרא מרחב נורמי.

האורך במרחב האוקלידי מקיים את הדרישות הטבעיות הבאות:

* אורך הוא תמיד חיובי, חוץ מאורכו של וקטור האפס, שהוא אפס.
* מתיחה של הווקטור בסקלר מכפילה גם את האורך בערכו המוחלט של אותו סקלר.
* מתקיים אי שוויון המשולש.

**Distance based anomaly detection approaches**

***זיהוי אנומליה על ידי איתור מרחק***

*בהגדרה, זיהוי אנומליה כרוך בכך שנתון מסויים "שונה" מאחרות. הגדרה זו מוגדרת בהכרח על ידי מערך הנתונים שעליו משווים את נקודת הנתונים.*

*חלק זה מתמקד באלגוריתמים לאיתור חריגות הנשענים על חישובי מרחק בין נקודות שונות לסק נקודות. הסמל D משמש לייצוג מערך הנתונים ה – N ממדי, המשוער כי נמצא במרחב רציף ללא הערכה אמיתית Rn. נקודות נתונים ב-D מסומנות על ידי תווים p,q. סמלים בעלי אותיות גדולות כמו P משמשים לציון קבוצות נקודות, כלומר . המרחק בין 2 נקודות מסומן על ידי d(p,q)*

1. *מדידה: עד כמה נקודה נתונה באופן לא נורמלי? זה דורש הפיכת נקודות לסולם חד ממדי.*
2. *אבסולוטי: האם נקודה נתונה היא אנומלית? זה דורש threshold אשר אומר שנקודה*

*היא אנומאלית אם .*

1. *רלטיבי: האם נקודה אחת יותר אנומלית מנקודה אחרת? זה מצריך השוואה בין נקודות. אז*

*אם נקודה יותר אנומאלית מנקודה אחרת .*

*נוכל לסמן את ההקשר עם הסימן אם .*

*Similarity Measures מדדי דמיון*

*לעתים קרובות אנו מוצאים דמיון במונחים של מידת מרחק. נשים לב שניתן למדוד דמיון על ידי השגת 'עקרון' בהתחשב ב'הפוך' של מידת מרחק. אמצעי דמיון נפוצים המיועדים לכלול כוללים אמצעים ישירים כמו האוקלידים, מינקובסקי ומהלאנוביס. משתמשים גם באמצעים נוספים, כמו דמיון קוסינוס ומדד ג'קרד; לעתים קרובות אמצעים עקיפים, כמו השכן המשותף הקרוב ביותר (SNN - מספר הנקודות הנפוצות בין k שכנים של שתי נקודות מייצג את הדמיון הרצוי בין הנקודות).*

מידע חריג במרחב חד מימד

כל נקודת מידע היא מספר יחיד. ישנם התפלגויות שונות לחריגים במרחב חד מימד:

Uniform Distribution – התפלגות אחידה

כאשר נתונים מתפלגים באופן אחיד על טווח סופי, הממוצע וסטיית התקן רק מאפיינים את טווח הערכים. אם בקרבת איזור של כל נקודת נתונים היא עשירה בנקודות כמו כל נקודה אחרת, אז אין חריגות בתנאי שהגבולות התחתונים והעליוניים של התחום רחוקים יחסית מהממוצע. אפשרות לאנומליה יכולה לקרות כשאיזור קטן מכיל פחות או יותר נקודות נתונים אשר מצופה בהתפלגות האחידה.

Normal Distribution - התפלגות רגילה

כאשר נתונים מתפלגים נורמאלית, צפיפות הנקודות יורד משמעותית ככל שאנחנו מתרחקים מהממוצע. כ 0.1% מהנקודות יותר מ- (3 סטיות סטאנדרטיות) מהממוצע, ורק % *מהנקודות אשר יותר מ6 סטיות סטאנדרטיות מהממוצע.*

נקודות אשר נמצאות מעבר למרחק מהממוצע נחשבים לחריגים. לעומת זאת קיומם של מספר נקודות הרחוקות מהממוצע זהו צירוף מקרים להתפלגות נורמאלית, מכאן ניתן לטעון כי אלה נקודות חריגות במידה והמספר שלהם גבוה מהמספר המצופה כשהנתונים מתפלגים נורמלית. לדוגמה אם 2% מנתוני המידע נמצאים מעל 3 סטיות סטאנדרטיות.

Other Unimodal Distributions – התפלגויות Unimodal אחרות

רבים מהתפלגויות Unimodal לא נורמאליים, כאשר יש גבול תחתון לטווח ערכי הנתונים, לדוגמה התפלגויות המכילות Log-normal ו- Gamma. אם מאפייני התפלגות ידועים, אחת מהדרכים לנסות למצוא Thresholds beyond אשר הוא מספר קטן (כ-1%) של נקודות נתונים. שוב, ניתן לטעון כי אוסף הנתונים הללו הינו אנומאלי אם מספרם גדול מחיזוי הסטטיסטיקה.

Multimodal Distributions – התפלגות מולטימליות

ההתפלגות לכמה מערכי נתונים כוללים מספר רב מצבים, שיתגלו רק כאשר הנתונים נבדקים מקרוב. היוריסטיקות כגון חוק אינו מועיל בהתפלגויות כאלה. ולכן, צריך חשוב על הנתונים אשר מורכבים מאוסף אשכולות של נקודות נתונים.

Clustering מהווה חלק חשוב בזיהוי אנומליות. נקודות אשר לא שייכות להתאגדות מסויימת של נקודות יכולות להיות מסומנות כנקודות חריגות. למעשה, נקודות אשר רחוקות מהתאגדויות שכנות נחשבות לוגית לאנומליות. הדבר החשוב במודל מסוג Clustering הוא ההבהרה מתי אוסף של נקודות צריך להיחשב כהתאגדות, ומה המשמעות של נקודה אשר נמצאת רחוק מספיק מהתאגדות זו (או מכמה התאגדויות).

ההגדרה הבלתי פורמלית של אשכול היא שמדובר באוסף של נקודות קרובות זה לזה.

Density-based cluster – התאגדות מבוססת צפיפות

זיהויים פופולארים מאוד בזיהוי מידע אנומאלי. אם המספר היחסי של נקודות (ליחידת מרחק) גדול משמעותית באיזור קטן מכל מערך הנתונים. ניתן לקחת בחשבון את הנקודות באיזור הזה כהתאגדות של נקודות. זו עדיין לא הגדרה מתימטית, מכיוון שהביטויים "איזור קטן" ו-"גבוה משמעותית" לא מוגדרות.

ההתפלגות של הצפיפות ניתנת לניתוח, (אם איננו מודאלי) נוכל לזהות את צפיפות הסף העליון כשיש צפיפות גבוהה אשר מספיקה לשקול איזור זה כהתאגדות של נקדות.

נשים לב שאותו מערך הנתונים יכול להכיל איזור אחד עם צפיפות גבוהה ואיזור אחר עם צפיפות נמוכה שגם נחשיב אותו להתאגדות.

Compare Intra & Inter groups distance – השוואת מרחקים פנים קבוצתיים עם בין קבוצתיים.

אספקט נוסף, הינו השוואה בין מרחקים בפנים קבוצתיים עם מרחקים בין קבוצתיים, ובכך לטעון שאוסף נקודות מהתאגדות קרובים יותר אחד לשני (יותר מהממוצע) מאשר מרחקים מהנקודות נמצאות מחוץ להתאגדות. לעומת זאת, על פי הגדרה זאת התפלגות המידע עלולה להוביל להתאגדות טרוויאלית אשר מאכלסת כמעט את כל הנקודות בהתאגדות אחת.

חלק מאלגוריתמי ה-Clustering מקצים כל נקודה להתאגדות מסויימת אך לא בהכרח.

אחרים, מניחים שמספר ההתאגדויות קבוע אך זה גם לא הכרחי.

כאשר משתמשים בגישות מבוססות התאגדויות לגילוי אנומליות, נקודות בעל גודל מינימאלי שבתוך ההתאגדויות בדרך כלל לא נחשבים לאנומליים. "גודל מינימאלי" הוא פרמטר שצוין חיצונית כגון, הסף שמבוסס על התפלגות גדלי ההתאגדויות במערך נתונים.

לא כל הנקודות שמחוץ להתאגדויות צריכים להיחשב כאנומאלים. נקודה אחת יכולה להיחשב "יותר אנומאלית" מנקודה אחרת אם היא רחוקה יותר מההתאגדות הקרובה ביותר – אשר מרחק זה מוערך בדרכים שונות.

מידע חריג במרחבים רב מימדיים

הגדלת מימדיות גורמת לסיבוכים נוספים. הסיבוכים נובעים מבחירת מידת המרחק. ייתכן שמדד המרחק האוקלידי אינו מתאים כיוון שהוא עשוי לשלב מספרים ממימדים שונים, אבל הבחירה בגורם קנה המידה המשותף אינו וודאי לכל המימדים. בניסיון להעריך איזו משתי נקודות היא קרובה יותר להתאגדות, למשל מידת המרחק שנבחר עשויה לקבוע את המנצח.

נורמליזאציה, לדוגמה, קנה מידה בהינתן מימדים שונים ככלי התאמה באותו התחום, נחוץ לעיתים, אבל מעלה את השאלה איזה סוג נורמליזציה הוא הטוב ביותר. ערכים קיצוניים עשויים לעוות את תוצאות הגישה לנורמליזציה.

סיבוך מעניין מתעורר כשאחד המימדים הרלוונטים הינו זמן, אשר מצריך התייחסות בצורה שונה. למעשה, ב"סדרת זמן" מעורבים 3 תכונות: הראשונה הינה תווית (זיהוי האובייקט), השניה מציינת זמן אבסולוטי, והשלישית היא ערך מספרי (נומרי). ערכי זמן הינם ערכים דיסקרטיים וההנחה היא שקיימות נקודות נתונים עבור כל תווית וכל נקודת זמן.

## תיאור המערכת העתידית

**איסוף המידע**

1. הסנפת חבילות בעמדת קצה.
2. שליחת המידע למאגר המידע ב – DB.

**עיבוד הנתונים**

1. רישום הנתונים לתוך מאגרי Packets - Features.
2. עיבוד חבילות ל- Flows – שיחות.
3. רישום ה Flows למאגרי Features – Flows
4. מאגר חריגות בהתאם למודל שנבנה.

**ויזואליזציה**

1. יצירת ממשק WEB ויזואלי לגראפים על סמך מאגרי הFeatures.
2. יצירת טבלה המציגה את החריגות שנמצאו על ידע המודל.

## בחירת Features

* + - 1. כמות חבילות ליחידת זמן   
         כמות גדולה של חבילות ליחידת זמן עשויה להעיד על פעילות חריגה.

**מודל – נתונים נאספים:**

* רשימה של זמני החבילות.
* כמויות של חבילות לפי פרק זמן.

**סימולציה – אוסף נתונים:**

הרצת תעבורה בפרוטוקול: Microsoft – ds אשר משתמש בפורט 445.

פרוטוקול זה משמש להעברת קבצים בין עמדות קצה

סימולציה נורמלית:

העברת קבצים בזמן רנדומלי בין 10-120 שניות באופן תמידי.

כל קובץ ינוע בגודל שבין 10-100 Bytes.

סימולציות אנומליות:

1. העברת קובץ יחיד בנפח שגדול פי 1000 מהסימולציה הנורמאלית.
2. העברת 100 קבצים בגודל של 10 Bytes בפרק זמן בין 0.1-2 שניות.

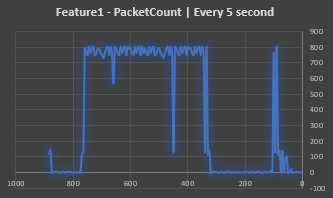
**למה זו יכולה להיחשב חריגה**

* כמות של חבילות למשך זמן ארוך מהממוצע (מכיוון שהקובץ גדול אז תהינה כמות רבה של חבילות לאורך זמן) נחשבת לחריגה מכיוון שזמן הפעילות גבוה מהזמן הממוצע.
* כמות גבוהה של חבילות בזמן קצר נחשבת לחריגה מכיוון שהכמות גבוהה מהממוצע.

**על מה היא מעידה?**

* מעידה על עומס ברשת – שיכול לגרום לאיטיות – בדרך כלל אין פעילות מסיבית, ולכן זאת תופעה פוטנציאלית שעלולה לקרות בזמן של חדירה.

חיזוי גרפי:



* + - 1. כמות הודעות ל- Flow

כמות קטנה/גדולה של חבילות בשיחה ליחידת זמן עשויה להעיד על פעילות חריגה.

**מודל – נתונים נאספים**זמן התחלת השיחה, יום התחלת השיחה, כתובת יוזם השיחה, כתובת יעד השיחה, פורט מקור, פור יעד, פרוטוקול שירות (Service), כמות החבילות הכללי בשיחה, כמות ה-Syn Flag בשיחה, כמות ה-Ack Flag בשיחה, כמות ה-Payload בשיחה, כמות ה - Fin בשיחה.

**סימולציה – אוסף נתונים**

סימולציה נורמלית:

הרצת תעבורה בפרוטוקולים : 80,3389,21,23.

העברת רצפי מחרוזות בגודל 2-10 מילים אשר בגודל 1-15 אותיות בזמן רנדומלי בין 1-10 שניות באופן תמידי.

סימולציות אנומליות:

1. הרצת תעבורה בפורט שלא היה קיים קודם לכן לדוגמה פורט מספר 4444.
2. העברת רצף מחרוזות יחיד בגודל של פי 10 מהסימולציה הנורמלית.
3. העברת 100 רצפי מחרוזות בגודל של מילה יחידה בפרק זמן בין 0.1-2 שניות.
4. שליחת רצף מחרוזת יחיד בגודל של אות בודדת.

**למה זו יכולה להיחשב חריגה**

1. שימוש בפורט שלא היה בו שימוש קודם לכן עלול להיחשב כחריג מכיוון שהוא חדש.
2. העברת מחרוזת גדולה נחשבת לחריגה מכיוון שתכיל חבילה בגודל אשר גדול מהממוצע.
3. כמות גדולה של מחרוזות ליחידת זמן עשויה להעיד על פעילות חריגה מכיוון ששיחה תכיל מספר רב של חבילות.
4. כמות קטנה של חבילות עם Payload בשיחה נחשבת לחריגה מכיוון שכמות החבילות שמכילות תוכן קטנה מהממוצע.

**על מה היא מעידה**

תוקף מנסה להסוות את פעילותו על ידי שימוש מופחת/מופרז בשיחות.

חיזוי גרפי:

* + - 1. גודל (bytes) חבילה ליחידת זמן:  
         העברת חבילות גדולות לאורך זמן עשויה להעיד על פעילות חריגה.

**מודל – נתונים נאספים**גודל החבילה ויחידת זמן.

**סימולציה – אוסף נתונים**

סימולציה נורמלית

בכל פרק זמן רנדומלי בין 10-120 שניות תעבור חבילה בגודל 10-100 bytes בין המחשבים.

סימולציות אנומליות

1. העברת חבילות בנפח שגדול פי 100 מהסימולציה הנורמאלית.
2. העברת 100 חבילות בגודל בפרק זמן בין 0.1-4 שניות.

**למה זו יכולה להיחשב חריגה**

* העברת חבילות גדולות נחשבת לחריגה מכיוון שבדרך כלל לא מעבירים חבילות מלאות למשך זמן.
* כמות גבוהה של חבילות בזמן קצר נחשבת לחריגה מכיוון שהכמות החבילות תיהינה גבוהה מהממוצע.

**על מה היא מעידה?**

* מעידה על עומס ברשת – שיכול לגרום לאיטיות – בדרך כלל אין פעילות מסיבית, ולכן זאת תופעה פוטנציאלית שעלולה לקרות בזמן של חדירה.

חיזוי גרפי:

* + - 1. הפרש הזמנים בין יצירת הקובץ לשימוש בקובץ.  
         הפרש זמנים קטן בין יצירת הקובץ עלול להעיד על קובץ חדש שנכנס למערכת ויזם תקשורת עם רכיב נוסף ברשת המקומית  
           
         חיזוי גרפי

**מודל – נתונים נאספים**הפרש הזמנים בין יצירת הקובץ לשימושו (בזמן מנורמל), למספר זיהוי חבילה.

**סימולציה – אוסף נתונים**

סימולציה נורמלית

הרצת תעבורה בתוכנות:

* NormalChatClient22.exe
* NormalChatClient23.exe
* NormalChatClient25.exe
* NormalChatClient80.exe

בכל 1-10 שניות ירוצו הקבצים לעיל.

סימולציות אנומליות

הרצת קובץ חדש שנרשם למערכת ומשתמש בפורט מספר 4445 להעברת מידע למחשב השני.

שם הקובץ יהיה MaskJPG.jpg אשר מכיל קובץ הרצה EXE.

**למה זו יכולה להיחשב חריגה**

* הפרש הזמנים בין יצירת הקובץ לשימוש בקובץ יכול להיחשב כחריג על סמך קטריון זמן אשר נמוך מהממוצע.

**על מה היא מעידה?**

* קובץ חדש שנוצר במחשב ויוצר תקשורת עם מחשבים אחרים ברשת הביתית, דבר אשר עלול להעיד על ניסיון התפשטות של נוזקה.

לצורך הדוגמה: תוקף או נוזקה שמנסה להשתמש בשיטת Lateral movement כדי להגיע לעוד רכיבים שונים.

|  |  |
| --- | --- |
| **קטגוריות** | **Features** |
| כמות חבילות ליחידת זמן | העברת קובץ יחיד בנפח שגדול פי 1000 מהסימולציה הנורמאלית. |
| העברת 100 קבצים בגודל של 10 Bytes בפרק זמן בין 0.1-2 שניות. |
| כמות הודעות ל- Flow | הרצת תעבורה בפורט שלא היה קיים קודם לכן לדוגמה פורט מספר 4444. |
| העברת רצף מחרוזות יחיד בגודל של פי 10 מהסימולציה הנורמלית. |
| העברת 100 רצפי מחרוזות בגודל של מילה יחידה בפרק זמן בין 0.1-2 שניות. |
| שליחת רצף מחרוזת יחיד בגודל של אות בודדת. |
| סריקת פורטים - שונים בין 4440-4540 |
| גודל (bytes) חבילה ליחידת זמן | העברת חבילות בנפח שגדול פי 100 מהסימולציה הנורמאלית. |
| מתקפת SynFlood - העברת 100 חבילות בגודל בפרק זמן בין 0.1-4 שניות. |
| הפרש הזמנים בין יצירת הקובץ לשימוש בקובץ | הרצת קובץ חדש שנרשם למערכת ומשתמש בפורט מספר 4445 |

סיכום Features:

## דרישות פונקציונאליות - כללי

### איסוף המידע

**תיאור פונקציונאלי כללי:** התהליך מתרחש באופן תמידי על ידי סוכן שיושב בעמדת הקצה.

ניטור המידע יהווה האזנה לרכיב התקשורת הפועל ברשת המקומית, על ידי בדיקת חיבורים (Local Area Network).

כל חבילה מנוטרת מתפרסרת למידע המתאים למערכת, ונרשמת לאחר עיבוד למאגרי המידע "packetsCountTenSecond" ו"Flows" בשרת ה-DB. בנוסף כל חבילה נרשמת לקובץ לוג file.

הקושי במימוש : הינו האזנה לתקשורת הפנימית ברשת ובחירת המידע הרלוונטי.

### עיבוד הנתונים

מצב למידה:

**תיאור פונקציונאלי HighLevel:** התהליך מתרחש בכל פרק זמן בשרת, על ידי טבלאות VIEW אשר מעבדות את הנתונים שנמצאים בתוך מאגרי המידע ה"Flows" ו""packetsCountTenSecond לתוך מאגרי ה- VIEWs :

PassiveAverageFlow"" ו – "PassiveAverageCountTenSecond".

מצב אקטיבי:

**תיאור פונקציונאלי HighLevel:** התהליך מתרחש בכל פרק זמן בשרת, על ידי טבלאות VIEW אשר מעבדות את הנתונים שנמצאים בתוך מאגרי המידע ה"Flows" ו""packetsCountTenSecond לתוך מאגרי ה- VIEWs :

ActiveAverageFlow "" ו – ""ActiveAverageCountTenSecond.

הקושי במימוש: הינו הפקת DATASET מתאים בהתאם למאגרי המידע השונים, נוסף על כך יצירת Flow מתוך איסוף החבילות.

### ויזואליזציה

**תיאור פונקציונאלי כללי:** התהליך מתרחש בעת הכניסה למערכת אשר יהווה את מסך הDashboard.

מסך הDashboard יכיל את מאגרי ה ה"Flows" ו""packetsCountTenSecond בצורות גרפיות.

בנוסף לכך, מסך הDashboard יכיל את טבלת ה"חריגים"

לצד המסך יהיה כפתור למסך כלל הגרפים במערכת, בעת לחיצתו יפתח מסך הגרפים אשר ישקף את הFeature-ים שנבחרו.

הקושי במימוש: יצירת ממשק Web שמושך את המידע מהמאגרים.

### זיהוי חריגות

**תיאור פונקציונאלי כללי:** התהליך מתרחש לאחר הזנת הנתונים למאגרי ה – Features. שימוש באלגוריתם אשר יבדיל בין הפעולות הנורמליות לפעולות החריגות. רישום במאגר "חריגים" את הנתונים שהתקבלו מהאלגוריתם לזיהוי החריגות.

הקושי במימוש : בחירת מודל מתאים בהתאם לFeatures.

## דרישות פונקציונאליות - מפורט

### איסוף המידע

**תיאור פונקציונאלי כללי:** התהליך מתרחש באופן ייזום על ידי המערכת בעת עליית מערכת ההפעלה על ידי סוכן שנמצא בעמדה.

ניטור המידע יהווה האזנה לרכיב התקשורת הפועל ברשת המקומית, על ידי בדיקת חיבורים (Local Area Network).

כל חבילה מנוטרת מתפרסרת למידע המתאים למערכת, נרשמת למאגר בזיכרון אשר מחשב בכל 10 שניות את כמות החבילות, וכמות הFlags, נרשמת בזיכרון למאגר Flows ולבסוף למאגרי המידע "packetsCountTenSecond" ו"Flows" בשרת ה-DB.

בנוסף כל חבילה נרשמת לקובץ לוג file בשם ייחודי (Logfile[Unix-Time]) בעמדת הקצה.

הקושי במימוש : הינו האזנה לתקשורת הפנימית ברשת ובחירת המידע הרלוונטי.

**משתמשי התהליך:** אחראי הרשת המקומית.

**הטריגר לתהליך:** עליית מערכת ההפעלה

**הקלטים של התהליך:**

1. עליית מערכת ההפעלה
2. הסנפת המידע הכללי מהחיבור הזמין – Local Area Network.
3. פרסור המידע לפי השדות להלן:

|unixTimeMicrosec | Time | Date | Day | TimePeriod | srcIpAddr | dstIpAddr | IpLen | ID | Ttl | IPChksum | srcPort | dstPort | Service | Seq | Ack | dataOfs | Flags | Window | TcpChksum |

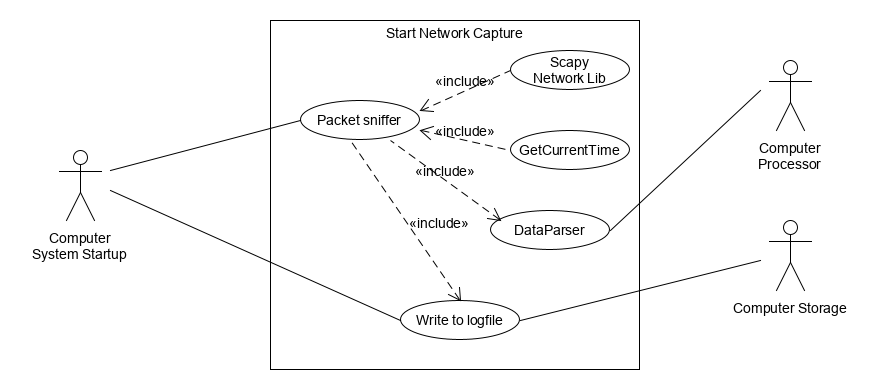
**הפלטים של התהליך:**

1. פרסור חבילות.
2. יצירת קובץ log file.
3. נרשמים כל הרשומות לקובץ לוג file בשם ייחודי (Logfile[Unix-Time]) בעמדת קצה.

**הנתונים המעורבים בתהליך:**

כתיבת נתונים למאגרי מידע רלוונטים.

UML Diagram



### עיבוד הנתונים

* + - * 1. Feature Packets in 10 seconds

**תיאור פונקציונאלי כללי:** התהליך מתרחש כאשר מתווספת חבילה חדשה, כתוצאה מכך מעבדים את החבילה לתוך מאגר ""packetsCountTenSecond ומונים את כמות החבילות, את הFlags השונים ואת גודל החבילות.

כל יום שבו מתבצע ניטור הרשת, תירשם למאגר View) "AverageCountTenSecond") גם הימים והשעות לצורך מעקב.

**משתמשי התהליך:** שרת הDB

**הטריגר לתהליך:** התהליך מתחיל כשנרשמת חבילה וכל 10 שניות.

**הקלטים של התהליך:**

1. רישום חבילה במערכת.
2. עיבוד המידע.
3. רישום למאגר הרלוונטי.

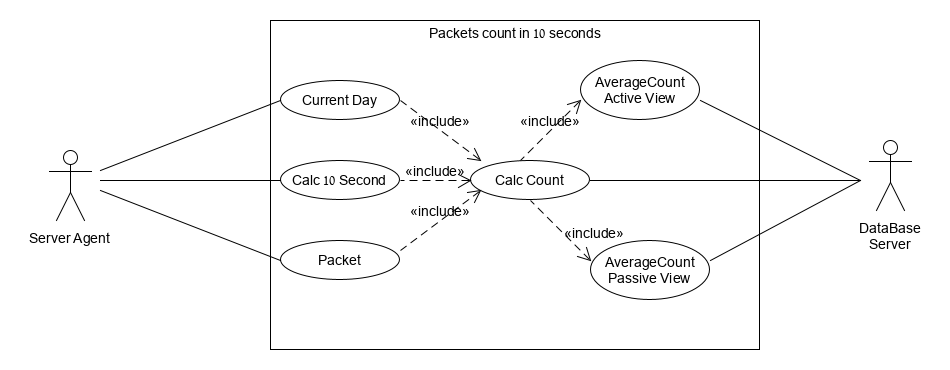
Feature 1 – packetsCountTenSecond | 10 second priod

(PacketCountTenID | TimePeriod | startTimeUnixMillisec | endTimeUnixMillisec | Day | packetCount | counterOfSy | counterOfAck | counterOfPa | counterOfR | counterOfRA | counterOfFin | packetsTotalSize)

**הפלטים של התהליך:**

כתיבת נתונים למאגרי מידע רלוונטים.

UML Diagram



SQL VIEW TABLE:

SELECT TimePeriod,

AVG([packetCount]) as [packetCount]

,AVG([counterOfSyn]) as [counterOfSyn]

,AVG([counterOfAck]) as [counterOfAck]

,AVG([counterOfPa]) as [counterOfPa]

,AVG([counterOfR]) as [counterOfR]

,AVG([counterOfRA]) as [counterOfRA]

,AVG([counterOfFin]) as [counterOfFin]

,AVG([PacketsTotalSize]) as [packetsTotalSize]

FROM [FinalProject].[dbo].[packetsCountTenSecond]

group by TimePeriod

השאילתה מציגה את ממוצעי הערכים לכל הימים שעברו בשבוע.

* + - * 1. חישוב מספרי בFLOWS

**תיאור פונקציונאלי כללי:** התהליך מתרחש בכל חבילה שעוברת כאשר מונים ומעדכנים את הFlow בהתאם לפרוטוקול תקשורת, ליוזם השיחה ולפורטים הרלוונטים בשיחה. אם מאגר הנתונים Flows ריק, נוסיף אליו את החבילה האחרונה שהתקבלה.

אחרת, נאסוף את החבילות הקשורות לFlow נמנה את כמויותהם, את כמות הFlags: הSynFlag,AckFlag,PayloadFlag,FinFlag ואת גודל הFlow, אחרי החבילה האחרונה במאגר ה - Flows ונספור את כמות החבילות, כמות

זיהוי הFlow יתבצע על ידי בדיקת יוזם השיחה – srcIpAddr, הSynFlag סיומו על ידי בדיקת FinFlag כאשר הSrcPort שווה לSrcPort או לdstPort שהיה קודם לכן.

* כאשר מתקבלת חבילה עם SynFlag – סימן שהתחיל ה – Flow בהינתן srcPort שונה.
* כאשר מתקבלת חבילה עם FinFlag – סימן שה - Flow הסתיים בהינתן srcPort שונה.

כל תאריך שבו מתבצע ניטור הרשת נרשם למאגר הימים והשעות לצורך מעקב.

**משתמשי התהליך:** שרת ה DB

**הטריגר לתהליך:** חבילה חדשה שמתקבלת במערכת.

**הקלטים של התהליך:**

1. רצף חבילות.
2. עיבוד המידע לפי המאגרים שלהלן:

Feature 2 – ServiceName | CountOfPackets and Flags counting

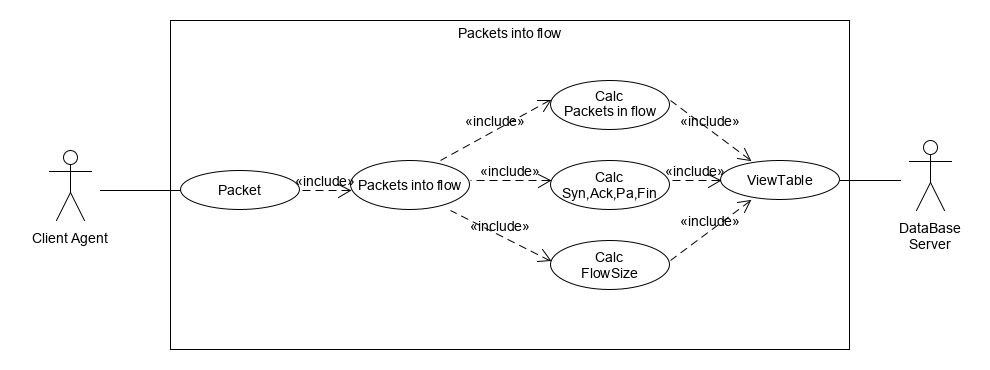
(FlowID | UnixStartTimeMillisec | UnixEndTimeMillisec | Time | Day | srcIpAddr | dstIpAddr | Service| srcPort | dstPort | counterOfPackets | counterOfSyn | counterOfPa | counterOfR | counterOfRA | counterOfFin | packetsTotalSize)

**הפלטים של התהליך:**

כתיבת נתונים למאגרי מידע רלוונטים.

**הנתונים המעורבים בתהליך:**

המידע המפורסר של החבילה.

UML Diagram

* + - * 1. מציאת קבצי הרצה

**תיאור פונקציונאלי כללי:** התהליך מתרחש באופן ייזום על ידי המערכת בעת עליית מערכת ההפעלה על ידי סוכן שנמצא בעמדה.

לכל חבילה מנוטרת, מזהים אם ישנה שיחה במצב ESTABLISHED לפי מספר srcPort ו- dstPort וגם בהתאם כתובות ה srcIpAddr,dstIpAddr. לכל שיחה כזו שולפים את הPID ולפיו מקבלים את שם הקובץ שהורץ ואת ניתובו בעמדת הקצה.

בנוסף לכך, מחשבים את הפרש הזמנים בין זמן יצירת הקובץ לזמן ההרצה של הקובץ.

**משתמשי התהליך:** שרת הDB

**הטריגר לתהליך:** עליית מערכת ההפעלה

**הקלטים של התהליך:**

פרסור המידע לפי השדות להלן:

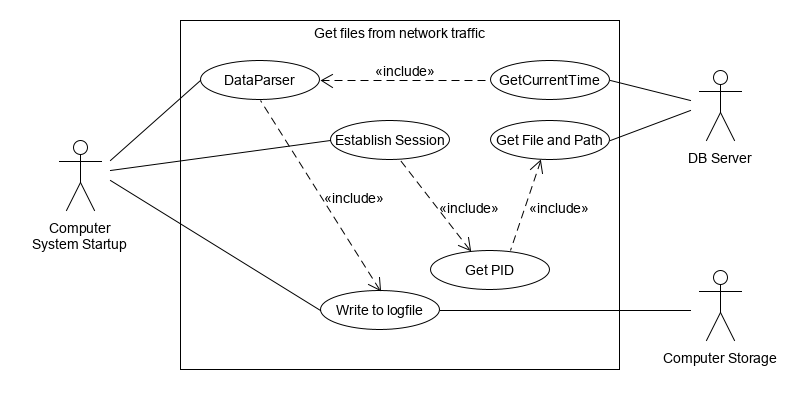
| unixTimeMicrosec | PID | filename | filePath | deltaTimeMillisec|

**הפלטים של התהליך:**

1. פרסור חבילות.
2. רשימה לקובץ log file.
3. נרשמים כל הרשומות לקובץ לוג file בשם ייחודי (Logfile[Unix-Time]) בעמדת קצה.

**הנתונים המעורבים בתהליך:**

כתיבת נתונים למאגרי מידע רלוונטים.

UML Diagram

### ויזואליזציה

**תיאור פונקציונאלי כללי:** התהליך מתרחש בעת כניסה לממשק WEB של המערכת.

מסך הDashboard יכיל את מאגרי ה ה"Flows" ו""packetsCountTenSecond בצורות גרפיות.

בנוסף לכך, מסך הDashboard יכיל את טבלת ה"חריגים"

לצד המסך יהיה כפתור למסך כלל הגרפים במערכת, בעת לחיצתו יפתח מסך הגרפים אשר ישקף את הFeature-ים שנבחרו.

הקושי במימוש: יצירת ממשק Web שמושך את המידע מהמאגרים השונים.

**משתמשי התהליך:** אחראי הרשת המקומית.

**הטריגר לתהליך:** התהליך מתחיל בעת כניסה למערכת.

**הקלטים של התהליך:**

מידע לפי המאגרים שלהלן:

כמות חבילות ליחידת זמן, כמות הודעות ל- Flow , גודל (bytes) חבילה ליחידת זמן, הפרש הזמנים בין יצירת הקובץ לשימוש בקובץ.

**הפלטים של התהליך:**

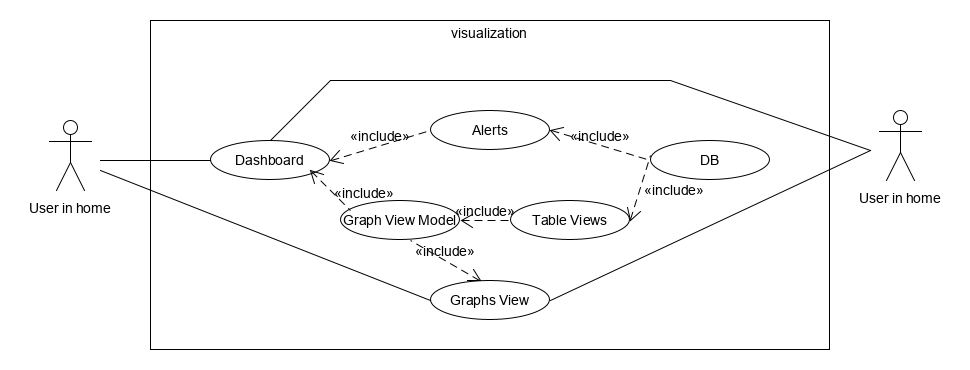
הצגת מאגרי מידע רלוונטים.

גרפים מופקים על ידי המאגרים.

הצגת טבלת "חריגים"

**הנתונים המעורבים בתהליך:**

מאגרי : packetsCountTenSecond , Flows, מאגר החריגות ו- מאגרי הVIEWS השונים.

UML Diagram

### זיהוי חריגות

**תיאור פונקציונאלי כללי:** התהליך מתרחש בכל פרק זמן קבוע של 30 שניות בצד השרת כאשר המערכת. התהליך יסמן את הערך "1" בכל שורה שבדק כדי שבכל פעם שירוץ הוא יעבור רק על רשומות עם שדות ריקים, בנוסף התהליך יסווג כל שורה בשדה הClassification אם היא חריגה או לא לפי המודל שיבחר על סמך ה - Features שנבחרו לעיל (יכולות להיות רשומות עם מספר features) ולכן יש צורך לשדה לכל Feature. במידה והרשומה פגעה בFeature מסויים יסומן הערך "1" על אותו השדה השייך ל- Feature. הסוכן יעשה שימוש באלגוריתם אשר יבדיל בין הפעולות הנורמליות לפעולות החריגות. לבסוף ירשום במאגר "חריגים" את הנתונים שהתקבלו מהתהליך לזיהוי החריגות.

**משתמשי התהליך:** סוכן בצד השרת.

**הטריגר לתהליך:** בכל פרק זמן של 30 שניות.

**הקלטים של התהליך:**

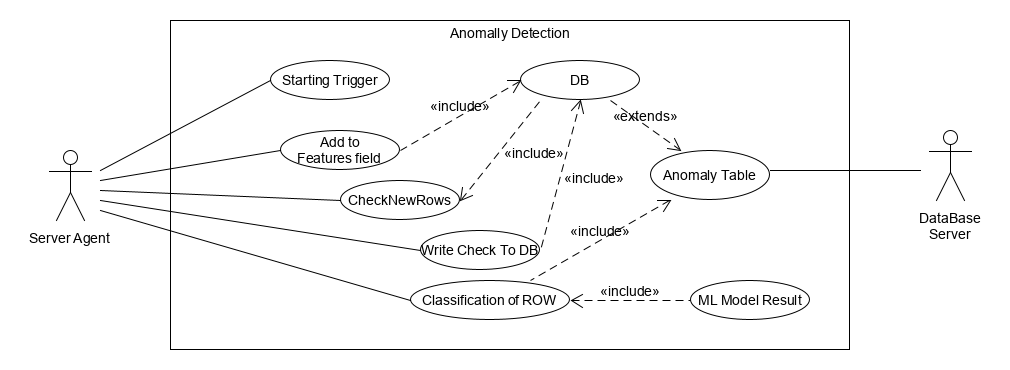
1. נתונים ממאגרי packetsCountTenSecond
2. נתונים ממאגרי Flows
3. נתונים ממאגרי הלמידה

**הפלטים של התהליך:**

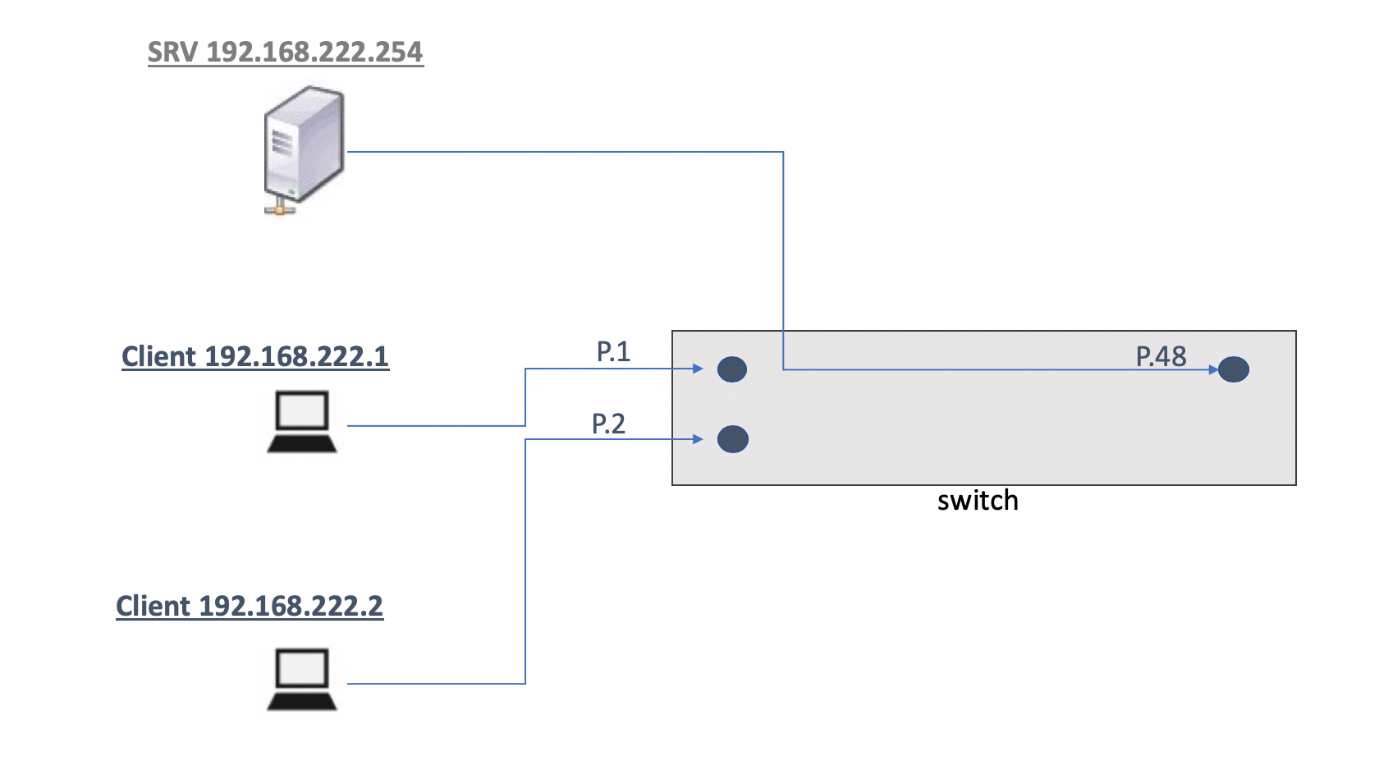
כתיבת נתונים למאגר החריגות.

**הנתונים המעורבים בתהליך:**

1. נתונים ממאגרי packetsCountTenSecond
2. נתונים ממאגרי Flows
3. נתונים ממאגרי הלמידה

UML Diagram

## תיאור ארכיטקטורה הכללית של המערכת



המערכת ממוקמת על שרת Windows Server 2019 עם Microsoft sql server 2014 ותתבצע תקשורת עם עמדות הקצה הממוקמות אצל הלקוח דרך הרשת הפנימית (LAN).

צד לקוח:

תהליך איסוף המידע יתחיל בצד הלקוח על ידי הסנפת חבילות בזמן אמת.

יצירת סימולציות (נורמליות, לא נורמליות) יבוצעו בצד הלקוח.

תהליך איסוף המידע יבוצע בשפת Python על ידי ספרייה שנקראתScapy .

צד שרת:

יכיל DB עם מאגרי מידע בהתאם לדרישות הפונקציונאליות ויאכלס בתוכו את המידע שנאסף על ידי הלקוח.

בנוסף, השרת יאפשר שירות WEB לויזואליות של המערכת.

שפת התכנות שבה נשתמש בשרת היא C# בטכנולוגית MVC , והDB שנשתמש בו יהיה MSSQL 2014.



אילוצים חשובים:

* 2 עמדות לצורך ביצוע הסימולציות.
* שרת אשר מסוגל לעבד נתונים בצורה מהירה.
* מתג תקשורת בכדי לאפשר תעבורה בין כלל הרכיבים של המערכת.

## הדרישות הלא פונקציונאליות

#######################

## חלוקת תפקידים

**הקמת מעבדה**

אלמוג, קובי וכנרת בעזרת פלד מנהל מערכות מידע במכללה.

**בחירת Features**

1. בחירת Features תוך שימוש במחקרים קיימים: קובי אלמוג וד"ר צחי.

**יצירת מודל**

1. יצירת מודל המבוסס על בחירת ה Features: אלמוג וד"ר צחי.

**צד לקוח**

* פיתוח סוכן מנהל את התהליך כולו- קובי
* פיתוח הסנפת/האזנה פקטות-אלמוג
* פיתוח כתיבה לDB – כנרת

בונוס- פיתוח הריגת תהליך וחסימה של תעבורה ספציפית.

**צד שרת**

* אפיון DB בטבלאות- כנרת ואלמוג
* יצירת פיתוח תקשורת מול סוכנים-קובי
* אינטראקציית אדם מחשב- אלמוג, קובי וכנרת.

# לוחות זמנים

## הספקת גרסאת אלפא למערכת

על הסטודנטים לעמוד בלוח זמנים צפוף אשר תהא תואמת את לוחות הזמנים הסימסטריאלים והלו"ז שנקבע על ידי מנחה הפרוייקט והינם כ-שלושה חודשים מתחילת העבודה   
וכן לעמוד בלוז הפורט להל"ן

## 

|  |  |
| --- | --- |
| שלב | זמן |
| מסמך דרישות מפורט | 9 שבעות |
| מסמך עיצוב תכנון | שבוע 12-13 |
| מסמך מפורט | תחילת סמסטר ב |
| PILOT | שבוע רביעי סמסטר שני |
| מוצר סופי | סוף שנת הלימודים. |

DB Raw Data

