Conestoga College

School of Applied Computer Science & Information Technology

SENG8081 - Case Studies Big Data

**Team Members:**

Paula Ramirez - 8963215

Vrunda Patel - 8989814

Pashmeen Kaur - 8992480

Jessica Christian - 8850898

Cyber Security Risk Analysis Using Big Data

18 October 2024

**Abstract**

This report gives an understanding of the implementation of big data techniques and solutions to handle cybersecurity risks. The main purpose of this is to analyze different threat categories such as, DDoS, Malware, Phishing, and Ransomware across various countries. As well some relevant information about each attack such as Attack vector, Geographical Location, Severity Score, indicators of compromise (IOCs) and other relevant factors.

To achieve this, we will apply big data methodologies, explore and categorize the data, to finally extract important conclusions on those cyber security risks. This document will detail the tools, process and techniques used in the analysis, following the six elements of big data. At the end, we will answer critical questions about cybersecurity trends and provide valuable insight about how to mitigate and prevent these threats.

**Table of contents**

Contents

[Introduction 3](#_Toc1427615030)

[Data Research and Integration 4](#_Toc2074857119)

[Data Collection 6](#_Toc850455093)

[Data Storage and Maintenance 6](#_Toc351022372)

[Data Quality 7](#_Toc803766124)

[Extension 8](#_Toc416858812)

[Proposed Allocation Project Team Roles 8](#_Toc586454681)

[Project Timeline 8](#_Toc182778089)

[References 10](#_Toc1159775206)

# Introduction

Since the digital era started cyber-attacks have existed and it has grown exponentially affecting security, privacy in multiples industries and companies and generating economic losses as well. To prevent this problem, it is a trend to analyze the different cyber-attacks to identify patrons and prevent futures threats. The Setracon Company that supports Enterprise Security Risk Management Operations said that before they used to wait for security problems before they did anything about them. This is called a reactive approach. (Slotnick, 2024).

This project is focused on implementing big data techniques to analyze various cybersecurity attacks. Through relevant data of cyber threat reports enhanced with Natural Language Processing (NLP), we will analyze some indicators, Threat Actor, Attack Vector, Geographical Location, and other relevant factors such as Severity Score, indicators of compromise (IOCs), and Risk Level Prediction.

Additionally, to cybersecurity data, we will integrate another monetary metrics to analyze possible correlations between monetary events such as GDP and emergence of cyber-attacks. This approach will allow us to create deep insight into how some external factors could influence the behavior of cyberattacks.

Finally with tools and defined process, we are seeking to categorize and conclude defense mechanism, given the decision makers valuable insight to mitigate risk and prevent futures attacks.

# Data Research and Integration

* **DATA RESEARCH**
* **NLP Based Cyber Security Dataset** **from Kaggle**

To carry out this project, we are considering the source **NLP Based Cyber Security Dataset** from Kaggle website. This dataset provides us with 1100 instances of cyber threat reports, including observations that were previously enhanced with NLP techniques to support research on Cyber Threat Intelligence (CTI) and give us pre-existing metrics with relevant information.

To understand this dataset, it is necessary to give a brief explanation of what it is NLP. NLP is a part of Artificial Intelligence focused on computers interactions and human language, which means that these kinds of machines can understand, interpret, and manipulate human language in a natural way. An example is detecting possible threats such as phishing or malware in emails; through these machine learning techniques, NLP can identify and classify emails if they use a suspicious tone or contain grammatical errors.

This dataset includes metrics related to NLP such as Risk Level Prediction, Predicted Threat Category, severity score, and other measures. The countries or region where the attack originated or was targeted and that we will analyze are USA, Russia, Global, North Korea, and Germany. (NLP Based Cyber Security Dataset, 2024).

A screenshot of a computer

Description automatically generated

Figure 1. Sample data Cyber Security

* **World Bank economic data (GDP per capita and Income Level).**

In addition to this data set, we are considering adding a library of python called “wbdata”. This library makes it easy to interact with World Bank databases to access economic metrics such as the income level and the Gross Domestic Product (GDP) of the countries.

A screenshot of a computer

Description automatically generated

Figure 2. Exploring Library Python

This integration could be useful to analyze demographic or economic data. By relating to cybersecurity incidents or the authors of the attacks in the affected countries with the economic stability of a region or the income level of the population; it is possible to conclude that attackers or weak economies could be influencing cyber-attacks.

A screenshot of a computer program

Description automatically generated

Figure 3. Exploring Library Python (2)

* **DATA INTEGRATION**

To integrate the cybersecurity datasets and economic data, a process was created to analyze the relationship between the severity of cyberattacks and the GDP per capita and income level metrics.

* **Strategy:**

The following steps describe the strategy used to carry out this analysis:

- Grouping the cybersecurity data and thus reaching the same level of granularity as the World Bank data source. A column called Standardized Country was created to unite the metrics from both datasets.

A screenshot of a computer

Description automatically generated

Figure 4 Grouping data

- After having the data integrated, the relationships of the metrics severity score, Risk Level Prediction and the number of cyberattacks with GDP and Income Level were analyzed. This correlation analysis was executed through the Python **corr()** function.

A screenshot of a computer

Description automatically generated

Figure 5 Correlations

- With sklearn, one of the most used libraries for machine learning in Python, a predictive model was created to analyze the influence of GDP and Income Level on the severity and quantity of cyberattacks.

A screenshot of a computer program

Description automatically generated

Figure 6 Linear Regression

* **Findings:**

During this process we faced challenges such as missing data for the country of North Korea and the difficulty of finding clear relationships between cybersecurity data and the economic metrics that we mentioned previously.

|  |  |  |
| --- | --- | --- |
| Correlation | Value | Interpretation |
| Correlation between GDP per capita and Severity Score | **-0.528**  **(52%)** | There is a moderate negative correlation between GDP per capita and the Severity Score.  This indicates that the higher the GDP per capita, the less severe the attacks tend to be. This leads to the hypothesis that these countries are better prepared to reduce the severity of the attacks. |
| Correlation between Income Level and Severity Score | **0.193**  **(19%)** | There is a weak positive correlation between Income Level and Severity Score, suggesting that countries with a higher income level may be associated with slightly more severe attacks. |
| Correlation between GDP per capita and Risk Level Prediction | **0.611**  **(61%)** | There is a moderate positive correlation. This indicates that the higher the GDP per capita, also face higher levels of risk prediction. |
| Correlation between Income Level and Risk Level Prediction | **0.519**  **(51%)** | There is a moderate positive correlation. This implies that countries with higher incomes also face higher levels of risk prediction. Attackers consider these countries to be high-value targets. |
| Linear Regression Model 'Risk Level Prediction' | **Coefficients: [8.09394661e-07 ,1.77573665e-02]** | The effect is very small, probably because GDP per capita has a high range |

* **Decisions:**

Following the analysis performed in Python, it was decided to remove the World Bank economic data source from our study:

- Correlations between economic metrics (GDP per capita and income level) did not prove to have a significant relationship with the cyberattack dataset.

- The sources did not have the same level of granularity. Economic metrics were aggregated at the country level, while cyberattack data had many observations per country.

- North Korea data was incomplete and several observations were grouped as “Global”.

- Regression model coefficients revealed the need for additional data preprocessing steps such as scaling to improve data interpretability and model performance.

A screenshot of a computer

Description automatically generated

Figure 7 Incomplete information



Figure 8 Global metrics

In our previous report, we anticipated that these challenges of missing relationships between metrics and data integrity issues might decide for the inclusion of this dataset. The analysis confirmed these findings and allowed us to make an informed decision to remove this source from subsequent steps of the project.

# Data Collection

At the same time that technology changes and some vulnerabilities are controlled, the attackers continue creating new attacks. To guarantee that our analysis is based on current and reliable information, it is necessary to implement an automatized process that recollected this information and to extract data periodically from sources.

* **EXTRACTION**

The main data of cybersecurity **NLP Based Cyber Security Dataset** that comes from Kagel website, could be updated in regular intervals (for example, trimestral or annual). For this, we will create a script on python that extract those data automatically and after that we integrated them to our storage system.

The script download and unzip the dataset from Kaggle directly local in Linux server. To access this service in a collaborative way, we will use the service directly from Google Cloud.

// Download data

wget https://www.kaggle.com/api/v1/datasets/download/hussainsheikh03/nlp-based-cyber-security-dataset

//unzip information

unzip nlp-based-cyber-security-dataset

A screen shot of a computer program

Description automatically generated

Figure 9 Script that download and unzip cyber dataset

Whit this automatic data recollection methodology, we will guarantee the quality and up to date data that we will use in our conclusion and analysis about cyber security attacks and their relationship with economic factors.

# Data Storage and Maintenance

Storage in a Big Data project is essential, so we must implement a strategy that allows us to store the data for later use. Our information does not contain sensitive data; however, it is necessary to consider the following:

* **UPDATING FREQUENCY:** The frequency with which we configure our extraction for the data source, “**NLP Based Cyber Security Dataset**” will be quarterly since we have seen that its update period is approximately this period.

A screenshot of a computer

Description automatically generated

Figure 10. Frequency Update Source NLP Cyber Security

* **STORAGE IN HFDS SYSTEM:** We have considered storing data in HDFS. This is an excellent option, especially, when considering the scalability, flexibility and tool use such Apache Spark. If the information grows in the future or we decide to include more sources, it could be the easiest best option because we wouldn’t limit our growth. Additionally, HDFS automatically handles data distribution, replication and fail tolerance.

The following code was added to script for storage the data set from local to Hadoop.

A screenshot of a computer

Description automatically generated

Figure 11 Storage in HDFS system

* **SIZE:** The information contains 1101 rows and the size is 249MB.

A screen shot of a computer

Description automatically generated

Figure 12 Size from HDFS

With all these details, we can conclude that the growth of the information is not accelerated.

* **DATA CLEANING WITH RDD PROGRAMMING IN APACHE SPARK**

It is crucial that this recollection process being enough flexible to adapt to new data sources that could be incorporated in the future. As well, data was preprocessed in Spark, where we included some rules to identify and management missing values and data errors; otherwise, we could affect our analysis.

A screenshot of a computer

Description automatically generated

Figure 13 Transformations in Apache Spark

* **STORING DATA IN HIVE**

We have created the schema in Hive to organize the data efficiently, enabling SQL-like queries over large data set stored in HDFS.

CREATE EXTERNAL TABLE IF NOT EXISTS cyber\_data\_pr (

Threat\_Category STRING,

IOCs ARRAY<STRING>,

Threat\_Actor STRING,

Attack\_Vector STRING,

Geographical\_Location STRING,

Sentiment\_in\_Forums FLOAT,

Severity\_Score INT,

Predicted\_Threat\_Category STRING,

Suggested\_Defense\_Mechanism STRING,

Risk\_Level\_Prediction INT,

Cleaned\_Threat\_Description STRING,

Keyword\_Extraction ARRAY<STRING>,

Named\_Entities ARRAY<STRING>,

Topic\_Modeling\_Labels STRING,

Word\_Count INT

)

ROW FORMAT DELIMITED FIELDS TERMINATED BY ','

LOCATION '/BigData\_PaulaRamirez/lab\_rdd'

TBLPROPERTIES('skip.header.line.count'='1');

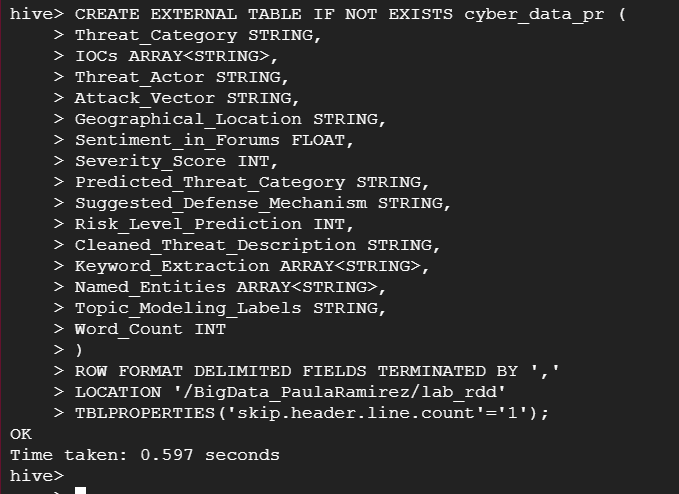


Figure 14 Creating Table on Hive

* **ANALYSIS ON HIVE**

**Sentiment Analysis**

Select threat\_actor, count(sentiment\_in\_forums)

from cyber\_data\_pr

group by threat\_actor ;

A screen shot of a computer

Description automatically generated

# Data Quality

Data comes from all sources and can be accurate or reliable to varying degrees. Data quality tried to track and maintain this. To maintain the data quality throughout the project we are going to use python. First, we are going to examine the structure of the dataset, data, columns and data types. Then we will check for missing values. There are different ways to handle missing values but mainly we will focus on two: 1. We will fill the missing values with mean, median or mode and 2. We will remove the rows and columns with too many missing values. We will check for the duplicate data and then remove them. We will ensure that all the data has correct data types. We will detect outliers using boxplots and remove the extreme outliers. We will also check for inconsistencies or incorrect data entries and correct them. We will ensure that the data is valid and has no negative values. To make sure that we are following each step of data quality we will document each data cleaning step.

Data Analysis and Visualization

We have decided to use Python and Tableau for data analysis and visualization. To analyze the data using aggregate functions and perform batch queries we will use Apache Hive. We will load the cleaned and processed data into tableau for creating interactive dashboards and visual reports to communicate insights effectively. Using python, we will do data manipulation and statistical analysis for which there are different libraries available such as Pandas, NumPy, Matplotlib and Statsmodels.

# Extension

The database we are using is continuously updating so right now its size is 240KB with 1100 records. So as per the data and regular updates, monthly growth of the data will be around 20% and at the end of a 1 year the data size would be increased by 2.13 MB. For this amount of data, we will use the same google cloud platform so we all can use it.

# Proposed Allocation Project Team Roles

We are a group of 4 people: Paula, Pashmeen, Jessica and Vrunda. First, we all discussed the tools that we have to use for analysis, cleaning, visualizing and communicating with the group. Vrunda will be responsible for planning all the steps and making sure that all the steps are going well according to the planning, and we are meeting the deadlines. Pashmeen has researched and decided the sources of the datasets. We all have looked into the different datasets from all the sources pashmeen selected and then came up with the final decision of using the NLP Based cyber security dataset. Paula and Jessica looked for other sources so we can find more insights. We all participated in writing this report for midterm project progress.

We are going to have regular team meetings so we can decide on work to do every week and the deadline to submit that. We will use GitHub to collaborate with each other and submit code.

# Project Timeline

|  |  |  |
| --- | --- | --- |
| **Date** | **Deliverable** | **Responsible** |
|  | Discussion of tools to use | Everyone |
| October 2 | Planning and responsibilities (each step) | Vrunda |
| October 8 | Decision of data sources websites | Pashmeen |
| October 9 | Choosing main dataset | Everyone |
| October 15 | Decision of other source to retrieve more insights | Jessica and Paula |
| October 18 | 1st Draft Circulated to Team | Everyone |
| October 19 | Final changes in the report and submission due by 10 PM | Vrunda, Paula |
| October 30 | Correlations Analysis | Paula Ramirez |
| November 12 | HiveSQL Analysis and Storage |  |

# References

Slotnick, J. A. (2024, September 24). *Security Data Intelligence is an organization’s Best Risk Management Tool*. Security Info Watch. <https://www.securityinfowatch.com/security-executives/article/55133716/security-data-intelligence-is-an-organizations-best-risk-management-tool>

*WBData*. (n.d.). *Welcome to wbdata's documentation!*  <https://wbdata.readthedocs.io/en/stable/>

Kempf, H. (2022, July 13). Public data for cybersecurity data science projects. *Medium*. <https://medium.com/@HunterKempf/public-data-for-cybersecurity-data-science-projects-3af6c0354025>

*NLP Based Cyber Security Dataset*. (2024, October 15). Kaggle. https://www.kaggle.com/datasets/hussainsheikh03/nlp-based-cyber-security-dataset