Silicon Tech Analytics

STA-CN-17031
Project

Submission instructions

- Cover sheet to be attached to the front of the project when submitted
- Project Questions to be attached to assignment when submitted
- All pages to be numbered sequentially

Project code	STA-CN-17031
Module title	CyberSecurity Dataset Big Data Analytics
Title	Big Data Analytics
Project number	031
Submission date	January 26, 2024
Additional enquiry	info@silicontechanalytics.com

STA-CN-17031 - Big Data Analytics

This Project is divided into two sections: (1) Big Data analytics on a real case study and (2) presentation.

Project evaluation comes from two main activities as follows:

1- Big Data Analytics report
2- Presentation (around 1000 words)

Big Data Analytics report

Projct Parts	Pts	Remarks *(breakdown of project for each sub-task)*	
Big Data Analytics using HIVE	30	*(10)*	Providing big data queries using HIVE.
		(10)	Using Built-in (Date, Math, Conditional, and String) Functions in HIVE.
		(10)	Visualizing the results of queries into the graphical representations and be able to interpret them
Big Data Analytics using Spark	50	*(15)*	Analyzing the dataset through statistical analysis methods.
		(35)	Designing single- and multi-class classifiers and evaluate and visualize the accuracy/performance.
Individual assessment	10	*(10)*	Find alternative solutions for high level languages and analytics approaches (**use references**), and Express findings from big data analytics with the relevant theories.
Documentation	10	*(10)*	Write down a scientific report.
	100		

Big Data Analytics using Hadoop and Spark

STA-CN-17031 – Big Data Analytics

Tasks:

(1) Understanding Dataset: UNSW-NB15

The raw network packets of the UNSW-NB15 dataset was created by the IXIA PerfectStorm tool in the Cyber Range Lab of the Australian Centre for Cyber Security (ACCS) for generating a hybrid of real modern normal activities and synthetic contemporary attack behaviours. Tcpdump tool used to capture 100 GB of the raw traffic (e.g., Pcap files). This data set has nine types of attacks, namely, *Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms*. The Argus and Bro-IDS tools are used and twelve algorithms are developed to generate totally 49 features with the class label.

a) The features are described here.
b) The number of attacks and their sub-categories is described here.
c) In this coursework, we use the total number of 10-million records that was stored in the CSV file (download). The total size is about 600MB, which is big enough to employ big data methodologies for analytics. As a big data specialist, firstly, we would like to read and understand its features, then apply modeling techniques. If you want to see a few records of this dataset, you can import it into Hadoop HDFS, then make a Hive query for printing the first 5-10 records for your understanding.

(2) Big Data Query & Analysis by Apache Hive

This task is using Apache Hive for converting big raw data into useful information for the end users. To do so, firstly understand the dataset carefully. Then, **make at least 4 Hive queries (refer to the table above)**. Apply appropriate visualization tools to present your findings numerically and graphically. Interpret shortly your findings.

Finally, take screenshot of your outcomes (e.g., tables and plots) together with the scripts/queries into the report.

Tip: The pts. for this section depends on the level of your HIVE queries' complexities, for instance using the simple *select* query is not supposed for full mark.

1 source: https://www.unsw.adfa.edu.au/unsw-canberra-cyber/cybersecurity/ADFA-NB15-Datasets/

(3) Advanced Analytics using PySpark

In this section, you will conduct advanced analytics using PySpark.

3.1. Analyze and Interpret Big Data

We need to learn and understand the data through **at least 4 analytical methods** (descriptive statistics, correlation, hypothesis testing, density estimation, etc.). You need to present your work numerically and graphically. Apply tooltip text, legend, title, X-Y labels etc. accordingly to help end-users for getting insights.

3.2. Design and Build a Classifier

a) Design and build a binary classifier over the dataset. Explain your algorithm and its configuration. Explain your findings into both numerical and graphical representations. Evaluate the performance of the model and verify the accuracy and the effectiveness of your model.

b) Apply a multi-class classifier to classify data into ten classes (categories): one normal and nine attacks (e.g., *Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode and Worms*). Briefly explain your model with supportive statements on its parameters, accuracy and effectiveness.

Tip: you can use this link (https://spark.apache.org/docs/2.2.0/ml-classification-regression.html) for more information on modelling.

(4) Individual Assessment

Discuss (1) what other alternative technologies are available for tasks 2 and 3 and how they are differ (use academic references), and (2) what was surprisingly new thinking evoked and/or neglected at your end?

Tip: add individual assessment of each member in a same report.

(5) Documentation

Document all your work. Your final report must follow 5 sections detailed in the **"format of final submission"** section (refer to the next page). Your work must demonstrate appropriate understanding of academic writing and integrity.

Points for the Presentation

Topic	Total Points	Remarks
Content	**50**	Covers topic in-depth with details.
Presentation design & layout features, Animations & transitions	**20**	Makes excellent use of fonts, colors, graphics, effects, transitions to enhance the presentation.
Length	**10**	Correct use of number of slides, Word Count (1000 words)?
Organization	**20**	The Data Analyst should present information in a logical, interestingsequence that the audience can follow.
	100	

This will be the second Submission which is located at a different submission link and here you will submit a presentation based on the report above. This will have a weight on conclusion of the project.

FORMAT OF FINAL SUBMISSION

- You need to prepare **one single file in PDF format** as your coursework within the following sections:
    1. Use **ONLY one** Cover Page
    2. Table of Contents
    3. Report of the tasks (it needs sub-sections for few tasks, accordingly)
    4. References (if any)

- And **one PDF file** for the presentation

SUBMISSION

single PDF into researcher@silicontechanalytics.com, by the end of Week 14