DEPARTMENT OF ENERGY TECHNOLOGY
AALBORG UNIVERSITY

PhD Public Defence

Title:	The Security Analysis of large Scale Streaming Data
Location:	AAU Esbjerg Campus, Room C1.119
Time:	Friday 10 August at 13.00
PhD defendant:	Sheeraz Niaz Lighari
Supervisor:	Associate Professor Dil Muhammad Akbar Hussain
Moderator:	Associate Professor Matthias Mandø
Opponents:	Associate Professor Daniel Ortiz-Arroyo, Department of Energy Technology, Aalborg University Esbjerg (Chairman) Professor Paolo Prinetto, Politecnico di Torino, Italy Associate Professor Dr. Sadiq Ali Khan, Chairman Department of Computer Science, University of Karachi, Pakistan

All are welcome. The defence will be in English.

After the public defence there will be an informal reception in the vestibule (C2) at Esbjerg Campus.

DEPARTMENT OF ENERGY TECHNOLOGY
AALBORG UNIVERSITY

Abstract:

This thesis, mainly involves the security investigation of large scale streaming data. To set up the ground, we began the examination of batch data utilizing KDDcup99 dataset. The KDDcup99 is widely used dataset by research community for threat detection. The project was started to analyse the KDDcup99 dataset using apache spark as a tool for security investigation, at that point it is examined by utilizing diverse machine learning algorithms like Support vector machine, Logistic relapse, Naïve bayes, Decision trees, Random Forest Tree which are cases of supervised algorithms. The project further utilized KMeans as an algorithm for unsupervised anomaly detection. The basic idea to analyze with various algorithms was to analyze them for finding the best algorithm for security analysis of extensive scale dataset. All these algorithms are incorporated as a part of spark machine learning library. In this project, we have further proposed a novel strategy for security investigation using hybrid model of rule based and clustering algorithm.

In the second part of the project, we investigate the streaming data created from the KDDcup99. A lot of work has been done in anomaly detection to the batch data yet recognizing oddities from streaming data by the by remains a generally accessible issue. In streaming data, the tasks related to find out the anomalies has become challenging with the passage of time because of the dynamic changes in data, which are produced by different methods applied in data streaming infrastructures. During the time spent on anomaly detection, above all else, it was first required to know the method for finding the normal conduct of information and afterward it was anything but difficult to know the dynamic conduct or change in the information. In this unique situation, clustering is an extremely noticeable strategy. The use of clustering technique is exceptionally basic to analyze the static information. Yet in the field of data mining, it is a key issue to analyze the streaming data. Therefore, the main focus of our project is to analyze the streaming data. For that, we are applying streaming form of KMeans clustering algorithm. The algorithm is analyzed both on single and distributed environment.

Besides as a use case to analyze the streaming data. We are exploring the latency value created by the sensors introduced in the Web Servers of Smart Metering Infrastructure. In our examination, Kafka Producer generates the sensor latency value. Kafka Consumer based on the Spark Streaming framework then analyzes the value. Moreover, in this project, we are also using the Gaussian distribution model to perceive the peculiarities in the sensor data. This model is widely used for anomaly detection.