Pipeline for Real-time Anomaly Detection in Log Data Streams using Apache Kafka and Apache Spark

Abstract

Anomaly detection is a standout amongst the most critical assignments so as to construct a system that is trustworthy and secure. The aim of anomaly detection is to detect significant deviation of the system behavior from that of the normal behavior. This approach is broadly used on static data, for instance on dumps of log data. Most systems require a real-time detection of anomalies with a specific end goal to lessen the harm that can be caused by the ignorance of an anomaly or detection at a later time. The recent implementations of the anomaly detection are mostly based on self-learning methods. Machine learning has brought about a significant transformation in the field of anomaly detection. One of the methodologies for anomaly detection depends on clustering algorithms. The implementation discussed in this paper utilizes a time-series evaluation approach for anomaly detection. The paper explains the pipeline built for anomaly detection and the visualization of the results.

References


Performance Prediction’, IEEE Second International Conference on Data Science in Cyberspace (DSC), Shenzhen, China, 26-29 June, 2017, pp. 500-504
20. Haitao Zhao, Shaoyuan Sun, Bo Jin, ‘Sequential Fault Diagnosis based on LSTM Neural Network’, IEEE Access, 30 Jan 2018, pp. 12929-12939
22. Qimin Cao, Yinrong Qiao, Zhong Lyu, ‘Machine learning to detect anomalies in web log analysis’, 3rd IEEE International Conference on Computer and Communications (ICCC), Chengdu, China, Dec 2017, pp. 519-523
31. Aniruddha Parvat, Jai Chavan, Siddhesh Kadam, Souradeep Dev, Vidhi Pathak, ‘A


Index Terms

Computer Science
Artificial Intelligence

Keywords

Anomaly detection, real-time, elastic stack, Long Short-Term Memory, Apache Spark