# Detection of Anomalies in Streaming Data using Streaming Analytics, Apache and Python: A Comparative Study.

Preeti Grover And

Sanjeev Kumar Prasad

School of Computer Science and Engineering ,Galgotias University Greater Noida, Uttar Pradesh, NCR, India preethi.grover\_phd18@galgotiasuniversity.edu.in

sanjeevkps2002@gmail.com

ORCID ID OF THE AUTHOR :- 0000-0002-0267-0783

#  Abstract:

 In the 21 century of digital era, there is an explosive growth of data from different sources such as: IOT Sensors, Securing Logs and web application, Stream Processing is part of Enterprise Data Architecture. Goal of the stream processing is to capture the unbounded stream of input data in real-time and process it with minimal latency to enable real-time consumption . Data streams, are an infinite and continuous flow of data from a source that arrives at a very high speed. Thus, streaming data is a subset of big data that addresses the velocity aspect of big data. In this paper we are discussing regarding the Anomaly Detection in Streaming Data Methods using Streaming Analytics and AI, comparing different Stream Processing Frameworks such as: Apache Spark Streaming, Apache Storm, Apache Flink, Apache Samza and Bytewax that is an open source, Python native, framework and distributed processing engine for processing data streams. Anomaly detection is the process of identifying these unusual patterns or behaviors in data. It is important because anomalies can indicate important events or problems, such as fraudulent activity, equipment failure, or security breaches. Social Network Analysis enables detection of an anomaly using K-Core and Neighborhood Metrics. Apache Spark provides a powerful platform for detecting network anomalies using big data processing and machine learning techniques. Distributed Streaming Architecture is the use of geographically distributed architectures for processing large data streams in real time , dedicated network of software components capable of ingesting and processing copious amounts of stream data from many sources as a stateful and stateless stream processing .

#  1. Introduction:

Streaming Analytics is the processing and analyzing of data records continuously rather than in batches. Data Streaming Processing allows organizations to deliver insights across massive datasets on a continuous basis.A number of Techniques, Algorithms and Framework exist and are used by Industries like: Google, Meta, Uber and many more, for detecting anomalies in streaming data pipelines. Anomalies such as: Json Schema change made by another department team that turns 6 columns into 600,is one such example of anomaly detection. In most scenarios, dynamic data is generated on a continuous basis and there streaming data processing is beneficial. Real-Time data streaming means collecting and ingesting a sequence of data from various data sources and processing that data in real time to extract its meaning. Data mining, website information analysis, data collection, and utilizing the information for business forecasting and product shaping are all components of social Network analytics. Evaluation of customer responses to support marketing progress and customer service decisions is the primary use of social media analytics. The emergence of big data from social media has brought about a new wave of excitement into the field of artificial intelligence and data analytics. Analyzing Stream Processing Techniques is better than traditional Data Mining Techniques for Processing huge volumes of Data. For instance, revealing market research information can be achieved through mining people’s opinions that results in improved business decision making.

 The most common applications of big data for social media are trend discovery, social media analytics, sentiment analysis, and opinion mining. Data must first be gathered and analyzed before anomaly detection can be used. The data may need to be cleaned up, organized, and the best algorithms and analysis approaches may need to be chosen. Statistical tools, machine learning algorithms, or other techniques are then used to identify the anomalies, which are subsequently reported or highlighted for more study. **Ex: sentiment-based analytics engine Simform built as an example of stream processing.**

**1(a)The Need to Process and Analyze Streaming Data:** High-velocity data from various sources (sensor data, stock-tick data, etc.) are ingested, processed, and analyzed by a streaming data analytics engine in real time or near real time. The output generated is displayed via dashboards, apps, or any other means**.**

****

**1(b)The Anomaly-Driven Process in Streaming Data are listed as below:**

**1)Analyze and visualize granular data:** Large data sets can be deconstructed using new analytical tools to reveal their underlying structure, enabling interpretation independent of averages and summary statistics. Semantic clustering algorithms can be used to analyse data such as consumer social media posts or competitor investment announcements, making the underlying patterns of thinking or behaviour explicit and understandable. Visualise the data and examine it to look for outliers and gaps in order to locate the abnormalities within these patterns.

 **2)Identify the anomalies that matter:** Numenta Anomaly Benchmark to Evaluate Real-time Anomaly Detectors. The Numenta Anomaly Benchmark (NAB) is a novel benchmark for evaluating algorithms for anomaly detection in streaming, online applications.

**3)Identify the Python Tools used for anomaly detection in streaming data:** **The PyOD library** is a comprehensive Python toolkit for detecting outlier observations in multivariate data.For Unsupervised anomaly detection in streaming data**, PySad Library is used**.

**4)Real-Time Data Streaming Tools:** These tools are :Apache Kafka, Apache storm, Amazon Kinesis, Azure Stream Analytics, Apache Flink, Apache Nifi , **Google Cloud Dataflow**, etc. Ex: In Google cloud environment we have 4 major components like: Ingest, Extract and Train, RealTime Prediction and Dashboard and Altering ,where we first ingest our unlabeled streaming data from multiple sources, than extract features using DataFlow and train Data and Model using K-MEANS Clustering Algorithm, in third step Data Flow will allow for real-time Online Reporting and lastly, we look at the dashboard to look at anomalies and alert .

**5)SNA Tools for Streaming Data Anomalies:** Streaming social network analysis collects real time data and creates an on-the-fly analysis of social relationships among people referred to in the stream data.

1. **Literature Review:** There are many Existing Anomaly Detection Algorithms and approaches to anomaly detection, and choosing the right one can be a challenging task.
	* One of the most widely used algorithms for anomaly detection is the **Local Outlier Factor (LOF)** algorithm.This algorithm uses the local density of points in a dataset to identify anomalies.We must first select a metric to gauge the density of points before we can utilise the LOF technique. The k-nearest neighbour (k-NN) distance, which calculates the separation between any point and its k closest neighbors, is the most popular option. Using this metric, the LOF algorithm then determines the local density of each point, classifying anomalous spots as those that have a local density that is noticeably lower than that of their neighbors.Another popular algorithm for anomaly detection is the Isolation Forest algorithm. This algorithm uses decision trees to identify anomalies, by isolating points that are difficult to reach in the decision tree.
	* Another well-liked method for anomaly identification is **one-class Support Vector Machines (SVMs).** This approach learns a decision boundary that separates the majority of the data from the anomalies using support vector machines. The fundamental tenet is that in most datasets, the vast majority of points are part of a single cluster, whereas anomalies make up a separate cluster. In order to identify points that are most likely to be anomalies, the One-class SVM algorithm leverages this attribute to train a decision boundary that divides the majority of the data from the outliers.
	* **The Elliptic Envelope algorithm is a final method for finding anomalies that relies on the notion that the data is regularly distributed.** This algorithm creates an ellipse out of the data and labels any locations outside of it as anomalous. The main premise is that most of the points in a dataset with regularly distributed data will fall inside the ellipse, whereas anomalies will fall outside of it. This characteristic is used by the Elliptic Envelope technique to locate anomalies by fitting an ellipse around the data and locating spots that fall outside of it.
	* **There is another Network Intrusion Detection Model with RNN** to explore new intrusion detection method in a Network Environment.
	* **Use of ML Algorithms to identify Outliners** like:Isolation Forest that attempts to isolate anomalies from rest of the data using an ensemble of decision trees. Due to its limitation like: it is based on unsupervised learning,where it requires no prior knowledge of the data or anomalies as what it is designed to detect.
	* **Use of Box Plot method for finding an outliner or anomaly detection** and In descriptive statistics, a box plot or boxplot is a method for graphically depicting groups of numerical data through their quartiles. To DROP the outlier which is due to incorrectly entered or measured data.

Below are some of the others Literature Articles Published with solutions to the Anomaly Detection In Streaming data as Below:-

 Table 1.1: Summarized Literature Review of Streaming Data Anomalies

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Author(s) |  | Definition |  Shortcomings |  |
| Gutierrez-Garcia, J.L., Sanchez-DelaCruz, E., Pozos-Parra, M.d.P. (2023) |  | Statisticial Based IDS/IPS |  |  |
| Shuvo, M.W.P., Hoq, M.N., Majumdar, S., Shirani, P. (2023). On Reducing Underutilization of Security Standards by Deriving Actionable Rules: An Application to IoT. |  | Knowledge Expert Based IDS/IPS |  |  |
|  |  |  |  |  |
| M.R. Amal, P. Venkadesh,H-DOCTOR: Honeypot based firewall tuning for attack prevention |  | HoneyPots  |  |  |
|  |  | Distributed IDS |  |  |
|  |  | Ml Based IDS/IPS |  |  |
| Lu T, Wang L, Zhao X. Review of Anomaly Detection Algorithms for Data Streams. Applied Sciences. 2023; 13(10):6353. <https://doi.org/10.3390/app13106353>  |  | Anomaly Detection Algorithms for Data Streams. |  |  |
| [Haolong Xiang](https://dl.acm.org/profile/99660602048) AND [Xuyun Zhang](https://dl.acm.org/profile/99660600949)Faculty of Science and *Engineering, Macquarie University, 2122, Sydney, Australia.* |  | Edge computing empowered anomaly detection framework |  |  |
|  |  |  |  |  |
| Streaming Linear Regression on Spark MLlib and MOACrossref DOI link: <https://doi.org/10.1145/2808797.2809374>Published Online: 2015-08-25Published Print: 2015-08-25 |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**Table 1. Anomaly Detection Techniques for various Applications**

**III. Methodology for streaming data anomalies detection :** Using Welford’s algorithm, a quartiles-based solution, a z-score metric-based solution, and a machine learning-based solution called Half-Space Trees (HST). The first 3 existing solutions are based on statistic indicators/metrics, while the last is from the machine learning field. Already there is a Framework called as Numenta Anomaly Benchmark (NAB) for Anomaly detection in streaming data.

**IV Proposed Tools ,Techniques and Algorithms for Anomaly Detection in Streaming Data:** Social network analysis (SNA) is quickly becoming a crucial tool for analyzing public sentiment and behavior. Social media posts are now being studied by researchers to forecast or attain outcomes. A comparative Study among the different tools for streaming data anomaly detection using Python and Social Network Analysis Tools are discussed and we are discussing why Python Tools are more useful than SNA Tools.

authenticity. It's crucial to spot anomalies on social media to stop bad behaviour like stalking, terrorist attack planning, and the dissemination of false information.

# V.Results and Discussions:

# VI.Conclusions:

 **Acknowledgement:** Financial support for this study was borne by Ms. Preeti Grover and there is no conflict of Interest in whatsoever manner between the authors. The authors wish to thank everyone who is involved in our peer reviewed and publication work.

**References:**

1. Priyanshi Singh, DeepikaAgrawal, SudhakarPandey et al. Anomaly detection and analysis in blockchain systems, 10 January 2023, PREPRINT (Version 1) available at Research Square [https://doi.org/10.21203/rs.3.rs-2414745/v1].
2. Intelligent Systems Design and Applications, 2021, Volume 1351,ISBN : 978-3-030-71186-3Hichem Dabbèchi, NahlaZaaboubHaddar, HaythamElghazel, Elyassami, S., Albloushi, S., Alnuaimi, M.A., Alhosani, O., Al Ali, H., Almarashda, K. (2022).
3. Intelligent Models for Mining Social Media Data. In: Saeed, F., Al-Hadhrami, T., Mohammed, E., Al- Sarem, M. (eds) Advances on Smart and Soft Computing. Advances in Intelligent Systems and Computing, vol 1399. Springer, Singapore. <https://doi.org/10.1007/978-981-16-5559-3_17>
4. Kim, Yoosin&Jeong, SeungRyul. (2015). Opinion-Mining Methodology for Social Media Analytics.KSII Transactions on Internet and Information Systems. 9. 391-406. 10.3837/tiis.2015.01.024.
5. Lua, Eng& Chen, Ruichuan&Cai, Zhuhua. (2011). Social Trust and Reputation in Online Social Networks. 2011. 811-816. 10.1109/ICPADS.2011.123.
6. Norjihan Abdul Ghani, Suraya Hamid, Ibrahim AbakerTargioHashem, Ejaz Ahmed, Big Social Media Data Analytics: A Survey, Computers in Human Behavior (2018), doi: 10.1016/j.chb.2018.08.039.
7. C. A. Clarke, E. Pfluegel and D. Tsaptsinos, "Multi-channel overlay protocols: Implementing ad-hoc message authentication in social media platforms," 2015 International Conference on Cyber Situational Awareness, Data Analytics and Assessment (CyberSA), 2015, pp. 1-6, doi: 10.1109/CyberSA.2015.7166118.
8. Praba, Lakshmi &GArumugam,.(2010). Message authentication code algorithm for IP-SEC. Computer Systems Science and Engineering. 25. 335-341.
9. Kazienko, P. (2018). Process of Social Network Analysis. In: Alhajj, R., Rokne, J. (eds) Encyclopedia of Social Network Analysis and Mining. Springer, New York, NY. https://doi.org/10.1007/978-1-4939- 7131-2\_244.
10. Chauhan, Tavishee and HemantPalivela. “Optimization and improvement of fake news detection using deep learning approaches for societal benefit.” Int. J. Inf. Manag. Data Insights 1 (2021): 100051.
11. Balaji T.K., Chandra SekharaRaoAnnavarapu, AnnushreeBablani,Machine learning algorithms for social media analysis: A survey,Computer Science Review,Volume 40,2021,100395,ISSN 1574- 0137,https://doi.org/10.1016/j.cosrev.2021.100395.
12. http://hdl.handle.net/10603/192361,Title of Thesis : Mining and Analysis of Social Networks Sites,AnupriyaJain:M K Sharma.
13. Hamzehi, M., Hosseini, S. Business intelligence using machine learning algorithms. Multimed Tools Appl 81, 33233–33251 (2022). https://doi.org/10.1007/s11042-022-13132-3.
14. https://learn.microsoft.com/en-us/power-bi/connect-data/service-tutorial-build-machine-learning-

model and Validating Binary Classification Machine Learning Models in Power BI (iterationinsights.com)

1. Xu, M., MacDonnell, M., Wang, A., & Elias, M. J. (2023). Exploring social emotional learning, school climate, and social network analysis. Journal of Community Psychology, 51, 84–102.
2. Meena, B., Sarwani, I.S.L., Archana, M., Supriya, P. (2020). Comparative Analysis of Apache Spark and HadoopMapReduce Using Various Parameters and Execution Time. In: Bhateja, V., Satapathy, S., Zhang, YD., Aradhya, V. (eds) Intelligent Computing and Communication. ICICC 2019.Advances in Intelligent Systems and Computing, vol 1034.Springer, Singapore.[https://doi.org/10.1007/978-981-15-](https://doi.org/10.1007/978-981-15-1084-7_70) [1084-7\_70](https://doi.org/10.1007/978-981-15-1084-7_70)
3. A statistical framework for handling network anomalies,https://dl.acm.org/doi/10.5555/3382225.3382380.
4. A. Rezaei, Z. M. Kasirun, V. A. Rohani and T. Khodadadi, "Anomaly detection in Online Social Networks using structure-based technique," 8th International Conference for Internet Technology and Secured Transactions (ICITST-2013), London, UK, 2013, pp. 619-622, doi: 10.1109/ICITST.2013.6750277.
5. M. Vouk, "Detecting and mitigating security anomalies," 16th IEEE International Conference on Tools with Artificial Intelligence, Boca Raton, FL, USA, 2004, pp. 4-, doi: 10.1109/ICTAI.2004.49.
6. H. C. Altunay, Z. Albayrak, A. N. Özalp and M. Çakmak, "Analysis of Anomaly Detection Approaches Performed Through Deep Learning Methods in SCADA Systems," 2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA), Ankara, Turkey, 2021, pp. 1-6, doi: 10.1109/HORA52670.2021.9461273.
7. Kumar S, Khan MB, Hasanat MHA, Saudagar AKJ, AlTameem A, AlKhathami M. An Anomaly Detection Framework for Twitter Data.Applied Sciences. 2022; 12(21):11059. <https://doi.org/10.3390/app122111059>.
8. Prarthi Jain, Seemandhar Jain, June 3, 2020, "Anomaly detection dataset", IEEE Dataport, doi: <https://dx.doi.org/10.21227/rt7n-2x60>.

22. http://hdl.handle.net/10603/303888,”Anomaly Detection and Analysis in Big Data” ,Garg, Sahil,Batra, Shalini,2018.

1. [Khamparia, A.](https://www.emerald.com/insight/search?q=Aditya%20Khamparia), [Pande, S.](https://www.emerald.com/insight/search?q=Sagar%20Pande), [Gupta, D.](https://www.emerald.com/insight/search?q=Deepak%20Gupta), [Khanna, A.](https://www.emerald.com/insight/search?q=Ashish%20Khanna) and [Sangaiah, A.K.](https://www.emerald.com/insight/search?q=Arun%20Kumar%20Sangaiah) (2020), "Multi-level framework for anomaly detection in social networking", [*Library Hi Tech*](https://www.emerald.com/insight/publication/issn/0737-8831), Vol. 38 No. 2, pp. 350-366. <https://doi.org/10.1108/LHT-01-2019-0023>.
2. A, K., S, P., D, G., A, K. & S, A. (2020). Multi-level framework for anomaly detection in social networking. Library Hi Tech, 38(2):350–366. doi: 10.1108/LHT-01-2019-0023.
3. Gutierrez-Garcia, J.L., Sanchez-DelaCruz, E., Pozos-Parra, M.d.P. (2023). A Review of Intrusion Detection Systems Using Machine Learning: Attacks, Algorithms and Challenges. In: Arai, K. (eds) Advances in Information and Communication. FICC 2023. Lecture Notes in Networks and Systems, vol 652. Springer, Cham. https://doi.org/10.1007/978-3-031-28073-3\_5.