DETEKSI PENCILAN PADA *DATA STREAM* POLUSI CAHAYA LANGIT MALAM DENGAN ALGORITMA EXACT-STORM

SKRIPSI

diajukan untuk memenuhi
salah satu syarat untuk memperoleh
gelar Sarjana Komputer

Oleh
Rahma Maulida
2003688

PROGRAM STUDI ILMU KOMPUTER
FAKULTAS PENDIDIKAN MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS PENDIDIKAN INDONESIA
2024

DETEKSI PENCILAN PADA *DATA STREAM* POLUSI CAHAYA LANGIT MALAM DENGAN ALGORITMA EXACT-STORM

Oleh
Rahma Maulida
NIM 2003688

Sebuah skripsi yang diajukan untuk memenuhi salah satu syarat dalam memperoleh gelar sarjana komputer di Fakultas Pendidikan Matematika dan Ilmu Pengetahuan Alam

RAHMA MAULIDA

2003688

DETEKSI PENCILAN PADA *DATA STREAM* POLUSI CAHAYA LANGIT MALAM DENGAN ALGORITMA EXACT-STORM

DISETUJUI DAN DISAHKAN OLEH PEMBIMBING:

Pembimbing I,

Prof. Dr. Lala Septem Riza, M.T.

NIP. 197809262008121001

Pembimbing II,

Dr. Judhistira Aria Utama, S.Si., M.Si.

NIP. 197703312008122001

Mengetahui,

Kepala Program Studi Ilmu Komputer

Dr. Muhammad Nursalman, M.T.

NIP. 197909292006041002

Deteksi Pencilan pada *Data Stream* Polusi Cahaya Langit Malam dengan Algoritma exact-STORM

Oleh
Rahma Maulida — rahma.m15@upi.edu
2003688

ABSTRAK

Polusi cahaya telah menjadi isu global yang menghalangi pengamatan astronomi dan mengganggu ekosistem. Data hasil sensor *Sky Quality Meter* (SQM) diambil secara sekuens sehingga menghasilkan data deret waktu yang sangat memungkinkan adanya pencilan saat pengamatan, sehingga deteksi pencilan dalam data SQM penting dilakukan untuk mengetahui apakah pencilan yang terdeteksi disebabkan oleh fenomena fisis atau gangguan dalam pengukuran. Penelitian berfokus pada pengembangan program deteksi pencilan berbasis *real-time streaming* menggunakan algoritma exact-STORM dengan *platform big data streaming* Apache Kafka. Untuk mencapai tujuan tersebut, metode yang digunakan dalam penelitian ini terdiri dari: (1) *data collection*, (2) perhitungan radius dengan metode Chebyshev, (3) *set up* Apache Kafka, dan (4) deteksi pencilan dengan algoritma exact-STORM. *Dataset* yang digunakan berasal dari Repositori Ilmiah Nasional (RIN) Dataverse yang dikumpulkan oleh Badan Riset dan Inovasi Nasional (BRIN) yang menyediakan data hasil sensor SQM dengan resolusi 1 menit yang dicatat secara kontinu setiap malam. Penelitian ini dapat membantu para ahli di bidang astrofisika untuk membuat program deteksi pencilan yang berbasis *real-time* secara *streaming*. Program ini secara efektif mendeteksi pencilan secara *real-time streaming* untuk data selama tujuh bulan dengan skor proporsi sebesar 51,11%, *error rate* sebesar 1,61%, dan akurasi sebesar 98,39%. Penelitian ini memvalidasi bahwa metode deteksi pencilan berbasis jarak dapat digunakan untuk mendeteksi pencilan secara *real-time* pada data kecerahan langit malam.

Kata Kunci: Apache Kafka, Astrofisika, *Data Stream*, exact-STORM, Polusi Cahaya, Deteksi Pencilan, *Streaming*

Outlier Detection in Night Sky Light Pollution Data Streams using the exact-STORM Algorithm

By

Rahma Maulida — rahma.m15@upi.edu

2003688

ABSTRACT

Light pollution has emerged as a global issue, obstructing astronomical observations and disrupting ecosystems. Data from the Sky Quality Meter (SQM) sensor results sequentially collected in time series data, making it a data stream and highly likely to have outliers. Therefore, outlier detection in SQM data is crucial to determine whether the detected anomalies are due to physical phenomena or measurement disturbances. This study focuses on developing a real-time streaming outlier detection program using the exact-STORM algorithm integrated with the Apache Kafka big data streaming platform. To achieve this objective, the research methodology comprises: (1) data collection, (2) radius calculation using the Chebyshev method, (3) Apache Kafka setup, and (4) outlier detection using the exact-STORM algorithm. The dataset used originates from the National Scientific Repository (RIN) Dataverse, collected by the National Research and Innovation Agency (BRIN), providing SQM sensor data recorded continuously every night with a resolution of 1 minute recorded continuously every night. This research can assist astrophysics experts in developing a real-time streaming outlier detection system. The program effectively detects outliers in real-time streaming for data over seven months, achieving a proportion score of 51.11%, an error rate of 1.61%, and an accuracy of 98.39%. This study validates that distance-based outlier detection methods can be effectively employed to identify outliers in real-time night sky brightness data.

Keywords: *Apache Kafka, Astrophysics, Data Stream, exact-STORM, Light Pollution, Outlier Detection, Streaming*

DAFTAR ISI

DAFTAR TABEL

DAFTAR GAMBAR

DAFTAR LAMPIRAN

DAFTAR PUSTAKA

Adewusi, A. O., Okoli, U. I., Adaga, E., Olorunsogo, T., Asuzu, O. F., & Daraojimba, D. O. (2024). Business intelligence in the era of big data: a review of analytical tools and competitive advantage. *Computer Science & IT Research Journal*, *5*(2), 415-431.

Aggarwal, C. C. (2017). *An introduction to outlier analysis* (pp. 1-34). Springer International Publishing.

Ahmad, S., Lavin, A., Purdy, S., & Agha, Z. (2017). Unsupervised real-time anomaly detection for streaming data. *Neurocomputing*, *262*, 134-147.

Akanbi, A. (2020, November). Estemd: A distributed processing framework for environmental monitoring based on apache kafka streaming engine. In *Proceedings of the 4th International Conference on Big Data Research* (pp. 18-25).

Akil, B., Zhou, Y., & Röhm, U. (2017, December). On the usability of Hadoop MapReduce, Apache Spark & Apache flink for data science. In *2017 IEEE International Conference on Big Data (Big Data)* (pp. 303-310). IEEE.

Anaf, J., Baum, F. E., Fisher, M., Harris, E., & Friel, S. (2017). Assessing the health impact of transnational corporations: a case study on McDonald's Australia. *Globalization and Health*, *13*, 1-16.

Angiulli, F., & Fassetti, F. (2007, November). Detecting distance-based outliers in streams of data. In *Proceedings of the sixteenth ACM conference on Conference on information and knowledge management* (pp. 811-820).

Aniello, L., Baldoni, R., & Querzoni, L. (2013, June). Adaptive online scheduling in storm. In *Proceedings of the 7th ACM international conference on Distributed event-based systems* (pp. 207-218).

Arasu, A., & Manku, G. S. (2004, June). Approximate counts and quantiles over sliding windows. In *Proceedings of the twenty-third ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems* (pp. 286-296).

Backhoff, O., & Ntoutsi, E. (2016, December). Scalable online-offline stream clustering in apache spark. In *2016 IEEE 16th International Conference on Data Mining Workshops (ICDMW)* (pp. 37-44). IEEE.

Ball, N. M., & Brunner, R. J. (2010). Data Mining and Machine Learning in Astronomy. *International Journal of Modern Physics D, 19*(7), 1049-1106. doi:10.1142/S0218271810017160.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (Vol. 3, No. 1). New York: Wiley.

Barringer, D., Walker, C. E., Pompea, S. M., & Sparks, R. T. (2011, September). Astronomy Meets the Environmental Sciences: Using GLOBE at Night Data. In *Earth and Space Science: Making Connections in Education and Public Outreach* (Vol. 443, p. 373).

Beck, K., Beedle, M., Van Bennekum, A., Cockburn, A., Cunningham, W., Fowler, M., ... & Thomas, D. (2001). Manifesto for agile software development.

Ben-Gal, I. (2005). Outlier detection. *Data mining and knowledge discovery handbook*, 131-146.

BenMark, G., Klapdor, S., Kullmann, M., & Sundararajan, R. (2017). How retailers can drive profitable growth through dynamic pricing. *McKinsey. com*.

Berry, R. L. (1976). Light pollution in southern Ontario. *Journal of the Royal Astronomical Society of Canada, Vol. 70, p. 97*, *70*, 97.

Bhatia, S., & Kumar, R. (2018). Review of graph processing frameworks. In *2018 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 998-1005). IEEE.

Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys (CSUR)*, *54*(3), 1-33.

Bloom, J. S., Richards, J. W., Nugent, P. E., et al. (2012). Automating discovery and classification of transients and variable stars in the synoptic survey era. *Publications of the Astronomical Society of the Pacific, 124*(921), 1175-1196. doi:10.1086/668468

Bock, R. K., Krischer, W., Bock, R. K., & Krischer, W. (1998). *The data analysis briefbook* (pp. 1-181). Springer Berlin Heidelberg.

Bolton, R. J., & Hand, D. J. (2001). Unsupervised profiling methods for fraud detection. Credit scoring and credit control VII, 235-255.

Boukerche, A., Zheng, L., & Alfandi, O. (2020). Outlier detection: Methods, models, and classification. *ACM Computing Surveys (CSUR)*, *53*(3), 1-37.

Breunig, M. M., Kriegel, H. P., Ng, R. T., & Sander, J. (2000, May). LOF: identifying density-based local outliers. In *Proceedings of the 2000 ACM SIGMOD international conference on Management of data* (pp. 93-104).

Carbone, P., Katsifodimos, A., Ewen, S., Markl, V., Haridi, S., & Tzoumas, K. (2015). Apache Flink™: Stream and Batch Processing in a Single Engine. *Bulletin of the IEEE Computer Society Technical Committee on Data Engineering, 38*(4), 28-38.

Carvalho, O., Roloff, E., & Navaux, P. O. (2017, December). A distributed stream processing based architecture for IoT smart grids monitoring. In *Companion proceedings of the10th international conference on utility and cloud computing* (pp. 9-14).

Casella, G., Fienberg, S., & Olkin, I. (2006). *Springer Texts in Statistics*. Design (Vol. 102), http://books. google.com/books?id=9tv0taI8l6YC.

Cavazzani, S., Ortolani, S., Bertolo, A., Binotto, R., Fiorentin, P., Carraro, G., ... & Zitelli, V. (2020). Sky Quality Meter and satellite correlation for night cloud-cover analysis at astronomical sites. *Monthly Notices of the Royal Astronomical Society*, *493*(2), 2463-2471.

Chen, M., Mao, S., & Liu, Y. (2014). Big data: A survey. *Mobile networks and applications*, *19*, 171-209.

Cheng, D., Zhou, X., Wang, Y., & Jiang, C. (2018). Adaptive scheduling parallel jobs with dynamic batching in spark streaming. *IEEE Transactions on Parallel and Distributed Systems*, *29*(12), 2672-2685.

Cinzano, P. (2005). Night sky photometry with sky quality meter. *ISTIL Int. Rep, 9*(1).

Cinzano, P., Falchi, F., Elvidge, C. D., & Baugh, K. (2000). The artificial night sky brightness mapped from DMSP satellite Operational Linescan System measurements. *Monthly Notices of the Royal Astronomical Society*, *318*(3), 641–657. https://doi.org/10.1046/j.1365-8711.2000.03562.x.

De Miguel, A. S., Aubé, M., Zamorano, J., Kocifaj, M., Roby, J., & Tapia, C. (2017). Sky Quality Meter measurements in a colour-changing world. *Monthly Notices of the Royal Astronomical Society*, *467*(3), 2966–2979. https://doi.org/10.1093/mnras/stx145.

Djenouri, Y., Belhadi, A., Lin, J. C. W., Djenouri, D., & Cano, A. (2019). A survey on urban traffic anomalies detection algorithms. *IEEE Access*, *7*, 12192-12205.

Edgeworth, F. Y. (1887). Xli. on discordant observations. *The london, edinburgh, and dublin philosophical magazine and journal of science*, *23*(143), 364-375.

Esling, P., & Agon, C. (2012). Time-series data mining. *ACM Computing Surveys (CSUR)*, *45*(1), 1-34.

Espey, B., & McCauley, J. (2014). Initial Irish light pollution measurements and a new Sky Quality Meter-based data logger. *Lighting Research & Technology*, *46*(1), 67-77.

Faid, M. S., Shariff, N. N. M., Hamidi, Z. S., Kadir, N., Ahmad, N., & Wahab, R. A. (2018). Semi empirical modelling of light polluted twilight sky brightness. *Jurnal Fizik Malaysia*, *39*(2), 30059-30067.

Falchi, F., Cinzano, P., Duriscoe, D. M., Kyba, C. C. M., Elvidge, C. D., Baugh, K., Portnov, B. A., Rybnikova, N., & Furgoni, R. (2016). The new world atlas of artificial night sky brightness. *Science Advances*, *2*(6). https://doi.org/10.1126/sciadv.1600377.

Fouladirad, M., Neal, J., Ituarte, J. V., Alexander, J., & Ghareeb, A. (2018). Entertaining data: business analytics and Netflix. *Int J Data Anal Inf Syst*, *10*(1), 13-22.

Fujimaki, R., Yairi, T., & Machida, K. (2005). An approach to spacecraft anomaly detection problem using kernel feature space. *In Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining* (pp. 401-410).

Gandomi, A., & Haider, M. (2015). Beyond the hype: Big data concepts, methods, and analytics. *International Journal of Information Management, 35*(2), 137-144. doi:10.1016/j.ijinfomgt.2014.10.007.

Garcia-Teodoro, P., Diaz-Verdejo, J., Maciá-Fernández, G., & Vázquez, E. (2009). Anomaly-based network intrusion detection: Techniques, systems and challenges. *computers & security, 28*(1-2), 18-28.

Garg, N. (2013). *Apache kafka* (pp. 30-31). Birmingham, UK: Packt Publishing.

Giles, D. K., & Walkowicz, L. (2020). Density-based outlier scoring on Kepler data. *Monthly Notices of the Royal Astronomical Society*, *499*(1), 524-542.

Golab, L., & Özsu, M. T. (2003). Issues in data stream management. *ACM Sigmod Record*, *32*(2), 5-14.

Green, R. F., Luginbuhl, C. B., Wainscoat, R. J., & Duriscoe, D. (2022). The growing threat of light pollution to ground-based observatories. *The Astronomy and Astrophysics Review, 30*(1), 1.

Grubbs, F. E. (1969). Procedures for detecting outlying observations in samples. *Technometrics*, *11*(1), 1-21.

Guo, Y., Rao, J., Cheng, D., & Zhou, X. (2016). ishuffle: Improving hadoop performance with shuffle-on-write. *IEEE transactions on parallel and distributed systems*, *28*(6), 1649-1662.

Gwadera, R., Atallah, M. J., & Szpankowski, W. (2005). Reliable detection of episodes in event sequences. *Knowledge and Information Systems, 7,* 415-437.

Hadida, A. L., Lampel, J., Walls, W. D., & Joshi, A. (2021). Hollywood studio filmmaking in the age of Netflix: a tale of two institutional logics. *Journal of Cultural Economics*, *45*, 213-238.

Han, J., Haihong, E., Le, G., & Du, J. (2011). Survey on NoSQL database. *Proceedings - 6th International Conference on Pervasive Computing and Applications*, 363-366. doi:10.1109/ICPCA.2011.6106531.

Han, S., Choi, W., Muwafiq, R., & Nah, Y. (2017, September). Impact of memory size on bigdata processing based on hadoop and spark. In *Proceedings of the international conference on research in adaptive and convergent systems* (pp. 275-280).

Hänel, A., Posch, T., Ribas, S. J., Aubé, M., Duriscoe, D., Jechow, A., ... & Kyba, C. C. (2018). Measuring night sky brightness: methods and challenges. *Journal of Quantitative Spectroscopy and Radiative Transfer*, *205*, 278-290.

Harinen, T., & Li, B. (2019). Using causal inference to improve the uber user experience. *Uber Engineering*, 1-11.

Hawkins, D. M. (1980). *Identification of outliers* (Vol. 11). London: Chapman and Hall.

Heigl, M., Anand, K. A., Urmann, A., Fiala, D., Schramm, M., & Hable, R. (2021). On the improvement of the isolation forest algorithm for outlier detection with streaming data. *Electronics*, *10*(13), 1534.

Hewage, T. N., Halgamuge, M. N., Syed, A., & Ekici, G. (2018). Big data techniques of Google, Amazon, Facebook and Twitter. *J. Commun.*, *13*(2), 94-100.

Hidayat, T., Mahasena, P., Dermawan, B., Hadi, T. W., Premadi, P. W., & Herdiwijaya, D. (2012). Clear sky fraction above Indonesia: an analysis for astronomical site selection. *Monthly Notices of the Royal Astronomical Society, 427*(3), 1903-1917.

Jafar, O. M., & Sivakumar, R. (2013). A study of bio-inspired algorithm to data clustering using different distance measures. *International Journal of Computer Applications, 66*(12).

Jiao, X., & Pretis, F. (2022). Testing the presence of outliers in regression models. *Oxford Bulletin of Economics and Statistics, 84*(6), 1452-1484.

Khan, S., Hussain, R. M., Baig, T. S., & Qasim, M. M. Ransomware Resilience: A Real-Time Detection Framework using Kafka and Machine Learning. *International Journal of Innovations in Science & Technology, 5*(4), 70-82.

Kim, E. J., & Brunner, R. J. (2016). Star-galaxy classification using deep convolutional neural networks. *Monthly Notices of the Royal Astronomical Society, 464*(4), 4463-4475. doi:10.1093/mnras/stw2672

Kleppmann, M., & Kreps, J. (2015). Kafka, samza and the unix philosophy of distributed data.

Knorr, E. M., & Ng, R. T. (1998, August). Algorithms for mining distancebased outliers in large datasets. In *Proceedings of the international conference on very large data bases* (pp. 392-403). Citeseer.

Kokate, U., Deshpande, A., Mahalle, P., & Patil, P. (2018). Data stream clustering techniques, applications, and models: comparative analysis and discussion. *Big Data and Cognitive Computing, 2*(4), 32.

Kolajo, T., Daramola, O., & Adebiyi, A. (2019). Big data stream analysis: a systematic literature review. *Journal of Big Data*, *6*(1), 47.

Kou, Y., Lu, C. T., & Chen, D. (2006). Spatial weighted outlier detection. In *Proceedings of the 2006 SIAM international conference on data mining* (pp. 614-618). Society for Industrial and Applied Mathematics.

Kreps, J., Narkhede, N., & Rao, J. (2011, June). Kafka: A distributed messaging system for log processing. *In Proceedings of the NetDB* (Vol. 11, No. 2011, pp. 1-7).

Kumar, S., Tiwari, P., & Zymbler, M. (2019). Internet of Things is a revolutionary approach for future technology enhancement: a review. *Journal of Big data*, *6*(1), 1-21.

Kunszt, P. Z., Szalay, A. S., & Thakar, A. R. (2001). The hierarchical triangular mesh. In *Mining the Sky* (pp. 631-637). Springer. doi:10.1007/978-3-642-56758-4_74.

Kurniawaty, R., Hidayat, M., & Ananda, F. S. (2024). The Influence of MPSAS Values and SQM Angles in Determining Fajr Time in a Mathematical Review. *JMEA: Journal of Mathematics Education and Application, 3*(1), 31-38.

Laney, D. (2001). 3D data management: Controlling data volume, velocity and variety. *META group research note*, *6*(70), 1.

Luginbuhl, C. B., Walker, C. E., & Wainscoat, R. J. (2009). Lighting and astronomy. *Physics Today, 62*(12), 32-37.

Malewicz, G., Austern, M. H., Bik, A. J., Dehnert, J. C., Horn, I., Leiser, N., & Czajkowski, G. (2010, June). Pregel: a system for large-scale graph processing. In *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data* (pp. 135-146).

Narkhede, N., Shapira, G., & Palino, T. (2017). *Kafka: The definitive guide: Real-time data and stream processing at scale*. O'Reilly Media, Inc.

Niri, M. A., Zainuddin, M. Z., Man, S., Nawawi, M. S. A. M., Wahab, R. A., Ismail, K., ... & Lokman, M. A. A. (2012). Astronomical determinations for the beginning prayer time of isha'. *Middle-East Journal of Scientific Research, 12*(1), 101-107.

Noble, C. C., & Cook, D. J. (2003). Graph-based outlier detection. In *Proceedings of the 9th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM Press (pp. 631-636).

Olteanu, M., Rossi, F., & Yger, F. (2023). Meta-survey on outlier and anomaly detection. *Neurocomputing*, *555*, 126634.

Park, J., Seo, Y., & Cho, J. (2023). Unsupervised outlier detection for time-series data of indoor air quality using LSTM autoencoder with ensemble method. *Journal of Big Data*, *10*(1), 66.

Penny, K. I., & Jolliffe, I. T. (2001). A comparison of multivariate outlier detection methods for clinical laboratory safety data. *Journal of the Royal Statistical Society: Series D (The Statistician), 50*(3), 295-307.

Phua, C., Alahakoon, D., & Lee, V. (2004). Minority report in fraud detection: classification of skewed data. *Acm sigkdd explorations newsletter, 6*(1), 50-59.

Pribadi, P., Pramudya, Y., Muchlas, M., & Okimustava, O. (2019, December). The IoT implementation on the night sky brightness measurement in Banjar using the sky quality meter. In *AIP Conference Proceedings* (Vol. 2202, No. 1). AIP Publishing.

Priyatikanto, R. (2022) "SQM_2020_2021.dat", *Site characteristics: Timau National Observatory*, https://hdl.handle.net/20.500.12690/RIN/A5XCJB/9BWAUT, RIN Dataverse, V1.

Priyatikanto, R., Mumpuni, E. S., Hidayat, T., Saputra, M. B., Murti, M. D., Rachman, A., & Yatini, C. Y. (2023). Characterization of Timau National Observatory using limited in situ measurements. *Monthly Notices of the Royal Astronomical Society*, *518*(3), 4073-4083.

Putra, Z. A. Y. (2023). *Deteksi Anomali Realtime Menggunakan Probabilistic Exponential Weighted Moving Average pada Data Stream dengan Apache*

Kafka Studi Kasus: Analisis Polusi Cahaya [S1 thesis, Universitas Pendidikan Indonesia]. https://repository.upi.edu/105379/.

Raisal, A. Y., Pramudya, Y., Okimustava, O., & Muchlas, M. (2017). The moon phases influence on the beginning of astronomical dawn determination in Yogyakarta. In International *Journal of Science and Applied Science: Conference Series* (Vol. 2, No. 1, pp. 1-7).

Rajkhowa, R. (2014). Light pollution and impact of light pollution. international *Journal of Science and Research (IJSR), 3*(10), 861-867

Reddy, K. S. S., & Bindu, C. S. (2017, February). A review on density-based clustering algorithms for big data analysis. In *2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud)(I-SMAC)* (pp. 123-130). IEEE.

Reunanen, N., Räty, T., & Lintonen, T. (2020). Automatic optimization of outlier detection ensembles using a limited number of outlier examples. *International Journal of Data Science and Analytics*, *10*, 377-394.

Richard, B. (2017). Hotel chains: survival strategies for a dynamic future. *Journal of Tourism Futures, 3*(1), 56-65.

Riza, L. S., Putra, Z. A. Y., Zain, M. I., Trihutama, F. Z., Utama, J. A., Samah, K. A. F. A., ... & Priyatikanto, R. (2024). Real-time anomaly detection in sky quality meter data using probabilistic exponential weighted moving average. *International Journal of Data Science and Analytics*, 1-18.

Russom, P. (2011). Big data analytics. *TDWI best practices report, fourth quarter, 19*(4), 1-34.

Sabharwal, R., & Miah, S. J. (2021). A new theoretical understanding of big data analytics capabilities in organizations: a thematic analysis. *Journal of Big Data*, *8*(1), 159.

Sagiroglu, S., & Sinanc, D. (2013, May). Big data: A review. In *2013 international conference on collaboration technologies and systems (CTS)* (pp. 42-47). IEEE.

Saputra, M. B., Danarianto, M. D., Murti, M. D., Alwan, M. A., & Yanti, R. J. (2022, February). Report on seeing, sky brightness, and meteorological properties measurements at Timau National Observatory site. In *Journal of Physics: Conference Series* (Vol. 2214, No. 1, p. 012013). IOP Publishing.

Sekar, R., Bendre, M., Dhurjati, D., & Bollineni, P. (2000). A fast automaton-based method for detecting anomalous program behaviors. In *Proceedings 2001 IEEE Symposium on Security and Privacy*. *S&P 2001* (pp. 144-155). IEEE.

Shamir, L. (2023). Outlier galaxy images in the Dark Energy Survey and their identification with unsupervised machine learning. *Astronomy and Computing, 43,* 100712.

Shieh, C. K., Huang, S. W., Sun, L. D., Tsai, M. F., & Chilamkurti, N. (2017). A topology-based scaling mechanism for Apache Storm. *International Journal of Network Management*, *27*(3), e1933.

Singh, K., & Upadhyaya, S. (2012). Outlier detection: applications and techniques. *International Journal of Computer Science Issues (IJCSI), 9*(1), 307.

Sommerville, I. (2011). Software Engineering. *Addison-Wesley*.

Sun, D., Zhang, G., Zheng, W., & Li, K. (2015). Key Technologies for Big Data Stream Computing.

Sun, P., Chawla, S., & Arunasalam, B. (2006). Mining for outliers in sequential databases. In *Proceedings of the 2006 SIAM international conference on data mining* (pp. 94-105). Society for Industrial and Applied Mathematics.

Thanh, B. N., Le, K. P., Huy, D. P., & Le Minh, D. N. (2024). Anomaly Detection in Smart Home Context based Machine Learning using Kafka.

Thein, K. M. M. (2014). Apache kafka: Next generation distributed messaging system. *International Journal of Scientific Engineering and Technology Research, 3*(47), 9478-9483.

Thiprungsri, S., & Vasarhelyi, M. A. (2011). Cluster Analysis for Anomaly Detection in Accounting Data: An Audit Approach. *International Journal of Digital Accounting Research, 11.*

Toliopoulos, T., Bellas, C., Gounaris, A., & Papadopoulos, A. (2020, June). PROUD: parallel outlier detection for streams. In *Proceedings of the 2020 ACM SIGMOD International Conference on Management of Data* (pp. 2717-2720).

Toshniwal, A., Taneja, S., Shukla, A., Ramasamy, K., Patel, J. M., Kulkarni, S., ... & Storm, M. (2014). Storm@ twitter. In *Proceedings of the 2014 ACM SIGMOD international conference on Management of data* (pp. 147-156).

Tran, L., Mun, M. Y., & Shahabi, C. (2020). Real-time distance-based outlier detection in data streams. *Proceedings of the VLDB Endowment*, *14*(2), 141-153.

Ur Rehman, A., & Belhaouari, S. B. (2021). Unsupervised outlier detection in multidimensional data. *Journal of Big Data*, *8*(1), 80.

Vercruyssen, V., Meert, W., Verbruggen, G., Maes, K., Baumer, R., & Davis, J. (2018). Semi-Supervised Anomaly Detection with an Application to Water Analytics. In *ICDM* (Vol. 2018, pp. 527-536).

Villar, V. A., Cranmer, M., Contardo, G., Ho, S., & Lin, J. Y. Y. (2020). Anomaly detection for multivariate time series of exotic supernovae. *arXiv preprint arXiv:2010.11194*.

Walker, M. F. (1970). The California site survey. *Publications of the Astronomical Society of the Pacific*, *82*(487), 672.

Warrender, C., Forrest, S., & Pearlmutter, B. (1999). Detecting intrusions using system calls: Alternative data models. In *Proceedings of the 1999 IEEE symposium on security and privacy (Cat. No. 99CB36344)* (pp. 133-145). IEEE.

Weigend, A. S., Mangeas, M., & Srivastava, A. N. (1995). Nonlinear gated experts for time series: Discovering regimes and avoiding overfitting. *International journal of neural systems*, *6*(04), 373-399.

White, T. (2012). *Hadoop: The definitive guide.* O'Reilly Media, Inc.

Wong, W. K., Moore, A. W., Cooper, G. F., & Wagner, M. M. (2003). Bayesian network anomaly pattern detection for disease outbreaks. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)* (pp. 808-815).

Yadranjiaghdam, B., Yasrobi, S., & Tabrizi, N. (2017, June). Developing a real-time data analytics framework for twitter streaming data. In *2017 IEEE International Congress on Big Data (BigData Congress)* (pp. 329-336). IEEE.

Yan, H., Sun, D., Gao, S., & Zhou, Z. (2018). Performance analysis of storm in a real-world big data stream computing environment. In *Collaborative Computing: Networking, Applications and Worksharing: 13th International Conference, CollaborateCom 2017, Edinburgh, UK, December 11–13, 2017, Proceedings 13* (pp. 624-634). Springer International Publishing.

Yeung, D. Y., & Chow, C. (2002). Parzen-window network intrusion detectors. In *2002 International Conference on Pattern Recognition* (Vol. 4, pp. 385-388). IEEE.

Zaharia, M., Chowdhury, M., Franklin, M. J., Shenker, S., & Stoica, I. (2016). Spark: Cluster Computing with Working Sets. *Communications of the ACM, 59*(11), 77-85. doi:10.1145/2934664

Zhang, Y., & Zhao, Y. (2015). Astronomy in the big data era. *Data Science Journal*, *14*, 11-11.

Zhao, L., & Akoglu, L. (2023). On using classification datasets to evaluate graph outlier detection: Peculiar observations and new insights. *Big Data*, *11*(3), 151-180.

Zhao, X., Zhang, J., & Qin, X. (2017). kNN-DP: Handling Data Skewness in kNN Joins Using MapReduce. *IEEE Transactions on Parallel and Distributed Systems*, *29*(3), 600-613.

Zhou, Z., Zhou, L., & Chen, Z. (2024). A Distributed Real-Time Monitoring Scheme for Air Pressure Stream Data Based on Kafka. *Applied Sciences*, *14*(12), 4967.

Zhuang, Z., Feng, T., Pan, Y., Ramachandra, H., & Sridharan, B. (2016, June). Effective multi-stream joining in apache samza framework. In *2016 IEEE international congress on big data (BigData Congress)* (pp. 267-274). IEEE.

Zikopoulos, P. C., & Eaton, C. (2011). *Understanding big data: Analytics for enterprise class hadoop and streaming data***.** McGraw-Hill Osborne Media.