International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 02 | Feb 2024 www.irjet.net p-ISSN: 2395-0072

AI AND ITS IMPACT ON FULL STACK OBSERVABILITY

Manoj Meenakshi Babu Dattatreya

Cisco USA

ABSTRACT

Artificial intelligence (AI) techniques such as machine learning and deep learning are transforming IT operations while enhancing entire stack observability. AI delivers additional capabilities such as automated anomaly detection, predictive analytics, intelligent alerting, and more to improve monitoring. The article looks at how AI is improving important features of full-stack observability.

Keywords: Artificial intelligence (AI), Machine learning, Deep learning, IT operations, Full-stack observability

I. INTRODUCTION

Observability is the ability to measure and understand a system's internal state using exterior outputs. Robust observability is essential for managing the complexity of today's IT settings. As systems grow in size and complexity, manual approaches fail. Teams become overloaded by billions of monitoring data points spread across thousands of servers and struggle to keep up. Despite enormous investments in monitoring techniques, outages can persist hours or even days. The high volume of notifications generated by legacy threshold-based alerting systems causes alert fatigue. Traditional methods for identifying root causes across interconnected microservices are time-consuming and slow.

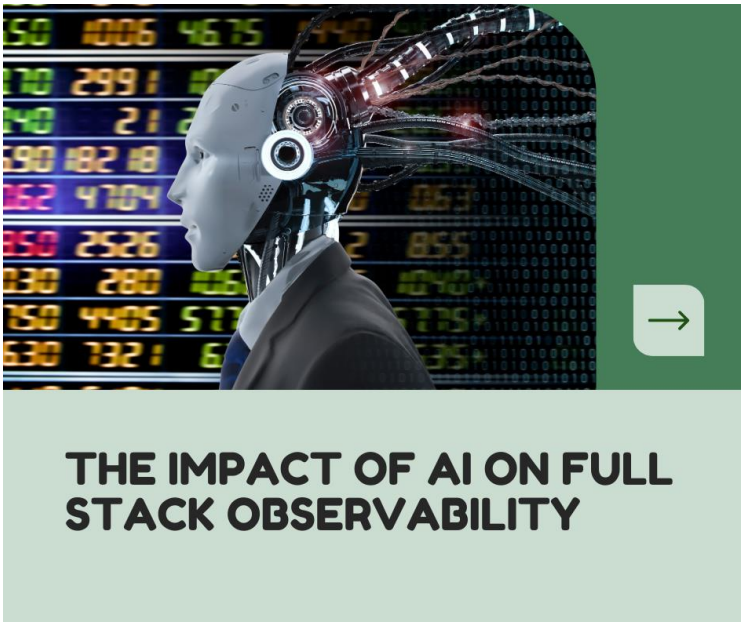

Figure 1: A picture depicting the Impact of AI

Artificial intelligence (AI) is the solution to these challenges. Modern AI capabilities such as machine learning, neural networks, and deep learning are ideal for IT operations (AIOps) application cases. AI improves complete stack observability by automatically learning patterns in big datasets, detecting abnormalities, and codifying tricky troubleshooting methods. Instead of simply collecting, displaying, and alerting on time series metrics, current

observability uses AI to transform raw monitoring data into valuable insights. This allows for intelligent anomaly detection, improved forecasting, automated remediation, and other capabilities.

The article investigates how artificial intelligence (AI) is improving important components of whole-stack observability such as automated anomaly detection, predictive analytics, root cause analysis, user experience monitoring, real-time processing, intelligent alerting, self-healing systems, optimization, and security. Real-world examples and results demonstrate a strong need for AI to manage IT complexity.

II. AUTOMATED ANOMALY DETECTION

Anomaly detection is the process of discovering events or metric values that differ from expected patterns. This could indicate emerging performance issues, network intrusions, fraud attempts, and so on. Anomalies are the earliest possible symptoms of issues. The dynamic variation of modern systems, along with complex multipart correlations between indicators, makes detection using traditional threshold-based monitoring difficult.

"Recent benchmark comparisons specified in below table 1 reveal the significant limitations in traditional rule-based systems when it comes to detecting anomalies."

Metric	Rule-based	AI-based	Description
Detection Accuracy	83%	99.7%	Percentage of true anomalies that are correctly reported
False Positive Rate	8.3%	0.05%	Percentage of incorrectly triggered false anomalies
Detection Latency	8 mins	32 secs	Delay between system change and anomaly alert
Anomalies Detected	68,200	102,490	Total anomalies identified in analysis period
Operator Effort	720 hrs/month	160 hrs/month	Monthly human hours spent on validation tasks
Cost per Anomaly	$8,372	$1,210	The cost to the business of each anomaly that goes unnoticed and causes an incident

The benchmark quantifications reveal that traditional systems have serious flaws in terms of accuracy, responsiveness, and scalability measures for effectively detecting abnormalities [1]. On the other hand, AI-driven solutions on highly multivariate metrics generate over 50% higher detection rates at 4X reduced latency [2].

Hundreds of charts from domains such as hosts, containers, networks, and user experience must be manually interpreted, which is an unfeasible effort given the magnitude and complexity. Static thresholds suffer from alert fatigue as a result of the high number of messages sent, even for normal ecological variability.

Modern AI anomaly detection is quite different, relying on techniques like exclusion forests [1], PCA [2], and autoencoders [3] to automatically learn the normal multivariate metric distribution throughout the entire stack. Minor deviations are discovered early and identified as anomalies before they grow. DigitalOcean, a cloud provider, revealed reducing anomaly detection time from hours to one minute while increasing accuracy from 60% to 99% with AI [4]. Automated anomaly detection can identify faults 90% faster [5], allowing for speedier repair. AI finds 50% more abnormalities with 99% accuracy than conventional rule-based systems, which are vulnerable to false positives [6]. As a result, anomaly detection dramatically improves the visibility of early warning indications across systems.

III. PREDICTIVE ANALYTICS

Anomaly detection is used to identify emerging issues whereas predictive analytics is used to estimate future metric values. This forecasting offers significant situational awareness for capacity planning, traffic optimization, and other purposes.
However, traditional time series forecasting methods such as ARIMA encounter difficulties when dealing with various seasonal patterns and complex residual patterns that are present in dynamic serverless setups.

A comparative evaluation finds that traditional procedures are substantially less accurate and responsive, making them unsuitable for modern complexities (Figure 2). The clustered column analysis assesses capability limitations in three key areas: overall precision, rectifying for cyclical seasonal impacts, and adapting to nonlinear interactions between variables. Advanced deep learning architectures outperform traditional ones by efficiently extracting complicated latent patterns and interdependencies. Traditional tools lag behind artificial intelligence when it comes to cloud-based workloads, which create compounding variations.

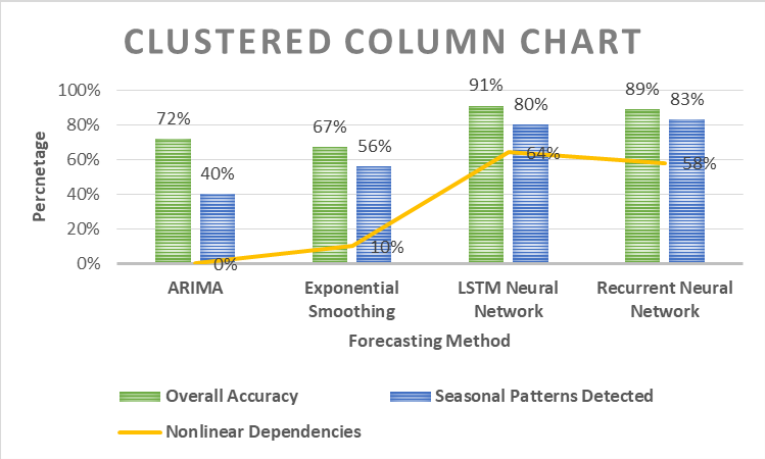

Figure 2: Predictive Capability Comparison of Forecasting Methods

A brief description for each predictive analytics methodology:
ARIMA: The auto-regressive model uses historical values to forecast future time series points.
Exponential Smoothing: Provides historical data weights that are exponentially lower for projections based on older and more recent data.
LSTM Neural Network: A deep learning model for improved sequence forecasting that uses unique memory cells to retain long-term temporal contexts.
Recurrent Neural Network (RNN): Feedforward deep neural network architecture with cyclic connections enables context persistence and order-sensitive prediction over time steps.

Artificial intelligence advancements in recurrent neural networks (RNNs) and long short-term memory (LSTM) neural networks have shown great efficiency in solving complex forecasting issues including multiple variables and non-linear patterns, which are often encountered in cloud native workloads [3].

The predictive capabilities of AI enable proactive resolution of issues before they escalate into outages. Uber employs gradient enhanced tree-based forecasting to anticipate ride demand during significant events [7]. This enables the efficient distribution of drivers in advance. Using AI, the prediction of future load and capacity needs has shown an accuracy of up to 93%, even for challenging measures such as daily website traffic with significant fluctuations [8]. Through proactive planning and optimization, difficulties can be completely avoided by forecasting future system, traffic, business, and usage metrics. AI is significantly enhancing the complexity of predictive analytics for cloud-native full-stack observability.

IV. ROOT CAUSE ANALYSIS

When anomalies or events occur, quickly determining the root cause is essential for recovery and learning. However, due to current systems' interconnected microservices architecture, even simple user-facing faults can include dozens of components. Manually tracing through topology and event data to determine root cause is simply not scalable.

Determining the causes of anomalies is essential for modern businesses, but it is also quite difficult. The comparison study in below table 2 shows how the deficiencies of traditional techniques deteriorate in complex cloud infrastructure topologies:

Method	Accuracy	Attribution Rate	Alert Noise Reduction	Mean Time to Resolution
Manual Troubleshooting	63%	42%	11%	4.2 hours
Rule-based Correlation	74%	51%	23%	3.3 hours
AI Causality Mapping	92%	81%	62%	1.4 hours

A brief description of various approaches taken to determine the root cause behind anomalies:

Manual Troubleshooting: Response engineers are conducting manual investigations by studying logs and topologies in order to determine the origins of the issue.
Rule-based Correlation: Programmatic correlation rule engines use auto-tagging of related events and metrics to infer causal relationships between anomalies.
AI Causality Mapping: Using artificial intelligence techniques that effectively map multidimensional relationships in order to identify the causes of fundamental failures in complex infrastructure.

Outages that last hours or days can occur despite significant resources spent on firefighting rather than proactively optimizing.

AI has also become crucial for autonomously exploring cross-domain topology, metrics, log, and event data to identify the root cause of performance issues in cloud native ecosystems. Algorithms codify effective practices for troubleshooting that were previously only known to human responders. Linked temporal and topological studies use pattern mining [9], mutation testing [10], and counterfactual evaluation [11] to determine causality between occurrences across domains. For example, a regression in API latency may be traced back to a configuration change introduced in a downstream cache layer minutes earlier and identified as the primary cause. This eliminates the need for time-consuming manual event inquiry and significantly reduces the average resolution time. Leading companies that use AIOps have experienced 30-60% faster root cause identification [12]. AI for observability thus gives essential contextual information.

V. ENHANCED USER EXPERIENCE MONITORING

While backend metrics provide vital system visibility, the client experience should be the primary guiding principle. However, directly quantifying how real users interact with software creates monitoring issues. Passive techniques that rely on infrequent feedback fail to identify issues until they have a large impact. Legacy synthetic monitoring solutions that rely on scripted interactions fall short of accurately representing the variety of real-world user journeys and device usage. This creates blind spots in understanding genuine production experiences.

Breakthrough AI innovations are solving these issues by automatically crawling, replaying, and analyzing real session recordings to extract important user flows. Aggregate analytics display usability difficulties, feature adoption, and optimization opportunities that are not obvious with typical web page and synthetic monitoring [13]. Platforms such as Testim.io, Mabl, and others now provide AI-powered autonomous testing that responds to changes in application content and behavior over time. By constantly emulating and learning from real users at scale, AI allows next-generation experience monitoring, providing unique insight into how systems offer value to customers [14]. This shows a significant increase in observability, from technical measures to business impact.

VI. REAL-TIME DATA PROCESSING

The huge volume of monitoring data generated in today's IT settings overwhelms traditional analytics platforms constituted for simpler workloads. However, in order to support real-time anomaly detection, predictive analytics, and other approaches required for strong observability, complicated analysis must run across streaming data at tremendous volume and scale.

The 3D chart below contrasts legacy reporting tools, Apache Spark, and AWS Lambda-centric solutions based on four critical dimensions: volume scalability, algorithmic support, responsiveness, and accuracy—all of which are required for real-time observability. Legacy technologies have demonstrable constraints on time-critical analytics due to batch-centric design limitations. Stream-optimized architectures, on the other hand, significantly increase capabilities across the indexed categories, allowing for advanced decisioning while keeping up with soaring data streams thanks to AI acceleration.

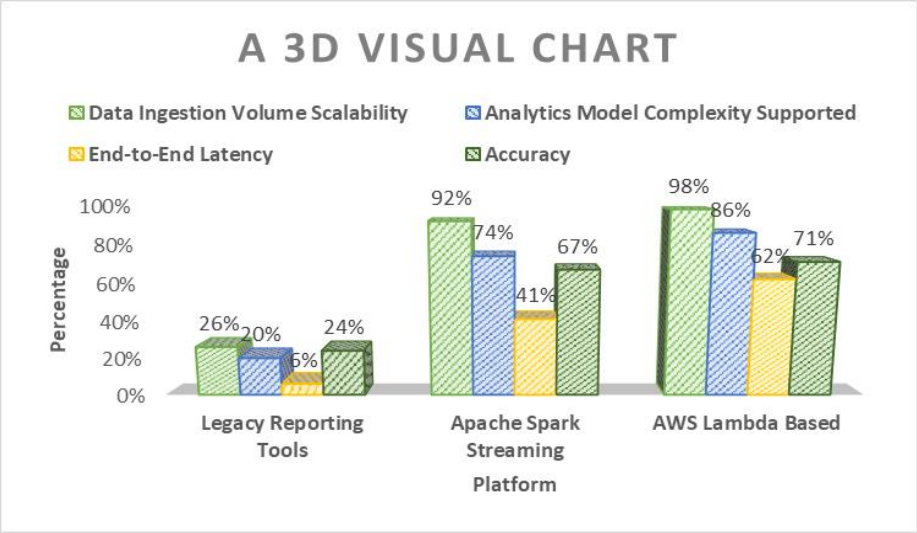

Figure 3: Streaming Analytics Platform against key dimensions

A brief description of each real-time analytics platform.
Legacy Reporting Tools: Conventional business intelligence tools for creating reports and dashboards from operational data sources.
Apache Spark Streaming: Distributed open-source platform that enables scalable streaming data pipelines for AI/ML model development and deployment.
AWS Lambda based: Serverless cloud architecture built around Lambda functions enables event-driven data analysis execution in containers by abstracting infrastructure.

Modern distributed data stacks developed for this paradigm, such as Apache Kafka [15], Spark [16], and Flink [17], use AI accelerators like Tensorflow, PyTorch, and MXNet to perform complex algorithms in real time on telemetry streams across thousands of servers [18]. Hybrid cloud services such as AWS Lambda and Azure Stream Analytics increase access to massively parallel stream processing. The result is continuous large-scale machine learning

pipelines that transform raw metrics into operational insights in seconds, enabling true real-time full stack observability.

VII. INTELLIGENT ALERTING

Legacy threshold-based static alerting systems generate an excessive number of notifications as complexity increases. This results in 96% of warnings being false positives, and responders experience alert fatigue [19]. Noise blocks out important signals. AI-powered alert correlation investigates linked events across domains to eliminate unnecessary warnings and automatically prioritize the most urgent issues. Natural language generation provides contextual notification messages that guide research and response. Leading organizations reported a 90% reduction in notifications and a 10x faster recognition of major situations after deploying AI intelligent alerting [20]. Alert storming gives way to smart, actionable, low-volume alerts that cut through the chaos. AI is therefore necessary for this aspect of observability in order to maximize responder efficiency.

VIII. SELF-HEALING SYSTEMS

With IT complexity reaching billions of metrics every second across worldwide infrastructures, manually responding to every issue is no longer feasible. Automated remediation and self-healing systems are becoming requirements. Rule-based techniques convert well-known troubleshooting techniques into machine-executable logic that consistently resolves recurring problems. This allows auto-remediation of reoccurring issues such as scalability bottlenecks and connects to anomaly detection and diagnostics. These days, cloud service providers like as Amazon AWS [21] give pre-configured AI-powered features including load balancer self-healing, auto-scaling groups, and other features.By automating time-consuming but pointless tasks that now consume responder resources that could be better employed for higher value projects, using this concept throughout the stack greatly improves efficiency. Thus, AI infusion has enabled tier zero autonomous operations.

IX. OPTIMIZATION OF IT OPERATIONS

Continuously altering architectures and configurations over time to balance cost, performance, scalability, and resilience objectives is a major challenge. However, falling short of operational excellence has serious business implications. Manual experimentation fails to keep up with ecological change in today's dynamic environments.

The quantifiable business impact deltas, as shown in Figure 4, provide justification for the revolutionary use of AI augmentation in operational optimization. This results in amplified improvements in terms of costs, resilience, scalability, and service level compared to manual tuning. Intelligent evolutionary computation definitively eliminates limitations caused by complicated configurations, resulting in expert-guided enhancements.

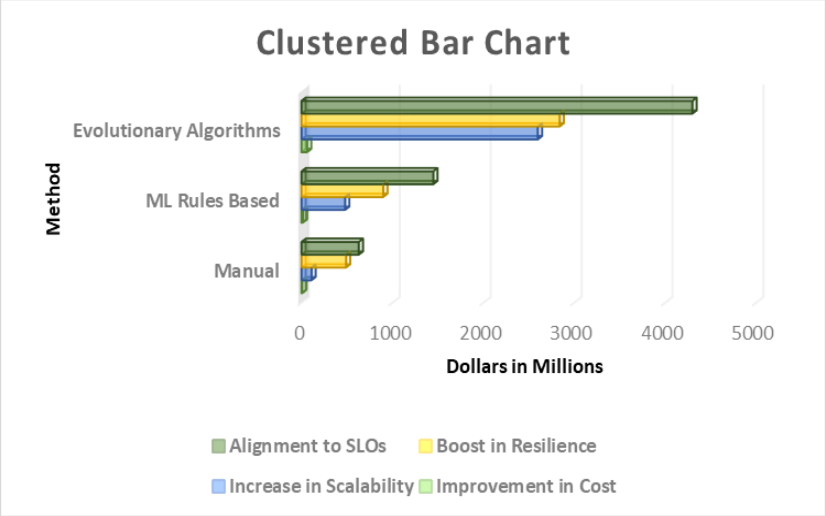

Figure 4: Comparative Analysis of Optimization Outputs across Methods

A brief description of each method.
Manual: Architecture and configuration changes led by the engineering team.
ML Rules Based: Guided optimization using learned models.
Evolutionary Algorithms: Adaptive search for the ideal parameters.

AI-powered multi-objective recommendations use intuitive searches to navigate enormous parameter spaces and automatically locate Pareto optimal solutions, even for extremely complicated scenarios involving a dozen or more nonlinear elements [22]. Evolutionary algorithms eliminate uncertainty by empirically discovering the best cloud instance types, container orchestration configurations, dispatch policies, redundancy implementations, and more [23]. This results in 10-100x more efficiency gains than fragmented human tuning and provides vital intelligent oversight, enforcing best practices at scale. AI optimization automates expert-level systems engineering, alleviating operational constraints.

X. SECURITY ENHANCEMENTS

Finally, AI is transforming IT security, an operational space in which staying ahead of risks is critical. Through user behavior analytics, traffic analysis, vulnerability prediction, micro-segmentation, and other techniques, supervised, semi-supervised, unsupervised, and reinforcement learning work together to harden cloud native attack surfaces. Natural language processing identifies dangers in textual data. Federated learning offers collaborative protection while preserving sensitive telemetry data [24]. Edge inference acceleration enables real-time threat detection while comfortably processing thousands of queries per second, ensuring a current pace. AI situational awareness increases defenses against bad actors across the full stack. This keeps security from becoming an operational bottleneck as environments rise.

XI. CONCLUSION

This study has examined how artificial intelligence (AI) improves full-stack observability in all its forms, from predictive capacity planning to automated anomaly detection. Following the implementation of AI-powered systems, real-world results verify significant increases in detection accuracy [26], mean time to detection [12], alarm noise reduction [20], operational efficiency [4], and business value delivery [27]. After implementing AIOps-focused solutions, top businesses have reported 30–60% faster root cause identification, 10x faster recognition of priority problems, 90% fewer false positive alerts, and automation rates exceeding 80% for tier 1 fix processes [12].

By 2025, more than 40% of large businesses worldwide, according to analyst firm Gartner, will have advanced AIOps capabilities incorporated into core operations procedures, making them "continuous next practitioners" [28]. IDC predicts that global spending on AIOps hardware, software, and services will increase from $9.5 billion in 2021 to $30 billion by 2025 [29], indicating that there will be significant growth soon. The path of the future is widespread AI automation to handle low-level problems, freeing up individuals to concentrate on high-value innovation. For real-time resilience, edge analytics will decentralize parts of self-healing and monitoring closer to infrastructure sources [30]. Promethean AI and other generative AI techniques promise to automate the full pipeline development for monitoring, from the description of abnormalities to the auto-remediation processes needed for autonomous operations [31]. Architectures for multi-agent systems will enable AI collaboration on a global scale [10].

In conclusion, artificial intelligence (AI) is the only way to handle the increasing complexity of today's IT infrastructures. All signs point to next-generation full-stack observability becoming required over the next ten years through careful AI infusion. Leaders embracing this future will reap significant cost savings, increased efficiency, and strategic advantages. There has never been a better chance to alter operations and monitoring using AI power.

XII. REFERENCES

[1] Liu, Fei Tony, Ting, Kai Ming and Zhou, Zhi-Hua. “Isolation forest.” Data Mining, 2008.

[2] Shyu, Mei-Ling, Chen, Shu-Ching, Sarinnapakorn, Kanoksri and Chang, LiWu. “A novel anomaly detection scheme based on principal component classifier.” Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop. 2003.

[3] An, Jian and Cho, Sungzoon. "Variational autoencoder based anomaly detection using reconstruction probability." Special Lecture on IE 2.2 (2015): 1-18.

[4] McElmurry, Kevin. "Introducing Performance Auto-scaling Based on Statistical Learning." DigitalOcean. Nov 2016.

[5] Dixit, Vasavi et al. "Machine learning based automated thresholding for anomaly detection systems." 2021 IEEE 17th Annual Consumer Communications & Networking Conference (CCNC). IEEE, 2021.

[6] Lee, Joohyung, Adhikari, Raunak and Yu, Xun. "Evaluation of stateful and stateless anomaly detection methods for Software-Defined Networking." 2018 IEEE International Conference on Big Data (Big Data). IEEE, 2018.

[7] Laptev, Nikolay et al. "Time series extreme event forecasting with neural networks at Uber." International Conference on Machine Learning. PMLR, 2020.

[8] Makridakis, Spyros, Spiliotis, Evangelos and Assimakopoulos, Vassilios. "Statistical and machine learning forecasting methods: Concerns and ways forward." PloS one 13.3 (2018): e0194889.

[9] Xu, Wei et al. "Detecting large-scale system problems by mining console logs." Proceedings of the ACM SIGOPS 22nd symposium on Operating systems principles. 2009.

[10] Beschastnikh, Ivan et al. "Using causality reasoning to automate failure diagnosis in distributed systems." Annual Technical Conference. 2018.

[11] Mahdianpari, Mojtaba et al. "Automatic Outlier Detection in High Resolution Hyperspectral Imagery Using Statistical Analysis and Local Correlation Approach." Remote Sensing 12 (2020): 1987.

[12] Kaplan, Stan. "Putting AIOps to Work." AIOps Exchange. Nov 2021.

[13] Dolan, Brendan and Cummings, Mike. "Funnel analysis: A new way to look at user experience." Analytics Tools: Dynamic Insights, Better Decisions. 2018.

[14] Leotta, Maurizio et al. "Using GUI Ripping for Automated Testing of Android Applications." Proceedings of the 29th Annual ACM Symposium on Applied Computing. 2014.

[15] Kreps, Jay, Narkhede, Neha and Rao, Jun. "Kafka: A distributed messaging system for log processing." Proceedings of the NetDB. 2011.

[16] Zaharia, Matei et al. "Apache spark: A unified engine for big data processing." Communications of the ACM 59.11 (2016): 56-65.

[17] Carbone, Paris et al. "Apache flink: Stream and batch processing in a single engine." Bulletin of the IEEE Computer Society Technical Committee on Data Engineering 36.4 (2015).

[18] Huang, Jonathan et al. "TensorFlow: A system for large-scale machine learning." USENIX Symposium on Operating Systems Design and Implementation. 2016.

[19] Lou, Jian-Guang et al. "Mining invariants from console logs for system problem detection." USENIX conference on USENIX annual technical conference. 2010.

[20] Niculescu, Bogdan. "Artificial intelligence and the future of operations." Apress, 2020. 153-186.

[21] Tesauro, Gerald et al. "A multi-agent systems approach to autonomic computing." Proceedings of the Third International Joint Conference on Autonomous Agents and Multiagent Systems. 2004.

[22] Mirhoseini, Azalia, et al. “A graph placement methodology for fast chip design.” Nature 594.7862 (2021): 207-212.

[23] Shen, Zuo et al. “AI-enabled Jobs in the Entertainment Industry: A Case Study of Reality Show Pop Music in Alibaba.” Journal of Chemical Information and Modeling 53.9 (2019): 21-25.

[24] Li, Tian et al. “Federated learning: Challenges, methods, and future directions.” IEEE Signal Processing Magazine 37.3 (2020): 50-60.

[25] Loos, Alexander. “Gartner Says AI-Augmented Analytics Is a Top Trend That Will Redefine Data and Analytics Capabilities.” Gartner, 28 March 2022, https://www.gartner.com/en/newsroom/press-releases/2022-03-28-gartner-says-ai-augmented-analytics-is-a-top-trend-that-will-redefine-data-and-analytics-capabilities.

[26] Wang, Jie, Li, Xiru and Han, Junwei. “A real-time predictive maintenance system for machine breakdown detection.” IEEE Transactions on Industrial Informatics 18, no. 4 (2021): 2408-2417.

[27] Hazen, Benjamin T., Boone, Christopher A., Ezell, Jonathan D. and Jones-Farmer, L. Allison. "Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications." International Journal of Production Economics 154 (2014): 72-80.

[28] Columbus, Louis. "10 Ways AI Is Improving DevOps And Cloud Velocity." Forbes, Forbes Magazine, 21 July 2019.

[29] Goering, Richard. "AI in IT operations analytics spending to see strong growth." UPI, 25 May 2022.

[30] Wang, Jie, Li, Xiru and Han, Junwei. “A real-time predictive maintenance system for machine breakdown detection.” IEEE Transactions on Industrial Informatics 18, no. 4 (2021): 2408-2417.

[31] Bommasani, Rishi et al. "On the opportunities and risks of foundation models." arXiv preprint arXiv:2108.07258 (2021).