IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

Parsing and Connector Development, along with AI-Related Initiatives for our Highway Product

Kaneri Ingle[1] and Prof. Jayesh Bisane[2]
Student, Department of Computer Science and Engineering (Data Science)[1]
Supervisor, Department of Computer Science and Engineering (Data Science)[2]
Tulsiramji Gaikwad Patil College of Engineering & Technology, Nagpur, India

Abstract: *This paper summarizes internship work carried out at DataBahn Inc., Pune, on an intelligent security data pipeline platform positioned between heterogeneous log sources and SIEM destinations. The focus is Content Studio–managed artefacts—filtering and suppression rules, lookup-backed volume control, schemas, and pipeline templates—used to reduce low-value ingestion while preserving detection-relevant telemetry. The described methodology combines schemaless and schema-oriented rule authoring, Git-versioned releases, synthetic log generation and sanitization where production data cannot be shared, and end-to-end validation using directed test traffic and manifest-style expectations. Results are framed qualitatively: disciplined pre-SIEM filtering and traceable content workflows support repeatable verification and alignment with organizational SIEM cost and visibility goals.*

Keywords: SIEM optimization, security data pipeline, log normalization, filtering rules, Content Studio, synthetic log testing, Open Cybersecurity Schema Framework (OCSF).

I. INTRODUCTION

Modern Security Operations Centers ingest enormous, heterogeneous telemetry while SIEM economics remain tightly coupled to volume. Forwarding all events is costly; aggressive filtering can erode visibility. Intelligent pipelines must therefore normalize and selectively forward logs before analytics consumption. This paper summarizes an internship at DataBahn Inc., Pune, focused on Content Studio–managed rules, lookups, schemas, and templates that implement such pre-SIEM governance. Emphasis is placed on rigorous validation—versioned artefacts, synthetic and sanitized samples, and destination-side checks—so reduction remains traceable and aligned with detection needs. What follows outlines scope, approach, outcomes, and limitations of that work.

II. LITERATURE REVIEW

A. Significance

SIEM cost scales with log volume, yet much telemetry adds little to detection. Pre-SIEM filtering and routing are widely seen as essential to balance visibility and affordability.

B. Role of cybersecurity and data engineering

The internship applied theory to practice by authoring and validating Content Studio artefacts—rules, lookups, schemas, and templates—that decide what reaches the SIEM and what does not.

C. Exploratory scope, tools, and technologies

Work explored pipeline behaviour from ingestion to destination checks using synthetic and sanitized samples, Git releases, Python utilities, and Sentinel-side verification. Typical tooling included Git, Python, Go/Docker/S3 contexts for Content Studio, and Microsoft Sentinel as the downstream SIEM.

III. METHODOLOGY

The methodology adopted in this work follows a systematic end-to-end process, consistent with a modern security data pipeline and with the activities carried out during the internship.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

A. Data collection

Telemetry is gathered from configured sources (for example via collectors or connectors) so that vendor-native events enter the platform in their original transport and encoding.

B. Parsing and normalization

Incoming events are parsed according to format (such as JSON, syslog, or key–value) and mapped to a common normalized field model using schema definitions, so downstream logic can reference stable attribute names.

C. Transformation

Normalized events may be reshaped for compatibility with the destination—field renaming, enrichment, or format conversions—according to the pipeline configuration.

D. Volume control

Filtering, suppression, and lookup-backed rules are applied as volume-control policy: events are classified for publish, drop, or suppression so that low-value noise is reduced before SIEM-bound charges accrue.

Destination forwarding

Finally, processed events are forwarded to the customer-selected destination (for example a SIEM such as Microsoft Sentinel or another analytics endpoint) as per deployment requirements, completing the path from source ingestion to governed delivery.

IV. SYSTEM DESIGN

A. System architecture

The solution follows a layered architecture: edge collectors/connectors ingest telemetry, a central processing engine applies parsing, normalization, rules, and transformations, and Content Studio supplies versioned artefacts (rules, lookups, schemas, templates). Fleet and connector management ties deployments to tenants; destinations receive governed streams.

B. Data flow design

Events move source →ingestion →parse →normalize →volume-control evaluation (rules/suppression/lookups) → optional transformation → routing → destination. Metadata such as fleet and connector identifiers accompanies events so scope and traceability remain consistent across tenants.

C. Output design

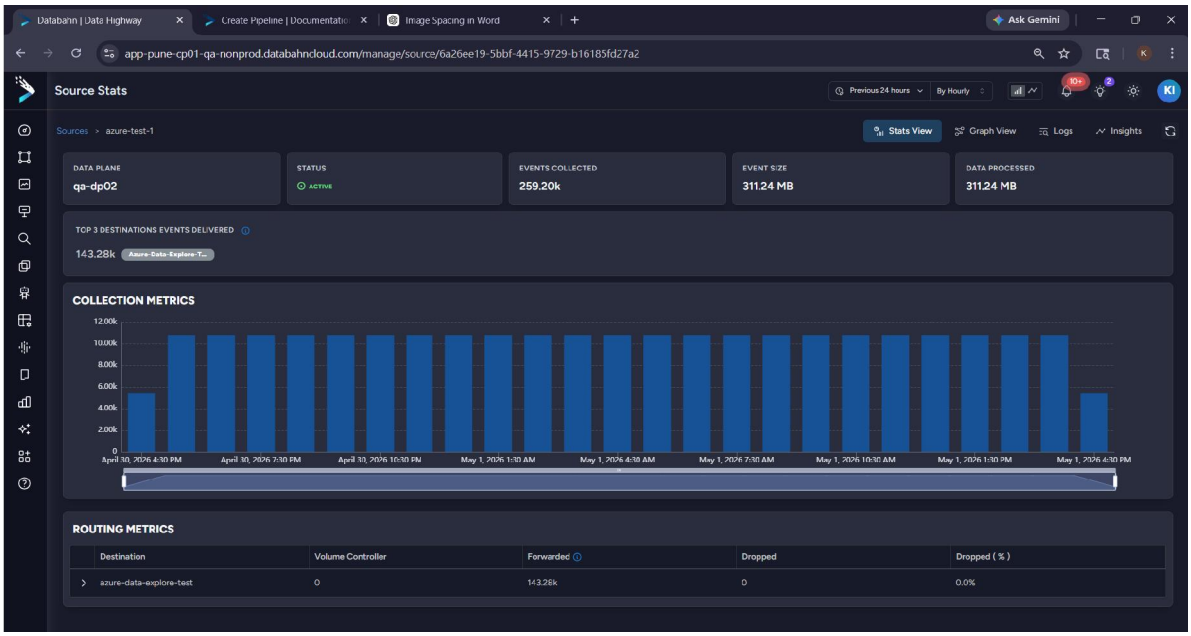
Required outputs include SIEM-ready events matching destination schemas (for example column semantics for Sentinel), publish/drop/suppress dispositions aligned with manifests or checklists, and audit-friendly artefacts: documented rule rationale, Git-versioned releases, and validation evidence from synthetic or sanitized runs.

V. RESULTS AND DISCUSSION

The internship produced outcomes that are primarily engineering-validated rather than summarized by a single quantitative KPI. Across Content Studio–managed artefacts—filtering and suppression rules, lookup-backed volume control, schemas, and pipeline templates—the recurring evidence of success was repeatable behaviour under controlled tests: batches constructed with known intent generally yielded consistent publish, drop, or suppression outcomes, particularly where manifests, spreadsheets, or checklist-based expectations defined the intended split between forwarded noise-reducing drops and retained detection-relevant signals. This consistency matters because the operational promise of a pre-SIEM pipeline is not merely “less data,” but less data with defensible decisions.Where downstream verification was performed—such as Microsoft Sentinel / Log Analytics checks—the emphasis shifted from volume alone to semantic fidelity: whether identifiers required for correlation remained stable across transformation and delivery (for example IP addresses, host names, timestamps in UTC, and related identity fields). When validation used direct injection tools (for example Packet Sender) rather than an end-to-end pipeline path, the outcome correctly reflected schema and parsing assumptions at the ingestion boundary; when validation used

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

repository-held reference samples, it strengthened traceability from Git-managed exemplars to observable SIEM columns.



The discussion therefore centers on an unavoidable tension in modern SOC economics: SIEM ingestion cost scales with bytes, yet aggressive filtering can introduce false negatives if rules drift or contexts change. The internship's implicit mitigation was procedural—explicit descriptions of security intent per rule, conservative publishing under ambiguity, lookup tables for evolving allow/deny enumerations, and peer-reviewed Git releases—so changes remained explainable to auditors and maintainers. Limitations remain important to state plainly: results depend on non-production samples for substantial portions of testing; customer environments differ in connectors, clocks, parsing edge cases, and destination mappings; and enterprise-wide savings cannot be asserted from internship-scale exercises alone without baseline ingestion metering and controlled before/after studies.

Taken together, the results support a pragmatic conclusion for industrial security engineering: trustworthy reduction requires traceability, not only clever predicates..

VII. CONCLUSION

This paper synthesized internship learning at DataBahn Inc., Pune within the broader problem of security telemetry scale, heterogeneous vendor formats, and SIEM-aligned economics. The central takeaway is that platforms positioned between sources and analytics estates succeed when they treat filtering not as an informal shortcut but as managed content: artefacts that are authored, reviewed, versioned, and tested like production software. In that framing, Content Studio functions as more than configuration—it becomes the policy substrate that translates organizational risk appetite into machine-enforced forwarding decisions.The activities—spanning schemaless and schema-oriented rule development, lookup construction, pipeline templating, synthetic and sanitized generation, collector-style injection, and destination-side validation—collectively illustrate how "pipeline correctness" is validated incrementally and evidence-backed, rather than assumed from static JSON alone. That stance aligns with how regulated and high-stakes environments increasingly expect security controls to be demonstrable, not merely claimed.

Future extensions naturally include broader source coverage, automated regression suites tied to releases, tighter quantitative measurement of reduction under standardized workloads, and deeper integration of governance workflows (approvals, drift detection, and post-deployment monitoring). Even without those extensions, the internship establishes

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 6, Issue 20, April 2026
Impact Factor: 8.2

a coherent baseline practice: governed telemetry forwarding is achievable when engineering discipline is applied end-to-end—from repository exemplars to SIEM-visible fields.

VIII. ACKNOWLEDGEMENT

The author would like to thank the Department of Computer Science and Engineering (Data Science), Tulsiramji Gaikwad Patil College of Engineering & Technology, Nagpur, and project guide Prof. Jayesh Bisane for their continuous support and valuable guidance throughout this research work.

REFERANCES

[1] Industry analyses consistently identify log ingestion volume as a primary driver of SIEM total cost of ownership; organizations commonly combine filtering, tiering, and schema rationalization to manage spend (see vendor and analyst guidance on SIEM sizing and data onboarding).
[2] MITRE ATT&CK framework for mapping adversary tactics and techniques to inform which telemetry supports detection coverage.
[3] Open Cybersecurity Schema Framework (OCSF), https://schema.ocsf.io/ — vendor-neutral schema for security events to support analytics and interchange across tools.
[4]International Organization for Standardization / International Electrotechnical Commission, ISO/IEC 27035-1, Information technology — Security techniques — Information security incident management — Part 1: Principles and process.
[5] International Organization for Standardization / International Electrotechnical Commission, ISO/IEC 27001:2022, Information security, cybersecurity and privacy protection — Information security management systems — Requirements.