DATA SOURCE EVALUATION

Data Sources Overview

Evaluating internal vs external data sources for domain-specific clustering.

This project applies on clustering VPS hosting real customer support data to uncover recurring issues and improve internal workflows to be able to generate actionable insights from real business data.

Primary data source:

Hosting chat logs from the support platform (≈ 5 000 records, 2021–2025),

Include timestamps, sender type, message text, and basic metadata.

Captures authentic user-engineer interactions across both simple and complex topics.

Supplementary source:

Email tickets (≈ 150 threads) – integrated into the main dataset with a *channel_type* flag to capture more complex support cases while preserving contextual consistency.

Excluded sources:

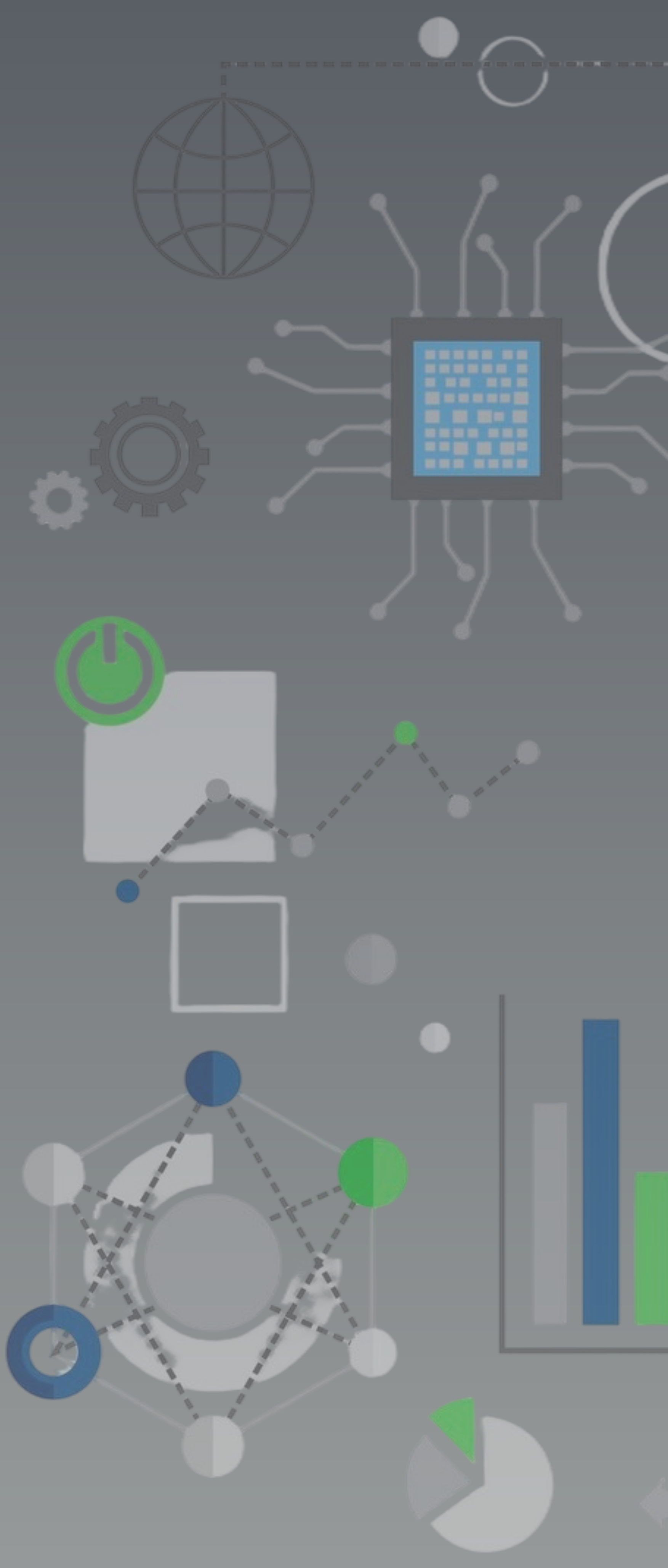
External help-desk datasets were not used, as each company's support domain has its own technical vocabulary, workflows, and customer context. Since the goal is to identify internal efficiency patterns within hosting's operations, mixing external data would reduce interpretability and business relevance.

Why this matters:

A domain-specific dataset ensures that clustering insights directly inform real process improvements — enhancing FAQ accuracy, engineer efficiency, and customer retention within company's unique support model.

Source	Type	Volume	Use Case
Chat Logs	Text (support platform)	~5000	Primary clustering dataset
Email Tickets	Text (IMAP export)	~150	Included in unified dataset (flagged by *channel_type*)
Public Datasets	External	–	Not used (domain mismatch)

From Raw Logs to Dataset



The initial data consist of unstructured chat and email logs exported from VPS hosting support platform. Each log contains entire conversation threads with timestamps, participants, and message text.

To transform this raw input into an analyzable dataset, the following preparation steps will be applied:

1 **Data consolidation** – merge chat and email records into a unified schema with consistent fields (ticket_id, timestamp_utc, sender_type, language, sentiment, channel_type, message_text, message_len, ner_entities).

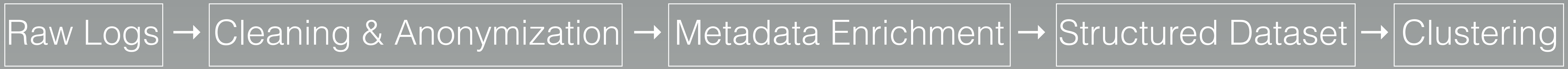
2 **Text cleaning & segmentation** – split long threads into individual messages or ticket summaries, remove system artifacts (signatures, boilerplate).

3 **Metadata enrichment** – detect and append derived attributes such as language, message length, or sentiment polarity.

4 **Anonymization** – mask personal data (names, emails, IPs) before further analysis.

5 **Filtering & timeframe selection** – keep focus on approximately the last 12–24 months of interactions to ensure relevance and manageable volume.

Result: a clean, structured dataset ready for feature extraction (TF-IDF + NER + sentiment) and clustering analysis.



Raw Logs → Cleaning & Anonymization → Metadata Enrichment → Structured Dataset → Clustering

Comparing Data Channels: Chat vs Email

Criterion	Chat Logs (~5000)	Email Tickets (~150)
Relevance	High – covers most recurring questions and quick fixes	High – includes detailed, multi-step technical issues
Accuracy	Moderate – informal phrasing, typos, boilerplate	High – more structured language and fewer noise artefacts
Completeness	High – continuous stream since 2021	Medium – gaps in archived older threads
Consistency	High – single export schema	Medium – varied formats and reply styles
Timeliness	High – live data from current stack	High – updated with recent escalations

To evaluate data suitability, both communication channels were compared across key quality dimensions. While chat logs offer higher volume and variety, email tickets capture deeper technical issues.

Together, they ensure balanced representation of everyday and complex support interactions.

Conclusion: combining both sources increases topic diversity and mitigates single-channel bias for clustering. Selected channels will be validated against real support processes to ensure maximum business value.

Data Quality & Limitations

Assessing raw data readiness for transformation into an ML-friendly dataset

Criterion	Assessment	Relevance for Hexcore dataset
Relevance	Measures how well the data reflects the business problem	Support logs are expected capture ticket topics and user-engineer communication: highly relevant for identifying recurring issues.
Accuracy	Evaluates correctness and noise level in the data	Some noise from greetings, boilerplate text, or repeated signatures; will be resolved via preprocessing and token filtering.
Completeness	Checks coverage and missing values	Few years of continuous data from both chat and email channels, minor gaps in older email archives.
Consistency	Assesses uniformity of structure and formatting	Chat and email logs will be merged into unified schema, timestamp and sender fields standardized.
Timeliness	Evaluates data recency and ongoing relevance	Logs cover 2021–2025 interactions, ensuring recency for model training

Limitations & Mitigation Strategies

- *Unstructured text*: will be handled via segmentation and TF-IDF vectorization.
- *Multilingual messages*: will be detected automatically; clustered separately to preserve semantic clarity.
- *Limited dataset size*: ~5k tickets — sufficient for exploratory clustering, but future scalability planned.
- *No labeled data for validation*: will be validated through manual review by Support Lead.
- *Privacy constraints*: anonymization of IPs, emails/names will be mandatory before text analysis.
- *Duplicate or reopened tickets:* will be deduplicated based on ticket_id, timestamps, and message similarity to prevent overrepresentation.
- *Engineer macros & templates*: will be filtered out during preprocessing to reduce repetitive system-generated text noise.
- *Channel imbalance (chat vs email)*: addressed through stratified sampling to ensure both communication types are proportionally represented in the final dataset.

Summary

Overall, the company’s support data provide sufficient breadth and domain specificity for transformation into an ML-ready dataset. Planned preprocessing and validation steps will ensure data reliability and ethical compliance before clustering analysis. The next stage will focus on manual validation of the most significant clusters and incremental dataset extension in line with the company’s support data growth.

Data Handling & Privacy Best Practices

Ensuring responsible data processing and reproducibility to maintain data integrity and protect user privacy, company's data preparation follows established privacy and data-cleaning standards.

Privacy & Security

- All personal identifiers (names, emails, IPs) anonymized before analysis.
- Encrypted storage; access limited to project maintainers.
- No data shared outside company; model outputs contain no raw text.

Data Cleaning

- Boilerplate text, greetings, and engineer templates removed during preprocessing.
- Language detection precedes tokenization to prevent cross-language noise.
- Duplicates and reopened cases merged by ticket_id and message hash.

Reproducibility & Governance

- Dataset versioning (v0.1 raw → v0.2 cleaned → v0.3 enriched).
- Documented preprocessing pipeline ensures traceability of all transformations.
- Compliance aligned with GDPR principles of data minimization and purpose limitation.