MANISH S N

manishsn.in contact@manishsn.in github.com/manish-003 linkedin

Backend engineer (Java, Linux/FOSS) building high-throughput content-inspection pipelines for Zoho Mail DLP: attachment parsing, metadata and embedded-content extraction, and real-time matching over large rule and dictionary sets. Apache Tika contributor.

EXPERIENCE

Zoho Corporation – MTS, Developer (Backend), Mail DLP May 2024 – Present
Estancia IT Park, Chennai *(Software Development Intern: Jan 2024 – May 2024)*

- **Redesigned the attachment processor** of the DLP engine, which inspects mail **data in transit**; own all attachment-related features, queries and improvements. Wrote custom Tika parsers and content handlers, and **contributed to Apache Tika** itself more than once in the process, to enable recovery of partially parsed documents, custom OCR, selective metadata handling, embedded link processing and encryption handling. My commits were accepted into the Apache Tika repository.
- **Identified a content-processing bottleneck:** (a monolithic YARA rule) and fixed it by splitting the rule across servers, cutting 4 MB / 300-SIT scan time from **39.6 s to 7.2 s (5.5×)** in benchmarks.
- **Built multilingual document fingerprinting** from scratch from the research literature, with a custom single-pass tokenizer (collector architecture) fusing lowercasing, stopword removal, char-gramming and hashing, with manual surrogate-pair handling, per-character break/ignore rules and script-aware boundaries (kana-to-kanji cuts for Japanese, char-gram fallback for long words and Chinese); median 65.4 ms versus 69.3 ms for an equivalent Lucene analysis chain on 3 MB.
- **Invented a modified, compressed Aho–Corasick automaton** for the keyword dictionary feature that matches directly on the compressed structure, without rebuilding the tree, covering 12,000 key phrases against 4 MB of content in real time with minimal RAM and stored-trie size; paper and patent filings in preparation.
- **Pioneered adoption** of new technologies: **led Docker adoption**, by containerizing the DLPEngine microservice and adding automated certificate rotation, connection draining and central logging, none of which existed before; driving the build **migration to Gradle**. Introduced **local LLM inference** for custom workflows and wrote an internal company article on running models locally: model and quantization selection, multimodal projector (mmproj) files, and MoE architectures.
- **Scouted new capabilities** for the team and proposed **ML-based DLP classification:** benchmarked Milvus vector lookup (53% F1), a fine-tuned DistilBERT head (83% micro F1) and an internal model fine-tuned on custom categories (88–96% accuracy, 0.02 MB/s on CPU), concluding GPU hardware is required; explored embedding models (BERT, DistilBERT, RoBERTa, GLiNER).

SKILLS

- **Languages:** Java, Python, C, C++, Bash, SQL
- **Content & search:** Apache Tika, Lucene, metadata/embedded-content extraction, document fingerprinting, Aho–Corasick, regex/YARA
- **Data & messaging:** HBase, Redis, ZooKeeper; pub/sub on Zoho's in-house Kafka-like queue (consumer registration, sync); ROS topics
- **Build & deploy:** Docker, Gradle, Ant, Git, Hg, Linux, nginx, AWS (in progress); self-hosted Calibre-Web and Vikunja on an Oracle Cloud ARM VPS
- **ML:** llama.cpp, embedding models (BERT, DistilBERT, RoBERTa), Vector DB (Milvus), fine-tuning

EDUCATION

B.E. Robotics & Automation, Anna University (Rajalakshmi Engineering College, founding batch), CGPA 8.51 (First Class with Distinction) 2020 – 2024

HONOURS & EXTRAS

- BAJA SAE India (2023–2024), Head of Electrical Team: redesigned the vehicle's electrical system from scratch; passed all technical inspections, 20 clean laps in the final round
- WorldSkills India, Cloud Computing (Skill 14) (currently competing)
- State 2nd, Thulir Vinadi Vina science quiz (2019); first prize, INSPIRE Internship Camp, DST (2018)