Sai Rohith Tanuku

United States | | tanukusai.sai52@gmail.com | https://www.linkedin.com/in/thisissairohith | https://www.github.com/rohith1125

SUMMARY

Software Engineer experienced in building backend services and data-intensive systems, including a C++20 engine for real-time database activity monitoring and distributed services handling more than 10 million daily transactions. Applies C++, Java, Python, SQL, and cloud technologies to implement and validate reliable systems, with measurable results including 99.95% transaction correctness and 20% lower peak database latency.

EDUCATION

University of Pittsburgh **Aug 2023 - Apr 2025**

Master of Science, Computer and Information Sciences (GPA: 3.957/4.00)

- **Coursework:** Algorithms, Data Structures, Machine Learning, Neural Networks, Cloud Computing

Manipal Academy of Higher Education **Aug 2019 - May 2023**

Bachelor of Science, Computer Science & Communication Engineering

- **Coursework:** Database, Operating Systems, SDLC, Computer Network, Data Mining, Web & App Dev

WORK EXPERIENCE

IBM | *Software Engineer* **Sep 2026**

- Develop the Guardium Sniffer, a C++20 engine that captures and decodes live Oracle and MySQL network traffic to support real-time database activity monitoring across enterprise deployments.
- Extend ANTLR3 and BNF-defined SQL grammars to parse and normalize queries across database dialects; use tcpdump and TCP/IP packet analysis to diagnose wire-protocol parsing failures, improving query coverage and troubleshooting accuracy.
- Build containerized Podman test environments connected to live Oracle and MySQL instances, then automate regression validation with Python and Bash.

Chicago Education Advocacy Cooperative | *Software Engineer* **May 2025 - Sep 2026**

- Engineered Java/Kotlin Spring Boot microservices and more than 20 REST APIs handling over 10 million daily transactions; implemented saga recovery and Redis idempotency to achieve 99.95% correctness under concurrent load, and optimized SQL/JPA queries across 300 million-plus monthly records to reduce peak database latency by 20%.
- Built Kafka event pipelines with idempotent consumers across 12-plus integrations, increasing throughput by 35% with no duplicate-charge incidents; tuned Prometheus monitoring and Istio to resolve gRPC stream exhaustion, raising throughput by 37% and reducing P95 latency by 12%.
- Built and deployed a production RAG assistant using LoRA-fine-tuned LLMs on AWS GPU infrastructure with Terraform and FastAPI; it served 500-plus engineers and more than 2,000 daily queries, reducing hallucinations by 41% and ticket-resolution time by 33%.

University of Pittsburgh | *Master of Data Science Course Facilitator and TA* **Jun 2024 - May 2025**

- Developed Python test suites for Coursera's autograder, reducing grading workload by 30%; also guided more than 140 graduate students through Python, SQL, and machine learning problems, contributing to a 15% increase in student satisfaction.

EAIESB | *Software Engineer Intern* **Jan 2023 - May 2023**

- Built and optimized distributed PySpark ETL pipelines on Azure to process millions of records daily, cutting processing time by 25%; added data-quality checks and monitoring for downstream analytics teams.
- Automated on-premises VM migrations to Oracle Cloud using the OCI Python SDK, reducing engineering migration effort by 60%; extended a deployment tool with new OCI resource types and CMDB API integration.

Larsen & Toubro | *Artificial Intelligence Research Intern* **Jun 2022 - Aug 2022**

- Designed a reusable, PyTorch-inspired deep-learning library with modular data-loading, training, and inference abstractions; used it to improve object-detection performance by 10% mAP across YOLO, RetinaNet, and EfficientDet.
- Built automated Python pipelines for satellite-imagery detection and LiDAR point clouds containing more than 500 million points, reducing manual inspection effort by 25% at 85% detection accuracy.

PROJECTS

Distributed Key-Value Storage System | C++, Stream Sockets

- Led backend development of a fault-tolerant distributed storage system for cloud storage and webmail services, using dynamic tablet partitioning to distribute data and load across storage nodes.
- Implemented primary-secondary replication, write-ahead logging, and checkpoint-based recovery so nodes could recover from failures without losing data.
- Added LRU-based memory eviction and inter-server communication over stream sockets to maintain efficient data access under memory pressure.

NL-to-SQL Generation with Fine-Tuned LLMs | PyTorch, Hugging Face, DeepSpeed

- Fine-tuned LLaMA-7B and DeepSeek-Coder on more than 120,000 natural-language-to-SQL pairs using DeepSpeed across multi-GPU A100 clusters, improving SQL query correctness by 30% over baseline models.
- Designed domain-adaptable training workflows for custom SQL dialects in finance, e-commerce, and healthcare, and built a frontend for interactive query generation.

Distributed Data Integrity Verification System | Cassandra, Ethereum, MapReduce

- Built a tamper-detection system combining Merkle trees, Cassandra storage, and Ethereum-anchored hashes; it detected 94% of tampering across more than 1,200 records.
- Parallelized verification with distributed MapReduce processing, achieving four times the performance of baseline methods.

TECHNICAL SKILLS

- **Programming Languages & Backend**: C#, C/C++, Java, Python, Kotlin, SQL, JavaScript, TypeScript, Shell Scripting, REST
- **Backend & API Development**: Microservices, REST APIs, Backend Development, Distributed Computing, System Design, Platform Engineering, Multi-threading, Caching, Load Balancing, APIs
- **Cloud, Big Data & Databases**: MongoDB, PostgreSQL, AWS, GCP, Docker, CI/CD, Spark, Cloud Networking
- **Frameworks, Libraries & AI**: Node.js, PyTorch, TensorFlow, MLOps, AI/ML Understanding
- **Networking & Automation**: Automation Testing, Linux, Networking Concepts, Networking Protocols