

Santhosh Muthukrishnan *Software Engineer | AI & Cloud Pipelines*

✉ ie.santhosh.m+c@gmail.com [in linkedin.com/in/thataintme](https://www.linkedin.com/in/thataintme) github.com/thataintme

Profile

Data-focused Backend Engineer with 2 years of experience in designing, building, and optimizing cloud-native systems and data infrastructure, including pipelines on AWS Lambda, SAM, and SQS/SNS. Built data preprocessing and modeling pipelines for multi-source ingestion within a Slurm-powered HPC research environment on LLMs. Known for materializing solutions for convoluted scenarios. Adept at solving complex problems, optimizing system performance, and delivering reliable, scalable infrastructure. Strong collaborator with excellent planning, problem-solving, and communication skills.

Seeking opportunities to contribute to data engineering, Agentic AI, and LLM-powered applications

Skills

Data Infrastructure: AWS Lambda, SAM, SNS, SQS, event-driven processing, dataset/annotation tooling, batch & async workflows

Databases: PostgreSQL, MongoDB, DynamoDB, Redis

Cloud & DevOps: AWS (Serverless, EC2, S3, Lambda, RDS), GCP, Git, Docker, Kubernetes, Jenkins, GitHub Actions

Backend: Java, Spring Boot, Spring MVC, Python, Flask, Node.js, Microservices, JWT, Hibernate

AI/ML: NLP, Deep Learning (LLMs and VLMs), High Performance Compute (Slurm), HuggingFace, RAG, LangChain

Data Analysis: R, MATLAB, PostGis, Scikit-learn, Pandas, NumPy

Testing: JUnit, Mockito, Selenium, Cypress

Professional Experience

AI Research Assistant

05/2026 – 08/2026 | Maynooth, Ireland

Maynooth University

- Designed and built **multi-source data ingestion pipelines** on a GPU-accelerated **HPC** cluster, using **Slurm** for **job scheduling**, queuing, and **resource allocation** across parallel training runs.
- Built preprocessing **pipelines** handling large-scale medical image data (Whole Slide Images, multi-GB per file), including patching strategies to make gigapixel inputs tractable for model training.
- Built evaluation pipelines to benchmark 8 **transformer** models and 4 **vision-language** models across 3 datasets. *Best-performing model reached **82.56%** accuracy on patch-wise prostate cancer gleason score classification.*
- Extended the research pipeline into a diagnostic web application for inference, letting users upload WSI/patch inputs and select from 12 trained models to generate predictions
- Optimized GPU utilization and job throughput on shared HPC resources, managing batch job scheduling to reduce queue wait times and compute waste.
- Engineered a reproducible experiment framework (data versioning, library versioning, config management, logging) to support rapid iteration across model-dataset combinations.

Postgraduate Student Researcher

05/2025 – 12/2025 | Dublin, Ireland

ADAPT Centre

- Built **multimodal** data pipelines fusing structured tabular data (demographic, biomarkers) with high-dimensional imaging data (3D MRI, extracted features), engineering format conversions (3D → 2D) to fit model input constraints
- Ran preprocessing and training workloads on a Slurm-based HPC cluster, handling job scheduling and resource allocation for compute-heavy imaging pipelines
- Engineered data pipelines to **overcome real-world clinical data collection constraints** such as sparse longitudinal data with unmatched parameters in multimodal records (missing values, misaligned timestamps, schema mismatches).
- Built and validated an XGBoost pipeline (feature engineering, training, evaluation), *achieving **$R^2=66.4\%$** in a full preprocessing → training → validation loop*
- Used auxiliary language models to auto-generate structured textual annotations at scale, building the pipeline that fed these into downstream model training

Codemantra

- Built the annotation tooling used by human clients to render PDF pages to images and hand-mark regions with structural tags (H1, H2, P, etc.), producing the labeled image-JSON pairs used by the AI team to train models.
- Owned multiple backend **microservices** and **cloud applications** in an AI-powered Intelligent Document Processing (IDP) product focused on PDF accessibility. Responsible for design, implementation, deployment, and maintenance.
- Built and maintained scalable APIs (**Spring Boot & Flask**) and cloud infrastructure (**AWS Lambda & SAM** with **SNS & SQS**) for dataset enumeration and inference: rendering new PDF files into page images, serving them to the AI models for classification, and maintaining data integrity across the enumeration, training, and inference systems.
- Optimized relational and non-relational database operations (**PostgreSQL**, MongoDB), improving query performance and overall system efficiency.
- Demonstrated strong involvement in product design and implementation strategy planning.
- Applied software design principles and patterns to improve code quality, scalability, and maintainability across services.

Notable achievements:

- Optimised PDF processing Lambda workloads by up to **12x**, eliminating timeouts on jobs that previously exceeded the 15-minute limit and cutting compute costs by **60%**.
- Optimised heavy local server processes, increasing operation speeds by **50%+**.
- Built a complete working pipeline of a PDF accessibility verifier application implementing popular accessibility standards.
- Reduced manual workload by providing automated workarounds for internal teams, offering a proven boost in ticket activity and closure by 30%.
- Implemented automated fallback procedures on local machines for cases that cross Lambda resource limits.

Projects

Cloud-Native PDF Asset Extraction System (Ongoing)

AWS Lambda, Amazon S3, DynamoDB, REST APIs, Python, Java

Developing a scalable **AWS**-based data pipeline for automated extraction of images, captions, and metadata from PDF documents. Built on an asynchronous event-driven processing architecture, leveraging **AWS Lambda** for processing, **DynamoDB** for textual storage, and **Amazon S3** for asset storage. Managed through backend Python services for PDF parsing, asset extraction, asset format conversion, metadata extraction, and cloud storage and retrieval workflows.

Built to support large-scale ingestion, indexing, and retrieval of document assets for future downstream retrieval. Applied cloud-native design principles, including serverless processing, decoupled services, and automated storage lifecycle management.

OpenCodeReasoning (Debugging focus)

QWEN-2.5, HuggingFace

A challenge to improve an LLM's ability to improve its coding and logical reasoning ability using Nvidia's OpenCodeReasoning dataset. Uses coding problem questions, buggy code, and critiques for said buggy code to improve the model's ability to catch and fix bugs.

Education

Master of Science (MSc) in Data Science and Analytics

08/2024 – 10/2025 | Maynooth, Ireland

Maynooth University

First-class honours 1:1, 81.8%

Relevant courses: *Machine Learning, Statistical Methods, Geospatial Analysis, Thesis on the utility of Vision-Language models in medical settings.*

Bachelor of Technology (BTech) in Computer Science Engineering

06/2018 – 06/2022 | Vellore, India

Vellore Institute of Technology

First class honours 1:1 87.5%

Relevant courses: *Software Engineering, Data Structures and Algorithms, Data Analytics, and Cloud Computing.*