G Santosh Kumar

+91-9738037623 — girisantoshkumar1999@gmail.com — g-santosh-kumar — gsantoshkumar1999 — gsantoshkumar.in

EDUCATION

Rajarajeswari College of Engineering, Bangalore, IN **2017 - 2021**
Bachelor of Technology in Electronics and Communications *8.69 CGPA*

SKILLS

Languages	Python, Java, SQL, JavaScript
Frameworks	FastAPI, Flask, Pydantic, TensorFlow, Keras, PyTorch, OpenCV, SciPy, Pandas, NumPy
APIs & Backend	REST API Development, Microservices Architecture, Uvicorn.
Data Engineering	Apache Airflow, BigQuery, Dataflow, Dataproc, Pub/Sub, Cloud Composer, Looker, Splunk
Databases	Firestore, BigQuery, Cloud SQL, Memorystore, Firestore Vector Store.
GenAI/LLMs	Gemini, Vertex AI platform, Agentspace (Google Enterprise), Document AI, RAG Systems, Vector Search
Agentic Frameworks	Pydantic AI.
Machine Learning	MLOps, Kubeflow, Vertex AI Pipelines, YOLOv8, Computer Vision, NLP, Recommendation Systems
Tools	Git, Docker, Docker Compose, Postman, VSCode, Pycharm, Jupyter Notebook
Cloud (GCP)	Vertex AI, Vision, Cloud Run, Cloud Functions, Cloud Build, Cloud Scheduler, Cloud Workflow, Pub/Sub, Eventarc, Dataflow, Dataproc, Cloud Tasks, Google Compute Engine, Cloud SQL, Artifact Registry, Cloud Composer, Looker, Model Registry, Document AI, Firestore, BigQuery, Google SSO, Memorystore, Google Live Model, Vertex AI for Commerce, Google Nano Banana, Google Veo3.

EXPERIENCE

- **Cloud Ambassadors** **Nov 2024 – Present**

  AI Engineer & GCP AI Architect

  – Architected and led full-stack development of **boxsand.ai** platform, deployed 8+ **GenAI** Applications (Video Intelligence, Podcraftor, Audiobook, Pixora, Medical Reports Analyzer, Social Media Sentiment Analyzer, Live Voice Agent, Pydanticai Coder Agent) using Vertex AI and microservice Architecture, accelerating customer lead generation by **40%** thereby enabling the company to achieve **GCP Premium Partner status within 6 months** and generating >**$250K in Cloud & GenAI attributed revenue**.
  – **Delivered 20+ AI/ML solutions** for enterprise clients across public safety, retail, and manufacturing sectors. Implemented computer vision systems (YOLOv8 + Vertex AI Vision) for **Sharjah Police** accident detection achieving **88% accuracy**, apparel defect detection for Lal10 reducing quality control time by **60%**, and recommendation engine for ChangePay increasing conversion rates by **18%**, generating **$150K+** cumulative project revenue.
  – Led multi-cloud migration initiative, containerizing and migrating **200+ serverless functions** (150 AWS Lambda, 50 Azure Functions) to GCP Cloud Run using Docker and Buildpacks. Achieved **55% reduction in compute costs** ($30K annual savings) and improved average response latency by **40%**, while maintaining 99.9% service availability.
  – Engineered production RAG and conversational AI systems for 6+ enterprise clients including ABP Network, Lissun, and Flipkart. Built real-time voice bot using Gemini Live Model, Vertex AI, and Firestore Vector Search processing **2M+ monthly customer interactions**. Implemented gRPC streaming with TTS/STT integration, achieving **¡500ms response time** and **99.5%** uptime across all deployments.
  – Established enterprise MLOps infrastructure using Vertex AI Pipelines and Kubeflow, implementing automated CI/CD workflows with Cloud Scheduler triggers, Model Registry versioning, and artifact management. Enabled zero-downtime model deployments with automated retraining pipelines, reducing model deployment time from **2 weeks to 2 days** and ensuring full auditability across **15+ production models**.
  – Deployed end-to-end **Agentspace** (Google Enterprise Search) implementations for 6 enterprise clients including Groww, Mamaearth, and Infoedge. Integrated third-party data sources (Confluence, ServiceNow, etc.,) with custom connectors, enabling unified enterprise search across **500K+ documents** and reducing information retrieval time by **70%**.

- **Accenture Solutions Pvt. Ltd.** **Nov 2021 - Oct 2024**

  Application Development Analyst — Python Data Analyst & Engineer

  – Designed and deployed automated ETL pipelines using Python and SQL, processing **500GB+ daily** customer and vehicle data. Optimized data transformation workflows reducing processing time by **25%** (from 8 hours to 6 hours) and eliminating manual data validation errors.

- Automated financial claim report generation and distribution system using Bash scripts and cron jobs, delivering **100+ weekly reports** to global dealer network. Eliminated 15 hours of manual work per week and reduced report delivery time from 2 days to 4 hours.
- Designed and deployed **10+ Splunk dashboards** aggregating application, functional, and error logs to monitor system performance across **5M+ daily transactions**. Integrated Microsoft Power Automate for alert workflows, reducing mean time to detection (MTTD) by **40%** and improving SLA adherence from 92% to **98%**.
- Developed data extraction and transformation workflows in Python and Java, processing XML data from SOAP web services and third-party DMS systems. Built real-time integration pipelines supporting **10K+ daily claim transactions with 99.8% data accuracy** and ¡2-minute processing latency.
- Performed root cause analysis for **120+ production incidents** using Splunk log analysis and Python debugging tools. Resolved critical data pipeline failures with average resolution time of 2 hours, preventing **$100K+** in potential claim processing delays.

PROJECTS

• **Podcraftor**: Full-stack application converting text into podcast episodes using Google TTS with SSML for natural audio. Automates transcript generation, voice configuration, music integration, and export with chapter-based customization.

• **Audiobook**: Intelligent system processing PDF/ePUB documents into audiobooks using Google TTS with SSML. Features automated chapter extraction, multi-voice support, background music integration, and customizable audio effects—achieving 95% user satisfaction in beta testing.

• **Pixora**: AI-powered image and video generation studio using Gemini 2.5 Flash, Imagen 4.0, and Veo 3.0 models. Supports text-to-image/video generation, image editing, composition, and drag-and-drop uploads.

• **Medical Docs Analyzer**: Healthcare document analysis system for interpreting handwritten prescriptions and discharge summaries using AI. Extracts KPIs, detects fraud, verifies legitimacy, and generates contextual questions to optimize insurance claim verifications.

• **Live Voice Agent**: Intelligent voice agent POC using Gemini Live Voice model with function calling capabilities. Developed for Kapture, Flipkart Mitra, and Gide.ai, enabling real-time voice interactions and automated customer support responses.

• **Gifinity**: Creative application generating sprite sheets using Gemini Nano (Banana model) and converting them to animated GIFs. Supports image uploads, camera capture, text prompts, and real-time previews with customizable speed controls.

AWARDS AND HONORS

- **ACE Award – Growth Catalyst Award – FY'24 Q1**
- **ACE Award – The Extra Mile Award in the Offshore – Team Category**
- **Spot Award Cluster 2**
- **Tech Expressway Academy – Merit Holder**
- **Brainiac Award**

CERTIFICATIONS

Generative AI Leader Certification	Google, May 2025
Data Engineering Foundations Specialization	IBM, Sep 2024
AWS Cloud Quest: Cloud Practitioner	AWS, Sep 2023
Splunk Core Certified Power User	Splunk, Jul 2023
Microsoft Certified: Azure Data Fundamentals	Microsoft, Jun 2023
Ultimate AWS Certified Cloud Practitioner	Udemy, Mar 2023
Techgenics Stream Training – JAVA	Accenture, May 2022
Machine Learning at Stanford University Online	Coursera, Aug 2020
Splunk Core Certified User	Splunk, Jul 2022
Quarter finalist in Texas Instruments Indian Innovation Challenge – 2019	Texas Instruments, 2019