Subhadip Mitra

Engineering Leader & AI Systems Architect

Singapore | +65 82501776 | contact@subhadipmitra.com
LinkedIn: linkedin.com/in/subhadip-mitra | GitHub: github.com/bassrehab

Last updated: December 20, 2025

Professional Summary

Senior Engineering Leader with 15+ years of experience bridging fundamental AI research and enterprise-scale system delivery. Currently leading Google Cloud's Data & Analytics practice for Southeast Asia while driving internal innovations on LLM inference efficiency and multi-agent systems. Proven track record of operating as a "Player-Coach": managing regional engineering portfolios while simultaneously architecting and patenting novel frameworks (UPIR, ARTEMIS, FTCS, Speculative Decoding).

Professional Experience

Google Cloud - Professional Services Organization: Data & Analytics Manager — Site Lead, PSO Southeast Asia *January 2021 – Present*

Built and lead Google Cloud's Data Analytics practice across Southeast Asia (7 countries) while serving as Site Lead overseeing regional delivery operations. Member of delta - Google Cloud's elite innovation and transformation team architecting enterprise AI solutions at scale across JAPAC.

Strategic Leadership & Delivery:

- Direct a $XXM+ Data Analytics delivery portfolio across JAPAC while simultaneously overseeing a $XXM+ cross-practice delivery portfolio as Site Lead for Southeast Asia
- Led critical interventions for strategic accounts across JAPAC including major financial services and consumer electronics manufacturers
- Executed high-value projects including 12K+ user analytics migrations, first Data & AI Centers of Excellence, and Data Monetization Platforms
- As co-Site Lead, scaled regional delivery excellence contributing to 100% annual revenue target attainment with 97% CSAT
- Spearheaded organization-wide agentic AI transformation across PSO JAPAC, driving operationalization and adoption for both external customer solutions and internal productivity gains

Technical Innovation & Research (Official IP):

- LLM Inference Efficiency: Research on speculative decoding, custom Triton kernels, and KV-cache compression strategies. Filed Google Technical Disclosure on hybrid compression systems for multi-tenant serving optimization.
- Intelligent Trust Engine (CatchMe): Developed industry-agnostic agentic AI system for enterprise-scale trust decisions across Finance, Healthcare, Insurance, Cybersecurity, and Supply Chain. Features APLS (self-learning pattern synthesis) and five-level cascade routing achieving 86% cost reduction with sub-50ms latency. Won Google Cloud PSO Hackathon JAPAC Regionals, qualified for World Finals. Two pending Google Technical Disclosures.
- Distributed Systems Automation: Created UPIR framework combining formal verification, program synthesis, and reinforcement learning for automated distributed systems development (274x speedup, 60% latency reduction).
- Context Processing Innovation: Developed FTCS (Field-Theoretic Context System) modeling context as interacting fields and ETLC framework reimagining data integration for GenAI era.
- Multi-Agent Frameworks: Published ARTEMIS framework for coordinated decision systems using adversarial debate protocols.

Standard Chartered Bank: Principal Engineer - Data & Analytics Transformation *January 2019 – January 2021*

Led enterprise-wide AI and data platform development serving 11 markets and 1200+ global users, delivering technical excellence while influencing C-suite data strategy.

- Delivered a Self-Service ML Platform that reduced model development time from 6 months to 1 week
- Designed credit risk AI models integrating alternative data sources, improving accuracy by 15%
- Modernized MarTech infrastructure, driving 30% increase in customer acquisition

Think Big Analytics (a Teradata company): Principal Data Engineer / Solution Architect *January 2017 – January 2019*

Architected enterprise-scale data solutions for Fortune 500 clients across APAC, designing scalable platforms with measurable business impact.

- Engineered 5 high-performance data lakes processing 1.2 PB/hour, achieving 20% optimization
- Built real-time fraud detection systems, reducing false positives by 60% and saving $XM annually
- Designed enterprise architectures supporting global Fortune 500 clients across APAC

Earlier Career: Software Engineering, Architecture and Technical Consulting Roles *January 2010 – January 2017*

Progressively advanced through roles in software development, systems integration, and technical consulting within financial services and algorithmic trading domains.

Research & Open Source Engineering

LLM Inference Efficiency Research
Reference implementations of acceleration techniques including speculative decoding (EAGLE, Medusa, tree speculation), KV-cache compression, and custom Triton kernels for transformer operations. Achieves 8.1x speedup with 88% peak bandwidth utilization on A100 GPUs.
GitHub: https://github.com/bassrehab/speculative-decoding, https://github.com/bassrehab/triton-kernels
Google Technical Disclosure - Pending

CatchMe - Intelligent Trust Engine
Industry-agnostic agentic AI system for enterprise-scale trust decisions across Finance, Healthcare, Insurance, Cybersecurity, and Supply Chain. Features APLS (self-learning pattern synthesis) and five-level cascade routing achieving 86% cost reduction with sub-50ms latency. Winner - Google Cloud PSO Hackathon JAPAC Regionals, qualified for World Finals.
Google Technical Disclosures - Pending: APLS & Cascade Routing

AI Metacognition Toolkit
Activation-level detection of sandbagging, deception, and situational awareness in LLMs. Linear probes achieve 90-96% accuracy across Mistral, Gemma, and Qwen models. Includes steering vectors for runtime behavior control (20% sandbagging reduction).
PyPI: https://pypi.org/project/ai-metacognition-toolkit/ | GitHub: https://github.com/bassrehab/ai-metacognition-toolkit | Docs: https://ai-metacognition-toolkit.subhadipmitra.com/

Steering Vectors for Agent Behavior Control
Runtime control of LLM agent behaviors through activation steering - modifying model outputs at inference without retraining. Achieves 65% uncertainty detection while maintaining 100% factual confidence. Features LangChain integration and multi-vector composition.
GitHub: https://github.com/bassrehab/steering-vectors-agents

Spark LLM Eval - Distributed Evaluation Framework
Enterprise-scale LLM evaluation framework on Apache Spark with statistical rigor. Features bootstrap

confidence intervals, significance testing, multi-provider support (OpenAI, Anthropic, Google), and Databricks integration.
GitHub: https://github.com/bassrehab/spark-llm-eval

UPIR - Automated Distributed Systems Synthesis
Framework combining formal verification, program synthesis, and reinforcement learning to automatically generate verified implementations from specifications. Achieves 274x speedup for complex systems with 60% latency reduction.
GitHub: https://github.com/bassrehab/upir | Technical Disclosure: https://www.tdcommons.org/dpubs_series/8852/

Publications & Technical Disclosures

ETLC: A Context-First Approach to Data Processing in the Generative AI Era
Google Cloud, May 2025
https://services.google.com/fh/files/blogs/etlc_full_paper.pdf

Field-Theoretic Context System (FTCS)
Google, Technical Disclosure Commons, May 2025
https://www.tdcommons.org/dpubs_series/8022/

ARTEMIS - Adaptive Multi-agent Debate Framework
Google, Technical Disclosure Commons, January 2025
https://www.tdcommons.org/dpubs_series/7729/

Data Monetization Strategy for Enterprises
BITS Pilani, December 2023
https://www.researchgate.net/publication/376557741

Education

MBA, Business Analytics — BITS Pilani
MTech, Software Systems — BITS Pilani

Technical Skills

Leadership: Enterprise Architecture, Technical Vision & Roadmaps, Digital Transformation, AI & Data Strategy, Cloud Architecture, C-Suite Advisory, Innovation Leadership

Data Engineering: Petabyte-Scale Data Platforms, Data Pipelines, Real-Time Processing, Data Mesh & Fabric, Apache Spark, Apache Hadoop, Apache HBase, Apache Flink, Apache Kafka, Apache Iceberg

AI & ML: Multi-Agent Systems, LLM Fine-tuning, RAG Systems, PyTorch, LangChain, LangGraph, LlamaIndex, Google ADK, MCP, A2A Protocol, MLflow, LLMOps, Vertex AI

Programming: Python, SQL, Scala, Triton, CUDA, Formal Verification, Program Synthesis, Distributed Systems

Cloud Platforms: Google Cloud Platform (GCP), BigQuery, Vertex AI, Dataflow, Cloud Composer, Cloud Run, Kubernetes, Terraform

Professional Affiliations

IEEE • ACM • Singapore Computer Society • Royal Institute of Navigation • IIT Madras