

Shiv Prakash

AI/ML Engineering Leader | ML Platforms, Agentic Systems, Data & MLOps | Ex-Walmart, Red Hat

Email – hello@shiv.io, Location - Bengaluru, India , linkedin.com/in/shivprakashin/

Engineering leader with 15+ years leading AI, ML, and data platform organizations at Walmart and Red Hat. Rare combination of org leadership and hands-on model ownership: built production systems spanning the full ML lifecycle, from training and evaluation through serving and inference optimization, including LLM fine-tuning and autonomous agents. Deep partner-of-record with Data Science organizations, converting models into realized business impact at enterprise scale. Scaled teams from 8 to 40+ across 6 geographies, with compliance built in (SOX, GDPR, CCPA, POPIA). Patent holder (2 US patents), MS in Machine Learning (Georgia Tech, in progress).

Recent Highlights

- Built and owned **production ML models end-to-end (STL, ARIMA, MHRNN demand forecasting and a return fraud model)** across training, scoring, versioning, inference, and post-processing (**MLFlow, KServe**)
- Delivered **6 autonomous agents (3 production, 3 beta)** including a **fine-tuned PHI-3** metadata agent with evaluation pipelines (**DeepEval, RAGAS**)
- Led ML-driven, CTO-sponsored **inventory optimization** realizing **\$75M** of a projected **\$500M** impact
- Re-architected the **demand forecasting platform (AIMS)** for **40% cost savings and 500% throughput gains**
- Led Next Gen **Data Mesh Data & Agentic AI Platform** for **200+ data products** and **20K users**
- Built customer-facing **conversational AI, chatbot, and self-service platform** across **5 global markets: contact rate -32%, CSAT +33.8%**
- Designed **tier-0 availability and inventory services** with **high TPS and sub-50ms latency**

Skills

- **Agentic AI & GenAI:** Multi-Agent Systems (MCP, A2A), Orchestration (LangGraph, CrewAI, LlamaIndex, FastMCP), RAG, Fine-tuning, & Vector Stores (Weaviate, Milvus, PGVector), Agent SDKs (OpenAI, Claude SDK, Google ADK, Vertex AI)
- **ML Modeling & MLOps:** Time-Series Forecasting (STL, ARIMA, MHRNN), Fraud/Risk Model Training & Scoring, Fine-tuning (PHI-3), Model Versioning & Serving (MLFlow, KServe, KubeFlow), Inference & Post-Processing Pipelines
- **LLMOps & Inference:** AI Gateway (LiteLLM), LLM Observability & Tracing (LangFuse), Content Safety/Guardrails (LlamaGuard), Evaluation Pipelines (DeepEval, RAGAS), KV Caching & Inference Optimization, LLM Serving
- **Data Platforms:** Data Mesh Strategy, Governance (Catalog, Quality), Discovery, Querying, Big Data & ETL (Spark, Kafka, Airflow, DBT, Fivetran), & Warehousing (Snowflake, BigQuery, Cassandra).
- **Cloud & Distributed Systems:** Google Cloud Platform (GCP), Azure, Kubernetes, Microservices (Spring Boot, Reactive, Java, Python, Go) & Event-Driven Architecture.
- **Databases & Observability:** Oracle RDBMS, Postgres, MySQL, Cassandra, Cosmos DB, Elasticsearch, Solr, Redis (Cache), Grafana, Prometheus, Splunk
- **Functional Domains:** Responsible AI & Governance (compliance-by-design for agentic systems: SOX, CCPA, GDPR, POPIA, Secure Code), Customer Experience & Support Automation (conversational AI, self-service, omni-channel comms), Supply Chain (Forecasting, Replenishment, Inventory, Logistics, Fulfillment, Care & Returns), Retail/eCommerce
- **Strategic Leadership:** AI Platform Strategy, Technical Roadmap, Executive Stakeholder Management, High-velocity/performance teams, Indirect management, Vision setting.

Experience Details

Independent Advisor | Agents & Data - AI Platform Consulting Architect (March 2026 – Present)

Advising early-stage AI startups (via a startup accelerator) on platform architecture and agentic system design - building agents, agentic harnesses, digital workforce foundations, and core Data & AI architectures, including model selection and evaluation frameworks.

Red Hat | Senior Manager – Data & AI (July 2025 – March 2026)

As the Data and AI Platform Lead for India, I built a high-performance engineering organization by sourcing top-tier AI talent from elite colleges and startups. I successfully **closed hiring for 11+ FTEs, including 6 intern conversions**, while **executing the AI acceleration roadmap**. Led a team of **15 direct reports, 6 interns & 7 indirect reports [28 total]**, including a mix of **Manager and Senior Architects** based in **Bangalore, Pune and Cork (Ireland)**.

My focus was solving for **Structured, Semi-Structured, and Unstructured Data, enabling Enterprise Deep Research through Multi-Agent Systems**. Managed strategic collaboration with leaders and partners across **APAC, EMEA, and the US**, and delivered unified global platform experiences.

Scaled **Data Access and Identity**, overseeing the end-to-end redesign of critical pillars including **Data Catalog, Discovery, Observability, and SRE Agentic automation** into a **simplified, scalable, and agent-native architecture** while **executing in rapid AI powered velocity**.

STRATEGIC PORTFOLIOS

Data Mesh Platform

Scaled to handle over 3 TB of curated data. Served as the central place for quarter end reconciliation company wide. **Reduced the cost of closing quarter from \$140k to \$8k. Launched version 1.2, 1.3 for platform and compliance certification for IBM ASCA & SOX.**

Data Mesh - Agentic Platform (Dataverse) & Digital workforce

Production Agents (3):

- **Dataverse Agent:** NL-to-SQL research system, multi-system context retrieval, and row-level security, 80% analytics reduction, duration reduced from 2 weeks to 3 days for report generation.
- **Platform Review Agent:** Governance automation, 30% review volume reduction
- **Metadata Enrichment Agent:** Fine-tuned PHI-3 model, 100% adoption in 2 weeks, built in line with CI/CD

Beta Agents (3):

- **SRE Agent:** On-call assistance with dynamic knowledge learning & auto provisioning.
 - **DSR Agent:** GDPR DSR Compliance request orchestration with full audit trails
 - **PBA Autonomous Agent:** Policy-based access (PBA) control and provisioning automation
- All agents ran on a centralized AI gateway - LiteLLM for unified model routing, LangFuse for observability and tracing, LlamaGuard for content safety guardrails.

AI Acceleration

Led the adoption of Gen AI tools like **Cursor, Claude CLI, Gemini Enterprise**. Increased the adoption rate to 100% in AI assisted productivity in the org of 129 associates. Led the evaluations for **models - Sonnet, Opus, Gemini Pro, Kimi (instruct) and Grok** for **coding and agentic integration**. Ran evals for **coding review agents - Code Rabbit, Claude, Gemini**.

Developer Experience

Designed & delivered **SOX compliant Programmable Dashboards**, Visualizations, and UX interfaces for AI interaction. Visualization solution solved for **next gen solution** replacing **Tableau, saving \$2.3 million**, **AI interface** acted as a **hub** for all **agent interactions and orchestrations**.

Walmart | Staff Software Engineer → Senior Software Engineering Manager II (August 2019 – June 2025)

Recruited into the Replenishment organization to solve critical operational challenges. I led the vision, formation, and execution of the Inventory Placement charter. Built the team from the ground up, **hiring 5 FTE and 5 interns**, orchestrating collaboration and alignment across **Data Science, Product, Engineering, and Business Leadership up to the SVP level** across geography. I led its entire engineering charter from Bangalore. Managed a layered team of total **22 FTEs including Sr Architects & Manager while driving global execution with partner teams**.

STRATEGIC PORTFOLIOS

Inventory placement

An **optimization system** that solved for **optimal inventory placement across Walmart Shipping Nodes**. **Realized \$75M of projected \$500M impact** in phase 1 & 2. It enabled next generation pixel-based demand forecasting and scaling next gen AI based automated FCs.

GenAI Placement Optimization Agent

Built assistant **agents** for **placements analysis and approval using multi-agent orchestrations**. Improved weekly analysis, partial or full approval of **placement decision** to under **4 hours from 48 hours**.

Ops and Engineering Excellence

Realized 31% reduction in operating cost for the existing replenishment lake via continuous monitoring of cost, migration, re-architecture, compression and optimization. Drove the engineering excellence horizontal for Replenishment. Identified and improved code, agile, incidents and productivity metrics. Improved metrics by **15% overall in code quality & builds**. **Managed 100% compliance to CVE and zero day exploits**.

Care and Returns squad and Centralized Comms platform

Led **5 backend engineers and 14+ cross functional team members in UX, UI, Mobile, Analytics, Big Data & QA**. Drove the roadmap for the product. Built Central comms platform which is **used in UK, Massmart(SA), CA, MX and Chile markets**. **GDPR, POPIA and CCPA compliant**. It **reduced the operations cost by 21.2%** and allowed Walmart to use bulk licenses for multiple comms channels with at least one channel view promise. **Improved CSAT score by 33.83%**

- Enhanced Integrated chatbot using inhouse conversational APIs (**Reduction in contact rate by 32%**)
- Maintained & Enhanced Refund and Self-Service experiences for UK Grocery & SA
- Led the **hiring of 10+ FTE in the org**.

US Walmart Stores Data Platform

Led the design, development, and delivery of automated Data Governance policies and systems, **serving 34 domain verticals** across stores. Coded and implemented key parts of the E2E solution. Established both **automated and static Schema Governance**, enabling **self-service schema onboarding, policy verification, and standardization, reducing onboarding time from 1.5 weeks to 2 days.**

Designed and developed a comprehensive **Data Catalog and Data Quality framework** to manage **data lineage, classification, discovery, and quality scoring for ingestion and data lakes.** This system reduced data discovery time maintained a clean and healthy data lake and facilitated effective data management and product development. **Daily ingestion of > 10TB.**

Led **SOX and CCPA compliance certification**, architecting and coding necessary changes across the entire store systems data platform. Additionally, developed and implemented a **Confluent-compliant Schema Registry**, ensuring faster, schema-compliant ingestion into the data lake, with a **99.99% uptime metric.** Contributed to a **Hot Store, Cold Store, ingestions router, control plane and Query routing layer.** Led cross-geo, cross function teams in India and US and drove the Roadmap.

Demand Forecasting Solution - AIMS & Data Platform

Built Automated Inventory Management System (**AIMS**), the **next gen demand forecasting platform.** Laid foundation for Functional and technical analytics platform. Made data modular and sharable. **On-boarded 2 more customers for our data in Walmart and improved processes.** Built a sharable library for ML and Cloud transfers, adopted by 6+ companywide teams. **Consolidated 5 systems into 1 platform, code bases from 35 repos to 3.** Improved developer productivity by 300%. Release time by 500%. It brought in **cost savings of 40% increasing throughput by 500%.**

Key win at the India Leadership and Global CTO level. Led cross-geo team in India and US and drove the Roadmap with cross functional teams.

Replenishment Systems

Built and Managed a 1 TB ES cluster for forecasting analytics and search query. Processed TB and Billion records for **replenishment processes** including – Ingestion from both streaming & static, aggregation and cleaning of Data. **Refreshed > 10TB of delta loads over week and 100GB daily.** Solved Forecasting for newly launched items, improving forecasting by upto 35% in some categories. Managed a collection of > 12 microservices which was a mix of Low latency, high TPS, Data access, batch and file uploads services.

Walmart | SDE 2 → Staff Software Engineer (October 2012 - August 2019)

Progressed from SDE 2(level 2) to Architect role(level 5) by August 2019, building **foundational inventory, replenishment and data platforms.**

High-Scale Distributed Systems: Architected Tier-0 **Global, Node-level inventory services** achieving sub-3ms latency at 13k TPS and an Ingestion Orchestrator processing 4 million updates/hr across 22M items. Built low latency, elastic **reservation services** under 50ms latency scaling up to 250k OPM.

Infrastructure Transformation: Led the "zero-to-one" migration of 1,200+ VMs to OpenStack/OneOps in under 2 months. Established the CI/CD backbone for 35+ service clusters, ensuring High Availability (HA) and Disaster Recovery (DR).

Performance Engineering: Built the IMS Mosaic microservices orchestration framework and optimized core workflow parsers by 500%, significantly driving down computational overhead.

Best Buy (Accenture BBTG) | Programmer → Senior Programmer (Aug 2010 – October 2012)

Retail systems modernization: Led Sterling SSFS upgrade (7.11 to 9.1), built **mobile carrier activation platform**, and optimized **omnichannel inventory availability**, expanded capability in **order managements systems for e-commerce and stores.**

Patents

- Chain of Fallback Customer Communications using Composite and Concurrent State Machines - US11973637B1
- Database System for Distribution Center Fulfillment Capacity Availability Tracking and Method - US20170091683

Education

- Masters in Science - Machine Learning (Part Time), [Georgia Institute of Technology, Atlanta, USA](#) (In Progress, 2028)
- Bachelors in Engineering - Computer Technology, [Yeshwantrao Chavan College of Engineering, Nagpur, INDIA](#)

Shiv Prakash, Bangalore, KA, IN