Joel Maria

Senior Staff Engineer | AI-Native Systems & High-Scale Architecture

joelmsanto@gmail.com

jmstechnologiesinc.com | *https://linkedin.com/in/joel-maria-960a7820* | *https://github.com/joelmsanto*

Boston, MA, 02114 | *+1 978 771 2261*

Senior Staff Engineer architecting AI-native production systems and high-scale distributed platforms. 16+ years designing event-driven, streaming-first infrastructures serving 50M+ users. Specializing in LLM orchestration, Retrieval-Augmented Generation (RAG), embeddings pipelines, vector search architectures, real-time inference integration, and cost-governed AI infrastructure across multi-tenant enterprise environments.

Work Experience

Senior Software Engineer, AI-Scale Platform Architecture Mar 2024 - Present

Upwork (Contract) | *Boston, MA*

- Led architecture for high-scale creator and gaming platforms serving 50M+ global consumers, designing distributed, event-driven systems across Node.js (TypeScript) and Python (FastAPI, asyncio).
- Architected and scaled horizontally distributed microservices and Kafka-based streaming pipelines powering real-time telemetry and mission-critical financial transactions with strong consistency guarantees and resilience under burst traffic conditions.
- Re-architected GraphQL APIs and data access layers, achieving 200% performance gains through resolver optimization, request batching, caching strategies, and query cost governance.
- Engineered non-blocking I/O, clustering, and async concurrency controls to maximize throughput while sustaining low-latency SLAs across high-traffic endpoints.
- Designed streaming-first telemetry contracts and enrichment pipelines enabling AI/ML-ready workloads for behavioral analytics, fraud detection, and adaptive monetization.
- Improved platform reliability and cost efficiency through structured observability (metrics, tracing, logs) and increased mobile performance by 35% via deep React Native Hermes optimization.

Lead Engineer – AI Voice & Distributed Banking Systems May 2021 – Jan 2024

U.S. Bank | *Boston, MA*

- Co-defined architecture for AI voice assistant powered by Node.js, Python, and TensorFlow inference services.
- Designed secure integration layer between mobile clients, ML inference services, and core banking systems.
- Implemented streaming API interactions for low-latency inference workflows.
- Designed scalable microservice communication model across distributed domains.
- Reduced application footprint by 20% through modular build strategies and dependency rationalization.
- Led 22 engineers building enterprise UI platform with 80%+ test coverage.

Fullstack Engineer – High-Scale Distributed SaaS Platform Oct 2019 – Dec 2020
Grax | Boston, MA

- Architected a Kubernetes-native Node.js platform handling 100M+ monthly API requests, enabling automated horizontal scaling and high-availability multi-tenant SaaS workloads.
- Led migration from REST polling to WebSocket streaming, reducing end-to-end latency by 50% and doubling real-time concurrency capacity.
- Optimized cloud infrastructure through container rightsizing, caching strategies, and batch processing, reducing operational costs by $67K/month.
- Designed event-driven data pipelines with BigQuery and implemented distributed OAuth2/JWT identity across services, establishing scalable retrieval and indexing patterns later leveraged for semantic and vector-based search architectures.

Senior Software Engineer – Distributed Identity & Microservices Jun 2016 – Mar 2019
Costar | Boston, MA

- Designed OAuth2 + JWT authentication systems supporting 5M+ daily logins.
- Built Backend-for-Frontend layer consolidating multiple domain services.
- Developed 60+ REST services with resilience patterns and observability hooks.
- Improved fault isolation and error handling across distributed services.

Fraud Detection Engineer – Streaming & Real-Time Systems Jul 2015 - Mar 2016
Fidelity Investment | Boston, MA

- Built real-time fraud detection services using Kafka, Redis, Django.
- Designed event-driven anomaly detection pipelines.
- Integrated external intelligence APIs for automated risk verification.

- Delivered real-time visualization dashboards for transaction monitoring.

Full-stack React.js Developer Mar 2014 - May 2015
Verizon Wireless Solutions | Boston, MA

- Designed distributed Node.js microservices supporting nationwide infrastructure systems.
- Developed cloud management modules for compute, storage, and network automation on Verizon IaaS platform.
- Enhanced platform scalability and data synchronization across distributed systems.

Senior Mobile Engineer (Contract Role) Nov 2013 - Feb 2014
Bank of America | Boston, MA

- Contributed to the modernization of hybrid mobile workflows (Cordova/PhoneGap) by implementing high-performance UI components using Backbone.js.
- Developed native bridge modules in Java and Objective-C to enable secure biometric authentication and OCR scanning features.
- Assisted in the transition of legacy Java services toward a RESTful architecture to support increased mobile concurrency for core banking modules.

Earlier Career Experience 2008–2013

- Senior Frontend & Full-Stack Engineering roles across fintech and enterprise environments, focusing on performance optimization, scalable UI systems, and early distributed architecture foundations.

Projects

Logistic and Last-Mile AI Platform Jan 2024 - Present
Upwork (Contract) | Boston, MA

- Architected a distributed AI-native logistics platform (React Native + Kubernetes) supporting real-time dispatch and routing across thousands of concurrent delivery flows.
- Designed Kafka streaming pipelines processing 50K+ GPS events/min, enabling sub-second anomaly detection and improving route deviation accuracy by 28%.
- Built a RAG-powered operational intelligence layer that reduced manual escalations by 35% through contextual retrieval and exception automation.

- Implemented embedding pipelines and low-latency vector search (<120ms), powering similarity-based decision augmentation for routing and driver behavior analysis.
- Engineered a production-grade LLM orchestration layer (prompt routing, guardrails, deterministic fallbacks), increasing automated resolution rates from 42% to 71%.
- Optimized LLM inference costs via selective invocation and token control strategies, reducing token usage by 38% and lowering monthly AI infrastructure costs by 31%.

Core Skills

AI & Distributed Systems: Event-Driven Architecture (EDA), Kafka & Kafka Streams, Retrieval-Augmented Generation (RAG), LLM Orchestration (Prompt Routing, Guardrails, Fallback Systems), Vector Embeddings & Similarity Search, Real-Time ML/LLM Inference Integration, Streaming APIs (WebSocket, SSE), High-Concurrency Microservices, API Orchestration Layers, Cost-Aware AI Infrastructure, Kubernetes, Serverless Architectures
Backend & Cloud: Node.js, TypeScript, Python, Django, Fastify, GraphQL Federation & Performance Optimization, REST, OAuth2 & JWT Distributed Auth, PostgreSQL, Distributed Data Design, AWS, GCP, Docker, CI/CD (GitHub Actions, Jenkins), Infrastructure Patterns for Horizontal Scalability
Frontend Platforms: React, React Native, Next.js (App Router, SSR/Streaming), Redux Toolkit, Mobile Runtime Optimization (Hermes), Backend-for-Frontend (BFF) Architecture

Education

Tech. University of Santiago Jan 2006 - Oct 2008
Computer Science

Certificates

Zend PHP 5 Certification Engineer
Zend Technologies