

Bastian Kuberek

Senior software engineer, 18+ years, who builds production AI and data systems end-to-end — LLM pipelines, agent platforms, and the product and infrastructure around them. Grew into full-stack owner and de-facto architect at AppliedXL, shipping production RAG, retrieval, structured-extraction, and evaluation systems over large document catalogs.

New York, NY • bkuberek@gmail.com • (631) 276-2221 • bkuberek.com • linkedin.com/in/bkuberek

Experience

Senior Software Engineer (Fullstack)

New York, NY (Remote)

AppliedXL

Aug 2024 – June 2026

- Grew over ~22 months into full-stack owner and de-facto architect at AppliedXL — a small, founder-led data-intelligence startup with an established product and codebase — taking end-to-end ownership across frontend, Python/LLM pipelines, AWS, payments, and incident response as contractors and the CTO rolled off, and building 8 services from scratch (several solo greenfield), working directly with the CEO and CTO.
- Built production RAG and research-agent systems over large document catalogs — a Neo4j knowledge-graph RAG backend, a LangGraph agent-orchestration stack, and an MCP server (9 tools, hexagonal architecture, Redis-backed SSE) that exposed previously UI-only clinical-trials intelligence to AI assistants (Claude, Cursor, ChatGPT), validated end-to-end by driving live tools to generate a research report.
- Took a carbon-market intelligence platform from hackathon prototype to production largely solo in ~9 weeks (834 commits, ~263,000 lines): an 11-registry scraper fleet ingesting 75,088 documents across 11,855 projects into a 6-stage document/LLM pipeline that parses PDFs, Excel, Word, and zipped bundles and extracts 77 structured fields, on Prefect 3 + AWS + Terraform with a React/TypeScript dashboard.
- Implemented structured extraction, entity extraction, deduplication, and normalization across messy source documents, with content-addressed S3 storage and multi-level change detection to skip unchanged documents and emit change events — avoiding re-scraping ~75,000 PDFs (18–24h per registry) and feeding downstream AI workflows.
- Built the News Engine agent/prompt platform end-to-end (React/TypeScript + Python + Auth0), giving non-engineers tools to version prompts, define agent flows, and run evaluations with rating/label review workflows — treating quality as a measurable cost/quality tradeoff and cutting evaluation cost by defaulting jobs to gpt-4o-mini.
- Ran AI observability and incident response with LangSmith, Grafana, and CloudWatch — root-caused production failures across the platform and cut the highest-volume LLM retry waste ~85%.
- Productionized a clinical-event extraction service (hexagonal Lambda) that pulls structured events from biotech press releases into a shared Postgres content store — live in production at 267,776 rows — using an is_primary consolidated-publishing model for entity resolution across duplicate extractions; feeds Kalshi's clinical-trial prediction market.
- Owned features from customer problem to production and iteration — building the customer-facing AI 'Analyst' discovery experience over Elasticsearch/OpenSearch retrieval and Stripe self-service billing — working directly with the CEO, product leadership, and enterprise customers as team ownership consolidated onto a solo builder.

Senior Software Engineer — Podcast Mission

New York, NY (Remote)

Spotify

May 2022 – Mar 2024

- Built backend services and audio-processing data pipelines for podcast creation workflows serving millions of creators, with distributed real-time transcoding on Java, Node.js, GCP, and Kubernetes.
- Led a 3-engineer WebRTC evaluation and proof-of-concept across iOS, Android, and Web, and held on-call for high-traffic podcast infrastructure (Joe Rogan releases), debugging distributed failures under peak load.

Senior Software Engineer — Data & Insights

New York, NY (Remote)

Spotify

Jan 2019 – Apr 2022

- Designed and built a company-wide data usage and discovery platform serving 1000+ data scientists, with distributed search and discovery over BigQuery, Elasticsearch, and Spanner that measurably improved the data leadership used for business-critical decisions.
- Drove software-stack standardization across 20+ engineering teams and migrated legacy systems into containerized microservices on Kubernetes, reducing technical debt and enabling independent team ownership.

Backend Engineer — Data Mission

New York, NY (Hybrid)

Spotify

Jan 2017 – Jan 2019

- Designed and built ‘Lexikon,’ a knowledge-graph system integrating user research and the data catalog — resolving entities and relationships across millions of records so data scientists could relate research to datasets, discover relevant work, and find topic experts.
- Owned end-to-end technical delivery of backend systems (Java, Python, GraphQL, gRPC, Elasticsearch, PostgreSQL, BigQuery), coordinating design decisions across infrastructure and product teams.

Software & Systems Engineer — Analytics / Internal IT

New York, NY (On-site)

Spotify

Oct 2011 – Dec 2016

- Built Spotify’s company-wide data-visualization and analytics platform (Python/Django), used for 6+ years, plus a web app running Hive/Hadoop queries with real-time progress tracking; later built and operated SSO/identity/access systems integrating Workday, LDAP, and Salesforce.

Web Developer (NYC agencies & startups)

New York, NY

Earlier Experience

Jan 2007 – Aug 2011

- Built and maintained web applications and product microsites for agency and startup clients (PHP/Symfony, JavaScript, MySQL), including SEO and analytics work that doubled page views.

Skills

AI / LLM / Agents: RAG, research agents & agentic workflows, MCP (Model Context Protocol), LangChain, LangGraph, LangSmith, OpenAI (GPT-4/4o/5.x) & Google Gemini, structured outputs, prompt versioning & evaluation frameworks, hexagonal/event-driven AI services

Retrieval, Search & Extraction: Elasticsearch / OpenSearch, Neo4j knowledge graph, embeddings, entity extraction / deduplication / normalization, structured document extraction (PDF, Excel, DOCX, zip), content-addressed storage, ranking & discovery

Languages & Frameworks: Python (async), TypeScript, JavaScript, React, Next.js, Node.js, FastAPI, Django, Pydantic, Java, SQL

Data & Pipelines: Prefect 3, PostgreSQL (RDS/Aurora), BigQuery, DynamoDB, Redis, Delta Lake, Polars, DuckDB, dbt — scalable ingest / normalize / enrich pipelines

Infrastructure & Observability: AWS (ECS, Lambda, EventBridge, SQS, S3, API Gateway), GCP, Terraform, Docker, Kubernetes, Grafana, CloudWatch, incident response & LLM ops, Auth0/OAuth 2.1, Stripe

Languages

Portuguese: Native

English: Proficient

Spanish: Advanced

Italian: Conversational

Turkish: Beginner