

Dibongo Ngoh

AI Forward Deployed Engineer/Architect

Dallas, TX | +1 (202) 907-6744 | dngoh365@gmail.com | dngoh.com | linkedin.com/in/bdhdjwkansh

PROFESSIONAL SUMMARY

Accomplished AI Forward Deployed Engineer/Architect with 16+ years of enterprise delivery experience, including deep specialization in production-grade AI systems, building AI agents and multi-agentic systems, and enterprise-scale generative AI platforms. Extensive hands-on portfolio spanning RAG, Vector DB, Agents, MCP, Evals, Guardrails, Orchestration and Observability, GPU inference optimization, and MLOps/LLMOps pipelines. Holds an extensive certification portfolio including AWS Solutions Architect Professional, AWS ML Specialty, Databricks ML Professional, SnowPro Advanced Architect, and CKA. Currently leads data and AI platform architecture at New York Life: GenAI application delivery (ClientMine), and enterprise AI enablement (Glean). Previously held Solutions Architect (AI/ML) roles at Accenture, 2nd Watch, and Rackspace. Delivers enterprise engagements across financial services, healthcare, insurance, and retail, and develops AI applications for business revenue growth and company productivity.

PROFESSIONAL EXPERIENCE

AI Delivery Engineering Architect | New York Life Insurance (TDAV)

03/2022 – Present

- Collaborate with product, security, business, C-level executives to understand business objects, and build AI Products and services that increases revenue and productivity.
- Built enterprise generative AI initiatives serving 50M+ users with 99.9% uptime; architected and deployed production RAG applications processing 100K+ daily queries with continuously updated knowledge bases using LangChain and Chroma.
- Built multi-agent systems using LangChain, LangGraph and LangSmith for automated customer service, reducing response time by 75% and improving satisfaction scores by 40%.
- Designed advanced prompt engineering strategies achieving 95% accuracy in complex reasoning tasks across Claude, GPT-4, and custom Llama models.
- Implemented semantic search solutions using Hugging Face models and vector databases, improving search relevance by 80%.
- Implemented conversation history in DynamoDB, real-time retrieval, contextual conversations for 2M+ concurrent users.
- Developed LangGraph workflows for complex multi-step AI processes, reducing manual intervention.
- Optimized LLM inference costs by 60% through strategic model selection, caching, and AWS Bedrock integrations.
- Built end-to-end MLOps pipelines on AWS EKS and Redshift, reducing model deployment time from weeks to hours.
- Created production-ready FastAPI services serving 50+ AI models with auto-scaling and load balancing.
- Implemented semantic search solutions using Hugging Face models and vector databases, improving search relevance by 80%.
- Established prompt engineering best practices and templates, standardizing LLM interactions across teams.
- Developed real-time data ingestion pipelines using Apache Kafka and Redshift for ML feature stores.
- Led architecture, SRB (security review board) readiness, and implementation planning for ClientMine 3.0 and ProspectMine, New York Life's GenAI client-intelligence applications; contributed to the production deployment that put ClientMine live for the field, and designed an extensible “mines” architecture to onboard future AI products.
- Designed ClientMine's persistence strategy across Databricks Lakebase and DynamoDB, JSON capability, latency validation, and chat-history durability, delivered the final Lakebase integration architecture.
- Led the PRISM migration to Databricks Apps: application architecture, work breakdown, table/object inventory, implementation sequencing, and delivery coordination across SI, platform, and business stakeholders.
- Designed PRISM's access-control transition from LDAP to Azure AD / Databricks group patterns with SCIM synchronization, and resolved Databricks Apps connectivity dependencies (application-domain routing, DNS/F5, network access validation).
- Re-architected the IBF platform on Databricks, defining workstreams for Salesforce ingestion, a Streamlit-based user experience, Tableau replacement, and agentic capabilities.
- Architected the enterprise Glean AI/search platform on NYL AWS with private networking, DMZ/private link, SaaS-option evaluation, SRB approval, SSO, and connector validation for supporting a pilot spanning 9 business units and 8 data connectors; built and validated the SharePoint connector.
- Designed an S3-to-Databricks ETL pipeline and the ALIP backend architecture for the Redshift-to-Databricks migration;
- Documented dbt, Databricks, and Harness CI/CD integration patterns for security review, and supported LDM/PDM Sales data-model and cloud-migration design work.

Solutions Delivery Architect - AI/ML | Ollion (formerly 2nd Watch)

09/2021 – 03/2022

- Trained and deployed deep neural networks using PyTorch and Hugging Face Transformers, increasing text-classification accuracy by 18% across 2.5M daily user queries.
- Architected and fine-tuned a custom computer vision model in PyTorch for real-time edge defect detection, cutting inference latency by 35% to 14ms.
- Engineered a distributed multi-GPU training pipeline in PyTorch and PyTorch Lightning, reducing model convergence time from 48 hours to 6 hours.

- Converted research-stage PyTorch models to optimized TensorRT / ONNX runtime environments, supporting 10,000+ QPS with 99.9% uptime.
- Lakehouse modernization. Led the customer-side migration of a legacy on-prem warehouse and Hadoop estate to a cloud Lakehouse, including Terraform-based infrastructure and orchestrated pipelines. Batch pipeline runtime fell from about 6 hours to 45 minutes, and infrastructure cost dropped about 30%.
- Built Data as a service data product that served the AI and ML team in deploying 80+ models into production for enterprise clients.

Solutions Architect - AI/ML | Rackspace Technology

03/2020 – 08/2021

- Retail demand forecasting. Built and productionized a demand-forecasting pipeline for a multi-location retailer and worked directly with merchandising teams to fold forecasts into replenishment decisions.
- Designed and deployed a customer-churn prediction model on AWS that identified at-risk accounts ahead of cancellation; worked with the client's product team to build targeted retention programs from the model output, giving them a data-driven retention process for the first time.

Senior Cloud Delivery Architect - AI/ML | ActiveSkyGate LLC

06/2019 – 03/2020

- Implemented end to end AI/ML solutions and applications on AWS Sagemaker for enterprise level business functions.
- Designed and shipped production neural networks using TensorFlow and TensorFlow Serving, boosting recommendation engine click-through rates (CTR) by 12%.
- Migrated legacy scikit-learn models to a scalable TensorFlow/Keras pipeline, increasing pipeline throughput by 45%.
- Implemented automated quantization and pruning in TensorFlow, shrinking model memory footprints by 60% for mobile deployment.

Senior Solutions Architect – AI/ML | Amazon Web Services (Professional Services)

04/2017 – 05/2019

- Built end-to-end MLOps pipelines on AWS EKS and Redshift, reducing model deployment time from weeks to hours.
- Created production-ready FastAPI services serving 50+ AI models with auto-scaling and load balancing.
- Developed real-time data ingestion pipelines using Apache Kafka and Redshift for ML feature stores.

Senior Database Software Engineer | Oracle Corporation

03/2016 – 03/2017

- Specialized in highly available and scalable systems; led customer operations (issue resolution) and projects for Database High Availability products and services for 400,000+ customers, driving revenue growth.
- Developed, deployed, tuned, and supported high-availability ODA and ExaData (Quarter and Full Rack) systems running Oracle 11g and 12c databases.

Data/ML Engineer | SoundHound AI (Amelia (formerly IPSoft))

01/2015 – 03/2016

- Specialized in NLP and conversational AI systems; designed and implemented chatbot frameworks using early transformer models and rule-based systems.
- Built custom NLP pipelines for document understanding and information extraction.
- Created automated model training and evaluation systems using Python and the boto3 SDK.
- Optimized database performance for ML workloads using PostgreSQL and Redis caching.

AI Research/Data Engineer | Sodexo

01/2013 – 12/2014

- Focused on computer vision and early deep learning applications; developed computer vision models for image classification and object detection.
- Implemented neural network architectures using TensorFlow and PyTorch.
- Built data processing pipelines for large-scale image and text datasets.

Data Analyst | Biodenta North America

01/2010 – 12/2012

- Built a foundation in software development and data systems; developed enterprise applications using Python and SQL databases.
- Built ETL systems for data warehousing and business intelligence.
- Implemented API integrations and microservices architectures.

AI/ML PORTFOLIO PROJECTS

- Argus — Built a LangGraph-based AI agent integrating Slack and Google Drive on AWS Bedrock (Claude), with observability via LangSmith; deployed successfully both locally and on ECS Fargate, then decommissioned ECS to control cost. Published to GitHub (iamdibo365/argus-agent) as an interview and portfolio asset; currently building a v2 multi-agent supervisor team version.
- DocuMind AI — Designed and built a full production agentic, multimodal RAG system on AWS using Terraform, Bedrock, OpenSearch, and Lambda.
- PolicyPilot — Authored a production-grade build guide for a Corrective RAG application on AWS Bedrock using Strands Agents, Agent Squad, OpenSearch Serverless, Step Functions, Guardrails, and AgentCore, delivered in both CDK and Terraform variants.
- PolicyPilot-N — Re-platformed PolicyPilot inference onto NVIDIA's stack (Triton Inference Server, TensorRT-LLM, NeMo Retriever NIMs), then extended with vLLM and Kubernetes; scaffolded Terraform + Kustomize for deployment on AWS EKS as hands-on proof of GPU inference architecture skills.
- GridGuard ML — Built a real-time energy anomaly detection system using XGBoost, MLflow, EKS, and Databricks, developed alongside an Xcel Energy interview process.

- VaultIQ — Built an MCP server for enterprise document Q&A.
- Enterprise Internal RAG — Built an internal RAG system connecting Slack, Google Workspace, Confluence, and Jira for enterprise knowledge retrieval.
- Built a production ETL pipeline on AWS and Databricks using a medallion architecture with Auto Loader and Unity Catalog.
- Built a freshness-aware RAG pipeline as an interview and portfolio asset for evaluating and prioritizing up-to-date retrieval results.
- Designed an air-gapped GenAI deployment architecture on Azure for a hospital scenario, originating from an Azure OpenAI L6 interview prompt.

TECHNICAL SKILLS

AI/ML & GenAI: RAG (naive, corrective, agentic, multimodal), multi-agent systems (LangGraph, CrewAI, AutoGen, Agent Squad, Strands Agents), prompt engineering, LLM integration (Claude, GPT-4, Llama), LLM fine-tuning (SFT/DPO/RFT), semantic search, enterprise AI search (Glean), NLP, computer vision, MLOps/LLMOps

Inference & GPU: AWS Bedrock, SageMaker, vLLM, SGLang, NVIDIA Triton Inference Server, TensorRT-LLM, NeMo Retriever NIMs, model caching and cost optimization

Cloud & Infrastructure: AWS (EKS, Lambda, DynamoDB, Redshift, Fargate), Azure, GCP, Terraform, CDK, Kubernetes, Kustomize, Docker, Harness, Azure AD / SSO / SCIM, private networking (DMZ, PrivateLink)

Data & Vector Stores: OpenSearch (incl. Serverless), Qdrant, Databricks (Apps, Lakebase, Auto Loader, Unity Catalog, medallion architecture), Snowflake, dbt, Fivetran, Apache Kafka, PostgreSQL, Redis

Frameworks & Languages: Python, FastAPI, Streamlit, boto3, LangChain, Hugging Face, TensorFlow, PyTorch, MCP (Model Context Protocol)

TECH STACK

LangChain, LangGraph, LangSmith, Langfuse, CrewAI, AutoGen, FastAPI, EKS, GKE, AKS, Docker, Jinja templates, Hugging Face, Ollama, FAISS, Chroma, Pinecone, Weaviate, PGVector, AWS Bedrock, AWS SageMaker, Vertex AI, GCP, Cursor, MLflow, Databricks, Snowflake, NVIDIA, Python, Streamlit, JavaScript, Flask, REST API, GitHub, Prometheus, Grafana, Trivy, New Relic, Datadog, CrowdStrike, Arize, Evidently AI, MCP

EDUCATION

Master of Business Administration, Information Systems
New Mexico Highlands University, NM | 2010 – 2013 | GPA 3.91

Bachelor of Science, Accounting
University of Buea, CMR | 2006 – 2009 | GPA 3.75

CERTIFICATIONS

- | | |
|--|--|
| <ul style="list-style-type: none">• AWS Certified ML Engineer – Associate• SnowPro Advanced Architect Certification• Databricks Certified Data Engineer Professional• AWS Certified Machine Learning – Specialty• AWS Solutions Architect Professional• Google Cloud Professional Data Engineer• AWS Business Professional• AWS Certified SysOps Administrator• AWS Certified Developer Associate• Oracle Database 12c Administrator Professional• Oracle Database 11g Administrator Associate | <ul style="list-style-type: none">• AWS Certified AI Practitioner• SnowPro Core Certification• Databricks Lakehouse Fundamentals• Certified Kubernetes Administrator (CKA)• HashiCorp Certified: Terraform Associate• AWS Certified Big Data – Specialty• AWS Technical Professional• AWS Certified Cloud Practitioner• AWS Solutions Architect Associate• Oracle Database 11g Administrator Professional |
|--|--|