KARTHIK SHINGTE

Senior AI/ML Engineer | Lead LLM & Agentic Systems Architect | GenAI Production Systems

karthikshingte21@gmail.com | +1 (774) 430-1375 | linkedin.com/in/karthikshingte

PROFESSIONAL SUMMARY

Senior AI/ML Engineer and technical lead with 12+ years building production GenAI and ML systems across healthcare (Kaiser Permanente), financial services (M&T Bank), and telecom (Lumen Technologies). Currently serving as technical lead for AI/ML initiatives at Kaiser Permanente, owning end-to-end architecture for LangGraph/LangChain multi-agent clinical intelligence systems, governing model deployment and compliance within a HIPAA-regulated environment, and coordinating cross-functional engineering teams spanning clinical informatics, data governance, risk, and platform engineering. Architected LangGraph/LangChain multi-agent clinical intelligence systems at Kaiser Permanente cutting irrelevant RAG context retrieval by ~40%, reducing care gap analysis latency by ~25%, and serving clinical analysts across KP's enterprise population health and care management programs.

Key engineering outcomes include fine-tuned BERT/RoBERTa models achieving 93% compliance classification accuracy at M&T Bank, sub-second fraud scoring via Kafka and Spark Structured Streaming, a 55% reduction in manual analyst review effort, and a 35% cut in distributed training time through parallelized hyperparameter tuning on AWS EC2. Deep AWS expertise (SageMaker, ECS, EKS, Lambda, Glue, Kinesis, CloudWatch) is complemented by Azure (OpenAI, Functions, Blob Storage) and GCP (Vertex AI, BigQuery, Cloud Run) experience, with strong proficiency in Docker, Kubernetes, Helm, GitHub Actions, MLflow, SageMaker Pipelines, and Kubeflow. Currently deepening LLM fine-tuning expertise (LoRA/QLoRA on domain-specific clinical corpora) and RAG evaluation methodology (RAGAS faithfulness and context recall scoring)

TECHNICAL STACK

Programming	Python 3.x, SQL, Java (Spring Boot), JavaScript (ES6+), Shell Scripting, Bash
Generative AI & LLMs	GPT-4, LLaMA 2, BERT, RoBERTa, Hugging Face Transformers, RAG Architecture, Prompt Engineering, Jinja2 Prompt Versioning, tiktoken Token Budgeting, SSE Streaming, Azure OpenAI, LLM Function Calling, Prompt Caching, RAGAS, TruLens, Guardrails AI, Meta LLaMA 2/3, Mistral 7B, Falcon, Anthropic Claude
Agentic AI	LangChain (ConversationalRetrievalChain, RetrievalQA), LangGraph (StateGraph, TypedDict), LangSmith, AutoGen, CrewAI, Tool-Using Agents, MCP Architecture, Multi-Step Reasoning Workflows, ReAct pattern, Chain-of-Thought prompting, LangSmith Evals
LLM Evaluation & Safety	RAGAS, TruLens, LangSmith Evals, Guardrails AI, Hallucination Detection, Faithfulness Scoring, Toxicity Filtering
Vector Databases	Pinecone, FAISS, Weaviate, ChromaDB
ML Frameworks	TensorFlow, PyTorch, Scikit-learn, Keras, XGBoost, LightGBM, LSTM, CNN, Faster R-CNN, SpaCy, NLTK, Gensim, SHAP, LIME, LoRA/QLoRA, PEFT, Hugging Face Trainer, vLLM, Sentence Transformers
NLP & Text Analytics	NER, Text Classification, Sentiment Analysis, Topic Modeling (LDA), Word2Vec, FastText, Tokenization, BERT Fine-Tuning, Transformer Architectures, Intent Classification, OCR, Text Chunking, Document Parsing, Document Extraction, Unstructured Data Processing
MLOps & Model Lifecycle	MLflow, Kubeflow, SageMaker Pipelines, SageMaker Model Registry, Model Versioning, Drift Detection, Champion-Challenger Testing, CI/CD for ML, Model Governance (SR 11-7), Explainability (SHAP, LIME), SageMaker Feature Store, Feast, Canary Deployments, Blue-Green Deployments, RLHF (exposure)
Data Engineering & ETL	Apache Airflow, AWS Glue, Prefect 2.x, Celery, PySpark, Spark SQL, Spark Structured Streaming, Kafka, Kinesis, CDC Pipelines, Incremental Loads, Schema Evolution, Idempotent ETL, Data Lineage, PyArrow, dbt (data build tool), Great Expectations, Delta Lake
Data Warehousing & BI	Snowflake, Amazon Redshift, Google BigQuery, Hive, Star/Snowflake Schema Design, Partitioning, Clustering, Power BI, Tableau, Matplotlib, Seaborn, Looker
Databases	PostgreSQL, MySQL, Oracle, SQL Server, DynamoDB, MongoDB, Redis, Cassandra, HBase — Schema Design, Query Optimization, Indexing, Stored Procedures, Connection Pooling, Replication, Performance Tuning, MongoDB Atlas, MongoDB Aggregation Pipelines
Cloud Platforms	AWS (SageMaker, ECS, EKS, Lambda, S3, RDS, Glue, EMR, CloudWatch, EC2, Kinesis, Step Functions, IAM), Azure (Blob Storage, Functions, SAS Tokens, Managed Identities, Azure OpenAI), GCP (Vertex AI, BigQuery, Cloud Run)
Big Data & Distributed	Apache Spark, PySpark, Spark SQL, Spark Streaming, Hadoop (HDFS, YARN), Hive, Amazon EMR, Databricks (exposure), Cloudera (exposure)
APIs & Web Frameworks	FastAPI (async/await, Pydantic v2, Middleware, DI, SSE), Flask, Django, React.js (Hooks, TypeScript, Redux), RESTful APIs, GraphQL, gRPC, OAuth2, JWT, SOAP
Containers & Infra	Docker, Kubernetes, AWS EKS, Helm, AWS Fargate, Amazon ECR, HPA (Horizontal Pod Autoscaler)
Messaging & Streaming	Apache Kafka, RabbitMQ, Amazon Kinesis, AWS SQS, GCP Pub/Sub
Monitoring & Observability	MLflow, Splunk, ELK Stack (Elasticsearch, Logstash, Kibana), Prometheus, Grafana, CloudWatch, Datadog
Testing & Quality	PyTest, Unittest, React Testing Library, JUnit, Postman, WireMock, SonarQube, TDD, Integration Testing, Contract Testing

CI/CD & Version Control	Git, GitHub, GitLab CI, Bitbucket, Jenkins, GitHub Actions, AWS CodePipeline, CodeBuild, CodeCommit
Methodologies	Agile/Scrum, DevOps, MLOps, DataOps, TDD, Data Mesh (exposure)

CERTIFICATIONS

- Amazon Web Services (AWS) – Certified Professional (Machine Learning Specialty / Cloud Practitioner)
- Microsoft Azure – Certified Professional (AI Engineer Associate / Data Engineer Associate)
- Google Cloud Platform (GCP) – Certified Professional (Data Engineer / ML Engineer)
- Databricks – Certified Professional (Data Engineering / Machine Learning)

PROFESSIONAL EXPERIENCE

Kaiser Permanente | **Senior AI/ML Engineer | Lead LLM & Agentic Systems Architect** November 2023 – Present

Oakland, CA (Remote)

- Serve as the technical lead for Kaiser Permanente's enterprise clinical AI initiative owning end-to-end solution architecture, driving key design decisions, and coordinating a cross-functional team of 4 data engineers, 2 ML engineers, and platform/DevOps collaborators to deliver production GenAI systems aligned with HIPAA, HITRUST, and enterprise security standards.
- Led architecture selection for KP's clinical AI platform, choosing LangGraph + LangChain for agent orchestration, FAISS and ChromaDB for vector retrieval, and FastAPI as the inference gateway-decisions documented in an ADR (Architecture Decision Record) that cut onboarding time for new ML engineers from ~2 weeks to ~4 days.
- Evaluated AutoGen and CrewAI as alternative multi agent orchestration frameworks; selected LangGraph for its explicit StateGraph control flow and TypedDict schema enforcement, which reduced nondeterministic agent behaviour in safety-critical clinical decision paths.
- Authored and enforced team engineering standards for prompt versioning (Jinja2), API contracts (Pydantic v2), async patterns (asyncio), and model lineage (MLflow) reducing PR review cycles by ~30% and cutting new-engineer onboarding from ~3 weeks to ~1 week.
- Mentored junior ML engineers on LangGraph StateGraph patterns, RAG pipeline design principles, FastAPI dependency injection, and production observability practices accelerating their time-to-first deployment on new features and reducing review cycles across the team.
- Led architecture and code reviews for all ML inference services and data pipeline changes, enforcing contract testing, parameterized PyTest coverage, and structured logging standards before production promotion.
- Conducted quarterly stakeholder presentations to clinical informatics and data governance leadership, translating complex AI system behaviors into business-impact narratives, driving sign-off on new model deployments and compliance documentation.
- Architected LangChain ConversationalRetrievalChain RAG pipelines over clinical and claims corpora, integrating FAISS and ChromaDB vector stores with tiktoken based token budgeting achieving approximately 40% reduction in irrelevant context retrieval for care gap identification and population health workflows.
- Designed and implemented multi step clinical decision workflows using LangGraph StateGraph with TypedDict schemas, enabling structured tool using agent execution for automated risk stratification, readmission risk prediction, and care gap detection across large-scale EHR datasets.
- Evaluated and benchmarked open-source foundation models including Meta LLaMA 3, Mistral 7B, and Falcon against Azure OpenAI for clinical AI use cases — selecting models based on HIPAA compliance posture, inference latency, and domain-specific RAG faithfulness scores.
- Built document AI pipelines to process unstructured clinical documents including OCR-based extraction from scanned records, text chunking strategies for token-aware RAG ingestion, section parsing of discharge summaries and pathology reports, and normalization workflows for downstream embedding generation and vector search.
- Configured Kubernetes HPA policies on EKS inference pods setting CPU/memory thresholds and custom metric scaling (Prometheus-based QPS triggers) to maintain p99 latency under 300ms during peak clinical analyst traffic windows.
- Built a Jinja2 based prompt versioning and A/B testing framework alongside an SSE streaming layer for real time LLM responses, improving prompt reproducibility and reducing end to end inference latency by approximately 25% for analyst facing clinical intelligence interfaces.
- Instrumented LangSmith tracing across all LangGraph agent workflows to capture token usage, latency, and tool call success rates per chain step enabling prompt-level debugging, faithfulness scoring via RAGAS, and iterative retrieval quality improvement cycles with the clinical informatics team.
- Integrated MongoDB Atlas as a flexible document store for unstructured clinical metadata, audit trail records, and ML pipeline intermediate outputs leveraging MongoDB aggregation pipelines for feature retrieval, Atlas Search for full-text clinical entity lookup, and TTL indexes for automated data expiry aligned with HIPAA retention policies.
- Evaluated MongoDB Atlas Vector Search alongside FAISS and ChromaDB for semantic retrieval over clinical corpora used MongoDB Atlas Vector Search for metadata-enriched embedding storage where structured clinical attributes (patient age band, diagnosis code, care program) needed to be co-queried alongside vector similarity scores in a single aggregation pipeline.
- Engineered unstructured clinical NLP pipelines processing discharge summaries, clinical notes, and pathology reports using transformer based extraction techniques, integrated with vector databases for semantic search, clinical entity extraction, and downstream analytics consumption.

- Designed and owned production-grade ETL/ELT pipelines using Python, Pandas, and Apache Airflow incorporating CDC based ingestion, incremental load strategies, idempotent processing, and schema evolution for large scale healthcare datasets including claims, EHR, pharmacy, and lab data.
- Engineered CDC based data pipelines connecting upstream HL7/FHIR clinical sources to Snowflake and PostgreSQL analytical layers, ensuring sub hour data freshness for real-time risk scoring, care management workflows, and regulatory reporting.
- Optimized columnar data transformations using PyArrow for multistage processing across millions of clinical records, improving pipeline throughput by approximately 35% for downstream ML feature generation, analytical aggregation, and BI reporting.
- Orchestrated production Airflow DAGs with retry logic, SLA monitoring, alerting, backfill strategies, and task dependency management, supporting multiple downstream consumers across real time inference and batch analytics workloads.
- Designed and optimized PostgreSQL and Snowflake schemas for healthcare analytics workloads implementing advanced SQL patterns including CTEs, window functions, multi table joins, and materialized views supporting real time ML feature serving and clinical reporting dashboards.
- Integrated Redis TTL based caching and pub/sub patterns to reduce database load by approximately 40% on high frequency ML feature lookup operations and real-time analytics endpoints, improving API response time under peak clinical workflow traffic.
- Developed and deployed ML inference services via FastAPI for risk stratification, readmission prediction, and care gap identification supporting real time REST and batch prediction modes with Pydantic v2 typed schemas, async handling, and structured error management.
- Implemented structured logging, distributed tracing, MLflow model versioning, and CloudWatch dashboards, maintaining full model lineage, drift detection alerting, and complete audit trails in compliance with HIPAA regulatory requirements.
- Owned and maintained CI/CD pipelines via GitHub Actions with comprehensive PyTest suites using fixtures, mocking, and parameterization ensuring automated regression coverage, production grade reliability, and full observability across all ML services.
- Containerized all ML inference services with Docker and managed Kubernetes (EKS/Helm) deployments, enabling zero downtime rolling updates and horizontal scaling for inference workloads with high peak traffic variability.
- Architected async first FastAPI microservices using asyncio, dependency injection, middleware, and Pydantic v2 BaseModel validation enabling high throughput, low latency ML inference serving and healthcare data APIs consumed by clinical operations and provider teams.
- Developed React.js frontend components with TypeScript, Hooks, and Context API for scalable state management — delivering interactive clinical analytics dashboards and ML result visualizations for operations, clinical, and executive teams.

Environment: Python 3.x, FastAPI, asyncio, LangChain, LangGraph, LangSmith, FAISS, ChromaDB, Pydantic v2, Redis, Pandas, PyArrow, Snowflake, PostgreSQL, Apache Airflow, AWS (S3, Glue, Lambda, EKS, CloudWatch, SageMaker), MLflow, Docker, Kubernetes, Helm, Kafka, PyTest, GitHub Actions, Jinja2, tiktoken, SSE, React.js, TypeScript, Prometheus, Grafana, Datadog

M&T Bank | **Python AI/ML Engineer** December 2021 – October 2023

Buffalo, NY (Remote)

- Designed and deployed NLP driven document intelligence pipelines using fine-tuned BERT and RoBERTa models to automate analysis of regulatory filings, audit reports, and policy documents achieving 93% compliance classification accuracy and reducing manual analyst review effort by 55% across Risk and Governance teams.
- Built BERT based intent classification and SpaCy NER models powering the internal service desk automation layer automatically routing ~82% of tier-1 support tickets and reducing average handle time from 14 minutes to under 4 minutes across internal banking operations.
- Built domain adapted text classification models to categorize customer complaints, dispute narratives, and servicing requests aligned with CFPB reporting requirements reducing triage time by 55% and improving SLA adherence within Customer Operations.
- Implemented Named Entity Recognition systems using SpaCy and transformer-based architectures to extract financial entities including account identifiers, merchant names, transaction references, and loan terms from unstructured servicing documents and scanned correspondence.
- Built LDA + transformer-based topic modeling pipeline processing ~50K monthly call center transcripts and NPS surveys surfacing 6 recurring risk themes that were escalated to executive leadership and incorporated into quarterly operational risk reporting.
- Developed and productionized fraud detection models leveraging LSTM networks and gradient-boosted ensemble techniques (XGBoost, LightGBM) to analyze transactional sequences and behavioral anomalies increasing SAR flagging recall by ~18% while holding false-positive rate below 3% across retail and commercial banking channels.
- Built real-time transaction monitoring using Apache Kafka and Spark Structured Streaming, enabling sub-second anomaly scoring for card and ACH transactions and processing millions of daily events across digital banking platforms.
- Designed and productionized time-series forecasting models using TensorFlow and PyTorch to predict portfolio-level credit risk trends, liquidity exposure, and delinquency probabilities under macroeconomic stress scenarios.
- Developed document verification and signature validation workflows integrating CNN-based image classification models to support remote onboarding and loan documentation validation in digital banking initiatives.
- Implemented parallelized hyperparameter tuning and distributed model training on AWS EC2 infrastructure, reducing model training time by 35% while optimizing performance across fraud detection and credit risk model portfolios.
- Conducted A/B testing and champion-challenger evaluations to benchmark fraud detection and credit decisioning improvements; implemented automated drift detection, stability tracking, and fairness evaluation frameworks meeting SR 11-7 regulatory compliance standards.

- Developed model fairness evaluation framework computing demographic parity, equalized odds, and disparate impact metrics across credit risk and fraud models per SR 11-7 producing model cards submitted to OCC examiners.
- Engineered distributed feature engineering pipelines using PySpark to process multi-million daily transaction records generating temporal, behavioral, and risk-based features supporting credit risk scoring and fraud detection model accuracy improvements.
- Integrated MongoDB Atlas as a document store for unstructured clinical metadata and audit trail storage, leveraging aggregation pipelines for ML feature retrieval and vector index queries alongside FAISS-based semantic search.
- Used MongoDB as the document store for unstructured servicing records, dispute narratives, and scanned correspondence ingested into NLP classification pipelines storing raw documents, extracted entities, and model inference outputs in a schema-flexible format enabling rapid iteration on NER and intent classification model features without relational schema migrations.
- Orchestrated batch and real-time ML workflows using Apache Airflow for automated data ingestion, feature recalculation, model retraining, and validation cycles across credit risk, compliance, and fraud detection systems.
- Designed reusable Pydantic v2 nested models to standardize schema definitions, enforce data contracts, and ensure consistency across banking microservices, REST APIs, and event-driven ETL pipelines.
- Designed and optimized relational database schemas in PostgreSQL, MySQL, and Oracle for high-volume banking transaction processing — implementing advanced indexing strategies, query plan analysis, stored procedures, triggers, and partitioning to achieve low-latency retrieval for risk analytics workloads.
- Built data mart and analytical layer structures supporting credit risk, fraud, and compliance reporting — implementing star-schema and snowflake-schema data models for efficient BI aggregation and regulatory reporting via Power BI and Tableau.
- Established end-to-end MLOps pipelines using Docker, Kubernetes, Jenkins, and MLflow for model versioning, CI/CD deployment, performance validation, and full audit traceability per SR 11-7 and Banking Model Risk Management guidelines.
- Collaborated with Risk, Compliance, and Data Governance teams to produce SHAP explainability reports, fairness analyses, model documentation, and audit artifacts required for regulatory sign-off and ongoing model validation cycles.
- Managed end-to-end data lineage, data quality validation, and reconciliation frameworks ensuring accuracy of training datasets, feature stores, and model scoring inputs for regulated ML systems subject to model risk management oversight.

Environment: Python, PySpark, TensorFlow, PyTorch, Scikit-learn, XGBoost, LightGBM, BERT, RoBERTa, SpaCy, LDA, MLflow, Docker, Kubernetes, Jenkins, Apache Airflow, Kafka, Spark Structured Streaming, AWS EC2, PostgreSQL, MySQL, Oracle, Snowflake, Power BI, Tableau, SHAP, LIME, LSTM, CNN, FastAPI, REST APIs, Pydantic v2

Lumen Technologies | Python Full Stack Developer – AI/ML July 2021 – November 2021

Monroe, LA (Remote)

- Designed and orchestrated scalable data pipelines using Prefect 2.x — implementing flows, tasks, YAML deployment configurations, work pools, scheduling, and block-based secrets management for secure and reliable telecom network data processing across high-volume OSS and BSS environments.
- Deployed ML anomaly detection and traffic forecasting models using Azure Machine Learning pipelines managing model registration, environment versioning, and scheduled inference jobs within Azure ML Studio, with model artifacts stored in Azure Blob Storage and inference endpoints exposed via Azure Functions.
- Automated ETL workflows ingesting structured and semi-structured data from network devices, OSS logs, CRM systems, and third-party monitoring tools into PostgreSQL and MySQL — implementing idempotent processing with incremental load strategies and data quality validation checks.
- Leveraged Pandas and NumPy for large-scale cleansing, transformation, and aggregation on network traffic, latency, and service reliability datasets supporting capacity planning, SLA reporting, and network operations analytics.
- Integrated Azure Blob Storage (azure-storage-blob SDK, SAS tokens, managed identities) for scalable cloud object storage, secure file handling, and data lake ingestion supporting telecom network telemetry and operational datasets.
- Developed ML models for anomaly detection, traffic forecasting, and predictive fault identification, integrating inference services into Flask and Django RESTful backend APIs to enable near real-time network risk scoring for NOC operations teams.
- Designed and optimized PostgreSQL and MySQL schemas supporting network performance, service provisioning, and customer analytics — implementing query optimization, indexing strategies, stored procedures, and views to improve SLA reporting and operational KPI performance.
- Engineered React.js and Redux real-time operational dashboards with custom hooks, centralized state management, and Axios integrations for NOC engineers monitoring live network KPIs, SLA risk indicators, and fault event streams.
- Containerized microservices using Docker and Jenkins CI/CD; integrated Kafka and RabbitMQ for asynchronous event streaming of network alerts, provisioning updates, and customer service events across distributed OSS/BSS platforms.
- Designed aggregation views, materialized result sets, and KPI reporting structures compatible with Power BI and Tableau, enabling real-time operational dashboards for network operations and service delivery stakeholders.
- Monitored application health and operational metrics using Splunk and ELK Stack, proactively identifying performance bottlenecks, log anomalies, and SLA risks across production telecom environments; collaborated with NOC teams to define alerting thresholds and escalation runbooks.
- Applied concurrency and task orchestration patterns to parallelize network data processing workflows, improving throughput and reducing end-to-end latency in high-volume telecom data pipeline environments.

Environment: Python 3.x, Flask, Django, React.js, Redux, REST APIs, Prefect 2.x, PostgreSQL, MySQL, Pandas, NumPy, Azure Blob Storage, Docker, Jenkins, Kafka, RabbitMQ, Power BI, Splunk, ELK Stack, PyTest, Agile/Scrum

Qualcomm | Python Developer January 2017 – November 2019

Delhi, India

- Developed scalable ETL pipelines to ingest, validate, transform, and load large volumes of device logs, modem trace files, firmware performance datasets, and test execution reports from CSV, JSON, log streams, and relational databases into analytics platforms supporting Snapdragon chipset validation programs.
- Automated scheduled batch processing and background data workflows using Celery and Apache Airflow to support nightly regression test result aggregation, silicon validation KPI generation, and performance benchmarking reporting delivered to hardware and firmware engineering teams.
- Implemented data validation and normalization frameworks ensuring accuracy, consistency, and completeness of telemetry data collected from embedded devices and lab test environments across chipset validation programs.
- Integrated asynchronous messaging systems (RabbitMQ, Kafka) for event-driven processing of device telemetry streams, automated failure alerting, and real-time regression monitoring across chipset and modem validation workflows.
- Designed Python-based RESTful backend services using Flask, Django, and FastAPI, integrating device diagnostics data, modem logs, and Snapdragon performance metrics with internal engineering analytics dashboards and automation systems.
- Deployed and managed backend applications on AWS (EC2, S3, RDS, Lambda); automated build, deployment, and environment provisioning using Jenkins CI/CD and Linux shell scripting, improving release efficiency for internal tooling.
- Designed and optimized relational database schemas in PostgreSQL and MySQL managing large-scale device performance datasets, modem diagnostics logs, and test execution records supporting silicon validation analytics and engineering reporting.
- Implemented JWT, OAuth2, and RBAC-based authentication for engineering portals handling proprietary chipset data and pre-release firmware; monitored production environments using ELK Stack and Splunk for high availability and rapid issue resolution.
- Implemented complex SQL queries, indexing strategies, query plan analysis, and stored procedures to improve data retrieval performance for engineering analytics dashboards, test regression KPI tracking, and firmware performance reporting.
- Contributed to microservices-based IoT device management platforms supporting engineering productivity and cross-team integration, improving backend service modularity and reusability across chipset engineering workflows.

Environment: Python 3.x, Flask, Django, FastAPI, Pandas, NumPy, SQLAlchemy, PostgreSQL, MySQL, AWS (EC2, S3, RDS, Lambda), Docker, Jenkins, Airflow, Celery, RabbitMQ, Kafka, REST APIs, JWT, OAuth2, ELK Stack, Splunk, PyTest, Linux, Bash, CI/CD

Helical IT Solutions | Junior Python Developer August 2013 – December 2016

Hyderabad, India

- Built scalable Python ETL applications using Flask and Django to extract, cleanse, validate, transform, and load structured and semi-structured datasets for ML training, data labeling, and enterprise analytics across multiple client engagements.
- Developed reusable, object-oriented Python modules for data validation, preprocessing, normalization, and quality-assurance checks, standardizing data workflows across large-scale annotation, labeling, and analytics projects.
- Automated high-volume batch data processing jobs using Celery, Cron jobs, and message queues (RabbitMQ, Kafka) with reliable scheduling, error handling, and retry mechanisms supporting annotation and reporting SLAs.
- Created Pandas and NumPy transformation workflows to normalize, aggregate, and enrich raw annotation, transactional, and operational data for downstream analytics teams, ML feature generation, and client deliverables.
- Designed idempotent data processing pipelines with reconciliation checks ensuring data accuracy, consistency, and reproducibility across analytical workflows and downstream ML training datasets.
- Designed and maintained PostgreSQL and MySQL database schemas supporting high-volume data labeling and annotation workflows; implemented optimized SQL queries, indexing strategies, stored procedures, and views to improve data retrieval for analytics and QA reporting.
- Integrated RESTful APIs and SOAP services with client systems, internal platforms, and third-party annotation tools for structured data exchange; implemented logging, monitoring, and exception-handling frameworks for production pipeline reliability.
- Deployed Python applications on Linux servers with Jenkins CI; developed PyTest and Unittest suites ensuring code stability, data accuracy, and regression prevention; participated in Agile/Scrum ceremonies for sprint planning and backlog grooming.

Environment: Python 2.7/3.x, Flask, Django, Pandas, NumPy, REST APIs, SOAP, PostgreSQL, MySQL, RabbitMQ, Kafka, Celery, Shell Scripting, Linux/Unix, Git/SVN, Jenkins, PyTest, Unittest, JSON, XML, Agile/Scrum

EDUCATION

Master of Science in Data Science | University of Massachusetts Dartmouth, MA, USA

Relevant coursework: Machine Learning, Deep Learning, NLP, Big Data Analytics, Statistical Modeling, Data Engineering, Cloud Computing

Bachelor of Engineering in Civil Engineering | Jawaharlal Nehru Technological University Hyderabad (JNTUH), India