NITHISH REDDY KONAKATI

+1 (682) 231-2750 | nithishkonakati@gmail.com | Portfolio | LinkedIn

PROFESSIONAL SUMMARY

- Data Engineer & Data Scientist with **6 years** of experience designing production-grade data platforms and machine learning solutions across healthcare, finance, and e-commerce.
- Designed batch and real-time **ETL pipelines** using **Spark**, **Airflow**, **NiFi**, **Kafka**, and **Flink** to process 100TB+ structured and semi-structured data.
- Developed and productionized ML models (**XGBoost**, **LSTM**, **BERT**, **CNN**) using **PyTorch**, **TensorFlow**, and **Scikit-learn** for fraud detection, forecasting, and NLP.
- Architected Lakehouse solutions using **Delta Lake**, **Iceberg**, and **Hive** integrated with **Snowflake**, **Redshift**, and **BigQuery** for scalable analytics and ML workloads.
- Led ML initiatives including LLM-based assistants, patient risk modeling, and credit risk classification using **LangChain**, **GPT-4**, and **Transformer models**.
- Delivered dashboards and automated reporting using **Power BI**, **Tableau**, **Looker**, and **Streamlit** to drive executive and operational decision-making.
- Deployed **CI/CD** and **MLOps** pipelines using **Docker**, **Kubernetes**, **Terraform**, **Jenkins**, **MLflow**, **DVC**, and **GitHub Actions** across **AWS**, **GCP**, and **Azure** environments.

TECHNICAL SKILLS

- **Languages:** Python, SQL, Scala, R, Java, Bash, PL/SQL, PowerShell
- **Data Engineering:** Apache Spark (PySpark), Airflow, Beam, NiFi, Kafka, Flink, dbt, Talend, Informatica, AWS Glue, Redshift Spectrum
- **Modeling & Warehousing:** Star Schema, Snowflake Schema, Data Vault 2.0, CDC (Debezium), Snowflake, Redshift, BigQuery, PostgreSQL, MySQL, Oracle, MongoDB, Cassandra, HBase
- **Lakehouse & Storage:** Delta Lake, Apache Iceberg, Hive, HDFS, Parquet, ORC, Lake Formation
- Cloud Platforms: AWS (S3, Glue, Redshift, EMR, SageMaker, Lambda), GCP (BigQuery, Dataflow, Vertex AI), Azure (Synapse, Data Factory)
- **ML & AI:** Scikit-learn, XGBoost, LightGBM, PyTorch, TensorFlow, MLflow, DVC, Hugging Face, LangChain, OpenAI (GPT-4), BERT, LSTM, CNN, SpaCy, Stats Models, SciPy
- **MLOps & DevOps:** Docker, Kubernetes, Jenkins, Terraform, GitHub Actions, Vertex AI Pipelines, Prometheus, Grafana, CI/CD
- **Governance & Metadata:** Great Expectations, Apache Atlas, Amundsen, Google Data Catalog, IAM, KMS
- **BI & Visualization:** Power BI, Tableau, Looker, Plotly, Seaborn, Streamlit, D3.js
- **Tools & IDEs:** Git, Jira, Apache Superset, ELK Stack, VS Code, PyCharm

PROFESSIONAL EXPERIENCE

Client: CVS Health, Texas **Oct 2023 – Present**

Role: Senior Data Engineer - ML Focus

- Designed and deployed real-time fraud detection pipelines using Apache Kafka, Flink, and Beam on GCP Dataflow, reducing detection latency from 8 minutes to under 1 minute.
- Designed and managed ETL pipelines using dbt, Apache NiFi, and Talend to ingest 80+ internal and third-party data sources into Snowflake, achieving 99.9% reliability.
- Engineered scalable data pipelines in Databricks and Spark, improving ETL performance and job run-time efficiency by 30%.
- Trained and deployed deep learning models (CNN, RNN) using TensorFlow and PyTorch to forecast patient engagement trends, improving accuracy by 15% over historical baselines.
- Partnered with data scientists to prepare and transform healthcare data for machine learning initiatives related to care optimization and risk prediction.
- Automated CI/CD workflows for data pipelines and ML deployments using GitHub Actions, Terraform, Jenkins, and Great Expectations, cutting manual deployment effort by 70%.
- Developed unit and integration tests to ensure data pipeline accuracy, integrity, and performance across production environments.
- Engineered a HIPAA-compliant data Lakehouse architecture using Delta Lake on Databricks to manage 100TB+ of structured and unstructured data from claims, EHRs, and wearable devices.

- Built secure cloud infrastructure across GCP and AWS, implementing IAM, KMS, and encryption-at-rest to meet data privacy and audit requirements.
- Delivered real-time insights through dashboards built in Tableau, Power BI, and Looker, tracking patient care metrics, resource usage, and operational performance across 25+ clinics.
- Collaborated with legal and compliance teams to enforce Lake Formation policies, audit logging, and data retention rules for sensitive health records.
- Integrated Apache Atlas for metadata tracking and data lineage, supporting data governance and regulatory audit needs.

Client: HSBC, India **May 2020 - July 2023**

Role: Data Scientist - ML & Data Engineering

- Built and tuned classification models using XGBoost and Random Forest to predict customer loan default risk, achieving 92% AUC on validation data and supporting early risk mitigation strategies.
- Developed scalable ETL frameworks using Apache Airflow, AWS Glue, and Redshift Spectrum to enable near-real-time analytics on S3-based data lakes, improving query performance and reducing processing time.
- Designed and deployed an LLM-based financial advisory assistant using LangChain and GPT-4 with a retrieval-augmented generation (RAG) pipeline, increasing client engagement with self-service tools by 50%.
- Analyzed patterns in credit card disputes and transaction failures using SQL, Pandas, and Seaborn, identifying root causes behind 20% of incidents and supporting operational improvements.
- Built NLP pipelines with SpaCy, BERT, and TensorFlow to extract structured insights from unstructured financial documents, improving data accessibility for downstream analytics by 40%.
- Enhanced dimensional data models using Star and Snowflake schema designs in Redshift and Snowflake, optimizing reporting workflows and accelerating compliance analytics.
- Automated data validation and transformation workflows using shell scripting and Python, reducing manual overhead and increasing pipeline efficiency.
- Implemented robust CI/CD pipelines for data and ML workflows using Docker, Kubernetes, Terraform, and GitHub Actions, enabling faster, reliable deployments.
- Contributed to data governance and lineage initiatives supporting compliance and audit readiness.
- Drove adoption of experiment tracking and model reproducibility practices by integrating MLflow and DVC into the model development lifecycle.

Client: Accenture, India **May 2019 - April 2020**

Role: Data Engineer

- Designed and maintained scalable data ingestion pipelines using PySpark, AWS Glue, and Talend to process 20M+ daily events, improving ETL runtime by over 30% through distributed processing and optimization.
- Built a real-time inventory monitoring pipeline using AWS Kinesis Firehose and Lambda, enabling automated stock alerts and enhancing supply chain visibility.
- Developed custom REST API connectors in Python to integrate CRM, vendor, and delivery platform data into a centralized data warehouse, improving accessibility for reporting and analytics teams.
- Worked with NoSQL databases like MongoDB to ingest semi-structured data and support downstream analytical workflows.
- Implemented CI/CD pipelines using Docker, Jenkins, and Terraform, reducing deployment overhead and improving environment consistency for data and ML workflows.
- Set up logging and monitoring infrastructure using Prometheus, Grafana, and CloudWatch, enhancing observability and accelerating incident resolution in production pipelines.
- Modeled datasets using Star schema design to support BI tools, improving query performance and reliability for executive reporting.
- Collaborated with data science teams by preparing structured datasets to support NLP and recommendation models, including sentiment and intent analysis.
- Delivered pre-aggregated and transformed datasets for customer segmentation initiatives, enabling targeted marketing and product personalization efforts.

CERTIFICATIONS

- Microsoft Azure Data Scientist Associate
- IBM Data Science Professional Certificate
- Machine Learning Specialization by DeepLearning.AI