ROHIT YADAV

📱 +91 9997498006 | 📧 roityadav@gmail.com | 📍 Noida, Uttar Pradesh | 🔗 Linkedin | 💻 rohity.com

PROFESSIONAL SUMMARY

10+ years of expertise in building and scaling end-to-end data products. I combine a strong background in large-scale data processing (PySpark, Snowflake, databricks) with advanced predictive modeling and statistical analysis. Recently, I have led the integration of LLMs into core data platforms, bridging the gap between traditional MLOps and modern AI to deliver high-impact, autonomous data solutions in cloud-native environments.

SKILLS

- **Machine Learning:** PyTorch, TensorFlow, Scikit-learn, XGBoost, Statsmodels, SciPy
- **Generative AI & LLMs** Hugging Face (Transformers, PEFT), LangChain, LlamaIndex, OpenAI API, Anthropic SDK
- **LLM Techniques**: RAG Pipelines, Fine-tuning (LoRA, QLoRA), Prompt Engineering, Agentic Workflows, RLHF
- **Vector DBs & Search:** Pinecone, Milvus, Weaviate, FAISS, ElasticSearch (Semantic Search)
- **Cloud & MLOps:**: AWS (SageMaker, Bedrock), Azure ML, MLflow, Datadog, Docker, Kubernetes
- **Data Engineering**: SQL, PySpark, Databricks, Snowflake, dbt, Pandas/Polars
- **Experimentation:** A/B Testing, Causal Inference, Bayesian Optimization, LLM Benchmarking (G-Eval, RAGAS)
- **Programming** Python, Bash, FastAPI (Deployment)

EXPERIENCE SUMMARY

Qventus Inc.	Lead Data Scientist	Aug-2024 to Present
Tata Consultancy Services Ltd.	Senior Data Scientist	Feb-2016 to Jul-2024

- **9+ years of experience in the end-to-end Machine Learning lifecycle**, specializing in sophisticated feature engineering and automated data pipelines (ETL/ELT) that transform raw, diverse sources into high-signal datasets for predictive modeling.
- **6+ years of expertise in Big Data AI Orchestration** using Azure Databricks, ADLS Gen2, and Azure Machine Learning to build and scale production-grade ML models in cloud-native environments.
- **4+ years as a Technical Lead** managing cross-functional teams of 9; bridging the gap between business stakeholders and technical execution to deliver complex AI solutions while mentoring junior scientists.
- **Advanced PySpark Developer** with deep expertise in the full SDLC; architecting high-performance model-serving layers and integrating real-time inference capabilities into enterprise applications.
- **Expert in High-Fidelity Data Architecture** using Medallion patterns (Bronze-Silver-Gold) and Unity Catalog to ensure model-grade data quality, rigorous schema governance, and reproducible research environments.
- **Specialist in Analytical Modeling & Feature Stores**, designing optimized star/snowflake schemas to power both real-time business intelligence and Gold-layer consumption for Deep Learning and LLM fine-tuning.
- **Proven track record in Performance Optimization** for AI/ML; re-engineered data ingestion workflows for large-scale datasets, **reducing model training latency by 70%** and accelerating the R&D-to-production cycle.
- **Strategic AI Architect** with a history of designing and delivering enterprise-scale analytics platforms, implementing best-in-class MLOps processes and rigorous validation methodologies.
- **Strategic Problem Solver** capable of translating abstract business challenges into viable technical models, with a recent focus on leveraging LLMs and RAG frameworks to unlock insights from unstructured enterprise data.

PROJECT EXPERIENCE

Project #1

Project Name	AI Operational Assistant
Duration	Jan-2025 to Present
Role	Senior AI Data Engineer
Client	Qventus Inc
Tech Stack	- Python, PySpark, Databricks (Delta Lake, Unity Catalog, Workflows, Feature Store, MLflow), SQL, AWS (S3, SageMaker, Lambda), Pandas, Scikit-learn, TensorFlow, LangChain, BERT/ClinicalBERT
Responsibilities	- Engineered data foundation using Databricks Medallion Architecture to fuel real-time ML models predicting OR case duration, delays, and staffing gaps — improving on-time starts by 18 % across 50+ hospitals. - Bronze Layer (Ingestion): Auto Loader + Kafka streams → raw EHR/OR logs → AI-ready event logs with 100 % fidelity, zero data loss. - Silver Layer (Refinement): PySpark + Clinical NLP pipeline (ClinicalBERT) → extracted 30+ entities (procedure, surgeon, equipment) from unstructured notes → F1-score: 0.92 → fed directly into prediction models. - Gold Layer (Consumption): Curated ML feature tables (e.g., gold.or_case_features) with engineered signals (historical duration, surgeon speed, room type) → consumed by online inference via Databricks Model Serving.

	- Databricks Feature Store → registered 50+ features with online/offline consistency → reduced model retraining drift from 12 % to <1 %. - Productionized 3 AI models using MLflow + Databricks Workflows: o Case Duration Predictor (XGBoost) → MAE reduced from 20 to 11 min o Delay Risk Classifier (TabNet) → AUC: 0.89 o OR Schedule Optimizer - Automated model monitoring: Drift detection (KS/Psi), data quality (DLT expectations), performance decay alerts → 99.7 % model uptime. - Led AI data governance: PHI tokenization, differential privacy, audit logs in Unity Catalog → passed HIPAA + SOC 2 AI addendum.
	-

Project #2

Project Name	Surgical Growth Solution
Duration	Aug-2024 to Dec-2024
Role	Senior Data Engineer
Client	Qventus Inc
Tech Stack	Snowflake, Databricks, DBT, ADF, AWS S3
Responsibilities	- Designed and automated end-to-end data ingestion pipelines integrating multi-source healthcare data into a unified Medallion architecture (Bronze → Silver → Gold), enabling a trusted analytical data layer for product teams and data science workloads. - Built scalable transformation workflows in Databricks to prepare curated datasets for advanced analytics and ML models, reducing manual data preparation time for downstream teams by 40%. - Developed production-grade DBT models (incremental, snapshots, staging, marts) with automated data quality tests, ownership metadata, and lineage tracking, improving pipeline reliability and auditability. - Refactored legacy SQL logic into modular DBT macros and reusable patterns, cutting code duplication by -30% and standardizing transformation logic across domains. - Diagnosed and optimized slow-running pipelines, improving Snowflake query performance and reducing compute costs by 20–25% through partitioning, clustering, and model re-architecture. - Led a cross-functional data migration effort and mentored a team of engineers while collaborating directly with product, analytics, and clinical stakeholders to translate requirements into technical execution.

Project #3

Project Name	Business Resilience Risk Intelligence
Duration	Jan-2024 to July-2024
Role	Microsoft Fabric Data Engineer
Client	Riskonnect Inc
Tech Stack	Microsoft Fabric (Lakehouse, Data Pipelines, Notebooks, Warehouse),, PySpark, Delta Tables, Striim, Git, Fabric Dataflows Gen2
Responsibilities	- Engineered an multi-tenant data platform on Microsoft Fabric, using multi-workspace isolation, RBAC, and Power BI RLS to enforce strict customer data boundaries across the entire pipeline. - Built metadata-driven ingestion that mirrored full multi-tenant transactional systems into OneLake using Striim CDC, with bronze-layer guardrails ensuring only onboarded customers were processed. - Designed and delivered high-performance Gold Delta models directly from Bronze, intentionally skipping the Silver layer, using PySpark ETL with partitioning, compaction, and Z-ordering to optimize query performance. - Operationalized Fabric multi-account scaling on capacity tier F6, designing workspace topologies, ingestion patterns, and cost-efficient workloads for large customer datasets. - Automated end-to-end ingestion → transformation → publish cycles using Fabric Data Pipelines, including parameterized orchestration, scheduled triggers, and CI/CD integration via Git branches. - Implemented a data quality and governance framework covering schema drift, referential integrity, row-count reconciliation, and onboarding validation before any exposure to semantic models. - Enabled cross-workspace analytics using Fabric Shortcuts to eliminate duplication while preserving tenant isolation and accelerating enterprise risk intelligence reporting.

Project #4

Project Name	AArete Payment Intelligence
Duration	Sep-2021 to Dec-2023
Role	Senior Data Engineer

Client	AArete
Tech Stack	Snowflake, Apache Airflow, Python, Pandas
Responsibilities	- Built 20+ analytical stored procedures in Snowflake analysing 100k+ daily claims, uncovering ~$29M annual savings for a 300,000-member plan, outperforming competitors 3:1 - Utilized Python (Pandas, Matplotlib) for validating SQL outputs and performing exploratory analysis on claims data to identify cost anomalies and savings trends - Built opportunity calculation engines processing 500k+ claims monthly, driving a 5% first-year reduction in claims waste and enabling clients to capture recoverable spend. - Partnered with product & engineering teams to refine payment rules, improving accuracy by 15% and reducing false positives by 25%. - Implemented monitoring with Apache Airflow, ensuring 99.5% reliability by tracking opportunity accuracy, processing time, and data quality. - Created optimized SQL views and data models for Power BI developers, collaborating on visualization logic to ensure accuracy, performance, and business relevance of dashboards.

Project #5

Project Name	Modernization of the reporting system by using Azure Data Lake and Power BI
Duration	Apr-2018 to Aug-2021
Role	Senior Data Engineer
Client	TCS Enterprise Analytics Platform
Tech Stack	Azure Data Factory, Databricks, Power BI
Responsibilities	- Reverse-engineered legacy reporting systems to design an enterprise-wide analytical data model, enabling migration from fragmented reports to a centralized Azure Data Lake architecture. - Modeled data using relational, dimensional, and hybrid schemas (OLTP → OLAP) with proper normalization/denormalization strategies to support scalable reporting and self-service BI adoption. - Established data modeling and metadata standards in partnership with the Data Governance team, driving consistent naming conventions and documentation across business units. - Unified data from Sales, Travel, Claims, HR, and Finance domains, eliminating redundant pipelines and creating a single, cross-functional source of truth for enterprise reporting. - Designed and implemented SCD Type 1 & Type 2 pipelines in ADF/Databricks to maintain full historical tracking and auditability of business-critical dimensions. - Enabled Power BI–based reporting on curated Delta Lake layers, reducing manual report refresh cycles and improving data availability SLAs for downstream analysts.

Project #6

Project Name	Jemma Migration
Duration	Aug-2016 to Mar-2018
Role	Data Engineer
Client	Prudential
Tech Stack	ADF, Databricks
Responsibilities	- Created truncate and incremental load JOB for movement of data from OLTP system to OLAP system. - Ensuring robustness of the pipeline by addressing possible failure scenarios. - Optimization of the data pipelines. - Ensuring data masking as per Data Privacy norms.

AWARDS AND ACHIEVEMENTS

1. Received the Contextual Master Award for architecting a large-scale expense data pipeline processing transactions from 500K+ users, enabling finance teams to run real-time spend analytics and strengthen continuous risk controls.
2. Winner at Division Level Speech Evaluation Contest - District 98, Toastmasters International
3. Served as VP Public Relations for TCS Nagpur Toastmasters Club
4. Conceptualized and designed magazine of TCS Nagpur Toastmasters from ground-up.

EDUCATION

Qualification	University	Grade
B. Tech	Dr. APJ Abdul Kalam Technical University (2015)	80% with Honors