RISHABH KUMAR KANDOI

AI Data Engineer

Master's in Data Science | Bachelor's in Computer Science & Engineering

Redmond, Washington, USA – 98052

+1 (585) 410-8739 | rishabhk8@gmail.com | LinkedIn | GitHub | Portfolio

PROFESSIONAL SUMMARY

Innovative Data Engineer with **5+ years** of experience, specializing in **AI-powered data solutions** and intelligent automation. Proficient in **Python, SQL, Azure (incl. Azure OpenAI), Langchain, and Machine Learning**, with a proven track record in developing **AI agentic systems** for data observability, optimizing complex **ETL** pipelines, and ensuring data integrity. Eager to leverage expertise in advanced AI applications to drive business efficiency and deliver impactful **data-driven** insights.

WORK EXPERIENCE

Microsoft | Redmond**,** WA**,** USA | **Data Engineer** | Contract via People Tech Group Ltd *Dec 2023 – Present*

- **Client Engagement: Microsoft (Mar 2024 – Present)**
  - Spearheaded the PoC and development of an **AI agentic tool** for data engineering (Python, **Langchain**, Azure OpenAI) to ensure data integrity across multi-million row **PostgreSQL** datasets.
    - Designed and orchestrated an AI prompt flow within **Azure AI Foundry** to autonomously discover **30+** logical business rules, translate natural language requirements into SQL, and proactively detect anomalies within data ingestion/transformation pipelines.
    - Automated the identification of diverse data quality issues (statistical & logical), delivering LLM-generated root cause analysis (RCA) and reducing manual validation efforts by an estimated **70%.**
  - Re-engineered SQL processes into multi-threaded **PySpark** notebooks, slashing processing time by **99.99% (from 3 days to 30 seconds)**, enabling daily identification of customer account linkages and providing sales teams with timely, informed decision-making capabilities for conversion and retention strategies.
  - Developed a **recommendation engine** using Fabric's **ML/AI capabilities**, leveraging vast datasets to predict trial-to-paid customer conversions with an **80% recall** and 70% precision, despite highly skewed data, and managed ML experiments for optimal model selection.
  - Implemented **Data Governance** and migrated Fabric workspace artifacts across Azure Subscriptions for **disaster recovery**.
  - Led migration of **3TB+** data from OnPrem & Azure SQL Server to **Microsoft Fabric**, streamlining processes and reducing overhead.
  - **Automated** 15+ manual tasks with Fabric/ADF Pipelines/Logic Apps, reducing manual intervention by 90% and integrating error handling & real-time email notifications.
- **Internal Initiatives: People Tech Group (Dec 2023 – Present)**
  - Led PoC for **Azure** Pipeline to compare 100+ **Power BI** reports, cutting manual comparison efforts by 60%.
  - Conducted PoC for MS Fabric's Copilot integration across 5+ services, enhancing automation and boosting workflow **efficiency**.

Freelancer | NY**,** USA | **Data Engineer / Data Scientist** *May 2023 – Nov 2023*

- Utilized Python (NumPy, Pandas, Seaborn, Matplotlib) for data cleaning and engineering, improving **data quality** and processing efficiency.
- Optimized **SQL** procedures, reducing database update time by 20%, increasing overall data processing speed.
- Designed efficient **Data Models** for NoSQL (MongoDB) solutions, optimizing schema design for faster data retrieval and enhanced insights.

Paytm Payments Bank | Noida**,** India | **Senior Software Engineer - Data** *Sep 2021 – Aug 2022*

- Optimized Java and SQL queries, processing **20M+ daily transactions**, with improved error handling and system resilience.
- Utilized **AWS Glue ETL**, boosting data analytic throughput by 25% through integration with S3 and Redshift.
- Led Money Transfer & Reconciliation team projects, enhancing user experience and increasing customer retention by 30%.
- Enhanced transaction notifications, reducing errors by 25% and improving accuracy by 20%.

BigBasket | Bangalore, India | **Software Engineer - Data** *Jan 2019 – Aug 2021*

- Managed end-to-end projects, optimizing Docker, Kubernetes, and Helm for **production** releases, increasing project efficiency by 30%.

- Implemented real-time **CDC** pipelines using **Debezium** to read MySQL bin-logs, reducing data latency by 25% and enabling quicker decision-making.
- Spearheaded a 500% speed increase in **cron job** execution, saving 20+ hours per week and enhancing operational efficiency.

EDUCATION

- University of Rochester | NY, USA | **Master of Science in Data Science |** GPA: 3.7/4.0 *Aug 2022 – May 2023*
- NIIT University | India | Bachelor of Science in Computer Science | **1st Rank Holder | GPA: 3.9/4.0** *Aug 2015 – Jul 2019*

TECHNICAL SKILLS

- **Programming Languages**: Python, R, SQL, C/C++, Java
- **Database Systems**: MySQL, PostgreSQL, MongoDB, Elasticsearch, Vector Databases
- **Big Data Tools**: Hadoop, MapReduce, Hive, Apache Spark, Pig
- **ETL & Infrastructure**: Databricks, Snowflake, Airflow, Kafka, Docker, Kubernetes, Data Modeling (Star/Snowflake), CDC (Change Data Capture), dbt, Terraform (IaC), Spark Structured Streaming
- **Cloud Platforms**: AWS, Azure (Data: ADF, Synapse Analytics, Fabric; AI: Azure OpenAI, CoPilot, Azure AI Studio/Foundry), GCP
- **Data Visualization**: Tableau, PowerBI (PowerApps, DAX), SSRS, Plotly, Matplotlib, Excel
- **Machine Learning/Deep Learning**: Langchain, LLMs (Azure OpenAI), Python (SciPy, Pandas, NumPy), Scikit-learn, TensorFlow, Keras, PyTorch
- **AI Development & MLOps**: LLM Application Development (Langchain, Azure OpenAI), Prompt Engineering, AI Agentic Systems, Data Observability Principles, AI-driven Root Cause Analysis (RCA)
- **Statistical Modeling**: A/B Testing, Generalized Linear Models, Clustering, Time Series Forecasting, Association Rules and Pattern Mining, Ensemble Models, Neural Network Models, Deep Learning.
- **Management & Monitoring Tools**: GitHub, JIRA, Grafana, Kibana, New Relic, Confluence, Datadog

LEADERSHIP SKILLS

- Managed and mentored a team of interns, fostering professional growth and driving successful project outcomes.
- Collaborated with cross-functional teams to deliver high-impact data solutions.
- Guided teams in data architecture design, ensuring compliance, best practices, and scalability.

PROJECTS

- **Trauma Detection (Healthcare)**
  Improved patient trauma level classification accuracy to **90%** (vs. manual classification **72%**), achieving **<5% False Negative Rate (FNR)** and **25% False Positive Rate (FPR)**. Utilized **EDA**, **sampling**, and **Ensemble Models** (ML) to outperform baseline metrics. Conducted statistical analysis to identify demographic influences.
- **Spam Page Detection (Big Data)**
  Developed a **PageRank algorithm** to detect spam pages in search engines, leveraging the **MapReduce Framework** in **Hadoop (HDFS)** for parallel data processing at scale.
- **Crime Rate Prediction (Time Series)**
  Predict crime rates across US cities using data from **Twitter**, **demographics**, and **Google Searches** related to mental health. Achieved **MSE of ~0.04**, outperforming existing studies. Identified high-impact crime categories requiring increased awareness.
- **Group Chat Text Segmentation (NLP)**
  Applied **Hierarchical Bayesian Topic Modeling** to segment **Slack** chat data, enhancing decision auditing and responsibility allocation. Involved data cleaning, tokenization, and similarity calculation to identify key topics and their relevance.

CERTIFICATIONS

- **Microsoft Certified: Fabric Analytics Engineer Associate (DP-600) | Fabric Data Engineer (DP-700)**
  Expertise in Microsoft Fabric Delta Lake, Pipelines, PowerBI, and Security.
- **Microsoft Learn Cloud Skills Challenge 2024 | Azure AI Engineer AI-102 (Expected Mar 2026)**
  Specialized in **GenAI** using **Microsoft Azure AI services** (OpenAI, AI Search, Vision, etc.).
- **Reltio Solutions Architect (Jan 2026)**
  Reltio Master Data Management (MDM) end-to-end designing the infrastructure.