Om Nankar

om.nankar@gmail.com | linkedin.com/om-nankar/ | github.com/omnankar10

Production-focused Data Scientist with 2+ years of experience engineering scalable ETL pipelines and deployment ready ML systems. Specialized in automating complex workflows using Agentic AI (CrewAI) and distributed computing (Spark/Kafka).

Skills

Languages: Python (Advanced), SQL, Bash/Shell, R, MATLAB
Data Engineering: PySpark, Apache Kafka, Airflow, dbt, PostgreSQL, neo4j
Machine Learning: PyTorch, LangChain, CrewAI, Scikit-learn, HuggingFace
Cloud & DevOps: AWS (Lambda, S3, Glue), Docker, Git/GitHub Actions, CI/CD pipelines
Visualization: Power BI, Tableau, Streamlit

Experience

Data Science Intern, Chista — Aug 2025 – Present

- Engineered an autonomous ingestion pipeline using **CrewAI (Agentic LLMs)** to parse unstructured healthcare data, **reducing client onboarding time by 40%** through automated schema mapping.
- Built a "Golden Layer" validation framework in **PostgreSQL**, implementing rigorous data quality checks (DQC) that **eliminated consistency errors** across complex formulary datasets.
- Deployed a Traceability Audit Tool (using **FastAPI/Streamlit**) allowing stakeholders to visualize model lineage and source data, directly **increasing user trust** in automated outputs.

Data Scientist, Wolters Kluwer — Jan 2023 – Aug 2024

- Developed an automated pricing engine in **Python**, replacing a legacy 7-day manual process with a 1-click execution pipeline that **reduced operational overhead by 70%**.
- Designed scalable ETL workflows to aggregate attrition and revenue metrics, feeding real-time **Power BI dashboards** used by the Finance Centre of Excellence for strategic planning.
- Optimized **SQL queries** for high-volume financial reporting, restructuring data models to improve report generation speed and accuracy for **10+ critical business metrics**.

Research Intern (Computer Vision), Symbiosis Centre for Applied AI — June 2021 – Aug 2023

- Developed deep learning pipelines (**PyTorch**) for medical imaging (MRI) and public safety object detection, resulting in **7 peer-reviewed publications**.
- Benchmarked detection architectures, optimizing model inference for edge cases in diverse datasets (dementia, Lyme disease).

Projects

Distributed Big Data Recommendation Engine (PySpark/Hadoop)

- Engineered a scalable ETL pipeline handling massive unstructured datasets using **PySpark** on a **Hadoop cluster**, performing complex transformations and **sentiment analysis** at scale.
- Built a **Hybrid Recommendation System** combining collaborative filtering and content-based approaches to predict user preferences with high precision.
- Optimized Spark jobs for **memory management and partition skew**, ensuring efficient processing of large-scale review data.

Real-Time Anomaly Detection System (Kafka & MLOps)

- Architected a streaming data pipeline using **Apache Kafka** for ingestion and **Flask** for serving, enabling **sub-second latency** for anomaly detection in sensor data.
- Deployed unsupervised learning models (**Isolation Forests**) exposed via REST APIs, containerized using **Docker** for consistent reproducibility across environments.
- Designed a **fault-tolerant consumer architecture** capable of handling high-throughput streams (simulating industrial sensor networks) without data loss.

Education

University at Buffalo – Masters of Science in Data Science — December 2025
Symbiosis Institute of Technology – Bachelor of Science in Information Technology — June 2023