Aditya Maksare

Data Engineer

adityamaksare@outlook.com | +918329001878 | LinkedIn | GitHub | Portfolio Website | Location: Pune, Maharashtra

PROFESSIONAL SUMMARY

Data Engineer with practical experience in building real-time streaming pipelines using Apache Kafka, PySpark Structured Streaming, and Python. Developed fraud detection system achieving 95%+ accuracy with sub-500ms latency through fault-tolerant ETL workflows and data quality validation. Proficient in windowed aggregations, LLM integration, and containerized deployment with Docker and MongoDB. Eager to apply skills in distributed stream processing and event-driven architectures to solve complex data engineering challenges.

TECHNICAL EXPERTISE

Data Engineering & Stream Processing:

Apache Kafka (Producer/Consumer APIs, Topic Management, Offset Handling) • PySpark Structured Streaming (Micro-batch Processing, Checkpointing, Watermarks) • Real-Time Data Pipelines • ETL/ELT Workflows • Stream Processing • Data Quality Validation • Fault-Tolerant Systems • Event-Driven Architecture • Windowed Aggregations • Dead Letter Queue (DLQ) Pattern

Programming & Databases:

Python (Data Processing, Scripting) • MongoDB (Aggregation Pipelines, Indexing, Document Modelling) • NoSQL Databases • SQL (Basic) • Data Modelling • Database Design

DevOps & Tools:

Docker (Basic) • Git • GitHub • Containerization Concepts

AI/ML Integration:

LangChain • Llama 3 (8B) • Ollama • Prompt Engineering • LLM-Powered Analytics • Generative AI

PROJECTS

AI-Powered Real-Time Fraud Detection System | GitHub

Technologies: Apache Kafka, PySpark Structured Streaming, Llama 3 (8B), MongoDB, LangChain, Docker, Flask, React.js

- Engineered **real-time streaming data pipeline** processing 14,400+ daily NSE stock transactions with **200-500ms latency**, achieving **95%+ fraud detection accuracy** using PySpark Structured Streaming and Apache Kafka
- Architected **fault-tolerant ETL pipeline** with PySpark checkpointing, exponential backoff retry logic (3 attempts), and MongoDB transaction persistence, ensuring **zero data loss** and **100% processing reliability**
- Implemented **LLM-powered classification** using Llama 3 (8B) with few-shot prompting to detect pump schemes, wash trading, and market manipulation patterns across 15 Indian stocks (₹400-₹12,000 price ranges)
- Developed **data quality framework** with schema validation, range checks, and Dead Letter Queue (DLQ) pattern for invalid records, improving pipeline reliability

Real-Time Weather Analytics Pipeline | GitHub

Technologies: Apache Spark, Apache Kafka, MongoDB, Docker, Python, PySpark Structured Streaming, Open-Meteo API

- Designed **end-to-end streaming ETL pipeline** ingesting real-time weather data from 5 Indian cities every 5 minutes, processing 1,440+ daily records with **multi-layer data quality validation**
- Implemented **3-layer stream processing architecture**: (1) Ingestion & parsing, (2) Data quality validation with DLQ routing, (3) Business logic with windowed aggregations and anomaly detection
- Built **windowed time-series aggregations** using 5-minute tumbling windows with 10-minute watermarks, computing avg/min/max temperature and windspeed metrics for real-time analytics
- Configured **multi-sink output strategy** writing validated data to both Kafka topics (clean-weather, weather-alerts) and MongoDB collections for downstream consumption

EDUCATION

Bachelor of Business Administration (Computer Applications)

Nivrutti Babaji Navale College of Commerce and Science, Lonavala, Maharashtra

Aug 2022 – Jun 2025 | **CGPA: 7.58/10**

CERTIFICATIONS

IBM Generative AI Engineering Specialization | (2024-2025)

Focus: LangChain, RAG, NLP, Prompt Engineering, Data Analysis, Flask, Machine Learning, Deep Learning