SERKAN COSKUN

Senior Data Engineer | AI/ML Engineer

sercostr@gmail.com | GitHub: github.com/sercostr | +44 7397 136049 | London, UK

PROFESSIONAL SUMMARY

Senior Data Engineer and AI/ML Engineer with 6+ years of expertise in designing and implementing enterprise-scale data pipelines, machine learning systems, and knowledge graph solutions for pharmaceutical manufacturing. Proven track record of building production-ready ML models, RAG systems, and real-time data platforms that drive operational efficiency and data-driven decision making at Pharmaceutical Digital Manufacturing.

Core Competencies:

- **AI/ML Engineering**: RAG Systems, LangChain, Vector Databases, Production ML Models, Predictive Analytics
- **Data Engineering**: Real-time ETL, Graph Databases (Neo4j), Snowflake, PostgreSQL, Data Quality Frameworks
- **Cloud & DevOps**: AWS (S3, ECS), Docker, Kubernetes, Airflow, CI/CD Pipelines
- **Programming**: Python, SQL, Cypher, PySpark, Pandas, Scikit-learn, XGBoost, TensorFlow

---

PROFESSIONAL EXPERIENCE

AI/ML Engineer | Pharmaceutical Digital Manufacturing

January 2023 - Present

Knowledge Graph & AI Systems

- **Architected enterprise Knowledge Graph platform** integrating 10+ data sources (SAP, LIMS, Snowflake) into Neo4j, enabling semantic search and graph-based analytics across pharmaceutical manufacturing operations
- **Built multi-source ETL pipelines** using Apache Airflow, processing 50K+ daily transactions with dynamic graph transformations using Cypher queries
- **Implemented document intelligence system** with automated PDF-to-knowledge extraction, generating embeddings and semantic chunks for 10,000+ technical documents

- **Developed Tech Transfer RAG application** integrating GDMS, Veeva, and Orbit document repositories, enabling natural language Q&A over manufacturing documentation
- **Created position-person-location resolution system** for workforce analytics, maintaining HR data consistency across multiple systems
- **Technologies**: Neo4j, Apache Airflow, Python, Cypher, AWS S3, Docker, LangChain, Sentence Transformers

ML Models & Predictive Analytics

- **Designed and deployed production ML models** for pharmaceutical batch manufacturing predictions using XGBoost, ExtraTreesRegressor, achieving 70%+ accuracy in delivery time estimation
- **Built real-time data refresh system** for batch tracking, processing 100K+ records daily from Snowflake UDH to PostgreSQL with less than 5 minutes latency
- **Implemented automated alert system** with schema validation, data quality checks, and anomaly detection, reducing data issues by 40%
- **Developed multi-model ML pipeline** predicting QA duration, testing times, and disposition activities across multiple manufacturing sites (Puurs, Kalamazoo, Freiburg)
- **Created model monitoring framework** detecting data drift and automatically triggering retraining workflows
- **Technologies**: Python, Scikit-learn, XGBoost, PostgreSQL, Snowflake, Oracle, SQL Server, SQLAlchemy, Pandas, NumPy

LangChain RAG Toolkit

- **Built production-ready RAG systems** using LangChain, OpenAI embeddings, and vector databases (Chroma, Pinecone, pgvector)
- **Created comprehensive AI engineering toolkit** with 10+ example projects demonstrating document processing, semantic search, SQL agents, and contextual compression
- **Developed advanced document loaders** supporting multiple formats (PDF, CSV, JSON, TXT) with metadata filtering and chunking strategies
- **Implemented SQL agent systems** enabling natural language queries to data warehouses, improving data accessibility for non-technical users
- **Designed educational framework** bridging data engineering concepts to AI/ML engineering for enterprise teams
- **Technologies**: LangChain, OpenAI API, Chroma, FastAPI, Streamlit, Vector Embeddings

Senior Data Engineer | Pharmaceutical Digital Manufacturing

January 2019 - December 2022

- **Architected batch tracking data platform** consolidating 8 source systems (SAP, LIMS, Snowflake) into unified PostgreSQL data warehouse
- **Built real-time data integration pipelines** using Python, Oracle cx_Oracle, and PostgreSQL, processing 500K+ transactions daily
- **Implemented data quality framework** with automated validation, error handling, and alerting using custom Python libraries
- **Created RESTful APIs** for batch data access, enabling downstream analytics applications and dashboards
- **Developed ETL logging and monitoring system** tracking 200+ daily ETL jobs with detailed performance metrics
- **Led migration from on-premises to cloud-based infrastructure** (AWS ECS, RDS), reducing operational costs by 30%
- **Mentored junior data engineers** on Python best practices, SQL optimization, and data pipeline design patterns

---

TECHNICAL PROJECTS

Intelligent Data Lineage System

- Built graph-based data lineage tracking using Neo4j, tracing data flow across 100+ tables and 50+ ETL jobs
- Implemented natural language interface for lineage queries using LangChain and custom agents
- **Impact**: Reduced data debugging time by 50%, improved compliance documentation

Automated Data Quality Monitoring

- Developed comprehensive alert system with 25+ data quality checks across 9 critical tables
- Implemented real-time email notifications with HTML-formatted issue summaries
- **Impact**: 40% reduction in data quality incidents, 3-hour improvement in issue detection time

Manufacturing Prediction Platform

- Created end-to-end ML platform predicting batch delivery times, QA durations, and testing schedules
- Built automated retraining pipeline with drift detection and A/B testing capabilities
- **Impact**: 70%+ prediction accuracy, enabling proactive manufacturing planning

CERTIFICATIONS

- AWS Certified Solutions Architect - Associate
- Neo4j Certified Professional
- Machine Learning Specialization (DeepLearning.AI)
- LangChain for LLM Application Development

KEY ACHIEVEMENTS

- **Published Internal Knowledge Graph**: First pharmaceutical manufacturing knowledge graph platform at Pharmaceutical
- **Production ML Models**: Deployed 12+ ML models serving 200+ daily predictions
- **Cost Optimization**: Reduced cloud infrastructure costs by 30% through architectural improvements
- **Team Leadership**: Mentored 5+ junior engineers, leading to 2 promotions
- **Innovation Award**: Recognized for AI/ML innovation in pharmaceutical manufacturing operations

---

TECHNICAL SKILLS

Languages: Python, SQL, Cypher, Bash, JavaScript

Databases: PostgreSQL, Snowflake, Neo4j, Oracle, SQL Server, Redshift

ML/AI: Scikit-learn, XGBoost, TensorFlow, LangChain, OpenAI, Vector Databases

Data Tools: Apache Airflow, dbt, Pandas, PySpark, SQLAlchemy

Cloud: AWS (S3, ECS, RDS, Lambda), Docker, Kubernetes

Visualization: Tableau, Power BI, Matplotlib, Plotly