SERKAN COSKUN

Senior Data Engineer

sercostr@gmail.com | GitHub: github.com/sercostr | +44 7397 136 049 | London, UK

PROFESSIONAL SUMMARY

Results-driven Senior Data Engineer with 6+ years of experience designing and implementing enterprise-scale data pipelines, ETL workflows, and data platforms in pharmaceutical manufacturing. Expert in building real-time data integration systems, graph databases, and cloud-native architectures that enable data-driven decision making. Proven track record of delivering high-performance, production-ready data solutions using Python, SQL, Snowflake, and Neo4j.

Core Competencies:

- Data Engineering: Real-time ETL, Data Warehousing, Data Quality Frameworks, Pipeline Orchestration, Data Governance
- **Database Technologies**: PostgreSQL, Snowflake, Neo4j, Oracle, SQL Server, Redshift
- Cloud Platforms: AWS (S3, ECS, RDS, Lambda), Azure (Data Factory, Blob Storage, Key Vault, Synapse), Docker, Kubernetes, Terraform
- Microsoft Ecosystem: Microsoft Fabric, Microsoft Purview, Azure Data Factory, Synapse Analytics
- **Programming**: Python, SQL, Cypher, Bash, PySpark
- Tools: Apache Airflow, dbt, Git, CI/CD, Jenkins, Alation

PROFESSIONAL EXPERIENCE

Senior Data Engineer | Pharmaceutical Digital Manufacturing

January 2019 - Present

Enterprise Data Platform Development

- **Architected real-time data refresh system** processing 100K+ records daily from Snowflake UDH to PostgreSQL with <5 minute latency, enabling near-real-time batch tracking analytics

- **Built automated data quality framework** with 25+ validation checks across 9 critical tables (udh_insp_dispositions, udh_prod_batches, udh_qc_results), reducing data issues by 40%
- **Designed multi-source ETL pipelines** integrating 8 source systems (SAP, LIMS, Snowflake, Oracle, SQL Server) into unified PostgreSQL data warehouse
- **Implemented alert system with email notifications** providing HTML-formatted daily monitoring reports and immediate error alerts
- **Created comprehensive ETL logging framework** tracking 200+ daily jobs with detailed performance metrics, error handling, and concurrency control
- **Optimized database performance** reducing query execution time by 60% through indexing strategies and SQL optimization
- **Technologies**: Python, PostgreSQL, Snowflake, Oracle, SQL Server, SQLAlchemy, Pandas, cx_Oracle, Docker

Knowledge Graph & Graph Database Engineering

- **Built enterprise Knowledge Graph platform** on Neo4j integrating 10+ data sources, managing 1M+ nodes and 5M+ relationships
- **Developed 50+ Airflow DAGs** orchestrating complex ETL workflows with dynamic dependencies, error recovery, and monitoring
- **Created graph transformation pipelines** converting relational data to graph structures using Cypher, enabling semantic queries and graph analytics
- **Implemented data lineage tracking system** providing end-to-end visibility of data flow across 100+ tables and 50+ ETL jobs
- **Designed HR position-person-location resolution system** maintaining workforce data consistency across multiple systems
- **Built document processing pipeline** extracting metadata and content from 10,000+ technical documents (GDMS, Veeva, Orbit)
- **Established graph database monitoring** with performance tracking, query optimization, and capacity planning
- **Technologies**: Neo4j, Apache Airflow, Python, Cypher, AWS S3, Docker, pandas, s3fs

Cloud Infrastructure & DevOps

- **Led migration from on-premises to AWS cloud infrastructure** reducing operational costs by 30%
- **Containerized 20+ data applications** using Docker and deployed to AWS ECS with automated CI/CD pipelines
- **Implemented infrastructure as code** using Terraform for AWS resource provisioning
- **Created monitoring dashboards** using CloudWatch and custom Python scripts for pipeline health tracking

- **Established disaster recovery procedures** with automated backups and tested recovery processes
- **Technologies**: AWS (S3, ECS, RDS, Lambda, CloudWatch), Docker, Kubernetes, Terraform, Jenkins

Data Quality & Governance

- **Developed data validation framework** with automated schema checks, null value detection, and business rule validation
- **Created data profiling tools** analyzing data patterns, distributions, and anomalies across 50+ tables
- **Implemented data cataloging system** documenting table structures, lineage, and business definitions
- **Established data quality SLAs** with automated alerting for threshold breaches
- **Built data reconciliation processes** ensuring consistency between source systems and data warehouse

Enterprise Data Governance & Microsoft Fabric Integration

- Architected end-to-end data governance pipeline on Microsoft Fabric integrating Microsoft Purview with Alation for enterprise-wide data cataloging and lineage tracking
- Built automated data ingestion workflows extracting metadata and business glossary from MSList and Transactional Compass systems into Microsoft Purview using Python notebooks
- Developed Purview-to-Alation synchronization pipeline ensuring bi-directional metadata flow and maintaining data consistency across 1000+ assets
- Implemented Azure-native data pipelines leveraging Azure Data Factory, Synapse Analytics, and Fabric notebooks for scalable data processing
- Established secure data infrastructure using Azure Key Vault for secrets management, Azure Blob Storage for staging data, and Storage Accounts with container-level access controls
- Created data quality checks within Fabric notebooks validating data integrity during ingestion and transformation phases
- Designed governance reporting dashboards in Power BI tracking metadata coverage, lineage completeness, and data quality metrics across pharmaceutical operations with interactive visualizations for stakeholder reporting
- Technologies: Microsoft Fabric, Azure Data Factory, Microsoft Purview, Alation, Azure Blob Storage, Azure Key Vault, Power BI, Python, PySpark, Synapse Analytics

---

KEY ACHIEVEMENTS

Data Governance Excellence: Implemented enterprise-wide governance platform cataloging 1000+ data assets, improving data discovery by 70%

- **Real-time Data Platform**: Reduced data latency from 24 hours to <5 minutes for critical manufacturing metrics
- **Cost Optimization**: Decreased cloud infrastructure costs by 30% through architectural improvements
- **Data Quality**: Achieved 99.5% data quality score across production datasets
- **Performance Improvement**: Optimized query performance by 60% through indexing and SQL refactoring
- **Team Leadership**: Mentored 5+ junior data engineers, leading to 2 promotions
- **Knowledge Graph**: Delivered first pharmaceutical manufacturing knowledge graph at Pharmaceutical
- **Pipeline Automation**: Automated 80+ manual data processes, saving 100+ hours monthly

---

TECHNICAL PROJECTS

Batch Tracking Data Warehouse

- Built comprehensive data warehouse consolidating 8 source systems into unified platform
- Processed 500K+ daily transactions with complex business logic and transformations
- **Impact**: Enabled real-time manufacturing visibility for 200+ stakeholders

Graph-Based Data Lineage

- Implemented Neo4j-based lineage tracking across 100+ tables and 50+ ETL jobs
- Created visual lineage explorer for impact analysis and debugging
- **Impact**: Reduced data debugging time by 50%, improved compliance documentation

Automated Data Quality Monitoring

- Developed comprehensive alert system with 25+ validation checks
- Implemented email notifications with HTML-formatted issue summaries

Enterprise Data Governance Platform - Microsoft Fabric & Purview

Built end-to-end data governance pipeline integrating MSList and Transactional Compass to Microsoft Purview, then syncing to Alation

- Implemented automated metadata extraction using Microsoft Fabric notebooks and Azure services
- Orchestrated data pipelines using Azure Data Factory with secure Azure Key Vault integration
- Developed interactive Power BI dashboards for governance metrics visualization and executive reporting
- Impact: Centralized data catalog for 1000+ data assets, improved data discovery by 70%, reduced compliance audit time by 50%

CERTIFICATIONS

- AWS Certified Solutions Architect - Associate
- Neo4j Certified Professional
- Snowflake SnowPro Core Certification
- Databricks Certified Associate Developer
- Apache Airflow Fundamentals

TECHNICAL SKILLS

Programming & Scripting

Python (Expert), SQL (Expert), Cypher, Bash, PySpark

Databases & Data Warehouses

PostgreSQL, Snowflake, Neo4j, Oracle, SQL Server, Redshift, pgvector

ETL & Orchestration

Apache Airflow, dbt, Custom Python ETL frameworks, Pandas, SQLAlchemy

Cloud & Infrastructure

AWS (S3, ECS, RDS, Lambda, CloudWatch), Azure (Data Factory, Blob Storage, Key Vault, Synapse Analytics), Microsoft Fabric, Microsoft Purview, Docker, Kubernetes, Terraform

Data Quality & Testing

Great Expectations, Custom validation frameworks, pytest, Data profiling tools

Version Control & CI/CD

Git, GitHub, Jenkins, GitLab CI, Docker Hub

Visualization & Monitoring

Tableau, Power BI, Grafana, CloudWatch, Custom dashboards

Additional Tools

Jupyter Notebooks, VS Code, DBeaver, pgAdmin, Neo4j Browser, Postman, Alation, Microsoft Fabric Notebooks