Sophia Valhari

Senior Data Engineer & Cloud Data Architect | Cloud-Native Data Platforms & Distributed Systems | Real-Time Data, Lakehouse & AI Platforms | AWS • Azure • GCP

sophia.walhaei@gmail.com

Profile

Senior Data Engineer & Cloud Data Architect with 9+ years of experience designing, building, and operating scalable, secure, production-grade data platforms for healthcare, finance, and enterprise SaaS organizations. Deep expertise in cloud-native lakehouse architectures, multi-terabyte ETL/ELT pipelines, and real-time streaming systems, supporting analytics, ML workloads, and business-critical reporting.
Proven track record of owning data platforms end-to-end in regulated, compliance-driven environments, improving pipeline reliability, reducing data latency by 50%, cutting runtimes by 45%, and maintaining 99.9% SLA reliability. Experienced in data governance, incident response, schema evolution, and cross-team delivery, with hands-on leadership transforming complex, failure-prone datasets into trusted, actionable business intelligence. Adept at leveraging Big Data technologies and multi-cloud platforms (AWS, Azure, GCP) to deliver resilient data systems that scale with business growth and operational demands.

Skills

Enterprise Data Engineering: ETL/ELT pipelines using dbt Core & Cloud, PySpark, Airflow, Talend, Dagster, Apache NiFi; implementing Change Data Capture (CDC), event-driven workflows, multi-terabyte batch & streaming processing, Data Mesh, Data Fabric, pipeline optimization

Software Development & Programming: Python, SQL/T-SQL, Scala, Java, R, Bash, YAML, JSON, REST, gRPC, Node.js; PostgreSQL, ClickHouse, DuckDB; backend services, API integrations, high-performance data processing

Business Intelligence & Analytics: Power BI, Tableau, Looker; SQL reporting, KPI dashboards, self-service BI enablement, executive and operational reporting, cross-functional analytics, data storytelling

Security, Governance & Compliance: HIPAA, SOC2, GDPR, FHIR; IAM, encryption, tokenization, HashiCorp Vault, Unity Catalog, access control, OpenLineage, Marquez, SIEM tools, anomaly detection, risk management, compliance reporting

Monitoring & Observability Systems: Prometheus, Grafana, Datadog, ELK Stack, Monte Carlo, SIEM tools; metadata management, operational metrics, anomaly detection, automated monitoring pipelines

Big Data & Real-Time Platforms: Hadoop, Apache Kafka (Streams, KSQL, Kinesis), Apache Flink, Apache Beam, Pub/Sub; Delta Lake, Iceberg, Hudi, ClickHouse, Druid, DuckDB; building scalable, fault-tolerant real-time systems

Cloud Platforms & Infrastructure: Azure (Synapse, Databricks, Data Lake, AKS), AWS (S3, Lambda, Glue, Redshift), GCP (BigQuery, Dataflow, Vertex AI); Snowflake, Palantir Foundry, Epic Clarity, Unity Catalog, Redpanda; multi-cloud deployments, cross-region replication, cloud cost optimization

AI & Machine Learning Engineering: Azure (Synapse, Databricks, Data Lake, AKS), AWS (S3, Lambda, Glue, Redshift), GCP (BigQuery, Dataflow, Vertex AI); Snowflake, Palantir Foundry, Epic Clarity, Unity Catalog, Redpanda; multi-cloud deployments, cross-region replication, cloud cost optimization

Data Modeling & Architectural Design: Star schema, snowflake schema, dimensional modeling, data vault, pipeline optimization, high-volume processing, real-time analytics, fault-tolerant architecture, scalable system design

Technical Leadership & Collaboration: Agile/Scrum, cross-functional leadership, strategic planning, stakeholder engagement, architecture reviews, data contracts, mentorship, team development, fostering high-performance culture

Professional Experience

Lead Data Engineer, West Monroe

- Led the architecture, development, and production operations of a multi-cloud enterprise data platform across Azure Databricks, AWS, and Snowflake, enabling high-volume ingestion, transformation, and consumption for mission-critical business domains.
- Designed and owned scalable, production-grade ETL/ELT frameworks using dbt, PySpark, Airflow, and Snowflake Streams/Tasks, reducing processing time by 45% while ensuring fault tolerance, reliability, and SLA adherence across real-time and batch workloads.
- Built event-driven, low-latency streaming systems using Kafka, Flink, and Redpanda to support near real-time decisioning, operational alerting, and downstream AI/ML workloads.
- Defined and enforced lakehouse architecture standards using Delta Lake, Databricks Unity Catalog, Iceberg, and metadata-driven automation, improving data reliability, lineage visibility, schema governance, and long-term maintainability across domains.
- Partnered with machine learning teams to enable AI-driven data platforms, supporting ML pipelines with MLflow, feature stores, real-time inference APIs, vector search, and RAG-based systems, accelerating experimentation and time-to-insight.
- Implemented end-to-end security, governance, and observability (IAM, encryption, tokenization, Datadog, Monte Carlo), reducing production incidents by 40% while meeting internal audit, regulatory, and risk-control requirements.
- Mentored and guided cross-functional engineers, established DevOps and platform standards (Terraform, GitHub Actions, Docker, Kubernetes), and led cloud cost optimization initiatives, reducing compute and storage spend by 25% while improving overall platform reliability.

Senior Data Engineer – Healthcare Analytics, Andela

- Engineered scalable, production-ready ETL/ELT pipelines for claims, eligibility, encounters, and EHR datasets using PySpark, SQL, Azure Data Factory, and Databricks, improving data freshness, stability, and SLA adherence for clinical and operational analytics.
- Designed and maintained healthcare data models (patient, provider, utilization, HEDIS, quality metrics, risk scores) ensuring accuracy, traceability, and regulatory alignment across BI and reporting teams.
- Developed automated HEDIS & CMS quality reporting pipelines, reducing manual processing by 70%, improving audit readiness, and ensuring timely year-end reporting for compliance-driven healthcare operations.
- Built enterprise data marts and semantic layers supporting population health, care management, cost-of-care analysis, and executive dashboards using Power BI, Looker, and dbt, accelerating data-driven decision-making across clinical and operational teams.
- Implemented data quality monitoring and governance frameworks using Great Expectations, Delta validation, and metadata-driven checks, reducing data defects by 40% across critical healthcare domains and ensuring production reliability.
- Collaborated with clinicians, analysts, and product teams to deliver predictive risk indicators, cost optimization models, and clinical insights, integrating AI/ML-driven analytics to accelerate decision-making across care coordination programs.
- Partnered with cross-functional teams to standardize healthcare data pipelines, maintain regulatory compliance, and enable secure data access, ensuring audit-ready, reliable, and high-impact analytics.

BI & Data Integration Engineer, Rangle.io

- Developed and optimized end-to-end ETL pipelines for finance, operations, and customer domains using SQL, SSIS, Talend, and Python, improving data reliability, refresh performance, and SLA adherence across enterprise reporting platforms.
- Designed data warehouse models (Star, Snowflake, conformed dimensions) enabling standardized KPIs, analytics consistency, and scalable self-service BI adoption, reducing time-to-insight for business units.
- Built interactive Power BI and Tableau dashboards for executives, automating performance reporting and reducing manual spreadsheet processes by 60%, improving decision-making speed and accuracy.
- Implemented data quality, validation, and reconciliation frameworks across ingestion layers, ensuring accuracy of critical metrics such as revenue, churn, product usage, and service-level KPIs, enhancing trust in enterprise reporting.

- Integrated APIs, flat files, and legacy on-prem systems into unified, production-ready data pipelines, accelerating data availability for analytics teams and reducing ETL overhead.
- Partnered with finance, product, and IT stakeholders to define metrics, streamline workflows, and deliver reliable, compliant data pipelines, supporting operational and strategic decision-making across the organization.

Education

Bachelor's in Computer Science