Afshae Re

Woodbridge, VA
afshaere@gmail.com

Summary

Senior Data Engineer with **7+ years of experience** designing, building, and operating **scalable, cloud-native data platforms** across **AWS, Azure, and GCP**. Expert in **Apache Spark, Kafka, Airflow, dbt, and SQL**, with hands-on experience delivering **batch and real-time data pipelines**, modern **lakehouse and data warehouse architectures**, and **analytics-ready datasets**.

Proven track record in **end-to-end data platform architecture**, migrating legacy systems to cloud-native solutions, and optimizing large-scale **ETL/ELT workflows** for performance and cost efficiency. Strong background in **data modeling, data reliability engineering, CI/CD automation, data observability, and governance**, enabling trusted BI, reporting, and machine learning use cases. Recognized for technical leadership, system design ownership, and building resilient, production-grade data systems.

Experience

Alphabridge

Senior Data Engineer — 12/2023 - Present

- Architected and owned **end-to-end data platform design**, delivering **batch and streaming pipelines** using **Apache Spark, Airflow, Kafka, and AWS S3**, processing **multi-terabyte datasets daily**.
- Led migration from **on-premise data warehouse to Snowflake**, redesigning ELT workflows and dimensional models, improving **query performance by 40%** and reducing **infrastructure costs by 25%**.
- Designed and deployed **real-time ingestion pipelines** using Kafka and Spark Structured Streaming to support **near-real-time dashboards, alerts, and operational analytics**.
- Built curated **analytics and feature engineering datasets** for downstream **machine learning and advanced analytics** use cases.
- Implemented **data quality checks, observability dashboards, pipeline monitoring, and SLA enforcement**, significantly improving data reliability.
- Established **CI/CD pipelines for data workloads**, enabling automated testing, controlled deployments, and faster release cycles.
- Provided technical leadership through design reviews, code reviews, and mentoring junior engineers.

Key Achievements

- Reduced critical ETL runtimes from **6 hours to under 2 hours** through Spark optimization, partitioning strategies, and query tuning.
- Improved platform stability and on-call readiness through proactive monitoring and standardized incident response practices.

AGIT

Data Engineer — 05/2020 - 11/2023

- Designed and maintained **scalable ELT pipelines** using **Python, SQL, AWS Glue, Airflow, and Amazon Redshift**.
- Built **analytics-ready datasets and dimensional data models** powering BI dashboards and ad-hoc analytics in **Power BI**.
- Automated ingestion from **REST APIs, relational databases, and flat files**, supporting diverse business domains.
- Implemented **incremental loading, CDC patterns, and data validation frameworks** to improve data freshness and trust.
- Collaborated cross-functionally with product, analytics, and business stakeholders to translate requirements into robust technical solutions.

Key Achievements

- Increased data availability SLA from **95% to 99.9%**.
- Reduced cloud data processing costs by **20%** through workload optimization and scheduling improvements.

Roche

Data Analyst 06/2018 - 04/2020

- Analyzed large datasets using **SQL, Python, and Excel** to deliver actionable business insights.
- Developed and maintained **Power BI and Tableau dashboards** supporting operational and executive reporting.
- Partnered with stakeholders to define metrics, KPIs, and analytical requirements.
- Performed **data validation, cleansing, and exploratory data analysis** to ensure accuracy and consistency.
- Supported data engineering initiatives by validating upstream pipelines and data outputs.

Key Achievements

- Automated recurring reports, reducing manual effort by **40%**.
- Improved data accuracy and stakeholder trust through standardized validation processes.

Skills

Programming, Scripting & Query Languages

- **Python (Advanced):** Pandas, NumPy, PySpark, REST API integration
- **SQL (Expert):** complex joins, CTEs, window functions, query optimization
- **Scala:** Spark-based data processing (intermediate)
- Java (basic), Bash / Shell scripting

Big Data & Distributed Processing

- **Apache Spark:** Spark SQL, DataFrames, RDDs, partitioning, caching, broadcast joins
- Hadoop ecosystem: **HDFS, Hive, YARN**
- Large-scale batch processing, fault-tolerant distributed systems

Streaming & Event-Driven Architectures

- **Apache Kafka:** topics, partitions, offsets, consumer groups, exactly-once semantics
- **Spark Structured Streaming**
- Event-driven and near-real-time analytics architectures

Cloud Platforms & Data Services

- **AWS:** S3, EC2, Glue, Lambda, EMR, Redshift, Athena, IAM, CloudWatch
- **Azure:** Data Factory (ADF), Synapse Analytics, ADLS Gen2
- **GCP:** BigQuery, Dataflow, Cloud Storage
- Cloud-native **data lakes and lakehouse architectures** (Bronze / Silver / Gold patterns)

Databases, Storage & File Formats

- **Data Warehouses:** Snowflake, Amazon Redshift, Google BigQuery
- **Relational:** PostgreSQL, MySQL
- **NoSQL:** MongoDB
- **Formats:** Parquet, ORC, Avro, JSON

ETL / ELT, Orchestration & Integration

- **Apache Airflow:** DAG design, scheduling, retries, SLAs, dependency management
- **dbt:** models, tests, snapshots, documentation
- AWS Glue, Informatica
- Incremental loads, CDC patterns, schema evolution

Data Modeling & Analytics Engineering

- Dimensional modeling (**Star & Snowflake schemas**)
- Fact & dimension tables, analytics data marts
- **Slowly Changing Dimensions (SCD Type 1 & Type 2)**

DevOps, CI/CD & Infrastructure

- Git, GitHub / GitLab, branching strategies
- CI/CD pipelines for data workloads
- Dockerized Spark and ETL jobs
- Infrastructure as Code (**Terraform – foundational**)

BI, Reporting & Semantic Layers

- Power BI, Tableau, Looker
- KPI definition, metric standardization
- Semantic layer design and self-service analytics

Data Quality, Reliability & Governance

- Data validation, reconciliation, anomaly detection
- **Data observability, pipeline monitoring, and alerting**
- Data SLAs / SLOs, incident response support
- Query tuning, indexing, partition pruning
- Data lineage, metadata management, access control (RBAC)
- Security, data privacy, and compliance fundamentals

Projects

Real-Time Analytics Data Pipeline

- Built an end-to-end **real-time data pipeline** using **Kafka, Spark Structured Streaming, and Snowflake**.
- Implemented event ingestion, transformation, and storage to power **near-real-time dashboards and alerting systems**.
- Optimized streaming workloads for **low latency, scalability, and fault tolerance**.

Cloud-Native Data Warehouse Platform

- Designed and implemented a **cloud-native data warehouse and lakehouse-style architecture** using **AWS and Snowflake**.
- Developed ELT workflows with **dbt and Airflow**, applying **star schema modeling** and analytics engineering best practices.
- Enabled executive dashboards and self-service analytics with improved **performance, governance, and reliability**.

Education

Bachelor's Degree in Computer Science