HASON SHOK

Lead Data Engineer | Principal Data Engineer | Cloud & Big Data Architect I (567) 654-3996 | hassonshok@gmail.com | Madera, CA

Summary

Results-driven **Lead / Principal Data Engineer** with **8+ years of experience** designing, building, and optimizing **enterprise-scale data platforms**, **lakehouse architectures**, and **real-time streaming systems** across **AWS, Azure, and GCP**. Deep expertise in **ETL/ELT pipelines**, **cloud-native data engineering**, **distributed systems**, **CDC**, **big data processing**, and **workflow orchestration**. Proven ability to **lead cross-functional engineering teams**, modernize **legacy data ecosystems**, and deliver **secure, scalable, high-performance analytics solutions**. Strong background in **SQL performance tuning**, **CI/CD automation**, **Infrastructure as Code (IaC)**, **data governance**, **data SLAs/SLOs**, **cost optimization (FinOps)**, and **regulatory compliance (HIPAA, GDPR)**.

Experience

Lead Data Engineer | 12/2022 - Present | AGIT

- Architected and led the implementation of a **20TB+/month cloud-native Lakehouse platform** using **Delta Lake** and **Databricks**, enabling both **batch and real-time analytics**.
- Designed **low-latency streaming pipelines (using Apache Kafka and Spark Structured Streaming to power operational and predictive dashboards.**
- Built **CDC-based ingestion frameworks** with **Debezium** and **Kafka Connect**, enabling **near real-time replication** into **Snowflake** and **Amazon Redshift**.
- Led **data platform engineering initiatives**, standardizing **ingestion**, **transformation**, **data quality**, and **observability layers** across teams.
- Defined and enforced **data SLAs and SLOs**, improving **pipeline reliability**, uptime, and stakeholder trust.
- Implemented **cost optimization (FinOps) strategies** across **Databricks**, **Spark**, and cloud storage, reducing compute and storage costs while maintaining performance.
- Designed **metadata management, data lineage, and data contract standards** to support **governance**, **auditability**, and scalable **data mesh-style ownership models**.
- Implemented **end-to-end CI/CD pipelines** using **GitHub Actions**, **Docker**, and **Terraform**, enabling fully automated, version-controlled deployments.
- Developed **data quality and observability frameworks** using **Great Expectations**, **Python**, and **Airflow**, reducing data incidents by over **60%**.
- Enforced **enterprise data governance and security controls** including **IAM/RBAC**, **encryption**, and **audit logging**, ensuring **HIPAA and GDPR compliance**.
- Mentored and led a team of data engineers, improving delivery velocity and system reliability by **30%+**.

Senior Data Engineer | 11/2020 - 11/2022 | Mountainise

- Designed and implemented **scalable ETL/ELT pipelines** on **AWS Glue**, **Lambda**, **EMR**, and **Snowflake**, integrating **50+ data sources**.
- Built **incremental and CDC ingestion pipelines** using **Kafka** and **Debezium**, reducing data latency to **under 2 minutes**.
- Optimized **Apache Spark workloads** using **partitioning**, **caching**, **broadcast joins**, and **columnar formats (Parquet, ORC, Avro)**, achieving **40% performance gains**.
- Automated orchestration, monitoring, and alerting using **Apache Airflow** and **dbt**, reducing pipeline failures by **50%**.
- Developed **dimensional data models** and **semantic layers** for **Tableau** and **Power BI**.
- Supported **feature-ready data pipelines** enabling downstream **ML and advanced analytics** use cases.

Data Engineer | 10/2018 - 10/2020 | SA Data Global

- Migrated legacy **SQL Server** and **SSIS** pipelines to **Azure Data Factory**, **Azure Databricks**, and **ADLS Gen2**.
- Designed **enterprise data warehouse schemas**, **materialized views**, and optimized **indexing strategies**.
- Tuned complex **SQL queries and stored procedures**, reducing report runtimes by **up to 70%**.
- Built **CI/CD pipelines** using **Azure DevOps** for automated testing and deployment.
- Developed **data reconciliation, validation, and audit frameworks** for financial datasets.

Associate Data Engineer | 04/2016 - 09/2018 | Blue Peak

- Built **data ingestion pipelines** using **Python** and **Apache Airflow** for APIs, flat files, FTP feeds, and relational databases.
- Worked extensively with the **Hadoop ecosystem (HDFS, Hive, MapReduce)**.
- Implemented **schema enforcement**, **data validation**, and **automated quality checks** across **40+ pipelines**.
- Supported **GCP migration initiatives** using **BigQuery**, **Dataflow**, **Pub/Sub**, and **Cloud Storage**.

Data Analyst | 03/2015 - 03/2016 | Weston Chase

- Developed **SQL reports** and **Power BI dashboards** for marketing and operations teams.
- Performed **data cleansing**, **transformation**, and **integration** for ETL workflows.
- Delivered **ad-hoc analytics** and actionable business insights.

Core Skills

Data Engineering & Architecture
ETL, ELT, Data Pipelines, Data Platforms, Data Lake, Data Lakehouse, Delta Lake, Data Warehousing, Data Modeling, Dimensional Modeling, Star Schema, Snowflake Schema, Slowly Changing Dimensions (SCD Type I & II), Schema Evolution, Partitioning, Indexing, Clustering, Materialized Views

Big Data & Distributed Processing

Apache Spark (PySpark, Scala), Spark Structured Streaming, Hadoop (HDFS, Hive, MapReduce), Apache Flink, Apache Beam, Dask, Distributed Computing

Streaming, CDC & Event-Driven Systems

Apache Kafka, Kafka Streams, Kafka Connect, Debezium, AWS Kinesis, Google Pub/Sub, Event-Driven Architecture, Near Real-Time Processing

Cloud Platforms

AWS (S3, Glue, EMR, Redshift, Athena, Lambda, Kinesis, IAM)
Azure (Azure Data Factory, ADLS Gen2, Databricks, Synapse Analytics)
GCP (BigQuery, Dataflow, Composer, Pub/Sub, Cloud Storage)

Orchestration & Analytics Engineering

Apache Airflow, dbt, Prefect, Dagster, Workflow Orchestration, Dependency Management, SLA Monitoring

Programming & Querying

Python, SQL, PySpark, Scala, Java, Bash, Shell Scripting

Databases & Storage

Snowflake, Amazon Redshift, BigQuery, PostgreSQL, SQL Server, MySQL, Oracle, MongoDB, Cassandra

DevOps, Platform & Automation

CI/CD, GitHub Actions, Jenkins, Azure DevOps, Terraform, Docker, Kubernetes, Infrastructure as Code (IaC), Data Platform Engineering

Data Quality, Observability & Reliability

Great Expectations, Data Validation, Data Profiling, Data Quality Frameworks, Anomaly Detection, Logging, Metrics, Monitoring, Alerting, Data SLAs, Data SLOs, Pipeline Reliability

Security, Governance & Compliance

IAM, RBAC, PHI/PII Masking, Encryption (At Rest & In Transit), Audit Logging, Metadata Management, Data Lineage, Schema Registry, Data Contracts, HIPAA, GDPR, Compliance

Advanced Analytics & ML Enablement

Feature Engineering Pipelines, Feature-Ready Data, Feature Stores (Offline/Online), ML Data Enablement

Analytics & BI

Tableau, Power BI, Superset, Semantic Layer, Analytics Engineering

Key Projects

Real-Time Patient Monitoring Platform (HIPAA-Compliant)
Kafka, **Spark Structured Streaming**, **Delta Lake**, **Databricks**, **Snowflake**, **AWS**, **Terraform**

- Processed **40K+ events per minute** from EHR, IoT, and lab systems.
- Reduced alert latency from **15 minutes to .**
- Implemented **PHI/PII masking**, **encryption**, **RBAC**, and **audit logging**.

Enterprise Data Lake Migration (On-Prem → Azure)
Azure Data Factory, **Databricks**, **ADLS Gen2**, **Synapse**, **PySpark**

- Migrated **15+ years of legacy data** to a modern **Azure Lakehouse**.
- Reduced nightly batch runtimes by **60%**.

CDC-Based Data Replication Framework
Debezium, **Kafka Connect**, **PostgreSQL**, **MySQL**, **Snowflake**

- Reduced replication lag from **45 minutes to .**

Data Quality & Observability Platform
Great Expectations, **Airflow**, **Python**, **Databricks**, **Grafana**

- Implemented **120+ validation rules**, anomaly detection, and SLA reporting.
- Reduced data quality incidents by **70%**.

Education

Bachelor's in Computer Science