OSAMA SHAHID

Senior Data Engineer | AWS • Azure Databricks • PySpark • SQL • Streaming & CDC • BI and ML Enablement
+91 8126688826 | osamazaidi75@gmail.com | linkedin.com/in/osama-shahid-391b84aa | github.com/osamazaidi75

15+ TB migrated at under 5 min cutover | 50M+ records/day | 8 business teams and 40+ dashboards | ~90% fewer schema incidents

SUMMARY

Senior Data Engineer with 4 years at TCS designing and running production data infrastructure that serves both business intelligence and machine learning workloads, on AWS (Glue, Lambda, Step Functions, EventBridge, DMS, Amazon MSK, S3, Athena, IAM) and Azure Databricks (PySpark, Delta Lake, Unity Catalog, ADF), in SQL, Python and PySpark. Depth in CDC and event driven streaming, dimensional modeling, warehouse schema design, data quality and governance, and infrastructure as code. Tech lead of a 4 engineer pod on life sciences data platforms for a Fortune 500 pharmaceutical client. AWS Certified Data Engineer – Associate.

EXPERIENCE

Tata Consultancy Services (TCS) Ltd. — IT Analyst (Senior Data Engineer, Tech Lead) **Nov 2022 – Present**
Client: Fortune 500 pharmaceutical (Life Sciences) | Noida, India | Production data platforms across AWS and Azure Databricks

Enterprise Data Migration and Real-Time Ingestion on AWS

- Migrated 15+ TB of biomedical and commercial data from SQL Server to AWS Aurora PostgreSQL with under 5 minutes of production cutover downtime by building CDC based ETL/ELT pipelines on AWS DMS, Glue (PySpark) and Step Functions, converting 120+ legacy stored procedures.
- Enabled near real time ingestion from 12 upstream source systems, processing 2M+ events/day at under 5 minute end to end latency, using event driven pipelines built on EventBridge, Python Lambda functions, Glue PySpark jobs and Kafka / Amazon MSK.
- Built ML backed pipelines classifying PII/PHI and extracting structured metadata from 50K+ documents/month with AWS Comprehend, Textract and CloudWatch alerting, feeding HIPAA and GxP compliance analytics.
- Cut deployment time from days to under 30 minutes across 4 AWS accounts with CloudFormation, Terraform, GitHub Actions and Docker packaged Python services.

Commercial Analytics Lakehouse on Azure Databricks: BI and ML Enablement

- Delivered governed, analytics ready data to 8 business teams and 40+ dashboards through bronze–silver–gold (medallion) pipelines in PySpark, Spark SQL, Delta Lake and Delta Live Tables with Unity Catalog governance, orchestrated via Azure Data Factory and Databricks Workflows.
- Modeled curated data into star schema fact and dimension tables served through Databricks SQL, Snowflake, Synapse and Athena — one query ready source of truth for reporting, statistical analysis and model inputs.
- Processed 50M+ records/day at 15 minute freshness by implementing incremental and streaming ingestion with Spark Structured Streaming and Auto Loader into Delta tables.
- Reduced Spark job runtime 45% and compute cost 30% by tuning partitioning, Z-ordering, broadcast joins and query plans on production workloads.

Data Integrity, Validation and Reliability

- Built data quality, validation and reconciliation frameworks covering every stage from ingestion to consumption, with automated test suites and documented standards adopted across 80+ pipelines.
- Cut schema related production incidents ~90%, from roughly 20/month to 2/month, by standardising schema evolution, data contracts and metadata driven ingestion.
- Reduced pipeline failures 40% across 30+ Apache Airflow DAGs through root cause analysis, retries, automated backfills, SLA alerting and operational monitoring.

Self-Service, Delivery and Leadership

- Shipped versioned REST APIs and Python automation services giving 6 enterprise applications self serve access to curated data.
- Partnered with product owners and business analysts to turn requirements into technical designs, delivering 25+ platform enhancements on schedule while leading a 4 engineer pod through mentoring and code reviews.
- Shortened build, test and documentation cycles 30% with GenAI assisted development (GitHub Copilot, LLM tooling) for PySpark code and unit tests.

SKILLS

Languages and SQL: SQL (data modeling, complex queries, stored procedures, performance tuning), Python, PySpark, Apache Spark, Spark SQL, JavaScript

AWS: Glue, Lambda, Step Functions, EventBridge, Amazon MSK, DMS, S3, Aurora PostgreSQL, Athena, IAM (roles and permissions), CloudWatch, CloudFormation, Comprehend, Textract

Azure and Databricks: Azure Databricks, Delta Lake, Delta Live Tables, Unity Catalog (access control, lineage, discovery), Databricks SQL, Databricks Workflows, Auto Loader, Azure Data Factory

Warehousing and Modeling: Dimensional modeling (star schema), warehouse schema design and evolution, Snowflake, Synapse, medallion architecture, object and document data stores

ETL, Streaming and Quality: ETL/ELT design, CDC and incremental loads, multi source integration, Kafka / Amazon MSK, Spark Structured Streaming, event driven architecture, REST APIs, validation and reconciliation frameworks, data contracts

Analytics and ML Enablement: BI and reporting enablement, curated datasets for analytics and ML consumers, ML based document and entity extraction (Comprehend, Textract), GenAI assisted development (GitHub Copilot, LLM tooling)

Orchestration and DevOps: Apache Airflow, GitHub Actions, Git, CloudFormation, Terraform, Docker, Agile, JIRA

CERTIFICATIONS AND EDUCATION

Certifications: AWS Certified Data Engineer – Associate • Microsoft Certified: Azure Database Administrator Associate (DP-300)

Education: B.Tech, Petroleum Engineering, Graphic Era University, Dehradun | 2018 – 2022