AKASH SAJJA

Senior Data Engineer
akashsajjawork@gmail.com
LinkedIn | United States
+1-928-514-8875

PROFESSIONAL SUMMARY

- Senior Data Engineer with 5+ years of experience designing and operating **production-grade data engineering systems** using **Python, SQL, Apache Spark, and cloud-native platforms**.
- Proven expertise in building scalable **ETL/ELT pipelines**, **feature-ready datasets**, and **data models** that support analytics and **AI-driven use cases**.
- Strong background in **healthcare data ecosystems**, including **claims, payments, and regulated datasets**, ensuring compliance with **HIPAA** and enterprise governance standards.
- Experienced in collaborating with **Product Managers, subject matter experts, and remote engineering teams** to rapidly prototype, iterate, and deliver high-value data solutions.

CERTIFICATIONS

AWS Certified : AWS Certified Developer Associate: Credential ID: c076ebe5f5eb475c9ae6a2df23656c8f - Link

TECHNICAL SKILLS

Languages: Python, SQL, Shell Scripting (Bash), R, Advanced SQL
Big Data: Apache Spark, Spark SQL, PySpark, Distributed Data Processing, TB-scale Data Pipelines, Kafka
Cloud – Azure: Azure Data Factory, Azure Data Lake Storage (ADLS Gen2), Azure Synapse Analytics, Azure SQL Database, Azure Monitor, Azure Key Vault
Cloud – AWS: Amazon S3, AWS Glue, AWS Lambda, Amazon EC2, AWS CloudWatch, IAM, AWS Step Functions
Data Warehousing: Azure Synapse, Amazon Redshift, Star/Snowflake Schema
Databases: SQL Server, MySQL, PostgreSQL, MongoDB, Redis
ETL & Integration: ETL/ELT Pipelines, Batch Processing, Incremental Loads, Schema Evolution, Workflow Orchestration, Apache Airflow, ADF Triggers
BI & Visualization: Power BI, Tableau, SSRS, Qlik, Looker, Grafana
DevOps: GitHub, Git, Jenkins, Docker, Kubernetes, Terraform, CI/CD pipelines
Streaming: Kafka, Spark Streaming, Event Hubs, Real-time Pipelines
Tools: JIRA, Confluence, Visual Studio, PyCharm, Postman, Swagger
Compliance: SOX, PCI-DSS, GDPR, HIPAA, Data Encryption, RBAC, Audit Logging

PROFESSIONAL EXPERIENCE

UnitedHealth Group — May 2025 – Present

Data Engineer (CPT)

- Designed, built, and maintained **production-grade ETL/ELT pipelines** using **Python, PySpark, and Spark SQL**, integrating **healthcare claims, payments, and member datasets** at 5–8 TB/day scale within a cloud data lake architecture.
- Developed **fault-tolerant, automated batch pipelines** using **workflow orchestration**, scheduling, and retries, achieving **99.5% pipeline availability**.
- Implemented **data quality, validation, reconciliation, and auditing frameworks**, reducing downstream data defects by **30–35%**.
- Optimized **distributed Spark workloads** using **partitioning, broadcast joins, and query tuning**, reducing processing latency by **25%**.
- Designed scalable **data models and feature-ready datasets** to support **analytics, reporting, and AI/ML downstream consumers**, enabling faster experimentation and reuse.
- Integrated **monitoring, logging, and alerting** to support **production data systems**, reducing incident resolution time by **40%**.
- Standardized reusable **ETL frameworks, configuration management, and CI/CD workflows**, reducing new pipeline development effort by **20%**.
- Collaborated cross-functionally with **Product Managers, subject matter experts, analysts, and engineers** to translate evolving business requirements into scalable **data products**.
- Built and maintained **data engineering foundations supporting AI and LLM-powered applications**, including scalable ingestion pipelines and curated feature datasets.
- Authored and maintained **technical documentation** covering pipeline architecture, data models, and operational workflows using **Confluence** and internal wikis.
- Prioritized pipeline enhancements based on **business value, stakeholder feedback, and cost efficiency**, iterating quickly to deliver measurable impact.

Johnson & Johnson Aug 2020 – Dec 2023

Data Engineer

- Enhanced large-scale financial data pipelines using Python, SQL, PySpark, and Apache Spark, processing 10+ TB/day of transactional and reference data.
- Designed and implemented **cloud-based data warehouse models** in **Amazon Redshift and Azure Synapse**, enabling analytics, reporting, and downstream **AI-enabled workloads**.
- Built scalable **ETL/ELT frameworks** using **Python, SQL, and Spark**, with incremental processing and schema evolution to support analytics and **data-driven decision systems**.
- Developed high-throughput ingestion pipelines using AWS Glue, S3, and Lambda, handling millions of records per day.
- Implemented CI/CD pipelines with Git, Jenkins, and Docker, improving deployment reliability and reducing release failures by 30%.
- Integrated monitoring, logging, and alerting to proactively detect pipeline failures and data quality issues.
- Supported production data systems through root cause analysis, impact assessment, and long-term remediation, reducing recurring incidents by 25%.
- Migrated legacy ETL workflows to cloud-native architectures, reducing infrastructure costs by 15–20%.
- Participated in data architecture and system design discussions, improving scalability, fault tolerance, and maintainability.
- Mentored junior engineers and reviewed code to ensure best practices, performance standards, and data reliability.
- Worked effectively in **remote, cross-functional teams**, coordinating with engineering, analytics, and product stakeholders across geographies.

Cyient Jul 2019 – Jul 2020

Software Engineer

- Developed backend and data processing components using **Python and SQL**, supporting ETL workflows processing **hundreds of thousands of records daily**.
- Built and enhanced **batch ETL jobs** for data extraction, cleansing, transformation, and loading, improving reporting availability by **15%**.
- Performed **query optimization, indexing, and performance tuning**, achieving **10–15% faster execution times**.
- Implemented **data validation and testing** using **Unit Test and PyTest**, reducing reprocessing and data inconsistencies by **20%**.
- Collaborated with senior engineers using **Agile practices and tools (JIRA, Confluence)** to improve pipeline stability and maintainability.

EDUCATION

Lewis University Jan 2024 – Dec 2025

Master of Science in Business Analytics