AKANKSH GATLA

akankshgatla@gmail.com • LinkedIn • GitHub • Portfolio

PROFESSIONAL EXPERIENCE

CAPITAL ONE — *Plano, TX*
Senior Data Engineer (Dec 2025 – Present)

- Engineered Standard Performance Reporting application integrating OneLake with AWS S3, orchestrating 20+ Glue Spark ETL jobs through Step Functions for automated daily processing.
- Implemented CI/CD pipelines with GitHub Actions, Jenkins, and Terraform, cutting deployment time by 40% using Blue-Green strategies for zero-downtime releases.
- Optimized Snowflake warehouse performance by 40% through query profiling, reengineering complex joins, and implementing materialized views for heavily accessed datasets.
- Enhanced query performance by 30% by architecting Snowflake Tables with compound indexing and micro-partitioning on 1.6M+ row datasets.
- Led cross-functional teams to build AWS Serverless Data Products (Lambda, Step Functions, S3, Glue), delivering MVP within 6-week sprints.
- Integrated Credit Bureau data feeds (Equifax, TransUnion, Experian) into 1600+ field schema using PySpark, applying SCD2/SCD1 for historical tracking.
- Strengthened security posture by remediating vulnerabilities, implementing RBAC IAM policies, and developing test frameworks achieving 95% coverage with pytest.

Technologies: AWS (S3, Glue, Step Functions, Lambda, Athena, Redshift), Snowflake, Databricks, PySpark, Terraform, Jenkins, GitHub Actions, pytest, behave

GENZEON CORPORATION — *Remote*
Data Integration Engineer (Aug 2025 – Nov 2025)

- Architected end-to-end ingestion pipelines processing HL7 feeds and EPIC FHIR databases into Azure Data Lake using Spark and Delta Lake for 350+ healthcare clients.
- Built intelligent automation framework in Databricks with PySpark and ML algorithms, reducing manual data mapping effort by 40% and improving data quality.
- Designed Medallion Architecture (Bronze/Silver/Gold) Databricks Jobs delivering curated views that improved downstream analytics performance by 30%.

Technologies: Azure Databricks, PySpark, Apache Spark, ADLS Gen2, Delta Lake, HL7, FHIR, Airflow, SQL, Python, Medallion Architecture

CAPITAL ONE — *Remote*
Data Engineer (Nov 2024 – Aug 2025)

- Engineered Standard Performance Reporting application integrating OneLake with AWS S3, orchestrating 20+ Glue Spark ETL jobs through Step Functions for automated daily processing.
- Implemented CI/CD pipelines with GitHub Actions, Jenkins, and Terraform, cutting deployment time by 40% using Blue-Green strategies for zero-downtime releases.
- Optimized Snowflake warehouse performance by 40% through query profiling, reengineering complex joins, and implementing materialized views for heavily accessed datasets.
- Enhanced query performance by 30% by architecting Snowflake Tables with compound indexing and micro-partitioning on 1.6M+ row datasets.
- Led cross-functional teams to build AWS Serverless Data Products (Lambda, Step Functions, S3, Glue), delivering MVP within 6-week sprints.
- Integrated Credit Bureau data feeds (Equifax, TransUnion, Experian) into 1600+ field schema using PySpark, applying SCD2/SCD1 for historical tracking.
- Strengthened security posture by remediating vulnerabilities, implementing RBAC IAM policies, and developing test frameworks achieving 95% coverage with pytest.

Technologies: AWS (S3, Glue, Step Functions, Lambda, Athena, Redshift), Snowflake, Databricks, PySpark, Terraform, Jenkins, GitHub Actions, pytest, behave

UNITY POPULATION HEALTH — *Remote*
Data Engineer (Aug 2023 – Nov 2024)

- Architected HIPAA-compliant ELT platform using Azure Data Factory and Databricks, automating 8-hourly EMR data extraction via RESTful APIs with OAuth 2.0.
- Implemented Medallion Architecture with PySpark and Delta Lake, applying SCD2 patterns and CDC for real-time clinical analytics and regulatory reporting.
- Built ML-powered automation system generating non-compliant patient cohorts for Value-Based Care, integrating with EMRs via HL7/FHIR APIs.
- Developed UDS reporting platform with Spark SQL and Power BI for HEDIS compliance, optimizing jobs through DataFrame caching and broadcast joins.

- Deployed applications with CI/CD pipelines via Jenkins, reducing deployment time by 30% while achieving 95% test coverage.

Technologies: Azure Data Factory, Databricks, PySpark, ADLS, Delta Lake, Power BI, DAX, Spark SQL, OAuth 2.0, HL7/FHIR, Jenkins, pytest

UNITY POPULATION HEALTH
Remote

Associate Data Engineer
(Sep 2020 – Jul 2022)

- Orchestrated complex ETL workflows using Airflow DAGs automating SQL Server processes with dynamic pipeline generation and SLA monitoring.
- Integrated OAuth 2.0 for EMR API access and conducted Risk Stratification Analysis using K-Means and Hierarchical clustering on clinical data.
- Developed PySpark analytics workflows in Databricks processing 3.8M patient records with partitioning and broadcast joins for predictive risk models.
- Architected AI-powered Patient Chatbot with NLP on Azure Bot Service, handling 5K+ monthly interactions with 92% satisfaction rate.
- Implemented feature engineering pipelines (Scaling, PCA, K-Means, DBSCAN) enhancing patient risk stratification accuracy by 35%.
- Automated Clustering Analysis for 350+ ICD-10 codes with Python and PySpark, creating Tableau dashboards for clinical decision support.
- Managed agile workflows using Jira and MS Teams for sprint planning and cross-functional collaboration.

Technologies: Apache Airflow, PySpark, Databricks, Azure Bot Service, Cognitive Services, SQL Server, Python, scikit-learn, NLP, Tableau, Jira

THE SPARK FOUNDATIONS
Remote

Junior Data Scientist
(Apr 2020 – Aug 2020)

- Optimized data pipelines with Airflow DAGs and PostgreSQL, accelerating retrieval by 40% and improving response time by 35% through indexing.
- Built Ensemble ML models (Random Forest, Gradient Boosting) achieving 84% accuracy with PCA and feature selection, visualized in Tableau dashboards.

Technologies: Apache Airflow, Python, PostgreSQL, scikit-learn, Random Forest, Gradient Boosting, PCA, Tableau, Pandas, NumPy

EDUCATION

UNIVERSITY AT BUFFALO
Buffalo, New York

Master of Science in Computer Science

Specialized in Big Data Analytics, Distributed Systems, and Machine Learning

LOVELY PROFESSIONAL UNIVERSITY
Punjab, India

Bachelor of Technology in Computer Science & Engineering

Specialized in Data Structures, Algorithms, Database Systems, and Cloud Computing

ACADEMIC PROJECTS

GLOBAL SUPER STORE

Automated ETL pipeline using AWS Lambda and Python to process Super Store API data, reducing manual data processing time by 60%, enabling real-time analytics and insights through Amazon S3, AWS Glue, and Athena for streaming data in the cloud.

HEALTH CARE APPLICATION

Engineered Healthcare prediction system with 89% F1 score with Random Forest Classifier, deployed on AWS EC2, via Putty featuring an interactive Flask based front end integrated with Tableau Dashboard's for patients.
Published: "Predictions And Analytics in Healthcare: Advancements in Machine Learning" IRJET Machine Learning for Healthcare Workshop (MLHC).

CERTIFICATIONS

Databricks Certified Data Engineer Associate | Microsoft Certified: Azure Data Engineer Associate