Akanksh Gatla

Exton, Pennsylvania | 716-808-2702 | akankshgatla@gmail.com | LinkedIn | GitHub | Portfolio

SUMMARY

Passionate Data Engineer with 5+ years of experience building scalable ETL/ELT pipelines and data products in fintech and healthcare. Expert in Spark, PySpark, AWS Glue, Azure Data Factory, Databricks and Snowflake, with a proven track record of optimizing transformations, reducing latency, and implementing robust CI/CD. Dedicated to applying expert knowledge on designing efficient data architectures and delivering analytics-ready datasets that drive business decisions.

TECHNICAL SKILLS

- **Languages and Core Skills:** Python 3.x, R, C/C++, SQL, T-SQL, Spark SQL, PySpark, HTML/CSS, Bash
- **Big Data Ecosystem:** Hadoop, HDFS, MapReduce, HBase, Apache Airflow, Apache Kafka, Apache Spark, Apache Flink, Apache NiFi, Apache Zookeeper
- **API & Cloud Services:** AWS (EC2, S3, DynamoDB, Lambda, Glue, Athena, AWS Pipeline, Redshift, Step Functions), Azure Data Factory, Data Lake, DevOps, IAM, Synapse, Data Lake, RBAC, BigQuery, Dataflow, Databricks
- **Data Visualization:** Tableau, Power BI, Excel, Google Looker Studio, NumPy, Pandas, Matplotlib, Seaborn
- **Version Control & Database:** GitHub, Git, PostgreSQL, SQL Server, MSSQL, MySQL, SQLite, Snowflake, MongoDB
- **Orchestration & CI/CD:** Apache Airflow, dbt, Jenkins , GitHub Actions, Terraform, Docker
- **API Development:** REST APIs, FastAPI, OAuth2.0
- **Data Modeling & Warehousing:** Star/Snowflake Schema, Dimensional Modeling, Data Vault, Fact and Dimension tables, Pivot Tables, Slowly Changing Dimensions, Change Data Capture (CDC), Partitioning, Clustering, Indexing, Materialized Views, Data Normalization/Denormalization, OLAP/OLTP, Delta Lake
- **Business Intelligence and Predictive Models:** Regression analysis, Decision Tree, Random Forest, Support Vector Machine, Neural Network, K-Means Clustering, KNN, Natural Language Processing, Principal Component Analysis (PCA)

WORK EXPERIENCE

Genzeon Corporation | *Data Integration Engineer* **Aug 2025 - Present**

- Engineered end-to-end ingestion pipelines for healthcare data from Flat files, HL7 feeds (via SFTP), and EPIC databases into Azure Data Lake (ADLS Gen2), enabling the processing for nearly 350+ healthcare clients.
- Automated Mapper file generation and transformation workflows in Azure Databricks, reducing manual data mapping effort by 40% and improving data accuracy across Coboodle and non-Coboodle clients.
- Designed and optimized Azure Databricks Jobs to clean, normalize, and translate clinical datasets, delivering Gold Layer views that improved downstream analytics performance by 30%.

Capital One | *Data Engineer* **Nov 2024 - Aug 2025**

- Developed a Standard Performance Reporting application that enhanced data processing efficiency by integrating OneLake sources with S3 buckets, organizing over 20 AWS Glue ETL scripts through Step Functions for daily execution.
- Automated CI/CD pipelines leveraging GitHub Actions, Jenkins pipelines to successfully deploy code base changes into QA and Prod environments AWS S3 buckets, cutting deployment time by 40% using Blue Green Deployment strategies for zero-downtime releases.
- Enhanced Snowflake view efficiency by 40% by conducting root cause analysis, data validation, and mapping complex upstream pipelines with heavily joined datasets using Databricks.
- Improved query performance by 30% and cut execution time from 1.5s to 1s by creating curated Snowflake Tables/Views using Compound Indexing and partitioning on over 1.6M rows of raw data, facilitating seamless analytics and reporting.
- Demonstrated cross-functional Leadership with Product and Analytics Teams to build Data Product, AWS Serverless Architecture workflows using AWS S3, Glue, and Step Functions, performed Cost Analysis to deliver MVP within 6 Weeks, and managed Jira Backlogs to ensure goal-driven delivery.
- Integrated external Credit Bureau data (Equifax, TransUnion, Experian) with Step Functions and Glue scripts to format it into a comprehensive 1600+ field structure using JSON schema and SQL, which improved data accuracy through SCD2 for history tracking and SCD1 for ongoing management.
- Corrected compliance and security vulnerabilities across repositories by updating package versions, optimizing S3 Bucket permissions, streamlining IAM policy management, and developing comprehensive test suites using pytest and behave, significantly enhancing system integrity and effectiveness.

Unity Population Health | *Data Engineer* **Aug 2023 - Nov 2024**

- Designed and implemented a HIPAA-compliant ELT system using Azure Data Factory and Databricks, automating 8-hourly data extraction from EMRs via RESTful APIs, and securely staging data in Azure Data Lake Storage, improving accessibility.
- Followed Medallion Architecture in Databricks Notebooks to apply SCD2, maintaining historical records with timestamps and partitions, orchestrated the output into silver and performed Change Data Capture with gold tables to support downstream analytics and reporting.
- Built an Automation system using Python to generate lists of non-compliant patients based on clinical, functional and operational checks defined by Value-Based Care, and display them on provider's Electronic Medical Records.
- Developed a Unified Data Systems reporting system using Spark SQL and Power BI DAX Query to create reports for healthcare clients focusing on clinical compliance handling long-running Spark Jobs using DataFrame, Views and memory tuning.
- Deployed applications on ADF with ADLS for storage, implementing CI/CD pipelines with Jenkins, reducing Deployment time by 30% while achieving 95% test coverage with pytest.

Unity Population Health | *Associate Data Engineer* **Sep 2020 - Jul 2022**

- Operated complex data workflows using Apache Airflow to automate ETL processes from SQL Server and perform Stored Procedure optimization for seamless data pipeline execution to automate repetitive tasks.
- Integrated OAuth 2.0 for accessing data from Electronic Medical Records (EMR) and conducted Risk Analysis of monitoring vital data using clustering algorithms to predict dominantly affected patient groups.
- Created code snippets in Databricks notebooks to handle 3.8 million rows with Spark SQL and Python to execute comprehensive predictive Risk analysis.
- Created a robust Patient Chatbot on Azure cloud for scalable and efficient deployment of the chatbot, enabling appointment booking, conducting mental health screenings and remote monitoring, serving 5K+ interactions/month.
- Performed Standard Scaling, PCA, Agglomerative and K-means clustering to identify patterns and trends in patient data based on Diagnosis, Lab, Immunizations and Screening tests enhance the understanding of patient risk profiles.
- Automated Clustering Analysis for 350+ ICD codes and draw conclusions with dashboard using Python script to visualize trends in vitals and lab tests, facilitating data-driven decision-making for healthcare providers.
- Managed project tasks and communication channels using Jira for task tracking and MS Teams for real-time collaboration, facilitating cross-functional teamwork and project delivery efficiency.

The Spark Foundations | *Junior Data Scientist* **Apr 2020 - Aug 2020**

- Improved data pipelines with Airflow DAGs, Python, and PostgreSQL, accelerating data retrieval by 40% and enhancing app response time by 35% through optimized database indexing and SQL script enhancements.
- Identified key variables using Feature Selection and PCA, and built an Ensemble Model with machine learning algorithms, achieving 84% accuracy with Random Forest Regressor. Created real-time dashboards in Tableau to visualize student KPIs, helping students identify areas for improvement

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 patient statices.
- "Predictions And Analytics in Healthcare: Advancements in Machine Learning" IRJET Machine Learning for Healthcare Workshop (MLHC).

EDUCATION

University at Buffalo, Buffalo, New York
Master Of Science – MS, Computer Science

Lovely Professional University, Punjab, India
Bachelor of Technology – BS, Computer Science & Engineering

CERTIFICATIONS

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