

# Samariddin Zarifov

[samariddinzarifov01@gmail.com](mailto:samariddinzarifov01@gmail.com) | [Raleigh, NC](#) | [linkedin.com/in/samariddin04](https://www.linkedin.com/in/samariddin04) | [github.com/Samariddin04/Projects](https://github.com/Samariddin04/Projects)

## Summary

Data Engineer with 5+ years building scalable data pipelines and analytics platforms on AWS, Azure, and GCP. Expert in PySpark, Databricks, and Snowflake, with hands-on experience designing streaming and batch pipelines, Delta Lake architectures, and CI/CD automation. Proven ability to optimize performance and cost, model data for BI/ML workloads, and deliver high-value outcomes across finance and enterprise analytics environments.

## TECHNICAL SKILLS

- **Big Data:** Databricks, Apache Spark (PySpark), Hadoop (MapReduce), Kafka, Delta Lake
- **ETL Tools:** ADF, Databricks, AWS Glue, Synapse Pipelines, DBT, Airflow, Fabric Spark Notebooks, SSIS, Alteryx
- **Cloud Platforms:** AWS (Glue, Redshift, S3, Lambda, Athena), Azure (Databricks, Data Factory, Synapse Analytics, ADLS), GCP (BigQuery, Dataflow)
- **Programming languages:** Python(Pandas, NumPy, Scikit-learn, Matplotlib, Seaborn), SQL, Scala, R, PowerShell
- **Databases & Warehousing:** Snowflake, Redshift, Synapse Analytics, PostgreSQL, MS SQL Server
- **BI & Analytics:** Power BI (DAX), Tableau, Looker, QuickSight
- **DevOps:** Azure DevOps, GitHub, Docker, Terraform
- **Certifications:** [AWS Solutions Architect Associate](#), [Professional Data Analyst](#), [Big Data with PySpark](#), [RegEx in Python](#), [Databricks Fundamentals](#), [Databricks Certified Data Engineer Associate \(In Progress\)](#)

## WORK EXPERIENCE

### Clearview CPA Advisors LLC

Senior Data Engineer

Pittsburgh, PA - Remote

07/2024–11/2025

- Lead 5+ concurrent data engineering projects including data lake modernization and real-time analytics, managing stakeholder engagement with C-level executives and ensuring 100% on-time delivery
- Architect 2 end-to-end data lake solutions using Databricks on Azure, integrating 10+ sources (ADLS, Snowflake, SQL Server) enabling analytics for 100+ users across 4 business units
- Build batch ETL pipelines in Databricks processing multi-source financial datasets into Delta Lake, automating daily and monthly data consolidation for 100+ users with 98% reliability and optimized query performance.
- Engineer scalable Spark (PySpark) workflows processing terabyte-scale data, implementing optimizations that improved query performance by 70% and reduced costs by 35%
- Design Snowflake data pipelines with automated orchestration, reducing storage costs by 40% and improving query response times by 65%
- Develop production DBT models for Snowflake and BigQuery, implementing data lineage and validation improving compliance by 85%
- Implement CI/CD pipelines using Azure DevOps and Terraform automating Databricks, ADF, and Synapse deployments, reducing deployment time by 75%
- Built AWS Glue ETL pipelines processing multi-year financial datasets stored in S3 (terabyte-scale) into Redshift, with Python-based validation scripts ensuring 95%+ reconciliation accuracy
- Provide technical leadership to 3 engineers through code reviews and best practices, reducing incidents by 60%

### William & Mary

Data Engineer

Williamsburg, VA

09/2023–06/2024

- Engineer end-to-end data lake using Databricks on AWS integrating 12+ sources, developing PySpark pipelines processing 5TB+ data for analytics
- Lead Flight Delay Analysis project using Databricks Spark analyzing 10M+ records, implementing machine learning model achieving 80%+ accuracy
- Build automated ETL pipelines using SQL, Python, R, and Alteryx reducing manual effort by 60% for semester reporting
- Implemented Snowflake Tasks for scheduled incremental loads and applied micro-partitioning and clustering keys to large Snowflake tables, reducing long-running scan time by 40%
- Create Tableau and Power BI dashboards with DAX for academic KPIs serving 25+ stakeholders
- Mentor 40+ students in data engineering and Spark optimization, improving comprehension by 25%
- Develop Gradient Boosting models for Virginia workforce analytics using AWS S3 and Python

- Partner with stakeholders to deliver real-time and batch pipelines processing 2TB+ data annually for telecom and retail analytics
- Architect enterprise data lake using Hadoop (MapReduce) and Spark integrating 7+ sources for 100+ users
- Optimize AWS Redshift warehouse implementing distribution keys and performance tuning, improving query speed by 80%
- Automate ETL using AWS Glue, Lambda, and Python reducing manual work by 90% and achieving 99% reliability
- Build Kafka streaming workflows processing 10K+ daily events into databases with <5 second latency
- Design Azure Data Factory pipelines with error handling improving efficiency by 55%
- Create CI/CD framework using GitHub and Terraform automating Glue and Redshift deployments, accelerating releases by 65%
- Develop 15+ Power BI dashboards with advanced DAX for operational metrics serving 100+ users
- Manage Databricks clusters with auto-scaling reducing compute costs by 35%

## EDUCATION

---

### College of William & Mary

**Williamsburg, VA**

*Master of Science, Business Analytics GPA 3.7*

**Key courses:** Database Management, Big Data, Machine Learning, Stochastic Modeling, Optimization, AI

**Awards & Scholarships:** MSBA E&G scholarship, “People’s Choice Award” in Virginia Datathon 2024

### Tashkent Institute of Finance

**Tashkent, Uzbekistan**

*Bachelor of Science, Economics and Finance GPA 3.9*

**Key courses:** Financial Analysis, Statistics, Econometrics, Accounting, Investment, Marketing Analytics

**Awards & Scholarships:** 1 of 3 full state scholarships at university for scoring 100% results in entrance exams

## PROJECTS

---

**Flight Delay Analysis and Prediction:** Built a distributed pipeline to analyze 10 years of flight data using PySpark on Databricks. Visualized insights with Pandas/Matplotlib and deployed H2O-based Random Forest model to predict delays with 80%+ accuracy. Normalized delay metrics by flight volume and identified high-risk routes for performance improvement.

**Boosting the Workforce in Virginia:** Engineered automated ETL + modeling pipeline using Gradient Boosting, identifying key factors impacting labor force participation. Provided policy recommendations to improve workforce engagement, influencing state-level decision-making.

**Automating Bank Reconciliations:** Achieved over 96% accuracy in automating financial transaction matching with machine learning models, significantly reducing manual effort and errors in the banking industry.