BHUVAN S

Senior Data Engineer

Denver, CO | bhuvansarakam@gmail.com | +1 (678) 989-6307

PROFESSIONAL SUMMARY

Results-driven Data Engineer with 10+ years of experience designing and delivering end-to-end data pipelines, cloud data platforms, and AI/ML solutions across Banking, Finance, and Insurance domains. Deep expertise building scalable ETL/ELT workflows on Azure Data Factory, Databricks, and Snowflake; engineering Python and SQL-based transformation layers; and deploying machine learning models that generate measurable business outcomes. Proven ability to architect modern data lakehouse and medallion-architecture pipelines, integrate multi-source financial data, and operationalize predictive analytics at enterprise scale.

TECHNICAL SKILLS

Cloud & Data Platforms:

Azure Data Factory (ADF) • Azure Databricks • Snowflake • Azure Synapse Analytics • ADLS Gen2 • Azure Blob Storage • Azure Event Hubs • AWS (S3, EC2, RDS, Redshift, EMR)

Languages & Query:

Python (PySpark, Pandas, NumPy, scikit-learn) • SQL (T-SQL, Snowflake SQL) • Spark SQL • DAX

Data Engineering & Orchestration:

ADF Pipelines & Linked Services • Databricks Delta Lake • Apache Spark • Apache Kafka • Apache Airflow • SSIS/SSRS • Azure DevOps CI/CD • Autosys • Medallion Architecture (Bronze/Silver/Gold)

AI / ML:

Generative AI & LLM Integration • Predictive Modeling • Feature Engineering • MLOps • Random Forest, Gradient Boosting, SVM, ANN, Decision Trees • NLP • Credit Risk Modeling (CECL, PD, EAD) • Fraud Detection

Visualization & Reporting:

Power BI • DAX Measures • Streaming Dashboards • Geospatial Analytics

PROFESSIONAL EXPERIENCE

Senior Data Engineer | **Janus Henderson Investors** – Denver, CO *Sept 2024 – Present*

- Designed and deployed an end-to-end ADF-to-Databricks-to-Snowflake medallion pipeline ingesting $73B AUM institutional data across Bronze, Silver, and Gold layers — enabling daily refresh of opportunity scoring datasets for 500+ funds and share classes.
- Built PySpark feature engineering pipelines in Databricks computing rolling 1/3/5-year performance metrics, Sharpe ratios, and win-rate signals across the institutional AUM pipeline; outputs loaded into Snowflake Gold tables consumed by ML scoring models that drove 15% higher win rates.
- Engineered real-time reconciliation pipelines using Azure Event Hubs and Databricks Structured Streaming to cross-validate MSCI BarraOne, FactSet, Aladdin, and Atlas data feeds — improving portfolio attribution accuracy by 30% and reducing manual reconciliation effort by 60%.
- Architected ADF pipelines with parameterized linked services to ingest vendor data from MSCI, FactSet, Aladdin, and Atlas into ADLS Gen2 staging, transforming via Databricks notebooks into Snowflake distribution tables serving 8 downstream applications — eliminating 30% data lag.
- Built a Python-based ML scoring framework on Databricks (scikit-learn + MLflow) for institutional opportunity identification; trained ensemble gradient boosting models on Snowflake feature tables and registered models in MLflow for scheduled batch inference pipelines via ADF.
- Optimized Snowflake data warehouse performance through clustering keys, query result caching, materialized views, and warehouse auto-scaling — reducing average query execution time by 40% and supporting 99.8% SLA compliance on critical data feeds.
- Developed Power BI dashboards connected to Snowflake Gold layer, incorporating AI-assisted opportunity signals and DAX KPIs — increasing sales team engagement 25% YoY and cutting data-to-insight latency by 25%.
- Migrated on-premises SQL Server ETL workflows to Azure-native ADF + Databricks + Snowflake stack, decommissioning legacy infrastructure and reducing hosting costs 20% while achieving 99.5% uptime.
- Maintained and enhanced Nasdaq API ingestion workflows in Databricks, automating daily pulls into Azure Blob Storage and enforcing data quality checks via Great Expectations — sustaining 99% data freshness for institutional analysis.

- Authored reusable Databricks notebook libraries for common transformation patterns (SCD Type 2, deduplication, schema drift handling) adopted across 3 data engineering teams, standardizing pipeline development and cutting new pipeline build time by 35%.
- Recognized with Shout Out at Quarterly Town Hall for architecting AI/ML-powered data pipelines that strengthened the firm's global data-driven analytics culture.

Cloud Data Engineer (Contract) | **Janus Henderson Investors** – Denver, CO *May 2023 – Sept 2024*

- Built the JH Institutional Opportunity Signal data product end-to-end: ADF ingestion from Bloomberg and internal APIs into ADLS Gen2, PySpark transformations in Databricks, Snowflake Gold layer storage, and ML scoring pipeline analyzing $73B AUM across 1, 3, and 5-year performance windows.
- Engineered Broad Market Scoring pipeline using ensemble ML models in Databricks, expanding opportunity identification coverage by 35% across global investable universes; pipeline ran nightly via ADF trigger with results surfaced in Snowflake for Power BI consumption.
- Developed Asset Allocation analytics pipeline in Databricks aggregating fund-level AUM by asset class from Snowflake, applying Python-based misallocation detection logic, and loading flagged records into a Snowflake reporting schema — improving portfolio optimization decisions by 20%.
- Designed Power BI real-time alert subscription reports on Snowflake opportunity data using DirectQuery and scheduled refresh, cutting sales team response time by 25% and improving CRM conversion rates.

Data Engineer | **EXL Services** – Bangalore, India *Jun 2020 – Aug 2021*

- Built ADF pipelines ingesting transactional data from on-premises Oracle sources into Azure SQL via Linked Services and Integration Runtimes; developed Python (Pandas/PySpark) transformation scripts to detect unearned cash discount patterns, recovering 15% revenue leakage for the Electronics domain.
- Designed moving-average and geospatial analytical models in Python on Databricks, publishing results to Power BI dashboards — improving domain stakeholder decision speed by 25%.
- Automated Azure SQL-to-Power BI data refresh using ADF scheduled triggers and REST API connectors, accelerating dataset refresh cycles from daily to hourly (80% faster).
- Migrated Oracle SQL batch workloads to SSMS/SSRS running on AWS RDS, using SSIS packages to extract, transform, and load data into Redshift; reduced ETL processing time 35% and enabled 40% faster digital transformation reporting.
- Constructed CECL credit risk pipelines using ADF and Azure Event Hubs to stream loan performance data into Databricks; built survival-based PD and EAD models in Python (scikit-learn/statsmodels), achieving sub-10% quarterly error rate on loss estimates.
- Developed segment-level historical loss rate models using Python mean-reversion techniques on Snowflake data; led a 3-analyst team to deliver production-ready models 2 weeks ahead of schedule.
- Built Decision Tree ML models in Python (scikit-learn) trained on Snowflake credit feature tables to optimize Credit Line Decrease policy — reducing loss rates by 20bps and improving approval rates from 90% to 51%.
- Engineered predictive regression models in Databricks to forecast CLD volume, EBIT, and RAR impacts, loading scenario outputs into Snowflake for executive reporting and enabling multi-million dollar policy decisions.

Data Engineer – Fraud Detection | **American Express** – Haryana, India *Jan 2017 – Jan 2018*

- Built Spark/Kafka/Airflow real-time fraud detection pipeline on AWS (EMR + S3) processing millions of daily card transactions; Python-based KModes clustering and stochastic gradient descent models reduced post-week fraud detection lag by 40% and improved detection accuracy by 18%.
- Engineered a SQL-based disposable email pattern recognition engine querying transaction metadata in Redshift, blocking 25% of synthetic account applications before card issuance.
- Deployed 5 ML classification models (Logistic Regression, Random Forest, SVM, ANN, Decision Trees) trained on AWS EMR/PySpark to predict seismic risk in coal mine data, achieving 92% accuracy vs. 78% baseline.
- Built AWS EC2/S3/Step Functions ETL automation pipelines orchestrating data extraction, Python transformations, and Redshift loading — automating 85% of manual data workflows end-to-end.
- Won American Express Data Science Championship for authoring data handbooks on credit card risk characteristics and systematic risky transaction detection methodologies.

EDUCATION

M.Tech & B.Tech in Mechanical Engineering (Intelligent Manufacturing) *2011 – 2016*
Indian Institute of Technology, Madras

CERTIFICATIONS & AWARDS

- Certified Data Scientist – International School of Engineering (INSOFE), 2016 – ML Algorithms, Statistics, Big Data Computing
- Big Data Foundations Level 1 – IBM
- American Express Data Science Championship Winner – Risk characteristics & fraud detection data handbooks (Nov 2017)
- Janus Henderson Quarterly Town Hall Shout Out – AI/ML data pipeline innovation strengthening global analytics culture