Manisha Yadav

Data Engineer

manishayv07@gmail.com | +1 (347) 580-8709

linkedin.com/in/manisha-yadav-31a526186

PROFESSIONAL SUMMARY:

Results-driven Data Engineer with 5+ years of experience building scalable, cloud-native data pipelines, real-time streaming solutions, and machine learning systems on AWS and Azure. Strong expertise in PySpark, SQL, and Python for large-scale ETL/ELT workflows, with hands-on experience in streaming technologies (Kafka, Spark Structured Streaming, AWS Kinesis) and modern data stack tools including dbt, Delta Lake, and Databricks. Experienced in developing and deploying ML models (Scikit-Learn, TensorFlow, PyTorch) for fraud detection and risk analytics, with production experience in AWS SageMaker. Domain expertise spanning financial services and healthcare (payers, providers, clinical data standards including FHIR, HL7, X12 837). Proven ability to design data lakes, Lakehouse architectures, and governed data warehouses (Snowflake, Redshift) while delivering end-to-end data and AI solutions in Agile environments that drive business-critical decisions.

PROFESSIONAL EXPERIENCE:

Data Engineer | Citigroup | New York, NY *May 2024 – Present*

- Designed and maintained scalable ETL/ELT data pipelines using PySpark, SQL, and AWS services (S3, Glue, EMR, Lambda) to process high-volume financial transaction data for analytics and reporting.
- Built batch and near-real-time data processing workflows using AWS Glue, EMR, and Kinesis, improving data availability and supporting fraud detection and risk analytics use cases.
- Optimized PySpark jobs through partitioning, caching, and parallel processing techniques, reducing pipeline runtime and improving performance for multi-terabyte datasets.
- Developed and managed data lake and warehousing solutions using Amazon Redshift and S3, implementing star and snowflake schemas for enterprise reporting and regulatory compliance.
- Automated workflow orchestration and scheduling for enterprise data pipelines, ensuring reliable ingestion, transformation, and delivery of critical business data.
- Collaborated with data scientists and analytics teams to build and maintain feature engineering pipelines for machine learning models used in customer risk scoring and fraud analytics.
- Integrated machine learning inference workflows into AWS-based data pipelines to support near real-time decision-making and operational reporting.
- Performed large-scale data analysis using Python, Pandas, and NumPy to identify data quality issues, analyze transaction patterns, and support business stakeholders with actionable insights.
- Implemented data validation, monitoring, and governance processes to ensure data accuracy, lineage, security, and compliance with financial regulatory standards.
- Created interactive dashboards and reporting solutions using Power BI and Tableau, enabling stakeholders and leadership teams to monitor KPIs and operational metrics.
- Worked closely with cross-functional teams, including business analysts, risk teams, and engineering teams in an Agile environment to deliver scalable cloud-based data solutions.

Environment: AWS (S3, Glue, EMR, Lambda, Redshift, Kinesis), PySpark, SQL, Python, Pandas, NumPy, Power BI, Tableau, ETL/ELT, Data Pipelines, Data Lakes, Data Warehousing, Machine Learning Integration, Agile, Git.

Data Engineer | Fuse machines | New York, NY *Sep 2022 - Dec 2023*

- Designed and implemented scalable Lakehouse architecture on Azure using ADLS Gen2 and Databricks, leveraging Medallion (Bronze/Silver/Gold) framework for enterprise data processing.
- Developed ETL/ELT pipelines using Azure Data Factory and PySpark in Databricks for ingestion and transformation of large-scale structured and semi-structured datasets.
- Optimized Spark-based data processing workflows using partitioning and query tuning techniques, improving data processing performance and reducing latency.
- Implemented Delta Lake capabilities including ACID transactions, schema evolution, and time travel to ensure reliable and consistent data pipelines.
- Migrated legacy data warehouse systems to Snowflake, improving scalability, query performance, and reducing maintenance overhead.
- Designed dimensional data models (star and snowflake schemas) to support enterprise reporting and analytical workloads.
- Integrated multiple data sources including REST APIs and enterprise applications, ensuring data consistency and high-quality ingestion.
- Developed and maintained interactive dashboards using Power BI and SSRS for business stakeholders and leadership reporting.

- Implemented role-based access control (RBAC) and Azure IAM policies to enforce secure and compliant data access.
- Collaborated with cross-functional teams in an Agile environment to deliver scalable data solutions supporting analytics and downstream AI/ML workloads.

Environment: Azure Data Factory, Azure Databricks, ADLS Gen2, Delta Lake, Snowflake, PySpark, SQL, Python, Power BI, SSRS, ETL/ELT, Data Lakes, Lakehouse Architecture, Medallion Architecture, Data Modeling, Azure IAM (RBAC)

Data Engineer | Cedar Gate Technologies| Kathmandu, Nepal *Jan 2020 – Aug 2022*

- Designed and developed scalable ETL pipelines using Python and AWS services such as Glue and Lambda to ingest and process healthcare datasets, including claims, clinical, and provider data from multiple sources.
- Built and maintained a centralized data warehouse using Amazon Redshift to support payer-provider analytics, enabling reporting, risk stratification, and value-based care insights.
- Implemented a data lake architecture on Amazon S3 to store structured and semi-structured data (CSV, JSON, XML), and enabled efficient querying using Athena and Redshift Spectrum.
- Automated schema discovery and metadata management using AWS Glue Data Catalog and Crawlers across numerous data sources.
- Optimized Redshift performance by configuring sort and distribution keys, leveraging materialized views, and utilizing Spectrum for large-scale query processing.
- Developed AWS Lambda functions for event-driven processing, workflow orchestration, and pipeline monitoring to enhance reliability and observability.
- Collaborated with analytics and BI teams to deliver curated datasets and data marts for reporting and dashboards using QuickSight.
- Created reusable SQL scripts and stored procedures for data transformation, cleansing, and deduplication of patient and provider records.
- Implemented secure data access controls using AWS IAM roles, encryption (KMS), and role-based permissions to ensure compliance with healthcare data standards.
- Orchestrated data workflows using Apache Airflow, incorporating scheduling, monitoring, and error handling mechanisms.
- Supported CI/CD processes using Git, Jenkins, and Terraform to enable automated deployment of data pipelines and infrastructure.
- Improved data processing efficiency and reduced latency by implementing incremental data loading strategies and optimizing ingestion workflows.

Environment: AWS (S3, Redshift, Glue, Lambda, Athena), Python, SQL, Airflow, Jenkins, Terraform, Git, QuickSight, JSON, CSV, XML, Healthcare Data (Claims & EHR).

TECHNICAL SKILLS:

- **Programming Languages:** Python, SQL, Scala, Unix/Shell Scripting
- **Data Engineering & Big Data:** PySpark, Apache Spark, Hadoop, Hive, HDFS, Databricks, Delta Lake, Snowflake, dbt, ETL/ELT Pipelines, Data Validation
- **Cloud Platforms:** AWS (S3, Glue, Lambda, Athena, EMR, EC2, Redshift, Kinesis), Azure (Azure Data Factory, ADLS Gen2, Azure SQL, Event Hubs), GCP (Big Query, Looker Studio)
- **Databases & Data Warehousing:** Oracle, MySQL, MongoDB, Hive, Amazon Redshift, Snowflake, Azure SQL, Big Query
- **Orchestration & DevOps:** Apache Airflow, Git, CI/CD, Docker, Jenkins, Terraform, Agile
- **Streaming & Messaging:** Apache Kafka, Spark Structured Streaming, AWS Kinesis
- **Data Governance & Security:** Unity Catalog, AWS Glue Data Catalog, Delta Lake, Data Lineage, RBAC, HIPAA Compliance
- **Healthcare Standards:** FHIR, HL7, X12 837 (Medical Claims), Payer/Provider/Clearinghouse Data
- **Visualization & Analytics:** Power BI, Tableau, Looker Studio, Alteryx, Pandas, NumPy
- **Machine Learning & AI:** Scikit-learn, TensorFlow, PyTorch, Feature Engineering, Model Deployment

EDUCATION:

- Master's in Artificial Intelligence | The Katz School of Science and Health at Yeshiva University| Manhattan, NY
- Bachelor of science in Computer science | Leeds Beckett University | Leeds, UK