Vikas Pabba

Email: pabba.vikas54@gmail.com
Mobile: +12108737446

Data Engineer

PROFESSIONAL SUMMARY:

- Data Engineer with 4+ years of experience built scalable ETL pipelines using AWS Glue, PySpark, and Airflow to process 1B+ weekly records, boosting SLA compliance by 21% and reducing latency 35%.
- Designed cost-efficient AWS S3 data lakes with Glue and Athena, reducing storage expenses by 40% and improving analytics query times through schema optimization and partitioning strategies.
- Automated streaming and batch ingestion from 50+ sources via Kafka, Spark, and Lambda, improving data availability timelines by 60% and supporting real-time analytics for enterprise platforms.
- Migrated 20+ legacy ETL jobs to AWS Glue and PySpark on Databricks, reducing manual processing by 90% and cutting compute costs by 55% with reusable code patterns.
- Developed advanced Redshift and PostgreSQL data models enabling 3x faster report generation, improved KPI accuracy, and optimized executive dashboards through pre-aggregated metrics and dimensional modeling.
- Tuned Spark jobs with caching, partitioning, and memory management, boosting large join performance by 3x and increasing overall pipeline throughput by 2.2x across multiple enterprise data flows.
- Integrated Salesforce, Oracle, and flat files into Snowflake via Azure Synapse pipelines, maintaining 99.9% data integrity and enabling unified analytics for business reporting and cross-functional teams.
- Designed schema evolution logic with AWS Glue DynamicFrames and implemented watermarking-based incremental loads, eliminating data duplication and supporting zero-downtime ingestion across dynamic data models.
- Built parameterized pipelines in Azure Data Factory, reducing execution time by 40% and achieving 80% code reuse across batch ingestion workflows for finance and marketing analytics.
- Enabled CI/CD automation for ADF and Databricks using Azure DevOps, reducing deployment errors by 70% and standardizing releases with version control, pipeline triggers, and template validation.
- Integrated CloudWatch and custom logging into Airflow DAGs and Glue jobs, enhancing observability, enabling root-cause analysis, and reducing job resolution time from hours to under 15 minutes.
- Created Power BI dashboards from Azure SQL and Synapse, improving marketing time-to-insight by 80% and enabling real-time tracking of lead conversion and campaign performance KPIs.

TECHNICAL SKILLS:

- **Cloud Platforms -** AWS (Glue, Lambda, Redshift, Athena, EMR, S3, CloudWatch), Azure (Data Factory, Synapse, Databricks, Data Lake Gen2, Key Vault, DevOps)
- **Data Engineering & ETL -** PySpark, SQL, Delta Lake, Airflow, AWS Glue, Azure Data Factory, SSIS, Kafka, Azure Functions, Redshift Spectrum
- **Programming & Scripting -** Python, PySpark, SQL, Bash, Shell Script
- **Data Warehousing & Databases -** AWS Redshift, Azure Synapse, PostgreSQL, SQL Server, MySQL, Oracle, Snowflake, MongoDB
- **Analytics & BI Tools -** Power BI, Tableau, AWS Athena, Excel, Redshift SQL Workbench
- **Automation & DevOps -** GitHub Actions, Azure DevOps, Terraform, CI/CD pipelines, YAML, Jenkins
- **Monitoring & Governance -** CloudWatch, Azure Monitor, Log Analytics, IAM, RBAC, Data Quality Validation, Schema Versioning, Confluence
- **Big Data & Processing Frameworks -** Apache Spark, Delta Lake, AWS EMR, Databricks
- **Other Tools & Concepts -** Data Modeling, Star Schemas, Z-Order Clustering, Versioned Datasets, Partitioning, A/B Testing, KPI Design, Metadata Management

PROFESSIONAL EXPERIENCE:

Cardinal Health **Feb 2025 – Present**

Data Engineer

Responsibilities:

- Designed scalable ETL pipelines in AWS Glue using PySpark to process 5TB daily, reducing latency by 40% and improving SLA compliance through ingestion automation from 25+ external vendor sources.
- Migrated legacy ETL to AWS Glue and Lambda, reducing operational overhead by 70% and enabling serverless data workflows with versioned S3 datasets for rollback, audit trails, and historical lineage tracking.
- Built cost-effective data lake on S3 integrated with Glue Catalog and Athena, improving SQL query turnaround by 60% and supporting real-time business reporting across analytics teams.
- Implemented CI/CD pipelines for Glue with GitHub Actions and Terraform, decreasing deployment time by 78% while enforcing release governance, template reuse, and parameterized configuration standards.
- Developed modular PySpark jobs with reusable Glue triggers, increasing reusability by 80% across ingestion domains and streamlining ML dataset delivery pipelines for downstream data scientists.
- Tuned Spark with broadcast joins, partition pruning, and memory caching, improving job performance by 3.5x and reducing costs through efficient compute and lifecycle management on S3 staging layers.
- Integrated Kafka with Lambda and DynamoDB Streams for operational alerting, reducing notification latency by 50% and improving incident response efficiency across critical health monitoring systems.
- Secured pipelines with IAM policies and KMS encryption in transit and at rest, ensuring 100% compliance with HIPAA, SOX, and internal enterprise data security standards.
- Built schema validation logic to block invalid records upstream, reducing downstream data quality issues by 90% and improving reliability for analytics and forecasting pipelines.
- Delivered Redshift optimization by designing partition strategies that improved BI dashboard performance from 40 seconds to under 10 seconds for high-frequency business-critical queries.
- Implemented alerting with CloudWatch and SNS across batch jobs, reducing mean-time-to-recovery from 2 hours to 15 minutes by enabling proactive remediation workflows.
- Mentored junior engineers in PySpark, Git, and AWS data tooling, accelerating onboarding and improving project delivery velocity across cross-functional data engineering pods.

PNC Financial **Nov 2023 – Jan 2025**

Data Engineer

Responsibilities:

- Designed scalable ingestion pipelines in Azure Data Factory to integrate 40+ data sources into Synapse, improving end-to-end processing reliability and cross-domain accessibility for structured and semi-structured data.
- Migrated legacy SSIS processes to Data Factory, reducing maintenance by 60%, increasing deployment consistency, and enabling modular pipeline design through parameterized datasets, lookups, and dynamic control flows.
- Developed PySpark-based data transformations in Databricks to cleanse financial datasets, improving reporting accuracy by 90% and validating data against defined contracts from business analysts.
- Applied Delta Lake optimizations like Z-order clustering and merge strategies, improving high-volume transaction query performance by 3x and reducing pipeline lag in analytics workloads.
- Designed star schemas and materialized views in Synapse, reducing query runtimes by 70% and supporting rapid BI reporting through pre-aggregated datasets and result-set caching.
- Automated regression testing with PySpark suites to validate data transformations during monthly releases, cutting QA effort by 50% and maintaining consistency across production pipelines.
- Built custom Azure Functions for real-time pre-validation tasks, improving upstream data quality while enhancing compliance via credential management with Azure Key Vault.
- Configured Monitor and Log Analytics to trace pipeline failures, SLA breaches, and transformation anomalies, improving response time and enabling full observability across ADF, Databricks, and Synapse workflows.

- Created interactive Power BI dashboards using Azure SQL and Synapse data, shifting executive reporting from weekly to daily and increasing real-time business insights for leadership teams.
- Implemented RBAC and resource locks in ADF and Azure Storage, ensured security audit compliance, and maintained architectural documentation in Confluence to support onboarding and knowledge retention.

eBay **Jun 2021 – Jul 2023**

Data Analyst

Responsibilities:

- Built Power BI and Tableau dashboards using Redshift models to visualize sales, conversions, and behavior trends, driving a 22% ROI increase and enabling 95% KPI tracking accuracy.
- Created automated ETL workflows with SQL-based data marts and Python in Lambda, reducing report generation time by 75% and saving 30 staff-hours monthly through job scheduling in Athena.
- Analyzed 10+ TB of transactional data in S3 and Redshift to identify upsell and cross-sell opportunities, enabling data-driven marketing strategies and improving campaign targeting precision.
- Performed exploratory analysis with Pandas and NumPy on user logs, uncovering engagement patterns and driving a 15% improvement in user retention through product team collaborations.
- Conducted A/B testing using Python statistical libraries to assess UI changes, which resulted in a 12% checkout completion uplift and informed future UX decisions with data-driven evidence.
- Implemented data profiling, audit checks, and anomaly detection in PySpark and SQL to ensure reporting pipeline integrity and reduce discrepancies in downstream metrics by 40%.
- Supported real-time ad hoc analysis in Athena and automated monthly Power BI executive reports, cutting manual effort by 90% and reducing turnaround time from 3 days to 2 hours.
- Collaborated with engineering teams to migrate workflows to AWS Glue and Redshift Spectrum, improving analytical scalability, reducing infrastructure costs, and aligning with enterprise-wide data modernization goals.

Certifications:

- Big Data Engineering Bootcamp with GCP and Azure Cloud

Educational Details:

Master of Science – Computer and Information Sciences - Texas A&M University – Kingsville

Bachelor of Technology – Electronics and Communication Engineering - Jawaharlal Nehru Technological University Hyderabad (JNTUH)