

ADEEBA | DATABRICKS DATA ENGINEER

872-325-5926 | begumadeeba20@gmail.com | [linkedin.com/in/adeeba-begum](https://www.linkedin.com/in/adeeba-begum)

Databricks Data Engineer with 5 years of hands-on experience across full-time roles, internships, and enterprise-scale projects. I am skilled in architecting and optimizing cloud-native data pipelines to handle high-volume, complex datasets. Proficient in building batch and streaming ETL workflows using Databricks, Apache Spark (PySpark/Scala), and AWS Glue, with advanced expertise in Delta Lake, Apache Airflow, and Terraform. Experienced in developing Java-based microservices and data processing components for low-latency and high-throughput analytics. Strong background in performance tuning, cost optimization, and real-time analytics enablement. Adept at implementing data governance and access policies with Unity Catalog, ensuring compliance and security for sensitive datasets. Collaborative team player with proven success in CI/CD automation, cross-functional coordination, and delivering reliable, high-quality data solutions in production environments.

WORK EXPERIENCE

Frontier Communications | Aws-Databricks Data Engineer, USA

July 2025 – Till Date

- Integrated streaming ingestion pipelines using Apache Kafka and AWS Kinesis and Databricks Structured Streaming, enabling near real-time data availability for reporting and analytics.
- Implemented data quality monitoring with Delta Live Tables (DLT), including built-in lineage tracking, schema enforcement, and automated alerts, reducing SLA breaches by 20%. Assisted in infrastructure automation with Terraform and CI/CD pipelines, reducing manual deployment efforts and improving delivery speed.
- Conducted performance benchmarking and workload optimization, recommending cluster sizing and caching strategies that reduced compute costs by 15%.
- Created internal knowledge-sharing documentation on best practices for Delta Lake optimization, Unity Catalog access policies, and Spark troubleshooting, improving team efficiency.
- Collaborated with InfoSec teams to implement fine-grained data access controls, encryption policies, and compliance measures for sensitive healthcare data.

Johnson & Johnson | Databricks Data Engineer Intern, USA

Dec 2024 – May 2025

- Developed and optimized PySpark- and Scala-based ETL pipelines on Databricks for ingestion of high-volume financial transaction datasets, improving processing speed by 18%.
- Built and deployed Java-based microservices to handle real-time financial data validation and enrichment, integrating with downstream analytics systems.
- Created streaming workflows using AWS Kinesis, Kafka, and Spark Structured Streaming (Scala API) for market data feeds, enabling sub-second latency analytics for trading dashboards.
- Implemented Delta Lake best practices, including Z-ordering, partition pruning, and schema evolution, ensuring data accuracy and query efficiency.
- Assisted in creating Terraform scripts for provisioning AWS infrastructure, streamlining environment setup for development and testing teams.
- Participated in data security audits, applying Unity Catalog policies and encryption standards to meet regulatory compliance requirements (e.g., SOX).

- Implemented Delta Live Tables (DLT) pipelines with built-in lineage tracking, schema evolution, and audit logging, improving transparency and reducing data recovery time by 25%.
- Enhanced data quality frameworks by adding schema drift detection, automated null handling, and duplicate resolution, leading to a 30% drop in downstream data errors.
- Collaborated with InfoSec teams to implement fine-grained access controls using Unity Catalog and encryption standards, ensuring compliance with enterprise security policies.
- Performed workload benchmarking and Spark job optimization, leveraging Z-ordering, partition pruning, and predicate pushdown, improving query performance and reducing unnecessary scans by 40%.
- Developed streaming ETL pipelines using Apache Kafka to ingest and process real-time data feeds, enabling low-latency analytics for operational dashboards.
- Participated in root cause analysis (RCA) sessions for high-severity incidents, driving long-term preventive solutions and improving overall pipeline reliability.
- Led internal “Data Engineering Guild” sessions, sharing best practices for Spark debugging, AWS resource tuning, and cost optimization, fostering a culture of knowledge sharing.

- Designed and optimized Azure Data Factory pipelines for ingesting data from multiple sources (SQL Server, APIs, flat files), implementing partitioning, bucketing, and predicate pushdown to improve ETL performance by 20%.
- Automated data quality validation using Python + Azure Databricks, verifying schema conformity and detecting anomalies before data landed in downstream analytics layers.
- Implemented Azure Blob Storage lifecycle policies and storage account tiering to reduce costs, alongside cost analysis for AWS S3, achieving a combined 15%+ reduction in cloud storage spend.
- Documented pipeline architecture, data flow diagrams, and migration steps in Confluence to accelerate onboarding and ensure alignment across development teams.
- Built scalable data pipelines leveraging Delta Lake on Azure Databricks, enabling ACID transactions, schema evolution, and time-travel queries for incremental updates.
- Supported cloud migration initiatives by configuring Azure Virtual Networks, Key Vaults, and RBAC policies, ensuring secure integration between on-premises systems and Azure resources.
- Gained hands-on experience with Azure Monitor and Log Analytics to track pipeline performance, optimize queries, and proactively detect failures.
- Assisted in setting up CI/CD pipelines using Azure DevOps (Repos, Pipelines, Artifacts) to automate deployment of data workflows and improve release cycles.

- Gained hands-on experience with Azure, AWS, and GCP by assisting in the setup of compute, storage, and networking resources for internal projects.
- Supported infrastructure automation using ARM templates, AWS CloudFormation, and Terraform, learning how to reduce manual provisioning effort.
- Assisted in building ETL workflows with Azure Data Factory and loading datasets into AWS S3 for analytics use cases.
- Helped configure IAM and RBAC policies across platforms to apply least-privilege access and strengthen cloud security.
- Monitored cloud resources with Azure Monitor, AWS CloudWatch, and GCP Operations Suite, documenting alerts and performance insights for senior engineers.
- Shadowed DevOps teams in setting up CI/CD pipelines with Azure DevOps and Jenkins, supporting automated deployment testing.

SKILLS

- **Programming Languages:** Python, SQL, PySpark, Scala (Spark API, functional programming), Java (microservices, data processing), Linux
- **Big Data Technologies:** Databricks, Apache Spark, Hadoop
- **Cloud Platforms:** AWS (S3, Glue, Lambda, Kafka, Redshift), Azure, GCP
- **ETL & Orchestration Tools:** AWS Glue, Apache Airflow, Terraform
- **Databases:** MySQL, PostgreSQL, MongoDB, Snowflake
- **Data Governance & Security:** Unity Catalog, Delta Lake
- **CI/CD & Version Control:** Jenkins, Docker, Kubernetes, Git, Maven, Tomcat, SonarQube.
- **Development & Collaboration Tools:** Eclipse, PyCharm, GitHub, Lucid chart, GitLab, Power BI, Tableau

EDUCATION

Saint Louis University, St Louis, MO

Master of Computer and Information Sciences

Galaxy Degree and PG College, Hyderabad, India

Bachelor's in Computer Applications