Rahul Negi

Dehradun,Uttarakhand,India-248001
7534925201 rahull.negii1@gmail.com rahulnegi1 rahulnegi001

Professional Summary

Databricks Certified Data Engineer with 2+ years of experience designing and deploying scalable, low-latency data pipelines using PySpark, Kafka, and Databricks on AWS/GCP. Proven track record in cloud migration, Delta Lake architecture, and automation with CI/CD and Airflow. Adept at optimizing ETL pipelines, validating large-scale ingestion via metadata-driven frameworks, and enabling real-time analytics for enterprise clients.

Certification

- Databricks Certified Data Engineer Professional
- Databricks Certified Data Engineer Associate
- Data Lakehouse Fundamentals
- Python for Data Science, AI and Development

Industry Experience

Data Engineer (Digivate Labs Pvt Ltd) **Sept 2023 – Present**

- Proven expertise in migrating legacy systems (Cloudera, Redshift, Vertica, etc) to modern platforms like Databricks, reducing operational latency and improving query performance.
- Strong hands-on experience developing robust ETL pipelines for data ingestion, transformation, and optimization using Kafka, Debezium, PySpark, and Databricks Auto Loader. Successfully orchestrated and scheduled these pipelines using Databricks Workflows for scalable, automated data processing.
- Proficient in working across cloud platforms such as AWS and GCP, leveraging services like Redshift, S3, Lambda, GCS, and Athena for data warehousing and analytics.
- Experienced in converting legacy MapReduce jobs to PySpark and optimizing SQL queries for performance in Databricks SQL.
- Excellent problem-solving skills and the ability to deliver long-term scalable solutions under tight deadlines.

Key Client Projects at Digivate Labs Pvt Ltd:

Snapdeal

- Led EMR modernization—migrated 35 TB of historical data from S3 to Databricks and rebuilt Spark workflows using scalable Delta Lake pipelines.
- Orchestrated metadata-driven PySpark ingestion with schema validation, partition logic, and transformation of nested List, Map, and Struct types.
- Delivered clean, transformed data into Aerospike (NoSQL) for low-latency downstream use; handled schema mismatches and expression-level failures with log4j metrics.
- Migrated 10,000+ tables from Vertica to Databricks with schema consistency, improving query performance by 35% using partitioning and UTC standardization.
- Owned daily client communication—managed requirement gathering, issue resolution, and progress reporting for alignment and delivery.

PayU

- Executed Redshift-to-Databricks migration, optimizing performance and validating compatibility across high-throughput reporting workloads. Replaced legacy ETL with Delta Live Tables to unify batch and streaming workflows.
- Migrated 200+ Redshift SQL workloads with dynamic parameterization to Databricks SQL; enabled 4,000+ scalable query executions with tuning.
- Ingested 2 TB of data from AWS S3 into Databricks using Auto Loader; validated schema, row counts, and nulls; streamed MySQL data via Kafka-Debezium with deduplication.
- Achieved 2–3x query performance gain by leveraging Delta Lake optimizations, including Liquid Clustering and Spark execution tuning.

Battery Smart

- Ported 10+ Python scripts from Apache Pinot to Databricks Workflows, removing dependency on Pinot.
- Rebuilt pipelines with PySpark, optimizing with partitioning, caching, and clustering for 35% performance gain.
- Designed robust error-handling with retry logic and Slack alerting, reducing manual support by 50%.

BookMyShow - PoC

- Conducted ETL workload profiling and resolved bottlenecks using partitioning, indexing, and caching—boosting query speed by 30%.
- Reduced resource usage and execution time by 20% via advanced PySpark UDF tuning and optimization techniques.
- Authored technical documentation and led KT sessions to ensure smooth handover with zero productivity loss.

Data Science Intern (AllHeart Web Pvt Ltd) **June 2023 – Aug 2023**

- Collaborated with cross-functional teams to gather requirements and analyze data; utilized Linux Bash scripting for large-scale data cleaning and preparation tasks.
- Developed and deployed data pipelines for Google Business Listings using Python; implemented data modeling, text classification, and web scraping workflows.

Technical Skills

Programming Languages: Python, R, SQL (MySQL, PostgreSQL, Databricks Spark SQL)
Cloud Platforms: AWS (S3, Redshift, RDS), Azure, Google Cloud Platform (GCP)
Big Data Processing: Apache Spark SQL, Hadoop, Apache Kafka, Databricks Data Lakehouse
Data Warehousing: Amazon Redshift, Google BigQuery
ETL/ELT Pipelines: Designing and building scalable ETL pipelines (batch and real-time ingestion), Data ingestion using tools like Apache Airflow, AWS Glue, Databricks Autoloader
Data Analysis & Visualization: Pandas, NumPy, Matplotlib, Seaborn, Tableau, Power BI, Databricks Dashboards
Data Formats: Parquet, Avro, ORC, JSON, CSV
Workflow Automation: CI/CD pipelines, version control (Git), automation with Airflow and Databricks
Tools: Jupyter, VS Code, Linux, Databricks

Education

Chandigarh University **2023**
Master's in Data Science *Mohali, Punjab*

Hemvati Nandan Bahuguna Garhwal University **2021**
Bachelor of Science *Dehradun, Uttarakhand*