Rahul Rajesh

Trivandrum, India | +91 6282162210 | rahulrajeshofficial15@gmail.com | linkedin.com/in/rahulrajesh15

Summary

Data Engineer with 2+ years of experience building and optimizing large-scale data pipelines and cloud-native architectures. Skilled in **PySpark, Kafka, Hive, and AWS (S3, Glue, Lambda, Redshift, EMR)** with a proven record of improving performance and scalability. Experienced in SQL optimization and Spark code tuning, reducing execution time by 85% and cutting storage by 40%. Strong background in workflow automation, cost optimization, and regulatory-compliant data management. Recognized for delivering reliable, production-grade solutions that accelerate analytics and support data-driven decision-making.

Experience

Acabes International (Arab Bank) Sep 2023 – Present
Data Engineer *Kochi, Kerala, India*

Tech: PySpark, Kafka, Informatica BDM, Hive, Impala, MongoDB, Python, Shell, AWS (S3, Glue, Lambda, EMR, Redshift).

- Designed and deployed scalable ETL pipelines using Informatica BDM and **AWS Glue**, increasing retention analytics efficiency by 20% and enabling faster decision-making.
- Migrated deeply nested MongoDB collections to Hive, designed optimized data models with struct and array types, reducing query runtime by 65% and storage by 40%, enabling faster analytics at scale.
- Optimized complex SQL queries and Spark code, reducing execution time by over 85% and significantly improving overall system performance and scalability.
- Automated structured CSV report generation from unstructured invoice text using regex-based Python scripts, eliminating 15+ hours/week of manual processing and reducing human error.
- Migrated 10+ workflows to PySpark on **AWS EMR**, achieving 4× faster execution and lowering infrastructure costs by $50K/year.
- Implemented Slowly Changing Dimension (Type 2) in Redshift to maintain historical customer records and ensure audit/regulatory compliance, improving accuracy of compliance reporting and customer lifecycle tracking.

Projects

Real-Time Fraud Detection System

- Built real-time fraud detection pipeline with **Kafka, PySpark Streaming, and AWS Kinesis**, processing millions of daily transactions and reducing fraud response latency.
- Processed data in **AWS S3 Data Lake**, orchestrated ETL pipelines with **Glue**
  Lambda, and designed Redshift warehouse to cut investigation time by 40% and enhance compliance reporting.
- Implemented ML-based anomaly detection with Python, reducing false positives by 25% and strengthening fraud monitoring accuracy.

Education

Younus College of Engineering and Technology Aug 2019 – Jul 2023
B.Tech. in Computer Science and Engineering (CGPA: 8.94/10) *Kollam, Kerala, India*

- University Topper in Semester 3 (10/10 SGPA).
- Completed a 6-month Data Science Internship by ICTAK, gaining practical exposure to predictive analytics and data visualization.

Technical Skills

Cloud: AWS (S3, Glue, Lambda, EMR, Redshift, CloudWatch)
Big Data & Streaming: PySpark, Spark SQL, Hadoop, Kafka, Hive, Impala
ETL/Data Pipelines: Informatica BDM, CDC, SCD-2
Databases & Storage: MongoDB, PostgreSQL, HDFS, Delta Lake
File Formats: Parquet, ORC, Avro, JSON, CSV
Programming: Python (Pandas, PyTest), SQL (Optimization), Shell Scripting
Data Modeling & Optimization: Dimensional Modeling (Star, Snowflake), Partitioning & Bucketing, Query Tuning

Publications

Enhanced network intrusion detection using deep reinforcement learning | Springer | 2023

- Developed RL-based model (DQN, PPO) in Python/PyTorch for real-time anomaly detection, achieving 95% F1-score and enabling proactive threat detection.
- Reduced false positives by 30% compared to signature-based methods, improving operational efficiency of security analysts.

Languages

- English
- Malayalam