Archit Raj

Bengaluru, KA | raj.ar@northeastern.edu | 9065704710 | LinkedIn | Portfolio | GitHub

Summary

Product-focused Data Engineer with 5 years architecting scalable real-time and batch data platforms using Python, SQL, PySpark, Kafka, dbt, Airflow, and Terraform across AWS and Databricks. Delivered outcomes across LLM-powered analytics and ML pipelines, including 40% latency reduction, 30% uptime improvement, and 15% Redshift cost savings. Build production-grade ETL and streaming systems that power product analytics, feature adoption, and executive decision-making across cross-functional engineering teams

Professional Experience

Senior Data Engineer - Abecedarian | Boston MA - USA **Sep 2023 - Jul 2025**

- Orchestrated production-grade ETL pipelines across batch and streaming systems using PySpark, dbt, AWS Glue, and Terraform, reducing refresh latency by 60% and enabling fully automated CI/CD-compliant deployments
- Prototyped a Kafka-based churn prediction engine ingesting 500K+ events per day, generating behavioral cohorts and LLM-ready signals for downstream analytics
- Architected a logistics pipeline handling ~2M daily events using PostgreSQL, S3, and Lambda, enabling real-time tracking of inventory flow and delivery delays
- Delivered customer analytics workflows in Databricks using PySpark and dbt, accelerating A/B testing, segmentation, and feature rollout analysis with 40% faster data availability
- Engineered a reusable onboarding framework combining REST API ingestion, relational joins, and KPI modeling, reducing ramp-up time by 40% and adopted across multiple teams
- Published technical articles on SQL tuning, Spark optimization, and ETL architecture, reaching over 1,500 readers and enhancing visibility within the data engineering community

Data Engineering Co-op (Backend Systems – Data & AI) - IBM | San Jose CA - USA **Jun 2022 - Dec 2022**

- Migrated Db2 modules by translating PL/SQL to Call-Level Interface and refining REST APIs, boosting interoperability and reducing integration time by 30%
- Tuned 1500+ SQL queries with indexing and partitioning, reducing latency by 50% and stabilizing CI/CD pipelines by integrating regression checks, increasing reliability by 15%
- Enhanced CI/CD in a regulated cloud environment by resolving Db2 test failures and engineering cleanup pipelines for 32TB+ geospatial data improving retrieval latency by 20% and reducing storage overhead by 15% using Python

Data Engineer - Vittude | Sao Paulo - Brazil **Jan 2020 - Aug 2021**

- Designed and maintained data infrastructure using PostgreSQL, SQL, Python (NumPy, Pandas, PySpark), Django, and dbt, automating 250+ workflows via REST APIs for ingestion, transformation, and reporting while ensuring regulatory compliance
- Implemented a real-time analytics pipeline using Amazon MSK, AWS Lambda, and S3 with SQL model transformations, achieving 99.4% ETL reliability and enabling feature usage tracking and experimentation
- Enhanced Redshift warehousing and Tableau dashboards, improving query speed by 40%, reducing storage costs by 15%, and lowering support issues by 20% through behavioral insights
- Led a data governance initiative to standardize schema design and lineage tracking, reducing duplication and improving onboarding for analysts and engineers
- Deployed alerting via CloudWatch and SNS to resolve pipeline failures, reducing downtime by 30% and improving SLA compliance

Data Engineer Intern - Vittude | Sao Paulo - Brazil **May 2018 - Jan 2019**

- Redesigned AWS Redshift warehousing and built ETL pipelines using Glue, Glue Crawler, and Lambda, accelerating ingestion by 20%, increasing processing speed, and lowered operational costs by 10%
- Developed financial dashboards in Power BI by integrating Redshift and S3, improving KPI reporting accuracy by 20% and saving 9 hours per week in manual reporting through near real-time visualization
- Scripted data validation checks using Python and SQL to flag anomalies and schema mismatches, reducing data quality issues by 30% and strengthening dataset reliability across teams

Projects

Stock Market Kafka Streaming with AWS (Kafka Stock Market - AWS) **Aug 2023 - Sep 2023**

- Built a high-throughput stock event pipeline using Kafka, S3, EC2, and Glue—streamed 500K+ events/sec, reducing latency by 40%

TrendWatch: YouTube Trending Video Analytics (YouTube Trending Vid-AWS) **May 2023 - Jul 2023**

- Ingested 10K+ YouTube records/day into an AWS pipeline and visualized trends in Tableau, reaching 150+ views in the first week

Education

MS in Data Analytics Engineering - Northeastern University - Boston MA, USA **Sep 2021 - Jul 2023**

BS in Management Information Systems - Iowa State University - Ames IA, USA **Aug 2016 - May 2020**

Skills

Programming: Python, SQL, Django, Flask, Pandas, NumPy, PySpark, DBT, Shell Scripting, RESTful APIs, Docker, Terraform
Databases: Relational-DBs (MySQL, PostgreSQL, Oracle, SQL Server, IBM DB2), NoSQL DBs (DynamoDB, MongoDB, Redis)
Data Warehouse & Big Data Frameworks: Redshift, Snowflake, Hadoop, Spark, EMR, Airflow, Databricks, Kafka, Lambda
Cloud, DevOps & CI/CD: AWS (S3, Step Functions), ECS, Prometheus, Elasticsearch, GitHub, JIRA, Asana, Obsidian, Draw.io
Analytics & Business Intelligence: Tableau, Power BI, Looker, QuickSight, Apache Superset, Qlik (QlikView/Qlik Sense)