Kavali Yashwanth Royal

Data Engineer — batch ELT, streaming and analytics engineering

Bangalore, India (open to remote) · yashwanthroyal9133@gmail.com · github.com/Yashu24-coder · LinkedIn · yashu24-coder.github.io

I build data pipelines end to end: land raw files, clean them once in code, model them into a tested schema, and refuse to publish a number that fails a quality check. Four public projects below — batch on AWS, streaming with Kafka and Spark, and dbt on top — each with its own test suite and CI.

PROJECTS

Retail data platform on AWS

github.com/Yashu24-coder/aws-retail-data-platform

Python · PySpark · AWS Glue · Athena · Redshift Serverless · Terraform

- Batch ELT over four inconsistent CSV exports: S3 bronze, Glue silver, star-schema gold, served through Athena and Redshift. 5,760 raw rows → 2,944 order-line fact rows across 14 months.
- Wrote 12 data-quality gates and 23 tests that fail the build rather than publish a bad number; every bucket, job, workgroup and alarm provisioned in Terraform.

Kafka order-events pipeline

github.com/Yashu24-coder/streaming-orders-pipeline

Kafka · Spark Structured Streaming · PySpark · Parquet

- Processed 41,236 events in 83 micro-batches with a 120-second watermark and 5-minute tumbling windows; 7.34% routed to a dead-letter queue with an explicit reason code each.
- Idempotent sink keyed by micro-batch id, so a replayed batch overwrites itself instead of double counting; 20 tests cover watermarking, deduplication and dead-letter routing.

SaaS revenue analytics with dbt

github.com/Yashu24-coder/dbt-saas-analytics

dbt · SQL · DuckDB · data modelling

- 13 models across staging, intermediate and marts, defining MRR, net new MRR and cohort retention once so the metrics reconcile by construction.
- Incremental fact table on a hashed key and a type-2 snapshot for plan changes; 27 data tests, including singular assertions for double-billing and uncollected revenue.

NYC taxi lakehouse on S3

github.com/Yashu24-coder/aws-nyc-taxi-data-engineering

AWS S3 · Glue · Athena · Terraform · pandas

- Bronze/silver/gold lakehouse over the public NYC TLC trip records, with Glue transformations, Athena analysis queries and Terraform for the whole stack.
- Data-quality rules quarantine unparseable timestamps, negative fares and duplicate rows; 5 tests and a CI workflow run on every push.

EDUCATION

Sri Venkateshwara College of Engineering and Technology

Bachelor of Technology — Computer Science (Artificial Intelligence & Machine Learning) · Graduated 2025 · CGPA 8.31 / 10

SKILLS

Languages Python, SQL, PySpark **Data** dbt, Kafka, Parquet, DuckDB **Cloud** S3, Glue, Athena, Redshift Serverless
Engineering Terraform, pytest, GitHub Actions, Docker, Git

All figures above come from synthetic or public sample datasets built to reproduce real-world defects; the pipelines, tests and CI are real and runnable from a clone.