Saket Kumar

Senior Data Engineer · 6+ Years · Bengaluru, India
+91-7484844919 | kumar.saket0021@gmail.com | linkedin.com/in/saketkr21/ | saketkumar.pages.dev

Professional Summary

Senior Data Engineer with **6+ years reducing compute cost, query latency, and CI/CD release cycles** on **Snowflake, Databricks, and BigQuery** data platforms. **Founding data engineer** on New Relic's Product Rating platform. Expert in **data & analytics engineering**, **Apache Spark tuning**, and **cross-warehouse migrations**. Portfolio of Snowflake compute savings on billing-critical pipelines – **$45K+ annually on the flagship ELT pipeline alone** – with **50% dbt model speed-ups** and platform scale from **data products serving 4M+ SKUs to 2,000+ production DAGs**. Creator of **dbt-polyglot** (PyPI): Open-source Python package transpiling SQL from Snowflake → Spark/Databricks. Multi-cloud (**AWS · GCP · Azure**); **GDPR/PII champion**.

Experience

- **New Relic** — Bengaluru, India
  Senior Data Engineer – Founding Member, Product Rating Data (India) — *July 2024 – Present*
  - **Open Lakehouse Migration (Snowflake → Iceberg)**: Led migration of ELT pipeline (440+ dbt models) from Snowflake to a fully open-source lakehouse – Iceberg on S3, **Spark Thrift** compute, **dbt-spark, Project Nessie catalog, Airflow 3**. Built automated model-parity validation (100% row/column checks) and mismatch-spike dashboard for zero-downtime cutover.
  - **Snowflake Cost & Performance**: Refactored critical dbt models on **Snowflake**, **reducing compute spend by $45.2k annually** while **improving query performance by 50%**; Achieved 100% data parity via automated validation frameworks.
  - **CI/CD for Data**: Designed a robust CI/CD pipeline using Jenkins and GitHub, automating linting, BDD / unit tests, dbt Cloud triggers, quality gates, and Slack alerts – **eliminating manual PR overhead** and **cutting deployment lead time by 85%**.
  - **Zero-Downtime Migration & Monetization Modeling**: Led migration of a billing-critical pipeline (10 GB/hour) from Airflow 1.0 to dbt Cloud + **Snowflake** with **100% data parity** and zero customer-visible downtime. Engineered a rating engine combining usage meters and complex billing rules – enabling accurate monetization of **15+ new product SKUs**.
- **Falabella** — Bengaluru, India
  Data Engineer — *June 2022 – May 2024*
  - **Revenue Impact – Fast Shipping Tags**: Built a high-performance data product on **BigQuery**, DataProc, Pub/Sub, and Looker Studio for **4M+ SKUs** across Falabella.cl's third-party marketplace (LATAM); drove a **50% lift in platform conversion rates** by delivering **97% accurate** insights.
  - **Serverless Ingestion Framework**: Designed a serverless Cloud Functions + Federated Queries ingestion layer that **reduced pipeline setup time by 80%** and eliminated Compute Engine maintenance overhead.
  - **Airflow Observability at Scale**: Built a custom Airflow monitoring dashboard for **2,000+ DAGs** (Airflow API, Cloud Functions, GCS, Docker) with automated alerts, root-cause analytics, and remediation suggestions – **improving MTTR by 60%** and pipeline availability to **99.9%**.
- **Infosys Limited** — Hyderabad, India
  Specialist Programmer – Clients: Walmart & Five Below — *Oct 2020 – June 2022*
  - **Enterprise ETL on Databricks & Spark (Walmart)**: Developed Spark-Scala + PySpark ETL pipelines on **Databricks** and GCP DataProc, ingesting massive datasets from GCS to **BigQuery** with automated data-quality and integrity gates; leveraged **Spark Structured Streaming** for near-real-time ingestion patterns.
  - **PII Encryption Framework (Five Below)**: Engineered a cross-org AES-256 / PGP encryption / decryption framework in Python, PySpark, Scala and Java processing **80 GB+ files** for end-to-end PII compliance across multi-language producers.

Technical Skills

- **Warehouses & Lakehouses:** Databricks, Snowflake, BigQuery, Delta Lake, Apache Iceberg, dbt (Core & Cloud)
- **Big Data, Streaming & Orchestration:** Apache Spark, PySpark, Spark Structured Streaming, Kafka, CDC (Debezium), Airflow, ETL / ELT, Great Expectations, dbt-expectations
- **Cloud & Infra:** AWS (Glue, S3, Athena, EMR), GCP (BigQuery, DataProc, Pub/Sub, Cloud Functions), Azure, Kubernetes, Docker, Terraform, MongoDB, Linux
- **Languages & Frameworks:** Python, SQL, NoSQL, Scala, Shell; FastAPI, Pandas, Pytest, REST APIs
- **DevOps, BI & Governance:** Jenkins, GitHub Actions, Prometheus, Grafana, BDD/Unit tests; Looker Studio, Tableau; GDPR pipelines, PII compliance, AES-256/PGP encryption, Data Quality Gates

Personal Projects

- **dbt-polyglot** — pypi.org/project/dbt-polyglot | github.com/Saketkr21/dbt-polyglot
  Open-source Python package on PyPI enabling dbt models authored in **Snowflake**, BigQuery, Redshift, T-SQL, or DuckDB to execute on **Spark / Databricks** or any other dialect unchanged via compile-time SQLGLOT transpilation with a Spark correctness fix-up layer.
- **lakehouse-lab** — github.com/Saketkr21/lakehouse-lab
  Self-authored 58-module Data Engineering curriculum: Spark performance (skew/OOM/AQE), **Apache Iceberg** & **Delta Lake**, **Kafka** + Structured Streaming, **Debezium CDC**, **dbt** quality with Great Expectations, and **Airflow**.

Continued on page 2...

Education

- **Netaji Subhash Engineering College** — Kolkata, India
  Bachelor of Technology - Electronics and Communication; GPA: 8.30 / 10 — *July 2016 - June 2020*

Certifications & Achievements

Certifications: dbt Fundamentals – dbt Labs (2024) · Azure Data Fundamentals DP-900 – Microsoft (2021) · Deep Learning Specialization – DeepLearning.AI (2020) · Machine Learning – Stanford / Coursera | Grade 95% (2019) · Python Advanced – Cutshort (2023) · AWS AI & ML Scholarship – Udacity (2026).
Leadership: Mentor junior engineers on dbt, Snowflake tuning, Spark performance, and SQL optimization.