Pronnoy Dutta

+91 7042768041 | pronnoy1998@gmail.com | Gurugram, India
LinkedIn | GitHub | Medium

Profile

Lead Data Engineer with 5.5+ years architecting cloud-native data platforms on **AWS** that process **200M+ records daily**, with a proven track record of cutting pipeline runtimes by **30%** and eliminating hundreds of hours of manual work annually. Expert in **PySpark, Spark optimization, and dimensional data modeling**; known for translating complex business requirements into production-grade, scalable solutions that directly accelerate analytical decision-making.

Technical Skills

Programming Languages: Python, SQL, Java, C#
Distributed & Big Data: Apache Spark, PySpark, Hadoop, Hive, YARN, Kubernetes
Cloud (AWS): S3, Glue, EMR, RDS, Redshift
Data Architecture: Data Pipelines, Data Lakes, Data Warehousing, Dimensional Modeling, SCD Type-2
Performance Engineering: Spark Optimization, Query Tuning, Partitioning, Resource Management
Orchestration & DevOps: Apache Airflow, Control-M, Git, CI/CD, Docker, Linux
Visualization: Tableau, Grafana
Certifications: AWS Security Specialty, AWS Solutions Architect Associate, CCNA

Experience

Axtria Mar 2022 – Present
Project Lead → Senior Associate → Associate → Analyst *Gurugram, India*

- Engineered cloud-native batch data pipelines on **AWS (S3, Glue, EMR, Redshift)** supporting business-critical analytics for a **50M-patient** commercial pharma dataset, achieving **99.9% SLA compliance** across all production runs.
- Drove a **30% end-to-end performance improvement** by migrating **15+ legacy Hive/SQL pipelines** to PySpark-based distributed processing, reducing daily pipeline wall-time from ∼9 hrs to ∼6 hrs.
- Reduced Spark processing time by **30%** on pipelines ingesting **200M records/day** by diagnosing data skew, tuning partitioning strategies, and right-sizing executor configurations — saving **˜$40K/yr** in EMR compute costs.
- Designed **SCD Type-2** dimensional data models supporting full historical auditability across **3+ years** of patient-level data, enabling downstream teams to eliminate ad-hoc reconciliation queries entirely.
- Automated Tableau refresh orchestration for **30+ dashboards** using a Python-based scheduling framework, eliminating **7 hours** of manual effort per cycle (**˜350 hrs/yr** saved).
- Served as primary escalation owner for production data incidents, leading RCA across **20+ P1/P2 events** and implementing monitoring guardrails that reduced repeat incidents by **60%**.
- Led and mentored a team of **4 data engineers**, owning sprint planning, code reviews, and architecture decisions for a platform serving **10+ downstream analytics consumers**.

Infosys Nov 2020 – Mar 2022
Systems Engineer *Remote*

- Architected cloud-based monthly sales analytics pipelines using **AWS S3, Python, and PySpark**, delivering data products for a retail client spanning **500+ stores** across 3 regions.
- Built a star-schema data warehouse with **8 fact/dimension tables** tracking revenue, billing frequency, and customer engagement KPIs — reducing ad-hoc reporting turnaround from days to hours.
- Developed KPI computation logic for **12+ business metrics** (revenue, churn, basket size) that became the single source of truth powering executive dashboards.

Education

Bharati Vidyapeeth College of Engineering, Pune — B.Tech, Computer Science July 2016 – June 2020
AWS Community Builders — Official Member, Security Team