# Anmol Mishra

anmol99993@gmail.com | LinkedIn | +91 8603045351

## PROFILE SUMMARY

Backend and Data Engineer with experience building production-scale batch and streaming data pipelines using Apache Spark, Kafka, Debezium, Trino, and AWS. Experienced in end-to-end data pipelines, CDC, analytical data modeling, cloud infrastructure, data quality, observability, and production incident response. Interested in distributed systems, database internals, performance engineering, and AI-powered data platforms.

## SKILLS

- **Programming:** Python, SQL, Scala, Java, Unix Shell
- **Data Engineering:** Apache Spark, Airflow, Kafka, Debezium, Trino, CDC, ETL/ELT, Data Modeling, Data Quality
- **Distributed Systems:** Distributed Processing, Replication, Concurrency, Consistent Hashing, Query Optimization
- **Cloud & Platform:** AWS, S3, EKS/Kubernetes, Docker, Terraform, Argo CD, CI/CD
- **Data & Storage:** PostgreSQL, Redis, Delta Lake, Iceberg, Parquet
- **Observability AI/ML:** Prometheus, Grafana, OpenTelemetry, CloudWatch, Humio, OpsRamp, RAG, LLMs, Spark-based AI/ML Pipelines

## EXPERIENCE

**Hewlett-Packard Enterprise** Jan 2023 – Present

***Software Engineer – Distributed Data Platform***

- Designed and operated production batch and streaming pipelines using **Apache Spark, Scala, SQL, Trino, Kafka, and Debezium**, processing **1+ TB/day** with reliable CDC and low-latency synchronization.
- Built end-to-end cloud-native data services across **ingestion, transformation, data modeling, metadata, storage, and query execution**, delivering **15+ BI data models** for Customer Success, FinOps, and Engineering.
- Improved pipeline performance, scalability, and reliability through **Spark partitioning, broadcast joins, AQE, resource tuning, Kubernetes, and AWS**, maintaining **95%+ production incident SLA compliance**.
- Built and maintained production observability across distributed workloads using **Prometheus, Grafana, CloudWatch, and Humio**, supporting incident diagnosis, root-cause analysis, alerting, and durable remediation.
- Improved data-platform reliability through **CI/CD, Kubernetes/Argo CD, infrastructure automation, documentation, and production troubleshooting** across cloud-native data services.

## OPEN SOURCE

- **Apache Spark – PR #58760:** Contributed to Spark SQL analyzer behavior for pipe SET column resolution, including planner changes and SQL test coverage. GitHub PR
- **AWS PyDeequ – PR #289:** Contributed data-quality functionality and supporting tests/documentation to PyDeequ, a Spark-based data-quality framework. **Merged upstream.** GitHub PR

## PROJECTS

**RaftDB – Mini Distributed Database** Java, Raft, LSM Tree, WAL, RocksDB, Apache Calcite, SQL

- Built a **distributed database from scratch** to understand the core systems behind production databases and query engines, rather than treating storage, consensus, recovery, and query optimization as black boxes.
- Implemented a fault-tolerant storage layer with **Raft consensus, leader election, log replication, WAL, snapshots, SSTables, Bloom Filters, and background compaction**, supporting crash recovery and automatic failover.
- Added a SQL query layer using **Apache Calcite** for parsing, relational planning, and cost-based optimization, with predicate pushdown, projection pruning, distributed joins, aggregations, and parallel query execution.

## EDUCATION

**Vellore Institute of Technology | 8.81/10 CGPA** 2019 – 2023 *B.Tech. Computer Science and Engineering*

Chennai, Tamil Nadu