

Prajwal Narendra Tidke

619-520-9475 | ptidke9@gmail.com | [LinkedIn](#) | [GitHub](#) | [Portfolio](#)

Professional Summary

Data Engineer with 5+ years of experience building high-scale batch and streaming pipelines processing multi-terabyte daily volumes across finance, manufacturing, and healthcare domains. Proficient in optimizing lakehouse architectures using AWS, Databricks, and Snowflake, with recent work building production data pipelines for RAG and LLM-powered applications.

Education

San Diego State University

MS in Big Data Analytics (Academic Excellence Award, CGPA: 4.0/4.0)

August 2024 – May 2026

San Diego, CA

Technical Skills

Languages: Python, Pandas, PySpark, Scala, SQL, Java

Data Engineering & Orchestration: Airflow, Spark, Kafka, Flink, Faust, Hive, Hadoop, dbt, SCD (Type 1 & 2), ETL/ELT, Medallion Architecture, Dimensional Modeling, Data Contracts, Data Quality, PostgreSQL, Oracle, DuckDB, Delta Lake, Apache Iceberg, Parquet, MinIO

Cloud & DevOps: AWS (S3, Glue, RDS, Athena, EMR, Lambda, SageMaker, CloudWatch), Azure, Databricks, Snowflake, Docker, Kubernetes, Terraform (IaC), BigQuery, CI/CD, FastAPI, Supabase, GitHub Actions, Grafana

AI Systems: RAG, LLMs, LLM Orchestration, LangChain, Vector Databases (ChromaDB), Azure OpenAI, Prompt Engineering, Claude Code, MCP, RAGAS

Experience

San Diego State University - LINC Lab

Graduate AI Data Engineer

San Diego, USA

February 2025 – May 2026

- Owned an ingestion pipeline parsing 1.5K+ multi-format clinical documents into ChromaDB via Google LangExtract, automating metadata extraction and removing **15 hours/week** of manual annotation.
- Designed a metadata-tagged chunking and semantic-filtering retrieval layer over ChromaDB and Azure OpenAI, cutting p95 query latency by **55%** (3.8s to 1.7s) for KALM.
- Deployed a RAGAS and LLM-as-a-judge evaluation harness gating answer quality at a **0.96** faithfulness threshold across a 200-query clinical test set, with HIPAA-aligned PII/PHI redaction.
- Designed and built a 5-stage deterministic state machine for KALM's crisis-response routing, with defined fallback states for high-risk inputs. [Link to KALM mental health companion](#)
- Ran a multi-run ablation study on a peer-knowledge augmentation layer, statistically validating and shipping the best configuration.

TresVista Analytics

Senior Data Engineer

Pune, India

October 2023 – August 2024

- Designed ELT pipelines on Databricks Medallion Architecture orchestrating portfolio analytics for a **\$550B+** AUM fund-of-funds client, improving pipeline processing reliability by **30%**.
- Architected an event-driven ingestion system (AWS SQS/SNS to S3) processing **1.2 TB/day**, and tuned high-cardinality tables with Delta Lake Z-Ordering to cut compute cost **\$42K/year** and median query latency by **38%**.
- Integrated 7 market and alternative-data feeds (S&P Global, Visible Alpha, Placer.ai, Similarweb) via a custom API connector, enforcing data contracts in PyTest CI/CD gates to cut weekly manual analysis by **45%** at **100%** SLA over 6 months.
- Mentored a team of 4 data analysts on best practices and delivery standards, recognized internally with the Best Mentor Award.

LTIMindtree

Senior Data Engineer

Pune, India

July 2021 – October 2023

- Directed technical planning, code reviews, and Agile delivery across 5+ enterprise clients, consistently meeting SLA targets.
- Led the migration of legacy enterprise data infrastructure to AWS Glue and Snowflake with SCD Type 2 history tracking, securing **\$650K/year** in operating savings and improving pipeline data freshness SLAs to **99.9%**.
- Built curated reporting datasets and refresh pipelines feeding Power BI (NA) and AWS QuickSight (EU) across SQL Server, Oracle, and Snowflake sources, standardizing metrics consumed by C-suite and senior management across both regions.

Data Engineer

July 2020 – July 2021

- Overhauled legacy batch ETL workflows, tuning Spark and Snowflake query performance via predicate pushdown and indexing strategies to accelerate monthly regulatory reporting timelines by **80%**.
- Built a schema-validation framework across 7+ downstream reporting pipelines, catching schema drift and cutting production data errors by **63%**.

Projects

Risk Pulse – Real-Time Portfolio Risk Intelligence Platform

[GitHub Link](#)

- Developed a real-time Kafka pipeline ingesting 8 equity symbols, computing 5-min VWAP, volatility, and drawdown alerts into a date-partitioned Parquet lake on embedded DuckDB; cut warm p50 from **11ms** to **0.5ms** by removing a cross-schema resolution cost.
- Modeled dbt marts with enforced data contracts (typed schemas, primary-key and not-null constraints), 10 data-quality tests, and column-level lineage, failing builds on schema drift before the serving layer.
- Built an Apache Iceberg table on MinIO as a snapshot of the Parquet lake and benchmarked Parquet vs Iceberg read latency; layered a ChromaDB and Azure OpenAI RAG service over the marts for natural-language risk queries.

Awards & Certifications

DataCamp - Certified Associate Data Engineer

Issued 2024

LTIMindtree - LTI GoMX Business Unit of the Year

Awarded 2022