

Pavan Badempet

Data Engineer | Python, PySpark & SQL

+91-7989256101 | pavan9b@gmail.com | linkedin.com/in/pavanbadempet | github.com/pavanbadempet | Hyderabad, India

Summary

Data Engineer with 3 years of experience building high-throughput distributed ETL data pipelines, Medallion Lakehouse platforms, and on-prem/cloud infrastructure across BFSI (Nomura Capital) and Automotive (Nissan) domains. Proficient in Python, SQL and PySpark performance tuning, dimensional modeling, workflow orchestration using Airflow and AutoSys, and data platform modernization with Databricks, AWS and Gen AI Integrated Development.

Skills

Languages: Python, Java, Scala, SQL

Big Data: Apache Spark, PySpark, Spark SQL, Databricks, Unity Catalog, Structured Streaming, Apache Iceberg, Delta Lake, Medallion Architecture, Dremio, MinIO, HDFS

Modeling: Dimensional Modeling (SCD Type 2, Star Schema), CDC, Schema Evolution, Data Quality

Cloud and Tools: AWS (S3, Lambda, Step Functions, EventBridge, SNS), Apache Airflow, AutoSys, Docker, Git, GitHub Actions, PostgreSQL, SQL Server, FastAPI, Streamlit

Work Experience

Tata Consultancy Services

Nov 2023 – Present

Data Engineer

Hyderabad, India

- Architected scalable batch extraction and transformation jobs using Spark and Spark SQL inside an Apache Iceberg Medallion Lakehouse at **Nomura Capital**, handling 1B+ records per month and serving 200K+ analytical queries daily.
- Optimized distributed computation via Adaptive Query Execution, broadcast joins, dynamic partition pruning, caching, and shuffle tuning, reducing SLA-bound runtime by 30%.
- Migrated workloads from legacy YARN and HDFS to Kubernetes with MinIO object storage, resolving dependency isolation challenges and lowering infrastructure spend by 25%.
- Spearheaded automation of orchestration schedules and deployment configurations, maintaining 99.9% on-time completion rate across 50+ recurring batch runs.
- Diagnosed failures via executor logs and DAG execution plans, constructed automated retries, fallback recovery handlers, and operational runbooks, cutting Mean Time to Recovery (MTTR) by 30%.
- Developed event-driven cloud capture solutions leveraging AWS Lambda, Step Functions, and EventBridge at **Nissan**, handling 100K+ events daily with idempotent delivery and schema validation via a Streamlit dashboard.
- Established programmatic quality validation rules and version-controlled table evolution handlers across 20+ source systems, improving output reliability to 99.5% for downstream consumers.

Projects

AI Healthcare System | Python, PySpark, Databricks, Delta Lake, Airflow, PostgreSQL, Docker, RAG

- Designed Batch and Streaming ETL workflows handling 1M+ records using Delta Lake Medallion Architecture and SCD Type 2, optimizing partitioned storage and MERGE operations to reduce runtime by 40%.
- Delivered FastAPI endpoints and PostgreSQL indexing to lower analytical query latency by 35%, containerized workloads with Docker and scheduled jobs using Airflow DAGs.
- Implemented a Retrieval Augmented Generation pipeline integrating Cloudflare AI embeddings with a local vector cache, compressing context payload size by 80% and achieving sub-100ms response times.

AI Recommendation Engine | Python, PySpark, Databricks, Structured Streaming, Delta Lake, Airflow, FastAPI

- Engineered batch and real-time ingestion flows using PySpark, SQL, Structured Streaming on Databricks for 20M+ records, enforcing data quality checks, Delta Lake checkpointing, and incremental loading.
- Built a low-latency serving API using FastAPI and vector retrieval engine with pgvector/FAISS indexing for candidate matching, enabling sub-50ms lookups across 100K+ items and automated scheduling by GitHub Actions.

Education

Guru Nanak Institutions Technical Campus

Aug 2019 – Sep 2023

Bachelor of Technology

Hyderabad, India

- **Major:** Computer Science and Engineering | CGPA: 8.2/10 | First Class with Distinction
- **Minor:** AI/ML | CGPA: 8.6/10
- **Scholarship:** Prime Minister's Scholarship (PMSS)