

YASH DANTALE

Data Engineer | PySpark & Apache Iceberg | AI-Ready
Pune, India | +91 9420638161 | yashdantale@gmail.com | linkedin.com/in/yashdantale

PROFESSIONAL SUMMARY

Data Engineer with **3+ years** of experience engineering scalable **Data Lakehouse** architectures on **AWS** and **Azure**. Specializes in **PySpark**, **Apache Iceberg**, and **SQL** optimization, with applied competency in **Generative AI** and **LLM** integration. **Delivered** enterprise-grade migrations and 'Golden Record' MDM solutions that improved pipeline performance by **30%**. **Automates** complex ETL workflows to ensure strict data governance in regulated pharmaceutical environments.

TECHNICAL SKILLS

Languages	Python (Advanced), SQL, PySpark, Scala, Bash Scripting
Big Data & Storage	Apache Spark, Apache Iceberg, Hadoop, AWS S3, Azure Data Lake Gen2
Cloud Platforms	AWS (Glue, EMR, Lambda), Microsoft Azure (Data Factory, Synapse Analytics), Databricks
AI & LLM Stack	LangChain, Vector Databases, RAG Architectures, AI Mathematics
ETL & Tools	Modak Nabu, Salesforce APIs, JDBC, REST APIs, Veeva EDC, SAAMA, Airflow, Microsoft Power BI

PROFESSIONAL EXPERIENCE

Cognizant Technology Solutions Data Engineer	Pune, India November 2022 - Present
Spearheaded the migration of 10TB+ of legacy pharmaceutical data to an Apache Iceberg architecture on AWS S3, enabling ACID transactions and time-travel queries for critical audit trails. Architected high-availability ingestion frameworks processing 10M+ records daily from Oracle (JDBC) and Salesforce, achieving 99.9% data uptime for downstream analytics. Engineered a 'Golden Record' identity resolution system using cross-reference mappings (XREFs), unifying disparate customer data sources into a single source of truth. Developed low-latency API connectors for Veeva EDC and SAAMA, reducing clinical trial data reporting lag by 40% and ensuring high-accuracy integration. Accelerated infrastructure deployment using Python scripts to dynamically generate SQL configurations, eliminating 15+ hours/week of manual setup and standardizing environments. Refactored legacy SQL scripts into modular PySpark pipelines (Medallion Architecture: Raw, Trusted, Curated), significantly improving data lineage and regulatory auditability. Optimized Spark partitioning and shuffle operations, resulting in a 30% reduction in end-to-end pipeline execution time and cloud compute costs. Orchestrated enterprise data governance workflows using Modak Nabu, reducing data quality errors by 25% through programmatic validation rules. Operationalized outbound data synchronization from the Data Lakehouse to Salesforce, ensuring seamless business operations with zero manual intervention. Awarded 'Owning It' (2025) for exceptional technical leadership and delivery of critical production releases under tight deadlines.	

EDUCATION & CERTIFICATIONS

Bachelor of Engineering in Electronics & Telecommunications Modern College of Engineering (SPPU) Honors in AI/ML	CGPA: 9.34/10.0 2018 - 2022
Databricks Certified Data Engineer Associate (Credential ID: 143606273) Microsoft Certified: Azure Data Fundamentals - GSI NVIDIA Technologies Curriculum - AI/ML Foundations	