Shubh Oswal

+1 (608) 298 8469 - soswal2506@@gmail.com - linkedin

TECHNICAL SKILLS

Programming Languages: Python, SQL, Scala, Java, Linux
Big Data/ETL: Spark SQL, PySpark, Spark, Databricks(Delta Lake, Unity Concepts), Kafka
Cloud: AWS (S3, EC2, Glue, Athena, Redshift), Azure (ADF, Synapse, ADLS Gen2, Azure SQL)
Data Stores: Snowflake, PostgreSQL, MySQL, NoSQL: MongoDB, Hadoop
ML Features: MLflow (tracking/registry concepts), feature engineering patterns; SageMaker
Dev/Tooling: Docker, Git, CI/CD, Airflow, Databricks Jobs, Jira, Splunk

WORK EXPERIENCE

Data Engineer
Lirik Inc, Milpitas, California *June 2025 - Present*

- Built analytics-ready curated datasets from payer/EHR/adjudication data (10M+ records) using Databricks, PySpark, Spark SQL, SQL, enabling downstream analytics and reporting.
- Improved data freshness from multi-day latency to less than 4 hours by building incremental upsert pipelines, optimizing Spark transformations, and validating outputs for reliable consumption.
- Reduced pipeline failure rates by 40% by orchestrating Airflow + Databricks Jobs with retries/SLAs/backfills and integrating AWS S3/Secrets Manager and Snowflake (stages, COPY INTO) while driving defect remediation.

Data Analyst
Recreation & Wellbeing, Madison, Wisconsin *June 2024 - August 2024*

- Built and deployed ETL workflows in SQL + Databricks notebooks, reducing manual data tasks by 40% and improving repeatability for analytics deliverables.
- Partnered with engineering and business stakeholders to deliver a Power BI dashboard used by 2,000+ stakeholders, supporting decision-making with validated metrics.
- Developed analytic data models for brand persona segmentation, improving campaign targeting effectiveness by 25%

Data Engineer Associate
Accenture, Mumbai, India *September 2021 - July 2023*

- Engineered end-to-end ingestion pipelines using Azure Data Factory and Databricks (PySpark) to process REST with transactional data into Delta tables (ADLS Gen2) supporting 100K records/day.
- Improved operational stability via monitoring/alerting and robust retry patterns (ADF & Azure Monitor), reducing repeat production incidents by 15%.
- Optimized Spark transformations and SQL models (Synapse/Azure SQL), improving analytics query performance by 10% and dashboard responsiveness

PROJECTS

- **Gen AI Analytics Co-Pilot**, Built a GenAI analytics co-pilot using LLM prompting + LangChain-style chains to translate natural-language questions into validated SQL over curated fact/dimension models for safe self-serve analytics. Implemented RAG (retrieval from a metric glossary/data catalog) with guardrails (read-only, complexity limits, semantic validation) and explainable outputs, reducing ad-hoc SQL requests.
- **Real-Time Stock Market Data Pipeline**, Built a real-time streaming data pipeline using Kafka and AWS to ingest and process 50K+ events/sec, landing data in an S3 data lake and enabling SQL analytics via Glue + Athena; added partitioning and checkpointing to ensure reliable, low-latency processing.

EDUCATION

University of Wisconsin Madison **Madison, Wisconsin, USA**
MS in Information *September 2023 - May 2025*
BMS College of Engineering **Bengaluru, Karnataka, India**
BE in Computer Science *August 2017 - August 2021*

EXTRACURRICULAR ACTIVITIES & CERTIFICATIONS

- **Graduate Data Club Lead** in Data Club - September 2023 - May 2025
- **Databricks Certified Generative AI Engineer Associate**
- **Databricks Certified Data Engineer Associate**