

Narayan Sahu

Data Engineer | Azure | Snowflake | Bangalore, India

Phone: 8456890730 | Email: narayansahu650@outlook.com | LinkedIn: linkedin.com/in/dataengineer-narayansahu | GitHub: github.com/Narayan-git

Summary

Data Engineer with 4+ years of experience in building scalable ETL pipelines, data warehousing, and reporting solutions using Azure Data Factory, SSIS, Python, SQL Server, Snowflake, and Power BI. Proven success in delivering mission-critical projects for UnitedHealth Group, including DSNP and DMEVS. Recognized for rapid learning, collaboration, and innovation with multiple performance awards for problem-solving, collaboration, and innovation. While my primary strength lies in data engineering, I have also contributed to AI-enabled workflows, including developing a RAG-based chatbot (Clinical Intelligence System) to support healthcare insights. Passionate about leveraging data to drive business insights and operational efficiency.

Technical Skills

Programming Language: Python, SQL

Cloud Platforms: Azure (Data Factory, Databricks, SQL Database, ADLS Gen2, Blob Storage)

Data Engineering: ETL/ELT, SSIS, Azure Data Factory, Airflow, Azure Databricks (PySpark)

Data Warehousing: Snowflake, Azure Synapse, Dimensional Modeling (Star & Snowflake Schema)

Databases: SQL Server, Azure SQL, Azure Cosmos DB, Oracle

Reporting & Analytics: Power BI, SSRS

Version Control & CI/CD: Git, Azure DevOps

Testing Tools: Pytest, SQL validation scripts

Job Scheduling: SQL Server Agent, ADF Triggers, Airflow

Certifications: Microsoft Azure Fundamentals (AZ-900)

Work Experience

Optum India (UnitedHealth Group), Bangalore, India

Mar 2024 - Present

Data Engineer

- Designed and optimized ETL pipelines using Azure Data Factory (ADF), SSIS, and Python, improving daily load performance by **30%**.
- Implemented data quality checks using Python (Pandas), SQL, and SSIS, increasing data reliability for downstream reporting by **95%**.
- Integrated and processed multi-source data from Snowflake, Azure Cosmos DB, and SQL Server, handling JSON, CSV, and Excel formats at a scale of over **1M records per day**.
- Reduced verification processing time by **80%** by rewriting legacy SSIS workflows and developing Python-based automation packages.
- Automated job orchestration using ADF triggers and SQL Server Agent, reducing manual intervention by **90%**.
- Created Power BI dashboards for monitoring pipeline failures, performance metrics, and SLA adherence.

Optum India (UnitedHealth Group), Bangalore, India

Jan 2022 - Feb 2024

Associate Data Engineer

- Developed ETL pipelines using SSIS and Python (Pandas, NumPy) to support state verification processes.
- Built complex SQL views, stored procedures, and triggers for business logic; improved query execution times by **25%** through indexing and optimization.
- Migrated on-prem workflows to Azure using ADF, ADLS Gen2, Azure SQL, and Logic Apps, improving scalability and reducing compute costs.
- Implemented data validation scripts and unit tests (Pytest + SQL) to ensure data consistency across environments.
- Collaborated in Agile/Scrum ceremonies and supported QA/UAT phases to ensure smooth releases.

Projects

DMEVS (Dual Special Needs Plan Medicaid Eligibility Verification System)

Optum India | Jan 2024 - Present

ADF • Python • SSIS • Power BI • SQL Server • Snowflake

- Developed secure ETL packages using SSIS and Python to reliably fetch and transfer data from SQL Server and Snowflake.
- Built complex SQL views, stored procedures, and triggers to process DSNP Medicaid eligibility data with high accuracy.
- Created Power BI dashboards to monitor ETL jobs, pipeline performance, and operational KPIs.
- Participated in Agile SCRUM ceremonies, requirement analysis, code reviews, and unit testing to ensure delivery quality.
- Supported Integration, QA, UAT, and Production environments for seamless project deployment.

Dual Special Needs Plan (DSNP)

Optum India | Feb 2022 - Dec 2023

SSIS • Python • SQL Server • Oracle • SSRS • Power BI

- Implemented ETL pipelines using SSIS and Python (Pandas, NumPy) to process state verification and Medicaid eligibility data.
- Built complex views, stored procedures, and triggers across SQL Server, Oracle, and Snowflake databases to meet business requirements.
- Developed SSRS and Power BI dashboards to monitor state verification workflows and job performance.
- Followed coding guidelines and best practices; supported Integration, QA, and UAT phases.

End-to-End SportsBar Order Data Pipeline (Databricks Lakehouse)

github.com/Narayan-git/DataEngineeringProjects

Azure • Databricks • Delta Lake • Apache Spark • Python • SQL

- Designed a production-grade **Medallion Architecture (Bronze–Silver–Gold)** pipeline in Databricks to unify multi-format order data from spreadsheets, WhatsApp exports, cloud drives, and APIs.
- Implemented scalable ingestion workflows using **Spark and Delta Lake**, enabling ACID transactions, schema enforcement, Change Data Feed (CDF), and daily incremental processing.
- Developed robust **data quality and standardization rules** including deduplication, multi-format date parsing, missing value handling, business rule validation, and reference data mapping.
- Engineered conformed **dimension and fact models** (customers, products, pricing, orders) and automated monthly aggregated sales metrics for downstream FMCG analytics.
- Orchestrated pipelines using **Databricks Workflows** and Azure Data Factory, implementing file archival, lineage tracking, and audit-friendly metadata features.
- Optimized pipeline performance with partitioning, clustering, and caching strategies, improving read latency and reducing compute cost across daily workloads.

RAG-Based Clinical Intelligence System

github.com/Narayan-git/DataEngineeringProjects

Azure OpenAI • RAG • Embeddings • BM25 • DeepEval • Python

- Built a **Retrieval-Augmented Generation (RAG)** pipeline leveraging a curated knowledge base of 100 NIH clinical documents as the exclusive evidence source for answering medical and genetic queries.
- Implemented **multi-strategy retrieval** including semantic search, hybrid retrieval, and BM25 ranking to ensure high recall and factual relevance across complex clinical topics.
- Integrated Azure OpenAI models (gpt-4o-mini_2024-07-18, text-embedding-3-small_1) using the 2025-01-01-preview API version with project isolation via PROJECT_ID.
- Engineered a **hallucination-resistant answer generation pipeline** by enforcing strict grounding to retrieved evidence chunks and restricting model outputs to verified clinical sources.
- Evaluated pipeline accuracy using **DeepEval** metrics (faithfulness, relevance, answer correctness), enabling measurable improvements in retrieval and generation reliability.
- Designed modular components for ingestion, chunking, embedding creation, vector storage, retrieval orchestration, and LLM-based answer synthesis to support scalable clinical intelligence applications.

Education

Master in Computer Application

2018 - 2021

Fakir Mohan University, Balasore

CGPA: 8.2/10

Bachelor of Computer Applications (BCA)

2015 - 2018

Berhampur University, Bhubaneswar

70%

Additional Highlights

- Strong understanding of SDLC, Agile/SCRUM, and Test-Driven Development (TDD)
- Excellent communication, problem-solving, and team collaboration skills
- Quick learner with a proactive approach to new technologies and challenges
- Certified through Optum AI Dojo program and recognized among the Super 30 by Optum's CTO.