NINAD N PATIL

United Kingdom | +44 7471566316 | pninad009@gmail.com | LinkedIn |GitHub

PROFESSIONAL SUMMARY

Data Engineer with nearly 4 years of experience who turns messy source feeds into fast, reliable analytics. At DRG, I built metadata-driven Azure Data Factory pipelines ingesting ~1M rows/week and cut daily refresh ~50% with monitored execution and alerting. I also improved PostgreSQL performance ~35% through partitioning and indexing, boosting report responsiveness. Outside work, my Databricks projects (Autoloader, Delta Live Tables, Unity Catalog) showcase production-ready Delta Lake patterns.

SKILLS

- **Cloud/Platform:** Azure (ADLS Gen2, Azure SQL Database, Key Vault), Databricks (DLT, Autoloader, Structured Streaming, Unity Catalog, SQL Warehouse)
- **Data Engineering:** ETL/ELT, CDC, SCD Types, Dimensional Modeling (Star/Snowflake), Delta Lake, Medallion Architecture, Data Quality (DLT Expectations), Lineage/Observability
- **Orchestration/Automation:** Azure Data Factory (pipelines, Mapping Data Flows, SHIR), Azure Logic Apps, GitHub Actions, Databricks Asset Bundles (CI/CD)
- **Programming/Query:** Python (Pandas, NumPy, PySpark, SQLAlchemy, Flask), SQL/T-SQL (CTEs, window functions, performance tuning), Bash, DAX, PostgreSQL, MySQL, REST/HTTP APIs, JSON/XML
- **Analytics/BI:** Power BI (DAX, RLS, Incremental Refresh, semantic models), Excel (Pivot Tables, VBA Automation), Grafana, Matplotlib, Seaborn
- **Data Acquisition**: REST APIs, BeautifulSoup, Selenium, Web Scraping, Managed Identity Authentication, Python Automation Scripts
- **Other:** Docker, Git/GitHub, Documentation & Runbooks

WORK HISTORY

Data Analyst and Engineer **May 2024 – Present**

The DRG, United Kingdom

- Designed and optimized robust, scalable database and **schema** models in collaboration with Finance and key stakeholders, ensuring **data integrity**, accessibility, and alignment with business needs.
- **Delivered** parameterized **ADF** pipelines processing **~1M+ rows/week** from EPOS, loyalty, and REST APIs; standardized schemas with **idempotent loads** and drift-tolerant ingestion.
- **Cut** daily refresh time **~50%** by **parallelizing** safe paths and removing redundant stages; reduced end-to-end runtime and increased on-time report readiness for Finance/Ops.
- **Migrated** MySQL to **Azure PostgreSQL** and **re-modeled** facts/dims with targeted **BTREE indexes** and date/location **partitioning**, reducing key report query times **~35%**.
- Automated daily data extraction from **RESTful APIs** and web scraping (**BeautifulSoup, Selenium**) and converting JSON, XML data into structured formats, improving database performance and usability.
- **Modeled** governed gold datasets for **Power BI** with **RLS** and **Incremental Refresh**, increased dashboard adoption and reduced refresh failures per week.
- **Partnered** with Marketing/Ops to analyze seasonality and run **A/B tests**; **translated** findings into campaign targeting and service planning.
- Implemented automated **Power BI** dashboards for real-time reporting and strategic decision-making, increasing quarterly revenue by 15% through enhanced sales insights.

Data Analyst **April 2021 - March 2022**

NSEIT, Nashik

- **Automated** SQL/Python ingestion and preprocessing with parameterized validations (types, duplicates, referential checks), **freeing ~10 hrs/week** and shortening release cycles.
- **Co-designed** a governed preprocessing framework with a **6-person** team (standardized schemas, data dictionary, access controls), **reducing rework** and handoffs across monthly releases.
- Automated recurring reconciliations and data-processing workflows, freeing **~10 hours/week** and shortening analysis turnaround for stakeholders.
- Generated comprehensive 20+ reports and dashboards to support informed decision-making by stakeholders.
- Monitored data integrity and quality, promptly addressing discrepancies and improving reporting reliability.

Data Analyst **March 2020 – March 2021**

Swami Vivekanand Education and Research Centre, **Mumbai**

- Integrated and standardized multi-source data with Python (Pandas/NumPy), cutting processing time by **~20%** and improving data reliability for downstream reporting.
- Built interactive **Power BI** dashboards for KPI tracking, enabling self-service analysis and faster decision cycles.
- Authored and optimized complex **SQL** to profile and segment customers, delivering a **~25%** uplift in decision-ready insights for stakeholders.
- Increased stakeholder engagement by **~30%** through targeted walkthroughs, clear documentation, and iterative dashboard improvements.

EDUCATION

- MSc in **Advance Computer Science with Data science (2:1)** **January 2023 – January 2024**
  University of Strathclyde**,** School of Computing Science
- Bachelor of **Computer Applications (2:1)** **May 2016 – March 2020**
  Savitribai Phule Pune university

PROJECTS

End-to-End Azure Data Engineering Lakehouse (GitHub Link)

- Medallion (Bronze/Silver/Gold) on **ADLS Gen2**; **ADF** pipelines are parameterized & metadata-driven with incremental loads, backfill controls, and CDC watermarking (JSON).
- **Databricks Autoloader + Structured Streaming** refine Bronze to Silver **Delta**; governance via **Unity Catalog** (metastore, external locations, access connector).
- **Delta Live Tables** curate Gold: **SCD2** dims (sequence_by) & **SCD1** upserted facts; **Expectations** + visual lineage DAG; analytics-ready for **SQL Warehouse/Power BI**.
- Ops & DevEx: Python utilities (cleaning, de-dupe), checkpointing & idempotent runs, empty-file cleanup, **Logic Apps** alerts, secrets via **Managed Identity/Key Vault**, CI/CD with **Databricks Asset Bundles** + GitHub

DataBricks Declarative Pipelines (GitHub Link)

- Built an end-to-end **Lakehouse** with **DLT** using a **Medallion** design, unifying **batch + streaming** with automatic DAG/lineage.
- Bronze: **streaming ingestion** with **Expectations** (warn/drop/fail) and **Append Flow** to consolidate multi-region sources; captured run-time **audit metrics** for observability.
- Silver: **Auto-CDC (Type-1)** upserts (keys + sequenceBy) with reusable Python utilities (casting, de-dup, common transforms) to keep curated tables incrementally accurate.
- Gold & Ops: **SCD2 dimensions** + Type-1 fact and a materialized business view (revenue by region×category); **Unity Catalog** governance, parameterized dev/prod configs, SQL seed scripts, and monitored alerts.

Azure Data Factory End-to-End (On-Prem + API + SQL): (GitHub Link)

- Built a hybrid ingestion layer in ADF with **Self-Hosted IR** for on-prem file shares, **HTTP/REST** for GitHub raw JSON, and Azure SQL incremental loads without watermark tables (state kept in **JSON** in the lake) enabling **idempotent re-runs** and drift-tolerant landings
- Orchestrated a Bronze, Silver, Gold flow: raw landings to Bronze, **Mapping Data Flows** to standardize/clean/type-cast into **Silver Delta** with **Alter Row** upserts on business keys, then curated **Gold** views (joins, aggregations, **dense rank, Top N**) optimized for BI.
- Centralized control via a **Parent pipeline** that chains on-prem, API, and SQL paths, parameterizes file lists/batch sizes, and runs safe steps in parallel reducing manual coordination and making backfills/replays predictable
- Shipped with **ARM publish artifacts** and documented runbooks for **repeatable CI/CD**, plus clear guardrails on schema drift vs. validation, partitioning options, and overwrite strategy for Gold snapshots.

Dockerized End-to-End ETL Pipeline with PostgreSQL, ChromaDB & Sentence-Transformers:

- Developed a Python-based ETL pipeline to extract and clean UK legislation Webpages into structured JSON.
- Stored data in PostgreSQL, generated embeddings with sentence-transformers, and indexed for vector search using ChromaDB.
- Built a CLI for semantic querying, retrieving relevant documents from ChromaDB.
- Containerized the complete solution (PostgreSQL, ChromaDB, ETL, CLI) with Docker for streamlined deployment.