Dhananjay G.

MSc Computer Science, University of Glasgow |ETL/ELT Pipeline Design| 4 years of experience | Data Engineer
London, UK | +44 7393568854 | guptadhananjay090@gmail.com | LinkedIn | GitHub

PROFESSIONAL SUMMARY

Google Cloud Certified Data Engineer with expertise in building **cloud-native architecture** across **Azure.** Known for excellence in **ETL , Python, SQL,** and **Pyspark,** delivering impact such as a **£1M+ revenue uplift,** and **40% faster support resolution** using GenAI tools. Led a **migration of 50M+ records,** improving data cleaning from **80% to 99.9%.** Received special letters of recommendation from industry leaders and academics for excellence in managing data.

RELEVANT SKILLS

- **Functional Competencies:** Data Modelling, Data Pipeline Development, Data Quality Improvement, Data Governance & Security Compliance, Mentoring & Coaching, Cross-Functional Collaboration, Stakeholder Engagement, Data Culture Advocacy, Continuous Learning
- **Technical Tools**:
    - **Programming:** Python (pandas, Numpy), SQL, PySpark, Bash
    - **Data Engineering Frameworks:** Apache Spark, SparkSQL, Airflow
    - **Cloud Platforms:** Azure (ADLS Gen2, Data Factory, Databricks,Analytics), GCP (BigQuery, Cloud Storage)
    - **GenAI Tools & Frameworks:** OpenAI, Gemini, LangGraph, RAG Architecture, Tokenisation
    - **Development & DevOps:** Git, GitLab, Docker, JIRA, Azure DevOps, CI/CD, Github
    - **Databases:** MS SQL (T-SQL), Vector database(Qdrant,Pinecone), PostGRES, NoSQL

WORK EXPERIENCE

Data Engineer, Agilitics   *Jan 2025 - Present*

- **Built and maintained an automated ELT pipeline for a client using Azure Data Factory and SQL Database,** ingesting over 3M+ finance records from an on-premises Oracle system into a cloud-based Azure data platform, supporting incremental loads (medallion architecture) in the staging layer.
- Engineered transformation and change-data capture (CDC) logic using T-SQL and MD5 hashing within Azure SQL and **Delta Lake (parquet format) on Azure Databricks,** ensuring high data fidelity across revenue and billing datasets, and enabling accurate, audit-ready finance reporting at scale.
- **Delivered a production-grade finance data mart and Power BI reporting solution** by modeling segmented customer- and company code-level metrics, powering dashboards for active customer counts, average rate per unit, and month-over-month general ledger trends, supporting financial planning and analysis across 20+ regional business units.
- **Designed and deployed a scalable retrieval-augmented generation (RAG) chatbot (FastAPI)** by leveraging **LangChain, OpenAI, and Qdrant,** enabling over 500 users to query internal documents (unstructured data) in natural language, **reducing support ticket resolution time by 40%.**

Data Engineer, Open Insight   *Apr 2022 - Jun 2023*

- **Designed and implemented 40+ business-critical KPIs** by analysing **400K user records for 2nd largest telecoms in Thailand** using **PySpark and Azure data lake storage(ADLS Gen2)** on **Azure Databricks,** directly informing targeting strategies that **generated £500K in incremental revenue.**
- **Boosted pipeline performance by 50%** by redesigning **data processing logic,** optimising joins, and applying strategic **caching** within Azure Databricks, reducing overall runtime from **35+ hours to under 12.**
- **Built automated, production-grade ETL workflows** that ingested data from **15+ complex sources,** applying **SQL (T-SQL) and Delta Lake transformations** to ensure scalability, data consistency, and **60% faster processing speeds.**
- **Developed end-to-end PySpark pipelines** to process over **400K telecom customer records,** enabling real-time KPI tracking, web (socials) and app customer insight, and **marketing dashboard refreshes via Power BI** with significantly improved timeliness.
- **Integrated KPI pipelines into GitLab CI workflows,** enabling continuous deployment and automated table refreshes in production; also **mentored two interns,** accelerating their upskilling and helping ensure consistent,

on-time delivery across sprints.

Associate Analyst, Mastech Infotrellis *Sep 2020 - Mar 2022*

- **Improved data accuracy from 80% to 99.9%** by **leading the end-to-end migration for DELL technologies of 50M+ records** (customer, address, credit) from Oracle databases (structured data) to a unified global system across EMEA, APJ, and LATAM (data warehouse), using robust validation and reconciliation methods.
- **Reduced batch processing time by 50% (4hrs to 2hrs)** by **automating ETL pipelines in Python and SQL(T-SQL)** with **REST API integration, flag-based incremental logic,** and **parallelised regional workflows** that scaled efficiently across large datasets.
- **Minimised post-migration issues and ensured schema compliance** by **designing transformation workflows** with clearly defined **business rules, data cleaning checks, and unit testing** integrated into **GitLab CI/CD** using **pytest.**
- **Resolved ad-hoc data issues and unblocked critical processes** by **collaborating closely with developers, testers, and analysts,** proactively addressing tasks like **custom customer data seeding** and migration bottleneck triage.
- **Accelerated environment rollout by 30%** by **managing sprint deliverables in Azure DevOps within an Agile setup,** aligning technical teams and stakeholders collaboration through structured planning, task tracking, and cross-functional coordination while creating technical documentation.

PROJECTS

- **LLM-Based Data Quality Analyser** *(GenAI, FastAPI, OpenAI, Hugging Face, Streamlit, AWS bedrock):*

Built a GenAI tool to analyse CSV and Parquet files stored in Amazon S3, detecting nulls, outliers, and structural anomalies using FastAPI and Pandas. Enabled **70% faster QA checks** by generating natural language summaries and automated fixes through a user-friendly Streamlit UI, with exportable reports for business teams.

- **Customer360** *(Azure Data Factory, Azure Databricks, ADLS Gen2, PySpark, SQL, Key Vault)*

Built an **event-triggered pipeline** to validate and route incoming CSV order files using ADF and Databricks, securely managing credentials via Azure Key Vault. Automated the ingestion and transformation of multi-source datasets using PySpark and SQL, generating **clean customer-level outputs** that supported advanced analytics and decision-making.

EDUCATION

MSc. Computer Science, University of Glasgow *Sep 2023 - Sep 2024*
Highlight: Developed scalable data pipelines and backend systems as part of academic projects, including a cloud-based ETL solution using Python, SQL, and Azure that processed 5M+ records, supporting real-time analytics use cases.

B.Tech. Computer Science, SRM University *Aug 2016 - Aug 2020*
Highlight: Built a mini data warehouse using MySQL and Python for a student performance tracking system, streamlining grade analysis and reducing report generation time by 60% for academic staff.

CERTIFICATIONS & RECOGNITIONS

- **Google Cloud Professional Data Engineer,** Issued by Google Cloud, July 2025
- **Database and SQL for Data Science,** Issued by IBM, 2024
- **Data Analysis with Python,** Issued by IBM, 2024
- **GitLab Certified Associate,** Issued by GitLab, 2024
- Special Letter of Recommendation Received from Dr. Megha, Professor, SRM University
- Special Letter of Recommendation Received from Dr. Tarun Yadav, Researcher, DRDO (Gov. of India)
- Special LinkedIn Recommendation Received from Louis Sebastian, Principal SW Engineer, DELL Technologies