Neeraj Kumar Mn

Work permit: Germany | **Phone:** (+49) 1774623827 (Mobile) | **Email address:** neerajde2000@gmail.com | **Portfolio:** Portfolio | **LinkedIn:** LinkedIn | **Github:** Github

ABOUT MYSELF

Data Engineer with around two years of experience developing and optimizing scalable ETL/ELT pipelines. Proficient in distributed data processing with PySpark and transforming complex data payloads into production-ready analytical structures.

SKILLS

Programming & Databases
Python (Pandas, NumPy), SQL, PostgreSQL, MySQL, MongoDB, Snowflake

Data Engineering & Big Data
Apache Spark, Hadoop, Kafka, ETL Pipelines

Cloud Platforms
Azure: Data Factory (ADF), ADLS Gen2, Synapse Analytics, Databricks, Microsoft Fabric | AWS: S3, Lambda, Redshift, Glue

Orchestration & Version Control
Apache Airflow, Docker, Git

Data Visualization
Power BI

WORK EXPERIENCE

3 JUN 2024 - 30 MAY 2025 - BANGALORE, INDIA

JUNIOR DATA ENGINEER (FULL-TIME) UPCONNECT LABS LLP-BANGALORE

- Optimized analytical query performance by 35% by refactoring complex SQL joins and indexing, significantly reducing database latency for cross-functional teams.
- Streamlined executive reporting workflows via Power BI dashboards, eliminating 40% of manual data gathering efforts for business stakeholders.
- Stabilized production pipelines by debugging complex PySpark failures, collaborating with team members to guarantee 99.9% data reliability for analytics.

1 MAY 2023 - 14 FEB 2024 - BANGALORE , INDIA

DATA SCIENTIST INTERN AI VARIANT

- Developed secure OAuth 2.0 authentication workflows with the Spotify API via Python's requests and base64 libraries to programmatically acquire system access tokens.
- Pattern Extraction (Regex): Isolated domain-specific references from 20,000 text entries to enhance structural data availability
- Boosted text processing speeds by 40% by deploying spaCy and NumPy for advanced tokenization, lemmatization, and stopword elimination.

PROJECTS

Olist Brazilian E-commerce Analytics Pipeline - E commerce / Data Engineering / Analytics [Azure Data Factory, Azure Databricks (PySpark, Spark SQL), Azure Data Lake Storage Gen2, Azure Synapse Analytics, Python, SQL, MongoDB]

- Orchestrated scheduled ingestions of 100K+ records from MySQL/MongoDB into ADLS Gen2 using ADF to ensure continuous data availability.
- Minimized manual integration by 95% by designing dynamic, parameterized ADF pipelines with Lookup and ForEach activities.
- Elevated data accuracy by 30% through rigorous PySpark deduplication, null-handling, and schema standardization in Azure Databricks.

Link github.com/Neerajmn28/BigDataProject1-Brazilian-Ecommerce

CarePlus Ticket Analytics Pipeline with AWS - Healthcare Analytics [AWS S3, AWS Lambda, AWS Glue, AWS Redshift, AWS Athena]

- Automated incremental data loads via S3 event triggers and Lambda, cutting overall ETL runtime by 50% while completely preventing duplicate records.
- Developed robust AWS Glue ETL jobs for dynamic schema updates and SQL transformations, boosting downstream data quality by 35%.
- Optimized data warehouse performance by converting raw datasets into Parquet format before loading into Amazon Redshift, significantly lowering storage costs.

Link https://github.com/Neerajmn28/Careplus-ETL-Pipeline-with-AWS

NYC Taxi Data Pipeline – Lakehouse Pipeline - Transportation / Public Data Analytics [Azure Data Factory, ADLS Gen2, Databricks, PySpark, Delta Lake, SQL, Python, REST APIs]

- Extracted 10K+ monthly records via API ingestion into ADLS Gen2 and implemented Azure Data Factory pipelines with parameterized datasets, loops, and conditional logic, reducing manual effort by 90%.||
- Processed and transformed data using Databricks + PySpark in a Bronze–Silver–Gold lakehouse, improving reliability for downstream analytics.
- Stored structured, cleaned datasets in Delta Lake tables, ensuring data readiness and accessibility for reporting and business insights.

LANGUAGE SKILLS

Other language(s): **ENGLISH C1** | **GERMAN B2**

EDUCATION & TRAINING

24 Feb 2025 - 24 Jun 2026 - BERLIN, GERMANY

MASTERS IN DATA SCIENCE- ARDEN UNIVERSITY

Level in EQF: 7