



# Suparna Chakraborty

 +91 9051198126 |  chakrabortysuparna07@gmail.com

[LinkedIN](#) | [Github](#)

## Professional Summary

Diligent and results-driven **Data Engineer** with extensive experience in **data migration, ETL development, and cloud-based data processing**. Proven expertise in **workflow automation** using Apache Airflow and PySpark. Skilled in data conversion across multiple systems, optimizing ETL pipelines, and ensuring **high-performance data operations**. A collaborative team player with strong problem-solving skills and a focus on delivering **scalable and efficient data workflows**.

## Technical Skills

- **SQL:** PostgreSQL, Databricks SQL, MySQL
- **Programming & Libraries:** Python (PySpark, SQLAlchemy, psycopg2, NumPy, Pandas, Matplotlib)
- **Cloud & Data Engineering:** Azure Data Engineering, Azure Data Fundamentals, Databricks
- **ETL & Workflow Orchestration:** Apache Airflow, SnapLogic, SAP BODS
- **Other Tools:** Advanced Excel, JavaScript, SAP, Oracle, Git & GitHub

## Certifications

- Microsoft Certified: **Azure Data Engineer Associate**
- Microsoft Certified: **Azure Fundamentals**
- Microsoft Certified: **Azure Data Fundamental**
- LinkedIn Learning: **PySpark, Azure Data Engineering, Azure Data Fundamentals, Airflow , Python , AWS**

## Projects

### ETL Pipeline Automation with Apache Airflow

- Designed and implemented an **end-to-end ETL pipeline** using Apache Airflow for workflow orchestration.
- Developed **PySpark scripts** for efficient data transformation and automated **CSV data ingestion** into SQLite.
- Configured **Airflow DAGs** for scheduling, monitoring, and error handling of data workflows.
- Utilized **Bash and Python** for post-processing and seamless integration tasks.
- Version-controlled the project using **Git and GitHub**, ensuring collaboration and code tracking.

## ETL Pipeline with Medallion Architecture for Real-Time IoT Data Processing

- Designed and implemented an **ETL pipeline using Azure IoT Hub** for real-time data ingestion from Raspberry Pi Azure IoT Online Simulator.
- Stored **raw data** in **Azure Data Lake** following **Medallion Architecture** (Bronze, Silver, Gold layers).
- Processed and transformed streaming data in **Databricks (PySpark)**, applying **data cleansing, aggregation, and enrichment**.
- Integrated the pipeline with **Azure Synapse Analytics** for scalable and optimized querying.

## Ad-hoc Activity on PySpark and Databricks

- Configured **PySpark in standalone mode** and worked on JSON data using **Databricks Community Edition**.
- Mounted Databricks notebooks with **Azure Storage Account** and worked with **Spark DataFrame API** and **Spark SQL**.
- Implemented schema parsing using **StructType**, UDFs, and `explode_outer` for JSON data cleansing.
- Connected with **Azure SQL DB** and **Cosmos DB**, performing **upsert operations and streaming data handling**.

## Experience

### PWC AC Bangalore – Data Engineer (*March 2022 – Present*)

- Worked on multiple **data migration projects** for prominent US-based clients.
- Designed **ETL solutions**, developed transformation logic, and validated data integrity in migration processes.
- Implemented **data verification scripts** to ensure completeness and accuracy of migrated datasets.
- Collaborated with stakeholders to optimize **data engineering pipelines** for better performance.

### PwC (IN) - Contractor – ETL Developer (*March 2021 – December 2021*)

- Developed **data pipelines** ensuring accurate ETL processes for US-based clients.
- Designed extract-transform approaches and managed **reconciliation of production data loads**.
- Implemented **performance tuning techniques** for optimized data processing workflows.

### Optimize IT Systems – Trainee (*July 2017 – December 2017*)

- Received training in **Java, Android, JavaScript, and SnapLogic**.

## Education

- **B.Sc. in Computer Science** – Calcutta University
- **M.Sc. in Computer Science** – Calcutta University