

Gaurav Thombal

+91 93250 59457 | thombalgaurav10@gmail.com | LinkedIn | GitHub | HackerRank | Pune, India

PROFESSIONAL SUMMARY

Aspiring Data Engineer with hands-on experience in building end-to-end ETL pipelines using Python, PySpark, Azure Data Factory, and Azure Data Lake Storage Gen2. Skilled in data ingestion, transformation, data quality, and implementing the Bronze–Silver–Gold Medallion Architecture. Experienced in working with REST APIs, SQL, Power BI, and GitHub through real-world portfolio projects. Passionate about building scalable data pipelines and eager to contribute to cloud-based data engineering solutions.

TECHNICAL SKILLS

Programming: Python, SQL **Big Data:** Apache Spark, PySpark **Databases:** SQL Server, MySQL, PostgreSQL
Cloud & Azure: Azure Data Factory (ADF), Azure Data Lake Storage Gen2, Azure Databricks, Azure Synapse Analytics, Microsoft Azure
Visualization: Power BI, Tableau **Version Control:** Git, GitHub
Concepts: ETL, ELT, Data Warehousing, Data Modeling, CDC, Medallion Architecture, Incremental Loading

PROJECTS

Amazon E-commerce Azure ETL Pipeline | [GitHub](#)

Tech Stack : Azure Data Factory, Azure Data Lake Storage Gen2, PySpark, Python, Google Colab, GitHub

- Designed and implemented an end-to-end ETL pipeline following the Medallion Architecture (Bronze–Silver–Gold) using Azure services and PySpark.
- Ingested Amazon e-commerce data from a GitHub repository into the Bronze layer using Azure Data Factory and stored it in Azure Data Lake Storage Gen2.
- Built a PySpark transformation pipeline in Google Colab to clean and standardize data by handling null values, removing duplicates, trimming whitespace, correcting data types, and validating records.
- Created multiple Gold layer business-ready datasets, including Daily Sales, Category Sales, Product Sales, Customer Sales, User Activity, and Category Daily Sales for analytics and reporting.
- Implemented a scalable data lake architecture that separates raw, cleaned, and aggregated data for downstream analytical workloads

PySpark ETL API Pipeline | [GitHub](#)

Tech Stack : Python, PySpark, REST API, Pandas, Parquet, Google Colab, GitHub

- Developed an end-to-end ETL pipeline using PySpark to extract data from a public REST API and transform it into an analytics-ready format.
- Implemented data cleaning, schema validation, null handling, duplicate removal, and data type conversion using PySpark DataFrame transformations.
- Stored transformed datasets in Parquet format to improve storage efficiency and query performance.
- Designed a modular ETL workflow following the Extract → Transform → Load (ETL) architecture with reusable pipeline components.
- Demonstrated large-scale data processing concepts using Apache Spark and PySpark.

Customer Segmentation & RFM Analysis Dashboard | [GitHub](#)

Tech Stack : Google BigQuery (SQL) | Power BI | Data Modeling

- Built an end-to-end Customer Segmentation Dashboard using the RFM (Recency, Frequency, Monetary) model to analyze customer purchasing behavior and identify high-value customer segments.
- Consolidated 12 monthly sales datasets into a unified dataset using SQL (UNION ALL) and implemented RFM scoring with window functions (NTILE, ROW_NUMBER) in Google BigQuery.
- Classified customers into business segments including Champions, Loyal Customers, Potential Loyalists, At Risk, and Lost Customers using SQL-based scoring logic.
- Developed an interactive Power BI dashboard to visualize customer segmentation, RFM metrics, and key business insights, enabling targeted marketing and customer retention strategies.

EDUCATION

Marathwada Mitra Mandal's College of Engineering, Pune

Bachelor of Engineering – Electronics & Telecommunication

Oct 2024 – Present

CGPA: 9.00/10

MIT Polytechnic, Pune

Diploma – Information Technology

Sep 2021 – Sep 2023

Percentage: 81.38%

CERTIFICATIONS

- Commonwealth Bank Introduction to Data Science Job Simulation** – Forage *Jan 2026*
- Deloitte Australia Data Analytics Job Simulation** – Forage *Feb 2026*
- EY Technology Risk Virtual Job Simulation** – Forage *Mar 2026*