

+91 8586959190
Noida, Uttar Pradesh, India
i.rahul.professional@gmail.com

Rahul Singh

Entry-Level Data Engineer

github.com/rahuldev
linkedin.com/in/rahul-singh1104

Aspiring Big Data Engineer with strong computer science fundamentals and hands-on experience building data pipelines using modern big data technologies. Skilled in designing **ETL/ELT** workflows, data ingestion, incremental loading, and batch processing on cloud computing platforms using **Apache Spark**, **PySpark**, **Hadoop**, **SQL**, and **Python** to support **analytics** teams.

SKILLS

Programming & Querying	Python, PySpark, SQL (Joins, CTEs, Window Functions), NoSql
Big Data & Processing	Apache Hadoop(HDFS (Distributed Storage), MapReduce (Batch Processing), Hive (SQL on Hadoop)) Apache Spark(PySpark, Databricks),
Cloud & Data Engineering	Azure Data Factory (ADF), Azure Data Lake Storage Gen2 (ADLS), Azure Synapse Analytics, REST APIs (HTTP ingestion), Azure Virtual Machines, Amazon Web Services (Working Knowledge) (EC2, EMR, S3), Google Cloud Platform (Working Knowledge) (Google Cloud Storage (GCS), Dataproc)
Data Architecture	Data Lakehouse (Bronze, Silver, Gold), Fact & Dimension Modeling, ETL / ELT Pipelines,
Data Formats	CSV, JSON, Parquet,
Tools	Git, GitHub, Linux (Basics),
Additional Skills	JavaScript (ES6+, async/await, promises), (Intermedate), Node.js (REST API development, middle-ware), React.js(hooks, component-based architecture),

PROJECTS

END-TO-END DATA ENGINEERING PIPELINE (E-COMMERCE DOMAIN)/PROJECT 01

Tech Stack: Azure Data Factory, ADLS Gen2, Databricks, PySpark, Synapse, SQL

- Designed **Bronze → Silver → Gold Lakehouse architecture** using Azure.
- Ingested **structured and unstructured data** from multiple sources including:
 - GitHub (CSV files), REST APIs (HTTP source),
 - MySQL database (https://filess.io), Azure SQL Database
- Designed and managed **data ingestion pipelines** using **ADF pipelines with Lookup + ForEach activities**, supporting scalable **pipeline design**.
- Implemented **data transformations and processing** in PySpark ensuring **data quality, consistency, and reliability**.
- Created curated **Silver layer** and **fact/dimension tables** in **Gold layer** using proper **data modeling**.
- Integrated **Azure Synapse Analytics** with ADLS Gen2 to support **reporting and analytics**.
- Version-controlled entire project using **GitHub**.
- Collaborated to design and maintain reliable data pipelines, ensuring data consistency and supporting reporting and analytics.

INCREMENTAL LOADING DATA PIPELINE USING WATERMARK STRATEGY/PROJECT 02

Tech Stack: SQL Database, Azure Data Factory / Python / Databricks, PySpark, ADLS

- Simulated an OLTP database and designed **incremental data pipelines** events using a **watermark-based CDC approach**.
- Extracted only **new and updated records** to optimize **data processing** and maintain **consistency**.
- Loaded raw incremental data into **Bronze layer**
- Performed data **cleansing, deduplication, and transformations** in **Silver layer** using PySpark. **analytics-ready datasets and fact tables** in **Gold layer** for reporting.
- Ensured **idempotent execution** and **pipeline reliability** across runs.
- Managed and supported end-to-end pipeline execution, assisting in monitoring and ensuring reliable data processing.

EDUCATION

Master of Compute Application(MCA), Amity University Online, 2023 - 2025

8.83 cgpa

Bachelor of Compute Application(BCA), ITS Mohan Nagar GZB 2020 - 2023

8.68 cgpa

CERTIFICATE

ShapemySkills - AWS Cloud computing certificate

IBM Python for Data Science Certificate

Data Engineering

Udemy: Master Big Data with Hadoop, Spark, Kafka & Cloud | Build Real-World Projects & Scalable Data Pipelines from Scratch

Web / Programming (Additional)

Udemy: The modern JavaScript course for everyone! Master JavaScript with projects, challenges and theory. Many courses in one!

Udemy: Master modern React from beginner to advanced! Next.js, Context API, React Query, Redux, Tailwind, advanced patterns