Abhay Singh *Data Engineer*

writetoabhaysingh21@gmail.com +91 9910356085 LinkedIn GitHub Gurugram, India

WORK EXPERIENCE

Data Engineer, *Xebia IT Architects* 02/2023 – Present | Gurugram, India

Key Development Projects:

1. Sales Promotion Data Migration (Scala & Spark)

- Designed and implemented a new **Scala-based ingestion framework** to migrate promotion sales data from a legacy JDBC system to an **SFTP-driven ingestion pipeline**.
- Built **reconciliation logic in Databricks using Spark and Delta Lake** to align schema, PKs, and business logic across systems.
- Developed **historical data backfill and comparison jobs** to validate 50M+ records, ensuring zero duplication and full referential integrity.
- Resolved **data mismatches** arising from divergent transformation logic and schema definitions between legacy and new platforms.

2. Legacy OPD Data Modernization

- Migrated multiple reporting datasets dependent on OPD legacy tables to newer standardized structures, improving **data accuracy, query performance, and maintainability**.
- Built **modular PySpark ingestion and transformation modules** supporting incremental updates and schema evolution.
- Enhanced **data mart reliability** by refactoring joins, harmonizing business rules, and eliminating redundant pipelines.

3. ADF Pipeline Update Automation

- Engineered internal automation tools that **reduced ADF release cycle time by 30%**.
- **Automated** Azure Data Factory pipeline updates by developing a Python tool that eliminated manual JAR deployments, reducing update time by **90%**.
- **Integrated** Azure DevOps REST APIs to auto-manage Git branches, modify JSON files, and raise pull requests, streamlining CI/CD operations.
- **Implemented** validation checks to prevent redundancy and improve deployment accuracy, leading to a **25% reduction in release bugs**.

Broader Contributions:

- **Refactored and optimized ETL pipelines** using Databricks, Apache Spark, and Azure Data Factory — improving performance by **25%** and cutting compute costs.
- Engineered internal email automation tool that **cut manual intervention by 16+ engineer hours/week**.
- **Optimized Azure cloud infrastructure** and Spark cluster utilization, leading to **$848K in cost savings**.
- Partnered with cross-functional teams to **enhance data platform reliability by 15%**, reducing downtime and data latency.

SKILLS: SQL, **Python**, **Scala, PySpark**, **Apache Spark**(Core, SQL, Streaming), **Apache Kafka, Databricks**, **Apache Airflow, MS Azure**(ADF, Databricks, Azure SQL, Azure DevOps)

EDUCATION

B.Tech. CSE specialization in AIML, *Sushant University* 07/2019 – 06/2023 | Gurgaon, India

PERSONAL PROJECTS

Real-Time Flight Monitoring System

- **Developed** a real-time data pipeline using Apache Spark and Delta Lake to ingest and process **1M+** daily flight records from the OpenSky API.
- **Designed** a modular Bronze–Silver–Gold data architecture with transformation logic including cleaning, deduplication, and enrichment.
- **Built** and scheduled **3 Airflow DAGs** to orchestrate ingestion (streaming), processing, and aggregation workflows **independently**.
- **Implemented** aggregations on enriched data to compute KPIs such as average velocity, altitude, and flight phase distributions every 5 minutes.

CERTIFICATIONS

Databricks Certified Data Engineer Associate
11/2023 - 11/2025

Databricks Lakehouse Fundamentals
01/2025 - 01/2026

AWARDS AND RECOGNITIONS

Galaxy Best Project Team (Tauras) Award, *Xebia IT Architects*

GEM Award (3x), *Xebia IT Architects*