Advait Khawase

https://advaitkhawase.vercel.app | advaitkhawase15@gmail.com | https://github.com/advaitkhawase15
https://www.linkedin.com/in/advait-khawase/| +91 7974983195

Summary

Results-driven Data Engineer with proven expertise in designing scalable data pipelines, real-time streaming solutions, and workflow automation across cloud and big data platforms. Demonstrated success in optimizing Spark-based ETL, deploying cloud-native architectures (AWS, Azure), and automating large-scale data integration for enterprise environments. Skilled in Apache Spark, Airflow, Kafka, and Databricks, with a track record of improving process efficiency and reliability through quantifiable impact on business operations.

Technical Skills

Big Data: Apache Airflow, Apache Spark, Apache Kafka, Hadoop, Presto, Databricks
Cloud: AWS Cloud(EC2, S3, AWS GLUE, Redshift), Azure(ADL, ADF)
Languages: C++, Python, PySpark, Shell Scripting, SQL
Databases: MySQL, Oracle, SAP HANA
Technologies & Tools: GitHub, Git, Informatica Powercenter, VS Code

Experience

Jio Platform Limited (JPL) Dec 2023 - Present
Data Engineer

- Developed 300+ data pipelines managing 6000+ tables in Databricks during cloud migration, collaborating with cross-functional teams to implement Medallion architecture for loading raw data into bronze and silver layers.
- Worked with Partner Center team for 3 months to develop 50+ Data pipelines handling over 1M+ records using Databricks and Informatica, achieving 15% improvement in ETL efficiency with zero downtime and quick response through email notifications.
- Set up Azure Data Factory during cloud migration and built 300+ pipelines for different teams and reusable templates, improving data flow efficiency and cutting development time by 80%.
- Engineered complex SQL queries for seamless integration into Informatica and ADF pipelines, resulting in an improvement in data transformation performance.
- Developed 100+ custom shell and Python scripts to automate the processing of Hive tables and Informatica jobs.
- Administered Apache Airflow, managing 500+ DAGs by configuring variables and resolving DAG issues, proactively resolved 200+ Informatica job failures, ensuring minimal pipeline downtime and quick response through failure email notification.

Projects

DocFlow - Unstructured Document Pipline for Jio, Retail and ENM

- Developed scalable data pipelines using shell scripting, Airflow, Azure Data Factory (ADF), and Databricks to ingest and process 10,000+ daily user files (PDFs, DOCXs, JPGs) from NAS storage into Databricks, enabling efficient data handling for Jio, Retail, and ENM teams.
- Engineered incremental file detection logic in Airflow and ADF workflows to identify and transfer new user documents (e.g., Driving licenses, offer letters) directly to Azure Data Lake Storage (ADLS), minimizing redundant processing.
- Orchestrated Databricks jobs to load files from ADLS into tailored, business-specific directory structures, facilitating seamless integration.

Import Clearance Data Pipeline

- Built a unified import-clearance datasource using Databricks for FCA team, combining and reshaping info from over 50+ source tables (with 10k to 1M+ records each), covering purchase orders, shipments, receipts, transport, customs, and duty calculations.
- Optimised runtime to around 1 hour on a modest Spark cluster via Parquet+Snappy, targeted repartitioning to control shuffle, selective caching and small table broadcasting.
- Automated the ETL pipeline using Databricks Workflows to enable reliable daily executions for loading processed data into the gold layer, incorporating automated failure alerts for timely issue resolution and operational continuity.

Education

Lakshmi Narain College of Technology and Science May 2019–May 2023
B.Tech in Computer Science and Engineering *CGPA: 8.7/10.0*

Achievements

GeeksForGeeks: Problem Solved - 300+ **Leetcode:** Problem Solved - 400+

AWS Certified Cloud Practitioner