Advait Khawase

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

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 Dec 2023 - Present

Data Engineer

- Developed a PySpark script for the Import Clearance, handling 40+ tables with 10k+ to 1Cr+ records per table. Optimised performance to complete execution in 1 hour while minimising Spark resource consumption.
- Worked with Partner Centre team for 3 months to develop 50+ Data pipelines through Spark/Hive/Airflow and Informatica PowerCenter, achieving 15% improvement in ETL efficiency with zero downtime and quick response through email notifications.
- Set up Azure Data Factory during cloud migration; built 100+ pipelines and 10+ reusable templates, improving data flow efficiency and cutting development time by 80%.
- Engineered complex SQL queries for seamless integration into Informatica PowerCenter and ADF piplines, resulting in improvement in data transformation performance.
- Developed 50+ custom Linux shell 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

HRDocFlow - Unstructured Document Pipline for Jio, Retail and ENM

- Created Informatica and Shell-based workflow to track, filter and move thousands of daily HR files(PDFs, DOCXs, JPGs) from NAS to HDFS across Jio, Retail and ENM teams.
- Scripted Airflow jobs to identify key HR documents by document ID(e.g., marksheets, offer letters) and transfer only the required files into Azure storage for further use.
- Used Databricks to load files from ADLs into business-wise volumes by processing D-3 data, ensuring no files are missed due to weekend or delay issues.

Real-Time Data Streaming Pipeline

- Designed and implemented a real-time data pipeline for new user tracking using Apache Airflow, Kafka, Spark, and Cassandra, with all components containerised using Docker.
- Automated data ingestion via a scheduled Airflow DAG, which pulled new user data from an API, published it to a Kafka topic, then processed and streamed it into Cassandra using Apache Spark with < 10sec latency.
- Managed Airflow metadata with PostgreSQL and monitored Kafka message flow and topic health using the Kafka Control Center.

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