APACHE NIFI & DATABRICKS: ORCHESTRATING DATA FOR STREAMING, ETL, AND AUTOMATION

I've been working with Apache NiFi for a while now, leveraging its capabilities to orchestrate and automate complex workflows. It's a powerful open-source orchestrator that is both cost-effective and highly flexible. With its intuitive visual interface, NiFi proves to be a powerful tool for streaming data, ETL processes, and workflow automation.

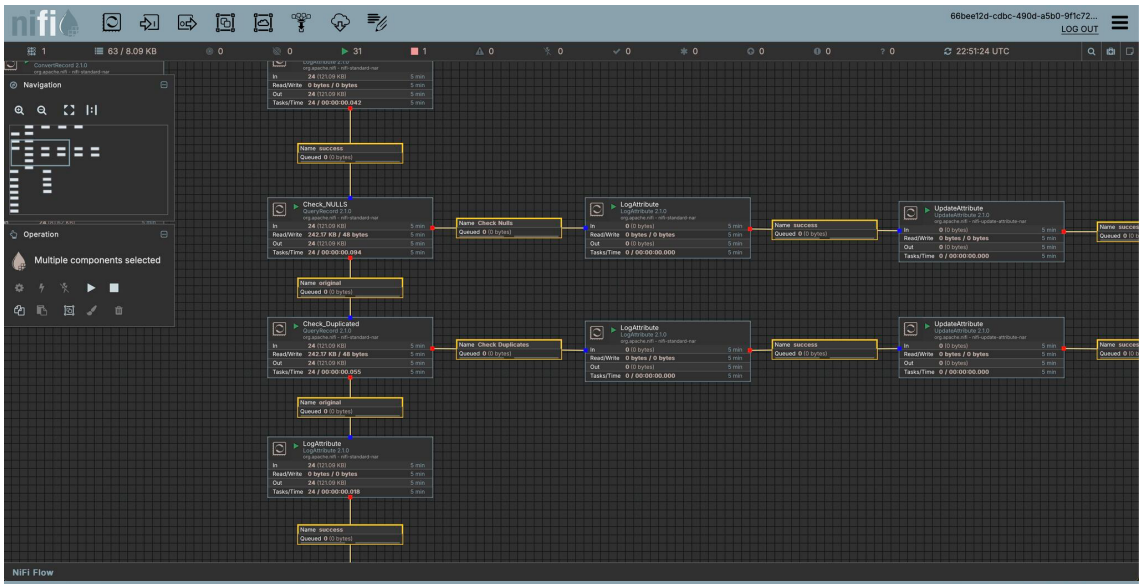
Here's an example of my latest project, integrating Kafka, Databricks, and Azure ADLS to efficiently process streaming data:
1. Connecting to Kafka – Ingesting real-time data streams.
2. Creating email alerts – Triggering notifications when the stream stops.
3. Storing a raw, unaltered copy of the data while simultaneously forwarding it for processing.
4. Merging Kafka data – Aggregating records based on a specific number of records, time, or size thresholds.
ETL processing:
5. Handling NULLs & Duplicates – Filtering, logging, and generating alerts for further analysis.
6. Filtering & Transforming Data – Using SQL queries to split records into different tables.
7. Renaming files dynamically – Example: file_name_${now():format('yyyy-MM-dd_HH-mm-ss')}_001.json.
8. Storing processed files in Azure Data Lake Storage (ADLS).
9. Triggering Delta Live Tables (DLT) in Databricks – Using the InvokeHttp processor to automate workflows.

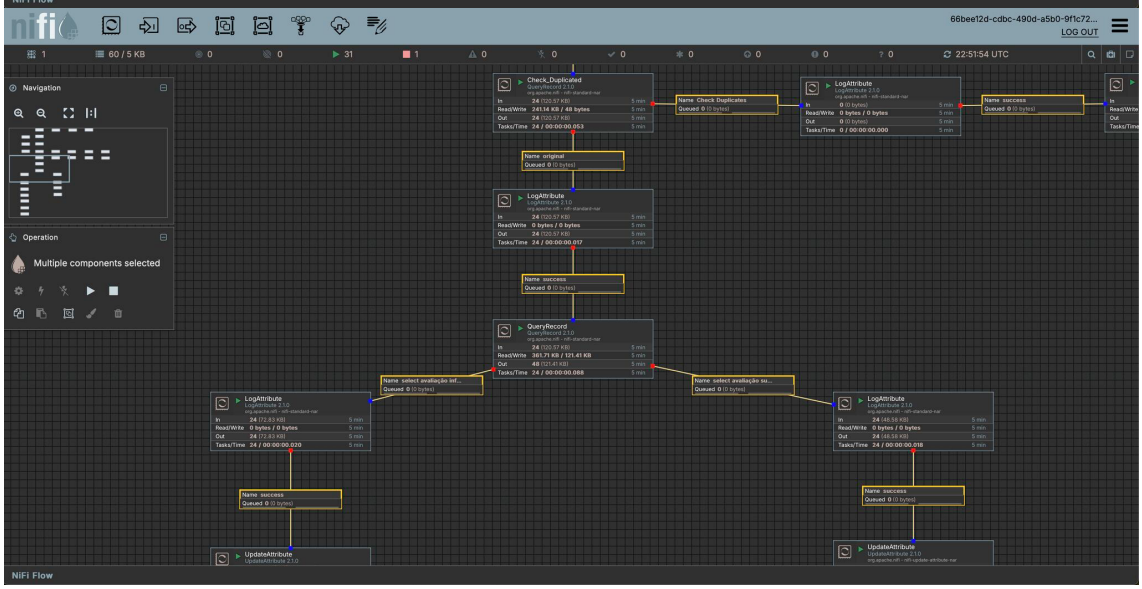
Once the Databricks pipeline is completed, a Webhook response notifies NiFi to proceed with the next steps, ensuring a fully automated data pipeline.

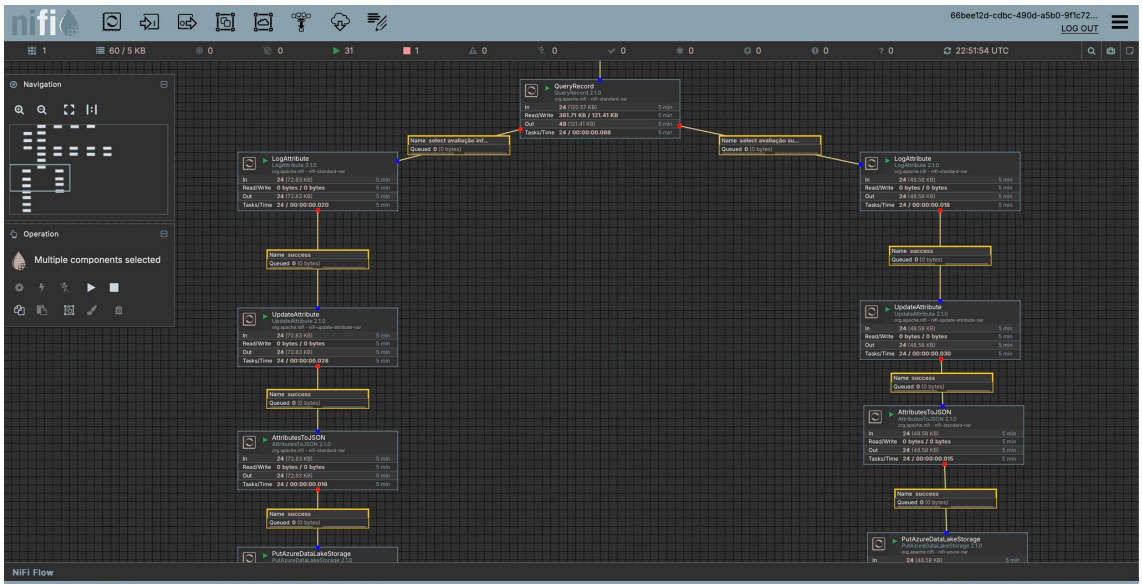
The combination of NiFi and Databricks is powerful. NiFi acts as an orchestrator, while Databricks handles the heavy lifting, leveraging powerful clusters and the benefits of Apache Spark and the Delta format.

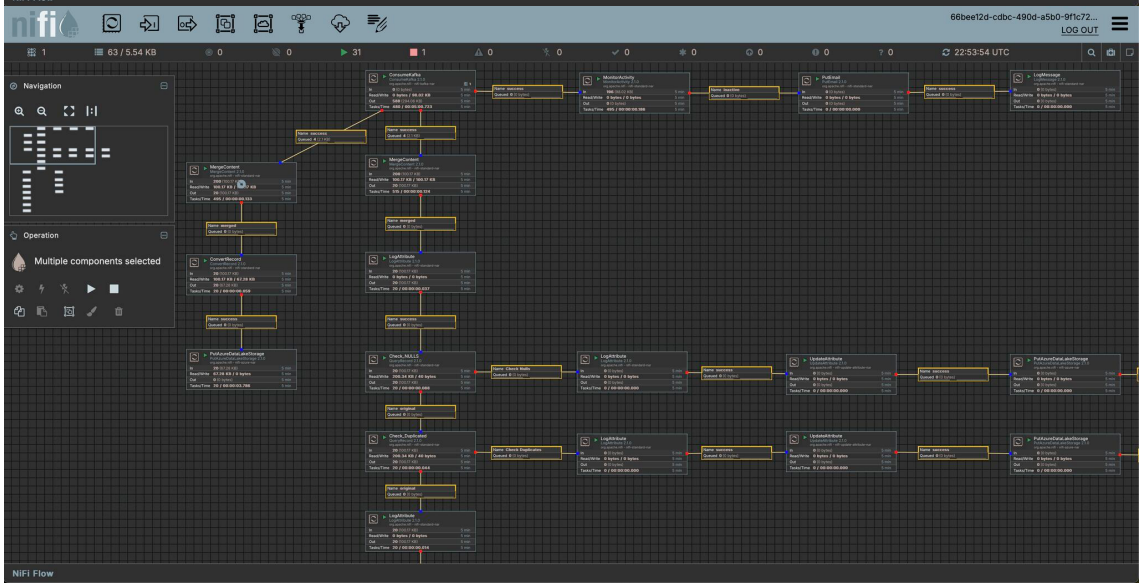
Why I use and recommend NiFi:
Open-Source & Cost-Effective – No licensing fees, full customization.
Intuitive UI – Drag-and-drop interface with real-time monitoring.
Scalability & Flexibility – Ideal for batch and streaming data.
Seamless Integration – Works smoothly with Kafka, Databricks, Azure, and other platforms.

This project highlights Apache NiFi's power in real-time data orchestration, making complex workflows simpler and more efficient.
#ApacheNiFi #Databricks #DataEngineering


66bee12d-cdbc-490d-a5b0-9f1c72...
LOG OUT
22:51:24 UTC
Navigation
Operation
Multiple components selected
Check_NULLS
Check_Duplicated
LogAttribute
UpdateAttribute
NiFi Flow


66bee12d-cdbc-490d-a5b0-9f1c72...
LOG OUT
22:51:54 UTC
Navigation
Operation
Multiple components selected
Check_Duplicated
LogAttribute
QueryRecord
UpdateAttribute
NiFi Flow


66bee12d-cdbc-490d-a5b0-9f1c72...
LOG OUT
22:51:54 UTC
Navigation
Operation
Multiple components selected
QueryRecord
LogAttribute
UpdateAttribute
AttributesToJSON
PutAzureDataLakeStorage
NiFi Flow


66bee12d-cdbc-490d-a5b0-9f1c72...
LOG OUT
22:53:54 UTC
Navigation
Operation
Multiple components selected
NiFi Flow