

Demystifying ETL Automation

ETL Pain Points Resolved with Flux

- Arul Dhesiaseelan, Flux Developer

One of our customers was discussing a requirement in which they launch their ETL tool every time they receive files from their trading partners. It does not surprise me when people use scripts to implement such use cases. While this may seem trivial in a happy world, it often involves some heavy lifting from the people involved in the process.

Our customer, I'll call him Joe the Trader, said, "We watch for available files from the folks we do business with. Once we see those files, we download them and perform some administrative work. After that, we run integrity checks on the data and, ultimately, orchestrate those files towards our ETL tool."

Joe might consider these tasks as mundane, but that's what makes them perfect candidates for automation. Automation not only makes this manageable, but brings additional benefits such as error handling and automatic recovery when the process deviates from the happy path. The challenges involved in error handling can be tedious and expensive.

"We sometimes need to wake up engineers in the middle of the night to fix an unanticipated problem in our workflow. But really, the problem's not unanticipated. We know it could happen. We just haven't taken the steps to automate how we respond to these late night problems," Joe told me on a recent call.

ETL Pain Points

An important requirement for ETL tools is the ability to interface with their SOA counterparts, streamlining data access across providers and consumers. There seems to be five major pain points that make ETL alien in an SOA environment when not addressed correctly:

1. Scripting is complex and error prone.

2. Data is spread across several data sources and processing is often time consuming.
3. Failure detection can often be cumbersome and very expensive.
4. Load balancing can be very tricky in a complex ETL environment.
5. Pinpointing hot spots and risk mitigation can be difficult in a heterogeneous setup.

Managing ETL Processes

ETL services are the tip of an SOA iceberg, but harbor a vast ecosystem catering to tools, processes, and best practices — making it a specialty. This isn't done without some challenges, but there are tools that can assist in these efforts and aid in productivity.

I spoke with another customer, Sabrina, whose team was struggling with their busy month-end processes. Analyzing spikes in their month-end ETL scripts and lack of monitoring tools made their situation worse. Sabrina acknowledged that isolating incidents was always difficult for their team when there were dependencies involved with other processes. This often kept their team busy in troubleshooting and sometimes manually running scripts, making them unproductive.

ETL processes require automation in an SOA world, but they also need a scalable solution that responds to dynamic load and automatic error resolution. Providing a centralized dashboard for monitoring SLAs and error notifications can reduce hours of staff time. One tool that can easily tackle the five common pain points and more is Flux.

Flux as an ETL Automation Tool

Flux has been a proven cross-platform workload automation solution for over a decade. The sophisticated scheduling and managed file transfer capabilities satisfy a majority of ETL tool automation requirements in SOA environments.

Through web services, Flux provides integration points for ETL tools such as SAP BusinessObjects Data Services, IBM DataStage, and Informatica that neatly interface with other applications. Flux's orchestration of ETL processes with scheduling, dependency, and error handling requirements

gives enterprises the scalable solution and automation needed to help meet productivity requirements.

Some of the key benefits of Flux for ETL tool automation include:

- Flux's drag-and-drop graphical designer helps ETL users manage complex job dependencies visually. Orchestrating ETL processes using a graphical designer makes ETL workflow design intuitive and less error prone.
- Improves performance by processing these large data sets in a parallel fashion.
- Notifies stakeholders of SLA violations, performs automatic error recovery based on business rules, and enforces SLAs on customers and data providers. (QoS and Governance)
- Helps design a scalable ETL environment by intelligently distributing load to servers based on availability and capability, efficiently utilizing resources, and lowering cost of ownership for enterprises.
- Easily pinpoints hot spots in ETL workflows by analyzing historic runs.
- Provides a comprehensive run history and audit trail of ETL processes. (Audit and Compliance)
- ETL workflows can be monitored from a centralized web dashboard. (Operations)

Bringing Workflow Capabilities to the ETL World

Flux brings workflow capabilities to the ETL world with its intuitive toolset and allows ETL developers and architects to focus on their core competencies and leave the complexities such as dependency management, error resolution, and resource optimization to Flux.