A Tipping Point for Automation in the Data Warehouse

www.stonebranch.com

stonebranch

Resolving the ETL Automation Problem

The pressure on ETL Architects and Developers to utilize automation in the design and management of their processes is increasing. In most organizations, the use of native ETL or operating system scheduling utilities along with scripting has provided the basis of orchestration inside the Data Warehouse. There are limitations to this approach as processes increase in complexity and scale, and as dependencies form between related technology domains such as ERP, ETL and Business Intelligence.

Partially due to poor marketing by vendors, Workload Automation has sat on the periphery of the toolset used to automate ETL processes. In contrast these solutions are widely deployed alongside the similarly process-heavy business applications that represent ETL data sources, but they have not realized broad based adoption as a 'must-have' Data Warehouse technology. That is changing with the requirement for ETL applications to help deliver centralized data that supports near real-time BI Analytics for business users. Workload Automation can improve management and oversight, decrease operational latency, and ultimately help ETL solutions scale to meet the growing demands of the underlying business they are meant to support.

'Stuff' Flows Downhill

Demand for Business Intelligence analytics has placed pressure directly on ETL to move faster with less potential for errors. A growing audience of business users, which are able to make requests for enterprise data, places the burden of improved accessibility on BI software. This in turn places data quality and service delivery requirements on Data Warehouses, and ultimately ETL, which is meant to orchestrate the availability of data to these demand-side applications. Increasing consumer side demands have provided a strong impetus to evaluate automation tools to improve data quality, service delivery, and operational efficiency. Starting with business SLA's and working backwards, it is clear the strain on ETL increases when demand from new data 'consumers' ramps up.

Performance Bottlenecks

There are two pressing issues that drive decision making in ETL use cases: Data accuracy, and speed of delivery. Data quality can be improved by ensuring that there is proper oversight into each of the complex actions taken against disparate enterprise data sources and the unified target data store. Speed of delivery on the other hand is impaired by processing latency. The delay between the completion of one task and the initiation of successors can have a significant impact on the ability to deliver data on time.

ETL Relies on a Brittle Form of Automation

The use of scripts as a means of filling scheduler feature gaps, and as a method of process integration and execution, is pervasive in even the most advanced IT organization. They provide a flexible way to manage the need to define and execute work. However, even the most talented developer cannot overcome their inherent problems: they are error prone, require ongoing maintenance and force manual approaches to monitoring and troubleshooting.

Embedded ETL Schedulers Lack Mature Automation Management Capabilities

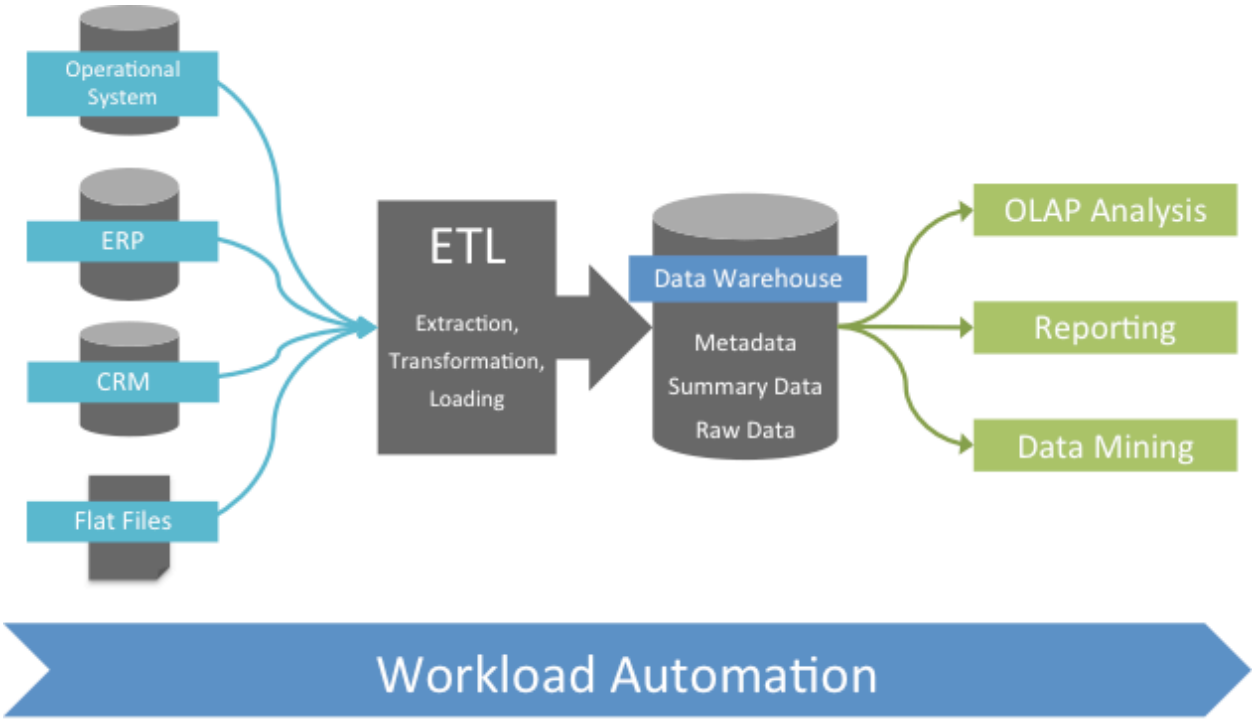
Determining the health and wellness of Data Warehouse processing can require significant manual intervention. When things go wrong with the execution of an ETL task or a script, there is very little information provided to the person tasked with monitoring process status. Manual troubleshooting is often required, provided that the operator is aware that something has gone wrong at all. This is not a scalable approach to operational oversight and failure recovery.

Business Processes Don't Stop at Technology Boundaries

One of the most important capabilities of Workload Automation is ease with which workloads can be defined and related across systems and applications. Business processes do not start and stop at notional technology boundaries. For example, ETL workloads can be initiated by the arrival of a file on the front end. Conversely, a report meant to be delivered to marketing would have a number of related actions that take place to prepare, normalize and present any associated data. ETL scheduling utilities were not built to manage these kinds of cross application dependencies.

A Definition of Workload Automation for the ETL Developer

Workload Automation is a core data center technology used to define, execute and monitor IT and Business processes based upon schedules and events. It is used pervasively to improve the management of siloed automation needs inside individual technology domains (such as ERP, DW, ETL, and BI), and also to define processes that span across these technologies. Workload Automation has been used to replace embedded scheduling tools that are provided with business applications, operating systems and databases, as well as the use of custom scripting. Given the complexity and scale of ETL processes (or job mappings), and their innate relationship with source and target data oriented systems, there is a compelling case to incorporate the use of Workload Automation solutions alongside them.



Simplify End-to-end Process Definition

Workload Automation products enable faster process definition. The better products on the market have advanced visual workflow oriented interfaces to allow users to drag and drop objects from a palate onto a canvas. Each step in a process can be modeled quickly and then modified by clicking to edit. In the same way connections and dependencies can also be defined visually, and then modified to enhance relationships for more complex dependency logic and scheduling criteria.

Workload Automation tools offer the capability to model an IT or business process from end-to-end. Users can drag and drop disparate workload objects in the same workflow and resolve the dependencies between them to carry out the overall goal of the process. Starting with events or actions related to source data, through data preparation and report delivery to end-users, ETL architects can help orchestrate truly intelligent automation. Workload Automation also makes it easier to break processes up into smaller discrete objects, which makes managing restart processing much easier.

End-to-end capabilities are more a case of a right, without the obligation. Developers can also author processes to focus only on data pulls from source data to the target data store. Alternatively they can take a demand side approach and focus workflow development on business user requests which take data from the warehouse or data mart and automate through the reporting tasks all the way to the end user who initiates a request. In this way, Workload Automation solutions have the potential to elevate ETL architects and developers as an even more critical asset in supporting business needs.

Eliminate Brittle Objects

Automation solutions have not completely eliminated the use of scripts, but rather reduced the reliance on them to resolve complex scheduling and event logic. Scripts can still be written to resolve Data Warehouse related integration problems. However, they are at least stored, executed and monitored along with the rest of the related workload objects in a way that ensures they are not causing significant operational headaches. The benefits of reducing the number of unmanaged scripts are reductions in maintenance time, reduced failure rates, simplified troubleshooting and reduced mean time to repair.

Improve Oversight

The supporting management and control capabilities offered in enterprise class Workload Automation products can provide significant value to ETL developers or operations teams tasked with monitoring, trouble shooting and repair. Some of these capabilities include:

- A monitoring console to replace manual status checking
- Failure notifications and intelligent error handling
- Source control and deployment for scripts and job definitions
- Auditing and logging for compliance purposes
- Multi-threaded and concurrent processing
- Workload balancing
- Auditing and compliance
- Automating reporting services

Managing failed processes is also significantly easier when you know where the process failed. Rather than manually checking individual files, renaming objects, and searching for changes, users know where a process was, and ultimately what to do to correct it.

Improve Processing Performance

Workload Automation can improve processing performance by reducing latency between steps in a workflow. Scheduling latency is not exclusively a data warehousing problem as many older style job scheduling systems exhibited the same problems as embedded ETL tools. Workload Automation can improve performance by eliminating processing latency between related jobs, time taken to recognize an event (file arrival), or checks on database updates. Workload automation solutions also allow for parallel or concurrent processing which is a limitation of scripting, OS schedulers, and some of the ETL schedulers.

New Projects Provide the Opportunity to Deploy Automation Technologies

If the benefits of Workload Automation to the data warehouse are clear, the next most obvious questions are when and how to deploy these products. If you ask experts in IT Automation, they will leave very little doubt that the best time to implement is when there is a high priority and funded IT project to attach to. Even though Workload Automation is considered a 'horizontal'

software category, it is most frequently adopted on the back of a single initiative. For example, in the late 1990's the solutions were implemented alongside ERP applications as a complimentary automation and orchestration tool. In the last few years, new ETL deployments have become a very common trigger for purchases. The automation solution is first used to manage the ETL-specific processes, and then extended and integrated with upstream and downstream applications, to manage process dependencies. Changes in related infrastructure will similarly drive adoption. As described previously, new Business Intelligence, Analytics, and Data Warehouse projects place pressure on ETL and can provide very substantial justification for automation.

New People Require New Approaches

New deployments of ETL are not the only trigger for adoption. In many cases once script heavy processes are defined and stabilized they are often transitioned to an operations team to manage. IT Operations in large enterprises are probably already using enterprise class workload automation products and demand the capabilities offered by these solutions. When they are tasked with taking on new internal clients, and their workloads, they typically require the use of a proper enterprise automation solution. They will work with the Data Warehouse teams to subsume and redefine their processes. They simply cannot afford manual status checking and limited visibility into the health and status of workloads they are monitoring. If this workload is going to be shifted to another team with less domain knowledge, and numerous competing internal customers, there should be an expectation that the way workloads are defined and managed will have to change.

What to do About Legacy ETL Processes?

With the resource demands placed on IT, it is extraordinarily difficult to build proper justification to tackle legacy operations problems. Updating legacy is a wholly uninteresting proposition to many CIO's. In the absence of project funding it is difficult to address brittle forms of automation unless an internal champion takes it upon themselves to look for a better solution. So given the value delivered by Workload Automation, what can be done to resolve problems with the way processes are currently defined?

In these circumstances it is best to take an incremental approach to migrating processes to Workload Automation. In our experience, there are often automation champions that emerge in an organization, that have the capacity to implement smaller deployments of WA, and as workloads are modified or changed, they migrate the definitions, scripts and logic associated with ETL, process by process over a longer time horizon. This is a low risk approach that is not burdened with lengthy business justification.

Formal Adoption of Big Data Technologies is on the Horizon

Technology shifts drive change in supportive infrastructure. Certainly the biggest pending shift for operations teams related to managing data access is how to deal with Big Data. For large enterprises NoSQL and Big Data-reliant applications are still very much in exploratory use, and remain outside the boundary of IT Operations to manage. The reality is those types of applications move faster, placing a bigger burden on existing technology categories like ETL and BI to access and make use of

the information they capture. For now, unless you are a social media software vendor, big data related queries are not directly related to core business (mission critical) data. The maturation of tools like Hadoop, and Pentaho (BI) and their use cases may remain peripheral to an enterprise data management strategy. That day is coming and when it does data warehouse infrastructure will need to support those demands.

Conclusion

Workload Automation has a critical role to play as ETL and process-driven data management technologies scale to meet both the voracious demands of new consumers of enterprise data and the incorporation of new technologies. It is a capability that will be leveraged as a standardized infrastructure layer to support both the vertical needs of Business Applications, ETL, DW and BI, as well as to resolve processes that span these domains.

The benefits will include decreased error rates, improved error handling, greater oversight, improved scalability of operations, and reduced cost of delivery. The justification to deploy can be driven off strategic initiatives or incremental improvements on legacy workloads. It's important to note that ETL Architects and Developers have a critical role to play in that adoption, and the necessary skill set to see it through.