BI and ETL Process Management Pain Points

Understanding frequently encountered data workflow processing pain points and new strategies for addressing them

What You Will Learn

Business Intelligence (BI) reporting and the extract, transform, and load (ETL) processes that support it are vital to equipping corporate decision makers with the information they need to maintain and improve business performance, successfully scale up as their businesses grow, and identify and implement new business initiatives and strategies. Getting current, accurate data in a timely manner is essential to helping ensure that the processes that move data from source to target systems run smoothly and quickly and that any errors are identified and resolved rapidly, without having IT staff waste valuable time finding and fixing issues manually.

In addition to keeping scheduled BI and ETL processes running smoothly, IT staff is regularly asked to provide unplanned reports that are vital for supporting business decisions. Scheduling these just-in-time reports is a complex process that not only consumes IT staff time but can also interfere with the successful execution of regularly scheduled production workflows.

The biggest challenge when scheduling BI and ETL jobs is achieving error-free, end-to-end integration between the processes that are distributed throughout the enterprise that supply the necessary data. The introduction of a system that can run automatically, manage its own dependencies, and does not need to be monitored at every step would help reduce time constraints and error-prone manual intervention. The overall benefits of an automated system are increased processing efficiencies and fewer mistakes.

New, sophisticated scheduling and process automation technologies provide all these benefits. They help by managing and fine-tuning vital enterprise data flows in an automated manner, eliminating error propagation and process stoppages, while freeing IT staff to address high-value strategic challenges.

This document examines five frequently encountered BI and ETL process management pain points that can affect the accuracy and timeliness of vital enterprise data delivery. This includes dashboards, interactive analytics, and standard reports, which are often used to make far-reaching business decisions. This document also discusses how each of these pain points can be successfully managed using process automation technology.
Pain Point One - The Race Against Time
Managing Batch Window Time Constraints

One of the biggest challenges facing an IT group is completing ETL processing and traditional batch-based BI jobs within the constraints of an ever-shrinking batch window. While there is a trend toward real-time BI, the vast majority of BI report generation today relies heavily on this “offline” window to complete these jobs. Several factors contribute to the difficulty in managing these processes in the time available.

The first factor is a severe lack of visibility into the various complex processes themselves. Typically, hundreds or even thousands of data sources are involved, with some traffic arriving from external sources, such as tax tables or data feeds from currency exchanges. This means there will always be bottlenecks, and it is very difficult to predict where they will occur. Additionally, preprocessing may be required on the incoming data feeds, involving other enterprise applications, before ETL jobs can commence. Finally, ETL and BI reporting processes lack integration, which makes coordinating them even more problematic. Unless these problems are carefully managed, the resulting reports, dashboards, or alerts can contain incorrect information or be generated too late to be of practical use.

The second factor to consider involves data feeds that are expected to arrive in a defined time window. Traditionally, this is handled using time-based schedules, which, for various reasons, are inefficient and ineffective. If feeds are expected to arrive at a certain time and one is late or corrupted, errors result. These errors then cascade through the rest of the process, propagating and often compounding the errors. To avoid this problem, time buffers are built into the schedule. These may lead not only to the loss of valuable cycles as machines idle but can also add so much time to a complex workflow that it cannot finish within the time allocated. Built-in delays can help prevent errors, but if the batch window is too small to accommodate all the buffers, the process may not complete in time. Conversely, if files arrive early, no advantage can be taken because of the static nature of the scheduling process.

Time-based scheduling not only increases processing time, it also consumes IT staff time. Many necessary preprocessing tasks may be outside the ETL process or even on a different distributed system, which adds to the complexity of managing them. These processes generate vital data that must be fed into the system before ETL jobs can commence, and it is easy to miss one or more critical data events, which can result in sending incorrect data downstream. This happens more frequently when dependencies within the enterprise are not interconnected. The complexity of the processes makes it difficult and time consuming for IT staff to identify all the dependencies and manage them effectively in a consistent manner.

New, sophisticated enterprise job scheduling technology provides a comprehensive, end-to-end view of these complex, interconnected workflows. Enterprise job scheduling is a distributed system that coordinates all aspects of the workflow, from data sourcing to report generation. Because of this total coordination and control, arbitrary delays do not need to be built into the system. The system allows IT staff to see exactly which processes are involved at every stage of the workflow, to allow all the dependencies to be met in an automated manner.

Figure 1 illustrates how an enterprise scheduler manages different parts of the entire data integration workflow and allows very complex dependencies to be easily mapped to the workflows. This mapping even allows batch jobs that are outside the BI system to be fully integrated and run in a coordinated manner.
Figure 1. Event-Based Batch Scheduling Using Enterprise Job Scheduling Coordinates Different Systems, Eliminates Arbitrary Delays, and Manages Dependencies Automatically

Pain Point Two - Cascading Errors and Painful Recovery
Eliminating Errors Caused by Improper Job Sequencing

Poorly managed dependencies lead to process errors that can cause jobs to fail, but these errors do not stop at the level of the job that has failed. They can cascade through the workflow and propagate in multiple places downstream. Consider what happens if a corrupted file enters the process, perhaps because the file was transferred before it was ready. The incomplete data from that file is then sent as input to a process - for example, the enterprise resource planning (ERP) system - which mixes good data with bad. This new corrupted data then goes into an ETL process, creating errors in the data warehouse. This erroneous data, in turn, results in incorrect reports and dashboards being incorrect and leading to bad business decisions. In this manner, even small errors can cascade throughout the system, causing increasing numbers of errors. When processes are managed manually, it is very easy to let bad data slip through the job processing system.

Starting processes before required predecessor processes have completed also causes errors. For example, if an ETL process starts before the ERP system has finished its job, more downstream errors are created. Jobs can also fail for various reasons, and these failures may not be detected, which also causes errors that cascade to report generation. In many cases, multiple errors have several different causes. These errors are even more difficult to detect and manage manually.

Another example of process errors that can occur due to poorly managed dependencies is if two jobs that update the same database table begin running concurrently. This can corrupt data or lock out the next job step from accessing the tables.

Essentially, errors commonly occur when data is not in the right place at the right time. Very complex workflows have too many “moving parts,” and data pathways can be easily broken. These errors can then cause additional errors in multiple places along the workflow, as shown in Figure 2. Vital IT staff time is required to detect the errors that caused the process to halt or complete erroneously and then to identify all the other downstream errors and rerun the process from the point of failure. It is a very difficult, frustrating, and time-consuming job.
Figure 2. Small Errors, if Propagated Throughout the Process, Can Cause a Cascade of Errors and Erroneous Reports, and Ultimately Lead to Bad Business Decisions

The real problem with these process-generated errors is that if they are not detected, as in the case in which data is loaded from an outdated or incomplete table but not detected by IT operations staff, bad data can appear in reports that ultimately lead to bad business decisions.

Pain Point Three - Ad Hoc Reporting

Managing Unplanned Reports in a Plan-Based Environment

Businesses are run based on the intelligence they collect, and the most successful business decisions are made using accurate, up-to-date information. Business managers constantly request data from IT in different configurations to support ongoing, strategic decisions. This, in turn, requires IT to generate ad hoc reports that can be time consuming and, if not set up correctly, result in outdated or inaccurate downstream information. Yet, ad hoc reporting is an essential part of what IT does, because it is a valuable business decision support tool that helps keep an enterprise competitive in an ever-changing market.

Setting up unplanned, custom reports is time-intensive, because the process is not always straightforward. Data sources must be very carefully examined to determine which processes need to be run in which order, with varying interdependencies, before data is in a consistent, updated state so that reports can be run. Manually mapping, running the processing steps correctly, and verifying data consistency before running the reports can take hours and create a significant delay in report generation. In addition, the time pressures that often accompany ad hoc report requests can make this process error prone.

Ad hoc report requests can also introduce problems for regularly scheduled workflows. Many IT managers cite ad hoc reporting as demanding more staff time than any other activity. Even more time is required for IT staff to identify and fix problems or rerun ad hoc reports that contain bad data. Of course, if bad data goes undetected, bad business decisions can result.

The key to taking full advantage of the power of ad hoc reports or just-in-time report generation is to set up the workflow correctly and automate the resulting process as much as possible. It is essentially a matter of being able to manage dependencies, which requires making transparent all the processes involved in generating an ad hoc report. Monitoring all the linked and dependent processes within the workflow is vital to helping ensure that an upstream process has completed before downstream report generation begins.

Sophisticated enterprise scheduling solutions allow IT to observe how each process in an ad hoc workflow relates to every other process. The entire end-to-end workflow is graphically illustrated and shows the real-time state of the ad hoc process as data moves downstream (Figure 3). An automated scheduling solution allows the entire process of
unplanned reporting to be efficiently and accurately managed: from data sourcing through generation and electronic delivery to the proper end user.

**Figure 3.** Business Intelligence Example: Ad Hoc Reports Are Easily Handled with an Enterprise Job Scheduler that Manages the Process Across Complex Interdependent Predecessor Processes, Report Generation, and Delivery

**Business Intelligence Example (ETL and Reporting)**

This solution makes it easier for IT to define and manage any number of ad hoc requests. Not only is the resulting data more likely to be accurate, the entire process takes far less time to run. IT staff simply needs to trigger the process that leads to the desired report generation and provide the delivery targets, and the report is run and delivered as soon as its prerequisite processes are complete and all dependencies are met. This saves valuable time that would otherwise be spent tracking the process, triggering steps manually, detecting whether data is accurate, determining the cause of any problems, and reconfiguring or restarting a workflow if necessary. The workflow itself takes far less time to run, because static delays can be managed with more fluidity.

**Pain Point Four - Service-Level Consistency**

**Managing Service-Level Agreements**

Service-level agreements (SLAs) are the universal benchmark of successful IT performance, and to achieve consistent SLA performance year after year, IT must have the resources to monitor SLA throughput on a day-to-day, hour-by-hour basis. Usually overall SLA performance slips a little each time an unforeseen problem halts a workflow and a job finishes late. Each time small errors or slippages occur, downstream effects accumulate as more and more jobs finish late or fail altogether. Often the complexity of the process makes it difficult to predict the result of one slippage or job failure on the overall process. With limited visibility into these effects, it is hard for IT managers to determine where to focus their resources.

Newer scheduling technologies provide what is probably the greatest single tool for improving SLA performance: the ability to monitor and manage end-to-end processes running in the enterprise from a single console. Process dependencies and entire enterprise workflows are clearly illustrated and can include the relative and granular states of all ongoing job steps (Figure 4).
One of the most important benefits of this console-based technology is that IT can be instantly alerted if an ongoing process will cause another downstream job to finish late. This predictive capability allows IT to make corrective adjustments before errors affect downstream workflows, which in turn greatly increases SLA performance.

Each workflow does not have to be constantly monitored to provide these benefits. The new technology includes automated alerting processes, which can send an alert to an on-call administrator or send an email to their smartphone, stating that an expected event did not or will not occur within a predicted time frame. This automated process expedites resolution steps by allowing immediate assessment and error correction. The alerting mechanisms often offer methods for automated error resolution, such as rerunning a process or sending diagnostic data through an automated escalation and recovery process.

**Pain Point Five - Resources**

**ETL Resource Conflict Management**

Because many batch processes might run at the same time as ETL and BI workflows, unexpected resource conflicts can arise. When seemingly unrelated processes run concurrently, there is always a chance that they might attempt to access the same resources at the same time. Some of the problems that result are delays due to waiting for resources to be released, simultaneous updates that cause data to be overlaid or corrupted, or outright step failures that halt a process, requiring unnecessary operator intervention.

Without coordinated queue management, critical workflows may be negatively affected by less critical processes. For example, a user may submit a resource-intensive ad hoc report request against a data store that needs to be refreshed with new data. A scheduled workload that normally would take an hour to run may take several hours if it has to compete with that report job. And, when all is done, the ad hoc report would be incorrect, because it is based on old or incorrect data.

It is critical that ETL processes be coordinated with all batch jobs that are scheduled to run in the same time frame. By placing ETL and BI within the broader enterprise scheduling context, IT can better manage resources and resource conflicts. Modern schedulers meet this challenge with resource and queue management features that include job limits and prioritization, resource locking, load balancing across servers, and automation of virtual resource management for jobs running in a virtualized environment.
Stop the Pain with Standardized, Automated Scheduling

**Workflow Automation: A Real-World Example**

Automated enterprise scheduling helps BI and batch operations staff address each of the pain points described above. This capability is illustrated by the real-world example of a midsized pharmaceutical company based in the United States that had successfully marketed a wide variety of branded and specialty generic products. As the enterprise grew, so did the number of strategic initiatives that needed to be addressed to manage the growth of the business. The company needed to increase its financial data warehousing capabilities to track direct sales and financial performance and to closely monitor customer performance against contracts. Also, weekly reports were required to be sent to an external field organization of hundreds of sales representatives.

**Barriers to Success**

The company discovered that too much manual processing was required to keep data flowing in a timely fashion. Not only was the manual processing and troubleshooting frustrating and unrewarding for IT staff, but it also consumed all the staff's available time. The company realized that to manage its growth, it would need a workflow automation solution to free vital staff time. It actually had a number of schedulers running in different parts of the process stream, but as is usually the case, each scheduler was unaware of what the others were doing.

The company’s most important requirements for an automation solution were that it provide a single, end-to-end view of all the processes running throughout the environment, automatically handle and resolve process dependencies, and be easy to use.

**Automated Solution Relieves the Pain**

The centralized enterprise scheduler solution selected is easy to use and provides broad cross-platform, cross-application support. The company believes that the time savings gained by refocusing IT resources on strategic initiatives will continue to increase over time. It also believes that it will benefit from a high degree of control over and operational stability of the environment as more and more processes are moved to the enterprise scheduler’s control.

The enterprise scheduler addresses all the company’s IT challenges:

- **Time:** By implementing this automated solution, the company is looking at about a 20 percent reduction in the time it takes IT to run jobs that include multiple steps. The centralized enterprise scheduling solution selected was able to eliminate all of the arbitrary delays built into the system to avoid conflicts. Elimination of these delays contributed to much more efficient throughput in workflow processes. This feature alone provided vital business information much faster than before, thus decreasing the delivery times of mission-critical BI reports.

- **Errors:** Because all dependencies can be incorporated into a schedule and because the enterprise scheduler automatically tracks them and provides alerts and data about problems, bottlenecks, and a range of execution errors, staff can quickly pinpoint issues and resume the processes without having to start them over again. Error reduction and the capability to proactively pinpoint issues and swiftly address them have contributed to faster and smoother process execution.

- **Ad Hoc reporting:** Using the graphical business views process generator, IT staff can quickly set up processes to run special reports that are not in the standard job schedule. IT staff can use the graphical workflow development environment to create single-instance reports, and it can determine how running these ad hoc reports affects regularly planned processes, because it can monitor performance of each step and all of the processes running in a given time frame. Because latency is removed through intelligent dependency management, processes run faster, ad hoc reports can be generated faster, and IT is seen as more responsive to business demands.
• Service-level consistency: The selected job scheduler prepares both the BI and ETL workflows, as well as other business processes, to scale more smoothly without compromising service levels. The scheduler can do this, because the demands of scaling, such as adding new jobs, databases, resources, services, reports, and end users, can be incrementally built on the enterprise scheduling foundation. Vital business information is prepared faster, because the scheduler can “predict” the downstream impact on a process when an upstream job encounters problems or runs late and can alert IT staff, who can then take appropriate action. And using centralized automation reduces error rates, adding more confidence in the reports generated by the IT group, which further increases service levels.

• Resources: Because the enterprise scheduler provides features to manage resource conflicts, IT staff has a better view of and control over ETL jobs and batch processes that run concurrently. As a result, conflicts can be eliminated and resources can be used according to priorities, or additional resources can be obtained automatically as needed to fulfill increased processing requests.

Essential Features That Help Address the Pain

The following are some of the features of the selected scheduling solution that are crucially important in helping relieve job processing pain points:

• Ease of use, with a highly intuitive user interface and script-free scheduling

• Broad coverage that allows users to incorporate various systems, applications, and technologies into the enterprise schedule

• End-to-end visibility of schedules and job execution

• Fine-tuned alerting and notification

• The ability to create complex schedules with many nested dependencies

Shifting IT Resources to Strategic Initiatives

The enterprise scheduling solution this company has selected gives it a powerful tool for supporting growth and allows it to shift resources to support critical business strategies. The solution also provides time savings and the ability to shift high-value IT talent and budgets to strategic initiatives.

The company is now prepared for future growth and the inevitable changes in the marketplace. It is ready to scale and adapt new systems and applications to support its BI needs, because it can easily incorporate almost any new data source required to support its ETL processes and strategic goals.

Better, faster, and more efficient ETL data integration is helping this enterprise support a welcome shift in focus to the strategic issues it needs to address. The robust enterprise scheduling solution is the remedy that this pharmaceutical company’s data center management had been seeking. Through better process intelligence, IT is now able to support better business decisions across the enterprise.

For More Information

To learn more about Cisco products and services, please visit http://www.cisco.com.