independIT Integrative Technologies GmbH
Bergstraße 6
D–86529 Schrobenhausen

schedulix!focus

Das schedulix Scheduling System im Data Warehouse Umfeld

Dieter Stubler Ronald Jeninga

November 25, 2016

Das schedulix Scheduling System im Data Warehouse Umfeld

Einführung

In Data Warehouse Umgebungen müssen täglich eine Vielzahl von Prozessen ausgeführt werden.
In vielen Unternehmen existiert zwar ein Enterprise Scheduling, welches häufig jedoch nicht zur Steuerung der einzelnen Prozesse im Data Warehouse herangezogen wird. Meist werden größere komplexe Batchprozesse vom Enterprise Scheduling angestoßen, die Ablaufsteuerung innerhalb dieser High Level Batchprozesse wird jedoch außerhalb der Enterprise Schedulings alternativ gelöst. Die Gründe dafür liegen meist in fehlenden Features des eingesetzten Enterprise Schedulings bzw. dessen komplexen und umständlichen Handhabung.
Die interne Steuerung der Data Warehouse Batchprozesse erfolgt meist durch den Einsatz von toolinterner Scheduling Funktionalität der eingesetzten Werkzeuge und/oder den Einsatz von Scripting (sh, perl, python, ...).

Toolinternes Scheduling

Der Einsatz von toolinternen Scheduling Funktionen ist problematisch, da häufig auch toolexterne Prozesse mit eingebunden werden müssen. Dies ist oft nicht möglich oder muss häufig über Workarounds gelöst werden. Sind in einer 'Best Of Breed' Umgebung mehrere unterschiedliche Tools für unterschiedliche Aufgaben im Einsatz, wird das Monitoring der in den verschiedenen Systemen aktiven Abläufe zu einem Problem. Soll ein Werkzeug ausgetauscht werden, muss auch die im Werkzeug realisierte Ablaufsteuerung neu implementiert werden.

Scripting

Noch problematischer ist hier der Einsatz von Scripting. Um ein Minimum an Stabilität zu erreichen, entstehen hier erhebliche Entwicklungsaufwände die häufig in der Entwicklung eines eigenen kleinen Scheduling Systems münden.

Notwendigkeit eines Scheduling Systems

Beide genannten und häufig anzutreffenden Lösungen sind weder effizient bzgl. der benötigten Enwicklungsressourcen, noch sind sie dazu geeignet einen mittel und langfristig stabilen Betrieb der Data Warehouse Umgebung zu gewährleisten. Der Betrieb einer Data Warehouse Umgebung ohne ein Scheduling System ist immer zu teuer und unzuverlässig.

Einsetzbarkeit von Enterprise Scheduling Systemen

Enterprise Scheduling Systeme haben ihre Wurzeln in der Steuerung von Produktionsabläufen auf Mainframesystemen. Diese Produktionsabläufe sind typischerweise sehr statisch, wenig komplex, stabil und erfordern im Allgemeinen wenig bis keine Operatoraktivitäten. Da Entwicklung und Produktion typischerweise voneinander getrennt sind, ist für die Enwickler eine Einarbeitung in das Enterprise Scheduling nicht notwendig.
In einer typischen Data Warehouse Umgebung ist dies jedoch anders. Data Warehouse Prozessabläufe ändern sich nahezu täglich. Fehler durch Ressourcenengpässe bei der Verarbeitung großer Datenmengen sind wesentlich häufiger. Eine strikte Trennung zwischen Entwicklung und Produktion existiert in den meisten Fällen nicht. Dies führt dazu, dass jeder Entwickler neben seiner eigentlichen Tätigkeit zusätzlich mit Überwachung und Betrieb seiner Data Warehouse Prozesse beschäftigt ist. Die komplexen Abhängigkeitsbeziehungen in Data Warehouse Abläufen stellen hohe Ansprüche an die Funtionalität des eingesetzten Scheduling Systems. Der Einsatz eines Enterprise Scheduling Systems zur Steuerung aller Data Warehouse (Teil)Prozesse ist deshalb häufig keine Lösung, da diese Systeme nicht an die Anforderungen eines Data Warehouse Betriebes angepasst sind.

Geforderte Eigenschaften eines Scheduling Systems im Data Warehouse Umfeld

Ein Scheduling System im Datawarehouse Umfeld muss daher unter anderem folgende Eigenschaften mitbringen:

- Der Umgang (Job/batch Definition, Execution, Monitoring, Operation) muss auch ohne Programmierkenntnisse leicht und schnell zu erlernen sein. Dazu müssen die dem System zugrundeliegenden Konzepte einfach, klar und verständlich sein.

- Das Scheduling System muss alle Features mitbringen um auch komplexe Aufgabenstellungen zur Ablaufsteuerung abbilden zu können, damit nicht auf Scripting Lösungen ausgewichen werden muss.

- Um auf neue Anforderungen im Data Warehouse schnell reagieren zu können müssen Abläufe jederzeit geändert werden können ohne bereits aktive Abläufe zu beeinflussen.

- Der Operator muss jederzeit in der Lage sein, Einfluss auf aktive Abläufe nehmen zu können, um auf Ausnahmesituationen geeignet reagieren zu können.

- Für Installation, Adminstration und Betrieb des Systems dürfen keine besonderen Systemprivilegien (root Rechte, etc) erforderlich sein.

- Die Definition, Ausführung, Überwachung und Problemanalyse muss ohne zusätzliche Softwareinstallation von jedem Arbeitsplatz aus möglich sein (Web Application Server).
- Da Data Warehouse Umgebungen immer dazu tendieren die Grenzen der Hardware auszuloten, muss das System geeignete Mechanismen zur Kontrolle der Systemressourcen zur Verfügung stellen um Fehler durch Ressourcenengpässe zu vermeiden bzw. zu reduzieren.
- Der Zugriff auf Abläufe bzgl. Definition, Monitoring und Operation sowie die Möglichkeit der Ausführung von Jobs in bestimmten Umgebungen (Jobserver) muss durch Privilegien abgesichert werden können.

Etablierte Enterprise Scheduling Systeme bieten obige Eigenschaften nicht oder nur teilweise und sind daher für den Einsatz in Data Warehouse Umgebungen nur sehr bedingt geeignet.
Im folgenden wollen wir auf einige der oben genannten Punkte etwas näher eingehen.

Bedienerfreundlichkeit

Die einfache Bedienbarkeit stand bei der Entwicklung des schedulix Scheduling Systems mit im Vordergrund. Ein Web Application Server stellt ohne Softwareintstallation am Arbeitsplatz ein einfach zu bedienendes, Browser basiertes GUI zur Verfügung, welches von der Ablaufdefinition, Scheduling, Monitoring, Operating bis zur Administration der Jobserver alle Bereiche abdeckt. Erfahrungen beim Kunden zeigen, dass mit einer kurzen Einführung von ca. 3 Stunden jedes Teammitglied in die Lage versetzt werden kann, das System produktiv einzusetzen. Die Möglichkeit für wiederkehrende, komplexere Aufgaben, Templates zu definieren auf Basis derer durch wenige Handgriffe diese Aufgaben gelöst werden können, macht eine tiefergehende Einarbeitung in das System für den Großteil des Teams unnötig.

Komplettes Feature Set

Data Warehouse Abläufe zeichnen sich typischerweise durch komplexe Abhängigkeitsbeziehungen und Ablauflogik aus. Damit zur Implementierung dieser Abläufe nicht auf Scriptinglösungen ausgewichen werden muss, muss das Scheduling System ein komplettes Feature Set mitbringen. Das schedulix Scheduling System wurde auf Basis der komplexen Anforderungen eines großen Data Warehouses entworfen und über mehrere Jahre Produktivbetrieb stetig verbessert.
Folgende Features des schedulix Scheduling Systems seien hier exemplarisch genannt:

- Beliebig viele Benutzerdefinierte Exit Status erlauben eine flexible Reaktion auf das Ergebnis eines Jobs. (nicht nur Failure oder Success!)

- Neben dem Exit Status können Jobs auch zusätzliche Ergebniswerte dem Scheduling System übergeben, welche im Monitoring angezeigt werden können, um dem Operator zusätzliche Informationen über Jobs zu geben.

- Auf Basis von Exit Status und Ergebniswerten kann die weitere Verarbeitung von Abläufen über Dependencies und Trigger gesteuert werden. Dazu gehören auch die bedingte bzw. wiederholte Ausführung von Ablaufteilen.

- Parameter und Ergebniswerte von Jobs können von nachfolgenden Jobs zur internen Steuerung herangezogen werden.

- Durch die Möglichkeit Abläufe hierachisch zu gliedern, wird die Definition von Dependencies und Triggern zur Ablaufsteuerung erheblich vereinfacht, die Wiederverwendbarkeit von (Teil)Abläufen verbessert und die Überschaubarkeit auch großer Abläufe gewährleistet.

- Beliebige Teile der Ablaufumgebung (Tabellen, Datenbereiche, DataMarts, Files, ...) können als Status und Eigenschafts behaftete Ressourcen abgebildet werden, um Abläufe abhängig von Status und Aktualität der benötigten Ressourcen steuern zu können. (z.B.: Erzeugen eines Reports nur wenn der benötigte Datenbereich 'VALID' ist und innerhalb des letzten Tages aktualisiert wurde).

- Der dynamische Submit erlaubt es Jobs andere Jobs bzw. Teilabläufe zur Laufzeit anzustoßen. Dies ermöglicht die programmatische Steuerung von Abläufen sowie die Paralleliserung von (Teil)Abläufen.

- Ein Warning System ermöglicht es den Operator auf mögliche Probleme aufmerksam zu machen, ohne den Ablauf zu beeinflussen.

- Ein vollständiges API erlaubt es alle Funktionen des schedulix Scheduling Systems programmatisch zu steuern.

- Durch die Ablage aller Definitions-, Konfigurations- und Laufzeitdaten in einem RDBMS sind jederzeit beliebige Benutzerdefierte Auswertungen möglich.

- ...

Eine komplette Aufstellung aller Features würde hier den Rahmen sprengen.
Die Vertreter der Enterprise Scheduling Systeme können hier meist nicht mithalten und erzwingen ein Ausweichen auf Scriptinglösungen. Dabei werden dann Aufgabenstellungen, welche sich nicht durch Features des Scheduling Systems lösen lassen in die Implementierung von Jobs verlagert. Ergebnis sind unflexible, nicht standartisierte, instabile und schwer zu wartende Individuallösungen.

Flexibilität

In Data Warehous Umgebungen müssen Abläufe beinahe täglich angepasst werden. Hinzu kommt, dass die Abläufe häufig auch eine sehr lange Laufzeit besitzen. Dies führt dazu, dass noch während eine Instanz eines Ablaufes aktiv ist, dieser änderbar sein muss. Damit dies nicht zu Problemen in bereits aktiven und teilweise abgearbeiten Ablaufen führt, dürfen sich Änderungen nur auf neu gestartete Abläufe auswirken. Das Scheduling System muss also in der Lage sein, zu einem Zeitpunkt meherer Versionen einer Ablaufdefinition behandeln zu können. Das schedulix Scheduling System versioniert deshalb alle Aspekte einer Ablaufdefinition und führt Abläufe in der Version aus, welche beim Start (Submit) des Ablaufes gültig war. Häufig müssen im Data Warehouse ad hoc Aktionen (Fehlerbereinigungen, Analysen, Migrationen, ...) umgesetzt werden. Da das Erzeugen von Jobs und Abläufen im schedulix Scheduling System nur wenige Handgriffe erfordern, können solche ad hoc Aktionen schnell und einfach unter die Kontrolle des Scheduling Systems gestellt werden und unterliegen damit der Ressourcenkontrolle des Systems. Die Risiken, dass solche ad hoc Aktionen, andere wichtige Produktionsabläufe stören, können dadurch deutlich reduziert werden.

Operator Funktionen

Um auf Ausnahmesituationen (die Regel in Data Warehouse Umgebungen) schnell, reagieren zu können stehen dem Operator Funktionen zum Eingriff in aktive Abläufe zur Verfügung.
Diese sind:

- Anhalten und Weiterlaufenlassen von (Teil)Abläufen (Suspend/Resume)
- Neustart von Rerun von gescheiterten (Restartable) Jobs (Rerun)
- Setzen des Exit Status von Jobs
- Verwerfen von (Teil)Abläufen (Cancel)
- Ignorieren von Abhängigkeiten
- Ignorieren von Ressourceanforderungen
- Verändern der Scheduling Priorität von (Teil)Abläufen
- Beenden laufender Jobs (Kill)
- Kommentieren von (Teil)Abläufen
- Zurücksetzen von Warnings
- Verändern der Verfügbarkeit und Menge von Ressourcen

- Starten/Stoppen von Jobservern

Alle Operatoreingriffe werden zur späteren Nachvollziehbarkeit in einem Audit Trail festgehalten.

Freiheit von Systemprivilegien

Typischerweise stehen den Mitarbeitern eines Data Warehouse Teams keine Systemprivilegien (root Rechte, ...) zur Verfügung. Benötigt das Scheduling System zum Betrieb bzw. Administration solche Privilegien, so kommt es bei nötigen Änderungen am System zu Zeitverzögerungen und zusätzlichen Kosten bei deren Umsetzung. Insbesondere gilt dies, wenn wie häufig anzutreffen, der Betrieb der Hardware und Betriebssystem an einem Outsourcer übergeben wurde. Das schedulix Scheduling System benötigt für Installation, Administration und Betrieb keine solchen Privilegien. Damit ist ein schneller und kostengünstiger Zugriff auf alle Aspekte des Scheduling Systems stets gewährleitet.

Schlußbemerkung

Auch wenn in diesem Dokument nicht auf alle Aspekte und Features des schedulix Scheduling Systems eingegangen wurde, sollte ein Eindruck über die Leistungsfähigkeit des schedulix Scheduling Systems in Data Warehouse Umgebungen vermittelt worden sein. Für weitere Fragen steht Ihnen unser Team gerne zur Verfügung.