SQL SERVER INTEGRATION SERVICES VS AZURE DATA FACTORY

Xenia DAMANCIUC

Departamentul Ingineria Software și Automatică, grupa TI-191 F/R, Facultatea Calculatoare, Informatică și Microelectronică, Universitatea Tehnică a Moldovei, Chișinău, Republica Moldova

Autorul corespondent: Xenia DAMANCIUC, e-mail: damanciuc.xenia@gmail.com; damanciuc.xenia@isa.utm.md

Conducător științific: Dorian SARANCIUC, DISA, FCIM, UTM

Rezumat. *Pentru dezvoltarea unei aplicații de succes este necesar să avem pus la bază un sistem care este considerat sursa centralizată de adevăr pentru informația predispusă. Întru realizarea acestui scop se folosește așa numitul proces ETL, care semnifică „extrage, transformă și încarcă”. În cadrul acestui articol veți avea posibilitatea să înțelegeți care sunt avantajele și dezavantajele prin implementarea unui ETL utilizând soluția SQL Server Integration Services în comparație cu Azure Data factory.*

Cuvinte cheie*: date, serviciu cloud, serviciu on-premise, ETL*

Introducere

Fie că stăm la baza dezvoltării unei aplicații scopul căreia va fi să colecteze date despre angajații unei companii și să le ofere altor aplicații interne sau externe. Este important de menționat faptul că este necesar de integrat o multitudine de aplicații, care oferă seturile de date în diferite formate. Pentru realizarea acestei solicitări se va implementa o soluție, scopul căreia va fi să extragă datele de la aplicațiile-producător, ulterior să poată procesa toate formatele predispuse, iar ca rezultat va încărca rezultatul final într-un depozit de date. Acest proces este numit ETL sau „extrage, transformă, încarcă”. Astfel, există diferite tipuri instrumente de implementare a acestuia:[1]

- *open-source* – instrumente ce sunt gata pentru a fi integrate cu alte sisteme. Sunt potrivite pentru organizațiile cu buget restrâns.
- *de procesare în serie* – instrumente ce vor realiza procesele în situațiile în care puterea de calcul este mai liberă.
- *bazate pe tehnologii cloud* – instrumente care pot colecta date de la orice serviciu cloud și le pot încărca în depozitul de date solicitate prin optimizarea eficientă conform cerințelor sistemului Dvs.
- *în timp real* – instrumente care permit obținerea datelor actualizate în timp real. Indiferent de volum, înregistrările relevante pot fi mutate instantaneu în destinațiile potrivite.

SQL Server Integration Services (SSIS) este o platformă care vă permite să creați soluții de integrare și transformare a datelor la nivel de întreprindere. SSIS este o tehnologie de depozitare a datelor care poate fi utilizată pentru extragerea, încărcarea și transformările de date, cum ar fi curățarea, agregarea și combinarea datelor [2].

Datele pot fi extrase și transformate din mai multe surse, inclusiv fișiere de date XML, fișiere plate și surse de date relaționale, și apoi încărcate în una sau mai multe destinații folosind Integration Services.

Azure Data Factory este serviciul ETL cloud al Azure pentru integrarea datelor fără server și transformarea datelor. Oferă o interfață de utilizare fără cod pentru crearea intuitivă și monitorizarea și gestionarea pe un singur panou.

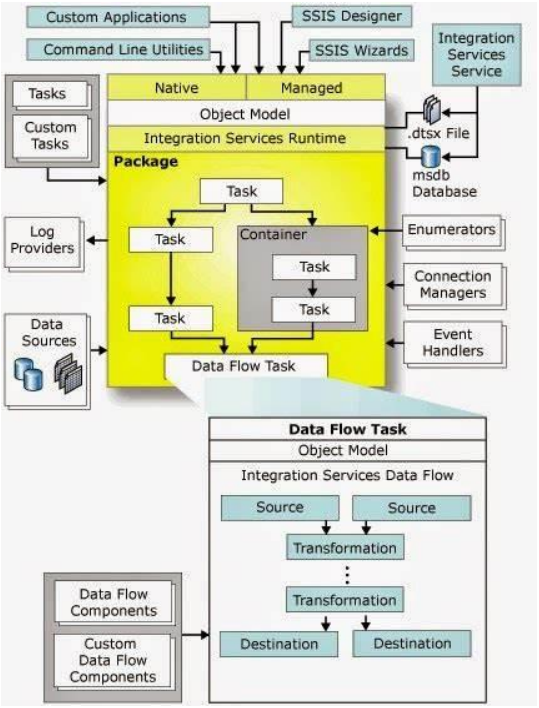
Prin urmare, vom studia avantajele și dezavantajele instrumentului de procesare în serie – SQL Server Integration Services (SSIS) în comparație cu instrumentul bazat pe tehnologie cloud – Azure Data Factory (ADF).

Părțile componente ale SSIS și ADF

SSIS este un instrument ETL (extract-transform-load). Este conceput pentru a extrage date din una sau mai multe surse, a transforma datele din memorie - în fluxul de date - și apoi a scrie rezultatele la o destinație.

SSIS are patru componente importante:

- *Flux de control* – Este partea de bază a pachetului SSIS. Aceasta stabilește ordinea tuturor componentelor, cum ar fi containerele, sarcinile etc.
- *Flux de date* – această componentă vă permite să extrageți, să transformați și apoi să încărcați datele într-o altă destinație.
- *Pachete* – colecție de control și flux de date. Containerele și sarcinile fluxului de date din fluxul de control și sursele, și destinațiile din fluxul de date sunt cunoscute în totalitate ca pachet.
- *Parametri* – Acestea sunt tipuri speciale de variabile. Ele ajută la ușurarea procesului de transmitere a valorilor de rulare către pachetele SSIS.



Figură 1 - Părțile componente ale unui SSIS

ADF, pe de altă parte, este mai mult un instrument ELT (extract-load-transform), adică pentru extragerea datelor dintr-o sursă și scrierea ulterioară într-o altă sursă. Există posibile transformări de date în timpul acestui transfer, cum ar fi conversia dintr-un format de fișier în altul, dar acestea sunt destul de limitate. Acesta poate fi comparat cu un flux de control în SSIS.

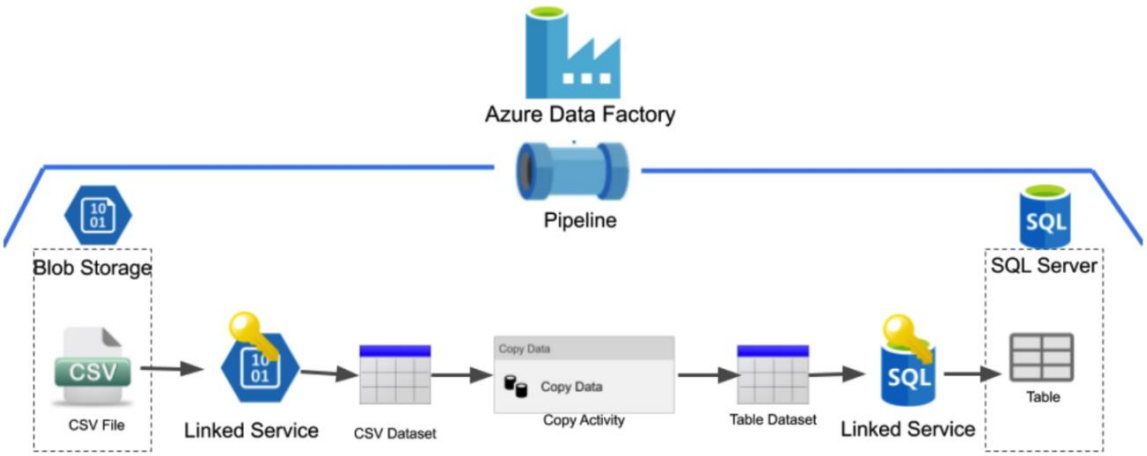
ADF la rândul său, de asemenea, constă din patru componente:

Pipeline – Conține sarcinile pe care doriți să le executați. Acesta definește fluxul de lucru complet, cum ar fi ce sarcini trebuie efectuate și în ce ordine.

Activitate – Aceștia sunt pașii individuali din interiorul unei pipeline, în care fiecare activitate îndeplinește o singură sarcină. Ele pot fi fie legate cu lanțuri, fie rulate în mod paralel și controlează fluxul în interiorul unui pipeline.

Seturi de date – Datele colectate sunt necesare ca intrare pentru procesul ETL. Acestea sunt vizualizări care reprezintă baze de date, fișiere sau mape.

Servicii legate – Acestea sunt practic șirurile de conexiune utilizate pentru a conecta sursele și serviciile de date și pentru a le autentifica.



Figură 2 - Părțile componente ale unui ADF

ADF vs SSIS

Ambele instrumente au descrieri similare, deoarece au fost create în același scop. Dar pentru a înțelege diferențele dintre ele, să le comparăm.

Curbă de învățare

Ambele instrumente sunt ușor de învățat și oferă posibilitatea îndeplinirii sarcinilor simple. Dar pentru a stăpâni orice software nou, este nevoie de mult timp și practică. SSIS este un software matur cu puține modificări majore în ultimii ani. De asemenea, acesta are la bază o documentație bine stabilită, precum probleme și răspunsuri comune deja analizate și discutate. Prin urmare, este ușor de învățat în comparație cu Azure Data Factory.

ADF, pe de altă parte, este încă în evoluție. Multe caracteristici și capabilități trebuie încă să fie lansate. În cazul apariției unei probleme este mai greu să fie găsită o soluție rapidă. Cu toate acestea, dacă aveți deja cunoștințe în SSIS, este mult mai ușor să-l învățați și să aplicați lucrurile deja cunoscute.

Viteza datelor

SSIS este un instrument ETL de procesare în serii. Acest lucru se realizează prin gruparea rândurilor care urmează să fie procesate în serii, apoi procesează fiecare serie sau pachet și actualizează fiecare grup așa cum este procesat.

ADF acceptă atât procesarea în serie, cât și procesarea fluxului (stream processing). Poate procesa date pe baza evenimentelor care au loc în conturile de stocare, cum ar fi ștergerea sau sosirea fișierelor. De asemenea, acceptă declanșatoare, unde puteți trece ora de început și de sfârșit pentru fiecare fereastră de timp din interogarea dvs.. Astfel, procesarea și returnarea datelor se va întâmpla între acel interval. În plus față de acestea, ADF acceptă și declanșatoare în serie.

Programabilitate

SSIS are un SDK de programare. Astfel, permite dezvoltatorilor să scrie propriul cod pentru definirea obiectelor de conexiune, sarcinilor, furnizorilor de jurnal și transformărilor. Dispune de un model de obiect programabil care permite dezvoltatorilor să creeze, să stocheze și să încarce pachete folosind BIML, precum și să creeze, să distrugă și să modifice oricare dintre obiectele conținute.

Pe de altă parte, ADF nu are un SDK de programare nativ, dar are automatizare folosind PowerShell fără a implica componente terțe. Pentru a rula conductele manual, puteți utiliza și metode precum .NET SDK, RestAPI-uri, Python SDK-uri.

Prețuri

SSIS vine ca parte a licenței SQL Server. Prețul este gratuit pentru edițiile Express și Developer, dar pentru Enterprise, costă 14.256 USD per nucleu.

Azure Data Factory oferă servicii la prețuri cu plata pe măsură. Acesta este calculat pe baza numărului de rulări de orchestrare a conductelor, a execuției fluxului de date și a depanării, precum și a numărului de operațiuni din fabrica de date, cum ar fi monitorizarea pipeline-ului.

Tabelul 1

Argumente PRO și CONTRA

Argumente/Serviciu	SSIS	ADF
PRO	• Capabil să gestioneze date dintr-o varietate de surse de date. • Oferă funcționalitate de transformare • C# sau VBA pot fi folosite extinderea funcționalului • Ușor de învățat și ușor de utilizat datorită documentației închegate • Capacitățile de depanare sunt excelente, în special în timpul execuției fluxului.	• Oferă o soluție fără server care elimină sarcinile banale de întreținere și actualizare a software-ului. • Suporta integrarea cu mai mulți conectori terți. • Acceptă interogări lungi și consumatoare de timp. • Foarte scalabil și rentabil. • Crearea pipeline-urilor este mult mai ușoară
CONTRA	• Raportul de execuție a pachetului poate fi văzut numai prin Management Studio. • Rularea mai multor pachete în paralel este dificilă. Evoluție lentă • Lucrul cu seturi de date nestructurate poate fi dificil.	• Mai puține funcții de transformare native în comparație cu SSIS. • Nu are instrumente de depanare. • Necesita menținerea unei strategii de facturare perfectă, altfel va duce la costuri excesive. • Lipsa flexibilității în comparație cu alte instrumente ETL, de ex. Componenta de script C#.

Concluzii

SSIS și ADF sunt ambele instrumente ETL foarte capabile. Fiecare poate reuși atunci când este utilizat corespunzător. Dacă volumul de lucru este în mare parte on-premises sau procesele ETL rulează în mod constant pe tot parcursul zilei, atunci SSIS este soluția potrivită. Dar dacă doriți să plătiți doar pentru resursele utilizate și cea mai mare parte a volumului de muncă este în cloud, atunci ADF este o alegere potrivită pentru dvs. Având în vedere factorii discutați mai sus și ceea ce este cerut de proiectul dvs., puteți decide care este instrumentul potrivit pentru lucrare, dacă SSIS, ADF sau un hibrid dintre cele două.

Bibliografie

1. Choosing Between SQL Server Integration Services and Azure Data Factory. [online] [accesat 22.11.2022]. Disponibil: https://www.mssqltips.com/sqlservertip/7094/azure-data-factory-vs-ssis-similarities-differences/
2. Comparing SSIS and Azure Data Factory. [online] [accesat 22.11.2022]. Disponibil: https://www.timmitchell.net/post/2020/07/16/comparing-ssis-and-azure-data-factory/