***Text prezentare DD***

//////////SLIDE 1//////////////ALEX

Pe vremuri, arhitectura depozitelor de date însemna că existau limite privind cantitatea de date care putea fi stocată și frecvența cu care aceste date puteau fi manipulate. Mai mult, instrumentele folosite pentru a popula un depozit de date local aveau capacitate limitată de a face față tipurilor de date complexe sau seturilor de date în flux, concepte care sunt acum predominante în peisajul aplicațiilor care alimentează depozitele de date. Întreprinderile necesită acum ca aceste surse să fie incluse în rapoartele lor, astfel încât depozitul de date trebuie să se modernizeze pentru a ține pasul. În prezent, Azure oferă multe instrumente și servicii pentru a ajuta la depășirea acestor probleme, dintre care multe pot fi integrate direct în ceea ce ar fi cunoscut acum ca un depozit modern de date.

În plus pe langa modernizarea bazei de date, instrumentele care operează, automatizează și completează depozitul de date trebuie, de asemenea, să țină pasul pentru ca soluția să se simtă coezivă. Acesta este motivul pentru care Azure oferă servicii excelente de integrare și automatizare care pot fi utilizate împreună cu tehnologiile bazei de date SQL. Prin aceste instrumente se poate obține mai mult cu mai puțin cod și mai putina confuzie, prin crearea de modele standard care pot fi aplicate generic la o varietate de probleme de procesare a datelor. Sarcinile obișnuite, cum ar fi copiile de rezervă ale bazelor de date, pot fi complet automatizate, ceea ce face ca problema recuperării în caz de dezastru să fie mai puțin îngrijorătoare. Cu cele mai recente caracteristici ale bazei de date SQL Azure, inteligența artificială este utilizată pentru a recomanda și aplica tuning și ajustări ale indecsilor pentru a asigura ca performanța bazei de date să fie la maxim. Acest lucru funcționează alături de detectarea avansată a amenințărilor, care asigură că bazele de date găzduite în Azure sunt mai sigure ca niciodată. În cele din urmă, companiile sunt din ce în ce mai interesate de big data și știința datelor, concepte care necesită atât prelucrarea unor cantități uriașe de date cât și menținerea unui grad bun de performanță. Din acest motiv, asa numitele lacuri de date au devenit mai populare și, mai degrabă decât să fie văzute ca un serviciu izolat, ar trebui privite ca un excelent însoțitor al depozitului modern de date.

Există câteva instrumente care vor face aceste tehnologii mai ușor de utilizat si în majoritatea cazurilor sunt compatibile cu mai multe platforme, cum ar fi Mac și Windows. Acestea sunt:

• Visual Studio: 2019 este versiunea curentă și este principalul (IDE) atunci când lucrați cu Azure și alte tehnologii bazate pe Microsoft.

• Instrumente de date Visual Studio SQL Server: acest supliment pentru Visual Studio oferă dezvoltatorilor posibilitatea de a crea proiecte de baze de date și alte proiecte legate de BI, cum ar fi Analysis Services.

• Microsoft Azure Storage Explorer: acest instrument ușor permite dezvoltatorilor să se conecteze la conturile de stocare în cloud și să le acceseze local pe computerul lor. Când lucrați cu lacuri de date, acest lucru poate fi foarte util.

• SQL Server Management Studio: dacă vă bazați pe un mediu Windows, atunci acesta este un instrument foarte puternic pentru monitorizare și gestionare a bazelor de date SQL pe care se merge de ani buni.

• Azure Data Studio: Aceasta este o versiune pe mai multe platforme a SQL Server Management Studio. În esență, acesta este locul ideal pentru gestionare și monitorizarea oricărui mediu Microsoft SQL.

/////////////SLIDE 2/////////ALEX

Platforma Microsoft Azure are o gamă largă de opțiuni de stocare a datelor la dispoziția utilizatorului, fiecare cu caracteristici și trăsături diferite care le fac potrivite pentru un anumit tip de date și scenariu. Având în vedere natura flexibilă și dinamică a cloud computing-ului, Microsoft a construit o platformă cuprinzătoare care asigură că toate varietățile de date pot fi satisfăcute. Recunoașterea necesității de a răspunde diferitelor tipuri de date se regaseste în ceea ce este cunoscut în lumea ingineriei datelor sub numele de „The 3 Vs” - volum, varietate și viteză. Orice combinație de volum, varietate și viteză poate fi rezolvată folosind o soluție de stocare în platforma Azure. Adesea, oamenii se referă si la un al patrulea V ca fiind „valoare”, ceea ce este un plus demn, deoarece valoarea se poate pierde adesea în volum. Pentru a utiliza un exemplu, dacă ar fi nevoie să stocați o cantitate mare de date de întreprindere care să fie un amestec complet de tipuri și dimensiuni de fișiere, ați utiliza un cont Azure Storage. Acest lucru vă va permite să vă organizați datele într-o structură clară și să vă măriți în mod eficient dimensiunea contului pe măsură ce aveți nevoie.

Trecându-ne la un depozit de date, știm că vom avea o cantitate mare de date bine structurate care trebuie să furnizeze rapid informații analitice. Pt asta avem nevoie de un motor SQL. În mod crucial, aici intră în joc al patrulea V, „valoare”. Seturile de date folosite pentru alimentarea unui depozit de date pot conține multe atribute care nu sunt deosebit de valoroase, iar bunele practici impun ca aceste atribute să fie tăiate înainte de a ajunge în depozitul de date. Regula de aur este că datele stocate într-un depozit de date trebuie să fie bine organizate și să aibă cea mai mare valoare. Un motor SQL face ca suprafața acestor date valoroase să fie ușoară și, în plus, nicio altă opțiune de stocare nu poate facilita îmbinarea seturilor de date pentru a produce valoare descoperită anterior la fel de ușor ca un motor SQL. Cu toate acestea, în procesul de decizie Azure se oferă două tipuri de motor SQL din care să alegem.

///////////SLIDE 3/////////////ALEX

Primul motor SQL pe care îl vom examina este Azure Synapse Analytics (fost Azure SQL Data Warehouse). Acest serviciu de procesare masiv paralelă (MPP) oferă scalabilitate, elasticitate și concurență, toate bazate pe motorul server Microsoft SQL. În timp ce Azure Synapse Analytics are o mulțime de optimizări care vizează stocarea datelor, există câteva motive pentru care a doua opțiune SQL, Azure SQL Database, poate fi mai potrivită. Baza de date SQL Azure este un sistem de tip OLTP care este optimizat pentru citiri și scrieri; cu toate acestea, are câteva caracteristici interesante care îl fac un candidat excelent pentru un mediu de depozitare de date. Apariția recentă a Azure SQL Database Hyperscale înseamnă că baza de date SQL Azure poate scala până la 100 TB și poate furniza noduri suplimentare de calcul numai în citire pentru a difuza date analitice.

/////////////////////////////////////////Slide5///////////DANI

**Arhitectura de Ingestie**

Pentru ca datele sa aiba valoare, trebuie sa fie mutate intr-un loc in care se poate utiliza aceasta valoare, iar metoda prin care sunt mutate datele utilizatorilor trebuie sa depinda de nevoile lor si de frecventa utilizarii datelor. Ingestia datelor reprezinta transportul datelor de la diferite surse intr-un mediu de stocare de unde pot fi accesate, utilizate iar mai apoi analizate.. Destinatia este de obicei un deposit de date, o baza de date sau un deposit de documente.

Extrage, transformã, si încarcã (ETL) este un proces în utilizarea bazei de date şi în special în depozite de date. Acesta descrie procesul de atragere si schimbare a datelor, astfel incat sa poata fi utilizate intr-un depozit de date analitice. Adesea acesta necesita numerosi pasi complexi, care implica curatarea datelor, transformarea datelor si integrarea lor. Astfel este necesar ca ETL-ul sa fie partitionat in straturi clar definite, care separa problemele de incarcare si curatare de problemele de transofrmare.

Primul strat este stratul brut in care datele nu sunt curatate sau filtrate. Acestea ajungand exact asa cum au plecat de la sursa.

Primul pas in mutarea datelor din formatul sursa bruta intr-un format curat, disponibil in depozitul de date ,este curatarea si standardizarea acestor date. Pirn curatarea datelor, se asigura ca inregisrarile necorespunzatoare nu sunt procesate in depozit.

Miscarea finala a datelor in sistem este de la locatia curata intr-un depozit de date analitice. Cerintele acestui strat necesita imbinarea, agregarea si interogarea datelor.

/////////////////////////////Slide6////////////////DANI

La inceputul unui proiect de depozitare a datelor, exista intodeauna o faza de planificare. O parte din aceasta faza ar trebui folosita pentru a intelege metodele prin care ajung datele si folosirea acestor cunostinte pentru a planifica modul in care aceste date vor fi introduse in depozit. Astfel ingerarea datelor poate fi efectuata in diferite moduri cum ar fi in timp real unde datele sunt extrase, procesate si stocate imediat ce sunt generate pentru luarea deciziilor pe moment, batch sau pe loturi cand ingestia are loc in loturi, datele fiind mutate la intervale programate, aceasta abordare fiind benefica pentru procesele repetabile, iar ultima fiind o combinatie a celor doua cunoscuta sub numele de arhitectura lambda. Cu siguranta ca exita mai multe metode dar acestea sunt cele mai populare.

////////////////////////////Slide7/////////////////////DANI

**Data Lake**

Termenul de data lake a fost folosit pentru prima data in anul 2010 de catre James Dixon care spunea ca un Data Lake este un depozit centralizat care permite stocarea tuturor datelor structurate si nestructurate la orice scara.

Azure Data Lake este o platforma cloud conceputa pentru a sprijini analiza datelor mari. Aceasta ofera stocare nelimitatea pentru orice tip de data de orice dimensiune. Azure vine cu trei forme de Data Lake:

Azure Data Lake Gen 1

Azure Data Lake Gen 2

Azure Storage

**Planificarea Enterprise Data Lake**

In mod obisnuit un Data Lake este utilizat pentru a stoca datele in forma cea mai bruta. Cu toate acestea, exista instrumente care pot realiza o curatare compexa in cadarul unui Data Lake.. Prin urmare este definit un director specific in acest scop, Iar odata ce datele ajung in acest director, acestea ar trebui sa fie imutabile. Dupa aceasta etapa pentru curatarea si transformarea datelor este folisit un motor SQL urmand ca mai apoi aceste date sa fie transferate si stocate.

//////////////////////////////////Slide7////////////////////////////DANI

//**Implementarea Enterprise Data Lake**

Dupa cum am spus si anterior, prima zona pentru datele de intrare ar trebui sa fie intr-un director brut al data Lake-ului. Traseul catre aceasta zona trebuie sa fie simplist si sa aiba o intretinere redusa, astfel incat sa existe putine bariere la intrare, iar datele sa fie rapid stocate in siguranta, intr-un loc de unde nu se vor pierde. In functie de modul in care se allege curatarea datelor, exista cateva considerente majore de evaluat, iar ruta catre aceasta zona poate varia foarte mult in functie de optiunile tehnologice ale platformei. Astfel exista mai multe moduri prin care se poate realiza curatarea:

Curatarea in cadrul unei baze de date folosind Azure SQL database si azure Synapse Analytics.(poza 171)

Curatarea in cadrul unui Data Lake folosind Spark SQL sau Python.

Curatarea in Azure Data Factory

Dupa curatarea datelor acestea sunt transferate catre un “depozit”. Acest depozit este zona in care datele curate sunt unite intree ele.

Sergiu

**(Slide 9) Capitolul 7: Logare, audit și reziliență**

Atunci când lucrurile nu merg bine în conducta de procesare a datelor, este posibil să nu fie vina platformei sau a dezvoltatorilor. Ar putea fi furnizorul sursă care își actualizează software-ul sau o pierdere intermitentă a conectivității la un serviciu Azure.

Toate defecțiunile pot avea capacitatea de a întrerupe depozitul de date și de a provoca pierderi de servicii a utilizatorilor.

În acest caz echipa de dezvoltare este obligața să știe cum se comportă platforma atunci când rulează fără probleme și, de asemenea, să înțelega cum se comportă atunci când lucrurile merg prost.

**1. Înregistrarea procesului de mișcare a datelor**

Baza de date Azure SQL este întotdeauna o recomandare care încorporează o schemă de înregistrare în aceeași bază de date care ar servi ca depozit de metadate.

Baza de date Azure SQL poate fi utilizată pentru fiecare activitate care ar trebui înregistrată pe platformă

Primele evenimente care vor fi înregistrate ar trebui să fie începutul și sfârșitul fiecărei conducte care se execută în Data Factory.

**2. Auditarea procesului de mișcare a datelor**

Există două cerințe principale de audit care permite masurarea aspectul normal al soluției. Acestea sunt

• Volumele de date: cantitățile de date care circulă prin platforma dvs.

• Timpi de procesare: Frecvența lucrărilor de ingestie și timpul necesar pentru a le finaliza

Timpii de procesare a auditului pot crea o imagine mai completă a capacității platformei. Atunci când utilizați Data Factory pentru a muta fizic datele, aceste informații de audit sunt colectate cu ușurință.

**3. Incorporarea rezilienței în procesul de mișcare a datelor**

Primul pas către reziliență este de a încorpora câteva verificări defensive de bază, permițând platformei să detecteze probleme.

Al doilea este atunci să poți acționa în mod autonom asupra acestor probleme, astfel încât să nu apară probleme din aval.

Azure Data Factory este utilă pentru o implementare rapidă și ușoară; sunt ușor limitative datorită informațiilor pe care le furnizează și a modului în care sunt afișate.

Aplicațiile logice care permit implementarea mai multor rezultate logice diferite la un anumit cod de eșec pot fi invocate folosind un apel API REST de la Data Factory.

**(Slide 10) Capitolul 8: Folosirea scripturilor și automatizării**

Un atribut comun al multor dezvoltatori este dorința de a face lucrurile rapid, consecvent și o singură dată. Pentru a răspunde acestei dorințe, scripturile și automatizarea sunt adesea folosite, deoarece oferă o metodă consistentă pentru a finaliza sarcini care apar în mod regulat.

**1. Puterea PowerShell-ului**

PowerShell este un limbaj de scripturi pentru administratorii de sistem și utilizatorii avansați care doresc să automatizeze rapid sarcinile comune în întreaga lor întreprindere.

**2. Scripturi utilizate în mod obișnuit**

Generarea de cod este un accelerator care permite proiectelor de depozite să iasă rapid de la sol. Există trei elemente care sunt necesare pentru a facilita o abordare de generare a codului:

• Contracte de date: tabele și proceduri SQL care dețin metadatele specifice entității

• Șabloane SQL: proceduri și tabele SQL predefinite care vor avea substituenți pentru înlocuirea textului

• Scriptul PowerShell: un script PowerShell care uneste celelalte două elemente pentru a crea numeroase implementări ale unui model în câteva secunde

**(Slide 11) Capitolul 9: Peste depozite de date moderne**

Privind peste depozitele de date moderne realizam ca există o gamă largă de produse BI pe piață, care oferă utilizatorilor o gamă largă de capabilități și vizualizări.

Sa luam Power BI deoarece acesta este instrumentul de vizualizare pentru orice date și, de asemenea, examinam alte produse Microsoft pentru date pe măsură ce părăsește depozitul de date, cum ar fi Azure Analysis Services și Azure Cosmos DB.

**1. Microsoft Power BI**

Microsoft Power BI (Power BI) este un produs de vizualizare a datelor.

Power BI folosește același motor analitic care este utilizat pentru Analysis Services, optimizând interogări analitice peste date tabulare utilizând procesarea în memorie.

Power BI este alcătuit din mai multe componente cheie:

• Power BI desktop: Instrumentul principal de dezvoltare pentru fișierele Power BI.

• Power BI report builder: Generatorul de rapoarte utilizat pentru crearea rapoartelor paginate.

• Power BI service: Portalul web pe care sunt publicate tablourile de bord și rapoartele.

**2. Azure Analysis Services**

Azure Analysis Services (AAS) este o implementare PaaS a produsului local care a venit ca parte a pachetului de instrumente de date SQL Server.

Azure Analysis Services oferă o capacitate de a scala modelul pentru a satisface cerințele de procesare și interogare.

**3. Azure Cosmo DB**

Cosmos DB este o bază de date NoSQL care oferă posibilitatea de a stoca documente JSON într-un mediu distribuit la nivel global, extrem de rezistent, care oferă 7 acorduri de nivel de serviciu de neegalat și timpi de latență extrem de mici, ceea ce îl face o platformă ideală pentru dezvoltarea web.