Architectuur in Beeld

Published on April 15, 2020

Jef Bergsma

Enterprise Architect with a focus on data and information

IT is geen ondersteunend proces meer

Data de kern is van een informatie verwerkende organisatie. Een bedrijfsproces verwerkt data met als doel informatie te ontsluiten waarmee beslissingen genomen worden die waarde creëren. Om van data, informatie te kunnen maken moet deze beschikbaar zijn en gebruikt kunnen worden. Hiervoor kennen we in het architectuurmodel met de functionele gebieden het functionele gebied Gegevens en besluiten persisteren.

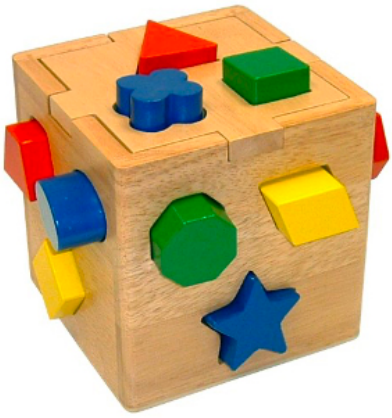
Het beschikbaar hebben van data is traditioneel het domein van databases waar de gegevens worden opgeslagen die door gebruikers in de verschillende transactiesystemen worden ingevoerd. Vanuit gebruikersperspectief is de data beschikbaar als de gegevens zijn ingevoerd. Vanuit technologisch perspectief is het invoeren van data maar een deel van het geheel. Data moet voldoen aan specificaties en passen in een voor gedefinieerd datamodel. Een database moet worden ingericht voor het juiste gebruik. Gaat het vooral om transacties, kleine wijzigingen door veel gebruikers met een hoge frequentie of gaat het meer om inlezen van grote datasets voor analyse doeleinden.

Data combineren

Zelfstandige datasets, behorende bij een transactiesysteem zijn meer en meer onderdeel van een groter dataset. Om informatie te kunnen geven wordt data gecombineerd met data uit andere bronnen, geanalyseerd en getoetst aan normen en regels om zo de informatie te kunnen bepalen waarmee de onderneming zijn waarde creatie kan realiseren. Welke producten gaan we promoten, welke relaties moeten we benaderen, wanneer gaan we over tot actie, hoe behalen we het hoogste rendement, etc..

Het samenbrengen van datasets kent vele uitdagingen. Een van de uitdagingen is om te komen tot een complete en consistente dataset. Aantallen in stuks, dozen en pallets mag je niet zomaar optellen om een totaalaantal te krijgen. Een waardebepaling van bezittingen in Azië, Europa en Amerika is niet alleen het optellen van de getallen, er moet bijvoorbeeld rekening worden gehouden met de geldende valuta.

Doorlooptijden geautomatiseerd bepalen voor processen die verschillende datum-notaties hanteren vraagt eerst om een datum-conversie. In tegenstelling tot mensen kan een computer deze omzetting niet impliciet in het achterhoofd doen maar moet hij hiervoor expliciet worden geïnstrueerd. De verschijningsvorm van de data, de datastructuur (relaties) en de syntax (eenheden en schrijfwijze) worden bepaald door de bron. Verschillende bronnen hebben verschillende verschijningsvormen. Als je de structuur van de bron kent en de verschijningsvorm van de dataelementen kun je ze inlezen en waar nodig omzetten in de structuur en de verschijningsvorm die nodig is.



De ene vorm is de andere niet

Structuur omzetten is relatief eenvoudig omdat het over een enkelvoudig dataelement gaat en hoe je die registreert. De datum 23 januari 2020 is eenvoudig om te zetten naar 23-01-2020 of 01/23/20 of welke vorm je dan ook nodig hebt. We spreken hier over een enkelvoudige transformatie.

Veel transformaties zijn complexer en vereisen meerdere enkelvoudige transformatie stappen, bijvoorbeeld een aantal aandelen in een Amerikaanse onderneming naar de waarde in Euro. Eerst moet je de waarde op een bepaalde datum bepalen, deze vervolgens vermenigvuldigen met het aantal aandelen en daarna het resultaat omrekenen naar de juiste

valuta waarvoor je de omrekenfactor moet hebben die dan weer afhankelijk is van een moment in de tijd. Daarbij kan de volgorde van de stappen variëren met hetzelfde resultaat.

Transformaties kunnen zeer complex worden waarbij meerdere transformatie regels op basis van meerdere gegevenssets moeten worden toegepast. Naast de data bron en het doelsysteem kunnen aanvullende datasets (referentie data) nodig zijn om de transformatie correct uit te kunnen voeren.

In de praktijk zien we veel handmatige, op Excel gebaseerde transformatie processen. Dit is vanuit architectuur geen structurele oplossing. Een op Excel gebaseerde transformatie is arbeidsintensief, persoonsafhankelijk, tijdrovend en foutgevoelig.

Dat vanuit een bron correct, snel en betrouwbaar overzetten naar een doelsysteem waar data gecombineerd wordt vereist een geautomatiseerde transformatie. Het inzetten van een tool beperkt de mensfactor tot het inrichten waarna het proces herhaald en betrouwbaar uitgevoerd kan worden. Het gehele proces duiden we aan als een ETL-proces. ETL staat daarbij voor:

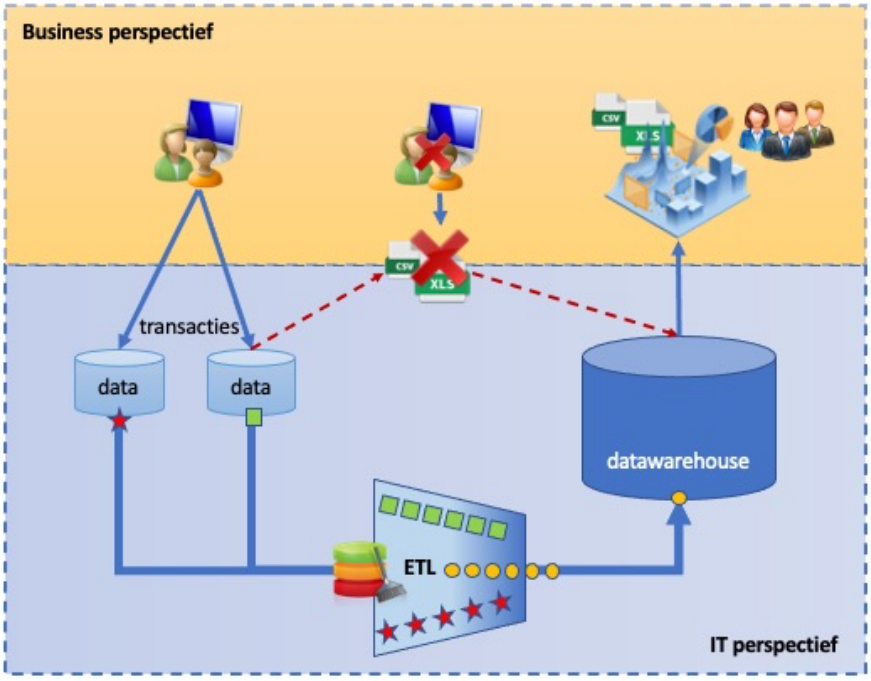
- *Extract; het uitlezen van de data uit de bron*
- *Transform; het uitvoeren van de transformatie van de bron structuur naar de doelstructuur*
- *Load; het opslaan van het resultaat in het doelsysteem.*

Verschillende datasets samenvoegen tot één grote dataset vereist een grote database waar het resultaat in opgeslagen kan worden. Hier ontstaat een totaalbeeld van de data waar een organisatie over beschikt, vaak inclusief historie. Deze dataverzamelplaats in een database noemen we een datawarehouse (DWH). Het datawarehouse vertegenwoordigt de Singel Point Of Truth, de zogenaamde SPOT. Van hieruit kunnen gebruikers analyses uitvoeren of de volledige werkelijkheid vanuit data perspectief of op een subset die relevant is voor de toepassing.

Centrale dataopslag

Een veel voorkomende oplossing voor het transformeren van data is de bron inlezen in Excel waar met allemaal formules en macro's de transformaties worden doorvoeren om het resultaat vervolgens in het DWH in te lezen vanuit het Excel bestand. Ook fouten in de brondata kunnen eenvoudig door de gebruiker handmatig of met formules worden hersteld.

Een architectuur voor een betrouwbaar DWH staat dit echter niet toe. Excel en handmatige handelingen zijn arbeidsintensief, vaak afhankelijk van een beperkt aantal personen en vooral foutgevoelig. Daarbij zijn de toegepaste transformatieregels vaak ondoorzichtig en is het beheer slecht geregeld. Dit staat op gespannen voet met het uitgangspunt dat Data in een DWH gevalideerd betrouwbaar moet zijn.



ETL proces als betrouwbare oplossing voor Excel transformaties

Het DWH moet voorzien in een informatiebehoefte. Als het datawarehouse niet op een gecontroleerde manier wordt gevuld met de juiste data gaat een gebruiker op zoek naar een eigen oplossing… Excel! Om dit te voorkomen voorzien we in de architectuur in een ETL-oplossing waarmee de transformaties centraal geregeld worden, het geheel onder beheer wordt geplaatst en de kwaliteit en continuïteit gewaarborgd is. Op deze manier verdwijnt de behoefte aan eigen Excel transformaties.

Gebruik van data in het datawarehouse is op basis van read only. Hiermee wordt voorkomen dat er onbedoelde wijzigingen in het datawarehouse worden gedaan. Ook het zelfstandig toevoegen van datasets (al dan niet geproduceerd met Excel) is dan niet mogelijk. De architectuur moet het mogelijk maken dat de gebruiker data visualisaties kan maken met tijdelijke databronnen waaronder Excel. Visualisatie behoort tot het functionele gebied Gegevens verwerken tot informatie en wordt daar verder uitgewerkt.

Een DWH-principe is dat data er wel in opgenomen wordt maar dat deze niet wordt aangepast of verwijderd. Daarmee wordt een representatie van de datahistorie opgebouwd van systemen die hun data met het datawarehouse delen. De werkelijkheid is echter complexer omdat data bewaartermijnen kent waarna het verwijderd of geanonimiseerd moet worden. Dit geldt natuurlijk behalve voor de transactie systemen ook voor het datawarehouse.