Stuurbare productiefabriek is effectief en drukt kosten

ETL aan de lopende band

Remko Wijnmaalen

De huidige ambachtelijke constructie van datalogistiek onder de meeste datawarehouse-oplossingen blijkt in de praktijk diverse problemen te veroorzaken. Zowel beheer als veranderprojecten zijn vaak te duur, nemen teveel tijd in beslag en leveren niet de kwaliteit op die verwacht wordt.

De voor de datalogistiek geschikte ETL-tools zijn tegenwoordig zeer krachtig en eenvoudig te gebruiken, maar ze worden in de meeste gevallen gebruikt als een soort 3GL programmeertaal. Dat betekent dat een te bouwen datalogistiekfunctie veelal op puur ambachtelijke wijze door IT'ers op maat ontworpen en gebouwd wordt.
De kracht van moderne ETL-tools komt pas tot zijn recht als we de datalogistiek vanaf de basis meer volgens een fabrieksmodel inrichten. Hierbij wordt een groot deel van het mechaniek van datalogistiek als een lopende band gestandaardiseerd, waarbij flexibiliteit wordt gewaarborgd door de uitgevoerde functionaliteit te sturen door middel van metadata. Dit valt te vergelijken met de huidige opzet van autofabricage, waarbij de robots die deel uitmaken van de lopende band een standaard functie uitvoeren met parameters voor de specificaties per individuele auto. Met andere woorden; de efficiëntievoordelen van een gestandaardiseerd productieproces worden zoveel mogelijk gecombineerd met flexibilisering van de individuele productiefuncties. Het is hoog tijd om de onnodig hoge maak- en beheerkosten te verlagen, de doorlooptijd van verandering te beperken en de flexibiliteit te vergroten, door de potentie van de moderne ETL-tools daadwerkelijk zichtbaar te maken.

Projecten binnnen een DWH-omgeving

Als een nieuwe behoefte rond managementinformatie opkomt, wordt meestal een nieuw project opgetuigd om op ambachtelijke wijze te voldoen aan het verzoek. Van inhoudelijke specificatie tot en met het uiteindelijke resultaat, maar ook zelfs de organisatie van het project zelf; alles is erop gericht om een maatwerkoplossing te realiseren. Het begint met de specificaties van wat er gebouwd moet worden. Deze zijn in meerdere vormen en detailniveaus beschikbaar. Denk bijvoorbeeld aan een functioneel ontwerp, business rules, een technisch ontwerp en uiteindelijk de programmacode zelf. Iedere vertaalstap voegt wel steeds meer detail toe, maar vergt ook overdracht van kennis, is tijdrovend, foutgevoelig en kent veel overlap. Als de ETL-programmatuur ook nog op maat blijkt te worden gebouwd is het arbeidsintensieve traject compleet. Daar komt nog bij dat de vertaalstappen vereisen dat verschillende disciplines nauw samenwerken.
De kenmerkende problemen bij deze praktijkaanpak zijn:

- Lange doorlooptijd. Misschien wel het grootste probleem is dat datawarehouseprojecten te lang duren. Door de ambachtelijke opzet van projecten kan de gemiddelde IT-afdeling veelal niet voldoen aan de totale vraag naar verandercapaciteit voor het opleveren van managementinformatie. Dit laatste heeft dan vaak tot gevolg dat eindgebruikers hun eigen oplossing verzinnen; wat de consistentie van managementinformatie niet ten goede komt;
- Hoge kosten. De arbeidsintensieve werkwijze betekent hogere kosten, onder andere doordat hergebruik van ETL-code zeer beperkt is. In plaats van het ontwikkelen en gebruiken van een bibliotheek van herbruikbare functionele modules, worden oplossingen voor standaard problemen steeds opnieuw gebouwd;
- Starheid. Het doorvoeren van veranderingen in bestaande ETL-code is lastig, omdat diepgaande kennis van de programmatuur schaars of zelfs niet meer aanwezig is. Vaak moet er een beroep worden gedaan op een kleine groep experts die toevallig bij voorgaande projecten betrokken waren. Het bijhouden van de documentatie, vooral bij productieproblemen, wordt nog wel eens overgeslagen, met als gevolg dat de documentatie na verloop van tijd nauwelijks nog toegevoegde waarde heeft;
- Te lage kwaliteit en servicegraad. Eenmaal opgeleverd blijkt dat de kwaliteit van de oplossing nogal eens tegenvalt. Van de geleverde managementinformatie is maar lastig uit te leggen of de inhoud ook echt klopt. Als de eindgebruiker vragen heeft over onverwachte waarden beginnen de problemen voor de beheerorganisatie pas echt. Voor maatwerkoplossingen ontbreekt namelijk vaak een eenduidige aanpak voor datakwaliteit en de gerealiseerde oplossing is vaak afhankelijk van de kunde en vindingrijkheid van de ontwerper of programmeur.

Beheer binnen een DWH-omgeving

Ieder ambachtelijk project draagt bij aan het vergroten van het aantal maatwerkcomponenten en hun bijbehorende documentatie

en verhoogt daarmee ongewild de complexiteit van de datawarehouse-omgeving als geheel. De beheerorganisatie is verantwoordelijk voor een stabiel productieproces waarbij het datawarehouse en de datamarts voorspelbaar, tijdig en correct gevuld worden. De toenemende hoeveelheid en complexiteit van maatwerk ETL-functies maakt dit steeds moeilijker te waarborgen.
Vragen van eindgebruikers zijn na verloop van tijd steeds lastiger te beantwoorden. Het ontbreekt beheerders vaak aan overzicht, bijgewerkte documentatie, toegang tot metadata en vanaf de basis ingebouwde middelen om problemen te analyseren en te verhelpen.
De beheerorganisatie is ook een van de belangrijkste partijen die zijn betrokken bij een project. Toch is de aandacht die de projectorganisatie aan toekomstig beheer besteedt vaak gering. Het beheerbaar maken en houden van de oplossing is vaak ondergeschikt aan het projectresultaat: de oplevering van management-informatie. Beheerders moeten het te vaak stellen met het alleen maar becommentariëren van door het project opgeleverde speci-ficaties.
Kortom, door de huidige maatwerk ETL-constructie en ambachtelijke projectaanpak raken beheerders veel tijd kwijt aan het volgen van projecten, verliezen ze het overzicht over de data-warehouse-omgeving en gaat de datakwaliteit langzaam achteruit door gebrek aan mensen en middelen om problemen te analyseren en te verhelpen. Uiteindelijk hebben deze problemen een negatief effect op het vertrouwen dat de organisatie heeft in het hele datawarehouse.

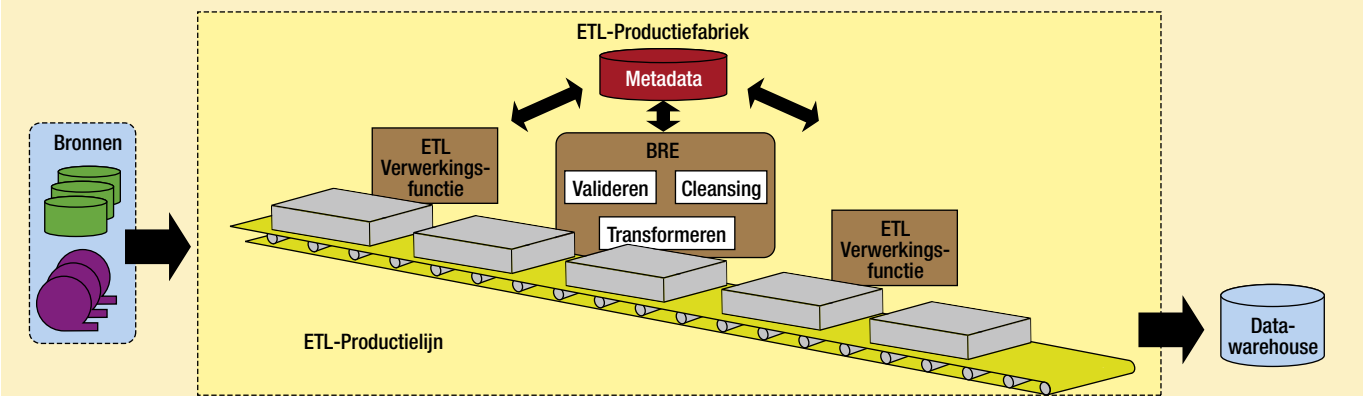
Fabriekconcept

Om verandering in de situatie te brengen moet het roer flink om. Dit zal alleen lukken als de datalogistiek rond het datawarehouse in de basis anders wordt opgezet. Zoveel mogelijk van het arbeidsintensieve maatwerk in de ETL-functies moet worden vervangen door een meer gestandaardiseerde productielijn met flexibel stuurbare productiefuncties. Dat is ook goed mogelijk, want bij de verwerking van gegevens naar het datawarehouse worden vaak vergelijkbare stappen uitgevoerd in min of meer dezelfde volgorde. De datalogistiek kan worden opgedeeld in functionele componenten die ieder bijdragen aan de mechanisatie van het ETL-proces van bron tot datawarehouse en van datawarehouse tot datamart. Gelukkig zijn de in de markt verkrijgbare ETL-tools ook zover dat deze aanpak prima te implementeren is: een ETL-productiefabriek, zie afbeelding 1.

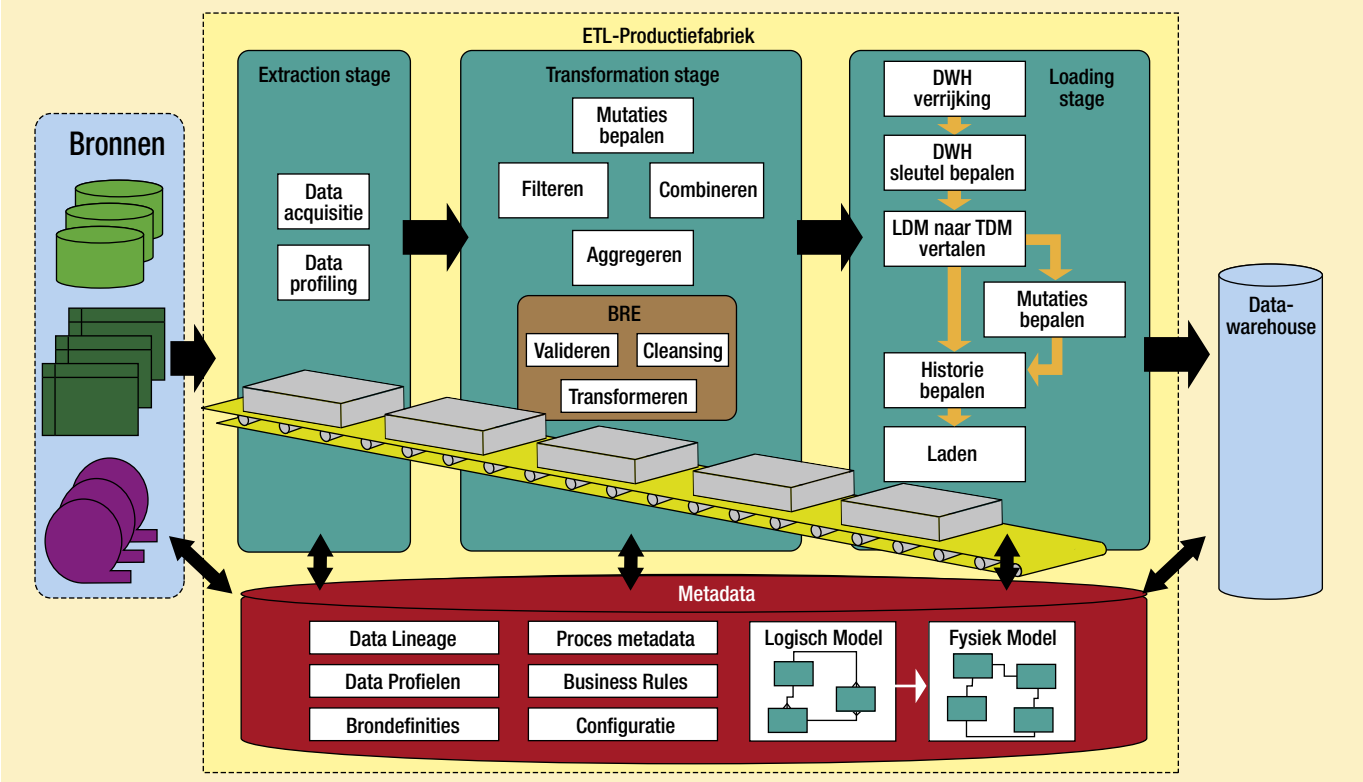
De opbouw van de ETL-productiefabriek kent de volgende basisonderdelen:
ETL-productielijn. Eén of meerdere productielijnen ('lopende band') die het mechaniek vormen van de datalogistiek onder het datawarehouse, voor de verwerking van gegevens zowel tussen bron en datawarehouse, als tussen datawarehouse en de diverse datamarts. Iedere productielijn bestaat uit een gestandaardiseerd verwerkingspatroon van functionele componenten waarmee uiteenlopende bronnen kunnen worden verwerkt, zonder dat het business logica bevat. De verwerking van de meeste bronnen kan vaak met één standaard patroon worden afgedekt. Voor de verwerking van de resterende bronnen kunnen eenvoudig aanvullende patronen worden gedefinieerd die specifieke behoeften afdekken. Binnen een patroon kunnen de functies eenvoudig worden geconfigureerd of zelfs aan en uit worden gezet. Het mechaniek van de fabriek blijft dus hetzelfde terwijl het proces per bron toch flexibel in te richten is.
ETL-verwerkingsfuncties. De productielijn voert langs verschei-dene 'machines' of 'robots' die ieder een stuk ETL-functionaliteit bieden. De lopende band kan diverse vertakkingen hebben die later weer samenkomen, maar de volgorde van de functies die wordt geboden staat in grote lijnen vast.
Metadata. Dit is de stuurinformatie voor de productielijn en de business logica voor haar ETL-functies. De metadata bestrijken alle specificaties die nodig zijn om de ETL-productiefabriek aan te sturen, zoals brondefinities, informatie over het logische en technische datamodel en business rules.

Opdeling ETL-productielijn in 'stages'

De ETL-productielijn is gewoonlijk ingericht in een drietal achtereenvolgende, functionele verwerkingsgebieden ('stages'), zie afbeelding 2. De stages worden gekenmerkt door het verschil in functionaliteit en de benodigde mate van flexibiliteit van het mechanisme.



Afbeelding 1: Overview van de ETL-productiefabriek.



Afbeelding 2: De drie stages in de ETL-productiefabriek.

Extraction stage. Als eerste dient een extractie-stage tot doel om data van de bronsystemen te ontvangen. In deze stage vinden nog geen bewerkingen plaats. Het ETL-verwerkingspatroon zelf is hierbij vrijwel altijd gelijk, maar staat nog wel enige variatie toe. De functionele componenten binnen het verwerkingspatroon worden met minimale verschillen in configuratie en volgorde gebruikt. Data-acquisitie is de eerste ETL-verwerkingsfunctie die nodig is in de extraction stage. Deze functie regelt het transport van data vanuit de omgeving van het bronsysteem en de ontvangst van data in de omgeving van het datawarehouse. Eventueel kan de functie nog aanvullende diensten leveren zoals het archiveren van meerdere versies van brondata.
Een tweede functie die veelal tot de extraction stage hoort is data profiling. Deze functie bepaalt de karakteristieken van brongegevens. Dit kan worden gedaan op verschillende detailniveaus en kan variëren van statistieken en tellingen tot het vaststellen van voorkomende waarden en de frequentie waarmee ze voorkomen. De dataprofielen kunnen daarnaast worden opgeslagen om inzicht te krijgen in het historisch 'gedrag' van een bronsysteem.

Transformation stage. De tweede stage bevat functies die per bron specifieke transformaties uitvoeren. Het patroon waarmee deze functies de ETL-productielijn vormen, is in deze stage het meest complex en vereist de meeste flexibiliteit. Dit deel van de fabriek bevat de kernfuncties die zorgen voor het integreren, combineren, herstructureren, berekenen, omvormen en afleiden van gegevens uit de bron. Voorbeelden van functies zijn het bepalen van mutaties, valideren en transformeren.

In deze stage van de fabriek kan prima een Business Rule Engine (BRE) worden ingezet. De BRE ondersteunt de definitie, de uitvoering en het beheer van regels die kunnen worden gebruikt om te valideren, te cleansen of te transformeren. De BRE maakt als goed afgebakende functie onderdeel uit van de ETL-productielijn. De BRE kan ook gebruikt worden door analisten, beheerders of zelfs eindgebruikers en zorgt er ook voor dat gebruikers de regels zelfstandig kunnen testen. Daarmee wordt de verantwoordelijkheid voor deze regels duidelijk belegd waar het thuis hoort: bij de business en niet bij IT.
Loading stage. De derde stage heeft tenslotte tot doel om de getransformeerde gegevens vast te leggen in het datawarehouse inclusief aanvullende metadata. De functies in de loading stage kennen in principe geen afhankelijkheid meer met een specifieke bron.
Denk bijvoorbeeld aan een functie die de data verrijkt met het moment van verwerking of met metadata waarmee de exacte bronaanlevering kan worden geïdentificeerd. Een tweede voorbeeld is een functie die het mogelijk maakt om objecten met een natuurlijke sleutel terug te vinden in het datawarehouse en een bijbehorende betekenisloze sleutel te bepalen. Andere functies in deze stage zijn nodig voor het correct vastleggen van historie, het bepalen van verschillen in gegevens ten opzichte van het datawarehouse en het laden van gegevens.

Gebruik metadata in ETL-productiefabriek

Voor de fabriek is een goede metadata repository van cruciaal belang. De fabriek bestaat weliswaar uit functionele componen-

ten die volgens een bepaald patroon aan elkaar gerelateerd zijn, maar het exacte gedrag van iedere functie wordt bepaald door stuurinformatie in de vorm van metagegevens, bijeengebracht en geïntegreerd in een repository.
De fabriek heeft verschillende soorten metadata nodig. Voorbeelden van relatief statische metagegevens zijn brondefinities, logische en fysieke gegevensmodellen van het datawarehouse en business rules. Een deel van deze metagegevens kan direct worden geïmporteerd uit *case tools*. Andere metadata zijn afkomstig uit bijvoorbeeld een BRE. Ook voor de configuratie van de productielijn van de fabriek moet een beperkte hoeveelheid metadata worden toegevoegd. Zo kan via metadata per bron het patroon van de productielijn nog worden geconfigureerd door bepaalde delen aan of juist uit te zetten. Ook is het mogelijk dat voor bepaalde functies nog een standaard instelling wordt gewijzigd.
Naast relatief statische metadata genereert de fabriek zelf ook metadata tijdens het uitvoeren van processen. Denk bijvoorbeeld aan productiemetadata over het verwerkingsproces zelf. Eenmaal gevuld voorziet de repository in veel van de informatiebehoeften van verschillende partijen. Alle soorten metadata zijn eenvoudig toegankelijk, zijn aan elkaar gerelateerd en zijn per definitie niet verouderd. Door van alle metadata ook nog versies bij te houden kan tevens eenvoudig worden voorzien in *data lineage* behoeften. Hierdoor wordt beheer een stuk eenvoudiger.

Projecten in ETL-productiefabriek

Het bouwen van een fabriek heeft vele voordelen. Daar staat uiteraard wel het nadeel van de initiële investering voor de bouw tegenover. Die investering kan echter wel snel weer worden terugverdiend, doordat de ETL-productielijn telkens volledig wordt hergebruikt en ook niet steeds opnieuw hoeft te worden getest tijdens ieder project. Daar komt nog bij dat er geen echte programmacode meer wordt ontwikkeld. De enige programmacode is immers de ETL-productielijn zelf. In plaats daarvan ligt de focus op het vastleggen van business logica en het uitvoeren van eenvoudige configuratieactiviteiten via de metadata repository. Hiermee wordt gelijk in de documentatiebehoefte voorzien. Het werk binnen projecten is hierdoor zeer efficiënt en tot een minimum beperkt, wat zorgt voor een korte doorlooptijd. Maar het gebruik van de fabriek resulteert ook in kosteneffectieve projecten. De metadata repository voorkomt de overhead van het opstellen van stapsgewijze documentatie tot programmacode.
De fabriek heeft als voordeel dat het mechaniek ervan een hoge mate van hergebruik introduceert, hetgeen de veranderkosten verder vermindert. Toch kan het voorkomen dat bepaalde bronnen zulke specifieke eisen stellen dat er aanvullende patronen van de fabrieksfuncties moeten worden gedefinieerd. Dit nadeel blijft echter beperkt omdat de verwerking van het merendeel van de bronnen met één patroon kan worden afgedekt.
Versiebeheer van de metadata zorgt ervoor dat oude versies van functionele componenten en business logica beschikbaar en benaderbaar zijn en blijven. Dit heeft als voordeel dat het inzichtelijkheid bevordert en laat potentieel ook toe dat oude versies van componenten en business rules worden gebruikt. De metadata repository heeft ook als bijkomend voordeel dat veranderingen eenvoudig en snel kunnen worden doorgevoerd. De ETL-productielijn hoeft hierbij niet te worden aangepast. Wijzigingen van brondefinities, business rules, configuratie of zelfs het business model van het datawarehouse hebben *geen impact* op de ETL-productielijn. Het testen van aanpassingen kan ook in veel gevallen beperkt blijven tot een lokale inspanning. Alleen indien er ingrijpende veranderingen worden doorgevoerd op functionele componenten van de fabriek zullen regressietesten noodzakelijk zijn.

Beheer in ETL-productiefabriek

De fabriek heeft verder nog als voordeel dat processen eenduidig zijn gedefinieerd voor alle bronnen die worden aangesloten op het datawarehouse. Vanaf het begin worden standaard voorzieningen ingebouwd voor validatie, cleansing, reconciliatie, auditability en de verwerking van kwalitatief slechte data. Deze voorzieningen zijn eenduidig en direct beschikbaar voor alle bronnen die worden verwerkt en genereren de bijbehorende metadata. Door toegang tot de geïntegreerde metadata kunnen beheerders vragen van eindgebruikers veel eenvoudiger beantwoorden. De metadata repository geeft immers inzicht in welke bronaanlevering is verwerkt en hoe, welke versie van validatie-, cleansing- en business rules zijn gebruikt en hoe de ETL-productielijn was geconfigureerd. Dit komt de dienstverlening en het vertrouwen in de kwaliteit van gegevens in het datawarehouse ten goede.
De standaardisatie van de ETL-productielijn heeft ook een gunstig effect op de beheerkosten. Alle business logica en de configuratie van de fabriek zijn via de geïntegreerde metadata beschikbaar. De samenhang tussen bronnen, business rules, het datawarehouse en alle verwerkingsprocessen blijft hierdoor inzichtelijk en overzichtelijk. Het gevolg is dat het beantwoorden van vragen van eindgebruikers, de analyse van problemen en het verhelpen ervan effectief en efficiënt kan plaatsvinden.

Conclusie

Het is hoog tijd om te stoppen met de ambachtelijke werkwijze waarbij voor veel managementinformatieverzoeken, maatwerk-ETL wordt geleverd. Een oplossing hiervoor is het eenmalig bouwen van een gestandaardiseerde, maar stuurbare ETL-fabriek met behulp van een modern ETL-tool. Ondanks de initiële investering brengt dit voordelen met zich mee die anders niet kunnen worden behaald. Een fabriek zorgt voor efficiënte en effectieve projecten met een korte doorlooptijd en lagere kosten. Tegelijkertijd brengt het de kwaliteit en beheerbaarheid van de datalogistiek op een hoger niveau. Hierdoor wordt het mogelijk om sneller en beter te voldoen aan de behoefte naar hoogwaardige managementinformatie; waar het de eindgebruiker tenslotte om gaat.

Remko Wijnmaalen is DWH & BI consultant bij FourPoints Intelligence.