STREAMLINING ANALYTICS

JAN MESKENS – DRIES VAN DROMME

Smals SECTIE ONDERZOEK 12/2013 • RESEARCH NOTE • 30

Abstract - Deze researchnote vat de infosessie 'Streamlining Analytics' samen, gepresenteerd in juni 2013. Streamlining Analytics bespreekt drie barrières die het introduceren van analytics in bestaande organisaties bemoeilijken: (1) een trage en complexe architectuur, (2) een niet-optimale aanpak van dataquality-problemen en (3) een atypisch projectkader waarbinnen analyticsprojecten uitgevoerd worden. Voor elk van deze barrières reiken we mogelijke oplossingen aan. Zo bespreken we manieren om een complexe data-architectuur te vereenvoudigen en methodes om dataquality-aspecten te behandelen binnen analyticsprojecten. Voorts behandelen we ook project frameworks die het mogelijk maken om de nodige resources en tijd van typische analyticstaken beter in te schatten.

Résumé - Cette Research Note résume la séance d'info 'Streamlining Analytics', présentée en juin 2013 et traitant de trois obstacles qui compliquent l'introduction de l'analytique dans des organisations existantes : (1) une architecture lente et complexe, (2) une approche non optimale des problèmes de qualité des données et (3) un cadre de projet atypique dans lequel sont exécutés les projets d'analytique. Pour chacun de ces obstacles, nous procurons des solutions possibles. Nous abordons ainsi des moyens de simplifier une architecture de données complexe et des méthodes pour traiter les aspects de qualité des données dans les projets d'analytique. De même, nous présentons des cadres de projet permettant de mieux estimer les ressources et le temps nécessaires pour des tâches d'analytique typiques.

Inhoud

1. Inleiding

In deze studie werd nagegaan hoe analyticsprojecten kunnen *gestroomlijnd worden*, in het bijzonder bij organisaties waar de techniek nog relatief nieuw is. Het stroomlijnen van analyticsprojecten richt zich voornamelijk op twee deelaspecten: het vlot inpassen van de technologie in een klassieke data-architectuur en het sneller opleveren van structurele resultaten dankzij het overwinnen van enkele typische barrières. Voor we hier dieper op ingaan introduceren we twee kernbegrippen: *predictive analytics* en de *data supply chain*.

1.1. Predictive analytics

Predictive analytics (PA) is een techniek die toelaat om gegevens te analyseren met als doel nieuwe inzichten te verwerven en op geautomatiseerde wijze modellen op te stellen, die ingezet kunnen worden om voorspellingen te maken over de toekomst. Dit proces verloopt meestal in twee fasen:

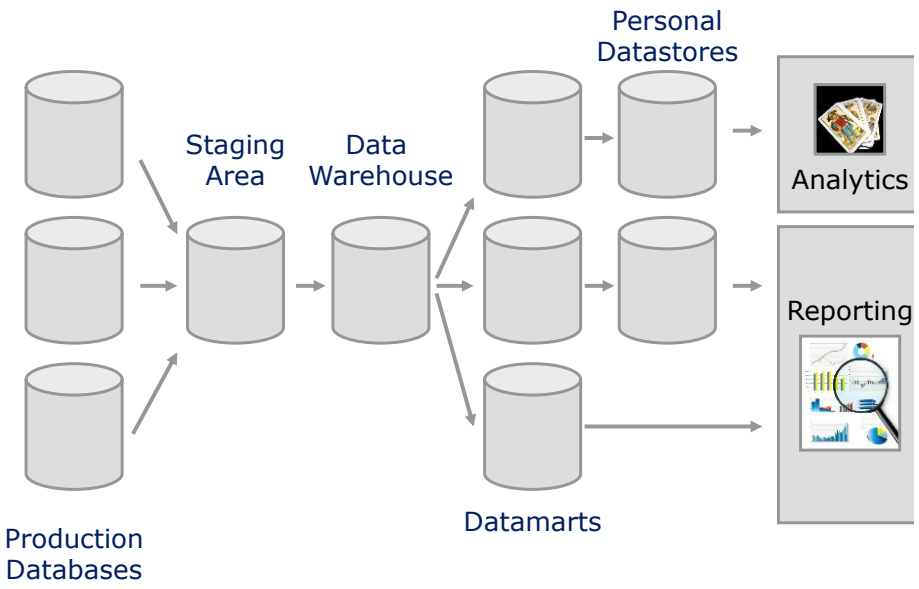
(1) er wordt een *predictief model* getraind op basis van de huidige data;

(2) dit model wordt geconfronteerd met nieuwe data met als doel deze data te classificeren volgens de lessen die getrokken zijn uit de trainingsfase.

Bovenstaand schema wordt al geruime tijd toegepast in heel diverse sectoren en toepassingsgebieden. In de financiële sector zijn er predictieve modellen getraind om mogelijke fraude met kredietkaarten te herkennen en automatisch kaarten te blokkeren wanneer het vermoeden sterk genoeg is. Binnen het domein van de interactieve marketing tracht men aan de hand van het koopgedrag van klanten in het verleden te voorspellen wat nieuwe klanten zouden willen kopen. Zo probeert men gericht advertenties te tonen aan potentiële kopers van bepaalde producten. De techniek kan ook beleidsondersteunend ingezet worden en vormt in die zin een pijler van moderne oplossingen voor business intelligence.

1.2. De data supply chain

Om een PA-project te doen slagen, is een belangrijke voorwaarde dat men over de juiste data kan beschikken. Deze data bevindt zich meestal in de productie-omgeving, waar de databanken rechtstreeks in verbinding staan met de eindgebruikerstoepassingen. PA-algoritmes loslaten op deze productiedatabanken is echter geen goed idee. Deze algoritmes zorgen namelijk vaak voor een explosie van IO-operaties of query's die de stabiliteit van de databank, en bijgevolg de eindgebruikerstoepassing, in gevaar kunnen brengen.



Figuur 1: De data supply chain

Om bovenstaande problemen te vermijden, wordt er vaak gebruikgemaakt van een "datasupplychain"-architectuur. Dit is een ketting van verscheidene databanken waar men data kopieert van een brondatabank naar een volgende in de ketting d.m.v. Extract-Transform-Load (ETL) scripts. **Figuur 1** geeft een typische architectuur van een data supply chain weer.

Centraal in dit schema staat het Data Warehouse (DWH), een databank die alle gegevens van de productiesystemen verzamelt in één globaal dataschema. Vanuit het DWH wordt data verdeeld over verscheidene *Datamarts*. Dit zijn kleinere databanken die slechts een subset van de gegevens uit het DWH bevatten en (doorgaans) één departement of reporting/analyticsapplicatie ondersteunen. Analytics- en reportingapplicaties maken ook vaak gebruik van zogenaamde *Personal Datastores*. Deze databanken bevatten data in een zodanig getransformeerde vorm zodat ze makkelijk te analyseren zijn door een welbepaald analyticsalgoritme. Deze trage en complexe architectuur komt zo veelvuldig voor in organisaties met enige geschiedenis, omdat de

performantie die met de technologie destijds haalbaar was geen andere mogelijkheid liet.

2. Barrières bij de introductie van analytics

De introductie van predictive analytics is meer dan alleen maar het toepassen van een welbepaald algoritme op een welgekozen dataset. Er moet ook nagedacht worden over hoe deze techniek kan toegepast worden binnen een organisatie en hoe PA past in bedrijfsarchitecturen. Concreet zijn er drie belangrijke barrières die de introductie van analytics bemoeilijken:

(1) analytics moet vaak ingebed worden in **een trage en complexe architectuur** zoals bv. de data supply chain. Data wordt in zulke architecturen meerdere malen gekopieerd en gedupliceerd waardoor analyticstoepassingen, die zich op het einde van de ketting bevinden, met veel vertraging pas de benodigde data verkrijgen;

(2) het **progressief verbeteren van datakwaliteit** doorheen de data supply chain is common practice in de meeste organisaties. Dit bemoeilijkt echter het invoeren van analytics, omdat tijdens deze incrementele dataquality-operaties data wordt gewijzigd, vaak zonder de nodige documentatie of zonder master data management. De betekenis van deze nieuwe data is vervolgens moeilijk te achterhalen door de PA-analist en zijn businessgesprekspartner, wat de interpretatie van resultaten bemoeilijkt;

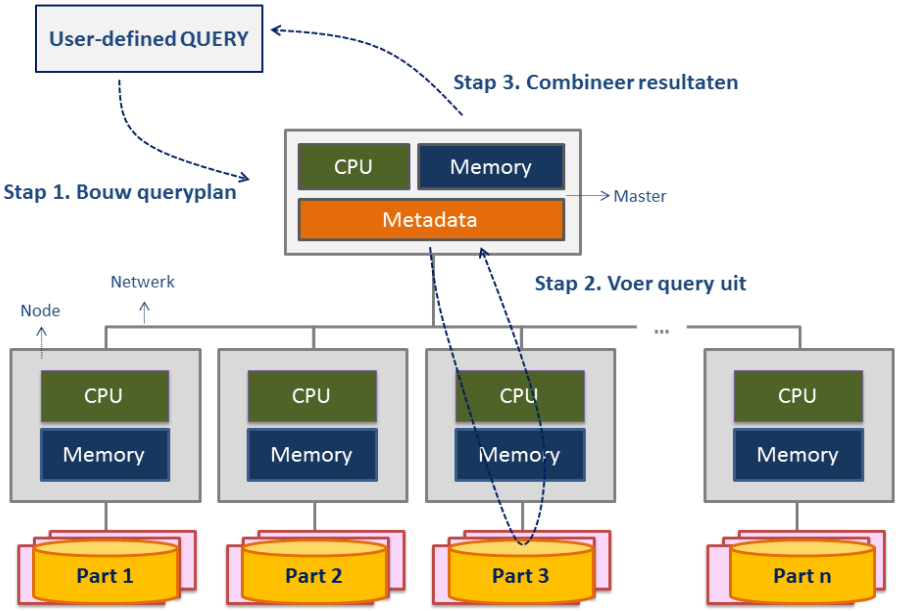
(3) analyticsprojecten worden vaak uitgevoerd in **een atypische projectstructuur met atypische projecttaken**. De typische rollen die deelnemen aan een PA-project zijn de PA-analisten, datamart-developers, DWH-teams, systeemploegen, functioneel analisten en de klanten. Er is veel communicatie nodig tussen deze teams om taken als data discovery, data preprocessing en model development uit te voeren. Dit maakt het moeilijk in te schatten hoeveel resources nodig zijn binnen een PA-project, hoe lang bepaalde taken duren, en hoeveel budget dient voorzien te worden.

3. Streamlining analytics

Ondanks de stijgende vraag naar analyticstoepassingen en -projecten blijkt het nog altijd moeilijk om met de voorgaande barrières om te gaan. In deze sectie bespreken we een aantal manieren om met deze barrières om te gaan en zo PA-projecten te stroomlijnen.

3.1. Hardware appliances voor analytics

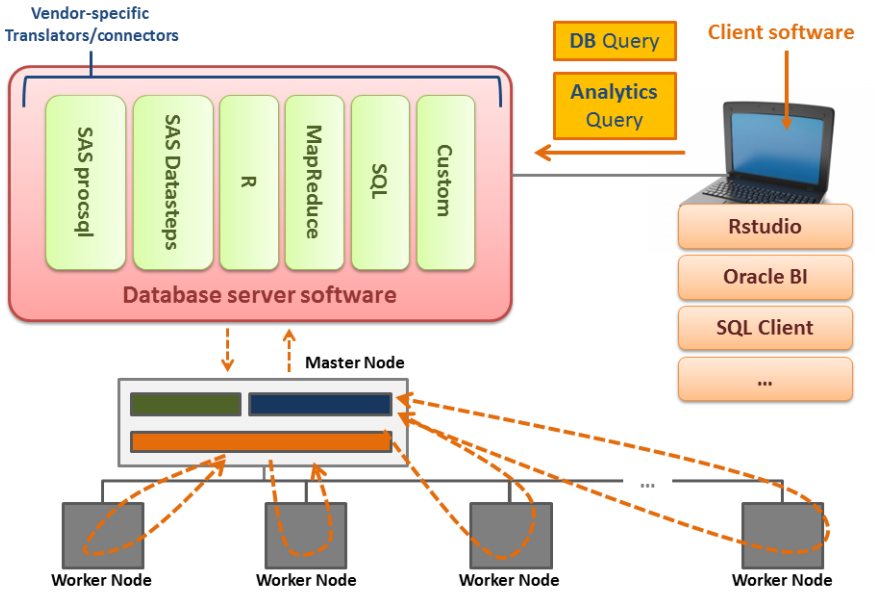
Een hardware appliance voor analytics – ook wel Data WareHouse (DWH) appliance genoemd – is een appliance die getuned is voor het snel uitvoeren van complexe query's. Bekende voorbeelden van DWH-appliances zijn EMC² Greenplum, Teradata Aster, IBM Pure Data for Analytics powered by Netezza, HP Vertica en SAP Hana.



Figuur 2: Globale architectuur van een DWH-appliance

DWH-appliances bestaan meestal uit geoptimaliseerde hardware in combinatie met een supersnelle databank. De hardware is meestal gebouwd volgens een **Massively Parallel Processing (MPP)** architectuur (zie Figuur 2). Dit houdt in dat er één *master node* is die het werk (i.e. de query) verdeelt over verscheidene *worker nodes*. Elk van deze worker nodes heeft een eigen (multi-core) processor, ram- en harddisk. De communicatie tussen de master node en worker nodes gebeurt via een supersnelle netwerkverbinding. In een MPP-architectuur worden query's uitgevoerd in drie stappen (zie ook Figuur 2):

1. Elke query wordt door de master node geanalyseerd en vervolgens wordt er een **queryplan** opgesteld. Het doel van dit queryplan is om te weten hoe de query kan verdeeld worden over de gedistribueerde architectuur.
2. A.d.h.v. het queryplan worden deeltaken door de master node doorgegeven aan de worker nodes. Dit is de daadwerkelijke **uitvoering van de query**.
3. Tot slot worden de verscheidene deelresultaten die de worker nodes opleveren **gecombineerd tot één globaal queryresultaat**.



Figuur 3: Architectuur van de DWH appliance database software

De database software die op een DWH-appliance draait is meestal opgebouwd uit een kernmodule, de eigenlijke databank, en verscheidene prioritaire connectors/translators (zie Figuur 3). Een verbindingsmodule maakt het mogelijk om de DWH-databank te verbinden met client-software-producten of om een bepaalde analyse of scriptingtaal uit te voeren in de DWH-appliance. Vaak voorkomende connectoren zijn SAS, R, MapReduce en SQL-dialecten.

Door het inzetten van DWH-appliances kan men de data supply chain gevoelig vereenvoudigen. Een appliance kan namelijk (een kopie van) de productiedatabanken laden en intern alle nodige transformaties doen. Hierbij dient men niet zelf, op voorhand, het geoptimaliseerd datamodel van het DWH te analyseren, specificeren en implementeren. Vervolgens kunnen de reporting- en analyticstoepassingen aan het andere eind van de ketting ook rechtstreeks connecteren op de appliance. Deze vereenvoudigde data supply chain heeft volgende voordelen:

- De supply chain wordt korter, wat de data sneller van de productiedatabanken naar de analyticstoepassingen brengt;
- De analyticstoepassingen kunnen gebruikmaken van de appliances. Dit geeft ook een enorme performantiewinst omdat de meeste appliances ondersteuning bieden om analyticsalgoritmes in de appliance uit te voeren i.p.v. op een aparte (tragere) server;
- Een vereenvoudigde ketting resulteert in minder (verspreide) dataquality-operaties doorheen deze ketting en biedt een ideaal aanhechtingspunt voor Master Data Management en Data Virtualisatie. Het is dus nog steeds belangrijk om ook in de vereenvoudigde ketting elke stap goed te documenteren.

3.2. Data quality: issues en tools

Gebrekkige datakwaliteit is een vaak voorkomend probleem in PA-projecten. Zo wordt analytics immers vaak toegepast om interessante groepen (risico's, nieuwe opportuniteiten, …) te identificeren, waarop applicatie en databank initieel niet gericht waren, en waar de datakwaliteit typisch slechter is. Datakwaliteit wordt soms ook misbruikt om aan controles te ontsnappen in de context van fraudebestrijding. Een voorbeeld hiervan is het keer op keer lichtjes afwijkend schrijven van een naam of adres zodat de afzonderlijke instanties van een onderliggende entiteit niet aan elkaar gelinkt kunnen worden.

Door het inzetten van dataquality-tools binnen analyticsprojecten kunnen we beter inspelen op deze problemen. Twee typische taken waarvoor we deze tools inzetten, zijn de volgende:

1) **Data profiling**, een formele en exhaustieve audit van de werkelijke inhoud van databanken, waarbij tools de aanwezige datakwaliteits-issues oplijsten en toegankelijk maken (voorkomende waarden, standaarden, vormpatronen, afwijkingen ervan, … t.e.m. businessrules);

2) **Fuzzy matching**, waarbij entiteiten uit verscheidene bronnen met elkaar kunnen gelinkt worden ondanks het ontbreken van een sleutel en ondanks de aanwezigheid van datakwaliteitsproblemen.

In een ideaal scenario (d.i. een apart datakwaliteitsproject) is het best practice om, eens de datakwaliteitsproblemen in kaart zijn gebracht en vervolgens oplossingen zijn voorgesteld (met behulp van de tools, in overleg met de business), deze problemen aan de bron aan te pakken, zodat ze niet meer of veel minder voorkomen. Zo kan elke afnemer, ook analyticsprojecten, daarvan de vruchten plukken. In het kader van een analyticsproject zelf echter, kunnen we hier niet van uitgaan en **kunnen we niet op die ideale situatie wachten**. We zullen er dus mee moeten omgaan en de hulp van zulke tools kunnen een enorme tijdswinst betekenen.

Network analytics is een specifieke vorm van predictive analytics waar men verscheidene entiteiten aan elkaar gaat linken tot een netwerk. Op basis van de eigenschappen van dit netwerk kan men dan zaken afleiden als 'welke entiteiten zijn het meest met elkaar verbonden?', 'welke entiteit wordt het meest gekruist in het netwerk?', etc. In combinatie met predictive analytics laat dit ook toe om nog complexere vragen te beantwoorden i.v.m. risicogroepen of interessegroepen. **Men kan erg creatief zijn** in de definitie van een link en in de combinatie van types links.

In de praktijk komt het vaak voor dat men entiteiten wenst te linken overheen verschillende databanken, soms met inbegrip van **externe databanken** (waarvan de kwaliteit niet bekend is en waarin andere standaarden gevolgd worden) en databanken waarin de **sleutel niet perfect betrouwbaar** is (zoals bv. in databanken waarin veel fuzzy

dubbels voorkomen). Om in zulke omstandigheden een zo volledig mogelijk netwerk op te stellen, is fuzzy matching uitermate geschikt gezien dit kan omgaan met uiteenlopende standaarden, kleine schrijffouten, verschillend gestructureerde adresvormen, etc.

3.3. Methodologie voor analyticsprojecten

De belangrijkste methodologieën die in de literatuur en in de praktijk bestaan voor analytics zijn sterk gelijkaardig. Het zijn iteratieve methodologieën, waarin feedback tussen de verschillende fasen, businesskennis en overleg met de business, centraal staan. Alle varianten kennen in een of andere vorm de volgende opbouw (hier gebruiken we de termen van CRISP-DM[1]): Business understanding, Data understanding, Data preparation, Modeling, Evaluation, Deployment.

Hoewel er een risico bestaat om te blijven cirkelen in de fasen Data understanding en Data preparation, kan de **iteratieve aard** aangegrepen worden om een ***Agile*-**projectoplevering te organiseren. Zo kan er bijvoorbeeld afgesproken worden om een milestone (per model, per interessegroep) in te richten na bv. drie maanden: op dat moment moet minstens een eerste modelresultaat ter validatie voorgelegd kunnen worden. Op die manier kunnen in iteraties telkens nauwkeurigere predictieve modellen opgeleverd worden of kan gekozen worden om prioriteit te geven aan een nieuw model (ander fenomeen, ander risico).

Tot slot werden aan het einde van de infosessie nog enkele belangrijke waarschuwingen meegegeven.

Skills. De individuele skills van de predictive analyst zijn bepalend voor de resultaten. Belangijkste knelpunten, die een bepaalde ervaring en denkwijze vergen om te overwinnen, zijn:

- de keuze van de juiste modelleertechniek,
- de extractie van *event*-, *behaviour*-, en *netwerk*-variabelen uit de ruwe data.

Succes en feedback. Succes van het project is **slechts meetbaar**, traceerbaar **als** ook de feedback (de resultaten van acties ten gevolge van het toegepaste model) deftig geregistreerd en opnieuw geëxploiteerd wordt. Dit vergt in feite nieuwe businessprocessen en registratiesystemen, die vaak in projectscope vergeten worden.

"You can't model what you don't have (examples for)". De meeste methodologieën en succesverhalen gaan ervan uit dat supervised-learning-technieken toepasbaar zijn. Dit veronderstelt dat voldoende voorbeelden beschikbaar zijn van het welbepaald type probleem waarvoor men een predictief model wenst te ontwikkelen. Deze voorbeelden, en hun aantal, bepalen de kwaliteit van de trainingsfase van een eerste (predictief) model.

[1] Cross-Industry Standard Process for Data Mining; KDD, Semma (by SAS®) and CRISP-DM: a parallel overview

In de praktijk beschikt men echter maar al te vaak over **te weinig** (juist getypeerde en van juiste metadata voorziene) **voorbeelden**. Dan dient men te vertrekken van hypothesen, zullen andere technieken ingezet worden en zullen de eerste resultaten slechts dienen om echte en voldoende specifieke voorbeelden te identificeren (mits feedbackregistratie, cfr. supra). Ook dit zijn nieuwe businessprocessen waarvoor de organisatie doorgaans niet klaar is.

4. Conclusie

Deze infosessie en bijbehorende management summary geeft een overzicht van vaak voorkomende barrières die opduiken bij analyticsprojecten en reikt manieren aan om hiermee om te gaan. Concreet bespreken we hoe hardware appliances de performantie en architectuur van een organisatie kunnen stroomlijnen, hoe je best kan omgaan met datakwaliteitsproblemen binnen een analyticsproject en geven we ook een methodologisch raamwerk mee voor analyticsprojecten. Uiteraard kunnen binnen organisaties die analyticsprojecten uitvoeren ook barrières bestaan die niet besproken zijn in deze infosessie. Hierbij denken we bijvoorbeeld aan de moeilijkheden om *operational BI* of *mobile BI* in te passen in een bestaande infrastructuur.

Een van de doelen van deze infosessie is om enkele nieuwe technologieën te introduceren (zoals de besproken hardware appliances) die nu of in de (nabije) toekomst van nut kunnen zijn binnen uw analytische organisatie. Uiteraard zijn er nog andere mogelijkheden om via minder ingrijpende maatregelen analyticsprojecten te stroomlijnen. Zonder exhaustief te zijn, denken we hierbij aan drie mogelijkheden:

- Overleg met uw huidige analytics-softwareleverancier of hij geen technieken aanbiedt die voor een beperkte licentiekost en migratie-effort reeds een mooie performantiewinst kunnen opleveren;
- Overweeg de upgrade van de bestaande analytics-server(s), misschien kan men voor een beperkte extra investering reeds een heel stuk performantiewinst boeken;
- Het upgraden van één softwarecomponent (bv. een nieuw type databank) in de data supply chain kan ook al een mooie snelheidswinst opleveren.

De sectie Onderzoek van Smals brengt met regelmaat verschillende publicaties uit over een hele waaier aan topics in de huidige IT-markt. U kan deze publicaties opvragen via het extranet http://documentatie.smals.be

Of u kan rechtstreeks contact opnemen met het secretariaat van de afdeling "Klanten & Diensten", op het nummer 02 787 58 88.