Federale Overheidsdienst FINANCIEN

Datamining

Summary

In de 2de helft van 2010 werd, in co-sourcing met SAS Institute, de installatie van een nieuwe SAS-serverinfrastructuur, een intensieve opleiding en de conversie van de bestaande dataminingflows van SPSS naar SAS afgerond. Vervolgens werd intern een 2-wekelijkse meeting behouden met minstens 1 dataminer per betrokken dienst. Dit transversaal forum heeft als doel: enerzijds het oplossen van resterende opstartproblemen, anderzijds het uitwerken van **nieuwe werkmethodes** en het afstemmen van de dataminingactiviteiten op de datawarehouseprojecten. De afspraken in dit document beschreven zijn **unaniem** goedgekeurd en werden regelmatig afgetoetst met andere betrokken projecten. Hierdoor ontstaat een duidelijkere aflijning tussen business, ICT en de (toekomstige) dienst encodage/decodage.

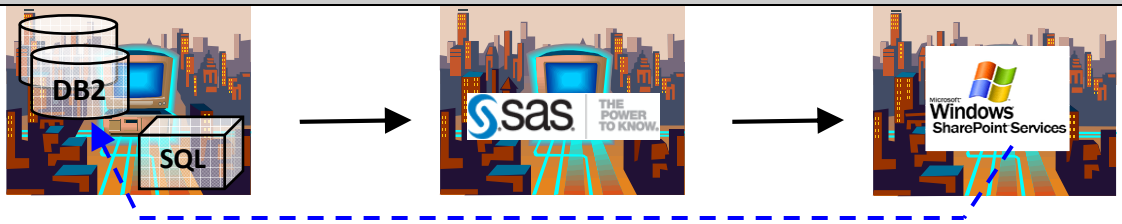
1. **Naamconventies**: de naamvereisten voor tabellen, kolommen,… sluiten nauw aan bij de standaarden van ICT, Supdev. Zij zullen toegepast worden voor alle nieuwe dataminingflows, én, tegen 30/06/2011, voor de kerntabellen van alle bestaande flows. Als dusdanig ontstaat een end-to-end consistentie: de naamgevingen zullen op termijn (quasi) identiek zijn tussen operationele systemen, datawarehouse en datamining.
2. **Documentatie**: een eenvoudig, pragmatisch standaardmodel qua documentatie van datamining-flows is goedgekeurd, teneinde 'business as usual' te kunnen onderhouden bij afwezigheid van een dataminer. Iedereen engageert zich een eerste versie van deze documentatie centraal en beveiligd beschikbaar te stellen tegen 30/06/2011 en ze minimaal halfjaarlijks bij te werken.
3. De Sharepoint-omgeving wordt als enige **outputomgeving** geponeerd voor datamining-resultaten. Het is een FodFin standaard als 'collaboration tool': gebruiksvriendelijk, garantie van backups, beveiligd, beschikbaar over het hele intranet en met tal van standaardfunctionaliteiten voor distributie van dataminingresultaten (input voor STIRON3, gerichte mailing, workflows,….). De opslag van dataminingscores vanuit Sharepoint naar het datawarehouse is ook mogelijk.
4. De strategie van de FOD Financiën voorziet een geleidelijke overgang van verouderde inputstromen naar het datawarehouse als **inputomgeving** voor de dataminers. Op termijn zullen daardoor enkel DB2 en, in mindere mate, SQL-server tabellen de standaardinput vormen. ICT, DCC draagt de verantwoordelijkheid om ze in SAS te declareren, zodanig dat ze direct toegankelijk zijn in de dataminingomgeving met de gepaste beveiligingen: toegang tot een gehele tabel of enkel toegang tot de niet-identificeerbare gegevens (vb. beveiliging van rijksregisternummer, de dataminer krijgt enkel een alternatieve, technische sleutel). Hiermee hopen we maximaal te kunnen inspelen op de privacybehoeften – nieuw wettelijk kader – en anticiperen we op een nieuwe workflow met de diensten 'encodage-decodage' en BEO.
5. Backups van de SAS-omgeving hebben een specifiek karakter. Daarom zijn de nodige richtlijnen gecommuniceerd teneinde een '**disaster recovery**' maximaal te garanderen. Anderzijds zijn de uurvensters vastgelegd tijdens dewelke de SAS-omgeving tijdelijk onbeschikbaar wordt gesteld. De maximale beschikbaarheid bedraagt op weekbasis 99,1%, tijdens de werkuren 100%. ICT, DCC voorziet de nodige **ondersteuning** voor de dataminingomgeving: declaratie en beveiliging van gegevens, uitrol van SAS client tools, interne support conform de standaard HP Open View,…

Een schematisch overzicht is terug te vinden op de volgende bladzijde.

Next steps:

1. Verdere afstemming van het datawarehouse en datamining, qua databronnen en beveiligingen.
2. Opmaak RACI-matrix (verfijning rollen van ICT, business en dienst encodage/decodage)
3. Logging en rapportering van de dataminingomgeving, in co-sourcing door het lichten van de extensie qua bijstand van het bestaand contract met SAS Institute.

Input	Datamining	Output



Geleidelijk: datawarehouse (2 releases)	SAS op server	MS Sharepoint op server
Geavanceerde beveiliging (LBAC)	Geavanceerde SAS beveiliging	Geavanceerde MS-beveiliging
Backups	Backups	Backups
Geïntegreerde metadata	Geïntegreerde metadata	Optioneel: metadata
Gedocumenteerd	Gedocumenteerd	Gedocumenteerd
Naamconventies	Naamconventies	Optioneel: naamconventies

1. Binnen het datawarehouse: **codering** van identificatiegegevens o.b.v. instructies encodage/decodage
2. Toegangsbeveiliging op kolombasis DB2 (LBAC) en in SAS op tabel- of projectbasis (Management Console)
3. **Decodering** van output (scorings) in Sharepoint → input voor mailing naar controleurs, STIRON3,…
4. Mogelijkheid tot opslag van output in het datawarehouse (historiek van scorings in DB2 !!)

Relation with datawarehouse

De strategie van de FOD Financiën voorziet een geleidelijke overgang van verouderde inputstromen naar het datawarehouse als inputomgeving voor de dataminers. Op termijn zullen DB2 en SQL-server tabellen de standaard vormen. DCC draagt de verantwoordelijkheid om ze in SAS Management Console te declareren, zodanig dat ze direct toegankelijk zijn in de SAS-omgeving met de gepaste beveiligingen.

Binnen het datawarehouse worden verschillende zones onderscheiden:

- ODS (Operational Data Store): een copie van de operationele gegevens, deze zijn in principe niet toegankelijk voor de dataminers.
- Staging files: intermediaire gegevensbestanden, combinaties, opschoning, signaletiek-gegevens,… Behoudens uitzonderingen (voor bv. evolutieve modellen die nood hebben aan statische technische sleutels) zullen dataminers niet connecteren op deze gegevens.
- Datamarts: opgeschoonde gegevens, meestal combinaties van staging files, ODS,…, meestal in een niet-genormaliseerde vorm. Deze zijn opgezet vanuit gebruikersstandpunt om maximaal de informatiebehoeften in te vullen. Daarnaast zouden ze ook aan de inputbehoeften van de dataminers moeten voldoen. Het combineren van bestanden op basis van adressen e.d. dient hierin reeds voorzien te zijn. Het is dus van alle belang dat deze datamarts ook bijkomende informatie bevat specifiek voor de dataminers. M.a.w. soms zullen er extra records of kolommen dienen meegeleverd te worden bovenop degene die dienstig zijn voor de datawarehouse-consumers.

Binnen het datawarehouse:

- Het al dan niet weerhouden van vervuilde business gegevens… We mogen aannemen dat, met de komst van SITRAN en de blijvende vernieuwingsprojecten van de FOD Financiën, vervuilde data eerder de uitzondering dan de regel wordt. Men oordeelt dat voor sommige bronnen het wenselijk is dat ongeldige business sleutels en niet-identificeerbare feiten worden weerhouden, voor andere bronnen is dit onwenselijk. Tijdens de analyse dienen businessverantwoordelijken geval per geval aan te duiden wat wél en wat niet dient weerhouden te worden binnen het kader van risico-analyse. Opmerking: “niet-identificeerbare feiten” betreft informatie die herhaaldelijk niet kon in verband gebracht worden met identificeerbare subjecten (bv. BTW-nummer niet teruggevonden).
- Er wordt gebruik gemaakt van technische sleutels (willekeurige nummers) met een identificatietabel ipv business sleutels (bv. KBO-nummers). Deze technische sleutels zijn eerder van het “statische” type. Dit wil zeggen dat ze a priori niet wijzigen in de tijd. De identificatietabellen (‘mini-signaletiek’) bevatten de business sleutels die geassocieerd worden met technische sleutels.
- Er is geen ‘security by design’ voorzien, men beperkt zich tot systeembeveiliging. Dit verhoogt wel de noodzaak aan monitoring, logging en auditing van de datawarehouse- en dataminingomgevingen.
- Specifieke privacy-behoeften zullen **ad hoc** op de datamarts (of views) worden toegepast (LBAC), waardoor bepaalde kolommen onzichtbaar worden voor de dataminers. Er kunnen encryptietechnieken toegepast worden opdat subjecten – personen, goederen,… – anoniem geïdentificeerd kunnen worden. De beschikbare sleutels voor de dataminers krijgen hier dan een dynamisch karakter.

Hiermee hopen we maximaal te kunnen inspelen op de privacybehoeften die zullen opgelegd worden in de toekomst (nieuw wettelijk kader).

Het is geenszins de bedoeling dat het datawarehouse de ‘kant en klare’ TAB (Table Analytique de Base) aanlevert. Wel zou de TAB-versie met enkele relatief eenvoudige flows binnen SAS Entreprise Guide moeten kunnen gegenereerd worden op basis van een finaal produkt van het datawarehouse, een datamart of view. Algemene vuistregel: vrij statische business regels qua gegevenstransformaties worden door het datawarehouse verwerkt, alsook complexe opschonings- en matching-problematieken. Dynamische transformaties (sterk veranderlijk in de tijd) alsook specifieke dataminingtransformaties (bv. een numerisch veld categoriseren) worden in SAS Eguide door de ‘data preparator’ uitgevoerd.