Federale Overheidsdienst FINANCIEN

Datamining

Résumé

Au deuxième semestre 2010, en co-sourcing avec SAS Institute, l'installation d'une nouvelle infrastructure de serveurs a été achevée, de même qu'une formation intensive et la conversion des flux de datamining existants de SPSS à SAS. Depuis, au niveau interne, une réunion a lieu toutes les deux semaines, avec au moins 1 dataminer par service concerné. Le but de ce forum transversal est le suivant : d'abord résoudre les problèmes de démarrage qui subsistent, ensuite élaborer de **nouvelles méthodes de travail** et adapter les opérations de datamining aux projets de datawarehouse. Les modalités convenues dont il est question dans ce document ont été adoptées à l'**unanimité**, et régulièrement confrontées aux autres projets concernés. Nous pouvons ainsi parler d'un alignement plus clair entre business, ICT et le (futur) service d'encodage/décodage.

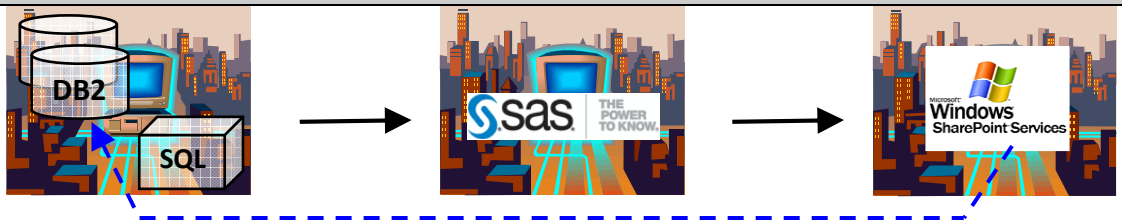
1. **Conventions de dénomination** : les règles de dénomination des tables, colonnes... correspondent étroitement aux normes ICT et Subdev. Elles seront appliquées à tous les nouveaux flux de datamining, et – pour le 30/06/2011 – aux tables essentielles de tous les flux existants. Cela donnera naissance à une cohérence générale : à terme, les dénominations seront (quasi) identiques dans les systèmes opérationnels, le datawarehouse et le datamining.
2. **Documentation** : un modèle standard simple et pragmatique de documentation des flux de datamining a été approuvé, afin de permettre la poursuite de l'activité normale en l'absence d'un dataminer. Chacun s'engage à produire une première version de la documentation, à la mettre à disposition en un point central sécurisé pour le 30/06/2011, et à l'actualiser au moins deux fois par an.
3. L'environnement SharePoint est le seul **environnement d'output** admis pour les résultats du datamining. Il s'agit d'un outil de collaboration standard du SPF Fin : convivialité, sauvegardes garanties, sécurité, disponibilité dans tout l'intranet, et nombreuses fonctions standard de distribution des résultats du datamining (input dans STIRON 3, mailing ciblé, workflows…). Le stockage des scorings du datamining dans le datawarehouse est également possible à partir de SharePoint.
4. La stratégie du SPF Finances prévoit une transition graduelle des anciens flux d'input vers le datawarehouse en tant qu'**environnement d'input** pour les dataminers. À terme, l'input standard se composera exclusivement de tables DB2, et dans une moindre mesure SQL Server. Il incombe à ICT DCC de les déclarer dans SAS, afin qu'elles soient directement accessibles dans l'environnement de datamining avec les sécurités adéquates : accès à une table complète ou seulement accès aux données non identifiables (p.ex. sécurisation du numéro de registre national, le dataminer ne reçoit qu'une clé technique alternative). De cette façon, nous espérons répondre le mieux possible aux impératifs de confidentialité (nouveau cadre légal) et nous anticipons un nouveau workflow avec les services 'encodage-décodage' et BEO.
5. Les sauvegardes de l'environnement SAS présentent un caractère spécifique. Pour cette raison, les directives ont été communiquées afin de garantir au maximum la possibilité de **disaster recovery**. Parallèlement, on a défini des créneaux horaires durant lesquels l'environnement SAS sera temporairement indisponible. La disponibilité maximale sera de 99,1% sur base hebdomadaire, dont 100% pendant les heures de travail. ICT DCC prévoit le **support** nécessaire pour l'environnement de datamining : déclaration et sécurisation des données, déploiement des outils client SAS, support interne conforme au standard HP OpenView…

On trouvera le schéma à la page suivante.

Ensuite :

1. Poursuite de l’adaptation mutuelle du datawarehouse et du datamining en termes de sources de données et de sécurisation.
2. Création d'une matrice RACI (affinement des rôles ICT, business et service encodage/décodage).
3. Journalisation et rapportage de l'environnement datamining en co-sourcing via l’exercice de l’option d’extension (assistance) du contrat actuel avec SAS Institute.

Input	Datamining	Output



Progressivement: datawarehouse
(2 releases)

- Protection avancée (LBAC)
- Sauvegardes
- Métadonnées intégrées
- Documentation
- Conventions de dénomination

SAS sur serveur

- Protection SAS avancées
- Sauvegardes
- Métadonnées intégrées
- Documentation
- Conventions de dénomination

MS SharePoint sur serveur

- Protection MS avancée
- Sauvegardes
- Option : métadonnées
- Documentation
- Option : conventions de dénomination

1. Dans le datawarehouse : **codage** des données d'identification suivant les instructions d'encodage/ décodage
2. Protection de l'accès suivant colonne dans DB2 (LBAC), et dans SAS suivant table ou projet (Management Console)
3. **Décodage** de l'output (scorings) dans SharePoint → input pour mailing aux contrôleurs, STIRON3…
4. Possibilité de stocker l'output dans le datawarehouse (historique des scorings dans DB2 !!)

Relations avec datawarehouse

La stratégie du SPF Finances prévoit une transition graduelle des anciens flux d'input vers le datawarehouse en tant qu'environnement d'input pour les dataminers. À terme, les tables DB2 et SQL Service constitueront la norme. Il incombe à DCC de les déclarer dans SAS Management Console, afin qu'elles soient directement accessibles dans l'environnement SAS avec les sécurités adéquates.

Dans le datawarehouse, on distingue les zones suivantes :

- ODS (Operational Data Store) : copie de données opérationnelles, qui, en principe, ne sont pas accessibles aux dataminers.
- Staging files : fichiers intermédiaires, combinaisons, nettoyage, données signalétiques… Sauf exceptions (p.ex. modèles évolutifs nécessitant des clés techniques statiques), les dataminers ne pourront se connecter à ces données.
- Datamarts : données nettoyées, en général des combinaisons de staging files, ODS…, le plus souvent sous forme non normalisée. Les datamarts sont configurés du point de vue de l'utilisateur afin de répondre le mieux possible aux besoins d'information. Simultanément, ils devront aussi répondre aux besoins d'input des dataminers. La combinaison des fichiers suivant les adresses, etc., doit déjà être prévue ici. Il est donc très important que les datamarts contiennent aussi des informations complémentaires spécialement destinées aux dataminers. Autrement dit, ils comporteront des enregistrements ou colonnes en plus de ce qui est déjà prévu pour les consommateurs du datawarehouse.

À l'intérieur du datawarehouse :

- Conserver ou non les données business polluées ? Nous pouvons supposer qu'avec l'arrivée de SITRAN et des autres projets de rénovation du SPF Finances, les données polluées seront l'exception plutôt que la règle. On juge que pour certaines sources, il est souhaitable de retenir les clés business invalides et les faits non identifiables, et qu'il en va autrement pour d'autres sources. Durant l'analyse, les responsables business doivent indiquer cas par cas ce qu'il convient de retenir ou non dans le cadre de l'analyse des risques. Remarque : Les 'faits non identifiables' concernent des informations qui, malgré plusieurs tentatives, n'ont pu être mises en relation avec des sujets identifiables (p.ex. numéro de TVA non trouvé).
- Il est fait usage de clés techniques (numéros aléatoires) avec une table d'identification au lieu des clés business (p.ex. numéro BCE). Ces clés techniques sont plutôt du type 'statique'. Cela signifie qu'elles ne changent pas dans le temps, du moins a priori. Les tables d'identification ('mini-signalétique') contiennent les clés business attachées aux clés techniques.
- Il n'existe pas de 'security by design' : on se limite à la sécurisation du système. Cela renforce la nécessité du monitoring, de la journalisation et de l'audit des environnements datawarehouse et datamining.
- Les besoins spécifiques en termes de confidentialité seront mis en œuvre **ad hoc** dans les datamarts (ou views) (LBAC). Certaines colonnes deviendront invisibles aux dataminers. On pourra aussi appliquer des techniques de cryptage pour pouvoir identifier les sujets (personnes, marchandises…) de façon anonyme. Les clés disponibles aux dataminers revêtiront ainsi un caractère dynamique.

Nous espérons pouvoir répondre par là aux exigences de confidentialité qui seront imposées à l'avenir (nouveau cadre légal).

Le datawarehouse n'est aucunement chargé de générer la TAB (Table Analytique de Base) 'prête à l'emploi'. Néanmoins, la version TAB devrait pouvoir être générée avec quelques flux relativement simples dans SAS Entreprise Guide, à partir d'un produit final du datawarehouse, d'un datamart ou d'une view. Principe général : les règles business relativement statiques de transformation des

données sont traitées par le datawarehouse, de même que les problématiques complexes de nettoyage et de matching. Quant aux transformations dynamiques (qui changent fortement dans le temps) et à des transformations spécifiques du datamining (p.ex. catégoriser un champ numérique), elles sont exécutées dans SAS Eguide par le 'data preparator'.