DATA WAREHOUSE SEFA – PROPOSTA DE CRIAÇÃO DE SANDBOX ANALÍTICA EXTERNA DA SEFA

I – INTRODUÇÃO

O presente documento tem por objetivo expor a proposta de nova abordagem no processo de implementação do *data warehouse* SEFA, implementação essa que está em andamento desde março de 2012. A proposta busca solucionar a necessidade imediata de acesso eficiente a dados corporativos sem no entanto prejudicar o esforço de construção de um ambiente analitico de qualidade na Secretaria da Fazenda, especificamente para bases de dados de documentos fiscais.

II – A QUESTÃO RAPIDEZ *VIS-A-VIS* GOVERNANÇA DE DADOS

O processo de implementação do *data warehouse* SEFA busca seguir a premissa de se utilizar as melhores práticas em construção de soluções analíticas. Não é ao acaso que o processo licitatório de aquisição da infraestrutura de DW contemplou também a contratação no mercado de serviços de implementação de *data warehouse*.

No entanto, a implementação de bases de dados em DW é um processo, via de regra, de médio/longo prazo, em especial quando as bases são de grande porte e alta complexidade, o que é o caso de algumas bases de dados da SEFA como a NF-e, EFD, Sintegra e Convênio 115. Isso porque existe toda uma sequencia de atividades necessárias para preparar uma base de dados em um modelo DW, a partir de métodos e práticas específicas, a fim de garantir qualidade na implementação e proporcionar boa governanaça dos dados analíticos.

Ao mesmo tempo, as áreas de negócio da organização (em especial a área de fiscalização) possuem uma significativa demanda reprimida de acesso amplo e eficiente a dados corporativos, a fim de poder melhor cumprir suas obrigações regimentais. Surge então um dilema: como disponibilizar a curto prazo dados corporativos complexos de forma ampla e eficiente e ao mesmo tempo implementar as melhores práticas de construção de ambientes analíticos?

III – SANDBOX ANALÍTICA: A SOLUÇÃO DO DILEMA

A experiência recente de implementação do *data warehouse* SEFA nos trouxe, dentre várias outras questões, um dilema entre a necessidade ***urgente*** de disponibilizar dados corporativos à área de fiscalização e a necessidade ***fundamental*** de desenvolver soluções analíticas corporativas sob as melhores práticas de implementação de *data warehouse*.

Um dilema similar levou a Teradata, empresa fabricante do *appliance* DW da SEFA, a criar o conceito de "**sandbox**", uma área dedicada no *appliance* de banco de dados que não está sujeita ao rigor metodológico de implementações analíticas[1].

A criação de uma *sandbox* deve ser fundamentada em necessidades de negócio tais como testes e exploração de novas abordagens de negócio, alocação temporária de dados aguardando implementaçao de uma aplicação, solicitações de dados especializados de curto prazo ou para execução de processo de análise exploratória de dados[2]. No caso específico da SEFA, a necessidade premente é a de disponibilizar de forma imediata os dados de documentos fiscais a alguns usuários avançados da área de fiscalização para permitir análises massivas com o fim de subsidiar projetos de fiscalização e auditorias.

Uma *sandbox* significa disponibilizar uma ou mais bases de dados no *appliance* DW, beneficiando-se do poder de armazenamento e processamento do equipamento porém sem preparar essas bases em um modelo de negócio (modelo analítico de DW).

A grosso modo, seria apenas "copiar" as bases de dados para dentro do *appliance* Teradata para rapidamente disponibilizar o uso a um grupo restrito de auditores fiscais. Isso difere do processo de criação de um *data warehouse*, pois este implica em todo um rito metodológico para transformar os dados em um modelo de negócio integrado aos demais modelos (como cadastro, GIAS, recolhimentos, documentos fiscais, apuração, etc.).

O conceito de *sandbox* fornece aos usuários uma parte do espaço em disco do *appliance*, possibilitando carga de dados para uso intensivo e realização de análise dos dados. Aproveita o paralelismo do banco de dados, permite o uso das mesmas ferramentas e utilitários disponíveis no ambiente produtivo DW. Também auxilia a área de TI a compreender melhor a complexidade dos dados para quando estes forem definitivamente movidos para um modelo de *data warehouse*.

No entanto, é importante que a abordagem de utilizar o conceito de sandbox **NÃO** se torne um substituto para a implementação dos dados em modelo DW. Uma *sandbox* é uma boa alternativa para, em caráter temporário, permitir o uso massivo de dados por usuários avançados, no entanto deve-se ter em mente que tal solução é **temporária** e não substitui o processo de implementação das bases de dados em modelo de *data warehouse*.

A *sandbox* corporativa é portanto um ambiente de curto prazo. É preciso que os usuários que acessarão a *sandbox* tenham em mente que a boa governança dos dados corporativos da SEFA exige que os dados disponibilizados nessa área temporária sejam, o mais rápido possível, objeto de transformação e modelagem para um ambiente DW corporativo ou seja, a *sandbox* nasceria já com um horizonte de descontinuação em favor de um modelo de *data warehouse*.

1 A esse respeito, ver Armstrong, Rob. **Think again: do your data warehousing efforts help or hinder creativity?** Teradata Magazine Online. [http://apps.teradata.com/tdmo/v08n04/viewpoints/FreshPerspectives/ThinkAgain.aspx]. Acessado em 18/12/2012.

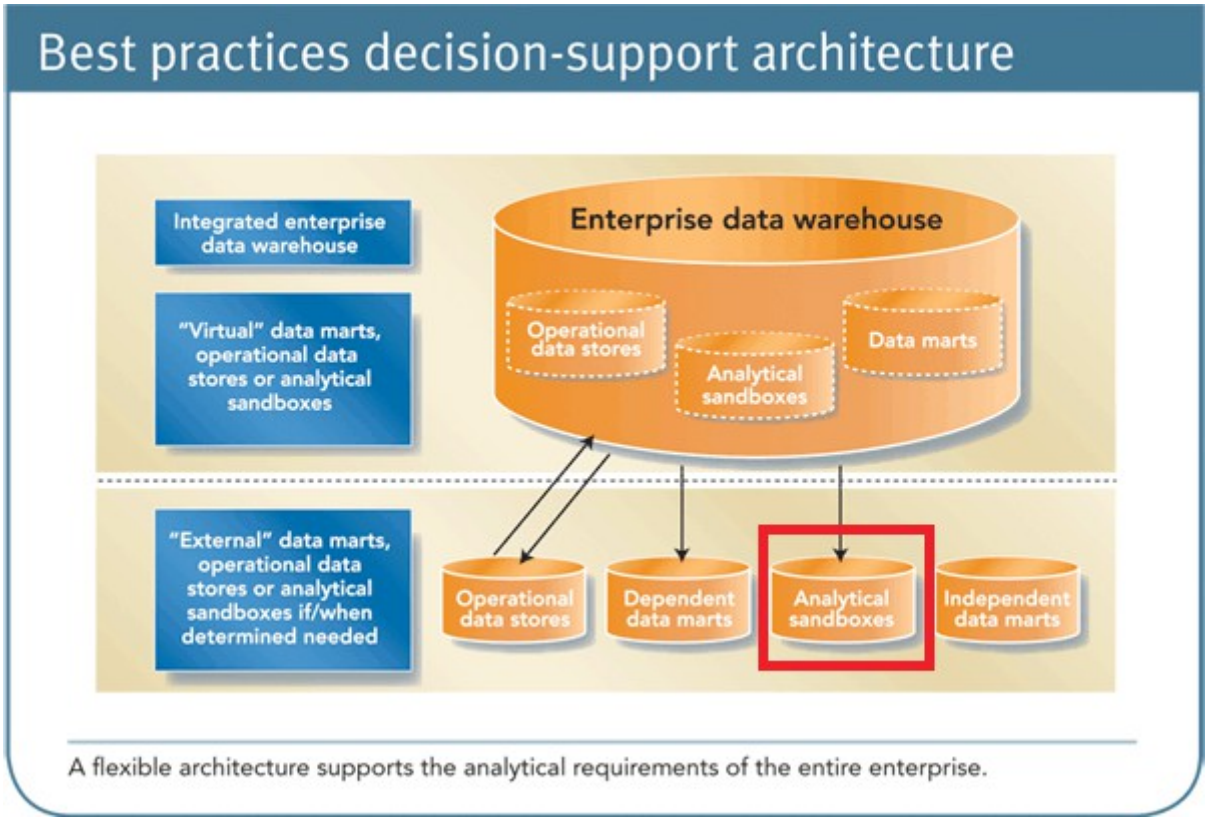
2 Smith, Debbie. **Deciphering the data mart mystery: a flexible decision-support architecture provides insight**. Teradata Magazine Online. [http://apps.teradata.com/tdmo/v08n03/viewpoints/WhyTeradata/DataMartMystery.aspx] Acessado em 18/12/2012.

IV – A PROPOSTA DE SANDBOX ANALÍTICA CORPORATIVA DA SEFA

Considerando o anteriormente exposto, a proposta seria a criação de uma área segregada do DW no appliance Teradata, de **10Tb** (Terabytes), para contemplar a carga das seguintes bases de dados:

- NF-e (tabelas de controle, emitente, destinatário/remetente, identificação, item, totais, local de entrega/retirada, cancelamento);
- EFD (tabelas C100, C170, Controle, Participante);
- CV115 (tabelas mestre, item, cadastro, controle);
- Sintegra (tabela "docresumounico").[3]

A sandbox proposta seria do tipo externo ao *data warehouse* ("***external analytical sandbox***"), conforme figura a seguir.



As bases de dados elencadas (NF-e, EFD, Sintegra e CV115) estariam disponíveis na *sandbox* analítica externa da SEFA para um grupo restritos de usuários, para subsidiar projetos de fiscalização e auditorias fiscais.

3 Deverá ser alinhado com a CRE/IGF exatamente quais são as tabelas necessárias para essa solução temporária, além de ser definido a origem da carga dos dados e temporalidade da carga.

Paralelamente ao uso dessa *sandbox*, continuam os trabalhos de implementação do *data warehouse* das áreas de assunto “Mercadorias Atacado” (NF-e, EFD) e “Comunicação/Energia” (CV115, EFD). Uma vez implementadas essas áreas de assunto no DW SEFA, a sandbox seria descontinuada[4].

Com isso é possível, ao mesmo tempo, permitir que a área de fiscalização utilize imediatamente a nova tecnologia de banco de dados Teradata adquirida pela SEFA para prospecção de dados de extrema relevância fiscal, ao mesmo tempo em que a CRE/AGTI e a CRE/IGF, em conjunto com a equipe Maxtera/Teradata e Celepar, desenvolvem o ambiente analítico de documentos fiscais, que a médio/longo prazo passará a ser a principal fonte analítica para os trabalhos de análise fiscal, como foi planejado desde o advento do projeto Phoenix.

Curitiba, 20 de dezembro de 2012

■ ■ ■

4 Está em andamento no âmbito da CRE/IGF uma discussão a respeito da pertinência ou não de manter a abordagem inicial de utilizar a base de dados do Sintegra em ambiente de *data warehouse*. Caso seja decidido por não alimentar o DW com dados do Sintegra, a *sandbox* externa analítica seria mantida, ainda que redimensionado seu tamanho, a fim de continuar abrigando a base de dados do Sintegra para suprir eventual necessidade de uso desses dados, até que se tornem efetivamente desnecessários.