

O talento do Talend

Quem procura uma solução ETL para Business Intelligence não deve deixar de conferir o Talend. Conectado a diversas fontes de dados, ele só não faz chover – ainda.

por Miguel Koren O'Brien de Lacy

Um dos componentes essenciais de uma solução para a inteligência de negócios (BI – *Business Intelligence*) é o módulo de integração de informações de diversas fontes de dados. Felizmente o mercado de Software Livre e de código aberto oferece algumas soluções bastante completas que são usadas dentro dos pacotes de BI ou em modo stand alone. As soluções de integração de dados são conhecidas como ETL (*Extract, Transform, Load*), pois a missão dessas soluções é integrar informações e prepará-las para a formação de tabelas no *data mart*. Normalmente os dados precisam ser manipulados depois de obtidos e logo inseridos nas tabelas dos bancos de dados em que reside o *data mart*. Ou seja, estamos falando em obter os dados, aplicar regras de negócios da empresa, facilitar as consultas posteriores ou validar a qualidade dos dados e entregá-los transformados ao destino final. Dentre as transformações normalmente aplicadas aos dados, podemos destacar a seleção de campos importantes para a análise, a filtragem dos registros necessários, a limpeza dos dados (“Sr.” ou “Senhor” passam a valer “S”), a geração dos campos calculados (por exemplo, “valor nf = quantidade * valor unitário”) etc. Mas essas soluções também podem ser consideradas integradores e manipuladores de dados para outras

necessidades, podendo substituir muitos sistemas de “interfaces” de dados que são desenvolvidos pelas áreas de TI de grandes empresas.

Originalmente, os sistemas corporativos foram desenvolvidos de forma isolada para cumprir alguma função específica, não sendo parte do objetivo de projeto a necessidade de compartilhar dados. Quando a quantidade desses sistemas começou a ser mais importante dentro das empresas, surgiram as primeiras necessidades de integração de dados, como a carga de pedidos aos sistemas de estoque e faturamento. Tais necessidades de integração foram implementadas diretamente por programas ou scripts específicos. Com a crescente complexidade das integrações requeridas, iniciou-se um encarecimento do desenvolvimento desses processos. Os sistemas tinham manutenção complexa e as lógicas de integração difíceis de documentar, auditar e entender. Por isso, as empresas começaram a implementar ferramentas prontas ou semi-prontas para essa necessidade. Essa foi a origem dos sistemas ETL.

Esses sistemas, porém, não possuem a popularidade que merecem em razão de certos fatores que impedem sua adoção. Entre estes fatores devemos destacar:

- ▶ Custo inicial;
- ▶ Complexidade no uso com curva de aprendizado muito empinada;

- ▶ Escalabilidade a necessidades de integrações pequenas.

A empresa Talend é fabricante do Talend Open Studio [1], apresentado como “o sistema ETL mais abrangente no mercado de software livre e código aberto”. O modelo de negócios da empresa é um dos clássicos para Software Livre. A empresa desenvolve um produto que licencia sem custo, usando licenciamento em modalidade livre, no caso a GPL [2], e oferece serviços e suporte comerciais. O software licenciado sob a GPL não possui todos os componentes que são oferecidos a clientes sob o licenciamento de subscrição do serviço, mas mesmo dessa forma é extremamente útil para as necessidades ETL de qualquer empresa. O *Talend Open Studio* foi liberado no mercado em 2005 pela empresa Talend, com bases na França, país com uma política oficial de apoio a Software Livre e Código Aberto. Analistas de TI como Forrester Research [3], IDC [4] e Bloor Research [5] posicionam o Talend como o melhor sistema para necessidades de integração de dados. O sistema BI em software livre *SpagoBI* [6][7] é um exemplo de uso do Talend como seu componente ETL padronizado.

O modelo de Software Livre é bem apropriado para esse tipo de solução, pois faz bom uso de desen-

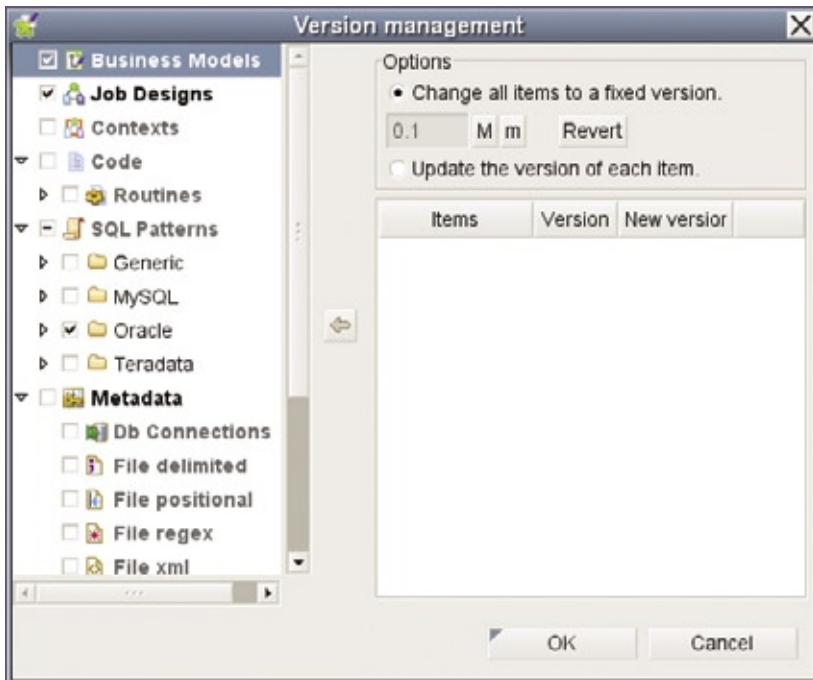


Figura 1 Gerenciamento de versões.

- ▶ Monitor de processos;
- ▶ Componentes modulares (atualmente mais de 400). Usando a API do sistema podem ser desenvolvidos componentes personalizados.

O gerenciador gráfico de processos ETL roda dentro do ambiente *Eclipse* [8][9], seguindo a filosofia de integração de ferramentas dentro desse ambiente para desenvolvimento.

Normalmente os primeiros três módulos são usados na ordem listada. O gerenciador de negócios é usado pelos usuários do sistema, enquanto o gerenciador de processos ETL e o gerenciador de metadados são usados pelos programadores. Mesmo que o repositório de projetos ETL com apoio a equipes de trabalho seja um componente oferecido comercialmente, o Talend Open Studio inclui a possibilidade de versio-

volvimento colaborativo, em que muitas pessoas contribuem com módulos para conexões a fontes de dados variadas. Mas o sucesso do Talend também mostra que o Software Livre não é apenas bem sucedido em soluções “comoditizadas”, como sistemas operacionais e utilitários de infra-estrutura. O Talend mostra que uma solução livre pode ser muito atrativa num mercado novo em formação, competindo com sistemas proprietários de grandes empresas como IBM, SAS e SAP, as quais têm preços de licenciamento de nível corporativo. Empresas como Oracle e Microsoft também competem nesse mercado, porém com uma visão mais restrita, focada nos produtos de base de cada uma: banco de dados Oracle e *SQL Server*, respectivamente.

Um forte apelo do Talend é a disponibilidade de uma ferramenta abrangente em que os próprios usuários da informação podem modelar as regras de negócios (fontes de dados, campos requeridos, regras de transformação etc.) para a integração dos dados. Hoje a área de TI de qualquer

empresa tem diariamente mais e mais dados, e os usuários têm necessidades de visualização que são mais urgentes a cada dia. Assim, é muito útil que os próprios usuários possam usar o sistema de forma não técnica para especificar as regras de integração.

Recurso e Tecnologia

O Talend é um sistema desenvolvido em Java que usa uma arquitetura modular formada por:

- ▶ Gerenciador gráfico de negócios (visão não técnica de necessidades de fluxo de dados);
- ▶ Gerenciador gráfico de processos ETL;
- ▶ Gerenciador de metadados (repositório para reutilização de objetos);
- ▶ Repositório de processos (módulo adicional da versão comercial);
- ▶ Interface web services a processos ETL;

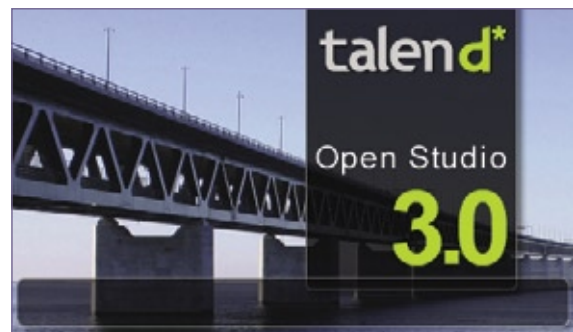


Figura 2 Tela de início do Talend.

namento do desenvolvimento dos processos. A figura 1 mostra o uso desse recurso.

O Talend é um sistema “gerador de código”, em oposição a um sistema de “caixa preta” – a qual é responsável pela execução dos processos ETL –, ou seja, o Talend não requer um servidor de execução de processos ETL. A vantagem da geração de código é que é mais simples integrar os processos ETL dentro de outros aplicativos e os modelos são de portabilidade muito mais flexível. Além de sistemas como SpagoBI, o

Talend é integrado em sistemas de empresas como Ingres, Teradata e JasperSoft. O mecanismo de geração de código permite a integração do código gerado pelo Talend dentro de outras soluções, além de ter maior portabilidade. Outra vantagem de mecanismos de geração de código é que mecanismos de integração “online”, conhecidos hoje como “integração operacional”, podem ser implementados com maior facilidade. O conceito de geração de código versus “caixa preta” de execução dos processos ETL é a maior diferença entre o Talend e muitos outros sistemas. O conceito de geração de código não foi popular no passado, possivelmente por causa do marketing das empresas comerciais destas soluções, mas as vantagens que o Talend apresenta usando este conceito são importantes.

No caso, o Talend gera código Java ou Perl e SQL para os processos ETL. Além de aplicar o conceito ETL, ele aplica também o conceito ELT (*Extract, Load, Transform*), o que significa que pode aproveitar a eficiência e performance nativas de bancos de dados SQL; ou seja, os dados são extraídos, carregados no destino e, somente depois, já dentro do banco de destino, é aplicada a lógica de transformação usando SQL e linguagem procedural de



Figura 3 Definição de repositório local inicial.

cada banco de dados. Para a funcionalidade ELT são suportados nativamente hoje os bancos de dados Oracle, MySQL e Teradata.

Operação

O sistema está disponível em versões para Linux, Unix e Windows. Para todas as plataformas é necessário ter Java e Perl no equipamento, considerando seu conceito de desenho que é de geração de código, ou seja, que os processos do Talend são executados pela máquina virtual Java (JVM) ou pelo interpretador Perl.

O download do sistema ainda é relativamente pesado (235 MB), mas com conexões de banda larga cada dia mais confiáveis e rápidas pode ser feito em poucas horas. Neste artigo exploraremos o Talend Open Studio para Linux em ambiente gráfico Gnome rodando em OpenSUSE 10.2. No caso de usá-lo em Windows 2000, existe um pré-requisito adicional que é a instalação do GDI requerido pelo Eclipse [10]. A partir do Windows XP, essa biblioteca está incluída nativamente. Vale a pena fazer o download também do pacote completo de documentação. Referente ao produto e à documentação, o Talend apresenta as vantagens de suporte por uma empresa comercial, pois a documentação é bastante completa. Outro destaque importante do Talend é o acesso a um sistema de ajuda online muito completo. Infelizmente, isso ainda não é norma na maioria dos sistemas de código aberto. O site do Talend oferece um wiki e uma área de tutoriais.

Simplesmente descompacte o pacote obtido por download e atribua propriedades de executável ao binário de início do sistema:



Figura 4 Geração de um novo projeto.

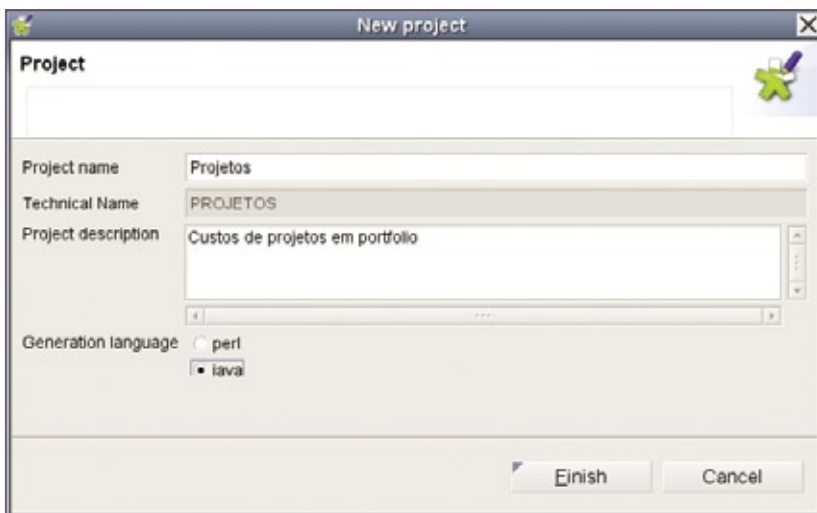


Figura 5 Detalhes de um novo projeto ETL.

```
chmod a+x TalendOpenStudio-linux-
↳gtk-x86
```

Selecione o binário de início apropriado para sua arquitetura de CPU. Para iniciar o gerenciador de processos, execute o binário:

```
./TalendOpenStudio-linux-gtk-x86 &
```

Isso exibirá a tela de abertura do sistema (figura 2).

Leia e aceite o contrato de licenciamento. Aparecerá finalmente a tela de conexão ao repositório de processos ETL (figura 3). No caso do Talend Open Studio, o repositório somente poderá ser local.

Com o repositório local definido, gere um projeto novo selecionando *Create a new local project* (figura 4).

Ao definir o título e a descrição do projeto, deve-se selecionar também a linguagem de geração do código. No exemplo para o artigo, escolhemos código em Java. A figura 5 mostra as opções de definição do projeto.

Finalmente já se pode entrar no projeto escolhido pela tela principal, bastando selecionar nela o projeto e pressionar o botão *open*. Após alguns momentos aparecerá a tela principal do sistema. Na primeira vez, será pedido ao usuário que re-

gistre o uso digitando seu email e selecionando o país. Este passo não é obrigatório, mas é recomendado para receber notificações de novas versões do sistema.

Antes de usar o Talend Open Studio, deve ser configurada a lista de elementos disponíveis para o desenho. Isso pode ser feito modificando-se as propriedades do projeto na

opção de “palette settings”. A figura 6 mostra a inclusão de conectores para MS SQL Server e PostgreSQL, que devem ser selecionados no caso de precisarmos desenhar um processo de integração de dados entre estes dois bancos de dados.

O desenho de um processo ETL requer várias etapas, incluindo desenho dos componentes de negócios, dos *jobs* de processamento, de conexões *JDBC*, acoplamento de módulos de processamento e outros elementos. A figura 7 mostra um exemplo de configuração de conexão *JDBC* ao banco Microsoft SQL Server. O driver *JDBC* utilizado vem junto com a instalação do Talend, no caso, o excelente *JTDS* [11]. A figura 8 mostra a tela do Talend Open Studio com a visão da paleta de ferramentas, componentes tratados e a edição de um elemento de negócios (*business object*).

Além das possibilidades de desenvolvimento lógico da solução dentro do ambiente gráfico, o Talend oferece

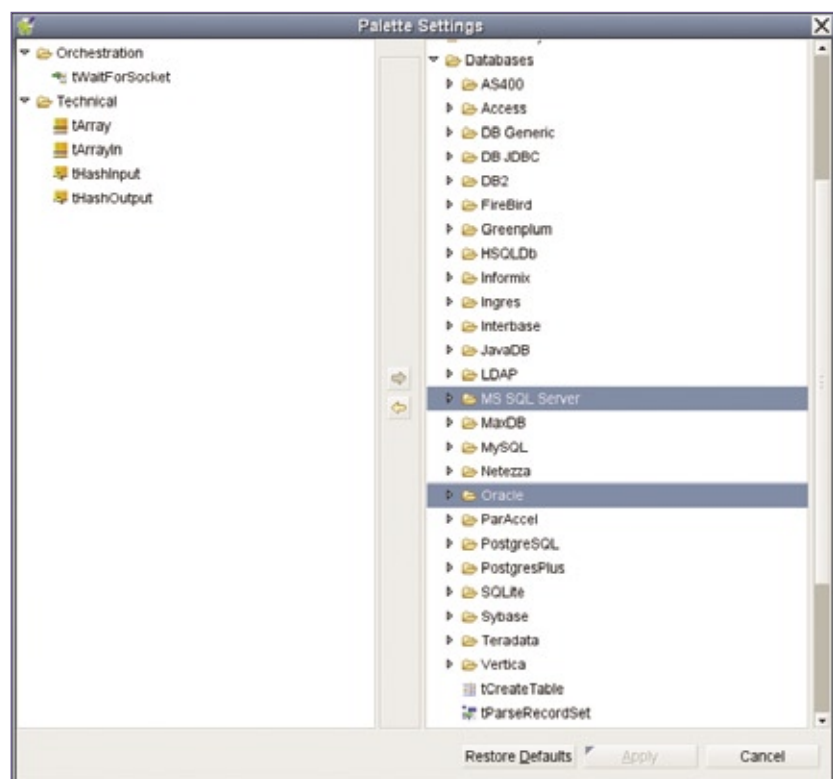


Figura 6 Formação da lista de ferramentas.

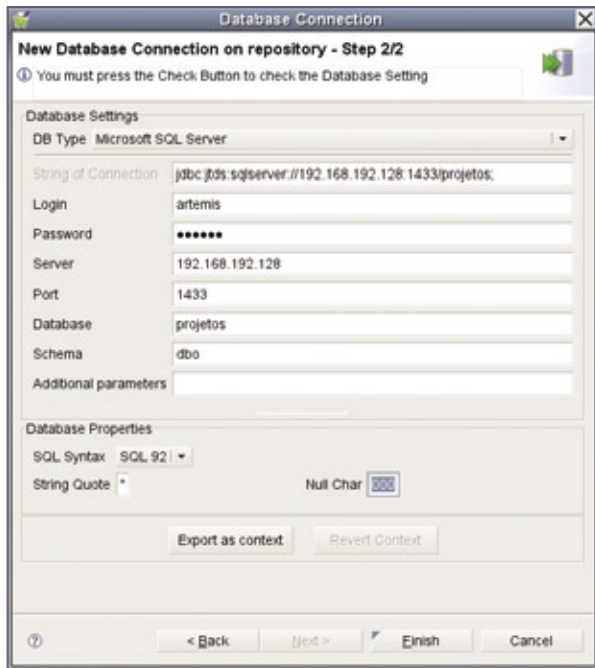


Figura 7 Definição de conexão JDBC.

ferramentas para preparar diagramas formatados para necessidades de documentação e apresentação. Por exemplo, podem ser escolhidas as fontes, as cores e os estilos de linhas para os objetos do diagrama, além de haver a possibilidade de alinhamento numa grade e a opção de layout automático dos componentes.

Entre as características que mais flexibilidade oferecem ao Talend está a biblioteca de componentes. Esses componentes estão divididos em grupos funcionais em que os de maior utilidade para uso corporativo podem ser os componentes preparados para acesso a outros sistemas, tais como *SugarCRM*, *Salesforce* e *SAP*. O Talend apresenta um destaque no mercado ao oferecer um conector ao SAP. A figura 9 mostra alguns componentes do *Forge* do sistema que podem ser baixados e instalados para aumentar a funcionalidade padrão do Talend.

Um destaque do mecanismo de componentes do Talend é a possibilidade de publicar os processos automaticamente, por exemplo, no servidor de BI SpagoBI. Outro componente de uso geral e que incrementa em

muito a complexidade de processos que podem ser definidos é o executor de scripts em *Groovy*. Além de poder desenvolver componentes específicos, a possibilidade de gerar scripts em *Groovy* para tratamento avançado dos dados permite que as transformações sigam regras muito especiais da empresa. A figura 10 mostra a definição de integração com o SpagoBI.

Vale a pena destacar algumas características conceituais do editor gráfico que são importantes para uso corporativo do Talend:

- ▶ Os desenhos são versionados com a data de cada versão;
- ▶ Cada objeto dos modelos recebe um estado que pode ser “não verificado”, “verificado” ou “validado” por default no caso

dos business objects. Entretanto, podem ser definidos outros estados usados na empresa;

- ▶ Existe uma área em cada projeto para reunir a documentação que pode conter objetivos, justificativas, exemplos de resultados desejados etc. O Talend também gera documentação técnica e de processo automaticamente;
- ▶ Separação de papéis entre usuários de negócios que descrevem suas necessidades a nível macro (definindo business objects) e desenvolvedores (definindo jobs e outros elementos);
- ▶ Notificação ou atualização automática do sistema quando disponibilizadas novas versões Talend ou dos componentes;
- ▶ Facilidade para execução de testes individuais e integrados de componentes;
- ▶ Importação e exportação de projetos completos ou elementos.

O Talend é um sistema muito completo, com o apoio de uma empresa comercial, um ecossistema de componentes muito amplo e com um ambiente de definição

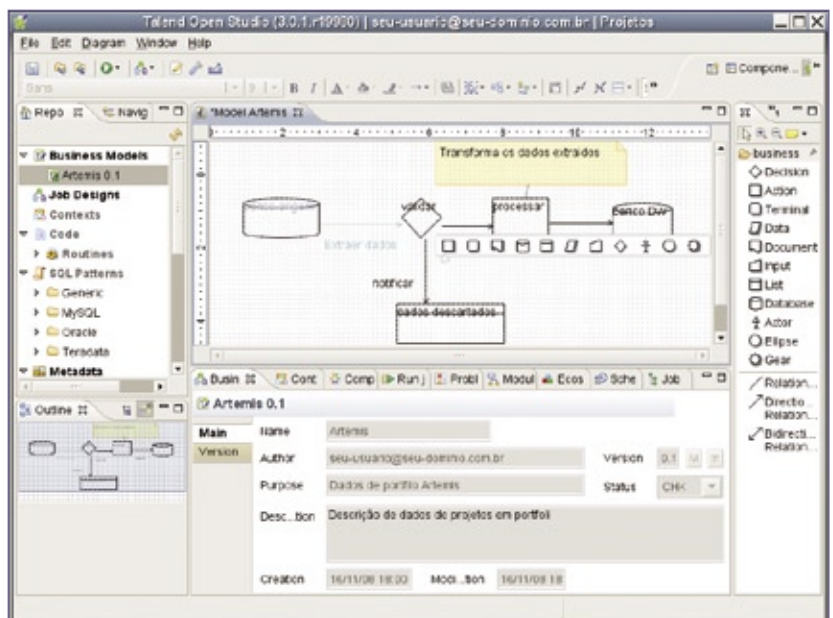


Figura 8 Tela do Talend Open Studio.

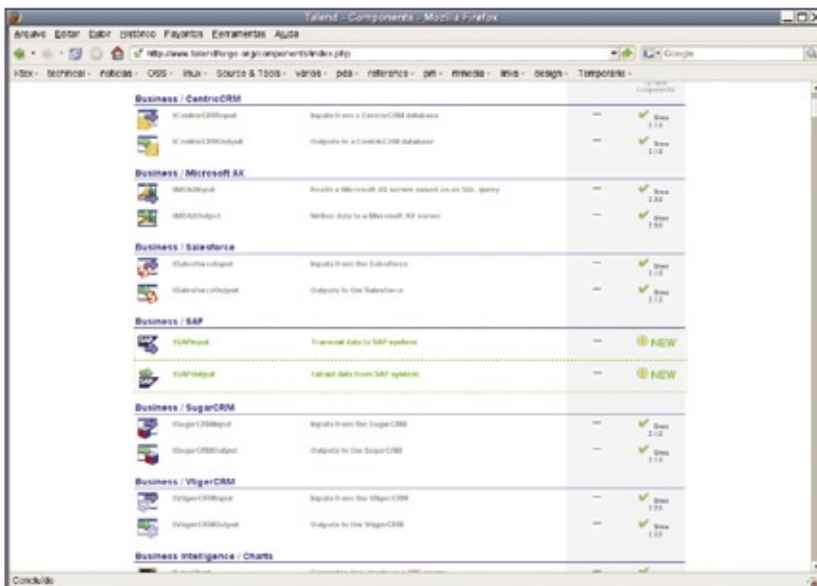


Figura 9 Forge de componentes mostrando aqueles específicos de algum sistema.

e desenvolvimento fácil de usar, além de poderoso e visualmente atraente. Assim, o projeto pode resolver a questão relatada no início do artigo sobre os motivos da falta de popularização de sistemas ETL no mercado.

Conclusão

Hoje, qualquer empresa tem necessidades de ETL e integração de dados, e em muitos casos é caro e complexo desenvolver soluções caseiras específicas. O Talend apresenta uma solução flexível com a qual estes processos podem ser

modelados e executados de forma confiável. Podem ser definidos e documentados com a participação dos usuários que são os principais interessados no resultado final. Naturalmente, esse tipo de divisão de atividades ou responsabilidades requer que todos os usuários, independentemente do papel que desempenham no sistema, sejam devidamente treinados. Em TI, é cada vez mais comum haver situações em que os sistemas são simples e poderosos para usar, mas os usuários devem ser treinados nos princípios da tecnologia apresen-

tada. Quem usa o Talend deverá estar ciente da documentação em inglês ou em francês, além do uso do sistema em inglês.

O uso do Talend dentro de uma empresa é facilitado pelas características corporativas do sistema que permitem, entre outras possibilidades, ter alta produtividade e visibilidade dos processos pelo apoio a componentes específicos, tais como conectores SAP, Salesforce ou customizados, e geração de documentação. Provavelmente será difícil encontrar um caso de uso que não possa ser atendido com comodidade pelo Talend. ■

Mais informações

- [1] Talend: <http://www.talend.com>
- [2] GPL versão 2: <http://www.gnu.org/licenses/old-licenses/gpl-2.0.html>
- [3] Forrester Research: <http://www.forrester.com>
- [4] IDC: <http://www.idc.com>
- [5] Bloor Research: <http://www.bloor-research.com>
- [6] SpagoBI: <http://spagobi.eng.it>
- [7] Miguel K.O. de Lacy, "Negócio inteligente": <http://www.lnm.com.br/article/1747>
- [8] Eclipse: <http://www.eclipse.org>
- [9] Linux Magazine 36 – Eclipse: <http://www.lnm.com.br/issue/1353>
- [10] GDI: <http://www.eclipse.org/swt/faq.php#nographicslibrary>
- [11] JTDS: <http://jtds.sourceforge.net>



Figura 10 Integração com o SpagoBI Server.