

Arquitetura de Data Warehouse e Data Marts

Aula 6: Processo ETL – Extração de Dados

Apresentação

Nesta aula serão apresentados o processo de ETL, as possíveis fontes de dados (estruturados) e (não estruturados), a periodicidade da extração dos dados, os métodos de extração, a documentação do processo de extração e a ferramenta de ETL.

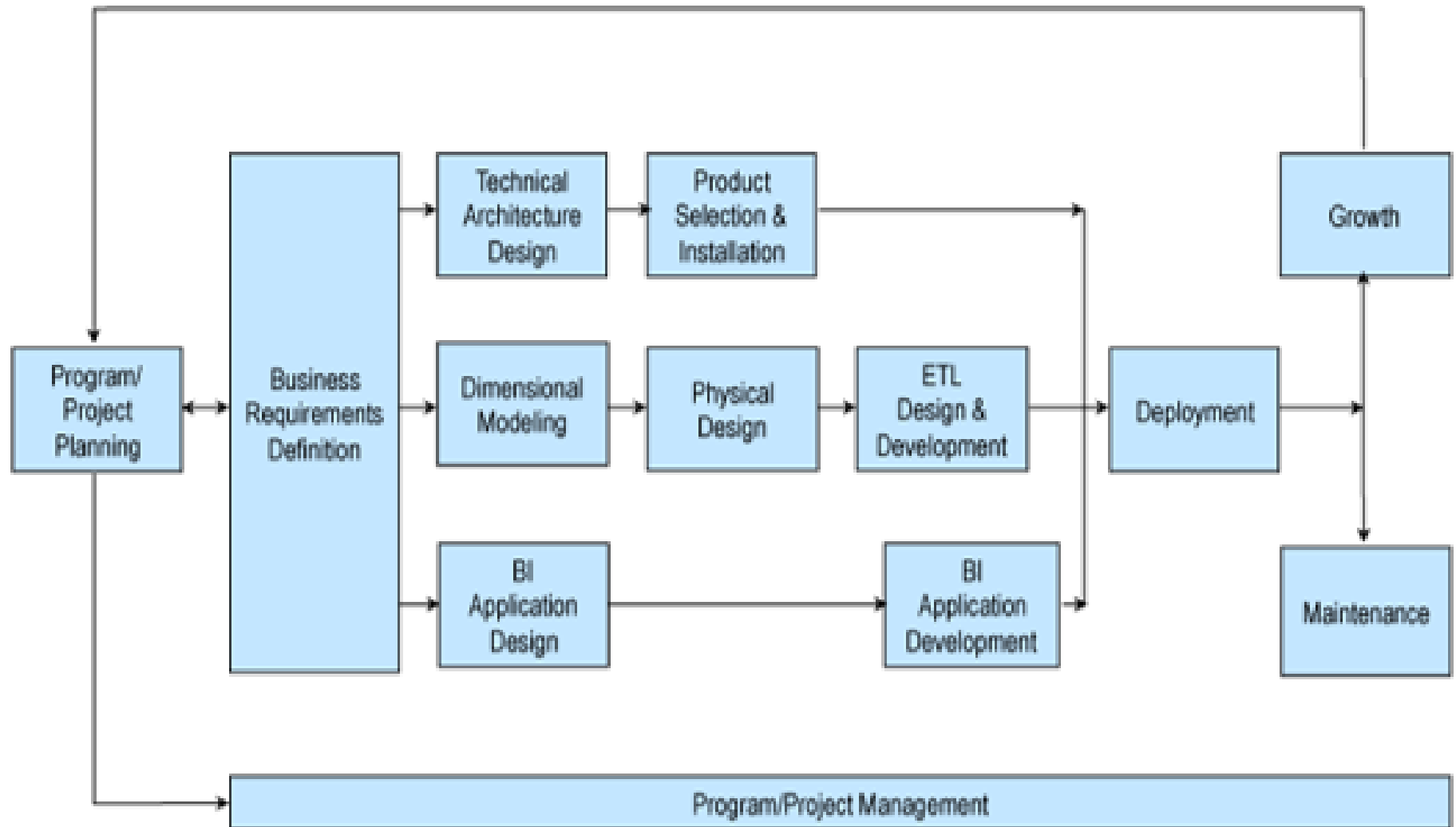
Objetivos

- Analisar o Processo ETL;
- Descrever as técnicas de extração de dados;
- Examinar as fontes de dados e ferramentas de ETL.

Processo ETL – Extração, Transformação e Carga dos dados

O Processo ETL (Extract, Transform and Load) é uma etapa importante e significativa em um projeto de Data Warehouse (DW). A construção do processo ETL segue a trilha de dados contida no ciclo de vida de projetos de DW/DM (Data Warehouse/Data Mart) logo após a definição do Projeto Físico do DW/DM, conforme ilustrado na Figura 1.

O processo ETL é responsável por extrair os dados dos sistemas origens ou arquivos onde os dados estão contidos, aplicar tratamentos necessários aos dados e carregá-los nas tabelas definitivas do DW/DM. Essa etapa é a mais onerosa para o projeto devido aos desafios encontrados na captação dos dados nos sistemas origens, na integração dos dados de variadas fontes, no ambiente de tráfego, nas equipes envolvidas, entre outros (KIMBALL, 2013) estima que o processo de ETL pode chegar a ocupar 70% do ciclo de desenvolvimento de um projeto de DW/DM.



 Ciclo de Vida de um Projeto de Data Warehouse por Kimball. Fonte: Kimball (2013).

A construção do processo ETL é baseado nos insumos coletados na etapa de Levantamento de Requisitos, no início do ciclo do de vida do projeto de DW. Certamente, ao avançar na trilha de dados, novos requisitos e necessidades podem ocorrer e eles devem ser adicionados aos documentos de requisitos do projeto e considerados no desenvolvimento do processo ETL.

Atenção

Conforme a evolução do desenvolvimento do processo ETL, as restrições de negócios são aplicadas aos tratamentos dos dados que entraram no processo e sua saída resultará em dados conforme sua natureza na origem e conforme as regras aplicadas durante a transformação. Todas essas informações precisam ser documentadas para atender às exigências de compliance/conformidade da organização. A rastreabilidade dos dados é de suma importância quando há a necessidade de comprovar que eles estão corretos e não foram adulterados durante o processo ETL.

A qualidade dos dados a serem extraídos das bases é um ponto a ser discutido. Para que o DW/DM apresente dados coerentes e confiáveis, é imprescindível que o sistema origem forneça dados com qualidade. Em muitos casos, o DW/DM acaba sendo um ambiente que valida os dados do sistema origem e os apontamentos de problemas encontrados devem ser direcionados para que a equipe responsável possa corrigi-los e melhorar a qualidade dos dados inseridos no sistema.

A segurança dos dados armazenados no DW/DM também é um tópico que deve receber atenção durante o projeto. Deve ser avaliado se todos os usuários que terão acesso ao DW/DM terão acesso a todos os dados. Caso existam restrições de acesso, controle de usuários ou controle de visualização, eles devem ser estudados para mapear como devem ser aplicados ao ambiente analítico.

A integração dos dados, que visa reunir e disponibilizar análises comuns sobre assuntos concentrados em sistemas diferentes, é fator muito importante no processo de ETL. Nesse momento, o entendimento realizado no levantamento de requisitos apoiará o desenvolvimento dos relacionamentos adequados para compor cenários de análises que, sem esse processo, é muito difícil de realizar.

A disponibilidade dos dados no ambiente analítico está relacionada com a prontidão deles no sistema origem e sua transferência para o DW/DM. As necessidades dos usuários podem ser variadas e isso reflete na carga dos dados. No entanto, essa carga depende de algumas variáveis que precisam ser conhecidas e atendidas para que o dado seja entregue conforme sua demanda.

Esse estudo deve ser realizado e discutido com os usuários para o bom entendimento e mapeamento dessas necessidades, pois uma alteração nesse sentido demanda grande esforço para ser atendida.

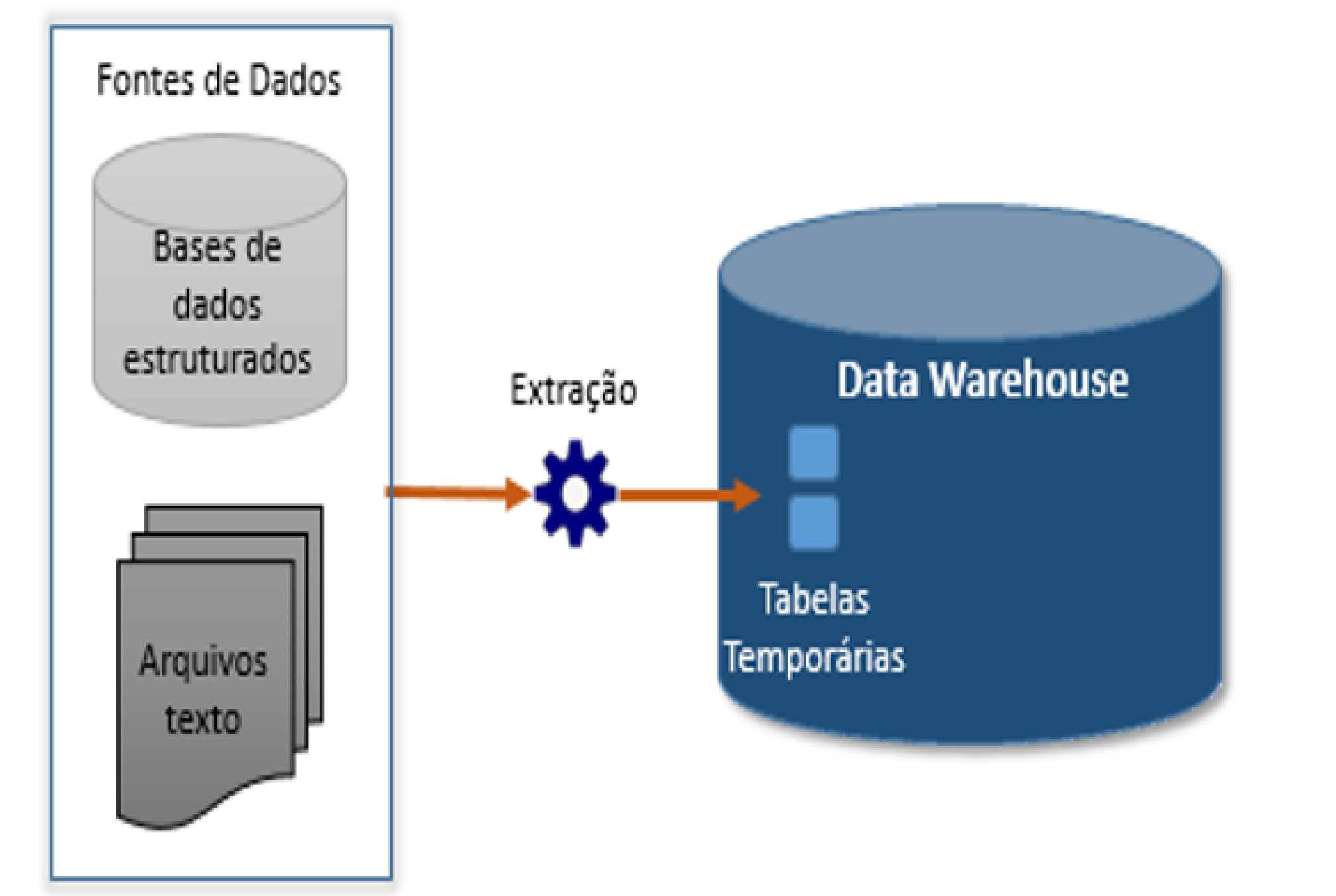
Nessa fase do projeto, geralmente as ferramentas que serão utilizadas foram discutidas pela equipe e já estão definidas. Atualmente, há diversas ferramentas para extração de dados e para construção da camada OLAP. O desenvolvimento dessa etapa deverá adequar-se aos requisitos de cada uma das ferramentas escolhidas, pois elas possuem diferentes formas de trabalho e essa escolha deverá basear-se nas condições e necessidades da organização.

Ralph Kimball divide o processo ETL em subsistemas e destaca quatro componentes principais: extração, limpeza e conformidade, entrega e gerenciamento dos dados. Nesta aula vamos trabalhar o primeiro componente e entender como ele é desenvolvido e o que é necessário para seu desenvolvimento.

Atenção! Aqui existe uma videoaula, acesso pelo conteúdo online

E (Extracting) - Extração de Dados

O primeiro passo do processo ETL é a extração dos dados. Essa é a porta de entrada para os dados oriundos dos sistemas transacionais ou sistemas origens, porque podemos ter dados obtidos de outros tipos de fontes de dados, como arquivos de texto. A figura ilustra a coleta dos dados das fontes de dados e o armazenamento nas tabelas temporárias.



Essa tarefa é executada com apoio do mapeamento dos dados realizado na fase de levantamento. Nesse momento, foi identificado se os dados existiam nos sistemas origem apontados pela área de negócio e confirmado pelo analista técnico responsável pelas fontes. As fontes devem ser confirmadas e a existência dos dados validada para que possam ser extraídos. Deve-se verificar se as regras para a extração serão atendidas pela disposição dos dados e se há as colunas de apoio para a leitura dos dados, como colunas de data de atualização dos registros.

Saiba mais

A primeira extração dos dados para o DW compreende um período de dados históricos existentes no sistema origem. A partir da primeira carga, o DW deverá receber os novos dados inseridos no sistema origem ou os dados que foram alterados. Tanto o processo de carga inicial quanto o de cargas incrementais devem ser preparados para a alimentação dos dados.


Devido ao volume de dados no DW, não é aconselhado que os dados sejam completamente apagados e carregados a cada carga realizada. Então, os dados novos ou alterados devem ser selecionados para que sejam levados no processo incremental de dados. Geralmente são usados alguns mecanismos para se saber que um registro foi alterado no sistema legado, por exemplo, na tabela de clientes do sistema origem deve haver uma coluna que grave a data de atualização do registro e o processo de extração irá verificar se a data de alteração corresponde ao período de movimento da carga.

Comentário

O processo de controle de alteração de registros não é uma tarefa simples de ser realizada e demanda muita atenção e estudo dos sistemas origens. Outra questão que precisa ser tratada com atenção é a necessidade de extrair dados de diferentes sistemas, cada qual com suas características e particularidades. As fontes de dados podem ser estruturadas ou não estruturadas e isso exige tratamentos diferentes para a leitura dos dados.

Fontes de dados (estruturados) e dados externos (não estruturados)

Veja mais informações sobre dados estruturados, semiestruturados e não estruturados.

 Clique nos botões para ver as informações.

[Dados estruturados](#)



Os dados estruturados são dados armazenados em estruturas planejadas para sua acomodação com uma semântica conhecida, tipo de dados, tamanho, ou seja, possuem um padrão definido. Os dados armazenados em um SGBD são considerados dados estruturados por apresentarem um padrão bem definido, com metadados.

[Dados semiestruturados](#)



Os dados semiestruturados que possuem alguma estrutura, mas não é crítico com tipos de dados, por exemplo. Exemplos de dados semiestruturados são arquivos XML, JSON, RDF e OWL.

[Dados não estruturados](#)




Os dados não estruturados são o oposto de um conteúdo com estrutura bem definida. Eles não possuem uma semântica explícita, não há padronização nos dados. São exemplos de dados não estruturados: arquivos de texto e páginas da internet. Com o crescente volume de dados disponibilizados na internet, há um despertar pela ingestão desses dados com o objetivo de complementar os dados do DW/DM para análises baseadas nesse rico repositório ou ainda construir um DW/DM baseado somente nesse tipo de dado.

A extração de dados estruturados apresenta-se simples frente aos outros tipos, pois os metadados facilitam o entendimento sobre o que será extraído e como deve ser tratado. A extração de dados não estruturados envolve tratamentos para a explicitação do significado do dado.

Exemplo

São exemplos de tratamentos: o processamento de dados em linguagem natural (PLN) e o uso de instrumentos semânticos. Após a extração dos dados, eles podem ser organizados e tratados para consumo nos sistemas analíticos.

Métodos de Extração

 Clique no botão acima.

Alguns processos de extração são desenvolvidos em linguagem procedural, que acessam arquivos criados pelos sistemas de origem dos dados, geralmente sistemas COBOL. As procedures eram criadas para acessar os arquivos, ler os dados contidos nos arquivos (vale ressaltar que esses arquivos possuem um padrão de identificação dos dados) e carregar os dados nas tabelas temporárias. Da mesma forma, a transformação dos dados e carga nas tabelas definitivas também ocorrem por procedures.

Esses processos precisam ser muito bem gerenciados para que o processamento dos dados ocorra de forma correta.

Então, devem ser desenvolvidos mapas para o acompanhamento das cadeias dos processos obedecendo sua ordem de execução, assim como as dependências de outros fluxos que precisam ser checados. Imagine carregar a venda sem informar ter o dado previamente carregado! Como existe a restrição dos dados entre o cliente e a venda do produto, haveria uma violação de integridade dos dados e, conseqüentemente, o registro seria perdido ou a carga paralisada.

Para auxiliar esse processo existem as ferramentas de ETL. Elas tornam mais simples a extração de diversas fontes de dados, facilitam a leitura de arquivos de texto semiestruturados, a aplicação de regras e validação dos dados, a construção dos fluxos (sequência de tarefas), entre outros. Seu ambiente gráfico traz um desenvolvimento mais intuitivo para os desenvolvedores, tornando o trabalho mais produtivo.

No mercado existem diversas ferramentas de ETL, cada uma com suas características, mas que possuem o mesmo objetivo: extrair de uma origem, transformar e carregar os dados em um ambiente destino. Alguns exemplos de ferramentas são: Informática Power Center, Oracle Warehouse Builder (OWB), SAS ETL Studio e Pentaho Data Integrator (PDI).

Nos exercícios, vamos utilizar a ferramenta Pentaho Data Integrator (PDI), disponível em: <https://sourceforge.net/projects/pentaho/files/>. Vamos conhecer a ferramenta mais à frente!

Documentação do processo ETL

Nessa fase é importante documentar tudo que está sendo feito para a extração dos dados, como de quais tabelas eles estão sendo copiados, de quais campos, quais serão os tratamentos aplicados a cada um dos campos e em quais tabela e campos eles serão adicionados.

A figura a seguir ilustra um exemplo para o documento de extração. O conjunto de colunas que compõe o grupo Extração de dados mapeia o endereço do dado que está sendo extraído. A coluna Origem informa se é uma tabela e o nome da tabela, a coluna Campo informa o nome do campo que será acessado, a coluna Tipo indica o tipo do dado do campo, a coluna tamanho informa o tamanho máximo do dado.

Extração dos dados				Carga Temporária		Transformação / Regras de Negócio	Carga dos Dados				
Origem	Campo	Tipo	Tamanho	Tabela	Campo	Regras de Negócio	Tabela	Campo	Tipo	Tamanho	Observação
Tabela Produto	codigo_produto	Integer		tmp_produto	cd_produto	Se o código do Produto estiver nulo, preencher com 1 - Não Informado	dim_produto	cd_produto	Integer	-	Código do produto no sistema origem
	produto	Varchar	45		ds_produto_tmp	-		ds_produto	Varchar	45	Descrição do produto no sistema origem
	fabricante	Varchar	100		ds_fabricante_tmp	-		nm_fabricante	Varchar	100	Nome do fabricante no sistema origem
					sk_produto	Verificar se existe na Dimensão Produto. Se existir preencher com a sk encontrada, senão inserir novo registro.		sk_produto	Integer	-	Chave primária do produto no DW
	-	-	-		id_produto	Chave primária da temporária.		-	-	-	-
	-	-	-								

 Documento de Extração de Dados. Fonte: O autor.

- O conjunto de colunas Carga Temporária informa qual tabela temporária receberá os dados extraídos. A coluna Tabela informa qual o nome da tabela temporária e a coluna Campo informa o nome do campo em que será inserido.
- A coluna Transformação/Regras de Negócio informa quais são os tratamentos a serem aplicados a cada uma das colunas contidas na tabela temporária para que os dados possam ser carregados na tabela definitiva.
- O conjunto de colunas Carga dos Dados informa a tabela destino e os campos em que os dados serão inseridos. A coluna Tabela informa o nome da Dimensão ou tabela Fato, a coluna Campo informa o nome do campo que receberá o dado, a coluna Tipo informa o tipo de dado que será inserido, a coluna Tamanho informa o tamanho máximo do dado que pode ser inserido e a coluna Observação contém dados importantes sobre o dado contido no campo.

Pentaho Data Integrator (PDI)

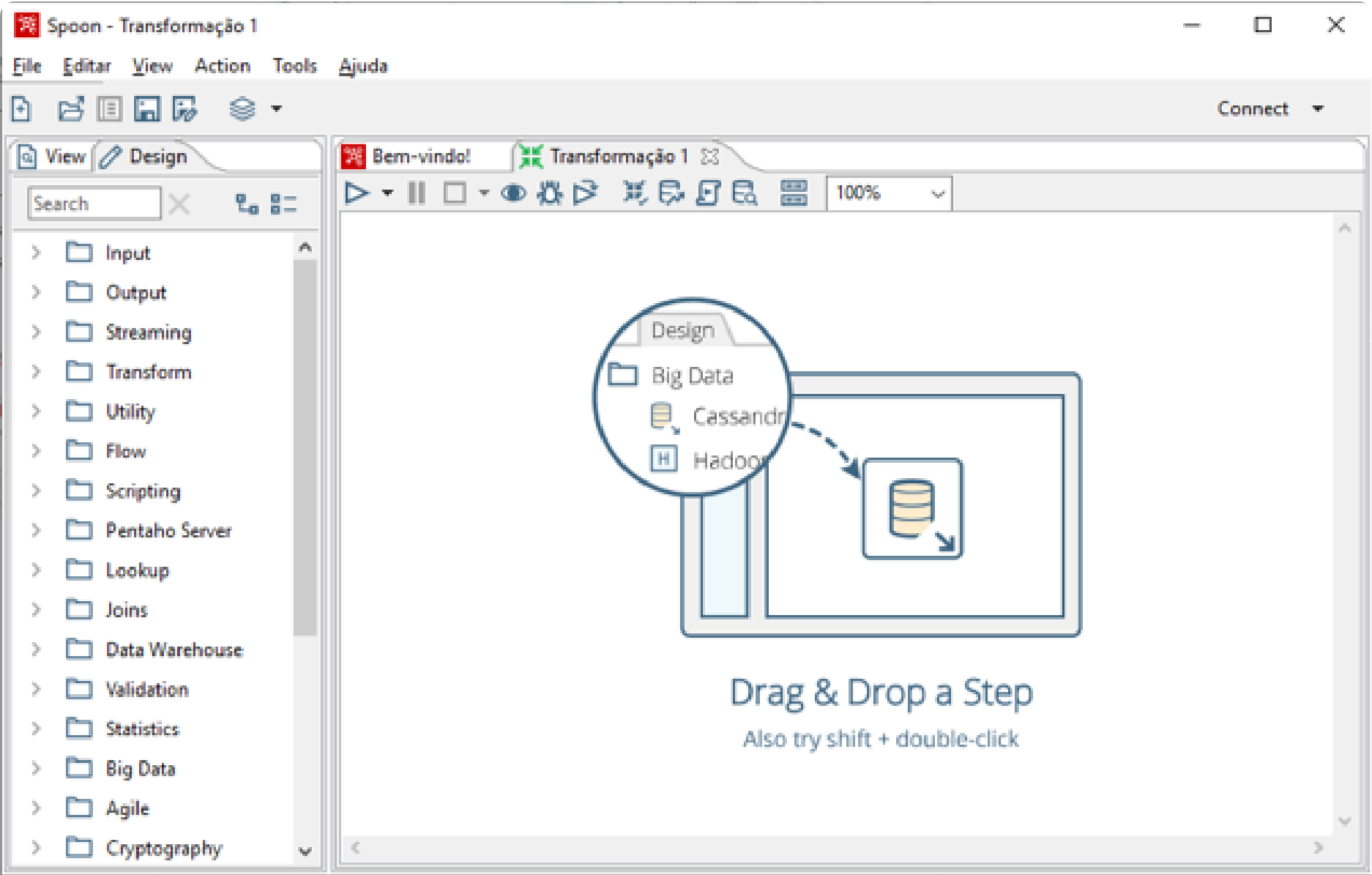
O primeiro passo a ser realizado na extração de dados é o acesso à base de dados na qual os dados estão armazenados. Após o acesso, os dados são selecionados conforme os critérios adotados, são copiados e inseridos nas tabelas temporárias, definidas na modelagem dos dados.

Agora vamos conhecer a ferramenta PDI. Ela é open source e faz parte do conjunto de ferramentas da plataforma Pentaho. Sua instalação é simples, bastando baixar o arquivo compactado com a denominação “pdi-ce-xxx-xxx.zip”, onde “xxx-xxx” é o número da última versão disponível, e descompactá-lo em algum local onde será executado, que pode ser em qualquer unidade de disco, inclusive em um pen drive.

Saiba mais

A sua interface gráfica é conhecida como Spoon, aberta com a execução de mesmo nome e é onde são criadas as transformações e os *jobs*. As transformações realizam as tarefas de tratamento de dados por meio de passos denominados steps e os *Jobs* reúnem as transformações, sequenciando as atividades que deverão ser executadas.

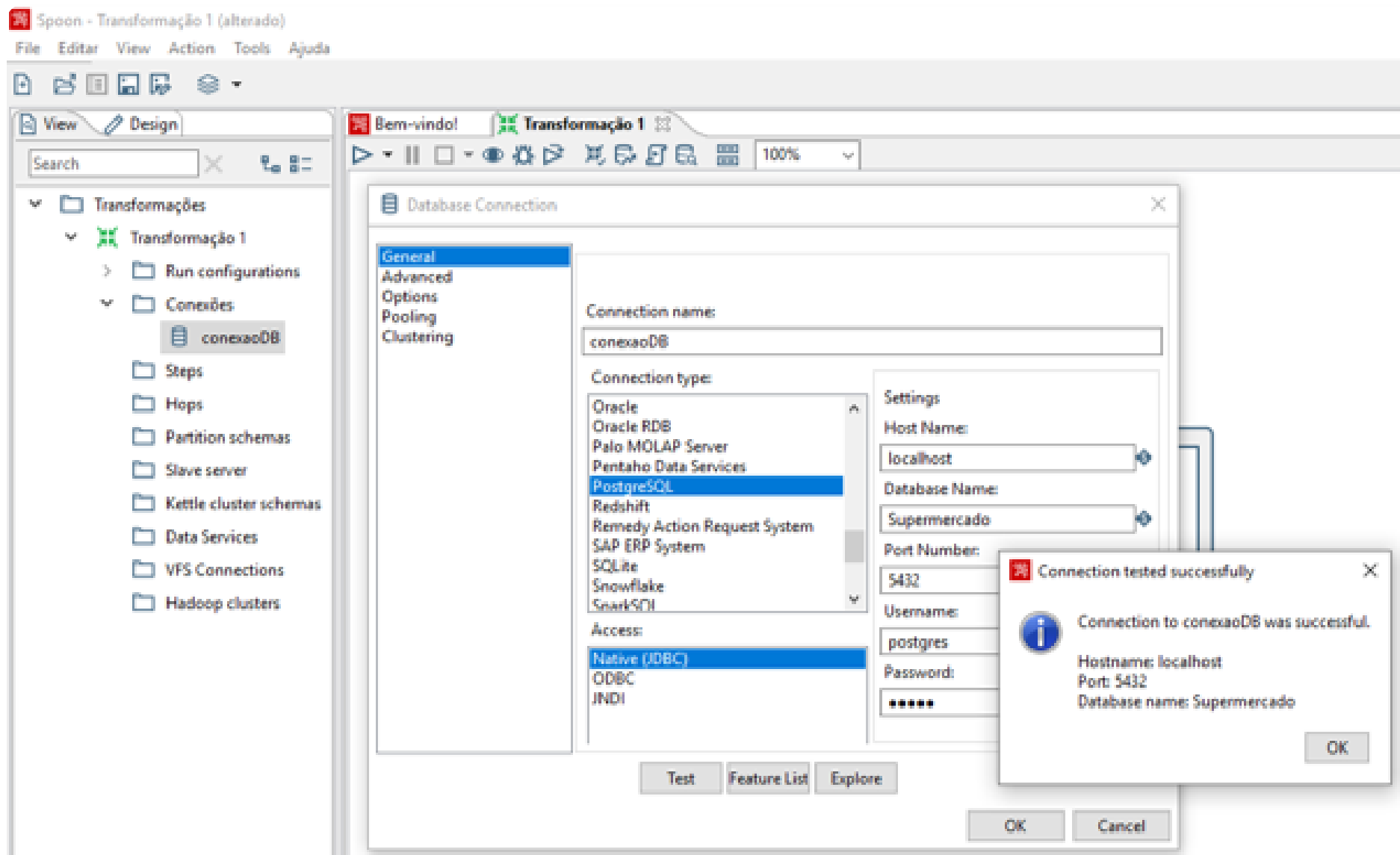
A figura ilustra a tela inicial do Spoon, em que no lado esquerdo estão os objetos chamados steps e, no lado direito, a área de trabalho em que os fluxos são criados.



Para iniciar o desenvolvimento é necessário estabelecer a conexão com a base de dados na qual os dados serão lidos e a base de dados que receberá os dados extraídos do sistema origem. Para isso, clique com o botão direito no objeto Transformações. No objeto Conexões, clique com o botão direito.

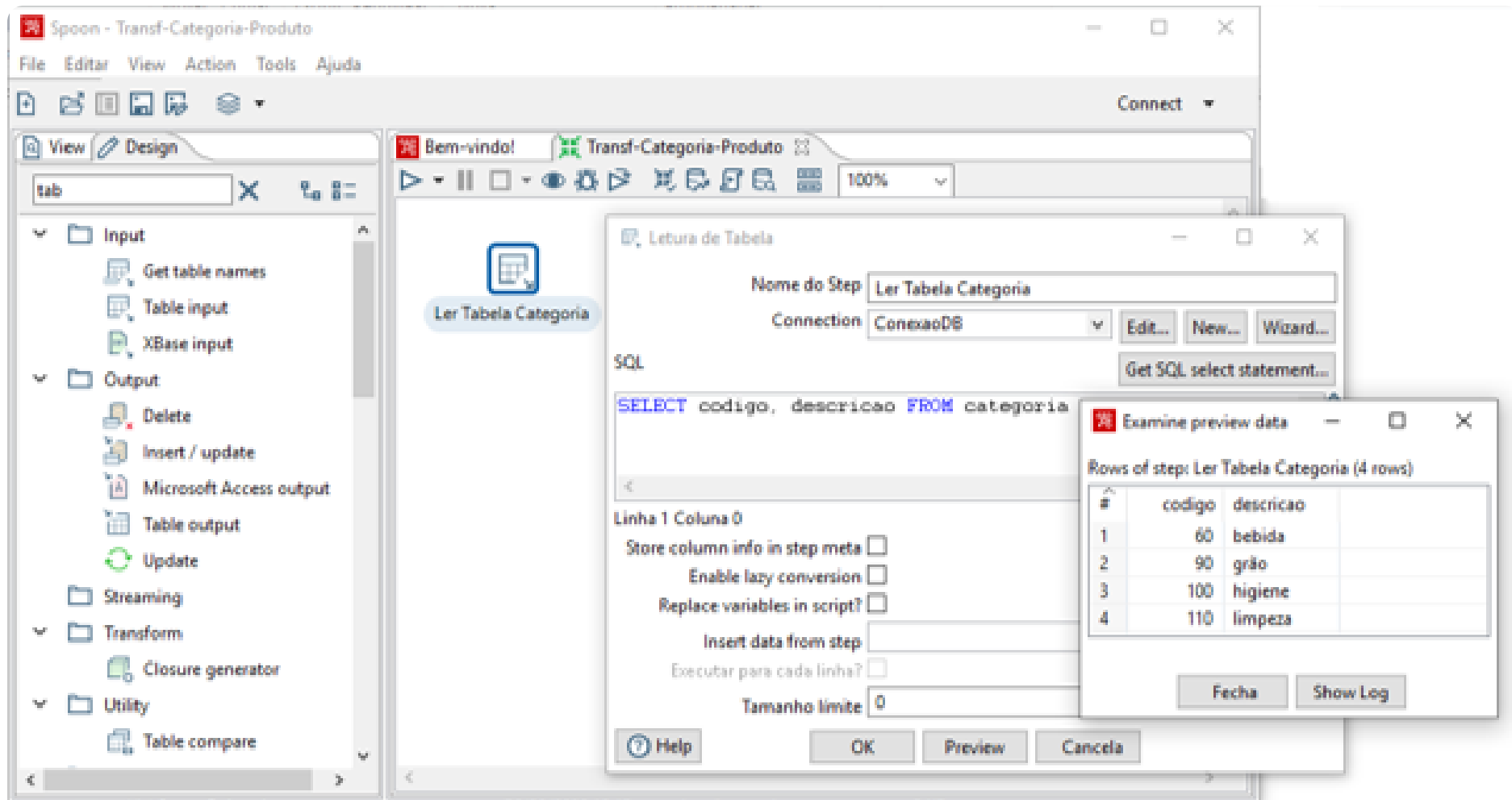
Em seguida, irá abrir a janela de criação da conexão, ilustrada na figura abaixo. Você deve informar os dados de acesso ao SGBD e o nome da base de dados.

Após o preenchimento dos dados, teste a conexão. Crie uma conexão para a base de dados origem (conexaoBD), onde os dados serão lidos e uma conexão para a base de dados destino (conexaoDW), onde os dados serão armazenados.

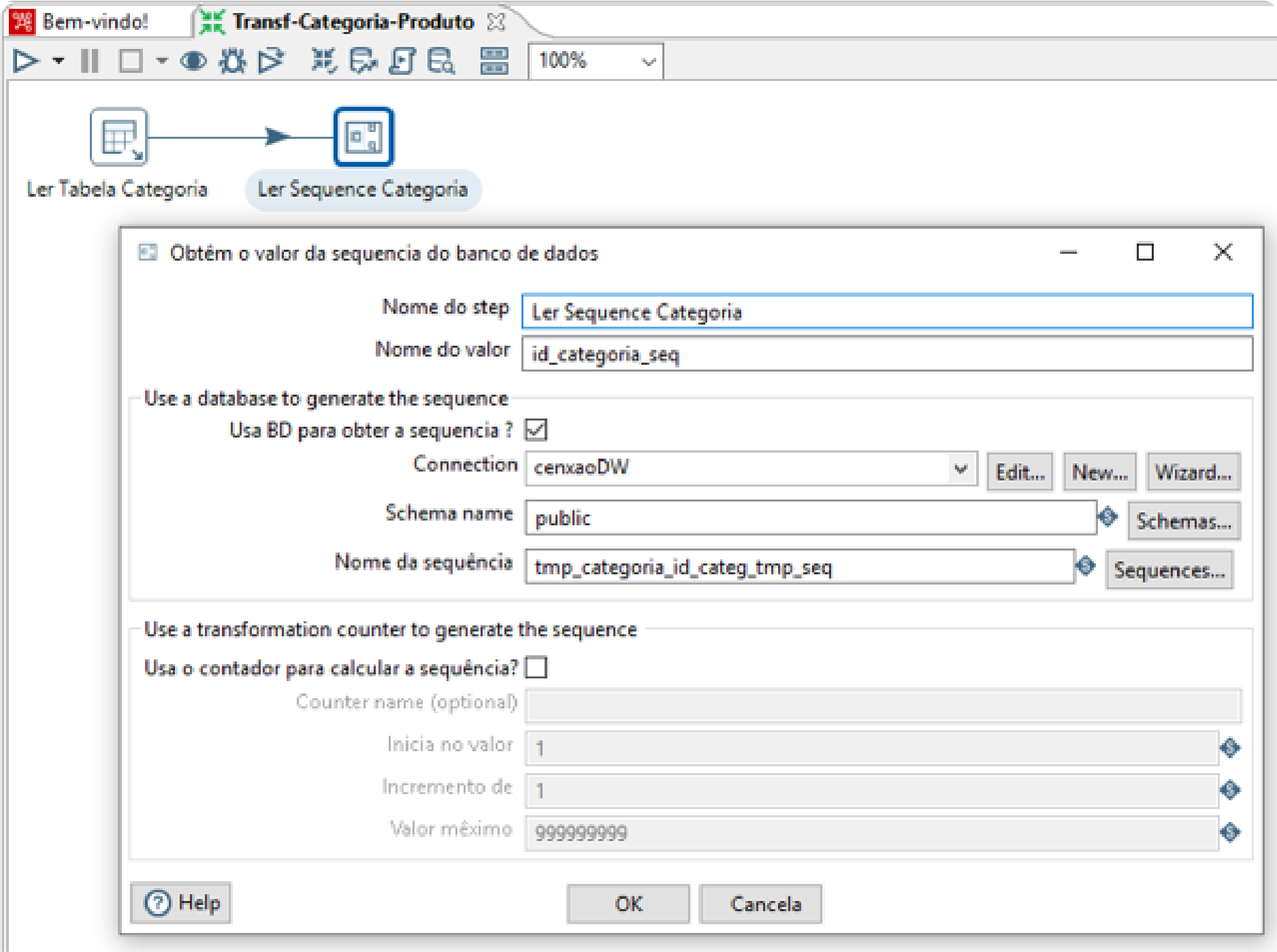


Após a conexão com a base de dados, a construção do fluxo de leitura dos dados pode ser iniciada. Isso é feito criando-se uma transformação (*transformation*), que é uma sequência de passos (*steps*) ligados por setas denominadas saltos (*hops*). O primeiro objetivo é ler os dados contidos na base do supermercado. Para isso, é necessário utilizar um step de leitura de dados. Então, adicione um step de Table Input, que pode ser encontrado na guia Design.

Conforme veremos na figura a seguir, abra o objeto e atribua um nome ao step, por exemplo Ler Tabela Categoria. Na combo Connection, escolha a conexão com a base de dados origem dos dados. No campo SQL, digite a query para a coleta dos dados. Caso deseje verificar o resultado obtido, clique no botão Preview.



Salve a transformação com o nome Transf-Categoria-Produto. Adicione um *step* Add *sequence* para que ele relacione a *sequence* criada para a tabela temporária. O campo Nome do Valor será o nome do campo referenciado no próximo *step*. Sselecione a conexão com a base de dados em que os dados serão inseridos e o nome do *step* Sequence criado na base de dados do DW.



Saiba mais

Os dados coletados no sistema origem serão inseridos na tabela temporária por meio do *step* Table Output. Ligue o *step* Sequence ao *step* destino para que seja possível visualizar o campo `id_categoria_seq`. Para isso, basta segurar a tecla Shift e arrastar até o outro *step*. Os *steps* devem ser ligados na sequência Tabela Origem > Sequence > Tabela destino.

Insira um nome para o *step*, selecione a conexão da base de dados do DW, selecione o nome da tabela destino. Marque a opção *Truncate Table*, para que o processo limpe a tabela temporária a cada nova execução e a opção *Specify database fields* para que seja possível relacionar os campos de origem (*Stream field*) e os campos de destino dos dados (*Table field*), na aba *Database fields*.

Bem-vindo!

Transf-Categoria-Produto

Ler Tabela Categoria

Ler Sequence Categoria

Carregar Temp Categoria

Saída a Tabela

Nome do Step

Carregar Temp Categoria

Connection

ConexaoDW

Edit...

New...

Wizard...

Target schema

public

Navega...

Target table

tmp_categoria

Navega...

Commit size

1000

Truncate table

☒

Ignore insert errors

☐

Specify database fields

☒

Main options

Database fields

Colunas a inserir:

#	Table field	Stream field	
1	id_categ_tmp	id_categoria_seq	
2	cd_categoria	codigo	
3	ds_categoria_tmp	descricao	

Get fields

Enter field mapping

Help

OK

Cancela

SQL



Após a execução do fluxo, os dados são carregados na tabela temporária Categoria.

Edit Data - postgres (localhost:5432) - DWSuper - tmp_categoria

File Edit View Tools Help

No limit

	id_categ_tmp [PK] serial	sk_categoria_tmp integer	ds_categoria_tmp character varying(45)	cd_categoria integer
1	7		bebida	60
2	8		grão	90
3	9		higiene	100
4	10		limpeza	110

Observe que nem todas as colunas da tabela temporária serão preenchidas nesse momento. A coluna sk_categoria_tmp dará suporte à validação dos dados e receberá a chave primária da dimensão Categoria, caso o elemento exista na tabela. Se não existir, ele será inserido, conforme a regra definida no documento de extração.

Dica

Com o conhecimento adquirido, construa os fluxos de leitura dos dados para as tabelas temporárias clientes, produto e para a tabela vendas.

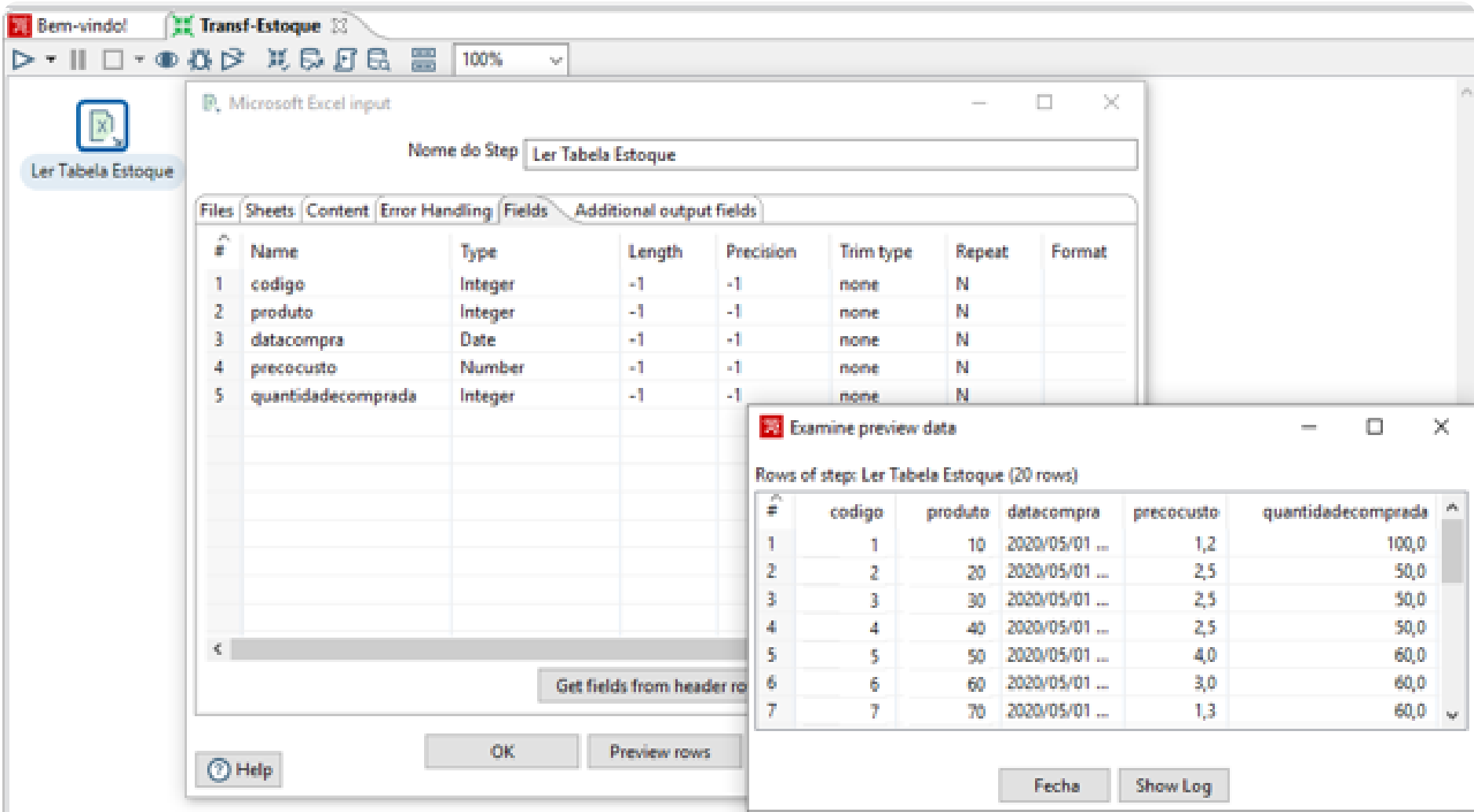
Para executar o exemplo apresentado no vídeo a seguir, deve se usar o arquivo dwsuper criado a partir da aula 3, agora renomeado a_dwsuper. Para criar o arquivo a_supermercado pode proceder da mesma forma usada na criação do arquivo dwsuper ou usar o script [BDsupermercado.txt](#) <https://stecine.azureedge.net/webaula/estacio/go0683/anexo/BDSupermercado.txt> como referência.

Atenção! Aqui existe uma videoaula, acesso pelo conteúdo online

Leitura de dados de um arquivo Excel

Para acessar um arquivo em formato Excel deve ser usado o step Microsoft Excel Input. Adicione esse step em uma nova transformação e salve com o nome Transf-Estoque.

Na caixa File ou directtory, escolha o arquivo Excel onde estão os dados do estoque e adicione para que seja visualizado na seção Selected Files. Na aba Sheets, clique no botão Get sheetname para selecionar a aba da planilha Excel onde estão os dados a serem coletados. Na aba Fields, clique no botão Get Fields from header row para que os campos da planilha sejam adicionados.



[Você pode utilizar a planilha Excel baseestoque.xlsx para a tabela origem Estoque disponibilizada em arquivo para download.](https://webaula/estacio/go0683/galeria/aula6/anexo/baseestoque.xlsx)
<<https://webaula/estacio/go0683/galeria/aula6/anexo/baseestoque.xlsx>>

Atenção! Aqui existe uma videoaula, acesso pelo conteúdo online

As ferramentas PAN e Kitchen

Conforme falado anteriormente, o Spoon é a interface gráfica que permite a construção dos fluxos de tarefas que serão executadas. O PAN é um programa que executa as transformações no ambiente de produção do DW/DM. E o Kitchen é um programa que executa os Jobs, que consistem em sequências de transformações. Ambas as ferramentas executam os fluxos em batch.

Periodicidade de Extração dos Dados

A periodicidade de extração dos dados foi definida nos passos anteriores e nesse momento ela será utilizada para scheduler o processo de ETL. De acordo com as definições, os Jobs serão executados automaticamente, assim como toda a cadeia sequenciada.

Atenção

A periodicidade de carga é definida com base nas necessidades da organização. Alguns sistemas necessitam que a carga seja efetuada várias vezes ao longo do dia para que as informações sejam as mais atuais possíveis, enquanto outras são carregadas mensalmente.

Quando a carga do DW/DM depende da liberação de outros sistemas, o processo deve checar se o antecessor foi liberado para que a carga seja iniciada. É importante respeitar as tarefas antecessoras para não haver problemas de inconsistências nos dados.

As ferramentas PAN e Kitchen

Para que esse fluxo funcione de forma adequada, há necessidade de controle de cargas, de precedências e ferramentas de gerenciamento de tarefas que auxiliem o acompanhamento das tarefas.

Estudamos nesta aula o processo ETL, seu objetivo, os métodos de extração, tipos de fontes de dados e conhecemos a ferramenta de ETL PDI e como criar um processo de extração de dados.

Agora, vamos fixar o entendimento!

Atividade

O Processo ETL é uma das etapas do projeto de Data Warehouse. Ele é responsável por:

- a) Definir as regras de negócio e carregar os dados nas tabelas destino do DW/DM.
 - b) Extrair os dados das fontes de dados origem, transformar os dados e carregar nas tabelas destino do DW/DM.
 - c) Retirar os dados das fontes de dados origem, transformar os dados e carregá-los nas tabelas destino do DW/DM e do sistema transacional.
 - d) Aplicar tratamento aos dados que estão nos sistemas origem antes de extrair para o DW/DM.
 - e) Extrair os dados das fontes de dados origem, transformar os dados e carregá-los nas tabelas destino do sistema origem.
-

Sobre as fontes de dados que alimentam o DW/DM é correto afirmar que:

- a) Os dados só podem ser extraídos de bases de dados estruturadas.
 - b) Os dados devem estar em um mesmo sistema origem.
 - c) Podem ser estruturadas e semiestruturadas.
 - d) Podem ser estruturadas, semiestruturadas e não estruturadas.
 - e) Podem ser semiestruturadas e não estruturadas.
-

O processo ETL pode ser feito por meio de linguagem procedural e por ferramentas que auxiliam o desenvolvimento das tarefas. A ferramenta de ETL:

- a) Aplica os tratamentos aos dados armazenados previamente por outro serviço de leitura de dados e carrega os dados na base de dados de destino.
- b) Conecta-se à base de dados origem, de onde os dados devem ser lidos, insere-os em uma área temporária, executa tarefas de tratamento, mas não carrega na base de dados destino.
- c) Conecta-se à base de dados origem, de onde os dados devem ser lidos, copia os dados insere-os em uma área temporária, executa tarefas de tratamento e carrega os dados na base de dados de destino.

- d) É limitada, realizando apenas os tratamentos nos dados que já estão inseridos na base de dados do DW/DM.
- e) Conecta-se à base de dados origem, de onde os dados devem ser lidos, insere-os em uma área temporária, mas dificulta o armazenamento dos dados na base de dados destino, pois não consegue realizar mais do que uma conexão com o SGBD.
-

Notas

Referências

KIMBALL, M. R. R. **The Data Warehouse Toolkit - The Definitive Guide to Dimensional Modeling**. 3. ed. Indianapolis: John Wiley Sons, 2013.

MACHADO, Felipe N. R. **Tecnologia e Projeto de Data Warehouse**. 6. ed. São Paulo: Érica, 2013.

PITON, R. **Data Warehouse Passo a Passo** – O guia prático de como construir um Data Warehouse do zero. Porto Alegre: Edição do Autor, 2018.

Próxima aula

- Processo ETL – A transformação dos dados e aplicação de regras de negócio;
- Processo ETL – A carga dos dados nas tabelas definitivas.

Explore mais

- Conheça mais sobre as ferramentas de ETL no artigo *A Survey on ETL Tools de 2016* (JEBEULA, T.), que compara algumas ferramentas e apresenta suas vantagens e desvantagens. <https://www.semanticscholar.org/paper/A-Survey-on-ETL-Tools-Jebeula/6e2861559d26c579c3c37ab56749dd71e82086fb>.