

cin.ufpe.br



UNIVERSIDADE FEDERAL DE PERNAMBUCO



PENTAHO DATA INTEGRATION

SEMANA 3

Jarley Nóbrega – jpn@cin.ufpe.br

Pentaho Data Integration



|| Agenda

Trabalhando com Banco de Dados

O Modelo de Dados da WCM

Desenvolvendo e Implementando um Datamart

Automação do Processo de ETL

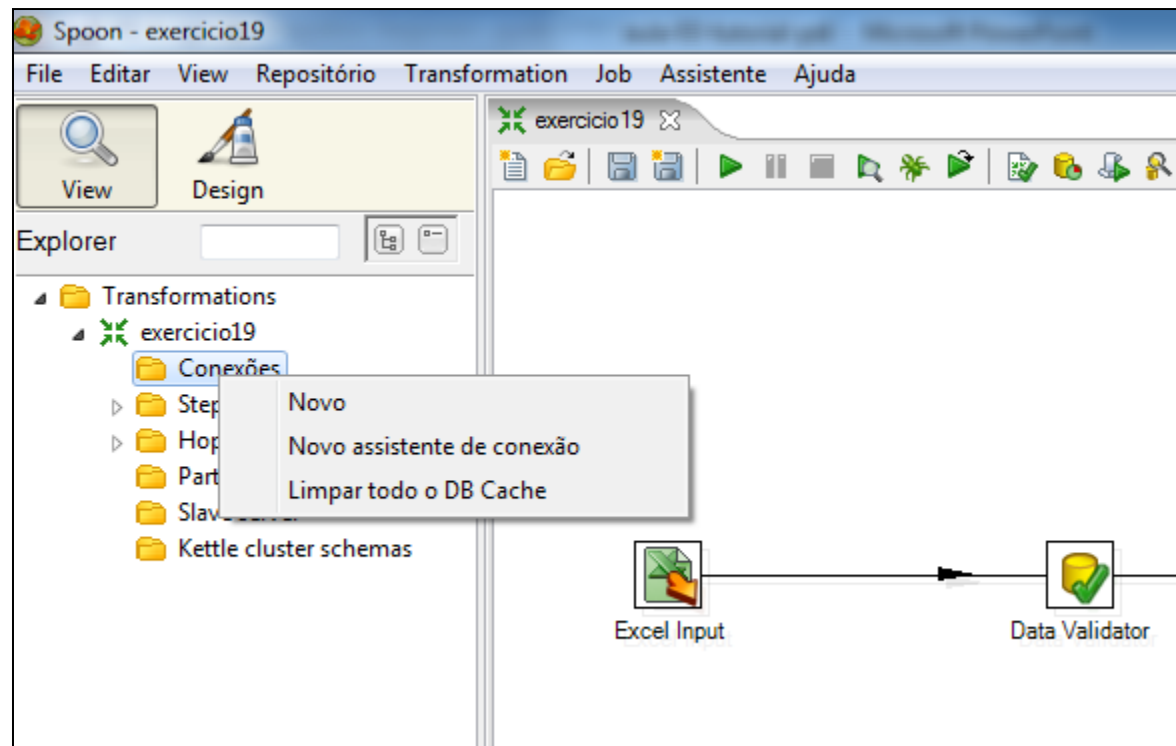




TRABALHANDO COM BANCO DE DADOS NO PDI

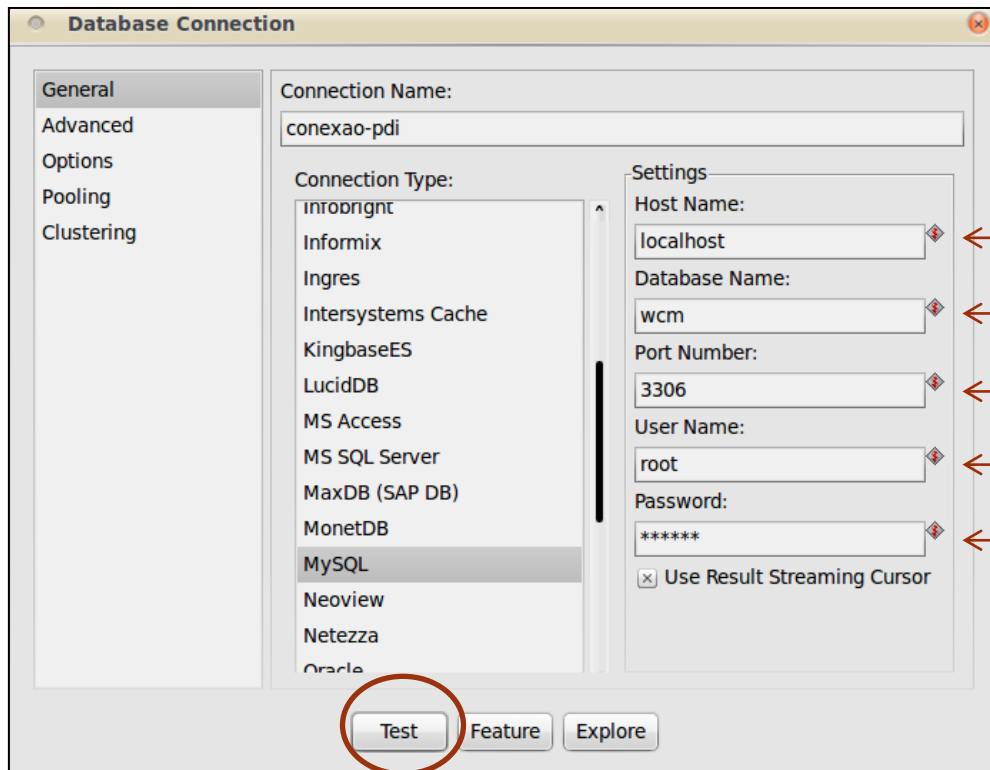
Trabalhando com banco de dados

- Como acessar um banco de dados no PDI
 - Criando uma conexão para uma transformação



Trabalhando com banco de dados

- Como acessar um banco de dados no PDI
 - Criando uma conexão para uma transformação



Nome/endereço do servidor

Nome do esquema do banco

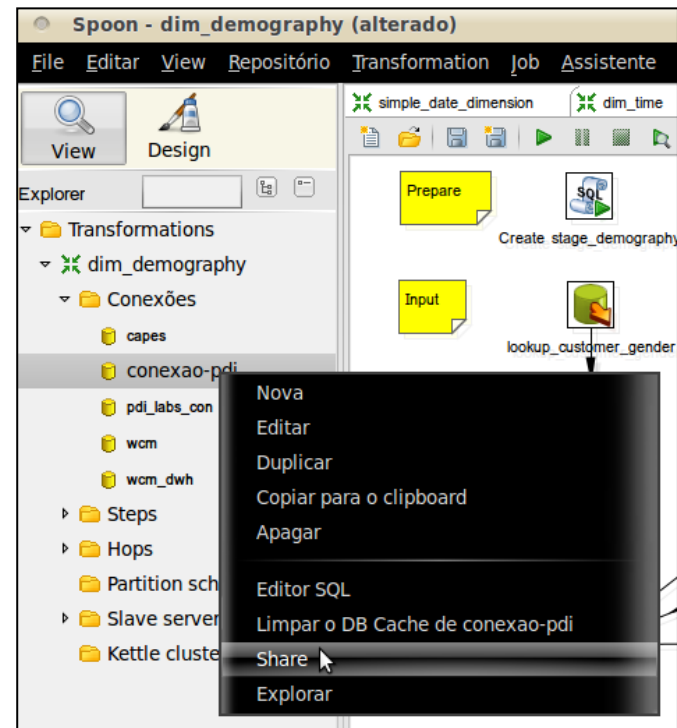
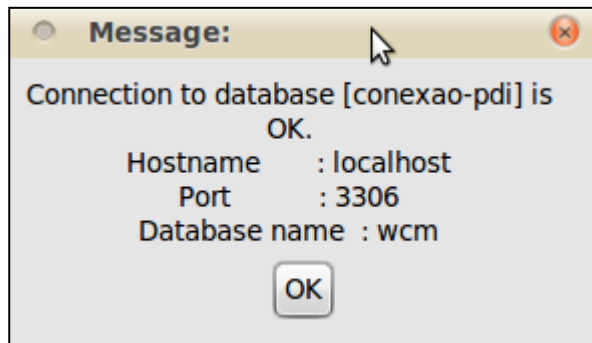
Porta do servidor (padrão: 3306)

Nome do usuário do banco

senha

Trabalhando com banco de dados

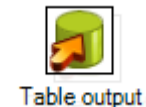
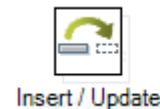
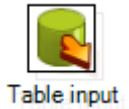
- Como acessar um banco de dados no PDI
 - ▣ Compartilhando uma conexão com todas as transformações e jobs



Trabalhando com banco de dados

Principais steps para leitura, armazenamento, atualização e remoção de registros

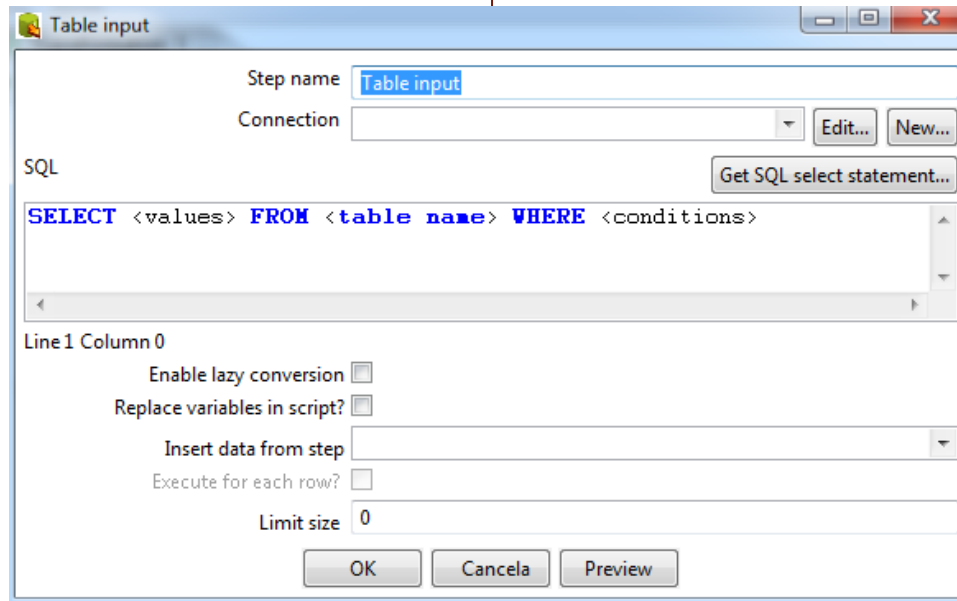
- Categoria Input
 - *Table input*
- Categoria Output
 - *Table output*
 - *Insert / Update*
 - *Delete*



Trabalhando com banco de dados

Configuração básica

■ *Table Input*



The screenshot shows a dialog box titled "Table input". It contains the following fields and controls:

- Step name:** A text field containing "Table input".
- Connection:** A dropdown menu with "Edit..." and "New..." buttons.
- SQL:** A text area containing the template: `SELECT <values> FROM <table name> WHERE <conditions>`. A button "Get SQL select statement..." is to the right.
- Line 1 Column 0:** A label for the first column.
- Enable lazy conversion:** A checkbox.
- Replace variables in script?:** A checkbox.
- Insert data from step:** A dropdown menu.
- Execute for each row?:** A checkbox.
- Limit size:** A text field containing "0".
- Buttons:** "OK", "Cancela", and "Preview" at the bottom.

Configuração básica

- *Table output*

The screenshot shows the 'Table output' dialog box in DBeaver. The 'Nome do Step' field contains 'Table output'. Below it are fields for 'Connection', 'Target schema', and 'Target table', each followed by 'Edit...' or 'New...' buttons. The 'Commit size' is set to '1000'. There are three unchecked checkboxes: 'Truncate table', 'Ignore insert errors', and 'Specify database fields'. At the bottom, there are two tabs: 'Main options' and 'Database fields'. The 'Database fields' tab is active, showing a table titled 'Fields to insert:' with columns '#', 'Table field', and 'Stream field'. The first row has the number '1' under the '#' column. To the right of the table are two buttons: 'Get fields' and 'Enter field mapping'.

Trabalhando com banco de dados

Principais steps para leitura, armazenamento, atualização e remoção de registros

- Categoria Scripting

- *Execute SQL Script*



- Categoria Lookup

- *Database join*
- *Database lookup*
- *Check if a column exists*



Database join



Database lookup



Check if a column exists

- *Database lookup*

IN1177 - Bar

Exercícios 20 e 21

- Criando um esquema de banco de dados
- Criando tabelas
- Carregando dados nas tabelas através de uma transformação

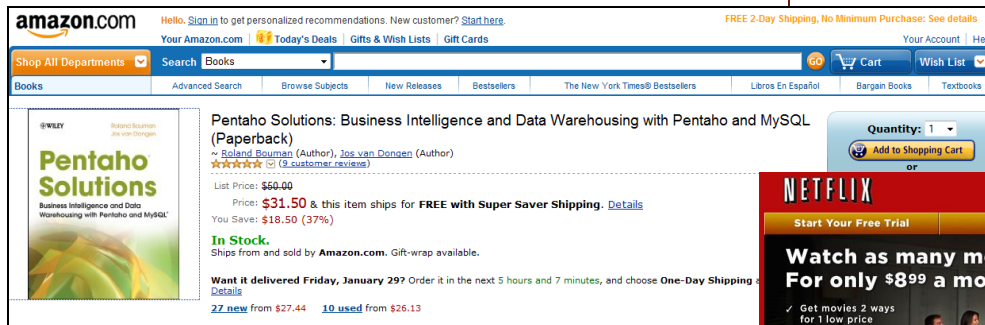


O MODELO DE DADOS DA WCM

World Class Movies

O Negócio da WCM

- Venda e locação de filmes pela Web
 - Concorrentes
 - Amazon.com
 - Netflix.com



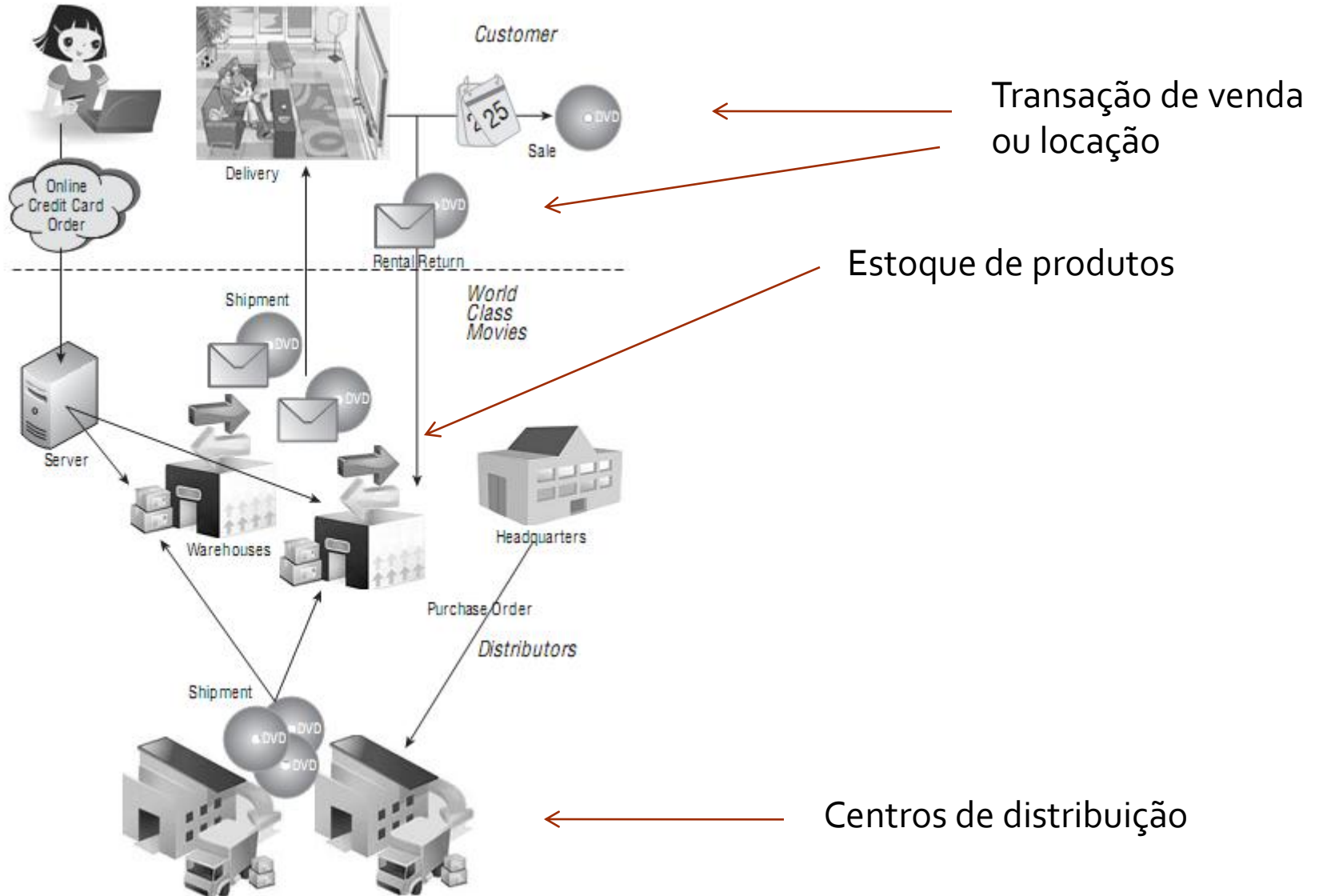
World Class Movies

Processos de negócios
integrados

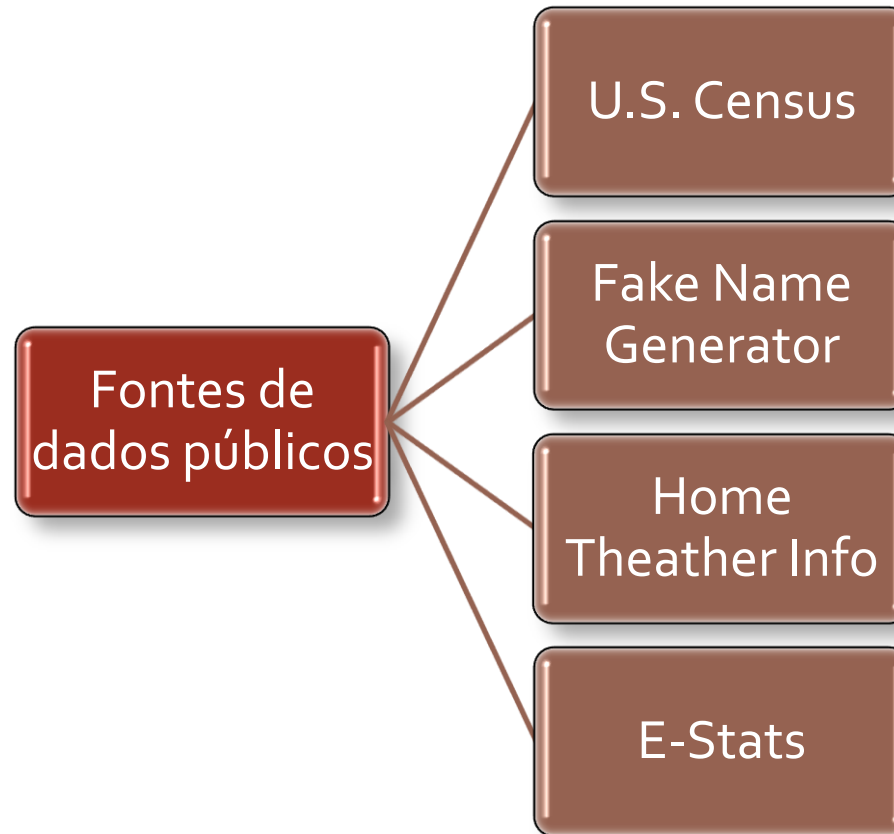
Gerenciamento
dos pedidos
dos clientes

Reposição do
estoque de
filmes

Fluxo do Processo da WCM

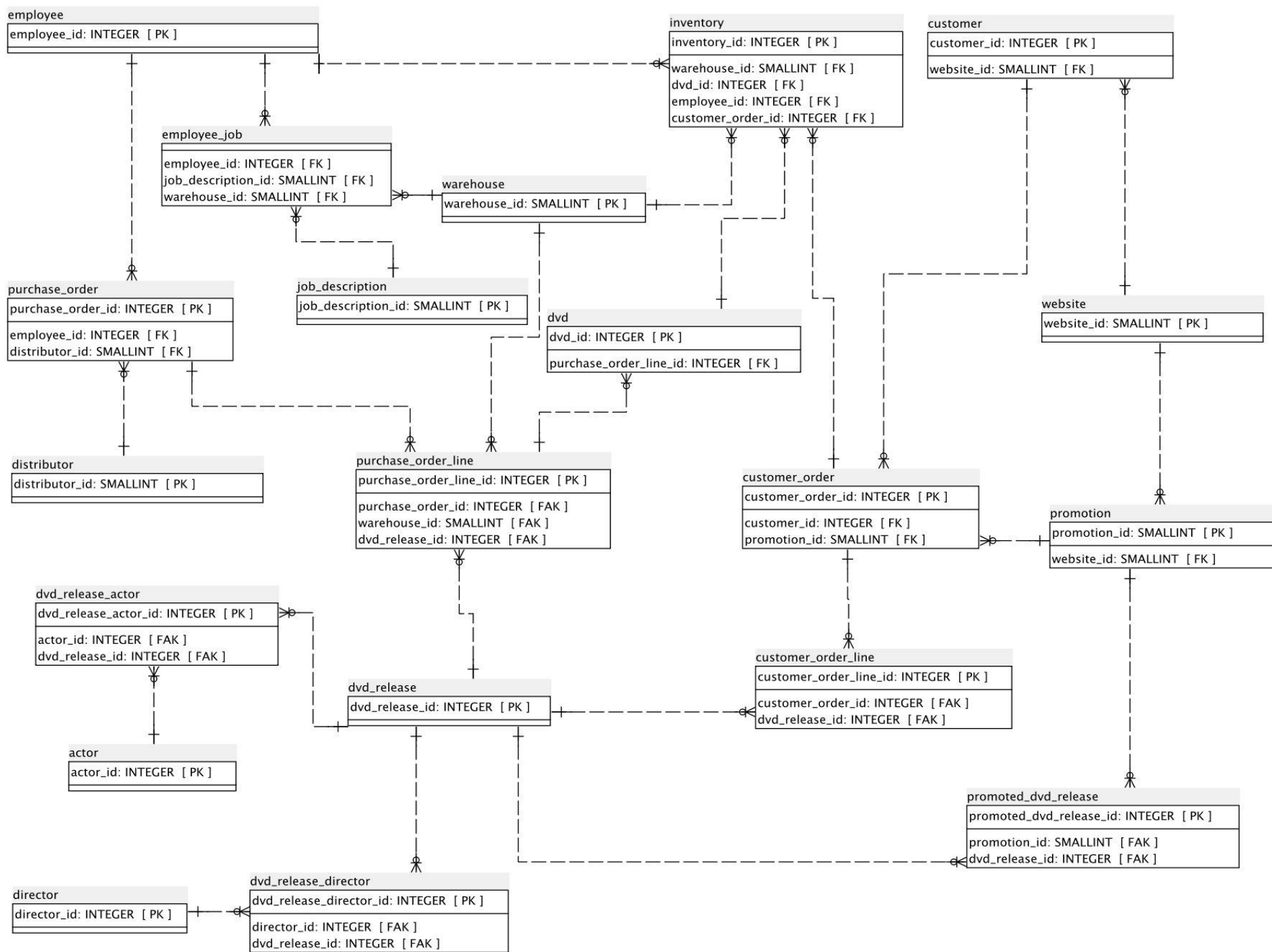


Aquisição e Geração de Dados



World Class Movies

- Principais papéis
 - Clientes, Produtos e Pedidos;
 - Dois tipos de pedidos:
 - Ordens de compra e pedidos de clientes
 - Outras entidades:
 - Centros de distribuição;
 - Empregados;
 - Clientes;
 - Distribuidores.
 - Pedidos gerados pelo site
 - Podem ter uma promoção associada



Questionamentos do Negócio

Departamento	Questionamentos
Finanças e Vendas	<ul style="list-style-type: none">• Qual o tempo de vendas por região, mês e categoria de filmes?• Que categoria de filmes gerou o maior volume de vendas de forma constante no tempo?• Qual é o nosso desempenho comparado com o mercado de entretenimento como um todo?• Nós estamos crescendo mais rapidamente ou mais lentamente que os nossos principais competidores

Questionamentos do Negócio

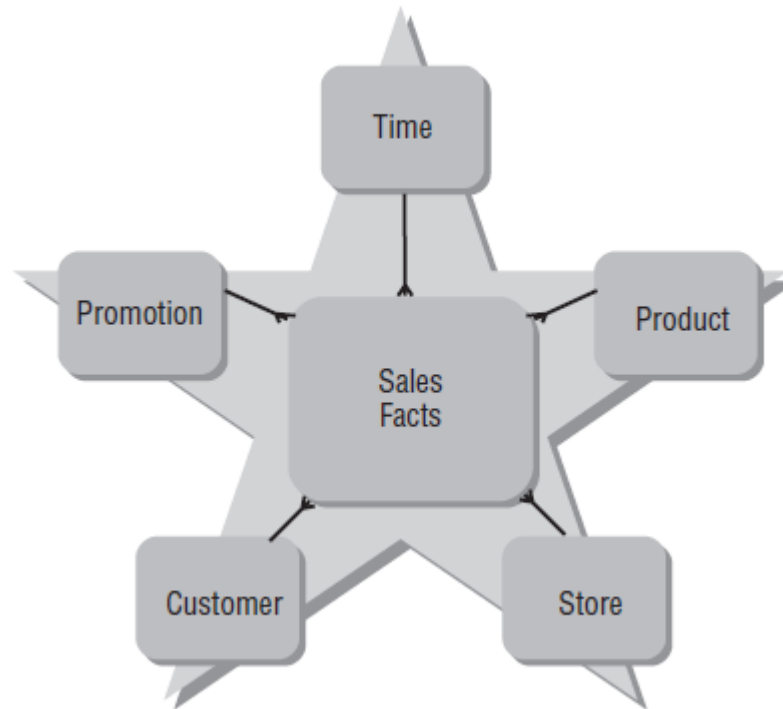
Departamento	Questionamentos
Logística	<ul style="list-style-type: none">• Qual o desempenho dos nossos distribuidores em termos de variedade de produtos, preço e prazo de entrega?• Como nós podemos aperfeiçoar os nossos custos de distribuição?

Questionamentos do Negócio

Departamento	Questionamentos
Marketing e Desenvolvimento de Produtos	<ul style="list-style-type: none">• Qual o tempo médio de relacionamento dos nossos 100 principais clientes comparados com os 100 menores?• Como nós podemos segmentar os nossos clientes baseados na análise RFM (<i>recency, frequency, monetary</i>)?• Nós temos dados dos clientes que podem ser usados para prever lucros ou prejuízos futuros?• Como nós podemos rastrear o ciclo de vida de um produto em determinados canais de venda?• Quais lançamentos de DVD geram maior valor de revenda baseado nas características do produto, como ator, diretor ou gênero do filme?

Modelagem do Negócio

- Esquema Estrela



(Bouman and Dongen, 2009)

Modelagem do Negócio

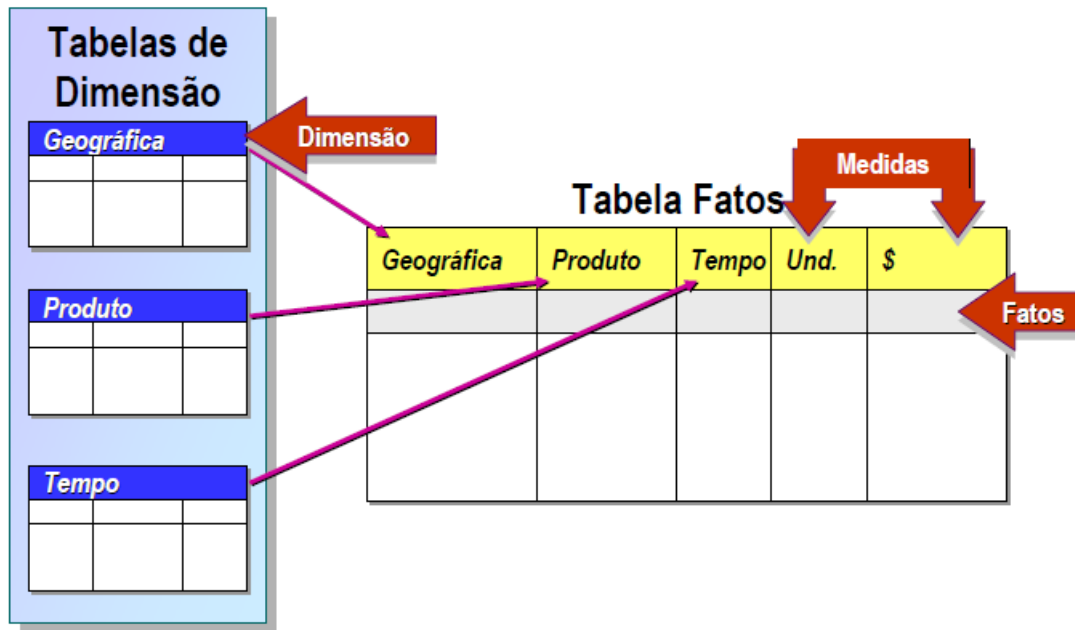
Alguns conceitos para
relembrar

- Tabela de fatos
- Tabela de dimensões
- Convenções de nomes
- Arquitetura em barramento
- Surrogate keys (chaves artificiais)
- Colunas de auditoria
- SCD tipos 1, 2, 3, 4, 5 e 6

Modelagem do Negócio

■ Tabelas de dimensões e fatos

Componentes do Modelo Dimensional



(IN1177 - Fidalgo, 2009)

Modelagem do Negócio

- Convenções de nomes de tabelas

Prefixo	Tabelas
STG_	Staging
HIS_	Arquivos históricos
DIM_	Dimensões
FCT_	Fatos
AGG_	Agregações
LKP_	Lookup

Modelagem do Negócio

- Slowly Changing Dimensions (SCD)
 - Tipo 1 - Overwrite

Existing situation

Customer_key	Customer_id	Customer_Name	Customer_City
1	22321	Humphries	Toronto

New situation

Customer_key	Customer_id	Customer_Name	Customer_City
1	22321	Humphries	Vancouver

(Bouman and Dongen, 2009)

Modelagem do Negócio

- Slowly Changing Dimensions (SCD)
 - ▣ Tipo 2 – Add Row

Existing situation

Customer_key	Customer_id	Customer_Name	Customer_City	Valid_from	Valid_to	Current_record
1	22321	Humphries	Toronto	1900-01-01	9999-12-31	1

New situation

Customer_key	Customer_id	Customer_Name	Customer_City	Valid_from	Valid_to	Current_record
1	22321	Humphries	Toronto	1900-01-01	2008-04-30	0
2	22321	Humphries	Vancouver	2008-05-01	9999-12-31	1

(Bouman and Dongen, 2009)

Modelagem do Negócio

- Slowly Changing Dimensions (SCD)
 - Tipo 3 – Add Column

Existing situation

Customer_key	Customer_id	Customer_Name	Customer_City	Customer_City_Old
1	22321	Humphries	Toronto	Toronto

New situation

Customer_key	Customer_id	Customer_Name	Customer_City	Customer_City_Old
1	22321	Humphries	Vancouver	Toronto

(Bouman and Dongen, 2009)

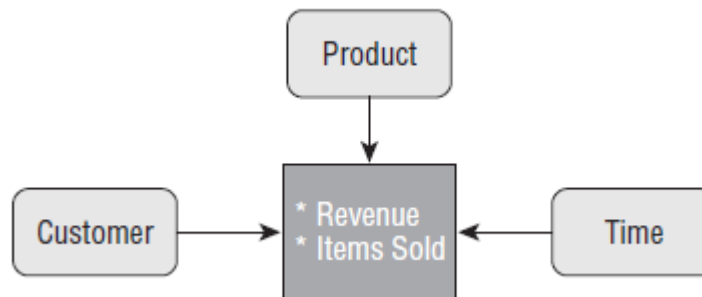


DESENVOLVIMENTO E IMPLEMENTAÇÃO DE UM DATAMART



Projetoando o modelo

- Após garantir que ...
 - Requisitos de negócio estão claros e entendidos;
 - As fontes de dados estão identificadas;
 - O conteúdo e a **qualidade** das fontes de dados foram identificados;
- Nosso modelo inicial



|| Construindo os Datamarts da WCM

- Duas conexões de banco no PDI
 - WCM: banco de dados operacional
 - WCM_DWH: data warehouse

Projetoando o modelo

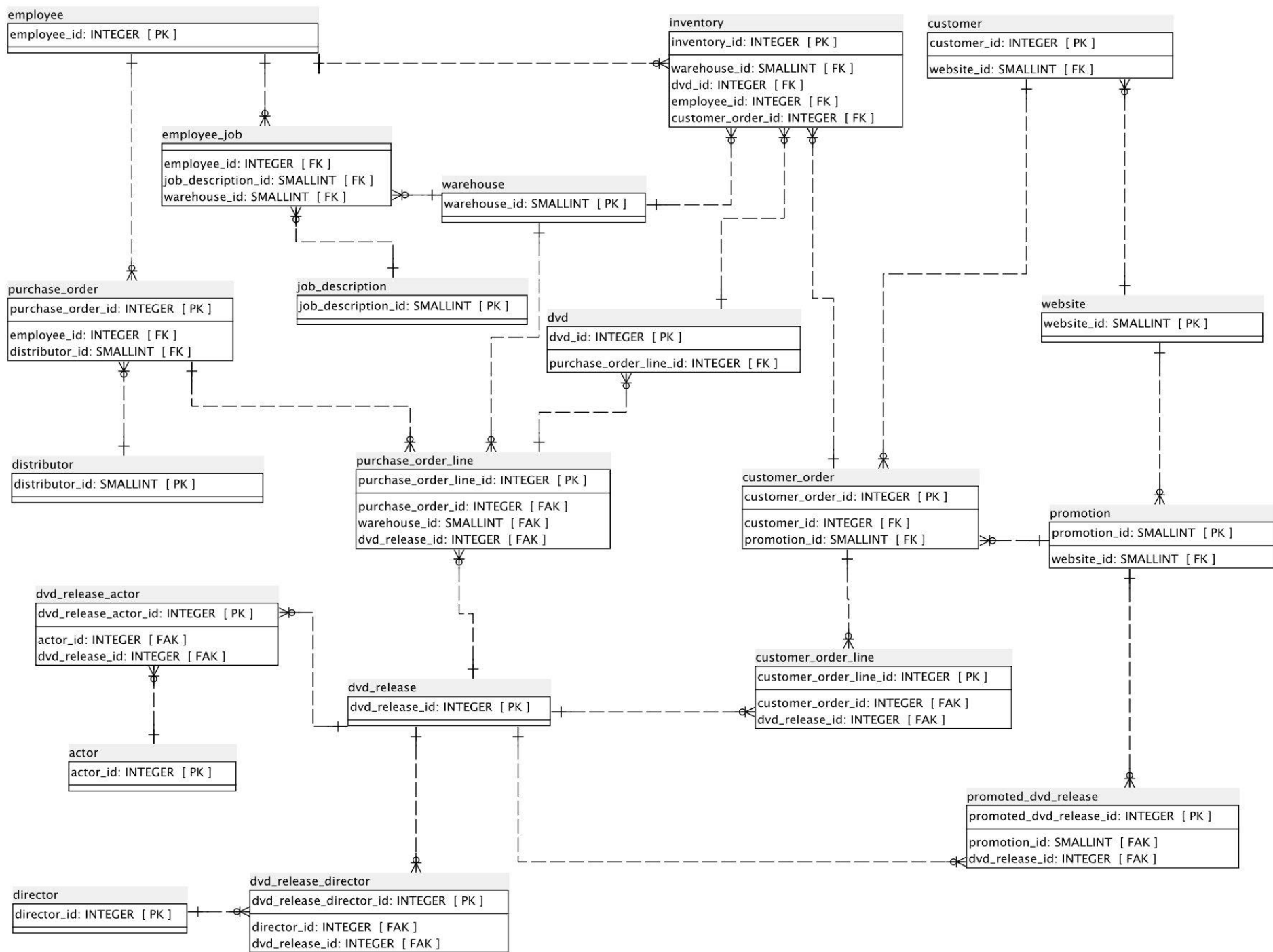
▪ Dimensão Customer (1ª versão)

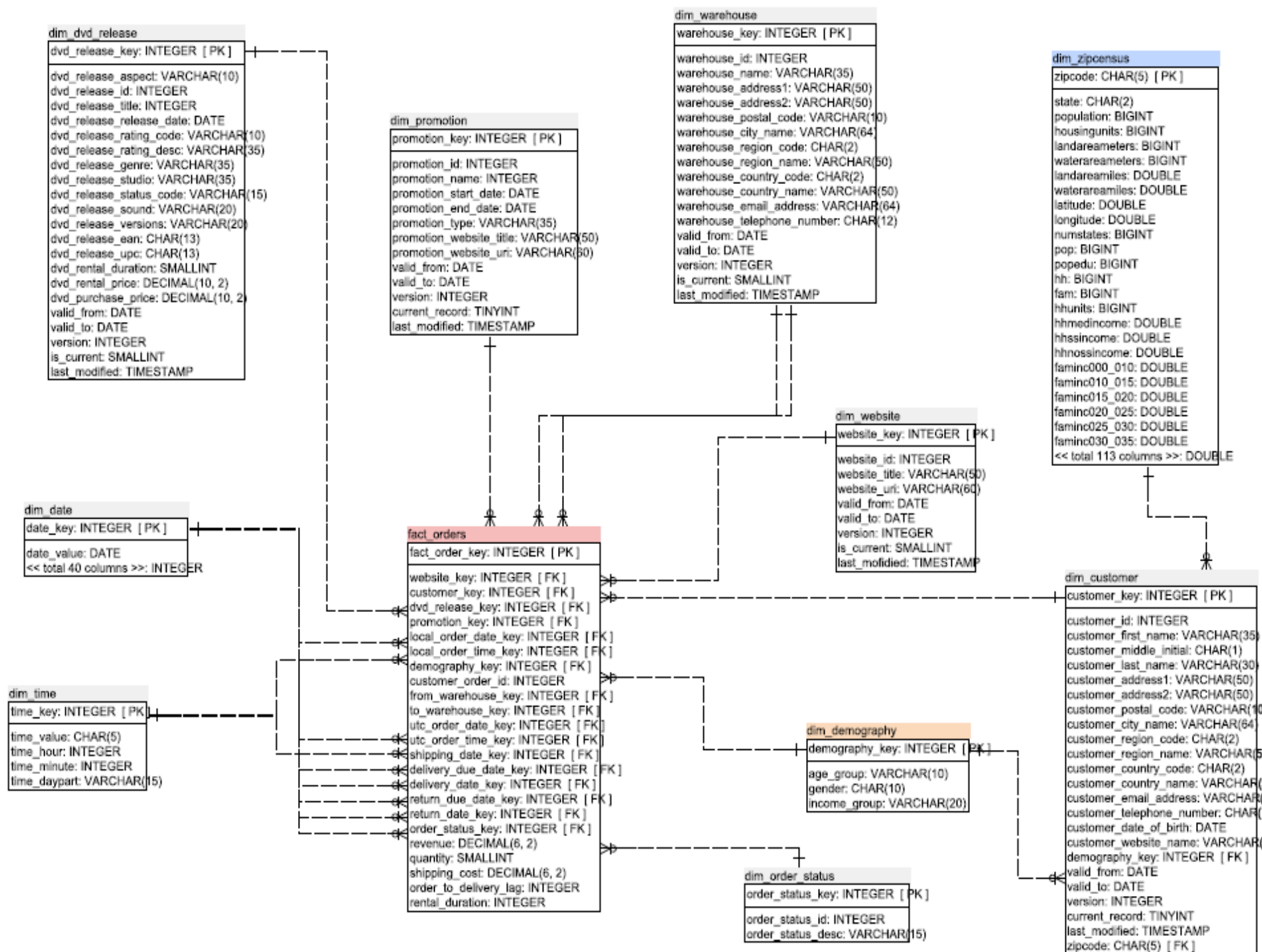
FIELD	ANALYTICAL	TYPE	LENGTH	SCD	DESCRIPTION
Customer_key	N	INT	4		Surrogate dimension key
Customer_id	N	INT	4		Original source system key
Customer_name	N	VARCHAR	63	2	Full name (first + middle + last)
Customer_city	Y	VARCHAR	64	2	Name of city
Customer_phone_number	N	CHAR	12	2	Telephone number
Customer_register_date_key	Y	INT	4	1	First registration date of customer

Projetoando o modelo

- Dimensão Time (1ª versão) - incompleta

FIELD	TYPE	LENGTH	DESCRIPTION	EXAMPLE
date_key	INT	4	Surrogate dimension key	20091123
date_value	DATE	4	Date value for the day	23-11-2009
date_julian	INT	4	Rounded Julian date	2455159
date_short	CHAR	12	Short text value for date	11/23/09
date_medium	CHAR	16	Medium text value for date	Nov 23, 2009
date_long	CHAR	24	Long text value for date	November 23, 2009
date_full	CHAR	32	Full-text value for date	Monday, November 23, 2009
day_in_week	TINYINT	1	Number of day in week	2
day_in_month	TINYINT	1	Number of day in month	23
day_in_year	SMALLINT	2	Number of day in year	327
is_first_day_in_month	TINYINT	1	1 for first day, 0 for other	0
is_first_day_in_week	TINYINT	1	1 for first day, 0 for other	0

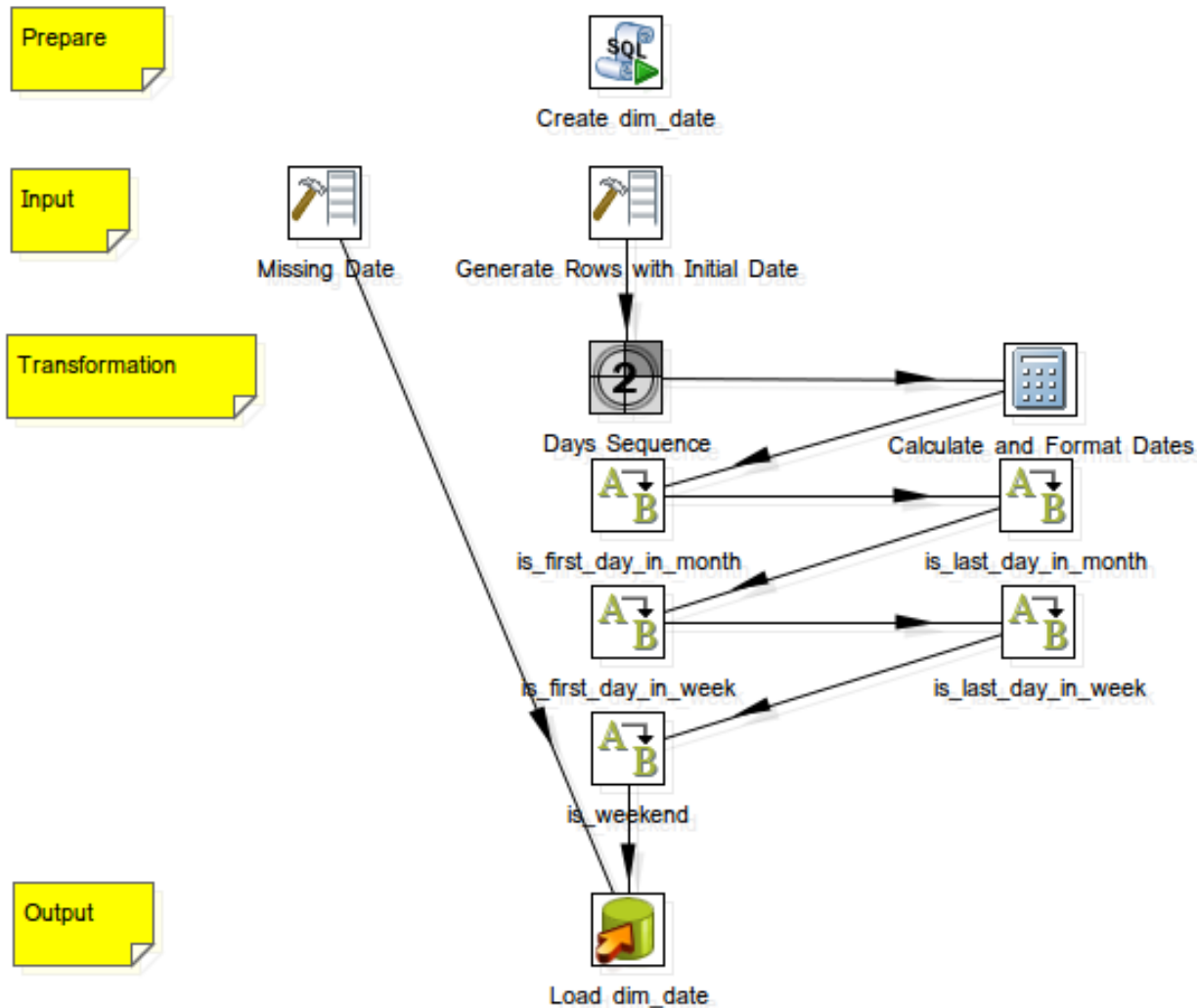




Carregando uma dimensão

- Dimensão Date
- Passos para carregar a dimensão
 - Criar a tabela de dimensão `dim_date`
 - Gerar uma linha para datas inválidas
 - Gerar uma linha por dia no calendário
 - Gerar uma sequência para os dias no calendário
 - Adicionar a sequência à data inicial e formatando as datas seguintes
 - Mapear dados numéricos ao texto gravado na tabela (`is_first_day_in_month`, `is_weekend`, etc.)
 - Carregar a tabela `dim_date`

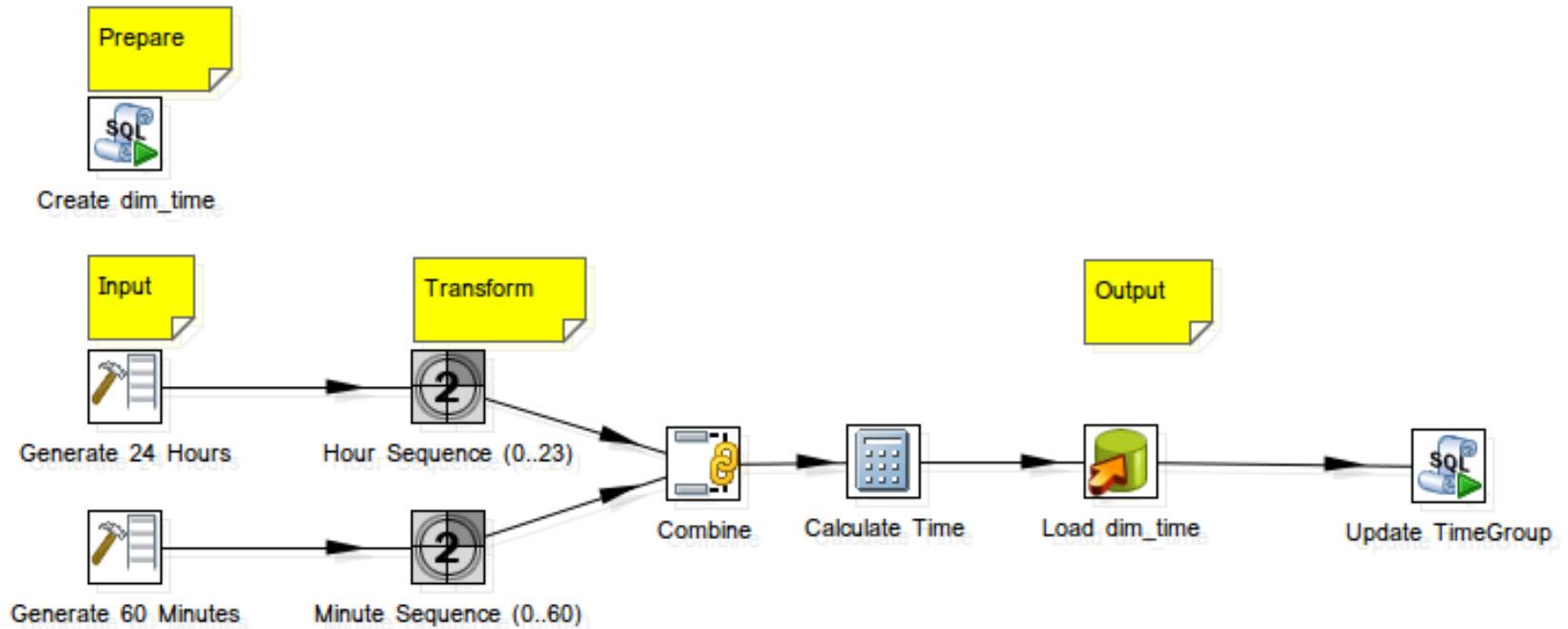
Carregando a dimensão *Date*



Carregando uma dimensão

- Dimensão Time
- Passos para carregar a dimensão
 - Criar a tabela de dimensão `dim_time`
 - Gerar as linhas das 24 horas e dos 60 minutos
 - Criar a sequência de horas e minutos
 - Fazer o produto cartesiano das linhas de horas e minutos
 - Fazer o *parsing* da hora/minuto em um campo do tipo Date
 - Carregar a tabela de dimensão `dim_time`

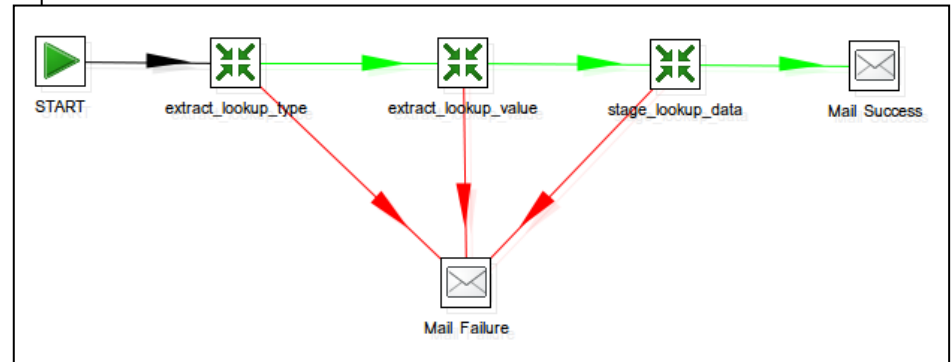
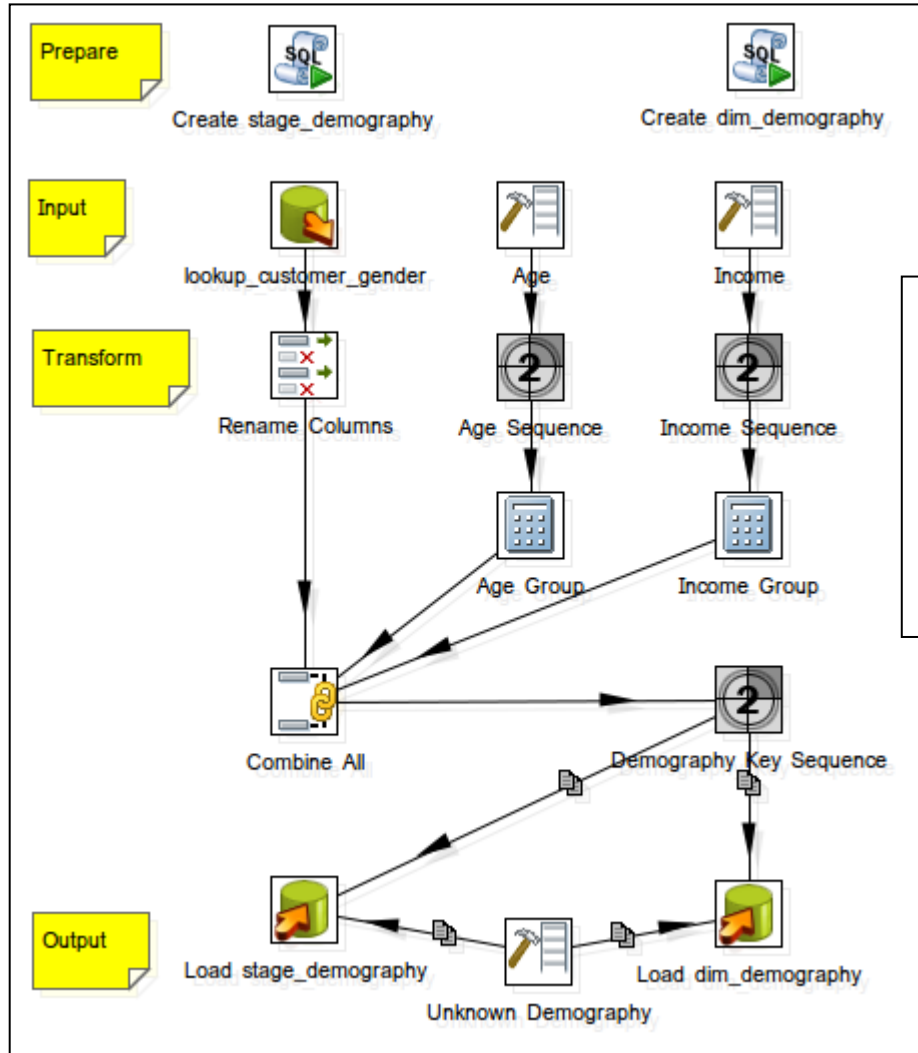
Carregando a dimensão *Time*



Carregando uma dimensão

- Dimensão Demography
- Passos para carregar a dimensão
 - Criar a tabela de dimensão `dim_demography`
 - Criar uma área de *staging* para fazer *lookup* com as tabelas banco operacional
 - Criar os dados de gênero, faixas de idade e renda mensal (inclui as *surrogates* de cada tabela)
 - Mapear os dados de gênero em Masculino e Feminino
 - Criar as faixas de idade e renda, combinando os dados através do produto cartesiano
 - Gerar as *surrogates* da tabela `dim_demography`
 - Criar uma linha para dados inválidos
 - Carregar os dados na tabela `dim_demography`

Carregando a dimensão *Demography*



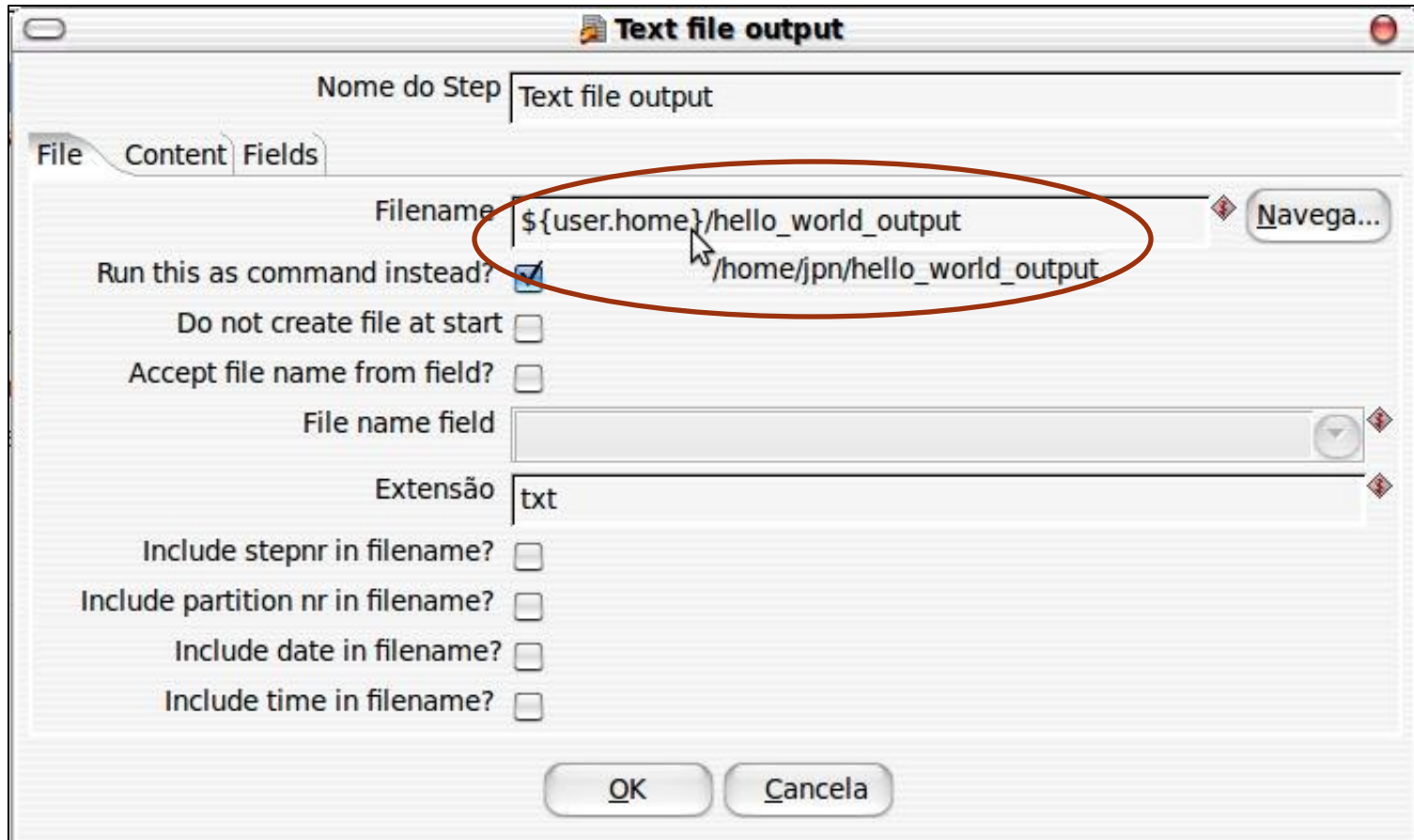


AUTOMAÇÃO DO PROCESSO DE ETL

ETL em um Ambiente de Produção

- Na prática...
 - A execução dos jobs e transformações não são feitas no Spoon
 - Requer muita **intervenção** “humana”
 - Distribuição e monitoramento do processo de ETL
 - Configuração de ambiente;
 - Coleta e análise de métricas da execução;
 - Notificação da execução.

Variáveis do Ambiente



|| Execução em Linha de Comando

Pan

- Execução de Transformações em linha de comando

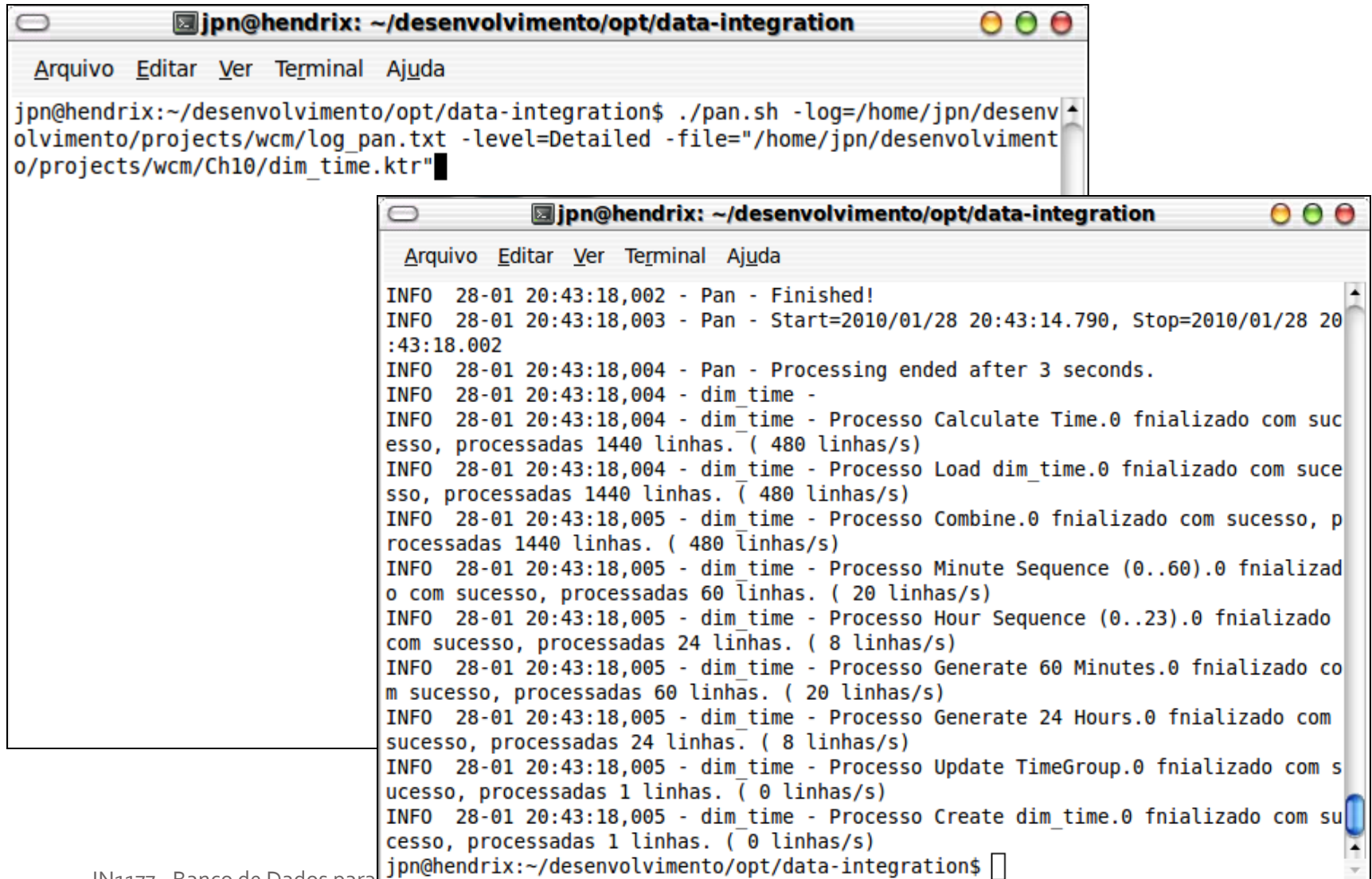
Kitchen

- Execução de Jobs em linha de comando

Carte

- Servidor para execução remota

Transformações da WCM com o Pan



```
jpn@hendrix: ~/desenvolvimento/opt/data-integration
Arquivo  Editar  Ver  Terminal  Ajuda

jpn@hendrix:~/desenvolvimento/opt/data-integration$ ./pan.sh -log=/home/jpn/desenvolvimento/projects/wcm/log_pan.txt -level=Detailed -file="/home/jpn/desenvolvimento/projects/wcm/Ch10/dim_time.ktr"
```

```
jpn@hendrix: ~/desenvolvimento/opt/data-integration
Arquivo  Editar  Ver  Terminal  Ajuda

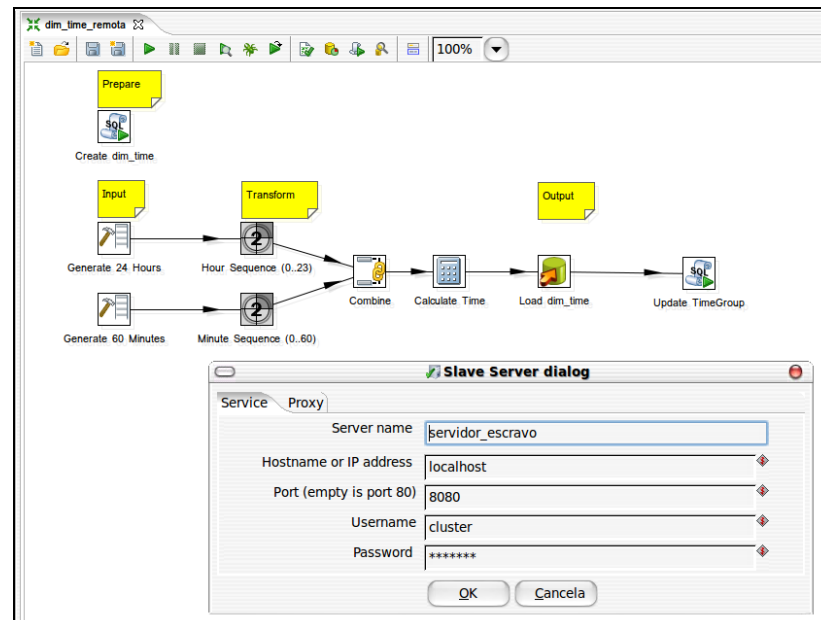
INFO 28-01 20:43:18,002 - Pan - Finished!
INFO 28-01 20:43:18,003 - Pan - Start=2010/01/28 20:43:14.790, Stop=2010/01/28 20:43:18.002
INFO 28-01 20:43:18,004 - Pan - Processing ended after 3 seconds.
INFO 28-01 20:43:18,004 - dim_time -
INFO 28-01 20:43:18,004 - dim_time - Processo Calculate Time.0 fnializado com sucesso, processadas 1440 linhas. ( 480 linhas/s)
INFO 28-01 20:43:18,004 - dim_time - Processo Load dim_time.0 fnializado com sucesso, processadas 1440 linhas. ( 480 linhas/s)
INFO 28-01 20:43:18,005 - dim_time - Processo Combine.0 fnializado com sucesso, processadas 1440 linhas. ( 480 linhas/s)
INFO 28-01 20:43:18,005 - dim_time - Processo Minute Sequence (0..60).0 fnializado com sucesso, processadas 60 linhas. ( 20 linhas/s)
INFO 28-01 20:43:18,005 - dim_time - Processo Hour Sequence (0..23).0 fnializado com sucesso, processadas 24 linhas. ( 8 linhas/s)
INFO 28-01 20:43:18,005 - dim_time - Processo Generate 60 Minutes.0 fnializado com sucesso, processadas 60 linhas. ( 20 linhas/s)
INFO 28-01 20:43:18,005 - dim_time - Processo Generate 24 Hours.0 fnializado com sucesso, processadas 24 linhas. ( 8 linhas/s)
INFO 28-01 20:43:18,005 - dim_time - Processo Update TimeGroup.0 fnializado com sucesso, processadas 1 linhas. ( 0 linhas/s)
INFO 28-01 20:43:18,005 - dim_time - Processo Create dim_time.0 fnializado com sucesso, processadas 1 linhas. ( 0 linhas/s)
jpn@hendrix:~/desenvolvimento/opt/data-integration$
```

|| Execução Remota com o Carte

- Qual a necessidade de execução remota?
 - Escalabilidade;
 - Scale-up: configuração de hardware
 - Scale-out: clustering
 - Disponibilidade;
 - Redução de tráfego e latência da rede.

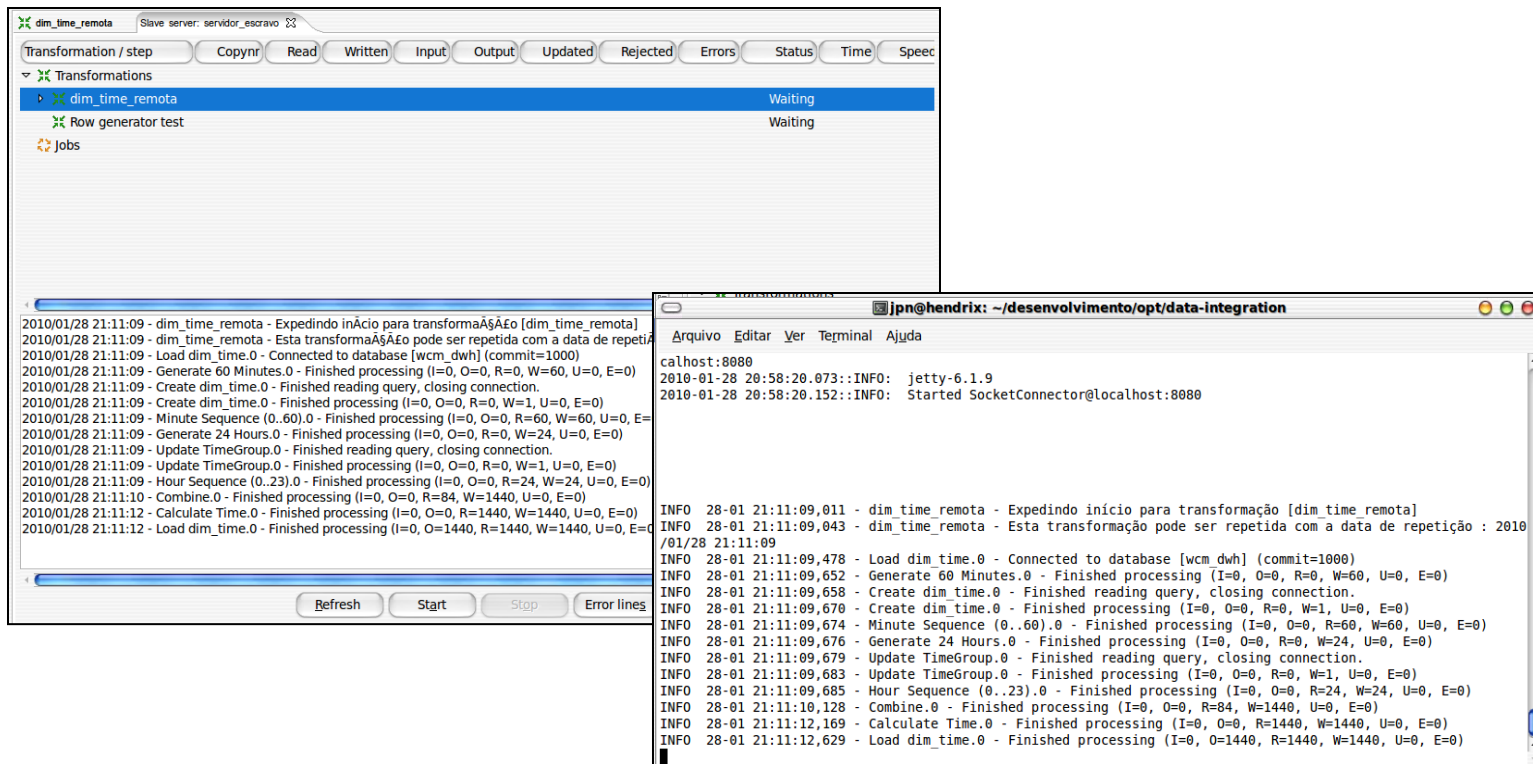
Execução Remota com o Carte

- Configuração necessária
 - Criar servidores de execução “escravos”
 - Alterar os jobs ou transformações para rodar remotamente



Execução Remota com o Carte

Exemplo: dim_time_remota



The screenshot displays the Carte interface for the 'dim_time_remota' transformation. The top panel shows the transformation's status as 'Waiting'. The bottom panel shows the log window with the following text:

```

2010/01/28 21:11:09 - dim_time_remota - Expedindo início para transformação [dim_time_remota]
2010/01/28 21:11:09 - dim_time_remota - Esta transformação pode ser repetida com a data de repetição : 2010/01/28 21:11:09
2010/01/28 21:11:09 - Load dim_time.0 - Connected to database [wcm_dwh] (commit=1000)
2010/01/28 21:11:09 - Generate 60 Minutes.0 - Finished processing (I=0, O=0, R=0, W=60, U=0, E=0)
2010/01/28 21:11:09 - Create dim_time.0 - Finished reading query, closing connection.
2010/01/28 21:11:09 - Create dim_time.0 - Finished processing (I=0, O=0, R=0, W=1, U=0, E=0)
2010/01/28 21:11:09 - Minute Sequence (0..60).0 - Finished processing (I=0, O=0, R=60, W=60, U=0, E=0)
2010/01/28 21:11:09 - Generate 24 Hours.0 - Finished processing (I=0, O=0, R=0, W=24, U=0, E=0)
2010/01/28 21:11:09 - Update TimeGroup.0 - Finished reading query, closing connection.
2010/01/28 21:11:09 - Update TimeGroup.0 - Finished processing (I=0, O=0, R=0, W=1, U=0, E=0)
2010/01/28 21:11:09 - Hour Sequence (0..23).0 - Finished processing (I=0, O=0, R=24, W=24, U=0, E=0)
2010/01/28 21:11:10 - Combine.0 - Finished processing (I=0, O=0, R=84, W=1440, U=0, E=0)
2010/01/28 21:11:12 - Calculate Time.0 - Finished processing (I=0, O=0, R=1440, W=1440, U=0, E=0)
2010/01/28 21:11:12 - Load dim_time.0 - Finished processing (I=0, O=1440, R=1440, W=1440, U=0, E=0)
  
```

The terminal window shows the following output:

```

Arquivo Editar Ver Terminal Ajuda
calhost:8080
2010-01-28 20:58:20.073::INFO: jetty-6.1.9
2010-01-28 20:58:20.152::INFO: Started SocketConnector@localhost:8080

INFO 28-01 21:11:09,011 - dim time remota - Expedindo início para transformação [dim time remota]
INFO 28-01 21:11:09,043 - dim_time_remota - Esta transformação pode ser repetida com a data de repetição : 2010/01/28 21:11:09
INFO 28-01 21:11:09,478 - Load dim time.0 - Connected to database [wcm_dwh] (commit=1000)
INFO 28-01 21:11:09,652 - Generate 60 Minutes.0 - Finished processing (I=0, O=0, R=0, W=60, U=0, E=0)
INFO 28-01 21:11:09,658 - Create dim time.0 - Finished reading query, closing connection.
INFO 28-01 21:11:09,670 - Create dim_time.0 - Finished processing (I=0, O=0, R=0, W=1, U=0, E=0)
INFO 28-01 21:11:09,674 - Minute Sequence (0..60).0 - Finished processing (I=0, O=0, R=60, W=60, U=0, E=0)
INFO 28-01 21:11:09,676 - Generate 24 Hours.0 - Finished processing (I=0, O=0, R=0, W=24, U=0, E=0)
INFO 28-01 21:11:09,679 - Update TimeGroup.0 - Finished reading query, closing connection.
INFO 28-01 21:11:09,683 - Update TimeGroup.0 - Finished processing (I=0, O=0, R=0, W=1, U=0, E=0)
INFO 28-01 21:11:09,685 - Hour Sequence (0..23).0 - Finished processing (I=0, O=0, R=24, W=24, U=0, E=0)
INFO 28-01 21:11:10,128 - Combine.0 - Finished processing (I=0, O=0, R=84, W=1440, U=0, E=0)
INFO 28-01 21:11:12,169 - Calculate Time.0 - Finished processing (I=0, O=0, R=1440, W=1440, U=0, E=0)
INFO 28-01 21:11:12,629 - Load dim_time.0 - Finished processing (I=0, O=1440, R=1440, W=1440, U=0, E=0)
  
```

Execução Remota com o Carte

Exemplo: dim_time_remota

Aplicativos Locais Sistema

Kettle transformation status - Mozilla Firefox

Arquivo Editar Exibir Histórico Favoritos Ferramentas Ajuda

http://localhost:8080/kettle/transStatus/?name=dim_time_remota

Mais visitados Getting Started Latest Headlines

Kettle transformation status

Status transformation : [dim_time_remota]

Transformation name	Status
dim_time_remota	Waiting

[Start this transformation](#)

[Prepare the execution](#)

[Cleanup this transformation](#)

Step name	CopyNr	Read	Written	Input	Output	Updated	Rejected	Errors	Active	Time	Speed	pr/in/out
Calculate Time	0	1440	1440	0	0	0	0	0	Finished	2.5	568.2	-
Load dim_time	0	1440	1440	0	1440	0	0	0	Finished	3.0	480.9	-
Combine	0	84	1440	0	0	0	0	0	Finished	0.5	2914.9	-
Minute Sequence (0..60)	0	60	60	0	0	0	0	0	Finished	0.0	1500.0	-
Hour Sequence (0..23)	0	24	24	0	0	0	0	0	Finished	0.1	480.0	-
Generate 60 Minutes	0	0	60	0	0	0	0	0	Finished	0.0	3749.9	-
Generate 24 Hours	0	0	24	0	0	0	0	0	Finished	0.0	585.3	-
Update TimeGroup	0	0	1	0	0	0	0	0	Finished	0.0	21.2	-
Create dim_time	0	0	1	0	0	0	0	0	Finished	0.0	29.4	-

[show as XML](#)

[Back to the status page](#)

[Refresh](#)



Perguntas?