

ESPECIALIZAÇÃO EM CIÊNCIA DE DADOS COM BIG DATA, BI E ANALYTICS

EXTRAÇÃO, ANÁLISE E GESTÃO DE DADOS

Alan Rezende do Amaral, Esp.

Apresentações

- Nome
- Formação
- O que faz?
- Experiência com BI?



Motivação

&

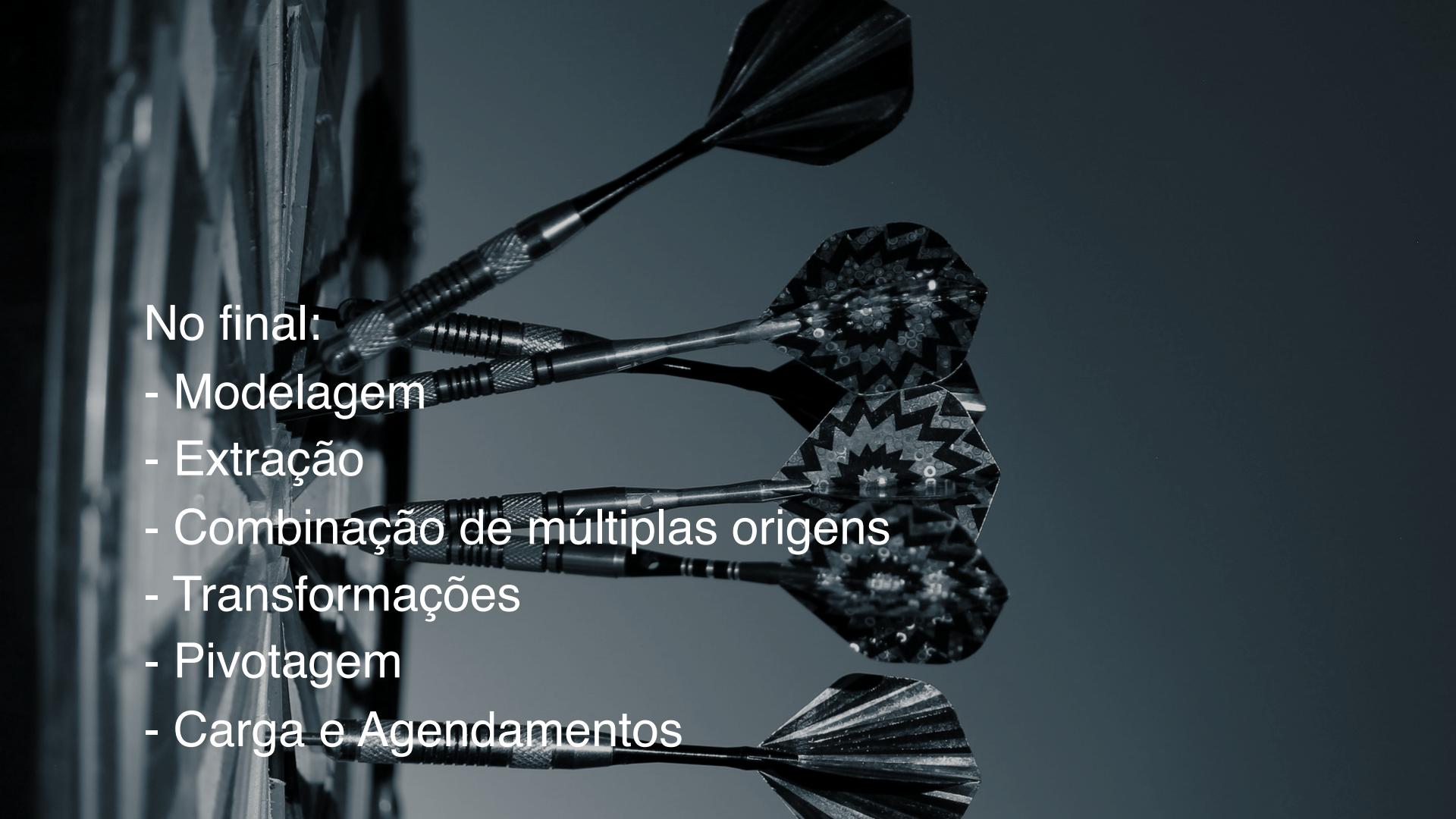
Expectativa



Objetivo

- Conceitos sobre gestão e governança de dados
- Muita prática





No final:

- Modelagem
- Extração
- Combinação de múltiplas origens
- Transformações
- Pivotagem
- Carga e Agendamentos

Gestão e Governança de Dados

Promovendo dados como ativo
de valor nas empresas

- Alinhado ao DAMA-DMBOK®
- Livro pioneiro em português



Apelo



Material

Gestão e Governança de Dados -
Promovendo dados como ativo de valor nas
empresas

- Bergson Lopes Rêgo

The Data Warehouse Toolkit

Second Edition

The Complete
Guide to
Dimensional
Modeling



Ralph Kimball
Margy Ross

Copyrighted Material

Material

The Data Warehouse Toolkit

The complete guide to dimensional modeling

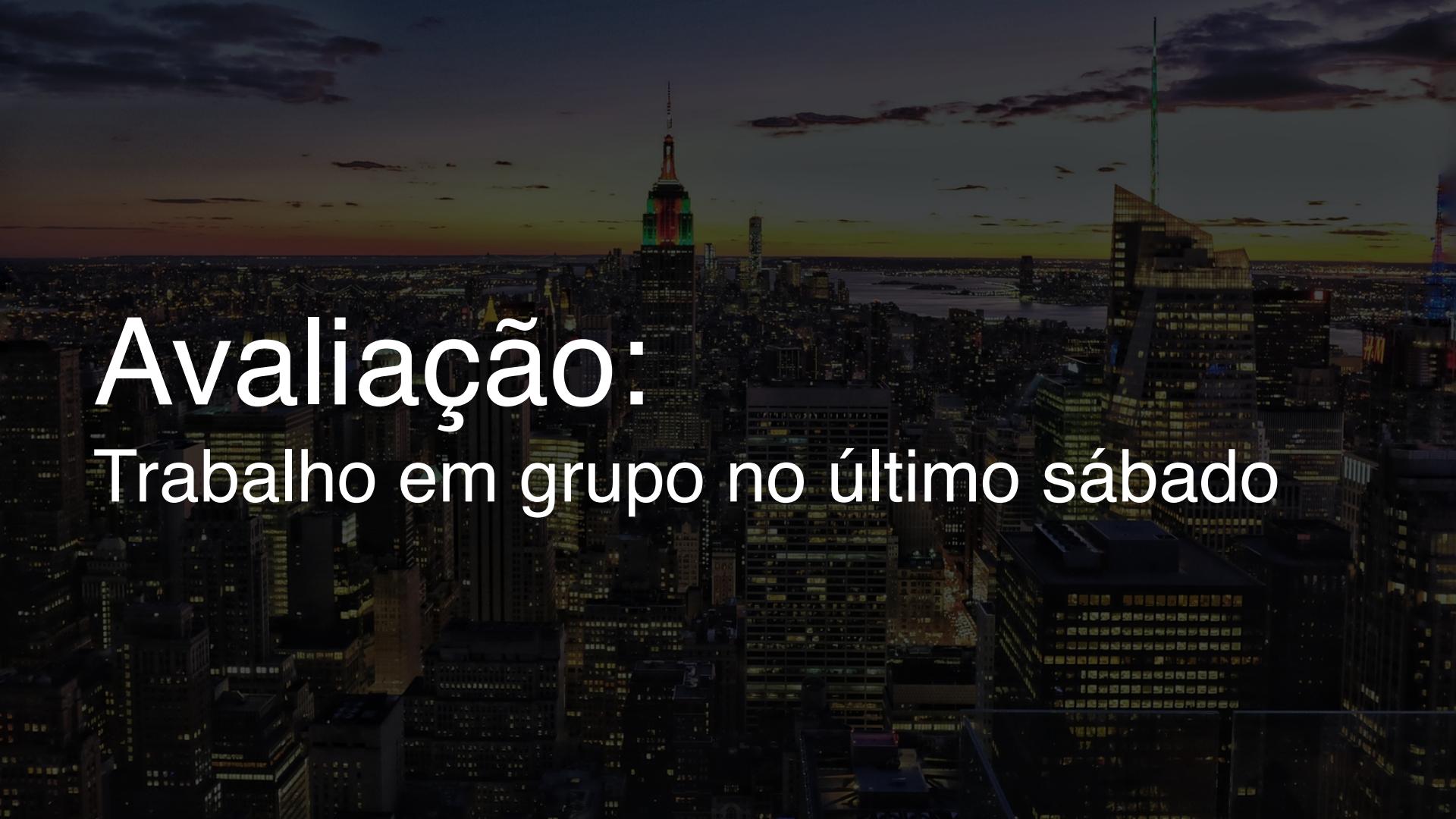
Ralph Kimball

Ferramentas

Pentaho Data Integration

PostgreSQL

Planilhas

The background image shows a panoramic view of a city skyline during sunset or sunrise. The sky is filled with warm, orange and yellow hues, with scattered clouds. In the center, the Empire State Building stands tall, its Art Deco spire reaching towards the top of the frame. To its right, the One World Trade Center is visible, its distinctive green spire reaching high into the sky. The city is densely packed with numerous skyscrapers, their windows glowing with lights from within. In the foreground, the tops of buildings and trees are visible against the bright sky.

Avaliação:
Trabalho em grupo no último sábado

Avaliação!



Perguntas:

- Conteúdo
- Avaliação à distância
- Correção e revisão

08/nov sexta 19h - 23h 10m (4h)	- Apresentação do curso, Introdução, breves conceitos sobre gerenciamento de dados. - Introdução ao Data Integration, Preparação do ambiente de desenvolvimento.
09/nov sábado 08h - 12h10m (4h)	- Laboratório 1: - Extração de dados simples
09/nov sábado 13h - 17h 10m (4h)	- Conceituar jobs X transformations - Laboratório 2: - Clean tables - CSV → Table Output - Job com as duas transformations -
22/nov sexta 19h - 23h 10m (4h)	- Laboratório 3: - Passagem de parâmetros - Iteração - Laboratório 4: - Pivotagem
23/nov sábado 08h - 12h 10m (4h)	- Boas práticas - Apresentação de cases - Tira dúvidas
23/nov sábado 13h - 17h10m (4h)	- Laboratório 5 (Trabalho):

VidaDePROGRAMADOR

.COM.BR

/* HISTÓRIA REAL
ENVIADA POR
FERNANDO ZAMBROTTA */

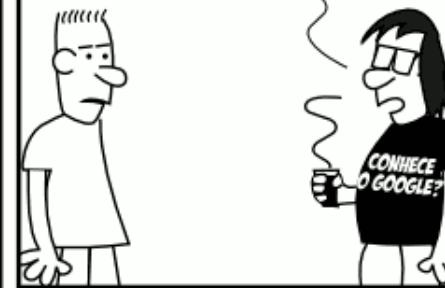


#112

CARA, EU ESTOU TENDO
MUITO TRABALHO NA EMPRESA
COM BANCO DE DADOS,
VOCÊ PODERIA
ME AJUDAR?



QUAL SISTEMA DE
BANCOS DE DADOS
VOCÊS ESTÃO
USANDO?



A GENTE USA O
EXCEL!



A photograph of a man walking away from the camera on a paved path in a park. He is carrying a red bag. The path is surrounded by green grass and trees. In the background, there are more paths and a bench. The overall atmosphere is calm and peaceful.

User experience
(or user behaviour)

Design

Dado, Informação, Conhecimento e Sabedoria



Qualidade desejada para os dados e metadados

“O simples fato de gerir e governar os dados não é suficiente para garantir o sucesso e o retorno financeiro”



Big Data

*“Decisões baseadas em emoções não são decisões”
(House of Cards)*

Nike

Parceria com empresa de TI
App para praticantes de running

Frequência de batimentos cardíacos;

Velocidade;

Quantidade de passos dados;

Distância percorrida e muitos outros

Integração com redes sociais

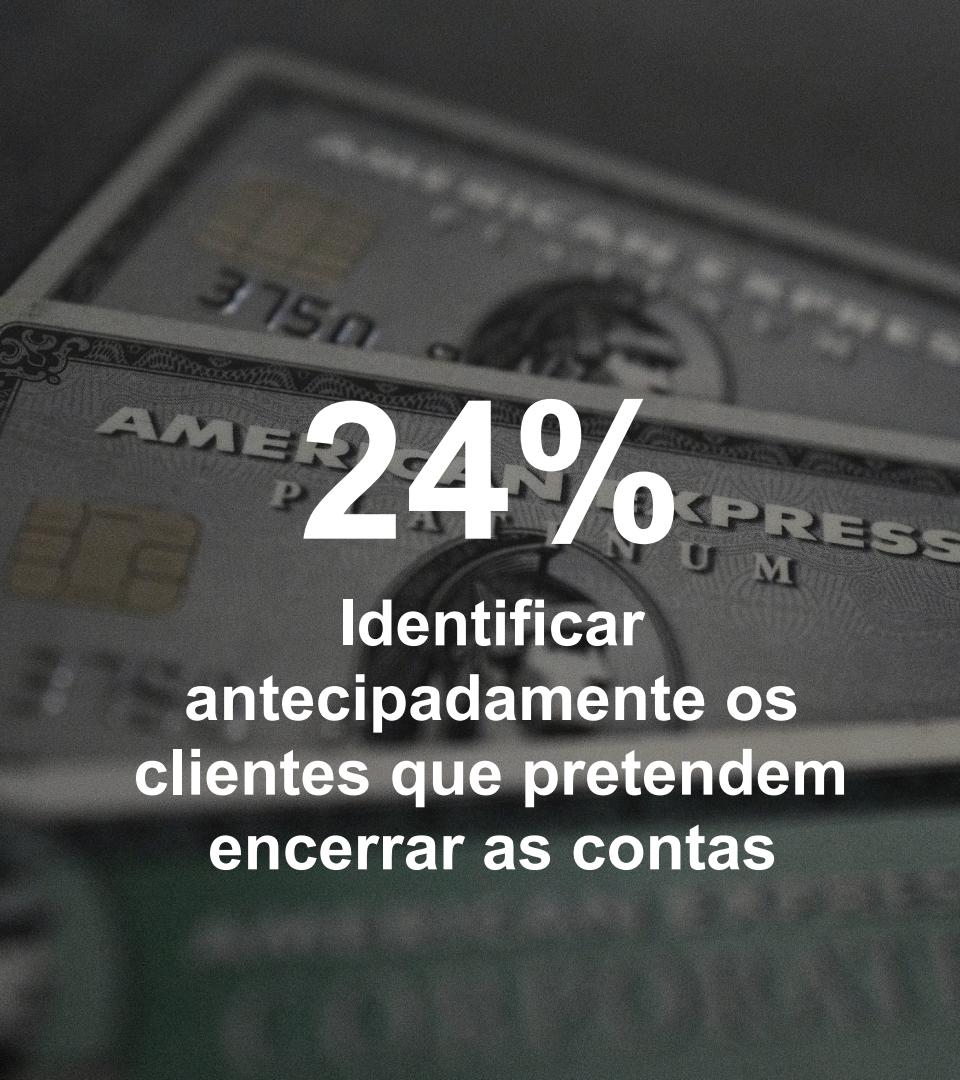
Competição entre os atletas (quem corre mais, mais rápido, melhor, etc...)



American Express

Objetivo de diminuir as taxas de cancelamento de seus clientes.

Modelos preditivos para analisar históricos de transações dos usuários de seus cartões de crédito, além de 115 variáveis, para prever potenciais churns.

A close-up, slightly blurred image of an American Express credit card. The card is dark with silver accents. The words "AMERICAN EXPRESS" are visible at the top, along with a card number starting with "3750".

24%

**Identificar
antecipadamente os
clientes que pretendem
encerrar as contas**

Os 5 V's do Big Data

Volume

Velocidade

Variedade

Veracidade

Valor

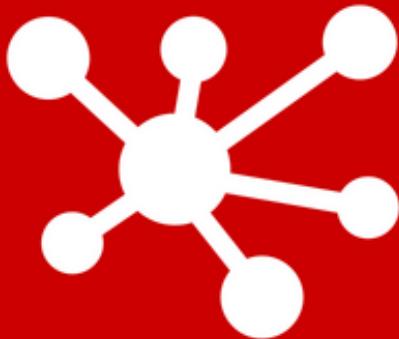
Segundo Jeff Magnusson, da Netflix, os dados devem ser:

“acessíveis, de fácil processamento, de fácil visualização e quanto mais tempo você demora para encontrá-los, menos valiosos eles se tornam”.

Preparação do Ambiente & Laboratório 1.1



Mãos à obra



Postgres

- Criar um banco de dados (UNI7)



Data Integration

- Verificar funcionamento PDI
- Extração simples
 - Arquivo CSV para uma tabela do Posgtres

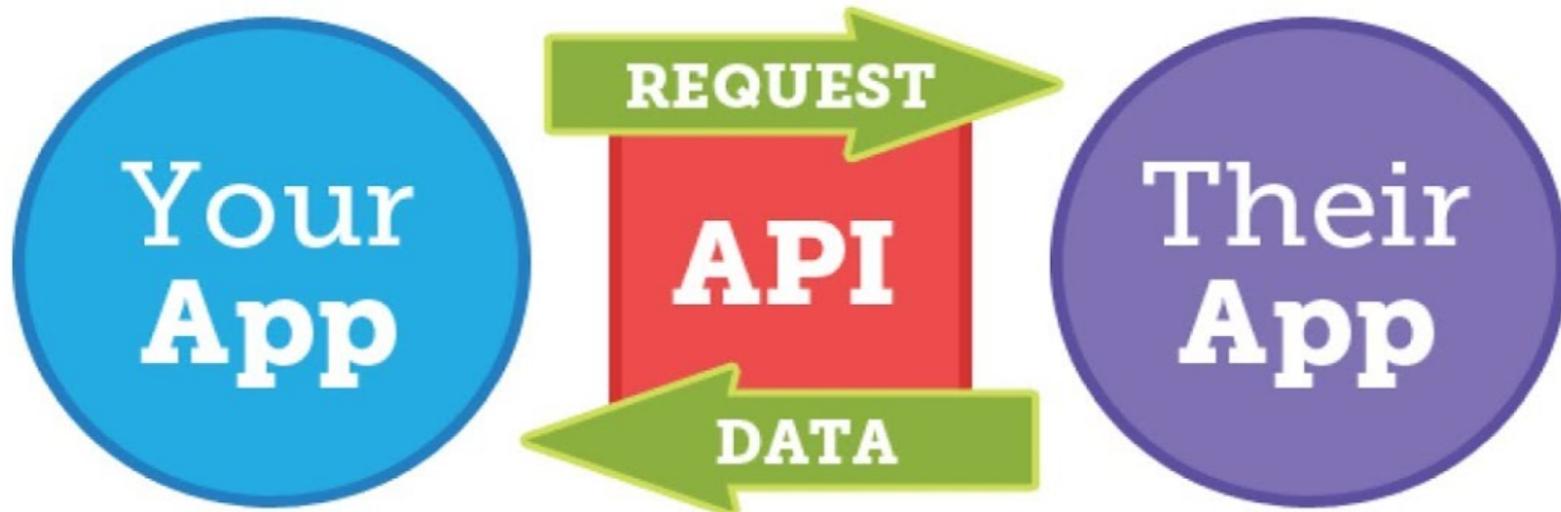
<http://alanamaral.com.br/upload>

UNI7/VI/aula1/Locacao.csv

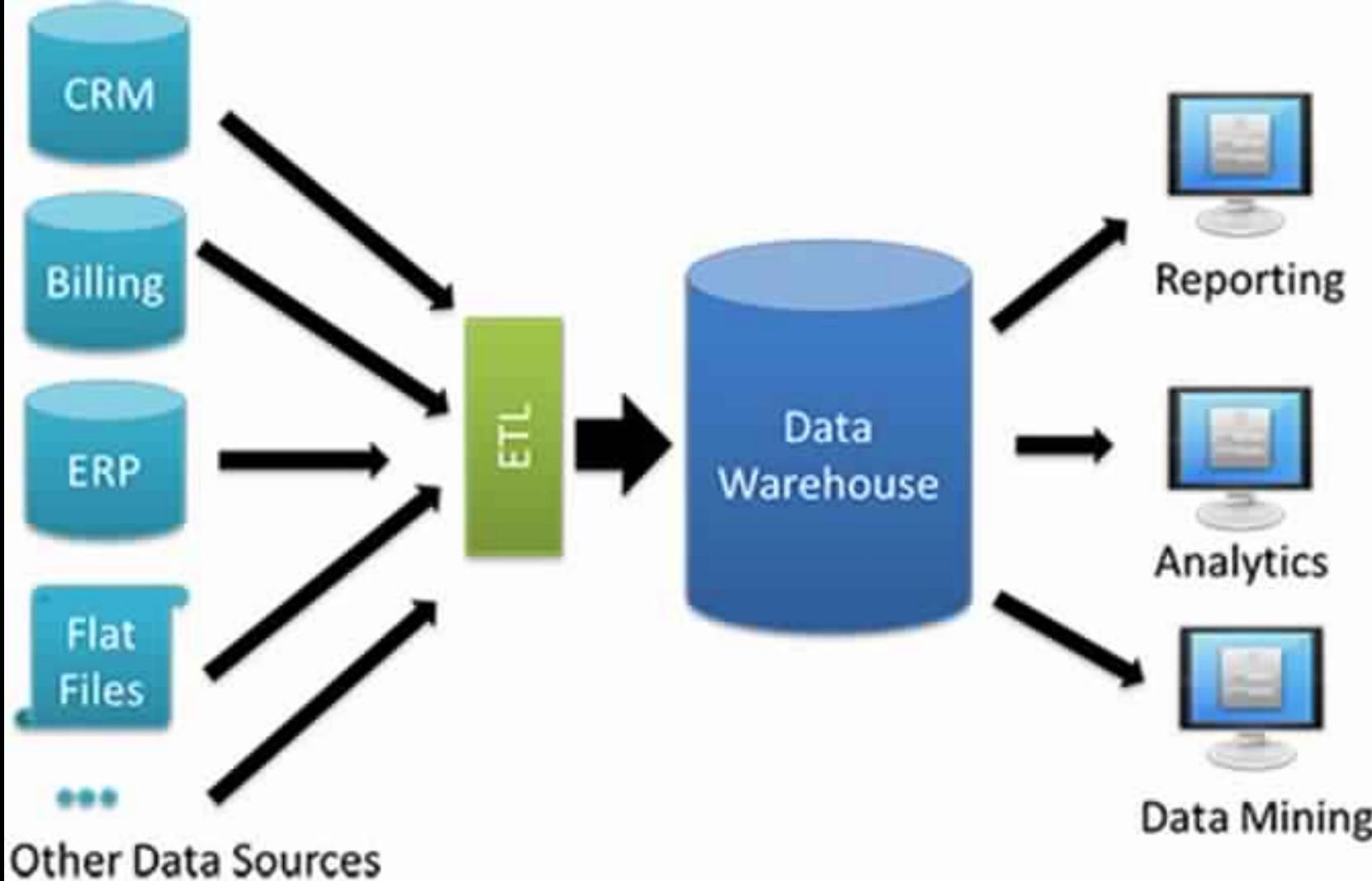
UNI7/VI/aula1/Loja.csv

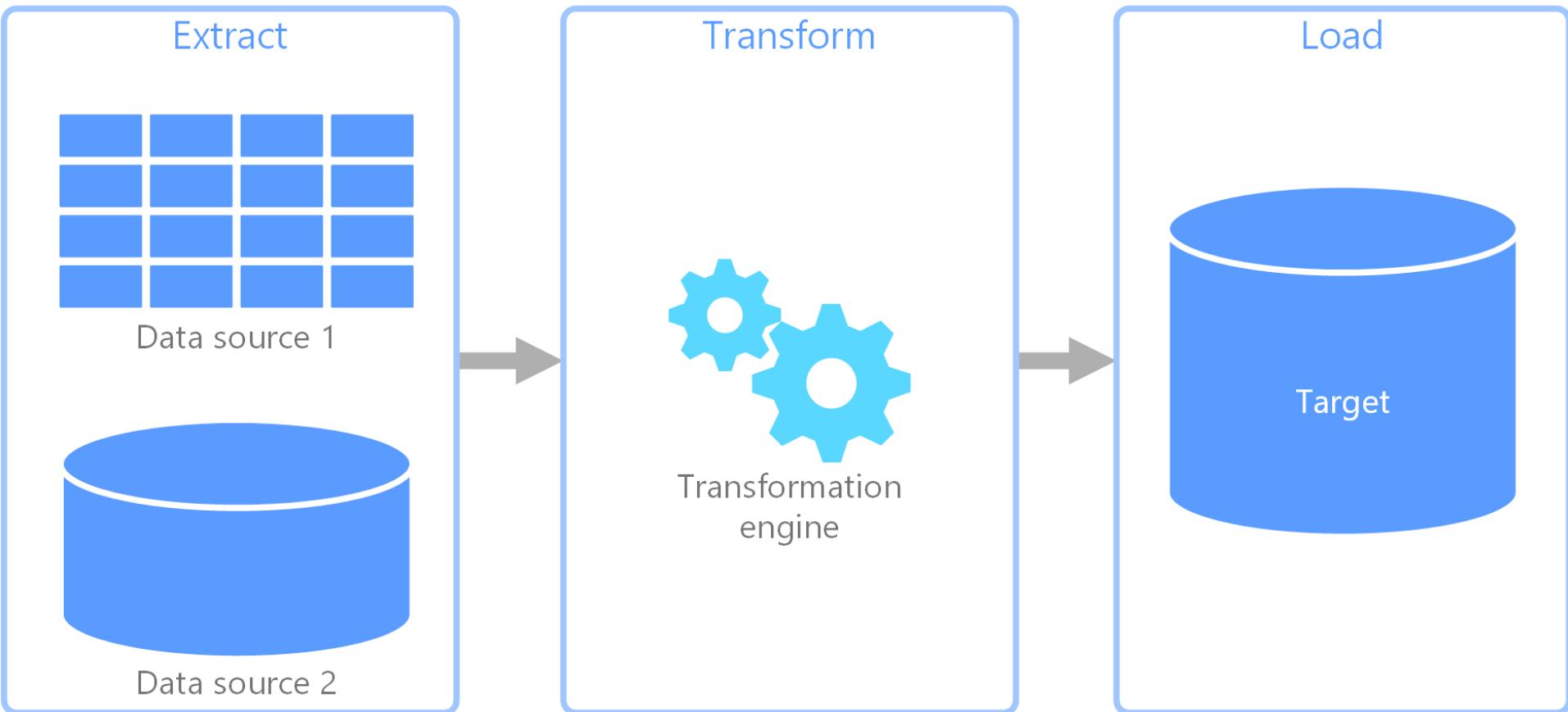
UNI7/VI/aula1/Vendedor.csv

What is an API?



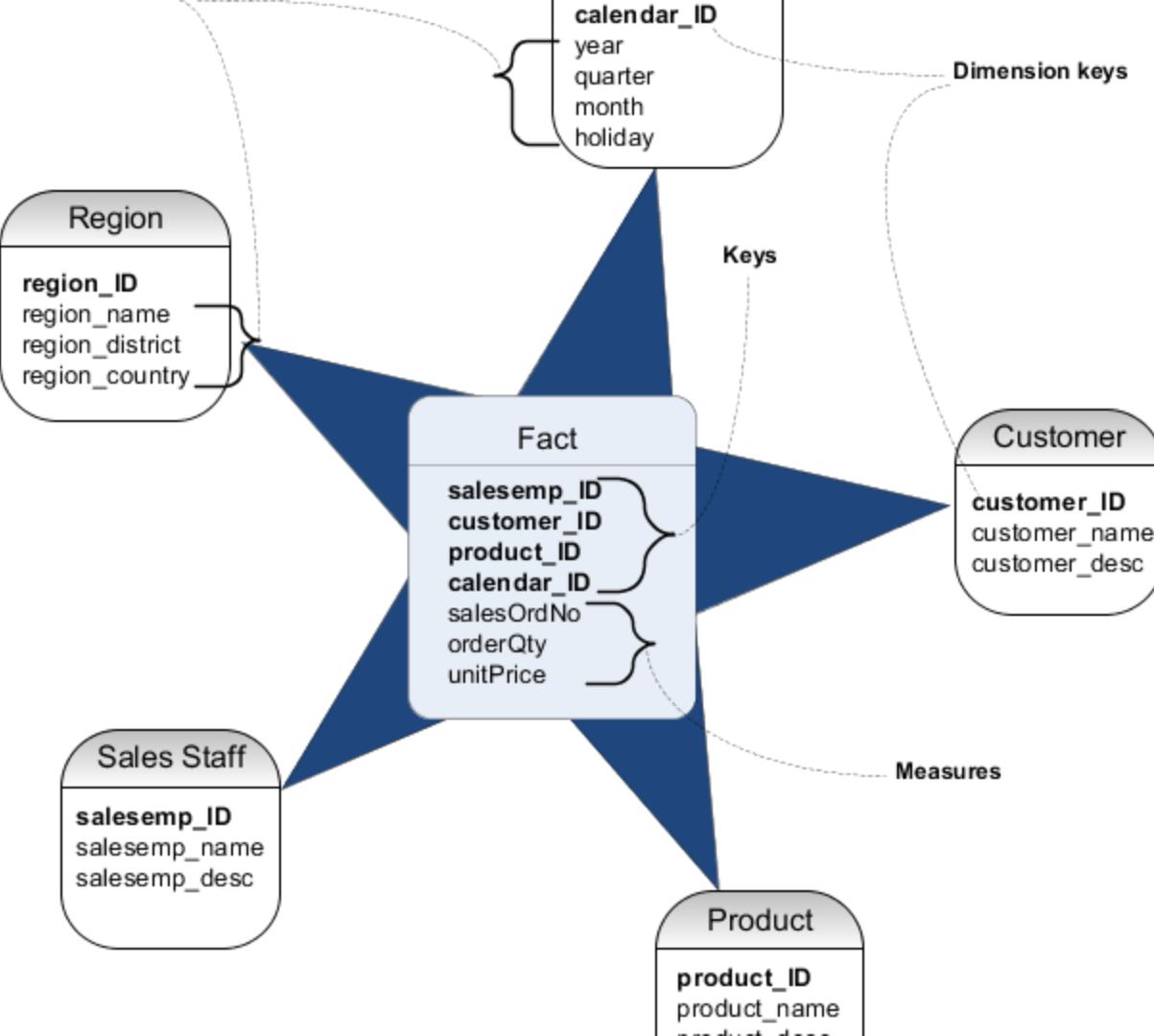
API Vs Web service





Modelagem Dimensional

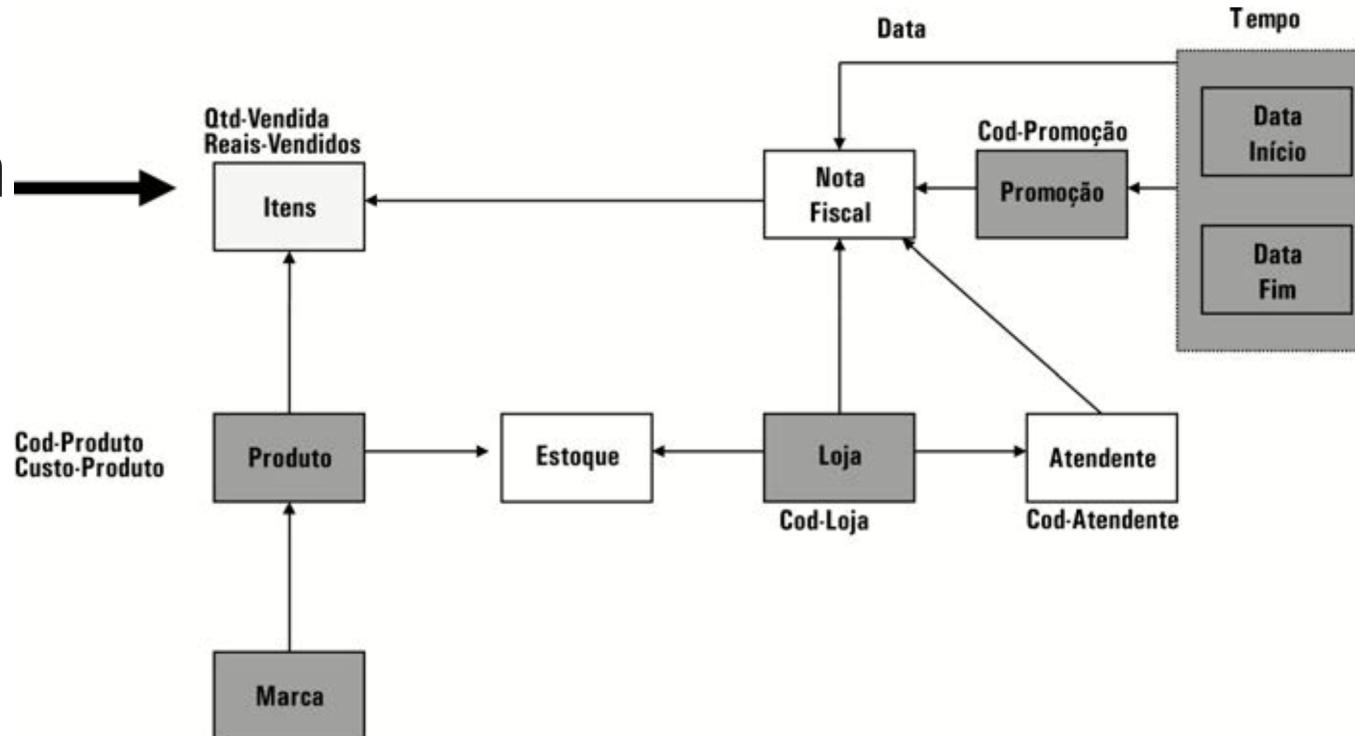
Star Schema



**Sua principal
característica é
a presença de
dados
altamente
redundantes,
melhorando o
desempenho**



Considere o
modelo de
dados de um
sistema
transacional





Agora veja como ele ficaria numa modelagem dimensional

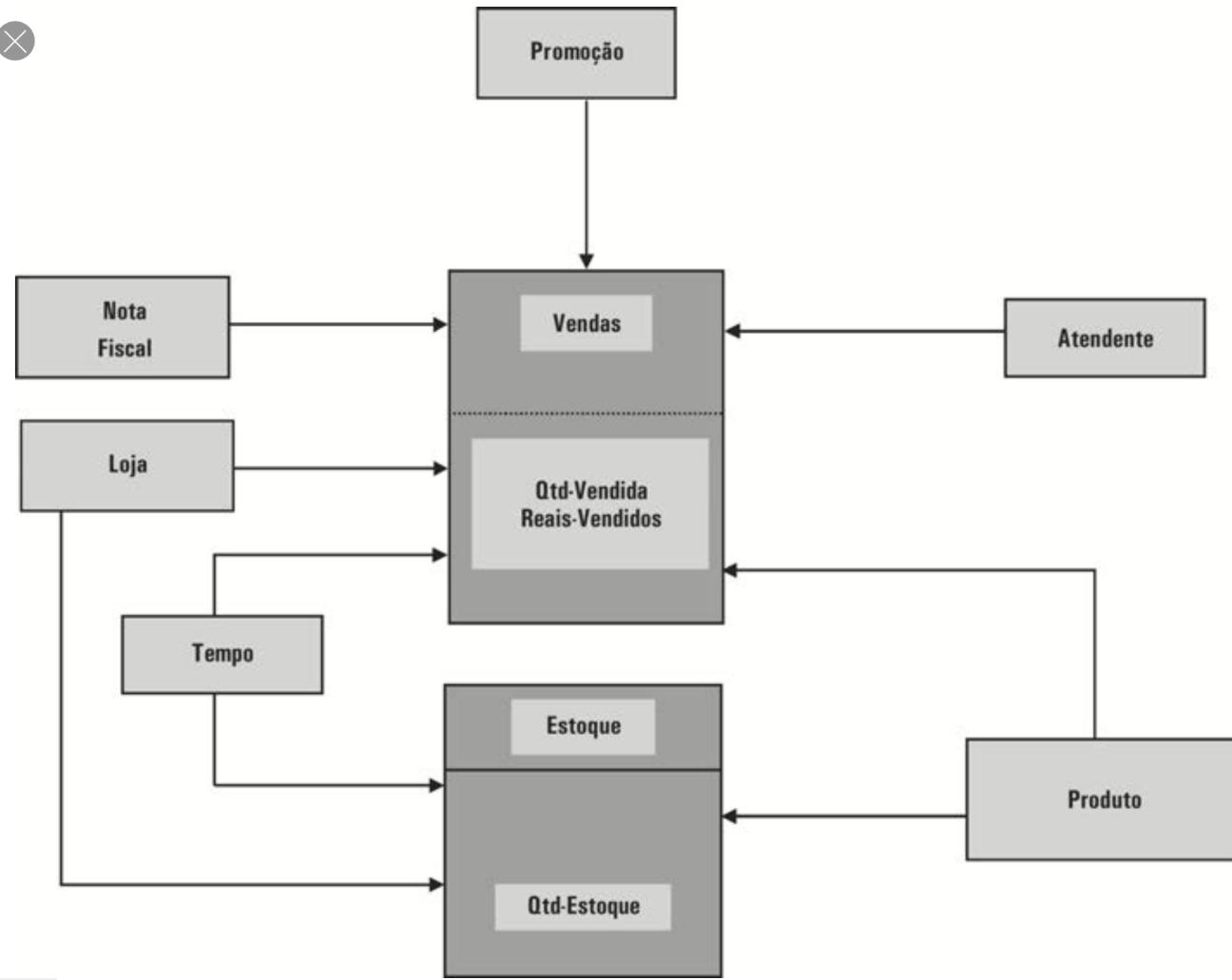


Tabela Fato

- Armazena o que ocorreu, é o fato propriamente dito
- **Essa tabela armazena 2 coisas:**
 - Os fatos ocorridos, ou seja, as métricas
 - As chaves para as dimensões
- **Por exemplo:**
 - Hoje eu tomei 5 cafés
 - Métrica: quantidade de cafés bebidos (5 nesse caso)
 - Dá para fazer uma análise com ela, por exemplo, no tempo: nos últimos 30 dias, que dias eu tomei mais café?

Tabela Dimensão

-
- **Qualificador**
 - “Elas vão qualificar, classificar ou descrever os dados que estão nas fatos, ou seja, as métricas”
 - **Por exemplo:**
 - Preciso medir as vendas.
 - Ótimo, mas pelo quê?
 - Pela empresa / pelo produto / pelas filiais / por dia
-



Hierarquia e grão

Hierarquia é o conjunto de atributos que possui uma ordem lógica do maior ao menor nível

O grão, também chamado de detalhe, é o menor nível da hierarquia da dimensão. É a informação base, o menor detalhe da informação.



Informação importantes

- Multi-plataforma sobre Java (linux, windows, mac os)
- Spoon (PDI Client): Interface de desenvolvimento intuitiva (arraste e solte)
- O Spoon gera os arquivos (scripts) do projeto (.kjb e .ktr)
- O projeto (Jobs e Transformations) precisam estar salvos antes da execução
- Kitchen e Pan executam os scripts (Jobs: .kjb e Transformations: .ktr)



Organização e repositório do projeto

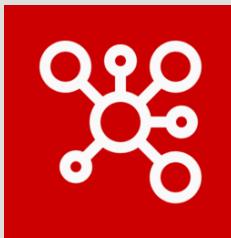
- Jobs são salvos com extensão .kjb
- Transformations são salvas com extensão .ktr
- Jobs e Transformations são scripts, salvos em arquivos XML
- Isso permite organizar o projeto em uma estrutura de diretórios
- É possível pesquisar padrões pois os arquivos são texto (XML)



Boas práticas:

- Variáveis para definir caminhos de arquivos, Jobs e Transformations
- Controle, registro e notificação de erros
- Estrutura de transação pode prevenir um desastre irreparável
- Sequenciar atividades que facilitem debug/teste
- Organizar a distribuição e conexão (hop) entre os steps
- Usar notas para registrar revisões, pendencias, etc...

Atalhos úteis Spoon



- Sugestão de variáveis (Windows: Ctrl+Space)
- Abrir step Job/Transformation:
Duplo clique com botão do meio do mouse (scroll/rodinha)
- Abrir step Job/Transformation:
Clique com botão direito do mouse > Open Referenced Object
- Adicionar hop (conexão) entre dois steps:
Shift + Clique no step origem + Clique no step destino
- Adicionar hop (conexão) entre dois steps:
Com o botão do meio do mouse, clicar na origem e arrastar até o destino

Uma mesma tarefa pode ser desenvolvida de forma diferente e alcançar o mesmo resultado

\$Art = ""; echo "hello"; class Art {
SELECT title [^o-ga-
alert("nice"); CODE IS POETRY Trans(
console.log("hi")); function #art {flo

Laboratório 1.2



Imagine o seguinte problema

- Queremos acompanhar o valor (e a quantidade) locado
 - Por mês/ano
 - De cada loja
 - De cada vendedor
 - Categoria



A wooden easel stands in a field, holding a blank, light blue rectangular canvas. The background features a calm lake, a dense forest along the shore, and a sky filled with soft, scattered clouds.

Dia 3

Detalhamento da Agenda

<p>Dia 3: Jobs X Transformations (conceitos) Laboratório</p>	<ul style="list-style-type: none">- Conceituar jobs X transformations- Laboratório 2:<ul style="list-style-type: none">- Clean tables- CSV → Table Output- Job com as duas transformations- Laboratório 2.1:<ul style="list-style-type: none">- Passagem de parâmetros- Iteração
<p>Dia 4: Laboratório</p>	<ul style="list-style-type: none">- Laboratório 3:<ul style="list-style-type: none">- Download- Job Completo- Laboratório 3.1:<ul style="list-style-type: none">- Pivotagem

O que nós já vimos até aqui?

- Alguns conceitos
- ETL
 - Input
 - Output
 - Merge
 - Sort
 - Group by
 - Calculator

Jobs X Transformations

What's the difference between transformations and jobs?

Q: In Spoon I can make jobs and transformations, what's the difference between them?

A: Transformations are about **moving and transforming rows from source to target**. Jobs are more about **high level flow control**: executing transformations, sending mails on failure, transferring files via FTP, ...

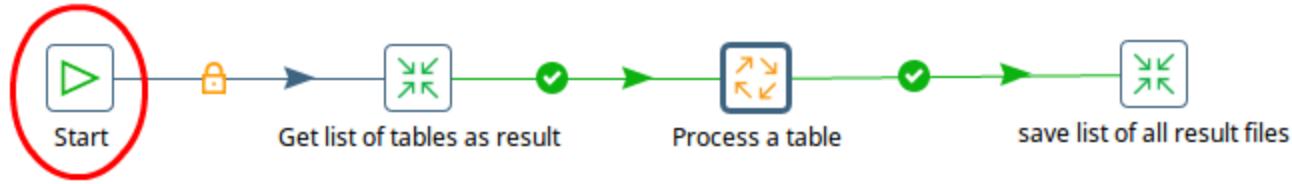


Another key difference is that all the steps in a transformation **execute in parallel**, but the steps in a job **execute in order**.

Transformations:
é onde a mágica acontece

Jobs:
sequenciamento de transformations

Jobs



Obrigatório

Laboratório

Imagine o cenário:

- Baixar orçamento de despesas do portal da transparência e carregar em uma tabela para o BI (OrcamentoDespesa)

<http://portaldatransparencia.gov.br/download-de-dados/orcamento-despesa/2018>

3 Laboratórios

Dia 3

1) Laboratório 2

- a) Preparação das tabelas
- b) Carga de um arquivo CSV para tabela

2) Laboratório 2.1

- a) Passagem de parâmetros
 - b) Iteração em lista
-

Dia 4

3) Laboratório 3

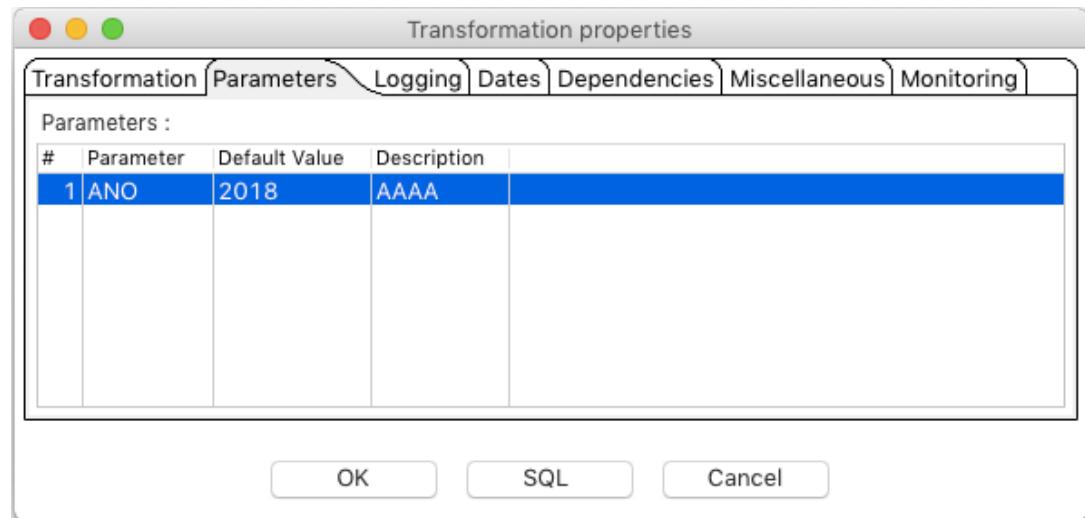
- a) Download do arquivo
- b) Job completo

Laboratório 2

- Clean tables
 - Limpar as tabelas de output para evitar registros duplicados
 - Componentes úteis:
 - Script
 - Delete
 - Outros
- CSV → Table Output
 - Fazer carga de uma tabela de DW a partir de um arquivo CSV
 - Componentes úteis:
 - CSV File Input
 - *String Operations**
 - *Replace in String**
 - Table Output
 - Add Constants
- Job

Laboratório 2.1

- Vamos modificar o nosso exercício do laboratório 2 para que o ano de referência do exercício seja passados por parâmetro
 - \${ANO}
- Iterar em uma lista de ANOs e passar isso como parâmetro



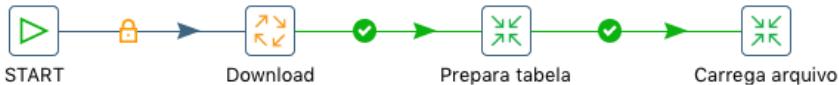
A wide-angle photograph of a tropical beach. The foreground shows light-colored, sandy beach. The middle ground is filled with the ocean, showing small, white-capped waves breaking near the shore. The water has a vibrant turquoise color. The background consists of a bright blue sky with scattered, wispy white clouds.

Dia 4

Laboratório 3

Job Completo

- Download do arquivo
- Prepara tabela
- Carrega arquivo



Dia 5

O que nós já vimos até aqui?

- Alguns conceitos
- ETL
 - Input
 - Output
 - Merge
 - Sort
 - Group by
 - Calculator
- Jobs
 - Passagem de Parâmetros
 - Iteração em Lista

Detalhamento da Agenda

Dia 5: Laboratório Cases	- Laboratório 5: - Pivotagem - Agendamentos
Dia 6: Trabalho	- Trabalho

Laboratório

- Pivotagem
 - Normalização
 - Desnormalização

Código	Vendedor	1	2	3	4	5	6	7	8	9	10	11	12
METAS	2017	25.740.000	30.205.000	29.000.000	25.960.000	24.650.000	27.750.000	27.175.000	26.700.000	27.045.000	26.720.000	28.120.000	27.715.000
6946	Shaun Seamons	75.000	150.000	380.000	430.000	290.000	285.000	420.000	125.000	60.000	105.000	475.000	70.000
7986	Omar Castelletti	180.000	330.000	105.000	185.000	180.000	330.000	190.000	130.000	425.000	385.000	255.000	245.000
6704	Mariann Laurant	100.000	485.000	405.000	335.000	135.000	335.000	455.000	420.000	280.000	120.000	245.000	50.000
1628	Abba Boeck	125.000	355.000	270.000	105.000	390.000	235.000	250.000	335.000	355.000	305.000	80.000	50.000
3728	Washington Macartney	175.000	475.000	385.000	80.000	450.000	450.000	345.000	295.000	295.000	190.000	340.000	490.000
6211	Kellia Champair	50.000	195.000	440.000	500.000	360.000	335.000	325.000	385.000	210.000	320.000	440.000	260.000
399	Cesar Pottie	175.000	470.000	370.000	415.000	395.000	85.000	490.000	250.000	195.000	250.000	190.000	455.000
9382	Linea Hallows	120.000	415.000	125.000	125.000	355.000	130.000	240.000	75.000	355.000	485.000	235.000	145.000
57	Shelia Rymill	365.000	130.000	355.000	360.000	205.000	290.000	130.000	380.000	405.000	220.000	445.000	195.000
2219	Tobit Bellwood	430.000	485.000	270.000	300.000	370.000	370.000	200.000	285.000	330.000	280.000	440.000	90.000
7581	Rip Barehead	455.000	320.000	275.000	190.000	215.000	345.000	205.000	210.000	360.000	170.000	345.000	175.000
4722	Andeee Bravington	375.000	450.000	115.000	85.000	440.000	130.000	350.000	95.000	165.000	95.000	225.000	250.000
1773	Delmore Biernacki	490.000	205.000	475.000	180.000	50.000	445.000	425.000	135.000	140.000	140.000	330.000	480.000
8046	Reynolds Kniveton	475.000	380.000	475.000	120.000	345.000	290.000	65.000	460.000	330.000	455.000	480.000	120.000
1481	Patric Gibbie	75.000	460.000	325.000	215.000	100.000	410.000	360.000	485.000	120.000	110.000	405.000	360.000
9348	Wynny Leader	200.000	295.000	225.000	125.000	445.000	160.000	500.000	330.000	355.000	495.000	445.000	480.000
5311	Isabella Stonary	135.000	300.000	70.000	260.000	75.000	165.000	315.000	215.000	145.000	195.000	420.000	255.000
471	Fae Tremeer	370.000	500.000	485.000	90.000	400.000	365.000	500.000	310.000	285.000	240.000	75.000	55.000
5070	Thoma Kuhle	360.000	145.000	475.000	290.000	50.000	340.000	120.000	295.000	345.000	115.000	80.000	140.000
255	Bren Bartaloni	125.000	75.000	425.000	230.000	420.000	120.000	310.000	165.000	410.000	450.000	80.000	85.000
4976	Frans Farncomb	115.000	85.000	455.000	175.000	65.000	255.000	445.000	485.000	230.000	260.000	135.000	455.000
5506	Quinta Breslau	320.000	185.000	470.000	140.000	295.000	295.000	90.000	65.000	470.000	485.000	60.000	440.000
1654	Anallise Veld	305.000	480.000	130.000	250.000	450.000	325.000	230.000	185.000	215.000	140.000	455.000	365.000
8850	Mervin Beazleigh	105.000	55.000	215.000	270.000	330.000	345.000	355.000	90.000	285.000	50.000	165.000	255.000
7945	Horten Rubinlicht	105.000	185.000	205.000	410.000	345.000	95.000	385.000	270.000	280.000	445.000	210.000	330.000
4942	Ellerey Ing	195.000	360.000	155.000	270.000	60.000	75.000	280.000	395.000	315.000	300.000	235.000	165.000
951	Redd Tocher	470.000	335.000	460.000	70.000	90.000	210.000	490.000	205.000	220.000	495.000	140.000	225.000
1277	Janene Yo	170.000	345.000	150.000	215.000	245.000	485.000	450.000	225.000	205.000	500.000	470.000	300.000
5390	Nikolos Tomkin	155.000	455.000	195.000	105.000	445.000	475.000	340.000	425.000	500.000	230.000	180.000	165.000

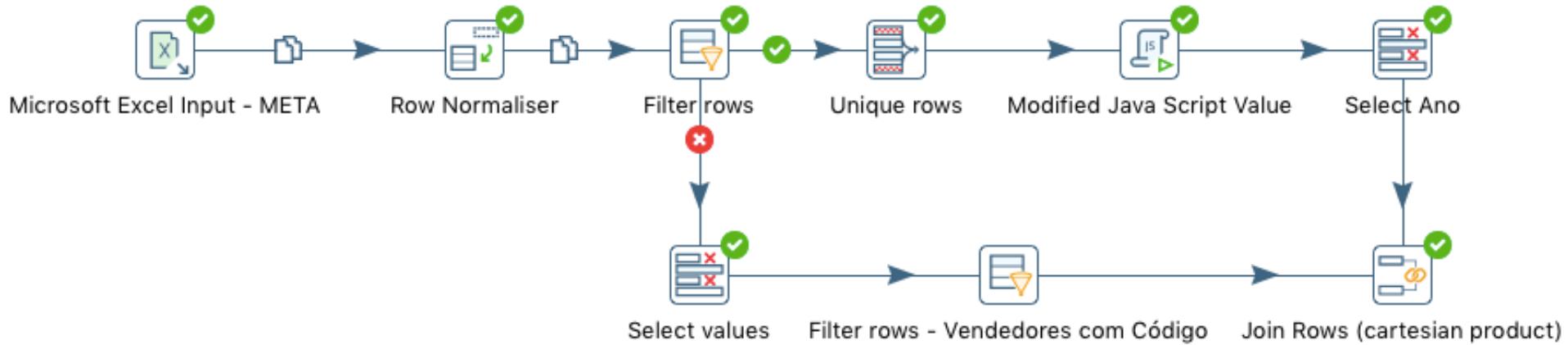
Tabela META

```
CREATE TABLE meta AS (
    vendedorcodigo TEXT,
    vendedornome TEXT,
    ano SMALLINT,
    mes SMALLINT,
    valor DOUBLE PRECISION
)
```

Exemplo do dado normalizado para um vendedor

Código	Vendedor	Ano	Mês	Valor
6946	Shaun Seamons	2.017	1	75.000
6946	Shaun Seamons	2.017	2	150.000
6946	Shaun Seamons	2.017	3	380.000
6946	Shaun Seamons	2.017	4	430.000
6946	Shaun Seamons	2.017	5	290.000
6946	Shaun Seamons	2.017	6	285.000
6946	Shaun Seamons	2.017	7	420.000
6946	Shaun Seamons	2.017	8	125.000
6946	Shaun Seamons	2.017	9	60.000
6946	Shaun Seamons	2.017	10	105.000
6946	Shaun Seamons	2.017	11	475.000
6946	Shaun Seamons	2.017	12	70.000

Sugestão de transformação



Row Normaliser

Step name

Row Normaliser

Type field

Mes

Fields

#	Fieldname	Type	new field
1	Janeiro	1	Valor
2	Fevereiro	2	Valor
3	Março	3	Valor
4	Abril	4	Valor
5	Maio	5	Valor
6	Junho	6	Valor
7	Julho	7	Valor
8	Agosto	8	Valor
9	Setembro	9	Valor
10	Outubro	10	Valor
11	Novembro	11	Valor
12	Dezembro	12	Valor

Type field: Nome da nova coluna que será normalizada

Fieldname: Nome da coluna recebida

Type: Valor da nova coluna (Type field)

new field: Nome do novo campo (recebe o valor do Fieldname)

Help

OK

Cancel

Get Fields

Agendamentos

- Hora programada
- Condicional



Kitchen.sh

Kitchen.bat

- Agendador de tarefas do Windows
- Crontab no Linux

Job.bat (.sh)

```
set JAVA_HOME=C:\Program Files\Java\jdk1.8.0_144
```

```
<PATH>\kitchen.bat /file:<PATH>\MyJob.kjb /level:Debugging > log.log
```

Exemplo (linux):

```
/opt/data-integration/kitchen.sh -level=Minimal -file /opt/vacuumFull.kjb > /opt/vacuumFull.log
```

Trabalho