

FIAP GRADUAÇÃO

**DISCIPLINA: PROJETO DE SISTEMAS APLICADO AS MELHORES PRÁTICAS EM
QUALIDADE DE SOFTWARE E GOVERNANÇA DE TI**

AULA:

23 – DATA QUALITY

PROFESSOR:

RENATO JARDIM PARDUCCI

PROFRENATO.PARDUCCI@FIAP.COM.BR

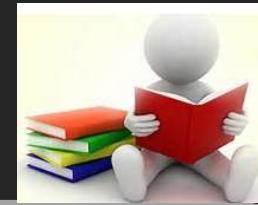
Renato Parducci - YouTube

AGENDA DA AULA

- ✓ CMMi nível 3 - VER/VAL
- ✓ MPS.br nível D - VER/VAL
- ✓ Técnicas e ferramentas para avaliar a qualidade dos dados

**Técnicas para avaliar a qualidade
dos dados de um software**

ESTUDO DE CASO SIMULADO



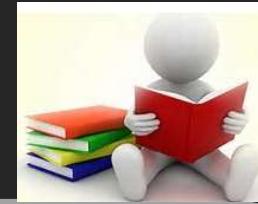
Em vários projetos que se seguiram na empresa de Dilan, foi possível notar a economia em projetos e o aumento da satisfação de clientes e da equipe da GD com a adoção das novas práticas de qualidade; mas Dilan percebeu uma lacuna:

Em projetos onde é necessário integrar sistemas legados (antigos) do cliente com os novos softwares que a sua software house desenvolveu, surgem muitas discussões sobre a qualidade final do projeto em função de surgirem dados incoerentes, inconsistentes e redundantes nas bases de dados.

Isso acontece especialmente em projetos de Business Intelligence e Data Science, que demandam cargas de dados dos sistemas legados para dentro dos novos sistemas desenvolvidos. Nesses casos, a má qualidade na origem é levada para dentro dos novos sistemas que por mais que não tenham bugs, ficam comprometidos e acabam sendo mal avaliados pelo cliente.

Consuelo foi acionada para tentar resolver esse problema.

ESTUDO DE CASO SIMULADO



Consuelo afirmou que esse tipo de situação se resolve com um processo de Data Quality (qualificação dos dados).

Segundo ela, existe uma máxima em Governança de Dados que fala:
“Melhor não possuir informação do que a informação estar errada”

Para explicar melhor como funciona o Data Quality, Consuelo preparou um curto treinamento que vem a seguir.

Qualidade de Dados

A produção, retenção, manutenção, entrega e gerenciamento do conteúdo informacional depende de processos bem definidos e gerenciados.

Esses processos precisam estar descritos e pessoas treinadas para cumpri-los, realizando atividades com ferramentas específicas.

Exemplo:

-Se uma carga de dados sobre vendas em lojas depende da execução de um programa de captura de movimento de caixa que roda em uma intranet, e esse programa é disparado manualmente pelo operador do datacenter entre as 21 e 22 horas de cada dia, conforme as lojas apontam que fecharam seus caixas no sistema de vendas, o operador precisa estar treinado e seguir a risca essas regras, as quais serão auditadas pelo seu coordenador.

Qualidade de Dados

Além da definição de processos adequados, a qualidade dos dados depende da **qualidade** das suas definições (**metadados/modelos descritivos e estruturas de dados**) que permitiram o armazenamento, aproveitamento, manutenções e consultas de forma mais fácil ou mais difícil, dependendo da sua concepção.

Qualidade de Dados

Resumindo...

A Qualidade de Dados pode ser subdividida em três ênfases:

- Qualidade dos Metadados;
- Qualidade do Conteúdo informacional;
- Qualidade dos Processos que sustentam o ciclo de vida informacional.



Qualidade de Dados - Metadados

Resumindo...

A Qualidade de Dados pode ser subdividida em três ênfases:

- Qualidade dos Metadados;
 - Qualidade do Conteúdo informacional;
 - Qualidade dos Processos que sustentam o ciclo de vida informacional.
- 
- Amplamente explorado em disciplinas de Design SW, Data Base Modeling e Quality

Qualidade de Dados - Metadados

Resumindo...

A Qualidade de Dados pode ser subdividida em três ênfases:

- Qualidade dos Metadados;
- Qualidade do Conteúdo informacional;
- Qualidade dos Processos que sustentam o ciclo de vida informacional.

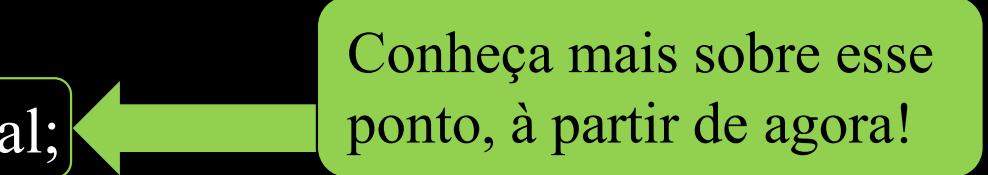


Amplamente explorado em todas as disciplinas do seu curso e reforçada em Compliance & Quality

Qualidade de Dados – Conteúdo

Resumindo...

A Qualidade de Dados pode ser subdividida em três ênfases:

- Qualidade dos Metadados;
- Qualidade do Conteúdo informacional; Conheça mais sobre esse ponto, à partir de agora!
- Qualidade dos Processos que sustentam o ciclo de vida informacional.

Qualidade de Dados – Conteúdo

Dados de BAIXA QUALIDADE – popularmente chamados de Dados Sujos

Dados Sujos são dados que descumprem a sua função informacional, ou seja, deixam de entregar a informação correta, realista e atualizada para que a empresa opere adequadamente.

Qualidade de Dados – Conteúdo

- E por que os dados ficam errados ou “sujos”? Existem muitas maneiras dos dados ficarem sujos, mas as principais categorias são:
- **Valores Default DUMMY:** Quando encontramos Defaults para os valores de colunas ou campos obrigatórios.
Exemplo: CPF com 999.999.999-9
- **Valores Default “INTELIGENTES”:** Quando os Defaults possuem significado.
Exemplo: Se a coluna IDADE contiver 000 o cliente é corporativo!
- <continua no próximo slide>

Qualidade de Dados – Conteúdo

- **Valores contraditórios:** Quando os valores de uma coluna ou campo são inconsistentes com os valores de outra coluna ou campo relacionado.
- Exemplo: o CEP informado não bate com a Cidade informada

- **Valores que estão fora do domínio:** Quando os valores encontrados não estão dentro do esperado para aquele campo.
- Exemplo: o sexo esperava F – Feminino M – Masculino e encontramos U

- **Valores em desacordo com a Regra de Negócio:** Se o a coluna desconto só terá valor superior a 20% quando o valor do produto ultrapassar R\$1000,00 e encontramos valorer de 30% para preço de produto de R\$500,00

Qualidade de Dados – Conteúdo



Falando da sua realidade...

Você já vivenciou um problema de falta de qualidade em dados?

Conte um exemplo de problema que enfrentou por não ter uma informação correta sobre algo ou alguém ter uma informação incorreta sobre você!

Qualidade de Dados – Conteúdo

O dado sujo pode existir em um único repositório consolidado ou em situações de distribuição.

Qualidade de Dados – Conteúdo

O dado sujo pode existir em um único repositório consolidado ou em situações de distribuição.

Seus impactos podem ser **desde pequenos transtornos que implicam em ações de correção baratas ou pequenas indenizações e multas até situações que podem gerar um colapso financeiro, moral e legal!**



Qualidade de Dados – Conteúdo

A Qualidade dos Dados em termos de conteúdo só será alcançada se as pessoas, atividades de trabalho e ferramentas empregadas na administração dos dados garantam:

- Harmonização dos conceitos e definições sobre como os dados são apurados, atualizados, usados e retidos;
- Limpeza/expurgo/desconsideração de dados que conflitem com as definições de harmonização (expurgo de dados com data de utilidade vencida ou que tenham erros de apuração/atualização) – processo de Data Cleansing.
- Correção de dados como alternativa para a limpeza, quando for possível a harmonização.

Qualidade de Dados - Processo

Resumindo...

A Qualidade de Dados pode ser subdividida em três ênfases:

- Qualidade dos Metadados;
- Qualidade do Conteúdo informacional;
- Qualidade dos Processos que sustentam o ciclo de vida informacional.



Vamos conhecer mais sobre o processo de garantia
de qualidade de dados

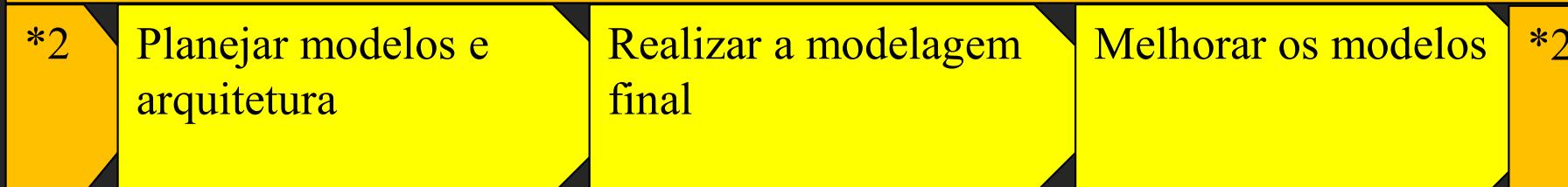
Qualidade de Dados - Processo

Seguindo as propostas do DAMA-DMBok*, devem existir processos com atividades, pessoas e ferramentas previstas para cumprir:

Qualidade de Conteúdo de Dados

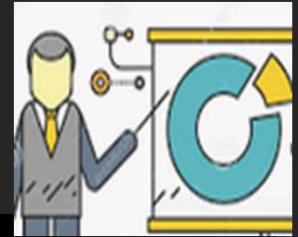


Qualidade do Modelo de Dados



*Principal referência na área de Governança de Dados

Qualidade de Dados - Processo



O primeiro ponto a tratar são **Parâmetros da Qualidade** que pedem que o Processo de Modelagem respeite:

- Identificação das pessoas corretas para participarem da modelagem as quais estejam capacitadas e sejam experientes em arquitetura de dados;
- Garantia de participação de pessoas chave que cuidam tecnicamente do desenvolvimento, manutenção, operação e suporte aos sistemas de informação legados;
- Aplicação das práticas de escolha de arquitetura que privilegiem o desempenho de entrega de informação, a segurança de acesso e a confiabilidade dos dados.

Qualidade de Dados - Processo



Considerando o Conteúdo Informacional, os **Parâmetros da Qualidade** pedem:

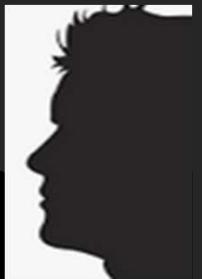
- 1) Acurácia: identificar quais as entidades da realidade da empresa que serão representadas através dos seus dados (pessoas, coisas, locais, etc.)
- 2) Completude: estabelece o conjunto de dados total necessário para explicar tudo o que a empresa precisa saber sobre a entidade real representada
- 3) Consistência: cruzar mais de uma fonte de informação antes de fechar uma definição sobre o conteúdo/significado dos dados
- 4) Atualidade: garantir que o conteúdo represente o estado mais atual da entidade que representa
- 5) Privacidade: possibilitar acesso para incluir, alterar, excluir e/ou consultar conteúdo, conforme um perfil pré-definido de pessoas/funções/cargos reesposáveis e habilitadas.

Qualidade de Dados - Processo

...

- 6) Razoabilidade: determinar o volume de transações sobre os dados de forma a dimensionar adequadamente a infraestrutura para garantir desempenho e disponibilidade
- 7) Integridade: referenciar dados entre si para representar as relações entre as entidades reais que representam
- 8) Unicidade: garantir que um dado não tenha conteúdo diferente em diversas instâncias, se esse dado representa a mesma entidade real – garantir também a inexistência de duplicação do mesmo conteúdo informacional
- 9) Validade: formatar e configurar regras de domínio sobre o dado de forma adequada.

Qualidade de Dados - Processo



O segundo ponto a tratar na Qualidade de Conteúdo é a **Perfilamento** (Data Profiling).

Ela trata da realização de **análises sobre os dados** e não a partir dos dados.

Exemplos de análises sobre os dados que permitem avaliar o arquivamento e uso de dados de forma a possibilitar a melhor administração possível dos repositórios de conteúdo:

- Existem 5.930 registros de clientes
- São registradas em média 294 notas fiscais de venda por dia
- Existem 92 registros de produtos com ID e nome e sem detalhes de unidade de medida e preço de reposição
- Das 14.005 linhas de registros de logradouros, existem 13.809 com CEP em branco
- O range de valores válidos para PercentualDesconto no cadastro de produtos vai de 0% a 7%

Qualidade de Dados - Processo

O objetivo do perfilamento é avaliar se existe algo estranho nos conteúdos. Algo que possa comprometer a utilidade.

Importante reforçar que não se trata de análise a partir dos dados (Data Analytics). **NÃO É OBJETIVO DA DATA PROFILING:**

- Apontar quem são os maiores consumidores de um produto
- Saber qual filial vende mais
- Apontar tendências de interesses de consumidores sobre determinados serviços
- Etc.

Qualidade de Dados - Processo



O terceiro ponto a tratar na Qualidade de Conteúdo é a **Análise de Dados (Data Analysis)**

Ela trata da realização de **análises do conteúdo em si, comparando-o com expectativas**.

Funciona auditando os registros de dados, verificando se o conteúdo nos repositórios respeita os parâmetros de qualidade.

Alguns exemplos de falta de qualidade verificada em processo de análise:

- Foram encontradas 12 linhas na tabela de clientes com a coluna NM_CLI (nome do cliente) com “XXX”
- O campo VL_VDA_LQD (valor da venda líquida) registrado na tabela REG_VDA está somando o imposto sobre produtos industrializados (IPI); quando não deveria.

Qualidade de Dados - Processo

O quarto e último ponto a tratar na Qualidade de Conteúdo é a **Limpeza e Ajuste de Dados (Data Cleansing)**

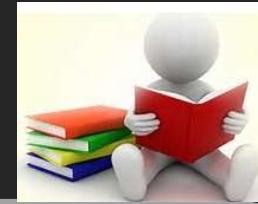


Ela trata da realização de **complementações, alterações ou expurgo de conteúdo** informacional para que o conteúdo atenda os parâmetros da qualidade de dados.

Essa atividade pode ser feita de forma manual ou usando ferramentas de bancos de dados e outros softwares utilitários.

Obs: intervenções manuais de limpeza, alteração ou complementação de dados só são recomendadas quando o volume de dados é pequeno.

ESTUDO DE CASO SIMULADO



Uma das pessoas que está recebendo este treinamento fez o seguinte questionamento a C, sobre um caso real que está acontecendo em um projeto na qual está atuando:

“Considerando os problemas a seguir, os quais eu estou enfrentando, quais seriam as ações de Cleansing que você proporia?

-Foram encontradas 12 linhas na tabela de clientes com a coluna NM_CLI (nome do cliente) com “XXX”

-O campo VL_VDA_LQD (valor da venda líquida) registrado na tabela REG_VDA está somando o imposto sobre produtos industrializados (IPI); quando não deveria.”

Qualidade de Dados - Processo



É importante que o processo de Cleansing realmente o ciclo de melhoria contínua da qualidade.

Para isso é importante ter indicadores sobre o nível de qualidade atual como:

- Número de chamados/ocorrências de necessidade de correção/complementação/exclusão de dados
- Número de ocorrências resolvidas

Qualidade de Dados - Manutenção

Uma vez limpados e ajustados os dados, é necessário manter a qualidade desses dados!



Para manter temos que ir aos produtores da informação e criar barreiras para entradas ruins.

Em sistemas próprios a estratégia é barrar. Por exemplo, não aceitar CPFs inválidos, incompletos, e-mails sem @, etc.

Mas nossa atenção precisa estar também em base de dados compradas, de má qualidade – muitas empresas adquirem arquivos de terceiros como tabelas de CEP dos correios, listas de clientes de empresas de marketing e esses dados podem vir com má qualidade.

Ainda precisamos estar atentos a manutenção dos sistemas, porque a medida que vão evoluindo, podem deixar de fazer validações importantes, alterarem regras de negócio que irão afetar a qualidade.

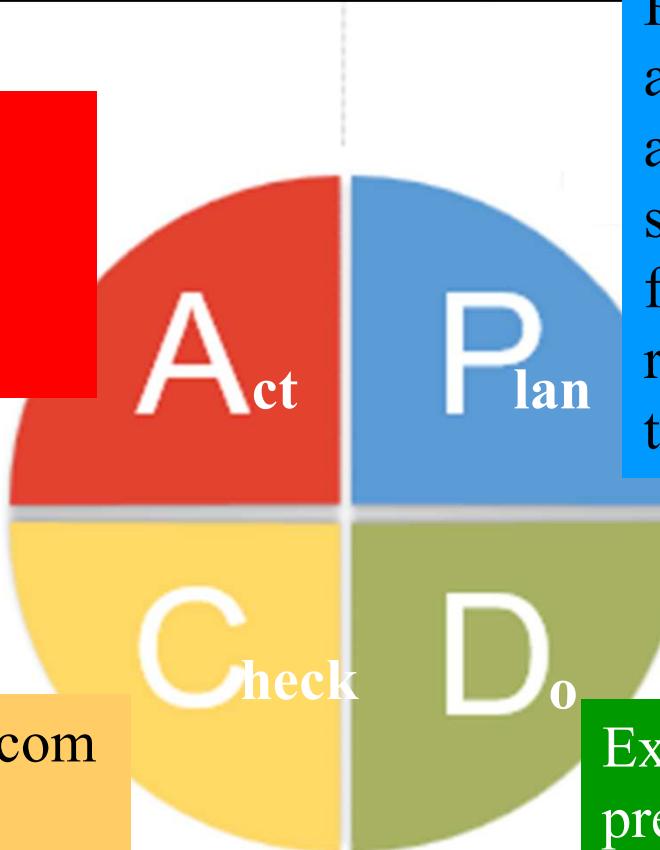
Qualidade de Dados

Considerações finais.

Lembre-se de que a Gestão da Qualidade de Dados prevê um ciclo permanente de melhoria que envolve...

Concluir sobre mudanças necessárias para melhorar e corrigir a forma de trabalho atual

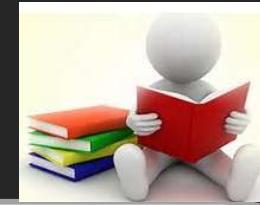
Planejar a implementação das ações necessárias, prevendo aquisições de equipamentos e software, pessoas e suas tarefas, forma de acompanhar e reportar riscos e resultados, tempo de trabalho e metas de entrega



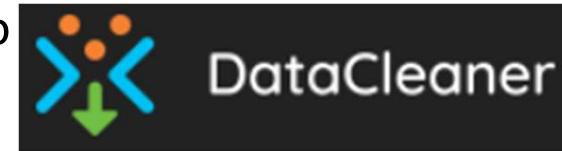
Verificar resultados obtidos com a forma de trabalho atual

Executar o plano, conforme previsto

ESTUDO DE CASO SIMULADO

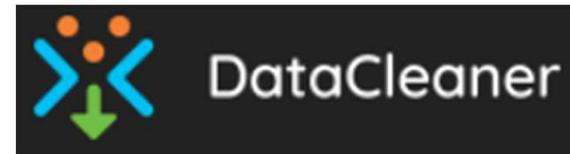


C trouxe uma ferramenta gratuita para a GD aprender a usar nos trabalhos de Data Quality – o



Siga as instruções de C, passo a passo, para aprender como usar essa ferramenta e otimizar os trabalhos de perfilamento, limpeza e análise de dados.

Qualidade de Dados - Automação



Execute a ferramenta
diretório de rede.

Abra o arquivo Customer.CSV e investigue possíveis problemas nos dados, de forma visual. Faça isso por 10 minutos.

Em seguida, vamos explorar alguns exemplos de complementação e limpeza de dados em nessa tabela de Clientes que já foi carregada no software DataCleaner e ver o que esse produto é capaz de fazer.

Veja no slide a seguir, as instruções sobre quais problemas de qualidade de dados você pode investigar...

AULA PRÁTICA

SOLUÇÃO EM SALA
DE AULA**Leia o estudo de Caso de Data Quality**Desafio: avaliar com
Data Cleaner

Observe o arquivo FIAP-QualidProjSW-Aula-20-customers-DataCleaner...

1. Descubra se existe chave única
2. Verifique se há nome ou sobrenomes nulos
3. Verifique se o nome da companhia possui somente letras
4. Verifique se há nome com apenas 1 caracter?
5. Avalie o gênero
6. O que você pode me dizer de país?
7. A data de nascimento é valida? Existe datas nulas? Existem datas com letras?
8. Todos os emails são válidos? Quais as análises que foram feitas?
9. A profissão está padronizada?

AULA PRÁTICA

SOLUÇÃO EM SALA
DE AULA

Agora, abra o arquivo de Clientes, que já está disponível na ferramenta (Customers Profiling)!

Demonstração do uso do Data Cleaner

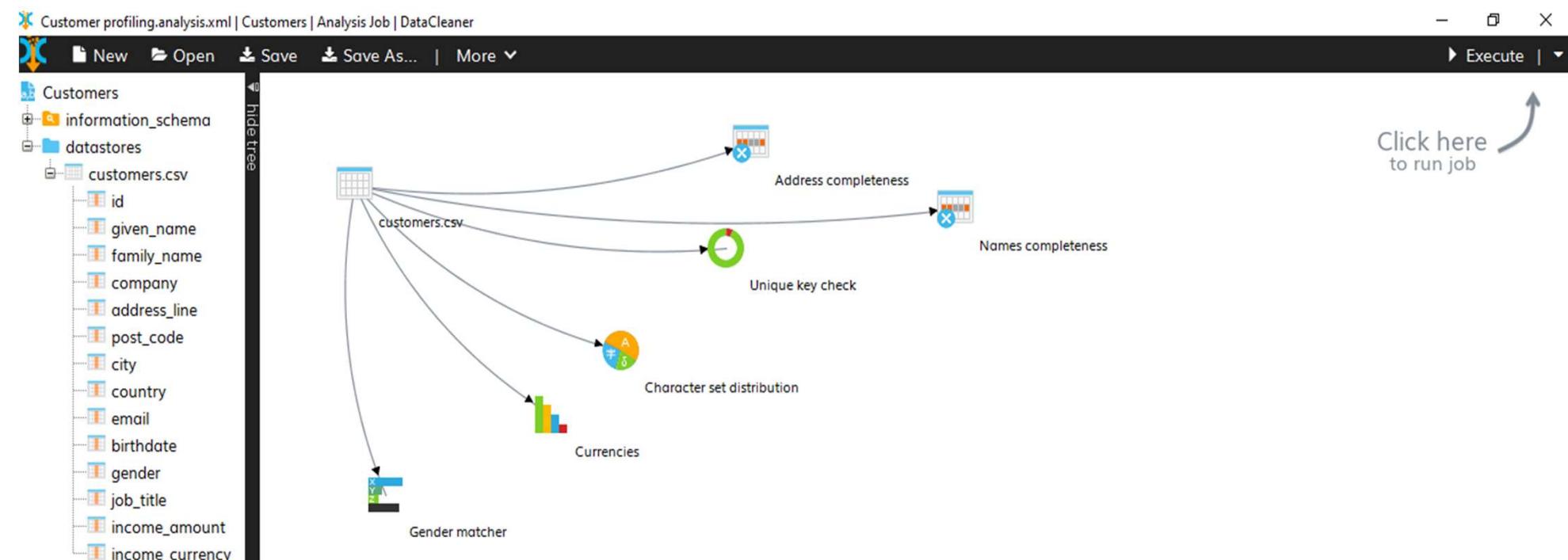
The screenshot shows the Data Cleaner application window. At the top, there's a toolbar with icons for New, Open, Save, Save As..., and More. Below the toolbar, a message says "Welcome to DataCleaner". On the left, there's a sidebar with a "Community Edition" notice and links to forums, Twitter, and LinkedIn. At the bottom, there are buttons for "Build new job", "Browse jobs", "Recent jobs", and "Manage datastores". A central "Open" dialog box is displayed, showing a list of XML files under the "jobs" folder. One file, "Customer profiling.analysis.xml", is highlighted with a red box and a red arrow pointing to it from the left side of the slide. The dialog also includes fields for "Nome do Arquivo:" and "Arquivos do Tipo:" set to "DataCleaner analysis files", and buttons for "Abrir" and "Cancelar".

AULA PRÁTICA

SOLUÇÃO EM SALA
DE AULA

Obtenha as respostas para a questão anterior, usando o exemplo carregado no Data Cleaner (arquivo de Clientes), disponibilizado pelo professor! Navegue na ferramenta e explore o que ela pode gerar automaticamente para você!

Demonstração do uso
do Data Cleaner



Qualidade de Dados - Automação

Conhecendo mais sobre o



- 1º) Crie o seu arquivo para avaliação em um SGBD ou crie um arquivo texto (com campos separados por vírgula – CSV).
- 2º) Acessando NEW no menu inicial da ferramenta, vá até Manage Data Storages.

A screenshot of the DataCleaner software interface. The title bar says "Manage datastores | DataCleaner". Below the title bar is a toolbar with icons for New (highlighted with a red box), Open, Save, Save As..., and More. A large red arrow points to the "Datastore Management" link under the toolbar. The main area shows "Existing datastores":

<input checked="" type="checkbox"/>		Customers Example CSV-file with representing customers' details	Edit	Query	Remove
<input type="checkbox"/>		Salesforce.com Example connection to SFDC - credentials not provided	Edit	Query	Remove
<input type="checkbox"/>		orderdb Example database for use with DataCleaner	Edit	Query	Remove

[Build job](#)

Qualidade de Dados - Automação

Conhecendo mais sobre o



3º) Escolha o tipo de arquivo que quer carregar.

The screenshot shows the DataCleaner application's Datastore Management screen. At the top, there is a toolbar with icons for New, Open, Save, Save As..., and More. Below the toolbar, the title "Datastore Management" is displayed next to a back arrow icon. A search bar labeled "Search/filter datastores" is positioned on the right side of the header.

The main area is titled "Existing datastores:" and lists three entries:

- Customers**: Example CSV-file with representing customers' details. This entry has a checked checkbox, an "Edit" button, a "Query" button, and a "Remove" button.
- Salesforce.com**: Example connection to SFDC - credentials not provided. This entry has an unchecked checkbox, an "Edit" button, a "Query" button, and a "Remove" button.
- orderdb**: Example database for use with DataCleaner. This entry has an unchecked checkbox, an "Edit" button, a "Query" button, and a "Remove" button.

At the bottom left, there is a "Build job" button. In the bottom right corner of the main window, there is a small image of a white boat on water.

At the bottom of the screen, there is a section titled "Register new:" with a red arrow pointing to it. This section contains a "Register new:" label and a row of icons representing various data sources and databases, including CSV, Excel, Microsoft Word, SharePoint, Google Sheets, MySQL, PostgreSQL, Oracle, and MongoDB. Below this row, there are more icons for various databases and systems, and a "More databases" button.

Qualidade de Dados - Automação

Conhecendo mais sobre o



4º) Dê um nome para o arquivo de dados e carregue na ferramenta.

The screenshot shows the DataCleaner application interface. On the left, the 'Datastore Management' screen lists existing datastores: 'Customers' (checked), 'Salesforce.com', 'orderdb', and a 'Build job' button. Below this is a 'Register new:' section with various icons for different data sources. On the right, a modal window titled 'CSV file datastore | DataCleaner' is open, showing configuration for a 'Comma-separated file'. The configuration includes:

- Datastore name: TabTesteFIAP
- Source: file (selected), e-Aula-18-PraticaDataQuality.csv (file path), Browse button
- Character encoding: UTF-8
- Separator: Comma (,)
- Quote char: Double quote (")
- Escape char: Backslash (\)
- Header line: 1
- Fail on inconsistent column count
- Enable multi-line values?

A red arrow points from the top right towards the 'Fail on inconsistent column count' checkbox. Below the configuration is a preview table of data:

id	given_...	family_...	com...	addres...	post_...	city	coun...	email	birth...
53	Paul	Taylor	Tesoro	8, Dyfrig ...	CF5 5AE	Cardiff	GBR	Paul.Ta...	2002-4...
188	Bibi	Asuncion	Nation...	16215 AL...	CA 926...	IRVINE	USA	Bibi.As...	1985.8...
189	Robert	Shanahan	Plains ...	124, Park ...	EN6 5EL	Potters...	GBR	Robert...	1991-6-5
191	Stephen	Sheridan	Exxon ...	Britten Dr...	EX328AQ	Barnst...	GBR	Stephe...	1996-1...
209	Gerner	Kristensen	Best Buy	Lindealle ...	3600	Frederi...	DNK	Gerner...	1969-9...
208	Sonja	Ermer	Comcast	Arndtstr. ...	23566	Lübeck	DEU	Sonja.E...	1990-1...
210	BOBBI	Miranda	Wells F...	757 THIR...	NY 100...	NEW Y...	USA	BOBBI...	1964-9-4

At the bottom of the modal are 'Register datastore' and 'Cancel' buttons.

Qualidade de Dados - Automação

Conhecendo mais sobre o



5º) Confirme o registro da sua tabela a ser analisada e crie um JOB de análise.

The screenshot shows the DataCleaner interface for managing datastores. At the top, there's a toolbar with "Manage datastores | DataCleaner", "New", "Open", "Save", "Save As...", and a "More" dropdown. Below the toolbar, the title "Datastore Management" is displayed with a back arrow. The main area is titled "Existing datastores:" and contains a list of datastores:

- Customers**: Example CSV-file with representing customers' details. Status: . Actions: Edit, Query, Remove.
- Salesforce.com**: Example connection to SFDC - credentials not provided. Status: . Actions: Edit, Query, Remove.
- orderdb**: Example database for use with DataCleaner. Status: . Actions: Edit, Query, Remove.
- TabTesteFIAP**: C:\Users\renat\Downloads\FIAP-DataGovernance-Aula-18-customers-Dat. Status: . Actions: Edit, Query, Remove.

At the bottom left, a large red arrow points to the "Build job" button, which is highlighted with a red border. This button is located at the bottom of the datastore list.

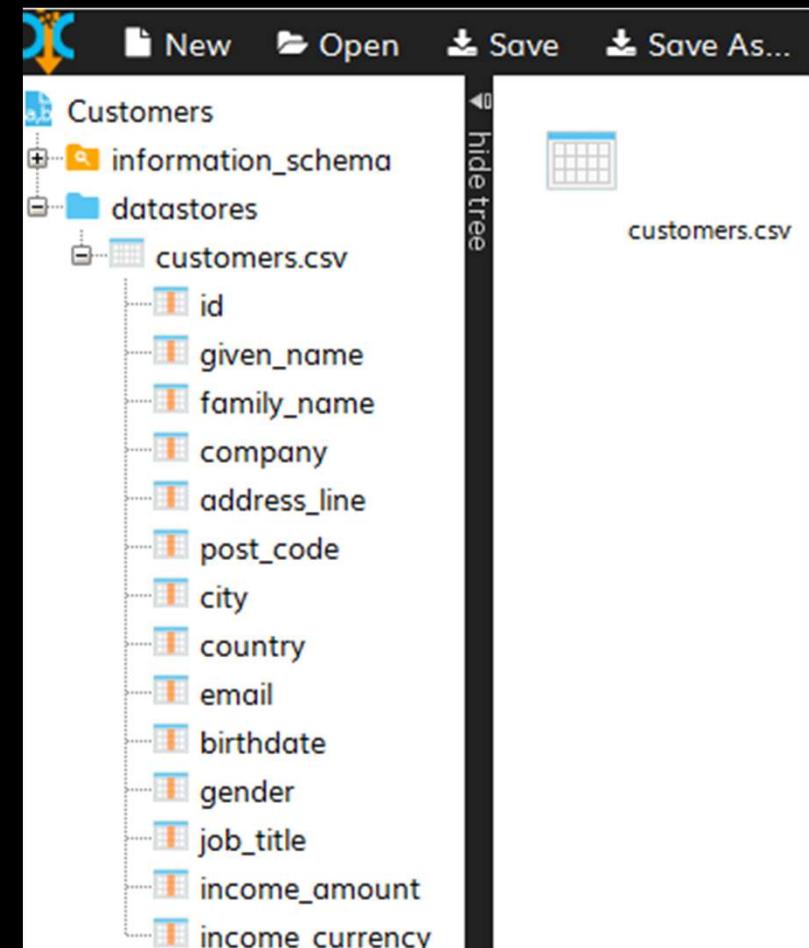
Qualidade de Dados - Automação

Conhecendo mais sobre o



DataCleaner

6º) Adicione as colunas para o profiling ao JOB (trabalho) de profiling que você está criando!



Qualidade de Dados - Automação

Conhecendo mais sobre o



DataCleaner

7º) Clique com o botão da direita do mouse na área em branco onde consta o seu arquivo para selecionar a operação que deseja fazer sobre os dados

The screenshot shows the DataCleaner application window. On the left, there's a sidebar with a tree view of datastores and a library section. The main area displays a CSV file named "FIAP-Datagovernance-Aula-18-PraticaDataQuality.csv". A context menu is open over the file, with the "Analyze" option highlighted and a red arrow pointing to it. The "Analyze" submenu is expanded, showing various data quality analysis tools like Date and time, Visualization, Number analyzer, Unique key check, etc. At the bottom of the main window, there's a message: "Start building ... Add components to your job to explore the library of available components". The taskbar at the bottom shows the DataCloud news channel and community edition.

Qualidade de Dados - Automação

Conhecendo mais sobre o



8º) Selecione por exemplo, checar a chave primária e depois, ligue a tabela à tarefa de profiling (clique com o botão da direita da tabela e adicione)!

The screenshot shows the DataCleaner interface with the following details:

- Left Panel:** Shows the project tree with a file named "FIAP-DataGovernance-Aula-18-PraticaDataQuality.csv" containing columns: id, given_name, family_name, company, address_line, post_code, city, country, email, birthdate, gender, job_title, income_amount, and income_currency.
- Middle Panel:** A tooltip for a right-clicked table row says: "Right-click the source table to link it to a task. This directs the flow of data from the table to the task." It lists options: "Link to ...", "Preview data", and "Remove table from source".
- Right Panel:** The "Unique key check" task configuration window.
 - Input columns:** A list of columns with "id" selected (radio button highlighted).
 - Required properties:** Buffer size set to 20000.

Two red arrows highlight the "Link to ..." option in the tooltip and the "id" selection in the "Input columns" list.

Qualidade de Dados - Automação

Conhecendo mais sobre o



9º) Execute a análise!

The screenshot shows the DataCleaner interface with the following elements:

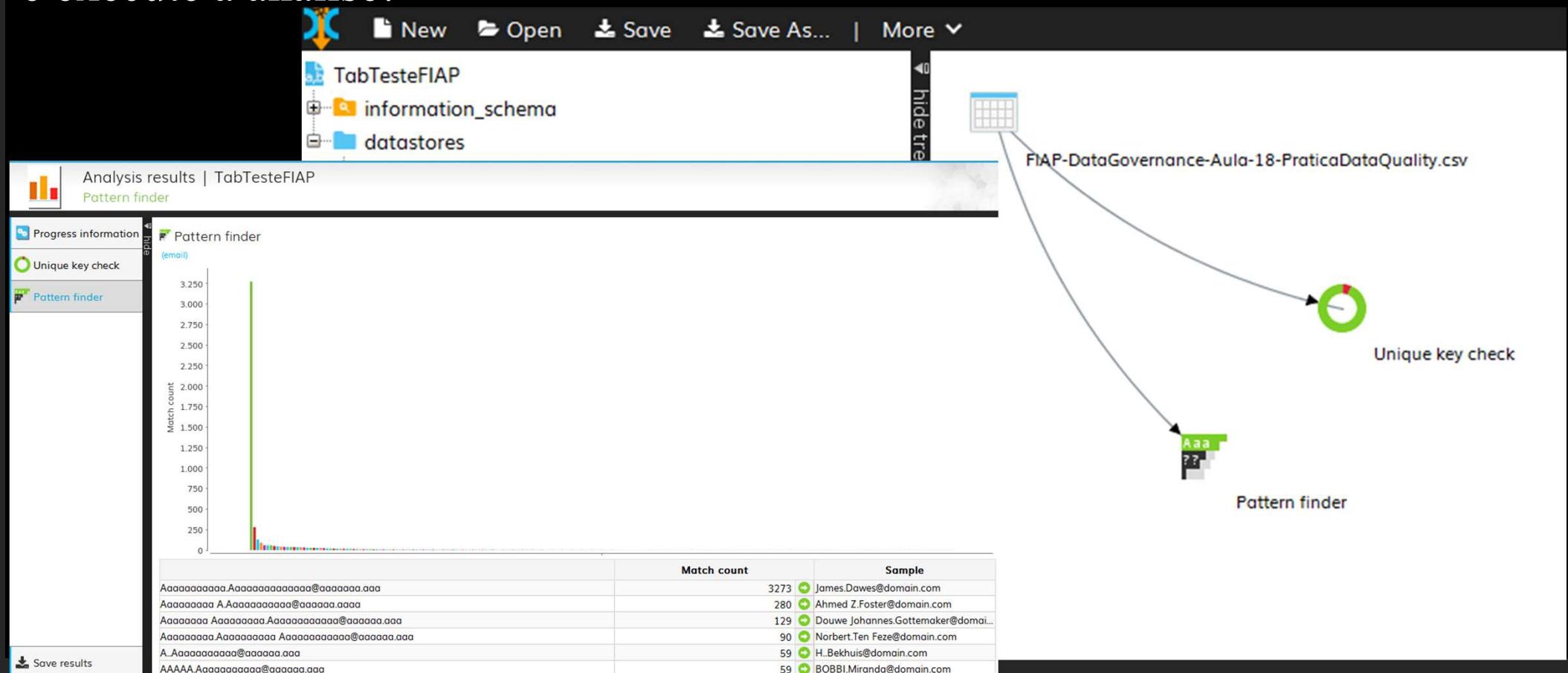
- Toolbar:** Includes "New", "Open", "Save", "Save As...", "More", "Execute", and other icons.
- Project Tree:** Shows a project named "TabTesteFIAP" containing "information_schema", "datastores", and a file named "FIAP-DataGovernance-Aula-18-PraticaDataQuality.csv". The CSV file is expanded, showing columns: id, given_name, family_name, company, address_line, post_code, city, country, email, birthdate, gender, job_title, income_amount, and income_currency.
- Job Graph:** A flowchart where the CSV file is connected to a "Unique key check" node.
- Callout:** A red callout box with a red arrow points to the "Execute" button in the toolbar, containing the text "Click here to run job".

Qualidade de Dados - Automação

Conhecendo mais sobre o



10º) Inclua uma análise de estrutura de formação do conteúdo do campo e-mail e execute a análise!



Qualidade de Dados - Automação

Conhecendo mais sobre o



11º) Inclua uma análise de distribuição de dados para verificar se existe um desbalanceamento de cadastros/concentração, usando o campo país!

The screenshot shows the DataCleaner application interface. At the top, there's a menu bar with options like New, Open, Save, Save As..., and More. Below the menu is a tree view showing a database connection to 'information_schema' and a data store named 'FIAP-Datagovernance-Aula-18-PraticaDataQuality.csv'. A 'Value distribution' tab is selected, displaying a chart and a table. The chart shows the value distribution of 'country' with a total count of 5115 and distinct count of 32. The table lists countries and their counts:

Value	COUNT(*)
GBR	1763
USA	1265
DEU	998
DNK	503
NLD	503
US	11
DE	10
GB	10
<blank>	9
DK	5
Denmark	4
NL	4
England	3
United States	3
germany	3
Deutschland	2
Great Britain	2
UK	2
netherlands	2
Danmark	1
Danemark	1
ENgland	1
GER	1
Netherlands, The	1

On the right side of the interface, there are three circular icons with arrows pointing to them: 'Unique key check', 'Pattern finder', and another 'Value distribution' icon.

Qualidade de Dados - Automação

Conhecendo mais sobre o



DataCleaner

12º) Por fim, avalie se existem campos em branco (não preenchidos)!

The screenshot shows the DataCleaner interface with a file named "FIAP-Datagovernance-Aula-18-PraticaDataQuality.csv" loaded. A "Completeness analyzer" icon is connected to the file, and a report titled "Incomplete records (116)" is displayed below. The report includes a "Save dataset" button. The data table lists 116 incomplete records from the CSV file, showing various columns like id, given_name, family_name, company, address_line, post_code, city, country, email, birthdate, gender, job_title, income_amount, etc.

id	given_name	family_name	company	address_line	post_code	city	country	email	birthdate	gender	job_title	income_amount	i
72	Ulla Bonde	Milbaek	Deere				DK	Ulla.Bond...	1901.10.4	M	President	874581.0	D
81	Ronald	Kamp	Hewlett-P...	De Hoven 61	7894 BP	ZWARTE...	NLD	Ronald.Ka...	1955-3-11	F	manager	606038.0	G
226	Christine	Matcham	Exxon Mo...	Silver Springs...			GBR	Christine....	1950-5-20	M	self	78491.0	G
508	Wilfred	JOHNSON	Sprint Ne...	Hamilton Vill...			GBR	Wilfred.JO...	1986-10-21	M	Data Scie...	45508.0	G
398	Wilfred	Gent	FedEx	Prospect Terr...			GBR	Wilfred.G...	1987-7-9	M	Data Anal...	52664.0	G
631	Bernard	Moore	Honeywell...	18, Wilson Ro...	SP4 8NL	Salisbury	GBR	Bernard....	1999-7-15	F	Owner	1037835.0	G
903	Keith	Wildegoo	News Corp.	Homecolne H...			GBR	Keith.Wild...	1993-1-14	F	Consultant	31770.0	G
612	Alan	Milsom	Ingram M...	Silver Springs...			GBR	Alan.Milo...	1999-3-17	F	Consultant	32825.0	G
1082	Amarily	Tenenbaum	Wells Fargo	600 S COLLE...	OK 74104-...	TULSA	USA	Amarily.T...	1994-4-11	F	consultant	1038832.0	G
764	Nick	PHILLIPS	Chevron	Prospect Terr...			GBR	Nick.PHIL...	1952-9-4	F	IT Manager	54389.0	G
1287	Gregory	Gribben	Abbott La...	Silver Springs...			GBR	Gregory.G...	2008-10-23	F	CIO	73249.0	G
667	Maria A	Jilley	Delta Air ...	Hamilton Vill...			GBR	Maria.A.Jil...	1927.1.22	F	Eng	59190.0	G
1439	Richard	Hanna	PepsiCo	Hamilton Vill...			GBR	Richard.H...	1957-9-17	M	Student	45517.0	G
1442	Alistair	Frost	Mondel��z	Prospect Terr...			GBR	Alistair.Fr...	1993-11-4	F	DBA	47683.0	G
1324	Mike	West	DuPont	Homecolne H...			GBR	Mike.West...	2013-10-23	M	title	1223325.0	G
1203	Alan	Palmer	HCA Hold...	Prospect Terr...			GBR	Alan.Palm...	1961-2-24	M	Architect	73079.0	G
1236	Marcel	Hayes	Target	2650 OCEAN...	NY 11235	BROOKLYN	USA	Marcel.Ha...	2004-7-24	M	Mrs.	n	G
1246	Bob	Hooper	World Fue...	Hamilton Vill...			GBR	Bob.Hoop...	1994-1-11	F	program...	46214.0	G
1501	Deborah	Gillitt	Bank of A...	Hamilton Vill...			GBR	Deborah....	1986-7-12	M	IT Architect	88468.0	G
1645	Curd	Jennewein	Amerisour...	Schultebeyrin...	49525	Lengerich	DEU	Curd.Jenn...	1995-2-1	M	Contractor	793208.0	G
1460	Bernhard	Schwefel	Hess	Krummstr. 45	40789	Monheim ...	DEU	Bernhard....	1982-6-16	M	Consultant	750204.0	G
1818	Ardell	Craig	Microsoft	FL 5,200 WHI...	NY 10591...	TARRYTO...	USA	Ardell.Crai...	1974-10-15	M	None	90579.0	G

Qualidade de Dados - Automação

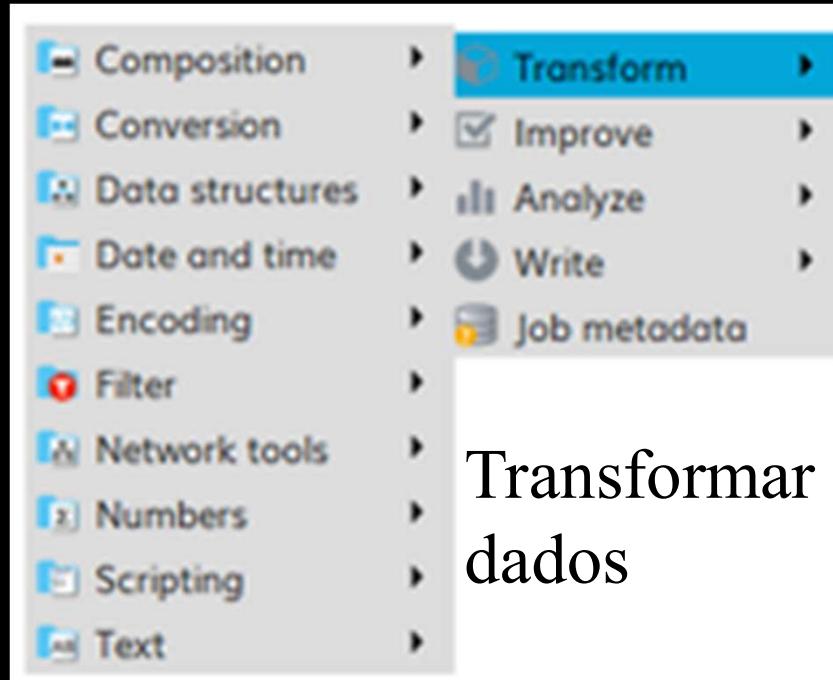
Conhecendo mais sobre o



DataCleaner

13º) Crie ações para sanear os dados!

Clicando com o botão da direita na área de desenho do JOB, você pode escolher as opções



Transformar
dados

Filtrar e escrever
um novo arquivo

- Create CSV file
- Insert into table
- Create staging table
- Delete from table
- Create Excel spreadsheet
- Update table

Transform
Improve
Analyze
Write
Job metadata

para tratar os
dados
isoladamente

Aplicar padrões para
melhorar
dados com templates
de referência.

Transform
Location
Reference data
Improve
Analyze
Write
Job metadata

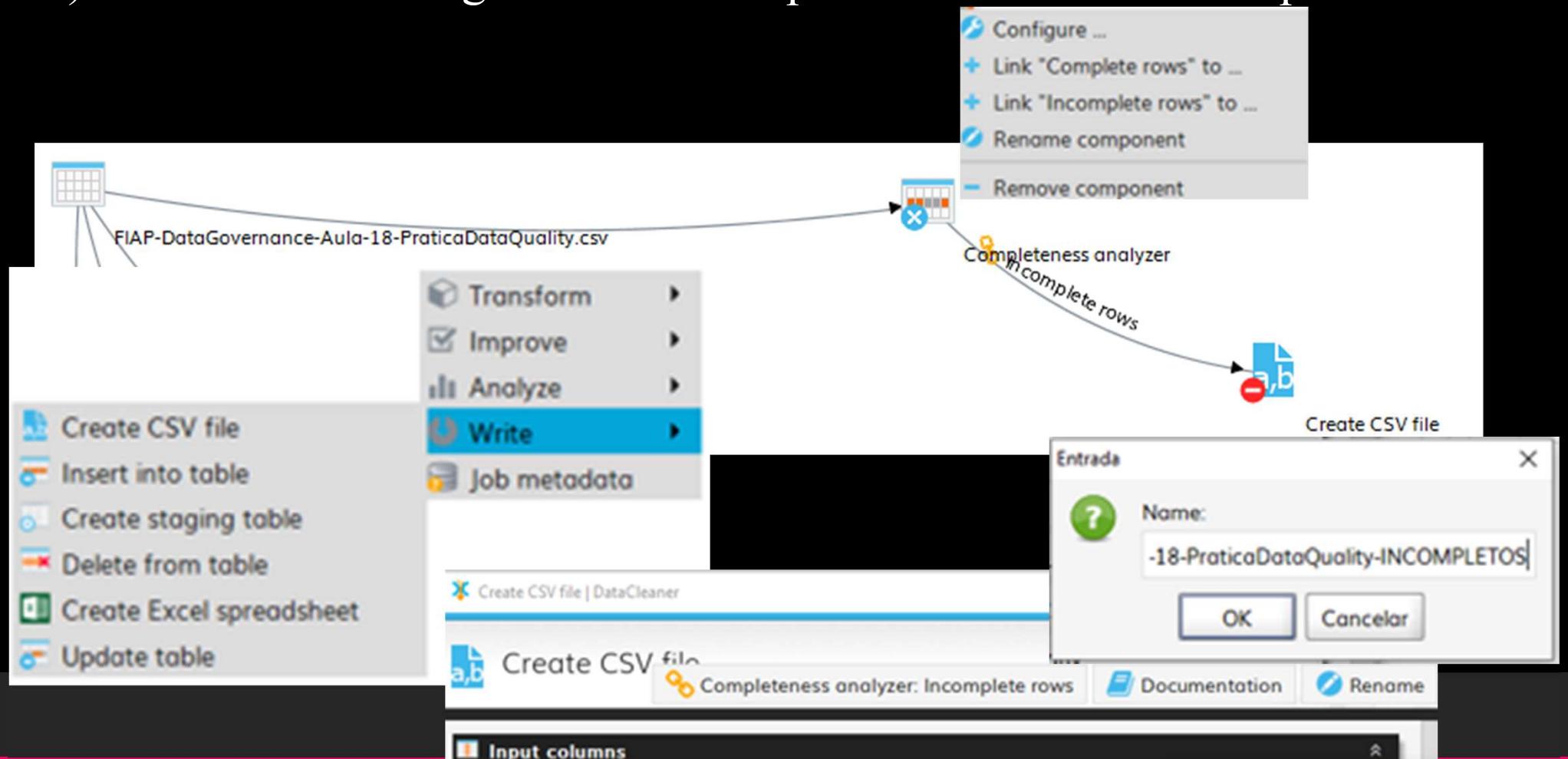
Qualidade de Dados - Automação

Conhecendo mais sobre o



DataCleaner

14º) Vamos isolar os registros com campos nulos em um CSV a parte!



Qualidade de Dados - Automação

Conhecendo mais sobre o



DataCleaner

15º) Você pode re-generar Id ...

The screenshot shows the DataCleaner interface with a modal dialog box titled "Generate ID". The dialog has two main sections: "Input columns" and "Required properties".

Input columns:

Column in scope:
A column which represent the scope for which the ID will be generated. If eg. a source column is selected, an ID will be generated for each source record. If a transformed column is selected, an ID will be generated for each record generated that has this column.

Search/filter columns:

- id
- given_name
- family_name
- company
- address_line
- post_code
- city
- country
- email
- birthdate
- gender
- job_title
- income_amount
- income_currency

Required properties:

Id type: Sequence
Offset: 0

Buttons at the bottom: Close (with an X icon)

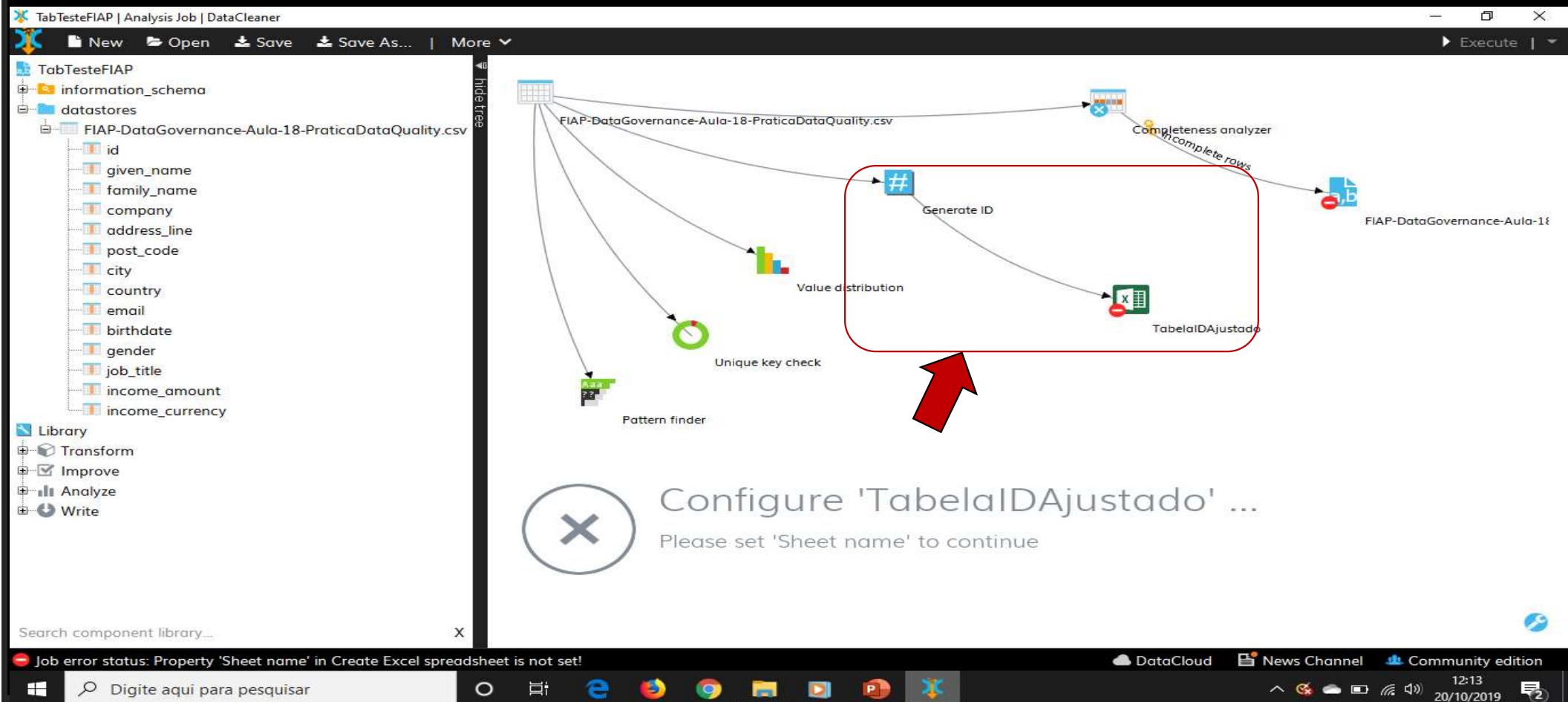
A large blue arrow points from the "id" radio button in the "Input columns" section to a large blue "# Generate ID" button on the right side of the slide.

Qualidade de Dados - Automação

Conhecendo mais sobre o



16º) ... e depois, gerar uma nova versão do arquivo.

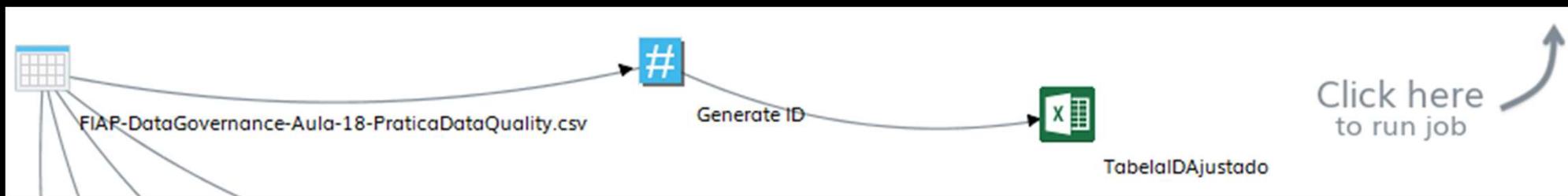


Qualidade de Dados - Automação

Conhecendo mais sobre o



17º) ... execute e observe o resultado da regeneração do ID!



A screenshot of Microsoft Excel showing the "TabelalDAjustado - Excel" sheet. The ribbon menu is visible at the top. The data in the spreadsheet is as follows:

Qualidade de Dados - Automação

Conhecendo mais sobre o



DataCleaner

É possível ainda, alterar valores com base em referências e muitos outros recursos.



DataCleaner é uma ferramenta poderosa, capaz de analisar com rapidez, uma grande variedade de bancos de dados Relacionais ou Não e trará esses dados, promovendo a qualidade na integração de dados para posterior uso em processos de tomada de decisão nas organizações.

Qualidade de Dados - Automação

São muitas as ferramentas disponíveis no mercado:



Ataccama



MasterDataOnline (MDO)



Informatica Data Quality



31 Ratings



Cloudingo



Dataloader.io



Talend Data Quality



21 Ratings



DemandTools



Syncsort Trillium DQ (formerly Trillium Software System)



SAP Agile Data Preparation



4 Ratings



TIBCO Clarity



Qualidade de Dados - Automação

Assim que puder, estude mais sobre elas, pesquisando na WEB!



RedPoint Data Management & Quality



Clear Analytics



VeriAS



StarDQ



Netlink Dataware



PeopleImport



DupeBlocker



RingLead DMS Cleanse



Validity Trust Assessments

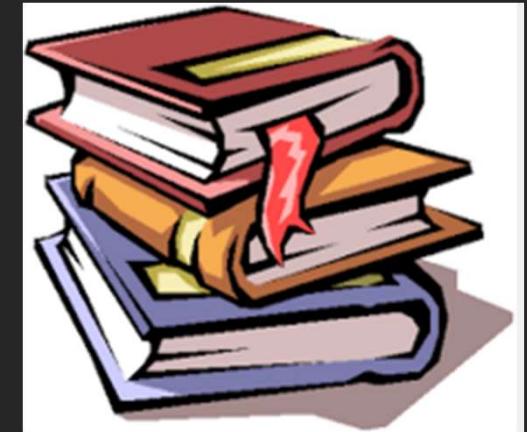




D Ú V I D A S

Material de aula estará no site após a aula.

BONS ESTUDOS!



Bibliografia

- RÊGO, Bergson Lopes. GESTÃO E GOVERNANÇA DE DADOS – Promovendo dados como ativo de valor nas empresas. São Paulo. Brasport, 2013.

DATA QUALITY

Fim

PROFESSOR:
RENATO JARDIM PARDUCCI

PROFRENATO.PARDUCCI@FIAP.COM.BR