Universidade Federal de Pernambuco

Centro de Informática

Graduação em Engenharia da Computação

Data profiling: Uma análise funcional de ferramentas gratuitas

Sylvia Marcella Dubeux Ratis

Recife

2017

Sylvia Marcella Dubeux Ratis

Data profiling: Uma análise funcional de ferramentas gratuitas

Trabalho apresentado ao Programa de Graduação em Engenharia da Computação da Universidade Federal de Pernambuco como requisito parcial para obtenção do grau de Bacharel em Engenharia da Computação.

Orientador: Robson do Nascimento Fidalgo

Recife

2017

Resumo

Sistemas de informação são críticos para dar suporte a decisões táticas, estratégicas e operacionais dentro de uma organização. A qualidade das decisões, no entanto, está diretamente relacionada à qualidade da informação fornecida para o gestor; esta, por sua vez, depende intrinsecamente da qualidade dos dados dentro do sistema. A geração de perfis de dados, ou *data profiling*, é uma maneira de demonstrar a existência de problemas nos dados que podem afetar o desempenho da organização. Este trabalho contém a realização de um estudo a respeito de *data profiling* e conceitos relacionados, bem como a identificação das operações mais comuns a serem realizadas sobre os dados na criação de um *data profile*. Os conhecimentos adquiridos são então utilizados na elaboração de uma análise funcional de três ferramentas comerciais gratuitas de *profiling*.

Agradecimentos

A Deus, em primeiro lugar, por ter me dado vida e me concedido força e coragem para concluir esta jornada. Dou graças a Ele por todos os momentos (bons e maus) que passei ao longo do caminho. A Ele seja toda honra e toda glória.

À minha família, pelo apoio incondicional (e pelos puxões de orelha necessários). Agradeço especialmente à minha mãe, Renata, por ter sido minha amiga fiel e companheira dedicada; por ter orado por mim, moldado o meu caráter e me ensinado os valores que carrego hoje comigo. Divido com ela, e com minha avó Sylvia (in memoriam), os louros dessa vitória. Agradeço também à minha "prima-generala" Júlia pela dedicação a me manter na linha na reta final desta jornada.

À equipe docente do Centro de Informática; especialmente aos professores Robson Fidalgo e Fernando Fonseca, por terem proporcionado as oportunidades que fizeram com que eu me encontrasse dentro do curso de graduação.

Aos amigos de CIn que estiveram ao meu lado durante toda a graduação, pelos conselhos, risadas, conversas e ocasionais noites sem dormir: Ryan Bernardo, Victor Brasileiro, Diego Queiroz, Carlos Azevedo, Felipe Lapenda e Bruno D'Ambrosio.

Aos amigos e colegas de Ávila Soluções pelo suporte na construção deste trabalho de graduação; em especial a Paulo Carvalho, que tornou-se muito mais do que um amigo ao longo do caminho, por ter me apoiado e acreditado em mim quando eu mesma não acreditava; e a Rodrigo Mateus, meu "pseudo-orientador", por ter duvidado e me incentivado a provar que ele estava errado.

A todos que contribuíram, direta ou indiretamente, para a minha formação, deixo aqui o meu muito obrigada.

Lista de Figuras

Lista de Tabelas

Sumário

1. Introdução

1.1. Contextualização

Sistemas de informação são cada vez mais críticos para dar suporte a decisões táticas, estratégicas e operacionais dentro de uma organização. Nos Estados Unidos, acima de 95% das organizações afirmam utilizar dados para impulsionar oportunidades de negócio, enquanto 84% acreditam que o uso de dados é uma parte fundamental da elaboração de estratégias de mercado. No entanto, a nível mundial, apenas 44% das organizações confiam nos dados à sua disposição para tomar decisões importantes de negócio, e mais da metade afirmam que essa falta de confiança ameaça a fidelidade de seus clientes finais (EXPERIAN DATA QUALITY, 2017).

A qualidade das decisões a serem tomadas, portanto, está diretamente relacionada à qualidade da informação fornecida para o gestor; esta, por sua vez, depende intrinsecamente da qualidade dos dados dentro do sistema. O gerenciamento da qualidade de dados torna-se mais importante à medida que cresce a dependência de tecnologias orientadas por dados. Singh e Singh (2010) consideram que, idealmente, os dados utilizados para tomada de decisão devem ser legíveis, compreensíveis, consistentes, relevantes e temporalmente oportunos. Apesar da importância de manter a qualidade de dados de um sistema, muitas organizações possuem pouca ou nenhuma iniciativa de governança ou gerenciamento de qualidade de dados, devido à visão de gestores que acreditam que se trata de um problema de TI, não de negócios, e portanto não estão dispostos a realizar investimentos nesse sentido (BLOOR RESEARCH GROUP, 2014).

O processo de geração de perfil de dados, ou *data profiling*, consiste em examinar os dados disponíveis em uma determinada fonte e coletar informações a respeito deles, produzindo metadados cuja análise é um passo importante para gerenciar a qualidade dos dados da fonte. Um cenário típico seria a varredura das tabelas de um banco de dados relacional para obter informações como tipos de dados, padrões de valores, completude e unicidade de colunas, ou até mesmo dependências funcionais e regras de associação (NAUMANN, 2014).

1.2. Motivação

Durante muito tempo, o processo de geração de perfil de dados foi considerado uma tarefa secundária e praticamente opcional dentro do processo de extração, transformação e carga de dados (ETL) de um *data warehouse*, objetivando apenas a identificação de pequenas anomalias a serem corrigidas nos dados antes da entrega do produto final. No entanto, Kimball (2004) considera que *data profiling* realizado no início de um projeto pode ter efeitos bastante significativos, inclusive determinando se o projeto deve ou não prosseguir.

Dentro do âmbito de qualidade de dados, *profiling* é tipicamente realizado como parte de um processo de limpeza de dados, revelando erros como formatos inconsistentes de dados ou valores ausentes em uma coluna. Perfis de dados também podem ser utilizados para medir e monitorar a qualidade geral de um conjunto de dados, e,g,, determinando a quantidade de registros que violam restrições estabelecidas previamente (NAUMANN, 2014). No caso de organizações que relutam em investir no gerenciamento da qualidade de seus dados, o uso de *software* gratuito de *data profiling* é uma maneira de demonstrar a existência de problemas que podem afetar o desempenho da organização, com pouco custo adicional envolvido. Existem diversas opções disponíveis para *download* gratuito: ferramentas de código aberto, versões de *software* proprietário com restrições de utilização, ou ainda produtos disponíveis para avaliação por um determinado período de tempo (BLOOR RESEARCH GROUP, 2014).

Abedjan, Golab e Naumann (2015) citam outros casos de uso para *data profiling*, a saber:

- A maioria dos Sistemas de Gerenciamento de Bancos de Dados (SGBD) coleta estatísticas simples a respeito de tabelas e colunas, como contagens de registros, valores únicos ou não nulos. Essas estatísticas são utilizadas pelo otimizador de consultas do SGBD para estimar o custo de planos de consulta e realizar diversos passos de otimização.
- Durante projetos de desenvolvimento, pesquisa ou administração de bancos de dados, não é incomum que profissionais se deparem com conjuntos de dados desconhecidos e não documentados. Perfis de dados podem ser utilizados para dar suporte à compreensão da estrutura, do conteúdo e da dimensão dessas fontes.

- Como uma extensão do item anterior, frequentemente é necessário integrar conjuntos de dados desconhecidos. Essa necessidade vem sendo ampliada nos últimos anos devido à abundância de dados abertos disponíveis na Web; organizações já enxergam o potencial de crescimento associado à integração desses dados abertos com os seus próprios. *Data profiling* não só fornece informações a respeito de cada conjunto de dados individualmente, mas pode revelar como e quão bem os dados de várias fontes podem ser integrados.
- A partir de uma instância de dados sem nenhum esquema formal associado, o processo de engenharia reversa da base de dados consiste na identificação de seus relacionamentos e atributos, além de metadados semânticos como chaves estrangeiras e cardinalidades. Todas essas informações podem ser obtidas através de *profiling*. O resultado do processo de engenharia reversa pode ser um modelo entidade-relacionamento ou um esquema lógico da base de dados, de modo a facilitar a manutenção, integração e criação de consultas sobre ela.
- A quantidade de informação além das bases de dados relacionais tradicionais tem crescido em volumes jamais vistos antes. Em 2015, estimava-se que mais de 80% do total de dados existentes no mundo eram não-estruturados. "*Big data*", ou grandes volumes de dados de tipos variados e baixa densidade de valores gerados muito rapidamente, não podem ser gerenciados de maneira tradicional; isso aumenta ainda mais a importância de compreender o conteúdo de tais dados antes de utilizá-los em projetos de integração, análise ou mineração de dados, e *data profiling* pode ser uma ferramenta importante nesse processo (CAI; ZHU, 2015).

Diante do exposto, observa-se que a realização de *data profiling* é uma atividade importante e frequente no cotidiano de profissionais e pesquisadores de tecnologia da informação. À medida que crescem o volume e a variedade de dados disponíveis, bem como a quantidade de usuários que os acessam, a criação eficiente e eficaz de perfis de dados torna-se um problema de gerenciamento de dados cada vez mais relevante, tanto na indústria como no ambiente acadêmico.

1.3. Objetivos

Este trabalho visa a realização de um estudo a respeito de *data profiling* e conceitos relacionados. Os conhecimentos adquiridos serão então utilizados na elaboração de uma análise funcional de ferramentas comerciais de *data profiling*. Para essa análise, serão consideradas três ferramentas gratuitas com presença de mercado significativa, selecionadas a partir do Quadrante Mágico de Qualidade de Dados do Gartner Group (2016) e do relatório de Ferramentas "Gratuitas" de *Profiling* do Bloor Research Group (2014).

1.4. Estrutura do Trabalho

Além do atual, este trabalho apresenta mais quatro capítulos. O capítulo 2 apresenta de maneira mais aprofundada os conceitos básicos relacionados a *data profiling,* bem como as operações mais comuns a serem realizadas sobre os dados na geração de um *data profile*. O capítulo 3 introduz as três ferramentas gratuitas de *profiling* a serem analisadas neste trabalho, enquanto o capítulo 4 mostra a análise funcional propriamente dita e seus resultados. O capítulo 5 apresenta as considerações finais e possíveis trabalhos futuros.

2. Geração de Perfil de Dados (*Data Profiling*)

Um perfil de dados, ou *data profile*, é um conjunto de informações técnicas a respeito dos dados em uma determinada fonte. Tais informações são chamadas metadados, ou "dados a respeito de dados". A geração de perfil de dados, ou *data profiling*, é o conjunto de atividades e processos realizados para determinar esses metadados, compondo uma síntese da qualidade, integridade e consistência dos dados de um sistema (SINGH; SINGH, 2010; ABEDJAN; GOLAB; NAUMANN, 2015).

2.1. Informação intrínseca e extrínseca

Schomm (2016) considera que perfis de dados são compostos por dois tipos de informação: intrínseca e extrínseca. Informação intrínseca são metadados que podem ser obtidos diretamente a partir de operações, chamadas "tarefas de *profiling*" (NAUMANN, 2014), realizadas sobre os dados, e.g., contagem de valores distintos em uma coluna. Tal informação é inerente aos valores presentes na fonte de dados e não requer nenhum conhecimento externo a seu respeito.

Informações extrínsecas dizem respeito a características dos dados que não podem ser obtidas diretamente a partir de seus valores; é necessário buscar esses metadados em outras fontes, como os criadores e/ou proprietários dos dados ou a documentação existente. No entanto, a obtenção de informação a partir dessas fontes pode ser difícil; o criador ou proprietário dos dados pode estar inacessível ou não se lembrar mais dos dados, enquanto que a documentação disponível pode ser incompleta, incorreta, desatualizada ou simplesmente mal elaborada (DOAN; HALEVY, 2005).

No contexto do presente trabalho, consideraremos como *data profiling* apenas o processo de extração de informação intrínseca.

2.2. *Data profiling* e *data mining*

Mineração de dados, ou *data mining*, é o processo de obtenção de padrões relevantes e conhecimento a partir de grandes volumes de dados (HAN; KAMBER; PEI, 2012). A princípio, não há uma distinção universalmente bem-aceita entre os processos de mineração e geração de perfil de dados; no entanto, alguns autores têm

definido ambos os conceitos de maneiras diferentes. Rahm e Do (2000) consideram que *data profiling* concentra-se na análise de instâncias de atributos individuais, enquanto *data mining* ajuda a encontrar padrões específicos de informações em grandes conjuntos de dados, e.g., relacionamentos entre vários atributos. Em suma, *profiling* seria a análise de colunas individuais, enquanto análises envolvendo múltiplos atributos seriam consideradas mineração.

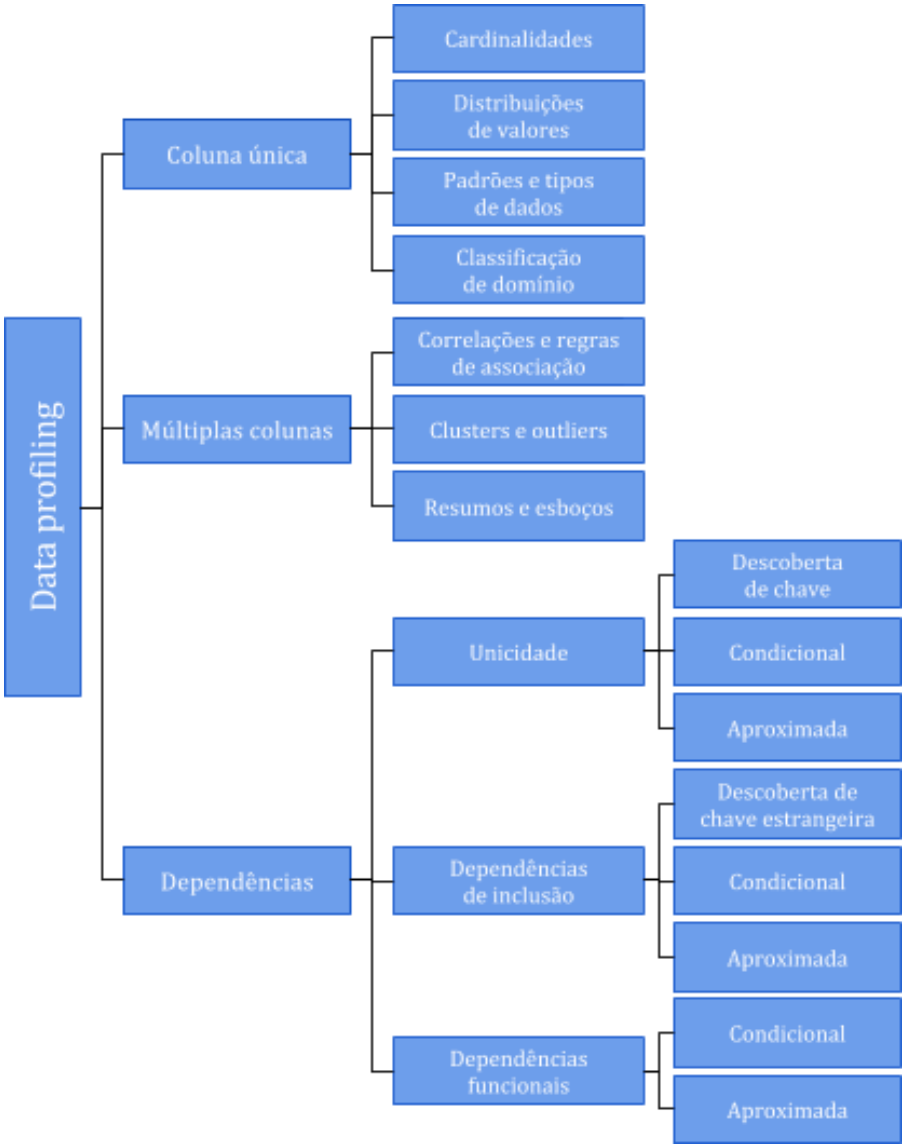
No entanto, Naumann (2014) oferece um critério de distinção diferente: enquanto *data profiling* trata da coleta de metadados técnicos para oferecer suporte ao gerenciamento de dados, *data mining* e *data analytics* objetivam a descoberta de resultados não óbvios para oferecer suporte ao gerenciamento de negócios. Em suma, um perfil de dados fornece informações a respeito de um conjunto específico de dados; tais informações são aplicáveis de maneira confiável apenas à instância de dados sobre a qual o perfil foi gerado. Já o processo de mineração gera *insights* aplicáveis a outras instâncias de dados de mesmo domínio, de modo a apoiar o processo de tomada de decisão a respeito de um negócio.

Essa distinção pelo objetivo da tarefa é mais flexível e considera que algumas tarefas tradicionalmente utilizadas em mineração de dados, como *clustering*, também podem ser utilizadas para descobrir informações a respeito de uma instância específica de dados, que é o propósito da criação de um *data profile* (ABEDJAN; GOLAB; NAUMANN, 2015).

2.3. Tarefas de *profiling*

Esta seção descreve as tarefas mais frequentes realizadas para extração de informação intrínseca de um banco de dados relacional. Para propósitos de organização, optou-se por analisar as tarefas de acordo com a classificação de tarefas de *profiling* proposta por Abedjan, Golab e Naumann (2015). Esta classificação divide as tarefas de acordo com o tipo de metadado que pode ser obtido, conforme a Figura 2.1.

Figura 2.1: classificação de tarefas tradicionais de *data profiling*.



Fonte: ABEDJAN; GOLAB; NAUMANN, 2015 (adaptado)

2.3.1. Perfil de coluna única

Como o nome sugere, as tarefas descritas nesta subseção coletam metadados a respeito de cada coluna em uma determinada tabela de maneira individual, sem considerar seu relacionamento com as demais. Este é o tipo mais simples de *profiling*.

Iniciaremos a descrição destas tarefas com a definição de cardinalidade. Cardinalidade é o número total de valores em um domínio (ELMASRI; NAVATHE, 2011). No contexto deste trabalho, estenderemos esta definição; a análise de **cardinalidades** engloba contagens diversas sobre os valores presentes em uma coluna (ABEDJAN; GOLAB; NAUMANN, 2015). Tais contagens incluem:

- O total de registros na coluna, que equivale à quantidade de linhas na tabela;
- O total de valores distintos na coluna;
- O total de valores nulos na coluna;
- A distribuição dos comprimentos em caracteres dos valores de uma coluna, isto é, comprimento máximo e mínimo e média de comprimento; e
- A seletividade da coluna, isto é, a razão entre a quantidade de valores distintos e o total de registros na coluna.

A análise de **distribuição de valores** pode ser considerada como uma extensão da análise de cardinalidades (ABEDJAN; GOLAB; NAUMANN, 2015), gerando uma série de descrições estatísticas adicionais a respeito dos dados em uma coluna. Tais descrições fornecem uma visão geral dos valores, sua frequência e a maneira como estão distribuídos (HAN; KAMBER; PEI, 2012). Serão considerados aqui:

- Os valores máximo e mínimo presentes na coluna;
- Medidas de tendência central: média, mediana e moda dos valores da coluna;
- Frequência dos valores da coluna, tanto individualmente como em grupos (histogramas);
- A constância da coluna, ou seja, a razão entre a frequência absoluta do valor mais frequente e o total de valores na coluna; e
- A verificação da lei de Benford (1938). A lei de Benford determina que, em um conjunto de valores numéricos ocorridos naturalmente, a distribuição de frequência do primeiro dígito d é aproximadamente $P(d) = \log_{10}(1 + {}^{1}/_{d})$; de modo que o dígito 1 tem cerca de 30% de chance de aparecer como o primeiro dígito, enquanto o dígito 9 tem menos de 5%. Essa verificação é útil para descobrir a existência de números forjados.

A análise de **padrões** fornece descrições do formato dos dados presentes na coluna analisada. Tais descrições podem ser feitas de várias maneiras; e.g., um número de telefone pode ser representado pela máscara "(dd) ddddd dddd" (em que

a letra “d” representa um dígito numérico), ou pela expressão regular “\([0-9]{2}\) [0-9]{5} [0-9]{4}”. Há duas análises possíveis a serem realizadas sobre os dados de uma coluna aqui: a detecção dos padrões mais frequentes encontrados e a verificação quanto à conformidade a um ou mais padrões determinados pelo usuário. Já tarefas que analisam **tipos de dados** consistem na extração de metadados sintáticos a respeito dos valores de uma coluna. A tarefa mais básica aqui é a detecção do tipo genérico de dados presentes na coluna, e.g., numérico, alfanumérico ou data. Uma análise mais sofisticada pode retornar o tipo concreto e específico do SGBD utilizado como fonte ou a ser utilizado como destino de dados, e.g., INT, VARCHAR ou TIMESTAMP. No caso de tipos numéricos, também é possível detectar o número máximo de dígitos e/ou casas decimais (ABEDJAN; GOLAB; NAUMANN, 2015).

Além das informações sintáticas obtidas nas tarefas do parágrafo anterior, é também interessante conhecer o significado dos valores em uma coluna. Tarefas de **classificação de domínio** objetivam obter informações semânticas a respeito dos valores de uma coluna. Está inclusa aqui a detecção da chamada “classe de dados”, isto é, um tipo de dados semântico e genérico, e.g., texto, código ou quantidade. Além disso, pode-se realizar a identificação do domínio concreto da coluna, ou seja, a entidade que o campo representa no mundo real, e.g., cidade, cartão de crédito ou e-mail (ABEDJAN; GOLAB; NAUMANN, 2015).

2.3.2. Perfil de múltiplas colunas

Esta subseção descreve análises realizadas sobre múltiplas colunas simultaneamente. Vale lembrar que, de acordo com a definição de *data mining* fornecida por Rahm e Do (2000) vista na seção 2.2., as tarefas descritas aqui seriam consideradas tarefas de mineração, não de criação de perfil.

A análise de **correlação** revela relacionamentos não-explícitos entre colunas numéricas; e.g., em uma tabela de “Funcionários”, os valores dos campos “salário” e “tempo de contrato” podem estar relacionados. Já regras de **associação** denotam relações entre valores contidos nas colunas; especificamente, valores de atributos diferentes que tendem a ocorrer juntos no mesmo registro (ABEDJAN; GOLAB; NAUMANN, 2015). A geração de regras de associação ocorre em dois passos: a descoberta de conjuntos de valores frequentes, ou seja, cuja frequência no conjunto

de dados ultrapassa um valor arbitrário; e, a partir destes, a geração de regras do tipo $A \rightarrow B$, onde B é um conjunto de valores frequente e a razão entre as frequências de A e B excede um determinado patamar (HAN; KAMBER; PEI, 2012).

Clustering é a tarefa de dividir dados em grupos (*clusters*) de objetos semelhantes (BERKHIN, 2006). Durante a análise de múltiplas colunas, é possível simplificar o modelo de dados e separar registros em grupos homogêneos utilizando algoritmos de *clustering*. Registros que não se encaixam em nenhum dos *clusters* determinados inicialmente podem ser considerados ***outliers***, isto é, objetos que não adequam-se ao comportamento geral dos dados (HAN; KAMBER; PEI, 2012). A existência de *outliers* pode indicar problemas de qualidade de dados, tornando a sua identificação uma tarefa de *profiling* interessante (ABEDJAN; GOLAB; NAUMANN, 2015).

Uma alternativa ao *clustering* é a criação de **resumos e esboços**, cujo objetivo é encontrar uma descrição compacta de um conjunto de dados. Tais descrições são úteis em situações com alto fluxo de dados de entrada, como por exemplo *streaming* (CHANDOLA; KUMAR, 2005; CORMODE et al., 2012). Além da criação de resumos, está incluso aqui também o cálculo do coeficiente de similaridade de Jaccard entre duas colunas A e B, $J(A, B)$, que é a razão entre o número de valores distintos que ambas têm em comum e o número total de valores distintos que aparecem em ambas; isto é, $J(A, B) = |(A \cap B)|/|(A \cup B)|$ (ABEDJAN; GOLAB; NAUMANN, 2015).

2.3.3. Dependências

Dependências são metadados que descrevem relacionamentos entre colunas (NAUMANN, 2014). Esta subseção abrange tarefas de detecção e verificação dos tipos de dependências descritos a seguir. A rigor, tais tarefas fariam parte da análise de múltiplas colunas; no entanto, a complexidade e a importância da identificação de dependências é tamanha que Abedjan, Golab e Naumann (2015) optaram por analisá-las separadamente.

Iniciaremos pela descrição do que chamaremos dependências de **unicidade** (DU). DUs descrevem combinações únicas de colunas, isto é, grupos de atributos cujos valores conjuntos são únicos para cada registro em uma tabela. Tais combinações podem ser utilizadas como chaves primárias da relação. A descoberta

de todas as DUs em uma relação é um problema NP-difícil, sendo um desafio na área obter bom desempenho na realização desta análise (ABEDJAN; QUIANÉ-RUIZ; NAUMANN, 2014).

Uma dependência de **inclusão** (DI) $A \subseteq B$, onde A e B são atributos, significa que todos os valores do atributo *dependente A* estão contidos no conjunto de valores do atributo *referenciado B*. Embora DIs sejam pré-requisitos para o estabelecimento de restrições de integridade referencial e herança entre relações, e portanto um passo importante na descoberta de chaves estrangeiras, sua existência apenas não é suficiente para determinar a corretude semântica desses relacionamentos (BAUCKMANN et al., 2007; ELMASRI; NAVATHE, 2011).

Uma dependência **funcional** (DF) descreve uma restrição semântica sobre os dados. Formalmente, uma DF é uma expressão do tipo $A \rightarrow B$, onde A e B são conjuntos de atributos, indicando que os valores de A determinam os valores de B; ou seja, quaisquer duas tuplas da relação que possuam o mesmo valor de A necessariamente possuem o mesmo valor de B. No contexto de *profiling*, interessa descobrir DFs não-triviais ($A \cap B = \emptyset$) e mínimas (não existe nenhuma DF $A \rightarrow C$ para qualquer $C \subset B$) (ELMASRI; NAVATHE, 2011; PAPENBROCK et al., 2015).

Antes de definir os dois últimos tipos de dependência analisados nesta subseção, faz-se necessário introduzir o conceito de dependência parcial. Dependência parcial é uma dependência que só é válida para um subconjunto dos dados; e.g., uma dependência que vale para 95% dos registros, com os 5% restantes fugindo à regra. Tendo sido detectada uma dependência parcial, seja ela DU, DI ou DF, é interessante saber se há uma condição que caracterize os registros para os quais ela ocorre. Dependência **condicional** é a combinação do predicado que descreve a dependência parcial com a descrição da sua condição de ocorrência (ABEDJAN; GOLAB; NAUMANN, 2015).

Finalmente, dependências **aproximadas** são dependências como as já definidas anteriormente, mas detectadas a partir de uma amostra ou resumo dos dados da relação analisada. Tais dependências podem ser consideradas como parciais para a relação completa (ABEDJAN; GOLAB; NAUMANN, 2015).

2.4. Considerações finais

Nesse capítulo, realizamos um estudo mais aprofundado a respeito do conceito de *data profiling* introduzido no capítulo 0, bem como outros conceitos adicionais importantes. Identificamos também aqui as operações sobre os dados, ou tarefas, mais comuns de *profiling*. Essas tarefas serão utilizadas posteriormente na elaboração de um estudo comparativo de ferramentas de *data profiling* em relação às funcionalidades disponibilizadas por cada uma para o usuário.

No próximo capítulo serão apresentadas as ferramentas analisadas neste trabalho.

3. Ferramentas Escolhidas

Neste capítulo serão apresentadas as ferramentas selecionadas para análise. Dois critérios foram levados em consideração nessa escolha: primeiramente, a presença de mercado das empresas fornecedoras destas ferramentas, como indicativo da sua relevância em ambientes comerciais. O segundo critério é a gratuidade das ferramentas, possibilitando a realização de *profiling* com pouco ou nenhum custo adicional para os clientes finais. A seção 3.1 detalha o processo de seleção, enquanto as seções posteriores introduzem brevemente cada uma das ferramentas escolhidas.

3.1. Processo de seleção

Para a avaliação da presença de mercado das ferramentas, levou-se em consideração o relatório técnico "Magic Quadrant for Data Quality Tools" (GARTNER GROUP, 2016); para o critério de gratuidade, o relatório "'Free' Data Profiling Tools" (BLOOR RESEARCH GROUP, 2014). Os dois relatórios serão descritos nas próximas subseções. Para análise neste trabalho, foram selecionadas as ferramentas fornecidas por empresas presentes em ambos.

3.1.1. Magic Quadrant for Data Quality Tools (GARTNER GROUP, 2016)

O Gartner Group disponibiliza anualmente relatórios que avaliam a presença de mercado das principais empresas fornecedoras de soluções de TI em diversas disciplinas. Tais relatórios, denominados "Quadrantes Mágicos", posicionam as empresas em um plano cartesiano, cujos eixos são capacidade de execução ("*ability to execute*") e completude de visão ("*completeness of vision*").

O critério de capacidade de execução avalia a competência das empresas em gerarem efeitos positivos sobre seu lucro, retenção de clientes e reputação, de modo a manterem-se competitivas no mercado. Para tanto, considera-se não apenas a qualidade do produto/serviço oferecido pela empresa, mas também fatores como execução de vendas e marketing, satisfação dos consumidores, responsividade ao mercado e viabilidade financeira da fabricante. Já a completude de visão mede a compreensão das empresas sobre como fatores de mercado podem ser explorados

na criação de oportunidades, avaliando critérios como inovação, modelo de negócio, compreensão do mercado e estratégia de vendas, marketing e oferta do produto.

Os quadrantes do plano cartesiano são, portanto, definidos a seguir:

- *Leaders*: capacidade de execução e completude de visão acima da origem. As empresas deste quadrante demonstram forte entendimento do mercado e alto grau de inovação, além de fornecerem ferramentas de alta qualidade direcionadas a domínios e casos de usos diversos. Estas empresas são consideradas grandes, bem estabelecidas no mercado e com presença multinacional.
- *Challengers*: capacidade de execução acima da origem, mas completude de visão abaixo da origem. Empresas neste quadrante possuem alta credibilidade e oferecem produtos de excelente qualidade, mas não demonstram o mesmo grau de inovação das empresas do quadrante de *leaders*.
- *Niche Players:* capacidade de execução e completude de visão abaixo da origem. As ferramentas oferecidas por empresas deste quadrante podem ter funcionalidades limitadas ou direcionadas a domínios, áreas geográficas ou segmentos de mercado específicos, resultando em uma presença de mercado menor.
- *Visionaries:* completude de visão acima da origem, mas capacidade de execução abaixo da origem. As empresas deste quadrante demonstram forte compreensão das tendências de mercado, mas possuem menor quantidade de usuários e/ou recursos do que grandes fornecedores.

Para o presente trabalho, foi considerado especificamente o quadrante mágico de ferramentas de qualidade de dados mais recente (GARTNER GROUP, 2016), exibido na Figura 3.1. Este relatório trata da disciplina de qualidade de dados como um todo, da qual *data profiling* é apenas um dos requisitos funcionais centrais. Os demais incluem: limpeza e padronização de dados, isto é, modificação e formatação de valores de modo a adequar-se a padrões, restrições ou regras de negócio; monitoramento contínuo para garantir que os dados continuem adequados ao longo do tempo; identificação de correspondências entre registros de um ou mais conjuntos de dados; e enriquecimento do valor de dados internos através do relacionamento com fontes externas de dados, como descritores geográficos.

Figura 3.1: Quadrante Mágico para Ferramentas de Qualidade de Dados.

Fonte: GARTNER GROUP, 2016

3.1.2. “Free” Data Profiling Tools (BLOOR RESEARCH GROUP, 2014)

O Bloor Research Group possui um relatório de tendências e posição atual de mercado, intitulado *Market Update*, específico para ferramentas de *data profiling* (BLOOR RESEARCH GROUP, 2013).

O *Market Update* atribui a cada empresa considerada uma posição em um diagrama *Bullseye*, composto de um conjunto de círculos concêntricos. A figura resultante é segmentada em três áreas, correspondentes a três categorias: *Champion*, *Innovator* e *Challenger*. As empresas que obtiveram maior pontuação geral estão mais próximas do centro. O analista define então uma pontuação de referência para uma empresa líder de segmento a partir de sua pontuação geral; as empresas que obtiverem pontuações acima desta referência ficam no segmento *Champion* do diagrama. As demais empresas são atribuídas à categoria *Innovator* se sua pontuação de inovação excede 2.5 e *Challenger* caso contrário. A posição exata em cada segmento é calculada com base na combinação de pontuações geral e de

inovação. Não há, no entanto, documentação acessível que defina claramente os critérios utilizados pelos autores para determinar estas pontuações.

Em 2014, os autores propuseram uma visão especializada do seu *Market Update* mais atual para *data profiling*, composta exclusivamente por empresas que oferecem soluções "gratuitas" (BLOOR RESEARCH GROUP, 2014). Estão inclusos aqui produtos de código aberto e ferramentas proprietárias, embora algumas destas últimas possuam restrições quanto a tempo de utilização ou quantidade de usuários, por exemplo. A Figura 3.2 mostra o diagrama *Bullseye* correspondente a este relatório.

Figura 3.2: diagrama *Bullseye* de ferramentas "gratuitas" de *data profiling*.

Fonte: BLOOR RESEARCH GROUP, 2014

Três das empresas presentes nele (Ataccama, Experian e Talend) estão presentes no *Market Update* original e obtiveram suas posições a partir dele, enquanto as demais foram adicionadas especificamente para o novo relatório. Essas três empresas também estão presentes no Quadrante Mágico de Qualidade de Dados do grupo Gartner para o ano de 2016, conforme visto na subseção 3.1.1. Por esse motivo, analisaremos neste trabalho as soluções gratuitas de *profiling* fornecidas por elas. Tais ferramentas serão apresentadas nas próximas seções.

3.2. Ataccama Data Quality Analyzer (DQ Analyzer)[1]

Posição da fabricante no Quadrante Mágico do Gartner Group: *Visionary*

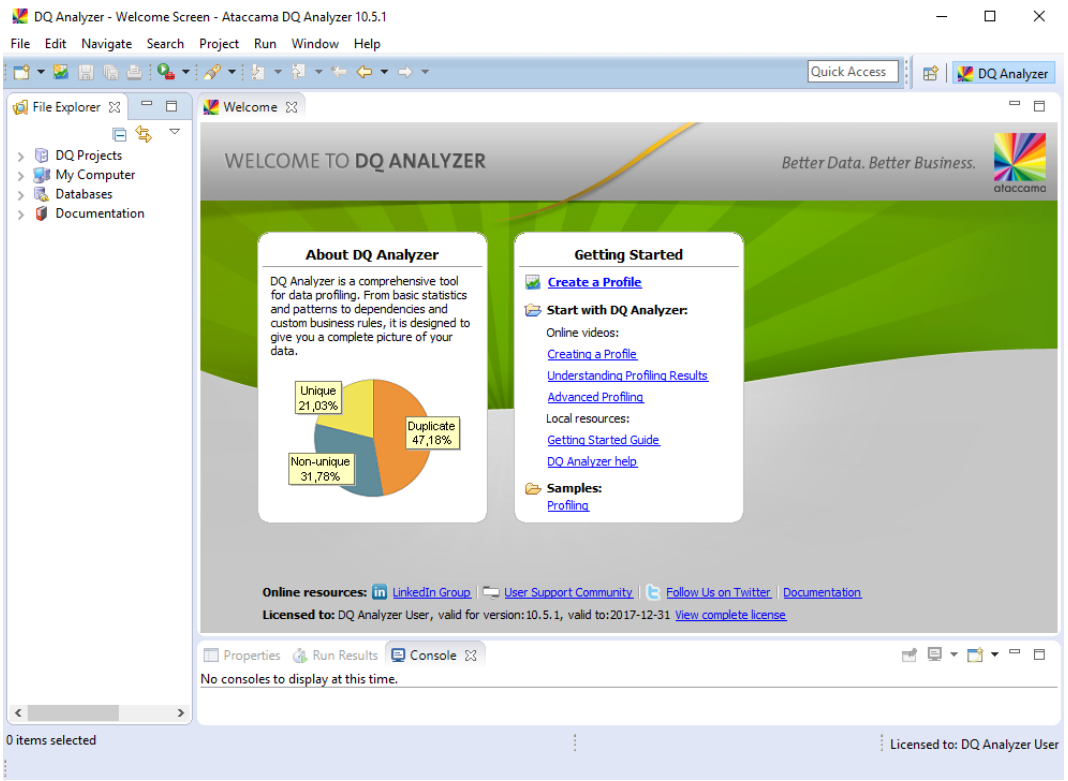
Posição da fabricante no diagrama do Bloor Research Group: *Challenger*

O DQ Analyzer é uma ferramenta de desenvolvimento proprietária construída sobre o framework Eclipse. Análises feitas utilizando o DQ Analyzer são inteiramente compatíveis com a solução completa (e paga) da Ataccama para qualidade de dados, o Data Quality Center; isto é, o usuário do DQ Analyzer pode importar seus resultados de *profiling* obtidos gratuitamente para o Data Quality Center e realizar operações diversas de correção, limpeza e padronização de dados.

O DQ Analyzer suporta nativamente conexões a planilhas do Microsoft Excel (arquivos no formato *.xls) e arquivos de texto delimitados ou de largura fixa (*.txt ou *.csv). Conexões a bancos de dados são realizadas através da API *Java Database Connectivity* (JDBC); sendo assim, é possível conectar a qualquer SGBD que possua um driver de conectividade JDBC.

A versão analisada da ferramenta foi a 10.5.1.

Figura 3.3: tela inicial do DQ Analyzer.



Fonte: A autora

[1] Disponível em: https://www.ataccama.com/products/data-discovery-and-profiling/dqa

3.3. Experian Pandora Free Data Profiler (Pandora Free)[2]

Posição da fabricante no Quadrante Mágico do Gartner Group (2016): *Challenger*

Posição da fabricante no diagrama do Bloor Research Group (2014): *Innovator*

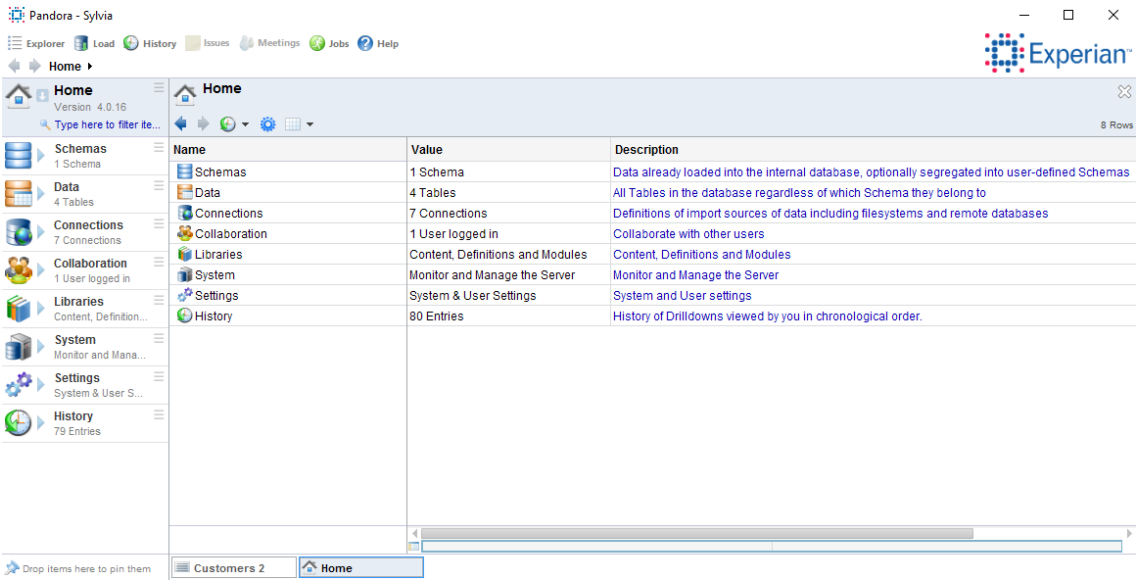
Pandora é a solução de qualidade de dados da Experian, direcionada para migração, gerenciamento de qualidade e *profiling* de dados. A empresa disponibiliza gratuitamente uma versão limitada do Pandora para uso pessoal, apenas com as funcionalidades de *profiling* e restrições de quantidade de tabelas e registros.

O Pandora possui uma política de licenciamento mais rigorosa, sendo necessário entrar em contato por e-mail com a Experian fornecendo o endereço MAC da máquina onde foi feita a instalação para obter uma licença (mesmo para a versão gratuita) antes de utilizar o produto. Tanto o Pandora como sua versão gratuita são o mesmo produto com as funcionalidades controladas por licença; dessa forma, os mesmos repositórios criados com o Pandora Free podem ser utilizados com a versão paga e totalmente funcional do *software*.

Assim como o DQ Analyzer, o Pandora Free dá suporte a importação de dados de arquivos de texto delimitados e planilhas, bem como conexão a bancos de dados através de drivers JDBC.

A versão analisada da ferramenta foi a 4.0.16.

Figura 3.4: tela inicial do Pandora Free.



Fonte: A autora

[2] Disponível em: https://www.edq.com/uk/solutions/experian-pandora/data-profiling/free-data-profiler/

3.4. Talend Open Studio for Data Quality (Open Studio)[3]

Posição da fabricante no Quadrante Mágico do Gartner Group (2016): *Visionary*

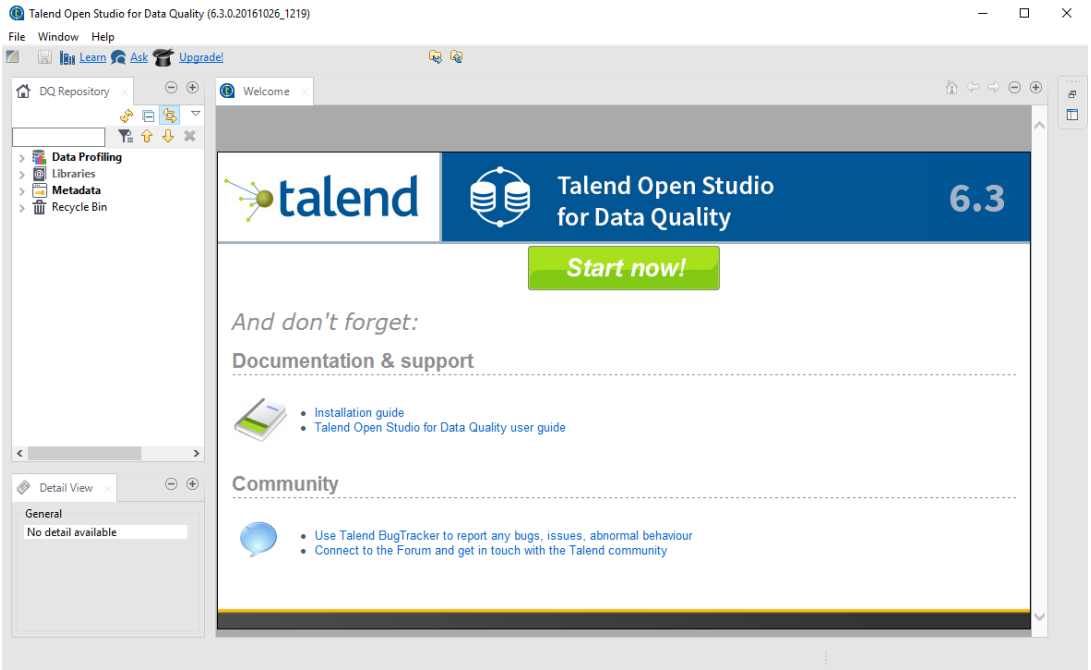
Posição da fabricante no diagrama do Bloor Research Group (2014): *Innovator*

Open Studio é a suíte de produtos gratuitos da Talend, composta por soluções voltadas para gerenciamento, integração, preparação e qualidade de dados. Para cada produto do Open Studio, a Talend possui um produto pago análogo, direcionado a empresas e com comodidades adicionais (e.g., suporte técnico e ferramentas administrativas). Neste trabalho, utilizaremos o termo "Open Studio" para nos referirmos ao Open Studio for Data Quality. O Open Studio é a única alternativa de código aberto analisada neste trabalho, sendo fornecido sob a versão 2.0 da licença Apache. Assim como o DQ Analyzer, o Open Studio foi construído sobre o framework Eclipse.

O Open Studio, de forma semelhante às duas ferramentas anteriores, possibilita conexões a bancos de dados através de drivers JDBC. A importação de arquivos de texto delimitados também é possível, mas não há suporte para importação de dados de planilhas do Excel.

A versão analisada da ferramenta foi a 6.3.0.

Figura 3.5: tela inicial do Open Studio.



Fonte: A autora

[3] Disponível em: https://www.talend.com/download/talend-open-studio/#t2

3.5. Considerações finais

Nesse capítulo, descrevemos os critérios e o processo de escolha das soluções gratuitas de *data profiling* a serem analisadas no presente trabalho. Adicionalmente, apresentamos brevemente cada uma das três ferramentas selecionadas.

No capítulo a seguir, será realizada a análise funcional das ferramentas introduzidas aqui, a partir das tarefas de *data profiling* introduzidas no capítulo 2.

4. Análise Funcional

Este capítulo contém uma análise das funcionalidades presentes em cada ferramenta de *data profiling* apresentada no capítulo 3, de acordo com as tarefas de extração de informações intrínsecas descritas na seção 2.3. Para testes, foi utilizada a base de dados de amostra Northwind em um ambiente com Microsoft SQL Server 2016 Developer Edition.

Cada tarefa considerada será dividida em *k* subtarefas, também definidas de acordo com a seção 2.3. A pontuação de cada ferramenta para uma funcionalidade será calculada por $^{n}/_{k} \times 100$, sendo *n* a quantidade de subtarefas da funcionalidade que a ferramenta é capaz de realizar. Ao final do capítulo, será feita a média aritmética das pontuações obtidas em cada funcionalidade para gerar a pontuação total da ferramenta.

Optou-se pela análise meramente funcional das ferramentas devido à inexistência, até o momento da realização deste trabalho, de um teste de referência (*benchmark*) voltado para ferramentas de *profiling* que leve em consideração outros critérios relevantes, como desempenho e completude da análise.

4.1. Cardinalidades

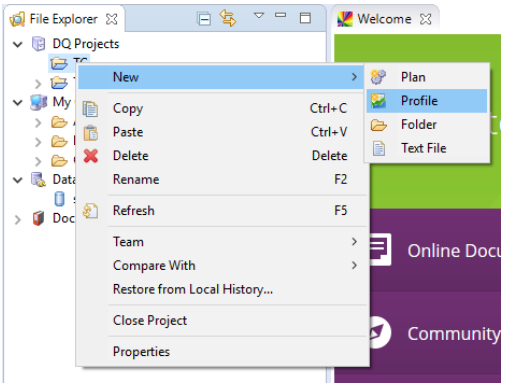
Esta funcionalidade é composta de contagens sobre os dados de uma coluna.

Contagens fazem parte do *profile* básico do DQ Analyzer, que pode ser criado a partir da opção "New" > "Profile" no menu de contexto aberto a partir de um projeto. Na janela de configuração que se abre, a opção "Standard analysis" deve ser selecionada. Nessa mesma janela, é possível escolher colunas específicas da tabela utilizada como entrada, habilitar *drill-through* (isto é, a visualização dos registros da tabela, no caso de um banco de dados) e criar um plano de *profile* para possibilitar análises mais avançadas, como pode ser visto na Figura 4.1.

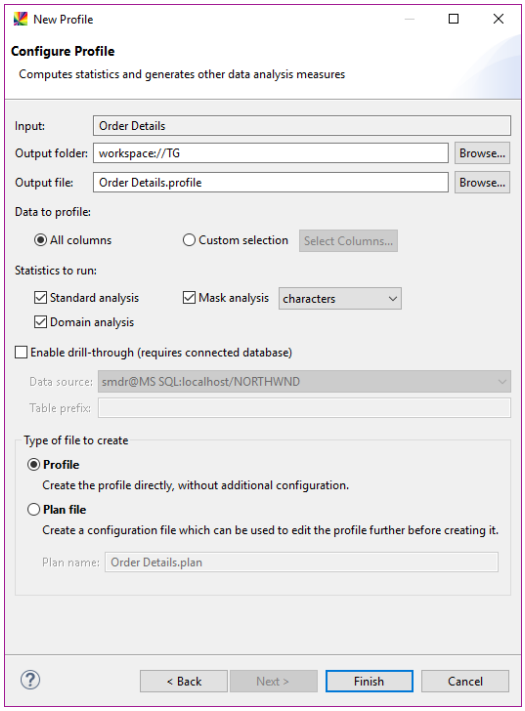
A Figura 4.2 mostra o resultado da análise básica do DQ Analyzer para uma coluna alfanumérica. Além do total de registros na coluna, é exibido, em tabela e gráfico de setores, o percentual de valores nulos e não-nulos. Valores não-nulos são por sua vez divididos em valores distintos (cujo percentual corresponde à seletividade da coluna) e duplicados. Os valores distintos são ainda separados em

únicos e não-únicos. A análise básica de colunas alfanuméricas também inclui uma tabela com a distribuição de comprimentos em caracteres dos valores. Logo, o DQ Analyzer recebe a pontuação de todas as subtarefas desta seção.

Figura 4.1: criação de novo *profile* no DQ Analyzer. (a) Menu de contexto; (b) Janela de configuração



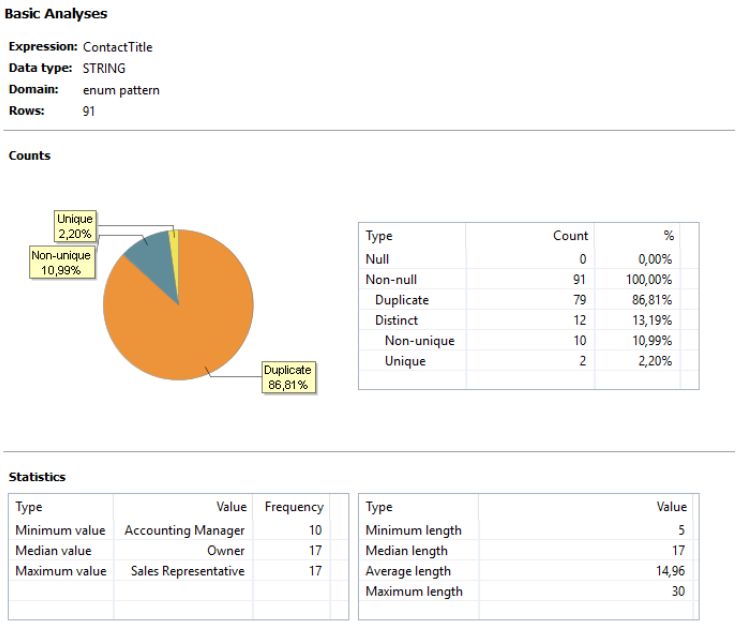
(a)



(b)

Fonte: A autora

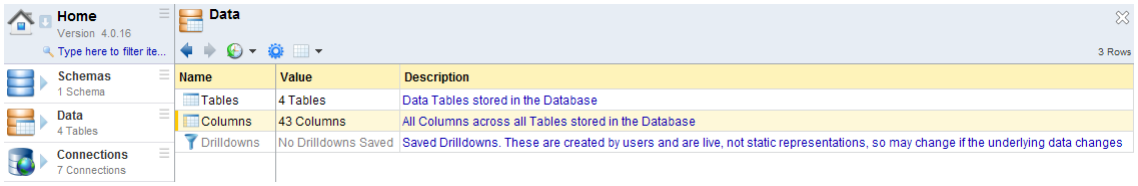
Figura 4.2: resultado de análise básica de colunas no DQ Analyzer (coluna alfanumérica).



Fonte: A autora

No Pandora Free, todas as informações relativas aos dados das tabelas de entrada podem ser visualizadas a partir da seção “Data”, a ser selecionada no menu do lado esquerdo da tela. Uma análise preliminar do conteúdo de todas as colunas das tabelas de entrada, inclusive contagens, é realizada já a partir da importação dos dados, não sendo necessário nenhum passo adicional para isso. Para visualizar os resultados desta análise, basta um duplo clique no link “Columns” da tabela “Data”, conforme a Figura 4.3. Também é possível selecionar apenas as colunas de uma tabela específica a partir do link “Tables”.

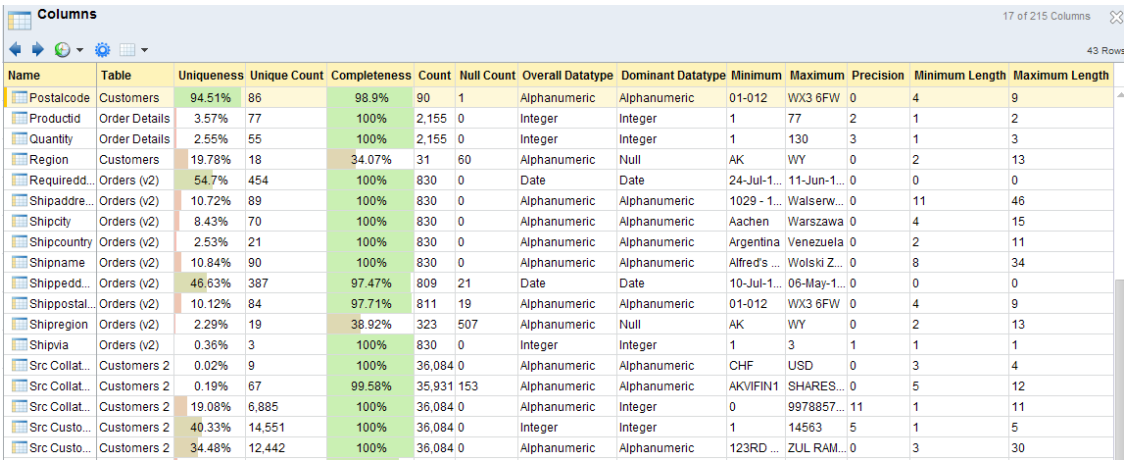
Figura 4.3: seção "Data" do Pandora Free.



Fonte: A autora

O resultado da análise preliminar de colunas pode ser visto na Figura 4.4. É possível visualizar o percentual de unicidade (seletividade) da coluna, a contagem de valores distintos e a contagem de valores nulos e não-nulos (cuja soma é o total de registros da coluna). No caso de colunas numéricas e alfanuméricas, pode-se visualizar também o comprimento mínimo e máximo em caracteres. A média de comprimento não é vista na análise preliminar, mas está inclusa no *profile* da coluna desejada.

Figura 4.4: resultado de análise preliminar de colunas no Pandora Free.

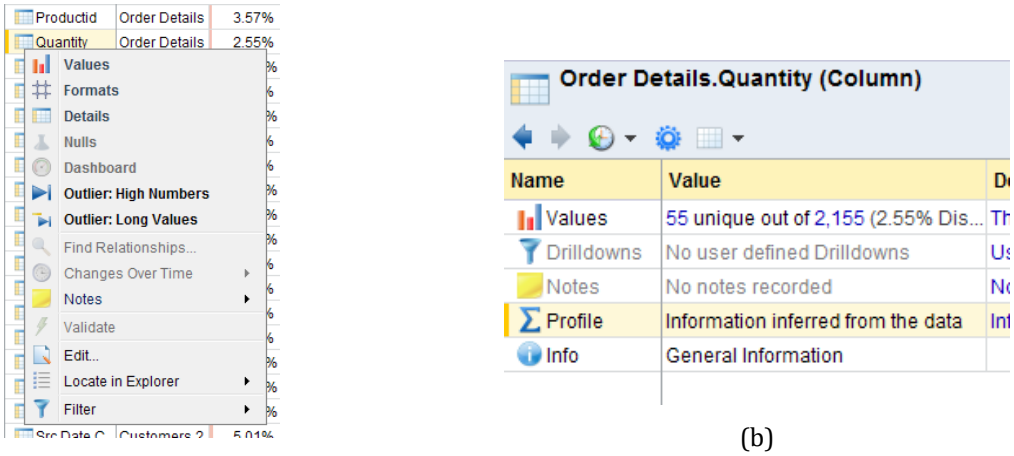


Name	Table	Uniqueness	Unique Count	Completeness	Count	Null Count	Overall Datatype	Dominant Datatype	Minimum	Maximum	Precision	Minimum Length	Maximum Length
Postalcode	Customers	94.51%	86	98.9%	90	1	Alphanumeric	Alphanumeric	01-012	WX3 6FW	0	4	9
Productid	Order Details	3.57%	77	100%	2,155	0	Integer	Integer	1	77	2	1	2
Quantity	Order Details	2.55%	55	100%	2,155	0	Integer	Integer	1	130	3	1	3
Region	Customers	19.78%	18	34.07%	31	60	Alphanumeric	Null	AK	WY	0	2	13
Requiredd...	Orders (v2)	54.7%	454	100%	830	0	Date	Date	24-Jul-1...	11-Jun-1...	0	0	0
Shipaddre...	Orders (v2)	10.72%	89	100%	830	0	Alphanumeric	Alphanumeric	1029 - 1...	Walserw...	0	11	46
Shipcity	Orders (v2)	8.43%	70	100%	830	0	Alphanumeric	Alphanumeric	Aachen	Warszawa	0	4	15
Shipcountry	Orders (v2)	2.53%	21	100%	830	0	Alphanumeric	Alphanumeric	Argentina	Venezuela	0	2	11
Shipname	Orders (v2)	10.84%	90	100%	830	0	Alphanumeric	Alphanumeric	Alfred's ...	Wolski Z...	0	8	34
Shippedd...	Orders (v2)	46.63%	387	97.47%	809	21	Date	Date	10-Jul-1...	06-May-1...	0	0	0
Shippostal...	Orders (v2)	10.12%	84	97.71%	811	19	Alphanumeric	Alphanumeric	01-012	WX3 6FW	0	4	9
Shipregion	Orders (v2)	2.29%	19	38.92%	323	507	Alphanumeric	Null	AK	WY	0	2	13
Shipvia	Orders (v2)	0.36%	3	100%	830	0	Integer	Integer	1	3	1	1	1
Src Collat...	Customers 2	0.02%	9	100%	36,084	0	Alphanumeric	Alphanumeric	CHF	USD	0	3	4
Src Collat...	Customers 2	0.19%	67	99.58%	35,931	153	Alphanumeric	Alphanumeric	AKVIFIN1	SHARES...	0	5	12
Src Collat...	Customers 2	19.08%	6,885	100%	36,084	0	Alphanumeric	Integer	0	9978857...	11	1	11
Src Custo...	Customers 2	40.33%	14,551	100%	36,084	0	Integer	Integer	1	14563	5	1	5
Src Custo...	Customers 2	34.48%	12,442	100%	36,084	0	Alphanumeric	Alphanumeric	123RD ...	ZUL RAM...	0	3	30

Fonte: A autora

Para visualizar o *profile* da coluna, basta selecionar “Details” no menu de contexto que se abre a partir do nome da coluna; em seguida, clicar no link “Profile” da tabela que se abre, conforme a Figura 4.5. O perfil gerado pode ser visto na Figura 4.6, e contém, entre outras informações, a média de comprimento da coluna. Portanto, o Pandora Free recebe a pontuação de todas as subtarefas desta seção.

Figura 4.5: criação de *profile* de coluna no Pandora Free. (a) Menu de contexto; (b) Tabela de detalhes de coluna



(a) (b)

Fonte: A autora

Figura 4.6: perfil de coluna do Pandora Free.

Profile for Orders (v2).Freight (Column)

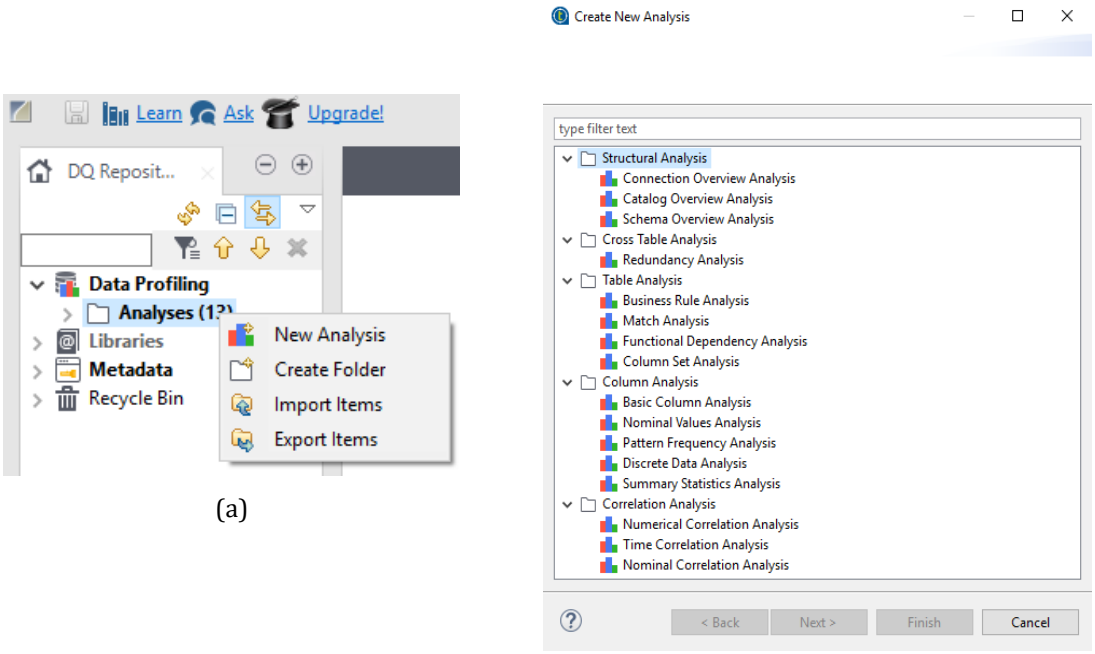
26 Rows

Name	Value	Description
Datatype	Decimal	The inferred Datatype based on the actual values within this Column
Dominant Datatype	Decimal	The most dominant Datatype for values in this Column
Nulls	0 (0% Distribution)	The number of rows in the Table where this Column has a missing value (null)
Unique Values	799 unique out of 830 (96.27% Dist...	The number of unique values in this Column and the distribution as a percentage of the number of rows in the Table for this column
Integers	6 unique out of 6 (0.72% Distribution)	The number of unique Integer values in this Column out of the total number of rows in the Table that are Integer and the percentage ...
Decimals	793 unique out of 824 (95.54% Dist...	The number of unique Decimal values in this Column out of the total number of rows in the Table that are Decimal and the percenta...
Minimum Value	0.02 (1 Occurrence)	The minimum value for this Column and the number of times it occurs
Maximum Value	1007.64 (1 Occurrence)	The maximum value for this Column and the number of times it occurs
Least Frequent Value	7 (1 Occurrence)	The value that occurs the least often in this Column and the number of times it occurs. The value shown is the first instance of a valu...
Most Frequent Value	0.56 (2 Occurrences)	The value that occurs the most often in this Column and the number of times it occurs. The value shown is the first instance of a valu...
Unique Formats	10 Formats	The number of unique formats in the Column
Least Frequent Format	9 (1 Occurrence)	The format that occurs least often in this Column and the number of times it occurs. The format shown is the first instance of a forma...
Most Frequent Format	99.99 (427 Occurrences)	The format that occurs most often in this Column and the number of times it occurs. The format shown is the first instance of a forma...
Precision	6 Digits	The maximum precision of all Integer, Decimal or Money values in this Column. Precision is the total number of digits to the left and r...
Scale	2 Digits	The maximum scale of all Integer, Decimal or Money values in this Column. Scale is the total number of digits to the right of the deci...
Shortest Length	1 Character	The shortest length of all values in this Column
Longest Length	7 Characters	The longest length of all values in this Column
Average Length	4 Characters	The average length of all values in this Column
Spaces	No Spaces (0%)	The number of rows in the Table where this Column has a missing Value (spaces)
Negative Values	No Negative Values (0%)	The number of unique negative values in this Column
Zeroes	No Zero Values (0%)	The number of rows in the Table where this Column has a Value that is Zero
Total Sum	64,942.69	The sum of all Integer, Decimal or Money values in the Column
Average	78.24	The average of all Integer, Decimal or Money values in the Column
Value Deviation	116.78	The Standard Deviation of all Integer, Decimal or Money values in the Column
Checksum	9ce9663d	Checksum of values calculated form the data
Encoding Errors	No Errors defined	Values that failed character set translation resulting in errors. This points to an invalid character set or corrupt data.

Fonte: A autora

Por fim, o Open Studio disponibiliza as funcionalidades desta seção dentro de sua análise de colunas, que pode ser selecionada na janela de criação de nova análise, conforme Figura 4.7.

Figura 4.7: criação de nova análise no Open Studio. (a) Menu de contexto; (b) Janela de nova análise



(a)

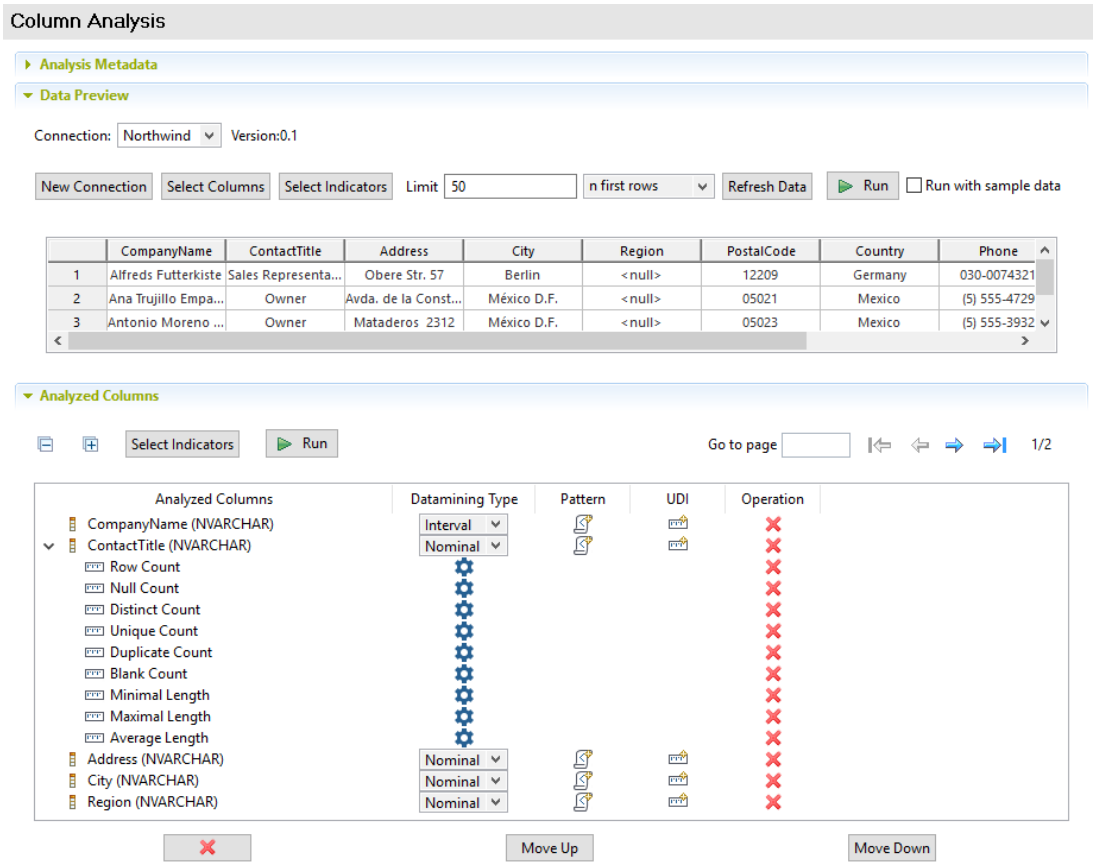
(b)

Fonte: A autora

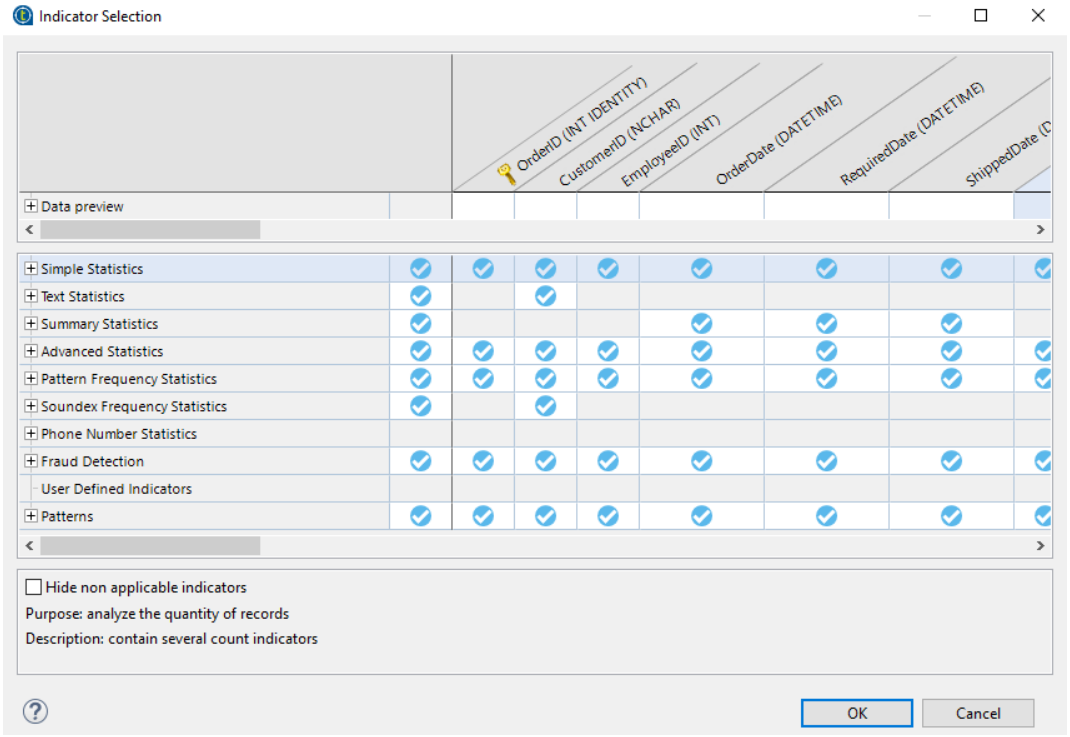
Todas as opções dentro de “Column Analysis” abrem a análise de colunas vista na Figura 4.8, onde é possível escolher as colunas a serem analisadas individualmente e as subtarefas a serem realizadas sobre elas (chamadas aqui de indicadores). A diferença é que a opção “Basic Column Analysis” não predefine subtarefas, enquanto as demais opções trazem alguns indicadores já selecionados de acordo com o tipo de análise. Contagens de linhas, valores distintos/duplicados e valores nulos/em branco são obtidas na janela de seleção de indicadores sob a categoria de “Simple Statistics”, enquanto a análise de distribuição de comprimento em caracteres está presente em “Text Statistics”.

A apresentação dos resultados das análises do Open Studio é feita através de tabelas e gráficos de coluna, como pode ser visto na Figura 4.9. A seletividade da coluna equivale ao percentual relativo à contagem de valores distintos. Sendo assim, a ferramenta recebe a pontuação relativa a todas as subtarefas da seção.

Figura 4.8: análise de colunas no Open Studio. (a) Janela de configuração da análise; (b) Janela de seleção de indicadores



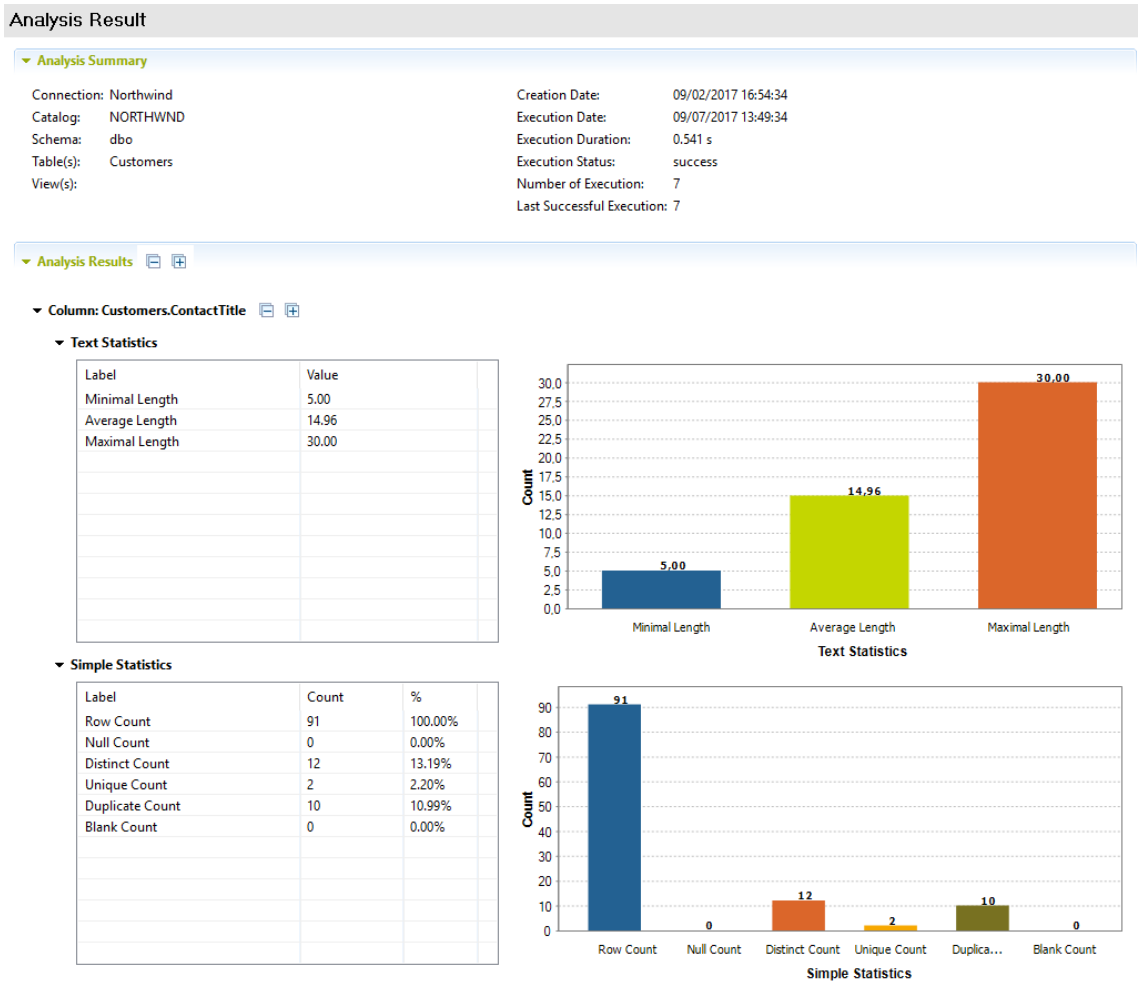
(a)



(b)

Fonte: A autora

Figura 4.9: resultado de análise de colunas do Open Studio para os indicadores "Text Statistics" e "Simple Statistics".



Fonte: A autora

A Tabela 4.1 exibe a pontuação obtida por cada ferramenta nesta seção.

Tabela 4.1: resultado da análise de ferramentas segundo o critério “Cardinalidades”.

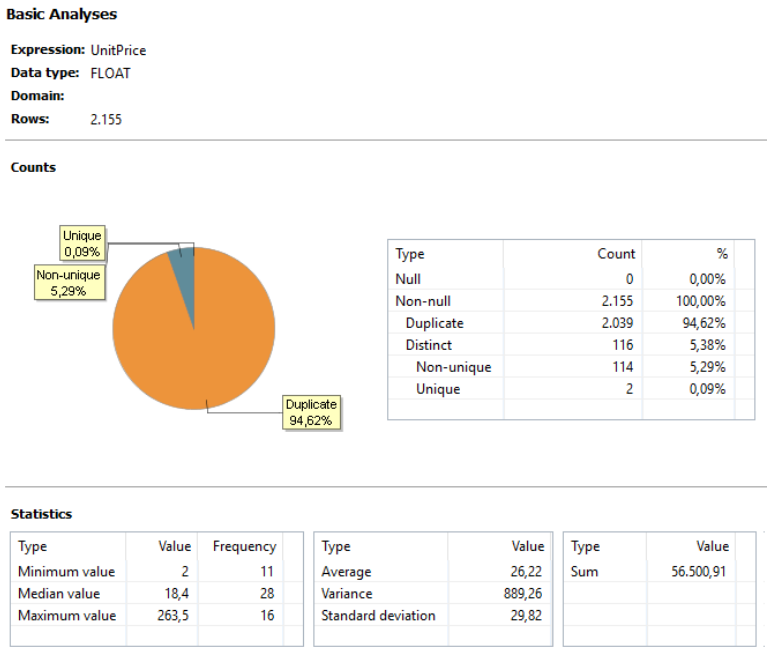
Subtarefa	DQ Analyzer	Pandora Free	Open Studio
Contagem de linhas	1	1	1
Contagem de valores distintos	1	1	1
Contagem de valores nulos	1	1	1
Distribuição de comprimento	1	1	1
Seletividade	1	1	1
Pontuação total (n/5 x 100)	**100**	**100**	**100**

4.2. Distribuição de valores

Esta funcionalidade trata de estatísticas mais avançadas a respeito dos valores de uma coluna.

O DQ Analyzer exibe valores mínimos e máximos da coluna dentro da análise básica já vista na Figura 4.2, assim como a medida de tendência central mediana. No caso de colunas numéricas, também são calculadas média, variância e desvio padrão, como pode ser visto na Figura 4.10. A opção "Standard Analysis" da janela de criação de novo *profile* também gera uma análise de frequência, vista na Figura 4.11, que exibe em uma tabela a frequência absoluta e relativa de todos os valores na coluna.

Figura 4.10: resultado de análise básica de colunas no DQ Analyzer (coluna numérica).



Basic Analyses

Expression: UnitPrice
Data type: FLOAT
Domain:
Rows: 2.155

Counts

Type	Count	%
Null	0	0,00%
Non-null	2.155	100,00%
Duplicate	2.039	94,62%
Distinct	116	5,38%
Non-unique	114	5,29%
Unique	2	0,09%

Statistics

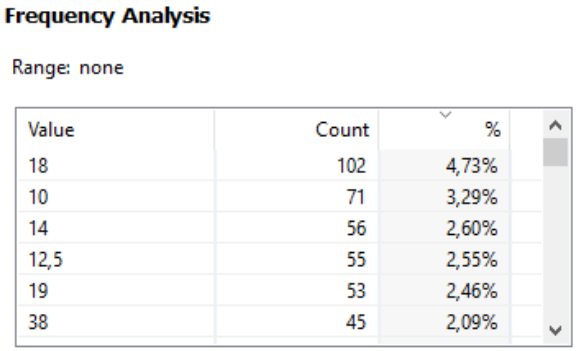
Type	Value	Frequency
Minimum value	2	11
Median value	18,4	28
Maximum value	263,5	16

Type	Value
Average	26,22
Variance	889,26
Standard deviation	29,82

Type	Value
Sum	56.500,91

Fonte: A autora

Figura 4.11: resultado de análise de frequência no DQ Analyzer.



Frequency Analysis

Range: none

Value	Count	%
18	102	4,73%
10	71	3,29%
14	56	2,60%
12,5	55	2,55%
19	53	2,46%
38	45	2,09%

Fonte: A autora

O Pandora Free inclui em seu *profile* de colunas (já visto na Figura 4.6) valores máximo e mínimo e medidas de tendência central (média e desvio padrão) da coluna analisada, além dos valores mais e menos frequentes; para uma análise mais detalhada da frequência, no entanto, é possível selecionar a opção “Values” no menu de contexto ou na tabela de detalhes de coluna da Figura 4.5. A tabela resultante pode ser vista na Figura 4.12, e contém a frequência absoluta e percentual de cada valor na coluna analisada.

Figura 4.12: análise de valores do Pandora Free.

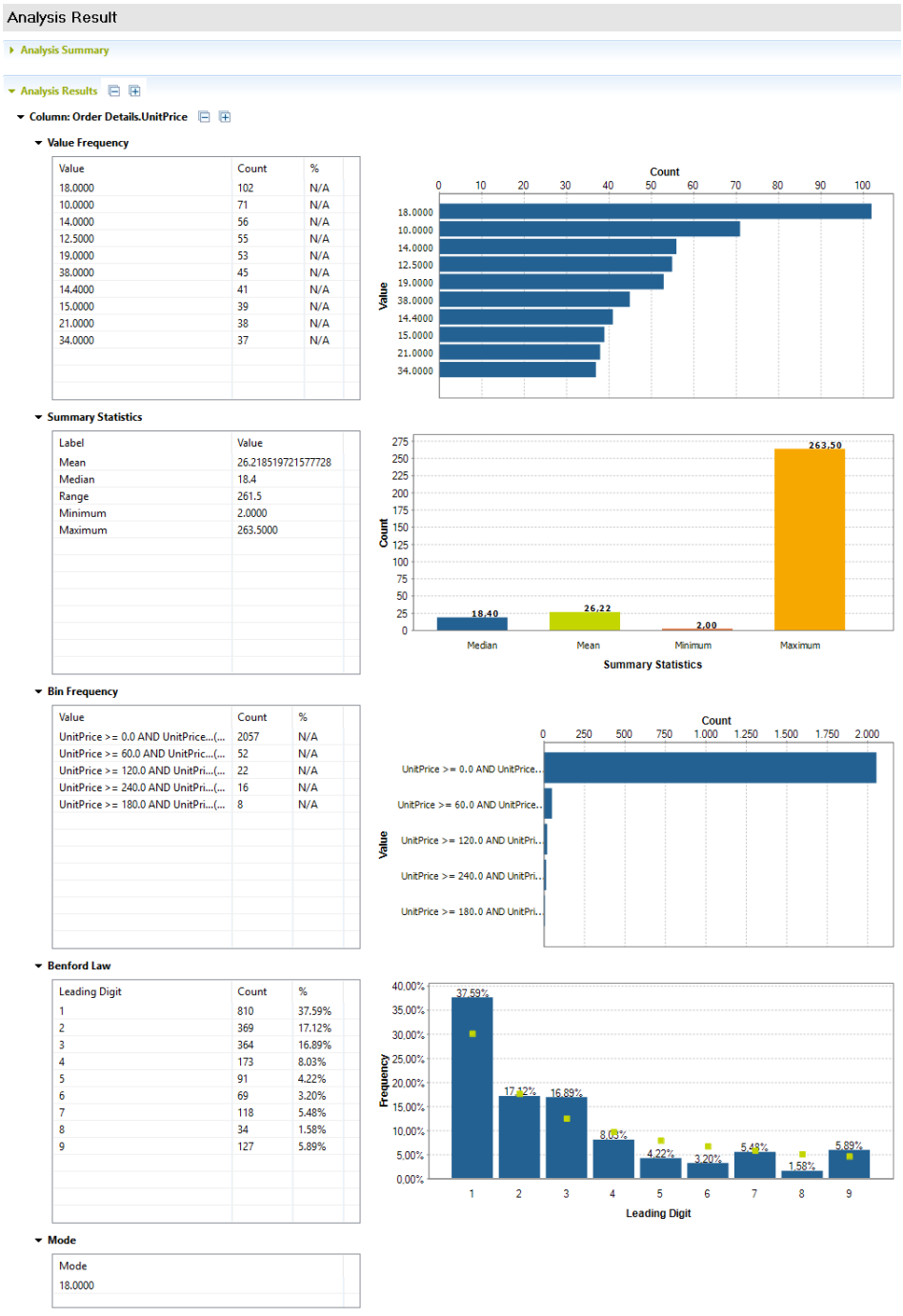
Values for Orders (v2).Shipcountry

Value	Rows	Distribution % ▼1	Value Length	Datatype	Format	Format Rows	Simple Format
Germany	122	14.7%	7	Alphanumeric	AAAAAAA	240	A
USA	122	14.7%	3	Alphanumeric	AAA	122	A
Brazil	83	10%	6	Alphanumeric	AAAAAA	268	A
France	77	9.28%	6	Alphanumeric	AAAAAA	268	A
UK	56	6.75%	2	Alphanumeric	AA	56	A
Venezuela	46	5.54%	9	Alphanumeric	AAAAAAAAA	62	A
Austria	40	4.82%	7	Alphanumeric	AAAAAAA	240	A
Sweden	37	4.46%	6	Alphanumeric	AAAAAA	268	A
Canada	30	3.61%	6	Alphanumeric	AAAAAA	268	A
Italy	28	3.37%	5	Alphanumeric	AAAAA	51	A
Mexico	28	3.37%	6	Alphanumeric	AAAAAA	268	A
Spain	23	2.77%	5	Alphanumeric	AAAAA	51	A
Finland	22	2.65%	7	Alphanumeric	AAAAAAA	240	A
Belgium	19	2.29%	7	Alphanumeric	AAAAAAA	240	A
Ireland	19	2.29%	7	Alphanumeric	AAAAAAA	240	A
Denmark	18	2.17%	7	Alphanumeric	AAAAAAA	240	A
Switzerland	18	2.17%	11	Alphanumeric	AAAAAAAAAAA	18	A
Argentina	16	1.93%	9	Alphanumeric	AAAAAAAAA	62	A
Portugal	13	1.57%	8	Alphanumeric	AAAAAAAA	13	A
Poland	7	0.84%	6	Alphanumeric	AAAAAA	268	A
Norway	6	0.72%	6	Alphanumeric	AAAAAA	268	A

Fonte: A autora

O Open Studio fornece as funcionalidades desta seção dentro de sua análise de colunas (já vista na Figura 4.8). O indicador “Summary Statistics” contém valores mínimo e máximo e duas medidas de tendência central, média e mediana. A moda pode ser obtida no indicador “Advanced Statistics”, bem como a distribuição de valores e a criação de histogramas Por fim, o indicador “Fraud Detection” realiza a verificação da Lei de Benford. O resultado da análise para estes indicadores está na Figura 4.13.

Figura 4.13: resultado de análise de colunas do Open Studio para os indicadores "Summary Statistics", “Advanced Statistics” e "Fraud Detection".



Fonte: A autora

Nem o DQ Analyzer nem o Pandora Free realizam verificação da Lei de Benford. Além disso, nenhuma das ferramentas analisadas possui um indicador explícito de constância da coluna; seria necessário realizar o cálculo manualmente a partir da frequência do valor mais comum e do total de valores na coluna.

A Tabela 4.2 contém os resultados da análise desta seção.

Tabela 4.2: resultado da análise de ferramentas segundo o critério “Distribuição de valores”.

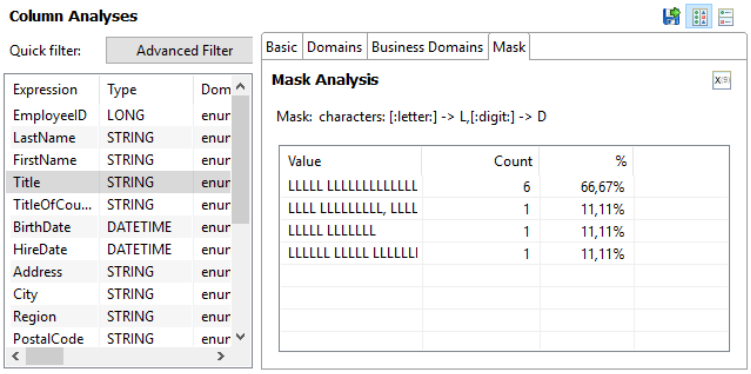
Subtarefa	DQ Analyzer	Pandora Free	Open Studio
Valor máximo e mínimo	1	1	1
Medidas de tendência central	1	1	1
Frequência de valores	1	1	1
Constância	0	0	0
Verificação da Lei de Benford	0	0	1
Pontuação total (n/5 x 100)	**60**	**60**	**80**

4.3. Padrões e tipos de dados

As funcionalidades desta seção referem-se à verificação e detecção de padrões e tipos de dados de uma coluna.

Para realizar a detecção de padrões no DQ Analyzer, a opção “Mask Analysis” precisa estar selecionada na janela de criação de novo *profile* (Figura 4.1). Há algumas opções pré-definidas de possíveis máscaras a serem aplicadas sobre os dados da coluna. Adicionalmente, a detecção de tipo de dados genérico (e.g., LONG, STRING, DATETIME) também é realizada automaticamente aqui. Ambos os resultados podem ser vistos na Figura 4.14.

Figura 4.14: resultado de análise de máscara do DQ Analyzer.

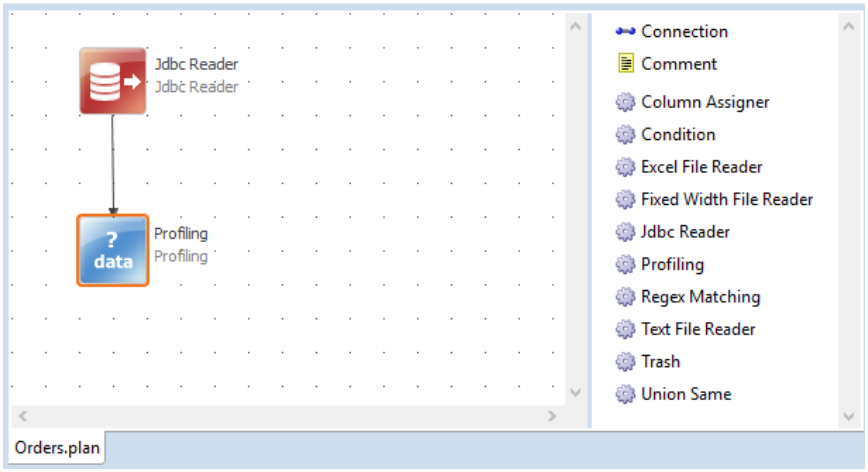


Fonte: A autora

Adicionalmente, é possível verificar padrões através de expressões regulares, através da criação de um plano de *profile*. Planos de *profile* possibilitam a criação de análises avançadas a partir de diversos operadores de dados, e podem ser criados selecionando a opção “New” > “Plan” no menu de contexto da Figura 4.1. A Figura

4.15 mostra um plano de *profile* simples, com apenas um operador de importação de dados e um de *profiling*.

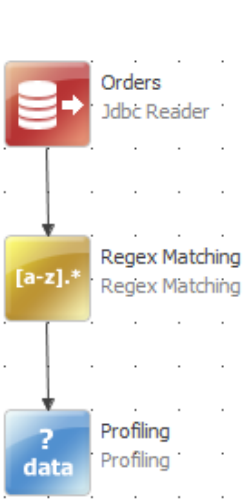
Figura 4.15: plano de profile do DQ Analyzer.



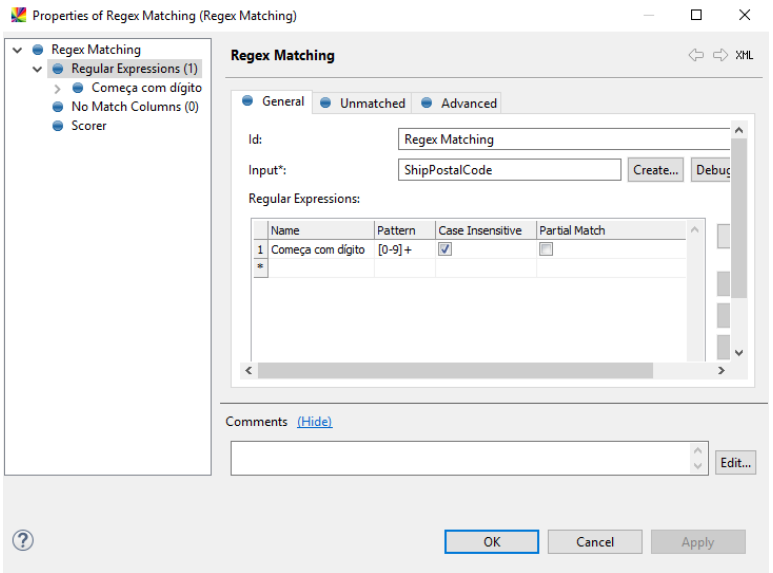
Fonte: A autora

Para a verificação de padrões, o DQ Analyzer disponibiliza um operador de casamento de expressões regulares. Este operador pode ser adicionado ao plano de *profile* e configurado com uma ou mais expressões regulares correspondentes aos padrões a serem verificados, conforme Figura 4.16.

Figura 4.16: análise de expressões regulares no DQ Analyzer. (a) Plano de profile; (b) Janela de configuração do operador "Regex Matching"



(a)

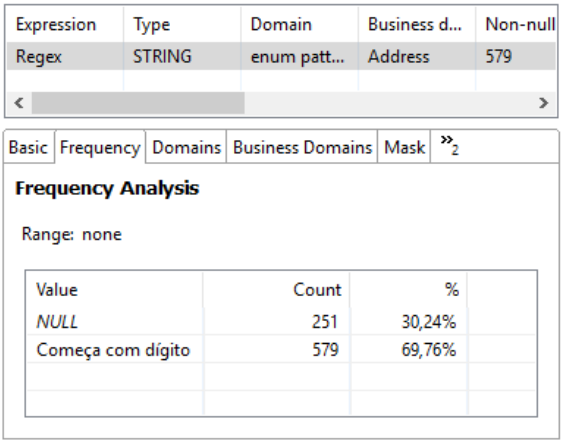


(b)

Fonte: A autora

O resultado desta análise é idêntico ao das análises de colunas vistas na Figura 4.2 e na Figura 4.11, mas com uma coluna calculada adicional que exibe o valor da coluna como o resultado da expressão regular. A análise de frequência desta coluna, vista na Figura 4.17, indica a quantidade de valores que podem ser descritos pela expressão regular.

Figura 4.17: resultado de análise de frequência para coluna de expressão regular no DQ Analyzer.



Fonte: A autora

No Pandora Free, o perfil de colunas (Figura 4.6) traz as informações de precisão e escala de um valor numérico, equivalentes respectivamente ao total de dígitos e de decimais, sendo a única ferramenta analisada que traz esta informação. Adicionalmente, a análise de valores (Figura 4.12) traz o tipo genérico de cada coluna. A opção “Formats” no menu de contexto da análise de colunas traz estas informações consolidadas em uma única visão, como pode ser visto na Figura 4.18.

Figura 4.18: análise de formatos do Pandora Free.

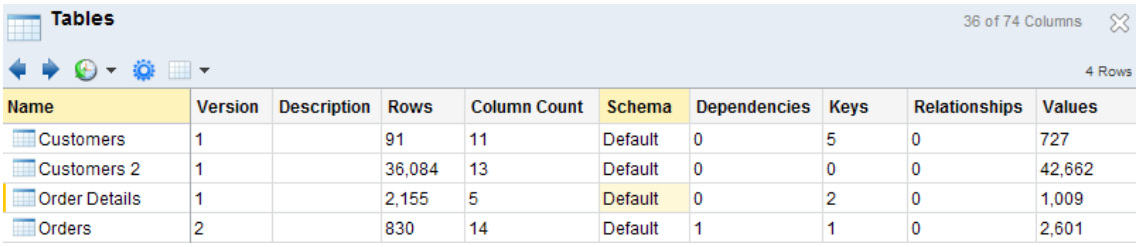
Σ Formats for Customers.Fax — 10 of 29 Columns — 20 Groups

Format	First Value	Simple Format	Values	Total Rows	Distribution %	Format Length	Format Type
(9)S99-99-99	(8) 34-93-93	(9)S9-9-9	1	1	1.1%	12	Alphanumeric
(9)S99.99.99.99	(1) 42.34.22.77	(9)S9.9.9.9	2	2	2.2%	15	Alphanumeric
(9)S999-9999	(1) 123-5556	(9)S9-9	10	10	10.99%	12	Alphanumeric
(99)S999-9999	(11) 555-2168	(9)S9-9	4	4	4.4%	13	Alphanumeric
(99)S999S9999	(91) 745 6210	(9)S9S9	2	2	2.2%	13	Alphanumeric
(99)S999S99S99	(02) 201 24 68	(9)S9S9S9	3	3	3.3%	14	Alphanumeric
(999)S999-9999	(171) 555-2530	(9)S9-9	17	17	18.68%	14	Alphanumeric
(999)S99S99S99S99	(071) 23 67 22 21	(9)S9S9S9S9	1	1	1.1%	17	Alphanumeric
99-999S9999	90-224 8858	9-9S9	1	1	1.1%	11	Alphanumeric
99-99S99S99	07-98 92 47	9-9S9S9	1	1	1.1%	11	Alphanumeric
99.99.99.99	20.16.10.17	9.9.9.9	9	9	9.89%	11	Alphanumeric

Fonte: A autora

É possível ainda realizar a verificação de padrões através da criação de colunas personalizadas. O link “Tables” da seção “Data” mostrada na Figura 4.3 abre uma análise preliminar das tabelas carregadas para o Pandora, como mostra a Figura 4.19. Ao clicar com o botão direito sobre uma tabela, é possível visualizar os registros da mesma através da opção “Rows” do menu de contexto. Clicando com o botão direito sobre uma coluna, é possível criar uma coluna personalizada a partir da opção “Insert Column” > “Custom...” do menu de contexto resultante, como pode ser visto na Figura 4.20.

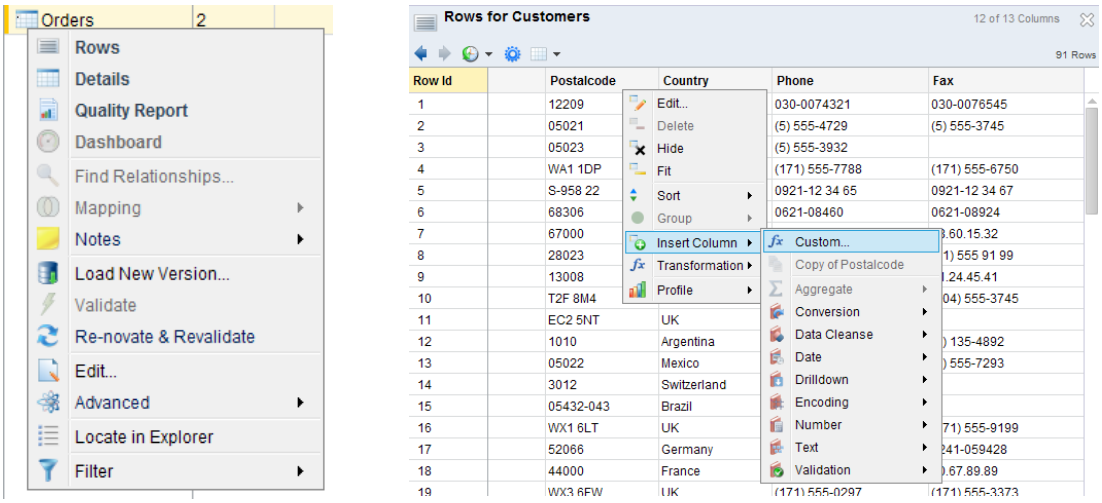
Figura 4.19: resultado de análise preliminar de tabelas no Pandora Free.



Name	Version	Description	Rows	Column Count	Schema	Dependencies	Keys	Relationships	Values
Customers	1		91	11	Default	0	5	0	727
Customers 2	1		36,084	13	Default	0	0	0	42,662
Order Details	1		2,155	5	Default	0	2	0	1,009
Orders	2		830	14	Default	1	1	0	2,601

Fonte: A autora

Figura 4.20: análise de registros de tabela do Pandora Free. (a) Menu de contexto de tabelas; (b) Tabela de registros e menu de contexto de colunas.



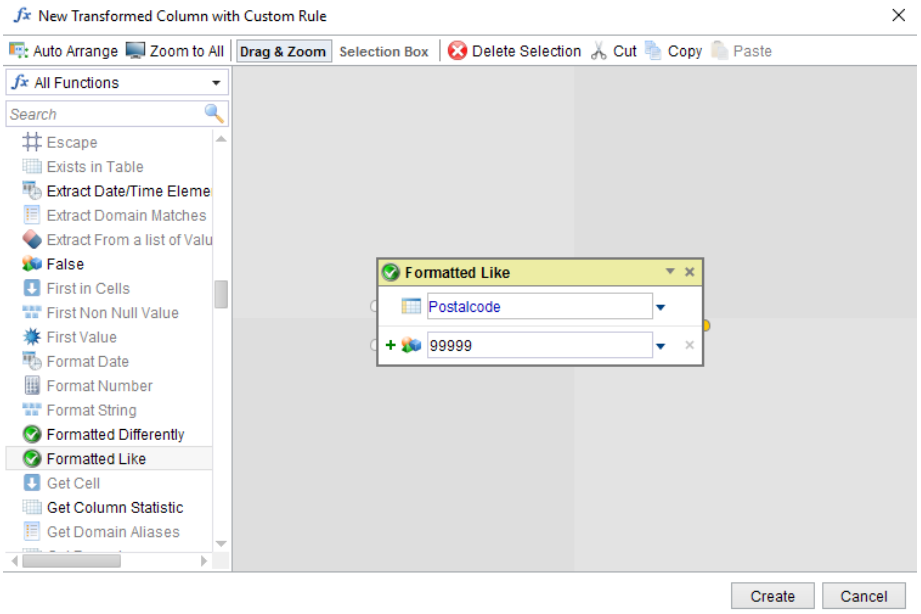
(a) (b)

Fonte: A autora

A janela de criação de coluna personalizada pode ser vista na Figura 4.21. Diversos operadores são disponibilizados aqui para a geração de colunas calculadas; no caso da verificação de padrões, uma possibilidade é a utilização do operador

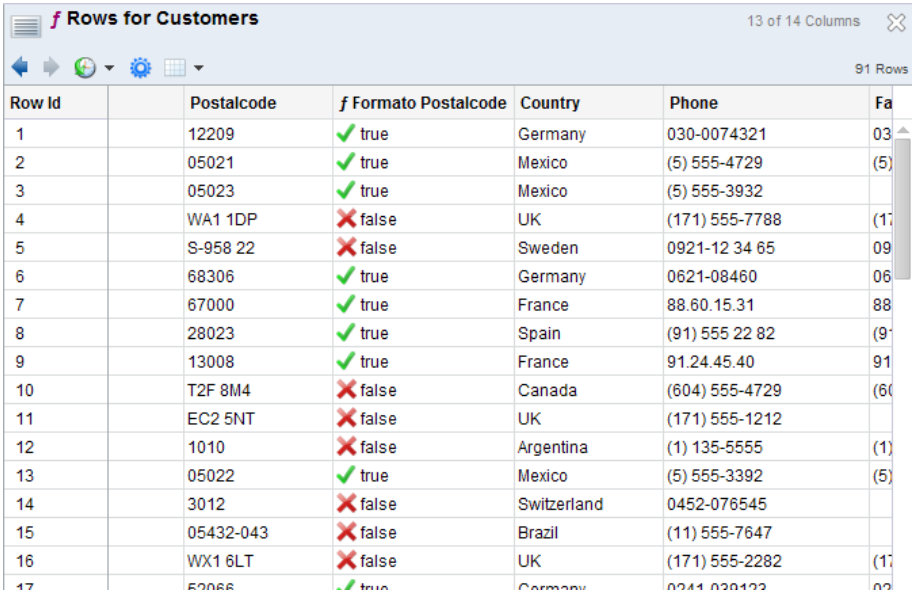
"Formatted Like", que gera uma coluna booleana informando se a célula da coluna analisada se encaixa em um padrão definido pelo usuário. A Figura 4.22 exibe essa nova coluna.

Figura 4.21: janela de criação de coluna personalizada no Pandora Free.



Fonte: A autora

Figura 4.22: análise de registros de tabela do Pandora Free com coluna personalizada.



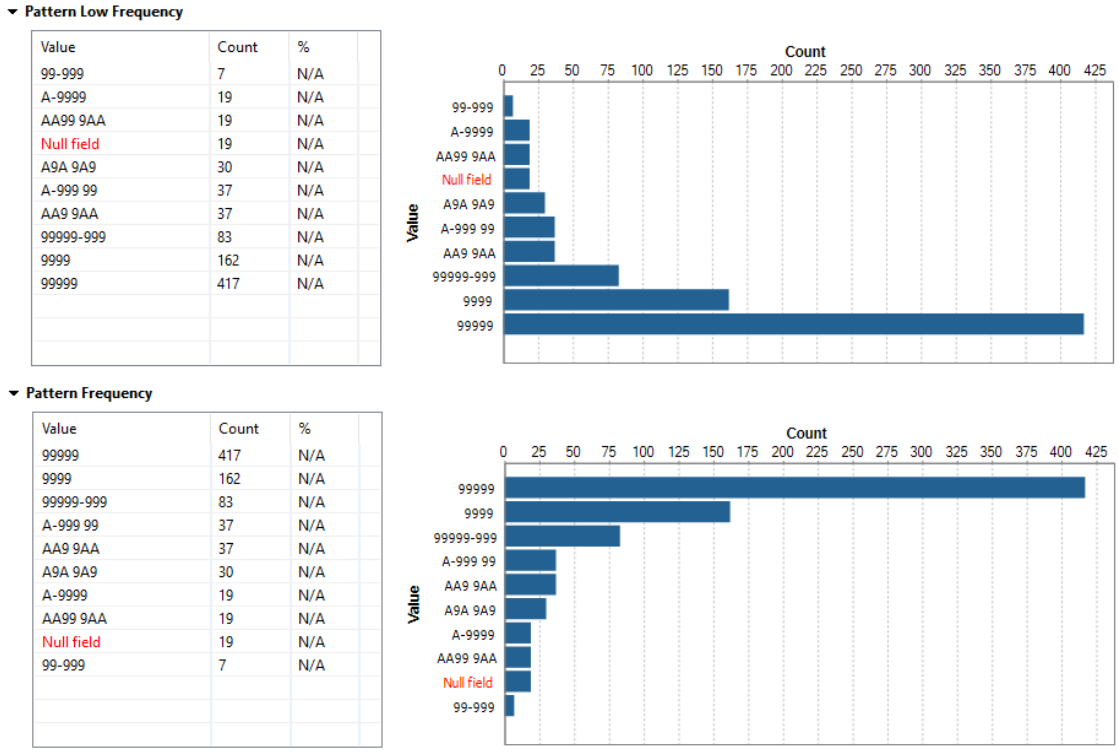
Row Id		Postalcode	ƒ Formato Postalcode	Country	Phone	Fa
1		12209	true	Germany	030-0074321	03
2		05021	true	Mexico	(5) 555-4729	(5)
3		05023	true	Mexico	(5) 555-3932	
4		WA1 1DP	false	UK	(171) 555-7788	(17
5		S-958 22	false	Sweden	0921-12 34 65	09
6		68306	true	Germany	0621-08460	06
7		67000	true	France	88.60.15.31	88
8		28023	true	Spain	(91) 555 22 82	(9
9		13008	true	France	91.24.45.40	91
10		T2F 8M4	false	Canada	(604) 555-4729	(6
11		EC2 5NT	false	UK	(171) 555-1212	
12		1010	false	Argentina	(1) 135-5555	(1)
13		05022	true	Mexico	(5) 555-3392	(5)
14		3012	false	Switzerland	0452-076545	
15		05432-043	false	Brazil	(11) 555-7647	
16		WX1 6LT	false	UK	(171) 555-2282	(17

Fonte: A autora

O Open Studio é o único das três ferramentas a detectar o tipo específico da coluna automaticamente, no momento da seleção de colunas para análise; isso pode ser visto na Figura 4.8. Ele também realiza detecção dos padrões mais e menos

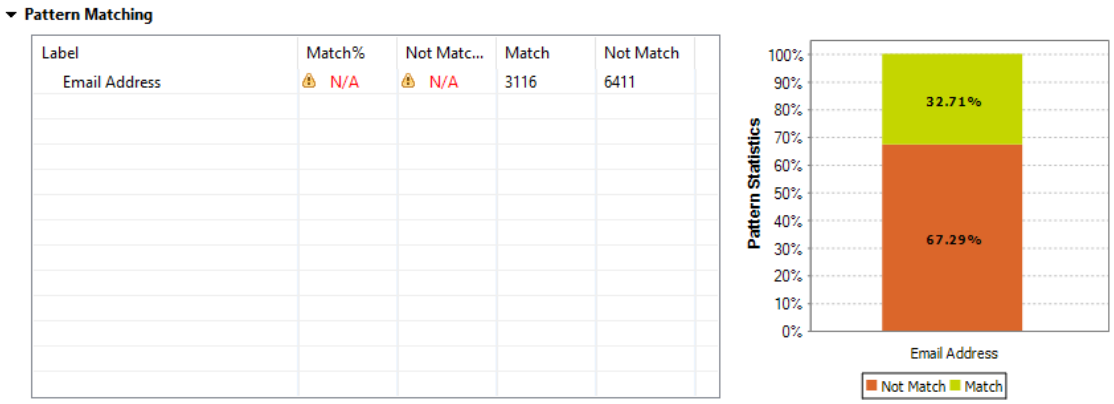
frequentes de uma coluna a partir do indicador “Pattern Frequency Statistics” de sua análise de colunas, disponibilizando o resultado desta análise da maneira vista na Figura 4.23. Adicionalmente, o indicador “Patterns” realiza a verificação de alguns padrões predefinidos a serem escolhidos pelo usuário. Os resultados de tal análise podem ser vistos na Figura 4.24.

Figura 4.23: resultado de análise de colunas do Open Studio para o indicador "Pattern Frequency Statistics" .



Fonte: A autora

Figura 4.24: resultado de análise de colunas do Open Studio para o indicador "Patterns" .



Fonte: A autora

A pontuação total das ferramentas para esta seção está na Tabela 4.3.

Tabela 4.3: resultado da análise de ferramentas segundo o critério “Padrões e tipos de dados”.

Subtarefa	DQ Analyzer	Pandora Free	Open Studio
Detecção de padrões	1	1	1
Verificação de padrões	1	1	1
Detecção de tipo genérico	1	1	1
Detecção de tipo específico	0	0	1
Detecção de dígitos/decimais	0	1	0
Pontuação total (n/5 x 100)	**60**	**80**	**80**

4.4. Classificação de domínio

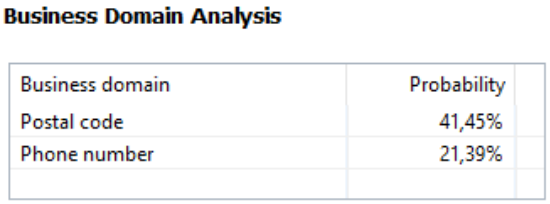
Esta funcionalidade consiste na identificação de domínio dos valores de uma coluna.

O DQ Analyzer possibilita tanto a identificação de classe de dados como de domínio concreto, bastando para isso marcar a opção “Domain analysis” na janela de criação de novo profile já vista na Figura 4.1. O resultado da análise de domínio de negócio, que determina possíveis domínios concretos aos quais a coluna pode pertencer e a probabilidade de que ela pertença a cada um, pode ser visto na Figura 4.25. Já a Figura 4.26 exibe o resultado da análise de domínio, que retorna a classe de dados detectada na coluna (i.e., *enum*, *pattern*, *integer*) e uma breve análise de contagens.

O Pandora Free não possui uma análise automática de domínio de uma coluna, concreto ou não. Quanto ao Open Studio, embora possamos considerar que sua análise de padrões (Figura 4.24) realiza uma verificação de domínio, as subtarefas analisadas aqui são de identificação; por esse motivo, a ferramenta também não obterá a pontuação referente a elas.

A Tabela 4.4 detalha o resultado da análise referente às subtarefas desta seção.

Figura 4.25: resultado de análise de domínio de negócio do DQ Analyzer.

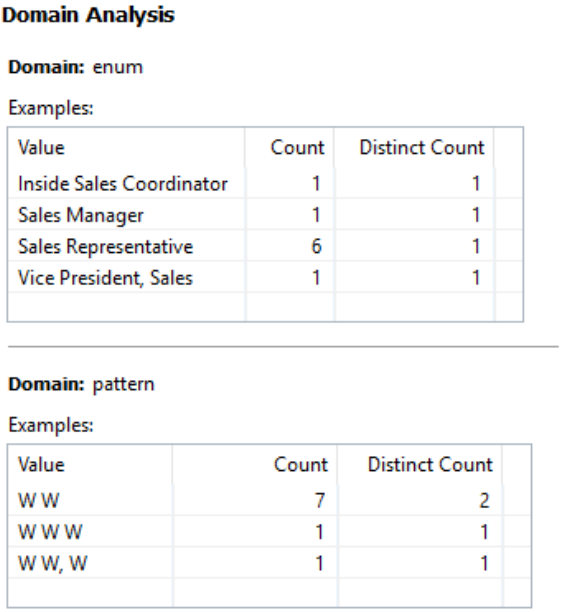


Business Domain Analysis

Business domain	Probability
Postal code	41,45%
Phone number	21,39%

Fonte: A autora

Figura 4.26: resultado de análise de domínio do DQ Analyzer.



Domain Analysis

Domain: enum

Examples:

Value	Count	Distinct Count
Inside Sales Coordinator	1	1
Sales Manager	1	1
Sales Representative	6	1
Vice President, Sales	1	1

Domain: pattern

Examples:

Value	Count	Distinct Count
W W	7	2
W W W	1	1
W W, W	1	1

Fonte: A autora

Tabela 4.4: resultado da análise de ferramentas segundo o critério “Classificação de domínio”.

Subtarefa	DQ Analyzer	Pandora Free	Open Studio
Detecção de classe de dados	1	0	0
Identificação de domínio concreto	1	0	0
Pontuação total (n/2 x 100)	**100**	**0**	**0**

4.5. Correlações e regras de associação

Esta seção trata da detecção de correlações entre valores numéricos e de regras de associação entre registros. Nenhuma das ferramentas analisadas aqui oferece a possibilidade de detecção de regras de associação. O DQ Analyzer e o Pandora Free também não descobrem correlações; já o Open Studio possui um grupo de análises de correlação (“Correlation Analysis”), conforme pode ser visto na Figura 4.7, com três possibilidades de combinações de atributos: nominal-numérico,

nominal-data e nominal-nominal. No entanto, para este trabalho, consideramos apenas correlações entre valores puramente numéricos, e o Open Studio não fornece esta funcionalidade; portanto, também não receberá a pontuação relativa a esta subtarefa. Dessa maneira, todas as ferramentas obtiveram uma pontuação total de zero neste critério, como pode ser visto na Tabela 4.5.

Tabela 4.5: resultado da análise de ferramentas segundo o critério "Correlações e regras de associação".

Subtarefa	DQ Analyzer	Pandora Free	Open Studio
Detecção de correlações	0	0	0
Regras de associação	0	0	0
Pontuação total (n/2 x 100)	**0**	**0**	**0**

4.6. *Clusters* e *outliers*

Esta seção abrange as funcionalidades de *clustering* e detecção de *outliers*. Nenhuma das ferramentas analisadas neste trabalho realiza *clustering*, e nem o DQ Analyzer nem o Open Studio possuem análises de detecção de *outliers*. O Pandora Free realiza análise de valores considerados *outliers* dentro de uma única coluna, como por exemplo valores maiores ou mais longos do que o normal, como pode ser visto nas opções do menu de contexto da Figura 4.5. No entanto, não é possível realizar esta análise sobre um grupo de colunas, como definido aqui; por este motivo, a ferramenta também não receberá a pontuação referente a esse critério. O resultado final pode ser visto na Tabela 4.6.

Tabela 4.6: resultado da análise de ferramentas segundo o critério "Clusters e outliers".

Subtarefa	DQ Analyzer	Pandora Free	Open Studio
Clustering	0	0	0
Detecção de *outliers*	0	0	0
Pontuação total (n/2 x 100)	**0**	**0**	**0**

4.7. Resumos e esboços

As tarefas desta seção consistem na criação de resumos e/ou esboços de dados. Nenhuma das ferramentas analisadas aqui oferece suporte à criação de resumos ou esboços, tampouco ao cálculo automático da similaridade de Jaccard, de

modo que todas receberão pontuação zero nas subtarefas deste critério, como pode ser visto na Tabela 4.7.

Tabela 4.7: resultado da análise de ferramentas segundo o critério “Resumos e esboços”.

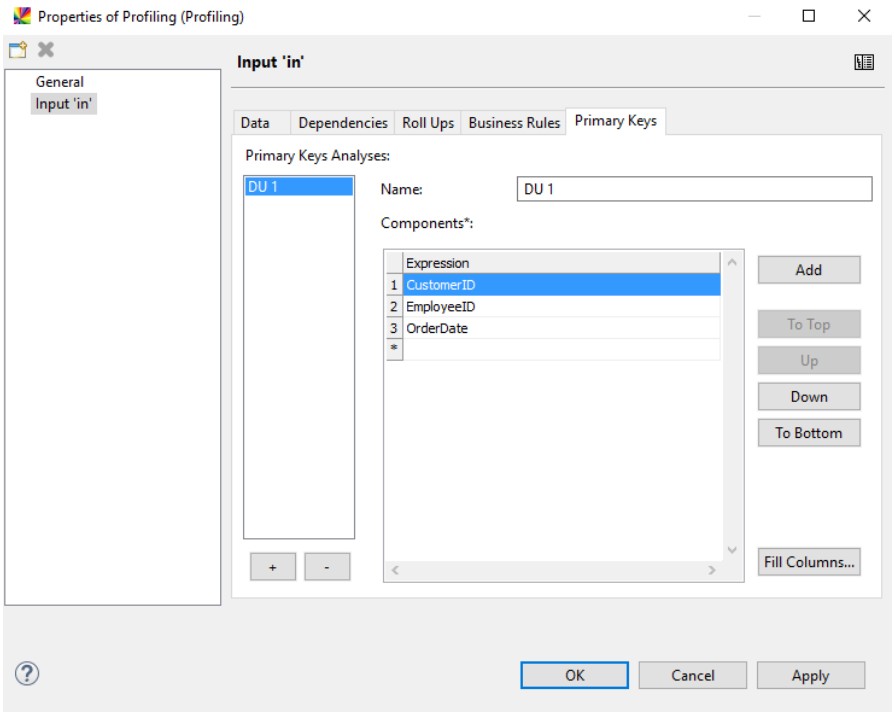
Subtarefa	DQ Analyzer	Pandora Free	Open Studio
Criação de resumos e/ou esboços	0	0	0
Similaridade de Jaccard	0	0	0
Pontuação total (n/2 x 100)	**0**	**0**	**0**

4.8. Dependências de unicidade

Este critério avalia subtarefas relativas à verificação e descoberta de combinações únicas de colunas.

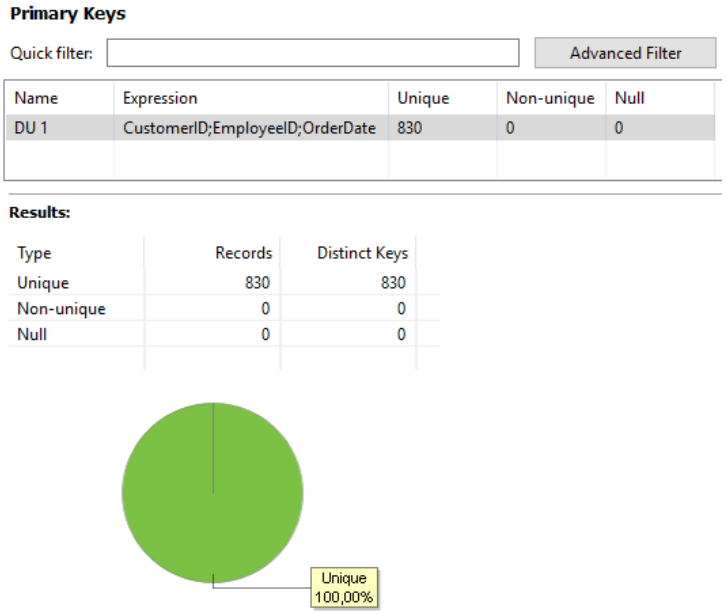
Análises de dependência no DQ Analyzer requerem a criação de um plano de *profile.* Para a verificação de dependências de unicidade, um plano simples como o da Figura 4.15 é suficiente. Um duplo clique no operador de *profiling* abre a janela de configuração vista na Figura 4.27. A aba “Primary Keys” permite a escolha de uma ou mais combinações de colunas, que serão verificadas para saber se são únicas. O resultado dessa análise é apresentado na <>.

Figura 4.27: janela de configuração do operador de *profiling* do DQ Analyzer, aba "Primary Keys".



Fonte: A autora

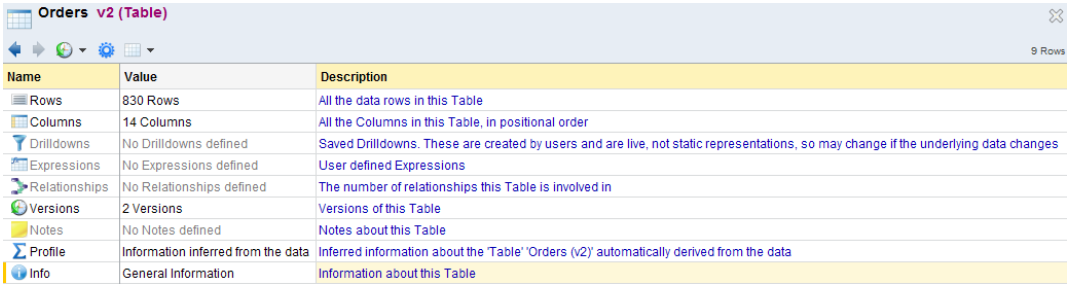
Figura 4.28: resultado de análise de chaves primárias do DQ Analyzer.



Fonte: A autora

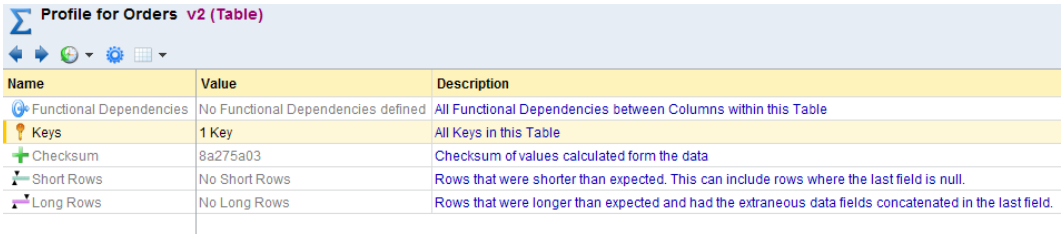
Para realizar a verificação de dependências no Pandora Free, é necessário visualizar primeiro os detalhes de tabela, selecionando a opção “Details” no menu de contexto da Figura 4.20. A janela resultante pode ser vista na Figura 4.29; um duplo clique na opção “Profile” abre o perfil de tabela mostrado na Figura 4.30.

Figura 4.29: janela de detalhes de tabela do Pandora Free.



Fonte: A autora

Figura 4.30: perfil de tabela do Pandora Free.

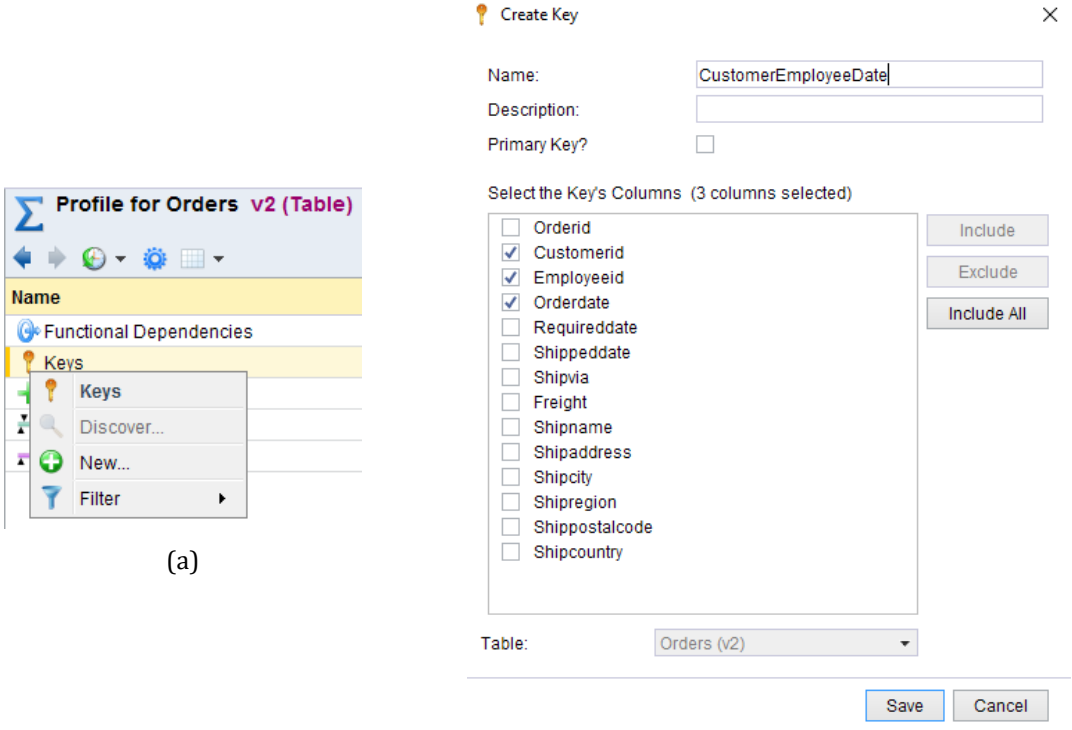


Fonte: A autora

Ao clicar com o botão direito sobre o registro “Keys”, um menu de contexto é aberto. A opção “New” abre a janela de criação de nova chave, como pode ser visto na Figura 4.31. Nesta janela é possível selecionar a combinação de colunas a ser verificada como única ou não. A descoberta de chaves, embora presente, não está disponível para utilização na versão gratuita da ferramenta.

A Figura 4.32 exibe o resultado da verificação de colunas-chave.

Figura 4.31: análise de chaves do Pandora Free. (a) Menu de contexto; (b) Janela de configuração



(a)

(b)

Fonte: A autora

Figura 4.32: resultado da análise de chaves do Pandora Free.

Keys for Orders (v2) (Table)

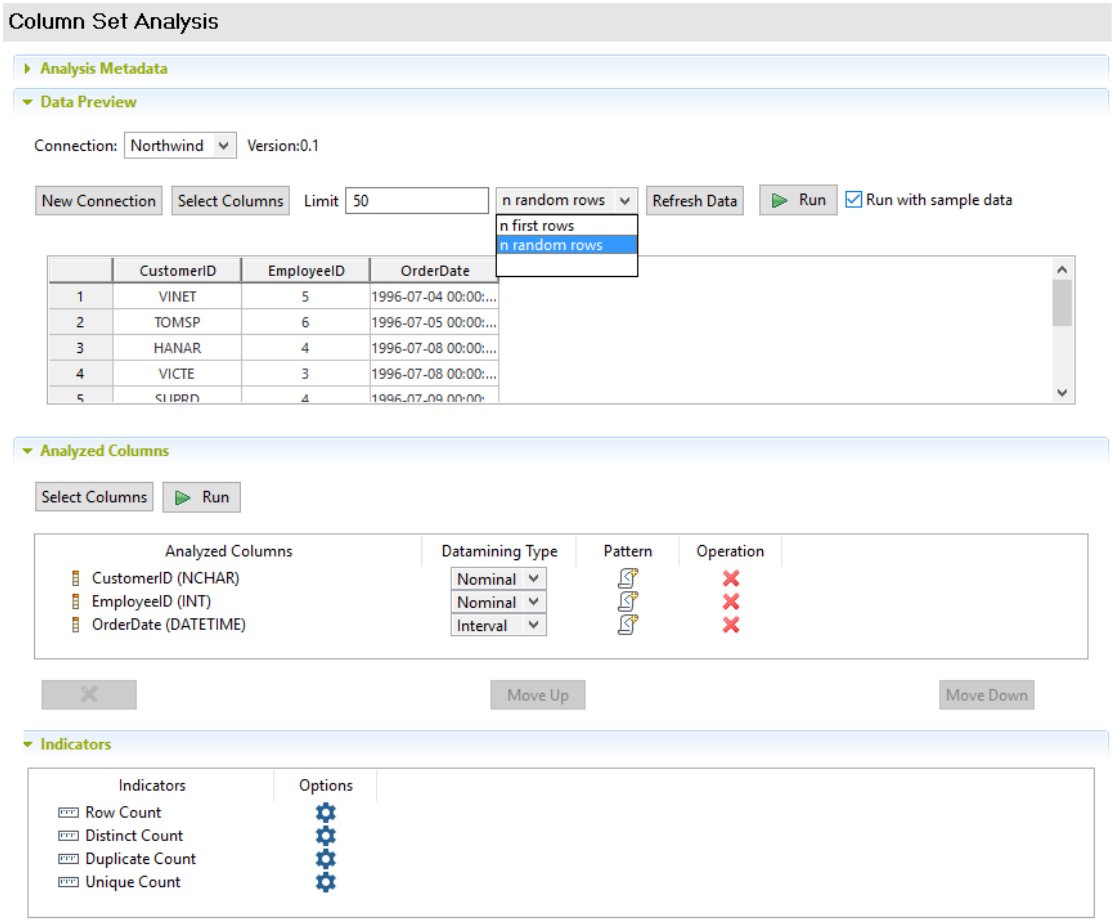
Business Name	Primary Key	Key Columns	Quality	Status	Rows In Error	Null References
Unnamed Key 274	No	Orderid	100%	Tested	0	0
Customeremployeedate	No	Customerid, Employeeid, Orderdate	100%	Tested	0	0

Fonte: A autora

O Open Studio verifica a existência de dependências de unicidade em sua análise de conjuntos de colunas, através da opção “Column Set Analysis” da categoria “Table Analysis”, na janela da Figura 4.7. A janela de configuração pode ser

vista na Figura 4.33. Ele possibilita também a avaliação de dependências de unicidade aproximadas ao permitir a realização de análises sobre uma quantidade predeterminada de linhas (primeiras ou aleatórias), ao marcar a opção “Run with sample data”.

Figura 4.33: janela de configuração de análise de conjunto de colunas do Open Studio.



Fonte: A autora

O resultado da análise é exatamente igual ao da análise de colunas simples para o indicador “Simple Statistics”, conforme visto na Figura 4.9, mas sendo referente ao grupo de colunas analisado; de modo que uma combinação de colunas é única se e somente se seu percentual de valores distintos for igual a 100%. Não é possível, no entanto, descobrir automaticamente combinações únicas de coluna.

Nem o DQ Analyzer nem o Pandora Free realizam avaliações de dependência aproximada; além disso, nenhuma das ferramentas analisadas aqui é capaz de descrever dependências de unicidade condicionais. As pontuações de cada ferramenta para esta seção podem ser vistas na Tabela 4.8.

Tabela 4.8: resultado da análise de ferramentas segundo o critério “Dependências de unicidade”.

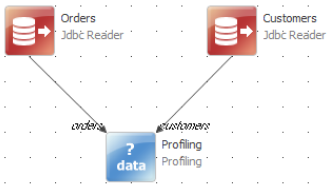
Subtarefa	DQ Analyzer	Pandora Free	Open Studio
Verificação de DU	1	1	1
Descoberta de DU	0	0	0
Dependência condicional	0	0	0
Dependência aproximada	0	0	1
Pontuação total (n/4 x 100)	**25**	**25**	**50**

4.9. Dependências de inclusão

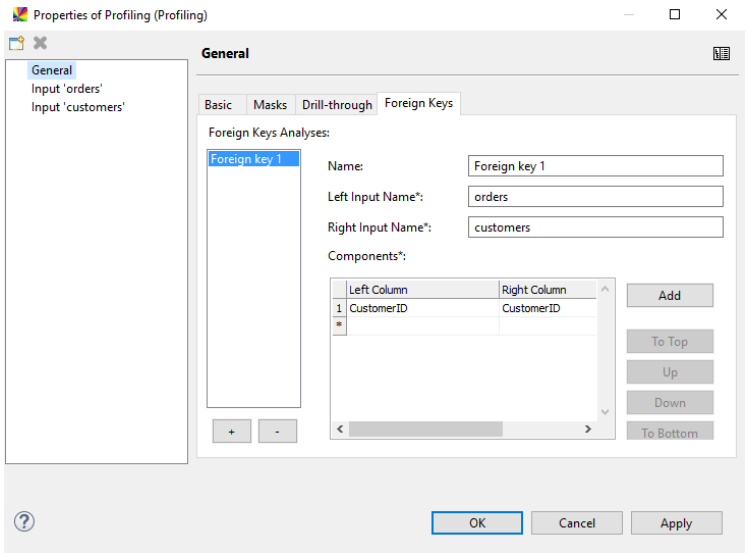
Esta seção avalia tarefas relativas à descoberta e verificação de dependências de inclusão, e consequentemente, de chaves estrangeiras.

Como dependências de inclusão são calculadas entre tabelas diferentes, o plano de *profile* do DQ Analyzer precisa possuir pelo menos dois operadores de entrada de dados. Nesse caso, a escolha das colunas a serem analisadas para verificar a existência de uma dependência de inclusão é feita a partir da aba “Foreign Keys” da janela de configuração (Figura 4.34). O resultado da análise gerada pode ser visto na Figura 4.35.

Figura 4.34: análise de chave estrangeira no DQ Analyzer. (a) Plano de profile; (b) Janela de configuração do operador "Profiling", aba “Foreign Keys”.



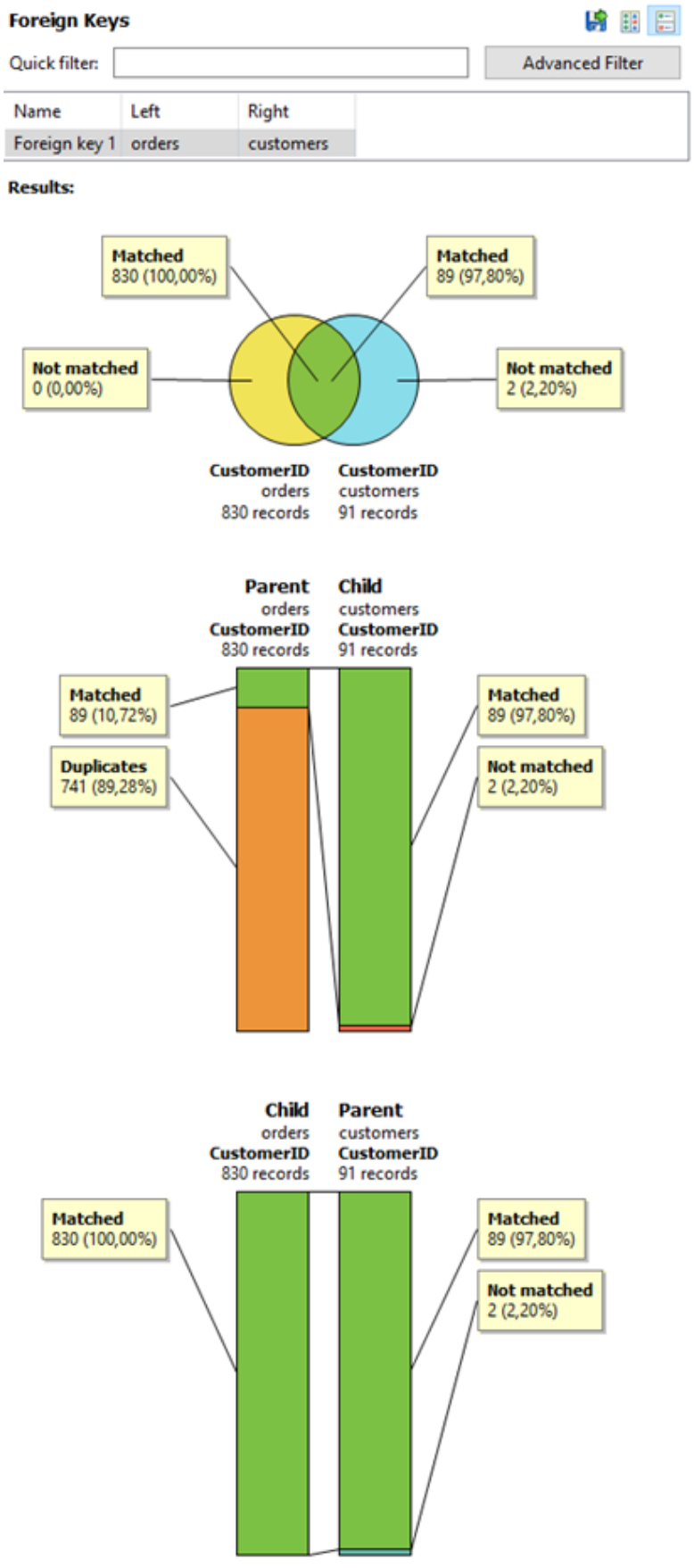
(a)



(b)

Fonte: A autora

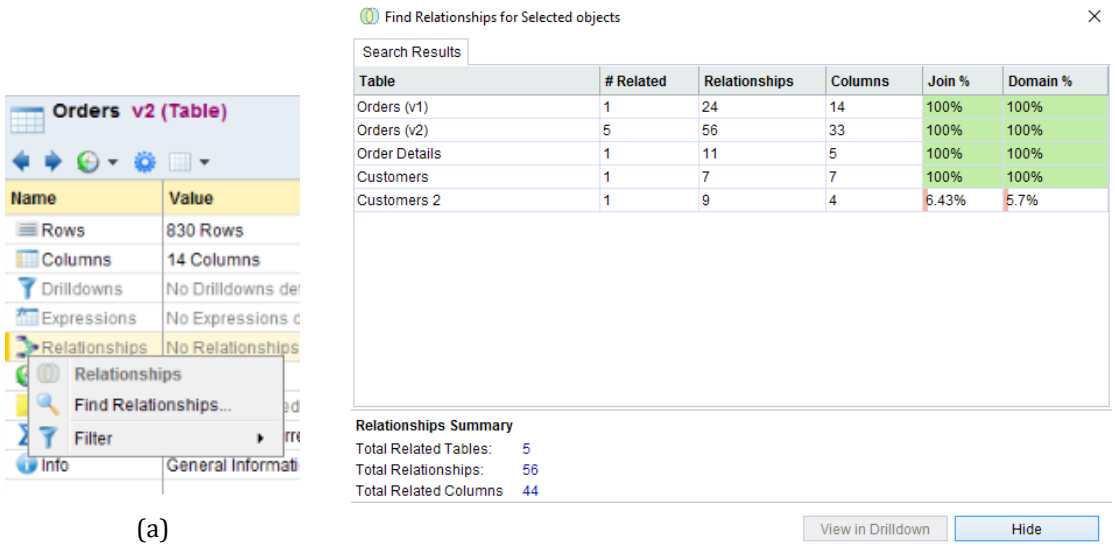
Figura 4.35: resultado da análise de chave estrangeira do DQ Analyzer.



Fonte: A autora

O gerenciamento de chaves estrangeiras no Pandora Free se dá através de relacionamentos entre tabelas, que podem ser visualizados a partir dos detalhes de tabela. No entanto, a versão gratuita da ferramenta não dá suporte à verificação de relacionamentos, e embora a opção “Find Relationships” esteja acessível no menu de contexto, a janela resultante não traz nenhum detalhe sobre as colunas que formam o relacionamento, como podemos ver na Figura 4.36; por esse motivo, a ferramenta não receberá a pontuação referente a esta subtarefa.

Figura 4.36: análise de relacionamentos do Pandora Free. (a) Menu de contexto; (b) Janela de configuração



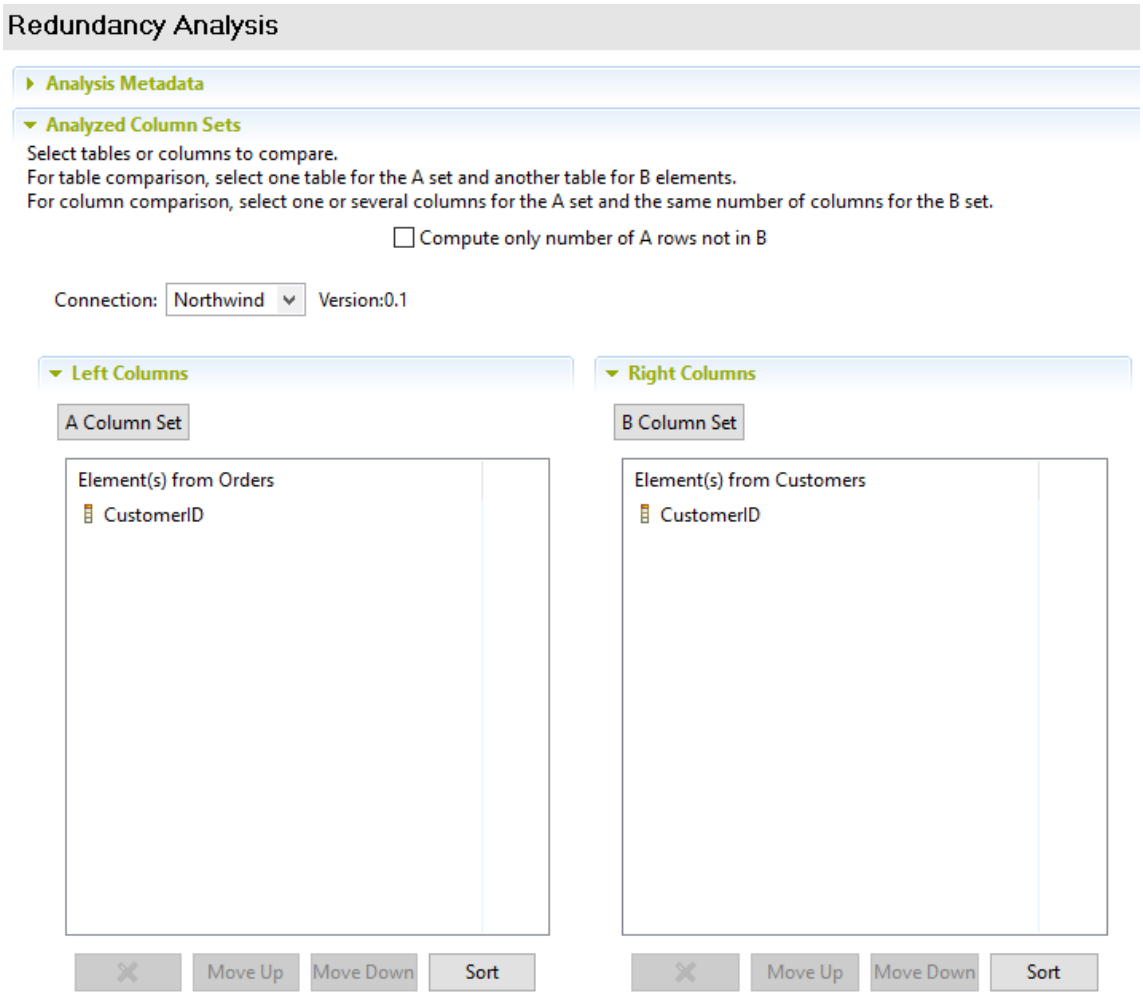
(a) (b)

Fonte: A autora

No Open Studio, a opção “Redundancy Analysis” encontrada na categoria “Cross Table Analysis” da janela da Figura 4.7 verifica os elementos de uma coluna que podem ser encontrados em outra, e vice-versa; de modo que é possível verificar dependências de inclusão a partir desta análise. A janela de configuração pode ser vista na Figura 4.37, enquanto o resultado da análise está na Figura 4.38.

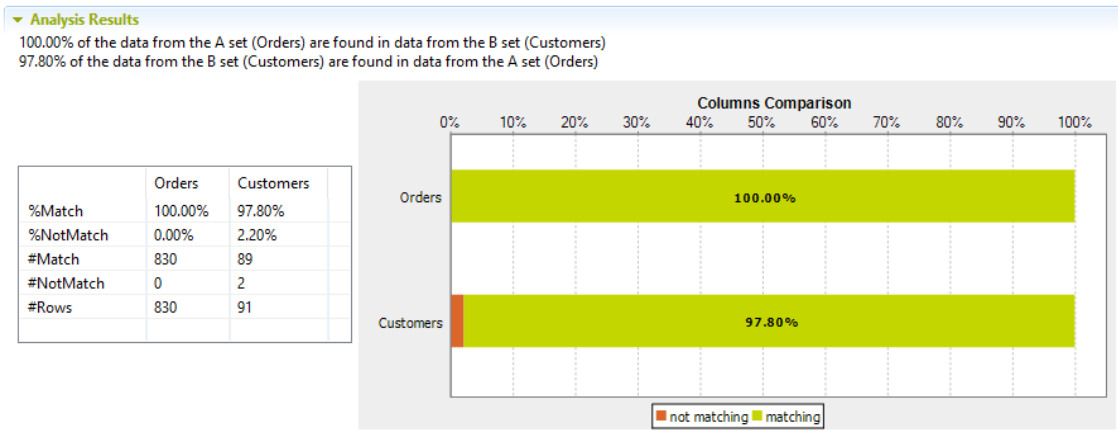
A pontuação final de cada ferramenta pode ser vista na Tabela 4.9. Nenhuma das ferramentas realiza detecção de dependências de inclusão, sejam totais, condicionais ou aproximadas.

Figura 4.37: janela de configuração de análise de redundância no Open Studio.



Fonte: A autora

Figura 4.38: resultado de análise de redundância no Open Studio.



Fonte: A autora

Tabela 4.9: resultado da análise de ferramentas segundo o critério “Dependências de inclusão”.

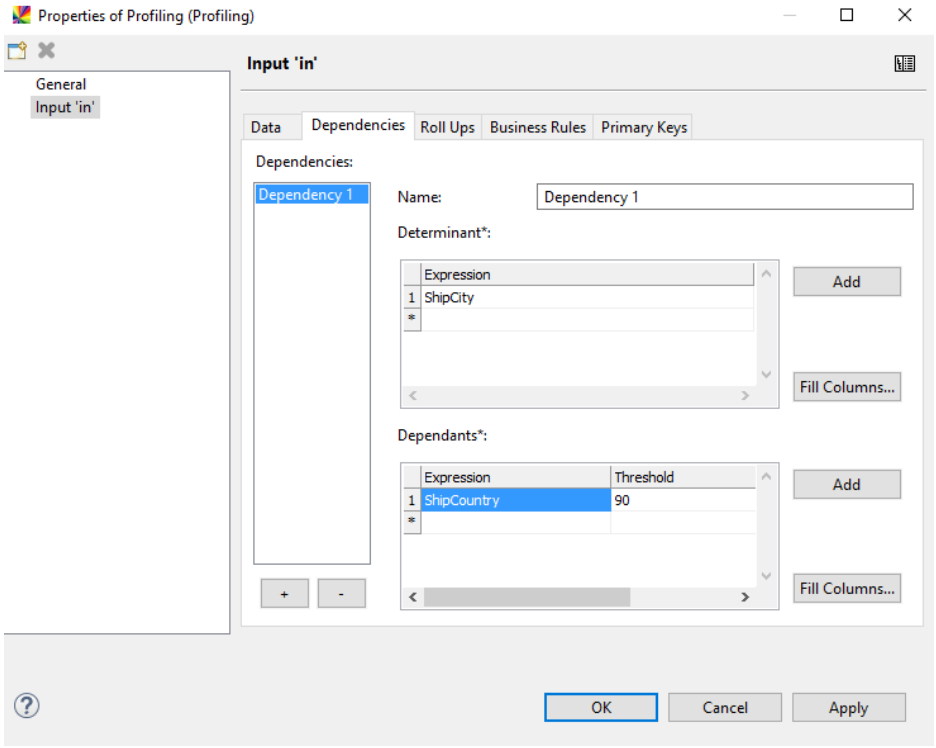
Subtarefa	DQ Analyzer	Pandora Free	Open Studio
Verificação de DI	1	0	1
Descoberta de DI	0	0	0
Dependência condicional	0	0	0
Dependência aproximada	0	0	0
Pontuação total (n/4 x 100)	**25**	**0**	**25**

4.10. Dependências funcionais

As funcionalidades desta seção referem-se à descoberta e verificação de dependências funcionais.

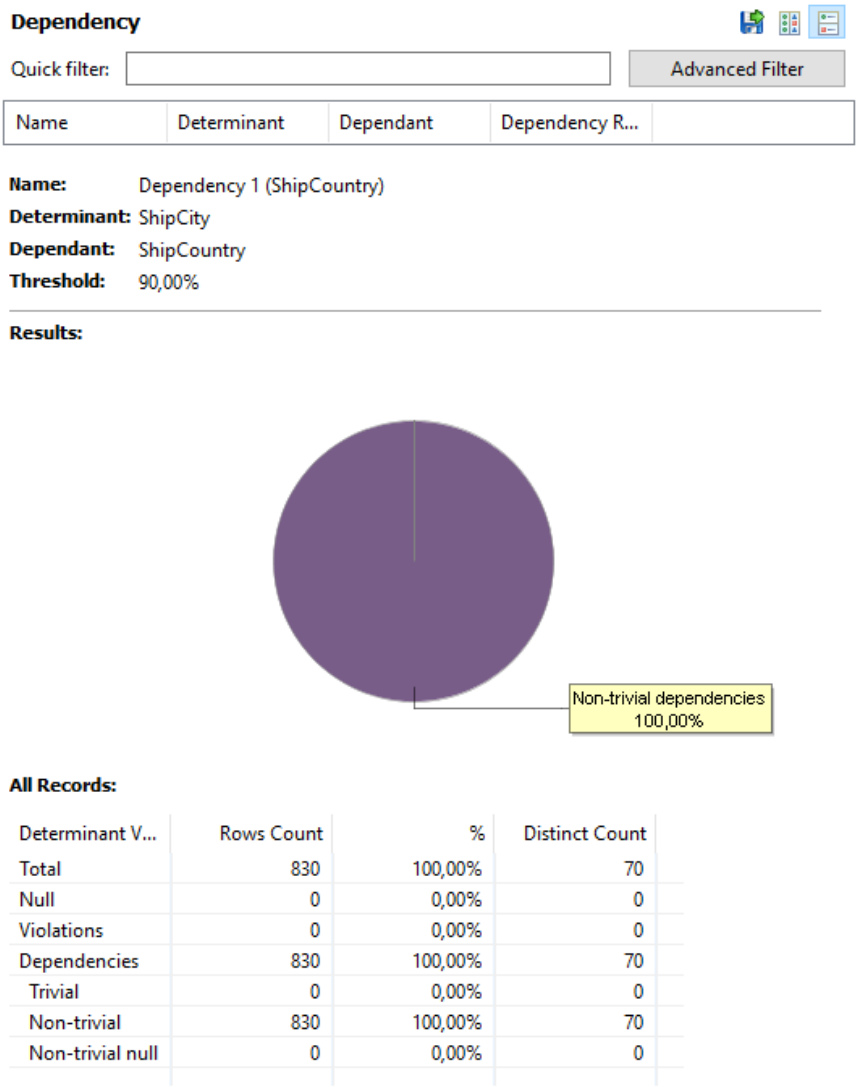
O DQ Analyzer, na aba “Dependencies” da janela de configuração de seu operador de *profiling* dentro de um plano de *profile*, permite escolher colunas determinantes e determinadas para verificação da existência de dependências funcionais, além de um limite percentual acima do qual a identificação da dependência é positiva. A Figura 4.39 traz a janela de configuração, e a Figura 4.40, o resultado da análise.

Figura 4.39: janela de configuração de análise de dependência funcional no DQ Analyzer.



Fonte: A autora

Figura 4.40: resultado da análise de dependências funcionais no DQ Analyzer.



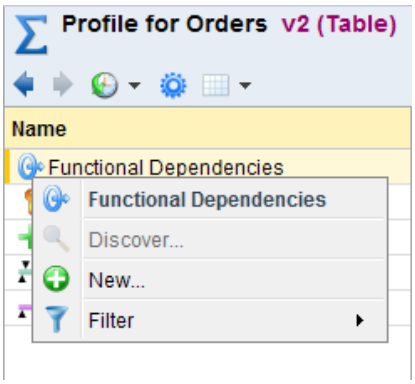
Determinant V...	Rows Count	%	Distinct Count
Total	830	100,00%	70
Null	0	0,00%	0
Violations	0	0,00%	0
Dependencies	830	100,00%	70
Trivial	0	0,00%	0
Non-trivial	830	100,00%	70
Non-trivial null	0	0,00%	0

Fonte: A autora

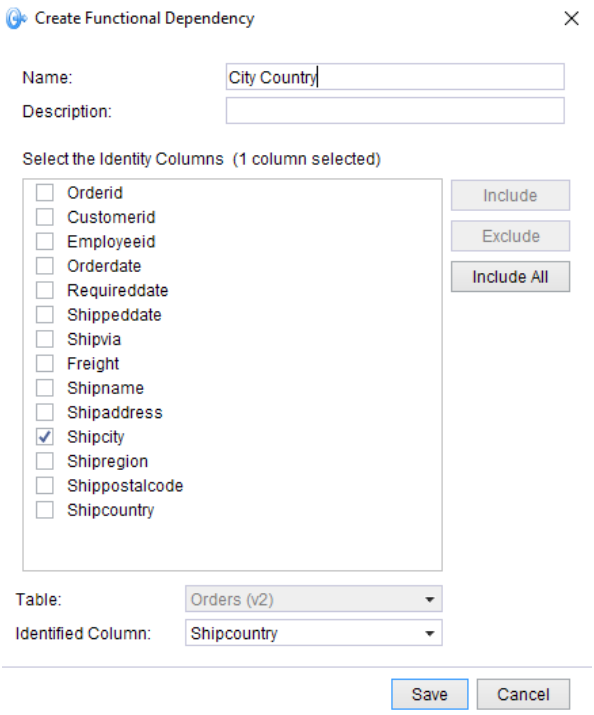
No Pandora Free, a verificação de dependências funcionais pode ser feita através do *profile* de tabela já visto na Figura 4.30. Clicando com o botão direito sobre o registro “Functional Dependencies”, abre-se um menu de contexto, onde é possível selecionar a opção “New” para abrir a janela de criação de dependências funcionais. Nesta janela pode-se selecionar uma ou mais colunas identificadoras e uma única coluna identificada por elas, de modo a verificar a existência ou não de uma dependência funcional entre elas. A janela de configuração e o menu de contexto podem ser vistos na Figura 4.43. A opção “Discover”, para descoberta automática de dependências funcionais, assim como na análise de dependências de unicidade, existe, porém não está disponível na versão gratuita.

A Figura 4.44 exibe o resultado da análise.

Figura 4.41: análise de dependências funcionais do Pandora Free. (a) Menu de contexto; (b) Janela de configuração



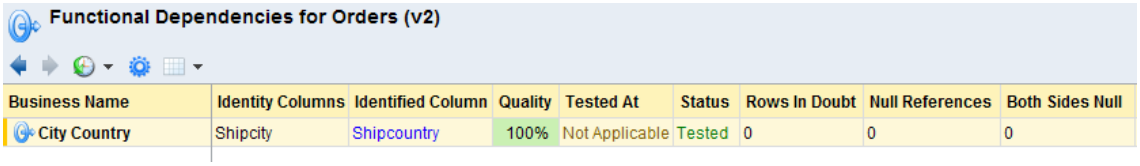
(a)



(b)

Fonte: A autora

Figura 4.42: resultado de análise de dependências funcionais do Pandora Free.

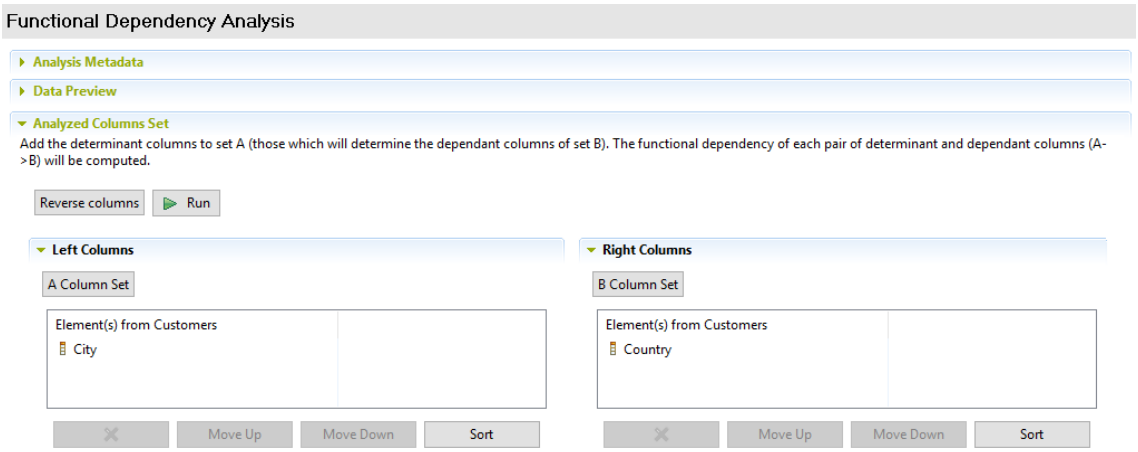


Fonte: A autora

Finalmente, a análise de dependências funcionais no Open Studio pode ser realizada a partir da categoria “Table Analysis”, opção “Functional Dependency Analysis” da Figura 4.7. A janela de configuração que se abre, presente na Figura 4.43, permite a escolha de colunas determinantes (lado esquerdo) e determinadas (lado direito). A análise computa a dependência funcional para cada par determinante-determinada de colunas, e seu resultado pode ser visto na Figura 4.44.

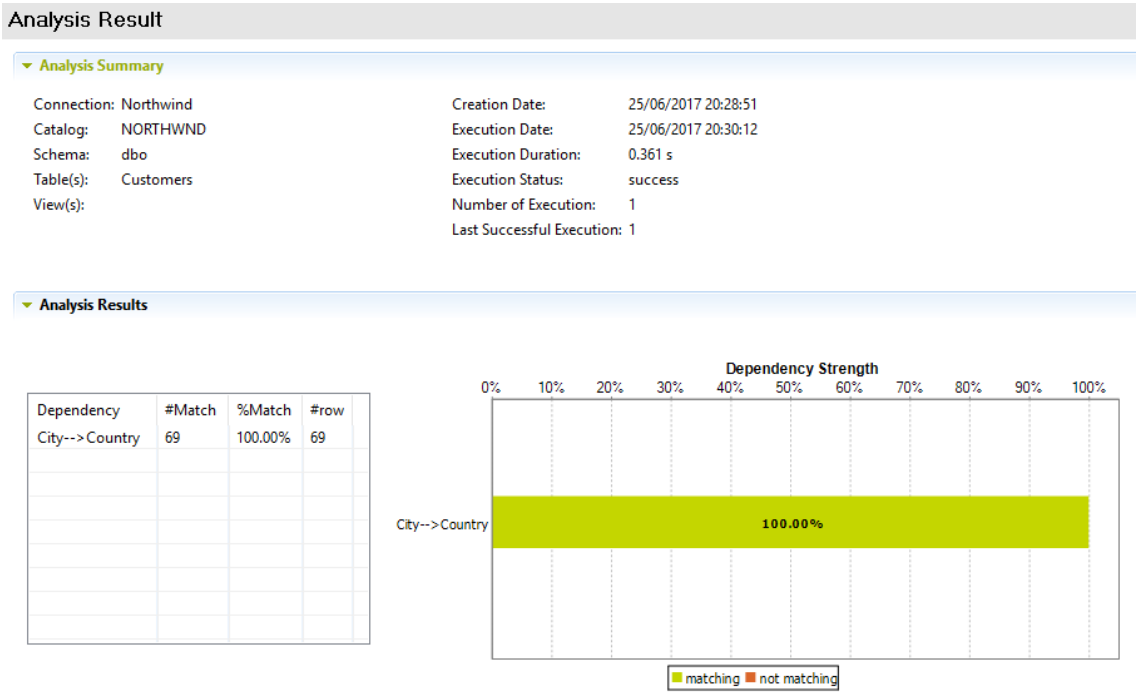
Assim como no caso das dependências de inclusão, nenhuma das ferramentas analisadas realiza detecção de dependências funcionais, sejam elas totais, aproximadas ou condicionais. A pontuação geral para esta funcionalidade está na Tabela 4.11.

Figura 4.43: janela de configuração de análise de dependências funcionais no Open Studio.



Fonte: A autora

Figura 4.44: resultado de análise de dependências funcionais no Open Studio.



Fonte: A autora

Tabela 4.10: resultado da análise de ferramentas segundo o critério “Dependências funcionais”.

Subtarefa	DQ Analyzer	Pandora Free	Open Studio
Verificação de DF	1	1	1
Descoberta de DF	0	0	0
Dependência condicional	0	0	0
Dependência aproximada	0	0	0
Pontuação total (n/4 x 100)	**25**	**25**	**25**

4.11. Resultados e considerações finais

A Tabela 4.11 exibe a média aritmética das pontuações obtidas por todas as ferramentas em cada critério.

Tabela 4.11: média aritmética das pontuações de cada ferramenta por critério.

Critério	DQ Analyzer	Pandora Free	Open Studio
Cardinalidades	100	100	100
Distribuição de valores	60	60	80
Padrões e tipos de dados	60	80	80
Classificação de domínio	100	0	0
Correlações e regras de associação	0	0	0
Clusters e *outliers*	0	0	0
Resumos e esboços	0	0	0
Dependências de unicidade	25	25	50
Dependências de inclusão	25	0	25
Dependências funcionais	25	25	25
Pontuação final (n/10)	**39,5**	**29,0**	**36,0**

Verificamos que, no que diz respeito a tarefas tradicionais de *profiling* em bases de dados relacionais, o DQ Analyzer oferece o maior número de funcionalidades dentre as ferramentas que analisamos, com o Open Studio não muito atrás. O Pandora Free obteve a menor pontuação geral; vale salientar, no entanto, que algumas das funcionalidades analisadas aqui eram disponibilizadas pela versão completa (e paga) da ferramenta Pandora, não tendo, portanto, sido consideradas aqui.

É possível observar que, no geral, todas as ferramentas obtiveram boas pontuações nos critérios referentes a análises de coluna única. Esse fato indica que a análise de colunas individuais é um segmento amadurecido dentro da disciplina de geração de perfil de dados. Em contrapartida, nota-se que todas as ferramentas obtiveram pontuação igual a zero nos critérios referentes a análise de múltiplas colunas, isto é, correlações, regras de associação, *clustering* e detecção de *outliers,* resumos e esboços; além disso, nenhuma ferramenta obteve mais de 50% de aproveitamento em nenhum dos critérios de análise de dependências. Isso sugere a necessidade das fabricantes de voltar suas atenções para os relacionamentos entre valores de colunas de modo a obter metadados mais completos e confiáveis.

Finalmente, é perceptível que as ferramentas analisadas obtiveram pontuações mais altas em funcionalidades de verificação do que de detecção, o que era esperado devido à maior complexidade das tarefas de detecção; no entanto, isso também indica uma carência a ser sanada nessas ferramentas se desejarem se manter competitivas, especialmente considerando o cenário de mercado voltado para grandes volumes de dados não-estruturados e advindos de diversas fontes que temos hoje.

5. Conclusão

O objetivo principal deste trabalho foi a realização de uma análise funcional de três ferramentas "gratuitas" de *data profiling*, a partir de um conjunto de tarefas tradicionais a serem realizadas sobre os dados de entrada a fim de obter metadados técnicos a seu respeito. Inicialmente, foi feito um estudo conceitual a respeito de perfis de dados, de modo a introduzir as funcionalidades consideradas na análise; em seguida, foram escolhidas e apresentadas três ferramentas gratuitas de *profiling* relevantes no mercado de qualidade de dados. Finalmente, realizou-se a análise funcional das ferramentas.

A partir dos resultados obtidos, pode-se concluir que comercialmente há ainda um grande potencial não explorado no que diz respeito à disciplina de geração de perfis de dados. Dois pontos específicos que merecem mais atenção no futuro foram identificados: detecção de regras "ocultas" nos dados, em detrimento da mera verificação de conformidade dos dados a regras já conhecidas; e aplicação de processos e tarefas muito utilizados em mineração de dados, como *clustering*, na obtenção de metadados técnicos.

5.1. Trabalhos Futuros

Este trabalho esteve restrito a uma análise meramente funcional de ferramentas comerciais; um possível trabalho futuro seria a proposição de um *benchmark* para geração de perfil de dados que leve em consideração outros critérios importantes como o desempenho (tempo e recursos de máquina utilizados) e a corretude da análise, sobretudo no que diz respeito às tarefas de *profiling* voltadas para detecção de regras e dependências.

Adicionalmente, seria interessante analisar o uso de ferramentas comerciais e gratuitas na realização de *profiling* aplicado a casos de uso reais, especialmente sobre bases de dados não-estruturados.

Referências

ABEDJAN, Z.; GOLAB, L.; NAUMANN, F. Profiling relational data: a survey. **VLDB Journal**, v. 24, n. 4, p. 557–581, 2015.

ABEDJAN, Z.; QUIANÉ-RUIZ, J. A.; NAUMANN, F. Detecting unique column combinations on dynamic data. In: 30th International Conference on Data Engineering, **Anais**...Chicago, IL, EUA: IEEE, 2014. Disponível em < http://ieeexplore.ieee.org/document/6816721/>. Acesso em: 05 fev. 2017.

BAUCKMANN, J. et al. Efficiently detecting inclusion dependencies. In: 23rd International Conference on Data Engineering, **Anais**...Istambul, Turquia: IEEE, 2007. Disponível em: <http://ieeexplore.ieee.org/document/4221822/>. Acesso em: 22 jan. 2017.

BENFORD, F. The Law of Anomalous Numbers. **Proceedings of the American Philosophical Society**, v. 78, n. 4, p. 551–572, 1938. Disponível em: <http://www.jstor.org/stable/984802>. Acesso em: 16 abr. 2017.

BERKHIN, P. A Survey Of Clustering Data Mining Techniques. In: KOGAN, J.; NICHOLAS, C.; TEBOULLE, M. **Grouping Multidimensional Data**, p. 25-71. Berlin, Heidelberg: Springer, 2006.

BLOOR RESEARCH GROUP. **Data Profiling and Discovery Market Update**. 2013. Disponível em: <http://www.bloorresearch.com/research/market-update/data-profiling-and-discovery-market-update-2013/>. Acesso em: 05 fev. 2017.

BLOOR RESEARCH GROUP. **"Free" Data Profiling Tools**. 2014. Disponível em: <http://www.bloorresearch.com/research/market-report/free-data-profiling-tools/>. Acesso em: 18 jan. 2017.

CAI, L.; ZHU, Y. The Challenges of Data Quality and Data Quality Assessment in the Big Data Era. **Data Science Journal**, n. 14, p. 2. 2015. Disponível em < http://datascience.codata.org/articles/10.5334/dsj-2015-002/>. Acesso em: 17 abr. 2017.

CHANDOLA, V.; KUMAR, V. Summarization - Compressing data into an informative representation. In: Fifth IEEE International Conference on Data Mining**, Anais...**Houston, TX, EUA: IEEE, 2005. Disponível em < http://ieeexplore.ieee.org/document/1565667/>. Acesso em: 16 abr. 2017.

CORMODE, G. et al. Synopses for Massive Data: Samples, Histograms, Wavelets, Sketches. In: **Foundations and Trends in Databases**, n. 9. Now Publishers Inc., 2012. Disponível em: <http://db.cs.berkeley.edu/cs286/papers/synopses-fntdb2012.pdf>. Acesso em: 17 abr. 2017.

DOAN, A.; HALEVY, A. Y. Semantic-Integration Research in the Database Community - A Brief Survey. **AI Magazine**, v. 26, n. 1, p. 83–94, 2005. Disponível em: <http://pages.cs.wisc.edu/~anhai/papers/si-survey-db-community.pdf>. Acesso em: 6 abr. 2017.

ELMASRI, R.; NAVATHE, S. B. **Sistemas de banco de dados**. 6ª ed. São Paulo: Pearson, 2011.

EXPERIAN DATA QUALITY. **The 2017 global data management benchmark report**. Disponível em <https://www.edq.com/resources/data-management-whitepapers/2017-global-data-management-benchmark-report/>. Acesso em: 01 mai. 2017.

GARTNER GROUP. **Magic Quadrant for Data Quality Tools**. [s.l: s.n.]. Disponível em: <https://www.gartner.com/doc/reprints?id=1-3MOFN4I&ct=161128&st=sb>. Acesso em: 12 abr. 2017.

HAN, J.; KAMBER, M.; PEI, J. **Data Mining: Concepts and Techniques**. 3ª ed. Waltham, MA, EUA: Morgan Kaufmann, 2012.

KIMBALL, R. **Kimball Design Tip # 59 : Surprising Value of Data Profiling**. n. 59, p. 1–2, 2004. Disponível em: <http://www.kimballgroup.com/wp-content/uploads/2012/05/DT59SurprisingValue.pdf>. Acesso em: 15 jan. 2017.

NAUMANN, F. Data profiling revisited. **ACM SIGMOD Record**, v. 42, n. 4, p. 40–49, 2014.

PAPENBROCK, T. et al. Functional Dependency Discovery: An Experimental Evaluation of Seven Algorithms. **Proceeding of the VLDB Endowment**, v. 8, n. 10, p. 1082–1093, 2015.

RAHM, E.; DO, H. H. Data Cleaning: Problems and Current Approaches. **IEEE Bulletin of the Technical Committee on Data Engineering**, v. 23, n. 4, p. 3–13, 2000.

SCHOMM, F. Data Profiling as a process: Bridging the gap between academia and practitioners. In: 28th GI-Workshop Grundlagen von Datenbanken, **Anais**...Nörten Hardenberg, Alemanha: 2016.

SINGH, R.; SINGH, K. A Descriptive Classification of Causes of Data Quality Problems in Data Warehousing. **IJCSI International Journal of Computer Science Issues**, v. 7, n. 3, p. 41–50, 2010.