UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

VINÍCIUS DE MEDEIROS SOARES

ANÁLISE COMPARATIVA DE FERRAMENTAS DE PERFILAMENTO DE DADOS

CAMPINA GRANDE - PB

2020

VINÍCIUS DE MEDEIROS SOARES

ANÁLISE COMPARATIVA DE FERRAMENTAS DE PERFILAMENTO DE DADOS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Carlos Eduardo Santos Pires.

CAMPINA GRANDE - PB

2020

S676a Soares, Vinícius de Medeiros.
Análise comparativa de ferramentas de perfilhamento de dados. / Vinícius de Medeiros Soares. – 2020.

11 f.

Orientador: Prof. Dr. Carlos Eduardo Santos Pires.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. Perfilamento de dados - ferramentas. 2. Qualidade de dados. 3. Dados – coleta. 4. Metadados. 5. Ciência de dados. 6. Ataccama DQ Analyzer. 7. DataCleaner. 8. DataMartist. 9. Oracle Enterprise Data Quality. 10. Talend Open Studio. I. Pires, Carlos Eduardo Santos. II. Título.

CDU:004.6(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

VINÍCIUS DE MEDEIROS SOARES

ANÁLISE COMPARATIVA DE FERRAMENTAS DE PERFILAMENTO DE DADOS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Carlos Eduardo Santos Pires
Orientador – UASC/CEEI/UFCG

Professor Dr. Nazareno Andrade
Examinador – UASC/CEEI/UFCG

Professor Dr. Tiago Lima Massoni
Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 2020.

CAMPINA GRANDE - PB

Análise Comparativa de Ferramentas de Perfilamento de Dados

Vinícius de Medeiros Soares
Universidade federal de Campina Grande
Campina Grande, Paraíba, Brasil
vinicius.soares@ccc.ufcg.edu.br

Carlos Eduardo Santos Pires
Universidade federal de Campina Grand
Campina Grande, Paraíba, Brasil
cesp@dsc.ufcg.edu.br

Resumo

A qualidade de dados é o valor atribuído às propriedades da informação disponibilizada e, certamente, impacta na eficiência de uma organização. O perfilamento de dados (PD) é a atividade de coletar dados sobre dados, ou seja, metadados, e é crucial do processo de gerenciar a qualidade de dados. Para isto, um analista de dados precisa recorrer ao uso de ferramentas especializadas. Esta pesquisa apresenta uma análise comparativa de cinco ferramentas de perfilamento de dados: Ataccama DQ Analyzer, DataCleaner, DataMartist, Oracle Enterprise Data Quality e Talend Open Studio. A avaliação é feita com base em uma classificação de tarefas de perfilamento de dados que elenca 27 pontos de análise. A pesquisa busca conhecer as ferramentas disponíveis e em quais tarefas as aplicações focam. O artigo destaca a ferramenta mais completa, dentre as avaliadas, para a tarefa de perfilamento de dados.

1. Introdução

Com o crescimento exponencial do volume de dados digitais, tratar da qualidade de dados é essencial em uma organização para os processos transacionais, operacionais e estratégias de marketing e negócios. De acordo com uma pesquisa de mercado sobre qualidade de dados, realizada pela empresa Gartner [1], dados ruins custaram, em média, 15 milhões de dólares às organizações em 2017.

A qualidade de dados pode ser definida como o valor que os dados atingem em relação a um conjunto de características, tais como: precisão, completude, consistência e relevância. O ponto inicial do trabalho de qualidade de dados é o perfilamento de dados, que auxilia a mensurar essas características. Portanto, o perfilamento de dados é a atividade de identificar e coletar informações sobre os dados.

Nesse ponto, um analista de dados assume a responsabilidade de gerenciar a qualidade de dados. Considerando essa situação, é comum recorrer a ferramentas de perfilamento para reunir informações sobre os dados, e poder continuar para o tratamento da qualidade de dados. Porém, a quantidade de opções disponíveis pode dificultar a escolha. Além disso, cada ferramenta possui um conjunto de funções específicas, que pode atender apenas uma parte dos processos que se deseja realizar.

Este trabalho teve como objetivo analisar e comparar ferramentas de perfilamento de dados estruturados, ou seja, dados armazenados em um banco de dados relacional, a fim de informar o analista de dados que utiliza essas ferramentas em seu meio de trabalho. As ferramentas analisadas foram Ataccama DQ Analyzer, DataCleaner, DataMartist, Oracle Enterprise Data Quality e Talend Open Studio. A comparação foi feita baseada em funcionalidades chave que consideram colunas individualmente e relações entre duas ou mais colunas, de uma ou mais tabelas. Para isto, um banco de dados com características relevantes para cada funcionalidade foi utilizado para a avaliação das ferramentas.

Este artigo está estruturado da seguinte forma: a seção 2 discute sobre os trabalhos relacionados; a seção 3 apresenta a metodologia; a seção 4 mostra os resultados e uma discussão sobre os mesmos; e a seção 5 oferece a conclusão da pesquisa.

2. Trabalhos Relacionados

Em [8], os autores definem perfilamento de dados como a atividade de coletar dados sobre dados, ou seja, metadados. Além disso, destacam a importância do perfilamento de dados para entender e explorar conjuntos de dados desconhecidos, otimizar consultas e realizar a integração e limpeza de dados. Os autores classificam e discutem as funcionalidades importantes do perfilamento de dados. Essa classificação foi utilizada como base para este estudo.

Em [7], algumas ferramentas de qualidade de dados são avaliadas considerando critérios de desempenho, além de funcionalidades de perfilamento, integração e limpeza de dados. O artigo elegeu a ferramenta DataCleaner como a que oferece o maior número de funcionalidades referentes à qualidade de dados. Porém, a análise é feita com base em pontos gerais, sem detalhar o que foi considerado em cada ponto avaliado. Um

desses pontos é se a ferramenta realiza "análise da estrutura da tabela", mas não especifica, por exemplo, se essa análise deve ser em relação aos tipos de dados das colunas, ou aos relacionamentos entre colunas de chave estrangeira e chave primária.

O presente artigo buscou especificar a análise em pontos mais objetivos, para determinar exatamente quais tarefas cada ferramenta realiza, diferentemente de [7], que apresenta um resultado mais genérico no que diz respeito às tarefas realizadas pelas ferramentas avaliadas.

Em [6], é feita uma análise das ferramentas considerando perfilamento de dados, medição da qualidade de dados e monitoramento contínuo da qualidade de dados. Para o perfilamento de dados, a pesquisa também utilizou a classificação das tarefas de perfilamento mostrada no livro "Data Profiling" [8] como base para seus pontos de análise.

No entanto, essa classificação foi atualizada na edição mais nova do livro no qual o presente artigo utilizou-se baseou, alterando algumas das categorias presentes na versão anterior. Ainda assim, mesmo para alguns dos pontos que se mantiveram iguais em ambas as classificações, houve divergência ao comparar os resultados finais da análise disponível em [6] e os resultados deste artigo.

3. Metodologia

O estudo compreende uma análise comparativa de ferramentas de qualidade de dados que apresentam funcionalidades referentes ao perfilamento de dados (PD). A análise consiste em observar quais requisitos de PD as ferramentas suprem. Esses requisitos são chamados de "tarefas" neste artigo. Os dados foram coletados por meio de observações, ao executar as funcionalidades de PD de cada ferramenta para um banco de dados pré-determinado.

O banco de dados utilizado na análise é composto por oito tabelas. Essas tabelas possuem características importantes que ajudam a observar cada uma das tarefas de PD nas ferramentas, tais como: colunas de chave primária e de chave estrangeira, colunas de valores numéricos e de valores alfanuméricos, presença de campos "null" e campos de textos vazios e campos com valores repetidos e dependências entre colunas.

Os dados apresentam problemas de qualidade, como valores repetidos, campos com valores "null" e campos de textos vazios. Essas características são utilizadas para observar o resultado apresentado para cada tarefa em ambientes não perfeitos. Valores repetidos, por exemplo, podem interferir na identificação de possíveis chaves primárias.

As ferramentas foram escolhidas utilizando os estudos [6] e [7] como base para a seleção. Foram escolhidas algumas ferramentas presentes em um ou ambos os estudos, a fim de comparar resultados obtidos com análises já publicadas.

As ferramentas DataCleaner, DataMartist, e Talend Open Studio estão presentes nos dois estudos considerados. A ferramenta Ataccama DQ Analyzer está presente em [7], e uma versão similar (em formato web) da mesma empresa é apresentada em [6]. Por fim, a ferramenta Oracle Enterprise Data Quality está presente em [6].

Outras ferramentas presentes nos estudos foram consideradas para inclusão nesta análise, mas foram descartadas por possuírem poucas funcionalidades relativas ao PD, ou por não possuírem versões gratuitas.

3.1 Tarefas de Perfilamento de Dados

O perfilamento de dados abrange um grande conjunto de tarefas. Para a análise, foi utilizada a classificação de tarefas apresentada no livro [8]. Essa classificação separa as tarefas que consideram colunas individuais e as tarefas que identificam dependências entre colunas.

As tarefas que analisam colunas individualmente, listadas na Tabela 1, são as mais simples, em termos de implementação. Podem ser agrupadas em: cardinalidades, que resumem metadados simples, como o número de linhas; distribuição de valores em uma coluna, como histogramas, e tipos de dados, padrões e domínio, como o tamanho de valores numéricos.

As tarefas que tratam de descoberta de dependências (entre colunas) analisam múltiplas colunas. Envolvem preparação dos dados, grande espaço de busca e geração e verificação de candidatos à dependência, portanto, são mais complexas para implementar. As tarefas mais utilizadas pelas ferramentas analisadas são "combinações únicas de colunas" (unique column combinations, UCCs), "dependências funcionais" (functional dependencies, FDs) e "dependências de inclusão" (inclusion dependencies, INDs).

Cardinalidades	Número de linhas
	Valores nulos
	Distinto
	Singularidade
Distribuição de valores	Histograma
	Extremos
	Constância
	Quartis
	Primeiro dígito
Tipos de dados, padrões e domínio	Tipo básico
	Tipo do dado
	Comprimento
	Tamanho
	Decimais
	Padrões
	Classe de dado
	Domínio
Dependências	Combinações de colunas únicas (UCCs)
	UCCs relaxadas
	Dependências de inclusão (INDs)
	INDs relaxadas
	Dependências funcionais (FDs)
	FDs relaxadas
Outras dependências	Dependências multivaloradas
	Dependências de ordem

	Dependências de correspondência
	Restrições de negação

Tabela 1: Classificação das tarefas de perfilamento de dados.

As Tabelas 2 e 3 são uma representação de um possível banco de dados que suporta registros de produtos (código do produto, nome, descrição, preço e quantidade no estoque) e vendas (código da venda, código do produto, quantidade de produtos na venda e o valor total dos produtos na venda). Essas tabelas possuem características importantes para exemplificar a função de cada tarefa de perfilamento, assim como as tabelas que foram usadas para a análise das ferramentas.

COD_PROD	NOME	DESC	PRECO	ESTOQUE
1	Chocolate	Doce de cacau	4,00	10
2	Caramelo	null	0,50	15
3	Pirulito	null	2,00	null
4	Chiclete	Goma de mascar	1,00	
5	Balinha		0,10	2

Tabela 2: PRODUTOS.

COD_VENDA	COD_PROD	QUANTIDADE	VALOR
1	3	3	12,00
3	2	10	5,00
3	2	2	1,00
4	null	null	
5	5		

Tabela 3: VENDAS.

Cada tarefa é detalhada a seguir, considerando as Tabelas 2 e 3 para exemplificação.

1. **Número de linhas:** refere-se ao número de linhas de uma tabela. Para a tabela “Produtos”, o número de linhas é 5.
2. **Valores nulos:** número ou porcentagem de valores nulos. Para a coluna “Estoque” em “Produtos”, o número de valores nulos é 1.

3. **Distinto:** número de valores distintos. Para a coluna “COD_VENDA”, em “Vendas”, há 4 valores distintos: 1, 3, 4 e 5.

4. **Singularidade:** número de valores distintos dividido pelo número de linhas. Para a coluna “COD_VENDA”, em “Vendas”, há 4 valores únicos: 1, 3, 4 e 5, para um total de 5 linhas, sendo assim a singularidade é de 0,80.

5. **Histograma:** geração de histogramas de frequência para as tarefas quantitativas.

6. **Extremos:** valores mínimos e máximos em uma coluna numérica. Para a coluna “Preco”, em “Produtos”, os extremos são 4,00 e 0,10.

7. **Constância:** frequência dos valores mais frequentes dividido pelo número de linhas (porcentagem do valor mais frequente). Para a coluna “COD_VENDA”, em “Vendas”, esse valor seria de 40%, pois o valor mais frequente aparece 2 vezes, em um total de 5 linhas.

8. **Quartis:** três pontos que dividem valores (numéricos) em quatro grupos iguais. Para a coluna “COD_VENDA”, em “Vendas”, o primeiro, o segundo e o terceiro quartis são 3, 3 e 4, respectivamente.

9. **Primeiro dígito:** é a distribuição do primeiro dígito em valores numéricos, para checar a lei de Benford. Essa lei define que o dígito 1 tem maior chance de aparecer como primeiro dígito. Enquanto dígitos menores têm menos possibilidade. O 9 tem a menor probabilidade de aparecer como primeiro dígito.

10. **Tipo básico:** define o tipo básico do dado: numérico, alfanumérico, data, tempo, etc. A coluna “Preco”, em “Produtos”, tem o tipo básico definido como numérico.

11. **Tipo de dado:** define o tipo de dados específico do DBMS (varchar, integer, etc.). A coluna “Nome”, em “Produtos”, é definida como varchar.

12. **Comprimento:** comprimento mínimo, máximo, mediano e médio dos valores em uma coluna. Para a coluna “Nome”, em “Produtos”, esses valores são: 7, 9, 8 e 8, respectivamente.

13. **Tamanho:** número máximo de dígitos em valores numéricos. Para a coluna “Quantidade”, em “Vendas”, o tamanho é 2.

14. **Decimais:** número máximo de decimais em valores numéricos. Para a coluna “Valor”, em “Vendas”, o número máximo de decimais é 2.

15. **Padrões:** geração de histogramas para exibir a frequência de padrões de formato dos textos de uma coluna.

16. **Classe de dado:** tipo de dado semântico genérico, como código, indicador, texto, data/tempo, quantidade, identificador. A coluna “Nome”, em “Produtos”, é identificada como texto.

17. **Domínio:** define a semântica do domínio de uma coluna, como cartão de crédito, primeiro nome, cidade, etc.

18. **UCC (identificação de chave primária):** Descoberta do conjunto de atributos que são únicos para aquela instância. Potencialmente podem formar a chave primária. As colunas “COD_PROD” e “Nome”, em “Produtos”, formam uma combinação válida para esta tarefa.

19. **UCCs relaxadas:** apenas um subconjunto dos dados atende à dependência. Pode especificar as condições que restringem o escopo.

20. **IND (identificação de chave estrangeira):** determina que todos os valores, ou combinações de valores, de um conjunto de colunas também aparece no outro conjunto de colunas. Ajuda a descobrir o conjunto de atributos que formam uma chave estrangeira. As colunas “COD_PROD”, em “Produtos”, e “COD_PROD”, em “Vendas”, formam uma combinação válida para esta tarefa.

21. **INDs relaxadas:** apenas um subconjunto dos dados atende à dependência. Pode especificar as condições que restringem o escopo.

22. **FD:** determina que para todo par de registros com o mesmo valor em uma coluna ‘x’, a coluna ‘y’ também terá o mesmo valor. As colunas “COD_PROD” e “Nome”, em “Produtos”, formam uma combinação válida para esta tarefa.

23. **FDs relaxadas:** apenas um subconjunto dos dados atende à dependência. Pode especificar as condições que restringem o escopo.

24. **Dependências multivaloradas:** semelhante à functional dependencies, mas um conjunto de valores é aceito, ao invés de um valor específico. (Comum em instâncias resultadas de joins.)

25. **Dependências de ordem:** determina que a ordem dos valores de uma coluna está alinhada ou inversamente alinhada a outra coluna. Se ordenarmos os registros com base em uma coluna 'x', eles também estarão ordenados pela coluna 'y'. As colunas "COD_VENDA" e "Valor", em "Vendas", formam uma combinação válida para esta tarefa. Apesar de não haver uma relação de dependência entre esses valores.

26. **Dependências de correspondência:** pode ser definida como a generalização de 'functional dependencies'. Utiliza o operador similar ≈, ao invés do operador igual =. E, estende a definição de uma para duas tabelas, a fim de comparar valores entre as duas. As colunas "COD_PROD" e "Nome", em "Produtos", formam uma combinação válida para esta tarefa.

27. **Restrições de negação:** descreve combinações de valores que são proibidas, de acordo com regras de negócio. Um exemplo seria que o valor de uma venda não pode ser diferente da multiplicação da quantidade da venda pelo preço do produto.

3.2 Ferramentas Avaliadas

O estudo analisou cinco ferramentas: Ataccama DQ Analyzer v11.1.1, DataCleaner v5.1.5, DataMartist v1.7.9, Oracle Enterprise Data Quality v12.2.1 e Talend Open Studio v7.3. Uma sexta ferramenta chegou a ser considerada para a análise (Experian Pandora), porém foi descartada pois não foi possível obter a chave de teste do software até o final deste estudo.

DQ Analyzer é uma ferramenta de perfilamento de dados, desenvolvida pela empresa Ataccama. A ferramenta oferece diversas funcionalidades (como a verificação de dependências) relevantes e suporta conectividade com diversas bases de dados, porém possibilita gerar menos visualizações gráficas que as outras ferramentas.

DataCleaner é uma ferramenta de qualidade de dados de código aberto, desenvolvida pela Human Inference. Pode ser usada para perfilamento, limpeza e integração de dados. Além disso, suporta conectividade com diversas bases de dados e permite exportar os dados facilmente. Apesar de ser destinada a usuários técnicos, a interface é bem intuitiva.

DataMartist é uma ferramenta de qualidade de dados de código aberto, desenvolvida pela empresa nModal Solution Inc.. Neste estudo, foi avaliada uma versão de teste gratuita por 30 dias, mas que possui todas as funcionalidades da versão paga. Ela suporta conectividade com várias bases de dados. A sua interface é menos rebuscada e apresenta um visual antigo em relação às outras ferramentas, mas é fácil de usar.

Oracle Enterprise Data Quality (EDQ) é uma ferramenta comercial da Oracle que permite a realização de perfilamento e limpeza de dados. Essa ferramenta é a que possui a maior diversidade de visualizações gráficas. Ela permite criar uma cópia dos dados de diversas bases, mas não permite manter uma conexão entre o projeto criado na ferramenta e as bases de dados. Como o usuário não consegue trabalhar diretamente com seu banco de dados, precisa manter uma atualização frequente da cópia do banco de dados a ser utilizada pela ferramenta.

Talend Open Studio for Data Quality é uma ferramenta gratuita desenvolvida pela empresa Talend. Ela suporta conectividade com diversas bases de dados e aceita diferentes formatos de arquivo para importação, como arquivos de texto e planilhas.

Adicionalmente, todas as ferramentas possuem manual de usuário, o que ajuda no processo de utilização.

4. Resultados

Nessa seção são apresentados os resultados da análise das ferramentas, de acordo com a Tabela 1. Para cada tarefa, existem três marcações possíveis: "x", quando a tarefa é realizada; "-", quando a tarefa não é realizada, e "parcial", quando a tarefa é realizada parcialmente. Os resultados são descritos com mais detalhes a seguir.

		Data Cleaner	Talend	DQ Analyzer	DataMartist	Oracle EDQ
Cardinali-dades	**Num de linhas**	x	x	x	x	x
	Valores nulos	x	x	x	x	x
	Distinto	x	x	x	x	x
	Singularidade	-	x	x	x	x
Distribuição de valores	**Histograma**	x	x	-	x	x
	Extremos	x	x	x	x	x
	Constância	-	x	x	x	x
	Quartis	x	x	parcial	-	-
	Primeiro dígito	-	x	-	-	-
Tipos de dados, padrões e domínio	**Tipo básico**	x	-	x	x	x
	Tipo do dado	x	x	-	-	-
	Comprimento	x	x	x	-	-
	Tamanho	x	-	-	-	x
	Decimais	-	-	-	-	-
	Padrões	x	x	parcial	x	x
	Classe de dado	-	x	-	-	x
	Domínio	-	x	-	-	x
Depen-dências	**C. de colunas únicas**	-	parcial	parcial	-	-
	UCCs relaxadas	-	-	-	-	-
	D. de inclusão	parcial	parcial	parcial	-	-
	INDs relaxadas	-	parcial	-	-	-
	D. funcionais	-	parcial	parcial	-	-
	FDs relaxadas	-	parcial	-	-	-
Outras depen-dências	**D. multivaloradas**	-	-	-	-	-
	D. de ordem	-	-	-	-	-
	D. correspondência	-	x	-	-	-
	Rest. de negação	-	-	-	-	-

Tabela 4: Cobertura das tarefas pelas ferramentas.

A Tabela 4 mostra quais tarefas cada ferramenta é capaz de realizar. É possível perceber que as ferramentas cobrem a maioria das tarefas mais simples, que consideram colunas individuais (em média, cada uma cobre 66% dessas tarefas). No que se refere às tarefas mais complexas, as de dependências, verifica-se que as mesmas são menos oferecidas (cobertura média de 20%). Sendo que as ferramentas DataMartist e Oracle EDQ não realizam nenhuma dessas tarefas.

Cardinalidades. Como mencionado na Seção 3.2, essas são as tarefas mais simples. Consistem apenas em contar o número de linhas, além de verificar o valor de cada uma para encontrar valores nulos e repetidos. Apenas a tarefa "singularidade" não é coberta por todas as ferramentas, sendo que a ferramenta DataCleaner não oferece. Ainda assim, é preciso considerar que a tarefa "singularidade" é apenas uma variação da tarefa "distinto", tal que o valor de "singularidade" corresponde ao valor de "distinto" dividido pelo número de linhas.

Distribuição de valores. Também são tarefas simples, mas suas coberturas são menores do que as do grupo anterior. "Extremos" está em todas as ferramentas. "Histograma", "Constância" e "Quartis" são cobertas pela maioria das ferramentas. A ferramenta DQ Analyzer não calcula os quartis para os dados avaliados, mas calcula os decis. "Primeiro dígito" é coberta apenas pela Talend, que realiza todas as tarefas deste grupo.

Tipos de dados, padrões e domínio. Neste grupo, "padrões" é a única tarefa coberta por todas as ferramentas. Porém, a DQ Analyzer não gera o histograma de padrões propriamente dito, mas apenas disponibiliza uma tabela com a

distribuição dos padrões. “Comprimento” e “tipo básico” são cobertas pela maioria das ferramentas. Quatro das tarefas restantes são cobertas por apenas uma ou duas ferramentas, e, “decimais” não é realizada por nenhuma delas.

Dependências. A partir deste grupo, a cobertura das tarefas é ainda menor e, em todos os casos, é feita parcialmente. É possível destacar Talend e DQ analyzer neste grupo. Elas realizam todas as descobertas de dependências principais deste grupo, e a Talend ainda realiza as versões “relaxadas” das descobertas de FDs e INDs. DataCleaner também consegue realizar a descoberta de INDs, enquanto Oracle EDQ e DataMartist não realizam nenhuma das tarefas deste grupo.

Em todos os casos, a descoberta de dependências não é feita sem que o usuário restrinja as colunas que deseja avaliar. As ferramentas apenas apontam o quanto uma dependência sugerida é verdadeira ou não.

Outras dependências. Assim como no grupo anterior, estas são tarefas mais complexas, que exigem mais de algoritmos especializados para sua realização. A descoberta de “dependências de correspondência” é realizada pela Talend, com consultas SQL. Nenhuma das outras tarefas é coberta pelas ferramentas analisadas.

Em [6], foi relatado que a ferramenta Talend conseguia realizar a tarefa de descoberta de INDs. Porém, neste artigo, foi identificado que a Talend também realiza as outras descobertas de dependências. Outra diferença foi na avaliação da DataCleaner, foi encontrado que ela consegue realizar a descoberta de INDs. O artigo [6] também verificou que todas as ferramentas realizam a tarefa “histograma” parcialmente, pois são gerados apenas histogramas de largura igual, e que as ferramentas não suportam histogramas de profundidade ou altura igual. Mas não foi considerado esse tipo de restrição neste artigo.

5. Conclusão

Neste artigo foi realizada uma pesquisa com cinco ferramentas de perfilamento de dados, considerando 27 requisitos, que vão de análises em colunas individuais à descoberta de dependências entre colunas. Cada ferramenta considerada possui suas vantagens e desvantagens. Todas possuem uma interface fácil de utilizar e disponibilizam manuais de usuário que facilitam o processo de aprendizagem da ferramenta.

Foi possível perceber que as ferramentas analisadas tinham, no geral, foco nas tarefas mais simples, de colunas individuais. Mesmo para aquelas que realizavam tarefas de descoberta de dependências, as tarefas de UCCs, FDs e INDs eram as principais. As ferramentas tendem a não fornecer funcionalidades referentes às dependências, pela maior complexidade das mesmas. Outro fator é que algumas descobertas de dependências são mais específicas, e não são necessárias para muitos trabalhos.

A Talend foi a ferramenta que apresentou mais funcionalidades dentro dos pontos avaliados, 20. Em seguida, vieram DQ Analyzer, com 13, DataCleaner e Oracle EDQ, com 12, e DataMartist, com 9. Além disso, Talend, DQ Analyzer e Data Cleaner foram as únicas que realizam algum tipo de atividade de descoberta de dependências.

Caso o objetivo do usuário seja apenas identificar dados avaliando colunas individualmente, em uma base de dados, qualquer uma das ferramentas proporciona uma quantidade aceitável de tarefas. Mas, considerando a parte de visualizações gráficas, a DQ Analyzer perde, ao não fornecer geração de histogramas para suas análises.

De todas as ferramentas avaliadas, a Talend é a mais completa. Ela oferece o maior número de funcionalidades referentes ao perfilamento de dados, além de ser a única que fornece pelo menos uma tarefa de cada grupo da classificação. Portanto, ela é a ferramenta indicada neste artigo como a melhor opção para o usuário que deseja adotar um processo de perfilamento de dados.

Existem muitas outras ferramentas que fornecem funcionalidades de perfilamento. Isso poderia invalidar a conclusão final, porém, por questões de tempo, não foi possível avaliar mais ferramentas. Para trabalhos futuros, é incentivado aumentar a quantidade de ferramentas avaliadas, além de considerar possíveis atualizações da classificação base das tarefas de perfilamento de dados.

6. Referências

[1] Cost of bad data. https://insidebigdata.com/2016/05/09/the-businesscosts-of-bad-data/

[2] DataCleaner. datacleaner.org/

[3] DataMartist. datamartist.com/

[4] DQAnalyzer. https://www.ataccama.com/product/dq-analyzer/download

[5] OracleEDQ. https://www.oracle.com/middleware/technologies/enterprisedata-quality.html

[6] Elisa Rusz, Lisa Ehrlinger and Wolfram Wöß. 2019. A SURVEY OF DATA QUALITY MEASUREMENT AND MONITORING TOOLS. (2019).

[7] Cihan Varol, Venkata Sai Venkatesh Pulla and Murat Al .2016. Open Source Data Quality Tools: Revisited. (2016).

[8] Felix Naumann, Ziawasch Abedjan, Lukasz Galaband and Thorsten Papenbrock. 2019. Data Profiling. Synthesis Lectures on Data Management. Morgan Claypool.