UNIVERSIDADE FEDERAL DO PARÁ

INSTITUTO DE CIÊNCIAS EXATAS E NATURAIS

FACULDADE DE COMPUTAÇÃO

Jairo Nascimento de Sousa Filho

Modelo de Trabalho de Trabalho Acadêmico da Faculdade de Computação e Programa de Pós-Graduação em Ciência da Computação.

Belém

2019

Jairo Nascimento de Sousa Filho

Modelo de Trabalho de Trabalho Acadêmico da Faculdade de Computação e Programa de Pós-Graduação em Ciência da Computação.

Monografia apresentada na Faculdade de Computação do Instituto de Ciências Exatas e Naturais como requisito parcial para obtenção do grau de Bacharel.

Universidade Federal do Pará

Orientador: Prof. Dr. Carlos Gustavo Resque dos Santos

Coorientador: Msc. Yvan Brito

Belém

2019

Jairo Nascimento de Sousa Filho

Modelo de Trabalho de Trabalho Acadêmico da Faculdade de Computação e Programa de Pós-Graduação em Ciência da Computação.

Monografia apresentada na Faculdade de Computação do Instituto de Ciências Exatas e Naturais como requisito parcial para obtenção do grau de Bacharel.

Conceito: Excelente!_____

Belém, 1 de janeiro de 2019.

BANCA EXAMINADORA

Prof. Dr. Carlos Gustavo Resque dos Santos - Orientador
UFPA

Nome Convidado 1
SIGLA INSTITUIÇÃO

Nome Convidado 2
SIGLA INSTITUIÇÃO

Escreva sua dedicatória aqui.

Agradecimentos

Os agradecimentos principais são direcionados à Gerald Weber, Miguel Frasson, Leslie H. Watter, Bruno Parente Lima, Flávio de Vasconcellos Corrêa, Otavio Real Salvador, Renato Machnievscz e todos aqueles que contribuíram para que a produção de trabalhos acadêmicos conforme as normas ABNT com LaTeX fosse possível.

Agradecimentos especiais são direcionados ao Centro de Pesquisa em Arquitetura da Informação da Universidade de Brasília (CPAI), ao grupo de usuários *latex-br* e aos novos voluntários do grupo *abnTeX2* e que contribuíram e que ainda contribuirão para a evolução do abnTeX2.

"Escreva sua epígrafe aqui"
(Fulano de Tal, 19XX)

Resumo

Segundo a (ABNT, 2003), o resumo deve ressaltar o objetivo, o método, os resultados e as conclusões do documento. A ordem e a extensão destes itens dependem do tipo de resumo (informativo ou indicativo) e do tratamento que cada item recebe no documento original. O resumo deve ser precedido da referência do documento, com exceção do resumo inserido no próprio documento. (. . .) As palavras-chave devem figurar logo abaixo do resumo, antecedidas da expressão Palavras-chave:, separadas entre si por ponto e finalizadas também por ponto.

Palavras-chave: latex. abntex. editoração de texto.

Abstract

This is the english abstract.

Keywords: latex. abntex. text editoration.

Lista de ilustrações

Lista de quadros

Lista de tabelas

Lista de abreviaturas e siglas

ABNT	Associação Brasileira de Normas Técnicas
abnTeX	ABsurdas Normas para TeX

Lista de símbolos

Γ	Letra grega Gama
Λ	Lambda
ζ	Letra grega minúscula zeta
$\in$	Pertence

Sumário

1 Introdução

Este documento e seu código-fonte são exemplos de referência de uso da classe abntex2 e do pacote abntex2cite. O documento exemplifica a elaboração de trabalho acadêmico (tese, dissertação e outros do gênero) produzido conforme a ABNT NBR 14724:2011 *Informação e documentação - Trabalhos acadêmicos - Apresentação*.

A expressão "Modelo Canônico" é utilizada para indicar que abnTEX2 não é modelo específico de nenhuma universidade ou instituição, mas que implementa tão somente os requisitos das normas da ABNT. Uma lista completa das normas observadas pelo abnTEX2 é apresentada em (ARAUJO, 2015a).

Sinta-se convidado a participar do projeto abnTEX2! Acesse o site do projeto em <http://www.abntex.net.br/>. Também fique livre para conhecer, estudar, alterar e redistribuir o trabalho do abnTEX2, desde que os arquivos modificados tenham seus nomes alterados e que os créditos sejam dados aos autores originais, nos termos da "The LATEX Project Public License"[1].

Encorajamos que sejam realizadas customizações específicas deste exemplo para universidades e outras instituições — como capas, folha de aprovação, etc. Porém, recomendamos que ao invés de se alterar diretamente os arquivos do abnTEX2, distribua-se arquivos com as respectivas customizações. Isso permite que futuras versões do abnTEX2 não se tornem automaticamente incompatíveis com as customizações promovidas. Consulte (ARAUJO, 2015b) para mais informações.

Este documento deve ser utilizado como complemento dos manuais do abnTEX2 (ARAUJO, 2015a; ARAUJO, 2015c; ARAUJO, 2015d) e da classe memoir (WILSON; MADSEN, 2010).

Esperamos, sinceramente, que o abnTEX2 aprimore a qualidade do trabalho que você produzirá, de modo que o principal esforço seja concentrado no principal: na contribuição científica.

Equipe abnTEX2

Lauro César Araujo

[1] <http://www.latex-project.org/lppl.txt>

2 Fundamentação Teórica

2.1 Dados Sintéticos

O conceito de geração de dados sintéticos vieram por volta de 1993, por Rubin. (RUBIN, 1993) Em suma, seu objetivo era tonar anônimo os domicílios que participaram do censo daquela época. A partir desse fato, confidencialidade dos dados se tornou muito necessário, o que ajudou na popularização dos dados sintéticos. Portanto, dados sintéticos foi definido como "qualquer dado produzido o qual possa ser aplicado a uma dada situação que não foi obtido por mensuração direta.". (EDUCATION, 2016)

A necessidade de dados sintéticos podem ser de várias formas, desde a escassez de dados reais ou indisponibilidade; para teste de dados não usuais; para evitar lidar com questões de privacidade dos dados; teste de aplicação sem precisar modificar dados da aplicação de produção; criar teste de estresse da aplicação com *Big Data* antes de criar versão para produção; bem como não precisar adicionar os dados de teste manualmente. (KUMAR, 2019)

A aplicabilidade dos dados sintéticos é ilimitada, e é bastante explorada por setores cujos dados são sensíveis como a financeiro (LOPEZ-ROJAS; AXELSSON, 2012) e de saúde. (BERGEAT et al., 2014) Também são muito bem aplicáveis para exaustívos testes de segurança, os quais são necessários vários casos de teste, já que o pesquisador tem controle suficiente processamento (fórmulas matemáticas ou regras de geração) e saída do dado, como um sistema de detecção de fraudes. (BARSE; KVARNSTROM; JONSSON, 2003)

2.1.1 Trabalhos Relacionados

2.1.1.1 *Synthetic Generation of High-Dimensional Datasets*

Albuquerque et al. (ALBUQUERQUE; LOWE; MAGNOR, 2011) descreveu um *framework* capaz de gerar dados sintéticos multidimensionais. O sistema (ver figura 1) recebe um *input* que representa algumas propriedades do conjunto de dados como número de dimensões, uma distribuição de dados padrão, tipo de dado de cada dimensão entre outros. A partir disso, é criada uma função densidade de probabilidade, com o fim de gerar um cojunto de dados padrão. Essas funções podem ser ajustadas e modeladas através de objetos. Também, essas funções podem ser de 1, 2 ou 3 dimensões. Adicionalmente, pode-se haver ruídos, para simular as irregularidades encontradas em conjunto de dados reais.

O framework apresentado também possui uma interface gráfica para auxiliar o usuário a configurar o conjunto de dados, bem como gerá-lo. Contudo, não foi encontrado uma interface para pré-visualização dos futuros dados gerados. Quanto aos tipos de dados, estes são restritos aos numéricos, quer sejam inteiros ou de ponto flutuante.

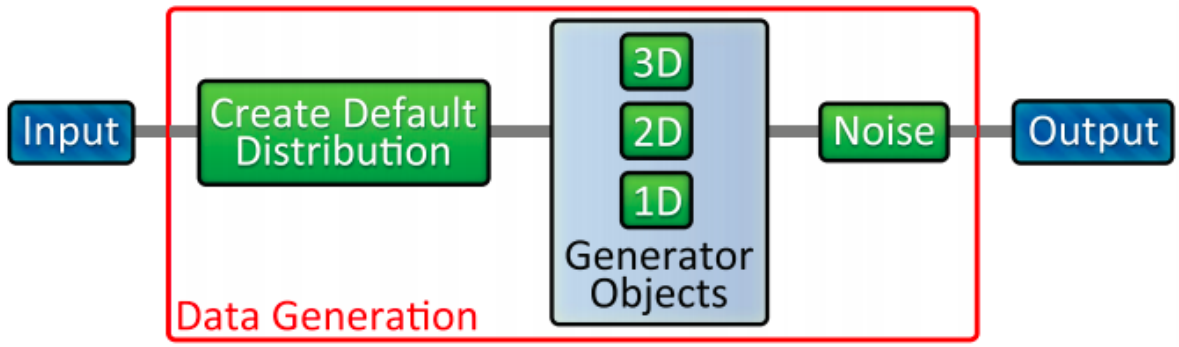

Figura 1. Visão Geral de geração de dados.

2.1.1.2 *SketchPadN-D: WYDIWYG Sculpting and Editing in High-Dimensional Space*

Wang et al. (WANG; RUCHIKACHORN; MUELLER, 2013) apresentou uma aplicação cujo principal diferencial é a capacidade de modelar, através de desenho, o comportamento das dimensões do conjunto de dados sintéticos. A priori, o usuário pode iniciar o processo de geração através do zero, de um conjunto de dados já existente, ou um conjunto de dados aleatório. A partir disso, o usuário visualiza os dados no gráfico - que pode ser as coordenadas paralelas ou o *scatterplot* - e pode modificá-lo através de cliques e arrastos. Por conseguinte, os dados podem ser gerados e isto também serve como retroalimentação do sistema. Na figura 2 é possível visualizar a visão geral do funcionamento do SketchPad.

2.1.1.3 *Synthetic Data Generator for Classification Rules Learning*

Liu (LIU et al., 2016) criou um gerador de dados sintéticos a partir de avaliação de regras de aprendizagem. O sistema funciona criando regras de aprendizagem - usando algoritmos de árvore de decisão como o ID3 - baseado em dados de entrada contruindo correlações entre os dados. Na figura 3 é possível visualizar uma árvore de decisão. Durante a leitura do conjunto de dados de entrada é feita a árvore de decisão e, concomitantemente, são geradas as regras de aprendizagem. Essas regras são utilizadas para gerar amostras de dados sintéticos.

2.1.1.4 *A prototype of synthetic data generator*

Garcia e Millán (GARCIA; MILLAN, 2011) criaram um sistema par gerar dados sintéticos pensado para desenvolvedores que buscam testar de forma eficiente e exaustiva

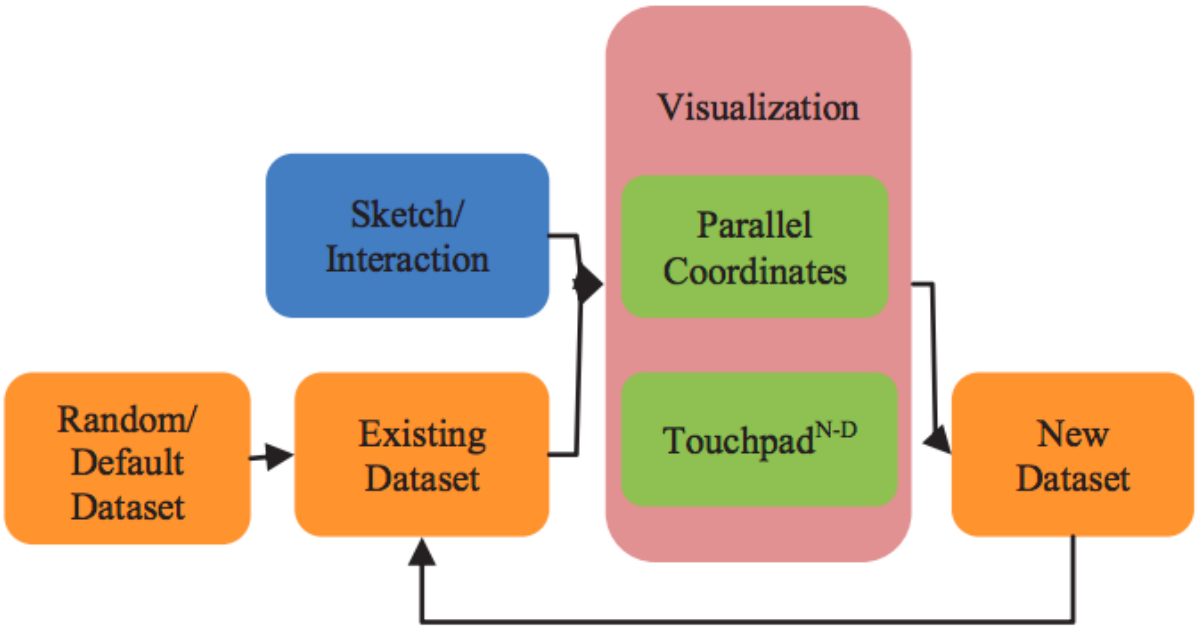

Figura 2. Visão Geral de geração de dados do Sketchpad.

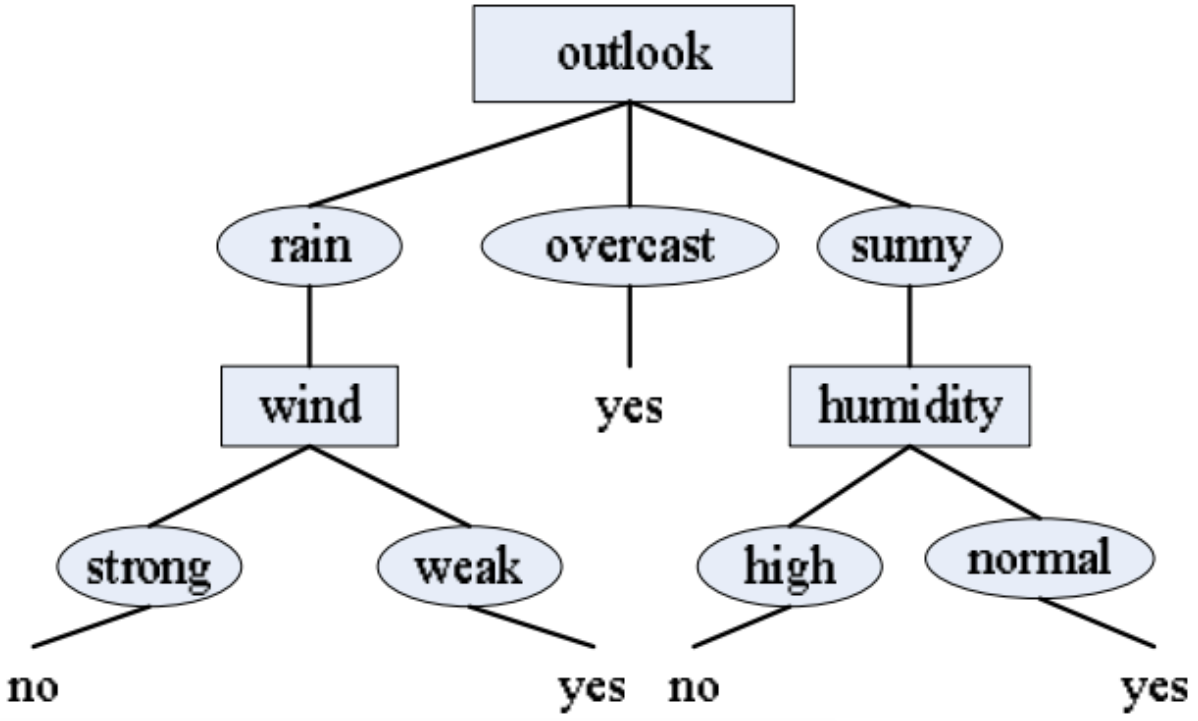

Figura 3. Exemplo de árvore de decisão para jogar tennis criado a partir de regras encontradas em um conjunto de dados.

a sua aplicação. Esses dados pode ser configurados (ver figura 4) de acordo com as preferências do usuário. As dimensões de dados seguem alguns padrões como a partir de fontes externas (Arquivos, Bibliotecas, Base de dados) Sequencial, Constante, Funcional, Intervalo ou Lista de valores.

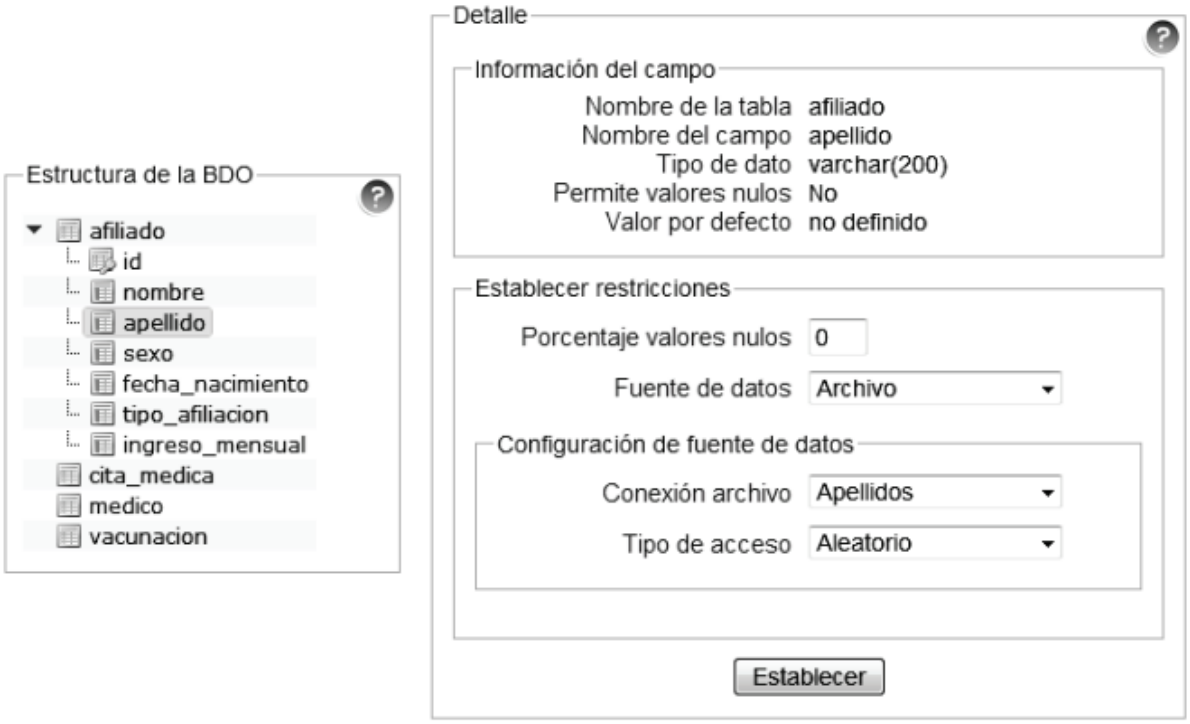

Figura 4. Exemplo da interface do usuário para configuração do gerador de dados.

2.1.1.5 *A methodology for synthetic household water consumption data generation*

Kofinas et al. (KOFINAS; SPYROPOULOU; LASPIDOU, 2018) criou uma metedologia para gerar dados sintéticos para simular consumo de água. A metodologia é avaliada através de algoritmos de validação - como a visualização dos resultados e fórmulas.

Como pode ser visto na figura 5, a geração dos dados é feita a partir de 2 fases. A fase 1 serve, basicamente, para investigar a distribuição dos dados. Esta fase, primeiramente transforma dados números em séries temporais de 30 segundos. Em alguns casos, não há registro, para isso, é criada uma tabela de incidentes e posteriormente uma probabilidade de existência de registro para que seja encontrada as classes usadas para construção do histograma de Pearson (DEAN; ILLOWSKY, 2009), por fim, são comparadas funções de distribuição com a atual com o fim de encontrar a que mais se aproxima.

Para a fase 2 cuida da geração de dados sintéticos propriamente. Basicamente, o sistema utiliza a distribuição criada na fase um para gerar os dados para 24h, respeitando as características diferenciadas para dias de semana e finais de semana.

2.1.1.6 DTM Data Generator

DTM Data Generator (SOFT., 2019) é uma plataforma de geração de dados sintéticos que existe de 1998. Esta possui suporte para geração de dados em arquivos, em banco de dados, também para *Big Data*. Possui suporte multiplataforma, através do modo *multiplatform runtime*, contudo é limitado quando comparado à versão Windows,

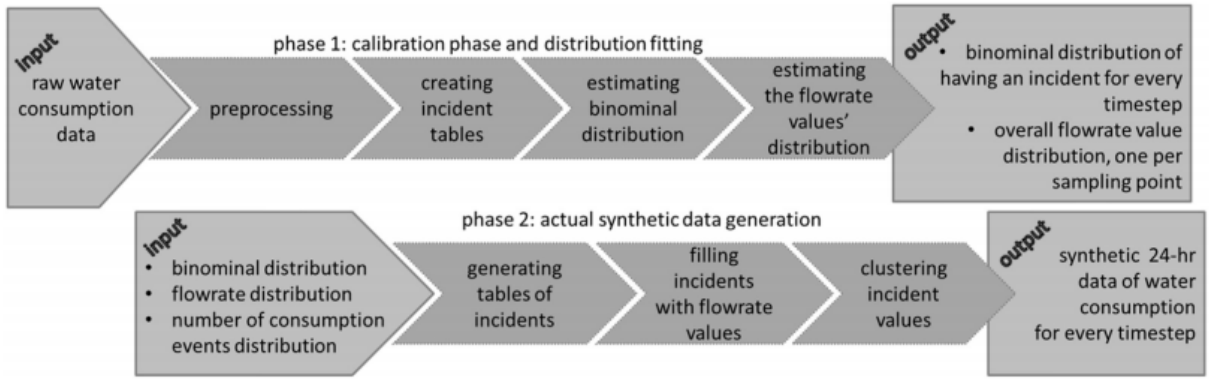

Figura 5. Fluxo de passos para geração dos dados sintéticos.

o qual suporta a versão para servidor também. É válido destacar que é um software essencialmente pago, isto é, existem versões gratuitas - demonstrações, para ser mais exato - mas limitadas. Além disso, há categorias de versões pagas, que vão desde limitações de geração (Standart - Professional) à vantagens mais técnicas (Professional - Entrerprise).

O DTM Data Generator possui uma vasta coleção de funcionalidades, as quais liberadas de acordo com as versões pagas. Adotando a versão mais cara, a lista de *features* é composta por geração de dados em JSON, XML, CSV ou geração por separador customizado. Também permite gerar dados por arquivo DSN (Database Source Name), gerar dados por linha de comando, e gerar um arquivo SQL para não seja necessário conexão com banco de dados.

É possível gerar cerca de 9.2 sextilhões de registros por *rule*, modos de atualizar dados existentes (adicionar, substituir e *Data Scrambling*), e suporte para bibliotecas de dados realistas. A plataforma disponibiliza entrada de dados através de SQL, XML, JSON, pela WEB através de HTTP ou FTP, XLSM, arquivos de texto e scripts em Python. Também é possível visualizar e testar os dados gerados, bem como gerá-los nos principais arquivos de texto (TSV, CSV, "DSV", JSON, XML) e banco de dados. (MS SQL Server, Oracle, DB2, MySQL, PostgreSQL, Informix, Sybase, SQLite e Firebird)

Há uma suíte de produtos relacionados fornecidos pela DTM soft. Além do gerador de dados, há o gerador de dados XML para teste de aplicação (DTM Test XML Generator); um gerador de planilhas Excel (DTM Data Generator for Excel) testador exaustivo - teste de estresse - de banco de dados (DTM DB Stress); Bem como editor, visualizador (DTM Data Editor), comparador e sincronizador de banco de dados (DTM Data Comparer) entre outros.

2.1.1.7 Red Gate SQL Data Generator

O SQL Data Generator (LTD, 2019) é um software que compõe uma suíte de ferramentas (chamada de SQL Toolbelt) da Red Gate. O software é exclusivo para o

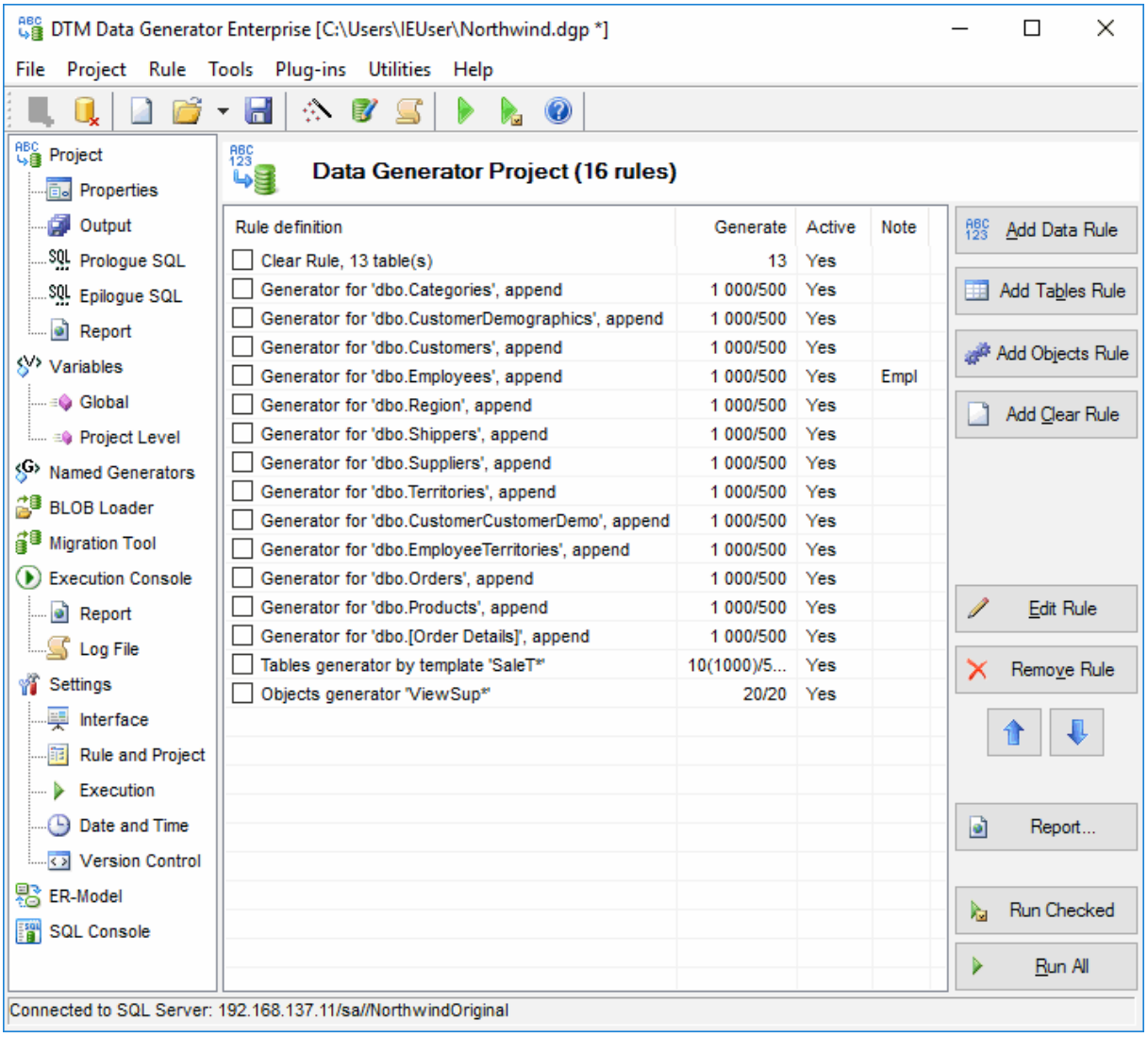

Figura 6. Usando o DTM Data Generator. Fonte: DTM Data Generator

ecossistema Windows, com suporte do Windows 7 ao 10, à versão para servidores do Windows, ao SQL Server (2008 ao 2017), .NET e Oracle. Este produto é distribuido através de licenças pagas e vitalícias, com atualizações gratuitas e, no mínimo 1 ano de suporte gratuito. Vale ressaltar que é possível testar o produto por 14 dias gratuitamente.

O SQL Toolbelt tem funcionalidades bem delimitadas e a função do Data Generator é popular um banco de dados. A população acontece ao escolher, primeiramente, uma tabela do banco. A partir disso, escolhe-se um gerador para cada coluna da tabela. Um gerador tem classificação fortemente baseada na realidade, isto é, possui geradores como palavras relacionadas à compras, pagamentos, pessoas (primeiro e último nome), dado geográficos e afins. Contudo, também disponibiliza a geração a partir de expressões regulares *Regex generator* e scripts de python. Por se tratar de banco de dados, também há checagem e tratamento de *constraints*, *Foreign keys* e *Dependencies*. O SQL Data Generator também permite lidar com arquivos XML, quer seja para geração de valores XML, como utilizar

como dados de entrada, além de mesclá-los com o *Regex generator*.

Quanto ao SQL Toolbelt oferecido pela Red Gate, ele conta com 2 modalidades, o completo com 14 programas e o *essentials* com 10. Entre os mais relevantes, pode-se citar o *SQL Data Compare*, *SQL Data Generator*, *SQL Test*, *SQL Backup Pro* e *SQL Scripts Manager*.

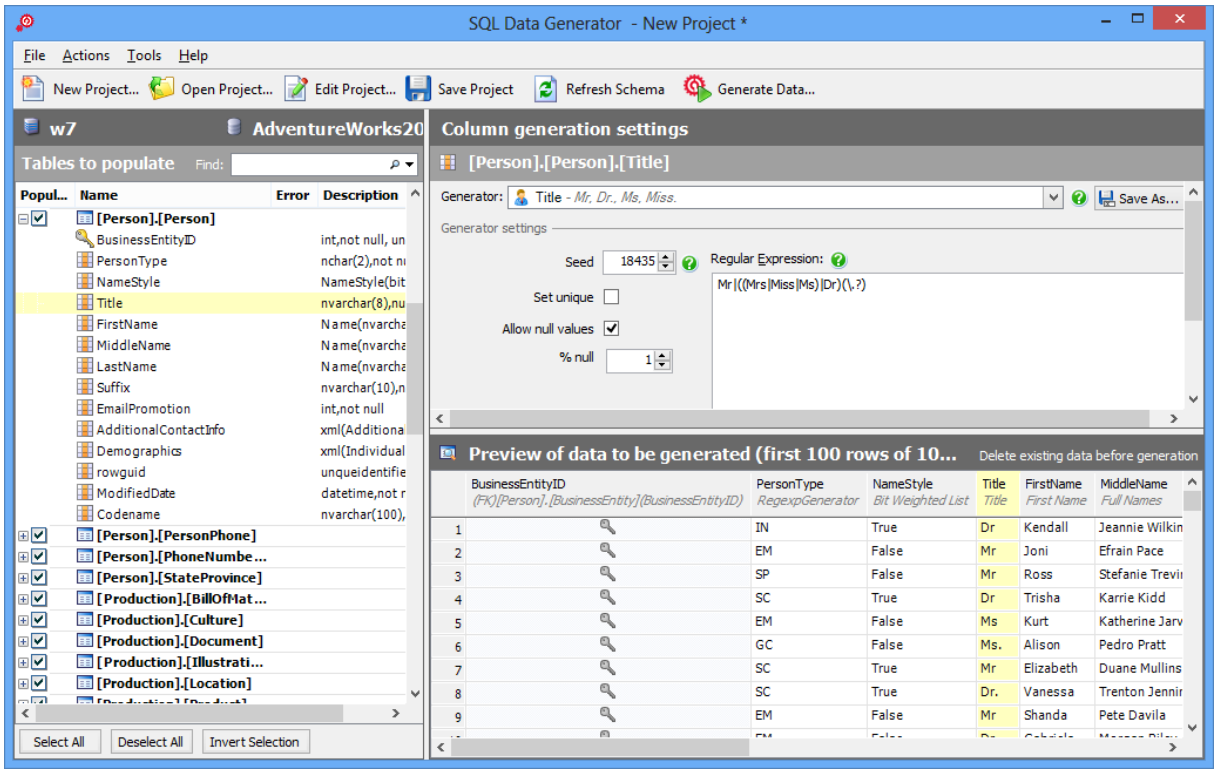

Figura 7. Usando o Redgate SQL Data Generator. Fonte: Red Gate SQL Data Generator.

2.1.1.8 Visual Studio (Premium) Data Generator

Microsoft Visual Studio (MICROSOFT, 2019) é um pacote de programas da Microsoft para desenvolvimento de *software*. Este é composto por 4 versões (*Express*, *Professional*, *Premium*, *Ultimate*), e a opção de gerar dados para teste está disponível a partir da versão *Premium*. O foco é permitir que verique o comportamento do banco de dados, sem relacioná-los com os dados da aplicação em produção.

Para gerar os dados de teste, deve-se utilizar os geradores de dados (*Data Generators*), que são correlacionados às tabelas do banco de dados. Os geradores podem ser dos mais primitivos (Binários, Inteiros, Data, *Float*), como de Imagem, Dinheiro, Expressão Regular, Categórico entre outros. Também é disponibilizado um Plano de Geração de Dados (*Data Generation Plan*), feito em XML, que contém informações do banco de dados, o tipo de dados de cada gerador e a quantidade de dados para ser gerado. Este plano serve basicamente para reutilização da lógica de teste.

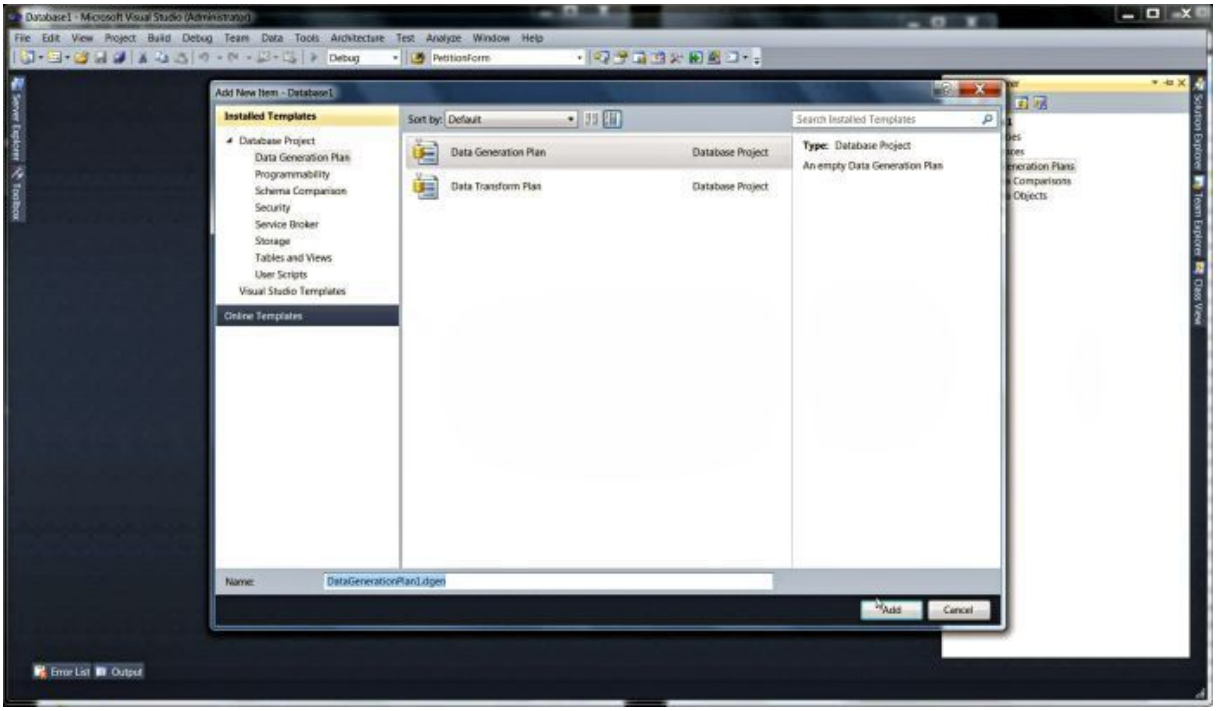

Figura 8. Usando o Microsoft Visual Studio. Fonte: anranik.

2.1.1.9 dbForge Test Data Generator

Test Data Generator (DEVART, 2018) é uma ferramenta GUI (*Graphical User Interface*) pela dbForge para gerar dados de teste para banco de dados SQL desde 1997. O software possui mais de 200 geradores predifinidos e configuráveis os quais permitem a geração de dados mais inteligentes, isto é, mais próximos da realidade, como nomes, localização, dados de saúde e afins. Quanto à compatibilidade, este é exclusivo do ecossistema Windows, com suporte à versão 7 ao 10, do Windows Server 2008 ao 2019 e ao SQL Server Azure, 2008 ao 2017. Além da GUI, também há o suporte para geração de dados a partir da linha de comando. O produto é distribuido sob licenças pagas e vitalícias, porém, com suporte ao cliente com tempo limitado e com 30 dias gratuitos para avaliação.

Para usar o dbForge Test Data Generator, é preciso fazer uma conexão com banco de dados. A partir disso, utiliza-se os *Data Generators* para determinar o comportamento dos dados para determinada coluna da tabela selecionada no Banco de dados. Os Geradores de dados podem ser do tipo emphBasics e emphAdvanced. Do primeiro tipo, são formas mais próximas dos dados primitivos, como datas, texto *lorem ipsum*, JSON, *ReGex*. Já o avançado conta com número de cartão de crédito, aniversário, número de conta bancária internacional, IPv4, *hash* de senhas. A geração de dados resume-se à população de banco de dados, não há uma forma de exportar os dados em arquivos como CSV e JSON.

Há um suíte exclusivo para SQL Server, contudo também para Oracle, MySQL, PostgreSQL entre outros. Neste suíte, há várias ferramentas que auxiliam na manutenção, mas não, necessariamente, a geração de dados, a exemplode um *previewer*. Destes, pode-se

citar um comparador de dados, criador de *querys*, um monitor - para supervisão do banco de dados - e afins.

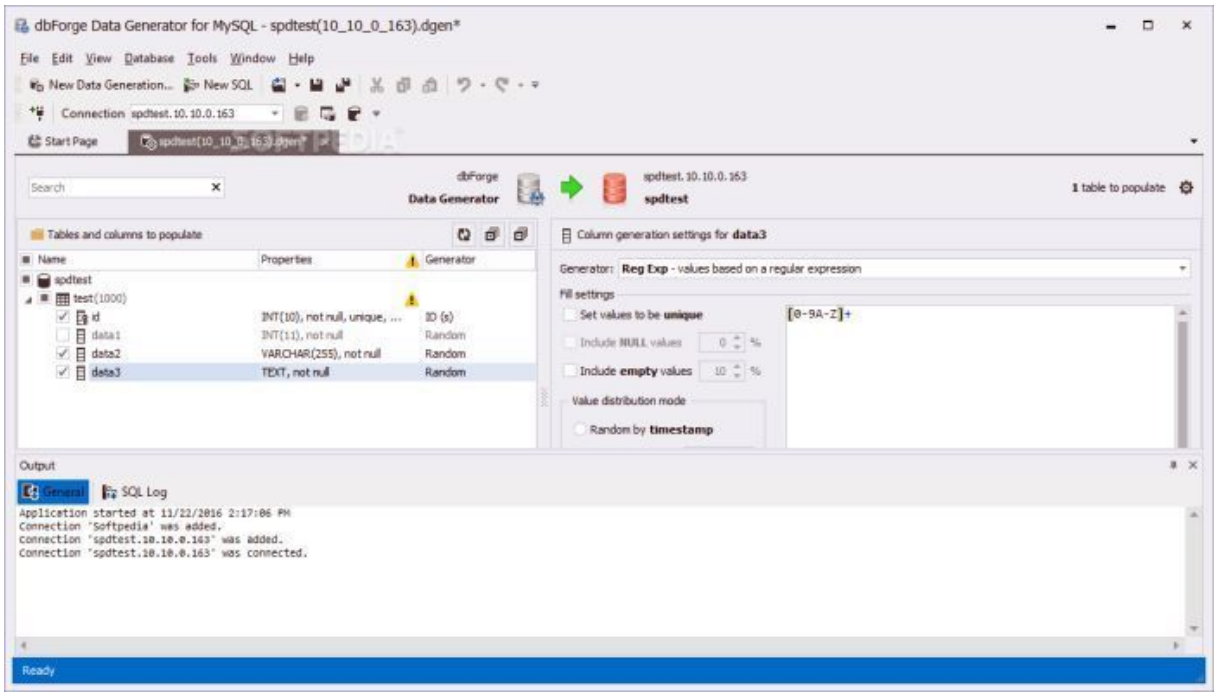

Figura 9. Usando o dbForge Test Data Generator. Fonte: anranik.

2.1.1.10 Mockaroo

Mockaroo (MOCKAROO, 2019b) é um *web site* e *framework* para desenvolver dados de teste. Há um total de 143 geradores, sendo a maioria considerados geradores realistas. Por ser um site, é possível acessá-lo por qualquer sistema operacional, dependendo apenas de conexão com a internet. O produto possui versões gratuita e pagas. - *Free*, *Silver*, *Gold*, *Enterprise* as quais variam no *host*, o qual pode ser do Mockaroo ou privado, máximo de registros por download, velocidade de download e preço.

Na tela inicial, é possível escolher o nome da coluna, o tipo de gerador, algumas opções - valores em branco e funções a partir dos dados. Ainda nesta tela, encontra-se o botão para *download* dos dados, pré-visualização dos mesmo (mas sem gráficos), algumas configurações como quantidade de linhas, formato dos dados para *download*, botão para clone ou deleção de banco de dados, e importação de dados csv/Excel ou SQL.

Outro serviço interessante do Mockaroo é Mockaroo APIs (MOCKAROO, 2019a). Este consiste em baixar dados programaticamente através de requisições REST (*Representational State Transfer*). As requisições podem ser feitas de 2 formas, a *Generate API* - gera os dados através de um banco de dados salvo e os envia pelo corpo de uma requisição - e *Mock APIs* que basicamente, simula um *back-end* como tratamento de parâmetros e

simulação de erros. É pensado para desenvolvimento ágil de aplicações *front-end*, isto é, sem perder muito tempo com o *back-end* a priori.

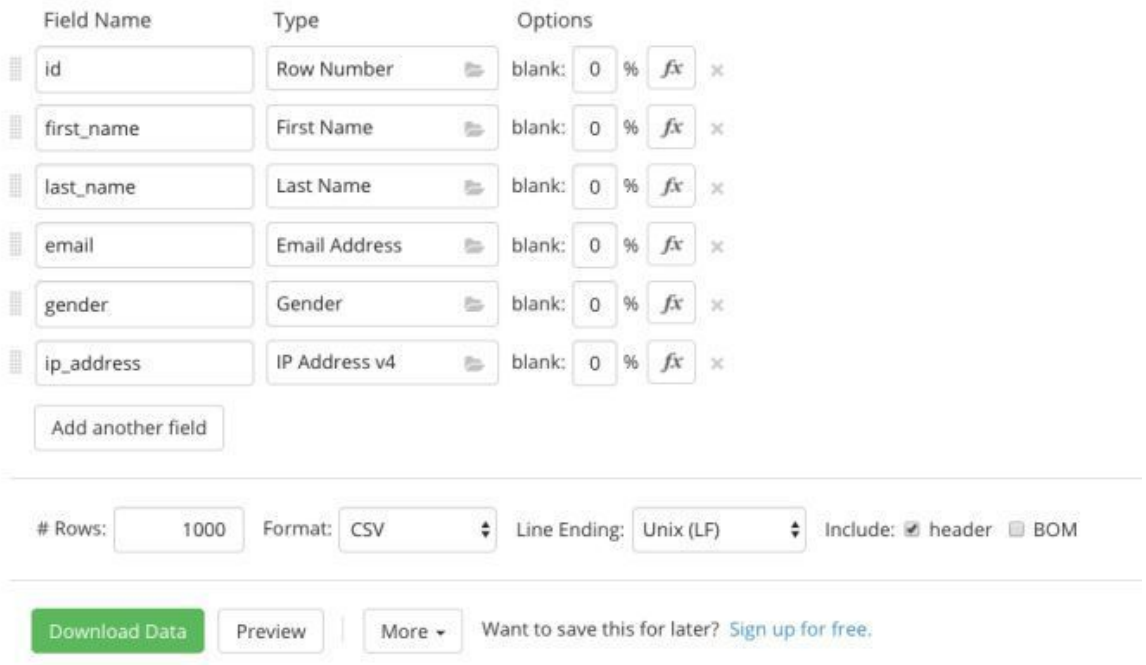

Figura 10. Usando o Mockaroo. Fonte: anranik.

2.2 Formato dos dados salvos

2.2.1 Arquivo

JSON (BRAY, 2017) (CROCKFORD, 2003) (Javascript Object Notation, ou em português Notação de Objecto Javacript),lançado em 2002, é uma formatação leve para troca de dados. O uso é facilitado tanto para seres humano quanto para máquina. O JSON é um formato de texto que é independente de linguagem, mas foi baseado no objeto provido do Javascript (ECMA-262, 1999).

Quanto aos tipos de dados suportados, o JSON (BRAY, 2017) é uma sequência de tokens. Os tipos de tokens aceitos é do tipo *object*, *array*, *string*, *number* e nomes literais como *false*, *true* e *null*.

CSV (SHAFRANOVICH, 2005) (comma-separated values, ou em português Valores Separados por Vírgula) é tipo de texto MIME (Internet Media) (FREED J. KLENSIN, 1996) que utiliza a encodificação de caracteres US-ASCII (HAUSENBLAS E. WILDE, 2014). Ao longo dos anos, seu uso foi consolidado para exportar dados entre vários softwares de tabelas (Microsoft suíte para Apple Suíte, por exemplo). A padronização do CSV demorou a ocorrer e por isso, vários outros estilos surgiram, a exemplo, o uso do CSV

com ponto-e-vírgula (;). Outros estilos foram criados a ponto de ser chamado de arquivo DSV (RAYMOND, 2003). Por conseguinte, outro estilo que teve notoriedade na troca de dados entre bancos de dados ou tabelas de dados foi o TSV (KORPELA, 2000). A idea é similar ao CSV, porém é utilizado uma tabulação em vez de vírgula.

2.2.2 Web Service

(GROUP, 2004) Um Web service é definido como um software sistema criado para suportar interoperabilidade entre máquinas através da rede computadores. Também possui uma interface descrita em um formato processável por máquinas (WSDL) e um protocolo para comunicação (SOAP). (GROUP, 2004) Essa era a arquitetura utilizada em 2004. Atualmente é predominante o uso de REST que em vez de exportar serviços como o SOAP, exporta os dados em si e não necessita do WSDL. (STACKIFY, 2017)

3 Arquitetura do projeto

3.1 Casos de uso do sistema

3.2 Ferramentas utilizadas

4 Protótipo

Este capítulo é dedicado em explicar mais sobre o prótótipo, seu fluxo de funcionamento, funcionalidades, mais detalhes sobre a interface do usuário entre outros. De modo geral, o protótipo é chamado de Blocks Data Generator e visa ser um gerador de dados sintéticos baseado em modelos de dados. Assim, o usuário pode manipular um ou mais modelos e cada modelo pode conter N dimensões, que por sua vez podem conter M geradores de dados encadeados. Os geradores de dados podem gerar dados numéricos, categóricos, temporais etc (haverá uma seção específica para geradores) e o resultado de um gerador pode servir de entrada para outro gerador através de operadores. Os operadores podem ou não aplicar uma operação matemática (soma, subtração, divisão, multiplicação) ao resultado do gerador anterior - a leitura de anterior e posterior é da esquerda para a direita, respectivamente. Junto com os operadores, também há outras propriedades que variam de acordo com o gerador (ver seção de Tipos de Geradores de Dados).

Ainda na modelagem das dimensões, é possível modificar seu nome, verificar o tipo do dado gerado pelo gerador, o ID e se está disponível para geração e visualização. Essa disponibilidade (chamado de *display*) foi feita para o caso de haver um modelo em que nem todas as dimensões sejam necessárias em determinado momento, mas também não queira perdê-las. Adicionalmente, é possível copiar e colar dimensões através de atalhos no teclado, bem como adicionar ou excluir dimensões, esta por só meio de um botão.

Também é disponibilizado um pré-visualizador de dados com apenas um gráfico - o coordenadas paralelas -, cujo volume de dados é independente do volume de dados para ser gerado, colorido e interativo. Além do *preview*, há uma integralização com um visualizador de dados mais elaborado e com mais opções de visualização. Outrossim, há um botão específico para gerar os dados em arquivos JSON, CSV, TSV ou por de requisições HTTP do tipo GET (*Web Service*). Vale ressaltar que quantidade de dados gerados, pré-visualizados, formato dos dados gerados e se contém a legenda dos dados no arquivo final é configurável em ambiente especializado, assim como para o *Web Service*.

4.1 Tipos de Geradores de Dados

4.1.1 Sequencial

Os geradores da categoria Sequencial geram valores encadeados dado um padrão. É possível gerar o próprio padrão a partir do gerador *Custom Sequence*, o qual você determina um valor Inicial (*Begin*), o valor Intervalar (*Step*), isto é, o qual vai ser incrementado ou decrementado dado uma Sentencia (*sentence*) customizada.

Contudo, já são predefinidos alguns geradores como o *constant*, o qual define um valor único de geração; o *counter*, funciona como um contador, onde-se é definido o valor Inicial e o Intervalar; o *Fixed Time Generator* gera um intevalo de tempo, onde-se define o valor inicial (*init*), Intervalar e a máscara (*mask*), isto é, como o tempo deve ser formatado; o *Sinusoidal Sequence* gera de acordo com a função senoidal, que, além do valor Inicial e Intervalar, há o 'a' de Amplitude, 'b' de frequência ângular e 'c' para representar a fase da onda.

4.1.2 Aleatório

A categoria aleatória de dados são as que contem mais geradores, pois são mais fáceis de se dissociar da realidade, pelo caráter aleatório, mas também de reaproximar, pelo caráter probabilístico. Esta categoria conta com geradores uniformes, isto é, a distribuição dos dados é equalizada; Também há um gerador de dados de tempo, parecido com o *Fixed Time Generator* com a diferença que o comportamento é definido pela fórmula de Poisson e que há mais duas configurações: unidade de tempo - a qual pode ser desde milissegundos a anos - e o lambda, advindo da fórmula. Há uma distribuição de poisson também, apenas com o lambda; É disponibilizado geradores de fórmulas clássicas com a normal (Gaussiana), Bernoulli, e Cauchy com seus devidos parâmetros. Além de números, também é possível gerar dados categóricos (*Categorical*), dadas as palavras inicialmente. Similarmente há o *Weighted Categorical* que possui valores de probabilidade para cada palavra, e já o *Categorical Quantity*, em vez de probabilidade, define quantas vezes cada palavra deve aparecer.

4.1.3 Funcional

A categoria funcional *Function* serve para gerar dados de acordo com outra dimensão chamado de *input*, isto é, facilita a correlação entre dimensões. Para dados numéricos, disponibiliza-se as função de primeiro *Linear Function* e segundo *Quadratic Function* grau, exponencial *Exponential Function*, logarítmica *Logarithm Function* e a *Piecewise Function*, cuja função é definida por subfunções, e no caso, é possível definir o gerador desejado até um determinado valor chamado de *Intervals* e depois pode-se escolher outro gerador.

Para dados categóricos, há a função categórica *Categorical Function* e a *TimeLaps Function* a qual funciona de forma semelhante ao gerador *Piecewise Function*, só que utiliza uma quantidade de tempo como limiar, chamado de *Laps* e também somente geradores de tempo como *input*.

4.1.4 Acessórios

Os geradores da categoria Acessórios (*Acessory*) foram pensados especialmente para serem concatenados com outros geradores, com o fim auxiliá-los. Entre os geradores acessórios, pode-se citar o *Missing Value*, o qual retira alguns dados do conjunto; o *Noise Generator* que adiciona dados fora do padrão, conhecido como ruído, com uma determinada probabilidade e intensidade* o *Constant Noise Generator* também adiciona ruídos, mas só que é um valor específico com determinada probabilidade de ser adicionado; o *Ranger Filter* permite ficar no conjunto de dados apenas os valores que estão entre os valores de início *Begin* e fim *End*; o *Linear Scale*; o *No Repeat* retira dados repetidos do conjunto; o *MinMax* define quais valores serão os maiores e menores de acordo com os parâmetros dados; o *Low-Pass Filter* o *Get Extra Value* pega os retornos extras dos geradores que retornam mais que um valor.

4.1.5 Geométrico

4.1.6 Baseado em dados reais

Para esta categoria, existe apenas um gerador, chamado como *Real Data Wrapper*. Basicamente, ele é criado automaticamente quando o usuário importa um conjunto de dados reais através de um CSV, por exemplo. Este gerador recebe tantos valores categóricos como numéricos e essa informação pode ser decidida automaticamente pelo gerador ou ser forçada pelo usuário. É possível gerar uma quande superior de dados do que do conjunto de dados real, para isso é feito um tratamento para dados faltantes. Esse tratamento é feito através de funções de geração, chamadas de *GenType*.

Essas funções pode ser do tipo *Standart*, que é pegar os dados do início ao fim de forma cíclica até chegar ao número desejado de registros. Também pode ser do tipo *Reverse* que ao invés do *Standart*, pega os dados do final ao início. É disponibilizado o modo aleatório (*Random*), e algumas variações. A primeira variação é o *QuartileRandom*, que divide o conjunto de dados em 3 marcos e a probabilidade de se pegar um dado daquele quartil é proporcional ao tamanho do marco. A leitura dos Marcos pode ser visualizada na figura **??** A segunda variação é o *AverageRandom* que utiliza o valor da média e da variancia - [Média - Variancia, Média + Variancia] de um conjunto de dados númerico ou utiliza os N valores categóricos mais frequentes com distribuição uniforme.

4.2 Modos de Geração de Dados

4.2.1 Padrão e *Streaming Data*

O protótipo, para uma quantidade limitada de registros e dimensões (5000 e 30 respectivamente) gera os dados agilmente, isto é, bloqueia a interface do usuário, para geração dos dados seja consistente com o modelo. O *Streaming Data* foi pensando para o *Big Data*. Então, quando foi passado o limiar da geração padrão dos dados, é criada uma cópia do modelo, para que o usuário seja livre para fazer alterações. E para fazer valer essa liberdade, o usuário tem o controle sobre o processo de geração por ver seu progresso e também poder cancelá-lo.

4.2.2 Web Service

Quanto ao *Web Service*, este foi pensado para facilitar o teste de aplicação. Cada modelo é independente, isto é, podem ser habilitados somente os modelos desejados para distribuição. É além da configuração por dentro do *software*, também é possível criar configurações temporárias para cada requisição, sem alterar as configurações o modelo. Os parâmetros disponíveis para configuração temporária pela URI são o nome do modelo, o formato dos dados e a quantidade de registro. É disponibilizado um aviso ao usuário quando um modelo está distribuindo dados via *Web Service* na aba do modelo. Um exemplo de URI para fazer requisição HTTP do tipo GET (<http://localhost:8000/?modelid=MODEL_r6w2ffk3.mva&nsample=100&format=csv>), nos quais "modelid"é o *ID* do modelo, "nsample"é a quantidade de registros desejado e "format"é o formato dos dados desejado.

4.3 Modos para Visualização de Dados

4.3.1 Preview

O pré-visualizador de dados foi criando pensando em oferecer uma visualização rápida e abrangente do modelo de dados. Para isso, foi escolhido o gráfico Coordenadas Paralelas, por conta de sua característica de visualização prática de dados multidimensionais. Também foram adicionadas algumas características extras como diferenciação por cores (mapa de calor para dados numéricos e cor única para dados categóricos); filtro de dimensão, para seja visualizado apenas o que for necessário; escolha de dimensão como referencial, isto é, a partir da dimensão escolhida, verificar como os dados se comportam nas outras dimensões. Isso pode ser ativado tanto clicando sobre o nome da dimensão, quanto através do *ComboBox* acima do *preview*; também é possível recarregá-lo e desativá-lo, para travamentos quando for trabalhar com *Big Data*, por exemplo.

4.3.2 Módulo de Visualização Externo e Integralizado

O módulo chamado VisTechLib é um conjunto de técnicas de visualização reutilizáveis. Ela pode chamada por dentro do Blocks e já pode consumir os dados do modelo atual. Dentre as visualizações disponíveis pode-se citar as Coordenadas Paralelas, *Scatterplot* Como diferencial, algumas funcionalidades são adicionadas como detalhe sob demanda, *zoom*, marcação de dados (*Highlight*), multiplas visualizações simultâneas, entre outras.

4.4 Estrutura de Interação Humano Computador

O protótipo possui uma interface gráfica para *Desktop* e segue um modelo conhecido como SPA (*Single Page Aplication*). Isso significa que há uma tela principal, e outras informações mais raras de serem consumidas aparecem através de *tabs*, *modals*, *alerts* e correlacionados.

4.4.1 Mensagens para o usuário

4.4.2 Atalhos do Teclado

4.4.3 Ajuda

5 Teste

5.1 Configuração do Teste

5.2 Resultado do Teste

6 Conclusão

Sed consequat tellus et tortor. Ut tempor laoreet quam. Nullam id wisi a libero tristique semper. Nullam nisl massa, rutrum ut, egestas semper, mollis id, leo. Nulla ac massa eu risus blandit mattis. Mauris ut nunc. In hac habitasse platea dictumst. Aliquam eget tortor. Quisque dapibus pede in erat. Nunc enim. In dui nulla, commodo at, consectetuer nec, malesuada nec, elit. Aliquam ornare tellus eu urna. Sed nec metus. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas.

Phasellus id magna. Duis malesuada interdum arcu. Integer metus. Morbi pulvinar pellentesque mi. Suspendisse sed est eu magna molestie egestas. Quisque mi lorem, pulvinar eget, egestas quis, luctus at, ante. Proin auctor vehicula purus. Fusce ac nisl aliquam ante hendrerit pellentesque. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Morbi wisi. Etiam arcu mauris, facilisis sed, eleifend non, nonummy ut, pede. Cras ut lacus tempor metus mollis placerat. Vivamus eu tortor vel metus interdum malesuada.

Sed eleifend, eros sit amet faucibus elementum, urna sapien consectetuer mauris, quis egestas leo justo non risus. Morbi non felis ac libero vulputate fringilla. Mauris libero eros, lacinia non, sodales quis, dapibus porttitor, pede. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos hymenaeos. Morbi dapibus mauris condimentum nulla. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Etiam sit amet erat. Nulla varius. Etiam tincidunt dui vitae turpis. Donec leo. Morbi vulputate convallis est. Integer aliquet. Pellentesque aliquet sodales urna.

Referências

ALBUQUERQUE, G.; LOWE, T.; MAGNOR, M. *Synthetic Generation of High-Dimensional Datasets*. Institute of Electrical and Electronics Engineers (IEEE), 2011. 2317–2324 p. Disponível em: <https://doi.org/10.1109/tvcg.2011.237>. Citado na página 29.

ARAUJO, L. C. *A classe abntex2: Modelo canônico de trabalhos acadêmicos brasileiros compatível com as normas ABNT NBR 14724:2011, ABNT NBR 6024:2012 e outras*. [S.l.], 2015. Disponível em: <http://www.abntex.net.br/>. Citado na página 27.

ARAUJO, L. C. *Como customizar o abnTeX2*. 2015. Wiki do abnTeX2. Disponível em: <https://github.com/abntex/abntex2/wiki/ComoCustomizar>. Acesso em: 27 abr 2015. Citado na página 27.

ARAUJO, L. C. *O pacote abntex2cite: Estilos bibliográficos compatíveis com a ABNT NBR 6023*. [S.l.], 2015. Disponível em: <http://www.abntex.net.br/>. Citado na página 27.

ARAUJO, L. C. *O pacote abntex2cite: tópicos específicos da ABNT NBR 10520:2002 e o estilo bibliográfico alfabético (sistema autor-data)*. [S.l.], 2015. Disponível em: <http://www.abntex.net.br/>. Citado na página 27.

ASSOCIAÇÃO BRASILEIRA DE NORMAS TÉCNICAS. *NBR 6028*: Resumo - apresentação. Rio de Janeiro, 2003. 2 p. Citado na página 11.

BARSE, E. L.; KVARNSTROM, H.; JONSSON, E. Synthesizing test data for fraud detection systems. In: IEEE. *19th Annual Computer Security Applications Conference, 2003. Proceedings.* [S.l.], 2003. p. 384–394. Citado na página 29.

BERGEAT, M. et al. A french anonymization experiment with health data. In: . [S.l.: s.n.], 2014. Citado na página 29.

BRAY, T. *The JavaScript Object Notation (JSON) Data Interchange Format*. 2017. Internet Engineering Task Force (IETF). Disponível em: <https://tools.ietf.org/html/rfc8259>. Acesso em: 31 jul 2019. Citado na página 38.

CROCKFORD, D. *ECMA-404 The JSON Data Interchange Standard.* 2003. Json.org. Disponível em: <https://json.org/json-pt.html>. Acesso em: 31 jul 2019. Citado na página 38.

DEAN, S.; ILLOWSKY, B. Descriptive statistics: Histogram. *Retrieved from the Connexions Web site: http://cnx. org/content/m16298/1.11*, 2009. Citado na página 32.

DEVART. *Data Generator for SQL Server*. 2018. Https://www.devart.com. Disponível em: <https://docs.devart.com/data-generator-for-sql-server/>. Acesso em: 21 ago 2019. Citado na página 36.

EDUCATION, M.-H. *The McGraw-Hill Dictionary of Scientific and Technical Terms, Seventh Edition (McGraw-Hill Dictionary of Scientific & Technical*

Terms). McGraw-Hill Professional, 2016. ISBN 0071608990. Disponível em: <https://www.amazon.com/McGraw-Hill-Dictionary-Scientific-Technical-Seventh/dp/0071608990?SubscriptionId=AKIAIOBINVZYXZQZ2U3A&tag=chimbori05-20&linkCode=xm2&camp=2025&creative=165953&creativeASIN=0071608990>. Citado na página 29.

FREED J. KLENSIN, J. P. N. *Multipurpose Internet Mail Extensions (MIME) Part Four: Registration Procedures*. 1996. Internet Engineering Task Force (IETF). Disponível em: <https://tools.ietf.org/html/rfc2048>. Acesso em: 31 jul 2019. Citado na página 38.

GARCIA, D.; MILLAN, M. A prototype of synthetic data generator. In: *2011 6th Colombian Computing Congress (CCC)*. IEEE, 2011. Disponível em: <https://doi.org/10.1109/colomcc.2011.5936311>. Citado na página 30.

GROUP, W. W. *Web Services Architecture*. 2004. Www.w3.org. Disponível em: <https://www.w3.org/TR/ws-arch/>. Acesso em: 02 ago 2019. Citado na página 39.

HAUSENBLAS E. WILDE, J. T. M. *ECMA-404 The JSON Data Interchange Standard.* 2014. Internet Engineering Task Force (IETF). Disponível em: <https://tools.ietf.org/html/rfc7111#page-3>. Acesso em: 31 jul 2019. Citado na página 38.

KOFINAS, D. T.; SPYROPOULOU, A.; LASPIDOU, C. S. A methodology for synthetic household water consumption data generation. *Environmental Modelling & Software*, Elsevier BV, v. 100, p. 48–66, fev. 2018. Disponível em: <https://doi.org/10.1016/j.envsoft.2017.11.021>. Citado na página 32.

KORPELA, J. *Tab Separated Values (TSV): a format for tabular data exchange*. 2000. Http://jkorpela.fi. Disponível em: <http://jkorpela.fi/TSV.html>. Acesso em: 31 jul 2019. Citado na página 39.

KUMAR, V. *15 Best Test Data Generation Tools In 2019*. 2019. Https://www.rankred.com. Disponível em: <https://www.rankred.com/test-data-generation-tools/>. Acesso em: 17 ago 2019. Citado na página 29.

LIU, R. et al. Synthetic data generator for classification rules learning. In: *2016 7th International Conference on Cloud Computing and Big Data (CCBD)*. IEEE, 2016. Disponível em: <https://doi.org/10.1109/ccbd.2016.076>. Citado na página 30.

LOPEZ-ROJAS, E. A.; AXELSSON, S. Money laundering detection using synthetic data. In: LINKÖPING UNIVERSITY ELECTRONIC PRESS. *The 27th annual workshop of the Swedish Artificial Intelligence Society (SAIS); 14-15 May 2012; Örebro; Sweden*. [S.l.], 2012. p. 33–40. Citado na página 29.

LTD, R. G. S. *SQL Data Generator*. 2019. Https://www.red-gate.com/. Disponível em: <https://www.red-gate.com/products/sql-development/sql-data-generator/>. Acesso em: 18 ago 2019. Citado na página 33.

MICROSOFT. *Generating Test Data for Databases by Using Data Generators*. 2019. Https://www.microsoft.com/. Disponível em: <https://docs.microsoft.com/en-us/previous-versions/visualstudio/visual-studio-2010/dd193262(v=vs.100)>. Acesso em: 19 ago 2019. Citado na página 35.

MOCKAROO. *Mockaroo APIs*. 2019. Https://www.mockaroo.com/. Disponível em: <https://www.mockaroo.com/api/docs>. Acesso em: 23 ago 2019. Citado na página 37.

MOCKAROO. *Mockaroo, realistic data generator*. 2019. Https://www.mockaroo.com/. Disponível em: <https://www.mockaroo.com/>. Acesso em: 23 ago 2019. Citado na página 37.

RAYMOND, E. S. *Data File Metaformats. Chapter 5. Textuality*. 2003. Http://www.catb.org. Disponível em: <http://www.catb.org/~esr/writings/taoup/html/ch05s02.html>. Acesso em: 31 jul 2019. Citado na página 39.

RUBIN, D. B. Statistical disclosure limitation. *Journal of official Statistics*, v. 9, n. 2, p. 461–468, 1993. Citado na página 29.

SHAFRANOVICH, Y. *Common Format and MIME Type for Comma-Separated Values (CSV) Files*. 2005. Internet Engineering Task Force (IETF). Disponível em: <https://tools.ietf.org/html/rfc4180#page-2>. Acesso em: 31 jul 2019. Citado na página 38.

SOFT., D. *DTM Database Tools*. 2019. Http://www.sqledit.com/. Disponível em: <http://www.sqledit.com/dg/index.html>. Acesso em: 17 ago 2019. Citado na página 32.

STACKIFY. *SOAP vs. REST: The Differences and Benefits Between the Two Widely-Used Web Service Communication Protocols*. 2017. Stackify.com. Disponível em: <https://stackify.com/soap-vs-rest/>. Acesso em: 02 ago 2019. Citado na página 39.

WANG, B.; RUCHIKACHORN, P.; MUELLER, K. SketchPadN-d: WYDIWYG sculpting and editing in high-dimensional space. *IEEE Transactions on Visualization and Computer Graphics*, Institute of Electrical and Electronics Engineers (IEEE), v. 19, n. 12, p. 2060–2069, dez. 2013. Disponível em: <https://doi.org/10.1109/tvcg.2013.190>. Citado na página 30.

WILSON, P.; MADSEN, L. *The Memoir Class for Configurable Typesetting - User Guide*. Normandy Park, WA, 2010. Disponível em: <http://mirrors.ctan.org/macros/latex/contrib/memoir/memman.pdf>. Acesso em: 19 dez. 2012. Citado na página 27.